Hindawi Publishing Corporation
Journal of Applied Mathematics
Volume 2014, Article ID 929768, 9 pages
http://dx.doi.org/10.1155/2014/929768

Research Article

Modeling and Testing Landslide Hazard Using Decision Tree

Mutasem Sh. Alkhasawneh,[1] Umi Kalthum Ngah,[1] Lea Tien Tay,[1] Nor Ashidi Mat Isa,[1] and Mohammad Subhi Al-Batah[2]

[1] *Imaging and Computational Intelligence (ICI) Group, School of Electrical & Electronic Engineering, Universiti Sains Malaysia, Engineering Campus, Nibong Tebal, 14300 Penang, Malaysia*
[2] *Department of Computer Science and Software Engineering, Faculty of Science and Information Technology, Jadara University, P.O. Box 733, Irbid 21110, Jordan*

Correspondence should be addressed to Mutasem Sh. Alkhasawneh; m_sh_ka1@yahoo.com

Received 9 September 2013; Accepted 9 December 2013; Published 4 February 2014

Academic Editor: Chong Lin

This paper proposes a decision tree model for specifying the importance of 21 factors causing the landslides in a wide area of Penang Island, Malaysia. These factors are vegetation cover, distance from the fault line, slope angle, cross curvature, slope aspect, distance from road, geology, diagonal length, longitude curvature, rugosity, plan curvature, elevation, rain perception, soil texture, surface area, distance from drainage, roughness, land cover, general curvature, tangent curvature, and profile curvature. Decision tree models are used for prediction, classification, and factors importance and are usually represented by an easy to interpret tree like structure. Four models were created using Chi-square Automatic Interaction Detector (CHAID), Exhaustive CHAID, Classification and Regression Tree (CRT), and Quick-Unbiased-Efficient Statistical Tree (QUEST). Twenty-one factors were extracted using digital elevation models (DEMs) and then used as input variables for the models. A data set of 137570 samples was selected for each variable in the analysis, where 68786 samples represent landslides and 68786 samples represent no landslides. 10-fold cross-validation was employed for testing the models. The highest accuracy was achieved using Exhaustive CHAID (82.0%) compared to CHAID (81.9%), CRT (75.6%), and QUEST (74.0%) model. Across the four models, five factors were identified as most important factors which are slope angle, distance from drainage, surface area, slope aspect, and cross curvature.

1. Introduction

Landslide is one of the most aggressive natural disasters that causes loss of lives and billions of dollars damages annually worldwide. They pose a threat to the safety of humankind lives and, the environment, resources, and property [1]. Landslide susceptibility is defined as the propensity of an area to generate landslides [2]. Assuming that landslides will occur in the future because of the same conditions that produced them in the past, susceptibility assessments can be used to predict the geographical location of future landslides [3–5]. With the characteristics of high incidence and extensive occurrence range, landslide research has aroused the attention of many scientists, some of whom have focused on landslide susceptibility mapping [6, 7]. Through scientific analysis of landslide susceptibility mapping, we can assess and locate risky landslide susceptible areas. Furthermore, it allows one to take the proper precautions to reduce the negative impacts of landslides [8].

Many studies have been conducted to detect landslides and to analyze the landslide hazard using the Geographic Information Systems (GIS) and remote sensing [9–13]. Recently, with the development of GIS data processing techniques, quantitative studies have been applied to landslide susceptibility analysis using various techniques. Such studies can be identified on the basis of the techniques used, such as probabilistic methods [14–18], logistic regression [19–21], and artificial neural network [22–25]. Most of these studies were aimed at increasing the accuracy of landslide prediction by finding suitable techniques for the respective study area. The objective of this study was to propose the best decision tree model to determine the most important factors which lead landslide susceptibility to occur. Decision tree is a popular classification technique and represents a good compromise between comprehensibility, accuracy, and efficiency [26].

Statistical decision tree models have been successfully used to classify and to estimate land use, land cover, and other geographical attributes from remote sensing data [27, 28]. Decision tree, having its origin in machine learning theory, is an efficient tool for classification and estimation. Unlike other statistical methods, decision tree makes no statistical assumptions, can handle data that are represented on different measurement scales, and is computationally fast [22]. Decision tree also has advantages that the estimation processes and order of important explanatory variables are explicitly represented by tree structures [29]. In addition, recent developments of computer technologies, algorithms of pattern recognition, and automatic methods of decision-tree design have enabled the use of decision tree models.

Pal and Mather [30] demonstrated the advantages of the decision tree for land cover classification in comparison with other classifiers, such as the maximum likelihood method and artificial neural networks. Saito et al. [2] used decision tree models to analyze a distribution of landslides that were almost suspended or dormant. They also indicated that decision tree models are useful for estimating landslide distributions. Bui et al. [31] compared the decision tree for landslide prediction in Vietnam. Decision tree showed a decent performance compared with Support Vector Machines (SVM) and Naive Bayes Models. Meanwhile decision tree showed a good ability in determinations of the important factors causing the landslide compared with other used models. Pang et al. [32] produced the landslide hazard mapping of Penang Island using decision tree Quinlan's algorithm C4.5. Twelve landslide causative factors were used in his study. Pradhan [33] used three models: decision tree, SVM, and adaptive neurofuzzy inference system (ANFIS) for producing the landslide hazard map for Penang Hill area. The decision tree showed a better performance compared with SVM and ANFIS classifier.

In this paper, four decision tree methods were used to build the optimum decision models including Chi-square Automatic Interaction Detector (CHAID), Exhaustive CHAID, Classification and Regression Tree (CRT) and Quick-Unbiased-Efficient Statistical Tree (QUEST). Twenty one factors were selected as the input variables of the decision trees. A data set of 137570 samples from Penang Island in Malaysia was used as examples for building the decision trees. The experiment contained ten rounds according to different partitions of training sets and test sets.

2. Decision Trees

A decision tree is a technique for finding and describing structural patterns in data as tree structures; a decision tree does not require the relationship between all the input variables and an objective variable in advance. This technique helps to explain data and to make predictions using the data [34]. A decision tree can also handle data measured on different scales, without any assumptions concerning the frequency distributions of the data, based on its nonlinear relationship [35]. Therefore, all variables were put into the decision tree model.

The main purpose of using the decision tree is to achieve a more concise and perspicuous representation of the relationship between an objective variable and explanatory variables. Namely, the decision tree can be visualized more easily; unlike neural networks, it is not a "black box."

The decision tree is based on a multistage or hierarchical decision scheme (tree structure). The tree is composed of a root node, a set of internal nodes, and a set of terminal nodes (leaves). Each node of the decision tree structure makes a binary decision that separates either one class or some of the classes from the remaining classes. The processing is carried out by moving down the tree until the terminal node is reached. In a decision tree, features that carry maximum information are selected for classification, while remaining features are rejected, thereby increasing computational efficiency [36]. The top down induction of the decision tree indicates that variables in the higher order of the tree structure are more important.

There are three tree types of decision tree: CRT, CHAID and Exhaustive CHAID, and Quest. The algorithms of the three types follow the following steps. Start tree building by assigning the node to classes, stopping tree building. Reach the optimal tree selection and perform cross-validation [37]. CART performs tree "Pruning" before producing the optimal tree selection, while CHAID method performs statistical tests at each step of splitting.

2.1. Classification and Regression Tree (CRT). CRT is a recursive partitioning method to be used both for regression and classification. CRT is constructed by splitting subsets of the data set using all predictor variables to create two child nodes repeatedly, beginning with the entire data set. The best predictor is chosen using a variety of impurity or diversity measures (Gini, towing, ordered towing and least-squared deviation). The goal is to produce subsets of the data which are as homogeneous as possible with respect to the target variable [38]. In this study, we used measure of Gini impurity that was used for categorical target variables.

Gini Impurity Measure. The Gini index at node t, $g(t)$, is defined as

$$g(t) = \sum_{j \neq i} p(j \mid t)(i \mid t), \tag{1}$$

where i and j are categories of the target variable. The equation for the Gini index can also be written as

$$g(t) = 1 - \sum_{j} p^2(j \mid t). \tag{2}$$

Thus, when the cases in a node are evenly distributed across the categories, the Gini index takes its maximum value of $1 - (1/k)$, where k is the number of categories for the target variable. When all cases in the node belong to the same category, the Gini index equals 0.

If costs of misclassification are specified, the Gini index is computed as

$$g(t) = \sum_{j \neq i} C(i \mid j)\, p(j \mid t)\, p(i \mid t), \tag{3}$$

where $C(i \mid j)$ is the probability of misclassifying a category j case as category i.

The Gini criterion function for split s at node t is defined as

$$\emptyset(s,t) = g(t) - p_L g(t_L) - p_R g(t_R), \tag{4}$$

where p_L is the proportion of cases in t sent to the left child node, and p_R is the proportion sent to the right child node. The split s is chosen to maximize the value of $\emptyset(s,t)$. This value is reported as the improvement in the tree [39].

2.2. Chi-Square Automatic Interaction Detector (CHAID) and Exhaustive CHAID. CHAID method is based the χ^2-test of association. A CHAID tree is a decision tree that is constructed by repeatedly splitting subsets of the space into two or more child nodes, beginning with the entire data set [40]. To determine the best split at any node, any allowable pair of categories of the predictor variables is merged until there is no statistically significant difference within the pair with respect to the target variable. This CHAID method naturally deals with interactions between the independent variables that are directly available from an examination of the tree. The final nodes identify subgroups defined by different sets of independent variables [41].

The CHAID algorithm only accepts nominal or ordinal categorical predictors. When predictors are continuous, they are transformed into ordinal predictors before using the following algorithm. For each predictor variable X, merge nonsignificant categories. Each final category of X will result in one child node if X is used to split the node. The merging step also calculates the adjusted p value that is to be used in the splitting step.

(1) If X has 1 category only, stop and set the adjusted p value to be 1.

(2) If X has 2 categories, go to step 8.

(3) Else, find the allowable pair of categories of X (an allowable pair of categories for ordinal predictor is two adjacent categories, and for nominal predictor is any two categories) that is least significantly different. The most similar pair is the pair whose test statistic gives the largest p value with respect to the dependent variable Y.

(4) For the pair having the largest p value, check if its p value is larger than a specified alpha-level α merge. If it does, this pair is merged into a single compound category. Then a new set of categories of X is formed. If it does not, then go to step 7.

(5) (Optional) if the newly formed compound category consists of three or more original categories, then find the best binary split within the compound category in which p value is the smallest. Perform this binary split if its p value is not larger than an alpha-level α split-merge.

(6) Go to step 2.

(7) (Optional) any category having too few observations (as compared with a user-specified minimum segment size) is merged with the most similar other category as measured by the largest of the p values.

(8) The adjusted p value is computed for the merged categories by applying Bonferroni adjustments [42, 43].

The CHAID algorithm reduces the number of predictor categories by merging categories when there is no significant difference between them with respect to the class. When no more classes can be merged the predictor can be considered as a candidate for a split at the node. The original CHAID algorithm is not guaranteed to find the best (most significant) split of all of those examined because it uses the last split tested. The Exhaustive CHAID algorithm attempts to overcome this problem by continuing to merge categories, irrespective of significance level, until only two categories remain for each predictor. It then used the split with the largest significance value rather than the last one tried. The Exhaustive CHAID requires more computer time [44]. Calculations of (unadjusted) p values in the above algorithms depend on the type of dependent variable. The merging step of both CHAID and Exhaustive CHAID sometimes needs the p value for a pair of X categories and sometimes needs the p value for all the categories of X. When p value for a pair of X categories is needed, only part of data in the current node is relevant. Let D denotes the relevant data. Suppose in D there are I categories of X, and J categories of Y (if Y is categorical). The p value calculation using data in D is given below. The null hypothesis of independence of X and the dependent variable Y is tested. To do the test, a contingency (or count) table is formed using classes of Y as columns and categories of the predictor X as rows. The expected cell frequencies under the null hypothesis are estimated. The observed cell frequencies and the expected cell frequencies are used to calculate Pearson chi-squared statistic or likelihood ratio statistic. The p value is computed based on the Pearson's chi-square statistic method. Consider

$$X^2 = \sum_{j=1}^{J} \sum_{i=1}^{I} \frac{\left(n_{ij} - \widehat{m}_{ij}\right)^2}{\widehat{m}_{ij}}, \tag{5}$$

where $n_{ij} = \sum_{n \in D} f_n I\ (x_n = i \wedge y_n = j)$ is the observed cell frequency and $\widehat{m}_{ij}$ is the estimated expected cell frequency for cell $x_n = i$, $y_n = j$ from independence model as follows. The corresponding p value is given by $p = \mathrm{pr}(X_d^2 > X^2)$ for Pearson's chi-square test where X_d^2 follows a chi-squared distribution with degrees of freedom $d = (J-1)(I-1)$, $\widehat{m}_{ij} = n_i \cdot n_j / n$, $n_i = \sum_{j=1}^{J} n_{ij}$, $n_j = \sum_{i=1}^{I} n_{ij}$, $n = \sum_{j=1}^{j} \sum_{i=1}^{I} n_{ij}$.

In step 8 the adjusted p-value is calculated as the p value times a Bonferroni multiplier. The Bonferroni multiplier adjusts for multiple tests. Suppose that a predictor variable originally has I categories, and it is reduced to r categories after the merging step. The Bonferroni multiplier B is the number of possible ways that I categories can be merged into r categories. For $r = I$, $B = 1$. For $2 \leq r < I$, use the following equation:

$$B = \sum_{v=0}^{r-1} (-1)^v \frac{(r-v)^I}{v!(r-v)^!}. \qquad (6)$$

2.3. Quick-Unbiased-Efficient Statistical Tree (QUEST). QUEST is a binary split decision tree algorithm for classification and data mining. QUEST can be used with univariant or linear combination splits. A unique feature is that its attribute selection method has negligible bias. If all the attributes are uninformative with respect to the class attribute, then each has approximately the same change of being selected to split a node [45].

The QUEST tree growing process consists of the selection of a split predictor, selection of a split point for the selected predictor, and stopping. In this algorithm, only univariant splits are considered. For selection of split predictor, it uses the following algorithm.

(1) For each continuous predictor X, perform an ANOVA F-test that tests if all the different classes of the dependent variable Y have the same mean of X, and calculate the p value according to the F statistics. For each categorical predictor, perform Pearson's χ^2-test of Y and X's independence, and calculate the p value according to the X^2 statistics.

(2) Find the predictor with the smallest p value and denote it X^*.

(3) If this smallest p value is less than α/M, where $\alpha \in (0, 1)$ is a user-specified level of significance and M is the total number of predictor variables, predictor X^* is selected as the split predictor for the node. If not, go to 4.

(4) For each continuous predictor X, compute Levene's F statistic based on the absolute deviation of X from its class mean to test if the variances of X for different classes of Y are the same, and calculate the p value for the test.

(5) Find the predictor with the smallest p value and denote it as X^{**}.

(6) If this smallest p value is less than $\alpha/(M + M1)$, where $M1$ is the number of continuous predictors, X^{**} is selected as the split predictor for the node. Otherwise, this node is not split [45].

3. Study Area

As shown in Figure 1, this study is focused on Penang Island which lies between 5°15′ to 5°30′ N latitude and 100°10′ to 100°20′ E longitude. The North Channel separates the study area from the mainland. It occupies an area of 285 km^2 and is one of the 13 states of Malaysia. The island is bounded to the north and east by the state of Kedah, to the south by the state of Perak, and to the west by the Strait of Malacca and Sumatra (Indonesia).

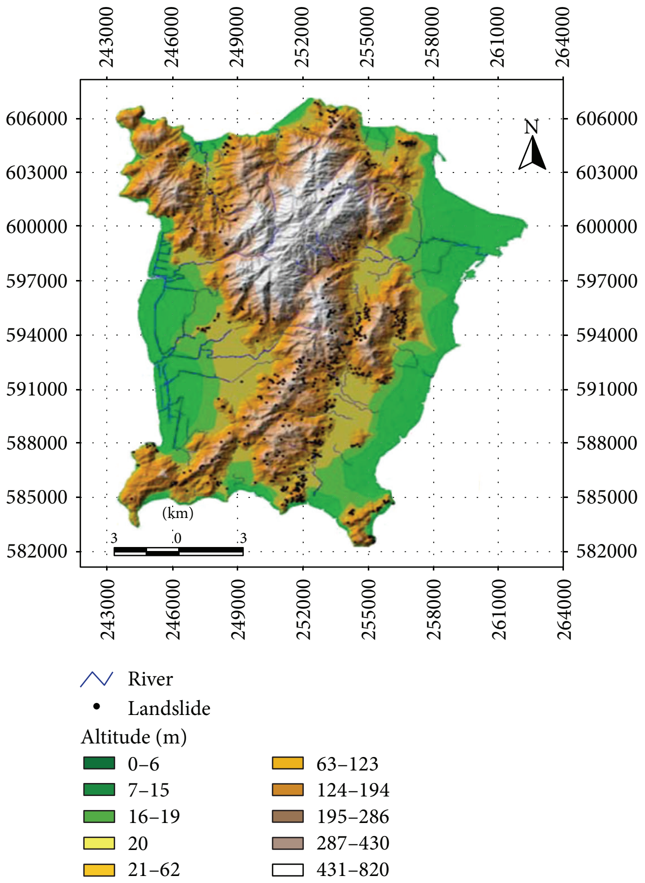

Figure 1: Study area map and landslide location map with hill shaded map.

Penang Island consists of both the island of Penang and a coastal strip on the mainland known as the Province Wellesley. This paper focuses only on the island, where frequent landslides occurred and threaten lives and damage properties [46, 47]. The heavy rain plays a major role in triggering the landslides in the study area. Data from the Malaysian Meteorological Department recorded that the rainfall amount varies approximately between 2254 mm and 2903 mm annually in the study area. Penang Island has a tropical climate with high temperature of 29°C to 32°C and humidity ranges from 65% to 96%. Topographic elevations vary between 0 m and 820 m above sea level. The slope angle ranges from 0° to 87° while 43.28% of island is flat. Geological data from the Minerals and Geosciences Department, Malaysia, show that Ferringhi granite, Batu Maung granite, clay, and sand granite represent more than 72% of the study area's geology. Vegetation cover consists mainly of forests and fruit plantations.

4. Data Collection

An effective intelligent system requires a comprehensive data set. Therefore 137570 samples of data were selected in this

Table 1: Number of nodes, terminal nodes, and order of importance variable.

Decision tree model	No. of nodes	No. of terminal nodes	Independent variable included "order of importance"
CHAID	317	254	$V_3, V_{16}, V_{15}, V_5, V_4, V_{21}, V_{17}, V_{13}, V_1, V_{12}, V_{19}, V_{18}, V_{10}$
Exhaustive CHAID	377	302	$V_3, V_{16}, V_{15}, V_5, V_4, V_{21}, V_9, V_{13}, V_1, V_{17}, V_{12}, V_{19}, V_{18}, V_{10}$
CRT	43	22	$V_3, V_{16}, V_{15}, V_5, V_4, V_2, V_{21}, V_6, V_8, V_9, V_{13}, V_1, V_{17}, V_{12}, V_{19}, V_{20}, V_{18}, V_{11}, V_{14}, V_7, V_{10}$
QUEST	55	28	$V_3, V_{16}, V_{15}, V_5, V_4, V_2, V_{21}, V_6, V_8, V_{13}, V_1, V_{17}, V_{12}, V_{19}, V_{18}, V_{11}, V_7, V_{10}$

analysis, where 68786 samples represent landslides and 68786 samples represent no landslides. Then, Digital Elevation Map (DEM) is used to extract 21 topographic factors. The DEM with five-meter resolutions of Penang Island was obtained from the Department of Survey and Mapping, Malaysia. The extracted factors are acronyms as V_1 (vegetation cover), V_2 (distance from the fault line), V_3 (slope angle), V_4 (cross curvature), V_5 (slope aspect), V_6 (distance from road), V_7 (geology), V_8 (diagonal length), V_9 (longitude curvature), V_{10} (rugosity), V_{11} (plan curvature), V_{12} (elevation), V_{13} (rain perception), V_{14} (soil texture), V_{15} (surface area), V_{16} (distance from drainage), V_{17} (roughness), V_{18} (land cover), V_{19} (general curvature), V_{20} (tangent curvature), and V_{21} (profile curvature). In the previous studies which have been done on Penang Island, only 14 factors (V_1 to V_{14}) were on the subject of investigation for landslide [48]. While, the factors V_{15} to V_{21} will be applied and investigated for the first time on the study area. Furthermore the 21 factors represent the available data of all factors that can cause the landslide in the study area. The intelligent system target (landslides history) is represented by 0 for no landslide and 1 for landslide. The data were normalized to range between 0 and 1, for each of the factors individually. A 10-fold cross-validation analysis was performed as an initial evaluation of the test error of the algorithms. Briefly, this process involves splitting up the data set into 10 random segments and using 9 of them for training and the 10th as a test set for the algorithm. Classification accuracy of each model was calculated as follows.

The accuracy of correctly classified landslide (1) is given by

$$\text{accuracy}(1) = \sum_{i=1}^{10} \frac{\text{number of correctly classified}(1)}{\text{number of }(1)}. \quad (7)$$

The accuracy of correctly classified no landslide (0) is given by

$$\text{accuracy}(0) = \sum_{i=1}^{10} \frac{\text{number of correctly classified}(0)}{\text{number of }(0)}. \quad (8)$$

The overall accuracy for decision tree model is given by

$$\text{overall accuracy} = \frac{\text{accuracy}(1) + \text{accuracy}(0)}{2}. \quad (9)$$

5. Discussion

Four tree algorithms, CHAID, Exhaustive CHAID, CRT, and QUEST, were applied to map landslide susceptibility hazard. The 4 trees construction is based on the entire sample of 137572 cases, a cross-validation with 10 folds, 0.05 adjustment of the probabilities, a minimum cases in parent node of 100, a minimum cases in child node of 50, and equal misclassification costs. The maximum number of levels is 3 for CHAID and exhaustive CHAID and 5 for CRT and QUEST.

The results for number of nodes, number of terminal nodes, and importance of independent variable produced by each model are presented in Table 1. The classification trees obtained show a tree with a total of 317 nodes that consist of 254 terminal nodes using CHAID, 377 nodes with 302 terminal nodes using exhaustive CHAID, 43 nodes with 22 terminal nodes using CRT, and 55 nodes with 28 terminal nodes using QUEST. An example of decision tree using CRT method is explained in Table 2. The tree has 43 nodes including the root node, 20 internal nodes, and 22 leaves (terminal nodes). Percentages in each category and in each joint category are presented in Table 2.

Also, the decision tree methods are used to analyze the relationships between landslide susceptibility and related factors. The normalized importance of factors in classification using CRT is shown in Figure 2. The top-down induction of the decision tree indicates that variables in the higher order of the tree structure are more important for analyzing landslide susceptibility. The tree structure demonstrates that important variables related to high landslide susceptibility catchments are ordered as follows: V_3 (slope angle), V_{16} (distance from drainage), V_{15} (surface area), V_5 (slope aspect), and V_4 (cross curvature).

The results for prediction accuracy produced by each model are presented in Table 3. The results show high classification accuracy for exhaustive CHAID algorithm as compared to other algorithms. It is found that the prediction accuracy for exhaustive CHAID is 82.0%, with sensitivity 72.3% and specificity 91.7%.

6. Conclusion

This study has analyzed landslide susceptibility in Penang Island, Malaysia, using ensemble learning with a decision-tree model. We can conclude that the decision tree clearly

Table 2: Tree table using CRT method.

Node	0		1		Total		Predicted category	Parent node	Primary independent variable		
	N	Percent	N	Percent	N	Percent			Variable	Improvement	Split values
0	68786	50.0%	68786	50.0%	137572	100.0%	1				
1	31018	95.8%	1347	4.2%	32365	23.5%	0	0	V_{15}	0.129	≤0.03725
2	37768	35.9%	67439	64.1%	105207	76.5%	1	0	V_{15}	0.129	>0.03725
3	29797	99.0%	314	1.0%	30111	21.9%	0	1	V_{16}	0.006	≤0.04580
4	1221	54.2%	1033	45.8%	2254	1.6%	0	1	V_{16}	0.006	>0.04580
5	15591	28.3%	39565	71.7%	55156	40.1%	1	2	V_{19}	0.010	≤0.02120
6	22177	44.3%	27874	55.7%	50051	36.4%	1	2	V_{19}	0.010	>0.02120
7	379	36.0%	674	64.0%	1053	0.8%	1	4	V_{17}	0.001	≤0.00958
8	842	70.1%	359	29.9%	1201	0.9%	0	4	V_{17}	0.001	>0.00958
9	6228	37.4%	10421	62.6%	16649	12.1%	1	5	V_3	0.003	≤0.01292
10	9363	24.3%	29144	75.7%	38507	28.0%	1	5	V_3	0.003	>0.01292
11	3727	75.5%	1207	24.5%	4934	3.6%	0	6	V_{16}	0.008	≤0.05104
12	18450	40.9%	26667	59.1%	45117	32.8%	1	6	V_{16}	0.008	>0.05104
13	237	28.0%	608	72.0%	845	0.6%	1	7	V_{12}	0.000	≤0.15625
14	142	68.3%	66	31.7%	208	0.2%	0	7	V_{12}	0.000	>0.15625
15	216	99.5%	1	0.5%	217	0.2%	0	8	V_{15}	0.000	≤0.02376
16	626	63.6%	358	36.4%	984	0.7%	0	8	V_{15}	0.000	>0.02376
17	1305	61.2%	829	38.8%	2134	1.6%	0	9	V_{17}	0.002	≤0.02659
18	4923	33.9%	9592	66.1%	14515	10.6%	1	9	V_{17}	0.002	>0.02659
19	3413	30.0%	7980	70.0%	11393	8.3%	1	10	V_{18}	0.001	≤0.05873
20	5950	21.9%	21164	78.1%	27114	19.7%	1	10	V_{18}	0.001	>0.05873
21	48	28.4%	121	71.6%	169	0.1%	1	11	V_{13}	0.001	≤0.10
22	3679	77.2%	1086	22.8%	4765	3.5%	0	11	V_{13}	0.001	>0.10
23	10966	47.6%	12083	52.4%	23049	16.8%	1	12	V_3	0.003	≤0.02510
24	7484	33.9%	14584	66.1%	22068	16.0%	1	12	V_3	0.003	>0.02510
25	68	98.6%	1	1.4%	69	0.1%	0	14	V_{12}	0.000	≤0.21875
26	74	53.2%	65	46.8%	139	0.1%	0	14	V_{12}	0.000	>0.21875
27	232	47.5%	256	52.5%	488	0.4%	1	16	V_{21}	0.000	≤0.4375
28	394	79.4%	102	20.6%	496	0.4%	0	16	V_{21}	0.000	>0.4375
29	758	79.2%	199	20.8%	957	0.7%	0	17	V_{15}	0.001	≤0.07812
30	547	46.5%	630	53.5%	1177	0.9%	1	17	V_{15}	0.001	>0.07812
31	3200	29.5%	7633	70.5%	10833	7.9%	1	18	V_{13}	0.001	≤0.50
32	1723	46.8%	1959	53.2%	3682	2.7%	1	18	V_{13}	0.001	>0.50
33	1828	22.1%	6450	77.9%	8278	6.0%	1	19	V_{17}	0.003	≤0.17214
34	1585	50.9%	1530	49.1%	3115	2.3%	0	19	V_{17}	0.003	>0.17214
35	5873	21.7%	21164	78.3%	27037	19.7%	1	20	V_{18}	0.001	≤0.46488
36	77	100.0%	0	0.0%	77	0.1%	0	20	V_{18}	0.001	>0.46488
37	2179	71.4%	873	28.6%	3052	2.2%	0	22	V_1	0.000	≤0.15385
38	1500	87.6%	213	12.4%	1713	1.2%	0	22	V_1	0.000	>0.15385
39	5193	59.2%	3577	40.8%	8770	6.4%	0	23	V_{18}	0.003	≤0.08809
40	5773	40.4%	8506	59.6%	14279	10.4%	1	23	V_{18}	0.003	>0.08809
41	7078	32.9%	14405	67.1%	21483	15.6%	1	24	V_{17}	0.001	≤0.52185
42	406	69.4%	179	30.6%	585	0.4%	0	24	V_{17}	0.001	>0.52185

Table 3: Classification accuracy produced by each model.

Decision tree model	Classification		
	Predicted (0)	Predicted (1)	Overall
CHAID	73.5%	90.3%	81.9%
Exhaustive CHAID	72.3%	91.7%	82.0%
CRT	61.4%	89.7%	75.6%
QUEST	54.4%	93.5%	74.0%

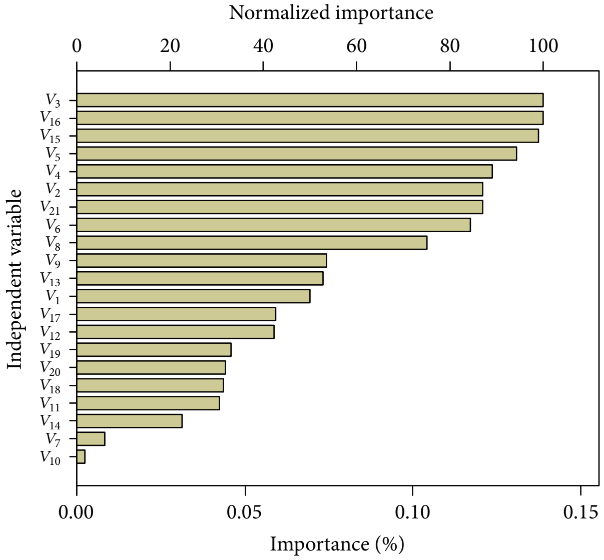

Figure 2: Normalized importance of factors using CRT method.

indicates the order of important variables and quantitatively describes the relationships among the occurrence of landslides, topography, and geology. The decision-tree model using the exhaustive CHAID algorithm showed greater accuracy than the other models, demonstrating the usefulness of the decision tree model for landslide hazard mapping. Accuracies were 82.0% for the exhaustive CHAID, 81.9% for the CHAID, 75.6% for the CRT, and 74.0% for the Quest algorithm. In this study, we determined factors that may be involved in landslide susceptibility, and the results can be used for landslide hazard mapping in other regions. Moreover, landslide hazard mapping map can be used to help mitigate hazards to people and facilities and as basic data for developing plans to prevent landslide hazards, such as in locating, monitoring, and facility sites. Further case studies and modeling are needed to better generalize the factors involved in landslide susceptibility.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] P. Aleotti and R. Chowdhury, "Landslide hazard assessment: summary review and new perspectives," *Bulletin of Engineering Geology and the Environment*, vol. 58, no. 1, pp. 21–44, 1999.

[2] H. Saito, D. Nakayama, and H. Matsuyama, "Comparison of landslide susceptibility based on a decision-tree model and actual landslide occurrence: The Akaishi Mountains, Japan," *Geomorphology*, vol. 109, no. 3-4, pp. 108–121, 2009.

[3] F. Guzzetti, A. Carrara, M. Cardinali, and P. Reichenbach, "Landslide hazard evaluation: a review of current techniques and their application in a multi-scale study, Central Italy," *Geomorphology*, vol. 31, no. 1–4, pp. 181–216, 1999.

[4] F. Guzzetti, P. Reichenbach, F. Ardizzone, M. Cardinali, and M. Galli, "Estimating the quality of landslide susceptibility models," *Geomorphology*, vol. 81, no. 1-2, pp. 166–184, 2006.

[5] F. Guzzetti, P. Reichenbach, M. Cardinali, M. Galli, and F. Ardizzone, "Probabilistic landslide hazard assessment at the basin scale," *Geomorphology*, vol. 72, no. 1–4, pp. 272–299, 2005.

[6] E. Yesilnacar and T. Topal, "Landslide susceptibility mapping: a comparison of logistic regression and neural networks methods in a medium scale study, Hendek region (Turkey)," *Engineering Geology*, vol. 79, no. 3-4, pp. 251–266, 2005.

[7] D. P. Kanungo, M. K. Arora, S. Sarkar, and R. P. Gupta, "A comparative study of conventional, ANN black box, fuzzy and combined neural and fuzzy weighting procedures for landslide susceptibility zonation in Darjeeling Himalayas," *Engineering Geology*, vol. 85, no. 3-4, pp. 347–366, 2006.

[8] S. He, P. Pan, L. Dai, H. Wang, and J. Liu, "Application of kernel-based Fisher discriminant analysis to map landslide susceptibility in the Qinggan River delta, Three Gorges, China," *Geomorphology*, vol. 171-172, pp. 30–41, 2012.

[9] A. Carrara, "Multivariate models for landslide hazard evaluation," *Journal of the International Association for Mathematical Geology*, vol. 15, no. 3, pp. 403–426, 1983.

[10] L. Ayalew and H. Yamagishi, "Slope failures in the Blue Nile basin, as seen from landscape evolution perspective," *Geomorphology*, vol. 57, no. 1-2, pp. 95–116, 2004.

[11] G. Metternicht, L. Hurni, and R. Gogu, "Remote sensing of landslides: an analysis of the potential contribution to geospatial systems for hazard assessment in mountainous environments," *Remote Sensing of Environment*, vol. 98, no. 2-3, pp. 284–303, 2005.

[12] D. E. Alexander, "A brief survey of GIS in mass-movement studies, with reflections on theory and methods," *Geomorphology*, vol. 94, no. 3-4, pp. 261–267, 2008.

[13] J. Remondo, J. Bonachea, and A. Cendrero, "Quantitative landslide risk assessment and mapping on the basis of recent occurrences," *Geomorphology*, vol. 94, no. 3-4, pp. 496–507, 2008.

[14] L. Luzi, F. Pergalani, and M. T. J. Terlien, "Slope vulnerability to earthquakes at subregional scale, using probabilistic techniques and geographic information systems," *Engineering Geology*, vol. 58, no. 3-4, pp. 313–336, 2000.

[15] S. Lee and K. Min, "Statistical analysis of landslide susceptibility at Yongin, Korea," *Environmental Geology*, vol. 40, no. 9, pp. 1095–1113, 2001.

[16] L. Donati and M. C. Turrini, "An objective method to rank the importance of the factors predisposing to landslides with the GIS methodology: application to an area of the Apennines (Valnerina; Perugia, Italy)," *Engineering Geology*, vol. 63, no. 3-4, pp. 277–289, 2002.

[17] S. Lee and U. Choi, "Development of GIS-based geological hazard information system and its application for landslide analysis in Korea," *Geosciences Journal*, vol. 7, no. 3, pp. 243–252, 2003.

[18] B. Neuhäuser and B. Terhorst, "Landslide susceptibility assessment using "weights-of-evidence" applied to a study area at the Jurassic escarpment (SW-Germany)," *Geomorphology*, vol. 86, no. 1-2, pp. 12–24, 2007.

[19] P. M. Atkinson and R. Massari, "Generalised linear modelling of susceptibility to landsliding in the central Apennines, Italy," *Computers and Geosciences*, vol. 24, no. 4, pp. 373–385, 1998.

[20] F. C. Dai, C. F. Lee, J. Li, and Z. W. Xu, "Assessment of landslide susceptibility on the natural terrain of Lantau Island, Hong Kong," *Environmental Geology*, vol. 40, no. 3, pp. 381–391, 2001.

[21] H. A. Nefeslioglu, T. Y. Duman, and S. Durmaz, "Landslide susceptibility mapping for a part of tectonic Kelkit Valley (Eastern Black Sea region of Turkey)," *Geomorphology*, vol. 94, no. 3-4, pp. 401–418, 2008.

[22] L. Ermini, F. Catani, and N. Casagli, "Artificial Neural Networks applied to landslide susceptibility assessment," *Geomorphology*, vol. 66, no. 1–4, pp. 327–343, 2005.

[23] S. Lee and I. Park, "Application of decision tree model for the ground subsidence hazard mapping near abandoned underground coal mines," *Journal of Environmental Management*, vol. 127, pp. 166–176, 2013.

[24] H. Gómez and T. Kavzoglu, "Assessment of shallow landslide susceptibility using artificial neural networks in Jabonosa River Basin, Venezuela," *Engineering Geology*, vol. 78, no. 1-2, pp. 11–27, 2005.

[25] C. Melchiorre, M. Matteucci, A. Azzoni, and A. Zanchi, "Artificial neural networks and cluster analysis in landslide susceptibility zonation," *Geomorphology*, vol. 94, no. 3-4, pp. 379–400, 2008.

[26] Y.-K. Yeon, J.-G. Han, and K. H. Ryu, "Landslide susceptibility mapping in Injae, Korea, using a decision tree," *Engineering Geology*, vol. 116, no. 3-4, pp. 274–283, 2010.

[27] R. Bou Kheir, J. Chorowicz, C. Abdallah, and D. Dhont, "Soil and bedrock distribution estimated from gully form and frequency: A GIS-based decision-tree model for Lebanon," *Geomorphology*, vol. 93, no. 3-4, pp. 482–492, 2008.

[28] N. J. Schneevoigt, S. van der Linden, H.-P. Thamm, and L. Schrott, "Detecting Alpine landforms from remotely sensed imagery. A pilot study in the Bavarian Alps," *Geomorphology*, vol. 93, no. 1-2, pp. 104–119, 2008.

[29] C.-S. Huang, Y.-J. Lin, and C.-C. Lin, "Implementation of classifiers for choosing insurance policy using decision trees: A case study," *WSEAS Transactions on Computers*, vol. 7, no. 10, pp. 1679–1689, 2008.

[30] M. Pal and P. M. Mather, "An assessment of the effectiveness of decision tree methods for land cover classification," *Remote Sensing of Environment*, vol. 86, no. 4, pp. 554–565, 2003.

[31] D. T. Bui, B. Pradhan, O. Lofman, and I. Revhaug, "Landslide susceptibility assessment in Vietnam using support vector machines, decision tree, and naïve bayes models," *Mathematical Problems in Engineering*, vol. 2012, Article ID 974638, 26 pages, 2012.

[32] P. K. Pang, L. T. Tien, and H. Lateh, "Landslide hazard mapping of penang island using decision tree model," in *Proceedings of the International Conference on Systems and Electronic Engineering (ICSEE '12)*, Phuket, Thailand, December 2012.

[33] B. Pradhan, "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*, vol. 51, pp. 350–365, 2013.

[34] M. Ture, F. Tokatli, and I. Kurt, "Using Kaplan-Meier analysis together with decision tree methods (C&RT, CHAID, QUEST, C4.5 and ID3) in determining recurrence-free survival of breast cancer patients," *Expert Systems with Applications*, vol. 36, no. 2, pp. 2017–2026, 2009.

[35] C. E. Brodley and M. A. Friedl, "Decision tree classification of land cover from remotely sensed data," *Remote Sensing of Environment*, vol. 61, no. 3, pp. 399–409, 1997.

[36] M. Xu, P. Watanachaturaporn, P. K. Varshney, and M. K. Arora, "Decision tree regression for soft classification of remote sensing data," *Remote Sensing of Environment*, vol. 97, no. 3, pp. 322–336, 2005.

[37] I. H. Witten and E. Frank, *Data Mining—Practical Machine Learning Tools and Techniques*, Elsevier, Amsterdam, The Netherlands, 2nd edition, 2005.

[38] R. J. Lewis, "An introduction to Classification and Regression Tree (CART) analysis," in *Proceedings of the Annual Meeting of the Society for Academic Emergenct Medicine*, San Francisco, Calif, USA, 2000.

[39] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, Wadsworth and Brooks, Montery, Calif, USA, 1984.

[40] J. A. Michael and S. L. Gordon, *Data Mining Technique: For Marketing, Sales and Customer Support*, Wiley, New York, NY, USA, 1997.

[41] D. B. V. Biggs and E. Suen, "A method of choosing multiway partitions for classification and decision trees," *Journal of Applied Statistics*, vol. 18, pp. 49–62, 1991.

[42] L. A. Goodman, "Simple models for the analysis of association in cross-classifications having ordered categories," *Journal of the American Statistical Association*, vol. 74, no. 367, pp. 537–552, 1979.

[43] G. Kass, "An exploratory technique for investigating large quantities of categorical data," *Applied Statistics*, vol. 29, no. 2, pp. 119–127, 1980.

[44] T. Hill and P. Lewicki, *Statistics: Methods and Applications, A Comprehensive Reference for Science, Industry, and Data Mining*, Stata Soft, USA, 2006.

[45] W.-Y. Loh and Y.-S. Shih, "Split selection methods for classification trees," *Statistica Sinica*, vol. 7, no. 4, pp. 815–840, 1997.

[46] H.-J. Oh and B. Pradhan, "Application of a neuro-fuzzy model to landslide-susceptibility mapping for shallow landslides in a tropical hilly area," *Computers and Geosciences*, vol. 37, no. 9, pp. 1264–1276, 2011.

[47] K. Lim Khai-Wern, T. Lea Tien, and H. Lateh, "Landslide hazard mapping of Penang island using probabilistic methods and logistic regression," in *Proceedings of the IEEE International Conference on Imaging Systems and Techniques (IST '11)*, pp. 273–278, May 2011.

[48] M. S. Alklhasawneh and U. K. Ngah, "Landslide susceptibility hazard mapping techniques review," *Journal of Applied Sciences*, vol. 12, pp. 802–808, 2012.