Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2012, Article ID 974638, 26 pages
doi:10.1155/2012/974638

Research Article

Landslide Susceptibility Assessment in Vietnam Using Support Vector Machines, Decision Tree, and Naïve Bayes Models

Dieu Tien Bui,[1,2] Biswajeet Pradhan,[3]
Owe Lofman,[1] and Inge Revhaug[1]

[1] *Department of Mathematical Sciences and Technology, Norwegian University of Life Sciences, P.O. Box 5003IMT, 1432 Aas, Norway*
[2] *Faculty of Surveying and Mapping, Hanoi University of Mining and Geology, Dong Ngac, Tu Liem, Hanoi, Vietnam*
[3] *Department of Civil Engineering, Spatial and Numerical Modelling Research Group, Faculty of Engineering, Universiti Putra Malaysia, Selangor, 43400 Serdang, Malaysia*

Correspondence should be addressed to
Dieu Tien Bui, buitiendieu@gmail.com, bui-tien.dieu@umb.no

Received 1 April 2012; Accepted 24 April 2012

Academic Editor: Wei-Chiang Hong

The objective of this study is to investigate and compare the results of three data mining approaches, the support vector machines (SVM), decision tree (DT), and Naïve Bayes (NB) models for spatial prediction of landslide hazards in the Hoa Binh province (Vietnam). First, a landslide inventory map showing the locations of 118 landslides was constructed from various sources. The landslide inventory was then randomly partitioned into 70% for training the models and 30% for the model validation. Second, ten landslide conditioning factors were selected (i.e., slope angle, slope aspect, relief amplitude, lithology, soil type, land use, distance to roads, distance to rivers, distance to faults, and rainfall). Using these factors, landslide susceptibility indexes were calculated using SVM, DT, and NB models. Finally, landslide locations that were not used in the training phase were used to validate and compare the landslide susceptibility maps. The validation results show that the models derived using SVM have the highest prediction capability. The model derived using DT has the lowest prediction capability. Compared to the logistic regression model, the prediction capability of the SVM models is slightly better. The prediction capability of the DT and NB models is lower.

1. Introduction

Vietnam is identified as a country that is particularly vulnerable to some of the worst manifestations of climate change such as sea level rise, flooding, and landslides. In the recent years, together with flooding, landslides have occurred widespread and recurrent in the northwest mountainous areas of Vietnam and have caused substantial economic losses and property damages. Landslides usually occurred during heavy rainfalls in the rainy season from May to October every year. In particular, in the Hoa Binh province during the rainy season of 2006 and 2007, large landslides occurred frequently due to heavy rainfalls. Most of these landslides occurred on cut slopes and alongside roads in mountainous areas. Landslide disaster can be reduced by understanding the mechanism, prediction, hazard assessment, early warning, and risk management [1]. Therefore, studies on landslides and determining measures to mitigate losses are an urgent task. However, the study on landslides in Vietnam is still limited except a few case studies [2–5]. Through scientific analyses of these landslides, we can assess and predict landslide prone areas, offering potential measures to decrease landslide damages [6, 7].

Spatial prediction of landslide hazard map preparation is considered the first important step for landslide hazard mitigation and management [8]. The spatial probability of landslide hazards can be expressed as the probability of spatial occurrence of slope failures with a set of geoenvironmental conditions [9]. However, due to the complex nature of landslides, producing a reliable spatial prediction of landslide hazard is not easy. For this reason, various approaches have been proposed in the literature. Review of these approaches has been carried out by Guzzetti et al. [10], Wang et al. [11], and Chacón et al. [12]. In the recent years, some soft computing approaches have been applied for landslide hazard evaluation including fuzzy logic [7, 13–20], neuro-fuzzy [3, 15, 21, 22], and artificial neural networks [6, 23–29]. In general, the quality of landslide susceptibility models is affected by the methods used [30]. For this reason, comparison of those methods with the conventional methods has been carried out using different datasets. Some researchers found that soft computing methods outperform the conventional methods [31–35]; however, other authors find no differences in overall predictive performance [36]. In general, soft computing approaches give rise qualitatively and quantitatively on the maps of the landslide hazard areas and the spatial results are appealing [37].

In more recent years, data mining approaches have been considered used for landslide studies such as SVM, DT, and NB [38, 39]. They belong to the top 10 data mining algorithms identified by the IEEE [40]. In the case of SVM, the main advantage of this method is that it can use large input data with fast learning capacity. This method is well-suited to nonlinear high-dimensional data modeling problems and provides promising perspectives in the landslide susceptibility mapping [41]. Micheletti et al. [42] stated that SVM methods can be used for landslide studies because of their ability in dealing with high-dimensional spaces effectively and with a high classification performance. In the case of DT, according to Yeon et al. [43] the probability of observations that belong to the landslide class can be used to estimate indexes of susceptibility. Saito et al. [44] used a decision tree model for landslide susceptibility mapping in the Akaishi Mountains (Japan) and stated that the decision tree model has appropriate accuracy for estimating the probabilities of future landslides. Nefeslioglu et al. [45] applied a DT in the metropolitan area of Istanbul (Turkey) with a good prediction accuracy of the landslide model. Yeon et al. [43] concluded that DT can be used efficiently for landslide susceptibility mapping. In the case of NB, although the method has been successfully applied in many domains [46]; however, the application in landslide susceptibility assessment may still be limited. NB is a popular and fast supervised learning algorithm for data mining applications based on the Bayes theorem. The main advantage of NB is that it can process a large number of variables, both discrete and continuous [47]. NB is suitable for large-scale prediction of complex and incomplete data [48]. The main potential drawback of this method is that it requires independence of attributes. However, this method is considered to be relatively robust [49].

The main objective of this study is to investigate and compare the results of three data mining approaches, that is, SVM, DT, and NB, to spatial prediction of landslide hazards for the Hoa Binh province (Vietnam). The main difference between this study and the aforementioned works is that SVM with two kernel functions (radial basis and polynomial kernels) and NB were applied for landslide susceptibility modeling. To assess these methods, the susceptibility maps obtained from the three data mining approaches were compared to those obtained by the logistic regression model reported by the same authors [2]. The computation process was carried out using MATLAB 7.11 and LIBSVM [50] for SVM and WEKA ver. 3.6.6 (The University of Waikato, 2011) for DT and NB.

2. Study Area and Data Used

2.1. Study Area

Hoa Binh has an area of about 4,660 km^2 and is located between the longitudes 104°48$'$E and 105°50$'$E and the latitudes 20°17$'$N and 21°08$'$N in the northwest mountainous area of Vietnam (Figure 1). The province is hilly with elevations ranging between 0 and 1,510 m, with an average value of 315 m and standard deviation of 271.5 m. The terrain gradient computed from a digital elevation model (DEM) with a spatial resolution of 20 × 20 m is in the range from 0° to 60°, with a mean value of 13.8° and a standard deviation of 10.4°.

There are more than 38 geologic formations that have cropped out in the province (Figure 2). Six geological formations, Dong Giao, Tan Lac, Vien Nam, Song Boi, Suoi Bang, and Ben Khe, cover about 72.8% of the total area. The main lithologies are limestone, conglomerate, aphyric basalt, sandstone, silty sandstone, and black clay shale. The ages of rocks vary from the Paleozoic to Cenozoic with different physical properties and chemical composition. Five major fracture zones pass through the province causing rock mass weakness: Hoa Binh, Da Bac, Muong La-Cho Bo, Son La-Bim Son, and Song Da.

The soil types are mainly ferralic acrisols, humic acrisols, rhodic ferralsols, and eutric fluvisols that account for 80% of the total study area. Land use is comprised of approximately 7.5% populated areas, 14.5% agricultural land, 52.6% forest land, 21% barren land and nontree rocky mountain, 0.4% grassland, and 4% water surface.

In the study area, there are heavy rainfalls with high intensity, especially during tropical rainstorms, and with an average annual precipitation varying from 1353 to 1857 mm (data shown for the period 1973–2002). The precipitation is most abundant during May to October with a rainfall that accounts for 84–90% annual precipitation. Rainfall usually peaks in the months of August and September with the average around 300 to 400 mm per month. The climate has a typical characteristic for the monsoonal region with a high humidity, being hot, and rainy. January is usually the coldest month with an average temperature of 14.9°C whereas the warmest month is July with an average temperature of 26.7°C.

Landslides occurred mostly in the rainy season when heavy rains exceeded 100 mm per day and continued for three days. Landslides also occurred when rainfall continued for five to seven days with rainfall larger than 100 mm for the last day. For example, landslides occurred in the Doc Cun and Doi Thai areas on September 2000 when the 7 days accumulated rainfalls were 308 and 383 mm, respectively. Many landslides occurred on 5 October 2007, in

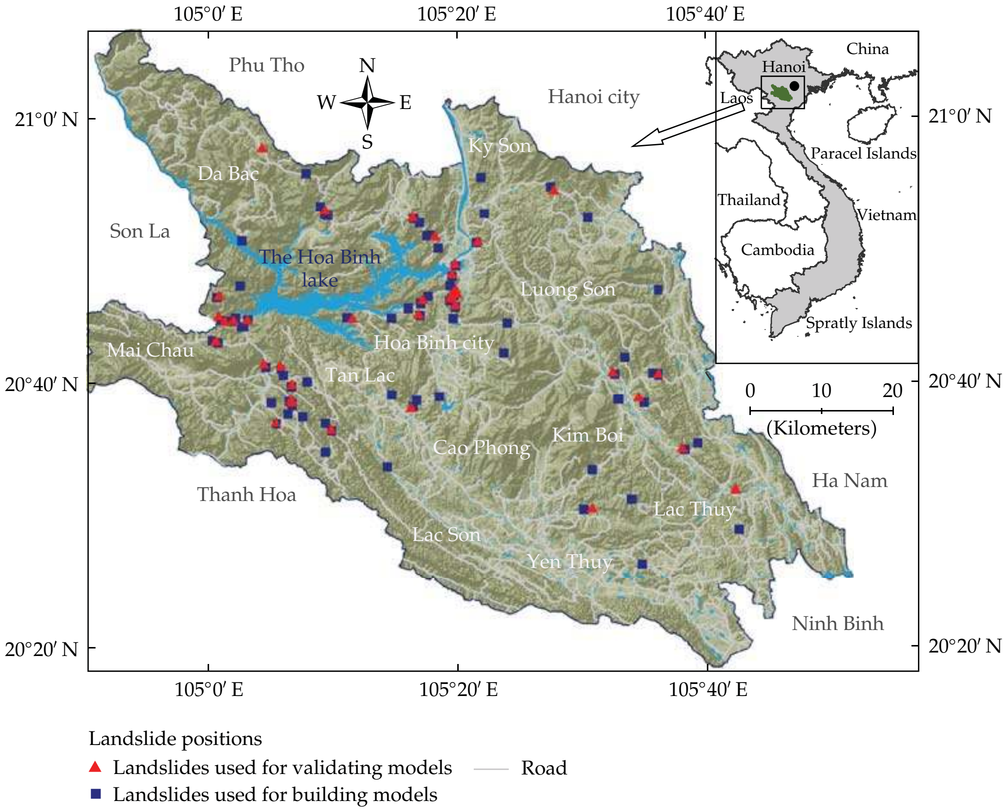


Figure 1: Landslide inventory map of the study area.

the Thung Khe, Toan Son, Phuc San, Tan Mai, Doc Cun, and surrounding areas with 3 days of accumulated rainfalls amounting from 334 to 529 mm.

2.2. Data

Landslides are assumed to occur in the future under the same conditions as for the past and current landslides [10]. Therefore, a landslide inventory map has been considered to be the most important factor for prediction of future landslides. The landslide inventory map portrays the spatial distribution of a single landslide event (a single trigger) or multiple landslide events over time (historical) [51]. For the study area, the landslide inventory map (Figure 1) constructed by Tien Bui et al. [2] was used to analyze the relationships between landslide occurrence and landslide conditioning factors. The map shows 118 landslides that occurred during the last ten years, including 97 landslide polygons and 21 rock fall locations. The size of the largest landslide is 3,440 m^2, the smallest is 380 m^2, and the average landslide size is 3,440 m^2.

Based on previous research carried out by Tien Bui et al. [2], ten landslide conditioning factors are selected to build landslide models and to predict spatial distribution of the landslides in this study. They are slope angle, slope aspect, relief amplitude, lithology, soil type, land use, distance to roads, distance to rivers, distance to faults, and rainfall.

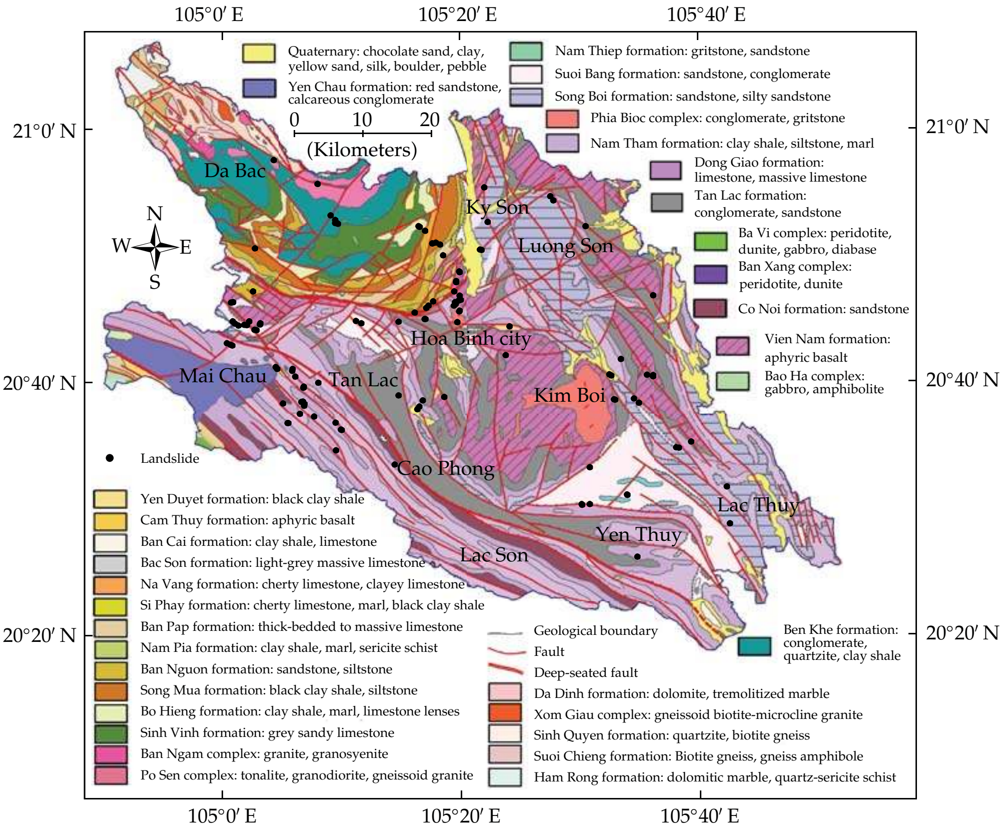

105°0′ E
105°20′ E
105°40′ E
21°0′ N
20°40′ N
20°20′ N
Quaternary: chocolate sand, clay, yellow sand, silk, boulder, pebble
Yen Chau formation: red sandstone, calcareous conglomerate
0 10 20
(Kilometers)
Nam Thiep formation: gritstone, sandstone
Suoi Bang formation: sandstone, conglomerate
Song Boi formation: sandstone, silty sandstone
Phia Bioc complex: conglomerate, gritstone
Nam Tham formation: clay shale, siltstone, marl
Dong Giao formation: limestone, massive limestone
Tan Lac formation: conglomerate, sandstone
Ba Vi complex: peridotite, dunite, gabbro, diabase
Ban Xang complex: peridotite, dunite
Co Noi formation: sandstone
Vien Nam formation: aphyric basalt
Bao Ha complex: gabbro, amphibolite
N
W
E
S
Da Bac
Ky Son
Luong Son
Hoa Binh city
Mai Chau
Tan Lac
Kim Boi
Cao Phong
Lac Thuy
Yen Thuy
Lac Son
Landslide
Yen Duyet formation: black clay shale
Cam Thuy formation: aphyric basalt
Ban Cai formation: clay shale, limestone
Bac Son formation: light-grey massive limestone
Na Vang formation: cherty limestone, clayey limestone
Si Phay formation: cherty limestone, marl, black clay shale
Ban Pap formation: thick-bedded to massive limestone
Nam Pia formation: clay shale, marl, sericite schist
Ban Nguon formation: sandstone, siltstone
Song Mua formation: black clay shale, siltstone
Bo Hieng formation: clay shale, marl, limestone lenses
Sinh Vinh formation: grey sandy limestone
Ban Ngam complex: granite, granosyenite
Po Sen complex: tonalite, granodiorite, gneissoid granite
Geological boundary
Fault
Deep-seated fault
Ben Khe formation: conglomerate, quartzite, clay shale
Da Dinh formation: dolomite, tremolitized marble
Xom Giau complex: gneissoid biotite-microcline granite
Sinh Quyen formation: quartzite, biotite gneiss
Suoi Chieng formation: Biotite gneiss, gneiss amphibole
Ham Rong formation: dolomitic marble, quartz-sericite schist

Figure 2: Geologic map of the study area.

The slope angle, slope aspect, and relief amplitude were extracted from a DEM that was generated from national topographic maps at the scale of 1 : 25,000. The slope angle map with 6 categories was constructed (Figure 3(a)). The slope aspect map with nine layer classes was constructed: flat, north, northeast, east, southeast, south, southwest, west, and northwest. The relief amplitude that presents the maximum difference in height per unit area [52] was constructed with 6 categories: 0–50 m, 50–100 m, 100–150 m, 150–200 m, 200–250 m, and 250–532 m. For the construction of the relief amplitude map, different sizes of the unit area were tested to choose a best one (20 × 20 pixels) using the focal statistic module in the ArcGIS 10 software.

The lithology and faults were extracted from four tiles of the Geological and Mineral Resources Map of Vietnam at the scale of 1 : 200,000. This is the only geological map available for the study area. The lithology map (Figure 3(b)) was constructed with seven groups based on clay composition, degree of weathering, estimated strength, and density [53, 54]. The distance-to-faults map was constructed by buffering the fault lines with 5 categories as: 0–200 m, 200–400 m, 400–700 m, 700–1,000 m, and >1,000 m. The soil type map (Figure 3(c)) was constructed with 13 categories. The land-use map (Figure 3(d)) was constructed with twelve categories.

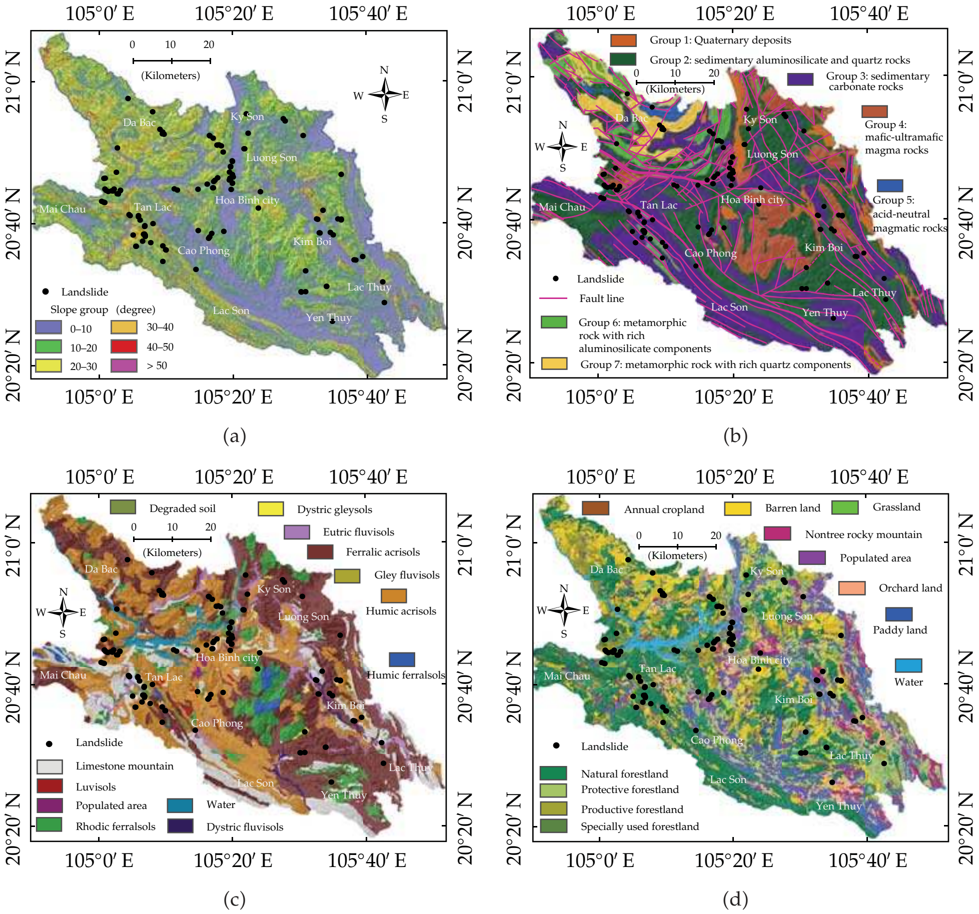


Figure 3: Landslide conditioning factor maps (a) slope, (b) lithology, (c) soil type, and (d) landuse.

A road network that undercut slopes was extracted from the topographic map at the scale of 1 : 50,000. A distance-to-roads map was constructed with 4 categories: 0–40 m, 40–80 m, 80–120 m, and >120 m. A hydrological network that undercut slopes was also extracted from the topographic map at the scale of 1 : 50,000. And then a distance-to-rivers map was constructed with 4 categories: 0–40 m, 40–80 m, 80–120 m, and >120 m.

The rainfall map was prepared using the value of maximum rainfall of eight days (seven rainfall days plus last day of rainfall larger than 100 mm) for the period from 1990 to 2010, using the Inverse Distance Weighed (IDW) method. The precipitation data was extracted from a database from the Institute of Meteorology and Hydrology in Vietnam.

3. Landslide Susceptibility Mapping Using SVM, DT, and NB Models

3.1. Support Vector Machines (SVM)

Support vector machines are a relatively new supervised learning method based on statistical learning theory and the structural risk minimization principle [55]. Using the training data, SVM implicitly maps the original input space into a high-dimensional feature space. Subsequently, in the feature space the optimal hyper plane is determined by maximizing the margins of class boundaries [56]. The training points that are closest to the optimal hyper plane are called support vectors. Once the decision surface is obtained, it can be used for classifying new data.

Consider a training dataset of instance-label pairs $(\mathbf{x}_i, \mathbf{y}_i)$ with $\mathbf{x}_i \in R^n$, $\mathbf{y}_i \in \{1, -1\}$, and $i = 1, \ldots, m$. In the current context of landslide susceptibility, $\mathbf{x}$ is a vector of input space that contains slope angle, lithology, rainfall, soil type, slope aspect, land use, distance to roads, distance to rivers, distance to faults, and relief amplitude. The two classes $\{1, -1\}$ denote landslide pixels and no-landslide pixels. The aim of the SVM classification is to find an optimal separating hyperplane that can distinguish the two classes, that is, landslides and no landslides $\{1, -1\}$, from the mentioned set of training data.

For the case of linear separable data, a separating hyperplane can be defined as

$$\mathbf{y}_i(\mathbf{w} \cdot \mathbf{x}_i + \mathbf{b}) \geq 1 - \xi_i, \tag{3.1}$$

where $\mathbf{w}$ is a coefficient vector that determines the orientation of the hyper plane in the feature space, $\mathbf{b}$ is the offset of the hyper plane from the origin, and ξ_i is the positive slack variables [57].

The determination of an optimal hyper plane leads to the solving of the following optimization problem using Lagrangian multipliers [58]:

$$\begin{gathered}
\text{Minimize} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \mathbf{y}_i \mathbf{y}_j (\mathbf{x}_i \mathbf{x}_j), \\
\text{Subject to} \sum_{i=1}^{n} \alpha_i \mathbf{y}_j = 0, \qquad 0 \leq \alpha_i \leq C,
\end{gathered} \tag{3.2}$$

where α_i are Lagrange multipliers, C is the penalty, and the slack variables ξ_i allows for penalized constraint violation.

The decision function, which will be used for the classification of new data, can then be written as

$$g(x) = \operatorname{sign}\left(\sum_{i=1}^{n} \mathbf{y}_i \alpha_i \mathbf{x}_i + \mathbf{b} \right). \tag{3.3}$$

In cases when it is impossible to find the separating hyper plane using the linear kernel function, the original input data may be transferred into a high-dimension feature space through some nonlinear kernel functions. The classification decision function is then written as

$$g(x) = \operatorname{sign}\left(\sum_{i=1}^{n} \mathbf{y}_i \alpha_i K(\mathbf{x}_i, \mathbf{x}_j) + \mathbf{b} \right), \tag{3.4}$$

where $K(\mathbf{x}_i, \mathbf{x}_j)$ is the kernel function.

The choice of the kernel function is crucial for successful SVM training and classification accuracy [59]. There are four types of kernel function groups that are commonly used in SVM: linear kernel (LN), polynomial kernel (PL), radial basis function (RBF) kernel, and sigmoid kernel (SIG). The LN is considered to be a specific case of RBF, whereas the SIG behaves like the RBF for certain parameters [60]. According to Keerthi and Lin [61], the LN is not needed for use when the RBF is used. And generally, the classification accuracy of the SIG may not be better than RBF [62]. Therefore in this study, only the two kernel functions, RBF and PL, were selected. According to Zhu et al. [63], the main advantage of using RBF is that RBF has good interpolation abilities. However, it may fail to provide longer-range extrapolation. On contrast, PL has better extrapolation abilities at lower-order degrees but requires higher order degrees for good interpolation. The formulas and their parameters are shown in Table 2.

The performance of the SVM model depends on the choice of the kernel parameters. For the RBF-SVM, the regularization parameter (C) and the kernel width (γ) are the two parameters that need to be determined, whereas C, γ and the degree of polynomial kernel (d) are three for the case of the PL-SVM. Parameter C controls the tradeoff between training errors and margin, which helps to control overfitting of the model. If values of C are large, that will lead to a few training errors, whereas a small value for C will generate a larger margin and thus increase the number of training errors [64]. Parameter γ controls the degree of nonlinearity of the SVM model. Parameter d defines the degree of the polynomial kernel.

The process of picking up the best pairs of parameters, which produce the best classification result, is considered to be an important research issue in the data mining area [65]. Many methods have been proposed, such as the heuristic parameter selection [66], the gradient descent algorithm [67], the Levenberg-Marquardt method [68], and the cross-validation method [69]. However, the grid search method that is widely used in the determination of SVM parameters is still considered to be the most reliable optimization method [70] and was selected for this study. Firstly, the ranges of all parameters with a step-size process were determined. Secondly, the grid search was performed by varying the SVM hyperparameters. Finally, the performance of every combination is assessed to find the best pairs of parameters. However, the grid search is only suitable for the adjustment of a small number of parameters due to the computational complexity [71].

3.2. Decision Tree (DT)

A DT is a hierarchical model composed of decision rules that recursively split independent variables into homogeneous zones [72]. The objective of DT building is to find the set of decision rules that can be used to predict outcome from a set of input variables. A DT is called a classification or a regression tree if the target variables are discrete or continuous, respectively [73]. DT has been applied successfully in many real-world situations for classification and prediction [74].

The main advantage of DT is that DT models have the capability of modeling complex relationship between variables. They can incorporate both categorical and continuous variables without strict assumptions with respect to the distribution of the data [75]. In addition, DTs are easy to construct and the resulting models can be easily interpreted. Furthermore, the DT model results provide clear information on the relative importance of input factors [76]. The main disadvantage of DTs is that they are susceptible to noisy data and that multiple output attributes are not allowed [77].

Many algorithms for constructing decision tree models such as classification and regression tree (CART) [78], chi-square automatic interaction detector decision tree (CHAID) [79], ID3 [80], and C4.5 [81] are proposed in the literature. In this study, the J48 algorithm [82], which is a Java reimplementation of the C4.5 algorithm, was used. The C4.5 uses an entropy-based measure as the selection criteria that is considered to be the fastest algorithm for machine learning with good classification accuracy [83]. Given a training dataset T with subsets $T_i, i = 1, 2, \ldots, s$, the C4.5 algorithm constructs a DT using the top-down and recursive-splitting technique. A tree structure consists of a root node, internal nodes, and leaf nodes. The root node contains all the input data. An internal node can have two or more branches and is associated with a decision function. A leaf node indicates the output of a given input vector.

The procedure of DT modeling consists of two steps: (1) tree building and (2) tree pruning [84]. The tree building begins by determining the input variable with highest gain ratio as the root node of the DT. Then the training dataset is split based on the root values, and subnodes are created. For discrete input variables, a subnode of the tree is created for each possible value. For continuous input variables, two sub-nodes are created based on a threshold that was determined in the threshold-finding process [81]. In the next step, the gain ratio is calculated for all the sub-nodes individually, and the process is subsequently repeated until all examples in a node belong to the same class. And those nodes are called leaf nodes and are labeled as class values.

Since the tree obtained in the building step may have a large number of branches and therefore may cause a problem of over-fitting [85], therefore, the tree needs to be pruned for better classification accuracy for new data. Two types of tree pruning can be seen: before pruning and after pruning. In the case of pre-pruning, the growing of the tree will be stopped when a certain criterion is satisfied, whereas in the post-pruning case the full tree will be constructed first, and then the ending subtrees will be replaced by leafs based on the error comparison of the tree before and after replacing sub-trees.

The information gain ratio for attribute A is as follows:

$$\text{GainRatio}(A, T) = \frac{\text{Gain}(A, T)}{\text{SplitInfo}(A, T)}, \tag{3.5}$$

where

$$\begin{gathered} \text{Gain}(A, T) = \text{Entropy}(T) - \sum_{i=1}^{s} \frac{|T_i|}{|T|} \text{Entropy}(T_i), \\ \text{SplitInfo}(A) = -\sum_{i=1}^{s} \frac{|T_i|}{|T|} \log_2 \frac{|T_i|}{|T|}. \end{gathered} \tag{3.6}$$

A DT can estimate the probability of belonging to a specific class and therefore the probability isused to predict the probability of landslide pixels. The estimated probability is based on a natural frequency at the tree leaf. However, the estimated probability might not give sound probabilistic estimates; therefore Laplace smoothing [86] was used in this study.

3.3. Naïve Bayes (NB)

An NB classifier is a classification system based on Bayes' theorem that assumes that all the attributes are fully independent given the output class, called the conditional independence assumption [48]. The main advantage of the NB classifier is that it is very easy to construct without needing any complicated iterative parameter estimation schemes [40]. In addition, NB classifier is robust to noise and irrelevant attribute. This method has been successfully applied in many fields [87].

Given an observation consisting of k attributes $\mathbf{x}_i, i = 1, 2, \ldots, k$ ($\mathbf{x}_i$ is landslide conditioning factor), $\mathbf{y}_j, j = \text{landslide, no landslide}$ is the output class. NB estimates the probability $P(\mathbf{y}_j/\mathbf{x}_i)$ for all possible output class. The prediction is made for the class with the largest posterior probability as

$$\mathbf{y}_{\text{NB}} = \underset{\mathbf{y}_j \in \{\text{Landslide, no-landslide}\}}{\operatorname{argmax}} P(\mathbf{y}_j) \prod_{i=1}^{n} P(\mathbf{x}_i/\mathbf{y}_j). \tag{3.7}$$

The prior probability $P(\mathbf{y}_j)$ can be estimated using the proportion of the observations with output class $\mathbf{y}_j$ in the training dataset. The conditional probability is calculated using

$$P\left(\frac{\mathbf{x}_i}{\mathbf{y}_j}\right) = \frac{1}{\sqrt{2\pi}\delta} e^{-(\mathbf{x}_i - \mu)^2/2\delta^2}, \tag{3.8}$$

where μ is mean and δ is standard deviation of x_i.

3.4. Performance Evaluation

The performances of the trained landslide models were assessed using several statistical evaluation criteria using counts of true positive (TP), false positive (FP), true negative (TN), false negative (FN).

TP rate (sensitivity) measures the proportion of the number of pixels that are correctly classified as landslides and is defined as TP/(TP + FN). TN rate (specificity) measures the proportion of number of pixels that are correctly classified as non-landslide and is defined as TN/(TN + FP). Precision measures the proportion of the number of pixels that are correctly classified as landslide occurrences and is defined as TP/(TP + FP). Overall accuracy is calculated as (TP + TN)/total number of training pixels. The F-measure combines precision and sensitivity into their harmonic mean and is defined as 2 $*$ Sensitivity $*$ Specificity/(Sensitivity + Specificity) [88].

In order to measure the reliability of the landslide susceptibility models, the Cohen kappa index (κ) [89–91] was used to assess the model classification compared to chance selection:

$$\kappa = \frac{P_C - P_{\text{exp}}}{1 - P_{\text{exp}}}, \tag{3.9}$$

where P_C is the proportion of number of pixels that are correctly classified as landslide or non-landslide and is calculated as (TP + TN)/total number of pixels. P_{exp} is the expected agreements and is calculated as ((TP + FN)(TP + FP) + (FP+TN)(FN+TN))/Sqrt(total number of training pixels).

A κ value of 0 indicates that no agreement exists between the landslide model and reality whereas a κ value of 1 indicates a perfect agreement. If κ value is negative, it indicates a poor agreement. A κ value in the range (0.80–1) is considered as indicator of almost perfect agreement while a value in the range (0.60–0.80) indicates a substantial agreement between the model and reality. For a value in the interval (0.40–0.60), the agreement is moderate and the values of (0.20–0.40) and <0.2 indicate over fair and slight agreement, respectively [92].

3.5. Preparation of the Training and the Validation Datasets

In this study, a total of ten landslide conditioning factors were used. They are slope angle, lithology, rainfall, soil type, slope aspect, landuse, distance to roads, distance to rivers, distance to faults, and relief amplitude. For each conditioning factor, a map is generated. These maps were then converted into a pixel format with a spatial resolution of 20 × 20 m. In the next step, frequency ratio values [93] were calculated for all categories based on the landslide grid cells. Based on these ratio values, each category was assigned an attribute number and then was rescaled in the range 0.1 to 0.9 (Table 1) using the Max-Min normalization procedure [94] as follows:

$$v' = \frac{v - \mathrm{Min}(v)}{\mathrm{Max}(v) - \mathrm{Min}(v)}(U - L) + L, \tag{3.10}$$

where v' is the normalized data matrix, v is the original data matrix, and U and L are the upper and lower normalization boundaries.

In landslide modeling, the landslide data should be split into two parts, training and validation datasets. Without the splitting, it would not be possible to validate the results [95]. In this study, the landslide inventory map with 118 landslide polygons was randomly split into two subsets: subset 1 comprised 70% of the data (82 landslides with 684 landslide grid cells) and was used in the training phase of landslide models; subset 2 is a validation dataset with 30% of the data (36 landslides with 315 landslide grid cells) for the validation and estimate the prediction accuracy of the resulted models.

All of the 684 landslide grid cells in the subset 1 were assigned the value of 1. SVM may seriously have negative effects on the model performance when the numbers of landslide and non-landslide grid cells in the training dataset are significantly unbalanced. Therefore, the same amount of no-landslide grid cells was randomly sampled from the landslide-free area and assigned the value of −1. In the cases of DT and NB classifiers, no-landslide grid cells were assigned to the value 0. Finally, an extracting process was conducted to extract values for the ten landslide conditioning factors to build a training dataset. This dataset contains a total of 1368 observations, ten input variables, and one target variable (landslide, no landslide).

3.6. Training of the Support Vector Machines, Decision Tree and Naïve Bayes Models and Generation of Landslide Susceptibility Indexes

3.6.1. Support Vector Machines (SVM)

In the case of SVM, the model selection with its optimal parameters searching plays a crucial role in the performance of the model. In this study, RBF and PL kernel functions were selected.

Table 1: Normalized classes of landslide conditioning factors used.

Data layers	Class	Class pixels (%)	Landslide pixels (%)	Frequency ratio	Attribute	Normalized classes
Slope angle (°)	0–10	42.82	0.20	0.005	2	0.26
	10–20	29.13	29.93	1.028	4	0.58
	20–30	20.25	54.75	2.704	5	0.74
	30–40	6.84	14.31	2.094	6	0.90
	40–50	0.93	0.80	0.862	3	0.42
	>50	0.04	0.00	0.000	1	0.10
Slope aspect	Flat (−1)	0.06	0.00	0.000	1	0.10
	North (0–22.5 and 337.5–360)	12.02	4.70	0.391	2	0.20
	Northeast (22.5–67.5)	14.56	11.81	0.811	6	0.60
	East (67.5–112.5)	12.06	7.81	0.648	5	0.50
	Southeast (112.5–157.5)	12.04	14.51	1.206	7	0.70
	South (157.5–202.5)	12.90	22.72	1.761	8	0.80
	Southwest (202.5–247.5)	14.60	26.33	1.804	9	0.90
	West (247.5–292.5)	11.31	7.11	0.628	4	0.40
	Northwest (292.5–337.5)	10.46	5.01	0.478	3	0.30
Relief amplitude (m)	0–50	27.00	1.10	0.041	1	0.10
	50–100	23.97	25.43	1.061	3	0.42
	100–150	22.98	41.04	1.786	6	0.90
	150–200	14.75	20.12	1.364	5	0.74
	200–250	7.06	8.41	1.190	4	0.58
	250–532	4.24	3.90	0.920	2	0.26
Lithology	Group 1	4.08	6.31	1.546	6	0.77
	Group 2	39.62	33.43	0.844	4	0.50
	Group 3	32.55	27.13	0.833	3	0.37
	Group 4	11.65	21.62	1.856	7	0.90
	Group 5	1.18	0.00	0.000	1	0.10
	Group 6	5.62	7.81	1.389	5	0.63
	Group 7	5.29	3.70	0.700	2	0.23
Land use	Populated area	7.53	14.01	1.862	10	0.75
	Orchard land	3.71	2.50	0.674	7	0.54
	Paddy land	9.17	4.10	0.448	5	0.39
	Protective forestland	8.58	20.32	2.368	12	0.90
	Natural forestland	31.91	15.62	0.489	6	0.46
	Productive forestland	11.72	22.62	1.930	11	0.83
	Water	3.97	1.00	0.252	4	0.32
	Annual crop land	1.60	0.20	0.125	3	0.25
	Nontree rocky mountain	4.08	7.21	1.767	9	0.68

Table 1: Continued.

Data layers	Class	Class pixels (%)	Landslide pixels (%)	Frequency ratio	Attribute	Normalized classes
	Barren land	16.95	12.41	0.732	8	0.61
	Specially used forestland	0.36	0.00	0.000	2	0.17
	Grass land	0.43	0.00	0.000	1	0.10
Soil type	Eutric fluvisols	3.49	6.11	1.751	12	0.83
	Degraded soil	0.03	0.00	0.000	3	0.23
	Limestone mountain	14.42	15.12	1.048	9	0.63
	Ferralic acrisols	36.53	43.84	1.200	10	0.70
	Rhodic ferralsols	8.97	3.40	0.379	7	0.50
	Humic acrisols	30.91	28.13	0.910	8	0.57
	Dystric fluvisols	0.73	2.80	3.828	13	0.90
	Dystric gleysols	0.39	0.60	1.524	11	0.77
	Luvisols	0.46	0.00	0.000	4	0.30
	Humic ferralsols	1.15	0.00	0.000	5	0.37
	Populated area	0.44	0.00	0.000	2	0.17
	Water	2.41	0.00	0.000	1	0.10
	Gley fluvisols	0.08	0.00	0.000	6	0.43
Rainfall (mm)	362–470	22.48	27.23	1.211	3	0.63
	470–540	46.40	35.84	0.772	2	0.37
	540– 610	22.18	9.01	0.406	1	0.10
	610–950	8.94	27.93	3.125	4	0.90
Distance to roads (m)	0–40	1.40	41.64	29.755	4	0.90
	40–80	1.68	21.52	12.788	3	0.63
	80–120	1.88	4.70	2.509	2	0.37
	>120	95.04	32.13	0.338	1	0.10
Distance to rivers (m)	0–40	3.86	14.41	3.731	4	0.90
	40–80	4.52	12.41	2.747	3	0.63
	80–120	4.82	8.31	1.725	2	0.37
	>120	86.80	64.86	0.747	1	0.10
Distance to faults (m)	0–200	18.09	24.02	1.328	5	0.90
	200–400	15.95	11.61	0.728	2	0.30
	400–700	19.89	24.22	1.218	3	0.50
	700–1,000	14.31	18.42	1.287	4	0.70
	>1,000	31.75	21.72	0.684	1	0.10

Table 2: RBF and PL kernels and their parameters.

Kernel function	Formula	Kernel parameters
RBF	$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2)$	γ
PL	$K(\mathbf{x}_i, \mathbf{x}_j) = (\gamma \mathbf{x}_i^T \mathbf{x}_j + 1)^d$	γ, d

Table 3: Degree of polynomial kernel versus area under the ROC curves in the training and validation datasets.

Degree of polynomial kernel	AUC	
	Training dataset	Validation dataset
1	0.9432	0.9524
2	0.9489	0.9560
3	**0.9575**	**0.9566**
4	0.9643	0.9556
5	0.9717	0.9435
6	0.9827	0.9046
7	0.9905	0.8767
8	0.9946	0.8314
9	0.9985	0.8067
10	0.9996	0.8133

The training process was started by searching the optimal kernel parameters using the grid-search method with cross-validation that can help to prevent overfitting. Since the numbers of landslide grid cells in the study area are not large, 5-fold cross-validation was used to find the best kernel parameters. The training dataset was randomly split into 5 equally sized subsets. Each subset was used as a test dataset for the SVM model trained on the remaining 4 data subsets. The cross-validation process was then repeated five times with each of the five subsets used once as the test dataset.

With the RBF kernel, the two kernel parameters of C and γ need to be determined. The procedure is as follows: (1) we set a grid space of (C, γ), where $C = 2^{-5}, 2^{-4}, \ldots, 2^{10}$ and $\gamma = 2^{10}, 2^{9}, \ldots, 2^{-4}$; (2) for each parameter, pairs of (C, γ) in the grid space, conduct 5-fold cross-validation on the training dataset; (3) choose parameter pairs of (C, γ) that have the highest classification accuracy; (4) use the best parameters to construct a SVM model for landslide prediction of new data. The best C and γ are determined as 8 and 0.25, respectively. The correctly classified rate is 91.1%.

With the PL kernel, the two kernel parameters of C and d need to be determined. Table 3 shows the results of training the SVM model using different d values. The result shows that when the values of d increase, AUC in the training dataset is increased as well. However, AUC in the validation dataset increases until d equals 3 and then decreases with the increasing of the d values. And therefore, the SVM model with three degrees of the polynomial kernel is selected. The accurately classified rate of SVM using PL kernel is 91.1%. The best C and γ are determined as 1 and 0.3536, respectively.

A detailed accuracy assessment for RBF-SVM and PL-SVM is shown in Tables 4 and 5. It could be seen that precision, F-measure, and TP rate are high (>90%) whereas FP rate is low (<10%). It indicates a high classification capacity for the training dataset for the two models. The Cohen kappa indexes are 0.822 and 0.823 for RBF-SVM and PL-SVM, respectively. It indicates a good agreement between the observed and the predicted values.

3.6.2. Decision Tree (DT)

In the case of DT, the first step is to determine the optimal value of the algorithm parameter such as the minimum number of instances (MNIs) per leaf and the confidence factor (CF).

Table 4: Detailed accuracy assessment by classes of RBF-SVM, PL-SVM, DT, and NB models.

Model	TP rate (%)	FP rate (%)	Precision (%)	*F*-measure (%)	Class
RBF-SVM	90.4	8.2	91.7	91.0	Landslide
	91.8	9.6	90.5	91.1	No landslide
PL-SVM	90.2	7.9	92.0	91.1	Landslide
	92.1	9.8	90.4	91.2	No landslide
DT	95.5	9.5	90.9	93.2	Landslide
	90.5	4.5	95.2	92.8	No landslide
NB	83.2	11.0	88.4	85.7	Landslide
	89.0	16.8	84.1	86.5	No landslide

Since a lower MNI is required to a leaf tree, the more branching will be created resulting in a larger tree. And thus, it may cause overfitting problem. In contrast, a higher MNI required per leaf will result in a narrow tree.

Figure 4 shows the MNI required per leaf versus the classification accuracy. In this test, the MNI required in a leaf was varied from 1 to 25 with a step of one, and the corresponding classification accuracies were obtained and plotted. The result shows that the highest classification accuracy is 92.8% corresponding to a MNI of 6. Therefore, the MNI per leaf of 6 was selected.

In order to explore the effect of the CF on the classification accuracy, the CF value was varied from 0.1 to 1 using a step size of 0.05. The corresponding classification accuracy was calculated. The result is shown in Figure 5. The result shows that the highest classification accuracy occurred with the CF of 0.35. Therefore CF of 0.35 was selected. With the two aforementioned parameters being determined, the decision tree model was constructed using the J48 algorithm. The probability of belonging to the landslide or the no-landslide classes for each observation was estimated using the Laplace smoothing. Using10-fold cross-validation, the decision tree model was constructed. The classified rate is 92.9%. The Cohen kappa index is 0.860. Detailed accuracy assessment of the decision tree model by class is shown in Tables 4 and 5. It could be observed that the TP rate, the precision, and the *F*-measure are greater than 90%. FP rates are 9.5% and 4.5% for the landslide and the non-landslide classes, respectively.

Figure 6 depicts the inferred DT model for landslide susceptibility assessment in this study. It could be observed that the size of the tree is 55 including the root node, 26 internal nodes, and 28 leafs (green rectangular boxes). In leaf nodes, value of 0.1 indicates the class of no landslide, whereas value of 0.9 indicates the landslide class. The number in the parentheses at each leaf node represents the number of instances in that leaf. It is clear that some instances are misclassified in some leaves. The number of misclassified instances is specified after a slash (Figure 6). The highest number of instances in a leaf node is 288, whereas the lowest number of instances in a leaf node is 7. The top-down induction of the tree shows that landslide conditioning factor in the higher level of the tree is more important. The relative importance of the landslide conditioning factor is as follows: distance to roads (81.5% in relative importance), slope (71.6%), land use (66.7%), aspect (61.1%), rainfall (61.5%), relief amplitude (61.6%), distance to rivers (60.1%), distance to faults (58.7%), lithology (57.7%), and soil type (52.8%).

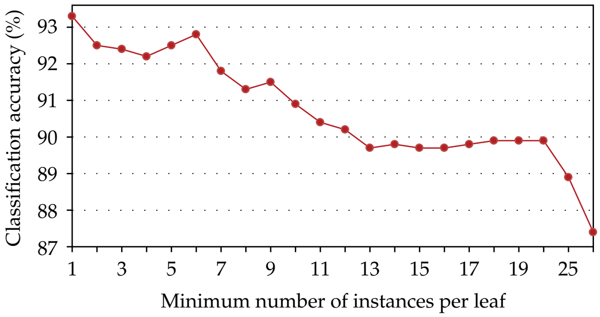


Figure 4: Minimum number of instances per leaf versus classification accuracy.

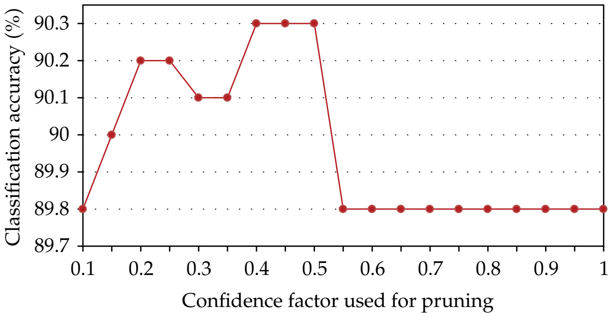


Figure 5: Confidence factor used for pruning versus classification accuracy.

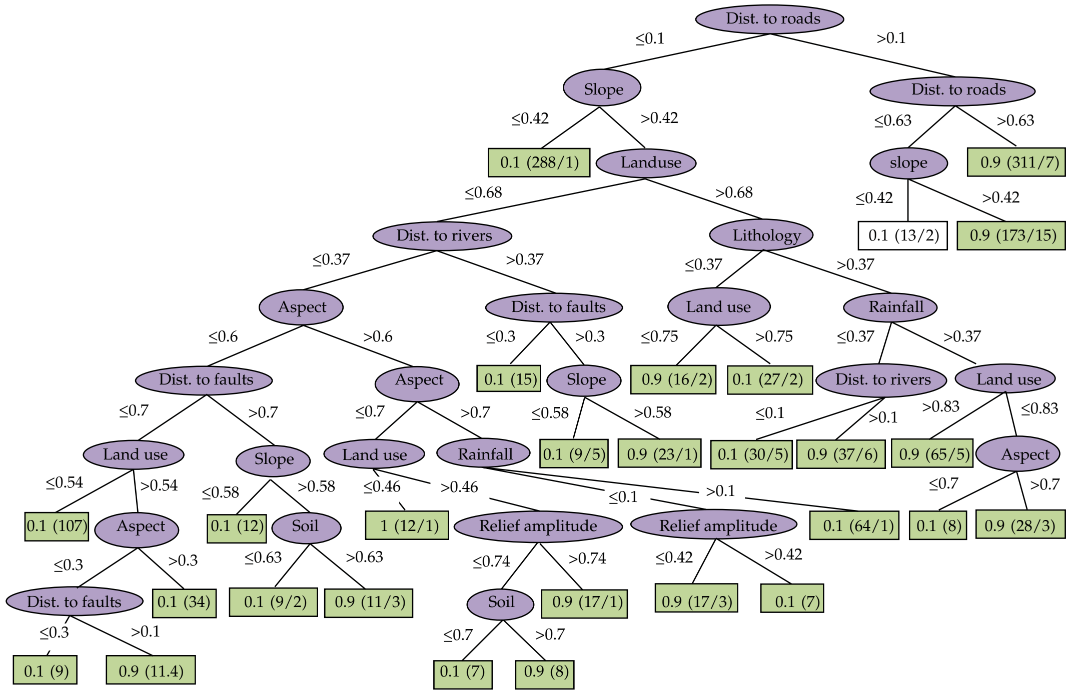


Figure 6: Decision tree model for landslide susceptibility assessment for the study area.

Table 5: Performance evaluation of RBF-SVM, PL-SVM, DT, and NB models.

Parameters	RBF-SVM	PL-SVM	DT	NB
Accuracy (%)	91.08	91.15	92.98	86.11
Cohen's kappa index	0.822	0.823	0.860	0.722

3.6.3. *Naïve Bayes (NB)*

In the case of NB classifier, the probability is first calculated for each output class (landslide, no landslide), and the classification is then made for the class with the largest posterior probability. The NB model was constructed using the WEKA software. The NB model obtained an overall classification accuracy of 86.1% in average. TP rate, precision, and *F*-measure are varied from 83% to 89%. The Cohen kappa index of 0.722 indicates that the strength of agreements between the observed and the predicted values is substantial. A summary result of the model assessment and performance is shown in Tables 4 and 5.

Once the SVM, DT, and NB models were successfully trained in the training phase, they were used to calculate the landslide susceptibility indexes (*LSIs*) for all the pixels in the study area. The results were then transferred into a GIS and loaded in the ARCGIS 10 software for visualization.

4. Validation and Comparison of Landslide Susceptibility Models

4.1. Success Rate and Prediction Rate for Landslide Susceptibility Maps

The validation processes of the four landslide susceptibility maps were performed by comparing them with the landslide locations using the success-rate and prediction-rate methods [95]. Using the landslide grid cells in the training dataset, the success-rate results were obtained. Figure 7 shows the success-rate curves of the four landslide susceptibility maps (obtained from RBF-SVM, PL-SVM, DT, NB models) in this study in comparison with the logistic regression model. It could be observed that RBF-SVM and logistic regression have the highest area under the curve, with AUC values of 0.961 and 0.962, respectively. They are followed by PL-SVM (0.956), DT (0.952), and NB (0.935). Based on these results we can conclude that the capability of correctly classifying the areas with existing landslides is highest for the RBF-SVM (equals to logistic regression), followed by the PL-SVM, DT, and NB.

Since the success-rate method uses the landslide pixels in the training dataset that have already been used for constructing the landslide models, the success-rate may not be a suitable method for measuring the prediction capability of the landslide models [96]. According to Chung and Fabbri [95], the prediction rate could be used to estimate the prediction capability of the landslide models. In this study, the prediction-rate results of the four landslide susceptibility models were obtained by comparing them with the landslide grid cells in the validation dataset. And then the areas under the prediction-rate curves (AUCs) were further estimated. The more the AUC value is close to 1, the better the landslide model.

The prediction-rate curves and AUC of the four landslide susceptibility maps are shown in Figure 8. The results show that AUCs for the four models vary from 0.909 to 0.955.

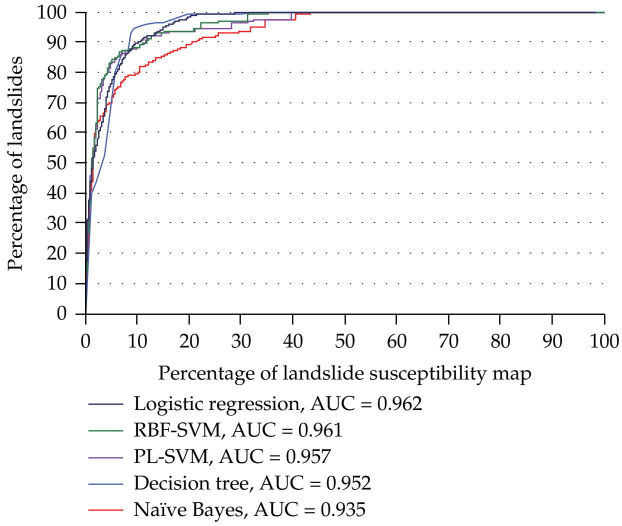


Figure 7: Success-rate curves and area, under the curves (AUCs) of RBF-SVM, PL-SVM, DT, and NB models in comparison with the logistic regression model.

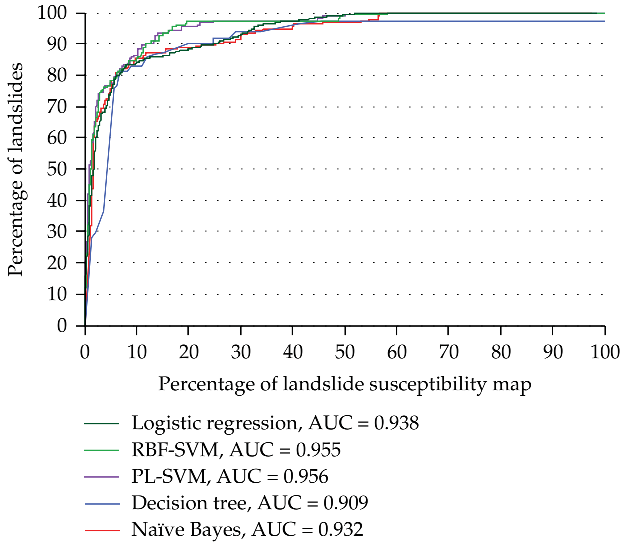


Figure 8: Prediction-rate curves and areas under the curves (AUCs) of RBF-SVM, PL-SVM, DT, and NB models in comparison with the logistic regression model.

It indicates that all the models have a good prediction capability. The highest prediction capability is for RBF-SVM and PL-SVM with AUC values of 0.954 and 0.955, respectively. They are followed by NB (0.935) and DT (0.907). Compared with the logistic regression (AUC of 0.938) that used the same data, it can be seen that the prediction capability of the two SVM models may be slightly better whereas the prediction capability of DT and ND is lower.

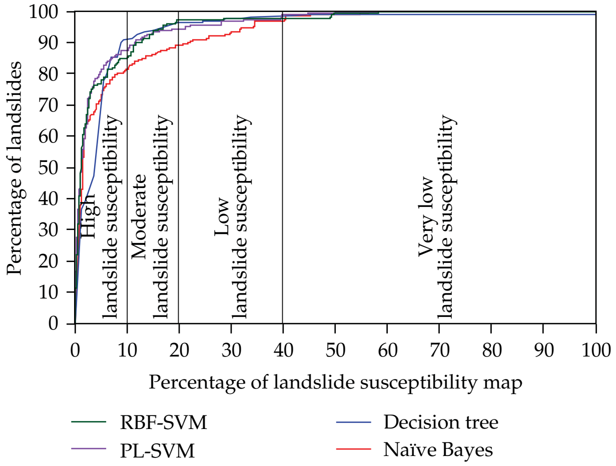


Figure 9: Percentage of landslides against percentage of landslide susceptibility maps using of RBF-SVM, PL-SVM, DT, and NB models.

4.2. Reclassification of Landslide Susceptibility Indexes

The landslide susceptibility indexes were reclassified into four relative susceptibility classes: high, moderate, low, and very low. In this study, the classification method proposed by Pradhan and Lee [8] was used to determine landslide susceptibility class breaks based on percentage of area: high (10%), moderate (10%), low (20%), and very low (60%) (Figure 9).

Landslide density analysis was performed on the four landslide susceptibility classes [97]. Landslide density is defined as the ratio of landslide pixels to the total number of pixels in the susceptibility class. An ideal landslide susceptibility map has the landslide density value increasing from a very low- to a higher-susceptibility class [32]. A plotting of the landslide density for the four landslide susceptibility classes of the four landslide susceptibility models (RBF-SVM, PL-SVM, DT, and NB) is shown in Figure 10. It could be observed that the landslide density is gradually increased from the very low- to the high-susceptibility class. Figure 11 shows landslide susceptibility maps using RBF-SVM, PL-SVM, DT, and NB models.

Table 6 shows the characteristics of the four susceptibility classes of the four maps of the study area. It can be observed that the percentages of existing landslide pixels for the high class are 87.2%, 87.5%, 90.7%, and 81.3% for RBF-SVM, PL-SVM, DT, and NB, respectively. In contrast, 80% of the pixels in the study areas are in the low- and very-low-susceptibility classes. These maps are satisfing two spatial effective rules [98], (1) the existing landslide pixels should belong to the high-susceptibility class and (2) the high susceptibility class should cover only small areas.

5. Discussions and Conclusions

This paper presents a comparative study of three data mining approaches SVM, DT, and NB for landslide susceptibility mapping in the Hoa Binh province (Vietnam). The landslide inventory was constructed with 118 polygons of landslides that occurred during the last ten years. A total of ten landslide conditioning factors were used in this analysis, including slope

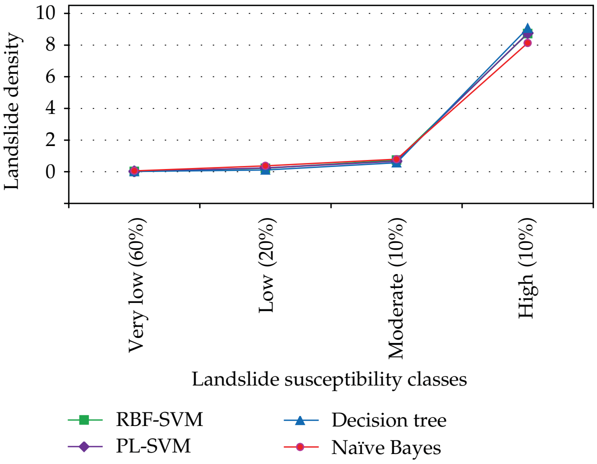


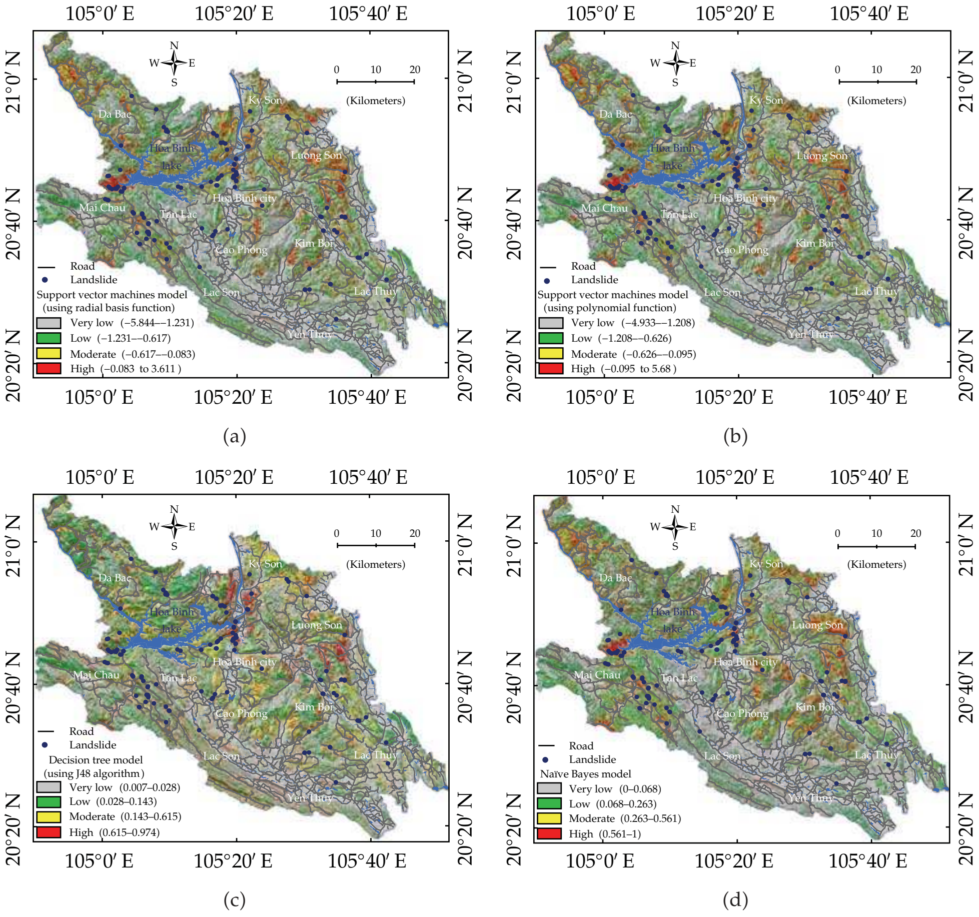
Figure 10: Landslide density plots of four landslide susceptibility classes of RBF-SVM, PL-SVM, DT, and NB models.

Table 6: Characteristics of the four susceptibility zones of the four landslide susceptibility models obtained from RBF-SVM, PL-SVM, DT, and NB models.

Landslide susceptibility classes	Percentage of area	Landslide density			
		RBF-SVM	PL-SVM	DT	NB
High	10.0	8.719	8.749	9.069	8.128
Moderate	10.0	0.740	0.660	0.571	0.791
Low	20.0	0.221	0.241	0.115	0.371
Very low	60.0	0.017	0.018	0.022	0.057

angle, lithology, rainfall, soil type, slope aspect, landuse, distance to roads, distance to rivers, distance to faults, and relief amplitude. For building the models, a training dataset was extracted with 70% of the landslide inventory, whereas the remaining landslide inventory was used for the assessment of the prediction capability of the models. Using the three data mining algorithms, SVM, DT, and NB, the landslide susceptibility maps were produced. These maps present spatial predictions of landslides. They do not include information “when” and “how frequently” landslides will occur.

In the case of SVM, the selection of the kernel function and its parameters play an important role in landslide susceptibility assessment. For the RBF function, the best kernel parameters of C and γ are 8 and 0.25, respectively. For the PL function, it is clear that the degree of polynomial function had significant effect in the model. The SVM model with a polynomial degree of 3 has the highest accuracy. The best kernel parameters of C and γ are 1 and 0.3536 respectively. In the case of DT, the probability that an observation belongs to landslide class using Laplace smoothing was used to calculate the landslide susceptibility index. For building the DT model, the selection of MNI per leaf tree and CF has largely affected the accuracy of the model. In this study, the best decision tree model is found with MNI per leaf tree as 6 and the CF as 0.35. Relative importance of landslide conditioning factors are as follows: distance to roads, slope angle, landuse, slope aspect, rainfall, relief amplitude, distance to rivers, distance to faults, lithology, and soil type. In the case of NB, the application for landslide modeling is relatively robust. This is not a time-consuming method,



(a)

(b)

(c)

(d)

Figure 11: Landslide susceptibility maps of the Hoa Binh province (Vietnam) using: (a) RBF-SVM; (b) PL-SVM; (c) DT; and (d) NB.

and techniques required to use are simple. The result of this study shows that NB gives relatively good prediction capability.

Qualitative interpretation of the high landslide susceptibility classes of the four maps shows that they agree quite well with field evidence and assumptions. High probability of landslides distributes in areas with active fault zones and road-cut sections. Using the success-rate and prediction-rate methods, the landslide susceptibility maps were validated using the existing landslide locations. The quantitative results show that all the landslide models have good prediction capability. The highest area under the success-rate curve (AUC) is for the RBF-SVM (0.961), followed by PL-SVM (0.956), DT (0.938), and NB (0.935). The highest prediction-rate result is for RBF-SVM and PL-SVM with areas under the prediction curves (AUC) of 0.954 and 0.955, respectively. They are followed by NB (0.932) and DT (0.903). When compared with the results obtained from the logistic regression (Figure 8), the prediction capabilities of the two SVM models are slightly better. On contrast, DT and NB models have lower accuracy. The quantitative results of this study are comparable to those obtained in other studies, such as Brenning [99] and Yilmaz [35]. The findings of this study agree with Yao et al [100] who states that SVM possesses better prediction efficiency than the logistic regression. Additionally, the findings also agree with Marjanović et al. [101], who reported that SVM outperformed the logistic regression and DT. Similarly, the results also agree with Ballabio and Sterlacchini [102], who concluded that SVM was found to outperform the logistic regression, linear discriminant, and NB.

The reliabilities of the landslide models were assessed using Cohen kappa index (κ). In this study, the kappa indexes are of 0.822, 0.823, and 0.860 for RBF-SVM, PL-SVM, and DT, respectively. It indicates an almost perfect agreement between the observed and the predicted values. Cohen kappa index is 0.722 for NB indicating substantial agreement between the observed and the predicted values. The reliability analysis results are satisfying compared with other works such as Guzzetti et al. [91] and Saito et al. [44].

Landslide susceptibility maps are considered to be a useful tool for territorial planning, disaster management, and natural hazards' mitigation. This study shows that SVMs have considered being a powerful tool for landslide susceptibility with high accuracy. As a final conclusion, the analyzed results obtained from the study can provide very useful information for decision making and policy planning in landslide areas.

Acknowledgments

This research was funded by the Norwegian Quota scholarship program. The data analysis and write-up were carried out as a part of the first author's Ph.D. studies at the Geomatics Section, Department of Mathematical Sciences and Technology, Norwegian University of Life Sciences, Norway.

References

[1] K. Sassa and P. Canuti, *Landslides-Disaster Risk Reduction*, Springer, New York, NY, USA, 2008.

[2] D. Tien Bui, O. Lofman, I. Revhaug, and O. Dick, "Landslide susceptibility analysis in the Hoa Binh province of Vietnam using statistical index and logistic regression," *Natural Hazards*, vol. 59, pp. 1413–1444, 2011.

[3] D. Tien Bui, B. Pradhan, O. Lofman, I. Revhaug, and O. B. Dick, "Landslide susceptibility mapping at Hoa Binh province (Vietnam) using an adaptive neuro-fuzzy inference system and GIS," *Computers & Geosciences*. In press.

[4] D. Tien Bui, B. Pradhan, O. Lofman, I. Revhaug, and O. B. Dick, "Spatial prediction of landslide hazards in Hoa Binh province (Vietnam): a comparative assessment of the efficacy of evidential belief functions and fuzzy logic models," *CATENA*, vol. 96, pp. 28–40, 2012.

[5] S. Lee and T. Dan, "Probabilistic landslide susceptibility mapping on the Lai Chau province of Vietnam: focus on the relationship between tectonic fractures and landslides," *Environmental Geology*, vol. 48, no. 6, pp. 778–787, 2005.

[6] S. Lee, "Landslide susceptibility mapping using an artificial neural network in the Gangneung area, Korea," *International Journal of Remote Sensing*, vol. 28, no. 21, pp. 4763–4783, 2007.

[7] B. Pradhan, "Use of GIS-based fuzzy logic relations and its cross application to produce landslide susceptibility maps in three test areas in Malaysia," *Environmental Earth Sciences*, vol. 63, no. 2, pp. 329–349, 2011.

[8] B. Pradhan and S. Lee, "Regional landslide susceptibility analysis using back-propagation neural network model at Cameron Highland, Malaysia," *Landslides*, vol. 7, no. 1, pp. 13–30, 2010.

[9] 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.

[10] 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.

[11] H. Wang, L. Gangjun, X. Weiya, and W. Gonghui, "GIS-based landslide hazard assessment: an overview," *Progress in Physical Geography*, vol. 29, no. 4, pp. 548–567, 2005.

[12] J. Chacón, C. Irigaray, T. Fernández, and R. El Hamdouni, "Engineering geology maps: landslides and geographical information systems," *Bulletin of Engineering Geology and the Environment*, vol. 65, no. 4, pp. 341–411, 2006.

[13] M. Ercanoglu and C. Gokceoglu, "Assessment of landslide susceptibility for a landslide-prone area (north of Yenice, NW Turkey) by fuzzy approach," *Environmental Geology*, vol. 41, no. 6, pp. 720–730, 2002.

[14] M. Ercanoglu and C. Gokceoglu, "Use of fuzzy relations to produce landslide susceptibility map of a landslide prone area (West Black Sea region, Turkey)," *Engineering Geology*, vol. 75, no. 3-4, pp. 229–250, 2004.

[15] B. Pradhan, E. A. Sezer, C. Gokceoglu, and M. F. Buchroithner, "Landslide susceptibility mapping by neuro-fuzzy approach in a landslide-prone area (Cameron Highlands, Malaysia)," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 12, pp. 4164–4177, 2010.

[16] S. Lee, "Application and verification of fuzzy algebraic operators to landslide susceptibility mapping," *Environmental Geology*, vol. 52, no. 4, pp. 615–623, 2007.

[17] A. Akgun, E. A. Sezer, H. A. Nefeslioglu, C. Gokceoglu, and B. Pradhan, "An easy-to-use MATLAB program (MamLand) for the assessment of landslide susceptibility using a Mamdani fuzzy algorithm," *Computers and Geosciences*, vol. 38, no. 1, pp. 23–34, 2011.

[18] B. Pradhan, "Application of an advanced fuzzy logic model for landslide susceptibility analysis," *International Journal of Computational Intelligence Systems*, vol. 3, no. 3, pp. 370–381, 2010.

[19] B. Pradhan, "Landslide susceptibility mapping of a catchment area using frequency ratio, fuzzy logic and multivariate logistic regression approaches," *Journal of the Indian Society of Remote Sensing*, vol. 38, no. 2, pp. 301–320, 2010.

[20] B. Pradhan, "Manifestation of an advanced fuzzy logic model coupled with Geo-information techniques to landslide susceptibility mapping and their comparison with logistic regression modelling," *Environmental and Ecological Statistics*, vol. 18, no. 3, pp. 471–493, 2011.

[21] 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.

[22] M. H. Vahidnia, A. A. Alesheikh, A. Alimohammadi, and F. Hosseinali, "A GIS-based neuro-fuzzy procedure for integrating knowledge and data in landslide susceptibility mapping," *Computers and Geosciences*, vol. 36, no. 9, pp. 1101–1114, 2010.

[23] S. Lee, J. H. Ryu, K. Min, and J. S. Won, "Landslide susceptibility analysis using GIS and artificial neural network," *Earth Surface Processes and Landforms*, vol. 28, no. 12, pp. 1361–1376, 2003.

[24] S. Lee, J. H. Ryu, J. S. Won, and H. J. Park, "Determination and application of the weights for landslide susceptibility mapping using an artificial neural network," *Engineering Geology*, vol. 71, no. 3-4, pp. 289–302, 2004.

[25] F. Catani, N. Casagli, L. Ermini, G. Righini, and G. Menduni, "Landslide hazard and risk mapping at catchment scale in the Arno River basin," *Landslides*, vol. 2, no. 4, pp. 329–342, 2005.

[26] 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.

[27] B. Pradhan, S. Lee, and M. F. Buchroithner, "A GIS-based back-propagation neural network model and its cross-application and validation for landslide susceptibility analyses," *Computers, Environment and Urban Systems*, vol. 34, no. 3, pp. 216–235, 2010.

[28] I. Yilmaz, "A case study from Koyulhisar (Sivas-Turkey) for landslide susceptibility mapping by artificial neural networks," *Bulletin of Engineering Geology and the Environment*, vol. 68, no. 3, pp. 297–306, 2009.

[29] B. Pradhan and M. F. Buchroithner, "Comparison and validation of landslide susceptibility maps using an artificial neural network model for three test areas in Malaysia," *Environmental and Engineering Geoscience*, vol. 16, no. 2, pp. 107–126, 2010.

[30] I. Yilmaz, "The effect of the sampling strategies on the landslide susceptibility mapping by conditional probability and artificial neural networks," *Environmental Earth Sciences*, vol. 60, no. 3, pp. 505–519, 2010.

[31] I. Yilmaz, "Landslide susceptibility mapping using frequency ratio, logistic regression, artificial neural networks and their comparison: a case study from Kat landslides (Tokat-Turkey)," *Computers and Geosciences*, vol. 35, no. 6, pp. 1125–1138, 2009.

[32] B. Pradhan and S. Lee, “Landslide susceptibility assessment and factor effect analysis: backpropagation artificial neural networks and their comparison with frequency ratio and bivariate logistic regression modelling,” *Environmental Modelling and Software*, vol. 25, no. 6, pp. 747–759, 2010.

[33] 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.

[34] H. A. Nefeslioglu, C. Gokceoglu, and H. Sonmez, “An assessment on the use of logistic regression and artificial neural networks with different sampling strategies for the preparation of landslide susceptibility maps,” *Engineering Geology*, vol. 97, no. 3-4, pp. 171–191, 2008.

[35] I. Yilmaz, “Comparison of landslide susceptibility mapping methodologies for Koyulhisar, Turkey: conditional probability, logistic regression, artificial neural networks, and Support Vector Machine,” *Environmental Earth Sciences*, vol. 61, no. 4, pp. 821–836, 2010.

[36] C. P. Poudyal, C. Chang, H. J. Oh, and S. Lee, “Landslide susceptibility maps comparing frequency ratio and artificial neural networks: a case study from the Nepal Himalaya,” *Environmental Earth Sciences*, vol. 61, no. 5, pp. 1049–1064, 2010.

[37] B. Pradhan, “Remote sensing and GIS-based landslide hazard analysis and cross-validation using multivariate logistic regression model on three test areas in Malaysia,” *Advances in Space Research*, vol. 45, no. 10, pp. 1244–1256, 2010.

[38] A. S. Miner, P. Vamplew, D. J. Windle, P. Flentje, and P. Warner, “A comparative study of various data mining techniques as applied to the modeling of landslide susceptibility on the Bellarine Peninsula, Victoria, Australia,” in *Geologically Active*, A. L. Williams, G. M. Pinches, C. Y. Chin, and T. J. McMorran, Eds., p. 352, CRC Press, New York, NY, USA, 2010.

[39] S. Wan and T. C. Lei, “A knowledge-based decision support system to analyze the debris-flow problems at Chen-Yu-Lan River, Taiwan,” *Knowledge-Based Systems*, vol. 22, no. 8, pp. 580–588, 2009.

[40] X. Wu, V. Kumar, Q. J. Ross et al., “Top 10 algorithms in data mining,” *Knowledge and Information Systems*, vol. 14, no. 1, pp. 1–37, 2008.

[41] S. B. Bai, J. Wang, G. N. Lu, M. Kanevski, and A. Pozdnoukhov, “GIS-Based landslide susceptibility mapping with comparisons of results from machine learning methods versus logistic regression in basin scale,” *Geophysical Research Abstracts, EGU*, vol. 10,A-06367, 2008.

[42] N. Micheletti, L. Foresti, M. Kanevski, A. Pedrazzini, and M. Jaboyedoff, “Landslide susceptibility mapping using adaptive Support Vector Machines and feature selection,” *Geophysical Research Abstracts, EGU*, vol. 13, 2011.

[43] 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.

[44] 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.

[45] H. A. Nefeslioglu, E. Sezer, C. Gokceoglu, A. S. Bozkir, and T. Y. Duman, “Assessment of landslide susceptibility by decision trees in the metropolitan area of Istanbul, Turkey,” *Mathematical Problems in Engineering*, vol. 2010, Article ID 901095, 2010.

[46] C. A. Ratanamahatana and D. Gunopulos, “Feature selection for the naive Bayesian classifier using decision trees,” *Applied Artificial Intelligence*, vol. 17, no. 5-6, pp. 475–487, 2003.

[47] W. Tzu-Tsung, “A hybrid discretization method for naïve Bayesian classifiers,” *Pattern Recognition*, vol. 45, no. 6, pp. 2321–2325, 2012.

[48] D. Soria, J. M. Garibaldi, F. Ambrogi, E. M. Biganzoli, and I. O. Ellis, “A “non-parametric” version of the naive Bayes classifier,” *Knowledge-Based Systems*, vol. 24, no. 6, pp. 775–784, 2011.

[49] J. Kazmierska and J. Malicki, “Application of the naïve Bayesian classifier to optimize treatment decisions,” *Radiotherapy and Oncology*, vol. 86, no. 2, pp. 211–216, 2008.

[50] C.-C. Chang and C.-J. Lin, *LIBSVM : a Library for Support Vector Machines*, ACM Transactions on Intelligent Systems and Technology, New York, NY, USA, 2011.

[51] B. D. Malamud, D. L. Turcotte, F. Guzzetti, and P. Reichenbach, “Landslide inventories and their statistical properties,” *Earth Surface Processes and Landforms*, vol. 29, no. 6, pp. 687–711, 2004.

[52] F. Vergari, M. Della Seta, M. Del Monte, P. Fredi, and E. Lupia Palmieri, “Landslide susceptibility assessment in the Upper Orcia Valley (Southern Tuscany, Italy) through conditional analysis: a contribution to the unbiased selection of causal factors,” *Natural Hazards and Earth System Science*, vol. 11, no. 5, pp. 1475–1497, 2011.

[53] T. T. Van, D. T. Anh, H. H. Hieu et al., *Investigation and Assessment of the Current Status and Potential of Landslide in Some Sections of the Ho Chi Minh Road, National Road 1A and Proposed Remedial Measures to Prevent Landslide from Threat of Safety of People, Property, and Infrastructure*, Vietnam Institute of Geoscience and Mineral Resources, Hanoi, Vietnam, 2006.

[54] F. Arikan, R. Ulusay, and N. Aydin, "Characterization of weathered acidic volcanic rocks and a weathering classification based on a rating system," *Bulletin of Engineering Geology and the Environment*, vol. 66, no. 4, pp. 415–430, 2007.

[55] V. N. Vapnik, *Statistical Learning Theory*, Wiley-Interscience, New York, NY, USA, 1998.

[56] S. Abe, *Support Vector Machines for Pattern Classification*, Springer, London, UK, 2010.

[57] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[58] P. Samui, "Slope stability analysis: a Support Vector Machine approach," *Environmental Geology*, vol. 56, no. 2, pp. 255–267, 2008.

[59] R. Damaševičius, "Optimization of SVM parameters for recognition of regulatory DNA sequences," *Top*, vol. 18, no. 2, pp. 339–353, 2011.

[60] S. Song, Z. Zhan, Z. Long, J. Zhang, and L. Yao, "Comparative study of SVM methods combined with voxel selection for object category classification on fMRI data," *PLoS ONE*, vol. 6, no. 2, Article ID e17191, 2011.

[61] S. S. Keerthi and C. J. Lin, "Asymptotic behaviors of Support Vector Machines with gaussian kernel," *Neural Computation*, vol. 15, no. 7, pp. 1667–1689, 2003.

[62] H.-T. Lin and C.-J. Lin, "A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods," Tech. Rep., National Taiwan University, Taipei, Taiwan, 2003.

[63] X. Zhu, S. Zhang, Z. Jin, Z. Zhang, and Z. Xu, "Missing value estimation for mixed-attribute data sets," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 1, pp. 110–121, 2011.

[64] R. Damaševičius, "Structural analysis of regulatory DNA sequences using grammar inference and Support Vector Machine," *Neurocomputing*, vol. 73, no. 4–6, pp. 633–638, 2010.

[65] S. Ali and K. A. Smith, "Automatic parameter selection for polynomial kernel," in *Proceedings of the IEEE International Conference on Information Reuse and Integration (IRI '03)*, pp. 243–249, Octobe 2003.

[66] D. Mattera and S. Haykin, "Support Vector Machines for dynamic reconstruction of a chaotic system," in *Advances in Kernel Methods*, pp. 211–241, MIT Press, Cambridge, Mass, USA, 1999.

[67] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee, "Choosing multiple parameters for Support Vector Machines," *Machine Learning*, vol. 46, no. 1–3, pp. 131–159, 2002.

[68] J. Platt, *Probabilistic Outputs for Support Vector Machines and Comparison to Regularized Likelihood Methods*, MIT Pres, Cambridge, Mass, USA, 2000.

[69] V. Cherkassky and F. Mulier, *Learning from Data: Concepts, Theory and Methods*, John Wiley and Sons, New York, NY, USA, 2007.

[70] L. Zhuang and H. Dai, "Parameter optimization of kernel-based one-class classifier on imbalance text learning," in *Pricai 2006: Trends in Artificial Intelligence, Proceedings*, vol. 4099, pp. 434–443, 2006.

[71] T. Mu and A. K. Nandi, "Breast cancer detection from FNA using SVM with different parameter tuning systems and SOM-RBF classifier," *Journal of the Franklin Institute*, vol. 344, no. 3-4, pp. 285–311, 2007.

[72] A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, and S. D. Brown, "An introduction to decision tree modeling," *Journal of Chemometrics*, vol. 18, no. 6, pp. 275–285, 2004.

[73] M. Debeljak and S. Džeroski, "Decision trees in ecological modelling," in *Modelling Complex Ecological Dynamics*, F. Jopp, H. Reuter, and B. Breckling, Eds., pp. 197–209, Springer, Berlin, Germany, 2011.

[74] S. K. Murthy, "Automatic construction of decision trees from data: a multi-disciplinary survey," *Data Mining and Knowledge Discovery*, vol. 2, no. 4, pp. 345–389, 1998.

[75] R. Bou Kheir, M. H. Greve, C. Abdallah, and T. Dalgaard, "Spatial soil zinc content distribution from terrain parameters: a GIS-based decision-tree model in Lebanon," *Environmental Pollution*, vol. 158, no. 2, pp. 520–528, 2010.

[76] G. K. F. Tso and K. K. W. Yau, "Predicting electricity energy consumption: a comparison of regression analysis, decision tree and neural networks," *Energy*, vol. 32, no. 9, pp. 1761–1768, 2007.

[77] Y. Zhao and Y. Zhang, "Comparison of decision tree methods for finding active objects," *Advances in Space Research*, vol. 41, no. 12, pp. 1955–1959, 2008.

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

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

[80] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.

[81] J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, Calif, USA, 1993.

[82] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, Los Altos, Calif, USA, 2nd edition, 2005.

[83] T. S. Lim, W. Y. Loh, and Y. S. Shih, "Comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms," *Machine Learning*, vol. 40, no. 3, pp. 203–228, 2000.

[84] J. H. Cho and P. U. Kurup, "Decision tree approach for classification and dimensionality reduction of electronic nose data," *Sensors and Actuators B*, vol. 160, no. 1, pp. 542–548, 2011.

[85] V. T. Tran, B. S. Yang, M. S. Oh, and A. C. C. Tan, "Fault diagnosis of induction motor based on decision trees and adaptive neuro-fuzzy inference," *Expert Systems with Applications*, vol. 36, no. 2, pp. 1840–1849, 2009.

[86] F. Provost and P. Domingos, "Tree induction for probability-based ranking," *Machine Learning*, vol. 52, no. 3, pp. 199–215, 2003.

[87] Z. Xie, Q. Zhang, W. Hsu, and M. Lee, "Enhancing SNNB with local accuracy estimation and ensemble techniques," in *Proceedings of the 10th international conference on Database Systems for Advanced Applications (DASFAA '05)*, L. Zhou, B. Ooi, and X. Meng, Eds., vol. 3453, p. 983, Springer, Beijing, China, April 2005.

[88] Y. Murakami and K. Mizuguchi, "Applying the Naïve Bayes classifier with kernel density estimation to the prediction of protein-protein interaction sites," *Bioinformatics*, vol. 26, no. 15, pp. 1841–1848, 2010.

[89] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[90] F. K. Hoehler, "Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity," *Journal of Clinical Epidemiology*, vol. 53, no. 5, pp. 499–503, 2000.

[91] 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.

[92] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.

[93] B. Pradhan and S. Lee, "Delineation of landslide hazard areas on Penang island, Malaysia, by using frequency ratio, logistic regression, and artificial neural network models," *Environmental Earth Sciences*, vol. 60, no. 5, pp. 1037–1054, 2010.

[94] C. M. Wang and Y. F. Huang, "Evolutionary-based feature selection approaches with new criteria for data mining: a case study of credit approval data," *Expert Systems with Applications*, vol. 36, no. 3, pp. 5900–5908, 2009.

[95] C. J. F. Chung and A. G. Fabbri, "Validation of spatial prediction models for landslide hazard mapping," *Natural Hazards*, vol. 30, no. 3, pp. 451–472, 2003.

[96] S. Lee, J. H. Ryu, and I. S. Kim, "Landslide susceptibility analysis and its verification using likelihood ratio, logistic regression, and artificial neural network models: case study of Youngin, Korea," *Landslides*, vol. 4, no. 4, pp. 327–338, 2007.

[97] S. Sarkar, D. P. Kanungo, A. K. Patra, and P. Kumar, "GIS based spatial data analysis for landslide susceptibility mapping," *Journal of Mountain Science*, vol. 5, no. 1, pp. 52–62, 2008.

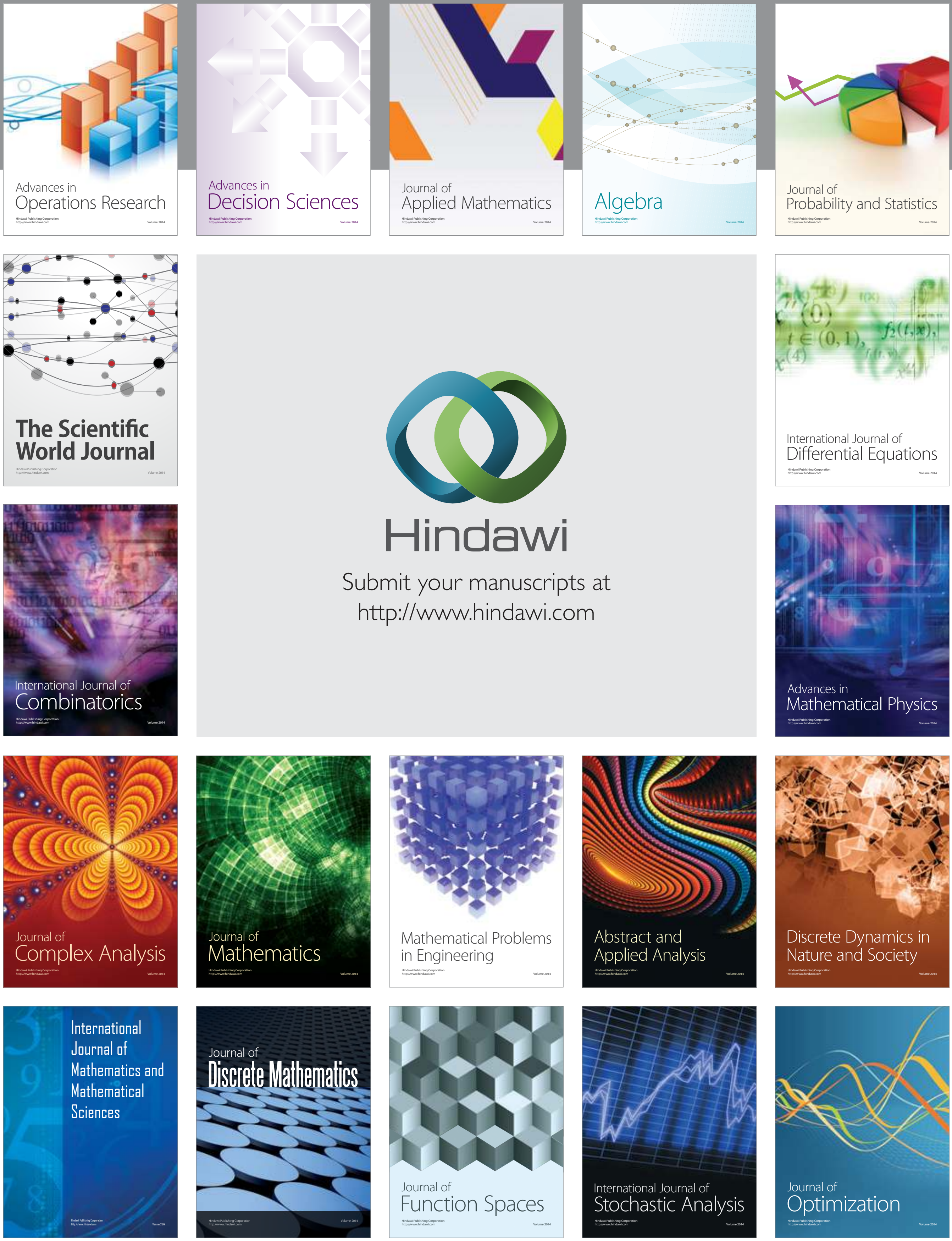
[98] T. Can, H. A. Nefeslioglu, C. Gokceoglu, H. Sonmez, and T. Y. Duman, "Susceptibility assessments of shallow earthflows triggered by heavy rainfall at three catchments by logistic regression analyses," *Geomorphology*, vol. 72, no. 1–4, pp. 250–271, 2005.

[99] A. Brenning, "Spatial prediction models for landslide hazards: review, comparison and evaluation," *Natural Hazards and Earth System Science*, vol. 5, no. 6, pp. 853–862, 2005.

[100] X. Yao, L. G. Tham, and F. C. Dai, "Landslide susceptibility mapping based on Support Vector Machine: a case study on natural slopes of Hong Kong, China," *Geomorphology*, vol. 101, no. 4, pp. 572–582, 2008.

[101] M. Marjanović, M. Kovačević, B. Bajat, and V. Voženílek, "Landslide susceptibility assessment using SVM machine learning algorithm," *Engineering Geology*, vol. 123, no. 3, pp. 225–234, 2011.

[102] C. Ballabio and S. Sterlacchini, "Support Vector Machines for landslide susceptibility mapping: the Staffora River Basin case study, Italy," *Mathematical Geosciences*, vol. 44, no. 1, pp. 47–70, 2012.


Advances in Operations Research
Advances in Decision Sciences
Journal of Applied Mathematics
Algebra
Journal of Probability and Statistics
The Scientific World Journal
International Journal of Differential Equations
International Journal of Combinatorics
Advances in Mathematical Physics
Journal of Complex Analysis
Journal of Mathematics
Mathematical Problems in Engineering
Abstract and Applied Analysis
Discrete Dynamics in Nature and Society
International Journal of Mathematics and Mathematical Sciences
Journal of Discrete Mathematics
Journal of Function Spaces
International Journal of Stochastic Analysis
Journal of Optimization