



A novel hybrid model of Bagging-based Naïve Bayes Trees for landslide susceptibility assessment

Binh Thai Pham¹ · Indra Prakash²

Received: 28 May 2017 / Accepted: 9 November 2017 / Published online: 25 November 2017
© Springer-Verlag GmbH Germany, part of Springer Nature 2017

Abstract Landslide susceptibility assessment was performed using the novel hybrid model Bagging-based Naïve Bayes Trees (BAGNBT) at Mu Cang Chai district, located in northern Viet Nam. The model was validated using the Chi-square test, statistical indexes, and area under the receiver operating characteristic curve (AUC). In addition, other models, namely the Rotation Forest-based Naïve Bayes Trees (RFNBT), single Naïve Bayes Trees (NBT), and Support Vector Machines (SVM), were selected for the comparison. Results show that the novel hybrid model (AUC = 0.834) outperformed the RFNBT (0.830), SVM (0.805), and NBT (0.800). This indicates that the BAGNBT is a promising and better alternative method for landslide susceptibility modeling and mapping.

Keywords Landslides · Machine learning · Naïve Bayes Trees · Bagging · GIS · India

Introduction

Landslide susceptibility assessment helps in the identification of potential landslide areas for better land use planning and management. Landslides are mostly impacted by topography, drainage, land cover, habitats, geological phenomenon (earthquakes, volcano), and weather conditions (rainfall) of the area. In recent years, many methods have been proposed and developed to predict landslides using various approaches such as a physical-based approach and statistical approach (Tien Bui et al. 2016a). While a physical-based approach is mostly impossible for assessment and prediction of landslides in large areas, a statistical approach is much more suitable and applicable for assessment of landslide susceptibility on a regional scale.

Many statistical methods have been developed for assessment of landslide susceptibility using Geographic Information Systems (GIS) in recent years. Methods such as the frequency ratio (Akgun et al. 2008; Lee and Sambath 2006), weight of evidence (Regmi et al. 2010), and evidential belief function (Althuwaynee et al. 2012; Lee et al. 2013) are known as

conventional statistical methods, whereas methods such as naïve Bayes (Pham et al. 2015; Tsangaratos 2016), artificial neural networks (Ermini et al. 2005; Pham et al. 2017c), neuro-fuzzy (Pradhan et al. 2010; Sezer et al. 2011), and Support Vector Machines (SVM) (Ballabio and Sterlacchini 2012; Pham et al. 2016a) are known as advanced machine learning methods and are used more efficiently and widely than conventional statistical methods. The advantage of machine learning methods is that they can handle better multi-variety and multi-dimensional data in uncertain or dynamic environments (Michalski et al. 2013). However, these machine learning methods can be further improved by using ensemble techniques that can create multiple algorithms to combine them for better results (Wang et al. 2012).

In recent decades, ensemble techniques such as Bagging (Breiman 1996), Rotation forests (RFs) (Rodriguez et al. 2006), MultiBoost (Webb 2000), and AdaBoost (Freund and Schapire 1995) have been widely used to solve a lot of real-world problems, including landslide prediction (Pham et al. 2017a). However, their application is still limited for the assessment of landslide susceptibility. Moreover, the performance of these ensemble techniques depends on the base classifiers used (Seni and Elder 2010). Therefore, the study of these ensemble techniques using different base classifiers is necessary for the assessment of landslide susceptibility.

In this study, the main objective is to propose a novel model, Bagging-based Naïve Bayes Trees (BAGNBT), which is a hybrid approach of a Naïve Bayesian Trees (NBT) classifier and Bagging (BAG) ensemble for the assessment of

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

¹ Department of Geotechnical Engineering, University of Transport Technology, 54 Trieu Khuc, Thanh Xuan, Ha Noi, Vietnam

² Department of Science & Technology, Government of Gujarat, Bhaskarcharya Institute for Space Applications and Geo-Informatics (BISAG), Gandhinagar, India

susceptibility of landslides. The Mu Cang Chai district located in northern part of Viet Nam was selected as the study area. Validation of the model was performed using the Chi-square (χ^2) test, statistical indexes, and area under the receiver operating characteristic (ROC) curve (AUC). In addition, other models, namely RF-based Naïve Bayes Trees (RFNBT), single Naïve Bayes Trees (NBT), and SVM, were selected for the comparison. AcrMap 10.2 and Weka 3.7.12 software were used for the data processing and modeling.

Methods used

Naïve Bayes Trees

NBT, which is a hybrid approach of naïve Bayesian and decision trees, was proposed by Kohavi (1996). Thus, the NBT takes advantages of both naïve Bayesian (which take into consideration evidence from many attributes to make the final decision) and decision trees (which is known as a very fast and comprehensive method for classification) to achieve better performance for classification (Kohavi 1996). In this study, the NBT is used as a base classifier in the ensemble framework to create the novel hybrid model for assessment of landslide susceptibility.

Basically, the NBT constructs a classification tree in which a decision tree is constructed at each node for splitting the datasets and a naïve Bayesian tree is built at each leaf to predict the variables (Kohavi 1996). The algorithm of the decision tree is based on the gain ratio (GR) values of variables as shown in Eq. 1 (Quinlan 1986); the Naïve Bayesian algorithm is shown in Eq. 2 (Murphy 2006):

$$GR(r, U) = \frac{Entropy(U) - \sum_{i=1}^m \frac{|U_i|}{|U|} Entropy(U_i)}{- \sum_{i=1}^m \frac{|U_i|}{|U|} \log_2 \frac{|U_i|}{|U|}} \tag{1}$$

$$t_{NB} = \underset{z_i}{argmax} PP(t_i) \prod_{i=1}^m \frac{1}{\sqrt{2\pi}\epsilon} e^{-\frac{(r_i - \sigma)^2}{2\epsilon^2}} \tag{2}$$

where U is the training dataset, r is the attribute used in the training dataset, m is the number of attributes, $PP(t_i)$ is the prior probability of the output variables $t_i = (1, 0)$, σ is the mean of r_i , and ϵ is the standard deviation of r_i .

Bagging ensemble

The BAG ensemble proposed by Breiman (1996) is a bootstrap aggregating method. It uses bootstrap replicates of the learning set to generate the multiple versions of a classifier and optimal learning datasets, and then these new version classifiers are combined using a plurality vote to create an

aggregated classifier for predicting the class (Breiman 1996). The BAG ensemble has been applied widely in many fields of medical science (West et al. 2005), computer sciences (Li 2007), and banking (Hsieh and Hung 2010). However, its application is still limited for landslide prediction. In this study, the BAG ensemble is combined with the NBT classifier to create the novel hybrid model for assessment of landslide susceptibility. Using the BAG algorithm, the overall probability of correct classification is shown as Eq. 3:

$$f(a) = \int \left[\sum_{z_i} F(z_i|a) P(z_i|a) \right] P_a d(a) \tag{3}$$

where $P(z_i|a)$ is the probability that the output class z_i is created by input attribute x , $F(z_i|a)$ is the relative frequency at which the output class z_i is predicted by input attribute x , and $P_x d(x)$ is the probability distribution of attribute x (Breiman 1996).

Rotation Forest

RRF, which is an efficient ensemble method, was applied effectively for landslide susceptibility assessment. This method was proposed by Rodriguez et al. (2006) using PCA (Principal Component Analysis) to extract features for generating optimal input datasets for classification tasks (Wold et al. 1987). In this study, the RF was selected for comparison as an ensemble technique using the base classifier of NBT for assessment of landslide susceptibility. The RF algorithm is based on the rotation matrix created using the base classifier and transformation method as follows (Eq. 4) (Rodriguez et al. 2006):

$$R_i = \begin{bmatrix} z_{i,1}^{(1)}, z_{i,1}^{(2)}, \dots, z_{i,1}^{(E_i)} & 0 & \dots & 0 \\ 0 & z_{i,1}^{(1)}, z_{i,1}^{(2)}, \dots, z_{i,1}^{(E_2)} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & z_{i,1}^{(1)}, z_{i,1}^{(2)}, \dots, z_{i,1}^{(E_k)} \end{bmatrix} \tag{4}$$

where $z_{i,1}^{(1)}, z_{i,1}^{(2)}, \dots, z_{i,1}^{(E_i)}$ are the coefficients of a matrix with the size of $E \times 1$, which is generated by randomly selecting from a set of instances, and $E = \frac{10}{T}$ is the number of instances for each subset T .

Support Vector Machines

SVM, proposed by Vapnik (1995), is a well-known model for classification and regression. In classification, the main principle of this model is to use statistical learning theory to find the linear hyper-plane for optimally splitting two variables. Kavzoglu et al. (2014) showed that SVM outperforms logistic regression model for mapping of shallow landslides. Tien Bui et al. (2016b) also observed that SVM outperforms other models such as kernel logistic regression and the logistic

Fig. 1 Mu Cang Chai district location map and landslide locations

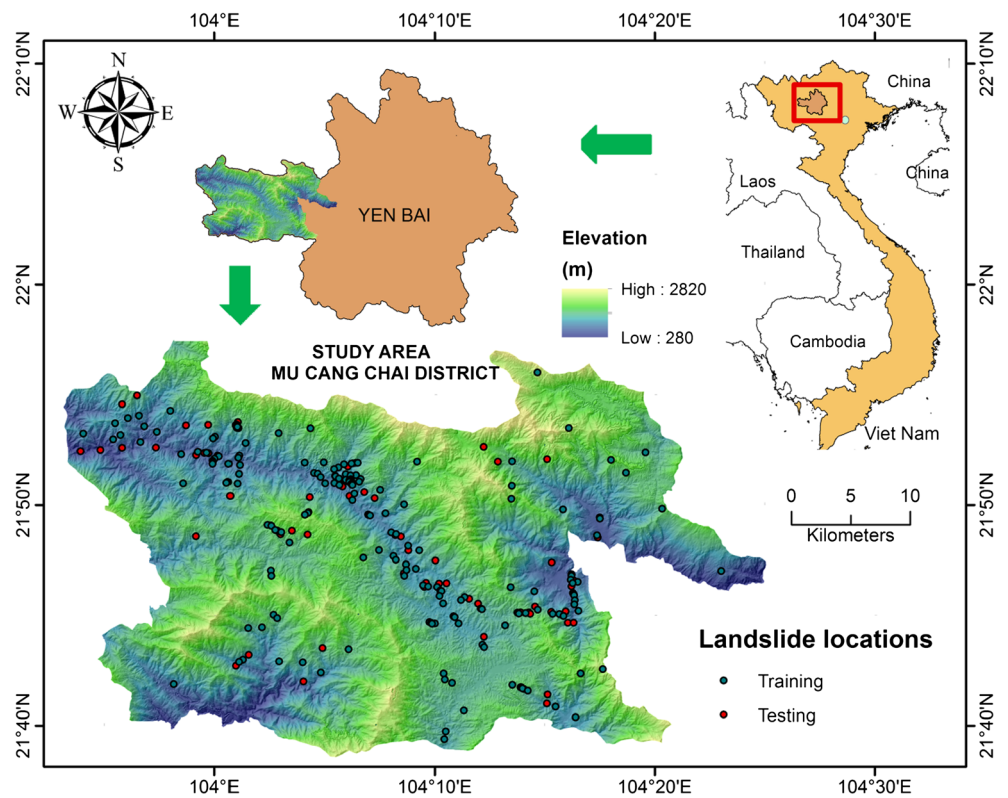
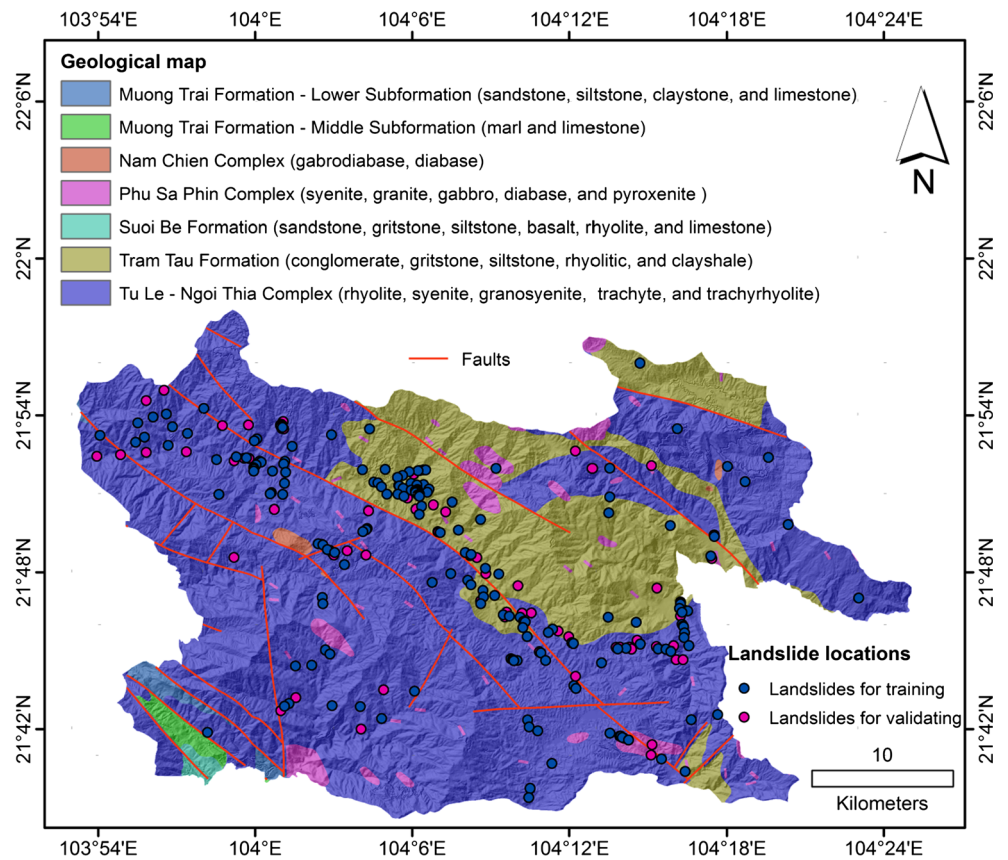


Fig. 2 Geological map of the Mu Cang Chai district



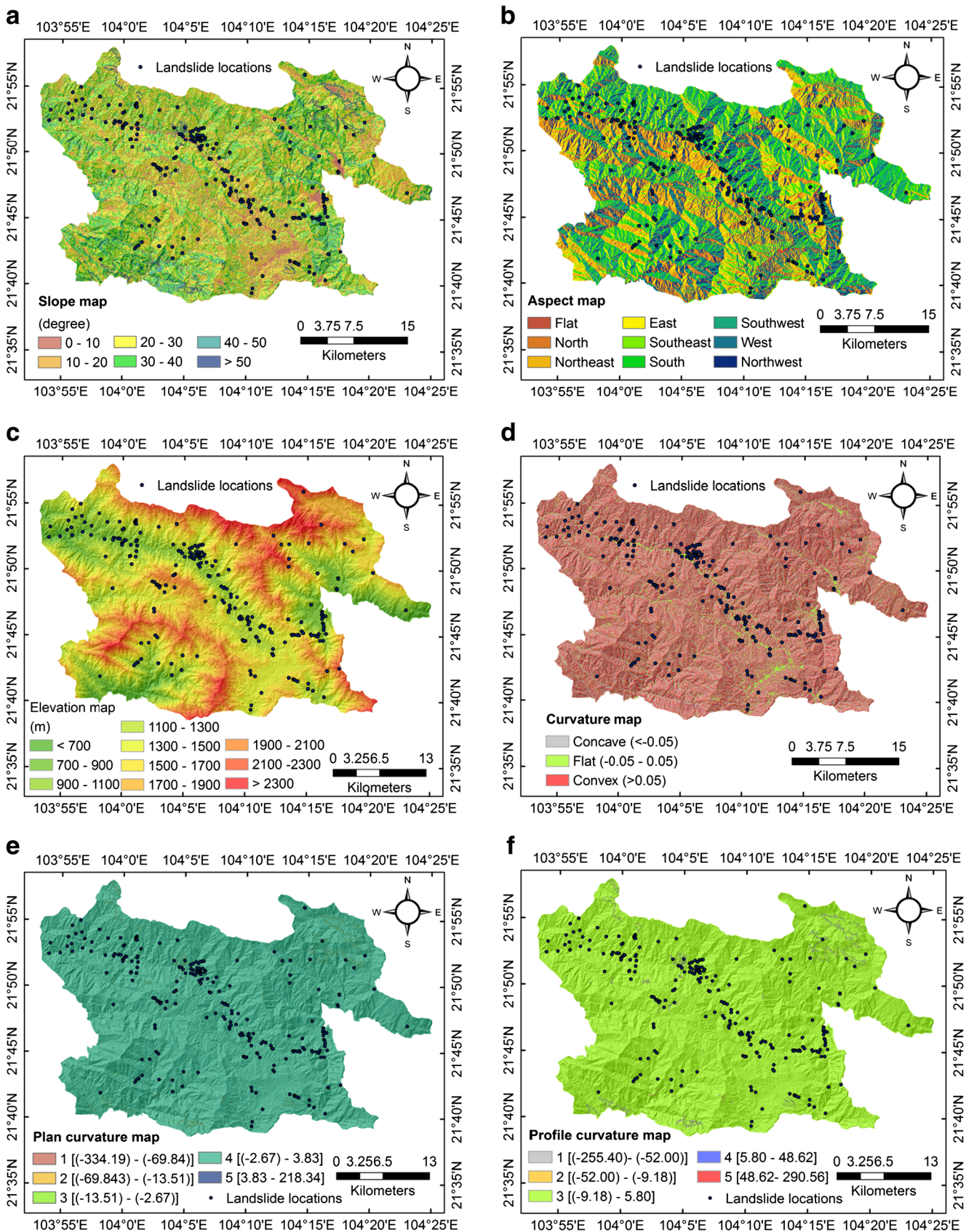


Fig. 3 Maps of landslide causing parameters: a slope; b aspect; c elevation; d curvature; e plan curvature; f profile curvature; g lithology; h rainfall; i land use; j distance to roads; k road density; l distance to rivers; m river density; n distance to faults; and o fault density

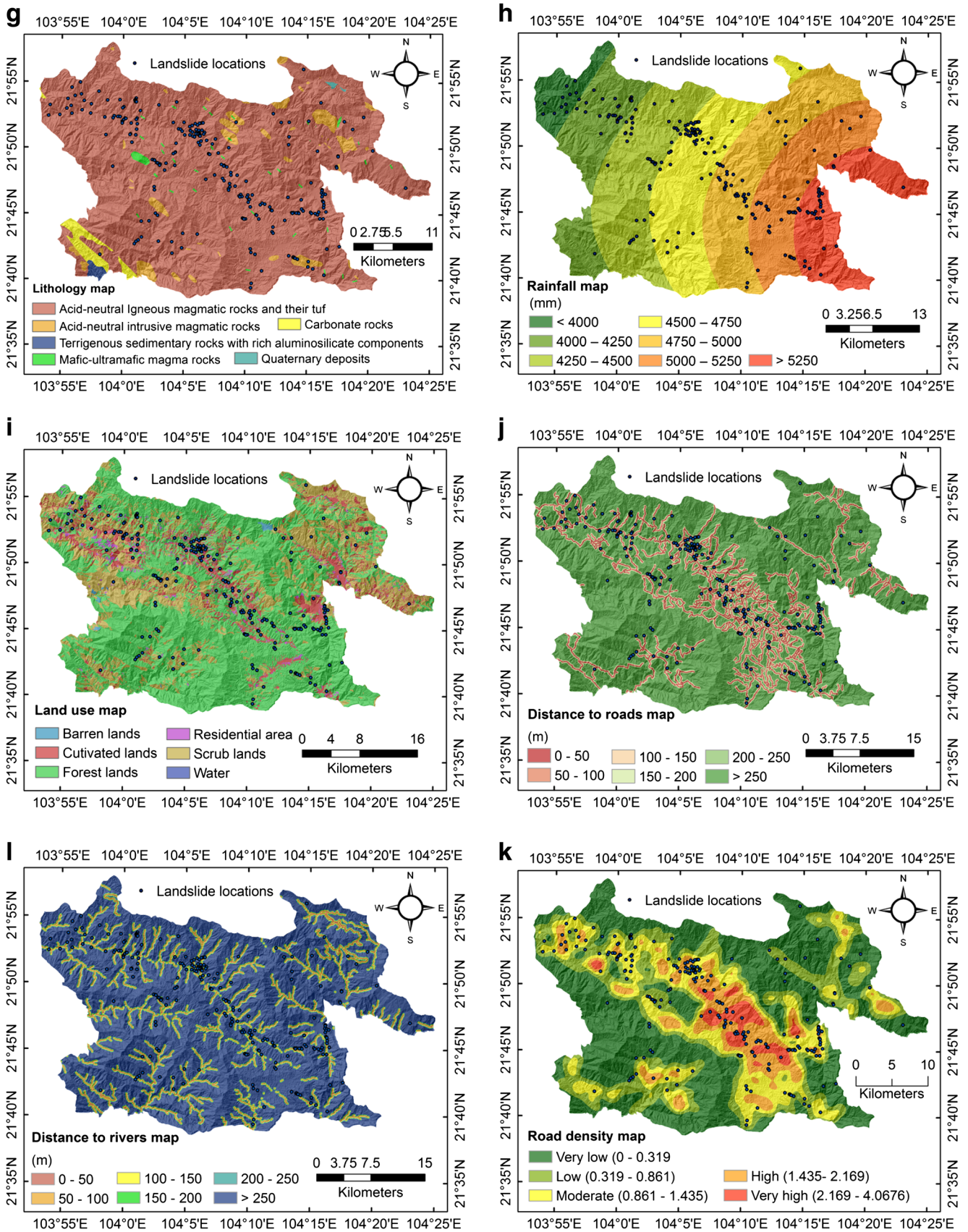


Fig. 3 (continued)

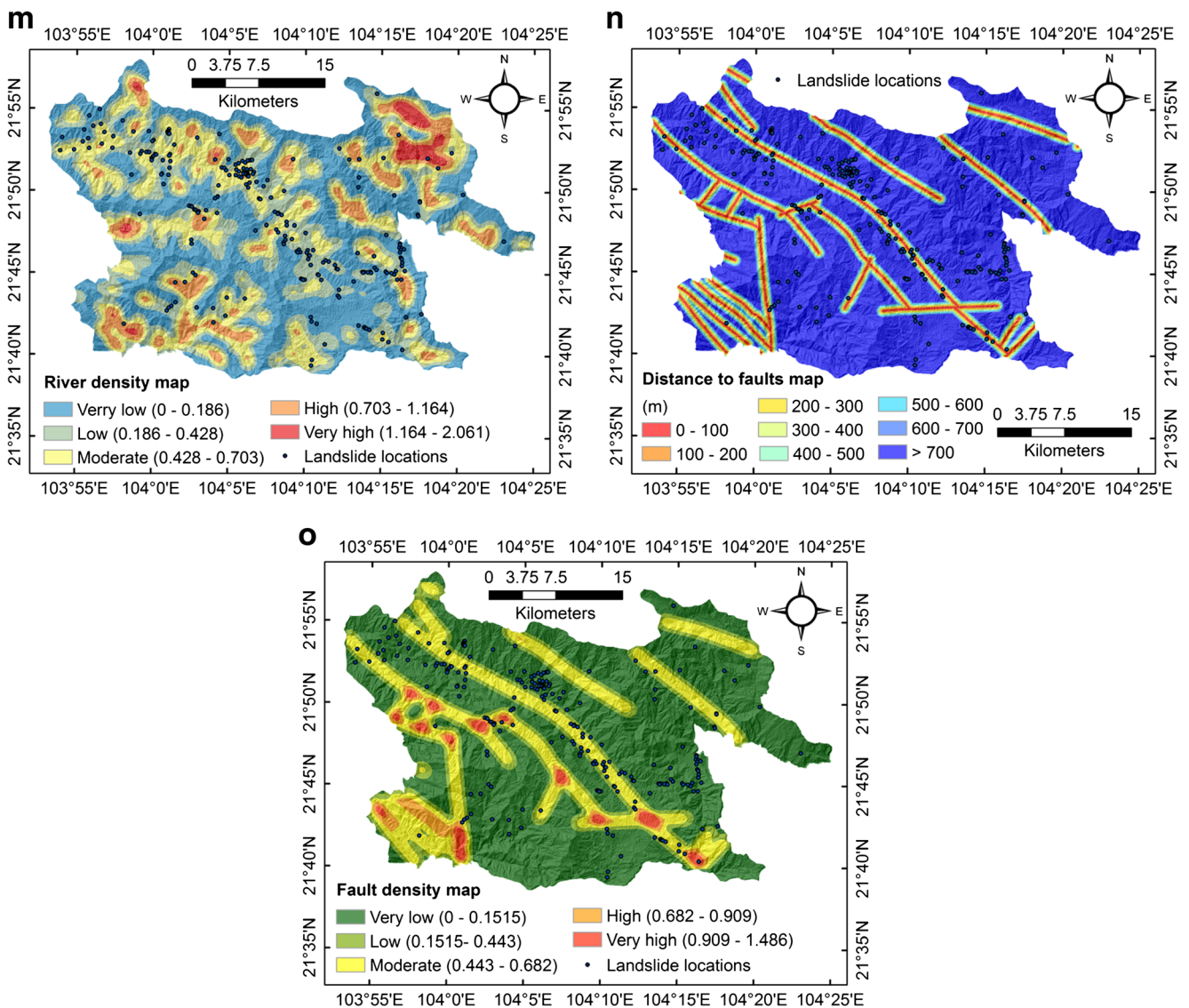


Fig. 3 (continued)

model tree. In this study, SVM is utilized for comparison with a novel proposed model. The SVM algorithm used in this study is expressed as Eqs. 5 and 6 (Vapnik 1995):

$$f(v_i) = \text{sign} \left[\sum_{i=1}^m \pi_i u_i \kappa(v, v_i) + q \right] \tag{5}$$

$$\kappa(v, v_i) = \exp \left\{ -\eta \|v - v_i\|^2 \right\} \tag{6}$$

where $f(v_i)$ is the decision function of the SVM algorithm of the training dataset (v_i, u_i) in which v_i are inputs and u_i are outputs, m is the number of instances in the inputs, π_i is the Lagrange multiplier, $\kappa(v, v_i)$ is the radial basis function (RBF) kernel, η is the kernel parameter, and q is the term of bias.

Validation methods

Methods such as the AUC, statistical indexes, and χ^2 test were utilized for validation of the models in this study.

The AUC is often used as a common index for validating models (Pham et al. 2017b; Shirzadi et al. 2017). The AUC is determined by analysis of the ROC curve plotted using pairs of statistical values, namely “sensitivity” and “100-specificity” (Feizizadeh et al. 2017). As the AUC equals 1, the performance of models is considered perfect. A higher AUC indicates better models.

Statistical indexes such as root mean squared error (RMSE), kappa (κ), negative predictive value (NPV), positive predictive value (PPV), specificity (SPF), sensitivity (SST), and accuracy (ACC) were utilized to validate the models (Pham et al. 2016a; Tien Bui et al. 2016a). Values of these

Table 1 Importance of the parameters for modeling using the Relief F technique

No.	Landslide-causing parameters	Average Merit	Standard deviation
1	Distance to roads	0.046	±0.005
2	Road density	0.039	±0.005
3	Elevation	0.029	±0.003
4	Rainfall	0.019	±0.003
5	Aspect	0.012	±0.003
6	River density	0.01	±0.002
7	Fault density	0.007	±0.001
8	Lithology	0.005	±0.001
9	Land use	0.005	±0.001
10	Distance to faults	0.004	±0.002
11	Slope	0.002	±0.002
12	Distance to rivers	0.001	±0.002
13	Plan curvature	0	±0
14	Curvature	-0.001	±0.003
15	Profile curvature	0	±0

indexes were calculated using confusion matrix values, namely false negatives (FN), true negatives (TN), false positives (FP), and true positives (TP). In general, higher PPV, NPV, SST, SPF, ACC, and κ indicate better models (Tien Bui et al. 2016b). In contrast, lower RMSE indicates better models.

The χ^2 test used for the comparison of difference of the models is based on the null hypothesis that there is no difference in the predictive capability of models and that the significant level ($p = 0.05$) is set up, and thus χ^2 values can be calculated (Tallarida and Murray 1987). The critical χ^2 statistic value for $p = 0.05$ (95% confidence level) with 1 degree of freedom is 3.84. Thus, if the significant level is smaller than 0.05 ($p < 0.05$) and the χ^2 value exceeds the threshold value of

3.841, then the null hypothesis is rejected, which means the difference of the models is significant (Kuncheva 2004).

Study area

The Mu Cang Chai district is located in the northern part of Viet Nam between the latitudes of 21°39'00"N to 21°50'00"N and longitudes of 103°56'00"E to 104°23'00"E, covering an area of about 1196.47 km² (Fig. 1). The area has a humid subtropical climate with 81% average humidity. Intensive torrential precipitation occurs during the rainy season in the months of August and September. Annual rainfall ranges from 3700 to 5490 mm with an average rainfall of 4630 mm. Seasonal temperatures in this area generally vary from 21 to 28 °C.

The topography of the area is highly variable, with elevation ranges from 280 to 2820 m. About 82% of the terrain has slope angles ranging from 10 to 40 °. Forest land occupies the largest part of the total land area (61.76%). Other land covers include scrub land, cultivated barren land, and built up areas.

Geologically, the area is occupied mainly by volcanic rocks of Tu Le–Ngoi Thia complex followed by Tram Tau formation. Intrusive igneous rocks of Phu Sa Phin complex are present in the area in patches. Other complexes and formations, namely Nam Chien complex (igneous rocks), Suoi Be formation (intercalation of sedimentary and igneous rocks), and Muong Trai formation (sedimentary rocks), occupy a small part of the area (Fig. 2).

The area is tectonically disturbed, having complex folded and faulted geological structures. Faults and litho-units are aligned in a northwest–southeast direction. Three main faults, namely Nam Co–Minh An, Phong Tho–Van Yen, and Nghia Lo, traverse the area. The major drainage is controlled by faults.

Fig. 4 Flow chart of the methodology adopted in the present study. BAG bagging, GIS Geographic Information Systems, LSM landslide susceptibility map, NBT Naïve Bayes Trees, ROC receiver operating characteristic

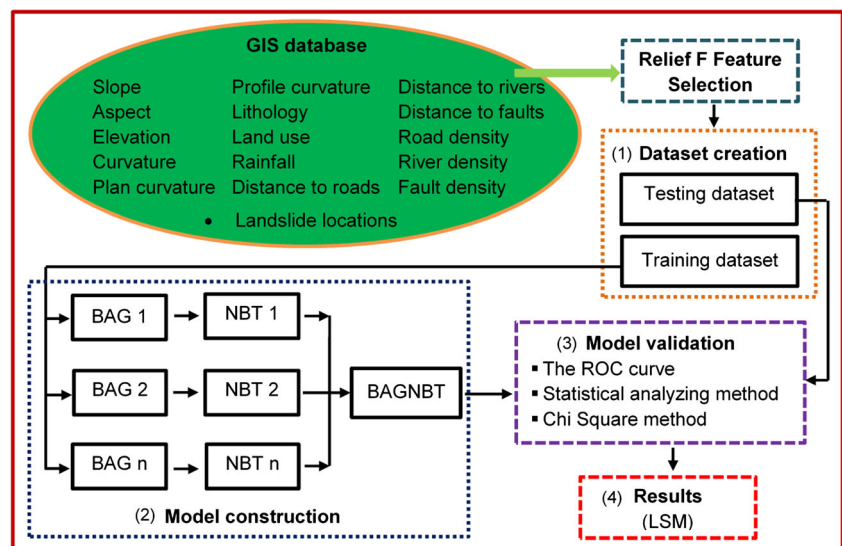
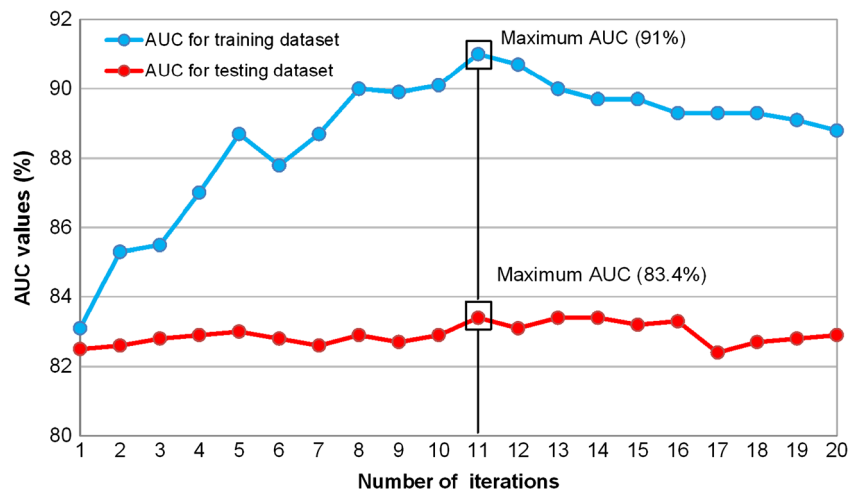


Fig. 5 Area under the receiver operating characteristic curve (AUC) of the Bagging-based Naïve Bayes Trees (BAGNBT) utilizing various numbers of iterations



Data used

Images, geological maps, road maps, and other relevant data for the area have been obtained from various sources including the Vietnam Institute of Geosciences and Mineral Resources, who have carried out surveys, assessment, and a landslide zoning warning study in the mountainous region of Vietnam.

Landslide inventory map

A landslide inventory map was constructed using a total of 248 historical landslide events identified by aerial photos of the year 2013 at a scale of 1:33,000 and Google Earth images after field checking the data. Types of landslide present in the area include rotational (124 events), debris slides (eight

events), translational (35 events), mixed (36 events), and toppling (45 events). The size of landslides varies from thousands of cubic meters to a few cubic meters. The largest landslide observed was of a volume of 100,000 m³, which occurred in February 2011 at Che Cu Na commune.

Landslide causing parameters

Fifteen landslide-causing parameters, namely slope, distance to faults, curvature, road density, profile curvature, aspect, plan curvature, river density, lithology, elevation, distance to roads, distance to rivers, rainfall, fault density, and land use, were selected for the assessment of landslide susceptibility (Fig. 3). Maps of these factors were generated for the analysis.

Maps of slope (Fig. 3a), aspect (Fig. 3b), elevation (Fig. 3c), curvature (Fig. 3d), plan curvature (Fig. 3e), and

Table 2 Validation of the models using the training dataset

No	Statistical indexes	BAGNBT	RFNBT	SVM	NBT
1	TP	147	143	138	134
2	TN	141	140	131	130
3	FP	27	32	36	41
4	FN	33	33	43	43
5	PPV (%)	84.48	81.71	79.31	76.57
6	NPV (%)	81.03	80.92	75.29	75.14
7	SST (%)	81.67	81.25	76.24	75.71
8	SPF (%)	83.93	81.40	78.44	76.02
9	ACC (%)	82.76	81.32	77.30	75.86
10	κ	0.655	0.652	0.546	0.523
11	RMSE	0.355	0.369	0.395	0.420

ACC accuracy, BAGNBT Bagging-based Naïve Bayes Trees, FN false negatives, FP falsepositives, κ kappa, NBT Naïve Bayes Trees, NPV negative predictive value, PPV positive predictive value, RFNBT Rotation Forest-based Naïve Bayes Trees, RMSE Root Mean Squared Error, SPF specificity, SST sensitivity, SVM Support Vector Machines, TN true negatives, TP true positives

Table 3 Validation of the models using the validating dataset

No	Statistical indexes	BAGNBT	RFNBT	SVM	NBT
1	TP	56	55	56	55
2	TN	54	53	51	50
3	FP	18	19	20	21
4	FN	20	21	21	22
5	PPV (%)	75.68	74.32	73.68	72.37
6	NPV (%)	72.97	71.62	70.83	69.44
7	SST (%)	73.68	72.37	72.73	71.43
8	SPF (%)	75.00	73.61	71.83	70.42
9	ACC (%)	74.32	72.97	72.30	70.95
10	κ	0.487	0.484	0.473	0.462
11	RMSE	0.414	0.419	0.424	0.426

ACC accuracy, BAGNBT Bagging-based Naïve Bayes Trees, FN false negatives, FP falsepositives, κ kappa, NBT Naïve Bayes Trees, NPV negative predictive value, PPV positive predictive value, RFNBT Rotation Forest-based Naïve Bayes Trees, RMSE Root Mean Squared Error, SPF specificity, SST sensitivity, SVM Support Vector Machines, TN true negatives, TP true positives

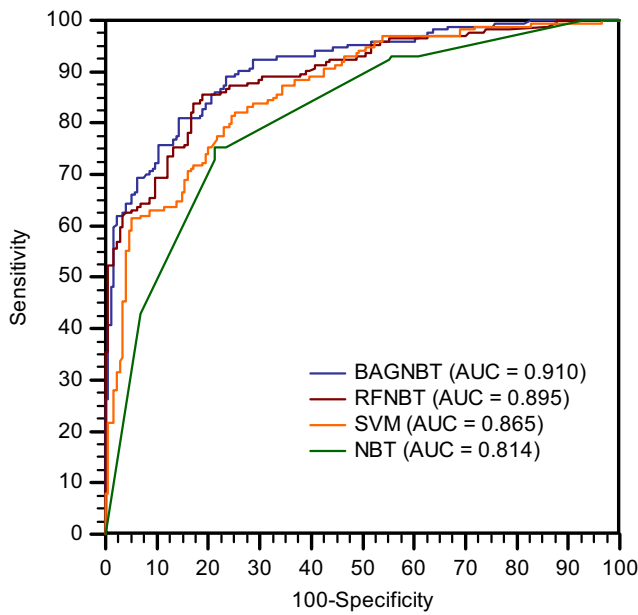


Fig. 6 Area under the receiver operating characteristic curve (AUC) of the models utilizing the training dataset. BAGNBT Bagging-based Naïve Bayes Trees, NBT Naïve Bayes Trees, RFNBT Rotation Forest-based Naïve Bayes Trees, SVM Support Vector Machines

profile curvature (Fig. 3f) were generated from a Digital Elevation Model (DEM) (20 m × 20 m) created from the contours extracted from the national topographical map (1:50,000 scale). Geological and mineral resources maps (1:50,000 scale) were used to extract lithology map of the study area (Fig. 3g). The rainfall map was constructed from the weather data (NCEP 2014) of the period 1984–2014 (Fig. 3h). Aerial

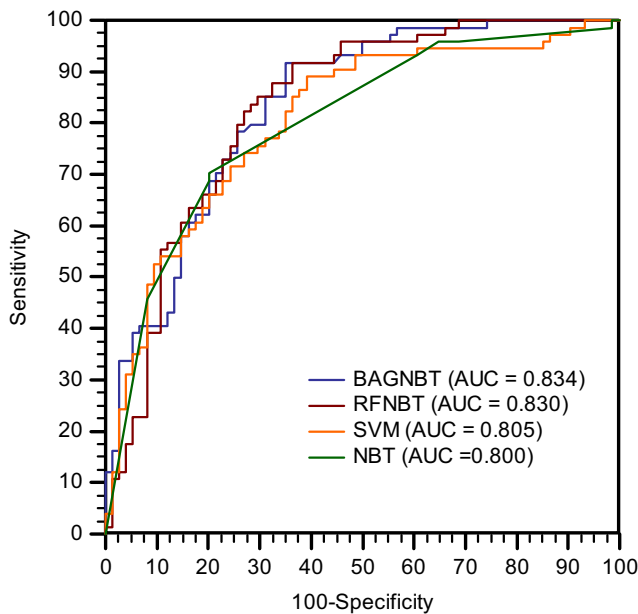


Fig. 7 Area under the receiver operating characteristic curve (AUC) of the models utilizing the validating dataset. BAGNBT Bagging-based Naïve Bayes Trees, NBT Naïve Bayes Trees, RFNBT Rotation Forest-based Naïve Bayes Trees, SVM Support Vector Machines

Table 4 Chi-square test of the models utilizing the training dataset

No.	Pairs	χ^2	<i>p</i> -Value	Significant
1	BAGNBT–RFNBT	265.465	< 0.0001	Yes
2	BAGNBT–SVM	265.465	< 0.0001	Yes
3	BAGNBT–NBT	251.792	< 0.0001	Yes
4	RFNBT–SVM	249.265	< 0.0001	Yes
5	RFNBT–NBT	248.702	< 0.0001	Yes
6	NBT–SVM	269.160	< 0.0001	Yes

BAGNBT Bagging-based Naïve Bayes Trees, NBT Naïve Bayes Trees, RFNBT Rotation Forest-based Naïve Bayes Trees, SVM Support Vector Machines, χ^2 Chi-squared

photos (1:33,000 scale) of year 2013 were used to identify landslide locations and to construct the land use map (Fig. 3i). Roads were created from the topographical map (1:50,000 scale), and then the distance to roads map (Fig. 3j) and road density map (Fig. 3k) were generated. Similarly, rivers were created from the topographical map (1:50,000 scale), and then the distance to rivers map (Fig. 3l) and river density map (Fig. 3m) were generated. In addition, faults were delineated from the national geological map (1:50,000 scale), and then the distance to faults map (Fig. 3n) and fault density map (Fig. 3o) were constructed.

Relief F technique for elimination and selection of landslide-causing parameters

Proper selection of landslide-causing parameters is very important to improve the effectiveness of modeling (Pham et al. 2016a). Therefore, it is necessary to remove irrelevant parameters to enhance the performance of the modeling. In this study, the Relief F feature selection technique was selected to test the importance of parameters used for modeling. This method is an efficient feature selection method proposed by Kira and Rendell (1992) to handle noise and complex datasets for modeling (Hall 2000). The main principle of Relief F is to select the feature using the landslide-causing parameters randomly, compute their nearest neighbors, and then adjust a

Table 5 Chi-square test of the models utilizing the validating dataset

No.	Pairs	χ^2	<i>p</i> -Value	Significant
1	BAGNBT–RFNBT	101.294	< 0.0001	Yes
2	BAGNBT–SVM	103.867	< 0.0001	Yes
3	BAGNBT–NBT	104.267	< 0.0001	Yes
4	RFNBT–SVM	104.915	< 0.0001	Yes
5	RFNBT–NBT	99.709	< 0.0001	Yes
6	NBT–SVM	94.174	< 0.0001	Yes

BAGNBT Bagging-based Naïve Bayes Trees, NBT Naïve Bayes Trees, RFNBT Rotation Forest-based Naïve Bayes Trees, SVM Support Vector Machines, χ^2 Chi-squared

weighting vector to give more weight to parameters that discriminate the case from neighbors of different classes (Wang and Makedon 2004). In general, as parameters are assigned higher weights, their predictive capability is higher, and vice versa. Parameters assigned weights of zero or sub-zero have no predictive capability for modeling and therefore such parameters should be removed from the dataset for modeling.

Feature selection results using the Relief F method in the study area are shown in Table 1. It shows that, out of 15 parameters, three, namely plan curvature (Average Merit [AM] = 0), curvature (-0.01), and profile curvature (0), have no contribution to landslide modeling as their weights equal zero or sub-zero. The other 12 parameters, distance to roads (AM = 0.046), road density (0.039), elevation (0.029), rainfall (0.019), aspect (0.012), river density (0.01), fault density (0.007), lithology (0.005), land use (0.005), distance to faults (0.004), slope (0.002), and distance to rivers (0.001), are important for landslide modeling. Thus, they were considered in the generation of final datasets for modeling.

Assessment of landslide susceptibility using the Bagging-based Naïve Bayes Trees (BAGNBT) model

The methodology of this study includes four main steps (Fig. 4):

- (1) *Creation of datasets:* Training and validating datasets were created from the landslide data collected from the study area. Of these datasets, the training dataset was employed to construct the models while the validating dataset was employed to evaluate the models. In total, there were 248 landslides recorded in the area, of which 174 landslides (70% of landslides) and 174 non-landslides were utilized for the training dataset and 74 landslides (30% of the remaining landslides) and 74 other non-landslides were employed for the validating dataset. In the datasets, landslides were assigned as “1” whereas non-landslides were assigned as “0” to facilitate the modeling process. In addition, 12 selected parameters (distance to roads, road density, lithology, elevation, distance to faults, rainfall, aspect, land use, river density, fault density, slope, distance to rivers) were sampled with training and validating landslide data to generate datasets for model study.
- (2) *Training the novel model:* the novel model of BAGNBT was constructed using the training dataset. In the modeling process, the BAG ensemble was first applied to optimize the input dataset for classification. The number of iterations was determined to be 11 for the best training of the BAGNBT model according to

Fig. 8 Maps of landslide susceptibility utilizing the models: a Bagging-based Naïve Bayes Trees (BAGNBT); b Rotation Forest-based Naïve Bayes Trees (RFNBT); c Support Vector Machines (SVM); and d Naïve Bayes Trees (NBT)

AUC analysis (Fig. 5). Simultaneously, the NBT classifier was applied to classify the classes of landslide and non-landslide for spatial prediction of landslides using the optimized sub-training datasets. In the final stage, the BAG ensemble was used to combine the generated NBT classifiers to construct the novel model.

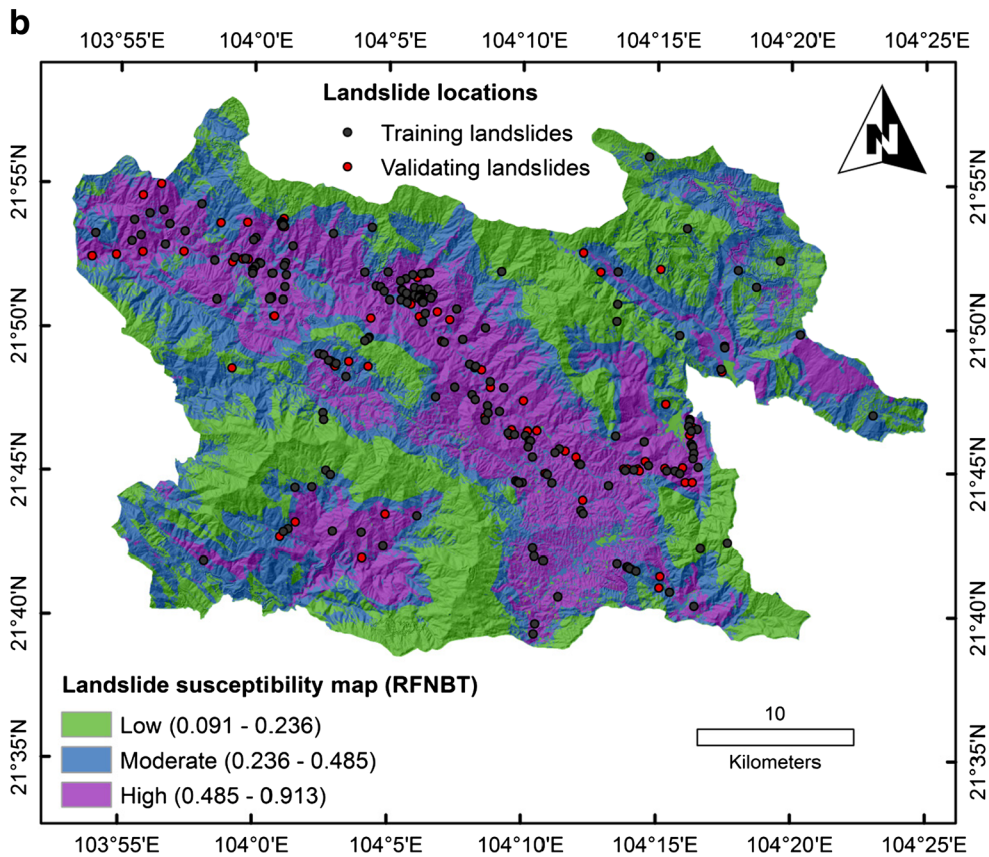
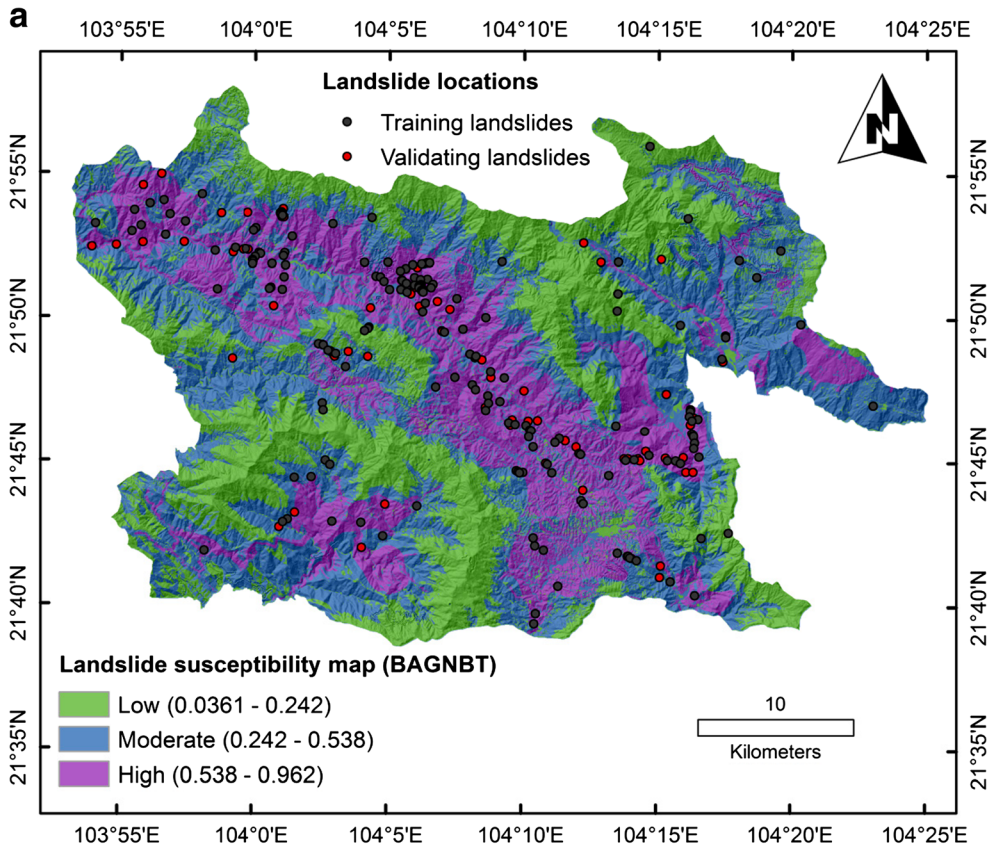
- (3) *Validating the novel model:* the novel model was validated using various methods, namely statistical indexes (NPV, PPV, SPF, SST, κ , ACC, and RMSE), the AUC, and χ^2 test. In addition, other methods such as RFNBT, SVM, and NBT were used for the comparison.
- (4) *Developing maps of landslide susceptibility:* Maps of landslide susceptibility were developed employing the BAGNBT, RFNBT, NBT, and SVM models. For developing the maps, indexes of landslide susceptibility were first created for pixels of the total study area by utilizing the applied models. Thereafter, the susceptible classes of the maps were classified based on the classification of those susceptible indexes. The geometrical interval method (Frye 2007) was employed in the present study for the classification as it is considered more efficient than other methods such as natural breaks, equal intervals, and standard deviation (Ayalew et al. 2004).

Results and discussion

Model validation

Validation of models was done using statistical indexes (Tables 2 and 3). Results of the training dataset show that the BAGNBT has the highest values of PPV (84.48%), NPV (81.03%), SST (81.67%), SPF (83.93%), ACC (82.76%), and κ (0.655) in comparison with the RFNBT, SVM, and NBT. In contrast, the BAGNBT has the lowest value of RMSE (0.355) in comparison with the RFNBT (0.369), SVM (0.395), and NBT (0.420). Similarly, as for the validating dataset, the results indicate that the BAGNBT has the highest values of PPV (75.68%), NPV (72.97%), SST (73.68%), SPF (75.00%), ACC (74.32%), and κ (0.487) in comparison with the RFNBT, SVM, and NBT. In contrast, the BAGNBT has the lowest value of RMSE (0.414) in comparison with the RFNBT (0.419), SVM (0.424), and NBT (0.426).

Validation results of the models using the AUC are shown in Figs. 6 and 7. Results from the training dataset indicate that



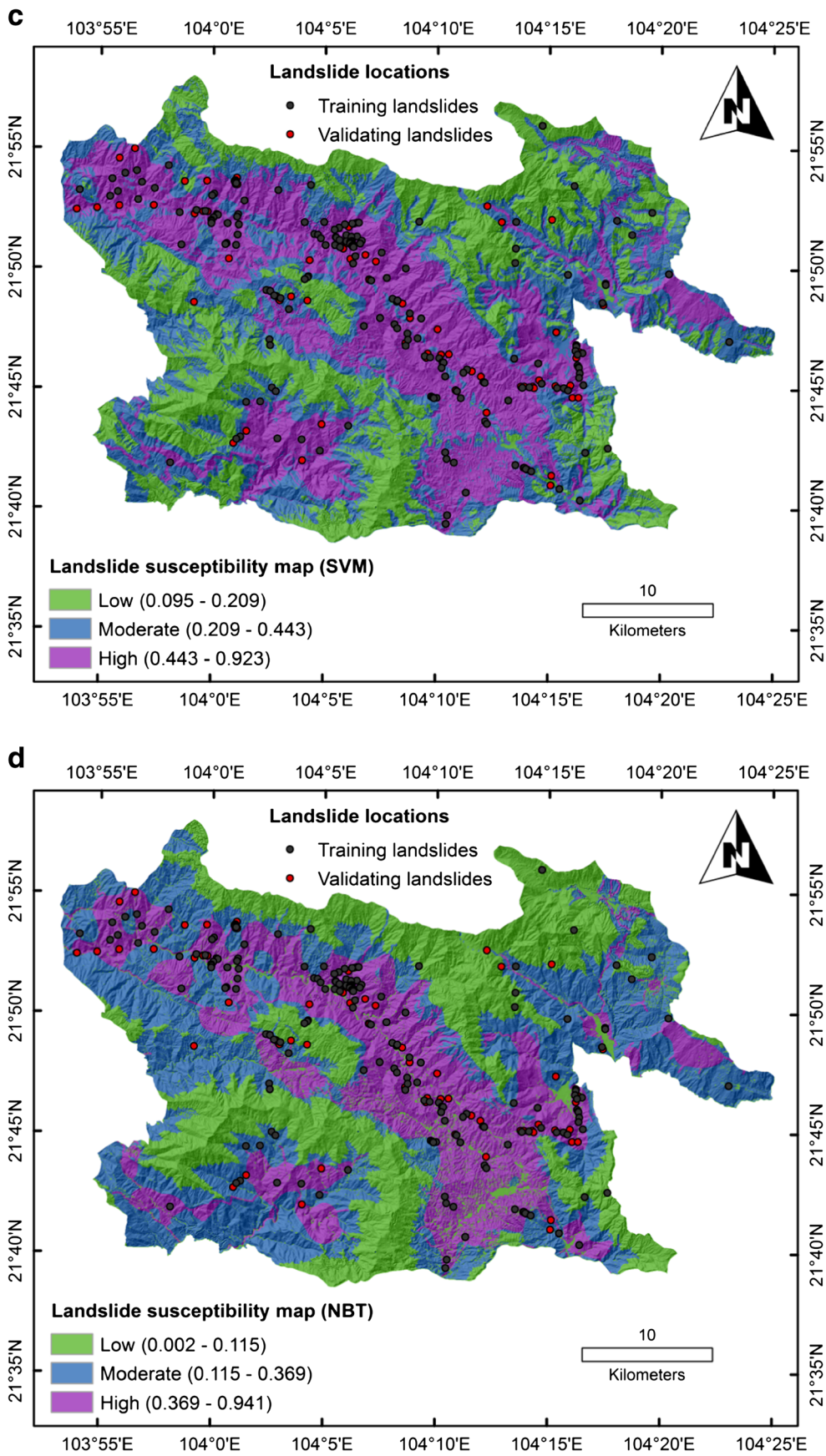


Fig. 8 (continued)

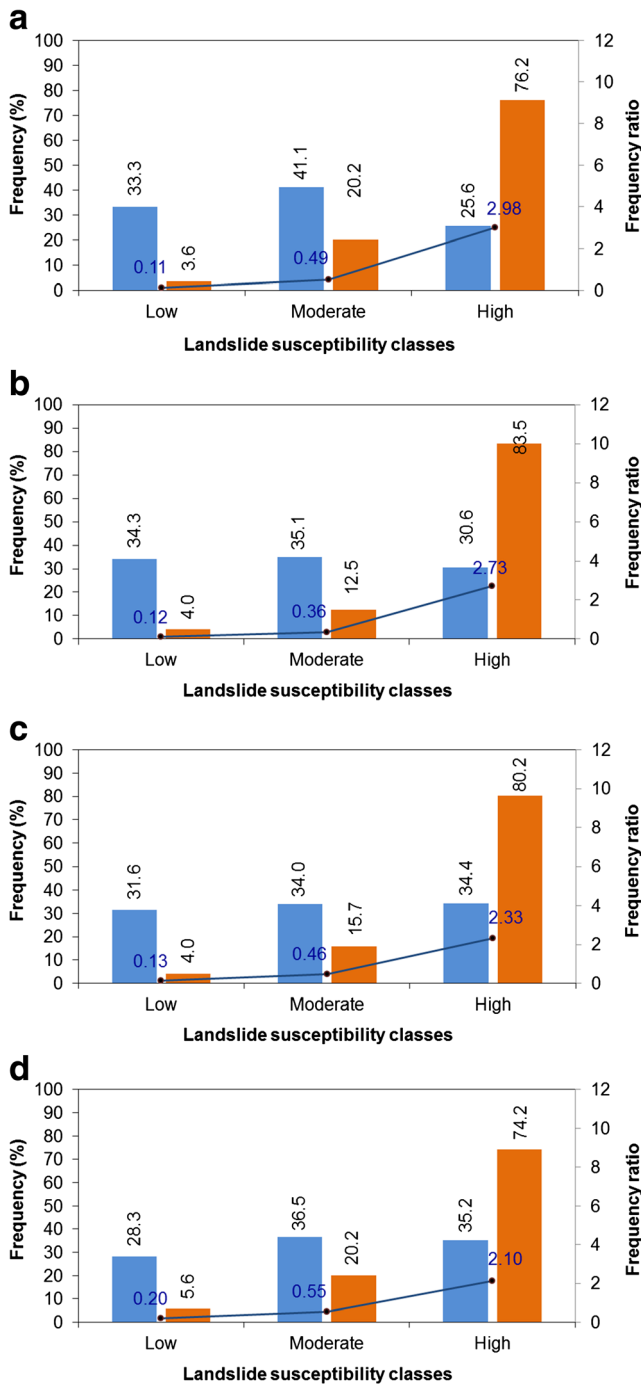


Fig. 9 Frequency analysis on the susceptibility maps utilizing the models: a Bagging-based Naïve Bayes Trees (BAGNBT); b Rotation Forest-based Naïve Bayes Trees (RFNBT); c Support Vector Machines (SVM); and d Naïve Bayes Trees (NBT)

the BAGNBT has a better AUC value (0.91) than the RFNBT (0.895), SVM (0.865), and NBT (0.814). Similarly, results from the validating dataset show that the BAGNBT has a better AUC value (0.834) than the RFNBT (0.830), SVM (0.805), and NBT (0.800).

Validation results of the models using the χ^2 test are shown in Tables 4 and 5. The results show that the χ^2 values of the

comparative pairs of the models are much higher than the threshold value of 3.841 for the training and validating datasets. Moreover, the *p*-values of the comparative pairs of models are much lower than the significant level of 0.05. Therefore, the differences in the predictive capability of the models are significant.

The results from the evaluation indicate that the applied models have good performance in assessing landslide susceptibility in this study. However, the BAGNBT has the highest performance in comparison with the RFNBT, SVM, and NBT. The analysis result is reasonable because the BAGNBT is an ensemble classifier approach of BAG and NBT. Of these, the BAG ensemble is able to reduce the variance of the prediction by creating optimal input data from the original dataset using combinations with repetitions for training the hybrid model (Breiman 1996), whereas the NBT classifier is also an efficient hybrid method that takes advantage of both efficient classifiers of the naïve Bayes and decision trees classifiers (Kohavi 1996). Comparison results also show that the RFNBT model is better than the SVM model as the RF ensemble uses the PCA which can help not only in reducing the dimensionality of complex datasets but also provides an easy way to train the datasets to get better performance from the model (Bro and Smilde 2014).

Landslides susceptibility maps

Maps of landslide susceptibility with three susceptibility classes (high, moderate, and low) were developed using the models shown in Fig. 8. Analysis of frequency was performed to evaluate the reliability of these maps (Fig. 9). The frequency ratio is defined as the ratio of the percentage of observed landslides and percentage of total area on each susceptible zone (Pham et al. 2016b). Results show that the frequency ratio value is the highest for the high class of susceptibility, followed by moderate and low values for all of the generated susceptibility maps. Evaluation results indicate that although all of the developed susceptibility map maps performed well, the map developed using BAGNBT outperforms those developed by other models.

Concluding remarks

Landslide susceptibility assessment was performed at the Mu Cang Chai district of Viet Nam applying the proposed hybrid model—a hybrid approach of the NBT classifier and the BAG ensemble. The AUC, statistical indexes, and χ^2 test were used for validation. In addition, other known models such as the RFNBT, SVM, and NBT were selected for comparison. Based on the feature selection method, 12 of 15 landslide-causing parameters, namely distance to roads, road density, elevation, rainfall, distance to faults, aspect, land use, river density, fault

density, lithology, slope, and distance to rivers, were selected for better modeling of landslide susceptibility.

Evaluation and comparison results of the models indicate that the novel hybrid BAGNBT model has the highest performance for landslide susceptibility assessment (AUC = 0.834) in comparison with the RFNBT (0.830), SVM (0.805), and NBT (0.800). Thus, the BAGNBT indicates as a promising and better alternative method for the assessment of landslide susceptibility. Maps of landslide susceptibility developed from the models would be helpful for proper land use planning and management.

Acknowledgements The authors express their sincere thanks to the Vietnam Institute of Geosciences and Mineral Resources for providing the data and to the Director of BISAG, Gujarat, India for providing facilities for this research work.

References

- Akgun A, Dag S, Bulut F (2008) Landslide susceptibility mapping for a landslide-prone area (Findikli, NE of Turkey) by likelihood-frequency ratio and weighted linear combination models. *Environ Geol* 54:1127–1143
- Althuwaynee OF, Pradhan B, Lee S (2012) Application of an evidential belief function model in landslide susceptibility mapping. *Comput Geosci* 44:120–135
- Ayalew L, Yamagishi H, Ugawa N (2004) Landslide susceptibility mapping using GIS-based weighted linear combination, the case in Tsugawa area of Agano River, Niigata Prefecture, Japan. *Landslides* 1:73–81
- Ballabio C, Sterlacchini S (2012) Support vector machines for landslide susceptibility mapping: the Staffora River basin case study, Italy. *Mathematical geosciences* 44:47–70
- Breiman L (1996) Bagging predictors. *Mach Learn* 24:123–140
- Bro R, Smilde AK (2014) Principal component analysis. *Anal Methods* 6: 2812–2831
- Ermini L, Catani F, Casagli N (2005) Artificial neural networks applied to landslide susceptibility assessment. *Geomorphology* 66:327–343
- Feizizadeh B, Roodposhti MS, Blaschke T, Aryal J (2017) Comparing GIS-based support vector machine kernel functions for landslide susceptibility mapping. *Arab J Geosci* 10:122
- Freund Y, Schapire RE A (1995). Decision-theoretic generalization of on-line learning and an application to boosting. In: European conference on computational learning theory, Springer, pp 23–37
- Frye C (2007) About the Geometrical Interval classification method <http://blogs.esri.com/esri/arcgis>
- Hall MA (2000) Correlation-based feature selection of discrete and numeric class machine learning
- Hsieh N-C, Hung L-P (2010) A data driven ensemble classifier for credit scoring analysis. *Expert Syst Appl* 37:534–545
- Kavzoglu T, Sahin EK, Colkesen I (2014) Landslide susceptibility mapping using GIS-based multi-criteria decision analysis, support vector machines, and logistic regression. *Landslides* 11:425–439
- Kira K, Rendell LA (1992). A practical approach to feature selection. In: Proceedings of the ninth international workshop on Machine learning, pp 249–256
- Kohavi R (1996). Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid. In: KDD, pp 202–207
- Kuncheva LI (2004) Combining pattern classifiers: methods and algorithms. Wiley, Hoboken
- Lee S, Sambath T (2006) Landslide susceptibility mapping in the Damrei Romel area, Cambodia using frequency ratio and logistic regression models. *Environmental Geology* 50:847–855
- Lee S, Hwang J, Park I (2013) Application of data-driven evidential belief functions to landslide susceptibility mapping in Jinbu, Korea. *Catena* 100:15–30
- Li C (2007). Classifying imbalanced data using a bagging ensemble variation (BEV). In: Proceedings of the 45th annual southeast regional conference, ACM, pp 203–208
- Michalski RS, Carbonell JG, Mitchell TM (2013) Machine learning: an artificial intelligence approach. Springer Science & Business Media, Berlin
- Murphy KP (2006) Naive bayes classifiers University of British Columbia
- NCEP (2014) Global weather data for SWAT <http://globalweathertamu.edu/home>
- Pham BT, Tien Bui D, Pourghasemi HR, Indra P, Dholakia MB (2015) 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. <https://doi.org/10.1007/s00704-015-1702-9>
- Pham BT, Pradhan B, Tien Bui D, Prakash I, Dholakia MB (2016a) A comparative study of different machine learning methods for landslide susceptibility assessment: a case study of Uttarakhand area (India). *Environ Model Softw* 84:240–250. <https://doi.org/10.1016/j.envsoft.2016.07.005>
- Pham BT, Tien Bui D, Prakash I, Dholakia MB (2016b) Rotation forest fuzzy rule-based classifier ensemble for spatial prediction of landslides using GIS. *Nat Hazards* 83:1–31. <https://doi.org/10.1007/s11069-016-2304-2>
- Pham BT, Bui DT, Prakash I (2017a) Landslide Susceptibility Assessment Using Bagging Ensemble Based Alternating Decision Trees, Logistic Regression and J48 Decision Trees Methods: A Comparative Study Geotechnical and Geological Engineering:1–15
- Pham BT, Khosravi K, Prakash I (2017b) Application and comparison of decision tree-based machine learning methods in landslide susceptibility assessment at Pauri Garhwal area. Uttarakhand, India *Environmental Processes*, pp 1–20
- Pham BT, Tien Bui D, Prakash I, Dholakia MB (2017c) Hybrid integration of multilayer Perceptron neural networks and machine learning ensembles for landslide susceptibility assessment at Himalayan area (India) using GIS. *Catena* 149(Part 1):52–63. <https://doi.org/10.1016/j.catena.2016.09.007>
- Pradhan B, Sezer EA, Gokceoglu C, Buchroithner MF (2010) Landslide susceptibility mapping by neuro-fuzzy approach in a landslide-prone area (Cameron highlands, Malaysia). *IEEE Trans Geosci Remote Sens* 48:4164–4177
- Quinlan JR (1986) Induction of decision trees. *Mach Learn* 1:81–106
- Regmi NR, Giardino JR, Vitek JD (2010) Modeling susceptibility to landslides using the weight of evidence approach: western Colorado, USA. *Geomorphology* 115:172–187
- Rodriguez JJ, Kuncheva LI, Alonso CJ (2006) Rotation forest: a new classifier ensemble method. *IEEE Trans Pattern Anal Mach Intell* 28:1619–1630
- Seni G, Elder JF (2010) Ensemble methods in data mining: improving accuracy through combining predictions. *Synthesis Lectures on Data Mining and Knowledge Discovery* 2:1–126
- Sezer EA, Pradhan B, Gokceoglu C (2011) Manifestation of an adaptive neuro-fuzzy model on landslide susceptibility mapping: Klang valley, Malaysia. *Expert Systems with Applications* 38:8208–8219
- Shirzadi A et al (2017) Shallow landslide susceptibility assessment using a novel hybrid intelligence approach. *Environmental Earth Sciences* 76:60
- Tallarida RJ, Murray RB (1987) Chi-square test. In: *Manual of Pharmacologic Calculations*. Springer, pp 140–142

- 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. *Environmental Earth Sciences* 75:1–22. <https://doi.org/10.1007/s12665-016-5919-4>
- Tien Bui D, Tuan TA, Klempe H, Pradhan B, Revhaug I (2016b) 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
- Tsangaratos P, Ilia I (2016) Comparison of a logistic regression and naive Bayes classifier in landslide susceptibility assessments: the influence of models complexity and training dataset size. *Catena* 145:164–179
- Vapnik VN (1995) *The nature of statistical learning theory*. Springer-Verlag, New York
- Wang Y, Makedon F (2004) Application of Relief-F feature filtering algorithm to selecting informative genes for cancer classification using microarray data. In: *Computational Systems Bioinformatics Conference, CSB 2004. Proceedings. 2004 IEEE*, 2004. IEEE, pp 497–498
- Wang H, Khoshgoftaar TM, Napolitano A (2012) Software measurement data reduction using ensemble techniques. *Neurocomputing* 92: 124–132
- Webb GI (2000) Multiboosting: a technique for combining boosting and wagging. *Mach Learn* 40:159–196
- West D, Mangiameli P, Rampal R, West V (2005) Ensemble strategies for a medical diagnostic decision support system: a breast cancer diagnosis application. *Eur J Oper Res* 162:532–551
- Wold S, Esbensen K, Geladi P (1987) Principal component analysis. *Chemom Intell Lab Syst* 2:37–52