EFFECTIVENESS OF MACHINE LEARNING ALGORITHMS IN LANDSLIDE SUSCEPTIBILITY MAPPING: A CASE STUDY OF TRABZON PROVINCE, TURKEY

Taskin Kavzoglu (1), Alihan Teke (1), Furkan Bilucan (1)

[1] Gebze Technical University, Departments of Geomatics Engineering, Gebze, Kocaeli, 41400, Turkey
Email: kavzoglu@gtu.edu.tr; a.teke2020@gtu.edu.tr; f.bilucan2020@gtu.edu.tr

KEY WORDS: Landslide Susceptibility, Random Forest, AdaBoost, Logistic Regression, Machine Learning

ABSTRACT: Determination of vulnerable zones to landslide is of utmost importance for disaster management and hazard mitigation. Therefore, one of the most significant processes in disaster planning is the production of accurate and up-to-date landslide susceptibility maps. The primary goal of this present work is to investigate the effectiveness of different machine learning algorithms considering random forest (RF), AdaBoost (AB), and logistic regression (LR) for generating landslide susceptibility map of Trabzon province, located in the northeast of Turkey. For this purpose, 12 most widely used landslide-conditioning factors (slope, elevation, plan curvature, profile curvature, slope length, topographical position index, topographical ruggedness index, topographical wetness index, stream power index, normalized difference vegetation index, distance to roads, and distance to rivers) were utilized to produce landslide susceptibility maps and the results were evaluated by utilizing receiver operating characteristic, area under curve (AUC), and overall accuracy (OA) according to confusion matrices. The validation results indicated that AUC obtained using RF, AB, LR methods were computed as 0.929, 0.910, and 0.744, respectively. Additionally, the statistical significance of the methods was evaluated using the McNemar's test and found that RF and AB methods produced similar results but their results significantly differ from that of the LR method.

1. INTRODUCTION

Landslides are defined as mass movements of soil or rock, and natural materials owing to the effect of gravity (Highland and Bobrowsky, 2008). Landslides are major geological disasters, causing property damages, economic devastation, casualties, and injuries. Furthermore, landslides lead to the destruction of natural resources and vegetation due to the deformations they create in the zone where they occur. Identification of landslide prone-areas ensures both protection of human life and avoidance of economic losses. Landslide susceptibility maps produced effectively, and precisely are functional tools that generate crucial information to planners and government agencies in terms of natural risk management, infrastructure, and land planning. In addition to these benefits, the effective use of these maps can significantly decrease harm capacity and other severe effects of landslides.

Considering the incidence of natural disasters in Turkey, landslides rank takes the second after the earthquakes and represents about 30% of the entire destruction (Hasekioğulları and Ercanoglu, 2012). In previous years, landslides have been experienced commonly and cause significant damages in the Eastern Blacksea region of Turkey. Specifically, the geological, topographic, and climatic structure of the region has triggered the landslide phenomenon (Yalcin, 2011). For this region, an increasing amount of landslides have occurred in the province of Trabzon, which resulted in many people losing their lives (Bayrak and Ulukavak, 2009).

Correspondingly the rapid developments in the geographical information system, remote sensing technologies, and soft computing techniques, there have been expanding concentrates on studies about landslide susceptibility. The performances of susceptibility zonation depend essentially on the methods utilized to produced landslide susceptibility models and the quality of the collected geo-environmental data. During the last three decades, the use of machine learning algorithms in the production of susceptibility maps has become the key instrument owing to their success and predictive performance against conventional statistical methods. Up to now, various machine learning algorithms including logistic regression (Ayalew and Yamagishi, 2005; Lee and Sambath, 2006), support vector machine (Pradhan, 2013), decision tree (Saito et al., 2009), artificial neural network (Gómez and Kavzoglu, 2005), random forest (Sahin et al., 2020), Naïve Bayes (Pham et al., 2016) have been intensively utilized for generating landslide susceptibility maps. Determination of landslide contributing factors is one of the quite critical and challenging tasks in the field of landslide susceptibility. However, there is no widespread consensus on the selection of contributing factors (Ayalew and Yamagishi, 2005). The main reason for this is that each study region has its specific characteristics (Van Westen et al., 2003). To be more specific, while any factor utilized in landslide susceptibility studies may be a contributing factor for a particular region, it may not be for another (Kavzoglu et al., 2015).

In this current work, the random forest, AdaBoost, and logistic regression algorithms were implemented to generate the landslide susceptibility model of Trabzon province, Turkey. Evaluation of effectiveness and performances of three machine learning algorithms were compared using the accuracy assessment metric in terms of overall accuracy, and area under curve (AUC). Furthermore, non-parametric McNemar's test was applied to determine the statistical significance of differences among the outcomes of employed machine learning techniques.

2. STUDY AREA AND DATASET

The study area is situated in the northeast part of Turkey (Fig. 1) and cover an area of about 4,685 km^2 and situated between 39° 15′ and 40° 15′ E and 41° 8′ and 40° 30′ N. In the study area, which has a mountainous geographical structure and irregular precipitation regime, precipitation is observed approximately all seasons and the mean annual precipitation amount per square meter is 830 mm. For the study area, one of the major elementary factors that trigger the landslide phenomenon is the high slope reaching about 85°. Additionally, since heavy precipitation and dense vegetation increase the speed of weathering, the study area has become vulnerable to landslide events (Yalcin, 2011). Furthermore, not only natural factors but also human attempts including infrastructure, superstructure, and deforestation have a significant role in landslide occurrences (Kavzoglu et al., 2014).

In the procedure of reliable and accurate production of landslide susceptibility maps, one of the key steps is the preparation of landslide contributing components. In this current work, slope, elevation, plan curvature, profile curvature, slope length, topographical position index (TPI), topographical ruggedness index (TRI), topographical wetness index (TWI), stream power index (SPI), normalized difference vegetation index (NDVI), distance to roads, and distance to rivers were considered as major contributing factors to generate landslide susceptibility models according to characteristics of the study area and analyzed the existing/collected data (Fig. 2). The NDVI map was produced from Landsat-8 satellite images and the Euclidean distance analysis was implemented to create distance to rivers and distance to roads layers. The remaining contributing factors have been generated from the 30 m spatial resolution SRTM elevation model using the GIS environment.

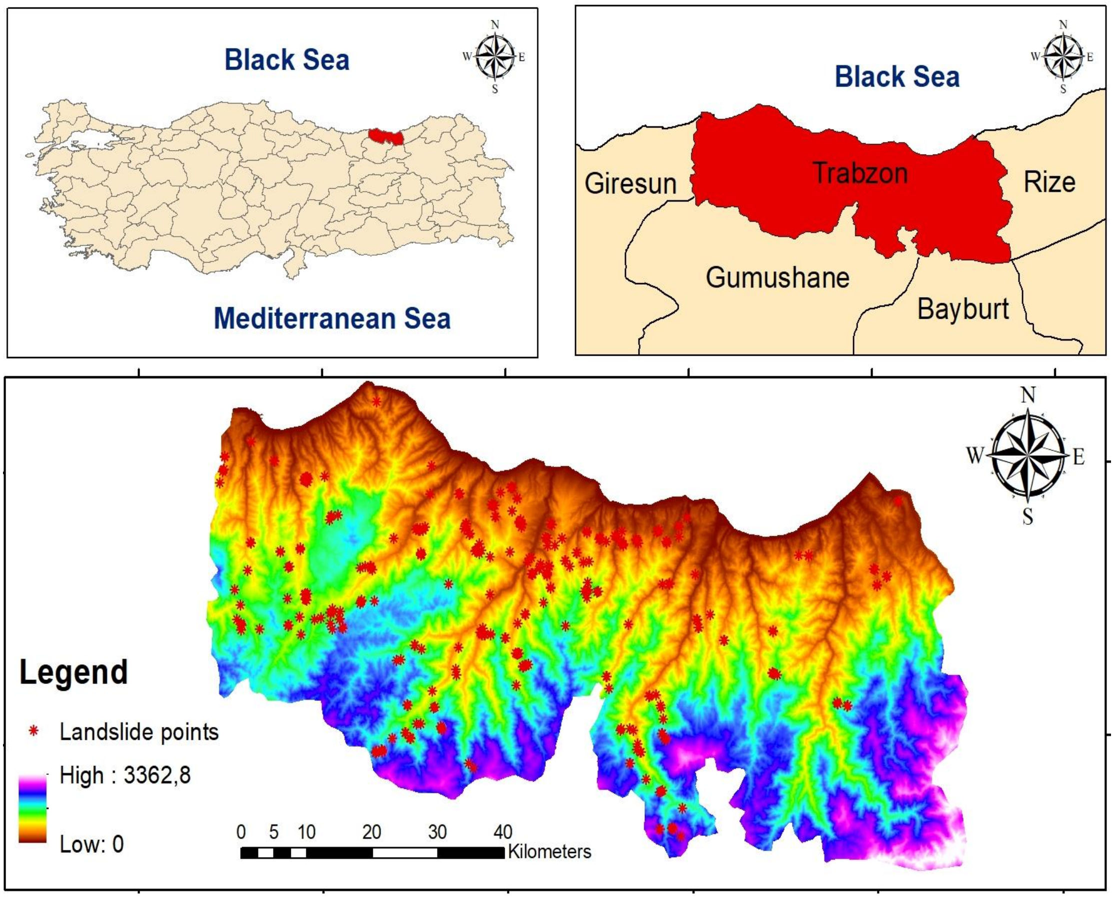
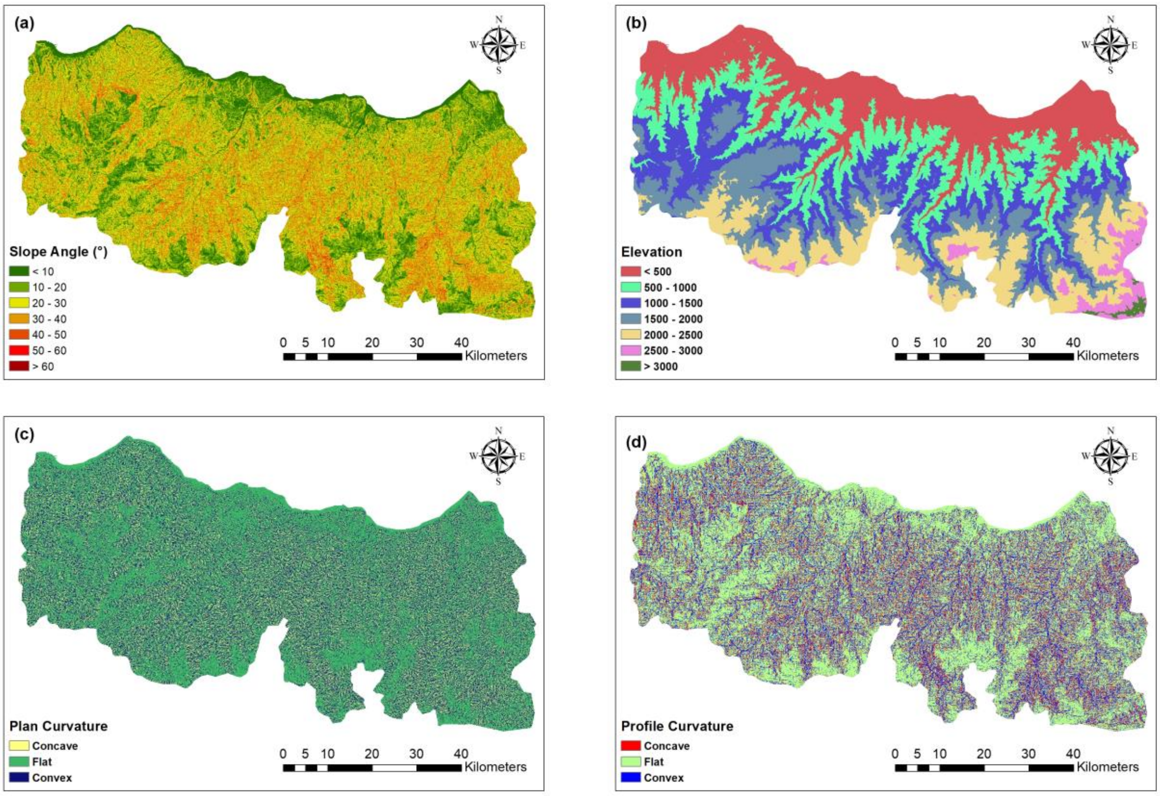


Figure 1. Locations of the landslide and study area, Trabzon province in Turkey



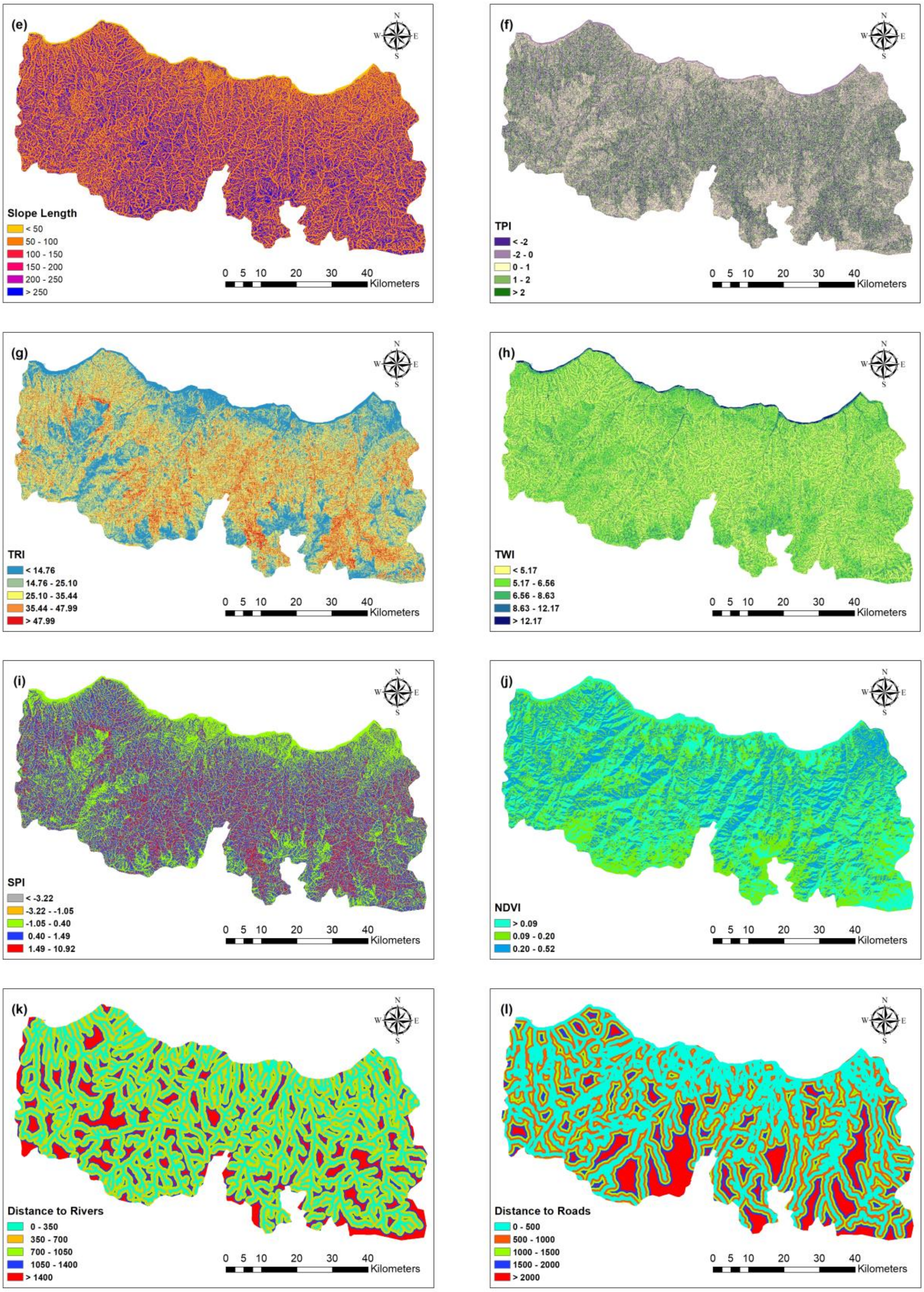


Figure 2. Landslide contributing factors; (a) Elevation, (b) Slope, (c) Plan curvature, (d) Profile curvature, (e) Slope length, (f) TPI, (g) TRI, (h) TWI, (i)SPI, (j) NDVI, (k) Distance to rivers, (l) Distance to roads

3. METHODOLOGY

3.1 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. To overcome the overfitting problem which is one of the most challenging issues of conventional machine learning methods, RF employs 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 compute the overall accuracy of the tree model. Ultimately, the majority voting rule is implemented in the prediction of the class labels of unknown samples (Kavzoglu, 2017).

3.2 AdaBoost (AB)

Known as Adaptive Boosting (AB), introduced by Freund and Schapire (1997) is one of the most powerful ensemble-learning methods that has been utilized to boost the predictive performance of the classifier techniques. The main idea behind the AB is to employ various weak classifiers on the same training data and then integrate these classifiers to create a stronger final classifier (Wu et al., 2020). The algorithm utilizes an adaptive resampling technique in the selection of training samples. The procedure of AB algorithm consists of three essential parts. Firstly, a subset of training data is created and equal weights are assigned to each sample of that subset. Secondly, the weights of misclassified samples are assigned to higher weights, while the weights of accurately classified samples remain constant. In the last phase, the weights of all samples in the training data set are normalized. This process is performed iteratively until it reaches the optimum performance of classifiers (Tien Bui et al., 2016)

3.3 Logistic regression (LR)

LR model is constructed with dichotomous target variables such as yes or no (i.e. 1 or 0, absence or presence) and single or multiple independent variables that determine the model. The core idea behind the LR method is to specify the optimum suitable model to characterize the connection between the target variable and independent variables. The advantage of the model is that the target variables can be measured in different types such as a nominal, ordinal, interval, or ratio scale (Yesilnacar and Topal, 2005), and also normal distribution is not necessarily required (Lee and Sambath, 2006).

4. RESULTS

The performances of the landslide susceptibility models can be validated by using landslide inventory data which were excluded in the stage of the construction of the model. Of 170 landslide polygons with known locations on the landslide inventory map, 119 (~70%) landslide locations were utilized to produce landslide susceptibility maps, whereas the rest 51(~30%) were utilized for the model assessment. The predictive performances of the landslide susceptibility models were computed by using the overall accuracy and the area under the ROC curve or simply AUC. In order to assess the effectiveness of the landslide susceptibility maps generated by the three methods, the overall accuracies of the models were calculated using the confusion matrix. The RF, AB, and LR algorithms yield classification results with overall accuracies of 93.960%, 92.277%, 78.323% respectively as shown in Table 1. To be more specific, while the RF method was superior to the AB method with minor variation, it resulted in a very high overall accuracy compared to the LR method. Among the methods utilized in landslide susceptibility assessment, LR resulted in the lowest overall accuracy.

Table 1. Landslide susceptibility mapping outcomes in terms of overall accuracy and area under curve values

Methods	**OA (%)**	**AUC**
Random Forest	93.960	0.929
AdaBoost	92.277	0.910
Logistic Regression	78.323	0.744

Another accuracy assessment metric that is employed to interpret the statistical reliability of model performances is the AUC value. In landslide susceptibility studies, it has been frequently used due to both its ability to be displayed graphical plotting in determining the effectiveness of the binary classification algorithms and good indicator in assessing the predictive performance of models (Kalantar et al., 2018). The highest AUC value, ranging between 0.5 and 1.0, indicates the optimum model. While the ROC curves and computed AUC values for the RF and AB methods were found to be 0.929 and 0.910, respectively, pointing out to the acceptable level of performance, AUC values for the LR method computed as 0.744, indicating reasonable discrimination ability. The results of AUC analysis obviously emphasized that RF and AB methods outperformed the LR method and the improvement in AUC values was approximately 18% and 17%, respectively.

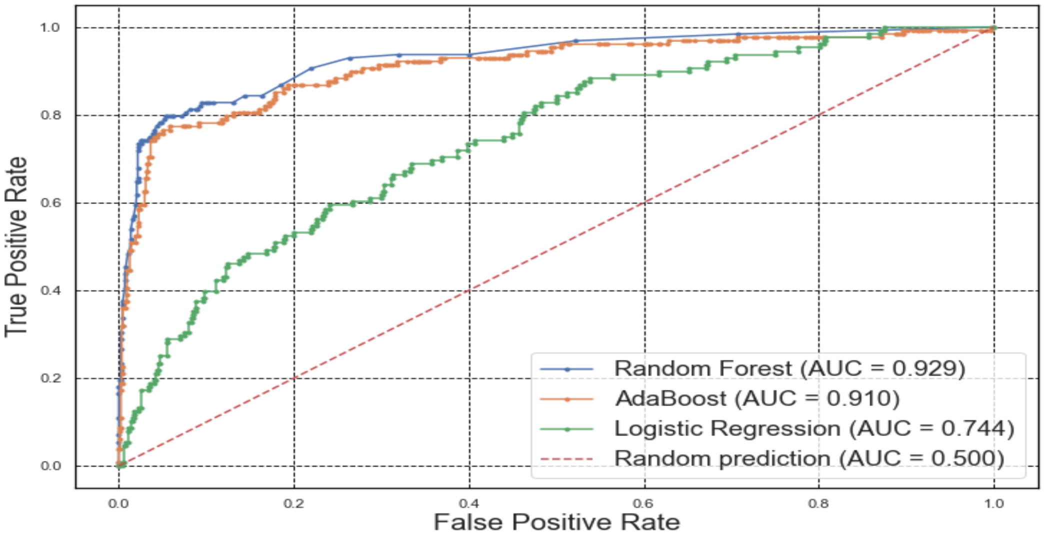
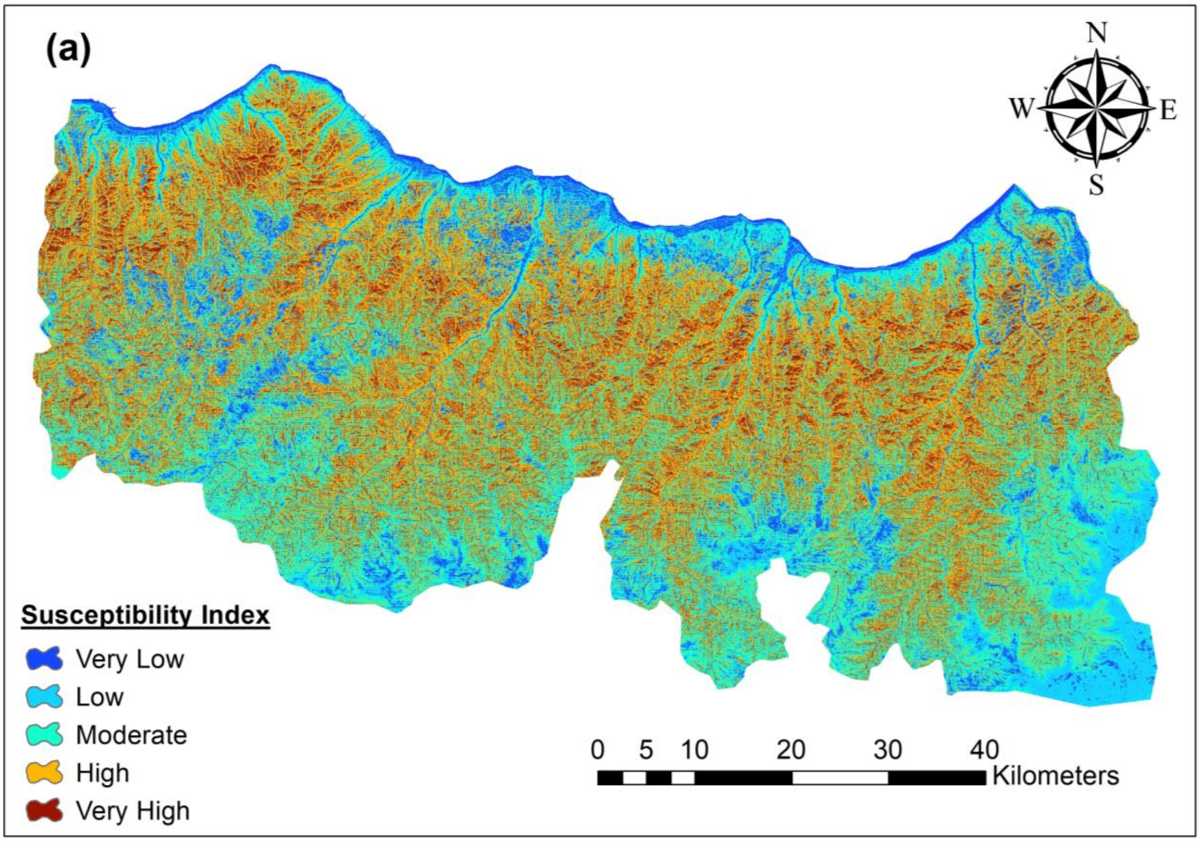
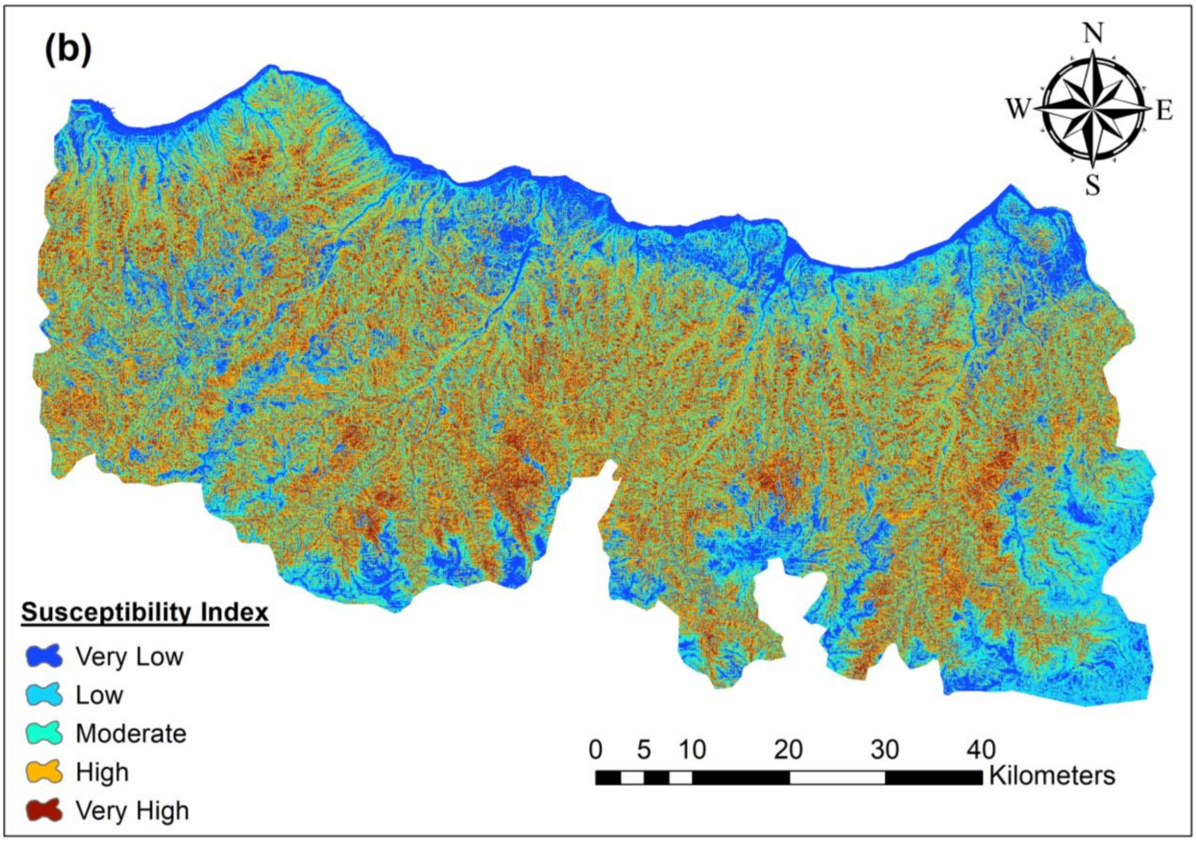


Figure 3. ROC statistics and AUC analysis estimated for RF, AB, and LR methods

Together with the standard accuracy assessment metrics, non-parametric McNemar's test was implemented to investigate the statistical significance of the accuracy effectiveness of the susceptibility maps produced with different methods. If the statistical value calculated as a result of McNemar's test is greater than the chi-square table value (3.84 at 95% confidence interval), it can be inferred that the difference between the results of the two classification methods is statistically significant. Considering the estimated statistical test values of RF and AB were compared, it was found that both methods produced similar results at the statistical significance level (3.02 < 3.84). When the performance of the LR method was compared with RF and AB methods, it was found that the LR method produced statistically different results.


(a)
N
W
E
S
Susceptibility Index
Very Low
Low
Moderate
High
Very High
0
5
10
20
30
40
Kilometers


(b)
N
W
E
S
Susceptibility Index
Very Low
Low
Moderate
High
Very High
0
5
10
20
30
40
Kilometers

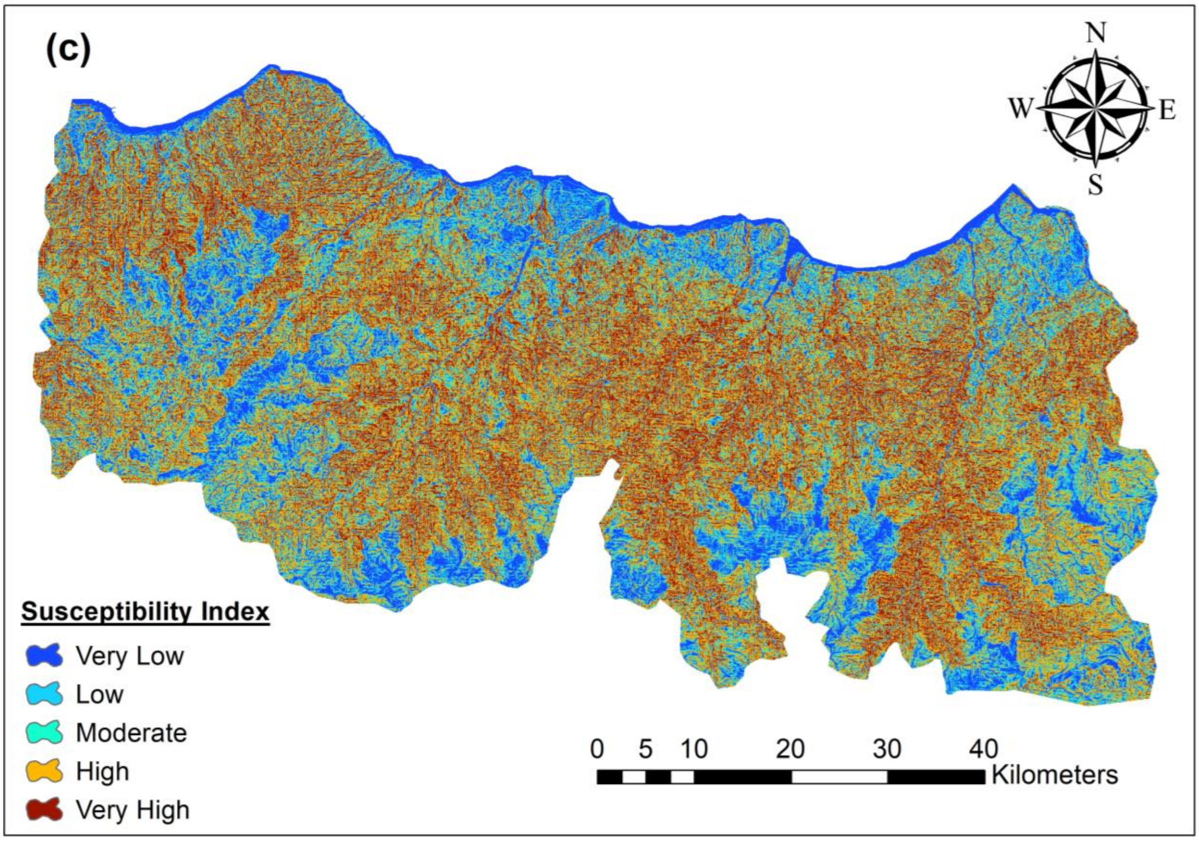


Figure 3. Landslide susceptibility maps produced by (a) RF, (b) AB, (c) LR methods

5. CONCLUSION

Landslide susceptibility mapping is a complicated task and comprises several procedures that include steps such as correct identification of contributing factors, selection of accurate measurement tools, evaluation of the output maps with correct metrics, and keeping them up-to-date. Until now, many techniques varying in huge scope for producing landslide susceptibility maps have been proposed and their effectiveness compared to each other in the literature. In this present work, the effectiveness of three machine learning techniques was assessed using statistical measures for the identification of landslide susceptibility of Trabzon province, situated in the northeast part of Turkey. These techniques were employed using 12 landslide conditioning factors. Standard accuracy assessment metrics together with the McNemar's test was implemented. According to the indication of the work, some significant inferences can be drawn for the performances of the machine learning techniques. Firstly, it was observed that RF and AB methods were more effective than the LR method in terms of overall accuracy and AUC value. Hence, it can be concluded that RF and AB methods considerably increase the predictive performance of the model outcomes. Secondly, according to McNemar's test results, it was determined that RF and AB methods produced similar results but significantly different results from the LR method. Finally, it can be clearly concluded that ensemble-learning techniques, namely the RF and AB methods, outperformed to the LR method, which a well-known conventional technique. To sum up, RF and AB were found to be functional methods for modeling landslide susceptibility compared with the LR method.

REFERENCES

Ayalew, L., and 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), pp. 15-31.

Bayrak, T., and Ulukavak, M., 2009. Trabzon Landslides. Electronic Journal of Map Technologies 1(2), pp. 20-30.

Breiman, L., 2001. Random forests. Machine Learning., 45(1), pp. 5-32.

Freund, Y., and Schapire, R. E., 1997. A Decision-Theoretic Generalization of On-Line Learning and Application to Boosting. Journal of Computer and System Sciences, 55(1), pp. 119-139.

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

Hasekioğulları, G. D., and Ercanoglu, M., 2012. A new approach to use AHP in landslide susceptibility mapping: A case study at Yenice (Karabuk, NW Turkey). Natural Hazards, 63(2), pp. 1157–1179.

Highland, L. M., and Bobrowsky, P., 2008. Introduction The Landslide Handbook-A Guide to Understanding Landslides. U.S. Geological Survey Circular 1325, Virginia.

Kalantar, B., Pradhan, B., Naghibi, S. A., Motevalli, A., and Mansor, S., 2018. Assessment of the effects of training data selection on the landslide susceptibility mapping: a comparison between support vector machine (SVM), logistic regression (LR) and artificial neural networks (ANN). Geomatics, Natural Hazards and Risk, 9(1), pp. 49-69.

Kavzoglu, T., Sahin, E. K., and Colkesen, I., 2014. Landslide susceptibility mapping using GIS-based multi-criteria decision analysis, support vector machines, and logistic regression. Landslides, 11(3), pp. 425–439.

Kavzoglu, T., Sahin, E. K., and Colkesen, I., 2015. Selecting optimal conditioning factors in shallow translational landslide susceptibility mapping using genetic algorithm. Engineering Geology, 192, pp. 101-112.

Kavzoglu, T., 2017. Object-Oriented Random Forest for High Resolution Land Cover Mapping Using Quickbird-2 Imagery. In: Handbook of Neural Computation, edited by Samui, P., Roy, S. S., and Balas, V. E., Amsterdam: Elsevier, pp. 607-619.

Lee, S., and Sambath, T., 2006. Landslide susceptibility mapping in the Damrei Romel area, Cambodia using frequency ratio and logistic regression models. Environmental Geology, 50(6), pp. 847–855.

Pham, B. T., Pradhan, B., Tien Bui, D., Prakash, I., and Dholakia, M. B., 2016. A comparative study of different machine learning methods for landslide susceptibility assessment: A case study of Uttarakhand area (India). Environmental Modelling & Software, 84, pp. 240–250.

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. Computers & Geosciences, 51, pp. 350–365.

Sahin, E. K., Colkesen, I., and Kavzoglu, T., 2020. A comparative assessment of canonical correlation forest, random forest, rotation forest and logistic regression methods for landslide susceptibility mapping. Geocarto International, 35(4), pp. 341-363.

Saito, H., Nakayama, D., and Matsuyama, H., 2009. Comparison of landslide susceptibility based on a decision-tree model and actual landslide occurrence: The Akaishi Mountains, Japan. Geomorphology, 109(3–4), pp. 108–121.

Tien Bui, D., Ho, T. C., Pradhan, B., Pham, B. T., Nhu, V. H., and Revhaug, I., 2016. GIS-based modeling of rainfall-induced landslides using data mining-based functional trees classifier with AdaBoost, Bagging, and MultiBoost ensemble frameworks. Environmental Earth Sciences, 75(14), pp. 1–22.

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

Wu, Y., Ke, Y., Chen, Z., Liang, S., Zhao, H., and Hong, H., 2020. Application of alternating decision tree with AdaBoost and bagging ensembles for landslide susceptibility mapping. Catena, 187(2020), Article 104396.

Yalcin, A., 2011. A geotechnical study on the landslides in the Trabzon Province, NE, Turkey. Applied Clay Science, 52(1-2), pp. 11-19.

Yesilnacar, E., and Topal, T., 2005. Landslide susceptibility mapping: A comparison of logistic regression and neural networks methods in a medium scale study, Hendek region (Turkey). Engineering Geology, 79(3–4), pp. 251–266.