



# A comparative study of support vector machine and logistic model tree classifiers for shallow landslide susceptibility modeling

Mousa Abedini<sup>1</sup> · Bahareh Ghasemian<sup>1</sup> · Ataollah Shirzadi<sup>2</sup> · Dieu Tien Bui<sup>3</sup>

Received: 15 May 2018 / Accepted: 20 August 2019 / Published online: 6 September 2019  
© Springer-Verlag GmbH Germany, part of Springer Nature 2019

## Abstract

The main aim of this study was to evaluate and compare the results of two data-mining algorithms including support vector machine (SVM) and logistic model tree (LMT) for shallow landslide modelling in Kamyaran county where located in Kurdistan Province, Iran. A total of 60 landslide locations were identified using different sources and randomly divided into a ratio of 70/30 for landslide modeling and validation process. After that, 21 conditioning factors, with a raster resolution of 20 m, based on the information gain ratio (IGR) technique were selected. Performance of the models was evaluated using area under the receiver-operating characteristic curve (AUROC), and also several statistical-based indexes. Results depicted that only eight factors including distance to river, river density, stream power index (SPI), rainfall, valley depth, topographic wetness index (TWI), solar radiation, and plan curvature were known more effective for landslide modeling using training data set. The results also revealed that the SVM model (AUROC = 0.882) outperformed and outclassed the LMT model (AUROC = 0.737). Therefore, analysis and comparison of the results showed that the SVM model by RBF function performed well for landslide spatial prediction in the study area. Eventually, the findings of this study can be useful for land-use planning, reducing the risk of landslide, and decision-making in areas prone to landslide.

**Keywords** Landslide · Machine learning · Decision tree · Goodness-of-fit · GIS · Iran

## Introduction

Understanding the mechanism of landslides and mapping of areas prone to landslide play a crucial role in disaster management, and it is possible to be used as a standard tool to support decision-making in different areas (Bui et al. 2016). Landslide is slope instabilities from the ingredients of hill slope. This phenomenon happens when the shear stress on slopes is greater than the shear strength of materials on the slope (Cruden 1991). Based on the studies conducted by European geotechnical thematic network (EGTN), just landslide accounts for about 17% of the world's natural disasters.

According to the reports of fatal events and financial damages of landslides in the literatures, recognizing the areas prone to landslide and determining their risk level are one of the important steps to be taken. Thus, in the last 2 decades, extensive research has been done on the methods of preparing maps for susceptibility and hazard mapping of landslides using tools and new technologies for the facilitation and acceleration of this field of inquiry. In general, it can be claimed that the landslide hazard map is a proper tool for crisis management in mountainous areas (Dahal et al. 2008). Therefore, to achieve a reliable and an accurate landslide susceptibility map, some quantitative methods should be tested and evaluated to better management of mountain areas.

Several methods and techniques have been developed for natural hazard susceptibility mapping such as landslides over the world. It can be classified as (1) expert knowledge-based models (Shirzadi et al. 2017b; Zhang et al. 2016), (2) bivariate and multivariate statistical-based models such as frequency ratio (FR) (Pham et al. 2015b; Shirzadi et al. 2017b), statistical index (SI) (Regmi et al. 2014), certainly factor (Hong et al. 2017a), logistic regression (LR) (Abedini

✉ Mousa Abedini  
Abedini@uma.ac.ir

<sup>1</sup> Department of Geomorphology, Faculty of Humanities, University of Mohaghegh Ardabili, Ardabil, Iran

<sup>2</sup> Department of Rangeland and Watershed Management, Faculty of Natural Resources, University of Kurdistan, Sanandaj, Iran

<sup>3</sup> Institute of Research and Development, Duy Tan University, Da Nang 550000, Vietnam

et al. 2017; Chen et al. 2019; Mousavi et al. 2011; Shirzadi et al. 2012; Tsangaratos and Ilia 2016), weights-of-evidence (WOE) (Chen et al. 2018b; Dahal et al. 2008; Shirzadi et al. 2017b; Xu et al. 2012), (3) machine learning models such as artificial neural network (ANN) (Pham et al. 2017; Shirzadi et al. 2017c), adaptive neuro-fuzz inference system (ANFIS) (Shirzadi et al. 2017c), decision tree (DT) (Khosravi et al. 2018; Thai Pham et al. 2018a, b), support vector machine (SVM) (Bui et al. 2016; Pradhan 2013; Pourghasemi et al. 2013; Tien Bui et al. 2018b; Yao et al. 2008), Bayesian logistic regression (BLR) (Chapi et al. 2017; Das et al. 2012; Tien Bui et al. 2018d), kernel logistic regression (KLR) (Chen et al. 2018c), logistic model tree (LMT) (Chen et al. 2018b, 2019), alternate decision tree (AD Tree) (Shirzadi et al. 2018; Tien Bui et al. 2018d), naïve bayes (He et al. 2019; Shirzadi et al. 2017b, c; Tien Bui et al. 2012), Bayes net (BN) (Tien Bui et al. 2018d), random forest (RF) (Doetsch et al. 2009; Hong et al. 2016; Huang and Zhao 2018), reduced error pruning tree (REPTree) (Pham et al. 2019), and hybrid models including machine learning and optimization algorithms (Abedini et al. 2018; Ahmadi et al. 2018; Bui et al. 2018; Chen et al. 2018a, b; Hong et al. 2017b, c, 2018a, b; Miraki et al. 2018; Nguyen et al. 2019; Pham et al. 2018a, b, 2019; Shafizadeh-Moghadam et al. 2018; Shirzadi et al. 2018; Tien Bui et al. 2018a, b, c).

However, evaluation and comparison of landslide using SVM and DT models are under researched over the world. Among this, Tien Bui et al. (2012) investigated landslide susceptibility in Vietnam using SVM, DT, and Bayes-based methods. Their finding revealed that the SVM model by RBF function had the highest prediction power (AUROC = 0.954), while the DT model had the lowest performance (AUROC = 0.903) in their study area. Pradhan (2013) compared the ability of decision trees, support vector machines, and fuzzy logic methods in assessing landslide susceptibility in Malaysia. Their results concluded that the percentage of AUROC for the obtained zoning map in fuzzy logic, SVM, and DT were 91.24, 91.67, and 88.36, respectively. Hong et al. (2015) predict landslide susceptibility maps using LR, DT, and SVM algorithms in Yihuang in China. Comparison and validation of the models were evaluated using AUROC index. According to their results, AUROC using training data set that shows the performance of the models in the LR model was 92.5%, and for the SVM, it was 88.8%, and for the DT, it was 95.7%. However, the predictive capability using validation data set in the LR model, the SVM, and the DT models were 81.1%, 84.2%, and 93.3%, respectively.

Although many models and methods have been employed for mapping landslide susceptibility around the world, however, it is still a debate on which model is the best performance. In this research, we compared a functional soft-computing benchmark model, a SVM, and a decision

tree soft-computing benchmark model, a LMT, for shallow landslide susceptibility modeling in a part of Kamyaran in Kurdistan province, Iran. Although they have been used for shallow landslide modelling in other regions over the world, however, they earlier have not been explored for the case study. Basically, the obtained result can be considerable as a reference for evaluating of the obtained results. Therefore, the primary aims of the current research are as follows; (1) selecting the most important and effective factors in landslide susceptibility assessment; (2) application of the tow data-mining models including LMT and SVM to achieve a reasonable landslide susceptibility map. It is mentioned that data preparation and processing was done using ArcGIS 10.2 and WEKA 2.7.12 software.

## Study area characteristics

The study area is located in the eastern part of the Kurdistan province, in the eastern longitude from 46° 47' 30" to 47° 00' 00", and the northern latitude from 34° 47' 00" to 34° 58' 30" with an area of approximately 516.44 km<sup>2</sup> (Fig. 1). In terms of topography and physical characteristics, the maximum, minimum, and average height above the sea level, respectively, was 2841, 1388, and 1757 m, and the difference between the highest and lowest elevation point was about 1453 m. Geologically, it included eight groups: (1) Quaternary period (Cenozoic era), (2) the end of the Cretaceous and the beginning of Paleocene (Mesozoic–Cenozoic era), (3) the end of Eocene and the beginning of Oligocene (Cenozoic era), (4) Paleocene–Eocene (Cenozoic era), (5) The end of the Cretaceous–the beginning of Paleocene (Mesozoic–Cenozoic era), (6) The end of Oligocene–Miocene (Cenozoic era), (7) Cretaceous era (Mesozoic era), and (8) Jurassic–Cretaceous (Mesozoic era) with the domination of alluvial deposits, limestone, sand, frequent shale, Flisch, sandstone, and conglomerate. Basically, about 70% of the landslides in the area have occurred during these formations. In addition, a broadly field survey indicated that the type of rotational slip (54.61%) and compound slip (31.53%), respectively, had accounted for the highest amount of landslides. The maximum length of landslide was 3043 m and the minimum length was 55 m. Meanwhile, the maximum width of landslide was 6320 m and the minimum width of landslide in this area was 87 m. According to Dumbarton (23.56) climate classification, this area had a Mediterranean climate. The average rainfall in the study area during 2000–2012 was about 560 mm per year and the average of annual temperature was about 13.69° C.

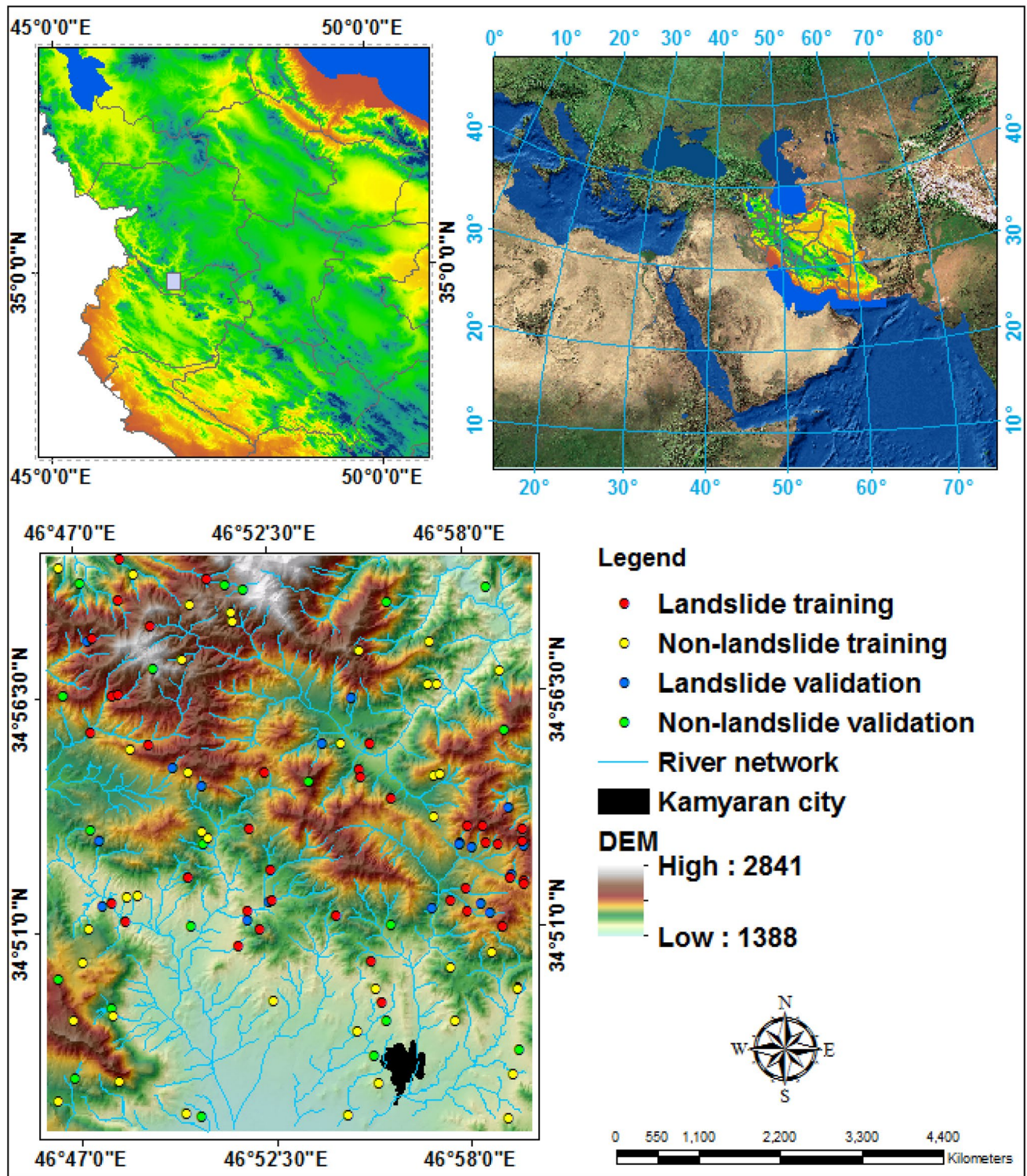


Fig. 1 Landslide location map of the study area



## Methodology

### Landslide inventory map

To study the relationship between the spatial prediction of landslides and the relevant conditioning factors, the existing landslide inventory map is required. Therefore, to come up with a detailed and reliable inventory map for the study area, two processes were utilized including extensive field surveys and accurate laboratory interpretations. First, the location of landslides collected from the Forests, Rangelands and Watershed Management Organization (Iran), and then, using field surveys, this location was checked and some characteristic of each landslide was recorded including length, width, and area of landslides using GPS and aerial photographs (1:40,000 scale), and satellite image interpretations (Fig. 2). Field surveys showed that type of landslides of the study area was mainly rotational sliding (54.61%), complex (31.53%), rotational falling (13.07%), and flow (0.79%), respectively. Maximum length of the landslides was 3043 m, while minimum length was 55 m. The landslide width varied from 87 m and 6320 m (Table 1). The largest landslide covered an area of about  $1.42 \times 10^7$  m<sup>2</sup>, whereas the smallest one was around  $2.52 \times 10^4$  m<sup>2</sup>. Landslides usually occurred

**Table 1** Main characteristic of the landslides in the study area

Statistical measure	Landslide length (m)	Landslide width (m)
Total range of length	55 ~ 3043	87 ~ 6320
Average	452.81	702.98
S.D	405.82	826.88
Skewness	3.69	3.69
kurtosis	18.86	18.75
Median	344.50	456.60
Mode	225.00	500.00

on slope materials with mixture soil (alluvium and gravel fans) and (mainly consist of Flisch, sandstone, and conglomerate formation) in the study area. In addition, results of field surveys concluded that the most important and effective factors for the landslide occurrence in the study area were alternative loose and dense soil layers (61.53), erosion and cutting the foot of slopes (29.23), tectonic (faults and fractures) (6.17), land-use change (3.07), respectively. Furthermore, about 62.30% of the landslides were inactive, while only about 37.69% of the landslides were active. In the study area, a total number of 60 landslides polygons were recognized and then classified into 70% (40 landslides) as training data set for modelling process and 30% (20 landslides) as validation data set for



**Fig. 2** Some photos of shallow landslides in the study area

validation process of the models (Shirzadi et al. 2018, 2019). In addition, a total number of 60 non-landslide locations were randomly selected over the world, and similar to the modeling step and validation check, they were categorized into two groups of 30% and 70%.

## Landslide conditioning factors

Identifying past and present locations of landslide occurrence is the first and most important step in the mapping of landslide susceptibility (Jiménez-Perálvarez et al. 2011). Landslide modeling is based on the statistical hypothesis that future landslides will occur under the same conditions as the past and present ones (Guzzetti et al. 1999). Landslide dispersion map is essential for understanding the effective factors that cause slope failure and change of their mechanism (Dai et al. 2002). In this study, geological map with a scale of 1:100,000 was obtained from the Iranian Geological Organization, and topographic map with a scale of 1:50,000 was obtained from the Armed Forces Geographical Organization, and rainfall data from nine meteorological stations located in the region were used to provide the distribution/inventory map of the

landslides in the study area. To come up with the landslide inventory map, specifying the exact place of landslides and creating a spatial data set for landslide risk in future studies are essential. At first, determining landslide zones and their locations according to the interpretation of aerial photographs and satellite images was provided from the Google Earth Images. After field surveying, the location of each landslide zones was recorded by a GPS device and the obtained information from Arc GIS 10.2 to prepare the landslide inventory maps. One of the basic steps in mapping landslide susceptibility is creating a data set and collecting the required data (Kavzoglu et al. 2015). For the selection of effective factors in landslide occurrence in the study area, almost the majority of variables involved in landslide were investigated as primary independent variables in the present study, and then, they were linked to the position of sliding zones in the inventory map (Table 2). In the next step, after identifying the conditioning factors affecting landslide in this area, through literature review and investigating features in landslide areas, data layers including 21 factors were recognized. Accordingly, slope angle, slope aspect, curvature, elevation above the sea level, profile curvature, plan curvature, solar radiation, valley depth (VD), Stream Power

**Table 2** Landslide conditioning factors and their classes

Factors	Classes
Slope (°)	(1) < 8; (2) 8–15; (3) 15–25; (4) 30–40; (5) 25–35; (6) 35–45; (7) > 45
Aspect	1) Flat; (2) north; (3) northeast; (4) east; (5) southeast; (6) south; (7) southwest; (8) west; (9) northwest
Elevation (m)	(1) 0–700; (2) 700–900; (3) 900–1100; (4) 1100–1300; (5) 1300–1500; (6) 1500–1700; (7) 1700–1900; (8) 1900–2100; (9) 2100–2300; (10) > 2300
Curvature	(1) High concave [(-76 to -7.2)]; (2) concave [(-7.1 to -2.4)]; (3) flat [(-2.3 to 0.8)]; (4) convex [(0.81–5.6)]; (5) high convex [(5.7–60)]
Plan curvature	(1) [(-16 to -1.8)]; (2) [(-1.7 to -0.53)]; (3) [(-0.52 to -0.28)]; (4) [(0.29–1.2)]; (5) [(1.3–10)]
Profile curvature	(1) [(-16.25 to -1.384641)]; (2) [(-1.384641 to -0.433258)]; (3) [(-0.433258 to 0.280279)]; (4) [0.280279–1.231662]; (5) [1.231662–13.95641]
Solar radiation	(1) 237000–517000; (2) 518,000–603,000; (3) 604,000–674,000; (4) 675,000–734,000; (5) 735,000–881,000
VD	(1) 0–50; (2) 50–100; (3) 100–150; (4) 150–200; (5) > 200
SPI	(1) 0–10; (2) 10–20; (3) 20–30; (4) 30–40; (5) 40–50; (6) 50–60; (7) > 60
TWI	(1) 0–10; (2) 10–20; (3) 20–30; (4) 30–40; (5) 40–50; (6) 50–60; (7) > 60
LS	(1) 0–10; (2) 10–20; (3) 20–30; (4) 30–40; (5) 40–50; (6) 50–60; (7) > 60
Land use	(1) Barren land, (2) dry-farming land, (3) farming land, (4) pasture, (5) residential area, (6) woodland, (7) water body
NDVI	(1) [(-0.26 to 0.056)]; (2) [(0.057–0.11)]; (3) [(0.12–0.17)]; (4) [0.18–0.28]; (5) [0.29–0.6]
Rainfall	(1) < 300; (2) 300–400; (3) 400–500; (4) 500–600; (5) 600–700; (6) 700–800; (7) > 800
Distance to faults (m)	(1) > 100; (2) 101–200; (3) 201–300; (4) 301–400; (5) 401–500; (6) 501–600; (7) 601–700; (8) > 700
Distance to roads (m)	(1) > 50; (2) 51–100; (3) 101–150; (4) 151–200; (5) 201–250; (6) > 250
Distance to rivers (m)	(1) > 50; (2) 51–100; (3) 101–150; (4) 151–200; (5) 201–250; (6) > 250
Faults density	(1) Very low (0–0.437); (2) Low (0.438–1.17); (3) moderate (1.18–1.93); (4) high (1.94–2.88); (5) very high (2.88–5.87)
Road density	(1) Very low (0–0.428); (2) Low (0.429–1.17); (3) moderate (1.18–1.87); (4) high (1.88–2.7); (5) very high (2.7–4.96)
River density	(1) Very low (0–0.00049); (2) low (0.0005–0.0013); (3) moderate (0.0014–0.0022); (4) high (0.0023–0.0032); (5) very high (0.0033–0.006)
Lithology	Group 1(Quaternary), group 2 (Late cretaceous–early Paleocene), group 3 (late Eocene-early Oligocene), group 4 (Paleocene-Eocene), group 5 (late cretaceous-Paleocene), group 6 (late Oligocene–Miocene), group 7 (cretaceous), group 8 (Jurassic-cretaceous)

Index (SPI), Topographic Wetness Index (TWI), and length slope were selected. Land-use map and Normalized Difference Vegetation Index (NDVI) were prepared by ETM<sup>+</sup> satellite image of the study area in 2 May 2005 (the Path and Row of this satellite images are 167 and 36, respectively) lithological map, distance from fault and fault density of the geological map in Kamyaran at a scale of 1:100,000. Rainfall map was prepared based on regression relationship between height, and long-term average of annual precipitation in nine rain-gauge stations inside and outside of the study area. Distance from drainage, drainage density, distance from the road network and road density maps, respectively, were made based on distance from the areas around the drainage and road network in the study area. Then, for statistical analysis of the data and algorithms, values of different classes related to each factor were entered into the WEKA 2.7.12. Ultimately, the final maps of landslide susceptibility in the study area were mapped using a combination of effective layers in landslide occurrence.

### Modeling using machine learning algorithm

#### Support vector machine (SVM) function

The algorithm of SVM was proposed by Vapnik (1998) based on Statistical Learning Theory (SLT) that follows Structure Risk Minimization (SRM). Indeed, it's an efficient learning system based on useful optimization theory that used inductive minimization principle of structural error leading to an overall optimal solution (Cristianini and Shawe-Taylor 2000). The main idea of SVM algorithm with a dual categorization and learning points changes the main entrance area to a higher dimensional space to find a suitable cloud page (Peng et al. 2014). Training points that are close to the desired page are called the support vector. When decision level is obtained, it can be used to estimate new data (Tien Bui et al. 2012). This method is a new class of models for the purpose of classification and prediction. Detailed explanations about two classes of SVM modeling in the study area are as follows:

Considering a set of linear separate training cells as Eq. (1):

$$X_i = (i = 1, 2, \dots, n). \tag{1}$$

Training cells included two classes of  $Y_i = \pm 1$ , and they were specified as the goals of SVM model to search for differentiate a hyper-plan of  $-N$  dimensional in two classes that were determined by their maximum gap. Mathematically, it can be said that as Eq. (2):

$$1/2 = \|W^2\|. \tag{2}$$

Subject to the limitations of the following Eq. (3):

$$1 \geq Y_i = ((W \cdot X_i) +). \tag{3}$$

Here,  $W^2$  a rule of normal is hyper-plan of a denominator and  $(.)$  specifies numerical production. By multi-coefficient of Lagrangian, the operational calculation value can be defined by the following equation:

$$-\sum_{i=1}^n \gamma_i ((W \cdot X_i) + b) - \|W\|^2 L = \frac{1}{2}. \tag{4}$$

In which  $\lambda_i$  is multi-coefficient of Lagrangian. This solution can be calculated through minimization of Eq. 1. Valuation of  $W$  and  $B$  variables was done by standard methods.

Therefore, Eq. (5) changes to the following:

$$-\frac{1}{vn} \sum_{i=1}^n \epsilon_i \|W\|^2 L = \frac{1}{2}. \tag{5}$$

$V(0, 1)$  was introduced for categorization (Hastie et al. 2002; Schölkopf et al. 2000). In addition, Vapnik (1998) introduced a kernel function to count non-linear boundary decision selection of kernel function in SVM model which is very important, although kernel functions of  $K(X_i, X_j)$  have been mostly used in the past. Just some of them were identified beneficial in a wide range of applications. Those that show these attributes are as follows (6), (7), (8), and (9):

Linear function:

$$K(X_i, X_j) = X_i^t \cdot X_j. \tag{6}$$

Polynomial function:

$$K(X_i, X_j) = (Y \cdot X_i^t \cdot X_j + r)^d, y > 0. \tag{7}$$

Radial basis function:

$$K(X_i, X_j) = e^{-y} (X_i - X_j)^2, y > 0. \tag{8}$$

Circular function:

$$\tan h = (y \cdot X_i^t \cdot X_j + r). \tag{9}$$

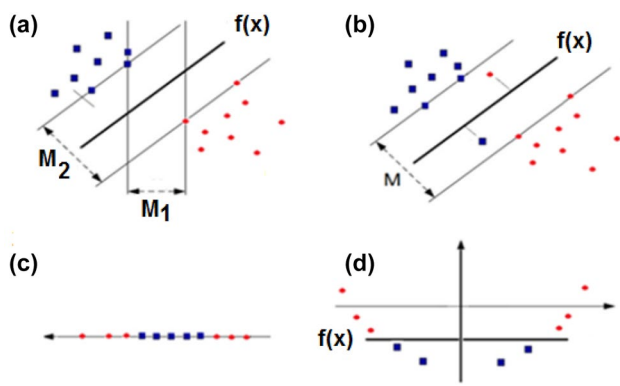
Therefore,  $r$ ,  $y$ , and  $d$  are kernel function parameters and they are entered manually. Sometimes core functions are used as the following parameters [Eq. (10)]:

$$y = 1/2\sigma^2. \tag{10}$$

Here,  $\sigma$  is an adjustable parameter that has the control of the kernel function.

If taken up, the exponential pattern becomes linear, and where there is the possibility of losing hyper-plans, if they are taken down, the decision boundary for the error in the training data becomes visible. In this study, +1 and -1, respectively, refer to landslide and slope stability of the place. In recent years, according to non-linear transmission along with a large special scale, this algorithm has been considered as one of the most popular methods for





**Fig. 3** **a** Maximum-margin classifier  $f(x)$  separating circles from squares in  $R_2$ ; **b** soft-margin classifier letting some points be misclassified; **c** linearly inseparable case in  $R_1$ ; **d** mapping original input space into feature space of higher dimension ( $R_2$ ). After mapping, classes get linearly separable

problem-solving in classification and regression (Kavzoglu et al. 2015). On the other word, the performance of SVM algorithm differs from other methods of separation and tries to construct a series of training points through the plan of differentiation. Figure 3 explains the performance of SVM algorithm: a) the maximum margin of separation in  $f(x)$  function, separating the circle points from square; b) classification of software margin which allows classification of some incorrect points; c) inseparable linear case; d) spatial mapping from the main entrance to higher dimensional space and separation of linear classes after mapping.

This algorithm follows the maximum margin of separation between classes  $M_2 > M_1$  and it's based on the construction of a classified cloud page between the maximum-margin ( $F(x)$  function in Fig. 3). If the target point is higher than the hyper-plan, it is classified as +1 (the square in Fig. 3a); otherwise, it is -1 (the circle in Fig. 3a). Since the set of classified data makes more noise, therefore, the SVM algorithm for finding  $f(x)$  function locates some points in the margin of the separation (Fig. 3b). In fact, the main idea of mapping relates to high dimensions and spaces which are shown to change non-linear case to the linear one (Fig. 3c and Fig. 3d).

**Logistic model tree (LMT) classifier**

A decision tree (DT) is a new generation of data-mining techniques which have been widely developed in the past 2 decades. This algorithm is a non-parametric method that considers the prediction of quantitative variables or classification of variables according to a set of predictor quantitative and qualitative variables (Pradhan 2013). In fact, DT is a hierarchical model of decision-making tools that recursively splits independent variables into homogeneous zones (Cho and Kurup 2011; Myles et al. 2004;

Pradhan 2013). DTs used to predict discrete variables are called classification trees, because they put the samples into categories or classes (Tien Bui et al. 2012). DTs used to predict continuous variables are called regression trees (Debeljak and Dzeroski 2009). The aim of a DT is to find a strategy to offer the obtained results of predictions from a set of input variables in the form of a series of rules (Han et al. 2011). The results of this model have been successfully used in many real-world conditions for classification and prediction issues (Murthy 1998). Among many classification algorithms including Bayes-based theorem algorithms and rule-based algorithms, DT algorithms is an efficient and effective one for the classification of large data sets (Murthy 1998).

Many algorithms for modeling of DTs have been developed including classification and regression tree (CART), best-first decision tree (BFT), ADT, LMT, RF, REP tree, random tree (RT), ID3, Chi-square automatic interaction detector decision tree (CHAID), C4.5, and so on. The only drawback of this model is that some decision tree methods can only predict and classify binary variables (yes and no or acceptance and rejection), and in some of them, when the number of instances of each class is low, error rate will be raised. In fact, there is sensitivity to noise, training set, and irrelevant characteristics in this model (Zhao and Zhang 2008). The progressive development of machine learning leads to emerge a new robust decision tree algorithm such as the LMT where leaf nodes are substituted with a regression function instead of a constant value. LMT is a mixture of C4.5 decision tree (Quinlan 1996) and logistic regression functions where the information gain is used for splitting, and the LogitBoost algorithm (Landwehr et al. 2005) is employed to fit the logistic regression functions at a tree node. To impede the problem of over-fitting of the ultimate LMT, the CART algorithm is utilized for pruning. The LogitBoost algorithm conducts additive logistic regression with least-squares fits for each class of  $C_i$  (landslide or non-landslide) (Doetsch et al. 2009). The posterior probabilities in the leaf nodes of the LMT are calculated by the linear logistic regression (Landwehr et al. 2005) as Eq. (11):

$$P(C/X) = \frac{\exp(L_c(X))}{\sum_{c=1}^c \exp(L_c(X))} \tag{11}$$

Here,  $C$  is the number of classes and the least-square fits  $L_C(x)$  are transformed, so that as Eq. (12):

$$\sum_{c=1} cL_c(X) = 0. \tag{12}$$

The flowchart of the methodology of the study area is shown in Fig. 4.

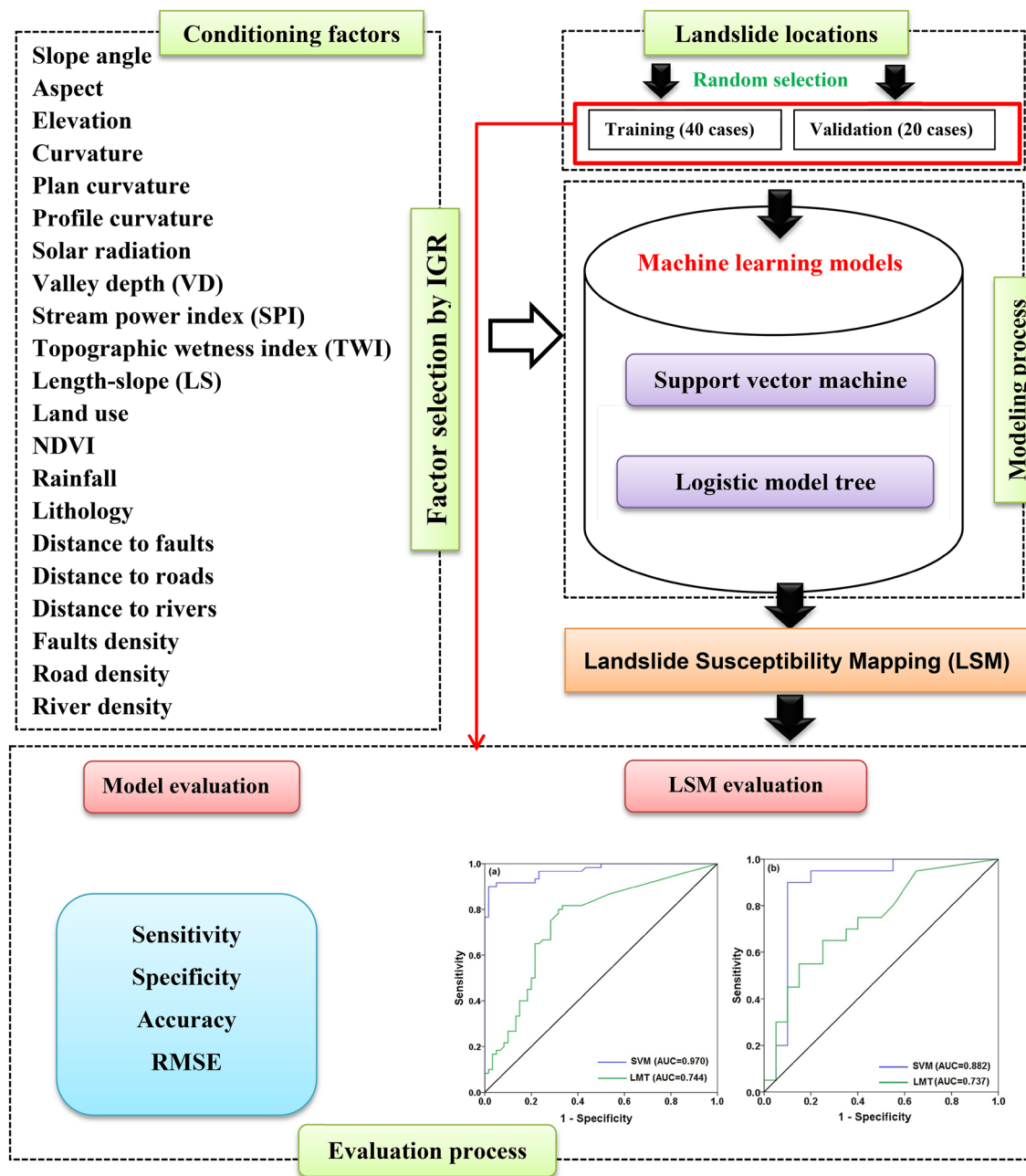


Fig. 4 The flowchart showing methodology of the landslide susceptibility analyses

### Factor selection using information gain ratio (IGR) technique

Landslide susceptibility evaluation is based on its determining factors (Costanzo et al. 2012). There are several methods to determine the predictive power of the factors affecting landslide occurrence such as Relief Significance (Ahmad and Dey 2005), Gain Ratio (Nithya and Duraiswamy 2014), and Information Gain Ratio (Bui et al. 2016; Chapi et al. 2017; Shirzadi et al. 2017b). In the

present study, the Information Gain Ratio (IGR), proposed for the first time by (Quinlan 1993), was used to determine the quantitative predictive power of the influencing factors. Higher IGR values indicate higher predictive power of an effective factor for the modeling. IGR technique was used to identify the most important factors among the 21 effective factors affecting landslide occurrence in the study area. If  $F$  is the training data with  $n$  input sample, and  $n(M_i, F)$  is the number of samples in the training data of  $F$



belonging to  $M_i$  class (landslide, non-landslide), then the following equation can be formulated:

$$\text{Info}(F, E) = \sum_{i=1}^2 \frac{n(M_i, F)}{|F|} \log_2 \frac{n(M_i, F)}{|F|}(F). \quad (13)$$

Given the factors affecting landslide occurrence, the amount of information required to divide F into the series ( $F_1, F_2, \dots, F_m$ ) is estimated through Eq. 14:

$$\text{Info}(F, E) = \sum_{j=1}^m \frac{F_j}{|F|} \text{Info}(F). \quad (14)$$

The IGR index for a specific effective factor, such as S factor (slope), is calculated by Eq. 15:

$$\text{Information gain ratio}(F, S) = \frac{\text{Info}(F) - \text{Info}(F, S)}{\text{Split info}(F, S)}, \quad (15)$$

where split info represents the information generated by dividing F of the training data into subset of  $l$  calculated by Eq. 16:

$$\text{Split Info}(F, E) = \sum_{j=1}^l \frac{F_j}{|F|} \log_2 \frac{F_j}{|F|}. \quad (16)$$

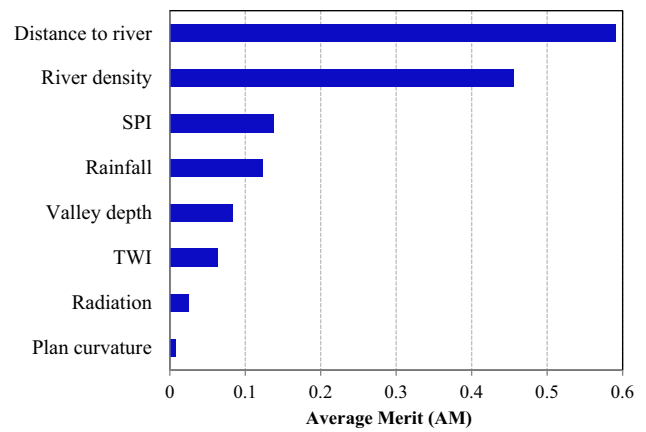
## Results and analysis

### Selection of landslide conditioning factors

To assess the predictive power of the landslide models, the factors affecting landslide were evaluated by IGR technique  $n$  the study area. Table 3 and Fig. 5 show the mean results of the IGR index for the 21 selected effective factors on the surface landslide occurrence in the study area. These results

**Table 3** Average information gain ratio for the landslide conditioning factors

Influencing factors	IGR	Influencing factors	IGR
Distance to river	0.591	Curvature	0.000
River density	0.456	Profile curvature	0.000
SPI	0.136	LS	0.000
Rainfall	0.123	Land use	0.000
VD	0.084	Lithology	0.000
TWI	0.063	NDVI	0.000
Solar radiation	0.025	Distance to fault	0.000
Plan curvature	0.008	Distance to road	0.000
Slope angle	0.000	Fault density	0.000
Aspect	0.000	Road density	0.000
Elevation	0.000	–	–



**Fig. 5** The predictive power of the most important landslide conditioning factors in this study

show that distance to river has the highest predictive capability with value (AM=0.591) for landslide model. It is due to that most of landslides have occurred in the proximity of rivers. While plan curvature has the lowest affecting on landslide occurrences (AM=0.008) in the study area. Other factors including river density (AM=0.456), SPI (AM=0.138), rainfall (AM=0.123), valley depth (AM=0.084), TWI (AM=0.063), and radiation (AM=0.025) also have significantly contribution to landslide models, respectively. In contract, 13 conditioning factors (slope angle, aspect, elevation, curvature, profile curvature, LS, land use, lithology, NDVI, distance to faults, distance to road, faults density, and road density) having average merit equal to “0” were removed from the landslide modeling. That was because of making a noise which negatively decreases the prediction power of modeling (Bui et al. 2016).

### Model result and analysis

#### Application of the SVM function algorithm

In this study, the SVM model was evaluated based on radial basis function (RBF) using WEKA 2.7.12 software. The optimal values of parameters for the SVM model are shown in Table 4. The probability of landslide occurrence (PLO) in the range between 0 and 1 for evaluation was transferred to ArcGIS 10.2. When the PLO was closer to 1, it had more probability for landslide occurrence and vice versa. Finally, the obtained landslide susceptibility map regarding the level of probability to landslides was classified in five classes of susceptibility (very low susceptibility, low susceptibility, moderate susceptibility, high susceptibility, and very high susceptibility) (Fig. 5a). The results from the output map in the study area showed that areas near the drainage were more prone to landslide, and according to IGR technique, the drainage factor also had the greatest impact on

**Table 4** Parameters of machine learning algorithms for landslide modelling in this study

Algorithm	Parameters
SVM	Build logistic model, true; C, 0.95; checks turned off, false; epsilon, 1.0E-12; filter type, normalize training data, kernel, RBF; number of folds, -1; number of random seed, 1; tolerance parameter, 0.001
LMT	The minimum number of instances at which a node is considered for splitting, 15; a fixed number of iterations for logitboost, -1

**Table 5** Description of statistic index based evaluations

Name	Formula
TP	–
TN	–
FP	–
FN	–
Sensitivity	$SST = \frac{TP}{TP+FN}$
Specificity	$SPF = \frac{TN}{FP+TN}$
Accuracy	$ACC = \frac{TP+TN}{TP+TN+FN+FP}$
RMSE	$RMSE = \left[ (1/m) \sum_{i=1}^m (e_i - \bar{e}_i)^2 \right]^{0.5}$

the occurrence. Accordingly, more than 70% of the study area was located in high susceptibility class (0.86–0.99) which referred to high potential of the study area to landslide occurrence.

### Application of the LMT classifier algorithm

In the present study, the LMT algorithm was used as a DT algorithm. The implementation of this algorithm was done by WEKA 2.7.12 software. The optimal values of parameters for the LMT model are shown in Table 4. Input and output variables depending on the intended purpose in a special version of DT, namely regression trees, were investigated (Pradhan 2013). DT algorithm offers a pruning mechanism in the induction stage to control tree growth, and in the next stage, the obtained output from landslide susceptibility was acquired according to the respective model. Finally, the observed and predicted values were obtained in the form of landslide susceptibility zoning map of the study area (Fig. 5b).

### Model performance and validation

In the landslide susceptibility assessment, two evaluation processes should be done. The first is for model evaluation and another evaluation is for susceptibility maps. In the model evaluation, four factors including sensitivity, specificity, accuracy, and RMSE were used as evaluation criteria (Table 5). In areas where landslides exist or are not present, the results of modeling criteria are either positive or

negative. This classification leads to the creation of four possibilities modes including true positive (TP), true negative (TN), false positive (FP), and false negative (FN). If the values of the above-mentioned criteria have the number of 1 in a model, then the model will be an appropriate/ideal model (Shirzadi et al. 2017b). The TP and FP are defined as the proportion of the number of pixels that are correctly classified as landslide and non-landslide, respectively. Meanwhile, TN and FN are the number of pixels classified correctly and incorrectly as non-landslide, respectively (Shirzadi et al. 2018). Hence, sensitivity is defined as the number of correctly classified landslides per total predicted landslides. Specificity is the number of incorrectly classified landslides per total predicted non-landslides (Pham et al. 2016; Shirzadi et al. 2017a, 2018). Accuracy is the proportion of landslide and non-landslide pixels which are correctly classified (Bennett et al. 2013). RMSE shows the error metric between the observed and estimated data of models (Bennett et al. 2013). Validation is an essential part of landslide susceptibility and landslide susceptibility maps without validation are worthless (Pradhan 2011). We evaluated performance of landslide models by Receiver-Operating Characteristic (ROC) curve technique which is a standard technique to perform such evaluation (Pham et al. 2016). ROC curve is built by plotting “sensitivity” value on the y-axis and “100-specificity” value on the x-axis. The sensitivity index indicates the number of landslide pixels correctly classified as “landslide” class. The specificity index indicates the number of non-landslide pixels correctly classified as “non-landslide” class (Pham et al. 2016; Shirzadi et al. 2017b). In this study, evaluation of landslide susceptibility with training data and validation check was done by the rate of success index and the rate of prediction. Currently, predictability of landslide susceptibility in the respective area was examined using the area under the curve ROC curve (AROC). Predictability of both sets of training data and validation data was obtained. With respect to assess the accuracy of a landslide susceptibility map, both training and validation data sets were used. Accordingly, when the training data set are used, the curve of assessing the accuracy is used for assessing the goodness-of-fit or performance of the models and when validation check data are used, assessing curve of accuracy in spatial map of prediction is applicable for power prediction or prediction accuracy. AUROC values are between 0.5 and 1, and the most ideal model has the largest area under the curve

(Shirzadi et al. 2017a). When a model cannot estimate the possible occurrence of landslides, AUROC value will be 0.5. Regarding the ROC curve, when the area under the curve is equal to 1, it shows the highest accuracy of the susceptibility map. Qualitative–quantitative correlation of AUROC and estimate evaluation is as follows: 0.9–1, excellent; 0.8–0.9, very good; 0.7–0.8, good; 0.6–0.7, average; 0.5–0.6, weak (Yesilnacar and Topal 2005).

Based on the results of Table 6 and using the most effective factors, the SVM and the LMT models for the training and validation data sets were exploited. The results indicated that the sensitivity, specificity, and accuracy criteria for the training data set in the SVM model were 0.951, 0.966, and 0.958, respectively. While for the LMT model, these values were 0.921, 0.965, and 0.942, respectively. On the other hand, these values for the validation data set in the SVM model were 0.850, 0.850, and 0.851, respectively, while for the LMT model, they were 0.810, 0.842, and 0.825, respectively. Also, in the training and validation data sets of the SVM model, the RMSE had the values of 0.058 and 0.126, respectively. For the LMT model, they were 0.126 and 0.185, respectively. These results indicated that all of these values in the SVM model were lower than those of the LMT model (Table 6).

### Preparation of landslide susceptibility maps and comparison

Landslide susceptibility mapping is the most significant issue in the spatial prediction of landslides (Pham et al. 2015a). Consequently, landslide susceptibility indexes (LSIs) for the LMT and the SVM models were gained and then reclassified based on the natural breaks method. Although all mathematical classifications such as equal interval, quantile, standard deviation, and geometrical interval were tried on classifying the LSIs, the natural breaks method was selected as the logical method due to conformity with the actual conditions of the environment. Eventually,

in this study, the LSMs were categorized into five classes including very low susceptibility, low susceptibility, moderate susceptibility, high susceptibility, and very high susceptibility (Fig. 6).

In this study, 70% of the training landslides were used and ROC curves were plotted for the used models (Fig. 7a). Based on Fig. 6a, according to training data set, the area under the curve using the SVM-RBF algorithm was about 0.970. It means that this algorithm is able to predict areas susceptible to landslides up to about 97%. However, for training data set, the value of the area under the ROC curve for LMT-DT algorithm was obtained to be about 0.744. It implied that the LMT had a performance of 74.4% for recognizing the landslides that maybe occurred in the future over the study area. The prediction accuracy of the models was obtained and plotted based on the validation data set (Fig. 7b). It indicated that the SVM model had a higher accuracy (AUC = 0.882) than the LMT (AUC = 0.737) model for spatial prediction of landslide sin the study area.

In addition to the AUC, the performance of the landslide models was statistically checked. There are two statistical test including Friedman and Wilcoxon signed-rank tests to determine the statistical differences between two or more models (Bui et al. 2016). The null hypothesis is that there is no any significant difference between the SVM and the LMT model for spatial prediction of landslides in the study area. Friedman tests only showed the significant differences among all models without any judgment as pairwise between two or more models (Bui et al. 2016). To check the performance between two or more models as pairwise, the Wilcoxon signed-rank test is generally applied in non-parametric tests. As in this study, only two models were used; we only applied the Friedman test to check the performance of the models. The result is shown in Table 7. It can be concluded that because of significant equaled to 0.000, there was a significant difference between the SVM and the LMT models for landslide susceptibility mapping. However, the obtained result was in agreement with the result of the ROC curve.

**Table 6** Model validation using several statistical evaluation measures

Index	Training dataset		Validation dataset	
	SVM	LMT	SVM	LMT
TP	58	58	17	17
TN	57	55	17	16
FP	2	2	3	3
FN	3	5	3	4
Sensitivity (%)	0.951	0.921	0.850	0.810
Specificity (%)	0.966	0.965	0.850	0.842
Accuracy (%)	0.958	0.942	0.851	0.825
RMSE	0.058	0.149	0.126	0.185

### Discussion and conclusion

Landslide susceptibility mapping plays an important role in providing a platform to decision-makers and authorities, particularly in landslide prone areas. Machine learning methods are more notoriously efficient in solving many real-world problems compared to conventional methods such as expert knowledge methods or analytic methods (Pradhan 2013). The present study comparatively examined machine learning algorithms, namely the SVM and the LMT for landslide susceptibility mapping in Kamyaran county where located in the province of Kurdistan, Iran. However, during the past 2



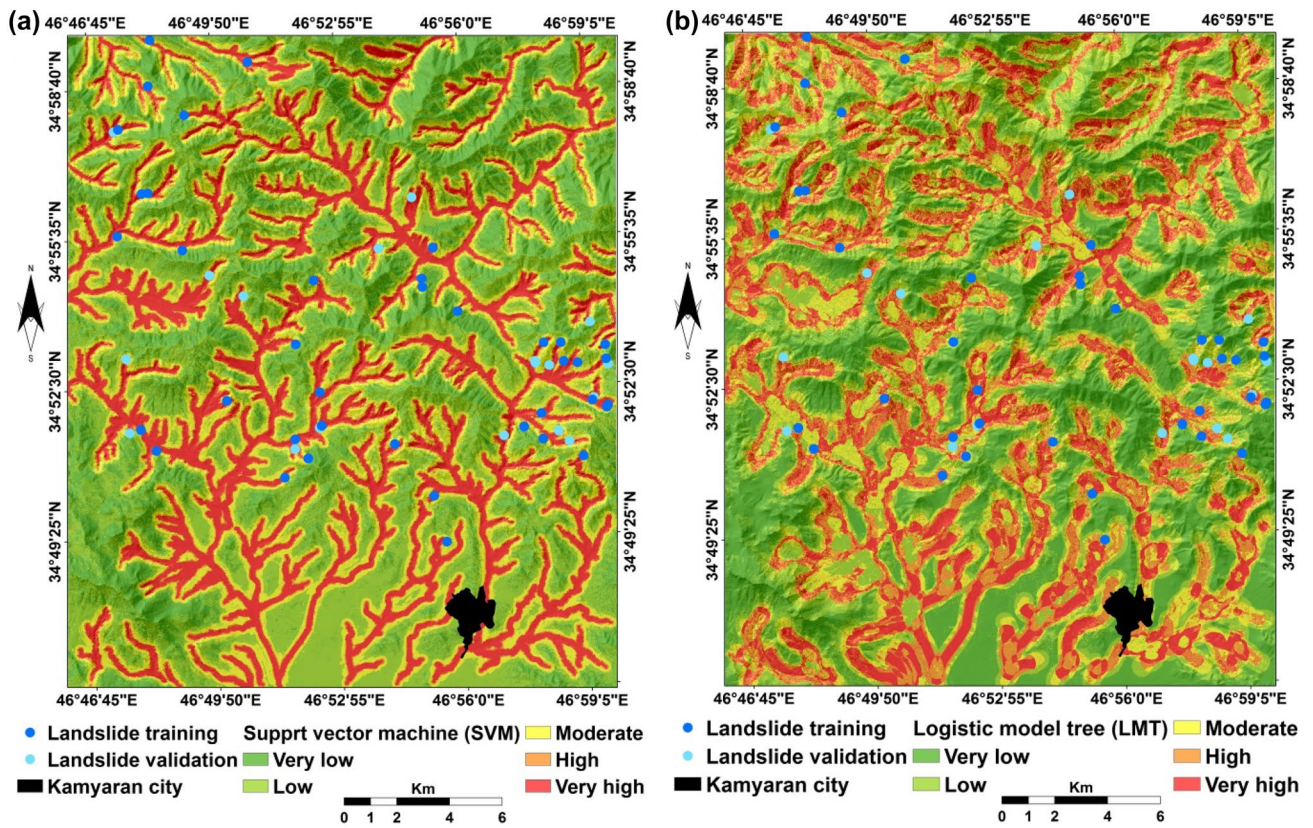


Fig. 6 Landslide susceptibility maps by machine learning models: a support vector machine and b logistic model tree

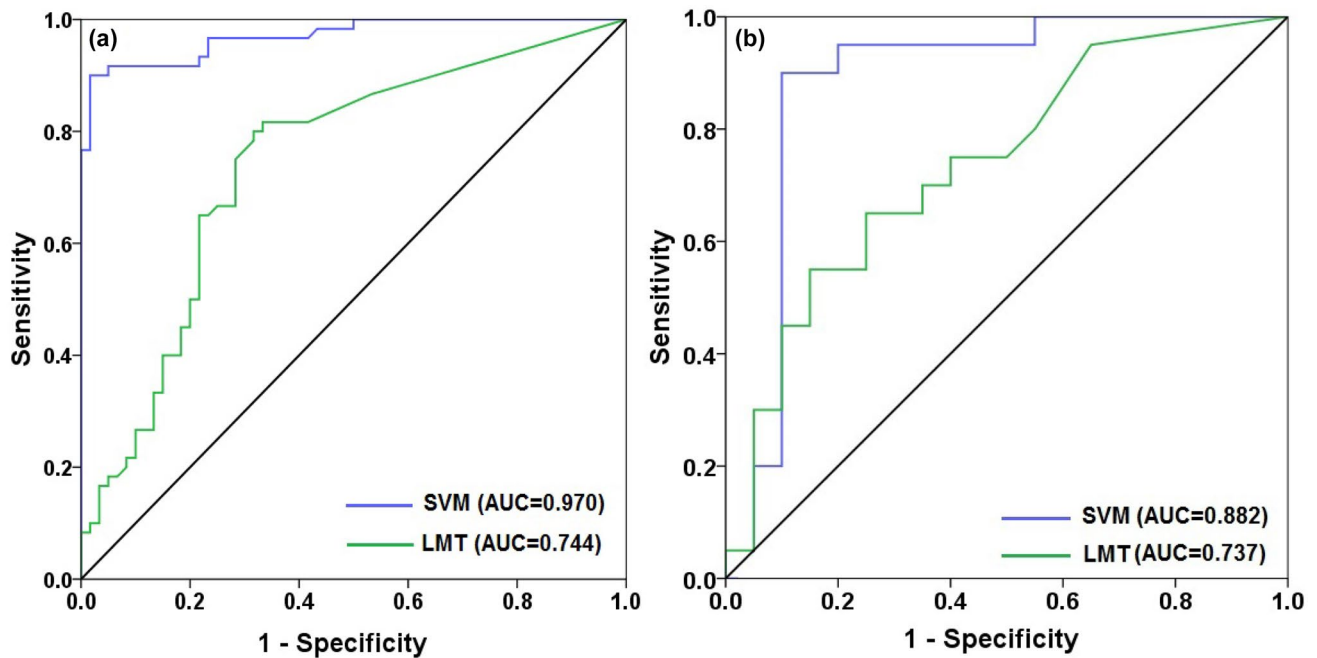


Fig. 7 Performance of landslide models with SVM and LMT models using: a success rate curve of the training dataset; b prediction rate curve of the testing dataset

**Table 7** Friedman test of the two landslide susceptibility models for the study area

No	Landslide models	Mean ranks	$\chi^2$	Significant
1	SVM	2.14	20.651	0.000
2	LMT	3.57		

decades, several methods and techniques have been explored and developed for the landslide modeling. However, till now, the models are limited only to a small number of studies.

Landslide susceptibility maps were prepared with a total of 60 landslide locations. Investigation of the results on the most effective factors among 21 known factors affecting landslide occurrence in the study area based on the AM index of the IGR showed that slope angle, slope aspect, elevation, slope curvature, profile curvature, LS, land use, lithology, NDVI, distance to fault, distance to road, fault density, and road density, due to having zero values were excluded from the final modeling process. However, the most important factor affecting landslide occurrence in the study area in both models was distance to rivers. Distance to river was the most important factor for landslide occurrence, because most of landslides were occurred near the rivers. This result is in concordance with Abedini et al. (2018) and Chen et al. (2017) who they reported that the distance to river was more significant factor for landslide occurrence. The over-performance of SVM is due to its robustness, reducing more noise and variance of training dataset, and also reducing more over-fitting problem than the LMT model in the study area. In addition, model validation process was performed using some statistical-based measures including precision, accuracy, and area under the ROC curve. Modelling process confirmed that the goodness-of-fit and performance of the SVM were found feasible higher than the LMT decision tree algorithm in the study area. Yao et al. (2008) compared the SVM and LR for landslide susceptibility mapping, and they concluded that the SVM was more accurate than the LR model. Tien Bui et al. (2012) compared the SVM, decision tree (DT), and Naïve Bayes (NB) algorithms for landslide modelling, and reported that the SVM was more powerful and performance than the other models. Ballabio and Sterlacchini (2012) also compared the SVM, the logistic regression, linear discriminate analysis, and naïve Bayes algorithms for spatial prediction of landslides. They confirmed the SVM outperformed other algorithms due to more decreasing the over-fitting and variance problems.

In addition, interpretation of landslide susceptibility maps showed that more areas with high and very high susceptibility values are located at the end of slope or where slope is close to the junction of rivers. Perhaps, the reason is the movement of subsurface water from rivers toward

surrounding slopes, creation of a humidity front, and reduction of soil shear strength in this area that it provides some landslides with less depth of failure.

Overall, both of the investigated landslide models showed acceptable performance for landslide susceptibility assessment. However, the SVM model appeared to have a comparatively better performance. Therefore, it can be employed to assess and develop more efficient landslide susceptibility maps for proper landslide hazard management. As a final conclusion, the results can provide very useful information for decision-making, land planning, crisis management, and normal risk reduction in landslide areas.

**Acknowledgements** The authors gratefully acknowledge of the Forests, Rangelands and Watershed Management Organization of Iran for preparing the report of landslide location in the study area, and are thankful to the members of geomorphology department of Mohaghegh Ardabil University and director of environmental management organization of Kurdistan. Ultimately, the authors would like to thank two anonymous honor reviewers and the editor for their helpful comments on the previous version of this manuscript.

## References

- Abedini M, Ghasemyan B, Mogaddam MR (2017) Landslide susceptibility mapping in Bijar city, Kurdistan Province, Iran: a comparative study by logistic regression and AHP models. *Environ Earth Sci* 76:308. <https://doi.org/10.1007/s12665-017-6502-3>
- Abedini M, Ghasemyan B, Shirzadi A, Shahabi H, Chapi K, Pham BT, Bin Ahmad B, Tien Bui D (2018) A novel hybrid approach of bayesian logistic regression and its ensembles for landslide susceptibility assessment. *Geocarto Int.* <https://doi.org/10.1080/10106049.2018.1499820>
- Ahmad A, Dey L (2005) A feature selection technique for classificatory analysis. *Pattern Recogn Lett* 26:43–56
- Ahmadlou M, Karimi M, Alizadeh S, Shirzadi A, Parvinnejhad D, Shahabi H, Panahi M (2018) Flood susceptibility assessment using integration of adaptive network-based fuzzy inference system (ANFIS) and biogeography-based optimization (BBO) and BAT algorithms (BA). *Geocarto Int.* <https://doi.org/10.1080/10106049.2018.1474276>
- Ballabio C, Sterlacchini S (2012) Support vector machines for landslide susceptibility mapping: the Staffora River Basin case study, Italy. *Math Geosci* 44:47–70
- Bennett ND, Croke BF, Guariso G, Guillaume JH, Hamilton SH, Jakeman AJ, Marsili-Libelli S, Newham LT, Norton JP, Perrin C (2013) Characterising performance of environmental models. *Environ Model Softw* 40:1–20
- Bui DT, Tuan TA, Klempe H, Pradhan B, Revhaug I (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
- Bui DT, Panahi M, Shahabi H, Singh VP, Shirzadi A, Chapi K, Khosravi K, Chen W, Panahi S, Li S (2018) Novel hybrid evolutionary algorithms for spatial prediction of floods. *Sci Rep* 8:15364
- Chapi K, Singh VP, Shirzadi A, Shahabi H, Bui DT, Pham BT, Khosravi K (2017) A novel hybrid artificial intelligence approach for flood susceptibility assessment. *Environ Model Softw* 95:229–245
- Chen W, Shirzadi A, Shahabi H, Ahmad BB, Zhang S, Hong H, Zhang N (2017) A novel hybrid artificial intelligence approach based on



- the rotation forest ensemble and naïve Bayes tree classifiers for a landslide susceptibility assessment in Langao County, China. *Geomatics, Natural Hazards and Risk* 8:1955–1977
- Chen W, Shahabi H, Shirzadi A, Hong H, Akgun A, Tian Y, Liu J, Zhu A-X, Li S (2018a) Novel hybrid artificial intelligence approach of bivariate statistical-methods-based kernel logistic regression classifier for landslide susceptibility modeling. *Bull Eng Geol Environ*. <https://doi.org/10.1007/s10064-018-1401-8>
- Chen W, Shahabi H, Shirzadi A, Li T, Guo C, Hong H, Li W, Pan D, Hui J, Ma M (2018b) A novel ensemble approach of bivariate statistical-based logistic model tree classifier for landslide susceptibility assessment. *Geocarto Int*. <https://doi.org/10.1080/10106049.2018.1425738>
- Chen W, Shahabi H, Zhang S, Khosravi K, Shirzadi A, Chapi K, Pham B, Zhang T, Zhang L, Chai H (2018c) Landslide susceptibility modeling based on gis and novel bagging-based kernel logistic regression. *Applied Sciences* 8:2540
- Chen W, Zhao X, Shahabi H, Shirzadi A, Khosravi K, Chai H, Zhang S, Zhang L, Ma J, Chen Y (2019) Spatial prediction of landslide susceptibility by combining evidential belief function, logistic regression and logistic model tree. *Geocarto International*:1–25
- Cho JH, Kurup PU (2011) Decision tree approach for classification and dimensionality reduction of electronic nose data. *Sensors Actuators B* 160:542–548
- Costanzo D, Rotigliano E, Irigaray Fernández C, Jiménez-Perálvarez JD, Chacón Montero J (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 Sci* 12(2):327–340. <https://doi.org/10.5194/nhess-12-327-2012>
- Cristianini N, Shawe-Taylor J (2000) An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press, Cambridge
- Cruden DM (1991) A simple definition of a landslide. *Bull Eng Geol Env* 43:27–29
- Dahal RK, Hasegawa S, Nonomura A, Yamanaka M, Masuda T, Nishino K (2008) GIS-based weights-of-evidence modelling of rainfall-induced landslides in small catchments for landslide susceptibility mapping. *Environ Geol* 54:311–324
- Dai F, Lee C, Ngai YY (2002) Landslide risk assessment and management: an overview. *Eng Geol* 64:65–87
- Das I, Stein A, Kerle N, Dadhwal VK (2012) Landslide susceptibility mapping along road corridors in the Indian Himalayas using Bayesian logistic regression models. *Geomorphology* 179:116–125
- Debeljak M, Dzeroski S (2009) Applications of data mining in ecological modelling. *Handbook of ecological modelling and informatics*. WIT Press, Southampton, pp 409–423
- Doetsch P, Buck C, Golik P, Hoppe N, Kramp M, Laudenberg J, Oberdörfer C, Steingrube P, Forster J, Mauser A (2009) Logistic model trees with auc split criterion for the kdd cup 2009 small challenge. In: *Proceedings of the 2009 International Conference on KDD-Cup 2009-volume 7*. JMLR org pp 77–88
- Guzzetti F, Carrara A, Cardinali M, Reichenbach P (1999) Landslide hazard evaluation: a review of current techniques and their application in a multi-scale study, Central Italy. *Geomorphology* 31:181–216
- Han J, Pei J, Kamber M (2011) *Data mining: concepts and techniques*. Elsevier, Amsterdam
- Hastie T, Tibshirani R, Friedman J (2002) *The elements of statistical learning: data mining, inference, and prediction*. biometrics. Springer, Berlin
- He Q, Shahabi H, Shirzadi A, Li S, Chen W, Wang N, Chai H, Bian H, Ma J, Chen Y (2019) Landslide spatial modelling using novel bivariate statistical based Naïve Bayes, RBF Classifier, and RBF network machine learning algorithms. *Sci Total Environ* 663:1–15
- Hong H, Pradhan B, Xu C, Bui DT (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
- Hong H, Pourghasemi HR, Pourtaghi ZS (2016) Landslide susceptibility assessment in Lianhua county (China): a comparison between a random forest data mining technique and bivariate and multivariate statistical models. *Geomorphology* 259:105–118
- Hong H, Chen W, Xu C, Youssef AM, Pradhan B, Tien Bui D (2017a) Rainfall-induced landslide susceptibility assessment at the Chongren area (China) using frequency ratio, certainty factor, and index of entropy. *Geocarto Int* 32:139–154
- Hong H, Ilia I, Tsangaratos P, Chen W, Xu C (2017b) A hybrid fuzzy weight of evidence method in landslide susceptibility analysis on the Wuyuan area, China. *Geomorphology* 290:1–16
- Hong H, Liu J, Zhu A-X, Shahabi H, Pham BT, Chen W, Pradhan B, Bui DT (2017c) A novel hybrid integration model using support vector machines and random subspace for weather-triggered landslide susceptibility assessment in the Wuning area (China). *Environ Earth Sci* 76:652
- Hong H, Liu J, Bui DT, Pradhan B, Acharya TD, Pham BT, Zhu A-X, Chen W, Ahmad BB (2018a) Landslide susceptibility mapping using J48 Decision tree with AdaBoost, bagging and rotation forest ensembles in the Guangchang area (China). *CATENA* 163:399–413
- Hong H, Panahi M, Shirzadi A, Ma T, Liu J, Zhu A-X, Chen W, Kougiass I, Kazakis N (2018b) Flood susceptibility assessment in Hengfeng area coupling adaptive neuro-fuzzy inference system with genetic algorithm and differential evolution. *Sci Total Environ* 621:1124–1141
- Huang Y, Zhao L (2018) Review on landslide susceptibility mapping using support vector machines. *CATENA* 165:520–529
- Jiménez-Perálvarez J, Irigaray C, El Hamdouni R, Chacón J (2011) Landslide-susceptibility mapping in a semi-arid mountain environment: an example from the southern slopes of Sierra Nevada (Granada, Spain). *Bull Eng Geol Env* 70:265–277
- 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
- Khosravi K, Pham BT, Chapi K, Shirzadi A, Shahabi H, Revhaug I, Prakash I, Bui DT (2018) A comparative assessment of decision trees algorithms for flash flood susceptibility modeling at Haraz watershed, northern Iran. *Sci Total Environ* 627:744–755
- Landwehr N, Hall M, Frank E (2005) Logistic model trees. *Mach Learn* 59:161–205
- Miraki S, Zanganeh SH, Chapi K, Singh VP, Shirzadi A, Shahabi H, Pham BT (2018) Mapping groundwater potential using a novel hybrid intelligence approach. *Water Res Manag*. <https://doi.org/10.1007/s11269-018-2102-6>
- Mousavi SZ, Kavian A, Soleimani K, Mousavi SR, Shirzadi A (2011) GIS-based spatial prediction of landslide susceptibility using logistic regression model. *Geomat Nat Hazards Risk* 2:33–50
- Murthy SK (1998) Automatic construction of decision trees from data: a multi-disciplinary survey. *Data Min Knowl Discov* 2:345–389
- Myles AJ, Feudale RN, Liu Y, Woody NA, Brown SD (2004) An introduction to decision tree modeling. *J Chemom* 18:275–285
- Nguyen VV, Pham BT, Vu BT, Prakash I, Jha S, Shahabi H, Shirzadi A, Ba DN, Kumar R, Chatterjee JM (2019) Hybrid machine learning approaches for landslide susceptibility modeling. *Forests* 10:157
- Nithya N, Duraiswamy K (2014) Gain ratio based fuzzy weighted association rule mining classifier for medical diagnostic interface. *Sadhana* 39:39–52
- Peng L, Niu R, Huang B, Wu X, Zhao Y, Ye R (2014) Landslide susceptibility mapping based on rough set theory and support vector machines: a case of the three Gorges area, China. *Geomorphology* 204:287–301



- Pham BT, Tien Bui D, Indra P, Dholakia M (2015a) A comparison study of predictive ability of support vector machines and naive bayes tree methods in landslide susceptibility assessment at an area between Tehri Garhwal and Pauri Garhwal, Uttarakhand state (India) using GIS In: National Symposium on Geomatics for Digital India and annual conventions of ISG and ISRS, Jaipur (India).
- Pham BT, Tien Bui D, Indra P, Dholakia M (2015b) Landslide susceptibility assessment at a part of Uttarakhand Himalaya, India using GIS-based statistical approach of frequency ratio method. *Int J Eng Res Technol* 4:338–344
- Pham BT, Pradhan B, Bui DT, Prakash I, Dholakia M (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
- Pham BT, Bui DT, Pourghasemi HR, Indra P, Dholakia M (2017) Landslide susceptibility assessment in the Uttarakhand area (India) using GIS: a comparison study of prediction capability of naïve bayes, multilayer perceptron neural networks, and functional trees methods. *Theor Appl Climatol* 128:255–273
- Pham BT, Shirzadi A, Bui DT, Prakash I, Dholakia M (2018a) A hybrid machine learning ensemble approach based on a radial basis function neural network and Rotation Forest for landslide susceptibility modeling: a case study in the Himalayan area, India. *Int J Sedim Res* 33:157–170
- Pham BT, Prakash I, Dou J, Singh SK, Trinh PT, Trung Tran H, Le Minh T, Tran VP, Kim Khoi D, Shirzadi A (2018b) A novel hybrid approach of landslide susceptibility modeling using rotation forest ensemble and different base classifiers. *Geocarto Int*. <https://doi.org/10.1080/10106049.2018.1559885>
- Pham BT, Prakash I, Singh SK, Shirzadi A, Shahabi H, Bui DT (2019) Landslide susceptibility modeling using reduced error pruning trees and different ensemble techniques: hybrid machine learning approaches. *CATENA* 175:203–218
- Pourghasemi HR, Jirandeh AG, Pradhan B, Xu C, Gokceoglu C (2013) Landslide susceptibility mapping using support vector machine and GIS at the Golestan Province, Iran. *J Earth Syst Sci* 122:349–369
- Pradhan B (2011) Use of GIS-based fuzzy logic relations and its cross application to produce landslide susceptibility maps in three test areas in Malaysia. *Environ Earth Sci* 63:329–349
- 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
- Quinlan JR (1993) C4. 5: Programming for machine learning. Morgan Kaufmann, Burlington
- Quinlan JR (1996) Bagging, boosting, and C4. 5. *AAAI/IAAI* 1:725–730
- Regmi AD, Devkota KC, Yoshida K, Pradhan B, Pourghasemi HR, Kumamoto T, Akgun A (2014) Application of frequency ratio, statistical index, and weights-of-evidence models and their comparison in landslide susceptibility mapping in Central Nepal Himalaya. *Arab J Geosci* 7:725–742
- Schölkopf B, Smola AJ, Williamson RC, Bartlett PL (2000) New support vector algorithms. *Neural Comput* 12:1207–1245
- Shafizadeh-Moghadam H, Valavi R, Shahabi H, Chapi K, Shirzadi A (2018) Novel forecasting approaches using combination of machine learning and statistical models for flood susceptibility mapping. *J Environ Manag* 217:1–11
- Shirzadi A, Saro L, Joo OH, Chapi K (2012) A GIS-based logistic regression model in rock-fall susceptibility mapping along a mountainous road: Salavat Abad case study, Kurdistan, Iran. *Nat Hazards* 64:1639–1656
- Shirzadi A, Bui DT, Pham BT, Solaimani K, Chapi K, Kavian A, Shahabi H, Revhaug I (2017a) Shallow landslide susceptibility assessment using a novel hybrid intelligence approach. *Environ Earth Sci* 76:60
- Shirzadi A, Chapi K, Shahabi H, Solaimani K, Kavian A, Ahmad BB (2017b) Rock fall susceptibility assessment along a mountainous road: an evaluation of bivariate statistic, analytical hierarchy process and frequency ratio. *Environ Earth Sci* 76:152
- Shirzadi A, Shahabi H, Chapi K, Bui DT, Pham BT, Shahedi K, Ahmad BB (2017c) A comparative study between popular statistical and machine learning methods for simulating volume of landslides. *CATENA* 157:213–226
- Shirzadi A, Solaimani K, Habibnejhad M, Kavian A, Chapi K, Shahabi H, Chen W, Khosravi K, Pham B, Pradhan BT (2018) Novel GIS based machine learning algorithms for shallow landslide susceptibility mapping. *Sensors* 18:3777
- Shirzadi A, Solaimani K, Roshan MH, Kavian A, Chapi K, Shahabi H, Keesstra S, Ahmad BB, Bui DT (2019) Uncertainties of prediction accuracy in shallow landslide modeling: sample size and raster resolution. *CATENA* 178:172–188
- Tien Bui D, Pradhan B, Lofman O, Revhaug I (2012) Landslide susceptibility assessment in Vietnam using support vector machines, decision tree, and Naive Bayes models. *Math Probl Eng*. <https://doi.org/10.1155/2012/974638>
- Tien Bui D, Khosravi K, Li S, Shahabi H, Panahi M, Singh V, Chapi K, Shirzadi A, Panahi S, Chen W (2018a) New hybrids of anfis with several optimization algorithms for flood susceptibility modeling. *Water* 10:1210
- Tien Bui D, Shahabi H, Shirzadi A, Chapi K, Alizadeh M, Chen W, Mohammadi A, Ahmad B, Panahi M, Hong H (2018b) Landslide detection and susceptibility mapping by AIRSAR data using support vector machine and index of entropy models in Cameron Highlands, Malaysia. *Remote Sens* 10:1527
- Tien Bui D, Shahabi H, Shirzadi A, Chapi K, Hoang N-D, Pham B, Bui Q-T, Tran C-T, Panahi M, Bin Ahmad B (2018c) A novel integrated approach of relevance vector machine optimized by imperialist competitive algorithm for spatial modeling of shallow landslides. *Remote Sens* 10:1538
- Tien Bui D, Shahabi H, Shirzadi A, Chapi K, Pradhan B, Chen W, Khosravi K, Panahi M, Bin Ahmad B, Saro L (2018d) Land subsidence susceptibility mapping in South Korea using machine learning algorithms. *Sensors* 18:2464
- Tsangaratos P, Ilia I (2016) Landslide susceptibility mapping using a modified decision tree classifier in the Xanthi Perfection, Greece. *Landslides* 13:305–320
- Vapnik V (1998) *Statistical learning theory*. Wiley, New York
- Xu C, Xu X, Lee YH, Tan X, Yu G, Dai F (2012) The 2010 Yushu earthquake triggered landslide hazard mapping using GIS and weight of evidence modeling. *Environ Earth Sci* 66:1603–1616
- Yao X, Tham L, Dai F (2008) Landslide susceptibility mapping based on support vector machine: a case study on natural slopes of Hong Kong, China. *Geomorphology* 101:572–582
- Yesilnacar E, Topal T (2005) Landslide susceptibility mapping: a comparison of logistic regression and neural networks methods in a medium scale study, Hendek region (Turkey). *Eng Geol* 79:251–266
- Zhang G, Cai Y, Zheng Z, Zhen J, Liu Y, Huang K (2016) Integration of the statistical index method and the analytic hierarchy process technique for the assessment of landslide susceptibility in Huizhou, China. *Catena* 142:233–244
- Zhao Y, Zhang Y (2008) Comparison of decision tree methods for finding active objects. *Adv Space Res* 41:1955–1959