Landslides (2016) 13:839–856
DOI 10.1007/s10346-015-0614-1
Received: 3 December 2014
Accepted: 15 July 2015
Published online: 8 August 2015

Ahmed Mohamed Youssef · Hamid Reza Pourghasemi · Zohre Sadat Pourtaghi · Mohamed M. Al-Katheeri

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

Abstract The purpose of the current study is to produce landslide susceptibility maps using different data mining models. Four modeling techniques, namely random forest (RF), boosted regression tree (BRT), classification and regression tree (CART), and general linear (GLM) are used, and their results are compared for landslides susceptibility mapping at the Wadi Tayyah Basin, Asir Region, Saudi Arabia. Landslide locations were identified and mapped from the interpretation of different data types, including high-resolution satellite images, topographic maps, historical records, and extensive field surveys. In total, 125 landslide locations were mapped using ArcGIS 10.2, and the locations were divided into two groups; training (70%) and validating (25%), respectively. Eleven layers of landslide-conditioning factors were prepared, including slope aspect, altitude, distance from faults, lithology, plan curvature, profile curvature, rainfall, distance from streams, distance from roads, slope angle, and land use. The relationships between the landslide-conditioning factors and the landslide inventory map were calculated using the mentioned 32 models (RF, BRT, CART, and generalized additive (GAM)). The models' results were compared with landslide locations, which were not used during the models' training. The receiver operating characteristics (ROC), including the area under the curve (AUC), was used to assess the accuracy of the models. The success (training data) and prediction (validation data) rate curves were calculated. The results showed that the AUC for success rates are 0.783 (78.3%), 0.958 (95.8%), 0.816 (81.6%), and 0.821 (82.1%) for RF, BRT, CART, and GLM models, respectively. The prediction rates are 0.812 (81.2%), 0.856 (85.6%), 0.862 (86.2%), and 0.769 (76.9%) for RF, BRT, CART, and GLM models, respectively. Subsequently, landslide susceptibility maps were divided into four classes, including low, moderate, high, and very high susceptibility. The results revealed that the RF, BRT, CART, and GLM models produced reasonable accuracy in landslide susceptibility mapping. The outcome maps would be useful for general planned development activities in the future, such as choosing new urban areas and infrastructural activities, as well as for environmental protection.

Keywords Landslide susceptibility mapping · Random forest · Boosted regression tree · Classification and regression tree · General linear model · Saudi Arabia

Introduction

Landslides are common erosion processes in the southwestern and western parts of Saudi Arabia, where many cities, highways, and roads are located along the Arabian shield, the main mountain ranges (Elkadiri et al. 2014; Youssef et al. 2014a, b, c). These mountain ranges are well known by the steep scarps that lead to the generation of spectacular landforms. In the major Wadi Basin particularly in the southwestern part of Saudi Arabia, the landslides are sudden mass movements that are usually initiated by intense precipitation, generated in rainstorms (Youssef et al. 2013, 2014c). In Saudi Arabia, during the last few decades, urban areas and many escarpment roads are quickly expanding toward the rugged mountainous and steep slopes, and accordingly, landslides occur more frequently (Youssef et al. 2012, 2013). In Saudi Arabia, different types of landslides were detected in different areas, including rock falls, rock and soil sliding, and debris flows (Youssef et al. 2012, 2013, 2014a).

Landslides often result in loss of human life and property and represent the most damaging natural hazards in the mountainous areas of different parts of the world. There are different landslide-conditioning factors that could be used to prepare the landslide susceptibility map for any area. These factors include lithology, lineaments, geomorphology, soil type and depth, slope angle, slope aspect, curvature, altitude, engineering properties of the lithological material, land use patterns, and drainage networks. Other external factors can play an essential part in triggering landslides, including heavy rainfall, earthquakes, volcanoes, and anthropogenic activities. Various studies have been carried out on landslide susceptibility assessment using remote-sensing and GIS techniques (e.g., Saha et al. 2005; Pradhan and Youssef 2010; Pradhan et al. 2010; Bednarik et al. 2012; Mohammady et al. 2012; Pourghasemi et al. 2012b, 2013a, b; Devkota et al. 2013; Xu 2013; Regmi et al. 2014).

Other types of studies in different areas in Saudi Arabia related to the landslide susceptibility assessment have been done, including landslide susceptibility mapping along the Al Hasher escarpment road using frequency ratio and index-of-entropy models (Youssef et al. 2014a). Elkadiri et al. (2014) studied the debris-flow susceptibility assessment using artificial neural networks and logistic regression models. Youssef et al. (2014b) assessed landslide susceptibility using ensemble FR and Logistic regression models for Fayfa Area, Saudi Arabia.

Various modeling approaches were applied to assess landslide susceptibility in any specific area which belongs to one of the three main groups: (1) heuristic, (2) deterministic, and (3) statistical (Committee on the Review of the National Landslide Hazards Mitigation Strategy 2004). Each of them has its own characteristics and disadvantages. Heuristic models rely mainly on the expert knowledge to assign weights to the various conditioning factors (e.g., Dai and Lee 2002; Dahal et al. 2008a, b). The heuristic models are highly subjective and depend on the site itself (Dai et al. 2001). Deterministic models are completely based on mathematical relationships that depend on the physical laws in which the relation between resisting and driving forces can be calculated for the mass movements. The most important data that are required for the deterministic models are engineering characteristics of the rocks and soils, slope geometry, discontinuity characteristics, and hydrological conditions (Yilmaz 2009). The main problem with the deterministic models is the need for intensive data from individual slopes, which makes these methods effective for studying only small areas (Ayalew and Yamagishi 2005). Statistical models were also used to analyze the landslide susceptibility (e.g., logistic regression, neural networks, index-of-entropy, GIS-based weighted linear combination, frequency ratio, general linear models, spatial multi-criteria evaluation, and neuro-fuzzy models Ayalew et al. 2004; Remondo et al. 2005; Abella and Van Westen 2007; Lee and Pradhan 2007; Mathew et al. 2009; Pradhan and Lee 2010; Devkota et al. 2013; Pourghasemi et al. 2013a, b; Schleier et al. 2014; Wu et al. 2014; Youssef et al. 2014a, b).

Other models such as random forest (RF), boosted regression tree (BRT), classification and regression tree (CART), and general linear (GLM) were applied in bio-informatics, genetics, eco-hydrological, ecological, and earth sciences (Pal 2005; Ham et al. 2005; Diaz-Uriate and de Andres 2006; Chen and Liu 2006; Gislason et al. 2006; Schröder et al. 2010). There are different new and powerful methods of boosted regression trees, multivariate adaptive regression splines, and maximum entropy methods for predicting the distribution of shallow landslides in tropical mountain rainforests in southern Ecuador. Catani et al. (2013) studied the influence of sensitivity and scaling issues in the landslide susceptibility mapping using random forests. They indicated that the unit (scale) and the training process strongly influence the classification accuracy and the prediction process. Paudel and Oguchi (2014) used the random forests in landslides susceptibility analysis for Tokamachi area, Niigata, Japan.

The Asir Region in Saudi Arabia is exposed to landslides and mass wasting episodes, which occur from time to time according to climatic and physiographic conditions. In recent years, several attempts of slope stability and landslide susceptibility mapping were applied in Saudi Arabia (Youssef et al. 2012, 2013, 2014a, b; Youssef and Maerz 2013; Elkadiri et al. 2014; Maerz et al. 2014).

In the current study, four data mining models were adapted to develop a landslide susceptibility map using remote-sensing and GIS techniques. These models are random forest, boosted regression tree, classification and regression tree, and general linear models, which were selected for a number of reasons, including being newly applied in the field of landslide susceptibility in Saudi Arabia, suitable for regional- and semi regional-scale applications, and relying mainly on remote-sensing datasets rather than extensive field surveys. We believe that the results obtained from our study provide a considerable contribution to the landslide literature. The landslide susceptibility maps can identify and delineate landslide-prone areas, so that planners and decision makers can choose favorable locations for development schemes, such as new urban areas.

Study area and geological setting

Wadi Tayyah is located in the Asir Region, Saudi Arabia (Fig. 1a). The study area covers an area almost 629.8 km^2 and is located between 17° 46′ 31″ and 18° 17′ 9″ N and 42° 14′ 55″ and 42° 48′ 30″ E (Fig. 1b). It represents a part of Abha Highland (which is related to the Arabian shield). The Shear escarpment highway, passes along this valley, descending from the top of the escarpment near the City of Abha down to the City of Mahail Asir, then to the coastal zone of western Saudi Arabia. This highway represents one of the most important highways in the area. It was constructed through this extremely difficult mountainous terrain almost 32 years ago. This road connects the Red Sea coastal areas (southern and western region of the Saudi Arabia) with the Asir and Najran Regions. It is used by private vehicles and light and heavy duty trucks. The road is located in the highly rigged mountainous area situated in the north of Abha City (Fig. 1b). The length of the Shear escarpment highway in the study area is about 16 km, measured from the top of the escarpment (2200 m above sea level (a.s.l.)) from east to the City of Mahail Asir (approximately 700 m a.s.l.). The altitude of the basin ranges from 221 to 2,988 m a.s.l. The study area is elongated in shape and dissected by many small wadis (valleys) that drain their waters toward Wadi Tayyah. The slope angles range from 0° to as much as 77.3°. The study area is heterogeneous in terms of terrain complexity (wadis and mountainous). Many urban areas (villages) are located inside the study as shown in (Fig. 1b).

The study area is mainly located in the different geologic units, which were digitized from the 1:250,000-scale Abha quadrangle geologic map (GM-75) (Greenwood 1985) (Fig. 1b). These geologic units are as follows: (1) alluvium and gravel—this unit includes wadi alluvium, dissected terraces, colluvium, and fan deposits. (2) The Bahah group—it is a major component in the western part of the Tayyah belt. It consists of a fault-bounded blocks (Greenwood 1985), including abundant volcanic greywacke, local boulder conglomerate, carbonaceous shale, slate, chert, bedded tuff, and interbeds of volcanic flow rock. These rocks are weakly to moderately cleavaged and highly cleavaged near faults. They are characterized by the presence of one cleavage (schistosity) which has steep dips toward east or west. Greywacke is massive to thinly bedded, including some sedimentary structures (grading, cross bedding, and lamina bedding). Massive greywacke forms thick beds from 1 to 3 m and interlayered with fine-grained and laminated bedded sections. The greywacke and inter-bedded slate are strongly metamorphosed to greenschist facies. Some intrusive rocks, including granodiorite and granite were encountered in the Tayyah belt. Near the intrusive rock, contact amphibolite-grade metamorphic rocks were encountered. (3) Jeddah group—it is inter-bedded with the Bahah group and bounded on the west by a major fault that separates it from the Ablah group. It includes basalt and andesite flows, pillow lava, flow breccia, and pyroclastic rocks. Dacitic pyroclastic rocks and volcanoclastic conglomeratic, coarse- to fine-grained greywacke and phyllite are inter-bedded with the volcanic rocks. These rocks are regionally metamorphosed to the greenschist facies, and sometimes to higher grades. Isoclinal to open folds are associated with the formation of the schistosity and imbricate faulting. Other complex folds were produced during the late deformation and the intrusion of plutonic rocks. (4) Ablah group—these rocks consist mainly of sedimentary rocks, including phyllite, calcareous phyllite, slate, fine gray wacke, brown weathering marble, subordinate quartzite, and a minor

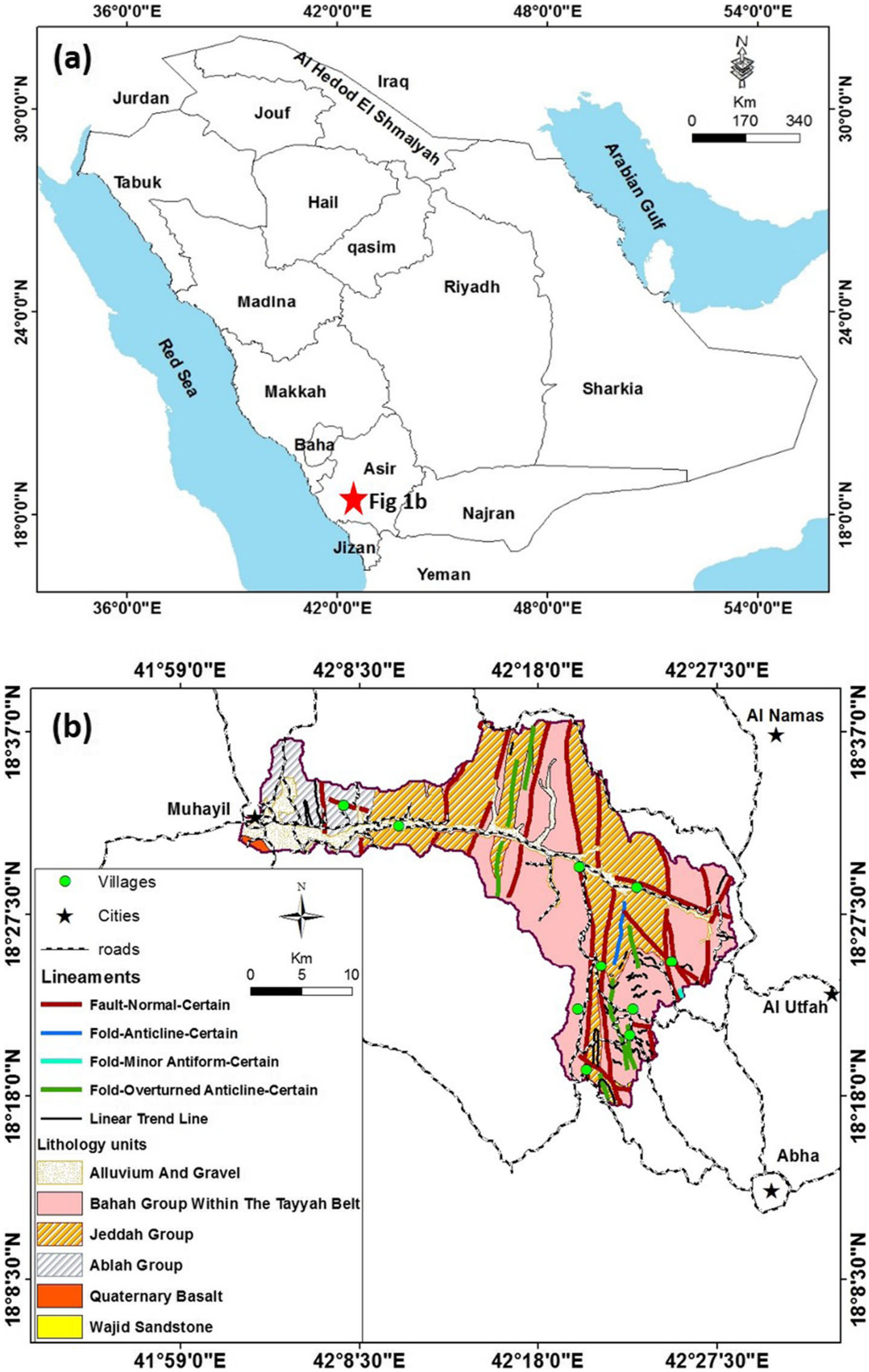

Fig. 1 **a** Location of the study area in Saudi Arabia map. **b** Geology and structural map of the study area (note: urban areas and roads in around the study area overly the geology and structural map)

constituent of metabasalt. The group is basically isoclinally folded and contains a strong cleavage or schistosity. The rocks have been metamorphosed to greenschist facies. 5) Quaternary basalt: It is olivine basalt, which forms two volcanic cones and surrounding flows. According to the geomorphological evidence, it is related to the Quaternary age. (6) Wajid sandstone—remnant of Wajid sandstone rests unconformable on Proterozoic rocks. It consists of light brown to reddish-brown quartz sandstone with some pebbles of quartz veins and quartzite at the base. The study area is crossed by numerous faults and intensively fractured rocks. Different types of structures such as faults, folds, and joints are encountered in the study area, and its surroundings according to geological map (Abha quadrangle GM-75, Greenwood 1985) (Fig. 1b). The geological map was verified by field investigation. The materials along the faults are highly crushed and weathered (Greenwood 1985). In addition, rocks close to the fault zone are highly shorn and jointed.

Data and methodology

The current research demonstrates the application of different data sources to be used in landslide susceptibility analysis. Figure 2 shows the steps of methodologies that were applied in the current study, including different data sources and data types, inventory map, extracted data, model building, and validation of models.

Step I. Data collection (data sources and data types) in which different data sources and types were adapted and used in the current study. First datasets are data related to field surveys that have been done for the study area at different times. Other data sources such as historical reports (collected from the civil defense authority, newspaper records, and interviews with local people) can give some ideas about the frequency of landslides in some areas, especially those that are close to urban areas and along the highways. Second data sources are satellite imageries, enhanced thematic mapper plus (ETM^{+}) with a spatial resolution of 15 m, QuickBird image with a spatial resolution of 0.6 m, and SRTM data (DEM, 90 m). Third datasets are the topographic maps of 1:10,000-scale. Fourth datasets include the meteorological data in which the historical records of the rainfall gauges that are located in and around the study area were used. A fifth dataset is the geological map (Abha quadrangle, GM-75) with a scale of 1:250,000. All the datasets used in the current study are in a digital format with a unified projection (UTM-Zone 38, WGS84 datum).

Step II. Preparing the inventory map. It is well known that landslides are more likely to occur under the same conditions that had been found on earlier landslides. Thus, a landslide inventory map for the study area represents an essential part for landslide susceptibility modeling. Understanding the relationship between the existing landslide distribution and the landslide-conditioning factors is a fundamental requirement for landslide susceptibility mapping (Ercanoglu and Gokceoglu 2004). A landslide inventory map was prepared according to the interpretation of different data types such as historical records, field investigation and surveys, interviews with people, and satellite images analysis (van Westen et al. 2006; Petley 2008). Different authors used geomorphological features to detect landslides from satellite imageries (De La Ville et al. 2002; Youssef et al. 2009). The different datasets were used to prepare the landslide inventory map as shown in (Fig. 2). Historical landslides have some specific morphological features that are easily identifiable with high-resolution imagery, especially in the three-dimensional models, and include breaks in the highly vegetated area and bare soil (Fig. 3a, b). Other features that can help in detecting landslides include the presence of flow materials along gullies, rims, and streams with different erosional features, flow tracks, and depositional fans (Elkadiri et al. 2014). In addition to these, circular and planar failures could be identified easily from the high-resolution satellite imagery according to different morphological features such as bare breaks and head and side scarps (Fig. 3a, b). Field observations were used to verify and collect some

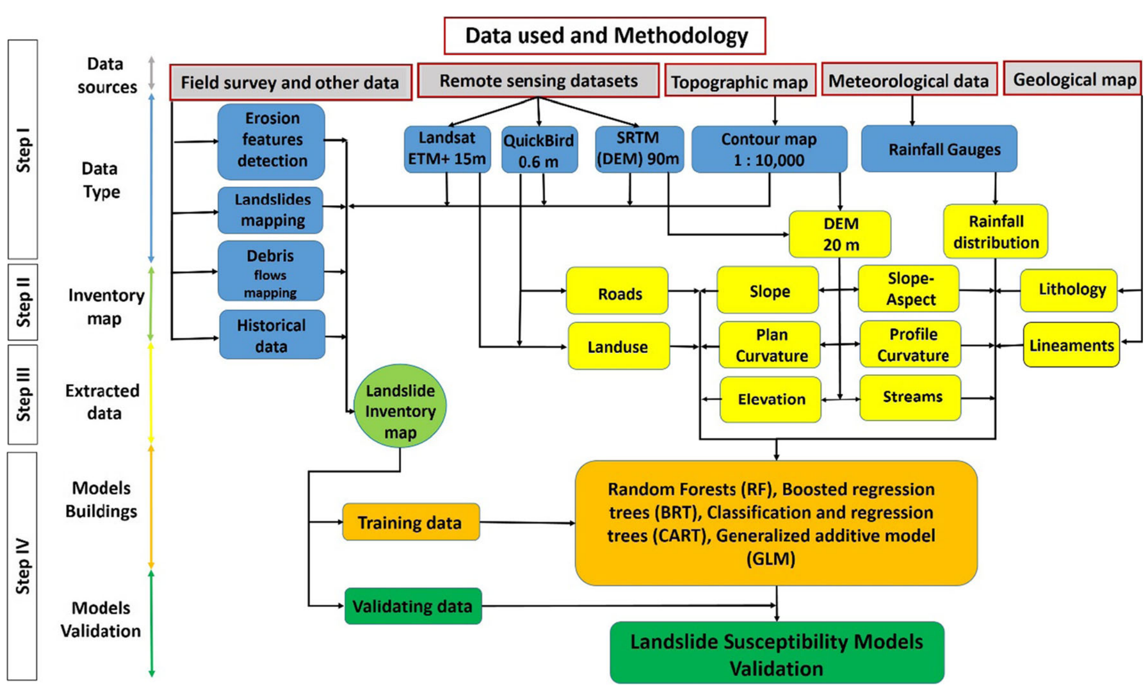

Fig. 2 Schematic diagram of the various datasets, their derived products, and their usage, illustrating the modeling process

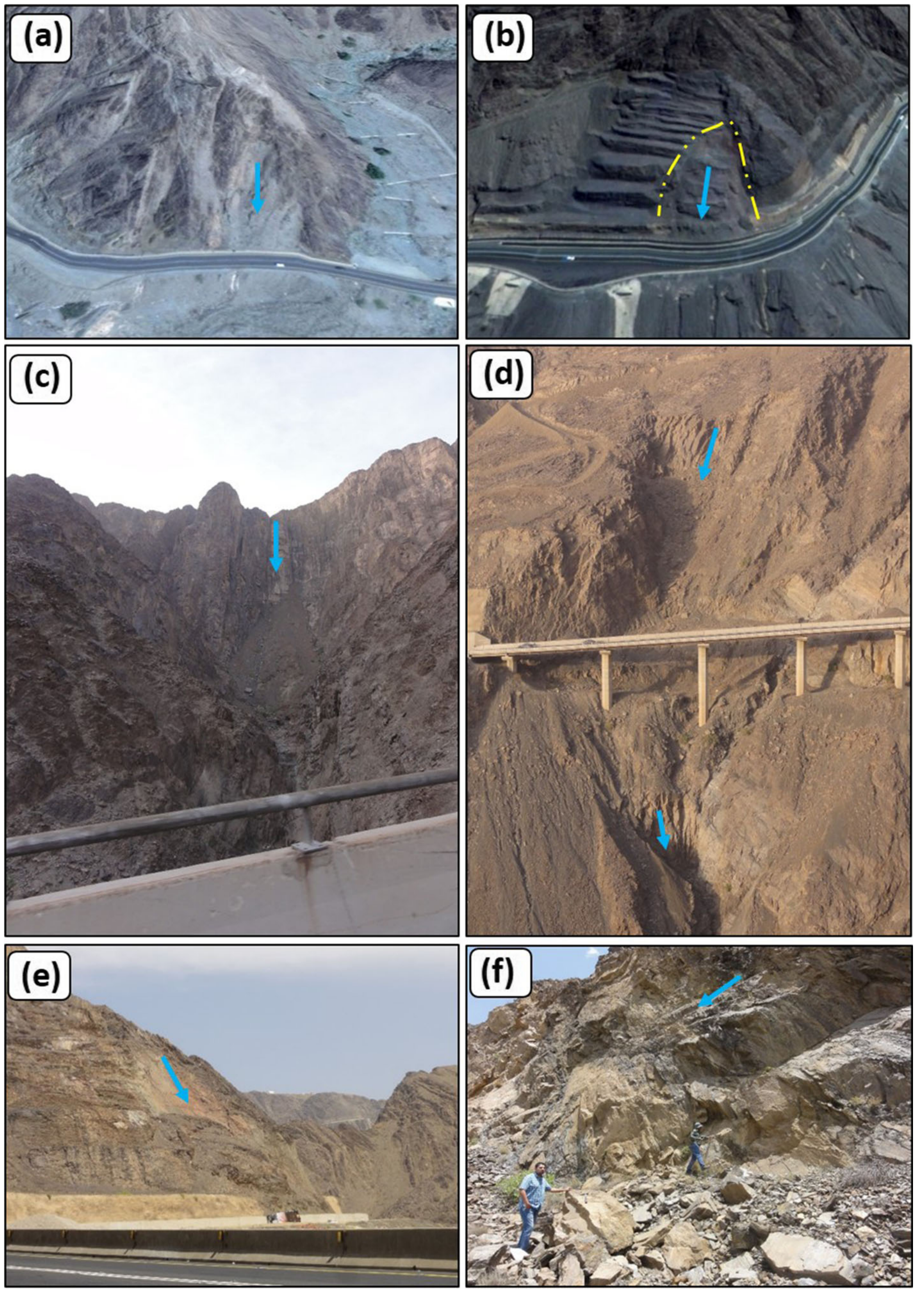

Fig. 3 **a**, **b** Three-dimensional high-resolution images used to detect different types of landslides: **a** planar failure and **b** circular failure. **c**–**f** Field shot of different types of landslides to help in preparing the landslide inventory map: **c** large structural control failure, **d** sliding and failures related to water erosion, **e** circular failure, and **f** large planar failure

fresh/new landslides in the study area (Fig. 3c–f). These data were collected and assembled together to create the landslide inventory map (Fig. 4). Using the previous methods and data sources, a total of 177 landslides were identified and mapped in the study area, among them about 25 landslides were visited in the field for verification purposes. Results indicated that all these locations are old landslides, which are characterized by different volumes ranging from a few cubic meters to about a few thousand cubic meters. The entire field investigated landslide sites were successfully verified, giving suggesting confidence in the applied technique. These landslide locations show mainly translational mass movements (along structures); some are related to rotational slides (circular failures) especially in highly fractured rocks, colluvium materials, and wadis terraces. Many landslides were detected along the faults that dissected the study area. Many structures, such as schistosity, fractures, and joints, with a dip angle of more than 30° toward these valleys facilitate many planar and wedge failures. The landslide locations were

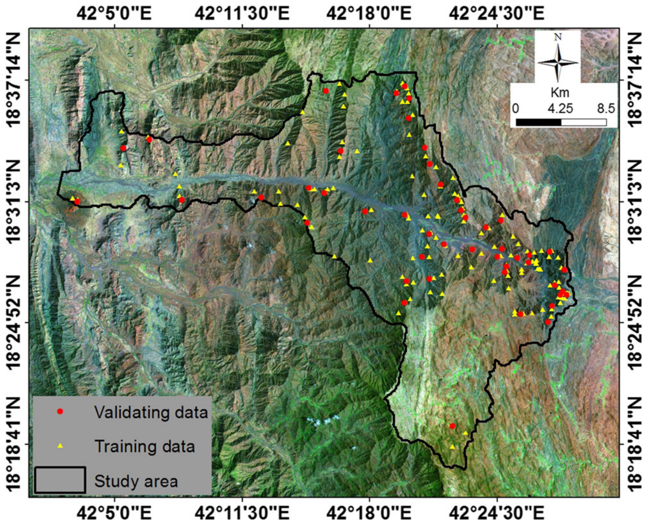

Fig. 4 Landslide location map with Landsat 8 false-color composite image (ETM+, 15 m; bands 7 (*red*), 4 (*green*), and 2 (*blue*)) of the study area

collected and digitized as point features. Out of the selected locations, about 75 % of the sites were used for model training, and the remaining (25 %) of the sites were used for validation purposes (Ohlmacher and Davis 2003; Chacon et al. 2006).

Step III. Landslide-conditioning factors: it is essential to determine the conditioning factors for landslide susceptibility mapping. Different types of databases were generated, compiled, and hosted in a geographical information system (GIS) for data interpretation and analysis. In the current study, 11 factors were used as conditioning factors. These include slope aspect, altitude, distance from faults, lithology, plan curvature, profile curvature, rainfall, distance from streams, distance from roads, slope angle, and land use (Fig. 5). All layers were transformed into a grid spatial database by 20×20-m pixel size and all the data in UTM coordinate system zone 38 with a datum of WGS 84. Five geomorphological layers were extracted from DEM using ArcGIS 10.2 software. These layers are slope aspect, altitude, plan curvature, profile curvature, and slope angle. Slope aspects are shown in classes of flat (−1), North (0°–22.5°; 337.5–360°), North-East (22.5–67.5°), East (67.5–112.5°), South-East (112.5–157.5°), South (157.5–202.5°), South-West (202.5–247.5°), West (247.5–292.5°), and North-West (292.5–337.5°) (Fig. 5a). Altitude value of the study area ranges from 455 to 2911 m, plan curvature from −51.9 to 72.9, profile curvature from −62.5 to 62.9, and slope angle from 0.00° to 81.3° (Fig. 5b–e). Landslide occurrence is likely affected by altitude where altitude is controlled by several geological and geomorphological processes (material types, wind action, rainfalls, and erosions) (Ayalew and Yamagishi 2005; Pourghasemi et al. 2013a, b). Lineaments, streams, and roads were extracted from the geological map, DEM, topographic map, Landsat images, and high-resolution satellite images. Distance maps were produced for lineaments, streams, and roads using the Euclidean Distance tool in ArcGIS 10.2 (Figs. 5f–h). The maps show that distance from lineaments ranges from 0 to 5749 m, distance from streams ranges from 0 to 4327 m, and distance from roads ranges from 0 to 12,798 m. The lithology units were extracted from the geological database where six main units were found including (1) Alluvium and Gravel, (2) Bahah Group within the Tayyah Belt, (3) Jeddah Group, (4) Ablah Group, (5) Quaternary Basalt, and (6) Wajid Sandstone (Fig. 5i). Rainfall data was extracted from analysis of SA113, SA138, A106, A107, A108, A118, A124, and A130 rainfall gauges surrounding the study area. Rainfall value ranges from 60.4 to 227.5 mm (Fig. 5j). Finally, the land use map was prepared from the interpretation of high-resolution and Landsat satellite images. Six land use types were extracted, including agricultural land, barren land, rocks with trees, urban and terraces, urban and agriculture, and soil with intense trees (Fig. 5k). In the current study, the landslide-conditioning factors were nominal, ordinal, and scale. Some factors are ordinal, such as slope angle, plan curvature, profile curvature, distance from lineaments, distance from roads, distance from streams, and rainfall, while elevation was in a ratio scale; however, after classification it transformed to ordinal scale. In addition, the nominal factors are lithology and land use, and some authors used slope aspect (which represents a specific type of factor) as a nominal factor (Pradhan 2010; Youssef 2015).

Step IV. Model building and model validation according to the relation between the landslides location and the different datasets. These models are RF, BRT, CART, GLM. The mentioned models will be discussed in the following paragraphs.

RF It is an ensemble-learning technique (Breiman 2001). It generates many classification trees that are aggregated to compute a classification (Breiman et al. 1984). Hansen and Salamon (1990) indicated that a necessary and sufficient condition for an ensemble of classification trees to be more accurate than any of its individual members is that the members of the ensemble perform better than random and are diverse. Random forests increase diversity among the classification trees by resampling the data with replacement and randomly changing the predictive variable sets over the different tree induction processes. The number of trees (k) and the number of predictive variables used to split the nodes (m) are two user-defined parameters required to grow a random forest. Predictive variables may be numerical or categorical; a translation to design variables is not needed. An unbiased estimate of the generalization error is obtained during the construction of a random forest. The proportion of mis-classifications (%) over all out-of-bag elements is called the out-of-bag (OOB) error. The OOB error is an unbiased estimate of the generalization error. Breiman (2001) proved that random forests produce a limiting value of the generalization error. As the number of trees increases, the generalization error always converges. The k needs to be set sufficiently high to allow for this convergence. The random forest technique estimates

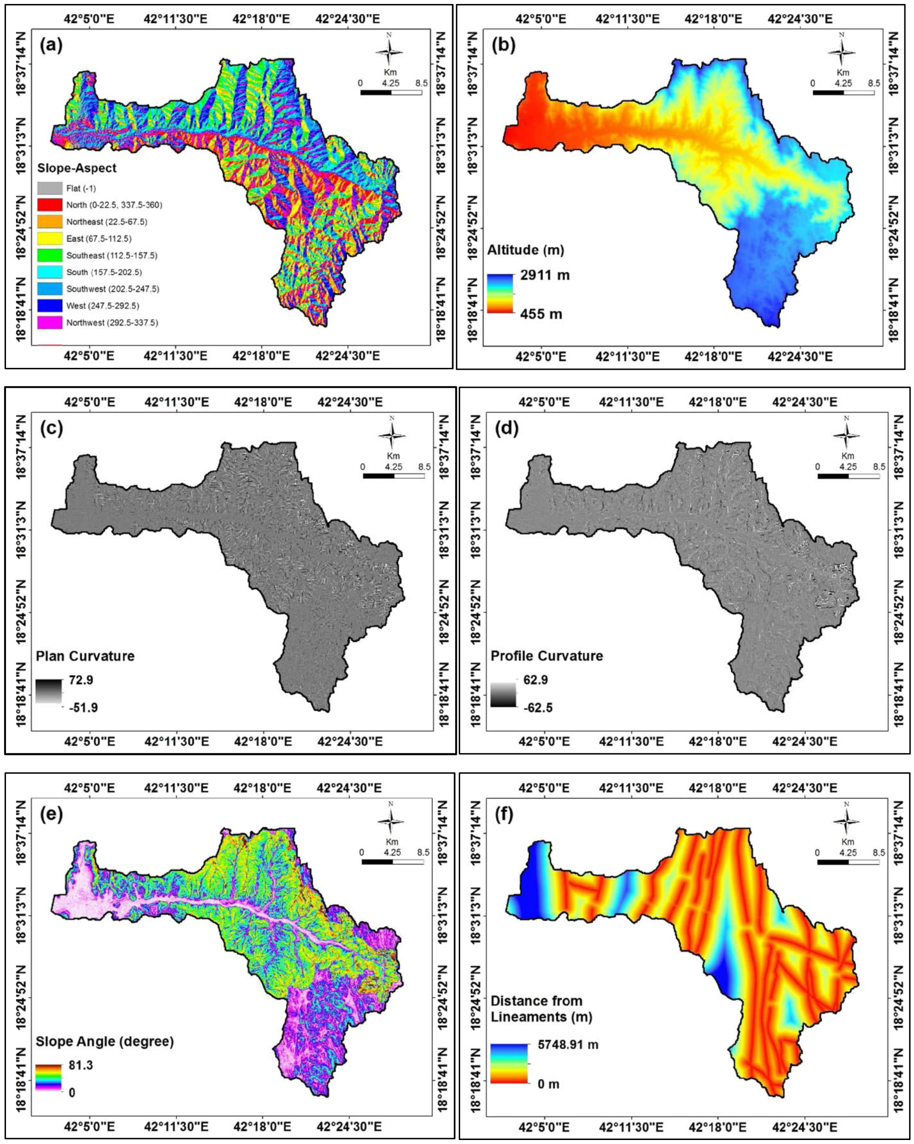

Fig. 5 Landslide-conditioning factor maps used in this study: **a** slope aspect map, **b** altitude map, **c** plan curvature map, **d** profile curvature map, **e** slope angle map, **f** distance from faults map, **g** distance from streams map, **h** distance from roads map, **i** lithology map, **j** rainfall map, and **k** land use map

the importance of a predictive variable by looking at how much the OOB error increases when OOB data for that variable are permuted while all other variables are left unchanged. The increase in OOB error is proportional to the predictive variable importance (Breiman and Cutler 2004). One of the main advantages of RF is the resistance to over training and growing a large number of random forest trees where it does not create a risk of over fitting (e.g., each tree is a completely independent random experiment). The RF algorithm data does not need to be rescaled, transformed, or modified. It has resistance to outliers in predictors and automatically handles the missing values (Breiman and Cutler 2004). In this study, for random forest modeling, the statistical package R version 3.8 was used. Using these guidelines, the number of trees in RF has been fixed to 1000 after a primary analysis and the *m* sampled at each node has been selected to be 3 to analyze the joint contribution of subsets of features while keeping a fast convergence during iterations. No calibration set is needed to regulate the parameters (Micheletti et al. 2014). Two types of error were assessed: mean decrease in accuracy and mean decrease in node impurity (mean decrease Gini) (Calle and Urrea 2010).

BRT It is a combination of statistical and machine learning techniques. It is one of the several techniques that aim to improve the performance of a single model by fitting many models and combining them for prediction (Schapire 2003). The more advanced use of the BRT is to model natural phenomena with nonlinear relationships. This model does not need prior data transformation or elimination of outliers, and can fit complex nonlinear relationships and automatically address interaction effects between predictors (Elith et al. 2008). BRT uses two algorithms namely regression and boosting. Decision trees represent information in a way that is intuitive and easy to visualize, and have several other advantageous properties. Trees are insensitive to outliers, and can modify missing data in predictor variables using

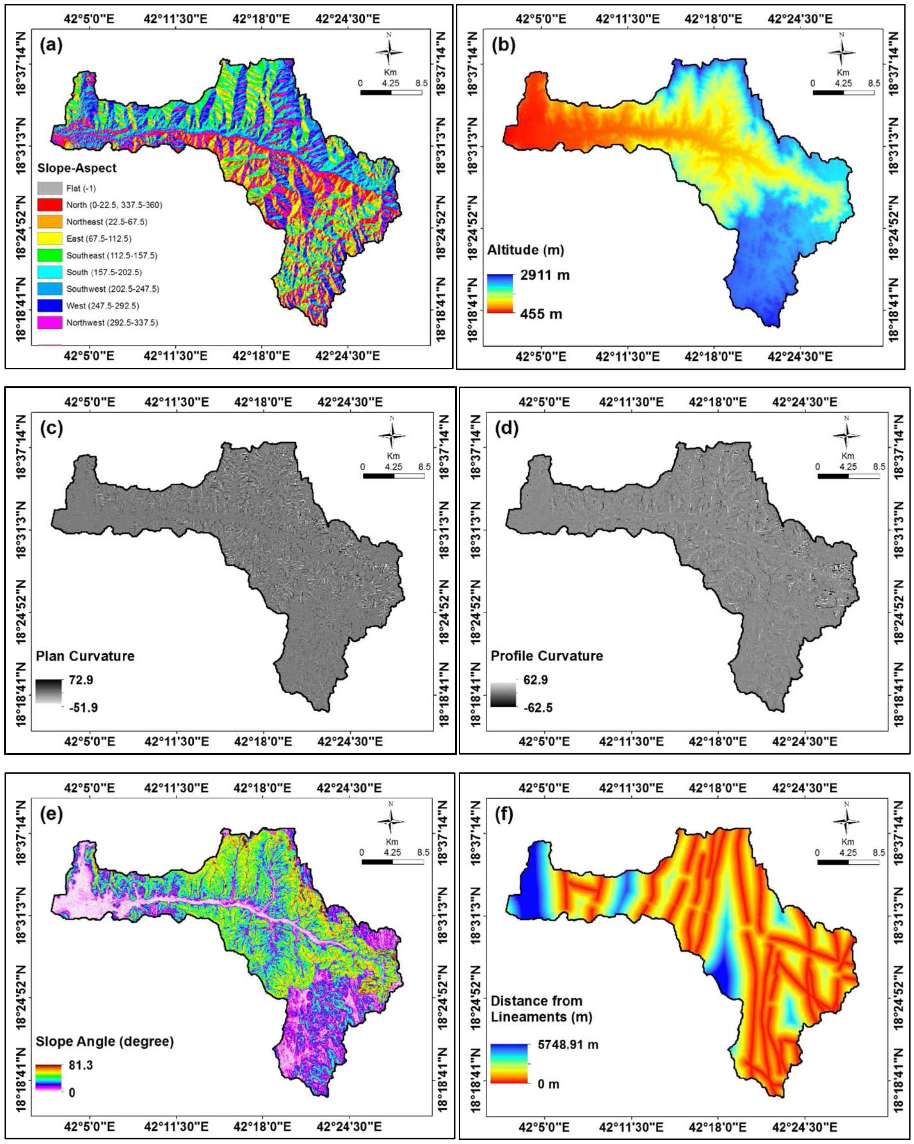

Fig. 5 (continued)

surrogates (Breiman et al. 1984; Elith et al. 2008). Boosting is a method for improving model accuracy, based on the idea that it is easier to find many rough rules of thumb than to find a single, highly accurate prediction rule (Schapire 2003). Fitting multiple trees in BRT overcome the biggest drawback of single tree models (their relatively poor predictive performance). Boosted regression trees were developed in R 3.8 statistical package with the help of the BRT extension for the gbm package (Ridgeway 2006), developed by Elith et al. (2008). Models were fitted using the gbm.step function, and the model was simplified by reducing the number of explanatory variables with the gbm.simplify function.

CART It is a rule-based algorithm that generates a binary tree through "binary recursive partitioning," a process that divides a node into yes/no answers as predictor values. Each division is based on a single variable, and the rule generated at each step minimizes the variability within each resulting subset, splitting them further based on the different relationships. According to literature review, CART was used only few times for landslide susceptibility (Nefeslioglu et al. 2010; Yeon et al. 2010; Felicísimo et al. 2012). CART is a technique that is easy and straight forward to interpret but too simple to describe many real-world situations (Elith et al. 2008). The predicted value of a "terminal" node is the average of the response values in that node (Breiman et al. 1984). CART is a popular technique because it represents information in a way that is intuitive and easy to visualize. Preparation of candidate predictors is simplified, because predictor variables can be of any type (numeric, binary, categorical, etc.), model outcomes are unaffected by monotone transformations and differing scales of measurement among predictors. Regression trees are insensitive to outliers and can accommodate missing data in predictor variables using surrogates (Breiman et al. 1984). The hierarchical structure of a regression tree means that the response to one input variable depends on values of inputs higher in the tree, so interactions between predictors are automatically modeled. Regression trees generally resulted in an over the complex decision tree that needs to be 'pruned' in order to convey only the most important information (i.e., the nodes that explain the largest amount of deviance) (McKenney and Pedlar 2003).

GLM It was obtained based on an extension of the general linear models, namely the GLM (McCullagh and Nelder 1989; Piccolo 1998; Federici et al. 2005, 2007; Giudici 2005; Greco et al. 2007; Falaschi et al. 2009). Generalized linear model extends the usual regression framework to cater for non-normal distributions (Payne 2012). Equation 1 summarized the mathematical (statistical) function (called LOGIT) for the GLM model (Bernknopf et al. 1988; Piccolo 1998):

$$Y = \Pr(Y = 1) = \frac{e^{C_0 + C_1X_1 ..... + C_nX_n}}{1 + e^{C_0 + C_1X_1 ..... + C_nX_n}} \quad (1)$$

where, $Y=\pi$, i.e., the probability of each condition factor of being unstable or stable given a certain combination of the instability factors (covariates) was modeled. LOGIT is used as a link function for modeling fractional response to handle data at extreme values of zero and 1. So, the GLM consisted of three elements: (1) a probability distribution for the response variable *Y*, (2) a linear predictor, and (3) a link function that provides the relationship between the linear predictor and the mean of the distribution function (Nikita 2014).

Results

Application of random forest

Aggregate OOB predictions are presented in Fig. 6 and Table 1 (confusion matrix). The OOB suggests that when the resulting model is applied to new observations, the answer will be in error

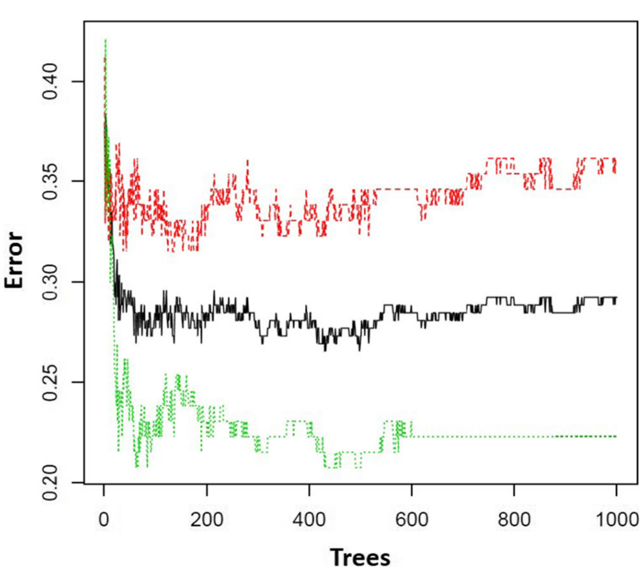

Fig. 6 The error rate of the overall RF model (*OOB* out of bag (*black line*); 0, absent landslide (*red line*); and 1, present landslide (*green line*))

25 % of the time. It is indicated that 75 % of the results are accurate, which is a reasonably good model. Overall measure of accuracy is then followed by a confusion matrix that records the disagreement between that final model's predictions and the actual outcomes of the training observations. The actual observations are the rows of this table, while the columns correspond to what the model predicts for an observation and the cell counts the number of observations in each variable (Williams 2011). The model predicts 1, and the observation was 0 for 30 observations. The finding showed that the model and the training dataset agree that for 72 of the observations, landslide absence is correctly predicted (78.3 %), and that of the actual 67 landslide locations, 48 (67.2 %) were correctly predicted. Results from variable selection random forest are shown in Fig. 7. This shows the 11 variables ordered by two specific importance measures (mean decrease accuracy and mean decrease Gini). Based on Fig. 7 and Table 2, the higher values indicate that the variable is relatively more important (Williams 2011). Table 2 indicated that slope angle is the most important variable (38.95 %), followed by land use (22.62 %), altitude (DEM; 15.77 %), and rainfall (8.34 %). In contrast, slope angle (18.76 %), altitude (9.80 %), distance from roads (7.82 %), and rainfall (7.79 %) had higher importance according to the Gini measure.

Application of boosted regression tree

Boosted regression trees were developed in the R statistical package using BRT extension for the gbm package (Ridgeway 2006), developed by Elith et al. (2008). Models were fitted using the gbm.step function, and the model were simplified by reducing the number of explanatory variables with the gbm.simplify function. The boosted regression trees method also depends on how many regression trees are produced, same as that in the random forest method (Fig. 7). The main difference of the BRT method from the RF method is that the boosted regression trees do not rely on bootstrap samples or randomized variable selection. The boosted regression trees method is best explained by examining the fitting algorithm first. Figure 8 shows how to calculate the weight value for each parameter. Table 3 shows the weighted value for each landslide-conditioning factor. According to Table 3, the highest value found for slope angle was about 34.1 %, followed by distance from road of 10.6 %, and the lowest value was for the lithology at 2.3 %.

Application of classification and regression trees

In the current work, the regression trees were built with the help of the R software and r part package. This generally results in a complex decision tree that needs to be "pruned" in order to convey only the most important information (i.e., the nodes that explain the largest amount of deviance) (McKenney and Pedlar 2003). Figure 9 shows the pruned regression tree that takes the full tree as the first argument and the chosen complexity parameter as the second. Based on Fig. 9, 1 is landslide occurrence and 0 is no landslide occurrence. Based on the figure, it was revealed that the most important factors were slope degree, slope aspect, distance from roads, and altitude, respectively. The weighted value for each landslide-conditioning factor is shown in Table 4 in which the highest value was for slope angle of about 39 % and the lowest value was for profile curvature and slope aspect at 1 % for each.

Table 1 Confusion matrix from RF model (0=no landslide, 1=landslide)

Number		Predicted		–
		0	1	Overall class error
Actual	0	82	48	0.37
	1	30	100	0.23

Application of generalized linear model

In this study, GLM was constructed using the R statistical package (version 2.8). A simple Gaussian family was specified as a link function for the normally distributed response data. Conditioning factors entered the models individually using a smoothing spline with only 2 degrees of freedom in a polynomial fit of degree 2 to avoid over fitting (Aertsen et al. 2009). Table 5 shows the results of the applying this model for each conditioning factor. According to Table 5, the observed relationships between landslide locations and each related factor using the GLM approach are presented. When a perfect linear relationship exists between the variables, the estimates for a regression model cannot be uniquely assessed (Pourtaghi et al. 2014). The term collinearity shows that two given predictors are near perfect linear combinations of both. The inclusion of more than two variables is called multi-collinearity, and model fitting with GLM is sensitive to collinearity across the independent variables (Hosmer and Lemeshow 2000; Pourtaghi et al. 2014). Furthermore, in this study, "Tolerance" (TOL) and the "variance inflation factor" (VIF) were used as the two important indices for multi-collinearity diagnosis (Zhu and Huang 2006; Pourtaghi et al. 2014). O'Brien (2007) expressed that tolerance of less than 0.20 or 0.10 and/or a VIF of 5 or 10 and above indicated a multi-collinearity problem. According to Table 5, the smallest TOL and highest VIF were 0.527 and 1.898, respectively. So, in this study, there is no extreme multi-collinearity between independent factors. Also, Table 6 indicated that slope aspect, altitude, distance from faults, plan curvature, rainfall, and distance from roads have negative effect in landslide susceptibility as they all have been negative β coefficients. In contrast, lithology, profile curvature, distance from streams, slope angle, and land use have positive β coefficients. Land use, lithology, and slope angle are the most important factors to landslide susceptibility, respectively (Table 6).

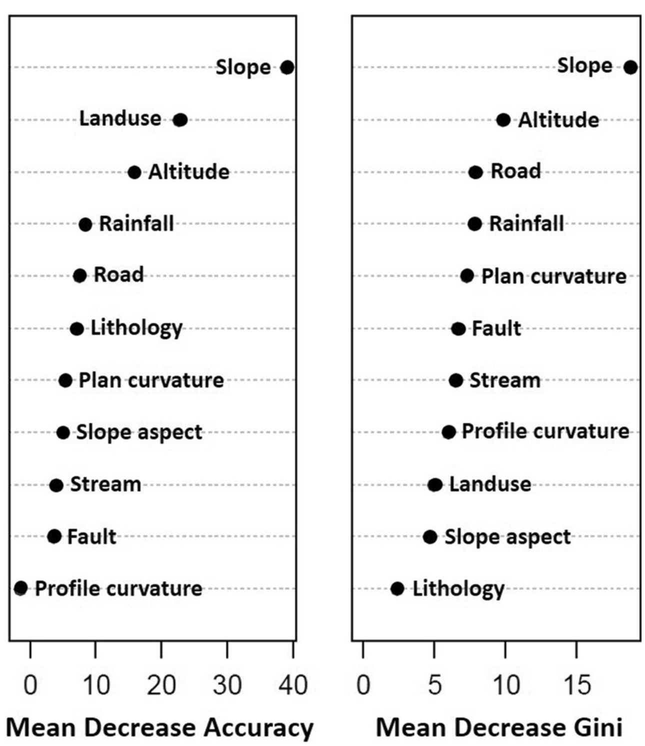

Fig. 7 Mean decrease accuracy and mean decrease Gini (sorted decreasingly from *top* to *bottom*) of landslide-conditioning factors as assigned by the random forest

Landslide susceptibility maps

The landslide susceptibility maps produced by four data mining models (RF, BRT, CART, and GLM) are represented in Fig. 10a–d. The obtained pixel values from these models were then classified based on the natural break classification scheme (Pourghasemi et al. 2012a, 2013a, b; Mohammady et al. 2012). These maps satisfied two spatial effective rules: (1) the existing landslide pixels should belong to the high-susceptibility class and (2) the high-susceptibility class should cover only small areas (Can et al. 2005; Bui et al. 2012). Finally, results revealed that very high landslide susceptibility map (LSM) class derived using the RF model covers 15.36 % of the total area; 28.05, 28.24, and 28.35 % of the total area are related to low, moderate, and high LSM zones, respectively (Fig. 10a); 43.72 % of the total area covered on the LSM map obtained from the BRT method is designated to be of low landslide susceptibility class; in contrast, 21.98, 19.44, and 14.85 % of the total area are related to moderate, high, and very high LSM zones, respectively (Fig. 10b). Also, 55.18, 18.92, 8.72, and 17.18 % of the total area, using the CART model, are covered with low, moderate, high, and very high LSM zones, respectively (Fig. 10c). According to the GLM model, 37.49 and 24.96 % of the study areas were classified as "low" and "moderate"" susceptibility, whereas 20.93 and 16.63 % of the areas were classified as "high" and "very high" susceptibility, respectively (Fig. 10d).

Validation of landslide susceptibility maps

Remondo et al. (2003) indicated that landslide validation must be of guidance in data collection and field practice for landslide mapping. Validation was used to carry out sensitivity analysis for individual variables and combinations of variables in which different map-making methods were tested (Chung and Fabbri 2003). Verification of the landslide susceptibility map produced from a model can be conducted using the receiver operating characteristics (ROC) curve (Akgun et al. 2012; Ozdemir and Altural 2013). The ROC curve is a useful method to determine the quality of deterministic and probabilistic detection and

Table 2 Relative influence of effective conditioning factors in RF model (0=no landslide, 1=landslide)

Factor	1	0	Mean decrease Accuracy (%)	Mean decrease Gini (%)
Slope aspect	−1.35	7.80	5.02	4.64
Altitude	13.47	8.43	15.77	9.80
Distance from faults	2.75	2.17	3.54	6.62
Lithology	−2.26	10..23	7.07	2.36
Plan curvature	9.19	−1.53	5.22	7.22
Profile curvature	2.79	−4.31	−1.54	5.94
Rainfall	3.39	7.91	8.34	7.79
Distance from streams	0.20	5.15	3.91	6.44
Distance from roads	8.65	2.02	7.50	7.82
Slope angle	33.65	24.06	38.95	18.76
Land use	16.10	20.83	22.62	4.92

forecast systems (Swets 1988). In the ROC curve, the sensitivity of the model (the percentage of existing landslide pixels correctly predicted by the model) is plotted against 1-specificity (the percentage of predicted landslide pixels over the total study area). The ability of the probabilistic model to predict the occurrence or non-occurrence of landslides reliability could be determined using the area under the ROC curve (AUC). A fit model has an AUC values above 0.5, and the quality of the model is increased by increasing the AUC values. However, values below 0.5 represent a random fit. Generally, to validate the model, success rate and prediction rate curves are used. These two techniques depend on a comparison between the existing landslide locations with the landslide susceptibility maps. The success rate method used the training landslide pixels that were used in establishing the landslide models. This method can help in determining how well the resulting landslide susceptibility maps have classified the areas of existing landslides. Another technique of validation is named the prediction rate curve which explains how well the model predicts the landslide. The prediction method is widely used by many authors (Mohammady et al. 2012; Akgun et al. 2012; Ozdemir and Altural 2013; Jaafari et al. 2014, Youssef et al. 2014a, b). In the current study, both success and prediction rate curves have been prepared to understand the effect of each model and their

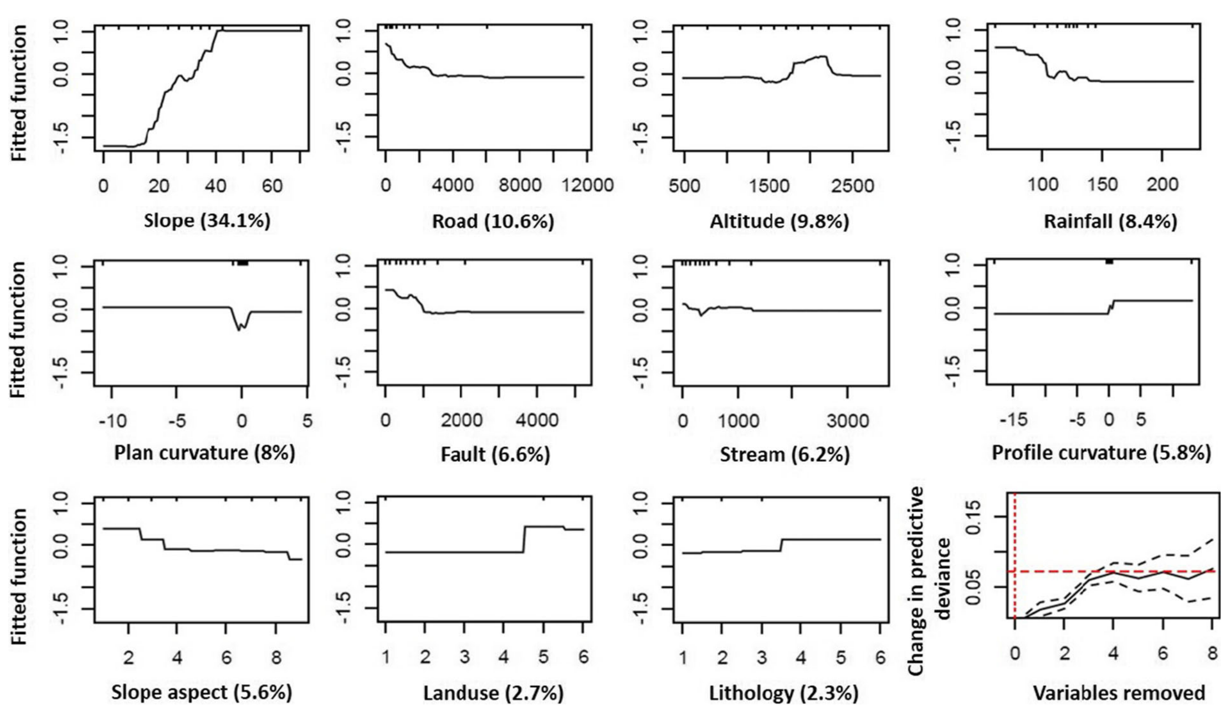

Fig. 8 Partial dependence plots of the predictor variables in the BRT model for predicting the site index of Wadi Tayyah Basin. The relative contribution of each predictor was reported *between brackets*. Rug plots at the *inside top of graph* showed distribution of sample sites along that variable, in deciles

Table 3 the weighted value for each landslide-conditioning factor in BRT model

Number	Factor	Weight
1	Slope aspect	5.56
2	Altitude	9.80
3	Distance from faults	6.56
4	Lithology units	2.29
5	Plan curvature	8.00
6	Profile curvature	5.81
7	Rainfall	8.44
8	Distance from streams	6.16
9	Distance from roads	10.56
10	Slope angle	34.15
11	Land use	2.68

validation as shown in Figs. 11a–d and 12a–d. In the success rate curves, the AUC values for the RF, BRT, CART, and GLM models are 0.783 (78.3 %), 0.958 (95.8 %), 0.816 (81.6 %), and 0.821 (82.1 %), respectively (Fig. 11a–d). In addition, the prediction

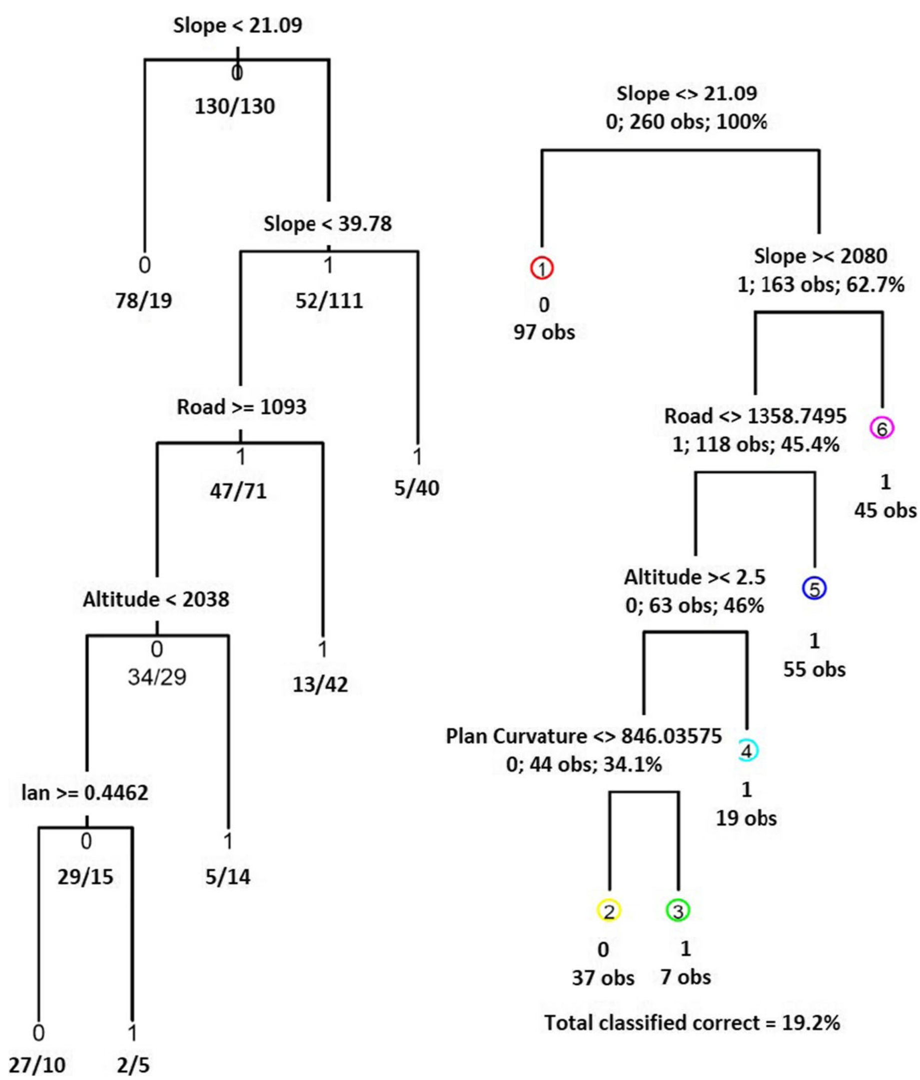

Fig. 9 Optimally pruned regression tree for the study area

Table 4 The importance of factors according to CART model

Number	Factor	Importance (%)
1	Slope	39
2	Altitude	17
3	Rain	9
4	Land use	8
5	Lithology	8
6	Distance from roads	7
7	Distance from faults	5
8	Plan curvature	3
9	Distance from streams	2
10	Profile curvature	1
11	Slope aspect	1

rate curve showed that the AUC values for the RF, BRT, CART, and GLM models are 0.812 (81.2 %), 0.856 (85.6 %), 0.862 (86.2 %), and 0.769 (76.9 %), respectively (Fig. 12a–d). It can be concluded that all these models give the success and prediction rate curve values above 0.7, showing that models for landslide susceptibility mapping in the study area are reasonable. These represent reasonable models for landslide susceptibility mapping in the study area. In addition, the results show that BRT gives the highest success rate followed by GLM, then CART, and finally RF models. However, for the prediction rate CART gave the highest value, followed by BRT, RF, and then GLM models.

Discussion

Stehman and Czaplewski (1998) indicated that the RF classifier can give the highest classification accuracy and the BRT had nearly similar results to RF, but the CART recorded the lowest overall accuracy and kappa coefficient. Some previous studies showed that RF and BRT classifiers could produce significantly higher accuracies compared with the CART method (Gislason et al. 2006; Cutler et al. 2007; Baatuuwie and Leeuwen 2011). They also indicated that the RF classifier could produce the highest accuracy compared with the CART classifier. In addition, other study showed that the accuracies of maps generated by the RF and BRT algorithms were not significantly different. Felicísimo et al. (2012) indicated that the CART is among the models that give highest prediction capability. They found that the CART gives AUC value of 0.77. The current research was in agreement with their results, since the success rate for CART was 0.816 and for the prediction rate the CART was the highest with a value of 0.862. Brenning (2005) indicated that the GLM model is an adequate method for the purpose of landslide susceptibility modeling, since it is able to compete with modern machine learning algorithms. This has been concluded when a

Table 5 Showing the GLM model parameters for the study area

Parameter	Collinearity statistics	
	Tolerance	VIF
Slope aspect	0.92	1.09
Altitude	0.53	1.90
Distance from faults	0.70	1.44
Lithology	0.70	1.43
Plan curvature	0.86	1.17
Profile curvature	0.84	1.20
Rainfall	0.56	1.77
Distance from streams	0.69	1.45
Distance from roads	0.57	1.74
Slope angel	0.77	1.30
Land use	0.78	1.28

Table 6 Coefficients of the conditioning factors in GLM model

Factor	Estimate (β)	Standard Error	z value	Pr(>|z|)
Intercept	−1.433e+00	1.885e+00	−0.760	0.4472
Slope aspect	−1.570e−01	6.624e−02	−2.371	0.0178*
Altitude	−1.865e−05	3.741e−04	−0.050	0.9602
Distance from faults	−4.013e−04	2.426e−04	−1.655	0.0980
Lithology	4.170e−01	1.608e−01	2.594	0.0095**
Plan curvature	−1.980e−01	2.188e−01	−0.905	0.3653
Profile curvature	4.548e−02	1.627e−01	0.280	0.7798
Rainfall	−2.556e−02	8.348e−03	−3.061	0.0022**
Distance from streams	3.313e−05	4.001e−04	0.083	0.9340
Distance from roads	−3.511e−05	7.191e−05	−0.488	0.6253
Slope angle	9.020e−02	1.414e−02	6.379	1.78e−10***
Land use	5.154e−01	3.615e−01	1.426	0.1539

Significant codes: 0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

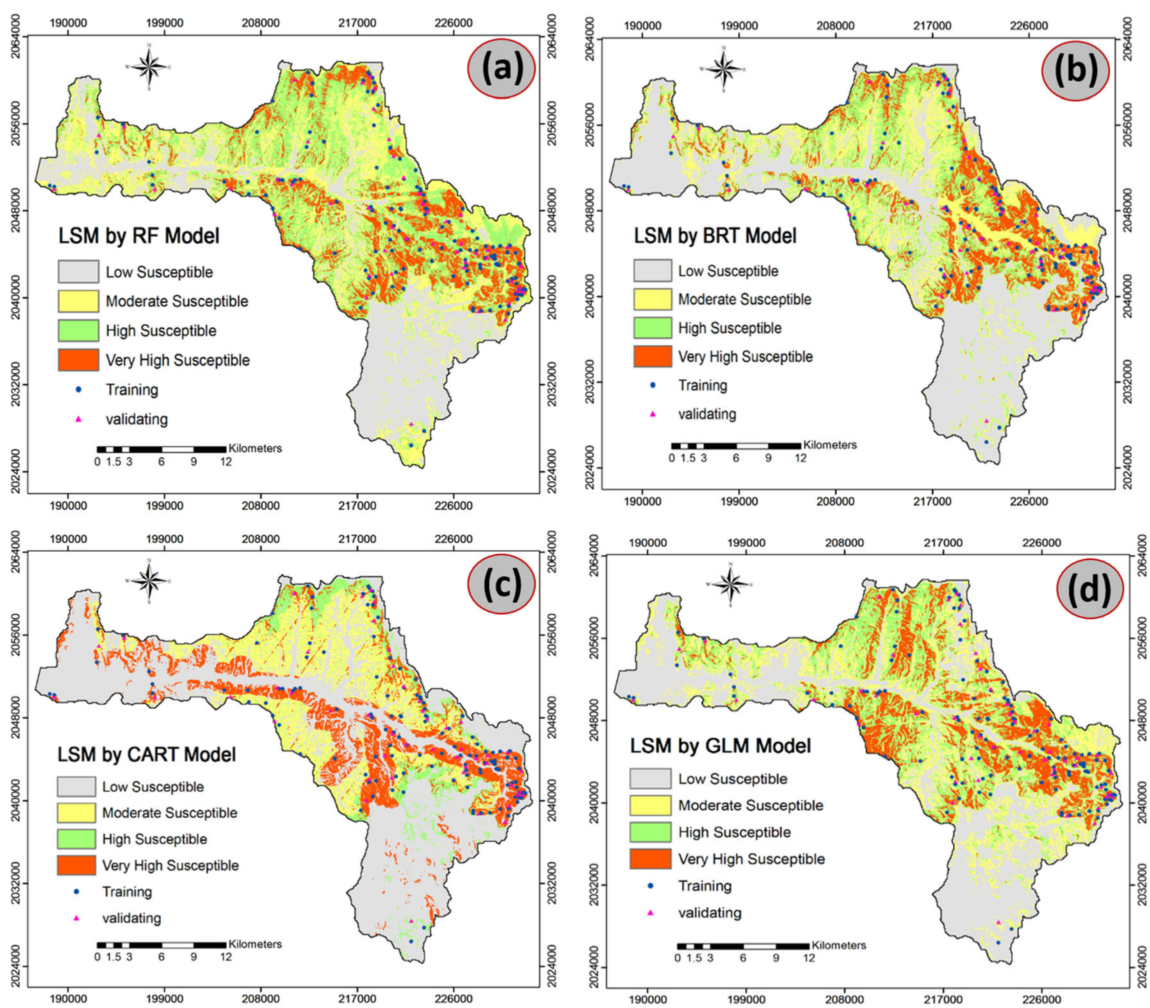

Fig. 10 Generated landslide susceptibility maps using **a** RF, **b** BRT, **c** CART, and **d** GLM

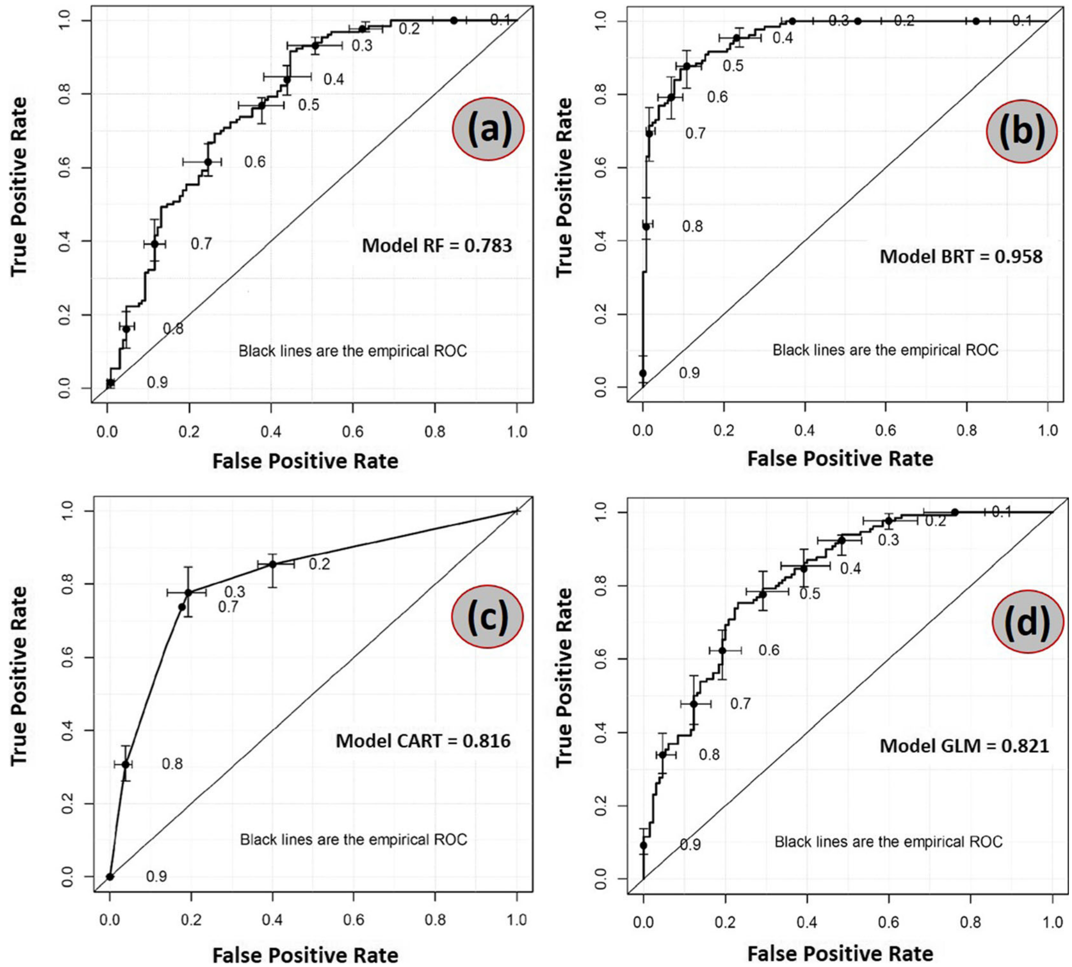

Fig. 11 Success rate curves for the susceptibility maps produced in this study for different models: **a** RF, **b** BRT, **c** CART, and **d** GLM

comparison between support vector machines, GLM, and bootstrap aggregated classification trees on the same area. Marmion et al. (2009) compared different ensemble techniques for improving the prediction of geomorphological maps using a single performance criterion (AUC). They showed that GAM, followed by BRT and GLM performed best models for susceptibility mapping. The results of the current work indicated that the GLM model represents the best performing model; it gives a success rate value of 0.821 and prediction rate value of 0.769. Vorpahl et al. (2012) compared different statistical techniques with model landslide susceptibility in Southern Ecuador. They indicated that RF and BRT perform best models in a 10-fold internal cross-validation. However, after a 40-fold external validation, they indicated that GAM and GLM with a stepwise backwards variable selection performed equally well. All models show a sufficient (AUC >0.7) up to excellent (AUC >0.9) performance on their training data. Finally, they concluded that rather simple models, such as the above, are similarly successful than complex machine learning techniques. Application of RF, BRT, CART, and GLM models in the current research showed that the AUC of success rate ranges from 0.783 to 0.958 and for the prediction rate ranges from 0.769 to 0.862 and were in agreement with Vorpahl et al. (2012) results. Finally; in a comparative study, capability performance of the four non-parametric tree-based algorithms was investigated for landslide susceptibility mapping using different landslide-related factors and the landslides location, as a case study in the Wadi Tayyah Basin, Asir Region, Saudi Arabia. The classifications were performed using four most commonly used non-parametric methods, i.e., RF, BRT, CART, and GLM algorithms due to their advantages against parametric methods. Comparison of the different model results was accomplished using two common methods' success and prediction rates. These comparisons indicated that for the success rate, the BRT give the highest success rate of 95.8 %, followed by GLM with a rate of 82.1 %, then CART with a success rate of 81.6 %, and finally RF with a success rate of 78.3 %. While for the prediction rate the results indicated that CART model gives the highest prediction rate of 86.2 %, followed by BRT with a rate of 85.6 %, then RF with a prediction rate of 81.2 %, and finally GLM with a prediction rate of 76.9 %.

Conclusions

In recent years, landslides have been considered to be the most critical natural hazards (serious threat to life and property) worldwide as well as in Saudi Arabia, and short- and long-

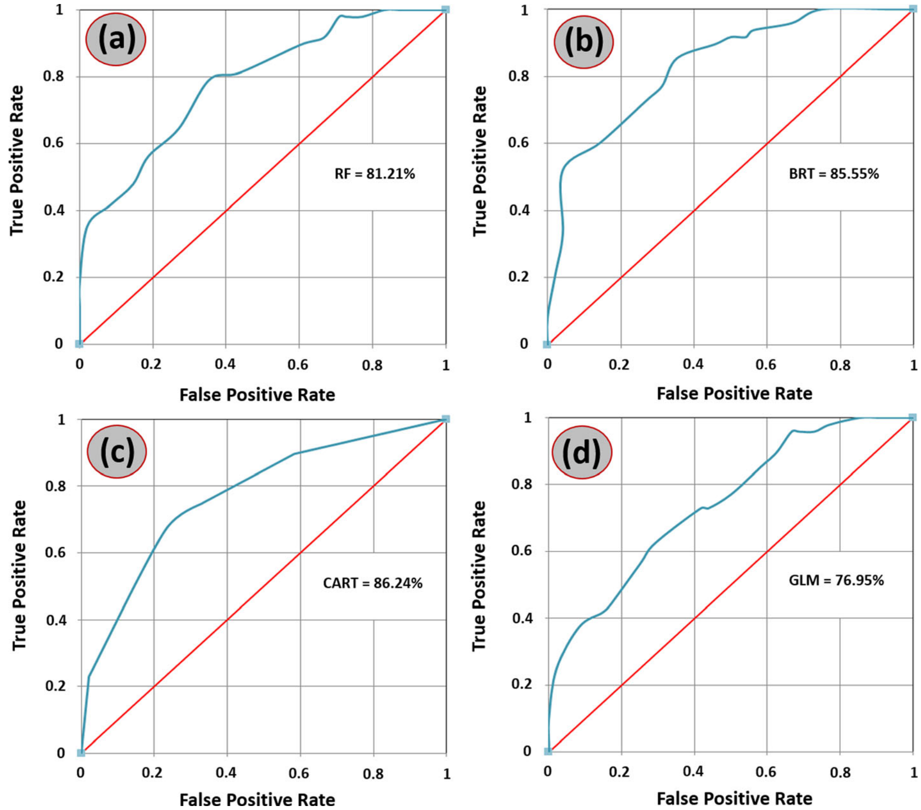

Fig. 12 Prediction rate curves for the susceptibility maps produced in this study for different models: **a** RF, **b** BRT, **c** CART, and **d** GLM

term solutions are required. Recently, landslide susceptibility map represents an essential method to delineate the landslide-prone areas. This can be achieved with the help of advanced statistical approaches that integrated in GIS environment. The main objective of this research was to use four data mining models named RF, BRT, CART, and GLM models, which are novel approaches to perform the landslide susceptibility mapping in the Wadi Tayyah Basin, Asir region, Saudi Arabia. Eleven landslide-conditioning factors (slope aspect, altitude, distance from faults, lithology, plan curvature, profile curvature, rainfall, distance from streams, distance from roads, slope angle, and land use) were prepared and used with the help of an inventory landslide data (training and validating data) to build the LSMs. In order to prove the prediction ability of the proposed models, both success rate and prediction rate curve of ROC were used to test the stability and prediction performance of the four landslide susceptibility maps. The AUC was calculated based on the test dataset (training data), which was randomly collected and the validating datasets. The AUC results showed that the success rates are 0.783 (78.3 %), 0.958 (95.8 %), 0.816 (81.6 %), and 0.821 (82.1 %), and the prediction rates are 0.812 (81.2 %), 0.856 (85.6 %), 0.862 (86.2 %), and 0.769 (76.9 %), respectively, for RF, BRT, CART, and GLM, respectively. Results and findings from this research illustrated that the mentioned models can adequately represent quantitative relationships between landslide occurrences and multiple spatial data factors (landslide-conditioning factors). Finally, these landslide susceptibility maps could be used as the preliminary basis by decision makers, planners, and engineers to avoid and/or minimize the damage and losses caused by existing and future landslides.

Acknowledgments

The authors would like to thank the editorial comments and anonymous reviewers for their helpful comments on the previous version of the manuscript.

References

Abella EAC, Van Westen CJ (2007) Generation of a landslide risk index map for Cuba using spatial multi-criteria evaluation. Landslides 4(4):311–325. doi:10.1007/s10346-007-0087-y

Aertsen W, Kint V, Van Orshoven J, Ozkan K, Muys B (2009) Performance of modelling techniques for the prediction of forest site index: a case study for pine and cedar in the Taurus mountains, Turkey. XIII World Forestry Congress, Buenos Aires, pp 1–12

Akgun A, Sezer EA, Nefeslioglu HA, Gokceoglu C, Pradhan B (2012) An easy-touse MATLAB program (MamLand) for the assessment of landslide susceptibility using a Mamdani fuzzy algorithm. Comput Geosci 38(1):23–34. doi:10.1016/j.cageo.2011.04.012

Ayalew L, Yamagishi H (2005) The Application of GIS-based logistic regression for landslide susceptibility mapping in the Kakuda–Yahiko Mountains, central Japan. Geomorphology 65:15–31. doi:10.1016/j.geomorph.2004.06.010

Ayalew L, Yamagishi H, Ugawa N (2004) Landslide susceptibility mapping using GIS-based weighted linear combination, the case in Tsugawa area of Agano River, Niigata Prefecture, Japan. Landslides 1(1):73–81. doi:10.1007/s10346-003-0006-9

Baatuuwie NB, Leeuwen ILV (2011) Evaluations of three classifiers in mapping forest stand types using medium resolution imagery: a case study in the Offinso Forest District, Ghana. African J Environ Sci Technol 5(1):25–36

Bednarik M, Yilmaz I, Marschalko M (2012) Landslide hazard and risk assessment: a case study from the Hlohovec-Sered landslide—area in southwest Slovakia. Nat Hazards. doi:10.1007/s11069-012-0257-7

Bernknopf RL, Brookshire DS, Shapiro CD (1988) A probabilistic approach to landslide hazard mapping in Cincinnati, Ohio, with applications for economic evaluation. Assoiate Geol Eng Bull 24:39–56

Breiman L (2001) Random forests. Mach Learn 45:5–32. doi:10.1023/A:1010933404324

Breiman L, Cutler A (2004) http://www.stat.berkeley.edu/users/Breiman/RandomForests/ccpapers.html

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Chapman & Hall, New York

Brenning A (2005) Spatial prediction models for landslide hazards: review, comparison and evaluation. Nat Hazards Earth Syst Sci 5:853–862. doi:10.5194/nhess-5-853-2005

Bui DT, Pradhan B, Lofman O, Revhaug I, Dick OB (2012) Landslide susceptibility assessment in the Hoa Binh Province of Vietnam: a comparison of the Levenberg-Marquardt and Bayesian regularized neural networks. Geomorphology. doi:10.1016/j.geomorph.2012.04.023

Calle ML, Urrea V (2010) Letter to the Editor: stability of random forest importance measures. Brief Bioinform 12(1):86–89. doi:10.1093/bib/bbq011

Can T, Nefeslioglu H, Gokceoglu C, Sonmez H, Duman TY (2005) Susceptibility assessments of shallow earthflows triggered by heavy rainfall at three catchments by logistic regression analysis. Geomorphology 72(1–4):250–271. doi:10.1016/j.geomorph.2005.05.011

Catani F, Lagomarsino D, Segoni S, Tofani V (2013) Landslide susceptibility estimation by random forests technique: sensitivity and scaling issues. Nat Hazards Earth Syst Sci 13:2815–2831. doi:10.5194/nhess-13-2815-2013

Chacon J, Irigaray C, Fernandez T, El Hamdouni R (2006) Engineering geology maps: landslides and geographical information systems. Bull Eng Geol Environ 65:341–411. doi:10.1007/s10064-006-0064-z

Chen XW, Liu M (2006) Prediction of protein–protein interactions using random decision forest framework. Bioinformatics 21(24):4394–4400. doi:10.1093/bioinformatics/bti721

Chung CJF, Fabbri AG (2003) Validation of spatial prediction models for landslide hazard mapping. Nat Hazards 30(3):451–472. doi:10.1023/B:NHAZ.0000007172.62651.2b

Committee on the Review of the National Landslide Hazards Mitigation Strategy (2004) Partnerships for reducing landslide risk. Assessment of the National landslide hazards mitigation strategy. Board on Earth Sciences and Resources, Division on earth and life studies, The National Academic Press, Washington, p 143

Cutler DR, Edwards TC, Karen J, Beard H, Cutler A, Hess KT, Gibson J, Lawler JJ (2007) Random forests for classification in ecology. Ecology 88(11):2783–2792

Dahal RK, Hasegawa S, Nonomura A, Yamanaka M, Dhakal S, Paudyal P (2008a) Predictive modeling of rainfall-induced landslide hazard in the Lesser Himalaya of Nepal based on weights of evidence. Geomorphology 102(3–4):496–510. doi:10.1016/j.geomorph.2008.05.041

Dahal RK, Hasegawa S, Nonomura A, Yamanaka M, Masuda T, Nishino K (2008b) GIS-based weights-of-evidence modeling of rainfall-induced landslides in small catchments for landslide susceptibility mapping. Environ Geol 54(2):311–324. doi:10.1007/s00254-007-0818-3

Dai FC, Lee CF (2002) Landslide characteristics and slope instability modeling using GIS, Lantau Island, Hong Kong. Geomorphology 42:213–228. doi:10.1016/S0169-555X(01)00087-3

Dai FC, Lee CF, Xu ZW (2001) Assessment of landslide susceptibility on the natural terrain of Lantau Island, Hong Kong. Environ Geol 40(3):381–391. doi:10.1007/s002540000163

De La Ville N, Diaz AC, Ramirez D (2002) Remote sensing and GIS technologies as tools to support sustainable management of areas devastated by landslides. Environ Dev Sustain 4(2):221–229

Devkota KC, Regmi AD, Pourghasemi HR, Yoshida K, Pradhan B, Ryu IC, Dhital MR, Althuwaynee OF (2013) Landslide susceptibility mapping using certainty factor, index of entropy and logistic regression models in GIS and their comparison at Mugling-Narayanghat road section in Nepal Himalaya. Nat Hazards 65(1):135–165. doi:10.1007/s11069-012-0347-6

Diaz-Uriate R, de Andres SA (2006) Gene selection and classification of microarray data using random forest. BMC Bioinformatics 7(3):1–13. doi:10.1186/1471-2105-7-3

Elith J, Leathwick JR, Hastie T (2008) A working guide to boosted regression trees. J Anim Ecol 77:802–813

Elkadiri R, Sultan M, Youssef A, Elbayoumi T, Chase R, Bulkhi A, Al-Katheeri M (2014) A remote sensing-based approach for debris-flow susceptibility assessment using artificial neural networks and logistic regression modeling. Selected topics in applied earth observations and remote sensing, IEEE J Sel Top Appl Earth Obs Remote Sens. doi:10.1109/JSTARS.2014.2337273

Ercanoglu M, Gokceoglu C (2004) Use of fuzzy relations to produce landslide susceptibility map of a landslide prone area (West Black Sea Region, Turkey). Eng Geol 75:229–250. doi:10.1016/j.enggeo.2004.06.001

Falaschi F, Giacomelli F, Federici PR, Puccinelli A, D'Amato Avanzi G, Pochini A, Ribolini A (2009) Logistic regression versus artificial neural networks: landslide susceptibility evaluation in a sample area of the Serchio River valley, Italy. Nat Hazards 50:551–569. doi:10.1007/s11069-009-9356-5

Federici PR, Puccinelli A, Cantarelli E, Casarosa N, D'amato Avanzi G, Falaschi F, Giannecchini R, Pochini A, Ribolini A, Bottai M, Salvati N, Testi C (2005) Uso di tecniche GIS nella valutazione della pericolosita' di frana nella valle del Serchio (Lu). Atti 9a Conferenza Nazionale ASITA 2:1059–1064

Federici PR, Puccinelli A, Cantarelli E, Casarosa N, D'amato Avanzi G, Falaschi F, Giannecchini R, Pochini A, Ribolini A, Bottai M, Salvati N, Testi C (2007) Multidisciplinary investigations in evaluating landslide hazard. An example in the Serchio River valley (Italy). Quat Int 171–172:52–63. doi:10.1016/j.quaint.2006.10.018

Felicísimo A, Cuartero A, Remondo J, Quirós E (2012) Mapping landslide susceptibility with logistic regression, multiple adaptive regression splines, classification and regression trees, and maximum entropy methods: a comparative study. Landslides. doi:10.1007/s10346-012-0320-1

Gislason PO, Benediktsson JA, Sveinsson JR (2006) Random forests for land cover classification. Pattern Recogn Lett 27:294–300

Giudici P (2005) Data mining: metodi informatici, statistici e applicazioni. McGraw Hill, Milano, p 401

Greco R, Sorriso-Valvo M, Catalano E (2007) Logistic regression analysis in the evaluation of mass-movements susceptibility: the Aspromonte case study, Calabria, Italy. Eng Geol 89:47–66. doi:10.1016/j.enggeo.2006.09.006

Greenwood WR (1985) Geologic map of the Abha quadrangle, sheet 18 F, Kingdom of Saudi Arabia, Ministry of Petroleum and Mineral Resources, Deputy Ministry for Mineral Resources GM-75 c, scale 1:250,000

Ham J, Chen YC, Crawford MP, Ghosh J (2005) Investigation of the random forest framework for classification of hyperspectral data. IEEE Trans Geosci Remote Sens 43(3):492–501

Hansen L, Salamon P (1990) Neural network ensembles. IEEE Trans Pattern Anal Mach Intell 12:993–1001

Hosmer DW, Lemeshow S (2000) Applied logistic regression, 2nd edn. Wiley, New York, p 392

Jaafari A, Najafi A, Pourghasemi HR, Rezaeian J, Sattarian A (2014) GIS-based frequency ratio and index of entropy models for landslide susceptibility assessment in the Caspian forest, northern Iran. Int J Environ Sci Technol 11(4):909–926

Lee S, Pradhan B (2007) Landslide hazard mapping at Selangor, Malaysia using frequency ratio and logistic regression models. Landslides 4(1):33–41. doi:10.1007/s10346-006-0047-y

Maerz NH, Youssef AM, Pradhan B, Bulkhi A (2014) Remediation and mitigation strategies for rock fall hazards along the highways of Fayfa Mountain, Jazan Region, Kingdom of Saudi. Arab J Geosci. doi:10.1007/s12517-014-1423-x

Marmion M, Hjort J, Thullier W, Luoto M (2009) Statistical consensus methods for improving predictive geomorphology maps. Comput Geosci 35:615–625. doi:10.1016/j.cageo.2011.04.012

Mathew J, Jha VK, Rawat GS (2009) Landslide susceptibility zonation mapping and its validation in part of Garhwal Lesser Himalaya, India, using binary logistic regression analysis and receiver operating characteristic curve method. Landslides 6(1):17–26. doi:10.1007/s10346-008-0138-z

McCullagh P, Nelder JA (1989) Generalized linear models, Second Editionth edn. Chapman and Hall/CRC, Boca Raton, p 532

Mckenney DW, Pedlar JH (2003) Spatial models of site index based on climate and soil properties for two boreal tree species in Ontario, Canada. Forest Ecol Manag 175:497–507

Micheletti N, Foresti L, Robert S, Leuenberger M, Pedrazzini A, Jaboyedoff M, Kanevski M (2014) Machine learning feature selection methods for landslide susceptibility mapping. Math Geosci 46:33–57. doi:10.1007/s11004-013-9511-0

Mohammady M, Pourghasemi HR, Pradhan B (2012) Landslide susceptibility mapping at Golestan Province Iran: a comparison between frequency ratio, Dempster-Shafer, and weights of evidence models. J Asian Earth Sci 61:221–236

Nefeslioglu HA, Sezer E, Gokceoglu C, Bozkir AS, Duman TY (2010) Assessment of landslide susceptibility by decision trees in the Metropolitan area of Istanbul, Turkey. Math Problems Eng 1–15, 901095. doi:10.1155/2010/901095

Nikita E (2014) The use of generalized linear models and generalized estimating equations in bioarchaeological studies. Am J Phys Anthropol 153:473–483

O'Brien RM (2007) A caution regarding rules of thumb for variance inflation factors. Qual Quant 41(5):673–690

Ohlmacher GC, Davis JC (2003) Using multiple logistic regression and GIS technology to predict landslide hazard in northeast Kansas, USA. Eng Geol 69:331–343. doi:10.1016/j.enggeo.2006.09.006

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

Pal M (2005) Random forest classifier for remote sensing classification. Int J Remote Sens 26(1):217–222. doi:10.1080/01431160412331269698

Paudel U, Oguchi T (2014) Implementation of random forest in landslide susceptibility study, a case study of the Tokamachi area, Niigata, Japan. Japan Geoscience Union Meeting, Pcaifico Yokohama, 28th April–2nd May, 2014

Payne R (2012) A guide to regression, nonlinear and generalized linear models in GenStat. VSN International, 5 The Waterhouse, Waterhouse Street, Hemel Hempstead, Hertfordshire HP1 1ES, UK, p 88

Petley DN (2008) The global occurrence of fatal landslides in 2007. Geophysical Research Abstracts, vol. 10, EGU General Assembly 2008, p 3

Piccolo D (1998) Statistica. Il Mulino, Bologna, p 969

Pourghasemi HR, Pradhan B, Gokceoglu C (2012a) Application of fuzzy logic and analytical hierarchy process (AHP) to landslide susceptibility mapping at Haraz watershed, Iran. Nat Hazards 63(2):965–996. doi:10.1007/s11069-012-0217-2

Pourghasemi HR, Pradhan B, Gokceoglu C (2012b) Remote sensing data derived parameters and its use in landslide susceptibility assessment using Shannon's entropy and GIS, AEROTECH IV-2012. Appl Mech Mater 225:486–491. doi:10.4028/www.scientific.net/AMM.225.486

Pourghasemi HR, Moradi HR, Fatemi Aghda SM (2013a) Landslide susceptibility mapping by binary logistic regression, analytical hierarchy process, and statistical index models and assessment of their performances. Nat Hazards 69:749–779. doi:10.1007/s11069-013-0728-5

Pourghasemi HR, Pradhan B, Gokceoglu C, Mohammadi M, Moradi HR (2013b) Application of weights-of-evidence and certainty factor models and their comparison in landslide susceptibility mapping at Haraz watershed, Iran. Arab J Geosci 6(7):2351–2365. doi:10.1007/s12517-012-0532-7

Pourtaghi ZS, Pourghasemi HR, Rossi M (2014) Forest fire susceptibility mapping in the Minudasht forests, Golestan province, Iran. Environ Earth Sci. doi:10.1007/s12665-014-3502-4

Pradhan B (2010) Landslide susceptibility mapping of a catchment area using frequency ratio, fuzzy logic and multivariate logistic regression approaches. J Indian Soc Remote Sens 38(2):301–320

Pradhan B, Lee S (2010) Regional landslide susceptibility analysis using back-propagation neural networks model at Cameron Highland, Malaysia. Landslides 7(1):13–30. doi:10.1007/s10346-009-0183-2

Pradhan B, Youssef AM (2010) Manifestation of remote sensing data and GIS on landslide hazard analysis using spatial-based statistical models. Arab J Geosci 3(3):319–326. doi:10.1007/s12517-009-0089-2

Pradhan B, Youssef AM, Varathrajoo R (2010) Approaches for delineating landslide hazard areas using different training sites in an advanced artificial neural network model. Geo-Spat Inform Sci 13(2):93–102

Regmi AD, Yoshida K, Pradhan B, Pourghasemi HR, Khumamoto T, Akgun A (2014) Application of frequency ratio, statistical index and weights-of-evidence models, and their comparison in landslide susceptibility mapping in Central Nepal Himalaya. Arab J Geosci 7(2):725–742. doi:10.1007/s12517-012-0807-z

Remondo J, González A, Díaz de Terán JR, Cendrero A, Fabbri A, Chung CJF (2003) Validation of landslide susceptibility maps; examples and applications from a case study in Northern Spain. Nat Hazds 30(3):437–449. doi:10.1023/B:NHAZ.0000007201.80743.fc

Remondo J, Bonachea J, Cendrero A (2005) A statistical approach to landslide risk modelling at basin scale: from landslide susceptibility to quantitative risk assessment. Landslides 2(4):321–328. doi:10.1007/s10346-005-0016-x

Ridgeway G (2006) Generalized boosted regression models. Documentation on the R package 'gbm', version 1.5-7, Available at: http://www.ipensieri.com/gregr/gbm.shtml

Saha AK, Gupta RP, Sarkar I, Arora MK, Csaplovics E (2005) An approach for GIS-based statistical landslide susceptibility zonation with a case study in the Himalayas. Landslides 2:61–69. doi:10.1007/s10346-004-0039-8

Schapire RE (2003) The boosting approach to machine learning: an overview. Nonlinear Estim Classif 171:149–171

Schleier M, Bi RN, Rohn J, Ehret D, Xiang W (2014) Robust landslide susceptibility analysis by combination of frequency ratio, heuristic GIS-methods and ground truth evaluation for a mountainous study area with poor data availability in the Three Gorges Reservoir area, PR China. Environ Earth Sci 71(7):3007–3023. doi:10.1007/s12665-013-2677-4

Schröder B, Vorpahl P, Märker M, Elsenbeer H (2010) Pitfalls in statistical landslide susceptibility modelling. Geophysical Res Abstracts 12:EGU2010-10786

Stehman SV, Czaplewski LR (1998) Design and analysis of thematic map accuracy assessment: fundamental principles. Remote Sens Environ 64:331–344

Swets JA (1988) Measuring the accuracy of diagnostic systems. Science 270:1285–1293

van Westen CJ, van Asch TWJ, Soeters R (2006) Landslide hazard and risk zonation—why is it still so difficult? Bull Eng Geol Environ 65:167–184. doi:10.1007/s10064-005-0023-0

Vorpahl P, Elsenbeer H, Marker M, Schroder B (2012) How can statistical models help to determine driving factors of landslides? Ecol Model 239:27–39

Williams G (2011) Data mining with Rattle and R (the art of excavating data for knowledge discovery series). New York, p 347

Wu XL, Ren F, Niu RQ (2014) Landslide susceptibility assessment using object mapping units, decision tree, and support vector machine models in the Three Gorges of China. Environ Earth Sci 71(11):4725–4738. doi:10.1007/s12665-013-2863-4

Xu C (2013) Assessment of earthquake-triggered landslide susceptibility based on expert knowledge and information value methods: a case study of the 20 April 2013 Lushan, China Mw6.6 earthquake. Di Adv 6(13):119–130

Yeon YK, Han JG, Ryu KH (2010) Landslide susceptibility mapping in Injae, Korea, using a decision tree. Eng Geol 116:274–283. doi:10.1016/j.enggeo.2010.09.009

Yilmaz I (2009) A case study from Koyulhisar (Sivas-Turkey) for landslide susceptibility mapping by artificial neural networks. Bull Eng Geol Environ 68(3):297–306. doi:10.1007/s10064-009-0185-2

Youssef AM (2015) Landslide susceptibility delineation in the Ar-Rayth Area, Jizan, Kingdom of Saudi Arabia, by using analytical hierarchy process, frequency ratio, and logistic regression models. Environ Earth Sci. doi:10.1007/s12665-014-4008-9, **Article on line first**

Youssef AM, Maerz N (2013) Overview of some geological hazards in the Saudi Arabia. Environ Earth Sci 70:3115–3130. doi:10.1007/s12665-013-2373-4

Youssef AM, Maerz NH, Hassan AM (2009) Remote sensing applications to geological problems in Egypt: case study, slope instability investigation, Sharm El-Sheikh/Ras-Nasrani Area, Southern Sinai. Landslides 6(4):353–360. doi:10.1007/s10346-009-0158-3

Youssef AM, Maerz HN, Al-Otaibi AA (2012) Stability of rock slopes along Raidah Escarpment Road, Asir Area, Kingdom of Saudi Arabia. J Geogr. doi:10.5539/jgg.v4n2p48

Youssef AM, Pradhan B, Maerz NH (2013) Debris flow impact assessment caused by 14 April 2012 rainfall along the Al-Hada Highway, Kingdom of Saudi Arabia using high-resolution satellite imagery. Arab J Geosci 1–11. doi:10.1007/s12517-013-0935-0

Youssef AM, Al-kathery M, Pradhan B (2014a) Landslide susceptibility mapping at Al-Hasher Area, Jizan (Saudi Arabia) using GIS-based frequency ratio and index of entropy models. Geosci J. doi:10.1007/s12303-014-0032-8

Youssef AM, Pradhan B, Jebur MN, El-Harbi HM (2014b) Landslide susceptibility mapping using ensemble bivariate and multivariate statistical models in Fayfa area, Saudi Arabia. Environ Earth Sci. doi:10.1007/s12665-014-3661-3

Youssef AM, Al-kathery M, Pradhan B, Elsahly T (2014c) Debris flow impact assessment along the Al-Raith Road, Kingdom of Saudi Arabia, using remote sensing data and field investigations. Geomat Nat Hazards Risk. doi:10.1080/19475705.2014.933130

Zhu L, Huang J (2006) GIS-based logistic regression method for landslide susceptibility mapping in regional scale. J Zhejiang Univ Sci A 7:2007–2017

A. M. Youssef
Geology Department, Faculty of Science,
Sohag University,
Sohag, Egypt

A. M. Youssef · M. M. Al-Katheeri
Geological Hazards Department, Applied Geology Sector,
Saudi Geological Survey,
P.O. Box 54141, Jeddah, 21514, Kingdom of Saudi Arabia

H. R. Pourghasemi (✉)
Department of Natural Resources and Environmental Engineering, College of Agriculture,
Shiraz University,
Shiraz, Iran
e-mail: hr.pourghasemi@shirazu.ac.ir

Z. S. Pourtaghi
Department of Environment Management Engineering, College of Natural Resources,
Yazd University,
Yazd, Iran