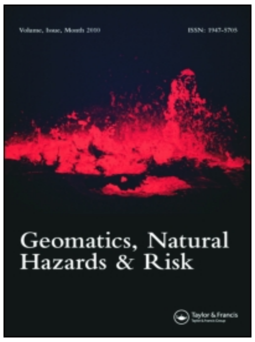

Geomatics, Natural Hazards and Risk

ISSN: 1947-5705 (Print) 1947-5713 (Online) Journal homepage: http://www.tandfonline.com/loi/tgnh20

A novel ensemble classifier of rotation forest and Naïve Bayer for landslide susceptibility assessment at the Luc Yen district, Yen Bai Province (Viet Nam) using GIS

Binh Thai Pham, Dieu Tien Bui, M.B. Dholakia, Indra Prakash, Ha Viet Pham, Khalid Mehmood & Hung Quoc Le

To cite this article: Binh Thai Pham, Dieu Tien Bui, M.B. Dholakia, Indra Prakash, Ha Viet Pham, Khalid Mehmood & Hung Quoc Le (2017) A novel ensemble classifier of rotation forest and Naïve Bayer for landslide susceptibility assessment at the Luc Yen district, Yen Bai Province (Viet Nam) using GIS, Geomatics, Natural Hazards and Risk, 8:2, 649-671, DOI: 10.1080/19475705.2016.1255667

To link to this article: https://doi.org/10.1080/19475705.2016.1255667

GEOMATICS, NATURAL HAZARDS AND RISK, 2017
VOL. 8, NO. 2, 649–671
http://dx.doi.org/10.1080/19475705.2016.1255667

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

A novel ensemble classifier of rotation forest and Naïve Bayer for landslide susceptibility assessment at the Luc Yen district, Yen Bai Province (Viet Nam) using GIS

Binh Thai Pham [a,b], Dieu Tien Bui [c], M.B. Dholakia[d], Indra Prakash [e], Ha Viet Pham[f], Khalid Mehmood[e] and Hung Quoc Le[f]

[a]Department of Civil Engineering, Gujarat Technological University, Ahmedabad , Gujarat, India; [b]Department of Geotechnical Engineering, University of Transport Technology, Thanh Xuan, Ha Noi, Viet Nam; [c]Geographic Information System Group, Department of Business Administration and Computer Science, University College of Southeast Norway, Bø i Telemark, Norway; [d]Department of Civil Engineering, LDCE, Gujarat Technological University, Ahmedabad, Gujarat, India; [e]Department of Science & Technology, Bhaskarcharya Institute for Space Applications and Geo-Informatics (BISAG), Government of Gujarat, Gandhinagar, India; [f]Vietnam Institute of Geosciences and Mineral Resources, Thanh Xuan, Hanoi, Vietnam

ABSTRACT

The objective of this study is to attempt a new soft computing approach for assessment of landslide susceptibility in the Luc Yen district, Yen Bai province (Viet Nam) using a novel classifier ensemble model of Naïve Bayes and Rotation Forest. First, history of 95 landslide locations was identified byfield investigations and interpretation of aerial photos. Also, the total ten landslide causal factors were selected (slope, aspect, elevation, curvature, lithology, land use, distance to roads, distance to rivers, distance to faults, and rainfall) to evaluate the spatial relationship with landslide occurrences. Information Gain technique is carried out to quantify the predictive capability of these factors. Second, landslide susceptibility assessment was carried out utilizing the novel classifier ensemble model. Finally, the performance of landslide model was validated using receiver operating characteristic curve technique, and statistical index-based evaluations. The novel classifier ensemble model indicates high prediction capability (AUC = 0.846) and relatively high accuracy (ACC = 78.77%). The study reveals that this model performs well in comparison to the other landslide models such as AdaBoost, Bagging, MultiBoost, and Random Forest. Overall, the novel classifier ensemble model is a promising method that could be used for landslide susceptibility assessment.

ARTICLE HISTORY
Received 29 October 2015
Accepted 27 October 2016

KEYWORDS
Landslides; GIS; Naïve Bayes; Rotation Forest; Viet Nam

1. Introduction

Landslide is known as one of the most serious natural hazards having devastating effects on human life and infrastructures (Tsangaratos et al. 2013; Alimohammadlou et al. 2014). All over the world, there were 2620 deadly landslide events occurred within 6 years from 2004 to 2010, killing a total of 32,322 people (Petley 2012). In Asia, approximately 18,000 people died and about 5.5 million people have been affected due to landslides during the period of 1950–2009, and the number of landslides in this region is relatively high in comparison to other regions of the world (EM-DAT 2010).

CONTACT Binh Thai Pham ✉ phambinhgtvt@gmail.com; binhpt@utt.edu.vn

Viet Nam is one of the top six countries frequently affected by natural disasters including landslides (Guha-Sapir et al. 2011). Over the years, landslides have occurred frequently in the country especially in north-western mountainous and hilly regions (Tien Bui 2012). However, only limited studies of landslides have been carried out in this region (Tien Bui et al. 2013; Tien Bui et al. 2015).

Landslide susceptibility assessment is considered as an appropriate solution for reducing landslide damages through proper land use planning (Fell et al. 2008). On regional scales, the assessment is based on the statistical assumption that landslide events in the future will occur under the same conditions that happened in the past (Guzzetti et al. 2005). Many methods and techniques have been developed for the landslide susceptibility assessment during last decade. These methods can be broadly grouped into two categories (1) qualitative methods and (2) quantitative methods (Guzzetti 2006). Qualitative methods are relatively subjective approaches which are based on expert's perspective for defining the parameters and giving weights (Castellanos Abella 2008). Quantitative methods are more objective which are based on criteria for selecting and assigning the weight for variables (Castellanos Abella 2008). Therefore, quantitative methods are preferable for landslide susceptibility assessment.

Many quantitative methods have been applied widely in landslide problems in recent years such as frequency ratio (Pham et al. 2015a, Youssef et al. 2015), evidential belief function (Jebur et al. 2015; Tien Bui et al. 2015), multi-criteria decision analysis (Gorsevski & Jankowski 2010; Dragićević et al. 2015), artificial neural networks (Conforti et al. 2014; Polykretis et al. 2015), support vector machine (Jebur et al. 2015; Ren et al. 2015; Pham et al. 2016a), decision tree (Lombardo et al. 2015, Tsangaratos & Ilia 2016), and logistic regression (Shahabi et al. 2015; Youssef 2015). These methods usually use new soft computing techniques that perform better than conventional methods and techniques (Pham et al. 2016d).

Even though these models have been applied successfully and efficiently in landslide susceptibility assessment, no model is totally perfect. Therefore, the improvement in these models is needed to achieve desire results. The performance of landslide models can be enhanced by using feature selection and ensemble frameworks (Tien Bui et al. 2014). The feature selection could quantify the predictive ability of landslide causal factors. Thereafter, the factors with non-predictive ability would have to be removed to improve the performance of landslide models (Martínez-Álvarez et al. 2013). Whereas, the ensemble frameworks that combine multiple classifiers to improve the performance of individual classifiers based on characteristics of the diversity (Kuncheva 2014).

Ensemble frameworks started in 1990s but received significant attentions of researchers in recent years. Ensemble techniques such as Bagging (Breiman 1996); AdaBoost (Freund & Schapire 1997); Random Subspace (Ho 1998); MultiBoost (Webb 2000); Random Forest (Breiman 2001); and Rotation Forest (Rodriguez et al. 2006) have been applied efficiently in improvement of the performance of individual classifiers for different problems. Out of these, Rotation Forest technique has resulted better outcomes (Rodriguez et al. 2006). Despite its merit, application of these ensemble frameworks for landslide models is still rare. Therefore, the main objective of present study is to attempt a novel classifier ensemble data mining approach for landslide susceptibility assessment at the Luc Yen district in Yen Bai province (Viet Nam). This method is a combination of Naïve Bayes classifier and Rotation Forest ensemble. These two methods are current state-of-the-art techniques but they have so far been seldom used for landslide models. In addition, the performance of the novel classifier ensemble model was also compared with other ensemble models such as Bagging, AdaBoost, MultiBoost, and Random Forest.

2. Study area

The study area of Luc Yen district (latitudes 21° 55′30″N to 22°17′30″N and longitudes 104°30′00′E to 105°53′33″E), which is located in the northeast of the Yen Bai province in Viet Nam, is affected by numerous landslides every year (Figure 1). It covers an area of about 810 km^2 that is 1.2% area of

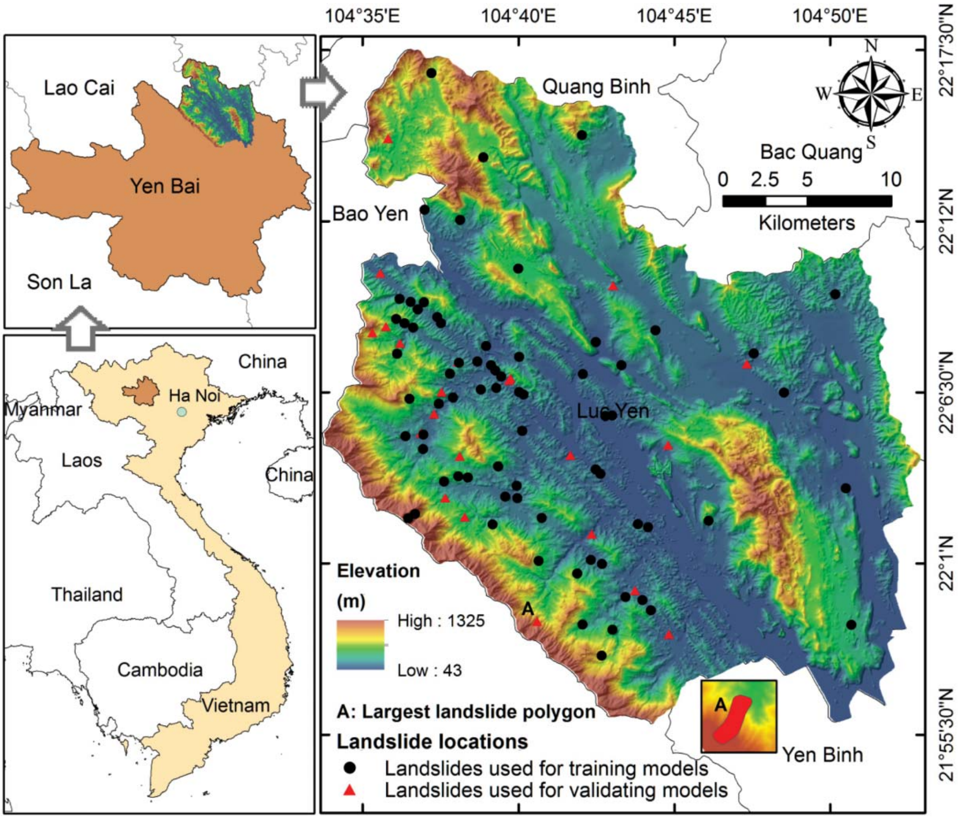

Figure 1. Landslide inventory map of study area.

the Yen Bai province. As per record, the population of Luc Yen district in the year 2010 was 10,3587 and average populated density was 120 people per km^2.

Luc Yen district is a mountainous region occupied by hills, small valleys, mounts, cliffs, and plains. The district is dissected by two dominant mountain ranges running in northwest-southeast direction namely Nui Voi and Large Rock mountains. Elevation in the area ranges from 43 to 1325 m above standard sea level, with an average elevation of 262m. Slope angles in the region vary from nearly flat to 81^o. Approximately 29.71% of the study area has very gentle slopes under 8°, and around 12.93% falls into slopes from 8^o to 15^o. Slopes in the range of 15°–25° occupy about 26.58% of the study area whereas 20.93% of the study area belongs to slopes of 25°–35°. Around 7.96% of the study area has slopes between 35° and 45°. Only 1.89% area is having slopes greater than 45° (Figure 6).

Geologically, there are eight main geological formations (Nui Voi, Ngoi Chi, Thac Ba, Phan Luong, An Phu, Tu Le, Ha Giang, and Nui Chua) in the study area. Different types of rocks (sedimentary, igneous and metamorphic) exist in the study area. Predominant rocks in the area are metamorphic (48%), whereas igneous rocks are occupying only 5.4% area. Alluvium and recent deposits are also present at places (Figure 6).

Different types of land use patterns have been observed in this area namely forest; barren; cultivation; grass; scrub, and residential area. Forest land occupies the largest area (68.07%), followed by barren and cultivation lands (15.09%), grass and scrub lands (7.36%), and residential area (4.5%). Water bodies occupy only 4.98% of the total area (Figure 6).

Luc Yen district is situated in the tropical monsoon region, thus regularly experiencing heavy rainfall during the months of June, July, and August. The annual average rainfall varies from 2500 to 3550 mm. Rainfall usually occurs with high intensity and over a short period of time, often triggering landslides,

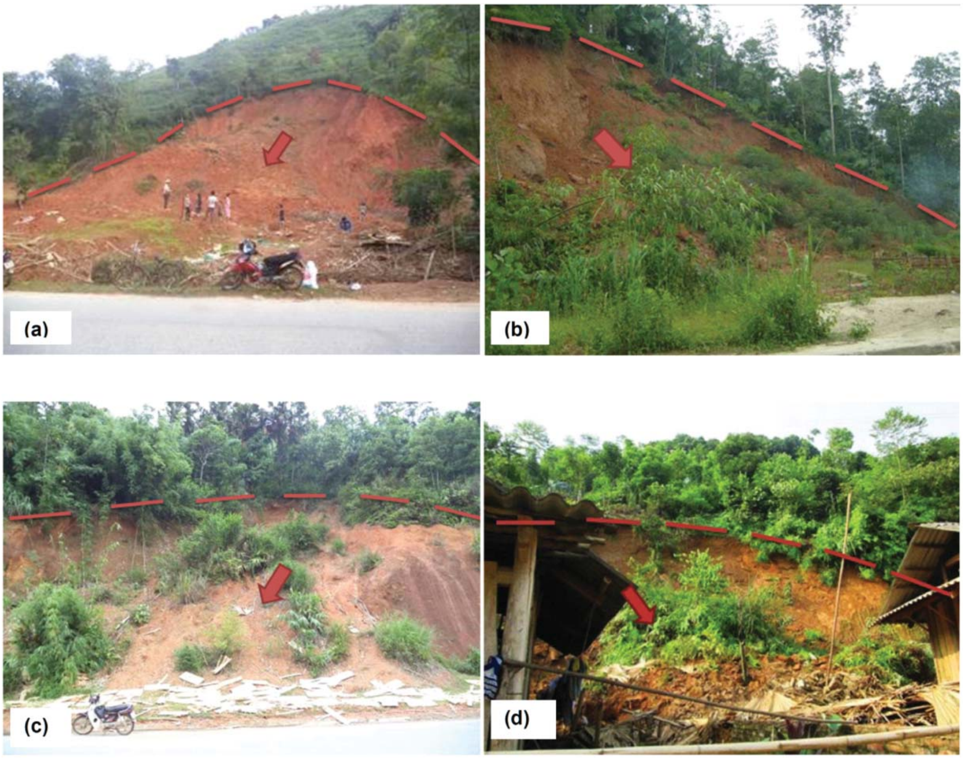

Figure 2. Examples of landslides in the study area: (a) and (b) landslides at the Tan Linh commune; (c) landslide at the Khanh Hoa commune; (d) landslide at Phuc Loi commune.

flooding, and causing erosion in the study area. The average daily temperature is 22°C. The temperature in the area varies from 2°C to 40°C. The average daily humidity ranges between 60% and 72%.

3. Materials and methodology

Landslide susceptibility analysis has been carried out in five main steps: (i) data collection from various sources, (ii) preparation of dataset, (iii) evaluation of prediction capability of landslide causal factors, (iv) assessment of landslide susceptibility using the novel classifier ensemble model, and (v) validation and comparison of landslide models.

3.1. Landslide inventory map

Preparation of landslide inventory map is considered as a primary and important step for landslide susceptibility assessment (Fell et al. 2008). The map indicates the location of landslide events that occurred in the past as well as in present. To construct a landslide inventory map, consultation of literature and interpretation of high-resolution satellite images/air photos are being done in conjunction with field investigation (Xu et al. 2012; Pradhan 2013).

Landslide inventory map in this study was constructed with the help of air photos (1:33.000) of the year in 2013 obtained from the Aerial Photo-Topography Company (Vietnam). Interpretations were carried out under a current national project in Viet Nam at the Vietnam Institute of Geosciences and Mineral Resources, namely ‘Survey, assessment and zoning of landslide warning in the mountainous region of Vietnam’. Field investigations were also carried out to check the interpretation results. Figure 2 shows photos of landslides in the study area that were taken during the field work phase.

Table 1. Landside causal factors and their classes employed in this study.

No.	Landslide causal factors	Classes
1	Slope (°)	(1) [0,8); (2) [8,15); (3) [15,25); (4) [25,35); (5) [35,45); and (6) $\geq$ 45
2	Aspect	(1) flat; (2) north; (3) northeast; (4) east; (5) southeast; (6) south; (7) southwest; (8) west; and (9) northwest
3	Elevation (m)	(1) [0,100); (2) [100,200); (3) [200,400); (4) [400,700); and (5) $\geq$ 700
4	Curvature	(1) concave (< -0.05); (2) flat [-0.05,0.05]; and (3) convex (> 0.05)
5	Lithology	(1) group 1; (2) group 2; (3) group 3; (4) group 4; (5) group 5; (6) group 6; and (7) group 7
6	Land use	(1) forests; (2) grass & scrub lands; (3) barren & cultivated lands; (4) residential area; and (5) water
7	Distance to Faults (m)	(1) [0,100); (2) [100,200); (3) [200,400); (4) [400,700); (5) [700,1000); and (6) $\geq$ 1000
8	Distance to Roads (m)	(1) [0,40); (2) [40,80); (3) [80,120); and (4) $\geq$ 120
9	Distance to Rivers (m)	(1) [0,40); (2) [40,80); (3) [80,120); and (4) $\geq$ 120
10	Rainfall (mm)	(1) [2770,2800); (2) [2800,2950); (3) [2950,3100); (4) [3100,3300); and (5) $\geq$ 3300

A total of 95 landslides that have been occurred during five year periods from 2008 to 2013 were identified and mapped to construct landslide inventory map (Figure 1). Field investigations along road and in the populated area revealed the biggest landslide event occurred in August, 2008 with volume of about 90,000 m^3 at the An Lac commune, and the smallest one is in July 2013 at volume of only 9 m^3 at the Dong Quan commune. From interpretation of air photos, the area of the largest landslide is about 664,158 m^2, whereas the smallest is approximately 22,821 m^2.

These landslides were classified into three types namely translational, rotational, and debris slides. In the study area, the number of translational landslides is 65 that are 68.4% of total landslides. The number of debris slides is 18 that equals to 19% of total landslides. Remaining 12 locations fall into rotational type of landslides that is approximately 12.6% of total landslide occurrences. Landslide locations were divided randomly into two parts, and then converted into raster data with the pixel size of 20×20 m for analysis. One part of 75% landslide locations (29,038 pixels) used for training process and another of 25% landslide locations (4979 pixels) utilized for validation process.

3.2. Landslide causal factors

Based on the analysis of the natural mechanism of landslides and the geo-environmental characteristics of the study area, a total of ten landslide causal factors (slope, aspect, elevation, curvature, lithology, land use, distance to roads, distance to rivers, distance to faults, and rainfall) were selected for landslide analysis in the present study. Moreover, these factors were reclassified into different classes for landslide spatial prediction which is based on the frequency analysis of landslides in this study and landslide studies (Table 1). Similar approaches have been adopted by other researchers in the identification of causal factors (Dai & Lee 2002; Pourghasemi et al. 2013; Tien Bui et al. 2016a).

3.2.1. Geomorphologic factors

It is well known that landslides are largely influenced by terrain types, therefore geomorphologic factors should be taken into account for landslide susceptibility assessment (Dou et al. 2014). In the present study, geomorphologic factors, i.e. slope; aspect; elevation; and curvature were obtained from a Digital Elevation Model (DEM) with a spatial resolution of 20 m. The DEM was generated from national topographic maps available on a scale of 1:50000 obtained from Vietnam Institute of Geosciences and Mineral Resources (Tien Bui et al. 2016b).

Slope is considered as one of the most important factors for slope instability analysis (Sadr et al. 2014), where slope is steeper there is high probability of slope failure (Dai et al. 2001). However, variations of soil thickness and strength should also be taken into account. The slope map was constructed with five categories (Tien Bui et al. 2014) namely 0–8°, 8–15°, 15–35°, 35–45°, and > 45° (Figure 3a). The distribution of landslide pixels on slope map is shown in Figure 6a. The slopes of 15–25^o occupy the highest percentage of landslide pixels (42.13%), followed by slopes of 25–35^o (29.66%), slopes of 8–15^o (16.31%), slopes of 35–45^o (10.46 %), slopes larger than 45^o (1.44%), respectively. There are no landslide pixels in flat slopes of 0–8^o.

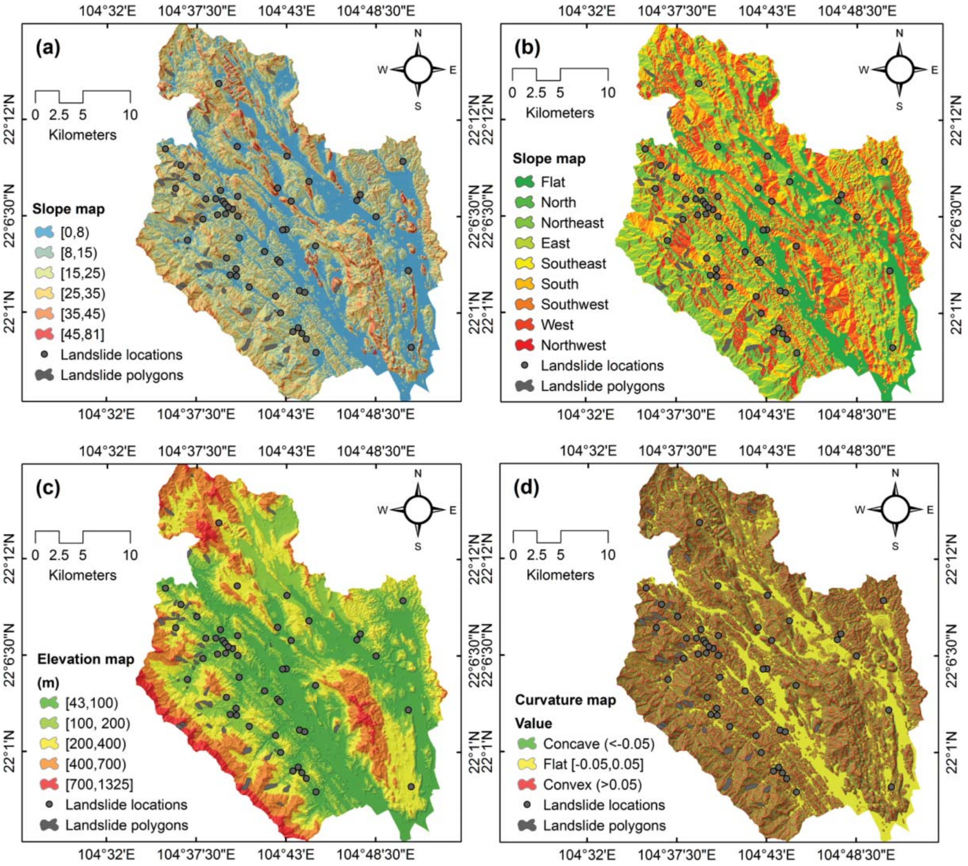

Figure 3. (a) Slope map, (b) aspect map, (c) elevation map, and (d) curvature map.

Aspect is also important factor influencing the slope instability as it controls topographic moisture due to impaction of solar radiation and rainfall (Sadr et al. 2014). The aspect map was constructed with nine classes (Tien Bui et al. 2014) such as flat (–1), north (0–22.5 and 337.5–360), northeast (22.5–67.5), east (67.5–112.5), southeast (112.5–157.5), south (157.5–202.5), southwest (202.5–247.5), west (247.5–292.5), northwest (292.5–337.5) (Figure 3b). The distribution of landslide pixels on the aspect map is shown in Figure 6b. There are no landslide pixels in flat class. The highest percentage of landslide pixels belongs to east class (27%), followed by southeast (22.2%), south (18.51%), northeast (15.65%), southwest (6.62%), north (6%), northwest (2.29%), and west (1.73%), respectively.

Elevation is known as one of the conditioning factors to landslide occurrences because degrees of weathering of rock depend on the elevation beside types of rocks and water conditions (Mika 2013). The elevation map was built with five intervals (Ercanoglu and Gokceoglu 2002) such as 0–100, 100–200, 200–400, 400–700, and > 700m (Figure 3c). The distribution of landslide pixels on the elevation map is shown in Figure 6c. The number of landslide pixels is the highest between elevation 200 and 400m (48.91%), followed by 400–700m (22.74%), 100–200m (20.34%), > 700m (7.01%), and 0–100m (1%), respectively.

Curvature is a factor that reflects the morphology of terrain surface representing changes in slope angles along a very small arc of the curve (Tien Bui et al. 2014), and thus be susceptible to slope instability. The curvature map (Figure 3d) was generated with three classes (Tien Bui 2012) such as concave (< –0.05), flat (–0.05–0.05), and convex (> 0.05). The distribution of landslide pixels on the curvature map is shown in Figure 6d. The landslide pixels only appear in concave (51.71%) and convex (48.29%) classes and no landslide pixels are shown in flat class.

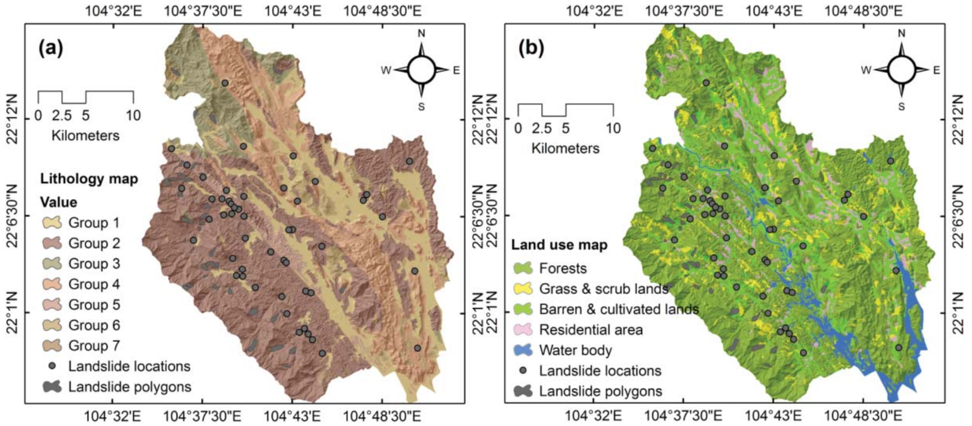

Figure 4. (a) Lithology map and (b) land use map.

3.2.2. Lithology

Lithology is one of the most important factors that influence the type and mechanism of the landslides because different types of rocks and soils are having different internal structures, mineral compositions, and thus susceptibility to landslide occurrences (Ercanoglu 2005).

In this study, the lithology map (Figure 4a) was constructed based on the Geological and Mineral Resources Map of the Luc Yen district on a scale of 1:50,000. Lithology was classified into seven groups (Table 2) based on mineral composition, degree of weathering, and estimated strength and density (Van et al. 2006; Tien Bui 2012). The distribution of landslide pixels on the lithological map is shown in Figure 6e. The highest percentage of landslide pixels falls in group 2 (78.38%) whereas the smallest percentage of landslide pixels (0.44%) is observed in group 7.

3.2.3. Land use

Land use pattern affects to landslide occurrences due to human intervention (Glade 2003). For instance, landslide occurs more frequently in barren area, and less frequently in forest and residential regions (Lallianthanga & Lalbiakmawia 2013). The land use map was generated from air photos on a scale of 1:33.000 using Envi 5.0 software with the maximum likelihood classification. A total of five land use classes were identified and grouped, i.e. forests, grass & scrub lands, barren & cultivated lands, residential area, and water bodies (Figure 4b). The distribution of landslide pixels on the land

Table 2. The components of lithological groups.

No.	Class	Main characteristics	Components
1	Group 1	Quaternary deposits	Alluvial sedimentary and Pluvial sedimentary: pebbles, stone, cobble, sand, silt, etc.
2	Group 2	Metamorphic rocks with rich aluminosilicate components	Quartz mica–schist, quartz sericite–schist, quartzite, and sericite–quartzite, etc.
3	Group 3	Terrigenous sedimentary and metamorphic rocks with rich quartz components	Gritstone, sandstone, siltstone, claystone, carbonates, alternated rhyolites, dacites, andesite sediments, quartz–mica sandstone, quartzitic sandstone, cherty shale, etc.
4	Group 4	Carbonate rocks	Limestone, dolomitized limestone, cherty limestone, clayish limestone
5	Group 5	Terrigenous sedimentary rocks with rich aluminosilicate components	Gritstone, sandstone, siltstone, claystone, carbonates, alternated rhyolites, dacites, andesite sediments
6	Group 6	Acid-neutral intrusive magmatic rocks	Rhyolite, dacite, felsite, and andesite rocks, plagioclase–granite, granophyre, granosyenite, granodiorite, diorite, and quartz–diorite
7	Group 7	Mafic-ultramafic magma rocks	Dunit,peridotit, pyroxenit, tremolite schist, artinolite schist, gabbro–pyroxenit, gabbro–amphibolit, gabbro–norit, gabbro–anorthosit, gabbro–diorit, gabbro–diabas, diabas, mafic bazan olivin, bazan tholeite, bazan dolerite, etc.

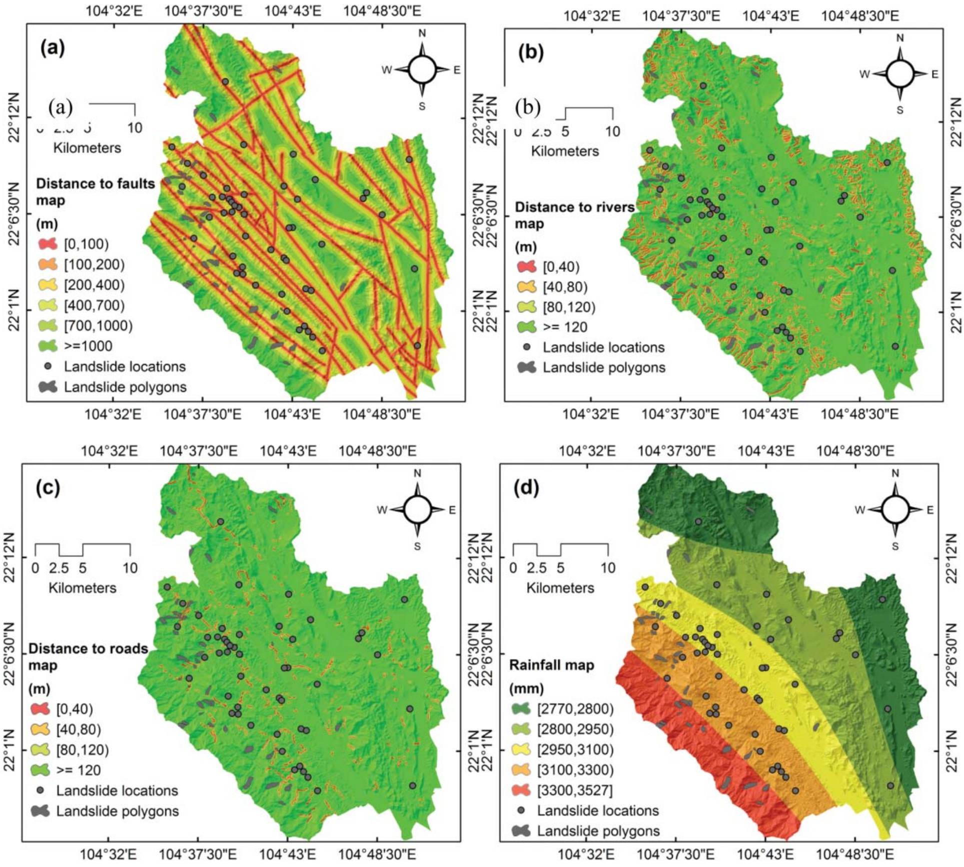

Figure 5. (a) Distance to faults map, (b) distance to rivers map, (c) distance to roads map, and (d) rainfall map.

use map is shown in Figure 6f. The highest percentage of landslide pixels in forests is 64%, following by grass and scrub lands (17.75%), barren and cultivation lands (16.99%), residential area (0.75%), respectively. There are no landslide pixels in water bodies.

3.2.4. Distance to features

Features such as faults, rivers, and roads should be taken into account for landslide susceptibility assessment (Tien Bui 2012). Faults are products of tectonic activities that break the continuity of soil or rock masses and are considered weak planes influencing slope stability. The fault lines were extracted from the Geological and Mineral Resources Map of the Luc Yen district at the scale of 1:50,000. The distance to faults map was then constructed with six classes by buffering these fault lines into study area (Tien Bui 2012) namely 0–100 m, 100–200 m, 200–400 m, 400–700 m, 700–1000 m, and $>$ 1000 m (Figure 5a). The distribution of landslide pixels on distance to faults map is shown in Figure 6g. The percentage of landslide pixels at a distance of 200–400 m is 22.57% and at 400–700 m is 24.91%. Lower percentage of landslide pixels has been observed at distances 0–100 m (9.96%) and 700–1000 m (9.77%).

The erosion of soil and rock masses caused by the activities of rivers has also influenced significantly landslide occurrences in the study area. The density of drainage affects moisture of terrain as more dense drainage pattern helps in accumulation of water, and thus making area more susceptible to landslide occurrence (Stevens & Wolfe 2012). In this study, river sections that undercut slopes larger than 15° were also extracted from national topographic maps on a scale of 1:50,000 (Tien Bui

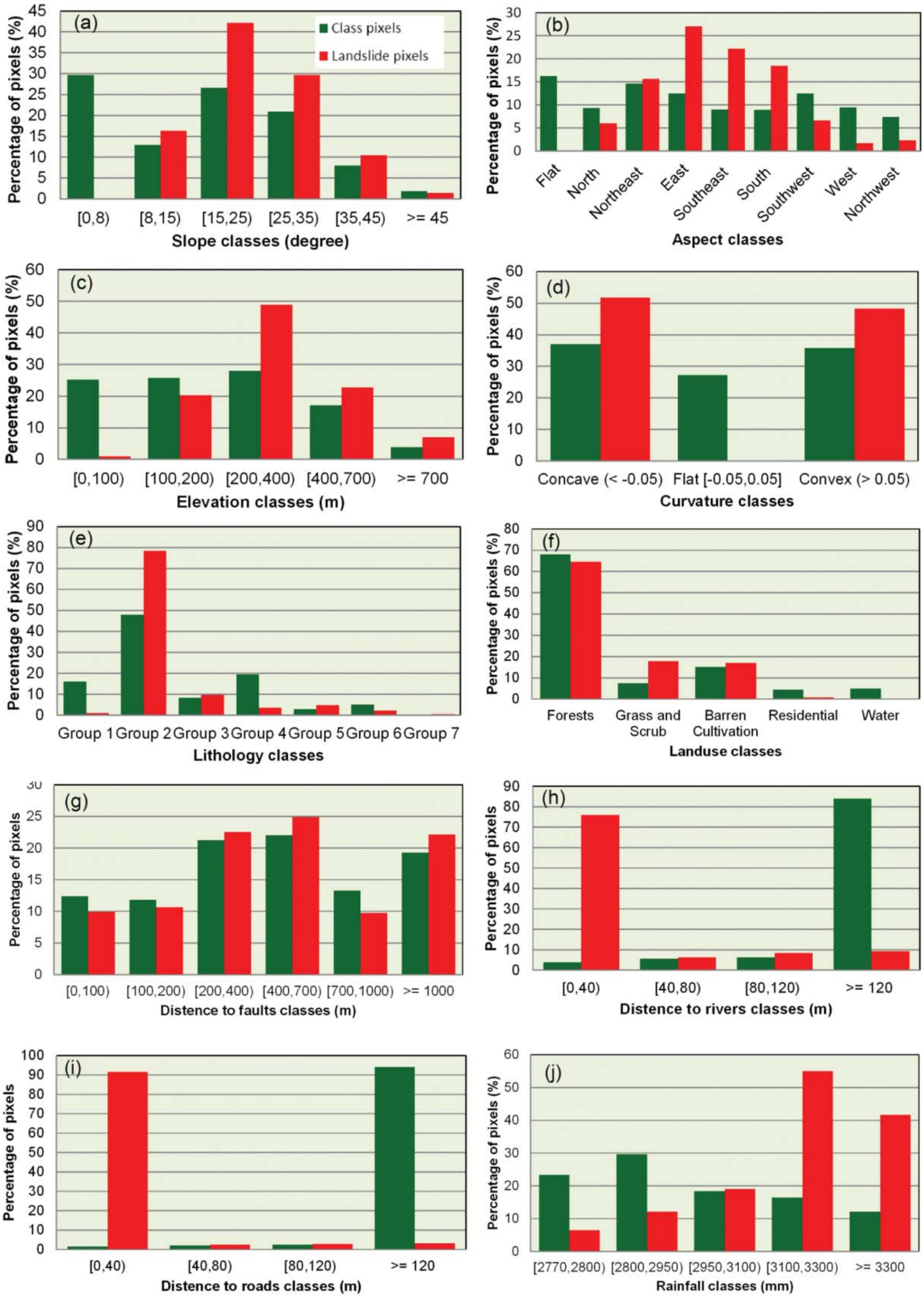

Figure 6. Distribution of pixels on landslide causal factors maps: (a) slope map, (b) aspect map, (c) elevation map, (d) curvature map, (e) lithology map, (f) land-use map, (g) distance to faults map, (h) distance to rivers map, (i) distance to roads map, and (j) rainfall map.

2012). Then the distance to rivers map was constructed with four categories: 0–40 m, 40–80 m, 80–120 m, and > 120 m (Figure 5b). The distribution of landslide pixels on the distance to rivers map is shown in Figure 6h. The percentage of landslide pixels is the highest at 0–40 m (91.58%), and very less in the rest categories, i.e. 40–80 m (6.4%), 80–120 m (8.39%), and > 120 m (9.23%).

Road sections in the mountainous and hilly regions that undercut slopes larger than 15°, breaking the continuity of soil or rock masses are considered to be susceptible to instability of slopes. The road networks were extracted from national topographic maps on a scale of 1:50,000. After that, the distance to roads map was constructed with four intervals (Tien Bui et al. 2015) such as 0–40 m, 40–80 m, 80–120 m, and > 120 m (Figure 5c). The distribution of landslide pixels on the distance to roads map is shown in Figure 6i. The percentage of landslide pixels is the highest at a distance between 0 and 40 m (75.98%), and very small at the distances: 40–80 m (2.41%), 80–120 m (2.89%), and > 120 m (3.12%).

3.2.5. Rainfall

Rainfall is considered to be a triggering factor that influences significantly to landslide occurrences (Shahabi et al. 2014). This is because rainfall affects the soil properties such as decreasing of soil shear strength. Rains also causes liquefaction of soil material and even flow of soil/ debris mass enhancing the susceptibility of soil masses to landslides (Highland & Bobrowsky 2008). In fact, landslide usually occurs during long-term intensive rainfall in the study area. The rainfall data during 30 years from 1984 to 2014 was extracted from the database of Climate Forecast System Reanalysis (CFSR) in Global Weather data for SWAT (NCEP 2014). The rainfall map was then generated with five classes namely rainfall less than 2800 mm, 2800–2950 mm, 2950–3100 mm, 3100–3300 mm, and greater than 3300 mm (Figure 5d). The distribution of landslide pixels on the rainfall map is shown in Figure 6j. The two highest percentages of landslide pixels are in the two highest rainfall classes, i.e. 3100–3300 mm (55%), > 3300 mm (41.73%). Lower percentage of landslide pixels has been observed in smaller rainfall, i.e. < 2800 mm (6.51%), 2800–2950 mm (12.11%), and 2950–3100 mm (19%).

3.3. Methodology

3.3.1. Feature selection of information gain

Information Gain method is one of the widely used techniques in feature selection for data mining (Tatsunori Mori 2002; Witten et al. 2011; Sharma & Dey 2012; Azhagusundari & Thanamani 2013). Although this method has demonstrated merits for spatial data analysis and modelling (Martínez-Álvarez et al. 2013), the application of this method in landslide studies is rare. The principle of this technique is based on evaluation of prediction ability and importance of the input variables (Sharma & Dey 2012; Azhagusundari & Thanamani 2013). The irrelevant or unimportant variables are then removed for learning process (Azhagusundari & Thanamani 2013). Consequently, the accuracy of results can be improved and the process of learning could be implemented more quickly (Doshi & Chaturvedi 2014).

Let z_i, $i = \vec{1}, n$ (z_i is the landslide causal factors); L_j, $j = \vec{1}, m$ (L_j is the out classes including landslide, non-landslide). The information gain value of each landslide causal factor is quantified based on the reduction measurement of the entropy (information) using the following equation:

$$\mathrm{InfoGain}(L, z_i) = \mathrm{IF}(L) - \mathrm{IF}_z(L), \tag{1}$$

where IF(L)is the entropy value of *L* that is the expected information needed to classify a landslide causal factor for *L* and is given by

$$\mathrm{IF}(L) = -\sum_{j=1}^{m} P(L_j)\, \log_2 P(L_j). \tag{2}$$

$\mathrm{IF_z(L)}$ is the information of L after integrating values of landslide causal factors z_i and is calculated by

$$\mathrm{IF_z(L)} \; = \; -\sum_{i=1}^{n} \frac{|\mathrm{L_i}|}{|\mathrm{L}|}\,\mathrm{IF(L_i)}, \tag{3}$$

where $|\mathrm{L_i}|/|\mathrm{L}|$ is the weight of the ith landslide causal factor and $\mathrm{IF(L_i)}$is the entropy of L corresponding to the ith landslide causal factor. As a remark, the factors with higher Information Gain value would have more important to landslide models. Also, the factors with zero Information Gain value are having no contribution to landslide models, thus it must be removed during dataset preparation.

3.3.2. Naïve Bayes classifier

Naïve Bayes classifier is one of the simplest soft computing methods which is based on the Bayesian theory and the maximum posteriori hypothesis (Rish et al. 2001). Naïve Bayes classifier uses a statistical hypothesis that all values of numeric attributes are independent and normally distributed in each class (Zhang & Su 2004). Naïve Bayes classifier has been applied effectively in many fields such as medical diagnosis (Domingos & Pazzani 1997), and management (Hellerstein et al. 2000). However, its application is still limited in landslide problems.

Let $t = t_i$, $i = 1, 2, \ldots, 10$ are the attributes of the 10 landslide causal factors, $\Gamma = \Gamma_j$, j = landslide, non-landslide that represent classified variables and outputs. The prediction using Naïve Bayes classifier is presented as follows:

$$\Gamma_{\mathrm{NBC}} = \underset{\Gamma_i = [\text{landslide,non landslide}]}{\operatorname{argmax}\ \mathrm{P}(\Gamma_i)} \prod_{i=1}^{10} \mathrm{P}(\mathrm{t_i} \,|\, \Gamma_i), \tag{4}$$

where $\mathrm{P}(\Gamma_i)$ is termed as the prior probability of Γ_j which can be estimated using the proportion of the observed cases with output class Γ_j in the training dataset. $\mathrm{P}(\mathrm{t_i}\,|\,\Gamma_i)$ is defined as the conditional probability which can be calculated as follows:

$$\mathrm{P}(\mathrm{t_i}\,|\,\Gamma_i) = \frac{1}{\sqrt{2\pi}\beta} e^{\frac{-(\mathrm{t_i}-\nu)^2}{2\beta^2}}, \tag{5}$$

where ν is mean and β is standard deviation of $\mathrm{t_i}$Naïve Bayes classifier is easy to construct, and has surprisingly good performance in classification. On the other hand, it is also shown as a method of poor probability estimation due to the conditional independence assumption (Zhang & Su 2004). Therefore, some researches have tried to improve its probability estimates (Friedman et al. 1997; Zadrozny & Elkan 2001). Additionally, the performance of Naïve Bayes classifier might be improved by using ensemble classifier framework (Pham et al. 2016c).

3.3.3. Rotation forest ensemble

Rotation Forest is a relatively new framework for creating classifier ensembles. It was first proposed by Rodriguez et al (2006). The basis of Rotation forest is that principal component analysis (PCA) is used to extract the features to create training datasets for learning base classifiers (Zhang & Zhang 2009; Koyuncu & Ceylan 2013). Rotation Forest ensemble has been utilized to solve several classification problems (Koyuncu & Ceylan 2013). The principal aim of Rotation Forest ensemble technique is to encourage same time individual accuracy and diversity (Rodriguez 2007). The success of Rotation Forest is relied on the rotation matrix created by the transformation methods and the base classifiers (Xia et al. 2014).

Suppose that $x = (x_1, x_2, \ldots, x_{10})$ is the vector of the 10 landslide causal factors and $y = (y_1, y_2)$ is the vector of landslide and non-landslide classes, X represents the training set. $C_{1,}$ C_2, …, C_L are classifiers in the ensemble, and by T which is landslide causal factor set. The steps for training classifier C_i are implemented as follows (Rodriguez et al. 2006; Rodriguez 2007; Zhang & Zhang 2009; Xia et al. 2014):

First, generating the rotation matrix $R_i{}^a$ by rearranging the matrix of R_i is as shown as follows:

$$R_i = \begin{bmatrix} a_{i,1}{}^{(1)}, a_{i,1}{}^{(2)}, \ldots., a_{i,1}{}^{(Q_1)} & 0 & \cdots & 0 \\ 0 & a_{i,1}{}^{(1)}, a_{i,1}{}^{(2)}, \ldots., a_{i,1}{}^{(Q_2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{i,1}{}^{(1)}, a_{i,1}{}^{(2)}, \ldots., a_{i,1}{}^{(Q_K)} \end{bmatrix}. \quad (6)$$

To make the matrix ofR_i, (i) T is split into K subsets with the number of the landslide causal factors for each subset is $Q = 10/K$. (ii) For classifier C_i, let $T_{i,j}$ be the jth, $j = 1, 2, \ldots$, K subset of the landslide causal factors. $X_{i,j}$ is landslide causal factors in $T_{i,j}$ from X. $X_{i,j}{}^{'}$ is randomly selected from $X_{i,j}$ with 75% size using bootstrap algorithm. After that, $X_{i,j}{}^{'}$ would be transformed to obtain the coefficients $a_{i,1}{}^{(1)}, a_{i,1}{}^{(2)}, \ldots., a_{i,1}{}^{(Q_i)}$, the size of $a_{i,1}{}^{'}$is Q x 1. (iii) Arrange a sparse rotation matrix $R_i$with the obtained coefficients

Then, the confidence is calculated for each class by the average combination method in the given test sample χ,

$$\mu_{\mathrm{k}}(\eta) = \frac{1}{\mathrm{L}} \sum_{\mathrm{i}=1}^{\mathrm{L}} \gamma_{\mathrm{i,k}}(\eta \mathrm{R_i}^{\mathrm{a}}), \qquad k = 1, 2, \ldots\ldots, \mathrm{c}, \quad (7)$$

where $\gamma_{i,k}(\eta R_i{}^a)$is the probability generated by the classifier C_i to the hypothesis that η belongs to class k.

Lastly, the η will be assigned to the class with the largest confidence.

3.3.4. The novel classifier ensemble model

In this study, the novel ensemble classifier model is generated by the combination of Naïve Bayes classifier and Rotation Forest ensemble. Rotation Forest ensemble was first applied to create the subsets of training. Thereafter, Naïve Bayes classifier was used to construct base classifiers from these subsets for classification. Methodological flow chart of the novel classifier ensemble model is shown in Figure 7. The advantage of the novel classifier ensemble model is that the training subsets are being optimized using Rotation Forest ensemble, and then these training subsets are utilized for training a base classifier of Naïve Bayes. Therefore, the novel classifier ensemble model could improve predictive capability of a base classifier of Naïve Bayes.

3.3.5. Statistical index-based evaluations

The five statistical indexes namely Positive Predictive Value (PPV), Negative Predictive Value (NPV), sensitivity, specificity, and accuracy (Pham et al. 2016b) were chosen to evaluate the performance of landslide models. Here, PPV indicates the probability of pixels that is classified correctly as ‘landslide’ class. NPV indicates the probability of pixels that is classified correctly as ‘non-landslide’ class. Sensitivity is the probability of landslide pixels that is classified correctly as ‘landslide’ class. Specificity is the probability of non-landslide pixels that is classified correctly as ‘non-landslide’ class. Accuracy is the proportion of landslide and non-landslide pixels that are correctly classified. These values were calculated using the values of confusion matrix (Dou et al. 2015) including true positive (TP), false positive (FP), true negative (TN), and false negative (FN) that was obtained through training and validating process in Weka software 3.6.11 version.

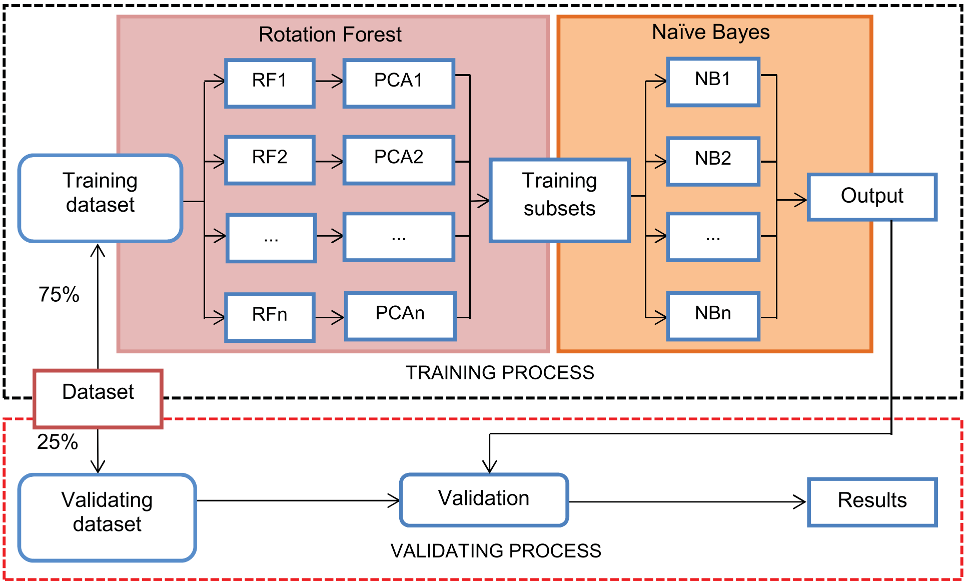

Figure 7. Methodological flow chart of the novel classifier ensemble model.

The overall performance of the landslide models is evaluated by using Receiver Operating Characteristic (ROC) curve technique. ROC curve is a graph with each point on it representing a pair of sensitivity and 100-specificity corresponding to a particular decision threshold (Fawcett 2006; Dou et al. 2014). The area under the ROC curve (AUC) indicates the goodness-of-fit of landslide models on the training data and prediction capability of landslide models using the validation data (Jones and Athanasiou 2005). The AUC value equals to '1' representing a perfect model whereas the AUC value equals to '0' indicating a non-accurate model. When the AUC value is closer to '1' the performance of landslide model is better (Walter 2002; Pourghasemi et al. 2012). According to Kantardzic (2011) the AUC values can be classified into different intervals with respective performance such as 0–0.7 (poor), 0.7–0.8 (fair), 0.8–0.9 (good), and 0.9–1.0 (very good).

4. Results

4.1. Feature selection using information Gain method

Utilizing Information Gain method with ten folds cross validation standard, the evaluation of predictive capability of ten landslide causal factors was carried out using training data. The Average Information Gain (AIG) value and its standard deviation for each factor were calculated and ranked (Table 3). In general, the total of ten landslide causal factors has contribution to landslide models (AIG > 0). Aspect shows the highest contribution to landslide models in the study area with AIG value is 0.189, followed by slope (AIG = 0.166), rainfall (AIG = 0.154), curvature (AIG = 0.15), lithology (AIG = 0.138), elevation (AIG = 0.13), land use (IG = 0.074), distance to rivers (AIG = 0.004), respectively. Distance to faults and distance to roads factors have the least contribution to landslide models with AIG of 0.002.

4.2. Model performance and validation

The performance of the novel classifier ensemble model for landslide susceptibility assessment is shown in Table 4 and Figure 8, 9. The results show that the novel model has a very high degree of the goodness-of-fit in the case of training data with 87.37% of predictive accuracy and 0.94 of area

Table 3. Predictive capability of the landslide causal factors to landslide models in this study area.

No.	Landslide causal factors	AIG	Standard deviation
1	Aspect	0.189	±0.001
2	Slope	0.166	±0
3	Rainfall	0.154	±0.001
4	Curvature	0.15	±0.001
5	Lithology	0.138	±0.001
6	Elevation	0.13	±0.001
7	Land use	0.074	±0.001
8	Distance to rivers	0.004	±0
9	Distance to faults	0.002	±0
10	Distance to roads	0.002	±0

Table 4. Performance of the novel classifier ensemble model using training and validation data.

No.	Parameter	Training dataset	Validation dataset
1	True positive	13812	3923
2	True negative	11559	3921
3	False positive	707	1056
4	False negative	2959	1058
5	PPV (%)	95.13	78.79
6	NPV (%)	79.62	78.75
7	Sensitivity (%)	82.36	78.76
8	Specificity (%)	94.24	78.78
9	Accuracy (%)	87.37	78.77

under ROC curve (AUC = 0.94). More specifically, the probability of pixels that are classified correctly as 'landslide' class is 95.13% (PPV = 95.13%) whereas the probability of pixels which are classified correctly as 'non-landslide' class is 79.62% (NPV = 79.62%). The probability of the landslide pixels are classified correctly to 'landslide' class is 82.36% (sensitivity = 82.36%). The probability of non-landslide pixels which are classified correctly as 'non-landslide' class is 93.20% (specificity = 82.36%).

The novel classifier ensemble model was validated using the validation dataset which has not been used during training process. The results indicate that the novel model has a good performance in landslide susceptibility assessment with 78.77% of predictive accuracy and 0.846 of the AUC value. Moreover, the probability of pixels is classified correctly as 'landslide' class is 78.79% (PPV = 78.79%). The probability of pixels which are classified correctly as 'non-landslide' class is 78.75% (NPV = 78.75%). The probability of landslide pixels are classified correctly to 'landslide' class is 78.76% (sensitivity = 78.76%). The probability of non-landslide pixels are classified correctly into 'non-landslide' class is 78.78% (specificity = 78.78%).

4.3. Reclassification of landslide susceptibility map

Landslide susceptibility map is the final result of landslide susceptibility assessment using the novel classifier ensemble model. In order to construct this map, landslide susceptibility indexes was extracted after the successful model training phase. By using ArcGIS software 10.2 each pixel inside the study area was assigned an unique susceptible index, and then the reclassification of landslide susceptibility map was carried out by ranking and grouping the landslide susceptibility indexes.

According to Pradhan and Lee (2010), the classification of landslide susceptibility classes can be implemented based on percentage of area of the region. At first, the susceptible indexes of all cells were sorted in descending order. And then, these indexes were grouped into several groups according to area percentage of the region. Moreover, Althuwaynee et al. (2014) proposed that the landslide susceptibility classes might be classified into five categories (very high, high, moderate, low, and not susceptible). In this study, landslide susceptibility map (Figure 10) was constructed with

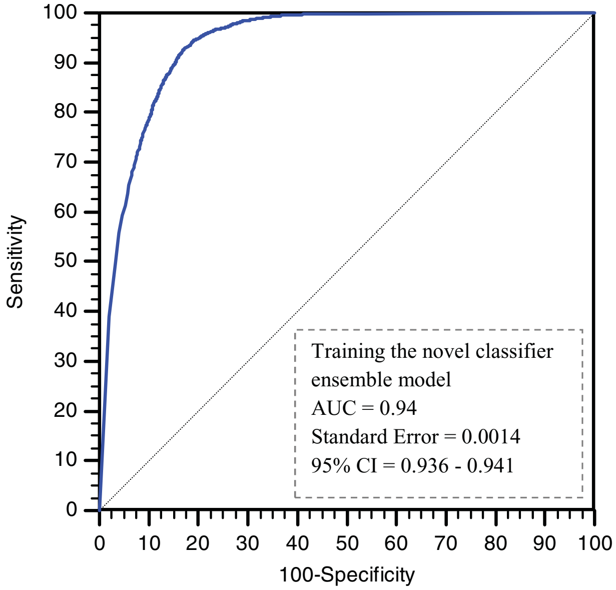

Figure 8. ROC curves and AUC analysis using the training dataset.

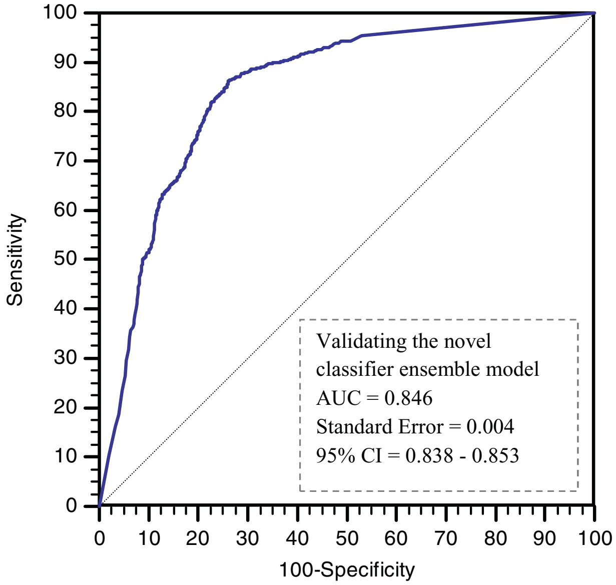

Figure 9. ROC curves and AUC analysis using the validation dataset.

into five classes on the base of area percentage of the region, namely: Not susceptible (50%), Low (20%), Moderate (15%), High (10%), and Very high (5%).

4.4. Model comparison

The performance of the novel classifier ensemble model was compared to other ensemble techniques using Naïve Bayes as a base classifier such as Bagging, AdaBoost, MultiBoost. These models are well

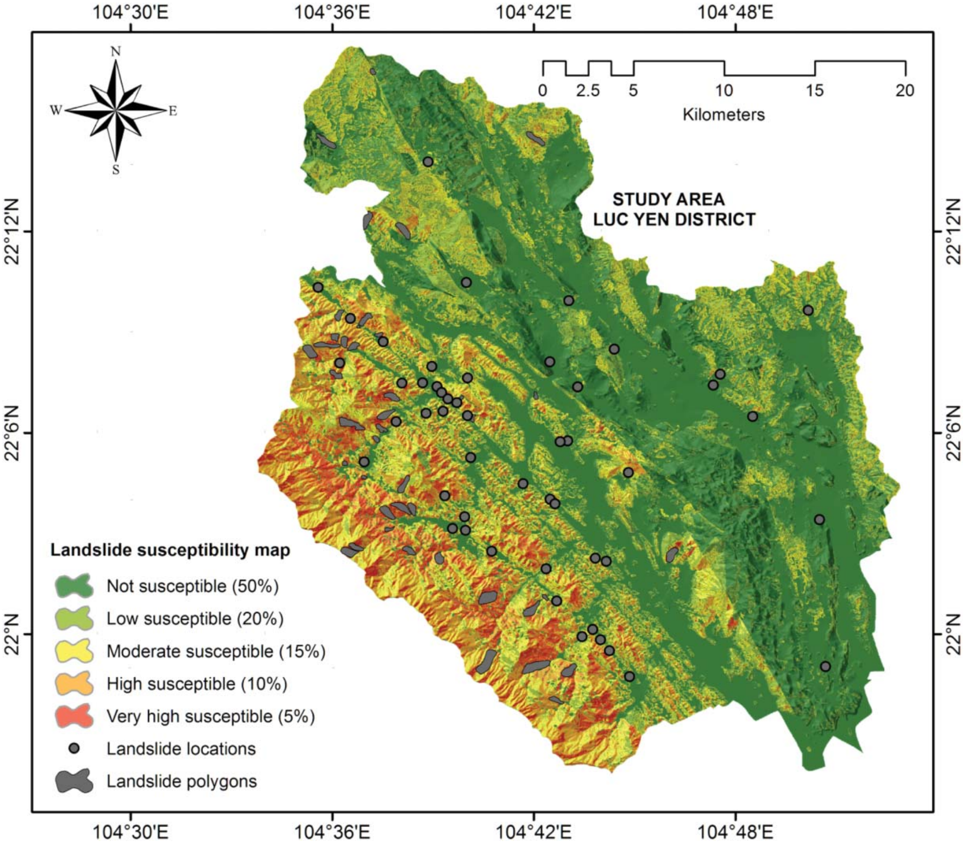

Figure 10. Landslide susceptibility map at the Luc Yen district using the novel classifier ensemble model.

known as boosting techniques that is one of the most important recent methodological developments in classification (Friedman & Tibshirani 2000). Additionally, an individual classifier ensemble of Random Forest was also taken into account for comparison.

Bagging is one of the earliest ensemble learning algorithms proposed by Breiman (1996). It is known as a bootstrap aggregation using the training dataset to generate multiple random subsets. After that, the Naïve Bayes classifier-based model is constructed on the base of each subset. The final classifier ensemble model is formed by integrating these classifiers.

AdaBoost is one of the most popular boosting algorithms for classification (Mease & Wyner 2008). AdaBoost was introduced by Freund and Schapire (1997), and it is known as an extremely effective adaptive boosting (Dietterich 2002). It creates the training subsets and assigns the weights for each subset through sampling process using base training set, and then the Naïve Bayes classifier uses these weighted subsets for classification.

MultiBoost is a combination of boosting and wagging techniques for reducing both variance and bias and avoiding the over-fitting (Geoffrey 2000). Using the training set, the subsets of training are built through random selection. These subsets are then assigned the weights through the boosting technique. Thereafter, the Naïve Bayes classifier model uses these subsets to produce the outcomes. However, the training process is to be continuous by resetting the weights of subsets according to the overall accuracy performance of the Naïve Bayes classifier model. Training process is finished if the optimal weights are assigned in training subsets to get the highest overall accuracy performance.

Random Forest is a combination of multiple decision tree classifiers that utilizes both bagging and random variable selection, it was proposed by Breiman (2001). In the beginning, the subsets of training are generated randomly from original training dataset using bootstrap aggregation approach,

Table 5. Model performance using training dataset.

No.	Parameter	AdaBoost	Bagging	MultiBoost	Random Forest
1	True positive	12623	13784	13719	14321
2	True negative	11594	10365	10451	13629
3	False positive	1896	735	800	198
4	False negative	2924	4153	4067	889
5	PPV (%)	86.94	94.94	94.49	98.64
6	NPV (%)	79.86	71.39	71.99	93.88
7	Sensitivity (%)	81.19	76.85	77.13	94.16
8	Specificity (%)	85.95	93.38	92.89	98.57
9	Accuracy (%)	83.40	83.17	83.24	96.26

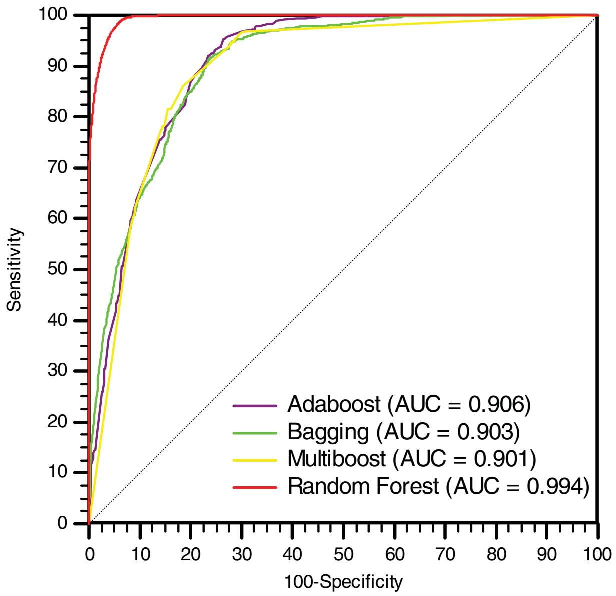

Figure 11. ROC curves and AUC analysis using the training dataset for the AdaBoost, Bagging, MultiBoost, and Random Forest models.

and then each of the individual decision trees is constructed from each subset (Díaz-Uriarte & De Andres 2006). Random Forest is an effective ensemble technique that could obtain good results with both low bias and low variance (Gislason et al. 2006).

Using training dataset, the performance of four landslide susceptibility models namely Bagging, AdaBoost, MultiBoost, Random Forest is shown in Table 5 and Figure 11. It can be clearly seen that these four models have high degree of the goodness-of-fit in landslide susceptibility assessment. Out of these models, the Random Forest model is the highest (Accuracy = 96.26%, AUC = 0.994), followed by the AdaBoost model (Accuracy = 83.40%, AUC = 0.906), the MultiBoost model (Accuracy = 83.24%, AUC = 0.903), and the Bagging model (Accuracy = 83.17%, AUC = 0.901), respectively. Overall, the novel classifier ensemble model has higher degree of the goodness-of-fit compared to Bagging, AdaBoost, MultiBoost. However it is less than the Random Forest model.

The validation of the four landslide models has been carried out using the validation dataset. The results are shown in Table 6 and Figure 12. The predictive accuracy of the MultiBoost model is highest (Accuracy is 79.3%), followed by the Bagging model (Accuracy is 79.03%), the AdaBoost model (Accuracy is 77.44%), and the Random Forest model (Accuracy is 67.53), respectively. The MultiBoost model and the Bagging model have higher accuracy comparing with the novel classifier ensemble model while the AdaBoost model and the Random Forest model have lower accuracy. Regarding to the area under ROC curves of these models, the Bagging model indicates the highest

Table 6. Model performance using validation dataset.

No.	Parameter	AdaBoost	Bagging	MultiBoost	Random Forest
1	True positive	3742	4351	4347	2226
2	True negative	3969	3519	3550	4499
3	False positive	1237	628	632	2753
4	False negative	1010	1460	1429	480
5	PPV (%)	75.16	87.39	87.31	44.71
6	NPV (%)	79.71	70.68	71.30	90.36
7	Sensitivity (%)	78.75	74.88	75.26	82.26
8	Specificity (%)	76.24	84.86	84.89	62.04
9	Accuracy (%)	77.44	79.03	79.30	67.53

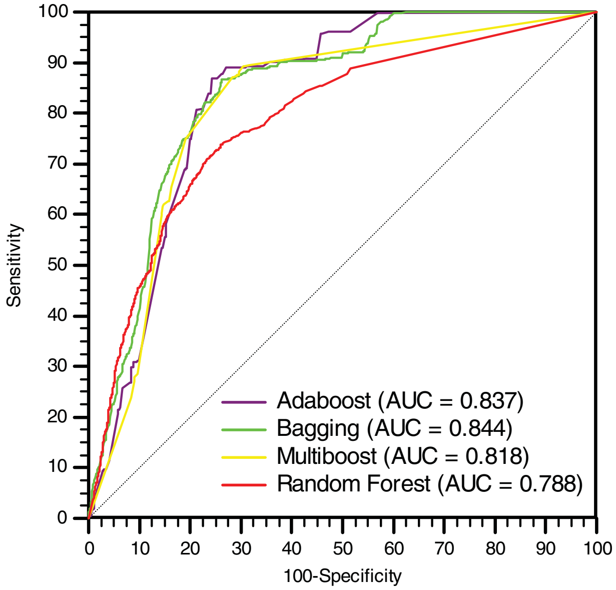

Figure 12. ROC curves and AUC analysis using the validation dataset for the AdaBoost, Bagging, MultiBoost, and Random Forest models.

(AUC = 0.844), following by the AdaBoost model (AUC = 0.837), the MultiBoost model (AUC = 0.818), respectively. The Random Forest model illustrates the smallest value of AUC (0.788) compared to other models. The AUC value of the novel classifier ensemble model is higher than all of four other models.

The performance capability of the novel classifier ensemble model has been further compared with four other landslide models using McNemar's test. It was proposed by Everitt (1992) as a statistical test based on the chi-square test value (χ^2) (Kuncheva 2004). This test compares the significance of differences between the landslide models. In case χ^2 value is greater than the critical value of 3.841459 and the level of significance (p) is less than 0.05, then the hypothesis of two significantly different models is correct. Thus the null hypothesis of two non-different models might be rejected (Dietterich 1998).

The results of the statistical test of prediction ability of the novel classifier ensemble model compared with other landslide models (AdaBoost, Bagging, MultiBoost, and Random Forest) are shown in Table 7. It could be observed that the statistical test of the novel classifier ensemble model vs. the AdaBoost model has the smallest chi-square value (38.823). It is dramatically higher than critical value of 3.841459. Furthermore, the p-value of all tests ($p < 0.0001$) is extremely lower than 0.05. Therefore, the novel classifier ensemble model has a difference with four other landslide models. This difference is statistically significant. It means that the performance of the novel classifier ensemble model is comparable to other landslide models.

Table 7. The performance of the novel classifier ensemble model (CEM) compared to other landslide models using McNemar's test.

No.	Pairwise comparison	χ^2	p	Significance
1	CEM vs. AdaBoost	38.823	<0.0001	Yes
2	CEM vs. Bagging	695.588	<0.0001	Yes
3	CEM vs. MultiBoost	665.719	<0.0001	Yes
4	CEM vs. Random Forest	2062.655	<0.0001	Yes

5. Discussions

Landslide susceptibility assessment has been done at Luc Yen district, Yen Bai province (Viet Nam) using the novel ensemble classifier model which is a combination of Naïve Bayes classifier and Rotation Forest ensemble. Naïve Bayes is an effective classifier. However, in the landslide problems, its performance is affected by independent assumption (Pham et al. 2016e). In contrast, Rotation Forest is a promising ensemble technique which could be used to improve the performance of individual classifiers (Pham et al. 2016e). Therefore, the ensemble classifier framework encompassing these two techniques could result better performance of landslide susceptibility assessment.

Landslide causal factors are usually used to prepare input data for running landslide models. Selection of these factors plays crucial role in getting qualitative output from the used model (Tien Bui 2012). Feature selection is an effective method in selection of variables in input data for modelling (Pham et al. 2015b) which can be used to realize the irrelevant or unimportant variables in the set of variables. Then these variables are removed to optimize the inputs for improving prediction accuracy of modelling (Dash & Liu 1997). In this study, the feature selection of Information Gain Method was selected to pick up the best landslide causal factors for the novel classifier ensemble model in landslide susceptibility assessment in the study area. Results show that all of the ten landslide causal factors (slope, aspect, elevation, curvature, rainfall, land use, lithology, distance to rivers, distance to faults, and distance to roads) are capable of prediction to landslide modelling. However, aspect and slope have the highest contribution to landslide models which is in agreement with other studies carried out by Sadr et al. (2014), and Van Den Eeckhaut et al. (2006).

Analysis results show the novel classifier ensemble model has the best degree of fit to landslide susceptibility assessment compared to other models on the base of the area under ROC curve. Moreover, its performance is dramatically higher than the AdaBoost model (1.33%), and the Random Forest (11.24%) model regarding to the predictive accuracy. However, it is slightly lower than the Bagging model (0.26%) and the MultiBoost (0.53%) model. Results of the present study are comparable with Rodriguez et al (2006) and Kavzoglu et al (2015) which showed that the Rotation Forest ensemble performs significantly better than other models such as AdaBoost and Random Forest; however, its performance is less than the MultiBoost ensemble, and quite similar to the Bagging ensemble. In comparison to other methods, the novel classifier ensemble model uses Naïve Bayes classifier which has abilities to deal with uncertainty and Rotation Forest ensemble which is more effective in dealing with small sample sizes, high-dimensional and complex data structures (Pham et al. 2016d).

Moreover, the present study proposed to use the McNemar's statistical test (Kavzoglu et al. 2015) for evaluation of the different significance of the novel classifier ensemble model and the other landslide models. McNemar's statistical test is known as one of the most powerful statistical tests for comparison (Roggo et al. 2003) which should be used to evaluate the performance of landslide models. Results (Table 7) show that the performance of the novel classifier ensemble model is different statistically with other models (AdaBoost, Bagging, MultiBoost, and Random Forest).

6. Conclusions

New methodological approach which combines the Rotation Forest ensemble and the Naïve Bayes classifier has been proposed for landslide susceptibility assessment at Luc Yen district of Yen Bai province (Viet Nam). This combined approach has not been carried out so far in other landslide studies. Performance of the novel landslide model was compared with other landslide models using current state-of-the art ensemble frameworks (AdaBoost, Bagging, MultiBoost, and Random Forest). In addition, feature selection method using the Information Gain Technique has been adopted to select the best landslide causal factors for running landslide models.

Results analysis proved that the novel classifier ensemble method is a promising technique that could be considered as an alternative for assessment of landslide susceptibility. Analysis also reveals that the performance of the novel model is comparable with other landslide models such as AdaBoost, Bagging, MultiBoost, and Random Forest. Moreover, while using this model, the Information Gain Technique should be used as a feature selection method to evaluate the importance of landslide causal factors for landslide susceptibility assessment. Additionally, this novel classifier ensemble method can be used for the evaluation of different types of landslides under varying geo-environmental conditions. Results of the present study could be helpful for the natural hazard management, planning and decision makings of the area affected by landslides.

Acknowledgement

The authors are thankful to the Vietnam Institute of Geosciences and Mineral Resources for sharing the data. The authors are also thankful to the Director, Bhaskarcharya Institute for Space Applications and Geo-Informatics, Gujarat, India for providing facilities to carry out this research work.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Binh Thai Pham http://orcid.org/0000-0001-9707-840X
Dieu Tien Bui http://orcid.org/0000-0001-5161-6479
Indra Prakash http://orcid.org/0000-0002-4309-0187

References

Alimohammadlou Y, Najafi A, Gokceoglu C. 2014. Estimation of rainfall-induced landslides using ANN and fuzzy clustering methods: a case study in Saeen Slope, Azerbaijan province, Iran. Catena. 120:149–162.

Althuwaynee OF, Pradhan B, Park H-J, Lee JH. 2014. A novel ensemble bivariate statistical evidential belief function with knowledge-based analytical hierarchy process and multivariate statistical logistic regression for landslide susceptibility mapping. Catena. 114:21–36.

Azhagusundari B, Thanamani AS. 2013. Feature Selection based on Information Gain. Int J Innov Technol Exp Eng (IJITEE). 2.

Breiman L. 1996. Bagging predictors. Machine Learn. 24:123–140.

Breiman L. 2001. Random forests. Machine Learn. 45:5–32.

Castellanos Abella EA. 2008. Multi-scale landslide risk assessment in Cuba. Enschede: ITC.

Conforti M, Pascale S, Robustelli G, Sdao F. 2014. Evaluation of prediction capability of the artificial neural networks for mapping landslide susceptibility in the Turbolo River catchment (northern Calabria, Italy). Catena. 113:236–250.

Dai F, Lee C. 2002. Landslide characteristics and slope instability modeling using GIS, Lantau Island, Hong Kong. Geomorphology. 42:213–228.

Dai FC, Lee CF, Li J, Xu JW. 2001. Assesment of Landslide susceptibility on the natural terrain of Lantau Island, Hong Kong. Environ Geol. 40:381–391.

Dash M, Liu H. 1997. Feature selection for classification. Intelligent Data Anal. 1:131–156.

Díaz-Uriarte R, De Andres SA. 2006. Gene selection and classification of microarray data using random forest. BMC Bioinformatics. 7:3.

Dietterich TG. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput. 10:1895–1923.

Dietterich TG. 2002. Ensemble learning. The Handbook Brain Theory Neural Networks. 2:110–125.

Domingos P, Pazzani M. 1997. On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learn. 29:103–130.

Doshi M, Chaturvedi SK. 2014. Correlation based feature selection (CFS) technique to predict student performance. Int J Comput Networks Commun (UCNC). 6.

Dou J, Chang K-T, Chen S, Yunus AP, Liu J-K, Xia H, Zhu Z. 2015. Automatic case-based reasoning approach for landslide detection: integration of object-oriented image analysis and a genetic algorithm. Remote Sens. 7:4318–4342.

Dou J, Oguchi T, Hayakawa YS, Uchiyama S, Saito H, Paudel U. 2014. GIS-based landslide susceptibility mapping using a certainty factor model and its validation in the Chuetsu Area, Central Japan. Landslide Science for a Safer Geoenvironment. Switzerland: Springer; p. 419–424.

Dragićević S, Lai T, Balram S. 2015. GIS-based multicriteria evaluation with multiscale analysis to characterize urban landslide susceptibility in data-scarce environments. Habitat Int. 45 (Part 2):114–125. International disasters database: http://www.emdat.be

Ercanoglu M. 2005. Landslide susceptibility assessment of SE Bartin (West Black Sea region, Turkey) by artificial neural networks. Nat Hazards Earth Syst Sci. 5:979–992.

Ercanoglu M, Gokceoglu C. 2002. Assessment of landslide susceptibility for a landslide-prone area (north of Yenice, NW Turkey) by fuzzy approach. Environ Geol. 41:720–730.

Everitt BS. 1992. The analysis of contingency tables. CRC Press.

Fawcett T. 2006. An introduction to ROC analysis. Pattern Recognit Lett. 27:861–874.

Fell R, Corominas J, Bonnard C, Cascini L, Leroi E, Savage WZ. 2008. Guidelines for landslide susceptibility, hazard and risk zoning for land-use planning. Eng Geol. 102:99–111.

Freund Y, Schapire R. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci. 55:119–139.

Friedman N, Geiger D, Goldszmidt M. 1997. Bayesian network classifiers. Machine Learn. 29:131–163.

Friedman TH, Tibshirani R. 2000. Additive logistic regression: A statistical view of boosting. Annals Statistics. 28:337–374.

Geoffrey IW. 2000 MultiBoosting: A technique for combining boosting and wagging. Machine Learn.

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

Glade T. 2003. Landslide occurrence as a response to land use change: a review of evidence from New Zealand. Catena. 51:297–314.

Gorsevski PV, Jankowski P. 2010. An optimized solution of multi-criteria evaluation analysis of landslide susceptibility using fuzzy sets and Kalman filter. Comput Geosci. 36:1005–1020.

Guha Sapir D, Vos F, Below R, Ponserre S. 2011. Annual Disaster Statistical Review 2011: The numbers and Trends. Cred: Brussels.

Guzzetti F. 2006. Landslide hazard and risk assessment. Bonn: University of Bonn.

Guzzetti F, Reichenbach P, Cardinali M, Galli M, Ardizzone F. 2005. Probabilistic landslide hazard assessment at the basin scale. Geomorphology. 72:272–299.

Hellerstein J, Thathachar J, Rish I. 2000. Recognizing end-user transactions in performance management. In Proceedings of AAAI-2000.

Highland LM, Bobrowsky P. 2008. The landslide handbook – a guide to understanding landslides. Reston (VA): USGS.

Ho TK. 1998. The random subspace method for constructing decision forests. Pattern Analysis and Machine Intelligence. IEEE Trans. 20:832–844.

Jebur MN, Pradhan B, Tehrany MS. 2015. Manifestation of LiDAR-derived parameters in the spatial prediction of landslides using novel ensemble evidential belief functions and support vector machine models in GIS. IEEE J Selected Topics Appl Earth Observations Remote Sens. 8:674–690.

Jones CM, Athanasiou T. 2005. Summary receiver operating characteristic curve analysis techniques in the evaluation of diagnostic tests. Soc Thoracic Surgeons. 365–375.

Kantardzic M. 2011. Data mining: concepts, models, methods, and algorithms. John Wiley & Sons.

Kavzoglu T, Kutlug Sahin E, Colkesen I. 2015. Selecting optimal conditioning factors in shallow translational landslide susceptibility mapping using genetic algorithm. Eng Geol. 192:101–112.

Kavzoglu T, Sahin EK, Colkesen I. 2015. An assessment of multivariate and bivariate approaches in landslide susceptibility mapping: a case study of Duzkoy district. Nat Hazards. 76:471–496.

Koyuncu H, Ceylan R. 2013. Artificial neural network based on rotation forest for biomedical pattern classification. Proceedings of the Telecommunications and Signal Processing (TSP), 2013 36th International Conference on; IEEE.

Kuncheva LI. 2004. Combining pattern classifiers. Chichester: Methods and Algorithms Wiley.

Kuncheva LI. 2014. Combining pattern classifiers: methods and algorithms. 2nd ed. Hoboken (NJ): John Wiley & Sons.

Lallianthanga RK, Lalbiakmawia F. 2013. Landslide Hazard Zonation Of Aizawl District, Mizoram, India using remote sensing and GIS techniques. Int J Remote Sensing Geosci (IJRSG). 2.

Lombardo L, Cama M, Conoscenti C, Märker M, Rotigliano E. 2015. Binary logistic regression versus stochastic gradient boosted decision trees in assessing landslide susceptibility for multiple-occurring landslide events: application to the 2009 storm event in Messina (Sicily, southern Italy). Nat Hazards. 79:1621–1648.

Martínez-Álvarez F, Reyes J, Morales-Esteban A, Rubio-Escudero C. 2013. Determining the best set of seismicity indicators to predict earthquakes. Two case studies: Chile and the Iberian Peninsula. Knowledge-Based Syst. 50:198–210.

Mease D, Wyner A. 2008. Evidence contrary to the statistical view of boosting. J Machine Learn Res. 9:131–156.

Mika. 2013. Weathering of Igneous Rocks. http://wwwgeomikacom/blog/2013/08/17/weathering-igneous/.

NCEP. 2014. Global Weather data for SWAT. http://globalweathertamuedu/home.

Petley D. 2012. Global patterns of loss of life from landslides. Geology. 40:927–930.

Pham BT, Bui DT, Prakash I, Dholakia M. 2016a. Evaluation of predictive ability of support vector machines and naive Bayes trees methods for spatial prediction of landslides in Uttarakhand state (India) using GIS. J Geomatics. 10:71–79.

Pham BT, Pradhan B, Tien Bui D, Prakash I, Dholakia MB. 2016b. A comparative study of different machine learning methods for landslide susceptibility assessment: a case study of Uttarakhand area (India). Environ Model Software. 10//;84:240–250.

Pham BT, Tien Bui D, Pham HV, Le HQ, Prakash I, Dholakia MB. 2016c. 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). J Indian Soc Remote Sens. 1–11.

Pham BT, Tien Bui D, Prakash I, Dholakia MB. 2016d. Hybrid integration of Multilayer Perceptron Neural Networks and machine learning ensembles for landslide susceptibility assessment at Himalayan area (India) using GIS. Catena. 2//;149, Part 1:52–63.

Pham BT, Tien Bui D, Prakash I, Dholakia MB. 2016e. Rotation forest fuzzy rule-based classifier ensemble for spatial prediction of landslides using GIS. Nat Hazards. 83:1–31.

Pham BT, Tien Bui D, Indra P, Dholakia M. 2015a. Landslide susceptibility assessment at a part of Uttarakhand Himalaya, India using GIS–based statistical approach of frequency ratio method. Int J Eng Res Technol. 4:338–344.

Pham BT, Tien Bui D, Pourghasemi HR, Indra P, Dholakia MB. 2015b. Landslide susceptibility assessment in the Uttarakhand area (India) using GIS: a comparison study of prediction capability of naïve Bayes, multilayer perceptron neural networks, and functional trees methods. Theor Appl Climatol. 122:1–19.

Polykretis C, Ferentinou M, Chalkias C. 2015. A comparative study of landslide susceptibility mapping using landslide susceptibility index and artificial neural networks in the Krios River and Krathis River catchments (northern Peloponnesus, Greece). Bull Eng Geol Environ. 74:27–45.

Pourghasemi HR, Mohammady M, Pradhan B. 2012. Landslide susceptibility mapping using index of entropy and conditional probability models in GIS: Safarood Basin, Iran. Catena. 97:71–84.

Pourghasemi HR, Pradhan B, Gokceoglu C, Mohammadi M, Moradi HR. 2013. Application of weights-of-evidence and certainty factor models and their comparison in landslide susceptibility mapping at Haraz watershed, Iran. Arabian J Geosci. 6:2351–2365.

Pradhan B. 2013. A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS. Comput Geosci. 51:350–365.

Pradhan B, Lee S. 2010. Delineation of landslide hazard areas on Penang Island, Malaysia, by using frequency ratio, logistic regression, and artificial neural network models. Environ Earth Sci. 60:1037–1054.

Ren F, Wu X, Zhang K, Niu R. 2015. Application of wavelet analysis and a particle swarm-optimized support vector machine to predict the displacement of the Shuping landslide in the Three Gorges, China. Environ Earth Sci. 73:4791–4804.

Rish I, Hellerstein J, Jayram T. 2001. An analysis of data characteristics that affect naive Bayes performance. New York (NY): IBM TJ Watson Research Center.

Rodriguez JJ. 2007. Rotation forest and random oracles: Two classifier ensemble methods. Proceedings of the Twentieth IEEE International Symposium on Computer-Based Medical Systems (CBMS'07); IEEE.

Rodriguez JJ, Kuncheva LI, Alonso CJ. 2006. Rotation forest: a new classifier ensemble method. In IEEE Transactions on Pattern Analysis and Machine Intelligence. 28:1619–1630.

Roggo Y, Duponchel L, Ruckebusch C, Huvenne J-P. 2003. Statistical tests for comparison of quantitative and qualitative models developed with near infrared spectral data. J Mol Struct. 654:253–262.

Sadr MP, Maghsoudi A, Saljoughi BS. 2014. Landslide susceptibility mapping of komroud sub-basin using Fuzzy logic approach. Geodynamics Res Int Bulletin. 2.

Shahabi H, Hashim M, Ahmad BB. 2015. Remote sensing and GIS-based landslide susceptibility mapping using frequency ratio, logistic regression, and fuzzy logic methods at the central Zab basin, Iran. Environ Earth Sci. 73:8647–8668.

Shahabi H, Khezri S, Ahmad BB, Hashim M. 2014. Landslide susceptibility mapping at central Zab basin, Iran: A comparison between analytical hierarchy process, frequency ratio and logistic regression models. Catena. 115:55–70.

Sharma A, Dey S. 2012. Performance investigation of feature selection methods and sentiment lexicons for sentiment analysis. IJCA Special Issue on Advanced Computing and Communication Technologies for HPC Applications. 3:15–20.

Stevens CW, Wolfe SA. 2012. High—resolution mapping of wet terrain within discontinuous permafrost using LiDAR intensity. Permafrost Periglacial Proc. 23:334–341.

Tatsunori Mori MK, Kazufumi Yoshida. 2002. Term weighting method based on information gain ratio for summarizing documents retrieved by IR systems. J Nat Language Proc. 9:3–32.

Tien Bui D. 2012. Modeling of rainfall-induced landslide hazard for the Hoa Binh province of Vietnam Aas [PhD thesis]. Norway: Norwegian University of Life Sciences.

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

Tien Bui D, Pham BT, Nguyen QP, Hoang N-D. 2016b. 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. Int J Digital Earth. 9:1–21.

Tien Bui D, Pradhan B, Revhaug I, Nguyen Ba D, Viet Pham H, Ngoc Bui Q. 2013. A novel hybrid evidential belief function-based fuzzy logic model in spatial prediction of rainfall-induced shallow landslides in the Lang Son city area (Vietnam). Geomatics, Nat Hazards Risk.

Tien Bui D, Pradhan B, Revhaug I, Nguyen DB, Pham HV, Bui QN. 2015. A novel hybrid evidential belief function-based fuzzy logic model in spatial prediction of rainfall-induced shallow landslides in the Lang Son city area (Vietnam). Geomatics, Nat Hazards Risk. 6:243–271.

Tien Bui D, Tien Ho C, Revhaug I, Pradhan B, Duy Nguyen B. 2014. Landslide susceptibility mapping along the national road 32 of Vietnam using GIS-based j48 decision tree classifier and its ensembles. In: Cartography from pole to pole. Berlin: Springer; p. 303–317.

Tsangaratos P, Ilia I. 2016. Landslide susceptibility mapping using a modified decision tree classifier in the Xanthi Perfection, Greece. Landslides. 13:305–320.

Tsangaratos P, Ilia I, Rozos D. 2013. Case event system for landslide susceptibility analysis. In: Claudio Margottini, Paolo Canuti, Kyoji Sassa, editors. Landslide science and practice. Berlin: Springer; p. 585–593.

Van Den Eeckhaut M, Vanwalleghem T, Poesen J, Govers G, Verstraeten G, Vandekerckhove L. 2006. Prediction of landslide susceptibility using rare events logistic regression: A case-study in the Flemish Ardennes (Belgium). Geomorphology. 76:392–410.

Van T, Anh D, Hieu H, Giap N, Ke T, Nam T, Ngoc D, Ngoc D, Thai T, Thang D. 2006. 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. Hanoi: Vietnam Institute of Geosciences and Mineral Resources.

Walter SD. 2002. Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data. Stat Med. 9:1237–1256.

Webb GI. 2000. Multiboosting: a technique for combining boosting and wagging. Machine Learning. 40:159–196.

Witten IH, Frank E, Mark AH. 2011. Data mining: practical machine learning tools and techniques. 3rd 765 ed. Burlington: Morgan Kaufmann.

Xia J, Du P, He X, Chanussot J. 2014. Hyperspectral remote sensing image classification based on rotation forest. Geosci Remote Sens Lett. 11:239–243.

Xu C, Dai F, Xu X, Lee YH. 2012. GIS-based support vector machine modeling of earthquake-triggered landslide susceptibility in the Jianjiang River watershed, China. Geomorphology. 145:70–80.

Youssef AM. 2015. Landslide susceptibility delineation in the Ar-Rayth area, Jizan, Kingdom of Saudi Arabia, using analytical hierarchy process, frequency ratio, and logistic regression models. Environ Earth Sci. 73:8499–8518.

Youssef AM, Al-Kathery M, Pradhan B. 2015. Landslide susceptibility mapping at Al-Hasher Area, Jizan (Saudi Arabia) using GIS-based frequency ratio and index of entropy models. Geosci J. 19:113–134.

Zadrozny B, Elkan C. 2001. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In Proceedings of the Eighteenth International Conference on Machine Learning. San Francisco (CA): Morgan Kaufmann; p. 609–616.

Zhang C-X, Zhang J-S. 2009. A novel method for constructing ensemble classifiers. Statistics Comput. 19:317–327.

Zhang H, Su J. 2004. Naive Bayesian classifiers for ranking. Machine Learn. 3201:501–512.