

Chapter 13

Machine Learning Techniques in Landslide Susceptibility Mapping: A Survey and a Case Study



Taskin Kavzoglu, Ismail Colkesen, and Emrehan Kutlug Sahin

Abstract Machine learning techniques have been increasingly employed for solving many scientific and engineering problems. These data driven methods have been lately utilized with great success to produce landslide susceptibility maps. They give promising results particularly for mapping large landslide prone areas with limited geotechnical data. This chapter surveys their use in landslide susceptibility analysis and presents a case study investigating their effectiveness with regard to a conventional statistical method, namely logistic regression. It starts with the importance of spatial prediction of future landslides from past and present ones and discusses the requirement of advanced techniques for landslide susceptibility mapping. A critical literature survey is given under five main categories including core algorithms and their ensembles together with their hybrid forms. An application is presented for machine learning application using bagging, random forest, rotation forest and support vector machines with their optimal settings.

13.1 Introduction

Having a primary role in the establishment and development of residential settlements, natural disasters have been a major research topic particularly for geoscientist and engineering professionals. Rapid population growth increases pressure on natural resources and the natural environment, and raises the consequent risk associated with human activities [1]. Prediction and risk assessment of natural disasters, which can be classified into two broad groups as hydro-meteorological and geophysical disaster, are vital for planning and mitigation studies that reduce the number of human and economic losses. It is a fact that the frequency and the scale of natural disasters have increased considerably, mainly as a result of the climate change and uncontrolled human-induced changes in the landscape

T. Kavzoglu (✉) · I. Colkesen · E. K. Sahin
Department of Geomatics, Gebze Technical University, Gebze-Kocaeli, Turkey
e-mail: kavzoglu@gtu.edu.tr

(e.g. deforestation and road construction), which weakens the soil stability. As a result, the cost related to the disasters has been an increase over time. According to the Centre for Research on the Epidemiology of Disasters, at least 17% of all fatalities from natural hazards are due to the landslides [2].

Landslides, a type of natural disaster causing severe human losses and property damage, are geological phenomena related to ground movements of rock fall, and debris flow. A landslide can be described as the movement of a mass of rock, debris, or earth down a slope, under the influence of gravity [3]. Landslides are the sudden onset disasters that are usually triggered by several factors, such as intense rainfall, snow melting, earthquakes, volcanic eruptions and land use changes undermining slope instability. It should be also mentioned that climate change resulting from the global warming is one of the driving forces for landslide since it causes increased temperature, higher intensity and frequency for rain events, and lower summer precipitations. Therefore, attention should be paid to disaster mitigation and contingency planning studies by the land-use planners and policy makers, pertaining to sustainable development and reducing the risk from potential landslide events. As underlined by [4], prevention from landslides may only possible to a limited degree. However, improved understanding of the causes may help to stop or limit the human actions that increase ground instability. Although more emphasis has been given to the investigation of possible landslide locations, the increasing trend is expected to continue for three major reasons: increased and uncontrolled urbanization, continued deforestation and increased precipitation caused by climate change [5].

Determining the location of landslide prone areas is of crucial importance for hazard management studies, which is generally conducted through landslide susceptibility analysis considering various meteorological and geo-environmental parameters. It is well-known that conventional methods of ground geotechnical survey are costly both in terms of time and money, also impractical for large regions although they produce more reliable landslide hazard maps. Landslide susceptibility mapping is based on a basic modelling concept that new landslides are most likely to take place at lands having similar geological, geophysical and environmental characteristics of the previous landslide locations. The susceptibility modelling approach, compared to the conventional ones, is rapid and cost-effective with no limitation for the size of the study area. In general, landslides are complicated geophysical processes related to geology, geomorphology and hydrogeology of the ground. They occur due to the existence of various factors that show site-to-site variation. In fact, susceptibility analysis includes a variety of uncertainties that make its modelling a difficult task requiring improved knowledge about the study area characteristics and advanced techniques to model inherent relationships from a complicated structure represented by various data types. The study of factors or conditions that cause slope instability and the triggering factors or processes is of primary importance in the analysis of landslide susceptibility [6].

Prior to any conceptualizing and modelling, dealing with the landslide phenomenology requires a profound understanding of the triggering and conditioning factors that are in control of the landslide process [7]. In the estimation of landslide susceptibility, causative (i.e. conditioning or preparatory) factors are usually

considered in modelling or learning the characteristics of the problem rather than triggering factors that are usually sudden hydrological or geotechnical changes. Causative factors are mainly related to geology, geomorphology, soil structure, road network, land cover type generally derived from remotely sensed images, weathering condition and hydrogeological condition of the study area. Determination of optimum number of conditioning factors for a study area is one of the most important and difficult task in landslide susceptibility assessment. Up to now, there are not any guidelines agreed universally for the determination of case-specific conditioning factors [8]. It should be pointed out that landslide conditioning factors showed variation with respect to the study area and its geographical locations. Thus, every study area has its own particular set of factors causing landslides [9]. In other words, a specific factor can be a causative one for one region but not contributing one in another region. Therefore, selection of causative factors for a particular landslide problem is a difficult task that is usually performed by the user's experience and availability of the data. In the literature, a large number of factors have been considered as causative factors in susceptibility analysis with varying degree of usage [10, 11]. Since the number of causative factors has recently increased, some data analysis techniques are certainly required to identify the directly related factors [12]. Increasing the number of causative factors appears beneficial to improve modelling quality, but it may reduce the prediction accuracy due to the involvement of highly correlated, redundant, sometimes irrelevant factors. This issue is likely to be a major problem for future studies having large number of factors at hand. Several solutions to this problem including the use of genetic algorithm, a priori ranking of factors, using certainty factor, testing various combinations of the factors have been lately investigated by researchers [8, 13, 14].

Since the 1970s, many scientists have proposed approaches to produce susceptibility maps showing the location of possible landslide locations. Methods used in landslide susceptibility mapping are diverse and numerous [15, 16]. They can be categorized into two groups as qualitative and quantitative methods. Qualitative methods, simple methods mainly applied in 1970s, are used with expert judgments and experience considering direct field measurements. Quantitative methods, on the other hand, are based on mathematically and statistically rigorous objective methodologies. They are intended to reduce the subjectivity of landslide susceptibility evaluation by incorporating statistical and geotechnical models. These methods have become popular due to their simple expressions of the dependent (i.e. landslides) and independent (i.e. conditioning factors) variables [17]. Quantitative methods can also be categorized into several subgroups: statistical, geotechnical and heuristic methods. While the statistical methods estimate the relationship between the causative factors and past landslides using bivariate and multivariate methods, deterministic methods also called geotechnical methods are case-specific ones using some factors measured on site, ignoring climate and human-induced factors. Logistic regression, weight of evidence and analytical hierarchy process are well-known statistical methods that are usually applied as benchmark methods when a new method or approach is proposed. Heuristic methods also known as machine learning techniques employ advanced algorithms to model the inherent complex relationship

through the analysis of causative factors for landslide and non-landslide locations. They introduce nonlinearity and do not assume any distribution for the dataset. Machine learning is an automated model building approach for data analysis that learns the underlying relationships or hidden insights in the data to construct analytical models. Thus, they can be used to produce accurate and repeatable results through iterative learning, despite not being explicitly programmed to do so. Up to now, many machine learning methods have been proposed and practiced in landslide susceptibility assessment, but the most popular ones have been the support vector machines, decision trees, artificial neural networks, and ensemble methods including bagging, random forest and rotation forest. Machine learning methods with some selected studies are listed in Table 13.1, showing that kernel and tree-based methods together with their hybrid versions are now popularly practiced in the current literature. While first application of machine learning methods were applied using the core algorithms of neural networks, decision trees and support vector machines, current studies have largely focused on their improved versions using their derivatives, hybrids or ensemble forms. These algorithms have been not only used for susceptibility assessment but also used for some other applications including detection of landslide locations [18], selection of relevant conditioning factors [8] and landslide displacement prediction [19].

Table 13.1 Review of machine learning methods applied in modelling landslide susceptibility

Category	Method	Citations
Neural networks	Adaptive neuro-fuzzy inference system	[14, 20–22]
	Back-propagation neural network	[23–29]
	Extreme machine learning	[30]
	Learning vector quantization	[31]
	Multivariate adaptive regression splines	[32–34]
	Radial basis function neural networks	[26]
	Self-organizing map	[35]
Fuzzy-based	Fuzzy clustering	[36, 37]
Hybrid	k-means with particle swarm opt.	[35]
	SVM-particle swarm optimization	[38]
	Wavelet packet-statistical models	[39]
Kernel-based	Gaussian process	[40]
	Kernel logistic regression	[26, 41]
	Support vector machines (SVM)	[7, 13, 14, 25, 26, 29, 42–44]
Tree-based	Bagging	[15, 45, 46]
	Boosting	[45, 46]
	Chi-squared automatic interaction detection	[47, 48]
	Decision trees	[8, 14, 28, 32, 49, 50]
	Functional tree	[46]
	Logistic model tree	[26]
	Random forest	[43, 51–53]
Rotation forest	[54]	

Machine learning methods have become increasingly ubiquitous throughout the hazard management and mitigation studies, especially in landslide susceptibility zonation. A comparative analysis of machine learning methods for the production of landslide susceptibility maps of Macka district of Trabzon in Turkey was carried out using available eight conditioning factors. For this purpose, the most popular machine learning algorithms, namely bagging, random forest (RF), rotation forest (RotFor) and support vector machines (SVM), used in susceptibility assessment. Moreover, their performances were compared with the conventional method, i.e. logistic regression (LR). For the evaluation of predictive ability of the bagging, RF, RotFor, SVR and LR models the root mean squared error (RMSE) and mean absolute error (MAE) were estimated from the differences between the predicted susceptibility index values derived from the models and known values of the test samples to determine the precision and bias of the predictions, respectively. Three common statistical measures, namely overall accuracy, receiver operating characteristic (ROC) curve and value of area under the ROC curve (AUC), was also calculated to compare their performances. The differences in model performances were analysed using Wilcoxon’s signed rank test.

13.2 Study Area

This study was conducted on Mackaregion of Trabzon, Turkey (Fig. 13.1). The study area covers approximately 855 km² rugged terrain, situated between 39° 19' and 39° 47' longitudes, and 40° 55' and 40° 36' latitudes. Due to its physiographic

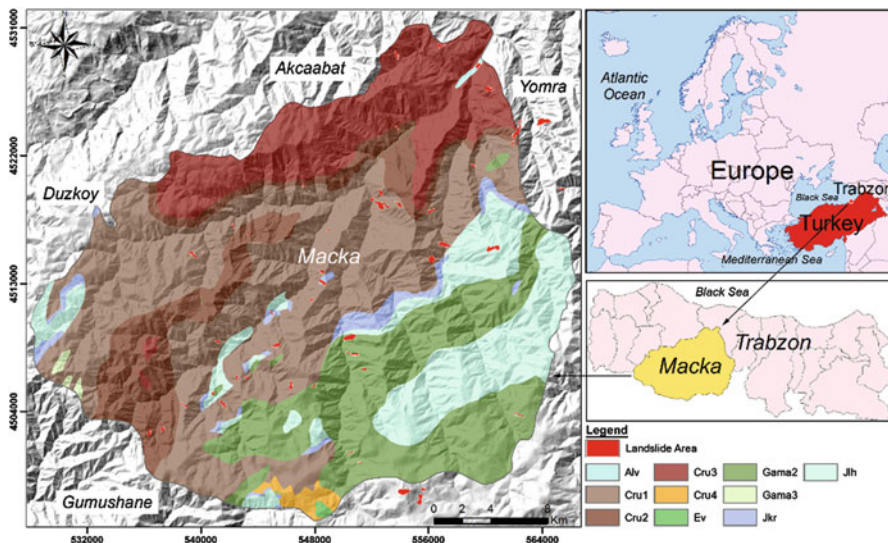


Fig. 13.1 Geological map of the study area and landslide inventory

Table 13.2 Detailed information related to factor maps

Major factors	Sub-factors	Sub-classes
Geology	Lithology	Jlh, Jcr, Cru1, Cru2, Cru3, Cru4, Ev, Gama2, Gama3, Alv
Geomorphology	Elevation (m)	80–576, 576–888, 888–1182, 1182–1.468, 1468–1742, 1742–2005, 2005–2266, 2266–2820
	Slope (°)	0–10.16, 10.16–16.41, 16.41–22.14, 22.14–27.09, 27.09–31.52, 31.52–35.95, 35.95–41.42, 41.42–66.42
	Aspect	Eight principal directions (N, NE, E, SE, etc.) and flat areas (–1°)
	Plan curvature	Concave, flat, convex
Hydrology	TWI	–0.83–1.07, 1.07–1.93, 1.93–2.78, 2.78–3.80, 3.80–5.21, 5.21–7.22, 7.22–10.07, 10.07–17.94
Land cover	Land use/cover	Urban, water, green tea, hazelnut, agriculture, deciduous, pasture, coniferous, soil/rock
	NDVI	–0.24–0.16, 0.16–0.26, 0.26–0.35, 0.35–0.42, 0.42–0.49, 0.49–0.55, 0.55–0.62, 0.62–0.76

conditions, land characteristics and climatic conditions such as above average rainfall and soil structure, the region has witnessed many landslides at different scales. The mean annual precipitation over the study area is approximately 200 mm and many of the recent landslides in the study area are triggered by heavy rainfalls. Elevations range from 80 to higher than 2800 m, and the slope angles reach 66°. The lithology map (Fig. 13.1) produced by the General Directorate of Mineral Research and Exploration covering ten types of geological formations was utilized. The study area is covered by Cru1 (basalt, andesite, lava and pyroclastic), Cru3 (basalt, andesite, lava and pyroclastic) and Gama2 (granite, granodiorite, quartz diorite and diorite) formations. Preliminary analysis shows that most of landslides occur under the Cru1, Gama2, Cru2 and Cru3 formations.

In this study, the landslide causative factors can be divided into four major groups as geomorphology, geology, hydrology and land cover. Since each causative factor map was produced from different sources, they were at different scales, so they were rescaled to 30 × 30 m pixel resolution. Moreover, digital elevation model (DEM) was produced from 1:25,000 scale topographic maps through digitization of contour lines. Elevation, slope, aspect, plan curvature and TWI factor maps were extracted from the DEM imagery. Detailed information related to the causative factors is presented in Table 13.2.

13.3 Landslide Inventory

Preparing a representative landslide inventory map is of crucial importance in all versions susceptibility mapping models. According to the basic assumption that future landslides will most likely happen in similar physiographic settings of the past

and present landslides [55]. Therefore, it is highly important to designate the location the past and current landslides correctly. In the literature, there is not any consensus on how landslide inventory maps should be prepared. Some sampling strategies have been proposed by researchers. In the present case study, a widely-used strategy based on the use of polygons representing the spatial location of a landslide was employed in the preparation of inventory map. In this study, landslide locations were taken from the map produced with “Turkish Landslide Inventory Mapping Project” by MTA Institute, Turkey. Totally 54 landslide (5144 pixels) and 23 non-landslide (1188 pixels) polygons were taken into consideration. Location analysis of the past landslides revealed that the north-east and south-west directions of the study area witnessed substantially more landslides. Spatial distribution of the landslides in the inventory map was statistically analysed, and found that the average landslide size was 85,908 m² and their acreage ranged from 6017 to 291,924 m². The minimum and maximum lengths of landslide are 117 m and 1092 m, respectively. Sixty percent of the landslides range from 120 to 500 m in length. Their width ranged between 65 and 451 m.

13.4 Methodology

Figure 13.2 shows the process adopted in this study for landslide susceptibility modelling. The process begins with the preparation of landslide inventory and causative factor maps. Subsequently, eight landslide conditioning factors (lithology, elevation, slope, aspect, plan curvature, topographical wetness index, land use/cover and NDVI) were considered for susceptibility mapping process using bagging decision tree, random forest (RF), rotation forest (RotFor) and support vector machine (SVM) methods. In addition, the logistic regression (LR) was utilized for comparison purpose.

13.4.1 Bagging Algorithm

Introduced by [56], bagging (or bootstrap aggregating) has been successfully applied to many classification and regression problems. Bagging algorithm aims to select a training sample using a bootstrap aggregating (a sample collected with replacement) from the original input training set and build a learning model. By means of bootstrap aggregating technique, it is possible to generate different training data sets and hence construct diverse learning algorithms in ensemble model. In other words, bagging decreases the overall prediction error or stabilizes individual weak learners (i.e. decision tree) by reducing variance. The resulting ensemble model created by bagging combines the predictions of multiple learning algorithms to make a final decision. Consider a given an input training data set containing n number of training examples, a sample of n training examples is generated by sampling with

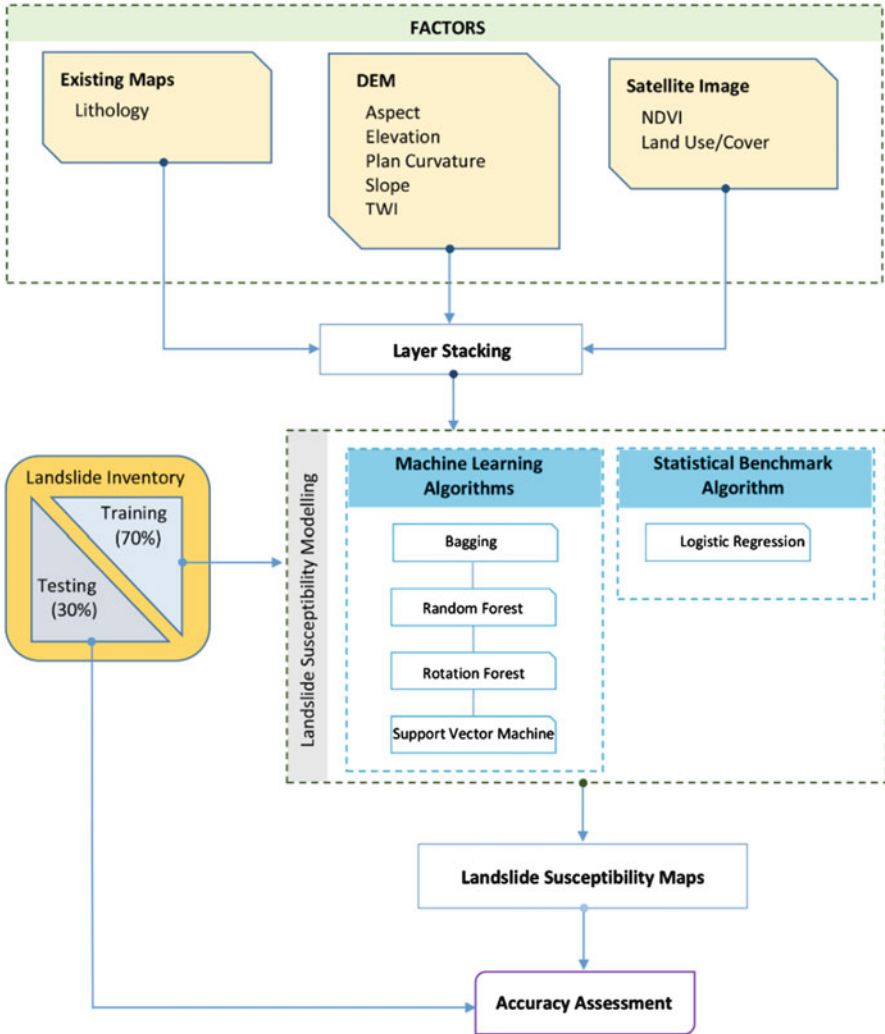


Fig. 13.2 Flowchart of the methodology used in this study

replacement. For training of each individual learning algorithm in ensemble model, this process is employed iteratively. Final prediction of a test sample is performed by combining a vote of the predictions of each individual learner (i.e. majority voting procedure). Although bagging is used to reduce the model variance, it is not successful in reducing the model bias. Thus, the trees in the ensemble model become correlated, limiting the level of error reduction. Therefore, it is advisable to determine components of the ensemble model to minimize the bias at the possible expense of variance [57].

13.4.2 Random Forest Algorithm

Random forest (RF) developed by [58] is widely-used ensemble learning algorithms successfully applied for classification, regression and feature selection purposes. RF is based on the idea that builds a set of decision trees, using randomly selected training samples through bootstrap aggregating strategy to make final a prediction. About two thirds of the selected samples known as in-bag samples are used for training of the decision tree with the remaining one third known as out-of-bag samples are used in an internal cross-validation to estimate the predictive accuracy of the constructed tree model. The output is decided by a majority voting. The underlying philosophy of RF is that the ‘strength’ of the trees is maintained while reducing the correlation between the trees in the forest. For the implementation of RF algorithm, two parameters (the number of trees and the number of variables) have to be set by the analyst. In order to construct a random forest ensemble model, two randomization processes are employed. First, training samples for each individual tree are randomly selected by applying bootstrap sampling strategy. Second, instead of selecting the best split, the tree inducer randomly samples a subset of the attributes and chooses the best one [59]. For this reason, RF can be viewed as an enhanced or generalized version of the bagging method that builds a randomized decision tree at each iteration.

13.4.3 Rotation Forest Algorithm

Rotation forest (RotFor) is an advanced ensemble learning algorithm, used to generate accurate and diverse classifiers [60]. RotFor applies a linear transformation method, principal component analysis (PCA), to the original feature subsets to project data in to a new feature space for each individual classifier in the ensemble model [61]. In the each iteration of the ensemble model construction process, the input features are randomly divided into k subsets. Then, PCA is applied to the each subset to extract the principal components of rational features. As a result, k sets of principal components are used to training of the each individual classifier of the ensemble model. To increase diversity, the bootstrap sampling strategy applied to the data created in each circle before the principal components transformations are applied.

13.4.4 Support Vector Machine Algorithm

Survey of literature reveals that support vector machine (SVM) has been one of the most popular kernel-based supervised learning algorithms, successfully applied to various fields. The main idea behind the SVM is to seek an optimal hyperplane that

provides maximum separation between linearly separable two classes. For non-separable cases, the data set is moved to a higher dimensional space using a kernel function to find the linear separation. When the SVM is used to model the complex non-linear problems including function approximation and regression estimation, the algorithm is often referred to as support vector regression (SVR) [62]. SVR algorithm assumes that each set of input features (i.e. landslide conditioning factors) has unique relation to its target variable (i.e. landslide susceptibility index). Thus, the SVR algorithm identifies the rules to estimate the target values of unknown test data samples from a set of inputs [63].

13.4.5 Logistic Regression Algorithm

Logistic regression (LR) is the most commonly applied multivariate analysis for producing landslide susceptibility maps. The LR method seeks relationship between a dependent variable (the presence or absence of landslides) and independent variables (i.e. conditioning factors). A linear fitting model is estimated describing the relationship between the dependent and independent variables. The LR process is associated with the probability of landslide phenomena to the “logit” Z (where $-\infty < Z < 0$ for higher odds of non-occurrence and $0 < Z < \infty$ for higher odds of occurrence) [64]. The LR function $Logit(p)$ equation is as follows:

$$Logit(p) = \log \left[\frac{p}{1-p} \right] \quad (13.1)$$

In this equation, p is the probability that the dependent variable ranging from 0 to 1, and $(p/(1-p))$ is the so-called odds or likelihood ratio. Using the logit transformations, the multiple linear regression equation can be written as:

$$Logit(pi) = \beta_0 + \sum_{i=1}^n \beta_i x_i \quad (13.2)$$

Where β_0 is the intercept, β_i indicates the coefficients measuring the contribution of independent variables x_i , and n show the number of independent variables.

13.5 Results

This study investigates the performance of the machine learning algorithms in comparison to logistic regression method for a study area in Turkey. For building landslide susceptibility models, training and test datasets including landslide and non-landslide samples were randomly selected from a landslide inventory map. The inventory data were randomly divided into training and testing datasets considering

70:30 sampling ratio. The same datasets were used to build regression models of the algorithms considered in this study. The Weka software (v.3.8) was utilized for bagging, RF, SVR algorithms, and SPSS (v.22) software was chosen for implementation of the LR method. Susceptibility index maps produced by the algorithms were reclassified into five common susceptibility levels by applying equal interval approach. For a successful susceptibility analysis using any parameter-based technique, it is crucially important to find and set optimal parameter values. Parameterization of machine learning algorithms considered here is explained as follows.

For the construction of bagging ensemble prediction model, decision tree algorithm was used as the base learner. The number of iteration is a critical user-defined parameter for the implementation of the bagging algorithm. A cross-validation strategy (i.e. fivefold) was applied to determine the optimum iteration number using a training dataset. The cross-validation result showed that optimum iteration number of bagging ensemble model was 30 for this study. In order to evaluate the predictive power of bagging, two standard statistical metrics namely, RMSE and mean absolute error (MAE) were also calculated from the test dataset, and the RMSE and MAE values of 0.303 and 0.186 were estimated respectively.

For the application of RF algorithm, the number of trees (n) and the number of input variables considered in each node split (k) are to be set by the user. The input data set consisted of eight landslide conditioning factors, hence the number of input variables (m) was set to be 3 (i.e. $k = \sqrt{m}$ variables at each split). On the other hand, out-of-bag (OOB) error results of RF ensemble model were used to determine the number of trees parameter. For this purpose, input data set was firstly classified using a large number of trees (i.e. 500 trees) to estimate changes in OOB error with increasing number of trees. The resulting graph showing the relations between OOB error and the number of trees (n) was given in Fig. 13.3.

It was observed that there was a sharp decline in OOB error from 0.183 to less than 0.05 as number of tree increased from 1 to 50. After that, OOB error continued to decrease slightly until the number of trees takes value 200. From this critical point to larger tree sizes, OOB error stays stable. For this reason, the number of trees (n) was set to be 200 for the current study. The predictive accuracy of RF model constructed with the user-defined parameters was tested using RMSE and MAE statistics, estimated as 0.290 and 0.193, respectively.

Two parameters of RotFor method had to be determined. In order to determine the optimum number of iterations, a cross-validation strategy was applied considering the training dataset and 70 iterations were estimated optimal. In the search for optimal number of splits (K), it was observed that changes in the parameter value have no effect on the prediction results. Consequently, it was set to 3 for the model building. RMSE and MEA were calculated as 0.329 and 0.256, respectively.

Radial basis function (RBF) kernel function was chosen in the implementation of the SVR algorithm. In SVR application, three parameters are needed to define from user-side. Meta-parameters of regularization parameter C , threshold value ϵ , and kernel width γ were determined by grid search method. As a result, ϵ value of 0.001, C value of 1.250, and γ value of 0.1 were determined as optimal. RMSE and MEA were calculated as 0.308 and 0.173, respectively.

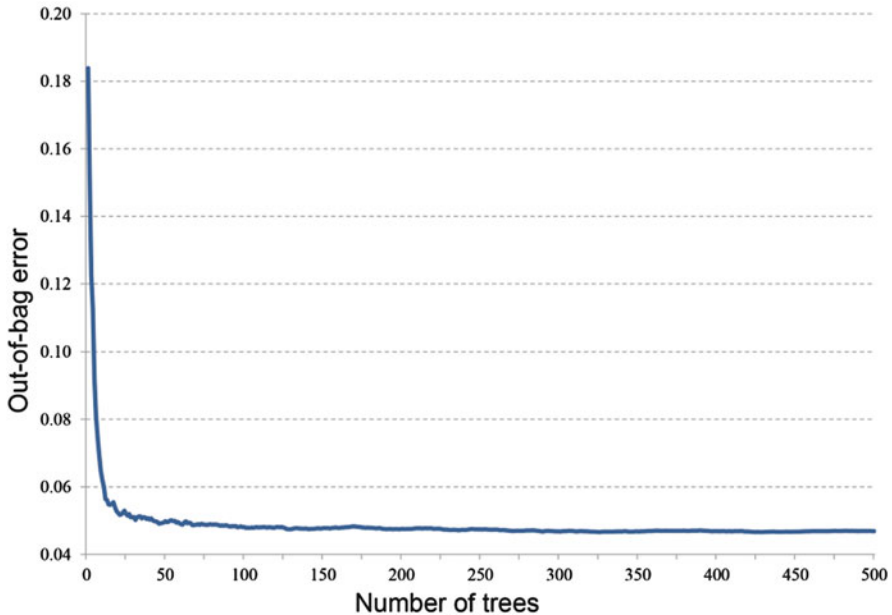


Fig. 13.3 Out-of-bag (OOB) error graph for random forest estimation

In performing the LR method, independent variables including the eight causative factors (lithology, elevation, slope, aspect, plan curvature, TWI, LULC and NDVI) and the dependent variable as landslide areas were used. The LR method was utilized to model the spatial relationship between the landslides and causative factors. The standard errors (SE), regression coefficients (β), Wald test statistic and associated p-values were estimated. Among all factors, slope was the most contributing factor since it had the highest coefficient value. Other effective factors were lithology and elevation. On the other hand, the coefficients estimated for NDVI and TWI were close to 0, indicating the minor impacts or weak relation to landslide occurrence. All causative factors had p-values lower than 0.1, indicating statistical significance between factors and the susceptibility to landslide at the 90% confidence level.

All methods with the above-mentioned parameter settings were applied to the multi-layer dataset to yield susceptibility maps. Histogram values were categorized into five susceptibility classes using quantile approach to obtain susceptibility maps (Fig. 13.4). The predictive powers of the landslide susceptibility models were measured using overall accuracies calculated using the test dataset. It should be noted that two susceptibility classes as very high and high level of the susceptibility map were considered as potential landslide occurrence and the rest (i.e., moderate, low and very low) were considered as non-landslide in accuracy assessment process. Overall accuracies for bagging, RF, RotFor, SVR and LR methods were estimated as 83.08%, 87.23%, 85.31%, 84.85% and 78.46%, respectively. Results revealed that the highest accuracy was estimated with the RF ensemble model (87.23%), whereas

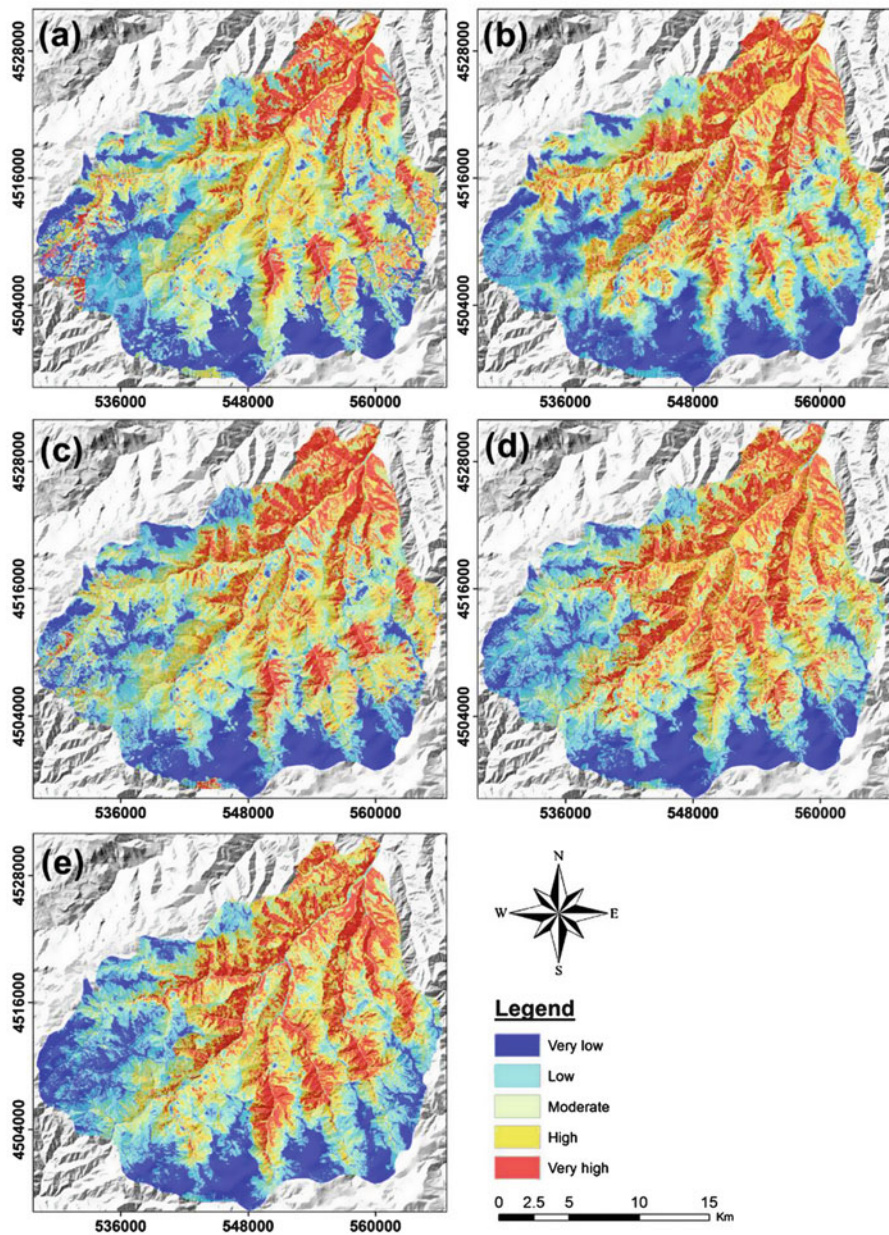


Fig. 13.4 Landslide susceptibility maps produced by (a) bagging, (b) random forest, (c) rotation forest, (d) support vector machine and (e) logistic regression

the lowest accuracy was calculated with traditional LR model (78.46%). RotFor and SVR algorithms showed similar performances and calculated accuracies of their predictive models were 85.31% and 84.85%, respectively. Results clearly showed that the machine learning algorithms produced more accurate results in comparison with the traditional LR method, and they improved the prediction accuracy up to 9% in terms of overall accuracy. This finding supported the results of some previous studies [65–68] whilst there are limited studies showed the effectiveness of LR method compared to machine learning algorithms [28, 41, 69].

In order to further evaluate the performances of the obtained susceptibility models, ROC curve and AUC value were applied. ROC curves are constructed by correctly classified pixels (sensitivity) and incorrectly identified pixels (1-specificity).

The estimated AUC values for bagging, RF, RotFor, SVR and LR models were 0.931, 0.963, 0.959, 0.955, and 0.868, respectively (Fig. 13.5). AUC values confirm the overall accuracy results for the method performances. From plotted the ROC curves, it was revealed that all susceptibility models produced acceptable results. When the estimated AUC values were analysed, the machine learning algorithms were much more effective for landslide susceptibility assessment, and the RF method produced the highest AUC value of 0.963, followed by RotFor (AUC = 0.959), SVR (AUC = 0.955) and bagging (AUC = 0.931). It is clear from the obtained results that the machine learning algorithms outperformed the standard LR model with higher AUC values.

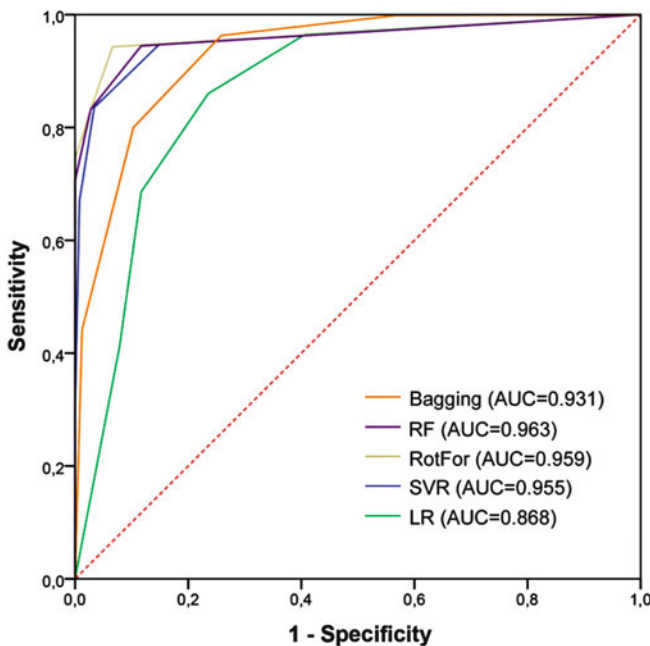


Fig. 13.5 Results of ROC curves and AUC values for the different methods

Table 13.3 Wilcoxon's signed-rank test statistic for landslide susceptibility models. Note that estimated values greater than the table value ($Z_{1, 0.05} = 1.96$) indicates statistical significance

	Bagging	RF	RotFor	SVR	LR
Bagging	–	8.464	5.104	2.954	9.913
RF		–	6.755	7.073	18.961
RotFor			–	2.860	11.824
SVR				–	10.741

In addition to the assessment of five susceptibility model performances using overall accuracy and AUC values, Wilcoxon's test was also employed to validate the significance of differences statistically. If the estimated statistic value is larger than critical table value ($Z_{1, 0.05} = 1.96$), the null hypothesis can be rejected with 95% confidence level. Calculated statistical test results were given as a matrix in Table 13.3. From the table, all estimated statistic values obtained by pairwise comparisons were greater than the critical table value. Therefore, it can be concluded that differences in the model performances was found to be statistically significant. In other words, performance of the RF method was statistically better than the other machine learning methods. This could be related to the RF characteristics of being non-parametric, capable of using continuous and categorical data, easy to parameterize, robust against overfitting, and not being sensitive to noise in the dataset.

13.6 Conclusions

Producing accurate and reliable landslide susceptibility maps representing the areas prone to landslides has been one of the most concentrated topics in hazard management. These maps are used as a base map in many global and regional studies on hazard management and planning. Therefore, reliability and accuracy of the landslide susceptibility maps has crucial importance in order to achieving the desired goals of the management plans. Landslide susceptibility mapping consists of complex and multi-stage steps including preparation of landslide inventories, selection of landslide causative factors, determining an appropriate prediction algorithm and accuracy assessment. Up to now, many methods or frameworks have been proposed so as to increase prediction accuracy of landslide models, and their performances have been investigated for susceptibility of landslides. In this study, four well-known machine learning algorithms, namely ensemble based bagging, random forest, rotation forest and kernel-based support vector machines, were employed in the process of susceptibility assessment. Their performances were compared to that of the LR method, which can be regarded as a conventional statistical approach.

Result of this study revealed some important findings. Firstly, when the estimated overall accuracies were analysed, the machine learning methods clearly outperformed conventional logistic regression methods (up to 9% improvement). ROC curves and related AUC statistics also supported the above finding. In addition,

performance differences were concluded as statistically significant based on Wilcoxon's test. Secondly, the results produced in this study also revealed that RF algorithm produced the best performance (overall accuracy of 87.23% and AUC value of 0.963) among the machine learning algorithms and the difference in their performances was statistically significant at 95% confidence interval. Thirdly, among the ensemble models based on bootstrap aggregating strategy considering accuracy results, the RF and RotFor algorithms was superior to bagging algorithm. This is an expected result since both the RF and RotFor methods are improved versions of the bagging. However, the processing time required to form an ensemble model is longer, particularly for RotFor method including PCA processing behind. From the findings of the study and the literature review, it was induced that the machine learning methods are invaluable tools for landslide susceptibility assessment, and they should be favoured over statistical methods particularly for cases including a large number causative factors and limited landslide locations at hand. On the other hand, literature survey reveals that instead of single usage of the methods, hybrid and ensembles of machine learning methods will play an important role in future studies for improvement in predictive power of landslide susceptibility evaluation.

References

1. Kreimer A, Munasinghe M (1991) Managing environmental degradation and natural disasters: an overview. In: Kreimer A, Munasinghe M (eds) *Managing natural disasters and the environment*. World Bank, Washington, DC, pp 3–6
2. Lacasse S, Nadim F (2009) Landslide risk assessment and mitigation strategy. In: Sassa K, Canuti P (eds) *Landslides – disaster risk reduction*. Springer, Berlin
3. Varnes DJ (1978) Slope movement types and processes. In: Schuster RL, Krizek RJ (eds) *Landslides: analysis and control*. Transportation Research Board, National Academy of Sciences, Washington, DC
4. Marsh SH (2000) Landslide hazard mapping: summary report. British Geological Survey Technical Report WC/00/11, 25p
5. Schuster RL (1996) Socioeconomic significance of landslides. In: Turner AK, Schuster RL (eds) *Landslides: investigation and mitigation*, Special Report 247. Transportation Research Board, National Research Council, National Academy Press, Washington, DC, pp 12–35
6. Dai FC, Lee CF, Ngai YY (2002) Landslide risk assessment and management: an overview. *Eng Geol* 64:65–87
7. Marjanović M, Kovacevic M, Bajat B et al (2011) Landslide assessment of the starč'a basin (Croatia) using machine learning algorithms. *Acta Geotech Slov* 2011:45–55
8. Kavzoglu T, Sahin EK, Colkesen I (2015) An assessment of multivariate and bivariate approaches in landslide susceptibility mapping: a case study of Duzkoy district. *Nat Hazards* 76:471–496
9. van Westen CJ, Rengers N, Soeters R (2003) Use of geomorphological information in indirect landslide susceptibility assessment. *Nat Hazards* 30:399–419
10. Hasekiogullari GD, Ercanoglu M (2012) A new approach to use AHP in landslide susceptibility mapping: a case study at Yenice (Karabuk, NW Turkey). *Nat Hazards* 63:1157–1179
11. Suzen ML, Kaya BS (2012) Evaluation of environmental parameters in logistic regression models for landslide susceptibility mapping. *Int J Digit Earth* 5:338–355

12. Smola A, Vishwanathan SVN (2008) Introduction to machine learning. Cambridge University Press, Cambridge, UK
13. Costanzo D, Rotigliano E, Irigaray C et al (2012) Factors selection in landslide susceptibility modelling on large scale following the GIS matrix method: application to the river Beiro basin (Spain). *Nat Hazards Earth Syst* 12:327–340
14. Pradhan B (2013) A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS. *Comput Geosci* 51:350–365
15. Brenning A (2005) Spatial prediction models for landslide hazards: review, comparison and evaluation. *Nat Hazards Earth Syst* 5:853–862
16. Marrapu BM, Jakka RS (2014) Landslide hazard zonation methods: a critical review. *Int J Civ Eng Res* 5:215–220
17. Bai SB, Wang J, Thiebes B et al (2014) Susceptibility assessments of the Wenchuan earthquake-triggered landslides in Longnan using logistic regression. *Environ Earth Sci* 71:731–743
18. Wan S, Yen JY, Lin CY et al (2015) Construction of knowledge-based spatial decision support system for landslide mapping using fuzzy clustering and KPSO analysis. *Arab J Geosci* 8:1041–1055
19. Lian C, Zeng ZG, Yao W et al (2013) Displacement prediction model of landslide based on a modified ensemble empirical mode decomposition and extreme learning machine. *Nat Hazards* 66:759–771
20. Pradhan B, Sezer EA, Gokceoglu C et al (2010) Landslide susceptibility mapping by neuro-fuzzy approach in a landslide-prone area (Cameron Highlands, Malaysia). *IEEE T Geosci Remote* 48:4164–4177
21. Choi J, Lee YK, Lee M et al (2011) Landslide susceptibility mapping by using an adaptive neuro-fuzzy inference system (ANFIS). In: 2011 I.E. international geoscience and remote sensing symposium (IGARSS):1989–1992
22. Aghdam IN, Varzandeh MHM, Pradhan B (2016) Landslide susceptibility mapping using an ensemble statistical index (Wi) and adaptive neuro-fuzzy inference system (ANFIS) model at Alborz Mountains (Iran). *Environ Earth Sci* 75:553
23. Lee S, Ryu JH, Lee MJ et al (2003) Use of an artificial neural network for analysis of the susceptibility to landslides at Boun, Korea. *Environ Geol* 44:820–833
24. Gómez H, Kavzoglu T (2005) Assessment of shallow landslide susceptibility using artificial neural networks in Jabonosa River Basin, Venezuela. *Eng Geol* 78:11–27
25. Yilmaz I (2010) Comparison of landslide susceptibility mapping methodologies for Koyulhisar, Turkey: conditional probability, logistic regression, artificial neural networks, and support vector machine. *Environ Earth Sci* 61:821–836
26. Bui DT, Tuan TA, Klempe H et al (2016) Spatial prediction models for shallow landslide hazards: a comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree. *Landslides* 13:361–378
27. Gorsevski PV, Brown MK, Panter K et al (2016) Landslide detection and susceptibility mapping using LiDAR and an artificial neural network approach: a case study in the Cuyahoga Valley National Park, Ohio. *Landslides* 13:467–484
28. Wang LJ, Guo M, Sawada K et al (2016) A comparative study of landslide susceptibility maps using logistic regression, frequency ratio, decision tree, weights of evidence and artificial neural network. *Geosci J* 20:117–136
29. Xu C, Shen LL, Wang GL (2016) Soft computing in assessment of earthquake-triggered landslide susceptibility. *Environ Earth Sci* 75:767
30. Vasu NN, Lee SR (2016) A hybrid feature selection algorithm integrating an extreme learning machine for landslide susceptibility modeling of Mt. Woomyeon, South Korea. *Geomorphology* 263:50–70
31. Pavel M, Nelson JD, Fannin RJ (2011) An analysis of landslide susceptibility zonation using a subjective geomorphic mapping and existing landslides. *Comput Geosci* 37:554–566

32. Felicísimo A, Cuartero A, Remondo J et al (2013) Mapping landslide susceptibility with logistic regression, multiple adaptive regression splines, classification and regression trees, and maximum entropy methods: a comparative study. *Landslides* 10:175–189
33. Wang LJ, Guo M, Sawada K et al (2015) Landslide susceptibility mapping in Mizunami City, Japan: a comparison between logistic regression, bivariate statistical analysis and multivariate adaptive regression spline models. *Catena* 135:271–282
34. Conoscenti C, Rotigliano E, Cama M et al (2016) Exploring the effect of absence selection on landslide susceptibility models: a case study in Sicily, Italy. *Geomorphology* 261:222–235
35. Wan SA (2013) Entropy-based particle swarm optimization with clustering analysis on landslide susceptibility mapping. *Environ Earth Sci* 68:1349–1366
36. Wang LJ, Sawada K, Moriguchi S (2013) Landslide susceptibility analysis with logistic regression model based on FCM sampling strategy. *Comput Geosci* 57:81–92
37. Alimohammadi Y, Najafi A, Gokceoglu C (2014) Estimation of rainfall-induced landslides using ANN and fuzzy clustering methods: a case study in Saen Slope, Azerbaijan province, Iran. *Catena* 120:149–162
38. Yu XY, Wang Y, Niu RQ et al (2016) A combination of geographically weighted regression, particle swarm optimization and support vector machine for landslide susceptibility mapping: a case study at Wanzhou in the three gorges area, China. *Int J Environ Res Publ Health* 13:487. <https://doi.org/10.3390/ijerph13050487>
39. Moosavi V, Niazi Y (2016) Development of hybrid wavelet packet-statistical models (WP-SM) for landslide susceptibility mapping. *Landslides* 13:97–114
40. Colkesen I, Sahin EK, Kavzoglu T (2016) Susceptibility mapping of shallow landslides using kernel-based Gaussian process, support vector machines and logistic regression. *J Afr Earth Sci* 118:53–64
41. Hong HY, Pradhan B, Xu C et al (2015) Spatial prediction of landslide hazard at the Yihuang area (China) using two-class kernel logistic regression, alternating decision tree and support vector machines. *Catena* 133:266–281
42. Yao X, Tham LG, Dai FC (2008) Landslide susceptibility mapping based on support vector machine: a case study on natural slopes of Hong Kong, China. *Geomorphology* 101:572–582
43. Goetz JN, Brenning A, Petschko H et al (2015) Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling. *Comput Geosci* 81:1–11
44. Hong HY, Pradhan B, Jebur MN et al (2016) Spatial prediction of landslide hazard at the Luxi area (China) using support vector machines. *Environ Earth Sci* 75:40
45. Bui DT, Ho CT, Revhaug I et al (2014) Landslide susceptibility mapping along the national road 32 of Vietnam using GIS-based J48 decision tree classifier and its ensembles. In: Buchroithner M, Prechtel N, Burghardt D (eds) *Cartography from pole to pole: selected contributions to the XXVth international conference of the ICA, Dresden 2013*. Springer, Berlin/Heidelberg, pp 303–317. https://doi.org/10.1007/978-3-642-32618-9_22
46. Bui DT, Ho TC, Pradhan B et al (2016) GIS-based modeling of rainfall-induced landslides using data mining-based functional trees classifier with Ada Boost, Bagging, and Multi Boost ensemble frameworks. *Environ Earth Sci* 75:1101
47. Althwaynee OF, Pradhan B, Park HJ et al (2014) A novel ensemble bivariate statistical evidential belief function with knowledge-based analytical hierarchy process and multivariate statistical logistic regression for landslide susceptibility mapping. *Catena* 114:21–36
48. Althwaynee OF, Pradhan B, Lee S (2016) A novel integrated model for assessing landslide susceptibility mapping using CHAID and AHP pair-wise comparison. *Int J Remote Sens* 37:1190–1209
49. Saito H, Nakayama D, Matsuyama H (2009) Comparison of landslide susceptibility based on a decision-tree model and actual landslide occurrence: the Akaishi Mountains, Japan. *Geomorphology* 109:108–121
50. Yeon YK, Han JG, Ryu KH (2010) Landslide susceptibility mapping in Injae, Korea, using a decision tree. *Eng Geol* 116:274–283

51. Catani F, Lagomarsino D, Segoni S et al (2013) Landslide susceptibility estimation by random forests technique: sensitivity and scaling issues. *Nat Hazards Earth Syst* 13:2815–2831
52. Trigila A, Iadanza C, Esposito C et al (2015) Comparison of logistic regression and random forests techniques for shallow landslide susceptibility assessment in Giampilieri (NE Sicily, Italy). *Geomorphology* 249:119–136
53. Were K, Bui DT, Dick OB et al (2015) A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. *Ecol Indic* 52:394–403. <https://doi.org/10.1016/j.ecolind.2014.12.028>
54. Pham BT, Bui DT, Prakash I et al (2016) Rotation forest fuzzy rule-based classifier ensemble for spatial prediction of landslides using GIS. *Nat Hazards* 83:97–127
55. Klose M (2015) Landslide databases as tools for integrated assessment of landslide risk. Springer Theses, Springer, Berlin
56. Breiman L (1996) Bagging predictors. *Mach Learn* 24:123–140
57. Aggarwal CC (2015) Data classification: advanced concepts. In: *Data mining: the textbook*. Springer, Cham, pp 345–387
58. Breiman L (2001) Random forests. *Mach Learn* 45:5–32
59. Rokach L (2016) Decision forest: twenty years of research. *Inform Fusion* 27:111–125
60. Rodriguez JJ, Kuncheva LI (2006) Rotation forest: a new classifier ensemble method. *IEEE T Pattern Anal* 28:1619–1630
61. Rokach L (2010) Ensemble-based classifiers. *Artif Intell Rev* 33:1–39
62. Schölkopf B, Smola AJ (2002) *Learning with kernels: support vector machines, regularization, optimization and beyond*. MIT Press, Cambridge, MA
63. Gleason CJ, Im J (2012) Forest biomass estimation from airborne LiDAR data using machine learning approaches. *Remote Sens Environ* 125:80–91
64. Mancini F, Ceppi C, Ritrovato G (2010) GIS and statistical analysis for landslide susceptibility mapping in the Daunia area, Italy. *Nat Hazards Earth Syst* 10:1851–1864
65. Kavzoglu T, Sahin EK, Colkesen I (2014) Landslide susceptibility mapping using GIS-based multi-criteria decision analysis, support vector machines, and logistic regression. *Landslides* 11:425–439
66. Wang YT, Seijmonsbergen AC, Bouten W et al (2015) Using statistical learning algorithms in regional landslide susceptibility zonation with limited landslide field data. *J Mt Sci-Engl* 12:268–288
67. Feng HJ, Yu JJ, Zheng JL et al (2016) Evaluation of different models in rainfall-triggered landslide susceptibility mapping: a case study in Chunan, southeast China. *Environ Earth Sci* 75 (21):1–15
68. Pham BT, Pradhan B, Bui DT et al (2016) A comparative study of different machine learning methods for landslide susceptibility assessment: a case study of Uttarakhand area (India). *Environ Model Softw* 84:240–250
69. Yilmaz I (2009) Landslide susceptibility mapping using frequency ratio, logistic regression, artificial neural networks and their comparison: a case study from Kat landslides (Tokat-Turkey). *Comput Geosci* 35:1125–1138