Vietnam Journal of Earth Sciences, 44(3), 327-342, https://doi.org/10.15625/2615-9783/16929

Vietnam Academy of Science and Technology
Vietnam Journal of Earth Sciences
http://www.vjs.ac.vn/index.php/jse

A new approach based on integration of random subspace and C4.5 decision tree learning method for spatial prediction of shallow landslides

Viet-Ha Nhu[1*], Tinh Thanh Bui[2], Linh Nguyen My[3], Hoe Vuong[4], Nhat Duc Hoang[5]

[1]*Department of Geological-Geotechnical Engineering, Hanoi University of Mining and Geology, Hanoi Vietnam*
[2]*Department of Prospecting and Exploration Geology, Hanoi University of Mining and Geology, Hanoi, Vietnam*
[3]*Department of Hydrogeoloy and Engineering Geology, Vietnam Institute of Geosciences and Mineral Resources, Thanh Xuan, Hanoi 100000, Vietnam*
[4]*Faculty of Surveying, Mapping and Geographic Information, Hanoi university of Natural Resources and Environment, Vietnam*
[5]*Faculty of Civil Engineering, Institute of Research and Development, Duy Tan University, Da Nang, 550000, Vietnam*

Received 02 August 2021; Received in revised form 21 October 2021; Accepted 12 February 2022

ABSTRACT

The research approaches a new machine learning ensemble which is a hybridization of Random subspace (RS) and C4.5, named RandSub-DT, for improving the performance of the landslide susceptibility model. This is based on the GIS database, including 170 landslide polygons and ten predisposing landslide factors, i.e., slope, aspect, curvature, TWI, land use, distance to road, distance to the river, soil type, distance to fault, and lithology. We carried out this study in the Ha Long and Cam Pha City areas which are important economic centers in the Quang Ninh province, Vietnam, where landslides seriously influence the daily life of the citizen causing economic damage. We then used **a** GIS database to construct and validate the proposed RandSub-DT model. The model performance was assessed using a confusion matrix and a set of statistical measures. The result showed that the RandSub-DT model with the classification accuracy of 90.34% in the training dataset and the prediction capability of 77.48% had **a** high performance for landslide prediction. This research proved that an ensemble of the C4.5 and RS provided a highly accurate estimate of landslide susceptibility in the research area.

Keywords: Landslide, random subspace, C4.5, GIS, Quang Ninh.

1. Introduction

Ha Long and Cam Pha City are two major economic centers of Quang Ninh province. However, frequently and continuously occur here, which seriously affect the citizen's daily life and cause significant economic losses for these areas, especially in the rainy season. It has been more and more seriously involved in recent years due to the impact of climate change in the state; the rains and storm activities are complicatedly happening because of extreme weather conditions. Therefore, the research

*Corresponding author, Email: nhuvietha@humg.edu.vn

and forecast of the landslide risk areas are necessary and urgent. This will assist local authorities to manage and plan for constructing infrastructure; to minimize the damage caused by landslides.

In recent years, applications of machine learning algorithms for predicting landslides have become popular (Akgun et al., 2012; Meng et al., 2016; Pham et al., 2016; Tien Bui et al., 2016; Bui et al., 2017; Gheshlaghi and Feizizadeh, 2017; Pham et al., 2018; Nhu et al., 2020; Nhu et al., 2021). These applications contribute to creating a landslide susceptibility map that helps identify areas with high landslide probability. Many machine learning methods have grown up to help with building landslide models. However, the demand for the development and application of new techniques and algorithms is still needed to enhance the quality and accuracy of landslide prediction. Recently, the performance of machine learning models has frequently developed with several ensemble machine learning methods, which has achieved some promising results (Bui et al., 2016; Pham et al., 2016; Pham et al., 2017; Nhu et al., 2021). Ensemble techniques utilize multiple algorithms to combine different machine learning methods that create hybrid models; the main advantage of these methods is that they can efficiently handle a large and complex number of the input to produce a reliable output.

In this article, the research purpose is to generate a *landslide susceptibility map* using a new machine learning ensemble approach that combines C4.5 (Quinlan, 2014) and Random subspace (RS) (Ho, 1995), it is named RandSub-DT, this could enhance the performance of the landslide model because C4.5 is a popular machine learning algorithm in studying landslide, whereas RS is a framework which has proven efficient in landslide modeling (Hong et al., 2017; Pham et al., 2017; Pham et al., 2018). This is the first time the combination of C4.5 and RS is responsible for the landslide study, resulting in a new effective prediction method to forecast landslide susceptibility with acceptable accuracy. Its development was under Python application and ArcGIS software.

2. Research area and data

2.1. Description of the research area

The research area belongs to Ha Long and Cam Pha city, Quang Ninh province. It is 180 km northeast from Hanoi capital with an area of about 563 km^2 and is limited by geographical coordinates: 20°40'00"-21°13'00"N and 106° 55'00"-107° 25'00"E (Binh Do Le, 1968; Hung Le, 1996; Nhu et al., 2020). The topography of the research area is strongly fragmented, fluctuating from medium to low mountainous terrain and the alternating coastal plains. The latitudes with the variation of the elevation from 00 to 829.1 m. The mountain ranges mainly extend along the northwest-southeast. The network of rivers and streams in the searching area is very developed, with two large rivers (the Man and the Dien Vong rivers) and the stream systems. A tropical monsoon climate characterizes the climate in the region. It is divided into two distinct seasons: the dry season (from October to April) and the rainy season (from May to October). The temperature fluctuates in a year from 5° to 40° Celsius, an average of 20 degrees (Binh Do Le, 1968; Hung Le, 1996).

The research area is one of the country's major industrial regions where mining, trade, and tourism industries play an essential and decisive role in the region's economic development. The area has a relatively developed transportation network, including roads and waterways. Highway 18A is the arterial route running along the length of this area. In addition, there are roads connecting

mines, districts, tourist areas, and villages to form a dense transportation network.

The research area has recorded landslide activities that often occur in steep slopes and mountainous regions with roads and residential regions passing through. Landslides arise due to both natural and human activities. Common human activities include socio-economic development, construction, urbanization, deforestation, etc. When the works were under construction, the prolonged heavy rain washed away the soil and rock above and formed mud streams that overflowed to the people's houses below (Figs. 1 and 2).

Figure 1. Landslides in Hong Ha ward (Ha Long city) on July 27, 2015. (Source of photos: www.baoquangninh.com.vn)

Figure 2. Handling landslides on Highway 18A in Cua Ong ward, Cam Pha city, July 8, 2017. (Source of photos: www.baoquangninh.com.vn)

2.2. Shallow landslide inventory

Notably, this study relies on RandSub-DT to model a machine learning-based spatial analysis of landslide occurrence. This underlying assumption of landslide susceptibility prediction is that the factors causing past landslides will continue to influence the likelihood of landslide occurrence in the future (Reichenbach et al., 2018). Hence, the authors gathered the geo-information on past landslides such as terrain, geological condition, and land use to construct a landslide inventory for the research area. This analysis used a landslide inventory with 170 shallow soil, and rock mixed soil slides from the NAFOSTED-Funded Landslide Project No105.08-2017.316 of Vietnam (Nhu et al., 2020).

The authors surveyed and identified the landslide using aerial photographs and Google Earth images 2020. We corrected the collected landslides and undertook the mapping to make a good database for modeling. We only consider rainfall-triggered landslides because no earthquake-triggered landslide was reported in the research area. These landslides occurred from 2015 to 2020.

2.3. Influencing factors

The factors such as topography, land use, lithology, soil type, and river network are the main ones that influence landslide

occurrent (Hue et al., 2004; Bui et al., 2017; Hung et al., 2017). Thus, we selected these factors as input for the landslide RandSub-DT model.

In this study, the slope is a vital factor for causing landslides. Because water flows from high points to low points by gravity, the slope plays an essential role in controlling surface flow, affecting the speed of the water flow and the time of the permeating water. The slope is larger; the permeability is lower, and vice versa. Besides, soil type and permeability directly affect landslides. When rainwater falls, some are absorbed into the topsoil, partly evaporated, partly retained by plant elements, and the remainder form surface runoff. Therefore, the water holding capacity of soil has a significant influence in causing landslides.

Aspect and curvature play a significant role in controlling the direction of the flow and the depth of the water. So that they affect the extent and intensity of landslides. TWI (Topographic Wetness Index) directly affects the landslide. The humidity value in the research area is higher, the land is easier to quickly reach saturation situation when it rains.

Vegetation plays a pretty important role in landslide mechanisms as removing soil moisture through evapotranspiration and providing root cohesion to the soil mantle. For land-use types being residential land construction land, the water permeability is not good, so the drainage capacity is good. With forest land and agricultural land, they prevent water well, reducing landslide risk. In contrast, coal mining land is an area that emerges as a dangerous landslide site. Because there are many coal mines and exploration activities, many coal companies have dumped rock and soil waste into the pile with a height from 30 to 70 m. Many households live close to the waste area due to coal mines, and landslides are threatening them. There are even some places where the foot of the dumpsite to some households is less than 50m, and there is no safe solution to ensure safety for families below the foot of the waste site.

The road distance is a fundamental cause for the landslide because many roads cannot take advantage of the available natural terrain due to high technical standards. Therefore, the route had to go through many hills, and the construction took many earthworks. Consequently, this one made a lot of the high road's slopes (up to 90 m) in complex geological conditions were alternating with strongly weathered rock. This is easy to cause landslides. Additionally, distance to the river is also one of the essential factors. Distance to river depends on the network of rivers and streams. The denser the river and stream network, the greater the accumulative capacity of the flow. The proximity to drainage lines of intensive gully erosion is an important factor in controlling landslides.

Faulting is one of the expressions of contributed tectonics to slope instability. Along the faulted zones are good places for the weathering process to develop and create a thick weathered crust. This makes advantages for landslides to arise and develop. Therefore, distance to fault is an important factor for predicting landslides.

The lithology is a prevalent factor for landslides. All the hardness, the durability of rock, and weathering products from the bedrock reflected the role of the lithology factor.

3. Background of the methods used

3.1. Random subspace framework (RSF)

The method is proposed firstly by Ho (Ho, 1995; Ho, 1998), which is an ensemble classifier technique. This method constructed a decision tree-based classifier that maintains highest accuracy on training data and improved on generalization accuracy as it

grows in complexity. The classifier consists of multiple trees constructed systematically by pseudo-randomly selecting subsets of components of the feature vector, that is, trees constructed in randomly chosen subspaces (Barandiaran, 1998). The training data is modified in the feature space. Thus, each training incidence $\boldsymbol{Z}i$ (i = 1,…, n) in the training sample set $Z = [Z_1; \ldots; Z_n]$ is defined as a p-dimensional vector $\boldsymbol{Z}_i = (z_{i1}, z_{i2}, \ldots, z_{ip})$, defined by p features. Then, randomly $r < p$ features from the p-dimensional data set Z are selected. Consequently, the modified training set $\widetilde{Z^b} = Z_1^b, \widetilde{Z_2^b}, \ldots, \widetilde{Z_n^b}$, is composed of r-dimensional training incidences. After this step, classifiers are built into the random subspaces $\widetilde{Z^b}$ and aggregated by utilizing a majority voting. Therefore, the implementation of RSF is in the following way:

(1) Repeat for b = 1, 2,…, B:

(2) Choose an r-dimensional random subspace $\widetilde{Z^b}$ from the original p-dimensional feature space $\boldsymbol{Z}$.

(3) Build a classifier $C^b(z)$ (with a decision boundary $C^b(z) = 0$) in $\widetilde{Z^b}$.

(4) Aggregate classifiers $C^b(z)$, b =1, 2,…, B, by utilizing majority voting for the final decision.

The RSF can benefit from using random subspaces to build and combine the classifiers. When the number of training incidences is comparatively tiny compared to the data dimension, we can solve the small sample size problem by building classifiers in random subspaces. The subspace dimension will be less than the original feature space, while the number of training incidence remains unchanged. Thus, the relative training sample size increases. Once the data have several redundant features, we can find a better classifier in random subspaces than in the original feature space. The aggregated decision of such classifiers might be better than a single classifier built on the original training set in the entire feature space (Skurichina et al., 2002).

There are parameters to be tuned for Random Subspace ensemble learning algorithms. After many experiments, we achieved the best results by applying the following values for parameters.

Classifier: represents the base classifier for application. We applied 11 different classifiers such as ANN, k-NN, SVM, RF, C4.5, Random Tree, REP Tree, LAD Tree, NB, Rotation Forest, and CART.

Numerations: represents the number of repetitions for application. The best performance comes for a set up to 10.

Seed: represents the number of seeds for application in a random way. The best performance comes with a seed = 1 in implementing the random subspace.

Subspace size: represents the size of each subspace. The best performance comes with a subspace = 0.5 in implementing the random subspace.

3.2. C4.5 decision tree learning

We eliminated the C4.5 decision tree algorithm tests for which training examples have the same result. Therefore, they do not appear in the decision tree if they do not have a minimum of two outcomes, which have a minimum number of instances. The given value for the minimum is 2, yet we can control it and raise it for tasks with noisy data. Candidate splits are taken into consideration if they cut a specific number of instances. After the subtraction, we might find that the information gain is negative. If we do not have attributes that have a positive information gain, which is a kind of pre-pruning, the tree will stop growing. This is indicated at this point since it could be unexpected to obtain a pruned tree, although post pruning is not active (Witten et al., 2016). The implementation used the default parameters.

4. Proposed methodology

4.1. Shallow landslide database for the research area

Figure 3 shows the constructed inventory map with the total number of 3730 pixels of landslide occurrences. They are randomly sampled and consist of 1865 pixels of non-landslide and 1865 pixels of landslide. This data is for model validation and training model to train the model. The samples in sets of two groups are 1008 and 2722 (in 70:30 ratio) (Nhu et al., 2020; Nhu et al., 2021), respectively.

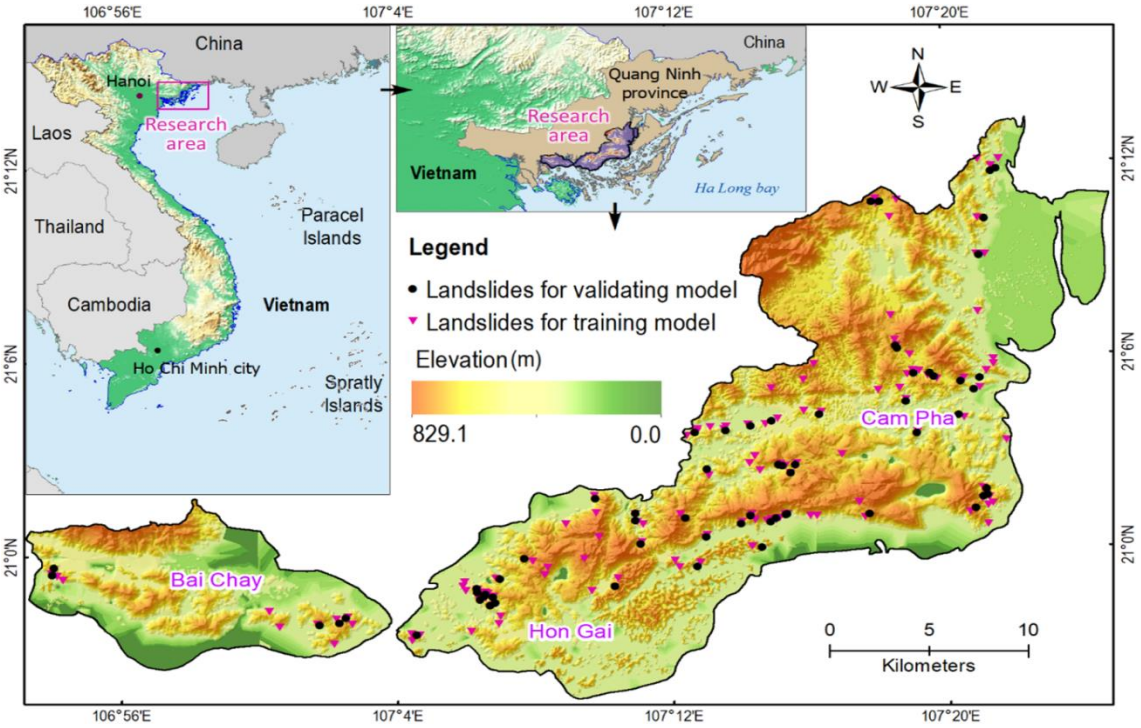

Figure 3. Location of Ha Long city (Bai Chay and Hon Gai) and Cam Pha city

To develop and ensure the accuracy of the landslide prediction model, it is essential to select the factors that cause the landslide. Accordingly, there are geo-environmental factors that contribute to landslides, including slope, aspect, curvature, TWI, land use, distance to road, distance to the river, soil type, distance to fault, and lithology (Hue et al., 2004; Yem et al., 2006; Bui et al., 2017; Hung et al., 2017) (Fig. 4).

Therefore, we created thematic maps for each mentioned factor. The extraction of four morphometric characteristics, Slope, Aspect, Curvature, and TWI, from the digital elevation model map (DEM) with a resolution of 25×25 m for this area, was based on the digital topographic maps 1:50.000 scale provided by the Ministry of Natural Resource and Environment of Vietnam. Accordingly, the Slope map was built with a slope ranging from 00 to 76.73 degrees (Fig. 4a). There are nine facing slopes used to create the Aspect map (Fig. 4b). The Curvature reflects the shape of the ground surface, which affects the occurrence of the landslide that changes from 45.1 to -35.5 degrees (Fig. 4c) (Shirzadi et al., 2017). The TWI in the area's soil changes between 4.6-24.1% (Fig. 4d). The land use

map is a part of the Status Land Use Project of the National Land Use Survey in Vietnam in 2010. We built the Landuse map (Fig. 4e) with 13 classes. The distance to the road map (Fig. 4f) and the distance to the river map (Fig. 4g) was extracted from the topology map at 1:50.000 scale (Ministry of Natural Resources and Environment, 2003). We also took the Soil type map (Figure 4h) from the Department of Agriculture and Rural Development of the Quang Ninh province with 10 soil types for the research area. Distance to faults was also a major factor in building the landslide model because this one makes the slope unstable (Brideau et al., 2009). To define optimal areas of impact zone along to the faults for landslide, we construct the buffer zones by the distance from 0 to over 1000 m along to the faults including 0-200, 200-600, 600-1000, and >1000 m (Fig. 4i). The geological map allows us to know information on underlying bedrock. This is an essential factor for landslide modeling (Ayalew and Yamagishi, 2005). The collection of the geological maps with eleven geologic units came from the General Department of Geology and Minerals of Viet Nam (Fig. 4j). More explanations of these factors can be found in Nhu et al. (2020). It should be noted that the data processing and coding were conducted using ArcGIS 10.4. and Weka 4.9 (Witten et al., 2016).

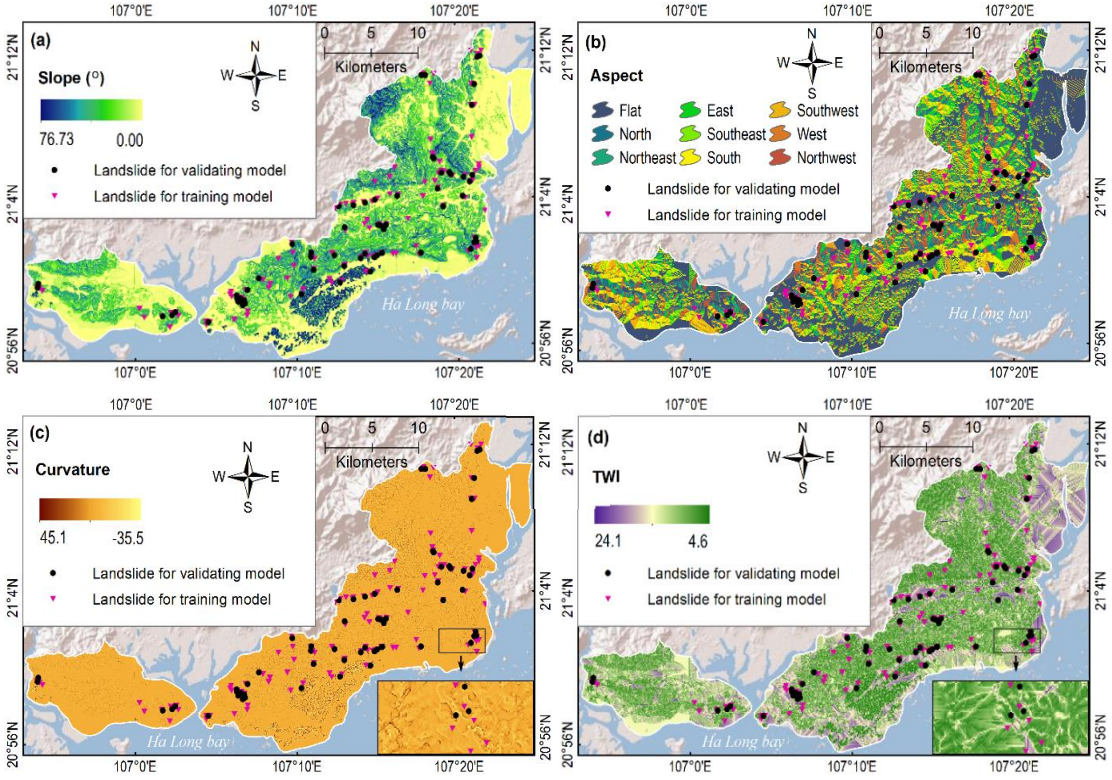

Figure 4. Landslide related factor; (a) Slope, (b) Aspect; (c) Curvature; (d) TWI; I Landuse; (f) Distance to road; (g) Distance to river. (h) Soil type; (i) Distance to fault; and (j) Lithology

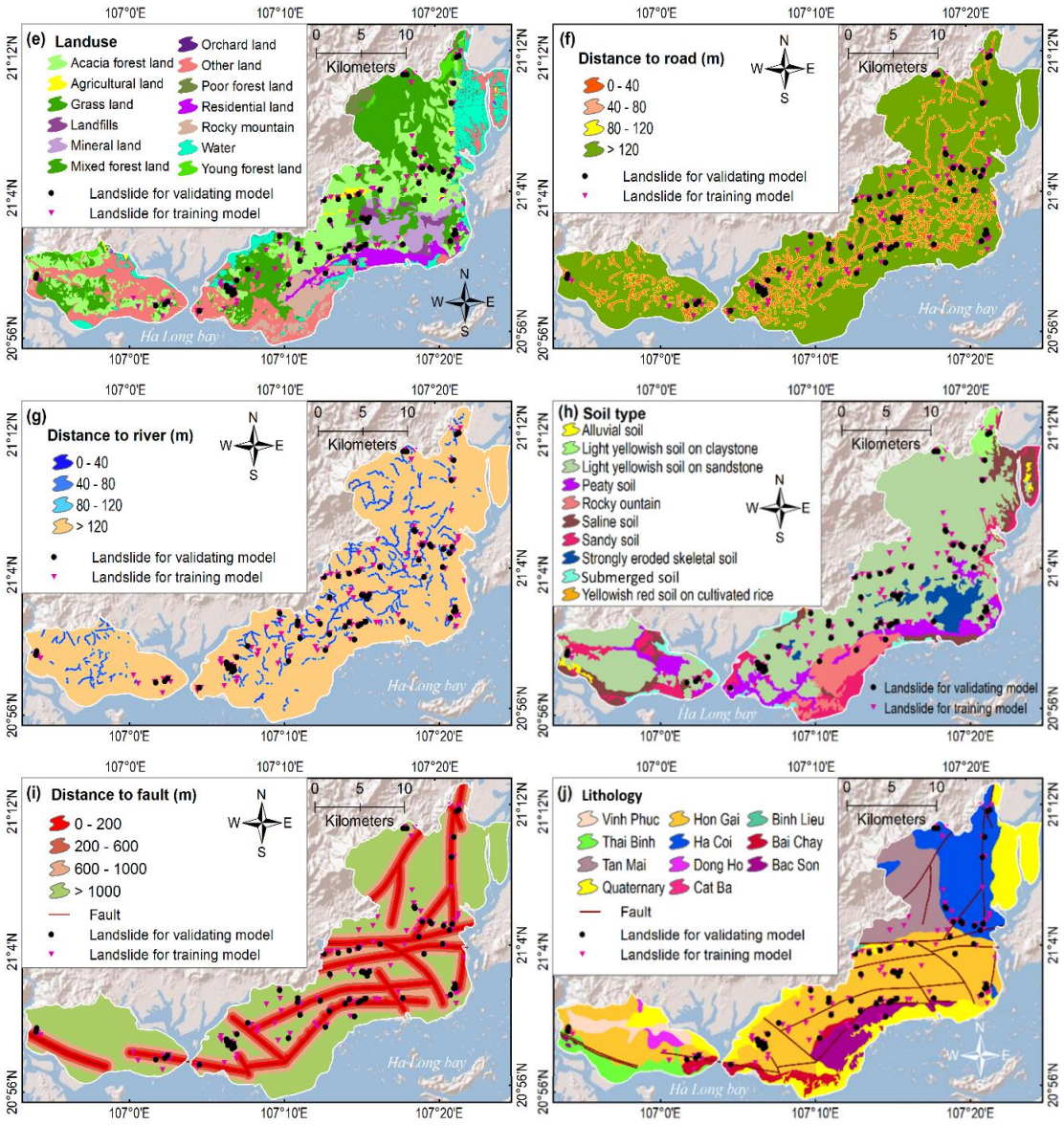

Figure 4. Cont.

4.2. Evaluation and verification of the shallow landslide data

In the beginning, the ArcCatalog application in ArcGIS software was used to establish a GIS database for building maps. This geodatabase file was utilized because of the capacity to host and process a considerable amount of geographic datasets with their various information types in just one document framework (Zeiller and Murphy, 2010). The GIS database consists of 170 landslide polygons and ten predisposing factors (slope, aspect, curvature, TWI, land use, distance to road, distance to river, soil type, distance to fault, and lithology). Then we transformed all the data to raster format with a 25 m to overcome the imbalance of absolute magnitudes (Dang et al., 2019. The categories of the ten predisposing factors were

coded and normalized (Bui Tien and Hoang, 2017; Truong et al., 2018).

In this landslide modeling, to evaluate the data Cross-validation technique was used that is proven efficient in data evaluation (Micheletti et al., 2014; Goetz et al., 2015). We split this data into training data and test data in the ratio of 70/30. Accordingly, we randomly chose 135 landslide polygons (70%, 2722 pixels) and used them for training the landslide models while using the remaining 58 landslides (30%, 1008 pixels) for the testing model. In this study, we applied the "on-off" classification method. The number of non-landslide pixels was also randomly picked in the not-yet landslide areas with slope angles less than 5° (Kavzoglu et al., 2014). Detailed discussions on sampling strategies can be found (Erener et al., 2017). In the next step, we constructed the training dataset and the validation dataset based on the values for all the pixels extracted from ten predisposing factors. Finally, the pixels were enciphered into "0-1" (Bui et al., 2017), in which the landslide pixels were assigned "1," and the non-landslide pixels were assigned "0".

To ensure the modeling result fits the objective, we built the landslide models using 10-folds cross-validation with the training dataset. We randomly split the training dataset into 10 equally sized subsets, using nine subsets for training, and one subset tested this model. This process was performed 10 times where each subset was once used as the testing dataset. The model was successfully trained using the training dataset with the 10-fold cross-validation procedure. The model was again validated using the validation dataset.

4.3. Shallow landslide model

In this study, we describe and present a new hybrid machine learning approach for Landslide Susceptibility Modeling for the first time in the research area: Ha Long and Cam Pha city, which is the combination between the C4.5 algorithm and Random subspace framework. The proposed approach is RandSub-DT. Figure 5 shows the Methodological concept of the proposed RandSub-DT model used in this study.

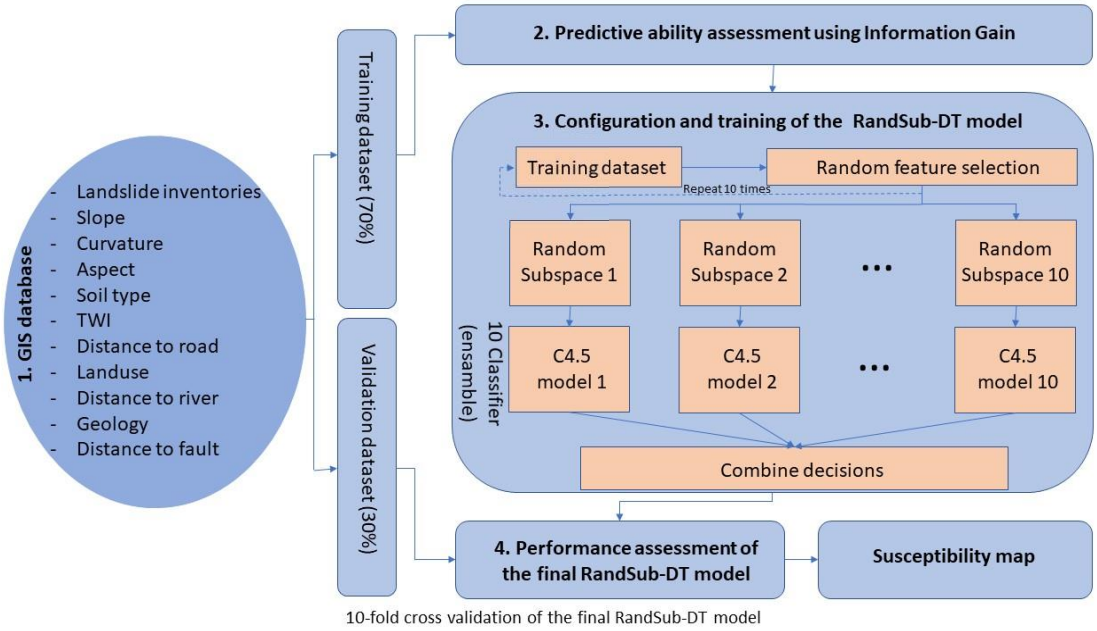

Figure 5. Methodological concept of the proposed RandSub-DT model used in this study

4.4. Model assessment and benchmark comparison

In this study, a confusion matrix was used to assess the RandSub-DT model on both the training and validation dataset because the model is considered as a binary system of pattern recognition (Bui et al., 2017). Based on the matrix, several statistical measures of the model are calculated, including Sensitivity (SEN), Specificity (SPE), Positive Predictive Power (PP2), Negative Predictive Power (NP2), Kappa index, and Classification Accuracy (CLA) (Bui et al., 2016). The SEN and SPE are the proportion of the landslide and non-landslide pixels concerning the correct prediction as landslide and non-landslide, respectively. The PP2 and NP2 are the exact predictive percentage of the model to landslide pixels and non-landslide pixels. CLA is the correct-prediction result of the model. If the CLA may not accurately classify the landslide pixels, more than an index for the assessment process was used. This is the Likelihood Ratio index (LLR) that assesses the trade-off of both SEN and SPE of landslide models. The higher the LLR value, the better the landslide model (Lagomarsino et al., 2015). The Kappa is the prediction performance of the model. The following formulas calculate the values of these indexes:

$$PP2 = \frac{TP}{TP + FP} \quad (1)$$

$$NP2 = \frac{TN}{TN + FN} \quad (2)$$

$$P_{exp} = \frac{(TP + FN)(TP + FP) + (FP + TN)(FN + TN)}{\sqrt{TP + TN + FN + FP}} \quad (3)$$

$$Kappa = \frac{CLA + P_{exp}}{1 - P_{exp}} \quad (4)$$

$$CLA = \frac{TP + TN}{TP + TN + FN + FP} \quad (5)$$

$$SEN = \frac{TP}{TP + FN} \quad (6)$$

$$SPE = \frac{TN}{TN + FP} \quad (7)$$

$$LLR = \frac{SEN}{1 - SPE} \quad (8)$$

Where TP (true positive) and TN (true negative) are the numbers of instances predicted correctly. FP (false positive) and FN (false negative) are the numbers of instances predicted erroneously. P_{exp} is the expected agreement (Pham et al., 2019).

According to many published articles, the ROC (Receiver Operating Characteristic) and the AUC (Lower Contour Area) are two necessary values to evaluate the performance of the RandSub-DT model (Lucà et al., 2011; Hoang and Tien Bui, 2016; Bui et al., 2017) entirely. The ROC curve is advantageous to confirm the predictive accuracy of models. The closer the curve is to the upper left corner, the better the performance of the slide model (Truong et al., 2018). AUC is employed for quantitative confirmation of models with excellent (AUC belong to 0.9-1), good (AUC belong to 0.8-0.9), fair (AUC belong to 0.7-0.8), and poor (AUC is less than 0.7) (Cantor and Kattan, 2000).

4.5. Compute shallow landslide susceptibility

Suppose the final RandSub-DT model is satisfied in the performance assessment check. In that case, it will go through the calculation of the susceptibility index for all the pixels of the research area. Next, using a Python application to convert these susceptibility indices to the ASCII raster format in ArcGIS. Finally, the landslide susceptibility map is classified into five classes: very high, high, moderate, low, and very low (Pradhan et al., 2010).

5. Result and discussion

5.1. Checking result of the shallow landslide data

Table 1 shows the result of the predictive ability evaluation of the ten predisposing factors. The 10-fold cross-validation was used

in this procedure to ensure the assessment result is stable, as suggested (Fushiki and Computing, 2011). The average merit (AM) of Distance to the river is the highest predictive value 0.907. It is reasonable because many landslides in this study area occurred near rivers and streams, at distances 0-80 m. Following Distance to the river is Distance to the road (AM of 0.768), Distance to a fault (AM of 0.641), Aspect (AM of 0.486), Curvature (AM of 0.442), and Lithology (AM of 0.431). In contrast, the figures of the others are moderately lower with TWI (AM of 0.259), Landuse (AM of 0.233), Slope (AM of 0.226), and Soil type (AM of 0.144) (Table 1).

Table 1. Predictive ability of ten landslide predisposing factors using Pearson technique and 10-fold cross validation techniques

No.	Predisposing Factors	Average Merit	Standard Deviation
1	TWI	0.259	0.135
2	Slope	0.226	0.166
3	Aspect	0.486	0.310
4	Distance to road	0.768	0.351
5	Lithology	0.431	0.236
6	Soil type	0.144	0.279
7	Distance to river	0.907	0.236
8	Distance to fault	0.641	0.358
9	Curvature	0.442	0.022
10	Landuse	0.233	0.229

Table 1 shows the good results with the high average merit values, which are essential factors. They are consistent with the characteristics of landslide occurrence in the research area, such as Distance to the river, Distance to the road, Distance to a fault. The research area, located along the coast with relatively complex geological tectonics, has mining activities. This is mentioned in (Van Den Eeckhaut et al., 2006; Costanzo et al., 2012). From the table, we could see that all predisposing factors point out predictive values relative to the building landslide model; therefore, we concluded that they are all relevant factors and are used in this analysis.

5.2. Model performance

Ten predisposing factors are essential to generate the database for building the RandSub-DT model. We trained the model using the training dataset with the 10-fold cross-validation technique. It is clear from the training result tables (Table 2 and 3) that the RandSub-DT model performed very well with the training dataset. According to the tables, the high relevant degree of the model to the dataset with the CLA value of 90.34% (Table 2). The agreeable degree of the model and the training dataset is good at 0.8068 (Kappa) (Table 2). Additionally, the percentage of the non-landslide pixels is correctly split with the SPE value of the RandSub-DT model is 98.67% (Table 2). In comparison, the figure for landslide pixels is slightly lower at 84.44% (Table 2). In contrast, the classified-pixels probability of the model to the landslide class is very high at 98.9% (PP2) (Table 2), compared to 81.78% of the figure for the non-landslide type (NP2) (Table 2).

Table 2. Performance measures of the RandSub-DT model using the training dataset

No.	Statistical index	Performance measures
1	True positive	1346
2	True negative	1113
3	False positive	15
4	False negative	248
5	Positive predictive value (%)	98.9
6	Negative predictive value (%)	81.78
7	Sensitivity (%)	84.44
8	Specificity (%)	98.67
9	Classification Accuracy (%)	90.34
10	Kappa	80.68
11	AUC	98.00

To assess the influence of landslide factors on the RandSub-DT-model building, we eliminated each element and then recalculated the corresponding classification accuracy value (CLA) for the evaluation process. The CLA-value decrease of the RandSub-DT model when removing one or more factors indicates the influence level of these factors on the model. The result from Table 3 reveals that the influence of the factors on the construction of the model ranges from 4.77% to 8.26%. Thus, it is reasonable and necessary to use all factors for this research.

Table 3. Contribution of the landslide predisposing factors to the RandSub-DT model

No.	Removing Factor	Classification Accuracy - CLA (%)
1	TWI	85.49
2	Slope	84.35
3	Aspect	83.69
4	Distance to road	84.35
5	Lithology	83.47
6	Soil	84.50
7	Distance to river	85.56
8	Distance to fault	83.25
9	Curvature	85.38
10	Landuse	82.07

After training the Randsub-DT model with the training dataset, we continued the estimation with the validation dataset and the result in Table 4. Noticeably, the prediction result is relatively high at 77.48% (CLA). The Kappa of the RandSub-DT is 0.5496, indicating that the model's prediction performance is better than other models. The exact predictive percentage of the model to landslide pixels is 76.98% (PP2), and the figure for non-landslide pixels is 77.96 (NP2). The proportion of the correctly predicted landslide pixels was 77.76% (SEN), and the model (SPE) correctly predicted 77.21% of non-landslide pixels

Table 4. Performance measures of the RandSub-DT model using the validation dataset

No.	Statistical index	Performance measures
1	True positive	388
2	True negative	393
3	False positive	116
4	False negative	111
5	Positive predictive value (%)	76.98
6	Negative predictive value (%)	77.98
7	Sensitivity (%)	77.76
8	Specificity (%)	77.21
9	Classification Accuracy (%)	77.48
10	Kappa	54.96
11	AUC	86.4

5.3. Shallow landslide susceptibility map

The establishment of the final RandSub-DT model is used to calculate the sensitivity index for all pixels of the research area. Accordingly, these susceptibility indices were converted to the ASCII raster format in ArcGIS using Python, which was finally followed by creating the landslide susceptibility map for the research area. Then, we used the Random subspace framework optimized by the C4.5 algorithm to calculate landslide susceptibility indices for the research area. All the influencing factors were converted to raster format and then fed to the RandSub-DT model to generate susceptibility indices called landslide probability index. The classification of these indexes came to light based on the influence level of the factors on landslide probability occurrence or susceptibility. Accordingly, Figure 6 shows the landslide susceptibility map for the area in the cartographical presentation with values ranging from 0 to 1).

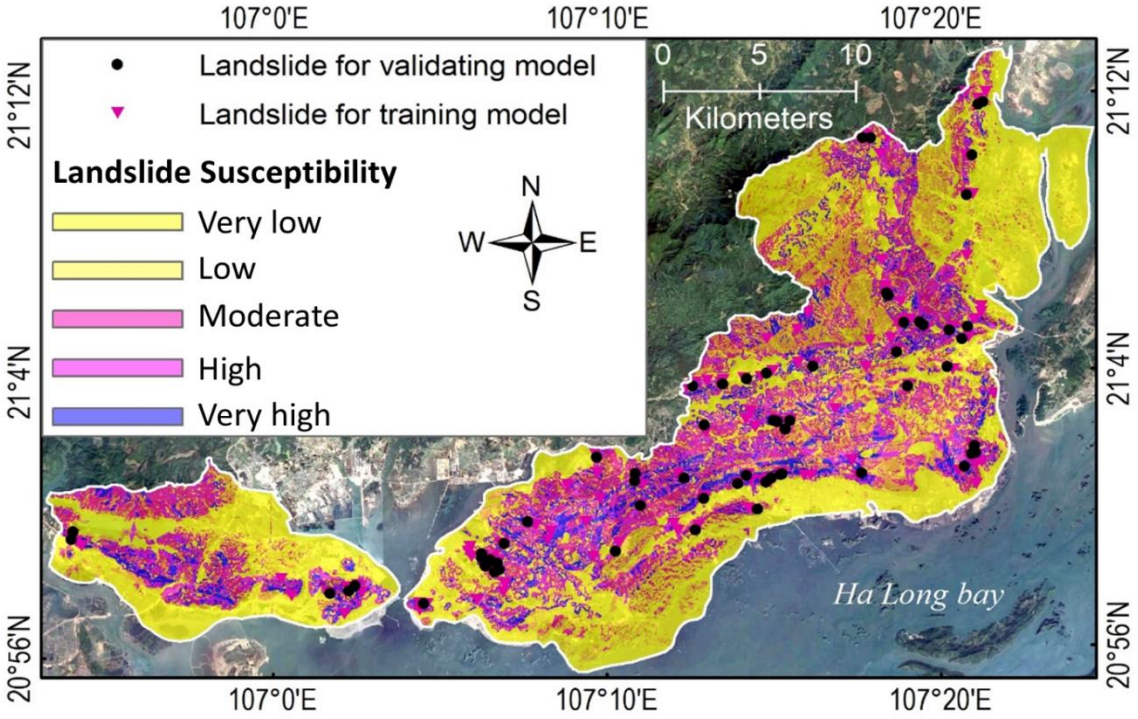

Figure 6. Shallow landslide susceptibility map generated by the proposed RandSub-DT model

6. Concluding remarks

This paper proposes a new modeling approach for landslide susceptibility mapping in Ha Long and Cam Pha city. This is a hybrid intelligence, called the RandSub-DT model. Up to now, no study has used this model for landslide research, except this study. The model generation relies on the GIS database, including 170 landslide polygons and ten predisposing landing factors. This GIS database aimed to construct and verify the RandSub-DT model. Using confusion matrices, we can check the quality of the final RandSub-DT model.

The results in this study proved that the new RandSub-DT model can perform well in landslide susceptibility mapping with high accuracy with the CLA is 90.34%. And the prediction result is good when the quantitative confirmation of models is excellent with AUC belonging to 0.9-1.0 on the train set. The impact on the validation dataset is 77.48% of accuracy, and the quantitative confirmation of models is good (AUC is 0.864).

The cartographical presentation of the research area shows the mapping result including five classes from the deficient class to the very high class: very high (8.4%), high (11.9%), moderate (15.1%), low (17.3%), and deficient (47.3%). The areas with a high probability of landslide cover 47 square kilometers. These areas should receive more attention in developing remedial measures for landslide prevention.

The main limitation of this study is that the parameters used in the RandSub-DT model are optimized; therefore, powerful optimization algorithms should be considered for searching and finding these parameters autonomously to improve the prediction performance of the landslide model.

Acknowledgements

This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 105.08-2017.316.

Reference

Akgun A., et al., 2012. An easy-to-use MATLAB program (MamLand) for the assessment of landslide susceptibility using a Mamdani fuzzy algorithm. Computers & Geosciences, 38(1), 23-34.

Ayalew L., H.J.G. Yamagishi, 2005. The application of GIS-based logistic regression for landslide susceptibility mapping in the Kakuda-Yahiko Mountains. Central Japan. Geomorphology, 65(1-2), 15-31.

Barandiaran, et al., 1998. The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(8), 1-22.

Binh Do Le, et al., 1968. The report of the industrial geological mapping in Hon Gai - Cam Pha area at 1:50.000 scale. General Department of Geology and Minerals of Viet Nam.

Brideau M.-A., M. Yan, D.J.G. Stead, 2009. The role of tectonic damage and brittle rock fracture in the development of large rock slope failures. Geomorphology, 103(1), 30-49.

Bui Tien D., N.-D. Hoang, 2017. A Bayesian framework based on a Gaussian mixture model and radial-basis-function Fisher discriminant analysis (BayGmmKda V1. 1) for spatial prediction of floods. Geoscientific Model Development, 10(9), 1-19.

Bui D.T., et al., 2017. A novel fuzzy K-nearest neighbor inference model with differential evolution for spatial prediction of rainfall-induced shallow landslides in a tropical hilly area using GIS. Landslides, 14, 1-17.

Bui D.T., et al., 2016. Spatial prediction models for shallow landslide hazards: a comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree. Landslides, 13, 361-378.

Bui D.T., et al., 2017. Spatial prediction of rainfall-induced landslides for the Lao Cai area (Vietnam) using a hybrid intelligent approach of least squares support vector machines inference model and artificial bee colony optimization. Landslides, 14, 447-458.

Bui D.T., et al., 2016. GIS-based modeling of rainfall-induced landslides using data mining-based functional trees classifier with AdaBoost, Bagging, and MultiBoost ensemble frameworks. Environmental Earth Sciences, 75, 1101.

Cantor S.B., et al., 2000. Determining the area under the ROC curve for a binary diagnostic test. Medical Decision Making, 20, 468-470.

Costanzo D., et al., 2012. Factors selection in landslide susceptibility modelling on large scale following the gis matrix method: application to the river Beiro basin (Spain). Natural Hazards and Earth System Sciences, 12, 327-340.

Dang V.-H., et al., 2019. Enhancing the accuracy of rainfall-induced landslide prediction along mountain roads with a GIS-based random forest classifier. Bulletin of Engineering Geology and the Environment, 78, 2835-2849.

Erener A., et al., 2017. Analysis of training sample selection strategies for regression-based quantitative landslide susceptibility mapping methods. Computers & Geosciences, 104, 62-74.

Fushiki T.J.S., Computing, 2011. Estimation of prediction error by using K-fold cross-validation. Statistics and Computing, 21, 137-146.

Gheshlaghi H.A., et al., 2017. An integrated approach of analytical network process and fuzzy based spatial decision-making systems applied to landslide risk mapping. Journal of African Earth Sciences, 133, 15-24.

Goetz J., et al., 2015. Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling. Computers & Geosciences, 81, 1-11.

Ho T.K., 1995. Random decision forests. Proceedings of 3rd international conference on document analysis and recognition. Proceedings of 3rd International Conference on Document Analysis and Recognition. IEEE Publishers, 278-282.

Ho T.K., 1998. The random subspace method for constructing decision forests. IEEE transactions on

pattern analysis machine intelligence, 20(8), 832-844.

Hoang N.-D., et al., 2016. A novel relevance vector machine classifier with cuckoo search optimization for spatial prediction of landslides. Journal of Computing in Civil Engineering, 30(5), 04016001.

Hong H., et al., 2017. A novel hybrid integration model using support vector machines and random subspace for weather-triggered landslide susceptibility assessment in the Wuning area (China). Environmental Earth Sciences, 76(19), 652.

Hue T., et al., 2004. Investigation and assessment of the types of geological hazard in the territory of Vietnam and recommendation of remedial measures. Phase II: A Study of the Northern Mountainous Province of Vietnam. Institute of Geological Sciences, Vietnam Academy of Science and Technology, Hanoi, 361pp (in Vietnamese).

Hung Le, et al., 1996. The report of the industrial geological mapping in Cam Pha, Quang Ninh at 1:50.000 scale. Hanoi, General Department of Geology and Minerals of Viet Nam.

Hung L.Q., et al., 2017. Landslide inventory mapping in the fourteen Northern provinces of Vietnam: achievements and difficulties. Workshop on World Landslide Forum. Springer Publishers, 501-510.

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

Lagomarsino D., et al., 2015. Quantitative comparison between two different methodologies to define rainfall thresholds for landslide forecasting. Natural Hazards and Earth System Sciences Discussions, 3(1), 891-917.

Lucà F., M. Conforti, G.J.G. Robustelli, 2011. Comparison of GIS-based gullying susceptibility mapping using bivariate and multivariate statistics: Northern Calabria, South Italy. Geomorphology, 134(3-4), 297-308.

Meng Q., et al., 2016. GIS-based landslide susceptibility mapping with logistic regression, analytical hierarchy process, and combined fuzzy and support vector machine methods: a case study from Wolong Giant Panda Natural Reserve. China. Bulletin of Engineering Geology and the Environment, 75, 923-944.

Micheletti N., et al., 2014. Machine learning feature selection methods for landslide susceptibility mapping. Mathematical Geosciences, 46, 33-57.

Ministry of Natural Resources and Enviroment, 2003. Mapping topology in Quang Ninh scale 1:50.000, Department of survey.

Nhu V.-H., et al., 2020. Effectiveness assessment of Keras based deep learning with different robust optimization algorithms for shallow landslide susceptibility mapping at tropical area. Catena, 188, 13. https://doi.org/10.1016/j.catena.2020.104458.

Nhu V.-H., et al., 2021. An approach based on socio-politically optimized neural computing network for predicting shallow landslide susceptibility at tropical areas. Environmental Earth Sciences, 80, 1-18. https://doi.org/https://doi.org/10.1007/s12665-021-09525-6.

Pham B.T., et al., 2019. A novel hybrid intelligent model of support vector machines and the MultiBoost ensemble for landslide susceptibility modeling. Bulletin of Engineering Geology and the Environment, 78, 2865-2886.

Pham B.T., et al., 2018. Bagging based Support Vector Machines for spatial prediction of landslides. Environmental Earth Sciences, 77, 146.

Pham B.T., et al., 2017. Landslide hazard assessment using random subspace fuzzy rules-based classifier ensemble and probability analysis of rainfall data: a case study at Mu Cang Chai District, Yen Bai Province (Viet Nam). Journal of the Indian Society of Remote Sensing, 45, 673-683.

Pham B.T., et al., 2016. Rotation forest fuzzy rule-based classifier ensemble for spatial prediction of landslides using GIS. Natural Hazards 83, 97-127.

Pham B.T., et al., 2017. Hybrid integration of Multilayer Perceptron Neural Networks and machine learning ensembles for landslide susceptibility assessment at Himalayan area (India) using GIS. Catena, 149, 52-63.

Pham B.T., et al., 2018. Spatial prediction of landslides using a hybrid machine learning approach based on

random subspace and classification and regression trees. Geomorphology, 303, 256-270.

Pradhan B., et al., 2010. Landslide susceptibility assessment and factor effect analysis: backpropagation artificial neural networks and their comparison with frequency ratio and bivariate logistic regression modeling. Environmental Modelling & Software, 25(6), 747-759.

Quinlan J.R., 2014. C4.5: programs for machine learning. Elsevier Publisher, 58-60.

Reichenbach P., et al., 2018. A review of statistically based landslide susceptibility models. Earth-Science Reviews, 180, 60-91.

Shirzadi A., et al., 2017. Rock fall susceptibility assessment along a mountainous road: an evaluation of bivariate statistic, analytical hierarchy process and frequency ratio. Environmental Earth Sciences, 76, 152.

Skurichina M., et al., 2002. Bagging, boosting and the random subspace method for linear classifiers. Pattern Analysis & Applications, 5, 121-135.

Tien Bui D., et al., 2016. Spatial prediction of rainfall-induced shallow landslides using hybrid integration approach of Least-Squares Support Vector Machines and differential evolution optimization: a case study in Central Vietnam. International Journal of Digital Earth, 9(11), 1077-1097.

Truong X., et al., 2018. Enhancing prediction performance of landslide susceptibility model using hybrid machine learning approach of bagging ensemble and logistic model tree. Applied Sciences, 8(7), 1046.

Van Den Eeckhaut M., et al., 2006. Prediction of landslide susceptibility using rare events logistic regression: a case-study in the Flemish Ardennes (Belgium). Geomorphology, 76, 392-410.

Witten I.H., et al., 2016. Data Mining: Practical Machine Learning Tools and Techniques (Morgan Kaufmann Series in Data Management Systems), 4th Edition. Morgan Kaufmann Publishers, 654pp.

Yem N., et al., 2006. Assessment of landslides and debris flows at some prone mountainous areas Vietnam and recommendation of remedial measures. State-level independent project (KC-08-01BS). Geology Institute - Vietnam Academy of Science and Technology, 145pp.

Zeiller M., J. Murphy, 2010. Modeling Our World: The ESRI Guide to Geodatabase Concepts, Second Edition. ESRI Press, 308pp.