ISSN: 2476-6909; ECOPERSIA. 2020;8(2):65-76

Successive Intermodal Ensembling: A Promising Approach to Improve the Performance of Data Mining Models for Landslide Susceptibility Assessment (A case study: Kolijan Rostaq Watershed, Iran)

ARTICLE INFO

Article Type
Original Research

Authors
Adineh F.[1] *MSc,*
Motamedvaziri B.[1]* *PhD,*
Ahmadi H.[2] *PhD,*
Moeini A.[1] *PhD*

How to cite this article
Adineh F, Motamedvaziri B, Ahmadi H, Moeini A. Successive Intermodal Ensembling: A Promising Approach to Improve the Performance of Data Mining Models for Landslide Susceptibility Assessment (A case study: Kolijan Rostaq Watershed, Iran). ECOPERSIA. 2020;8(2):65-76.

[1]Forest, Range & Watershed Management Department, Natural Resources and Environment Faculty, Science and Research Branch, Islamic Azad University, Tehran, Iran
[2]Reclamation of Arid & Mountainous Regions Department, Agriculture Faculty, University of Tehran, Karaj, Iran

Correspondence
Address: Reclamation of Arid & Mountainous Regions Department, Agriculture Faculty, University of Tehran, Karaj, Iran
Phone: -
Fax: -
bmvaziri@gmail.com

Article History
Received: September 19, 2019
Accepted: December 07, 2019
ePublished: May 19, 2020

ABSTRACT

Aims In the present study, random forest (RF) and support vector machine (SVM) were used to assess the applicability of ensemble modeling in landslide susceptibility assessment across the Kolijan Rostaq Watershed in Mazandaran Province, Iran.

Materials & Methods Both models were used in two modeling modes: 1) A solitary use (i.e., SVM and RF) and 2) Their ensemble with a bivariate statistical model named the weights of evidence (WofE) which then generated two more models, namely SVM-WofE and RF-WofE. Further, the resulting maps of each stage were dually coupled using the weighted arithmetic mean operation and an intermodal blending of the previous stages.

Findings Accuracy of the models was assessed via the receiver operating characteristic (ROC) curves based on which the goodness-of-fit of the SVM and the SVM-WofE models were 0.817 and 0.841, respectively, while their respective prediction accuracy values were found to be 0.848 and 0.825. The goodness-of-fit of the RF and the RF-WofE models respectively was 0.9 and 0.823, while their respective prediction accuracy values were found to be 0.886 and 0.823. The goodness-of-fit and prediction power of SVM and SVM-WofE ensemble were respectively 0.859 and 0.873. The same increasing pattern was evident for the ensemble of RF and RF-WofE where their goodness-of-fit and prediction power increased, respectively, up to 0.928 and 0.873. Moreover, the goodness-of-fit and prediction power of RF-SVM ensemble were increased up to 0.932 and 0.899, respectively. The results of the averaged Kappa values throughout a 10-fold cross-validation test as an auxiliary accuracy assessment attested to the same results obtained from the ROC curves.

Conclusion Successive intermodal ensembling approach is a simple and self-explanatory method so far as the context of many data mining techniques with a highly complex structure has been simply benefitted from the weighted averaging technique.

Keywords Random Forest; Support Vector Machine; Spatial Modeling; Weights of Evidence

CITATION LINKS

[1] Landslides-cause and ... [2] Landslide hazard and risk zonation ... [3] Spatial prediction models for ... [4] Landslide susceptibility mapping ... [5] Landslide susceptibility mapping using ... [6] Manifestation of LiDAR-derived parameters ... [7] Application of wavelet analysis and a ... [8] Spatial prediction models for shallow ... [9] Review on landslide susceptibility ... [10] Assessment of land subsidence ... [11] Newer classification and regression ... [12] Conditional variable importance ... [13] Hillslope characteristics as ... [14] How can statistical models ... [15] Landslide susceptibility estimation by ... [16] Random forests and evidential belief ... [17] GIS-based groundwater potential ... [18] Investigation of general indicators ... [19] Groundwater potential mapping using ... [20] Land subsidence susceptibility assessment ... [21] Sensitivity analysis of effective factors ... [22] Antecedent rainfall thresholds for ... [23] A caution regarding rules of thumb ... [24] Random ... [25] Letter to the editor: Stability of random ... [26] Letter to the editor: On the stability and ... [27] Tutorial on support vector machine ... [28] Text categorization with support vector ... [29] Knowledge-based analysis of microarray gene ... [30] Support vector machines and kernel methods ... [31] An assessment of support vector machines for land ... [32] Support vector machines for predicting distribution ... [33] A gentle introduction to support ... [34] A gentle introduction to support ... [35] Support vector machines-an ... [36] Landslide susceptibility assessment using SVM machine ... [37] Integration of geological datasets for ... [38] Geo-information tools for landslide ... [39] Landslide susceptibility assessment ... [40] Land-cover change model validation ... [41] The caret ... [42] A test of transferability for ... [43] Binary logistic regression versus ... [44] Presence-only approach to assess landslide ... [45] Applied logistic ...

Copyright© 2020, TMU Press. This open-access article is published under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License which permits Share (copy and redistribute the material in any medium or format) and Adapt (remix, transform, and build upon the material) under the Attribution-NonCommercial terms.

Introduction

A landslide is the downward movement of slope materials owing to the pull of gravity and different predisposing and triggering factors [1]. Landslides are among the most frequent natural events and occur in different types and intensity which can impose sizeable casualties and socio-economic losses. Over the decades, hundreds of conceptual and numerical models have been developed with different computational algorithms and many studies have been devoted to landslide susceptibility assessment. Based on a general consensus, these models can be set in four different categories [2]: 1) Inventory-based approaches, 2) Heuristic and expert knowledge-based approaches which estimate landslide potential from data on preparatory variables based on expert knowledge and opinions, 3) Statistical-probabilistic data-driven approaches including multivariate statistical methods such as logistic regression or bivariate statistical methods such as information value, frequency ratio, fuzzy logic, certainty factor, and weights of evidence, and 4) Site-specific deterministic approaches such as factor of safety.

Nowadays, data mining models have gained due attention in the machine learning community due particularly to coping with data scarcity, being scale-invariant, and handling a wide range of the conditioning factors of a phenomenon with different characteristics. Several studies have been devoted to this field using different data mining models such as maximum entropy, boosted regression tree, classification and regression tree, support vector machine (SVM) [3-10], general linear model, random forest (RF) [11-20], stability index mapping (SINMAP) [21], and some other detailed analyses on the climatic thresholds of rainfall-triggered landslides [22]. These models have shown better performances than other bivariate or multivariate statistical methods. On the other hand, ensemble models have shed light on the ways to generate an integrated model with more accurate results. In particular, they can efficiently integrate models to attain high performance in terms of their goodness of fit and predictive skill.

Ensemble models are meta-algorithms that are designed to decrease variance, bias, or to improve the predictive skill of single machine learning models. In generation of ensemble models, the single models (known as base learners) have to be as accurate as possible and as diverse as possible. Although the foundation of ensemble methods was first designed for machine learning models, landslide-related studies have been disseminated in recent years suggesting the ensemble of statistical and data mining models as a new horizon to ensemble modeling. On the other hand, landslide science and the inferences therein contained have long been known to be relative and rather local which cannot be generalized to different scales, especially different watersheds. Also, there have not been adequate studies in this field to support such claims, which prompted us to conduct further study in this area. Moreover, the computational algorithm of the ensembling process in different literature has remained rather unclear and somewhat bounded by a certain family of models. By considering these research gaps, the current study sets out to attain two main objectives: 1) To test the boosting role of a bivariate statistical model (i.e., the weight of evidence) when ensembled with data mining models (i.e., SVM and RF) and 2) To propose a simple yet practical type of model ensembling technique which is underpinned by a successive integration of different models.

Materials and Methods

The methodological flowchart of this study is presented in Figure 1 and thoroughly described in the following sections.

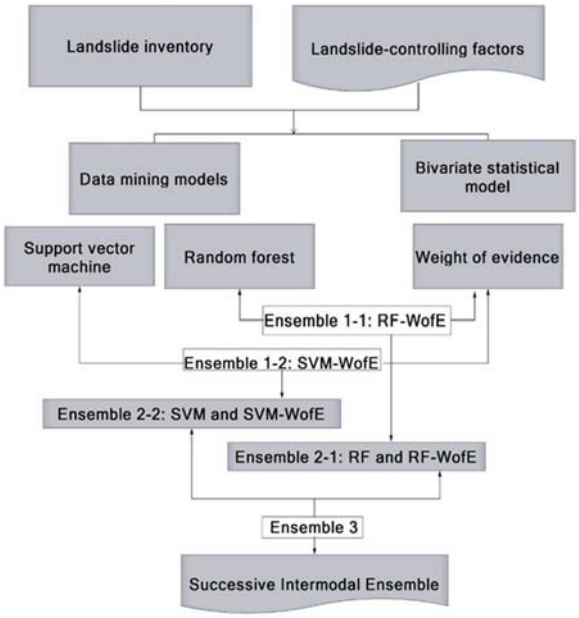

Figure 1) Methodological flowchart adopted in the present study

Study area

The Kolijan Rostaq Watershed is one of the mountainous landslide-prone basins in the Mazandaran Province in Iran, which extends for an area of 255km^2. It is located between 52°34′ E to 52°42′ E longitudes and 36°12′ N to 36°25′ N latitudes, UTM (Universal Transverse Mercator) Zone 40. Elevation ranges between 72m and 1451m a.s.l. (Figure 2).

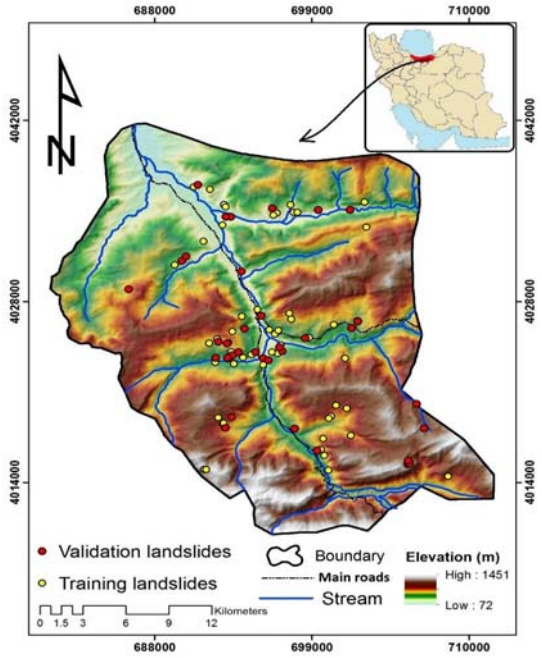

Figure 2) Geographical location of the study area in the Mazandaran Province in Iran

According to reports of Mazandaran's FRWO Office (Forest, Range and Watershed Management Organization), the average annual rainfall varies between 585 and 830mm. Forests account for about 77% of the study area, followed by agricultural lands (19.4%), orchards (2.2%), and the remaining area is distributed among other land uses such as residential areas, rangelands, and natural/artificial reservoirs. From a geological viewpoint, most of the area is covered by the $M_{m,s,l}$ group (71.6%), followed by P_{el} (8.1%), and P_{lc} (7.7%), and the remaining area is shared between Q_m, K_{2l1}, J_k, and K_{2l2} groups. Table 1 details more information regarding these geological groups.

Data compilation

Landslide inventory is a pivotal part of any spatial modeling endeavor. In this study, 90 landslides were recorded as point features by using the available geoinformatics (Google Earth, a handheld GPS device, archived organizational data, and local information) during extensive field surveys and its final digitized map was generated in ArcGIS 10.3. Further, the sample points were partitioned into two sets of training (60%) and validation samples (40%).

As previously mentioned, landslides occur in consequence of the synergic interaction between geo-environmental, topological, climatic, and man-made factors (i.e., predictors). Although there is no general consensus on which set of predictors could contribute the most to the modeling process, field surveys and archived data paved the way to select more relevant factors. Also, it was tried to select these factors from different categories mentioned above to the extent that the available data allowed such inclusion. Data heterogeneity of each factor across the study area was another criterion that was taken into account in the factor selection process. By doing so, ten causative factors were selected namely elevation, slope percentage, slope aspect, plan curvature, average annual rainfall, proximity to roads/streams/ faults, lithological formations, and land use. Further, to cope with data multicollinearity, the variance inflation factor (VIF) was used. The VIF values above 5 indicate critical multicollinearity and a strong correlation among the factors which will accordingly cause bias in the results of the models [23]. The tolerance index is the reciprocal of VIF which also indicates the multicollinearity issue. A tolerance value of less than 0.2 indicates problematic multicollinearity (Table 2). Table 2 depicts multicollinearity analysis among landslide-controlling factors using the tolerance and VIF tests.

Table 1) Description of the lithological formations in the study area

Formation code	Materials	Age	Era
$M_{m,s,l}$	Marl, calcareous sandstone, sandy limestone, and minor conglomerate	Miocene	Cenozoic
P_{el}	Medium to thick-bedded limestone	Paleocene/Eocene	Cenozoic
P_{lc}	Polymictic conglomerate and sandstone	Pliocene	Cenozoic
Q_m	Swamp and marsh	Quaternary	Cenozoic
K_{2l1}	Hypocrite bearing limestone	Late Cretaceous	Mesozoic
J_k	Conglomerate, sandstone, and shale with plant remains and coal seams	Middle Jurassic	Mesozoic
K_{2l2}	Thick-bedded to massive limestone	Late Cretaceous	Mesozoic

Table 2) Multicollinearity analysis between landslide-controlling factors using the tolerance and VIF tests

Factors	Std. Error	Sig.	Collinearity Statistics	
			Tolerance	VIF
Slope aspect	0.030	0.100	0.890	1.123
Elevation	0.000	0.061	0.251	3.990
Proximity to faults	0.000	0.390	0.614	1.629
Lithological formation	0.032	0.189	0.691	1.448
Land use	0.033	0.000	0.741	1.350
Plan curvature	0.057	0.224	0.901	1.110
Average annual rainfall	0.001	0.028	0.400	2.497
Proximity to streams	0.000	0.408	0.514	1.947
Proximity to roads	0.000	0.010	0.602	1.661
Slope percentage	0.002	0.758	0.793	1.261

Random forest

Random forest (RF) was first introduced and developed by Breiman [24]. It adopts classification and regression trees (CARTs) and has been widely used in environmental science [11, 12, 13] and natural hazard modeling, in particular, landslide susceptibility mapping [14, 15]. Random forest is a non-parametric multivariate classification algorithm that is based on averaging the results of many decision trees and gives an accurate classification for a wide range of datasets [15]. Its algorithm uses a bootstrapping technique to select a subset of observations as the training set by taking advantage of random binary trees and the remaining data are excluded as out-of-bag (OOB) [15, 16, 24]. The error of classifying the data as landslide and non-landslide categories is computed by out-of-bag (OOB) error via comparing the out-of-bag predicted responses against the true responses. More mathematical details can be found in Breiman [24], Calle and Urrea [25] and Nicodemus [26]. In the current study, the RF model was implemented in the "randomForest" package.

Support vector machine (SVM)

The support vector machine is a supervised machine learning technique that was first introduced by Boser, Guyon, and Vapnik in 1992 [27]. The SVM uses classifiers that are a geometrical (vector-based) representation of the data in which the data are plotted based on their features [28-32]. Normally, a straight line (i.e., a classifier) is set between two sets of points in a 2D space (input space, R2). However, divers set of points with unorganized and complex features are hardly separable with only a straight line. Hence, SVM sets a line between datasets with a maximum margin on either side of its surroundings using quadratic programming [33, 34]. When facing a highly unorganized pattern, a linear decision surface in a two-dimensional space would not be practical anymore. Therefore, the SVM transforms the data into a higher-dimensional space using a kernel trick and then categorizes the data by a separator termed as the hyperplane [34, 35]. A kernel is a dot product in a feature space. The most popular kernel functions called Gaussian kernel (also known as the radial basis function) are used in this study. More details can be found in Yao *et al*. [4] and Marjanović *et al*. [36].

Weights-of-evidence

Weights of evidence (WofE) is a bivariate statistical model that is based on calculating the landslide density in the classes of a predictor. The WofE was first expounded by Bonham-Carter *et al*. [37] and was initially applied to mine exploration. Later on, Van Westen [38] adopted the technique to analyze landslide susceptibility. In this method, the positive and negative weights (Wi+ and Wi-) are assigned to each pixel of the classes within the factor map. The Wi+ refers to the importance of the presence of a class to landslide occurrence. Positive values of Wi+ in a particular class portray a condition that is highly prone to landslide occurrence and vice versa [10, 38, 39]. The Wi- refers to the importance of the absence of a class for landslide occurrence. The Wi+ and Wi- values closer to zero indicate that there is no relationship between the presence/absence of a class and landslide occurrence. In order to quantify the relationship between the conditioning factor classes (i.e., a predictor) and landslide occurrence, a contrast factor was calculated (Equation 1).

Equation (1) $$C_i = W^+ - W^-$$

If the contrast factor is zero, the spatial overlap of the class within a particular predictor and landslide occurrence would then be out of chance which suggests the lack of any positive or negative relationship. Positive and negative contrast values, respectively have a connotation of the positive and negative association between the spatial pattern of factor classes and landslides occurrences across the study area. Lastly, the rate values were assigned to factors from which the secondary maps were generated.

Model Ensembling

In this study, model ensembling follows a successive integration process which undertakes models both solely and in an intermodal fashion. The first ensembling stage couples each model with the bivariate statistical model (e.g., SVM and WofE). For the latter, after assigning the rates to each factor's class, the secondary maps together with the landslide inventory map are fed into the data mining models. The second ensembling stage regards coupling each data mining model (e.g., SVM) with its previous ensemble (e.g., SVM-WofE). From this stage forward, ensembling follows the weighted arithmetic mean of the models' results. The area under the receiver operating characteristic curve (AUROC) is considered as the weighting element. That is, once the susceptibility map of the first ensemble stage and the solitarily used data mining models are derived, they will put into the ROC test and the resulting AUROC values will then be calculated and used as the weighting value. Prior to this process, the best models at the previous stage (despite being solitary or ensemble) are selected based on their AUROC values. The weighted arithmetic mean-based ensembling process can be formulated as the following expression:

Equation (2)

$$Ensemble_2 = \frac{\sum_{i=1}^{n} AUROC_i \times LSM_i}{\sum_{i=1}^{n} AUROC_i}$$

Where $AUROC_i$ and LSM_i, respectively, are the AUROC value and landslide susceptibility map of the i^{th} model. The third ensembling stage adopts an intermodal process as it integrates the previously selected superior models so as to further improve their goodness-of-fit and prediction power, following Equation 2.

Performance analysis

The ROC curve is the basis of model selection in the successive process of model ensembling as well as a baseline to test whether the latter has been successful or not. The ROC curve plots the sensitivity termed as true positives (i.e., correctly determining the landslide areas by the model) on the y-axis against the 1-specificity termed as false positives (incorrectly classifying a non-landslide area as a landslide by the model as opposed to what observed in reality) on the x-axis [40]. The area under the ROC curve (AUROC) is a representative of the model's goodness-of-fit and prediction power based on which an AUROC value of 1 represents perfect performance, while values close to 0.5 represents a model that is performed purely out of random chance. It is noteworthy that the considered AUROC values for the weighted arithmetic mean ensembling process relies on the validation set since the latter is more informative and refer to the spatial transferability of the models' results, although the final judgment of models' performance is based on the AUROC values derived from both training and validation sets. In addition, to the AUROC values derived from the conventional 60:40% data balance (i.e., training and validation datasets), in this study, it is used a 10-fold cross-validation technique through which the average value of Cohen's Kappa was calculated for each model as a metric of its accuracy. The latter was implemented in the *caret* R-package (short for classification and regression training) [41]. In the k-fold cross-validation process, the data are randomly split into *k* segments in which the segment-*k* is held out for validation while the 1-*k* is used for training the models, ensuring that each segment is used once for validating the results of the models [10, 42-44].

Findings

Table 2 presents in detail the VIF and tolerance values for different factors. The highest and the lowest values, respectively, correspond to elevation and plan curvature, still, all values are less than 5 (i.e., the acceptable range) which indicates that all the factors can be used in the modeling process and no factor removal is required.

Calculated rates of the factors' classes obtained from the WofE model are presented in Table 3. Regarding the slope aspect, the highest rate corresponds to the north-facing slopes. As for the plan curvature, flat areas and concave curvatures, respectively, have the lowest and the highest values. The relationship between elevation and landslide occurrence was rather indirect, which is conceivable from lacking a definite pattern in the rates of this factor. The areas nearby the faults had the highest rates. Although the $M_{m,s,l}$ formation accounts for most of the study area, the J_k gained the highest rate. Areas nearby the streams and the roads had the highest rate. Rates of slope classes gave a discernible pattern where steeper slopes had the highest rate. Agricultural lands and residential areas had the highest rate among other land use classes. However susceptibility rates of average annual rainfall classes gradually increased towards upper classes, the trend faces a decline where a lower rate for the highest values of rainfall (i.e., 797-835mm) was evident.

Table 3) Calculated rates of the classes based on the algorithm of the weight of evidence model

Factor	Class	Areal percentage of landslides	Areal percentage of the class	W-	W+	C
Slope aspect	Flat	3.333	2.042	-0.013	0.490	0.503
	North	40.000	28.055	-0.182	0.355	0.536
	East	16.667	21.663	0.062	-0.262	-0.324
	South	18.889	24.239	0.068	-0.249	-0.318
	West	21.111	24.001	0.037	-0.128	-0.166
Plan curvature	Concave	15.556	12.457	-0.036	0.222	0.258
	Flat	61.111	68.350	0.206	-0.112	-0.318
	Convex	23.333	19.193	-0.053	0.195	0.248
Elevation (m)	67-300	23.333	15.163	-0.101	0.431	0.532
	300-600	55.556	42.794	-0.252	0.261	0.513
	600-900	13.333	30.731	0.224	-0.835	-1.059
	900-1200	4.444	9.801	0.058	-0.791	-0.849
	>1200	3.333	1.511	-0.019	0.791	0.810
Proximity to faults (m)	0-100	18.889	14.087	-0.058	0.293	0.351
	100-200	7.778	13.301	0.062	-0.537	-0.598
	200-300	12.222	11.931	-0.003	0.024	0.027
	300-400	11.111	10.365	-0.008	0.069	0.078
	>400	50.000	50.316	0.006	-0.006	-0.013
Lithological formation	$M_{m,s,l}$	1.111	2.300	0.012	-0.728	-0.740
	P_{el}	1.111	3.366	0.023	-1.108	-1.131
	P_{lc}	0.000	2.287	0.023	-12.235	-12.258
	Q_m	62.222	71.631	0.286	-0.141	-0.427
	K_{2l1}	3.333	8.061	0.050	-0.883	-0.933
	J_k	24.444	7.668	-0.201	1.160	1.360
	K_{2l2}	7.778	4.688	-0.033	0.506	0.539
Proximity to streams (m)	0-100	17.778	9.837	-0.319	0.592	0.911
	100-200	20.000	10.532	-0.345	0.641	0.987
	200-300	13.333	11.437	-0.235	0.153	0.389
	300-400	20.000	12.320	-0.325	0.485	0.810
	>400	28.889	55.874	0.210	-0.660	-0.870
Proximity to roads (m)	0-100	43.333	26.441	-0.609	0.494	1.103
	100-200	23.333	18.226	-0.310	0.247	0.557
	200-300	10.000	13.194	-0.169	-0.277	-0.109
	300-400	12.222	9.712	-0.239	0.230	0.469
	>400	11.111	32.427	0.067	-1.071	-1.138
Slope (%)	0-5	11.111	14.480	0.039	-0.265	-0.303
	5-15	46.667	50.340	0.071	-0.076	-0.147
	15-30	32.222	29.911	-0.034	0.074	0.108
	30-45	6.667	4.901	-0.019	0.308	0.327
	>45	3.333	0.369	-0.030	2.201	2.231
Land use	Reservoir	0.000	0.015	0.000	-7.220	-7.220
	Orchard	4.444	2.174	-0.023	0.715	0.739
	Forest	32.222	76.690	1.068	-0.867	-1.935
	Agriculture	57.778	19.363	-0.647	1.093	1.741
	Rangeland	0.000	0.078	0.001	-8.861	-8.862
	Residential	5.556	1.040	-0.047	1.676	1.722
Average annual rainfall (mm)	585.7-675	4.444	6.628	-0.017	-0.400	-0.383
	675-725	10.000	12.324	-0.051	-0.209	-0.158
	725-765	17.778	19.249	-0.109	-0.080	0.030
	765-797	46.667	31.196	-0.485	0.403	0.888
	797-835	21.111	30.603	-0.096	-0.371	-0.275

Figure 3 portrays the susceptibility maps derived from different solitary and ensemble models. The preliminary graphical check indicates that almost all the models unanimously introduced the central parts of the basin as highly susceptible areas to landslide occurrence and this pattern extends towards the lower section of the watershed. The first analysis attests that all the models, either solitarily used or ensembled, are well performed. According to Hosmer and Lemeshow [45], AUC values ranged from 0.8 to 0.9 indicate a good model while values greater than 0.9 signify an excellent model. As shown in Diagram 1, the AUC values of all the models are placed between the abovementioned ranges which attest to the high goodness-of-fit and prediction power of the RF and SVM models, although some differences in their produced susceptibility maps and the AUROC values exist. In the second ensembling stage (Diagrams 1c-f), the proposed weighted arithmetic mean-based ensemble increased both the learning capability and prediction power.

In the third ensembling stage (Diagram 2), the learning capability and prediction power reached their peaks.

Figure 4 and Table 4, respectively, present the distribution and areal extent of the susceptibility classes derived from the proposed successive intermodal ensembling method based on which about 15% of the study area is identified as highly susceptible to landslide occurrence and should be taken into account for further risk assessment and mitigation practices.

Table 5 presents in detail the averaged Kappa values of each model throughout the 10-fold cross-validation procedure in which the highest and the lowest Kappa values, respectively, correspond to the SVM and the ensemble of RF: RF-WofE and SVM: SVM-WofE. The Higher Kappa value indicates a higher accuracy of the model.

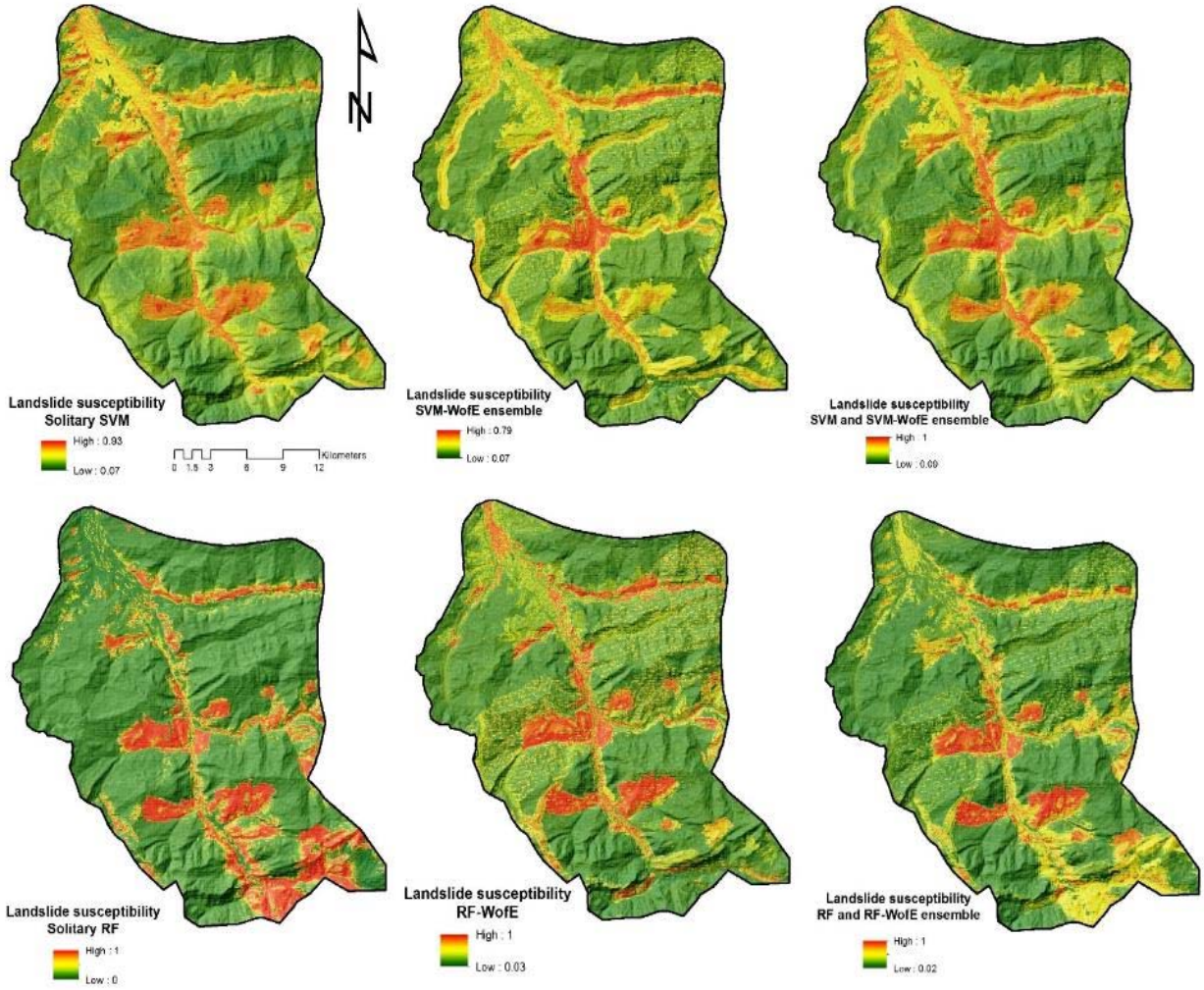

Figure 3) Landslide susceptibility maps obtained from the different ensembling process

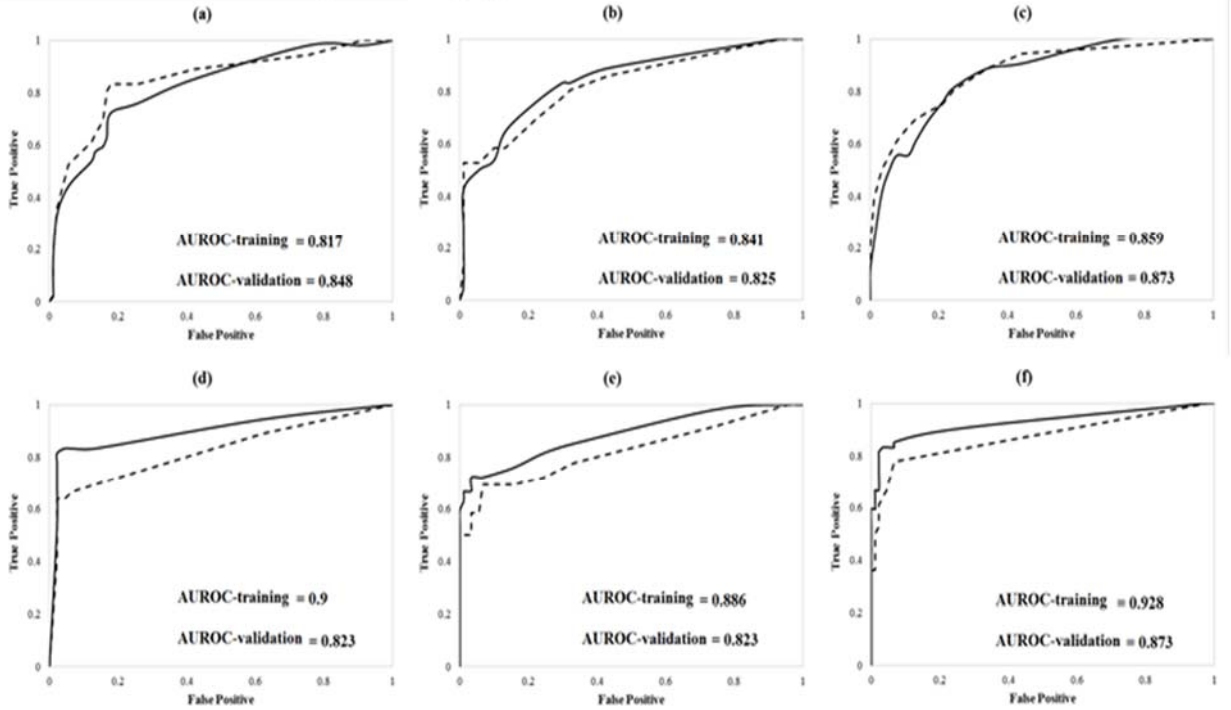

Diagram 1) The ROC curves generated at the training and validation stages for different models; a) SVM; b) SVM-WofE; c) ensemble of SVM and SVM-WofE (SVM: SVM-WofE), d) RF; e) RF-WofE; f) ensemble of RF and RF-WofE (RF: RF-WofE)

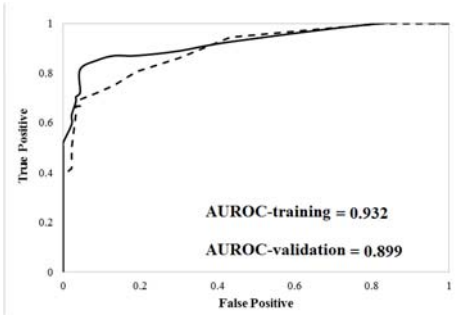

Diagram 2) The ROC curve generated at the training and validation stage for the ensemble of RF: RF-WofE and SVM: SVM-WofE based on the successive intermodal ensemble approach

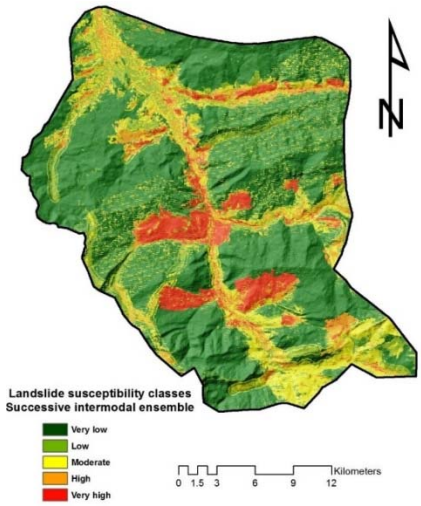

Figure 4) Landslide susceptibility classes generated by the ensemble of RF: RF-WofE and SVM: SVM-WofE based on the successive intermodal ensemble approach

Table 4) Areal extent of the susceptibility classes derived from the successive intermodal ensemble of RF and SVM models

Landslide susceptibility class	Area (ha)	Percentage
Very low	35838.27	58.10
Low	7441.11	12.06
Moderate	9351.54	15.16
High	4877.82	7.91
Very high	4172.67	6.76

Table 5) The Kappa values averaged throughout the 10-fold cross-validation procedure for each model

Data mining model	Average Kappa obtained from a 10-fold cross-validation
SVM	0.2286
SVM-WofE	0.3761
RF	0.4008
RF-WofE	0.395
Ensemble of SVM: SVM-WofE	0.4031
Ensemble of RF: RF-WofE	0.4207
Ensemble of RF: RF-WofE and SVM: SVM-WofE	0.621

Discussion

The highest VIF values are normally associated with factors that are highly identical and strongly correlated. For instance, when rainfall maps are generated merely from gradient equations, a high VIF value between the rainfall and elevation is highly probable. Moreover, DEM-derivatives, if obtained from simple equations, tend to show high VIF values with

their source (i.e., DEM). Such multicollinearity, however, was not discerned in this study.

Results regarding slope aspect were expectable since north-facing slopes receive more precipitation and contain more moisture as it can accordingly lead to slope failure. Regarding concave curvatures, they can contain more moisture which can increase the shear force and the probability of slope failure. The relationship between elevation and landslide occurrence becomes more transparent through such factors as lithology, vegetation, or rainfall, which in turn, leaves the elevation factor as more of a proxy. Through these interpretations, the plausible reasons for the gained rates for different classes of elevation can rest in the fact that higher altitudes are covered with more resistant formations and lower altitudes receive the precipitation far sooner than it reaches uplands. The areas nearby the faults having the highest rates indicate that seismic activities can also be another triggering factor that has been contributing to landslide occurrences in the study area.

Considering the materials of the lithological formations presented in Table 1, the main reason for the J_k gaining the highest rate stems from the presence of shale and coal seams. Shale forms an impermeable surface that impedes the infiltration processes and forces the infiltrated water to flow along the shale surface. The latter perpetually forms a surface of rupture which will eventually lead to slope failure. Coal seams can also initiate the process of landsliding via lubricating the soiled materials in the interlayers. Areas nearby the streams are prone to the undercut of the slope caused by streamflow which leads to frequent shallow landslides nearby. High susceptibility to landsliding in areas near the roads roots in the unsupervised man-made slopes and loose road foundations.

Regarding the rates of slope classes, it can be conceived that steeper slopes exert extra forces to slope materials which, in turn, makes them less resistant. Agricultural activities and artifacts cause an artificial rise in the water table and, in some cases, uneven soil compactness across the area. Besides, once the farmlands are left behind as bare lands, vegetation cover at the soil surface is gradually diminished and leaves the surface unprotected against intense rainfalls. Residential areas by triggering the adjacent slopes can redistribute the old landslides and initiate new slope failures. Furthermore, buildings located on slopes with slight-to-steep inclinations can trigger more landslides due to the increased overburden pressure. Rates of average annual rainfall classes gradually increase towards upper classes which stems from the triggering role of intense or long-lasting rainfalls. The lower values of the rate at the top class of rainfall (i.e., 797-835mm) can be justified due to the gentle slopes of the lower altitudes.

Although the susceptibility maps of different models show somewhat similar patterns, they can exhibit different performances. The latter is evident in Diagram 1. Slight differences in the AUC values of different models stem from the ensembling process and seeking through these changes can pinpoint some valuable insights as follows. Regarding the ROC plots of the solitarily used SVM and its ensemble with the WofE (Diagrams 1a and 1b), it is evident that although the first ensembling process (i.e., SVM-WofE) increases the learning capability of SVM, it does not distinctly affect its prediction power, but, instead, it faces a discernible decline. As for the RF and RF-WofE (Diagrams 1d-e), the first ensembling process shows even more disappointing results where although it keeps the prediction power intact, the learning capability faces a decline. This contradicts with the inferences of some studies reported that the ensemble of data mining models with bivariate statistical methods can be successful. In fact, the latter reminds that the inference of landslide susceptibility assessment is rather local and cannot be generalized over other areas, that is, different models that are fed with different data configurations (i.e., distinct geological/geomorphological arrangements attributed to different geographical locations) can result in different performances. The second ensembling stage (Diagrams 1c-f), shows promising results regarding the practicality of the proposed successive intermodal ensembling method which has improved both the prediction power and the learning capability of the models. The third ensembling stage (Diagram 2) puts even more emphasis on the promising result of this technique where the learning capability and prediction power reach their peaks, compared to those of the previous stages. Table 5 also attests to the abovementioned statements in which the 10-fold cross-validation also shows

that the overall accuracy of the models increases once they are ensembled based on the proposed integration algorithm.

It is noteworthy that the success of the proposed ensembling method is highly indebted to its constituent models in which the strengths of different models have synergistically boosted the performance of their ensemble so that each model's shortcomings are covered by the strength of its counterpart. In this regard, as opposed to single decision trees, the RF model proposes the averaged results of many decision trees in such a way that the outcome has less variance, less overfitting, high flexibility, and accordingly high accuracy and generalization power. These strengths are supported by the advantages of the SVM model which provides a robust pattern recognition algorithm to solve landslide susceptibility as a complex problem. Moreover, regardless of the obscure performance of the WofE once combined with the data mining models and by putting the excellent performance of the final ensemble model in perspective, it is fair to say that the ability of the WofE model in finding any random correlation between the landside occurrence and the presence/absence of different factors would provide an algorithm for data mining models to render the raw inputs into more interpretable pieces of evidence of the studied phenomenon. On a last note, using the successive intermodal ensembling technique in the pursuit of achieving high modeling performance is advised.

Conclusion

The current path of landside susceptibility modeling necessitates generating more efficient models in terms of high learning and prediction skills. In an effort to test the previously reported promising ensemble of data mining and bivariate statistical methods, in this study, the RF and SVM data mining models were fused into the WofE algorithm. Also, a successive intermodal ensembling approach was proposed which accompanies data mining and bivariate statistical models in a successive manner until its product excels the previous performance records. The main take-home messages of this study are as follows. As opposed to previous studies that underlined promising results for the ensemble of data mining and bivariate statistical models, the latter can be specific to a certain area, but the results of the present study showed the opposite. Instead, the proposed successive intermodal ensembling approach which lies in simple weighted arithmetic mean of the models' outputs can discernibly increase the performance of the models including both the learning/fitting capability and prediction/generalization power. It is a simple, self-explanatory, and highly beneficial method so far as the context of many data mining techniques with a highly complex structure has been simply benefitted from such weighted averaging technique, nonetheless its application in model ensembling has been neglected by the scholar. In the current study, the intermodal part of the suggested approach refers to the fact that successive ensembling can increase the performance of the models to a certain degree, but an intermodal ensembling mode (i.e., among different distinct models) can ameliorate previous performances towards excellency.

Acknowledgments: None Declared by Authors
Ethical permissions: None Declared by Authors
Conflicts of interests: None Declared by Authors
Authors' Contribution: Fatemeh Adineh (First author), Introduction Writer/Main Researcher (25%); Baharak Motamedvaziri (Second author), Methodologist (25%); Hasan Ahmadi (Third author), Statistical Analyst (25%); Aboalfazl Moeini (Fourth author), Discussion Writer (25%)
Funding/Support: None Declared by Authors

References

1- Radbruch-Hall DH, Varnes DJ. Landslides-cause and effect. Bull Int Assoc Eng Geol. 1976;13(1):205-16.
2- Van Westen CJ, Van Asch TW, Soeters R. Landslide hazard and risk zonation-why is it still so difficult?. Bull Engi Geol Environt. 2006;65(2):167-84.
3- Brenning A. Spatial prediction models for landslide hazards: Review, comparison and evaluation. Nat Hazards Earth Syst Sci, Copernic Publ Eur Geosci Union. 2005;5(6):853-62.
4- Yao X, Tham LG, Dai FC. Landslide susceptibility mapping based on support vector machine: A case study on natural slopes of Hong Kong, China. Geomorphology. 2008;101(4):572-82.
5- Pourghasemi HR, Jirandeh AG, Pradhan B, Xu C, Gokceoglu C. Landslide susceptibility mapping using support vector machine and GIS at the Golestan Province, Iran. J Earth Syst Sci. 2013;122(2):349-69.
6- Jebur MN, Pradhan B, Tehrany MS. Manifestation of LiDAR-derived parameters in the spatial prediction of landslides using novel ensemble evidential belief functions and support vector machine models in GIS. IEEE J Sel Top Appl Earth Obse Remote Sens. 2014;8(2):674-90.
7- Ren F, Wu X, Zhang K, Niu R. Application of wavelet analysis and a particle swarm-optimized support vector machine to predict the displacement of the Shuping

landslide in the Three Gorges, China. Environ Earth Sci. 2015;73(8):4791-804.
8- Bui DT, Tuan TA, Klempe H, Pradhan B, Revhaug I. Spatial prediction models for shallow landslide hazards: A comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree. Landslides. 2016;13(2):361-78.
9- Huang Y, Zhao L. Review on landslide susceptibility mapping using support vector machines. CATENA. 2018;165:520-9.
10- Mohammady M, Pourghasemi HR, Amiri M. Assessment of land subsidence susceptibility in Semnan plain (Iran): A comparison of support vector machine and weights of evidence data mining algorithms. Nat Hazards. 2019;99(2):951-71.
11- Prasad AM, Iverson LR, Liaw A. Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. Ecosystems. 2006;9(2):181-99.
12- Strobl C, Boulesteix AL, Kneib T, Augustin T, Zeileis A. Conditional variable importance for random forests. BMC Bioinform. 2008;9(1):307.
13- Bachmair S, Weiler M. Hillslope characteristics as controls of subsurface flow variability. Hydrol Earth Syst Sci. 2012;16(10):3699.
14- Vorpahl P, Elsenbeer H, Märker M, Schröder B. How can statistical models help to determine driving factors of landslides?. Ecol Model. 2012;239:27-39.
15- Catani F, Lagomarsino D, Segoni S, Tofani V. Landslide susceptibility estimation by random forests technique: Sensitivity and scaling issues. Nat Hazards Earth Syst Sci. 2013;13(11):2815-31.
16- Pourghasemi HR, Kerle N. Random forests and evidential belief function-based landslide susceptibility assessment in western Mazandaran Province, Iran. Environ Earth Sci. 2016;75(3):185.
17- Naghibi SA, Pourghasemi HR, Dixon B. GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran. Environ Monit Assess. 2016;188(1):44.
18- Pourtaghi ZS, Pourghasemi HR, Aretano R, Semeraro T. Investigation of general indicators influencing on forest fire and its susceptibility modeling using different data mining techniques. Ecol Indic. 2016;64:72-84.
19- Golkarian A, Naghibi SA, Kalantar B, Pradhan B. Groundwater potential mapping using C5. 0, random forest, and multivariate adaptive regression spline models in GIS. Environ Monit Assess. 2018;190(3):149.
20- Mohammady M, Pourghasemi HR, Amiri M. Land subsidence susceptibility assessment using random forest machine learning algorithm. Environ Earth Sci. 2019;78(16):503.
21- Zarei P, Talebi A, Alaie Taleghani M. Sensitivity analysis of effective factors in hillslopes instability; a Case Study of Javanrud region, Kermanshah province. Ecopersia. 2018;6(4):259-68.
22- Nafarzadegan AR, Talebi A, Malekinezhad H, Emami N. Antecedent rainfall thresholds for the triggering of deep-seated landslides (case study: Chaharmahal & Bakhtiari Province, Iran). Ecopersia. 2013;1(1):23-39.
23- O'brien RM. A caution regarding rules of thumb for variance inflation factors. Qual Quant. 2007;41(5):673-90.
24- Breiman L. Random forests. Mach Learn. 2001;45(1):5-32.
25- Calle ML, Urrea V. Letter to the editor: Stability of random forest importance measures. Brief Bioinform. 2010; 12(1):86-9.
26- Nicodemus KK. Letter to the editor: On the stability and ranking of predictors from random forest variable importance measures. Brief Bioinform. 2011;12(4):369-73.
27- Jakkula V. Tutorial on support vector machine (svm). Washington DC.; 2006.
28- Joachims T. Text categorization with support vector machines: Learning with many relevant features. European Conference on Machine Learning, 1998 April 21-23, Chemnitz, Germany. Berlin: Springer; 1998.
29- Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, et al. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Natl Acad Sci. 2000;97(1):262-7.
30- Cristianini N, Scholkopf B. Support vector machines and kernel methods: The new generation of learning machines. AI Magazine. 2002;23(3):31.
31- Huang C, Davis LS, Townshend JR. An assessment of support vector machines for land cover classification. Int J Remote Sens. 2002;23(4):725-49.
32- Guo Q, Kelly M, Graham CH. Support vector machines for predicting distribution of Sudden Oak Death in California. Ecol Model. 2005;182(1):75-90.
33- Statnikov A. A gentle introduction to support vector machines in biomedicine: Theory and methods. Singapore: World Scientific; 2011.
34- Statnikov A, Aliferis CF, Hardin DP, Guyon I. A gentle introduction to support vector machines in biomedicine. Singapore: World Scientific Publishing Company; 2013.
35- Kecman V. Support vector machines-an introduction. In: Support vector machines: theory and applications. Berlin: Springer; 2005. pp. 1-47
36- Marjanović M, Kovačević M, Bajat B, Voženílek V. Landslide susceptibility assessment using SVM machine learning algorithm. Engi Geol. 2011;123(3):225-34.
37- Bonham-Carter GF, Agterberg FP, Wright DF. Integration of geological datasets for gold exploration in Nova Scotia. Photogramm Eng Remote Sens. 1988;54(11):1585-92.
38- Van Westen CJ. Geo-information tools for landslide risk assessment: An overview of recent developments. Landslides Eval Stab. 2004;1:39-56.
39- Kornejady A, Ownegh M, Rahmati O, Bahremand A. Landslide susceptibility assessment using three bivariate models considering the new topo-hydrological factor: HAND. Geocarto Int. 2018;33(11):1155-85.
40- Pontius Jr RG, Schneider LC. Land-cover change model validation by an ROC method for the Ipswich watershed, Massachusetts, USA. Agric Ecosyst Environ. 2001;85(1-3):239-48.
41- Kuhn M. The caret package. R Foundation for Statistical Computing [Internet]. Vienna; 2012 [cited 2019, June 20].
42- Lombardo L, Cama M, Maerker M, Rotigliano E. A test of transferability for landslides susceptibility models under extreme climatic events: Application to the Messina 2009 disaster. Nat Hazards. 2014;74(3):1951-89.
43- Lombardo L, Cama M, Conoscenti C, Märker M, Rotigliano EJ. Binary logistic regression versus stochastic gradient boosted decision trees in assessing landslide susceptibility for multiple-occurring landslide events: Application to the 2009 storm event in Messina (Sicily,

southern Italy). Nat Hazards. 2015;79(3):1621-48.
44- Lombardo L, Fubelli G, Amato G, Bonasera M. Presence-only approach to assess landslide triggering-thickness susceptibility: A test for the Mili catchment (north-eastern Sicily, Italy). Nat Hazards. 2016;84(1):565-88.
45- Hosmer DW, Lemeshow S. Applied logistic regression. New York: JohnWiley& Sons; 2000.