Baltic J. Modern Computing, Vol. 10 (2022), No. 2, 224-240
https://doi.org/10.22364/bjmc.2022.10.2.09

Ensemble Conditioning Factor Selection with Markov Chain Framework for Shallow Landslide Susceptibility Mapping in Lake Sapanca Basin and its Vicinity, Turkey

Taskin KAVZOGLU, Alihan TEKE

Gebze Technical University, Faculty of Engineering, Department of Geomatics Engineering, Gebze, Turkey

{kavzoglu, a.teke2020}@gtu.edu.tr

Abstract. The selection of landslide predisposing factors is usually permeated by a certain level of subjectivity, which sometimes adversely affects the performance of the established predictive models. Although filter-based feature selection algorithms have been extensively utilized for discarding the irrelevant factors from the geospatial database, they extremely suffer from statistical biases. Another important limitation is the uncertainty about which feature selection method to choose from among the wide array of available options. In this study, an ensemble feature selection strategy, namely the Markov Chain framework, was suggested to seek an optimal factor subset from filter-based factor selection results. To achieve this objective, 21 landslide conditioning factors were initially considered and seven well-known filter-based feature selection techniques were utilized to determine the factor importance scores. The proposed ensemble approach produced an optimal feature subset consisting of seven conditioning factors using a scree plot analysis after eliminating 14 factors (i.e., reduced by about 66%). The random forest (RF) algorithm was then utilized for predicting the landslide susceptibility by using both the optimal factor subset and all factors. The validation results indicated that overall accuracy (OA) and area under curve (AUC) obtained by using the optimal subset were computed as 90.983% and 94.561%, respectively. The RF algorithm fed by the optimal subset outperformed the scenario in which the whole dataset was used by more than 6% in terms of AUC. The performance differences were also confirmed by McNemar's test, and thus statistical differences for all cases were ascertained.

Keywords: Markov Chain, Ensemble Feature Selection, Elbow Point Detection, Random Forest, Landslide Susceptibility

1. Introduction

All over the world, natural disasters lead to casualties, immense economic losses as well as deterioration of ecological balance. Besides costing human lives, they take a heavy toll on residential settlements, industrial development, and agricultural areas owing to the instantaneous deformations they create. Compared to other catastrophes, landslides place a critical position in terms of the damages they inflict. According to the Center for

Research on the Epidemiology of Disasters (CRED), landslides account for at least 17% of whole natural disaster fatalities globally (Lacasse and Nadim, 2009). To avoid the serious consequences of landslides, the production of robust and up-to-date susceptibility maps delineating the probable spatial distribution of landslide and non-landslide regions is of great importance (Kavzoglu et al., 2014). However, evaluation of landslide susceptibility engenders thought-provoking challenges to researchers since they have multi-dimensional mechanisms, unstable characteristics, and non-linear behaviors (Kavzoglu et al., 2020; Sakellariou and Ferentinou, 2001; Van Asch et al., 2007). It is hence significant to reveal the key factors underlying their occurrences.

In the literature, a large proportion of landslide contributing factors has been employed for susceptibility mapping. However, the selection of optimal landslide causative factors is still the subject of research, and there are still no globally agreed clear frameworks or guidelines (Ayalew and Yamagishi, 2005). The main reason for this could be explained by the particular characteristics of the study sites under consideration (Van Westen et al., 2003). More specifically, while any factor utilized in landslide susceptibility mapping may be a contributing factor for a certain area, it may not be for another (Kavzoglu et al., 2015). On the other side, superfluous and irrelevant contributing factors will diminish the reliability, the predictive accuracy of the algorithm, and thus increase instability (Teke and Kavzoglu, 2021). Furthermore, the model with the overabundance of data will be more prone to the adverse consequences of the overfitting issue due to the curse of dimensionality (known as the Hughes phenomenon). In this context, feature selection techniques, particularly filter-based ones, have been intensively employed to overcome the aforementioned difficulties due to their ability to improve the model performance thereby discarding redundant attributes from the dataset. However, even when implemented in the same dataset, such techniques might produce different importance scores and ranking lists for each factor as they consider different features and relations inherent in the dataset, such as information theory, distance measurements, or entropy. As a result, considering solely a single feature selection algorithm may cause not only the ideal feature subset to still having trivial attributes, but also fallacious inferences due to the biased prediction.

While conventional feature selection approaches are typically biased towards selecting features with high-dimensionality, ensemble feature selection methods provide benefits to mitigate and compensate for such biases (Neumann et al., 2016; Sarkar et al., 2014). The principal concept behind ensemble feature selection is aggregating the results of the different individual feature selection methods to achieve more effective and stable outcomes. Broadly speaking, they take the outcomes of several variations of the feature ranking and convert the multiple ranked feature lists into an individual rank list (Wald et al., 2012). However, the essential prerequisite is to select the correct technique to combine or aggregate these ordered rank lists. In this context, rank aggregation methods enable the union of data from distinct sources. With the use of aggregation techniques, multiple lists are transformed into a single order list, which contributes to a more accurate, stable, and robust result. Another issue faced in this process is to determine how many parameters will be selected from the combined list; that is, the specification of the threshold value. In the current literature, researchers generally tend to select a certain number of features or a subset of features comprising a particular percentage of the dataset (Pradhan and Sameen, 2017; Tanyu et al., 2021), or top-ranked important factors are selected considering a predetermined cut-off threshold (Lee et al., 2020; Thai Pham et al., 2018). However, there exists no theoretical support to identify the critical threshold in most studies.

In this study, a Markov Chain modeling framework-based ensemble feature selection strategy was proposed for the first time to seek an optimal landslide predisposing factor subset and overcome deficiencies of single feature selection techniques. Firstly, seven well-known filter-based feature selection techniques including gain ratio (GR), information gain (IG), symmetrical uncertainty (SU), Chi-Square (χ^2), Pearson correlation coefficient (PCC), Fisher-score (FS), and Gini-index (GI) were utilized to measure the importance of each factor and seven different ranking lists were produced, which were aggregated by benefitting from Markov Chain modeling strategy to transform multiple lists into single ranking order. Afterward, a scree plot analysis was implemented to seek the optimal predisposing factor subset. Eventually, both the whole dataset and optimal subset were utilized to generate a landslide susceptibility map using Random Forest (RF) algorithm. The predictive performances of the resultant maps were evaluated with two accuracy metrics including overall accuracy (OA) and area under curve (AUC) score.

2. Study Area and Dataset

2.1. Description of Study Area

The current work was carried out in the Sapanca Basin and its surrounding, encompassing land of about 945 km^2, situated on the Catalca-Kocaeli section of the eastern Marmara Region of Turkey where urbanization and industrialization are rapidly rising (Colkesen and Kavzoglu, 2018) (Fig. 1). Topographically, the study area has elevations varying from 40 m to 1637 m and slope gradients up to about 70°. Tectonically, it is located on a tectonic hole that was originated and dominated by the dextral strike-slip tectonics of the North Anatolian Fault Zone (NAFZ). The NAFZ is a strike-slip fault causing some fatal earthquakes within the last century. Geologically, in the south of the basin and its surroundings, Paleozoic metamorphic (schists, marbles, and gneissic quartzite) and Mesozoic limestone and marbles and rocks belonging to the metaophiolite units (peridotite, gabbro, amphibolite, metalava) are found, while in the north, Upper Cretaceous-Paleocene aged flysch deposits (limestone, marl) and Eocene aged units (conglomerate, sandstone, marl, limestone) with flysch characteristics (Esenli, 1995). Having a moderately deep lake ecosystem, Sapanca Lake is nourished by 15 seasonally varying stream flows (Gürbüz and Gürer, 2008). Climatologically, in the Sapanca lake basin, a transitional climate, which is influenced by the Mediterranean and the Black Sea climates, is dominant (Ceylan, 1999).

2.2. Landslide Inventory

In the research area, one of the most important hypotheses acknowledged in the process of developing landslide susceptibility maps is that historical landslide activities that happened in particular areas will likely occur in areas with similar features. As a result, landslide inventory maps are regarded as an essential tool for subsequent phases as they provide critical information such as the kind, location, and area of landslides. The landslide data used in this study were collected within the scope of the Landslide Inventory Project of Turkey, which aims to produce inventory maps that reveal the spatial distributions, types, and activities of mass movements on a national scale to be used in the prevention of natural disasters caused by mass movements. Provided by the

General Directorate of Mineral Research and Exploration (GDMRE), Turkey, landslide zones in polygon format were utilized for constructing the landslide inventory map. There exists a total of 190 landslide zones that were documented as polygon features, as illustrated in Fig. 1. The total terrain exposed by landslide events is 8.27 km^2 in extent, with the lowest and greatest landslide areas being 2,363 m^2 and 139,499 m^2, respectively.

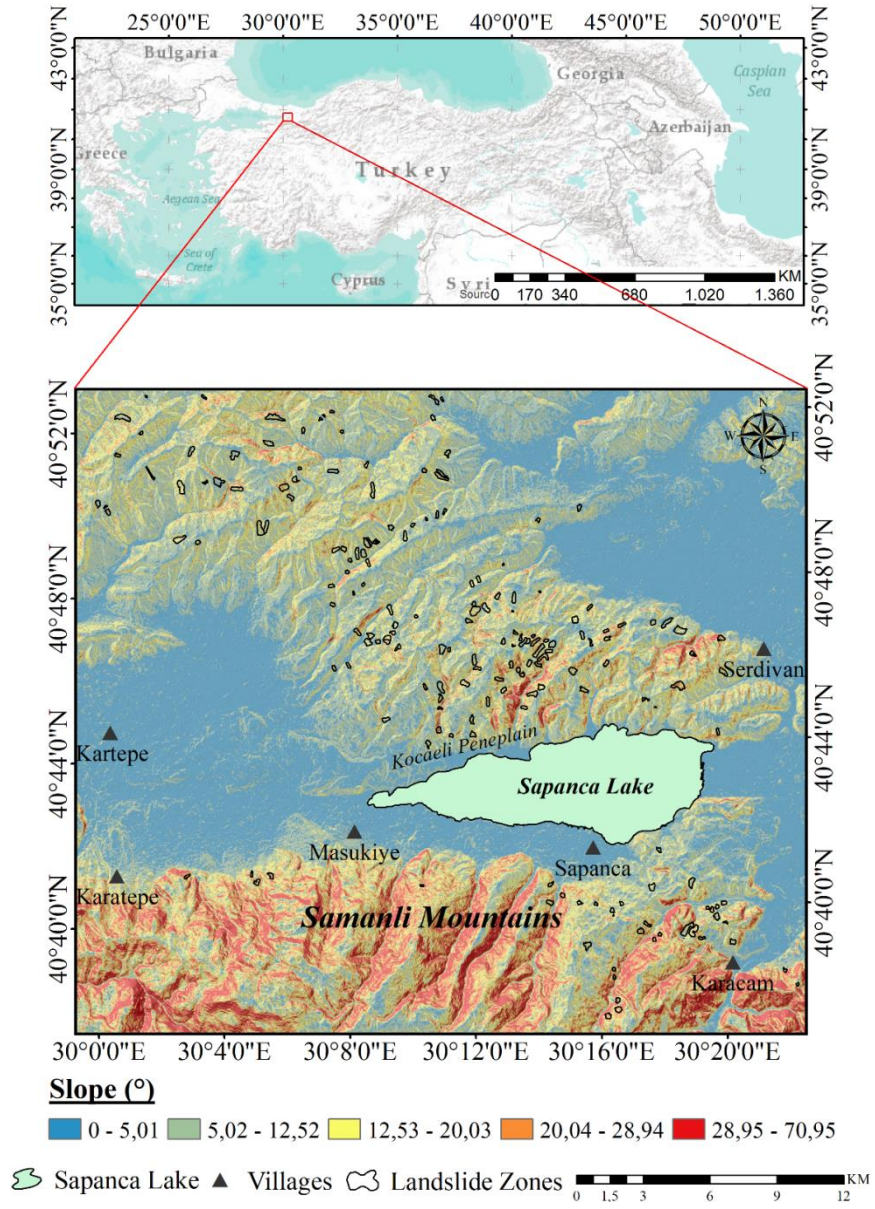

Figure 1. Location map of the study area and landslide inventory.

In addition to landslide samples, determining non-landslide samples is one of the stages required to build the landslide inventory map. In this study, the strategy proposed by Gómez and Kavzoglu (2005) was adopted. The strategy is based on the assumption that non-landslide samples should be selected from 100% landslide-free zones, just as landslide instances are collected from 100% risky areas. The proposed method is based on the idea that landslides are unlikely to occur in terrains with less than a 5% slope and river channels (Colkesen et al., 2016; Kavzoglu and Teke, 2022). To prevent any biases and conclusions from the imbalance dataset, a number of non-landslide pixels equal to the total landslide cases were gathered using this process.

Table 1. Data source and scale/resolution information of landslide predisposing factors.

Major Factors	Sub-Factors	Source	Scale/Resolution
Geology	Lithology	General Directorate of Mineral Research and Exploration of Turkey (http://www.mta.gov.tr)	1:100,000
	Distance to lineaments	Landsat-8 Operational Land Imager (OLI) multispectral image, (https://earthexplorer.usgs.gov/)	30 m
Topographical	Elevation	Shuttle Radar Topography Mission (SRTM- https://earthexplorer.usgs.gov/)	30 m
	Aspect Curvature Plan curvature Profile curvature Slope Slope length TPI TRI Valley depth	DEM	
Hydrological	Distance to rivers Drainage density	Digitized existing river networks	30 m
	TWI	DEM	
Environmental	Distance to roads Road density	Digitized existing road networks	30 m
	LULC NDVI	Landsat-8 Operational Land Imager (OLI) multispectral image, (https://earthexplorer.usgs.gov/)	
	Soil Type Soil Depth	Ministry of Agriculture and Forestry	1:25,000

2.3. Landslide Predisposing Factors

According to analyzed geo-environmental data as well as characteristics of the study area, a total of 21 landslide conditioning factors (Table 1), namely aspect, curvature, distance to lineaments, distance to rivers, distance to roads, drainage density, elevation, lithology, land use/land cover (LULC), normalized difference vegetation index (NDVI), plan curvature, profile curvature, road density, slope, slope length, soil depth, soil type, topographic position index (TPI), topographic roughness index (TRI), topographic wetness index (TWI), and valley depth was initially taken into consideration to investigate their effectiveness and to produce a reliable, robust, and up-to-date landslide susceptibility map.

Procured by Shuttle Radar Topography Mission (SRTM), the digital elevation model (DEM) was utilized to produce a total of 11 landslide thematic maps (i.e., aspect, curvature, elevation, plan curvature, profile curvature, slope, slope length, TPI, TRI, TWI, valley depth). The lithology map of the basin and its vicinity, composing 13 units, was provided by the GDMRE. Indicating potential tectonic activity, lineaments were extracted by using Landsat 8 OLI imagery and a thematic map of distance to lineaments was produced using the Euclidean distance function. The Euclidean distance function was also used to calculate distances to roads and rivers, and road density was calculated using existing road network data. Drainage pattern was extracted from DEM and drainage density maps were prepared using the line density tool in ArcGIS. Landsat 8 OLI (Operational Land Imager) data acquired in 2021 was used to create the LULC. Based on the analyzed existing/collected, data it was decided that six types of LULC classes cover the bulk of the study site, which are including water, urban, cultivated lands, non-cultivated areas, forest, and road. Delineating the characteristics of the density and healthiness of green vegetation, NDVI was generated using the Red and NIR bands of Landsat-8 OLI. The soil map was also digitized which was supplied by the Ministry of Agriculture and Forestry, Republic of Turkey.

3. Methodology

Prediction of landslide susceptibility is a notable practice for providing valuable information to the local stakeholders and authorities, land-use planners, and government agencies, which is required for many local to global-scale studies. In this current work, initially, 21 landslide causative factors were considered for producing landslide susceptibility maps of the Sapanca Lake Basin of Turkey. Later, six filter-based feature selection methods were implemented to measure the importance score of each factor; thus, seven ordered rank lists were produced. To combine the ranking lists, Markov Chain-based ensemble feature selection strategy was applied. Determination of the number of critical factors, which is one of the research questions of this study, was done by scree-plot analysis (i.e., elbow point detection) in the Python programming language. Lastly, landslide susceptibility maps were produced using both the whole dataset and optimal factor subset through the ensemble learning-based RF algorithm. Two performance evaluation metrics (i.e., OA and AUC) and a statistical significance test (i.e., McNemar's test) were applied to assess the predictive performances of produced thematic maps.

3.1. Filter-based Feature Selection Methods

Focusing on different characteristics of the dataset (e.g., distance, consistency, information, dependency, similarity measures), filter-based feature selection techniques work without benefitting any inductive learning algorithm to evaluate the attributes. The generic working mechanism of such techniques incorporates two major stages. In the first stage, feature importance/relevance scores are calculated based on certain criteria. In the latent stage, the features with higher scores are selected for inducing the learning algorithm and low-scoring attributes are discarded from the dataset. It should be also noted that these techniques have two kinds of evaluation schemes: univariate and multivariate (Aggarwal et al., 2014). In the former one, each attribute is ranked separately from the rest of the feature space whereas the second one employs a batch method to assess characteristics. Despite their promising prospects, filter methods suffer from several deficiencies, practically related to the issues faced in working principles in their nature. They, particularly univariate types, neglect the interaction with the learning algorithm. To elaborate further, each attribute is handled individually, implying that it might cause the building of models with poor predictive performances (Saeys et al., 2007). Besides, one of the most critical problems is the determination of the optimal feature subset, that is, the uncertainty about how to determine the cut-off (i.e. threshold) value to distinguish the relevant and irrelevant features in the dataset, which is one of the focal points of this work.

Gain Ratio (GR) is a filter-based feature selection algorithm developed to surmount the issue of IG's tendency to select attributes with high quantities of distinct values even if it is not more informative (Quinlan, 1993). To eliminate the bias, GR assesses the attributes by dividing the IG of the anticipated feature by the entropy of the observed feature. Like other entropy-based feature selection approaches, GR has been employed in several studies to compute the importance values of predisposing factors utilized in determining landslide susceptibility (Fallah-Zazuli et al., 2019; Tanyu et al., 2021).

Information gain (IG) is a widely utilized criterion to determine the limits of a feature's importance in the domain of ML and information theory (Quinlan, 1993). IG is calculated based on the term of entropy, varying between 0 and 1, characterized as a measure of uncertainty in a system. On a fundamental basis, the IG measures the discrimination potential of features. That is, as the IG value rises, the discrimination ability of features also rises. It is mainly investigated how much information is acquired about a class when a specific feature is utilized. More specifically, information is gained under a rule that supports alleviating variance and indicates the significance of parameters. This approach has been utilized in identifying the optimal predisposing factors in various landslide susceptibility modeling studies (Park and Kim, 2019; Pham et al., 2017).

Symmetrical uncertainty (SU) is one of the most robust filter-based feature selection techniques that is particularly utilized in high-dimensional datasets. Similar to GR, its main motivation is to compensate IG's bias against attributes with overabundance values since the IG value is divided by the sum of the entropies of the random variable (Kannan and Ramaraj, 2010). Another key point of SU is its potential to ensure a common measure of bonds between the features irrespective of the form of the essential distributions.

Chi-square (χ^2) is an algorithm based on chi-square statistics and evaluates the deviation of each feature from the expected distribution according to class labels. In this approach, the resemblance between two variables is calculated, and estimations on whether the variables are correlated with each other are conducted. It can be also used to determine whether variables are suitable for representing the data. The feature selection method based on χ^2 statistics consists of two basic steps. Firstly, chi-square statistics of the features are calculated according to the classes of the target variable. Secondly, according to the significance level determined, chi-square statistics are analyzed with the principle of chi-merge. Discretization of the features is then carried out iteratively until inconsistent features are found in the dataset.

Pearson correlation coefficient (PCC) assesses the worth of a feature by estimating the correlation (Pearson's) between the target and the class. Correlation coefficients are used to calculate the correlation between a selection of qualities and their respective classes, as well as the inter-correlations between the features. The importance of a collection of features increases as the correlation between features and classes develops yet decreases as the inter-correlation grows.

Fisher-score (FS) is mainly based on the idea of acquiring a set of variables for which the proximities of instances with different labels should be as far as possible, or vice versa. The main focus of its working principle tries to satisfy two conditions. The first condition is that the distance between the class centers should be the maximum and the second one is that the distribution of all classes within itself should be minimized.

Gini-index (GI) is a metric that is also used to determine the best splitting criteria for features in decision trees and essentially measures the impurity of the variables in a given dataset. It is a class-based measurement approach associated with information gain, resulting from fixing a specific variable. The GI is independently calculated for each feature. A feature with a low impurity value in a dataset corresponds to the optimal candidate whereas a feature with a high impurity is considered an irrelevant feature, therefore, the features that contain the discriminative information have small GI values.

3.2. Markov Chain Framework for Rank Aggregation

The practice of merging the ordered preferences of multiple lists, also known as rank aggregation, is a helpful tool for data mining applications. Rank aggregation is the unsupervised equivalent of regression, with the purpose of finding an aggregate ranking that minimizes the distance to each of the provided ranked lists (Sculley, 2007). It is basically the process of combining the rating results of entities from different ranking systems to get a robust one. In this study, Markov Chain-based rank aggregation, which is an unsupervised learning method, was utilized to combine factor ranking lists obtained from seven filter-based feature selection methods. Symbolizing the features in different lists as nodes in a graph, Markov Chain-based rank aggregation is convenient for base rankers (Liu et al., 2007; Sculley, 2007). It assumes that the entities have a Markov Chain, and the ranking relationships between items in the ranking lists describe Markov Chain transitions. In order to rank entities, the stationary distribution of the Markov Chain is used. The aggregate ranks of the lists are computed (or approximated) by computing the stationary distribution on the Markov Chain.

3.3. Elbow Point Detection

The relative costs of increasing an adjustable parameter of a data point might not be worth the performance benefit in return. The same concept is also evident in the results obtained from filter-based feature selection algorithms. One of the most central questions here is to identify how many parameters will be selected from the ranking list; in other words, the specification of the threshold value. In this study, the threshold point that determines the optimal number of features (i.e. factor subset) was determined using the “Kneedle” algorithm proposed by Satopää et al. (2011). The "Kneedle" technique employs curvature as a mathematical measurement of how much a function deviates from a horizontal plane. In this point, the authors highlighted that the greatest curvature captures the leveling off effects operators employ to detect knees. The method "Kneedle" discovers useful data points in continuous data sets that demonstrate the optimal balance of intrinsic trade-offs called "knees" (curves with negative concavity) or "elbows" (curves with positive concavity) based on the mathematical concept of curvature for continuous functions.

3.4. Random Forest (RF)

Random forest (RF), introduced by Breiman (2001), is a robust ensemble-learning algorithm that has been frequently implemented in many domains of multi-task purposes including classification, regression, unsupervised learning, and feature selection. RF, which is a decision tree-based method, is applied by combining many decision trees. RF employs the statistical resampling bootstrapping technique in the model training phase. Each tree in the forest is trained using about 2/3 of the samples namely, in-bag samples, and the remaining 1/3 samples namely out-of-bag samples are utilized to calculate the overall accuracy of the tree model (Kavzoglu, 2017). Ultimately, the majority voting rule is implemented in the prediction of the class labels of unknown samples. Due to its flexibility to various tasks, easy parameterization, working mechanism with both categorical and continuous data, and accuracy level achieved, the RF has captured increasing awareness in many studies associated with landslide susceptibility mapping (Dou et al., 2019; Teke et al., 2021; Youssef et al., 2016).

4. Results and Discussion

The reported study fundamentally aims to realize two main significant objectives. The first one is to convert multiple ranking order lists into a single list to compensate for the biased results caused by the unstable nature of filter-based feature selection techniques. The second is to specify the critical cut-off value to find the optimal factor subset. From the first perspective, initially, the importance score of each conditioning factor was estimated using the seven filter-based feature selection approaches. This allows a more detailed analysis of some aspects of the geospatial data and the overall nature of the filter-based feature selection. By carefully examining the data, results clearly revealed that all filter-based feature selections used in the work yielded different ordered ranking lists, as illustrated in Table 2. For instance, profile curvature was one of the top 6 factors in the GR, IG, SU, χ^2, and GI methods although it was ranked 19th and 21st in PCC and FS, respectively. Nevertheless, there exist also some similarities regarding factor importance rankings, the most obvious of which is that slope is ranked first in all factor

selection results. According to the calculated ranking results, the slope was the most significant landslide-explanatory factor, having the highest importance in all individual filter-based feature selection algorithms. Also, TRI took second place in terms of importance rankings except for the GI algorithm.

Table 2. Factor rankings obtained by individual filter-based feature selection techniques. (*Note that MC indicates ranking list obtained with Markov Chain-based ensemble framework).

Factor	GR	IG	SU	χ^2	PCC	FS	GI	MC
Slope	1	1	1	1	1	1	1	1
TRI	2	2	2	2	2	2	3	2
Profile Curvature	6	3	4	3	19	21	4	3
Elevation	8	4	8	4	4	4	6	4
Plan Curvature	4	5	3	5	21	19	7	5
Lithology	7	7	6	7	10	10	11	6
Aspect	10	6	9	6	14	12	10	7
Soil Depth	5	8	5	8	15	15	13	8
Soil Type	3	10	7	10	5	8	15	9
Distance to Lineaments	9	9	10	9	3	3	12	10
Drainage Density	13	11	12	12	6	5	8	11
Slope Length	11	12	11	11	8	9	16	12
NDVI	15	14	15	14	7	6	14	13
TWI	14	13	14	13	12	14	2	14
LULC	12	15	13	15	11	13	17	15
Road Density	16	16	16	16	13	11	9	16
Distance to Rivers	17	17	17	17	9	7	18	17
TPI	19	18	19	18	18	20	20	18
Curvature	18	19	18	19	20	18	21	19
Valley Depth	21	20	21	20	17	16	5	20
Distance to Roads	20	21	20	21	16	17	19	21

From the second perspective, the critical threshold value was captured by the scree-plot analysis, which is a visual and graphical evaluation tool. The analysis allows seeking the critical cut-off value in which the difference between factor scores decreases and becomes insignificant. The process is applied not only visually, but also automatically and mathematically in the Python programming language to seek the optimal feature subset in this work. With the application of the Markov Chain framework, it was observed that the difference between the stationary probabilities of factors taking place after the aspect in the final ranking list gradually decreased and become insignificant (Fig. 2). More clearly, this analysis implied that the relative costs to increase the number of factors are no longer worth the corresponding performance benefit. Therefore, it was determined that the optimal subset consisted of seven factors (i.e., slope, TRI, profile curvature, elevation, plan curvature, lithology, aspect), which corresponds to approximately 33% of the entire dataset. According to the results obtained with the Markov Chain-based ensemble feature selection, the factor with the highest stationary probability was found to be slope with 0.250, followed by TRI (0.162), profile curvature (0.114), elevation (0.084), plan curvature (0.065), lithology (0.052), aspect (0.036). 6 out of 7 factors within the optimal subset were topographical

parameters while there was only a factor (lithology) associated with the geological structure, which are compatible with many previous studies (e.g., Kamp et al., 2008; Kavzoglu et al., 2021; Kincal et al., 2009). On the other hand, none of the anthropogenic, environmental and hydrological features were included in the optimal dataset, and thus, the other 14 landslide contributing factors (i.e., soil depth, soil type, distance to lineaments, drainage density, slope length, NDVI, TWI, LULC, road density, distance to rivers, TPI, curvature, distance to roads, and valley depth) were discarded from the data set.

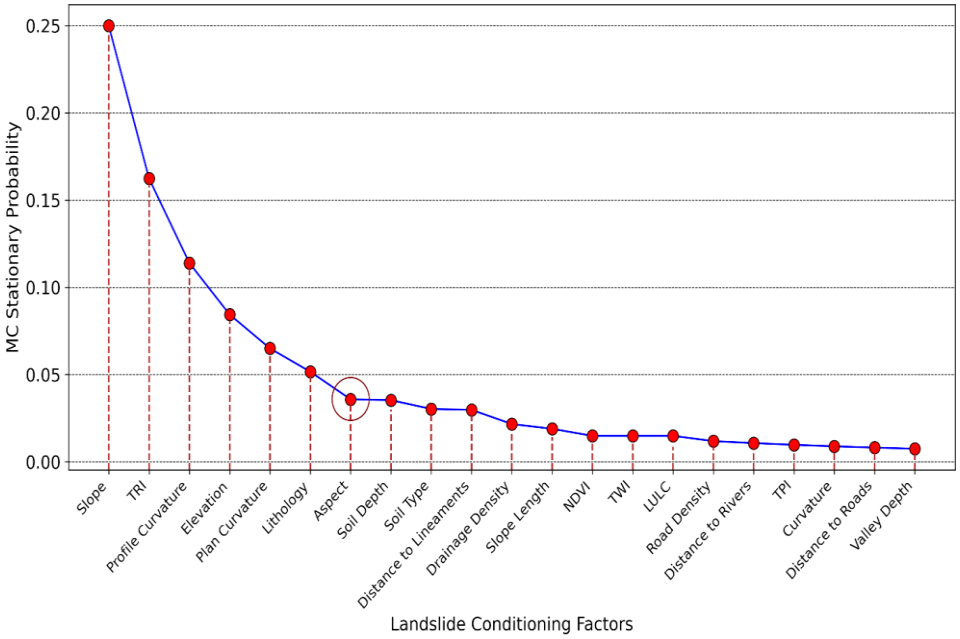

Figure 2. Stationary probability of each factor estimated with Markov Chain framework.

In this study, results were analyzed using two accuracy metrics, namely overall accuracy (OA) and area under the ROC curve (AUC) (Fig. 3). OA simply indicates the proportion of the accurately estimated instances to the whole instances. Likewise, the AUC value has been commonly employed owing to both its capability to be exhibited visual representation and appropriate measures for assessing the predictive achievement of algorithms in landslide susceptibility mapping studies. The AUC value ranges from 0.5 to 1. The predictive performances of the models can be called fair (if the AUC value is between 0.7 and 0.8), good (if the AUC value is in the range 0.8-0.9), or excellent (if the AUC value is between 0.9-1) (Cantor and Kattan, 2000). According to the outcomes of these evaluation metrics, when the optimal factor subset was utilized, the RF had the OA and AUC scores of 90.983% and 83.509%, respectively. However, when the whole data set was employed, the OA and AUC scores were computed as 83.509% and 88.143%. Consequently, it should be mentioned that the predictive performance of the RF model with the optimal dataset can be described as excellent.

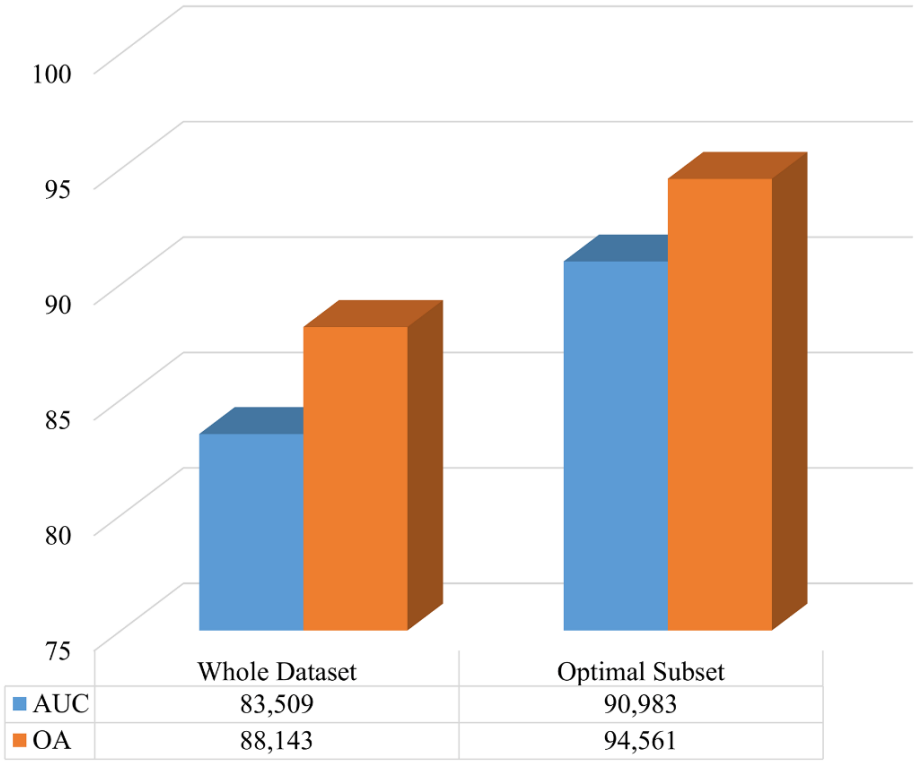

Figure 3. Performance analysis of RF algorithm constructed with all factors and optimal conditioning factors.

Apart from the accuracy assessment metrics, performance differences between the models were statistically measured by using McNemar's test to make impartial and sound comparisons. If the computed statistical value is higher than the threshold level (3.84 for a 95% confidence interval), it can be said that the difference in performances from the point of model accuracy is statistically significant. When the results of the RF algorithm using the optimal subset and all dataset was compared, the chi-square value signifying the measure of the statistical significance between two independent models was calculated as 11.236. Thus, it can be clearly stated that the model generated by optimal factors produced statistically superior results compared to the entire dataset since the estimated statistical test values were higher than the critical table value.

When the landslide susceptibility maps obtained using the optimal subset was visually examined (Fig. 4), it was observed that the north-eastern part of the basin and the areas to the west of Sapanca Lake had very low/low landslide susceptibility. This could be explained by the low slope gradients and topographic elevations of these zones, and thus, these sections have denser residential areas compared to the others. On the other hand, it was found that landslide susceptibility was higher on the northern and north-eastern slopes of the Samanli Mountains and in the northwest of the Kocaeli Peneplain section. This could be explained by the general geomorphological characteristics (e.g., deep valleys on north-facing slopes of the mountains.

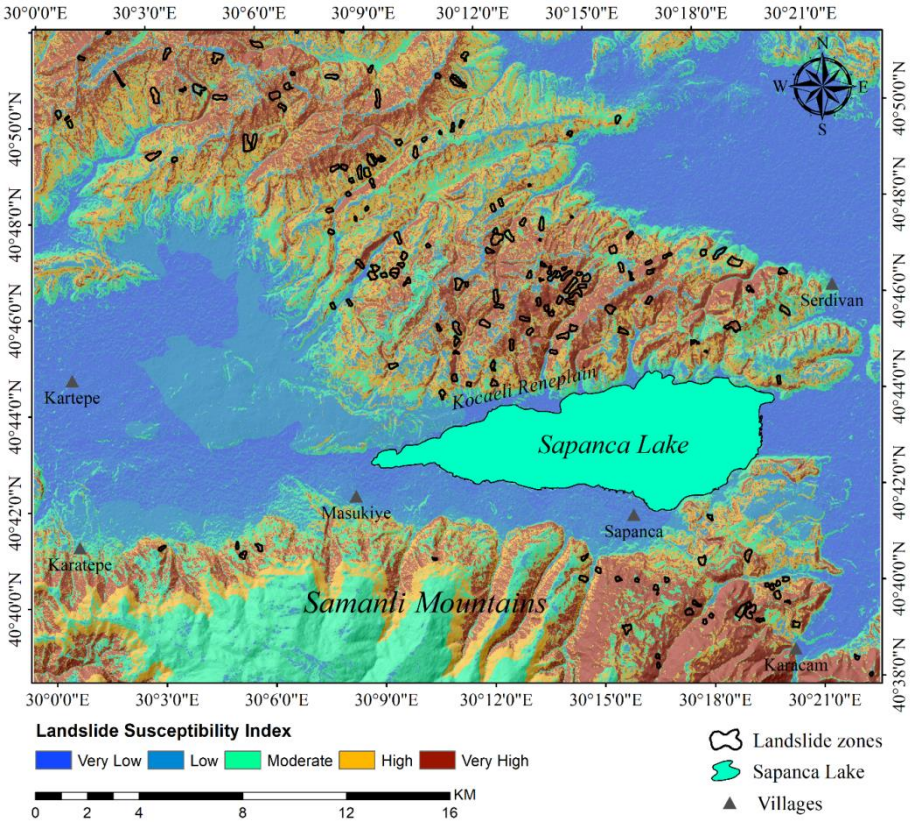

Figure 4. Landslide susceptibility map obtained by using the optimal factor subset.

5. Conclusions

In this study, a Markov Chain-based ensemble feature selection and elbow point detection strategies were proposed to minimize the bias originating from the use of individual filter-based feature selection methods and identify the most relevant predisposing factors from the whole dataset. According to the findings obtained from the experiments, the most significant conclusions matching the core objective of this work are collectively given here. Firstly, the predictive performance of the landslide susceptibility map produced with the optimal factor subset resulted in an improvement of approximately 7% and 6% in terms of OA and AUC score, respectively, when compared to the scenario in which the whole dataset was employed. In addition, based on McNemar's test, the differences were found to be statistically significant. Secondly, with the scree-plot analysis, the entire dataset was reduced by about 66%, producing landslide susceptibility maps with higher performance and alleviating the computational complexity of the model. It should be also worth mentioning that slope, TRI, profile curvature, elevation, plan curvature, lithology, and aspect were found to be the most important factors according to the results of the proposed Markov Chain modeling framework strategy, which is compatible with the findings of many previous studies.

This current work sheds new light on the determination of optimal landslide conditioning factors for landslide susceptibility mapping practices by not only considering an ensemble feature selection methodology but also proposing the application of the "Kneedle" algorithm. In a nutshell, the findings will be beneficial for the decision-maker and policy-maker individuals in the study area to mitigate potential harms and provide a natural environment for biological diversity.

List of Abbreviations

Area Under Curve (AUC)
Center for Research on The Epidemiology of Disasters (CRED)
Chi-Square (χ^2)
Digital Elevation Model (DEM)
Fisher-score (FS)
Gain Ratio (GR)
General Directorate of Mineral Research and Exploration (GDMRE)
Gini-index (GI)
Information Gain (IG)
Land Use/Land Cover (LULC)
Markov Chain (MC)
Normalized Difference Vegetation Index (NDVI)
North Anatolian Fault Zone (NAFZ)
Operational Land Imager (OLI)
Overall Accuracy (OA)
Pearson Correlation Coefficient (PCC)
Random Forest (RF)
Shuttle Radar Topography Mission (SRTM)
Symmetrical Uncertainty (SU)
Topographic Position Index (TPI)
Topographic Roughness Index (TRI)
Topographic Wetness Index (TWI)

References

Aggarwal, C. C., Kong, X., Gu, Q., Han, J., Yu, P. S. (2014). Feature selection for classification: A review Data Classification: Algorithms and Applications, 571–605.

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**(1–2), 15–31.

Breiman, L. (2001). Random forests. *Machine Learning*, **45**(1), 5–32.

Cantor, S. B., Kattan, M. W. (2000). Area under the ROC curve for a binary diagnostic test. *Medical Decision Making*, **20**, 468–470.

Ceylan, M. A. (1999). The relief and the precipitation specials in the Sapanca lake basin (In Turkish). *Türk Coğrafya Dergisi*, **34**, 643-659.

Colkesen, I., Kavzoglu, T. (2018). Selection of optimal object features in object-based image analysis using filter-based algorithms. *Journal of the Indian Society of Remote Sensing*, **46**(8), 1233–1242.

Colkesen, I., Sahin, E. K., Kavzoglu, T. (2016). Susceptibility mapping of shallow landslides using kernel-based Gaussian process, support vector machines and logistic regression. *Journal of African Earth Sciences*, **118**, 53–64.

Dou, J., Yunus, A. P., Tien Bui, D., Merghadi, A., Sahana, M., Zhu, Z., Chen, C. W., Khosravi, K., Yang, Y., Pham, B. T. (2019). Assessment of advanced random forest and decision tree algorithms for modeling rainfall-induced landslide susceptibility in the Izu-Oshima Volcanic Island, Japan. *Science of the Total Environment*, **662**, 332–346.

Esenli, V. (1995). *Sapanca gölü ve havzasının hidrojeokimyası ile dip sedimanlarının mineralojik ve jeokimyasal incelenmesi*, PhD thesis, Istanbul Technical University, Istanbul, Turkey.

Fallah-Zazuli, M., Vafaeinejad, A., Alesheykh, A. A., Modiri, M., Aghamohammadi, H. (2019). Mapping landslide susceptibility in the Zagros Mountains, Iran: a comparative study of different data mining models. *Earth Science Informatics*, **12**(4), 615–628.

Gómez, H., Kavzoglu, T. (2005). Assessment of shallow landslide susceptibility using artificial neural networks in Jabonosa River Basin, Venezuela. *Engineering Geology*, **78**(1–2), 11–27.

Gürbüz, A., Gürer, Ö. F. (2008). Anthropogenic affects on lake sedimentation process: A case study from Lake Sapanca, NW Turkey. *Environmental Geology*, **56**(2), 299–307.

Kamp, U., Growley, B. J., Khattak, G. A., Owen, L. A. (2008). GIS-based landslide susceptibility mapping for the 2005 Kashmir earthquake region. *Geomorphology*, **101**(2008), 631–642.

Kannan, S. S., Ramaraj, N. (2010). A novel hybrid feature selection via Symmetrical Uncertainty ranking based local memetic search algorithm. *Knowledge-Based Systems*, **23**(6), 580–585. https://doi.org/10.1016/j.knosys.2010.03.016

Kavzoglu, T. (2017). Object-Oriented Random Forest for High Resolution Land Cover Mapping Using Quickbird-2 Imagery. *Handbook of Neural Computation*, 607–19. Elsevier Inc.

Kavzoglu, T., Sahin, E. K., Colkesen, I. (2015). Selecting optimal conditioning factors in shallow translational landslide susceptibility mapping using genetic algorithm. *Engineering Geology*, **192**, 101–112. https://doi.org/10.1016/j.enggeo.2015.04.004

Kavzoglu, T., Sahin, E. K., Colkesen, I. (2014). Landslide susceptibility mapping using GIS-based multi-criteria decision analysis, support vector machines, and logistic regression. *Landslides*, **11**(3), 425–439. https://doi.org/10.1007/s10346-013-0391-7

Kavzoglu, T., Teke, A. (2022). Predictive Performances of Ensemble Machine Learning Algorithms in Landslide Susceptibility Mapping Using Random Forest, Extreme Gradient Boosting (XGBoost) and Natural Gradient Boosting (NGBoost). *Arabian Journal for Science and Engineering*. https://doi.org/10.1007/s13369-022-06560-8

Kavzoglu, T., Teke, A., Bilucan, F. (2020). *Effectiveness of Machine Learning Algorithms in*

Landslide Susceptibility Mapping: A Case Study of Trabzon Province, Turkey. Asian Conference on Remote Sensing (ACRS), November 2020.

Kavzoglu, T., Teke, A., Yilmaz, E. O. (2021). Shared Blocks-Based Ensemble Deep Learning for Shallow Landslide Susceptibility Mapping. *Remote Sensing*, **13**(25), 4776. https://doi.org/https://doi.org/10.3390/rs13234776

Kincal, C., Akgun, A., Koca, M. Y. (2009). Landslide susceptibility assessment in the İzmir (West Anatolia, Turkey) city center and its near vicinity by the logistic regression method. *Environmental Earth Sciences*, **59**(4), 745–756. https://doi.org/10.1007/s12665-009-0070-0

Lacasse, S., Nadim, F. (2009). Landslide Risk Assessment and Mitigation Strategy. Bridge Engineering Handbook, Second Edition: Substructure Design, 315–359. https://doi.org/10.1201/b15621

Lee, D. H., Kim, Y. T., Lee, S. R. (2020). Shallow landslide susceptibility models based on artificial neural networks considering the factor selection method and various non-linear activation functions. *Remote Sensing*, **12**(7), 1194. https://doi.org/10.3390/rs12071194

Liu, Y. T., Liu, T. Y., Qin, T., Ma, Z. M., Li, H. (2007). Supervised rank aggregation, 16th International World Wide Web Conference, WWW2007, 481–490.

Neumann, U., Riemenschneider, M., Sowa, J. P., Baars, T., Kälsch, J., Canbay, A., Heider, D. (2016). Compensation of feature selection biases accompanied with improved predictive performance for binary classification by using a novel ensemble feature selection approach. *BioData Mining*, **9**(1), 1–14. https://doi.org/10.1186/s13040-016-0114-4

Park, S., Kim, J. (2019). Landslide susceptibility mapping based on random forest and boosted regression tree models, and a comparison of their performance, *Applied Sciences (Switzerland)*, **9**(5), 942.

Pham, B. T., Bui, D. T., Dholakia, M. B., Prakash, I., Pham, H. V., Mehmood, K., Le H. Q. (2017). A novel ensemble classifier of rotation forest and Naïve Bayer for landslide susceptibility assessment at the Luc Yen district, Yen Bai Province (Viet Nam) using GIS. *Geomatics, Natural Hazards and Risk*, **8**(2), 649–671.

Pradhan, B., Sameen, M. I. (2017). Landslide Susceptibility Modeling: Optimization and Factor Effect Analysis. In B. Pradhan (Ed.), *Laser Scanning Applications in Landslide Assessment* (pp. 115–132), Springer International Publishing. https://doi.org/10.1007/978-3-319-55342-9_6

Quinlan, J. T. (1993). *C4.5: Programs for Machine Learning.* Morgan Kaufmann Publishers.

Saeys, Y., Inza, I., Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, **23**(19), 2507–2517.

Sakellariou, B. M. G., Ferentinou, M. D. (2001). GIS-Based Estimation of Slope Stability. *Natural Hazards Review*, **2**(1), 12–21.

Sarkar, C., Cooley, S., Srivastava, J. (2014). Robust feature selection technique using rank aggregation. *Applied Artificial Intelligence*, **28**(3), 243–257.

Satopää, V., Albrecht, J., Irwin, D., Raghavan, B., (2011). Finding a "kneedle" in a haystack: Detecting knee points in system behavior. *Proceedings - International Conference on Distributed Computing Systems*, 166–171. https://doi.org/10.1109/ICDCSW.2011.20

Sculley, D. (2007). Rank aggregation for similar items. *Proceedings of the 7th SIAM International Conference on Data Mining*, 587–592. https://doi.org/10.1137/1.9781611972771.66

Tanyu, B. F., Abbaspour, A., Alimohammadlou, Y., Tecuci, G. (2021). Landslide susceptibility analyses using Random Forest, C4.5, and C5.0 with balanced and unbalanced datasets. *Catena*, **203**(April 2020), 105355. https://doi.org/10.1016/j.catena.2021.105355

Teke, A., Kavzoglu, T. (2021). Determination of Effective Predisposing Factors Using Random Forest-Based Gini Index in Landslide Susceptibility Mapping. *2nd International Geoinformation Days (IGD)*, May 2021.

Teke, A., Yilmaz, E. O., Kavzoglu T. (2021). Comparative Assessment of Deep Learning and Machine Learning Learning Models in Shallow Landslide Susceptibility. *International Symposium on Applied Geoinformatics*, December 2021.

Thai Pham, B., Tien Bui, D., Prakash, I. (2018). Landslide susceptibility modelling using

different advanced decision trees methods. *Civil Engineering and Environmental Systems*, **35**(1–4), 139–157.

Van Asch, T. W. J., Malet, J. P., van Beek, L. P. H., Amitrano, D. (2007). Techniques issues and advances in numerical modelling of landslide hazard. *Bulletin de La Societe Geologique de France*, **178**(2), 65–88. https://doi.org/10.2113/gssgfbull.178.2.65

Van Westen, C. J., Rengers, N., Soeters, R. (2003). Use of Geomorphological expert knowledge in indirect landslide hazard assessment. *Natural Hazards*, **30**, 399–419.

Wald, R., Khoshgoftaar, T. M., Dittman, D. (2012). Mean aggregation versus Robust Rank Aggregation for ensemble gene selection. *Proceedings - 2012 11th International Conference on Machine Learning and Applications*, ICMLA 2012, 1, 63–69.

Youssef, A. M., Pourghasemi, H. R., Pourtaghi, Z. S., Al-Katheeri, M. M. (2016). Landslide susceptibility mapping using random forest, boosted regression tree, classification and regression tree, and general linear models and comparison of their performance at Wadi Tayyah Basin, Asir Region, Saudi Arabia. *Landslides*, **13**, 839–856. https://doi.org/10.1007/s10346-015-0614-1

Received May 26, 2022, accepted June 17, 2022