A Comparative Assessment Between the Application of Fuzzy Unordered Rules Induction Algorithm and J48 Decision Tree Models in Spatial Prediction of Shallow Landslides at Lang Son City, Vietnam

Dieu Tien Bui, Biswajeet Pradhan, Inge Revhaug and Chuyen Trung Tran

Abstract The main objective of this study is to investigate potential application of the Fuzzy Unordered Rules Induction Algorithm (FURIA) and the Bagging (an ensemble technique) in comparison with Decision Tree model for spatial prediction of shallow landslides in the Lang Son city area (Vietnam). First, a landslide inventory map was constructed from various sources. Then, the landslide inventory was randomly partitioned into 70 % for training the models and 30 % for the model validation. Second, six landslide conditioning factors (slope, aspect, lithology, land use, soil type, and distance to faults) were prepared. Using these factors and the training dataset, landslide susceptibility indexes were calculated using the FURIA, the FURIA with Bagging, the Decision Tree, and the Decision Tree with Bagging. Finally, prediction performances of these susceptibility maps were carried out using the Receiver Operating Characteristic (ROC) technique. The results show that area under the ROC curve (AUC) using training dataset has the largest for the Decision Tree with Bagging (0.925) and the FURIA with Bagging (0.913), followed by the Decision Tree (0.908) and the FURIA (0.878). The prediction capability of these

D. Tien Bui (✉) · I. Revhaug
Department of Mathematical Sciences and Technology, Norwegian University of Life Sciences, P.O. Box 5003IMT, N-1432 Aas, Norway
e-mail: buitiendieu@gmail.com

D. Tien Bui
Faculty of Surveying and Mapping, Hanoi University of Mining and Geology, Dong Ngac, Tu Liem, Hanoi, Vietnam

C. Trung Tran
Faculty of Information Technology, Hanoi University of Mining and Geology, Dong Ngac, Tu Liem, Hanoi, Vietnam

B. Pradhan
Faculty of Engineering, Department of Civil Engineering, University Putra Malaysia, 43400 Serdang, Selangor Darul Ehsan, Malaysia

P. K. Srivastava et al. (eds.), *Remote Sensing Applications in Environmental Research*, Society of Earth Scientists Series, DOI: 10.1007/978-3-319-05906-8_6,

models was estimated using the validation dataset. The highest prediction was achieved using the FURIA with Bagging (AUC = 0.802), followed by the Decision Tree (AUC = 0.783), the Decision Tree with Bagging (AUC = 0.777), and the FURIA (AUC = 0.773). We conclude that the FURIA with Bagging is the best model in this study.

Keywords GIS · Landslide susceptibility · Remote sensing · FURIA · Decision tree

1 Introduction

Landslides are one of many types of natural processes and when threaten mankind they will represent as hazard (Glade et al. 2005). Globally, landslides cause thousands of deaths and injuries, and the direct and indirect costs of landslide damages go up to many billions of USD annually (Roberds 2005). Climate changes and its anticipated consequences are expected to lead to an increase in natural hazards including landslides, resulting in loss of lives and infrastructure damages (Korup et al. 2012).

Landslide damages can be reduced if we understand the mechanisms of occurrence, prediction, hazard assessment, early warning, and risk management (Sassa and Canuti 2008). Landslide hazard assessment can help authorities to reduce landslide damages through proper land use management for infrastructural development and for environmental protection (Tien Bui et al. 2013a). The spatial prediction of landslides is considered as one of the most difficult aspects in the assessment of landslide hazard. For this reason, various methods and techniques have been proposed and they range from simple qualitative techniques to sophisticated mathematical models (Chung and Fabbri 2008). Good overview of these methods including their disadvantages and advantages can be seen in Guzzetti et al. (1999) and Chacon et al. (2006).

In recent years, with the development of geographical information systems (GIS) and computer sciences, some new methods such as neural networks, fuzzy logic, and neuro-fuzzy have become new solutions for landslide modelling with good prediction capabilities (Pradhan et al. 2010; Sezer et al. 2011; Pourghasemi et al. 2012; Tien Bui et al. 2012c; Akgun et al. 2012; Althuwaynee et al. 2014). Although a series of methods and techniques have been proposed and implemented, no agreement has been reached so far on which method is the best one for landslide susceptibility mapping. It is clear that the quality of landslide susceptibility models is influenced both by the methods used and the sampling strategies employed. In more recent years, data mining and ensemble-based approaches have received much attention in many fields including landslide studies (Tien Bui et al. 2013b). They are reported having an improvement of the prediction performance of models (Rokach 2010; Tien Bui et al. 2013c).

The main objective of this study is to investigate potential application of the Fuzzy Unordered Rules Induction Algorithm (FURIA) with Bagging (an ensemble

technique) in comparison with the Decision Tree model, for spatial prediction of shallow landslides at Lang Son city area (Vietnam). FURIA is a fuzzy rule based classification system that combines advantages of RIPPER (Cohen 1995) and fuzzy logic. FURIA and its ensemble have not been used in landslide modelling. The computation process was carried out using WEKA ver.3.6.6, MATLAB 7.11, and ArcGIS 10.

2 Study Area and Spatial Database

2.1 Study Area Characteristics

The study area that includes the Lang Son city and the Dong Dang town (Fig. 1) is located in the northeast mountainous province of Lang Son (Vietnam). It covers an area of about 168 km^2 and lies between longitudes 106°41′34″E and 106°48′32″E, and latitudes 21°49′43″N and 21°57′13″N. Slopes in the study area are from 0° to 84°, around 66 % of the study area has slopes steeper than 15°. The elevation ranges from 194 to 800 m a.s.l with a mean of 328 m.

The study area is comprised of approximately 45.2 % forest land, 21.5 % paddy land, 20.4 % barren land, and 5.7 % crop land, whereas settlement areas cover about 6.9 %. The soil types are mainly ferralic acrisols, dystric gleysols, rhodic ferralsols, and eutric fluvisols that account for 95.2 % of the total study area. Eleven lithologic formations are recognized in the region and six of them account for 80 % of the study area. They are Na Khuat, Tam Lung, Khon Lang, Lang Son, Tam Danh, and Mau Son formations. The main lithologies are marl, siltstone, tuffaceous conglomerate, griltstone, sandstone, basalt, and clay shale. Approximately 16 % of the study area is covered by Quaternary deposits that mainly contain granule, grit, breccia, boulder, sand, and clay.

Landslides in the study area mainly occurred during extreme rainfall events and tropical rainstorms. With the rapid development of economics in the province for the last two decades, the expansions of the infrastructures and the settlements which are shifted into the mountainous regions, have increased slope disturbance. In addition, the deforestation is still continuing leading to potential increase of landslides.

2.2 Spatial Database

2.2.1 Landslide Inventory

In the study area, the landslides were mainly rainfall-triggered shallow soil slides and debris flows. Rock fall was reported in some very few cases and is not included in this study. No information on earthquake-induced landslides has been reported so far. The landslide inventory map for this study was constructed from several sources:

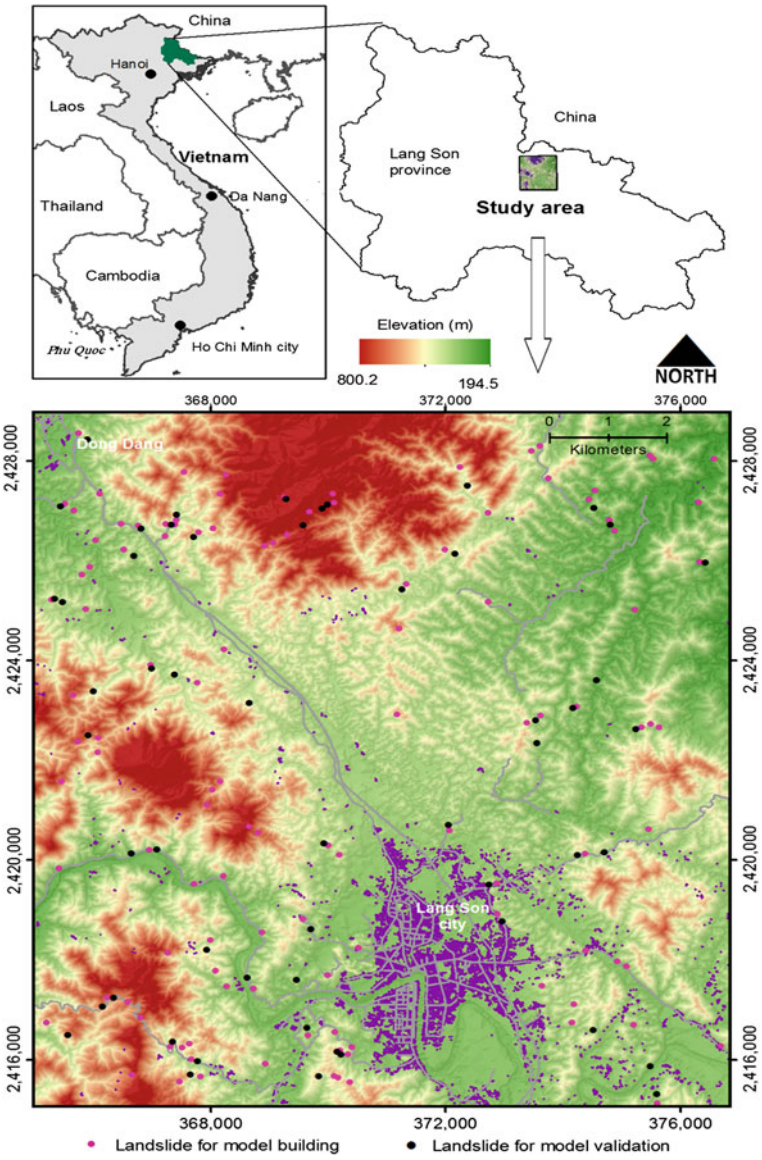

Fig. 1 Study area location map showing the landslide inventory

(1) Landslides that occurred before the year of 2003 detected by the interpretation of aerial photographs and field survey data. The aerial photographs have a resolution of about 1 m. The aerial photographs were acquired by the Aerial Photo—Topography Company 2003; (2) the landslide inventory map of 2006 (Tam et al. 2006); (3) the landslide inventory map of 2009 (Truong et al. 2009); (4) Some recent landslides

were identified during field works. A total of 172 landslides depicted by polygons (Fig. 1) were identified and registered in the inventory map, including 86 rotational slides, 52 translational slides, and 34 debris flows.

2.2.2 Digital Elevation Model and Derivatives

In this study area, the digital elevation model (DEM) was generated from National Topographic Maps at scales 1:5,000 for the Lang Son city and 1:10,000 for the surrounding areas. The DEM has 5 m resolution. Slope and aspect were extracted from the DEM. In the case of the slope map, six categories were constructed (Fig. 2a), whereas nine layer classes were constructed for the aspect map (Fig. 2b).

2.3 Lithology and Distance to Faults

The lithological map was constructed with seven groups: conglomerate, basalt, quaternary, siltstone, limestone, sandstone, and tuff (Fig. 2c). The distance-to-faults map (Fig. 3b) was constructed by buffering the fault lines. Five fault buffer categories were constructed: 0–100, 100–200, 200–300, 300–400, and >400 m. The lithology and fault lines were extracted from four tiles of the Geological and Mineral Resources Map of Vietnam at 1:50,000 scale (Quoc et al. 1992; Truong et al. 2009).

2.4 Land Use and Soil Type

Land use was extracted from the land use status map from 2010 of the Lang Son province. The scale of the land use status map is 1:50,000 and this map is a result of the Status Land Use Project of the National Land Use Survey in Vietnam. A total of nine classes were constructed for the land use map (Fig. 2d). Regarding the soil type map, a total of eight layers were constructed for analysis (Fig. 3a). The soil types were extracted from the national pedology map at scale 1:100,000.

3 Methodology

3.1 Training and Validation Dataset

The landslide inventory and six conditioning factor maps (slope, aspect, lithology, landuse, soil type, and distance to faults) were converted to a grid cell format with spatial resolution of 5 m. Assuming $N(LS)$ is the total number of grid cells in the

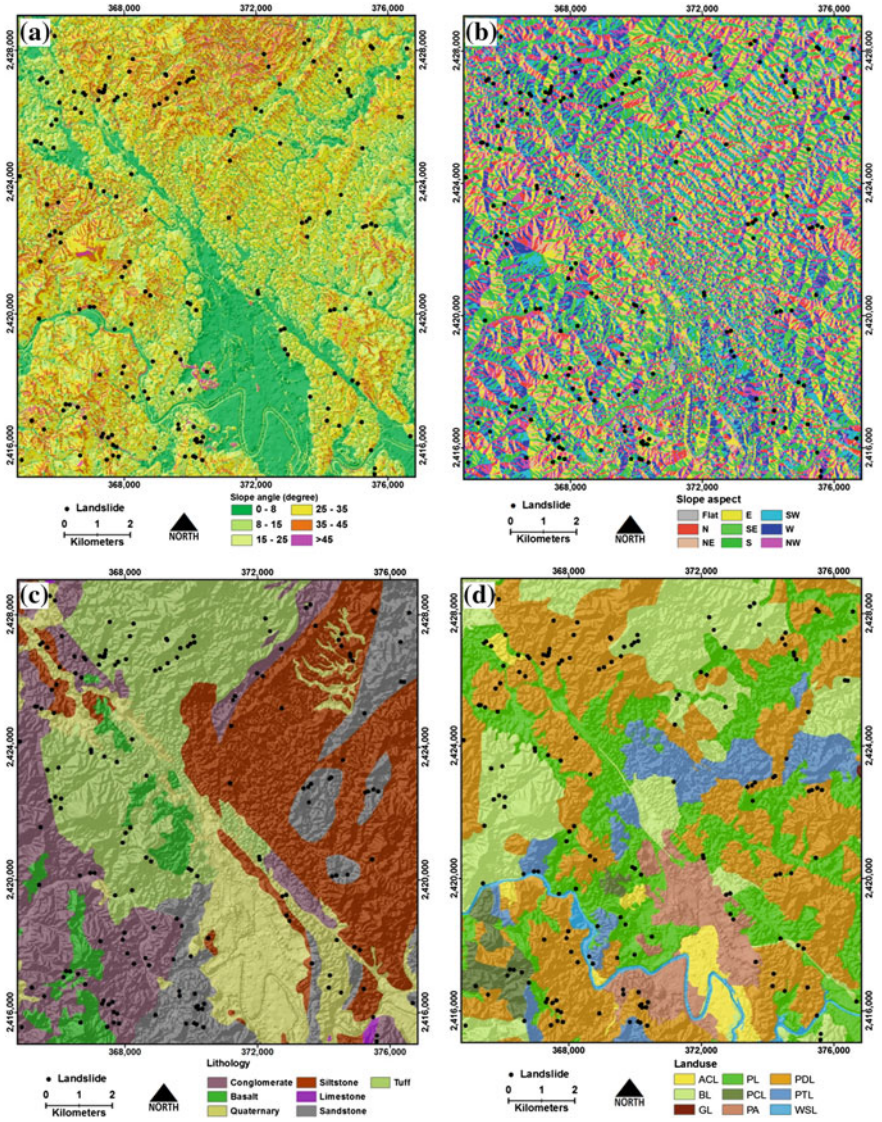

Fig. 2 **a** Slope; **b** Aspect; **c** Lithology; **d** Land use (*ACL* annual crop land; *BL* barren land; *GL* grass land; *Pl* paddy land; *PCL* perennial crop land; *PA* populated area; *PDL* productive forest land; *PTL* protective forest land; *WSL* water surface land)

study area and the training dataset D has $N(D)$ total number of landslide grid cells. We define F_{ij} as the j-th layer class of the landslide conditioning factors F_i and $N(F_{ij})$ is the total number of grid cells in the class F_{ij}. By overlaying the landslide grid cells in the training dataset on each of the six landslide conditioning maps, the

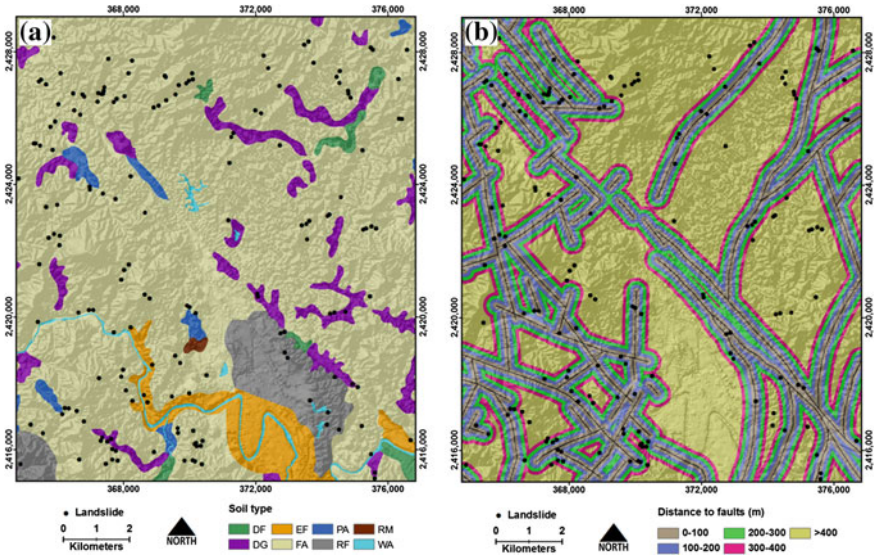

Fig. 3 **a** Soil type; **b** Distance to faults; (*DF* dystric fluvisols; *DG* dystric gleysols; *EF* eutric fluvisols; *FA* ferralic acrisols; *PA* plinthic acrisols; *RF* rhodic ferralsols; *RM* rocky mountain; *WA* water area)

number of grid cells in F_{ij} overlapping with the landslide grid cells $\mathrm{N}(\mathrm{T} \cap \mathrm{F_{ij}})$ was determined. Then, each category of the six maps was assigned to an attribute value that was calculated using the following equation

$$Attribute(F_{ij}) = \frac{W_{F_{ij}}}{\sum_{j=1}^{n} W_{F_{ij}}} \tag{1}$$

where

$$W_{F_{ij}} = \frac{N(D \cap F_{ij})/N(D)}{[N(F_{ij}) - N(D \cap F_{ij})]/[N(LS) - N(D)]} \tag{2}$$

The numerator in Eq. (2) is the proportion of landslide pixels that occur in the factor class, whereas the denominator is the proportion of non-landslide pixels in the factor class.

In landslide susceptibility modeling, a landslide inventory is suggested to be partitioned into two subsets (Chung and Fabbri 2003), one subset will be used for building the landslide models whereas the other will be used for model validation. In general, the partition of landslide inventories using temporal distribution is considered to be the best method (Chung and Fabbri 2008). However, the dates for the past landslide are unknown; therefore we randomly split the landslide inventory map in a 70/30 ratio for building and validation of the model, respectively.

Resulting in a training dataset that contains 117 landslide locations (3,793 landslide grid cells), that was used for building models, and a validation dataset with 55 landslide locations (1,664 landslide grid cells). Landslide pixels were assigned a value of 1.

The same number of grid cells was randomly sampled from the no landslide areas and were assigned a value of 0. A total of 3,793 no-landslide grid cells were generated for the training data and 1,664 no-landslide grid cells for the validation data. At the final step, the values of the six landslide conditioning factors were extracted to build the training and validation datasets. The training and validation datasets contain 7,586 and 3,328 observations, one dependent variable, and six independent variables (the six landslide conditioning factors) (Table 1).

3.2 Fuzzy Unordered Rules Induction Algorithm

FURIA is a fuzzy rule based classification system proposed by Huhn and Hullermeier (2009). This algorithm is an extension of a state-of-the-art rule learner called RIPPER (Cohen 1995) in which fuzzy and unordered rules are to be used instead of conventional rules and rule lists, respectively.

Suppose that we have a training dataset D that have instance-label pairs (x_i, y_i) where i is the i-th training instance, $\boldsymbol{x_i} \in R^n$, and $y_i \in \{1, 0\}$. In the current context, x_i is the vector of input of the six landslide conditioning factors: slope, aspect, lithology, land use, soil type, and distance to faults. The two classes of {1, 0} denote landslide and no-landslide pixels. RIPPER divides the training dataset into two subsets a growing set and a pruning set. The first one will be used for growing the rules whereas the second one is used for pruning. At the first step, rule sets will be generated and learned using the growing set. Each rule to be grown by greedily adding antecedents until the rule is satisfied. All possible combinations of landslide conditioning factors were tested and the final one with the highest value of FOIL's Information Gain (IG) (Quinlan and Cameron-Jones 1993) was selected.

$$IG_r = p_r[\log_2(p_r/(p_r + n_r)) - \log_2(p/(p + n))] \tag{3}$$

where p_r and n_r are the number of positive and negative instances cover by the rule, whereas p and n are the number of positive and negative instances cover by the default rule.

For avoiding over-fitting, the rule pruning process was carried out by simplifying the rules. All of the learned antecedents will be pruned if the antecedents maximizing V_r. Finally, the rule optimization process was carried out.

$$V_r = p_r/(p_r + n_r) \tag{4}$$

FURIA combines advantages of RIPPER and fuzzy logic, and the rule order in the rule list is not important and there is no default rule (Trawinski et al. 2011). Rules

Table 1 Attribute classes of landslide conditioning factors used in the FURIA, FURIA with bagging, decision tree, decision tree with bagging

Data layers	Class	Number of pixels in class	Landslide pixels	W_{Fij}	$\sum W_{Fij}$	Attribute
Slope (degree)	0–8	39,813,500	0	0.0000	46.0394	0.000
	8–15	17,114,025	39	0.5599	46.0394	0.012
	15–25	35,543,725	796	5.5028	46.0394	0.120
	25–35	47,169,725	1,549	8.0691	46.0394	0.175
	35–45	25,972,525	1,204	11.3908	46.0394	0.247
	>45	2,455,275	205	20.5169	46.0394	0.446
Aspect	Flat (−1)	7,881,175	0	0.0000	45.3072	0.000
	North	18,304,150	33	0.4430	45.3072	0.010
	Northeast	21,121,050	159	1.8497	45.3072	0.041
	East	20,045,575	262	3.2115	45.3072	0.071
	Southeast	19,713,650	959	11.9535	45.3072	0.264
	South	20,360,850	1,326	16.0028	45.3072	0.353
	Southwest	22,3707,00	847	9.3034	45.3072	0.205
	West	20,089,900	198	2.4217	45.3072	0.053
	Northwest	18,181,725	9	0.1216	45.3072	0.003
Lithology	Conglomerate	30,238,950	1,086	8.8247	30.9657	0.285
	Basalt	7,775,900	54	1.7063	30.9657	0.055
	Quaternary	26,912,475	204	1.8625	30.9657	0.060
	Siltstone	45,368,200	1,081	5.8547	30.9657	0.189
	Limestone	248,450	0	0.0000	30.9657	0.000
	Sandstone	18,722,500	597	7.8351	30.9657	0.253
	Tuff	38,802,300	771	4.8823	30.9657	0.158
Land use	Annual crop land	4,499,225	64	3.4952	34.3324	0.102
	Populated area	10,566,150	155	3.6045	34.3324	0.105
	Protective forest land	12,473,925	322	6.3429	34.3324	0.185

(continued)

Table 1 (continued)

Data layers	Class	Number of pixels in class	Landslide pixels	$W_{F_{ij}}$	$\sum W_{F_{ij}}$	Attribute
	Productive forest land	61,029,725	1,325	5.3346	34.3324	0.155
	Paddy land	39,295,850	584	3.6517	34.3324	0.106
	Barren land	33,753,675	1,304	9.4928	34.3324	0.276
	Perennial crop land	3,974,925	39	2.4108	34.3324	0.070
	Water surface land	2,498,475	0	0.0000	34.3324	0.000
	Grass land	36,825	0	0.0000	34.3324	0.000
Soil type	Ferralic acrisols	133,741,975	3,285	6.0353	20.2541	0.298
	Dystric gleysols	10,298,575	90	2.1473	20.2541	0.106
	Plinthic acrisols	2,195,825	4	0.4476	20.2541	0.022
	Water area	1,788,350	0	0.0000	20.2541	0.000
	Dystric fluvisols	2,082,500	0	0.0000	20.2541	0.000
	Eutric fluvisols	8,029,575	220	6.7323	20.2541	0.332
	Rhodic ferralsols	9,744,950	194	4.8916	20.2541	0.242
	Rocky mountain	247,025	0	0.0000	20.2541	0.000
Distance to faults (m)	0–100	28,343,050	837	7.2563	32.2505	0.225
	100–200	25,966,100	1,292	12.2264	32.2505	0.379
	200–300	22,463,775	402	4.3972	32.2505	0.136
	300–400	17,874,600	399	5.4849	32.2505	0.170
	>400	73,481,250	863	2.8858	32.2505	0.089

for each label class were induced separately using one-versus-the rest strategy. FURIA transforms the crisp rules of RIPPER into fuzzy rules using the trapezoidal membership function (Huhn and Hullermeier 2009). In this function, each fuzzy interval is specified by four parameters and is written as $I = (T_1, T_2, T_3, T_4)$.

$$I(x) = \begin{cases} 1 & T_2 \leq v \leq T_3 \\ \frac{v - T_1}{T_2 - T_1} & T_1 \leq v \leq T_2 \\ \frac{T_4 - v}{T_4 - T_2} & T_3 \leq v \leq T_4 \\ 0 & \text{else} \end{cases} \tag{5}$$

For an instance $v_i = (x_{i1}, \ldots, x_{i6})$, the fuzzy membership function can be expressed as

$$\mu(v_i) = \prod_{j=1}^{6} I_j(x_j) \tag{6}$$

The fuzzification of a single antecedent of a rule is only relevant to a subset $D_k \in D$, and then D_k is divided into two subsets D_{k+} and D_{k-}. The quality of the fuzzification is checked to choose the best one using the purity rule criteria as mentioned in Eq. (7)

$$pur = \frac{p_i}{p_i + n_i}; \quad p_i = \sum_{v \in D_{k+}} \mu A_i(v); \quad n_i = \sum_{v \in D_{k-}} \mu A_i(v)\,;\, A_i \in I(x) \tag{7}$$

Fuzzy rules were constructed for class y_i and a certainty degree CD_i for the consequence. The final decision for output is based on the largest V value as

$$V = \sum_{i=1}^{m} \mu_{rule(i)}(v) * CD_i \tag{8}$$

Finally, the rule generalization procedure is carried out to obtain the final fuzzy rule list. A detailed explanation can be seen in Huhn and Hullermeier (2009, 2010).

FURIA was training using stratified 10-fold cross-validation. First, the training dataset was randomly partitioned into 10-folds of equal size. Then, in each run, 9-folds were used for fitting the model whereas the remaining fold was used to assess the performance. The procedure is repeated ten times and results are averaged. In this study, the fuzzy aggregation operator of the product T-Norm (used as fuzzy AND) was selected to combine rule antecedents. This is because FURIA product was reported significant better than FURIA-min (the minimum of T-Norm) (Huhn and Hullermeier 2010). Since the selection of number folds for the training data used for pruning is significant affecting the model accuracy, a test was therefore performed with different folds versus classification accuracy. The result shows that 4-folds used for pruning and the rest for growing the fuzzy rules have the highest classification accuracy. Other parameters were set as default in WEKA. Finally, the

FURIA model with 45 rules was constructed for landslide susceptibility in this study. The overall accuracy was 84.84 %. The details for the accuracy by class and performance by the FURIA model are shown in Tables 4 and 5.

3.3 Decision Tree

Decision tree classifiers are hierarchical models composed of a root, internal nodes, leaf nodes, and branches, and have been considered one of the most popular classification methods in data mining. The goal of decision tree modeling is to generate a tree structure that contains a set of rules using the training dataset. The tree structure has the capability to predict the output for a new similar dataset with good accuracy. Once a decision tree model is constructed; it can process new data by following a path from the root node to the leaves and values for the new data will be obtained. Since the output for pixels in landslide susceptibility modeling present continuous values, the decision trees are called regression trees. The key advantage of decision trees is that they are easy to construct. In addition, the results from decision trees are readily interpretable with clear information of the contribution of the variables on the model results. However, decision trees do not allow for multiple outputs and are susceptible to noisy data (Zhao and Zhang 2008).

Various algorithms for constructing decision trees have been successfully developed such as classification and regression tree (CART) (Breiman et al. 1984), Chi-square Automatic Interaction Detector decision tree (CHAID) (Michael and Gordon 1997), ID3 (Quinlan 1986), C4.5 (Quinlan 1993), and J48 (Witten and Frank 2005). However the C.45 algorithm has been considered as the fastest algorithm for machine learning with good classification accuracy (Lim et al. 2000). In this study, the J48 algorithm, which is a Java re-implementation of the C4.5 algorithm, was used. The detailed description of the C4.5 algorithm can be seen in Quinlan (1993). Only a short description of decision tree is discussed here. There are two steps in the decision tree construction, the first one is the tree building and the second one is the tree pruning. The first step of the tree building process is to find the input landslide conditioning factor with the highest gain ratio using the training data set, and then select as the first internal node called root node. In the next step, the training dataset was split based on the root values, and sub-notes were created. Then, the gain ratio was estimated for each sub-node. The variable with the highest gain ratio is selected, and the recursive partitioning of the training data set is continued until all instances in the training dataset are assigned to leaf nodes or no remaining variables in which the training data can be further split. In some cases, the resulting tree may be obtained with a large number of branches, and thus the tree may over-fit the training dataset with a perfect classification, but the model has a poor classification performance for a new dataset. Therefore, the tree pruning was carried out by removing unessential nodes but with the classification accuracy still remaining (Breiman et al. 1984).

Table 2 Minimum number of instances per leaf

No.	Minimum number of instances per leaf	Classification accuracy (%)	
		Training dataset	Validation dataset
1	1	88.43	69.35
2	2	88.18	68.66
3	4	87.37	68.78
4	6	86.77	68.90
5	8	85.92	71.06
6	9	85.65	70.97
7	10	85.41	70.64
8	12	85.42	71.00
9	13	85.17	72.09
10	14	85.03	73.29
11	15	84.89	72.71
12	16	84.91	72.45
13	18	84.46	72.29
14	20	84.31	71.51

In this study, the first step in constructing decision tree models is to determine the parameters that influencing the classification accuracy of the resulting tree. The type of pruning is based on sub-tree rising. Laplace smoothing was used here to improve probabilistic estimates at leaves (Tien Bui et al. 2012a, 2013b; Tehrany et al. 2013). A test was carried out to find the most suitable parameters for the study area based on the classification accuracy. The results are shown in Tables 2, 3. The results show that the best values for minimum number of instances per leaf and the confident factor are 14 and 0.15 respectively. The selection number of fold of training data used for reduce-error pruning does not affect the accuracy of the model.

Using the training data set and the above mentioned parameters, decision tree model was trained using with stratified 10-fold cross-validation. The 10-fold cross-validation was preferred to be used in order to ensure that the decision trees generalize beyond the training data (Breiman et al. 1984). Finally, the decision tree model was constructed for landslide susceptibility. The size of the tree is 133. The tree has the root node, 65 internal nodes, and 67 leaves. The classification accuracy is 86.82 %. The more detail of accuracy by class and performance of the decision tree model is shown in Tables 4, 5.

3.4 Bagging

Bagging known as bootstrap aggregation, is one of the earliest ensemble algorithms proposed by Breiman (1996). Bagging is a method that uses bootstrap sampling to generate multiple subsets from the training dataset. Each subset is called a bootstrap sample created by sampling the training dataset of the same size with replacement. In the next step, each of the subset will be used to construct a

Table 3 Confidence factor used for pruning

No.	Confidence factor	Classification accuracy (%)	
		Training dataset	Validation dataset
1	0.05	84.47	72.47
2	0.10	84.79	73.17
3	0.15	85.03	73.29
4	0.20	85.10	73.07
5	0.25	85.04	73.07
6	0.30	85.01	72.86
7	0.35	84.99	72.86
8	0.40	85.05	72.98
9	0.45	85.14	72.98
10	0.50	85.15	72.98
11	0.55	85.18	72.92
12	0.60	85.18	72.92

Table 4 Performance of the FURIA, the FURIA with Bagging, the decision tree, and the decision tree with bagging

No.	Parameters	FURIA	FURIA with bagging	Decision tree	Decision tree with bagging
1	Accuracy (%)	84.84	86.38	86.82	87.50
2	Kappa index	0.697	0.728	0.736	0.750
3	MAE	0.150	0.148	0.211	0.209
4	RMSE	0.361	0.331	0.322	0.308

Table 5 Accuracy assessments by classes of the FURIA, the FURIA with bagging, the decision tree, and the decision tree with bagging

Model	True positive rate (%)	False positive rate (%)	F-measure (%)	Class
FURIA	0.920	0.223	0.859	Landslide
	0.777	0.080	0.837	No-landslide
FURIA with bagging	0.945	0.217	0.874	Landslide
	0.783	0.055	0.852	No-landslide
Decision tree	0.921	0.184	0.875	Landslide
	0.816	0.079	0.861	No-landslide
Decision tree with bagging	0.918	0.168	0.880	Landslide
	0.832	0.082	0.869	No-landslide

classifier based model. Then, the final model is determined by aggregating all the based classifiers (Fig. 4).

Using the training data set, the FURIA with Bagging and the Decision tree with Bagging models were trained. The parameters setting for the above two models are

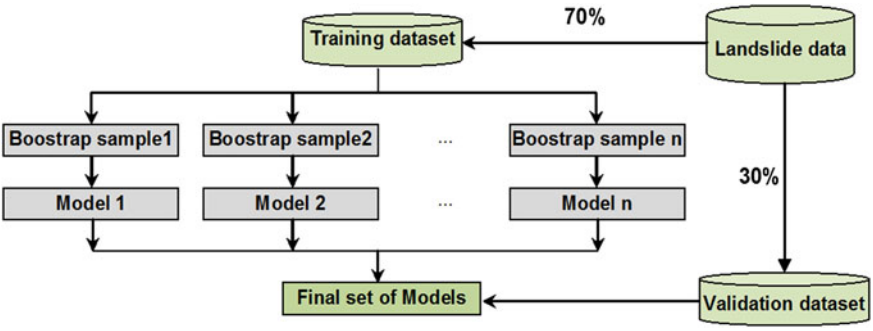

Fig. 4 General framework of the bagged FURIA and the bagged decision tree in this study

remaining the same as in Sects. 3.2 and 3.3. The models were trained and the final results were obtained. The classification accuracy is 86.38 % and 87.50 % for the FURIA with Bagging and for the Decision Tree with Bagging, respectively. The results from the trained models are shown in Tables 4 and 5.

3.5 Generation of Landslide Susceptibility Maps

The successfully trained models were then applied to calculate landslide susceptibility indexes for all the pixels in the study area. The obtained results were converted into a GIS format and loaded in ArcGIS.10. The landslide susceptibility maps were visualized by mean of four susceptibility classes based on the percentage of area (Pradhan and Lee 2010a, b): high (10 %), moderate (10 %), low (20 %), very low (60 %). For the purpose of visualization, only two landslide susceptibility maps that were produced from the FURIA with Bagging and Decision Tree with Bagging models are shown (Figs. 5, 6).

The landslide densities (Kanungo et al. 2008) analysis was carried out for the landslide susceptibility maps by overlaying the four susceptibility zones with the landslide inventory map. Ideally, the density value should increase from very low to high susceptibility zones. The graph of the density analysis for the two models in this study is shown in Fig. 7. The result shows that that there is a gradual increase in landslide density from the very low susceptible zone to the high susceptible zone.

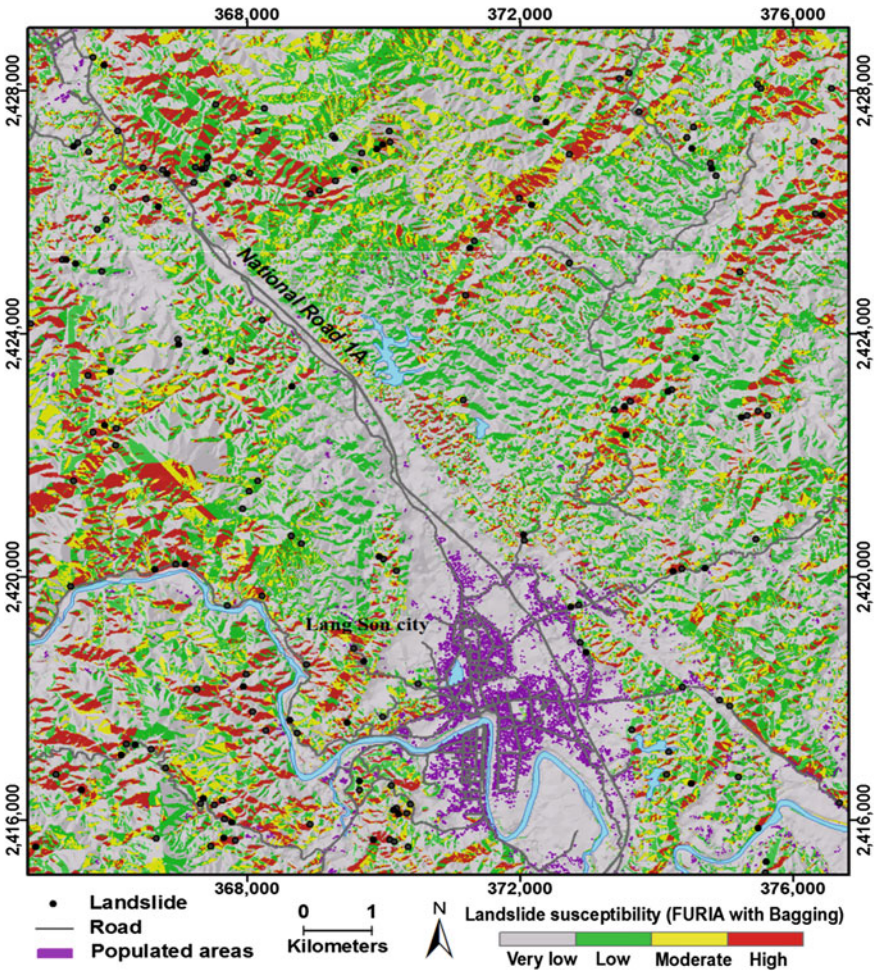

Fig. 5 Landslide susceptibility map of the Lang Son city areas using the FURIA with bagging model

4 Validation and Comparison of Landslide Susceptibility Models

4.1 Model Performance and Evaluation

The performance measurement of four landslide susceptibility models (FURIA, FURIA with Bagging, Decision tree, Decision tree with Bagging) were assessed using several statistical evaluation criteria (Tien Bui et al. 2012a) as follows:

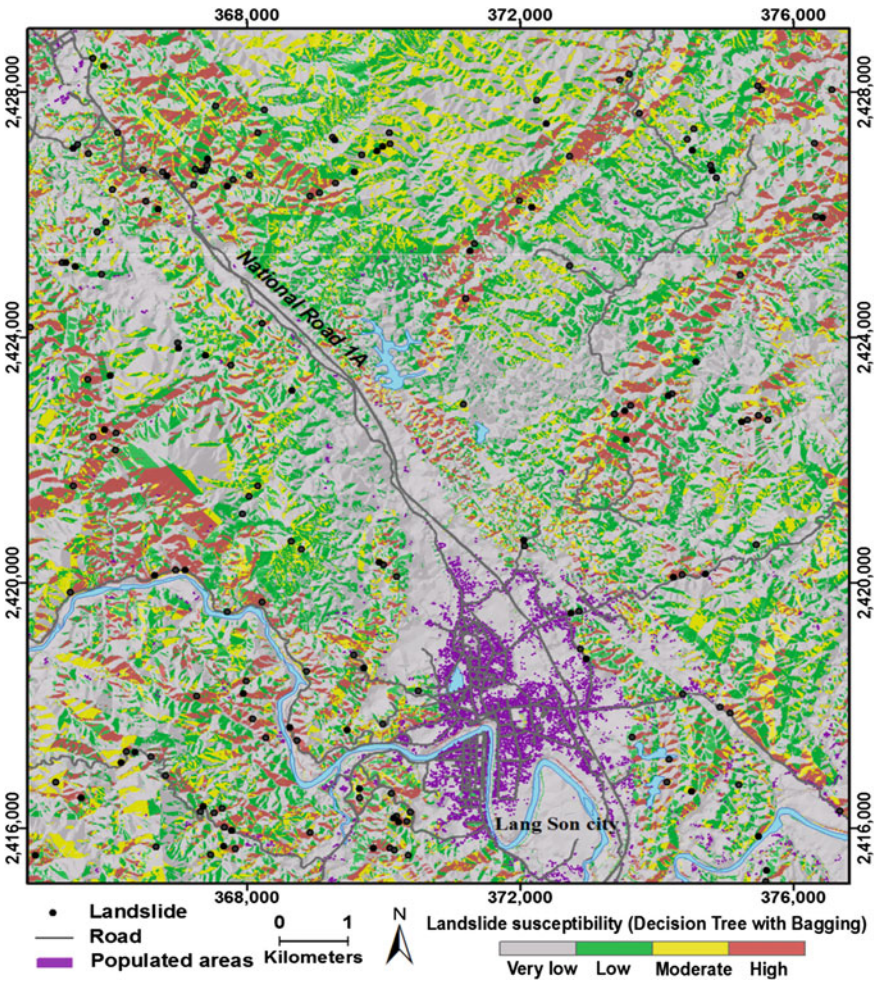

Fig. 6 Landslide susceptibility map of the Lang Son city area using the decision tree with bagging model

$$Sensitivity = TP/(TP + FN) \tag{9}$$

$$Specificity = TN/(TN + FP) \tag{10}$$

$$Accuracy = (TP + TN)/(TP + TN + FN + FP) \tag{11}$$

$$F - measure = 2 * Sensitivity * Specificity/(Sensitivity + Specificity) \tag{12}$$

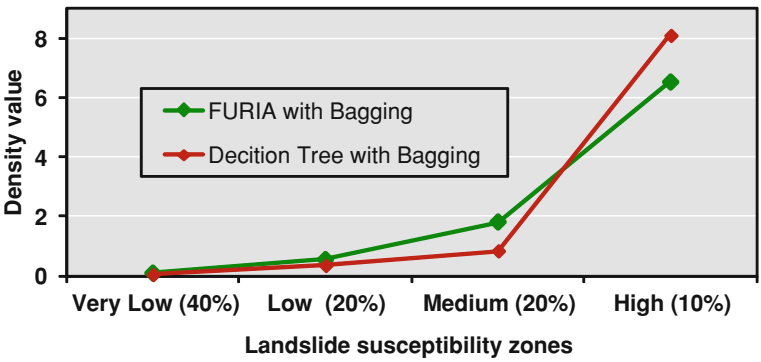

Fig. 7 Density plots of four landslide susceptibility classes of the FURIA with bagging and the decision tree with bagging models

$$Root\ mean\ squared\ error\ (RMSE) = \mathrm{Sqrt}\left(\frac{(pred_1 - act_1)^2 + \cdots + (pred_n - act_n)^2}{n}\right) \tag{13}$$

$$Mean\ absolute\ error\ (MAE) = \frac{|pred_1 - act_1| + \cdots + |pred_n - act_n|}{n} \tag{14}$$

$$\begin{aligned} Kappa\ index\ (\kappa) &= \frac{P_C - P_{exp}}{1 - P_{exp}} \\ \text{where } P_C &= (TP + TN)/(TP + TN + FN + FP) \\ P_{exp} &= [(TP + FN)(TP + FP) + (FP + TN)(FN + TN)]/Sqrt(TP + TN + FN + FP) \end{aligned} \tag{15}$$

True positive (*TP*) rate measures the proportion of number of pixels that are correctly classified as landslides. True negative (*TN*) rate measures the proportion of number of pixels that are correctly classified as non-landslide. False negatives (*FN*) are the number of landslide pixels classified as non-landslide pixel. True negatives (*FN*) are the number of non-landslide pixels classified as landslide pixels. Precision measures the proportion of the number of pixels that are correctly classified as landslide occurrences. *F-measure* combines precision and sensitivity into their harmonic mean. *Act* is the actual target value whereas *pred* is the predicted value. P_C is the proportion of number of pixels that are correctly classified as landslide or non-landslide. P_{exp} is the expected agreements.

It could be observed that there is a high and almost equal in term of classification accuracy for the three models, FURIA with Bagging, the Decision Tree, and the Decision Tree with Bagging (Table 4). Accuracy assessment by classes (Table 5) shows that the rate of correctly classified landslide pixels is higher than those for non-landslide pixels for all models.

Table 6 The range of the kappa index and the corresponding agreement between the model and reality (Cohen 1960)

No.	Kappa index	Agreement
1	0.80–1	Almost perfect
2	0.60–0.80	Substantial
3	0.40–0.60	Moderate
4	0.20–0.40	Fair
5	0–0.2	Slight
6	≤0	Poor

The reliability of the susceptibility models was measured using the Kappa index (Guzzetti et al. 2006). Kappa indexes for the FURIA, FURIA with Bagging, the Decision Tree, and the Decision Tree with Bagging are 0.697, 0.728, 0.736, and 0.750 respectively. It indicates a substantial agreement (Table 6) between the observed and the predicted values. The reliability analysis results are satisfying compared with other works such as Saito et al. (2009) and Tien Bui et al. (2012a).

4.2 Model Validation

The prediction capability of the susceptibility models were evaluated using ROC curves. A ROC curve is used to plot sensitivity/1-specificity with different thresholds. Compared to the to success and prediction rate curves (Chung and Fabbri 2003), ROC curves are considered not sensitive, by keeping in mind of the considerable difference between landslide and non-landslide pixels. Therefore ROC curves are considered as more appropriate evaluation and validation tool for landslide models (Van Den Eeckhaut et al. 2009).

The area under the ROC curve (AUC) is used as an important measurement of the landslide model performance. A landslide model will be considered a preferred model if it has a larger AUC value than other models. A perfect model will have an AUC of 1 whereas a random model has an AUC of approximately 0.5.

In this study, ROC curves and AUCs were prepared for each landslide model in two cases: the first one used the training dataset and the second one used the validation dataset. Since in the first case the same landslide pixels that have already been used to construct the landslide models, therefore, the ROC curve and AUC is only measured the degree of model fit of the model with the training dataset. The result (Fig. 8 and Table 7) shows that the highest degree of fit has the Decision Tree with Bagging (AUC = 0.925), followed by the FURIA with Bagging (AUC = 0.913), the Decision Tree (AUC = 0.908), and FURIA (AUC = 0.878). The prediction capabilities of the landslide models were obtained in the second case. This case uses the validation dataset that has not been used in the training phase and can provide the validation and explain how well the model and the conditioning factors predict the existing landslides (Pradhan and Lee 2010c). The result (Fig. 9 and Table 8) shows that the FURIA with Bagging has the highest prediction capability (AUC = 0.802). The remaining models have almost equal prediction capability (AUC from 0.773 to 0.783).

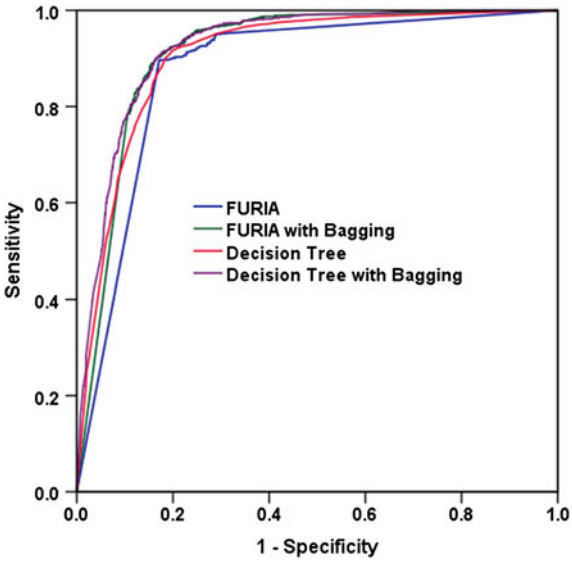

Fig. 8 ROC curves based on the training dataset for the FURIA, the FURIA with bagging, the decision tree, and the decision tree with bagging

Table 7 Area under the curves (*AUC*) based on the training dataset for the FURIA, the FURIA with bagging, the decision tree, and the decision tree with bagging

No.	Landslide model	AUC	Std. error	95 % CI	
				Lower bound	Upper bound
1	FURIA	0.878	0.004	0.869	0.886
2	FURIA with bagging	0.913	0.004	0.906	0.920
3	Decision tree	0.908	0.004	0.901	0.915
4	Decision tree with bagging	0.925	0.003	0.919	0.931

4.3 Relative Contribution of the Conditioning Factors

The relative contribution of each conditioning factor on the susceptibility models can be estimated by excluding the factor in the models and then the classification accuracy was estimated. It is clear that the highest accuracy was obtained when all of the six factors are used (Table 9). Aspect has the highest contribution to the models whereas soil type has the lowest contribution. More details are shown in Table 9.

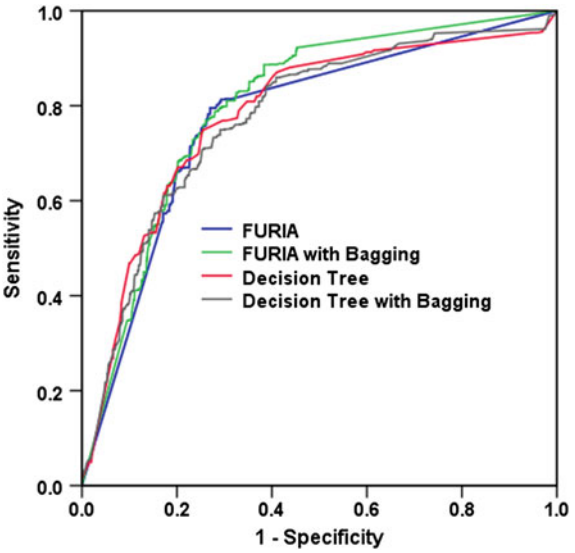

Fig. 9 ROC curves based on the validation dataset for the FURIA, the FURIA with bagging, the decision tree, and the decision tree with bagging

Table 8 Area under the curves (*AUC*) based on the validation dataset for the FURIA, the FURIA with bagging, the decision tree, and the decision tree with bagging

No.	Landslide model	AUC	Std. error	95 % CI	
				Lower bound	Upper bound
1	FURIA	0.773	0.008	0.757	0.790
2	FURIA with bagging	0.802	0.008	0.786	0.817
3	Decision tree	0.783	0.008	0.767	0.799
4	Decision tree with bagging	0.777	0.008	0.761	0.793

Table 9 Relative contribution of the conditioning factors

No	Conditioning factors	Classification accuracy (%)	
		FURIA with bagging	Decision tree with bagging
1	Minus slope	82.61	83.54
2	Minus aspect	79.08	79.54
3	Minus lithology	83.39	83.60
4	Minus landuse	82.64	83.68
5	Minus soil type	85.61	85.75
6	Minus distance to faults	83.07	83.10
7	All	86.38	87.50

5 Conclusion

Over the last two decades, various methods and techniques for the landslide modeling have been used and discussed, however, the FURIA model and Bagging technique are seldom been applied and a comparison between FURIA with

Decision Tree and their Bagging has not been carried out so far. Decision Tree models have only been applied in a limited number of studies. The recent development in geographic information systems (GIS) and computer science allows users to apply these techniques with huge GIS data (Pradhan 2013).

In general, there are three main steps used for the landslide susceptibility modeling in this study, data preparation, susceptibility analyses, and validation and comparison. In the first step, the landslide inventory map with 172 landslide polygons was constructed. Among them, approximately 70 % (117 cases) was selected for the training models, whereas the remaining 30 % (55 cases) were used for model validation. And then, landslide conditioning factors were determined. All maps were prepared with a spatial resolution of 5 m. In the next step, a total of four models were constructed. The validation result show that the FURIA with Bagging (AUC = 0.913) and the Decision Tree with Bagging (0.925) have the highest degree of fit with the training data. They are followed by the Decision Tree (AUC = 0.908), and FURIA (AUC = 0.878). Regarding the prediction capability, the FURIA with Bagging has the highest value (AUC = 0.802), the other models have almost equal prediction capability (AUC is around 0.77).

It is well known that the selection of sampling strategy influences the prediction capability of landslide models (Yilmaz 2010). As shown in Chung and Fabbri (2008), the temporal partitioning of landslides is considered to be the best method. However, the temporal partitioning method is not suitable for this study due to unknown dates of landslide occurrence. Therefore the randomly split method was used. The main disadvantage of this method is that it may cause an overestimated of prediction capability of future landslides if spatial separation between training and validation landslides are small (Brenning 2005).

The selection of conditioning factors are an important task for the assessment of landslide susceptibility and may impact on the overall prediction performance for landslide susceptibility models (Pradhan 2013). Although no agreement on universal guidelines has been reached for the selection of conditioning factors (Tien Bui et al. 2012b), the factors related to topography, geology, soil types, hydrology, geomorphology, and land use are considered to be the most commonly used in landslide analyses (Van Westen et al. 2008). Therefore six landslide conditioning factors (slope, aspect, lithology, distance to faults, landuse, and soil type) were selected for this study.

As a final conclusion, all the models exhibit reasonably satisfactory performance. However, we may conclude that the FURIA with Bagging is considered to be the best one from this study. And it is important to note that the performance of these landslide models depends not only on the methods but also on sampling strategy followed, as well as the quality of the data used. Therefore, the quality of the susceptibility maps produced by the four models can be improved if the quality of the data used increases. The analyzed result obtained from this study is valid for shallow landslides. These maps may be useful for natural hazard management policy, planning and decision-making in landslide prone areas.

Acknowledgement This research was supported by the Geomatics Section, Department of Mathematical Sciences and Technology, Norwegian University of Life Sciences, Norway.

References

Akgun A, Sezer EA, Nefeslioglu HA, Gokceoglu C, Pradhan, B (2012) An easy-to-use MATLAB program (MamLand) for the assessment of landslide susceptibility using a Mamdani fuzzy algorithm. Comput Geosci 38:23-34

Althuwaynee OF, Pradhan B, Park HJ, Lee JH (2014) A novel ensemble decision-tree based CHi-squared automatic interaction detection (CHAID) and multivariate logistic regression models in landslide susceptibility mapping. Landslides (Article online first available). http://dx.doi.org/10.1007/s10346-014-0466-0

Breiman L (1996) Bagging predictors. Mach Learn 24:123–140

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Wadsworth, Belmont

Brenning A (2005) Spatial prediction models for landslide hazards: review, comparison and evaluation. Nat Hazards Earth Syst Sci 5:853–862

Chacon J, Irigaray C, Fernandez T, El Hamdouni R (2006) Engineering geology maps: landslides and geographical information systems. Bull Eng Geol Environ 65:341–411

Chung C-J, Fabbri AG (2008) Predicting landslides for risk analysis—spatial models tested by a cross-validation technique. Geomorphology 94:438–452

Chung CJF, Fabbri AG (2003) Validation of spatial prediction models for landslide hazard mapping. Nat Hazards 30:451–472

Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Measur 20:37–46

Cohen WW (1995) Fast effective rule induction. In: Machine learning: proceedings of the twelfth international conference. Morgan Kaufmann, Lake Taho

Glade T, Anderson M, Crozier MJ (2005) Landslide hazard and risk. Wiley, London

Guzzetti F, Carrara A, Cardinali M, Reichenbach P (1999) Landslide hazard evaluation: a review of current techniques and their application in a multi-scale study, Central Italy. Geomorphology 31:181–216

Guzzetti F, Reichenbach P, Ardizzone F, Cardinali M, Galli M (2006) Estimating the quality of landslide susceptibility models. Geomorphology 81:166–184

Huhn J, Hullermeier E (2010) An analysis of the FURIA algorithm for fuzzy rule induction. In: Koronacki J, Raś Z, Wierzchoń S, Kacprzyk J (eds) Advances in machine learning I, vol 262. Springer, Berlin, pp 321–344

Huhn J, Hullermeier E (2009) FURIA: an algorithm for unordered fuzzy rule induction. Data Min Knowl Disc 19:293–319

Kanungo D, Arora M, Gupta R, Sarkar S (2008) Landslide risk assessment using concepts of danger pixels and fuzzy set theory in Darjeeling Himalayas. Landslides 5:407–416

Korup O, Gorum T, Hayakawa Y (2012) Without power? Landslide inventories in the face of climate change. Earth Surf Proc Land 37:92–99

Lim TS, Loh WY, Shih YS (2000) A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. Mach Learn 40:203–228

Michael JA, Gordon SL (1997) Data mining technique: for marketing, sales and customer support. Wiley, New York

Pourghasemi H, Pradhan B, Gokceoglu C (2012) Application of fuzzy logic and analytical hierarchy process (AHP) to landslide susceptibility mapping at Haraz watershed. Iran Nat Hazards 63:965–996

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 (2010a) 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

Pradhan B, Lee S (2010b) Landslide susceptibility assessment and factor effect analysis: backpropagation artificial neural networks and their comparison with frequency ratio and bivariate logistic regression modelling. Environ Model Softw 25:747–759

Pradhan B, Lee S (2010c) Regional landslide susceptibility analysis using back-propagation neural network model at Cameron Highland, Malaysia. Landslides 7:13–30

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 (1993) C4.5 programs for machine learning. Morgan Kaufmann, San Mateo

Quinlan JR (1986) Induction of decision trees. Mach Learn 1:81–106

Quinlan JR, Cameron-Jones RM (1993) FOIL: a midterm report. In: European conference on machine learning. Springer, Berlin

Quoc NK, Dan TH, Hung L, Huyen DT (1992) Geological map. In: Binh Gia group (ed), Vietnam Institute of Geosciences and Mineral Resources, Hanoi

Roberds W (2005) Estimating temporal and spatial variability and vulnerability. In: Hungr O, Fell R, Couture R, Eberhardt E (eds) Landslide risk management. Taylor and Francis, London

Rokach L (2010) Ensemble-based classifiers. Artif Intell Rev 33:1–39

Saito H, Nakayama D, Matsuyama H (2009) Comparison of landslide susceptibility based on a decision-tree model and actual landslide occurrence: The Akaishi Mountains, Japan. Geomorphology 109:108–121

Sassa K, Canuti P (2008) Landslides-disaster risk reduction. Springer, Berlin, p 650

Sezer EA, Pradhan B, Gokceoglu C (2011) Manifestation of an adaptive neuro-fuzzy model on landslide susceptibility mapping: Klang valley, Malaysia. Expert Syst Appl 38:8208–8219

Tam VT, Tuy PK, Nam NX, Tuan LC, Tuan ND, Trung ND et al (2006) Geohazard investigation in some key areas of the northern mountainous area of Vietnam for the planning of socio-economic development. Vietnam Institute of Geosciences and Mineral Resources, Hanoi, p 83

Tehrany MS, Pradhan B, Jebur MN (2013) Spatial prediction of flood susceptible areas using rule based decision tree (DT) and ensemble bivariate and multivariate statistical models. J Hydrol 504:69-79. http://dx.doi.org/10.1016/j.jhydrol.2013.09.034

Tien Bui D, Pradhan B, Lofman O, Revhaug I (2012a) Landslide susceptibility assessment in Vietnam using support vector machines, Decision tree and Naïve Bayes models. Math Prob Eng. doi.10.1155/2012/974638

Tien Bui D, Pradhan B, Lofman O, Revhaug I, Dick O (2013a) Regional prediction of landslide hazard using probability analysis of intense rainfall in the Hoa Binh province, Vietnam. Nat Hazards 2:707–730

Tien Bui D, Ho TC, Revhaug I, Pradhan B, Nguyen DB (2013b) "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, edited by Buchroithner M, Prechtel N, Burghardt D, 303-17. Springer Berlin Heidelberg

Tien Bui D, Pradhan B, Lofman O, Revhaug I, Dick OB (2012b) Landslide susceptibility assessment in the Hoa Binh Province of Vietnam: a comparison of the Levenberg–Marquardt and Bayesian regularized neural networks. Geomorphology 171–172:12–29

Tien Bui D, Pradhan B, Lofman O, Revhaug I, Dick OB (2012c) Landslide susceptibility mapping at Hoa Binh province (Vietnam) using an adaptive neuro-fuzzy inference system and GIS. Comput Geosci 45:199–211

Trawinski K, Cordon O, Quirin A (2011) On designing fuzzy rule-based multiclassification systems by combining furia with bagging and feature selection. Int J Uncertainty Fuzziness Knowl Based Syst 19:589–633

Truong PD, Nghi TH, Phuc PN, Quyet HB, The NV (2009) Geological mapping and mineral resource investigation at 1:50 000 scale for Lang Son area. Northern Geological Mapping Division, Hanoi

Van Den Eeckhaut M, Reichenbach P, Guzzetti F, Rossi M, Poesen J (2009) Combined landslide inventory and susceptibility assessment based on different mapping units: an example from the Flemish Ardennes, Belgium. Nat Hazards Earth Syst Sci 9:507–521

Van Westen CJ, Castellanos E, Kuriakose SL (2008) Spatial data for landslide susceptibility, hazard, and vulnerability assessment: an overview. Eng Geol 102:112–131

Witten IH, Frank E (2005) Data mining: practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, Los Altos

Yilmaz I (2010) The effect of the sampling strategies on the landslide susceptibility mapping by conditional probability and artificial neural networks. Environ Earth Sci 60:505–519

Zhao Y, Zhang Y (2008) Comparison of decision tree methods for finding active objects. Adv Space Res 41:1955–1959