A COMPARISON BETWEEN THREE CONDITIONING FACTORS DATASET FOR LANDSLIDE PREDICTION IN THE SAJADROOD CATCHMENT OF IRAN

B. Kalantar[1,*], N. Ueda[1], H. A. H. Al-Najjar[2], V. Saeidi[3], M. B. A. Gibril[4], A. A. Halin[5]

[1] RIKEN Center for Advanced Intelligence Project, Goal-Oriented Technology Research Group, Disaster Resilience Science Team, Tokyo 103-0027, Japan - (Bahareh.kalantar, naonori.ueda)@riken.jp
[2] Centre for Advanced Modelling and Geospatial Information Systems (CAMGIS), Faculty of Engineering and IT, University of Technology Sydney, 2007 NSW Sydney, Australia - husam.al-najjar@student.uts.edu.au
[3] Dept. of Mapping and Surveying, Darya Tarsim Consulting Engineers Co. Ltd., 1457843993 Tehran, Iran- saeidi@daryatarsim.com
[4] Research Institute of Sciences and Engineering, University of Sharjah, 27272 Sharjah, UAE - mbgibril@sharjah.ac.ae
[5] Dept. of Multimedia, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Serdang, 43400 Selangor, Malaysia - alfian@ieee.org

Commission III, ICWG III/IVa

KEYWORDS: Landslide Susceptibility, Parameter Selection, GIS, Machine Learning, Factor Optimization

ABSTRACT:

This study investigates the effectiveness of three datasets for the prediction of landslides in the Sajadrood catchment (Babol County, Mazandaran Province, Iran). The three datasets (D1, D2 and D3) are constructed based on fourteen conditioning factors (CFs) obtained from Digital Elevation Model (DEM) derivatives, topography maps, land use maps and geological maps. Precisely, D1 consists of all 14 CFs namely altitude, slope, aspect, topographic wetness index (TWI), terrain roughness index (TRI), distance to fault, distance to stream, distance to road, total curvature, profile curvatures, plan curvature, land use, steam power index (SPI) and geology. D2, on the other hand, is a subset of D1, consisting of eight CFs. This reduction was achieved by exploiting the Variance Inflation Factor, Gini Importance Indices and Chi-Square factor optimization methods. Dataset D3 includes only selected factors derived from the DEM. Three supervised classification algorithms were trained for landslide prediction namely the Support Vector Machine (SVM), Logistic Regression (LR), and Artificial Neural Network (ANN). Experimental results indicate that D2 performed the best for landslide prediction with the SVM producing the best overall accuracy at 82.81%, followed by LR (81.71%) and ANN (80.18%). Extensive investigations on the results of factor optimization analysis indicate that the CFs distance to road, altitude, and geology were significant contributors to the prediction results. Land use map, slope, total-, plan-, and profile curvature and TRI, on the other hand, were deemed redundant. The analysis also revealed that sole reliance on Gini Indices could lead to inefficient optimization.

1. INTRODUCTION

Landslides are a type of natural disaster that can have detrimental effects on human livelihood, which includes the destruction of properties, undesirable changes to the environment as well as human casualties (Chen et al. 2018). The damages incurred interfere with many economic and social activities. Various factors can be linked to the cause of landslides, where many are beyond human control. These include melting of glaciers, excessive rainfall, mining activities, volcanic eruptions and earthquakes (Mousavi et al. 2011; Dou et al. 2015;) Therefore, the ability to predict landside occurrences is exceptionally vital, especially for disaster mitigation and management. It would also be beneficial if the contributing factors could be identified according to their importance, which would greatly facilitate and expound the benefit of landslide prediction.

Topographical, geological, and hydrological datasets are active conditioning factors for landslide prediction, but each carries different levels of importance (Mahalingam et al. 2016; Afungang et al. 2017). The prioritization of such factors depends on the characteristics of the study area; hence there is no guideline for any particular factors selection/consideration (Chen et al. 2018).

Dou et al. (2015) emphasized the effect of factors optimization prior to landslides susceptibility mapping in order to reduce noise and uncertainty. Concerning that, Afungang et al. (2017) optimized conditioning factors selection using the Information Value Model, where they ended up with the six factors that include slope, curvature, aspect, land use, geomorphology, and lithology, to map landslides prone areas. Mahalingam et al. (2016) investigated LiDAR-derived datasets to map landslide-prone areas using six different machine learning models. Their results suggest that slope was the most crucial conditioning factor in all models, though the relative contribution of other factors varied across each model. Dou et al. (2015) investigated the optimal number and types of causative factors in statistical models. Their findings indicated that reducing 15 factors to 6 critical factors (slope angle, slope aspect, drainage density, lithology, distance to geological boundary, and distance to faults) results in more accurate landslides predictions. In another study, Pradhan and Lee (2010) proposed an adaptive neuro-fuzzy inference system (ANFIS) to examine the importance of the eight landslide conditioning factors of altitude, slope, lithology, distance from road, distance from drainage, distance from fault, plan curvature and vegetation index or NDVI (study area: Cameron Highlands, Malaysia). The authors adopted an incremental strategy by beginning with four factors for landslide susceptibility mapping and then increasing the number of factors

* Corresponding author

one by one. Their results indicate that distance to fault was least influential and curvature was most influential for prediction, respectively. The drawback of this work however, is that they did not rank the importance of each factor before modelling. In a similar work by Sezer et al. (2011), they evaluated ANFIS by beginning with three factors and made their way up to 7 factors. The Receiver Operating Characteristic (ROC) from the first model increased from 67.38% to 98.52% in the 5th model. Again, the significance of each factor was not clearly elaborated. The authors, however, highlighted that plan curvature was the least important factor, while the lithology factor had increased the ROC up to 10%.

It is quite apparent from the literature that researchers are striving to choose the best factors, along with the suitable modelling technique (Al-Najjar et al. 2019; Kalantar et al. 2019, Nguyen et al. 2019). The process of finding the optimal factor combination and appropriate modelling approach is crucial as different factor combinations and model selection can lead to different results. For instance, adding or removing conditioning factors can cause desirable (or undesirable) prediction accuracy values of the selected model (Kalantar et al. 2019).

This study meant to investigate the same theme where we look at a total number of 14 landslide conditioning factors to determine the best combination that yields the best prediction. In particular, the Variance Inflation Factor (VIF), Gini importance, and Chi-square were used to evaluate the effectiveness of the factors under consideration. Consequently, three datasets are created, i.e., D1, which includes all 14 conditioning factors; D2, which is a dataset (based on D1) that is reduced using factor analysis and importance; and D3 containing DEM derivatives (morphometrics factors). Different modelling techniques (i.e., supervised machine learning) were used namely Support Vector Machine (SVM), Artificial Neural Network (ANN), and Logistic Regression (LR).

2. STUDY AREA AND DATA USED

For this work, the study area chosen is Sajadrood catchment, which is located in Babol county within the Mazandaran Province of Iran (Figure 1a). The coordinates for this catchment are approximately in the north latitudes 36°9′ and 36°10′ and east longitudes 52°30′ and 52°40′ with a coverage area of approximately 118.8km^2. The population is estimated to be around 26,809 people (2006 census). The study area consists of dense forests, agriculture areas and paddy fields (Figure 1b). According to the Iranian Meteorological Organization, Sajadrood's temperature ranges between -3°C (February) to 38°C (August) with a long-term average temperature of 17.1°C. The climatic condition of the catchment is cold and mild mountainous and receives heavy rainfall throughout the year, with an annual average precipitation of 680 mm.

The study has various types of geological formations as shown in Figure 1c. Using a 1: 25,000-scale topographic map of Sajadrood, we generated a 10m Digital Elevation Model (DEM) as the primary data source for landslide susceptibility mapping. In this study, 227 landslide inventory points were collected from satellite imagery and field surveys by the Geological Survey of Iran. 70% of the landslide inventories were randomly used to train three supervised machine learning models, namely the SVM, ANN, and LR. The remaining 30% of the landslide inventory points were reserved for testing the machine learning models.

In this work, 14 conditioning factors (Figure 1) derived from the DEM and topographic databases (using ArcMap 10.3) are considered, namely altitude, slope, aspect, topographic wetness index (TWI), terrain roughness index (TRI), stream power index (SPI), distance to fault, distance to stream, distance to road, land use, total curvature, profile curvature, plan curvature and geology. These factors are chosen due to their availability and also since they were also used in relevant works such as that by (Nguyen et al. 2019).

2.1 Landslide Conditioning Factors Preparation

Selection and preparation of conditioning factors are done according to the works of (Kalantar et al. 2018), which are briefly explained in this section. A region's altitude variation has considerable influence on landslide susceptibility. We, therefore, classified altitude into the five classes using the natural break scheme. Resultantly, the altitude factor ranges from the minimum height of 74 meters to a maximum of 1500 meters (Figure 1d). A crucial factor that triggers landslides as a source of stress and instability in steep areas is the slope. The slope angle map is hence separated into 5 interval classifications: (i) 0°-8.4°, (ii) 8.5°-13°, (iii) 14°-17°, (iv) 18°- 23°, and (v) 24°-48° (Figure 1e). Slope Aspect influences vegetation growth and moisture level of the soil (due to rainfall), wind, and solar radiation. We categorized aspect into the 9 classes (i) flat, (ii) north, (iii) northeast, (iv) east, (v) southeast, (vi) south, (vii) southwest, (viii) west, and (ix) northwest (Figure 1f).

Topographic wetness index (TWI) measures the tendency of runoff and the position where water converges. Terrain Roughness Index (TRI), on the other hand, indicates slopes that are concave and convex upward, while Steam Power Index (SPI) measures the intensity and erosive power of slope surface runoff. The calculations for these three indices are as follows:

$$SPI = A_s \tan\beta \quad (1)$$

$$TWI = \log_e\left(\frac{A_s}{\tan\beta}\right), \quad (2)$$

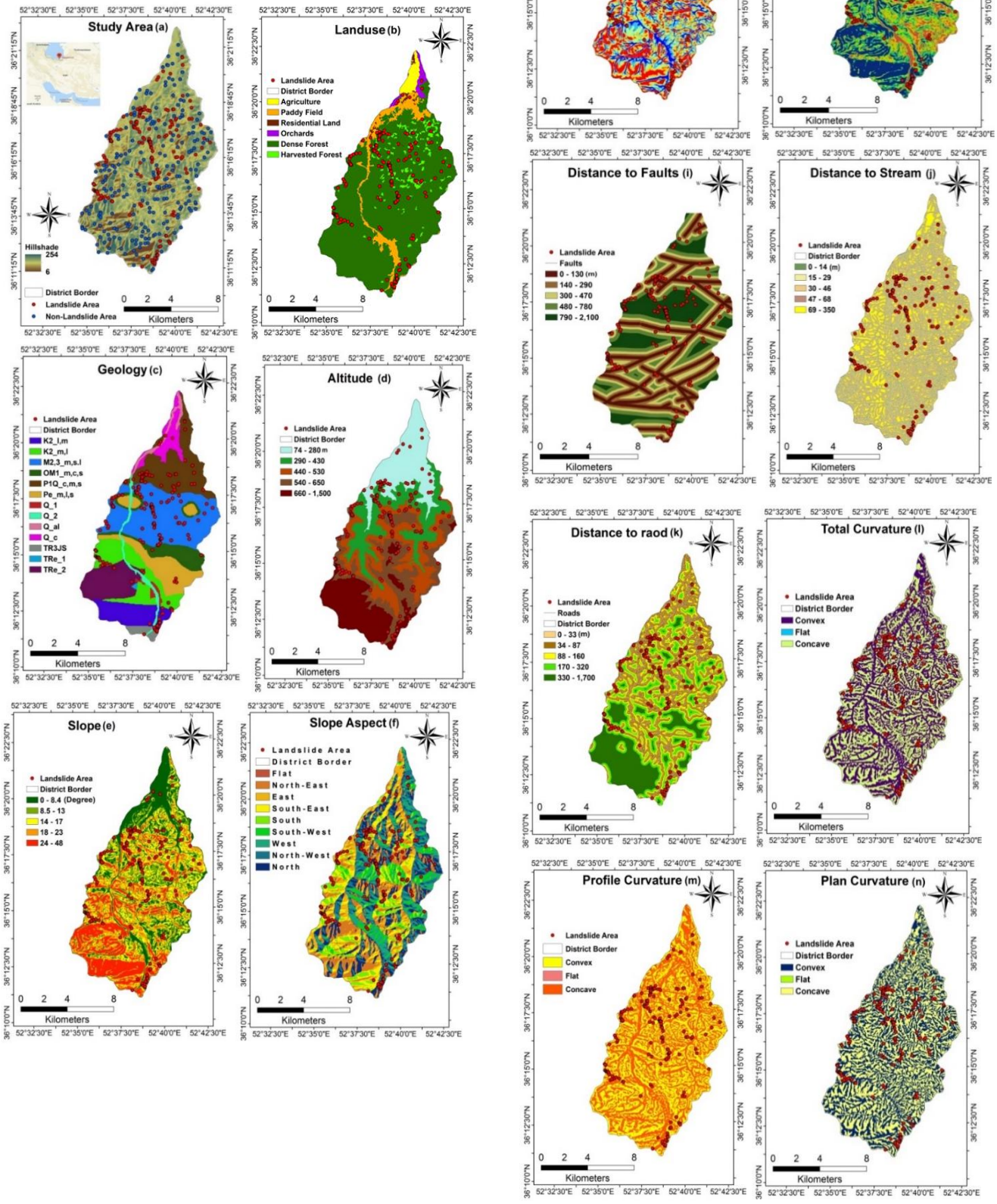
$$TRI = \sqrt{[x](\max^2+\min^2)} \quad (3)$$

where A_s = area of catchment in m^2
β = gradient of the slope in radians
max, min = largest and minimum value of a pixel I nine rectangular altitude neighbourhoods.
SPI, TWI, and TRI are then classified into five classes (Figure 1o, g, h).

Landslides commonly occur along faults, rivers, and roads, mainly as a result of soil erosion and human activities. In this work, we follow the classification done by Hong et al (2018) and Golkarian et al. (2018). The distances to faults, streams, and roads were separated into five classes using the Euclidean distance function in ArcGIS (Figure li, j, k). Different land use types can be a sign of human activities and/or environmental changes, which can influence ground shape and stability. In this work, it was discovered that the land use map of the study area contained six land use categories, namely (i) agriculture, (ii) paddy field, (iii) residential land, (iv) orchards, (v) dense forest, and (vi) harvested forest. We used supervised classification from the Landsat Enhanced Thematic Mapper (2017) image with an accuracy of 90%.

While surface curvature reflects the shape of the ground surface affecting soil runoff, the profile curvature affects water velocity flow that drains the surface, which also influences erosion and deposition. Plan curvature reflects slopes steepness (horizontal

plane) that influences surface runoff characteristics. Total curvature is the surface's curvature, which is by definition, equals to the sum of the profile and plan curvatures. Extra details regarding curvatures (which include equations and formulas), can be found in the literature (Alkhasawneh et al. 2013). In this work, total, profile, and plan curvature maps were classified into three categories: (i) concave, (ii) flat, and (iii) convex (Figure 1 l, m, n).



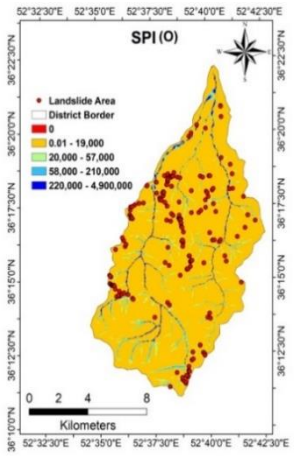


Figure 1. (a) Study area, (b) Land use map, (c) Geology map, (d) Altitude, (e) Slope, (f) Aspect, (g) Topographic Wetness Index (TWI), (h) Terrain Roughness Index (TRI), (i) Distance to Fault, (j) Distance to Stream, (k) Distance to Road, (l) Total Curvature, (m) Profile Curvature, (n) Plan Curvature, (o) Stream Power Index (SPI).

3. METHODOLOGY

The datasets used in this research are shown in Table 1. Dataset D1 includes 14 conditioning factors, D2 is a reduced-size dataset of D1. The eight conditioning factors were derived by applying three-factor optimization techniques namely Variance Inflation Factor (VIF), Gini importance indices, and Chi-square. Lastly, the third dataset D3 includes only DEM-derived factors. Figure 2 simplifies the methodological flowchart of this research.

Dataset	Conditioning Factors
D1	Altitude, Slope, Curvature, Plan Curvature, Profile Curvature, TWI, TRI, SPI, Distance to Stream, Distance to Road, Distance to Fault, Land use, Lithology, and Aspect
D2	Altitude, TWI, SPI, Distance to Stream, Distance to Road, Distance to Fault, Lithology, Aspect
D3	Altitude, Slope, Curvature, Plan Curvature, Profile Curvature, TWI, TRI, SPI, and Aspect.

Table 1. Three datasets of conditioning factors.

3.1. The importance of Factor Analysis

Selecting suitable conditioning factors is essential to produce accurate landslide susceptibility maps. Multicollinearity, outliers, and spatial variations of conditioning factors are issues that necessitate factor analysis in susceptibility assessment. This type of analysis enables the removal of redundant factors, which makes constructing and training any model simpler (Kalantar et al. 2017). In this work, the highly related features discard approach was adopted. Mainly, an estimation of variance-inflated factor (VIF) was used:

$$VIF = \frac{1}{1- R^2} \quad (4)$$

where R' = the multi-correlation coefficient between features. VIF values that are 5 or 10 and higher suggest highly correlated factors. Such features are deemed unsuitable and are consequently removed from consideration (O'Brien 2007).
In addition to factor analysis, other techniques to handle data redundancy are the Chi-Square Factor Optimization and Gini Importance methods. A higher Chi-square value is responsible

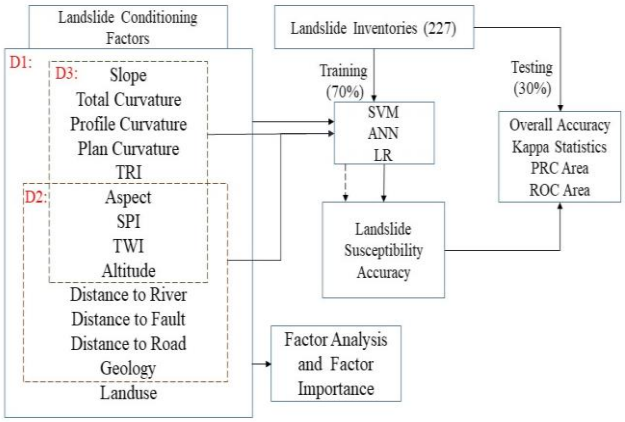


Figure 2. Flowchart of this research.

for the more critical prediction factor to detect the landslides. In this work, the p-value was evaluated against a 0.05 level of significance, which allows the establishment of the significant relationship between landslide occurrence and the particular conditioning factors. Also, the Gini coefficient and Cramer's V statistics (both ranging from 0 to 1) are computed for each factor. For the Gini coefficient, a value of 0 indicates that all the variables are equal. A value of 1, on the other hand, denotes inequality among the variables. In contrast, Cramér's measures the correlation between landslide conditioning factors. Here, 0 implies no correlation whereas 1 shows a perfect correlation. Therefore, the highest value of Cramer's reveals the highest correlation between the factors while the highest value of the Gini coefficient represents a lower correlation.

3.2 Models

3.2.1 Support Vector Machine (SVM)

The SVM is a machine learning algorithm based on statistical learning theory. It was initially meant for binary classification problems but can be extended to multi-class classification as well.

The SVM operates in a higher dimensional feature space, which is obtained by using a specific kernel function. The intuition behind the algorithm is to discover an optimal separating hyperplane between the positive and negative classes by calculating the maximum margin to the nearest training examples (Cortes and Vapnik 1995). The positive class is annotated as +1, whereas the negative class as -1.

In this work, intuitively, the positive class refers to landslide whereas the negative class to non-landslide. Specifically, the algorithm is given a set of n labelled training examples $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ with $x_i \in \mathrm{R}^n$(where i=1,…,n) and $y_i \in \{+1. -1\}$. In this work, x_i represents each of the abovementioned conditioning factors. Depending on the type of data, the SVM's performance is determined by the choice of the kernel function. Commonly used kernels are the RBF (radial basis function), polynomial, sigmoid, and linear. In this work, we opt for the linear kernel due to its simplicity. Overall, the linear separating hyperplane of the SVM can be written as follows

$$\frac{y_i(\omega \cdot x_i + b)}{\geq 1} - \delta_i \quad (5)$$

where ω = the coefficient vector, which decides the separating hyperplane's final orientation. The variable b is hyperplane's offset from the origin and the slack variable δ_i caters for penalizing any constraints violation (Cortes and Vapnik 1995).

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-V-3-2020-625-2020 | © Authors 2020. CC BY 4.0 License.

3.2.2 Linear regression (LR)

Similar to the SVM, LR is a binary classifier as well. In the context of this research, LR's main objective is to identify the optimal coefficients associated with each independent variable (i.e., conditioning factor) by discovering relationships with the dependent variables (Ozdemir and Altural 2013), which in this work are landslide vs. no-landslide.

The LR does assume a normal distribution (Pradhan and Lee 2010), and the independent variables are annotated as 0 and 1 to reflect landslide and no-landslide, respectively. Since LR calculates its output based on the Sigmoid (or Logistic) function, the output is a probability value. Specifically, LR determines the probability of a class based on the following

$$z = \theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n \quad (6)$$

$$g(z) = \frac{1}{1-e^{-\theta^T x}} = \Pr(y = 1|x; \theta) \quad (7)$$

where θ denotes the linear model parameters, which are the coefficients representing the weight contribution of each conditioning factor x in, the function g(z) is the logistic function that calculates the probability of whether the input values correspond to the positive class y=1, indicating a landslide. In this work, g(z)>0.5 is considered to be in a positive class.

3.2.3 Artificial neural network (ANN)

In contrast to statistical models, the ANN is independent of any data's statistical distribution hence does not require the calculation of any statistical variable (Pradhan and Lee 2010). ANNs also have the ability to generalize even when dealing with imperfect/incomplete data for nonlinear problems (Tian et al. 2019). In this study, a Multi-Layer Perceptron (MLP) ANN was trained and learning the weights is achieved using the back-propagation algorithm. The MLP is a widely used architecture that consists of three main components, namely the input layer (input data), output layer (provides prediction results), and one or more hidden layers that interconnect the input and output (Aditian et al. 2018). As with any machine learning model, training the MLP-ANN begins with random weight assignments for each neuron. Learning occurs by a continuous update of each of the weights and stops upon reaching acceptable training accuracy. The updating of the weights is basically performed via the minimization of a particular error function that calculates the difference between the predicted and the actual output values. To gain more insight into the algorithm, readers can be directed to the following literature (Kim et al. 2014)

3.3 Accuracy Assessment

The metrics used for classifier evaluation is Overall Accuracy (OA), Kappa Statistics, Receiver Operating Characteristics (ROC), and Prediction Rate Curve (PRC) area. Overall accuracy (OA) determines the proportion of sites that have been correctly mapped. It is obtained by dividing the total number of pixels that are correctly classified by the total number of pixels. OA is expressed as a percentage. According to Shafii and Price (2001) and Viera and Garrett (2005), Cohen's Kappa interprets the degree of agreement between observed and predicted values. A Kappa of 1 indicates the best agreement in the model. ROC stands for the Receiver Operating Characteristics curve. Based on Tsangaratos and Ilia (2016), the ROC plots the true positive rate (i.e. the rate at which the model correctly predicts landslide) against the false positive rate (i.e. the rate at which the model predicts landslide as non-landslide). The area under the curve (AUC) calculates the area under the ROC, which indicates a classifier's overall accuracy. An area of 0.5 indicates weak, whereas one as flawless (Beguería 2006). The Prediction Rate Curve is a plot where the vertical y-axis is the success rate (i.e. truly detected landslides), and the horizontal x-axis is the total positive landslide-prone areas. It is also used to determine the prediction prowess of a classifier (Beguería 2006; Pourghasemi and Rossi 2019); they were similar to the ROC, the area varies from 0 to 1.

4. RESULTS

First of all, the importance of each conditioning factor was investigated by analysing the VIF and Gini importance indices. The former is shown in Table 2 whereas the latter in Table 3. From Table 2, VIF values less than 10 indicate low correlation, whereas VIFs above 10 suggest higher correlation. It can also be seen that most of the Tolerance values are higher than 0.1, indicating less correlation between the factors (exceptions being for land use and aspect).

In Table 3, higher Chi-square values with a p-value less than 0.05 indicate that the factor is significant for landslide prediction. Specifically, Chi-square analysis highlights distance to road, altitude, and geology more than any of the other factors. Land use is seen as the least important factor.

Variable	Summary statistics and Multicollinearity	
	Tolerance	VIF
Altitude	0.67	1.81
Slope	0.98	23.29
Total Curvature	1.00	3.02255E+13
Plan Curvature	1.00	1.11475E+13
Profile Curvature	1.00	1.18828E+13
TWI	0.63	1.65
TRI	0.98	23.51
SPI	0.30	1.100
Distance to Stream	0.24	1.06
Distance to Road	0.39	1.18
Distance to Fault	0.18	1.03
Land use	0.07	1.00
Geology	0.57	1.48
Aspect	0.03	1.00

Table 2. Variance Inflation Factor (VIF) analysis results.

The results of the Gini importance indices include information value (IV), Cramér's V and Gini coefficient values. Higher IV values can be seen for distance to road, geology and TWI, which can be translated to "strong" predictors for mapping landslide-prone zones. Cramér's V, on the other hand, shows all factors expect distance to road, having values less than 0.3. This indicates an insignificant correlation (except for distance to the road at 0.69). The Gini coefficient values indicate a slight correlation between all factors (all values ~ 0.5). The degree of correlation, however, was higher for distance to the road (0.26), which is a value close to zero.

As previously mentioned, three datasets (D1, D2, and D3) are considered in order to see which one would provide the best representation for landslide susceptibility mapping. Note that D1 consists of all the 14 conditioning factors. The intuition behind D2 and D3 is to see whether a reduced set of CFs can also achieve good accuracies. Hence, for the purpose of optimization, redundant factors are removed prior to modelling, which is consistent with the work in Mousavi et al. (2017). The differences between VIF, Gini, and Chi-square led us to choose the most

common factors. Consequently, slope, total curvature, plan curvature, profile curvature, TRI, and land use were removed from the datasets to create the D2. D3 included only DEM-derived factors.

The three classification models SVM, ANN, LR, were constructed based on the three datasets. In Table 4, it can be seen that all three algorithms performed equally well in mapping the spatial distribution of landslide-prone areas. It appears that SVM performed best with an overall accuracy (OA) of 82.81% using dataset D2 as compared with ANN (OA = 80.18%) and LR (OA = 81.71%). Dataset D1, which contains all 14 factors, showed inferior performance in OA. Additionally, the accuracy for all models dropped to the value of 62.55%, 66.96%, and 60.35 for SVM, ANN, and LR model, respectively, using D3, which indicates that the DEM-generated dataset is not a suitable representation.

Generally, the Kappa statistics in Table 4 showed "substantial agreement" between the observations (ground truth or inventory map) and predictions (landslide susceptibility map) for the three models using the D1 and D2 datasets, while the degree of agreement abruptly declined to "fair agreement" for all models utilizing the D3. This is in line with the explanations provided by (Viera and Garrett 2005). This again shows that DEM-derived data alone from D3 is insufficient for training the classifiers. The AUC (Table 4) shows promising levels for all three models when considering D1 and D2.

For instance, using the D1, a maximum AUC of 0.88 was obtained by ANN and LR, whereas, LR model performed better using D2 at 0.89. Besides, all three models only reported: "moderate accuracy" for D3. Likewise, the prediction power and success rate for true positives for the three models were evaluated by PRC (Table 4) and the results indicated the best performance was obtained by ANN (0.88) using D1. Prediction performance, however, was for the SVM (0.58) when using D3. Finally, for this study, all accuracy assessments and validation techniques agreed that applying SVM, ANN and LR using D3 were unreliable for accurate landslide susceptibility mapping, while almost all three models performed better when using D1 and especially D2. When looking at processing time, the LR model using D1 performed ~2.67 times faster than when it was exploited in the D2 dataset. The LR technique has been implemented within 0.03 seconds using optimized factors (D2) and is ranked the fastest algorithm to compare with SVM (0.14s) and ANN (0.27s). Again, it confirmed the importance of factors optimization for a broad set of variables and conditioning factors in landslide-prone zones where we are dealing with large sets of data and variables.

For a better understanding of each conditioning factor and reliability of our predictions, we omitted each factor in time from model. Table 5 indicated that just removing the distance to the road had a significant effect on the level of agreement between the observations and predicted landslide areas so; the final map may seem unreliable without this particular factor, this confirmed uncertainty associated with Cramér's V and Gini coefficient results, as well.

Accuracy	Factors	Models		
		SVM	ANN	LR
Overall Accuracy (OA)	D1	82.15	81.05	81.71
	D2	82.81	80.18	81.71
	D3	62.55	66.96	60.35
Kappa Statistics	D1	0.64	0.62	0.63
	D2	0.65	0.60	0.63
	D3	0.25	0.33	0.20
PRC area	D1	0.77	0.88	0.87
	D2	0.78	0.88	0.86
	D3	0.58	0.70	0.63
ROC area	D1	0.82	0.88	0.88
	D2	0.83	0.88	0.89
	D3	0.71	0.71	0.63

Table 3. Accuracy assessment and validation of SVM, ANN, and LR.

5. DISCUSSION

The increased measures of VIF for slope, total curvature, plan and profile curvature, and TRI has been detected as collinearity and redundancy in the datasets. Tolerance values less than 0.1 also indicated the presence of multicollinearity in land use and aspect. The Chi-square method, on the other hand, categorized distance to the road and land use as the best and worst factors, respectively. In contrast, Gini indices values obtained controversy results as Cramér's V and Gini coefficient concluded that distance to the road was a redundant variable, whereas IV evaluated distance to the road with a higher degree of inequality as an influential factor. Bergsma (2013) noted that Cramér's V could be biased when Chi-square increases and the result may overestimate the degree of association. To ensure that distance to the road is essential, we examined its absence in the SVM, ANN and LR (Table 5) and computed the Kappa Index. As a result, Kappa decreased dramatically when the distance to the road is removed. In all, this indicated that distance to the road was indeed a very critical factor (in line with Mousavi et al. 2011).

The accuracies of the models were evaluated using the datasets D1, D2, and D3. Mainly, all three models performed well using D1 and D2 datasets. The SVM, using the optimized factors (i.e. D2), outperformed others based on overall accuracy and Kappa. This implies that the redundancy removal in factor optimization leads to better classification performance. The LR algorithm shows identical accuracy and Kappa using D1 and D2 due to the corresponding coefficient matrix with data evaluation and exclusion of nature during the logistic regression process (Mousavi et al. 2017). For this reason, as well, the evaluation results for VIF and Chi-square were in agreement with the LR coefficient matrix to eliminate data redundancy. Validation of the ANN algorithm using PRC shows the highest prediction accuracy and performed significantly well compared to SVM and LR. Due to this, we foresee ANN to be a reliable alternative when dealing with uncertain, noisy and insufficient conditioning factors. The AUC finally validated that all three algorithms performed well, while LR shows the best overall performance using the D2 dataset.

Two experiments by Pradhan and Lee (2010) and Sezer et al. (2011), which was discussed earlier in this article, applied the ANFIS algorithm with almost the same conditioning factors for susceptibility mapping in different study areas. In comparison with these works, the significance of conditioning factors was diverse, and even the most important factor considered by one research was labelled as the least important factor by the other one. Although in our study, we had only four practical factors in common with this researches, we could obtain a good level of accuracy using other conditioning factors, as well.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-V-3-2020-625-2020 | © Authors 2020. CC BY 4.0 License.

all 14 factors and calculated kappa for SVM, ANN, and LR Factors	Chi-square method		Gini Indices method		
	Chi-square	p-value	Gini Coefficient	Information value	Cramer's V
Distance to Road	217.4873	0.000000	0.261312	2.522338	0.690923
Altitude	86.5748	0.000000	0.428559	0.240013	0.377996
Geology	75.3112	0.000000	0.447492	0.436883	0.324061
TWI	48.0967	0.000146	0.456976	0.415725	0.293340
TRI	33.7034	0.090079	0.461493	0.320196	0.277514
Aspect	31.9614	0.059079	0.465126	0.218050	0.264099
Distance to Fault	30.9730	0.028995	0.470014	0.275576	0.244894
Plan Curvature	28.7033	0.121309	0.470073	0.082453	0.244651
Distance to Stream	28.1970	0.134651	0.472638	0.034174	0.233932
Total Curvature	26.1546	0.345367	0.478608	0.185279	0.206841
Slope	25.6101	0.373241	0.480964	0.165356	0.195118
Profile Curvature	20.8180	0.794739	0.482680	0.142213	0.186118
SPI	16.9545	0.151316	0.487801	0.099749	0.156199
Land use	7.7209	0.562497	0.492528	0.047146	0.122243

Table 4. The importance of factors using Gini importance and chi-square techniques.

All Conditioning Factors Except	Method		
	SVM	ANN	LR
Altitude	0.6564	0.5771	0.652
Slope	0.6388	0.6476	0.6388
Curvature	0.6432	0.5859	0.6344
Clan Curvature	0.6432	0.6123	0.6344
Profile Curvature	0.6432	0.6123	0.6344
TWI	0.6476	0.6167	0.6211
TRI	0.6388	0.6564	0.6388
SPI	0.6344	0.5595	0.6344
Distance to Stream	0.6388	0.6211	0.6388
Distance to Road	0.2423	0.4405	0.1586
Distance to Fault	0.6432	0.5947	0.6344
Land use	0.6256	0.6432	0.6432
Lithology	0.6652	0.5595	0.6476
Aspect	0.6432	0.5551	0.6476

Table 5. Cohen's Kappa Index for SVM, ANN, and LR techniques of landslide susceptibility by removing one conditioning factor in a time.

Thus, prior factor optimization in our research led to avoiding over learning the algorithms, heavy calculation, and modelling, especially when dealing with a large area and several conditioning factors.

6. CONCLUSION

Three supervised learning models (SVM, ANN, and LR) were constructed based on each dataset. The primary objective was to determine which dataset was most representative for landslide prediction. The first dataset D1 considered 14 conditioning factors; the second dataset D2 had a reduced set of 8 factors, while the third dataset D3 included only DEM-derived factors. VIF, Chi-square, and IV Gini index firmly prioritized the conditioning factors where there is no standard guideline to rank these factors, and it is highly subjective to the characteristics of the study area. Factor optimization ultimately highlighted distance to the road; altitude and geology were significant contributing factors, slope, plan and profile curvature that seemingly affects erosion process more than other factors in many similar studies (Pradhan and Lee 2010) were found to be insignificant factors for this case study. For this particular area, the importance of distance to road indicated that most of the predictions and landslides had been identified in the areas close to the roads. So, road construction may potentially trigger the hazards more than other factors. Predominantly, the SVM model obtained the best accuracy and kappa of 82.81% and 0.65, followed by LR (81.71%) and ANN (80.18%) using D2. The same scenario goes with D1, as well, and SVM (82.15%) achieved the best result even though LR had a hidden factor optimization layer. For this study, SVM was confirmed as the best classifier for mapping the susceptible landslides. Again, none of the algorithms reached a supportable level of accuracy using D3 although ANN behaved more effectively with this incomplete dataset.

To put it briefly, the availability of data from different remote sensing sources lead to deal with massive data and conditioning factors to predict the landslide hazards; therefore, the quality and speed of modelling necessitate factor optimization, in advance. The outcome of this research emphasized that the importance of landslide causative factors differs from one site to another, and it could be remarkably changed by human activities (Kalantar et al. 2019); also, the choice of optimizer could directly affect the optimization results. The site dependency of landslide conditioning factors and the choice of optimizers emphasize that even a pre-used group of conditioning factors for a particular zone might not be successfully applied to another region. Therefore, for a reliable result, the use of all available datasets in a study area is highly beneficial, besides, without proper optimization algorithms, one cannot omit a factor even it was tagged insignificant by some other researchers. Especially for this study, road construction was the main source of improper human activities in residential areas with lower altitude. Thus, it is recommended to use more than one optimizer prior to classification. Moreover, for those governmental organizations and private sectors involving in road construction, it is suggested that more attention is needed during transport network construction and maintenance in Sajadrood due to geology and unstable soil type. Lastly, this promoted the importance of landslide mitigation and early warning system to decrease casualties and losses where construction is inevitable.

REFERENCES

Aditian A, Kubota T, Shinohara Y .2018. Comparison of GIS-based landslide susceptibility models using frequency ratio, logistic regression, and artificial neural network in a tertiary region of Ambon, Indonesia. *Geomorphology*. 318, 101–111.

Afungang R. N, de Meneses Bateira C.V, Nkwemoh C. 2017. Assessing the spatial probability of landslides using GIS and informative value model in the Bamenda highlands. *Arabian Journal of Geosciences.* 10(17), 384.

Alkhasawneh M. S, Ngah U. K, Tay L.T, Isa M, Ashidi N, Al-batah M. S .2013. Determination of important topographic factors for landslide mapping analysis using MLP network. *The Scientific World Journal.*

Al-Najjar, H. A. H., Kalantar, B., Pradhan, B., & Saeidi, V. (2019, October). Conditioning factor determination for mapping and prediction of landslide susceptibility using machine learning algorithms. *In Earth Resources and Environmental Remote Sensing/GIS Applications X* (Vol. 11156, p. 111560K). International Society for Optics and Photonics.

Beguería S. 2006. Validation and evaluation of predicitive models in hazard assessment and risk management. Nat. Hazards. 37, 315-329.

Bergsma W. 2013. A bias-correction for Cramér's V and Tschuprow's T. *Journal of the Korean Statistical Society*. Vol. 42, no. 3, 323-328.

Chen W, Pourghasemi H.R, Naghibi S.A. 2018. Prioritization of landslide conditioning factors and its spatial modeling in Shangnan county, China using GIS-based data mining algorithms. Bull. Eng. *Geol. Environ*. 77, 611–629.

Cortes C, Vapnik V. 1995. Support-Vector Networks CORINNA, *IEEE Expert*. vol. 7, no. 5, pp. 63–72.

Dou J, Tien Bui D, Yunus A.P, Jia K, Song X, Revhaug I, Xia H, Zhu Z. 2015. Optimization of causative factors for landslide susceptibility evaluation using remote sensing and GIS data in parts of niigata, Japan. *PLoS ONE*. 10, e0133262

Golkarian A, Naghibi S.A, Kalantar B, Pradhan B. 2018. Groundwater potential mapping using C5.0, random forest, and multivariate adaptive regression spline models in GIS. *Environ. Monit. Assess.* 190.

Hong H, Pradhan B, Sameen M.I, Kalantar B, Zhu A, Chen W. 2018. Improving the accuracy of landslide susceptibility model using a novel region-partitioning approach. *Landslides*. 15, 753–772.

Kalantar, B., Pradhan, B., Naghibi, S. A., Motevalli, A., & Mansor, S. (2018). 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). *Geomatics, Natural Hazards and Risk*, 9(1), 49-69.

Kalantar, B., Ueda, N., Lay, U. S., Al-Najjar, H. A. H., & Halin, A. A. (2019, July). Conditioning Factors Determination for Landslide Susceptibility Mapping Using Support Vector Machine Learning. *In IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium* (pp. 9626-9629). IEEE.

Kim D. H, Kim Y. J, Hur D. S. 2014. Artificial neural network based breakwater damage estimation considering tidal level variation. *Ocean Engineering*. 87, 185-190.

Mahalingam R, Olsen M.J, O'Banion M.S. 2016. Evaluation of landslide susceptibility mapping techniques using lidar-derived conditioning factors (Oregon case study). *Geomatics*, 7, 1–24.

Mousavi S.M, Golkarian, A, Amir Naghibi S, Kalantar B, Pradhan B. 2017. GIS-based Groundwater Spring Potential Mapping Using Data Mining Boosted Regression Tree and Probabilistic Frequency Ratio Models in Iran. *AIMS Geosci*. 3, 91–115.

Mousavi S.Z, Kavian A, Soleimani K, Mousavi S.R, Shirzadi, A. 2011. GIS-based spatial prediction of landslide susceptibility using logistic regression model. Geomat. *Nat. Hazards Risk*. 2, 33–50.

Nguyen H, Mehrabi M, Kalantar B, Moayedi H, Abdullahi M. A. M. 2019. Potential of hybrid evolutionary approaches for assessment of geo-hazard landslide susceptibility mapping. Geomatics, *Natural Hazards and Risk*. 10(1), 1667-1693.

Ozdemir A, Altural T. 2013. A comparative study of frequency ratio, weights of evidence and logistic regression methods for landslide susceptibility mapping: Sultan mountains, SW Turkey. *J. Asian Earth Sci*. 64, 180–197.

Pourghasemi H. R, Rossi M. 2019. Natural Hazards GIS-Based Spatial Modeling Using Data Mining Techniques, *Advances in Natural and Technological Hazards Research*, *Springer*.

Pradhan B, Lee S. 2010. Landslide susceptibility assessment and factor effect analysis: Backpropagation artificial neural networks and their comparison with frequency ratio and bivariate logistic regression modelling. *Environ. Model. Softw*. 25, 747–759.

Sezer E.A, Pradhan B, Gokceoglu C. 2011. Manifestation of an adaptive neuro-fuzzy model on landslide susceptibility mapping: Klang valley, Malaysia. *Expert Syst. Appl*. 38, 8208–8219.

Shafii B, Price W. J. 2001. Application of Bayesian Methods for Assessing Detection Accuracy in Remote Sensing Ground Truth, *International Journal of Applied & Experimental Mathematics*. no.1.

Tian Y, Xu C, Hong H, Zhou Q, Wang D. 2019. Mapping earthquake-triggered landslide susceptibility by use of artificial neural network (ANN) models: an example of the 2013 Minxian (China) Mw 5.9 event. *Geomatics, Natural Hazards and Risk*. vol 10, no. 1, 1-25.

Tsangaratos P, Ilia I. 2016. Comparison of a logistic regression and naïve bayes classifier in landslide susceptibility assessments: The influence of models complexity and training dataset size. *Catena* 145, 164–179.

Viera A.J, Garrett J.M. 2005. Understanding interobserver agreement: the Kappa statistic. *Family Medicine*. 237, 360–363.