GENETIC ALGORITHM BASED FEATURE SELECTION FOR LANDSLIDE SUSCEPTIBILITY MAPPING IN NORTHERN IRAN

Z. Nikraftar [a] , S. Rajabi-Kiasari [a,1], S. T. Seydi [a]

[a] School of Surveying and Geospatial Engineering, College of Engineering, University of Tehran, Tehran, Iran.
(xahir.nikraftar, saeed_rajabi, seydi.teymoor)@ut.ac.ir

KEYWORDS: Genetic algorithm, Iran, Landslide Susceptibility, Machine Learning, Feature Selection, GIS

ABSTRACT:

Recognizing where landslides are most likely to occur is crucial for land use planning and decision-making especially in the mountainous areas. A significant portion of northern Iran (NI) is prone to landslides due to its climatology, geological and topographical characteristics. The main objective of this study is to produce landslide susceptibility maps in NI applying three machine learning algorithms such as K-nearest neighbors (KNN), Support Vector Machines (SVM) and Random Forest (RF). Out of the total number of 1334 landslides identified in the study area, 894 (≈67%) locations were used for the landslide susceptibility maps, while the remaining 440 (≈33%) cases were utilized for the model validation. 21 landslide triggering factors including topographical, hydrological, lithological and Land cover types were extracted from the spatial database using SAGA (System for Automated Geoscientific Analyses), ArcGIS software and satellite images. Furthermore, a genetic algorithm was employed to select the most important informative features. Then, landslide susceptibility was analyzed by assessing the environmental feasibility of influential factors. The obtained results indicate that the RF model with the overall accuracy (OA) of 90.01% depicted a better performance than SVM (OA=81.06%) and KNN (OA=83.05%) models. The produced susceptibility maps can be productively practical for upcoming land use planning in NI.

1. INTRODUCTION

Landslides as responsible for at least 17% of all fatalities from natural hazards worldwide (Lacasse and Nadim, 2009) threaten human lives and environmental ecology. Different factors such as rainfall, earthquakes, and erosion of slope can trigger Landslides (Liu et al., 2013). Human activities such as deforestations and constructions are further causes of landslides in hilly areas. According to the (Iranian Landslide working party (ILWP), 2007) about 187 people have been killed in Iran by landslides and losses resulting from mass movements to the end of September 2007 have been estimated at 126,893 billion Iranian Rials (almost $12,700 million dollars) using the 4900 landslide database (Iranian Landslide working party (ILWP), 2007). However, the Northern provinces of Iran including Guilan, Mazandaran and Golestan are one of the most critical places vulnerable to landslide problems. The landslides observed and found in this area include old and new landslides (Shahabi et al., 2014).

The assessment of landslide hazard and risk has recently become a topic of interest for both geoscientists and the local administrations. By increasing availability of high-resolution spatial data sets, GIS, remote sensing, and computers with large and fast processing capacity, It has been feasible to partially automate the landslide hazard and susceptibility mapping process and thus minimize fieldwork (Tangestani, 2009). For modelling landslide susceptibility, a variety of algorithms have been proposed by researches in the literature. Nevertheless, in most cases, machine learning approaches performed better compared to other conventional analytical and expert opinion based methods (Zhou et al., 2018). For instance, artificial neural network (Chen et al., 2017b; Wang et al., 2019), random forest (Pourghasemi and Rahmati, 2018; Dou et al., 2019), support vector machine (Xu et al., 2012; Kumar et al., 2017), K-nearest neighbor (KNN) (Miloš Marjanovic et al., 2009; Chang et al., 2011), logistic regression (Hong et al., 2015; Zhou et al., 2018) and decision tree (Pradhan, 2013) models have been extensively used for analyzing landslide susceptibility and achieved high prediction accuracies. Most of the aforementioned studies confirmed the central role of geological factors (lithology, structure, and weathering), topographical factors (slope, elevation, aspect, etc.), soil parameters (soil depth and soil type), land use/cover and hydrologic conditions (rainfall) in generating accurate landslide susceptibility maps. Furthermore, other features, such as slope length, topographical wetness index (TWI), topographic position index (TPI), the vertical distance to the nearest channel network, relative slope position and valley depth have been reported to play important roles in landslide susceptibility modeling (Chauhan et al., 2010; Costanzo et al., 2012; Pourghasemi et al., 2013; Yilmaz et al., 2013; Massimo Conforti et al., 2014; Samia et al., 2017; Vargas-Cuervo et al., 2019).

Due to the variety of the landslide related parameters, it is not well clarified which combination of parameters would produce the best solution for a given landslide susceptibility problem. In addition, when all available parameters are used, it is more likely that correlated and redundant information to be considered, which may reduce the accuracy of a resulting map. To prevail over this problem, feature selection or dimensionality reduction techniques can be applied. They have been successfully used in many research areas; including environmental modelling, machine learning, data mining, statistics, pattern recognition, and remote sensing. Genetic Algorithm (GA) has been intensely employed for feature selection purposes. In regard of landslide susceptibility assessment, the application of the GA has been limited to optimization of algorithms (Chen et al., 2009; Liu et al., 2013; Kavzoglu et al., 2015; Chen et al., 2017a).

The main objective of this study is to seek the best combination of factors by integrating Machine learning approaches such as KNN, SVM and RF and applying feature selection regarding a GA in northern Iran. Performance analyses were conducted and evaluated based on differences in overall accuracies (OA).

[1] Corresponding author

2. STUDY AREA AND DATASET

The study area is located in the north of Iran including Mazandaran, Guilan and Golestan Provinces (Fig. 1). Due to its topographical and climatic conditions, it is highly prone to landslide activities. The intense rainfall and high slope gradients are the two most landslide predisposing factors. Digital elevation model (DEM), remotely sensed imagery and geological maps of the study area were used to create maps of the explanatory variables. Digital Elevation Model (DEM) with a 30m resolution was produced from topographic map applying a triangulated irregular network (TIN) model using digitized contours in ArcGIS software. The DEM was utilized for generating thematic maps of aspect, distance from faults, flow accumulation index1, rainfall_kriging and slope1. The SAGA software was also employed to produce maps of convergence index1, longitudinal curvature, LS factor1, plan curvature1, profile curvature1, relative slope position1, stream power index1 (SPI1), topographic wetness index1 (TWI1), topographic wetness index2 (TWI2), valley depth1, vertical distance to channel network1, catchment area1, catchment slope1, closed depressions1 and cross-sectional curvature1 factors while Landsat Thematic Mapper (TM) image was used to extract Normalized difference vegetation index (NDVI) maps of the study area through Google earth engine.

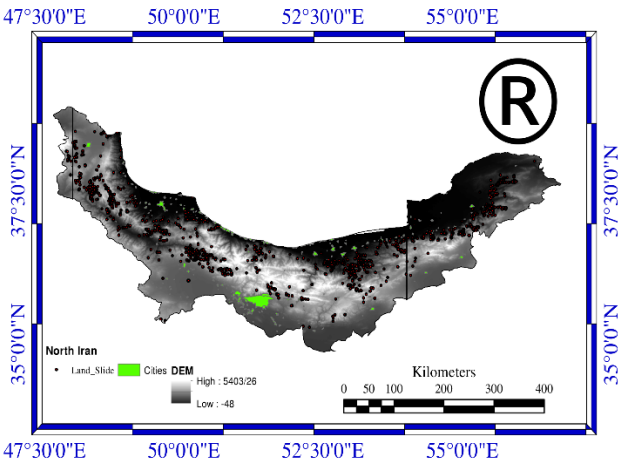

Figure 1. Location of the study area.

3. METHODOLOGY

3.1 Support Vector Machines

Support vector machine (SVM) is a supervised machine learning algorithm which is commonly used for classification purposes and is based on the statistical learning theory (Vapnik, 1998). Due to the acceptable result, The SVM algorithm converts to the most popular classifier method among remote sensing analysis. The main idea behind SVM is to find a hyperplane that maximizes the margin between the two classes (Vapnik, 1998). When the data are not linearly separable, SVM uses the kernel trick. A kernel is a dot product in a feature space. In the current study, we used one of the most popular kernel functions called radial basis function (RBF). Detailed explanations are given in (Yao et al., 2008).

3.2 K-nearest neighbors

The K-Nearest Neighbors (KNN) is a non-parametric classification method. KNN algorithm is one of the case-based learning methods, which keeps all the training data for classification. The KNN classifies a sample by assigning it a label as the most frequently represented among the K nearest samples; this means that a decision is made by examining the labels on the K-nearest neighbors and taking a vote. More details are found in (Witten et al., 2011).

3.3 Random Forest

RF has been proved to be the state-of-the-art ensemble classification technique that is a collection or ensemble of Classification and Regression Trees (Breiman, 2001) trained on datasets of the same size as a training set, called bootstraps, created from a random resampling on the training set itself. The RF algorithm provides a unique combination of prediction accuracy and model interpretability among popular machine learning methods.

3.4 Genetic Algorithm

Dimension reduction (DR) techniques are of great importance to diminish the complexity and redundancy of problems especially when issuing too many features as inputs. The DR applied in four main groups including feature extraction (FE), feature selection (FS), wrapper and filter. In the FS method, the informative bands are selected; therefore they do not alter the structure of data. FS is one of the most common methods for DR purposes. There are many methods based on Metaheuristic Optimization Algorithm (MOA). The most frequently asked of these methods is GA, an artificial Intelligence-Inspired technique. GA is a search heuristic algorithm that is inspired by biological evolution in nature as well as Darwin's evolution theory. This algorithm was introduced by (Holland, 1975). The GA works based on the creation of an initial population and evaluates them, finally selects the best chromosome based on defined criteria. The GA has a series of operators including selection, cross over and mutation. GA seeks to optimize the best bands according to the cost function. More details about the fundamentals of GA, their process and applications can be accessed in (Holland, 1975). For the application of the GA, optimal setting of parameters is a critical step for the success of the process. In this study, GA parameters were set according to Table 1.

Table 1. Parameters for GA

Size Initial Population	15
Length of Chromosome	24
Rate of Crossover	0.8
Rate of Mutation	0.01
Crossover	Two-point

3.5 Model Application

In this study, for collecting dataset and implementing methods, ArcGIS 10.0.2 software, SAGA GIS software (version 6.4.0), Google earth engine, and Matlab 2013a have been applied.

4. RESULTS AND DISCUSSION

Collecting 21 potentially landslide triggering factors in NI, this study aims to find the best performing approach plus the best input combination of features with the aid of GA. Out of 1334 landslides identified in the study area, 894 (≈67%) and 440 (≈33%) cases were used for the model calibration and validation, respectively. Thus, landslide susceptibility assessment was implemented based on integrating GA and three machine learning methods including KNN, SVM, and RF in the Matlab environment. Table 2 shows the selected factors by GA for each

of the applied models. 5 factors of flow accumulation index1, convergence index1, longitudinal curvature, stream power index1 (SPI1) and topographic wetness index2 (TWI2) were excluded from the models. The results of the different methods of landslide susceptibility mapping were evaluated to ensure the selection of a beneficial method and to improve the prediction accuracy of the landslide susceptibility map. Landslide susceptibility maps were tested based on the known landslide locations within the study area. For visual and easy interpretation of the areas, the resulting landslide map was classified into four susceptibility classes (Fig. 2).

Table 2. Selected factors in each of the KNN, SVM and RF models by GA

Factors	**KNN**	**SVM**	**RF**
Aspect	×		
distance from faults		×	×
flow accumulation index1			
convergence index1			
longitudinal curvature			
LS factor1		×	
NDVI		×	×
plan curvature1	×		
profile curvature1	×		
rainfall_kriging	×	×	×
relative slope position1		×	×
slope1		×	
stream power index1 (SPI1)			
topographic wetness index1 (TWI1)	×		×
topographic wetness index2(TWI2)			
valley depth1	×	×	×
vertical distance to channel network1	×	×	×
catchment area1			×
catchment slope1		×	×
closed depressions1	×	×	
Cross-sectional curvature1	×	×	

To evaluate the performance of the proposed framework, measurements of the overall accuracy (OA) was used. The OA value is the ratio of the number of correctly classified grid cells to the total number of grid cells, calculated as follows:

$$OA = \frac{a}{b} \times 100\% \quad (1)$$

Where a and b refer to the numbers of the correctly classified landslide or non-landslide grid cells and the total number of grid cells in the validation set, respectively. Obviously, a higher OA value implies better classification in precision.

The validation results revealed that the applied models had good accuracies (>0.8) in predicting future landslides in NI. However, the RF model with the OA of 90.01 performed highly better than KNN (OA=83.05) and SVM (OA=81.06) models. Furthermore, the Percentages of different landslide susceptibility classes by the employed methods were represented in Fig. 3. In regard to the RF model, the areas classified as low susceptibility cover 39.86% of the total area. Moderate and high-susceptible classes cover 10.35%, and 8.12% of the total area, respectively. The non-landslide class covers 41.66% of the study area. The areal extents of these sub-classes for KNN model were found to be 30.6%, 6.39%, 10.38%, and 52.61%, correspondingly, whereas landslide susceptibility map produced based on SVM, 34.5% of the study area has low susceptibility, and the moderate, high, and no landslides zones from 8.14%, 7.08%, and 50.26% of the study area, respectively.

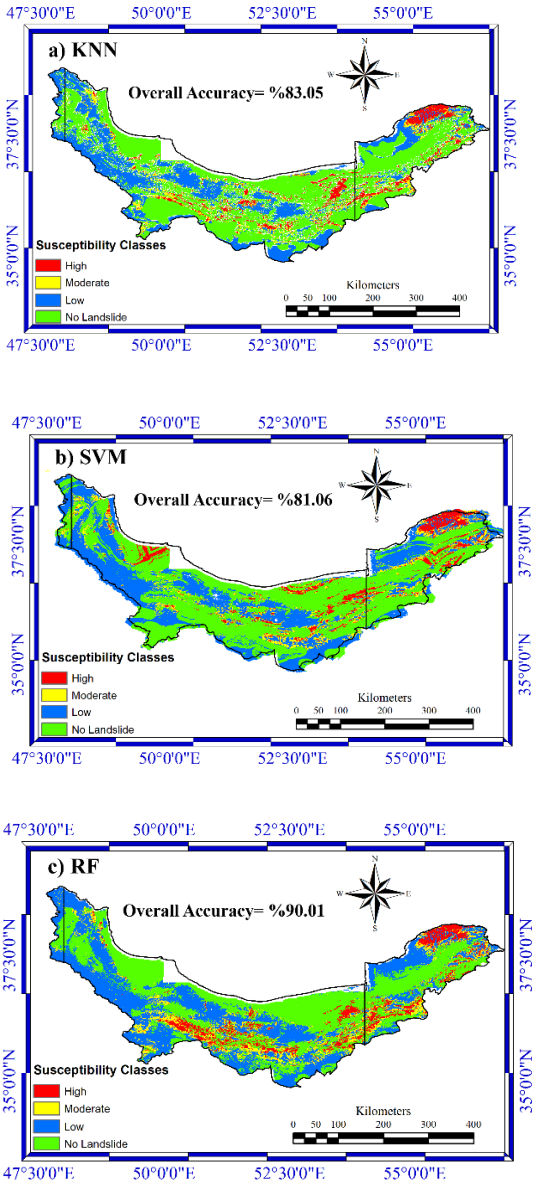

Figure 2. Landslide susceptibility maps generated using a) KNN b) SVM c) RF models based on GA optimization in NI.

5. CONCLUSION

Landslide susceptibility mapping plays an important role in providing a platform to decision-makers and authorities, particularly in landslide-prone areas. The current study dealt with the landslide susceptibility mapping using three machine

learning models namely KNN, SVM, and RF, in the critical northern Iran area. Hence, a set of landslide-controlling factors in a large amount of information were selected via the implementation of GA for each of the models. The results show that the RF model is a good estimator of landslide susceptibility in the study area.

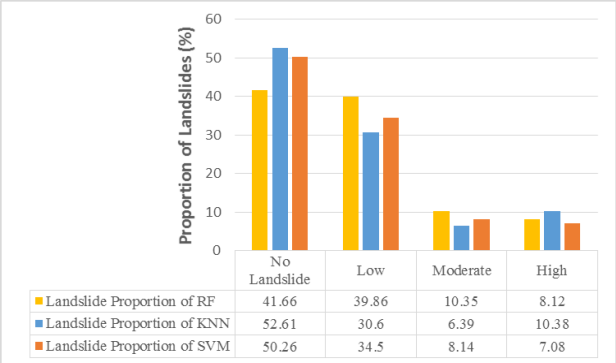

Figure 3. Percentages of different landslide susceptibility classes.

Moreover, The GA is proved to be extremely appropriate for the parameter identification in the slope stability analysis. The results from the implementation of three models also verified that landslides potentially will take place in the eastern part of NI, namely Northern Golestan. According to the RF model as the best performing model, about 8.12% of the study area is located in high susceptibility class which can be a matter of great interest to decision-makers and the local authorities for formulating land use planning strategies and implementing pragmatic measures.

6. REFERENCES

Breiman, L., 2001. Random Forests. Mach. Learn. 45, 5–32.

Chang, K.-T., Hwang, J.-T., Liu, J.-K., Wang, E.-H., Wang, C.-I., 2011. Apply two-hybrid methods on the rainfall-induced landslides interpretation. Presented at the 19th International Conference on Geoinformatics, IEEE, Shanghai, China.

Chauhan, S., Sharma, M., Arora, M.K., Gupta, N.K., 2010. Landslide Susceptibility Zonation through ratings derived from Artificial Neural Network. Int. J. Appl. Earth Obs. Geoinformation 12, 340–350.

Chen, W., Panahi, M., Pourghasemi, H.R., 2017a. Performance evaluation of GIS-based new ensemble data mining techniques of adaptive neuro-fuzzy inference system (ANFIS) with genetic algorithm (GA), differential evolution (DE), and particle swarm optimization (PSO) for landslide spatial modeling. CATENA 157, 310–324.

Chen, W., Pourghasemi, H.R., Kornejady, A., Zhang, N., 2017b. Landslide spatial modeling: Introducing new ensembles of ANN, MaxEnt, and SVM machine learning techniques. Geoderma 305, 314–327.

Chen, Y.-R., Chen, J.-W., Hsieh, S.-C., Ni, P.-N., 2009. The Application of Remote Sensing Technology to the Interpretation of Land Use for Rainfall-Induced Landslides Based on Genetic Algorithms and Artificial Neural Networks. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 2, 87–95.

Costanzo, D., Rotigliano, E., Irigaray, C., Jiménez-Peŕalvarez, J.D., Chacon, J., 2012. Factors selection in landslide susceptibility modelling on large scale following the gis matrix method: application to the river Beiro basin (Spain). Nat. Hazards Earth Syst. Sci. 12, 327–340.

Dou, J., Yunus, A.P., Bui, D.T., 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. Sci. Total Environ. 662, 332–346.

Holland, J.H., 1975. Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence. The U. of Michigan Press.

Hong, H., Pradhan, B., Xu, C., Bui, D., 2015. Spatial prediction of landslide hazard at the Yihuang area (China) using two-class kernel logistic regression, alternating decision tree and support vector machines. CATENA 133, 266–281.

Iranian Landslide working party (ILWP), 2007. , Iranian landslides list. Forest, Rangeland and Watershed Association. Iran.

Kavzoglu, T., Sahin, E.K., Colkesen, I., 2015. Selecting optimal conditioning factors in shallow translational landslide susceptibility mapping using genetic algorithm. Eng. Geol. 192, 101–112.

Kumar, D., Thakur, M., Dubey, C.S., Shukla, D.P., 2017. Landslide susceptibility mapping & prediction using Support Vector Machine for Mandakini River Basin, Garhwal Himalaya, India. Geomorphology 295, 115–125.

Lacasse, S., Nadim, F., 2009. Landslide Risk Assessment and Mitigation Strategy. Landslides – Disaster Risk Reduct. 31–61.

Liu, S.-H., Lin, C.-W., Tseng, C.-M., 2013. A statistical model for the impact of the 1999 Chi-Chi earthquake on the subsequent rainfall-induced landslides. Eng. Geol. 156, 11–19.

Massimo Conforti, Stefania Pascale, Gaetano Robustelli, Francesco Sdao, 2014. Evaluation of prediction capability of the artificial neural networks for mapping landslide susceptibility in the Turbolo River catchment (northern Calabria, Italy). CATENA 113, 336–350.

Miloš Marjanovic, Branislav Bajat, Miloš Kovacevic, 2009. Landslide Susceptibility Assessment with Machine Learning Algorithms. Presented at the International Conference on Intelligent Networking and Collaborative Systems, IEEE, Barcelona, Spain.

Pourghasemi, H., Pradhan, B., Gokceoglu, C., Moezzi, K.D., 2013. A comparative assessment of prediction capabilities of Dempster–Shafer and Weights-of-evidence models in landslide susceptibility mapping using GIS. Geomat. Nat. Hazards Risk 4, 93–118.

Pourghasemi, H.R., Rahmati, O., 2018. Prediction of the landslide susceptibility: Which algorithm, which precision? CATENA 162, 177–192.

Pradhan, B., 2013. A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS. Comput. Geosci. 51, 350–365.

Samia, J., Temme, A., Bregt, A., Wallinga, J., Guzzetti, F., Ardizzone, F., Rossi, M., 2017. Characterization and quantification of path dependency in landslide susceptibility. Geomorphology 292, 16–24.

Shahabi, H., Khezri, S., BinAhmad, B., Hashim, M., 2014. Landslide susceptibility mapping at central Zab

basin, Iran: A comparison between analytical hierarchy process, frequency ratio and logistic regression models. CATENA 115, 55–70.

Tangestani, M.H., 2009. A comparative study of Dempster-Shafer and fuzzy models for landslide susceptibility mapping using a GIS: An experience from Zagros Mountains, SW Iran. J. Asian Earth Sci. 35, 66–73.

Vapnik, V.N., 1998. Statistical Learning Theory. Wiley.

Vargas-Cuervo, G., Rotigliano, E., Conoscenti, C., 2019. Prediction of debris-avalanches and -flows triggered by a tropical storm by using a stochastic approach: An application to the events occurred in Mocoa (Colombia) on 1 April 2017. Geomorphology 339, 31–43.

Wang, Y., Fang, Z., Hong, H., 2019. Comparison of convolutional neural networks for landslide susceptibility mapping in Yanshan County, China. Sci. Total Environ. 666, 975–993.

Witten, I.H., Frank, E., Hall, M.A., 2011. Data Mining: Practical Machine Learning Tools and Techniques, 3rd ed, The Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann.

Xu, C., Dai, F., Xu, X., Lee, Y.H., 2012. GIS-based support vector machine modeling of earthquake-triggered landslide susceptibility in the Jianjiang River watershed, China. Geomorphology 145–146, 70–80.

Yao, X., Tham, L.G., Dai, F.C., 2008. Landslide susceptibility mapping based on Support Vector Machine: A case study on natural slopes of Hong Kong, China. Geomorphology 101, 572–582.

Yilmaz, I., Marschalko, M., Bednarik, M., 2013. An assessment on the use of bivariate, multivariate and soft computing techniques for collapse susceptibility in GIS environ. J. Earth Syst. Sci. 122, 371–388.

Zhou, C., Yin, K., Cao, Y., Ahmed, B., Li, Y., Catani, F., Pourghasemi, H.R., 2018. Landslide susceptibility modeling applying machine learning methods: A case study from Longju in the Three Gorges Reservoir area, China. Comput. Geosci. 112, 23–37.