Assessment of the effects of training data selection on the landslide susceptibility mapping: a comparison between support vector machine (SVM), logistic regression (LR) and artificial neural networks (ANN)

Bahareh Kalantara, Biswajeet Pradhan a,b, Seyed Amir Naghibic, Alireza Motevalli c and Shattori Mansora

aDepartment of Civil Engineering, Faculty of Engineering, University Putra Malaysia, Serdang, Malaysia; bSchool of Systems, Management and Leadership, Faculty of Engineering and Information Technology, University of Technology Sydney, New South Wales, Australia; cDepartment of Watershed Management Engineering, College of Natural Resources and Marine Sciences, Tarbiat Modares University, Noor, Iran

ABSTRACT
Landslide is a natural hazard that results in many economic damages and human losses every year. Numerous researchers have studied landslide susceptibility mapping (LSM), each attempting to improve the accuracy of the final output. However, few studies have been published on the training data selection effects on the LSM. Thus, this study assesses the training landslide random selection effects on support vector machine (SVM) accuracy, logistic regression (LR) and artificial neural networks (ANN) models for LSM in a catchment at the Dodangeh watershed, Mazandaran province, Iran. A 160 landslide locations inventory was collected by Geological Survey of Iran for this investigation. Different methods were implemented to define the landslide locations, such as inventory reports, satellite images and field survey. Moreover, 14 landslide conditioning factors were considered in the analysis of landslide susceptibility. These factors include curvature, plan curvature, profile curvature, altitude, slope angle, slope aspect, distance to faults, distance to stream, topographic wetness index, stream power index, terrain roughness index, sediment transport index, lithology and land use. The results show that the random landslide training data selection affected the parameter estimations of the SVM, LR and ANN algorithms. The results also show that the training samples selection had an effect on the accuracy of the susceptibility model because landslide conditioning factors vary according to the geographic locations in the study area. The LR model was found to be less sensitive than the SVM and ANN models to the training samples selection. Validation results showed that SVM and LR models outperformed the ANN model for all scenarios. The average overall accuracy of LR, SVM and ANN models are 81.42%, 79.82% and 70.2%, respectively.

KEYWORDS
Landslide; SVM; ANN; LR; remote sensing; training data; GIS

1. Introduction
Landslide results in significant economic damage and human losses annually. This phenomenon is described as a ground range wide movement, which includes rock falls, deep slope failures and
shallow debris flow (Varnes 1978). A quick look at the economic damage and human losses would be enough to encourage managers and planners to operate inclusive projects on landslide. Modelling landslide through appropriate models could provide reliable landslide susceptibility maps, which could offer appropriate information to managers and land-use planners for development plans and land-use management.

The number of landslide locations in the training data-set and sampling strategy are important factors that significantly affect the quality of landslide susceptibility maps. Regarding the number of samples to be used in training a susceptibility model, although this could depend on the availability of landslide inventories in the given area, the training and testing data-sets selection can affect the model prediction capability. Dhakal et al. (2000) denoted that different landslide and non-landslide group samples generated randomly have effects on the critical factors identification, such as geology and model prediction capability. They found that stratified random sampling method achieved better accuracy than random sampling without stratification. In another paper, Remondo et al. (2003) discussed random sampling strategies as being more suitable and logical than study area spatial split, and that spatial validation is more logical than temporal validation. They argued this fact by mentioning that the entire landslide inventory (analysis and validation sets) corresponded to the same period and conditions. Temporal validation compared analysis population with population corresponding to a later period. In addition, Hjort and Marmion (2008) analysed the sample size effects on the accuracy of geomorphological model. They analysed different sample size groups ranging from 20 to 600 samples based on generalized linear model, generalized additive model, generalized boosting method and artificial neural network (ANN) in two data settings, i.e. independent and split-sample approaches. They found that accuracy increases with the increase in sample numbers in all models, and the robust predictions level was reached with 200 observations. Overall, they suggested that a few hundred observations are needed in geomorphological modelling. Heckmann et al. (2014) investigated the sample size effects on logistic regression (LR) model accuracy for predicting debris flow spatial distribution. They discussed two main problems of having inadequate sample size for training landslide susceptibility models. First, a sample size has to be large enough so it can cover the landslide conditioning factors variability within the study area. Second, the sample size must not be too large, because this will likely violate the independent observations assumption because of spatial autocorrelation.

Sampling strategy was also found to have effects on the landslide susceptibility models both in terms of prediction accuracy and generalization capability. Three sampling strategy main types are identified in the literature such as scarp, seed cell, and point. Yilmaz (2010) suggested that the scarp-sampling method could achieve higher accuracy than other sampling methods for ANN-based landslide susceptibility mapping (LSM). Paulin et al. (2010) analysed the pixel size effects on the cartographic representation of the different landslide types. In general, they found that as pixel size increased, the landslide lost cartographic representation. However, they also suggested that different models could achieve results similar to other models but under different pixel sizes. Baeza et al. (2010) studied the effect of sample size and type on the accuracy of landslide susceptibility map. Their results showed that using at least half of the samples for training a discriminant analysis is necessary. In addition, they found that sample type had no significant influence on model performance. More recently, Hussin et al. (2016) developed landslide susceptibility models with different sampling strategies (i.e. scarp centroid, points populating the scarp, entire scarp polygon). They showed that highest success rates were obtained when sampling shallow landslides as 50-m grid-points, and debris flow scarps as polygons.

In the past decades, different models have been applied for LSM in many countries and climatic conditions. These models could be classified into several categories, including statistical, expert-based and data mining methods. In different studies, frequency ratio (Lee and Pradhan 2007; Pradhan and Lee 2010a; Nourani et al. 2014; Shahabi et al. 2014; Umar et al. 2014; Hong et al. 2016b), weights of evidence (Pourghasemi et al. 2012a, 2012b), ANN (Pradhan and Lee 2010b; Zare et al. 2013; Conforti et al. 2014; Nourani et al. 2014), decision tree (Tien Bui et al. 2012; Pradhan
evian, evidential belief function (Hong et al. 2016b), LR (Regmi et al. 2014; Pradhan and Lee 2010a; Marjanović et al. 2011; Devkota et al. 2013; Pourghasemi et al. 2013a; Nourani et al. 2014; Shahabi et al. 2014; Umar et al. 2014; Hong et al. 2016b), maximum entropy (Constantin et al. 2011; Pourghasemi et al. 2012b) and support vector machine (SVM) (Marjanović et al. 2011; Ballabio and Sterlacchini 2012; Pourghasemi et al. 2013a; Peng et al. 2014; Hong et al. 2016b) have been used. More recently, some researchers used data mining algorithms including boosted regression tree, classification and regression tree, and random forest models (Miner et al. 2010; Stumpf and Kerle 2011; Vorpahl et al. 2012; Catani et al. 2013; Micheletti et al. 2014; Youssef 2015) have been implemented for landslide susceptibility modelling.

The literature review revealed that several studies have employed different models for LSM, but landslide inventory spatial variation has been less considered frequently. With respect to the spatial variations of landslide conditioning factors, the differences between the landslide samples quality and the landslide locations spatial distribution, training data selection is expected to have effects on model performance. Therefore, this study assesses the effects of random training landslide selection on the accuracy of the SVM, LR and ANN models. The models were selected for this study based on previous studies (Pradhan and Lee 2010a; Ballabio and Sterlacchini 2012; Devkota et al. 2013; Pourghasemi et al. 2013a; Nourani et al. 2014; Shahabi et al. 2014; Umar et al. 2014) that have used these models successfully in different areas for LSM. To carry out this task, five training and validation data-set scenarios were generated randomly. Then, SVM, LR and ANN models were trained and built for each scenario to evaluate the training sample effects in spatial distribution terms. The number of samples was kept at 70% of the whole landslide inventory data.

2. Study area

The study area is situated in Dodangeh watershed, Sari County, Mazandaran Province, Iran. It is approximately at the zone of 36° 2’ 44.56” N latitude and 53° 14’ 34.78” E longitude, and covers an approximate area of 1754 km² (Figure 1) with a population of 8469 (2012 census). The highest area temperature ranges between 29°C and 31°C between June and August, and the mean precipitation varies between 35 and 105 (mm/month). According to Emberger climagram, climate in the area is cold and wet (altitude of 800–2000 meters) and cold and semi-arid (altitude of 2000–2800 meters). The study area is covered mostly by vegetation including medium density forest, mixed forest, mixed agriculture, orchard and mixed orchard. The study area is covered with various types of

Figure 1. Study area (left image) landslide inventory map and (right image) Iran states.
lithological formations including Tertiary, Cretaceous and Quaternary types of formations. The M23msl (Marl, calcareous sandstone and siltstone, silty marl/sandy limestone, mudstone) and PIQcs (Conglomerate, sandstone, silt stone, silty marl) covers about 74.9% of the study area. The Quaternary deposit covers 5.11% of the study area and include Qal (Recent loose alluvium in the river channels), Q2 (Young alluvial fans and terraces, river terraces; mainly cultivated), Qsw (Swamp).

3. Data used

The landslide inventory data are considered the most vital factor in foreseeing the future incidents (Tien Bui et al. 2012). For this investigation, 160 landslide locations were detected in the Dodangeh watershed. Different methods were implemented to define the landslide locations such as inventory reports, satellite images and field surveys. Most landslides are shallow rotational with a few being translational. Fourteen conditioning factors were considered in the analysis of landslide susceptibility based on the literature review and data availability (Sangchini et al. 2016; Pradhan et al. 2017). These factors include curvature, plan curvature, profile curvature, altitude, slope angle, slope aspect, distance to faults, distance to stream, topographic wetness index (TWI), stream power index (SPI), terrain roughness index (TRI), sediment transport index (STI), lithology and land use. These factors were transformed into a vector spatial database by employing ArcMap 10.3. A 10-m digital elevation model (DEM) was exploited from a 1:25,000-scale topographic map. Employing the mentioned layer, the slope degree, slope aspect, altitude, curvature, SPI, TWI, TRI and STI were generated.

3.1. Slope

Slope has considerable influence on slope stability, which can be explained as the form between any section of the surface and a horizontal datum. Many studies have shown that slope is a main conditioning factor in LSM (Dai and Lee 2002; Ayalew et al. 2004; Gómez and Kavzoglu 2005). Slope angle was produced employing the 10-m DEM. This layer was classified into (1) $0^\circ - 8^\circ$, (2) $8^\circ - 15^\circ$, (3) $15^\circ - 22^\circ$, (4) $22^\circ - 30^\circ$ and (5) $30^\circ - 66^\circ$ (Figure 2(a)).

3.2. Aspect

Aspect related factors including exposure to sunlight, winds and precipitation are vital factors in instigating landslides (Dai et al. 2001). In this work, DEM was employed to compute the aspect classes. In the next stage, aspect was categorized into nine groups of eight directions and flat (Figure 2(b)).

3.3. Altitude

Altitude is affected by several geomorphologic and geological processes, and thus, it is an influential conditioning factor in landslide analyses (Gritzner et al. 2001; Dai and Lee 2002; Ayalew and Yamagishi 2005). In this investigation, altitude was prepared using the Dodangeh watershed 10-m DEM, where the watershed altitude varied between 210 and 1240 m (Figure 2(c)).

3.4. Curvature (plan and profile)

Plan curvature is clarified as the counter line curvature made by a horizontal plane crossing with the surface. Water convergence and divergence during the flow, which affect slope erosion processes, can be determined using plan curvature (Ercanoglu and Gokceoglu 2002; Oh and Pradhan 2011). This layer was generated by the ArcGIS spatial analysis tool (Figure 2(e)). Plan curvature represents the flow acceleration and erosion/deposition rate. Profile curvature affects the flow velocity variation
Figure 2. Landslide conditioning factor maps: (a) slope, (b) aspect, (c) altitude, (d) curvature, (e) plan curvature, (f) profile curvature, (g) sediment transport index (STI), (h) topographic wetness index (TWI), (i) TRI, (j) stream power index (SPI), (k) distance to fault, (l) distance to stream, (m) land use, (n) lithology.
Figure 2. (Continued)
down the slope (Talebi et al. 2007). These layers were generated by the ArcGIS spatial analysis tool (Figure 2(f)).

### 3.5. Sediment transport index (STI)

STI represents the slope failure and deposition procedures (Moore and Wilson 1992). There is a direct relationship between the STI and the water accumulation at the bottom of the catchment as well as the erosion amount (Pourghasemi et al. 2013a). STI can be calculated as below

\[
STI = \left( \frac{A_s}{22.1} \right)^{0.6} \left( \frac{\sin \beta}{0.0896} \right)^{1.3},
\]

where \( \beta \) denotes the slope at a certain pixel and \( A_s \) refers to the upstream area. In STI (0–550) five classes were constructed using quantile algorithm for analysis as shown in Figure 2(g).

### 3.6. Topographic wetness index (TWI)

This factor was developed in a rainfall-runoff model called TOPMODEL (Beven and Kirkby 1979). At any watershed location, TWI is described by the flow accumulation theoretical measure, where soil moisture affects the slope material, which results in pore water pressure and diminishes soil stability (Gómez and Kavzoglu 2005; Yilmaz 2009), and could be regarded as an influential factor in landslide studies. Thus, TWI has been employed in many landslide studies (Gökceoglu and Aksoy 1996; Yilmaz 2009; Hong et al. 2016a), and can be computed as

\[
TWI = \log_e \left( \frac{A}{b \tan \beta} \right),
\]

where \( A \) denotes flow accumulation in square meters, \( b \) refers to the pixel width through which water flows in meters and \( \beta \) (radian) represents the slope (Beven and Kirkby 1979). In this study, TWI was reclassified into five categories namely, (1) –0.83–2.1, (2) 2.2–2.9, (3) 3–3.7, (4) 3.8–4.8 and (5) 4.9–17 using quantile algorithm (Figure 2(h)).
3.7. **Terrain roughness index (TRI)**

TRI shows the concave and convex upward slopes (Althuwaynee et al. 2014), and was calculated as follows:

\[ TRI = \sqrt{|x| \left(max^2 - min^2\right)}, \]  

(3)

where max denotes the pixel largest value in nine rectangular altitude neighbourhoods, and min denotes the minimum value (Riley 1999). TRI was categorized into five classes employing the quantile algorithm: (1) 0–56, (2) 57–85, (3) 86–110, (4) 120–150 and (5) 160–250 (Figure 2(i)).

3.8. **Stream power index (SPI)**

SPI represents flowing water erosion power according to the presumption which relates discharge to definite catchment area. This factor was determined by the formula introduced by Moore and Wilson (1992)

\[ SPI = \frac{A \tan \beta}{b}, \]  

(4)

where \(A\) is the specific catchment area and \(\beta\) is the local slope gradient computed in degrees. In this study, SPI was reclassified into five categories namely, (1) 0–219, (2) 220–985, (3) 986–2520, (4) 2530–6020 and (5) 6030–14 000 (Figure 2(j)) using quantile algorithm.

3.9. **Buffer zones (distance to fault and stream)**

The landslides occurred mainly along the fault and decreased sharply with distance from it. A buffer map related to the seismogenic fault with a 5-km buffer zone interval is shown in Figure 2(k). Moreover, the distance to the stream was derived in two steps. First, the streams were delineated using flow accumulation and converted into a vector format. In the second step, a Euclidean distance analysis was conducted to produce the distance to the stream factor. In addition, the distance to fault was derived using the same distance analysis and existing topographic maps of the study area (Figure 2(k,l)).

3.10. **Land use**

Land use is affected by human activities and alterations in the environment. Its importance in LSM has been referred to by several publications (e.g. Restrepo et al. 2003). In this research, the Landsat Enhanced Thematic Mapper (2006) image was used to produce land-use layer by employing a supervised classification algorithm with an accuracy of 86%. These classes include agriculture, dry farming, high density forest, medium density forest, mixed forest/orchard, mixed agriculture/orchard, orchard, mixed orchard/agriculture, sandy/dune and urban (Figure 2(m), Table 1).

| Code | Land use               |
|------|------------------------|
| I1   | Agriculture            |
| DF   | Dry farming            |
| F1   | High density forest    |
| F2   | Medium density forest  |
| FO   | Mixed forest/orchard   |
| IO   | Mixed agriculture/orchard |
| O    | Orchard                |
| OI   | Mixed Orchard/agriculture |
| SD   | Sandy/dune            |
| U1   | Urban                  |
3.11. Lithology

Lithology can be regarded as one of the most vital conditioning factors in LSM because strength and the rocks and soils permeability are influenced directly by lithological characteristics (Kavzoglu et al. 2014). Thus, several studies have considered this factor in their analyses (Dai and Lee 2002; Ayalew and Yamagishi 2005). The lithology map (Figure 2(n)) was generated by Geological Survey of Iran with 1:100,000 scale (Table 2).

### Table 2. Types of geological formation of the study area.

| Code | Class | Formation | Lithology | Geological age |
|------|-------|-----------|-----------|---------------|
| Mms  | A     | Continental series | Grey-red marls; thin-medium bedded sandstone & calcareous sandstone intercalations | Tertiary |
| K2ml | B     | Upper red formation equivalent | Marl, silty marl, calcareous marl, marly limestone, limestone | Cretaceous |
| M23msl | C    | Upper red formation equivalent | Marl, calcareous sandstone and siltstone, silty marl, sandy limestone, mudstone | Tertiary |
| Mm1  | D     | Qom formation equivalent | Homogeneous ochre coloured marls | Tertiary |
| Mmg  | E     | – | Undivided Mg and Mm1 rock units | Tertiary |
| OM1mg | F    | Qom formation equivalent | Marl, silty marl, gypsum, sandy limestone, locally minor andesitic lava | Tertiary |
| Pel  | G     | – | Medium-thick bedded cream-grey limestone & sandy limestone; locally thin marl layers | Tertiary |
| Pelm | H     | – | Alternations of grey limestone and marble | Tertiary |
| Pemls | I    | – | Silty marl, marl, sandy-marly limestone, calcareous sandstone, conglomerate | Tertiary |
| Pes  | J     | – | Light brown sandstone, thin marl layers | Tertiary |
| PIQcs | K    | Continental series | Conglomerate, sandstone, silt stone, silty marl | Tertiary |
| Q2   | L     | – | Young alluvial fans and terraces, river terraces; mainly cultivated | Quaternary |
| Qal  | M     | – | Recent loose alluvium in the river channels | Quaternary |
| Qsw  | N     | – | Swamp | Quaternary |
| Residential area | O | | | |

4. Methods for landslide susceptibility zonation

In the current work, LSMs were produced employing GIS-based SVM, LR and ANN models. The procedure applied was completed in three steps. First, five landslide inventory random scenarios were generated (Figure 3). Then, three approaches (i.e. SVM, LR and ANN) were used to model the correlation between landslide conditioning factors and landslide incident in each scenario using Weka 3.8 software. Finally, the accuracy of all approaches in different scenarios was evaluated by the ROC curve and confusion matrix to assess the sample selection effect on the model accuracy. Figure 4 shows the proposed method general methodology.

4.1. Support vector machine (SVM)

SVM has been employed widely in different classification and regression problems because of its effectiveness in working with linearly non-separable and high dimensional data-sets (Kavzoglu et al. 2014; Mountrakis et al. 2011). Statistical learning theory can be considered as the basic theory of SVM (Cortes and Vapnik 1995). Ballabio and Sterlacchini (2012) mentioned two kinds of errors in landslide modelling. The first is data linearity, which is an assumption in classic models. The second error is model overfitting (Ballabio and Sterlacchini 2012), which reduces model generality and its applicability in larger regions. These two errors can reduce model performance and subsequently
LSM accuracy. The SVM reduces both the error test and model complexity. In addition, some non-linear kernel types in this model which could resolve the first mentioned error in landslide modelling. SVM performance is affected by kernel function, which could be one of the linear, radial basis function, sigmoid and polynomial (Bui et al. 2012; Naghibi 2017). The SVM classifier attempts to
find an optimal separation hyper plane that could differentiate the two classes (Hong et al. 2016b). In case of linear separable, the hyper plane could be determined by

\[
y_i (w^T x_i + b) \geq 1 - \delta_i,
\]

where \( w \) refers to the coefficient vector that determines the hyper plane orientation in the feature space, \( b \) represents the offset of the hyper plane from the origin and \( \delta_i \) is the positive slack variables (Cortes and Vapnik 1995). The optimal hyper plane definition resolves an optimization problem as

\[
\text{Minimize } \sum_{i=1}^{n} a_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j y_i y_j (x_i, x_j),
\]

\[
\text{Subject } \sum_{i=1}^{n} a_i y_i = 0, \quad 0 \leq a_i \leq C
\]

where \( a_i \) is the Lagrange multipliers and \( C \) refers to the penalty. In this work, the SVM model RBF kernels were implemented as follows (Pourghasemi et al. 2013b):

\[
\text{RBF : } K (x_i, y_i) = \left( -y \|X_i - X_j\| \right), \quad y > 0,
\]

where \( K(x_i, y_i) \) denotes the kernel function; \( y \) is the gamma term in the kernel function for RBF, sigmoid and polynomial kernels; \( d \) denotes the degree term; \( r \) shows the bias term; \( y, d \) and \( r \) are the parameters that should be optimized in the modelling process.

### 4.2. Logistic regression (LR)

LR is one of the most widely employed statistical models in LSR. It clarifies the relationship between a categorical variable and some dependent factors, which could be categorical, continuous or binary variables (Hong et al. 2016a). The benefit of the algorithm is that variables do not require normal
distribution (Pradhan and Lee 2010c). Independent variables in the LR could be designated as 0 and 1, denoting the landslide absence and presence. Model output varies between 0 and 1 and represents landslide susceptibility. The LR is based on the logistic function $P$, determined as

$$P = \frac{\exp(z)}{1 + \exp(z)},$$

(8)

where $P$ denotes the probability related to a certain observation, and $z$ could be defined as

$$Z = \beta_0 + \beta_1X_1 + \beta_2X_2 + \cdots + \beta_nX_n,$$

(9)

where $\beta_0$ denotes the intercept of the algorithm, $\beta_i$ is the coefficient representing the independent variables contribution $X_i$, and $n$ denotes the number of conditioning factors (Atkinson and Massari 1998).

4.3. Artificial neural network (ANN)

The ANN is a computational procedure with capability to acquire, present, and calculate mapping from data multivariate space to another (Garrett 1994). The goal of the ANN model is to build a method for predicting outputs from input factors that have not been used in the modelling process (Lee et al. 2003). The most prominently employed neural network method is the back-propagation learning algorithm (Lee et al. 2003). The standard ANN model comprises three layers, namely, input layer (i.e. landslide conditioning factors), hidden layers and output layer (i.e. landslide susceptibility). The ANN determines a certain weight for each input factor and after multiplying, sums the product, and employs a non-linear transfer function to build results (Lee et al. 2003).

This study employed two neural networks architectures namely, the multilayer perception (MLP) and radial basis function (RBF), where both are vigour classifiers (Yilmaz and Kaynar 2011). The main difference between RBF and MLP is that MLP is more general, while RBF is considered as more localist learning and is associated with input data (Yilmaz and Kaynar 2011). The MLP can separate non-linear data (Lee et al. 2003), exploit helpful information from imprecise data, and produce acceptable results even when the inputs are not perfect (Ermini et al. 2005; Kanungo et al. 2006).

5. Validation of the susceptibility map

The quality of a landslide susceptibility model is usually verified by using landslide inventory data that is not used during building the model. Of the 160 landslides identified, 112 ($\approx$70%) locations were used for the landslide susceptibility maps, while the remaining 48 ($\approx$30%) cases were used for the model validation. The landslide susceptibility analysis was performed using both confusion matrix and receiver operating characteristics (ROC).

5.1. Confusion matrix

The confusion matrix includes information on model classification accuracy (Provost and Kohavi 1998), and provides useful information on model performance. The matrix columns display the instances in an estimated class while the rows show the instances in an actual class (Sammut and Webb 2011). This matrix helps experts to determine whether the system is misidentifying the two classes (i.e. mislabelling one as the other) or correctly classifying them (Stehman 1997). This matrix was also employed to evaluate the algorithm performance in other investigations (Naghibi et al. 2016).
5.2. Receiver operating characteristic (ROC)

The ROC provides a graphical plot in which the binary classifier method performance is determined. This curve has been used in many studies to evaluate the algorithm performance in LSM and groundwater potential mapping (Tien Bui et al. 2012; Ozdemir and Altural 2013; Hong et al. 2016a; Naghibi et al. 2017), where the true positive rate is plotted against the false positive rate at several threshold settings (Huang et al. 2002; Naghibi and Moradi Dashtpagerdi 2016; Mousavi et al. 2017). The area under this curve determines the employed model accuracy in LSM. The area under curve changes from 0.5 to 1, and a 1 value indicates perfect performance, while a 0.5 value denotes weak model performance. In this study, to apply the ROC curve in this research, 47 non-landslides were generated by a random algorithm and then employed to calculate area under the curve using the ArcGIS software.

6. Results and discussion

In this study, five training data selection scenarios were generated and analysed randomly to evaluate their effects on the accuracy of SVM, LR and ANN models for LSM. Subsequently, the landslide susceptibility maps (LSMs) were classified into five classes (i.e. very low, low, moderate, high and very high) using the quantile classification scheme method (Hong et al. 2016a, 2016b). Figure 5 shows the maps produced by the SVM model for each scenario. In general, by looking at the spatial distribution of the landslide susceptibility zone, significant training data selection effects can be observed. The SVM model fits the best hyperplane that can separate landslides from non-landslides effectively, although some difficulty could be encountered in fitting the hyperplane when the predictors (landslide factors) are not separable, indicating that the training sample selection can have

Figure 5. Landslide susceptibility using SVM (a) Scenario 1, (b) Scenario 2, (c) Scenario 3, (d) Scenario 4, (e) Scenario 5.
effects on the SVM accuracy because landslide conditioning factors vary according to the geographic locations in the study area. SVM also works well with high dimensional data, and in the current study, 14 landslide conditioning factors were used. Categorical factors such as land use and lithology can have lesser effects on SVM accuracy. The reason is that when selecting different landslide training data subsets, the number of selected landslides may not vary from one class to another. However, in continuous factors such as slope and altitude, more effect can be observed because the values vary continuously.

In contrast, the LR model was found to be less sensitive to training data selection. Figure 6 shows that results of the LR model are similar in terms of spatial distribution, but are different from those generated by the SVM model. The distribution of the susceptibility zones produced by LR varies in the study area and those produced by the SVM. Both methods attempted to fit a linear equation to the landslide inventory data, and could generally place the landslide locations to the high and very highly susceptible classes. Similarly, the susceptibility maps produced by the ANN model had susceptibility zone spatial distributions nearly similar to those produced by the LR model (Figure 7).

The random landslide training data selection affected the parameter estimations of the SVM, LR and ANN algorithms. Table 3 shows the landslide conditioning factors estimated parameters in each scenario. In general, it can be observed that both coefficient value and positive/negative correlation of the factors with the dependent variable (landslide presence/absent) were affected. For example, the TWI factor coefficient in the first scenario was $-0.083$ and $0.028$ in the second scenario. Similarly, the TRI coefficient estimated by the LR varied from one scenario to another. The factor slope and altitude coefficients had consistent negative and positive signs. The results suggested that to estimate the importance of a factor, further analysis (such as optimization, parameter fine-tuning, new kernel function, etc.) is required, and cannot only rely on the estimated coefficients by the SVM, LR or ANN models. For this purpose, researchers often use the frequency ratio and other factor optimization methods, such as random forest and chi-square test. Different coefficient values
may produce similar accuracy, but the final outputs can be affected and the importance of the factors may not be reliable.

The three-model validation using the ROC curves and confusion matrix are shown in Tables 4 and 5. For all scenarios, the LR outperformed the SVM and ANN models. The average overall accuracy of LR, SVM and ANN models are 81.42%, 79.82% and 70.20%, respectively. The ROC curves showed that the LR algorithm average success rate is higher than that of SVM algorithm by ~6%, and the ANN model by 8%. In addition, the LR model prediction rate is higher than that of SVM and ANN models at an average of 10%. This quantitative assessment indicates that the LR and ANN models are less sensitive to training data selection than the SVM model in the study area. The advantages of the LR model include the final model output interpreted as a probability and not an exact prediction value, as in the case of other models. The LR has less parameters to fine tune, while the SVM model requires optimizing the kernel function, the penalty and gamma parameters. The ANN model requires finding the best network architecture that can predict the output classes accurately. In addition, the LR model training is more efficient than that of the SVM and ANN, especially in large-scale data.

Further model evaluation was made by producing landslide density graphs, as shown in Figure 8. The number of landslides in each susceptibility class was calculated and plotted in a two-dimensional chart, where its horizontal axis showed the susceptibility class and the number of landslides was presented in the vertical axis. The results showed that the LR model predicted the highest landslide percentage (42%) in the very high class. In addition, the SVM and ANN models predicted 13% and 38% of landslides in the very high susceptibility class.

Previous LSM works using the LR, SVM and ANN methods found several important findings related to the current work. Xu et al. (2012) showed that the training data and kernel function selections had significant effects on SVM modelling, and optimization is required for accurate LSM. The
| Table 3. Effect of random sampling on coefficient estimate. |
|----------------------------------------------------------|
| Scenario 1 | Scenario 2 | Scenario 3 | Scenario 4 | Scenario 5 |
| SVM | LR | ANN | SVM | LR | ANN | SVM | LR | ANN | SVM | LR | ANN | SVM | LR | ANN |
| TWI | -0.083 | 0.189 | 0.032 | 0.028 | 0.188 | 0.050 | 0.27 | -0.009 | 0.010 | 0.204 | 0.076 | 0.044 | -0.389 | 0.205 | 0.029 |
| TRI | 0.450 | -0.208 | -0.012 | 0.352 | 0.05 | -0.033 | 0.43 | -0.174 | -0.079 | 0.406 | -0.033 | -0.053 | 0.473 | -0.079 | -0.081 |
| Curvature | 0.172 | 0.015 | -0.007 | 0.254 | -0.006 | -0.035 | 0.503 | -0.261 | -0.039 | 0.234 | -0.192 | -0.018 | 0.084 | -0.045 | -0.003 |
| STI | 1.436 | -0.985 | -0.201 | 1.544 | -1.165 | -0.218 | 1.111 | -0.553 | -0.120 | 1.012 | -0.859 | -0.153 | 1.211 | -0.730 | -0.098 |
| SPI | -1.426 | 0.821 | 0.187 | -1.318 | 1.006 | 0.186 | -0.923 | 0.571 | 0.157 | -1.396 | 0.999 | 0.124 | -1.135 | 0.630 | 0.083 |
| Slope | -0.604 | 0.542 | -0.092 | -0.531 | 0.415 | 0.113 | -0.499 | 0.388 | 0.126 | -0.160 | 0.274 | 0.098 | -0.663 | 0.333 | 0.088 |
| Profile curvature | 0.392 | -0.304 | -0.047 | 0.308 | -0.297 | -0.029 | 0.385 | -0.304 | -0.052 | 0.466 | -0.394 | -0.028 | 0.255 | -0.207 | -0.016 |
| Plan curvature | 0.290 | -0.297 | -0.008 | 0.317 | -0.322 | 0.018 | 0.135 | 0.024 | 0.044 | 0.187 | -0.077 | 0.0327 | 264 | 0.304 | -0.167 | -0.009 |
| Lithology | 0.008 | -0.006 | 0.006 | -0.305 | 0.040 | 0.013 | -0.007 | -0.002 | 0.014 | -0.180 | 0.005 | 0.009 | 0.332 | -0.0223 | 0.003 |
| Land use | -2.920 | 0.699 | 0.091 | -3.142 | 0.661 | 0.102 | -3.173 | 0.691 | 0.093 | -3.308 | 0.679 | 0.068 | -2.971 | 0.6345 | 0.102 |
| Distance to stream | -0.242 | 0.002 | 0.0004 | -0.247 | 0.002 | 0.0042 | -0.151 | 0.001 | 0.0034 | 0.003 | 0.0009 | 0.0038 | -0.578 | 0.0016 | 0.000438 |
| Distance to fault | -0.240 | -0.0001 | -0.0001 | -0.133 | -0.0003 | -0.0010 | -0.352 | -0.0001 | -0.000174 | 0.236 | -0.0005 | -0.0001957 | -0.292 | -0.0002 | -0.0001 |
| DEM | 2.059 | -0.668 | -0.120 | 1.948 | -0.748 | -0.100 | 1.863 | -0.570 | -0.081 | 1.975 | -0.631 | -0.121 | 1.816 | -0.5684 | -0.077 |
| Aspect | 0.102 | -0.0002 | -0.0001 | 0.142 | 0 | -0.00008 | 0.088 | -0.00001 | 0.000000675 | 0.265 | -0.00009 | -0.000335 | 0.550 | -0.00015 | -0.0002 |
| Intercept | -0.985 | 0.781 | - | -0.958 | 0.725 | - | -1.152 | 0.775 | - | -1.329 | 1.191 | - | -0.923 | 0.7863 | - |
landslide training samples, often selected by a random process, to achieve relatively consistent results optimization of the SVM model hyperparameters by cross validation approach, is required. Marjanović et al. (2011) compared the SVM and LR models for LSM, where the authors pointed out that the training phase of the SVM model is time-consuming and requires considerable memory. However, they suggested that even with these disadvantages, SVM is suitable for LSM if the training pixels are carefully selected. In contrast, Pham et al. (2017) found that the SVM model outperformed the LR model. Even then, Wang et al. (2016) found that the LR model outperformed the ANN model, and selecting appropriate landslide samples for training the models can improve model accuracy. The SVM, LR and ANN are widely used for LSM, and are concluded to have good predictive capabilities with certain limitations. Previous works also suggested optimizing the training data-set as well as the model parameters for better LSM. In addition, the integration of these models into a hybrid model was also found to improve the base model accuracy (Pham et al. 2017).

One SVM model drawback is the considerable time required for application (Naghibi et al. 2017), while LR requires lesser time. To improve the SVM model performance, different researchers have

| Table 4. Effect of random sampling on model accuracy. |
|---------------------------------|----------------|----------------|----------------|----------------|----------------|
|                          | Scenario 1 | Scenario 2 | Scenario 3 | Scenario 4 | Scenario 5 |
|---------------------------|------------|------------|------------|------------|------------|
| SVM Overall accuracy (%)  | 81.25      | 83.93      | 81.70      | 82.59      | 80.80      |
| LR                         | 81.70      | 82.59      | 81.70      | 82.59      | 80.80      |
| Kappa statistic            | 0.625      | 0.679      | 0.634      | 0.652      | 0.616      |
| ROC area landslides        | 0.813      | 0.872      | 0.817      | 0.872      | 0.817      |
| Non landslides             | 0.813      | 0.872      | 0.817      | 0.872      | 0.817      |
| Average                    | 0.813      | 0.872      | 0.817      | 0.872      | 0.817      |
| ROC area landslides        | 0.813      | 0.872      | 0.817      | 0.872      | 0.817      |
| Non landslides             | 0.813      | 0.872      | 0.817      | 0.872      | 0.817      |
| Average                    | 0.813      | 0.872      | 0.817      | 0.872      | 0.817      |

| Table 5. Effect of random sampling on ANN model accuracy. |
|---------------------------------|----------------|----------------|----------------|----------------|----------------|
|                          | Scenario 1 | Scenario 2 | Scenario 3 | Scenario 4 | Scenario 5 |
|---------------------------|------------|------------|------------|------------|------------|
| Overall accuracy (%)      | 74         | 71         | 67         | 68         | 71         |
| ROC area                  | 0.79       | 0.77       | 0.74       | 0.73       | 0.75       |

![Figure 8. Landslide density graph of the SVM, LR and ANN models.](image)
suggested creating ensemble models by using its results and other statistical methods. For instance, SVM was combined with frequency ratio in Tehrany et al. (2015) and has been reported to have high efficacy. The high LR method efficacy could be related to its nonparametric nature (Nandi and Shakoor 2010), which allows the model to work with skewed distribution data, a phenomenon that can also be observed in the natural environment and landslide conditioning factors.

The current work suggests the careful input data pre-processing before the actual modelling. The outliers, highly correlated factors, and noisy data should be pre-processed to reduce final training data selection product effects. However, the LR model suggested a comparison of several models may be a better solution for accurate modelling because knowledge domain is different from one area to another.

5. Conclusions

Landslide susceptibility assessment methods have progressed significantly over the past decade; however, random training data selection effect is not well-investigated. This study contributes to literature by investigating the accuracy effect and parameter estimates of the SVM, LR and ANN models under five training data-set scenarios for LSM. In these scenarios, the data-sets were created using random sampling algorithms. Overall, the results showed that LR and ANN models had better performances than the SVM model in terms of sensitivity to training samples. The findings indicated that both the coefficient value and positive/negative factors correlation with the dependent variable (landslide presence/absent) were affected.

The main challenge in landslide susceptibility assessment is having large-scale and accurate landslide inventory data. With the limited number of landslide locations, models such as ANN can be difficult to build and optimize. Training models with a small number of training data often lead to having models with less generalization capacity. Hence, producing relatively acceptable landslide susceptibility maps with limited data and appropriate landslide location selection as well as optimizing model hyperparameters and their architecture are highly suggested. Future works should focus on improving the model generalization so that models built with large data-sets can be used in data-scarce environments.

Disclosure statement
No potential conflict of interest was reported by the authors.

ORCID
Biswajeet Pradhan http://orcid.org/0000-0001-9863-2054

References
Althuwaynee OF, Pradhan B, Park HJ, Lee JH. 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.
Atkinson PM, Massari R. 1998. Generalised linear modelling of susceptibility to landsliding in the central Apennines, Italy. Comput Geosci. 24(4):373–385.
Ayalew L, Yamagishi H. 2005. The application of GIS-based logistic regression for landslide susceptibility mapping in the Kakuda-Yahiko Mountains, Central Japan. Geomorphology. 65(1):15–31.
Ayalew L, Yamagishi H, Ugawa N. 2004. Landslide susceptibility mapping using GIS-based weighted linear combination, the case in Tsugawa area of Agano River, Niigata Prefecture, Japan. Landslides. 1(1):73–81.
Baeza C, Lantada N, Moya J. 2010. Influence of sample and terrain unit on landslide susceptibility assessment at La Pobla de Lillet, Eastern Pyrenees, Spain. Environ Earth Sci. 60(1):155–167.
Ballabio C, Sterlacchini S. 2012. Support vector machines for landslide susceptibility mapping: the Staffora River Basin case study, Italy. Math Geosci. 44(1):47–70.
Beven KJ, Kirkby MJ. 1979. A physically based, variable contributing area model of basin hydrology/Un modèle à base physique de zone d’appel variable de l’hydrologie du bassin versant. Hydrolog Sci J. 24(1):43–69.

Bui DT, Pradhan B, Lofman O, Revhaus I, Dick OB. 2012. Application of support vector machines in landslide susceptibility assessment for the Hoa Binh province (Vietnam) with kernel functions analysis. International Congress on Environmental Modelling and Software. 226. http://scholarsarchive.byu.edu/iemssconference/2012/Stream-B/226

Catani F, Lagomarsino D, Segoni S, Tofani V. 2013. Landslide susceptibility estimation by random forests technique: sensitivity and scaling issues. Nat Hazards Earth Syst Sci. 13(11):2815–2831.

Conforti M, Pascale S, Robustelli G, Sdao F. 2014. Evaluation of prediction capability of the artificial neural networks for mapping landslide susceptibility in the Turbolo river catchment (northern Calabria, Italy). Catena. 113:236–250.

Constantin M, Bednarik M, Jurchescu MC, Vlaicu M. 2011. Landslide susceptibility assessment using the bivariate statistical analysis and the index of entropy in the Sibiciu basin (Romania). Environ Earth Sci. 63(2):397–406.

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

Dai FC, Lee CF. 2002. Landslide characteristics and slope instability modeling using GIS, Lantau Island, Hong Kong. Geomorphology. 42(3):213–228.

Dai FC, Lee CF, Li J, Xu ZW. 2001. Assessment of landslide susceptibility on the natural terrain of Lantau Island, Hong Kong. Geomorphology. 42(3):327–343.

Devkota KC, Regmi AD, Pourghasemi HR, Yoshida K, Pradhan B, Ryu IC, Dhitul MR, Althuwaynee OF. 2013. Landslide susceptibility mapping of the slopes in the residual soils of the Mengen region (Turkey) by deterministic stability analyses and image processing techniques. Eng Geol. 44(1–2):147–161.

Dhakal AS, Amada T, Aniya M. 2000. Landslide hazard mapping and its evaluation using GIS: an investigation of sampling schemes for a grid-cell based quantitative method. Photogramm Eng Remote Sens. 66(8):981–989.

Ercanoglu M, Gokceoglu C. 2002. Assessment of landslide susceptibility for a landslide-prone area (north of Yenice, NW Turkey) by fuzzy approach. Environ Geol. 41(6):720–730.

Ermini L, Catani F, Casagli N. 2005. Artificial neural networks applied to landslide susceptibility assessment. Geomorphology. 66(1):327–343.

Garrett JH. 1994. Where and why artificial neural networks are applicable in civil engineering. J Comput Civ Eng, ASCE. 8(1994):129–130.

Gökçeöglu C, Aksoy H. 1996. Landslide susceptibility mapping of the slopes in the residual soils of the Mengen region (Turkey) by deterministic stability analyses and image processing techniques. Eng Geol. 44(1–4):147–161.

Gómez H, Kavzoglu T. 2005. Assessment of shallow landslide susceptibility using artificial neural networks in Jabonosa river basin, Venezuela. Eng Geol. 78(1):11–27.

Gritzner ML, Marcus WA, Aspinall R, Custer SG. 2001. Assessing landslide potential using GIS, soil wetness modeling and topographic attributes, Payette River, Idaho. Geomorphology. 37(1):149–165.

Hjort J, Marmion M. 2008. Effects of sample size on the accuracy of geomorphological models. Geomorphology. 102(3):341–350.

Hong H, Naghibi SA, Pourghasemi HR, Pradhan B. 2016a. GIS-based landslide spatial modeling in Ganzhou city, China. Arab J Geosci. 9(2):1–26.

Hong H, Pradhan B, Bui DT, Xu C, Youssef AM, Chen W. 2016b. Comparison of four kernel functions used in support vector machines for landslide susceptibility mapping: a case study at Suichuan area (China). Geomat Nat Hazards Risk. 1–26. https://doi.org/10.1080/19475705.2016.1250112

Huang J, Negnevitsky M, Nguyen DT. 2002. A neural-fuzzy classifier for recognition of power quality disturbances. IEEE Trans Power Deliv. 17(2):609–616.

Hussin HY, Zumpano V, Reichenbach P, Sterlacchini S, Micu M, van Westen C, Bălteanu D. 2016. Different landslide sampling strategies in a grid-based bi-variate statistical susceptibility model. Geomorphology. 253:508–523.

Karunungan DP, Arora MK, Sarkar S, Gupta RP. 2006. A comparative study of conventional, ANN black box, fuzzy and combined neural and fuzzy weighting procedures for landslide susceptibility zonation in Darjeeling Himalayas. Eng Geol. 85(3–4):347–366.

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(3):425–439.

Lee S, Pradhan B. 2007. Landslide hazard mapping at Selangor, Malaysia using frequency ratio and logistic regression models. Landslides. 4(1):33–41.

Lee S, Ryu JH, Lee MJ, Won JS. 2003. Use of an artificial neural network for analysis of the susceptibility to landslides at Boun, Korea. Environ Geol. 44(7):820–833.

Marjanović M, Kovačević M, Bajat B, Voženilek. V. 2011. Landslide susceptibility assessment using SVM machine learning algorithm. Eng Geol. 123(3):225–234.
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(2):725–742.

Remondo J, González A, De Terán JRD, Cendrero A, Fabbri A, Chung CJF. 2003. Validation of landslide susceptibility maps; examples and applications from a case study in Northern Spain. Nat Hazards. 30(3):437–449.

Restrepo C, Vitousek P, Neville P. 2003. Landslides significantly alter land cover and the distribution of biomass: an example from the Ninole ridges of Hawai`i. Plant Ecol. 166(1):131–143.

Riley SJ. 1999. Index that quantifies topographic heterogeneity intermountain. J Sci. 5(1–4):23–27.

Sammut C, Webb GI, editor. 2011. Encyclopedia of machine learning. Springer US. doi:10.1007/978-0-387-30164-8

Shahabi H, Khezri S, Ahmad BB, Hashim M. 2014. Landslide susceptibility mapping at central Zab basin, Iran: a comparison between analytical hierarchy process, frequency ratio and logistic regression models. Catena. 115:55–70.

Steinm SV. 1997. Selecting and interpreting measures of thematic classification accuracy. Remote Sens Environ. 62(1):77–89.

Stumpf A, Kerle N. 2011. Object-oriented mapping of landslides using random forests. Remote Sens Environ. 115(10):2564–2577.

Talebi A, Uijlenhoet R, Troch PA. 2007. Soil moisture storage and hillslope stability. Nat Hazards Earth Syst Sci. 7(5):523–534.

Tehrany MS, Pradhan B, Jebrur MN. 2015. Flood susceptibility analysis and its verification using a novel ensemble support vector machine and frequency ratio method. Stoch Environ Res Risk Assess. 29:1149–1165.

Tien Bui D, Pradhan B, Lofman O, Revhaug I. 2012. Landslide susceptibility assessment in Vietnam using support vector machines, decision tree, and Naïve Bayes models. Math Probl Eng. 2012(2012), Article ID 974638, 26. http://dx.doi.org/10.1155/2012/974638

Umar Z, Pradhan B, Ahmad A, Jebrur MN, Tehrany MS. 2014. Earthquake induced landslide susceptibility mapping using an integrated ensemble frequency ratio and logistic regression models in West Sumatera province, Indonesia. Catena. 118:124–135.

Varnes DJ. 1978. Slope movement types and processes. Special report, Vol. 176. NW Washington, DC: Transportation Research Board.

Vorpahl P, Elsenbeer H, Mäker M, Schröder B. 2012. How can statistical models help to determine driving factors of landslides? Ecol Model. 239:27–39.

Wang LJ, Guo M, Sawada K, Lin J, Zhang J. 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(1):117–136.

Xu C, Dai F, Xu X, Lee YH. 2012. GIS-based support vector machine modeling of earthquake-triggered landslide susceptibility in the Jianjiang river watershed, China. Geomorphology. 145:70–80.

Yilmaz I 2009. A case study from Koyulhisar (Sivas-Turkey) for landslide susceptibility mapping by artificial neural networks. Bull Eng Geol Environ. 68(3):297–306.

Yilmaz I. 2010. The effect of the sampling strategies on the landslide susceptibility mapping by conditional probability and artificial neural networks. Environ Earth Sci. 60(3):505–519.

Yilmaz I, Kaynar O. 2011. Multiple regression, ANN (RBF, MLP) and ANFIS models for prediction of swell potential of clayey soils. Expert Syst Appl. 38(5):5958–5966.

Youssef AM. 2015. Landslide susceptibility delineation in the Ar-Rayth area, Jizan, kingdom of Saudi Arabia, using analytical hierarchy process, frequency ratio, and logistic regression models. Environ Earth Sci. 73(12):8499–8518.

Zare M, Pourghasemi HR, Vafakhah M, Pradhan B. 2013. Landslide susceptibility mapping at Vaz Watershed (Iran) using an artificial neural network model: a comparison between multilayer perceptron (MLP) and radial basic function (RBF) algorithms. Arab J Geosci. 6(8):2873–2888.