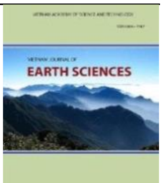
Vietnam Journal of Earth Sciences, 42(3), 237–246, Doi: 10.15625/0866-7187/42/3/15047

(VAST)

Vietnam Academy of Science and Technology

Vietnam Journal of Earth Sciences

http://www.vjs.ac.vn/index.php/jse



Landslide susceptibility mapping using Forest by Penalizing Attributes (FPA) algorithm based machine learning approach

Tran Van Phong[1], Hai-Bang Ly[*2], Phan Trong Trinh[1], Indra Prakash[3], Dao Trung Hoan[4]

[1]Institute of Geological Sciences, Vietnam Academy of Sciences and Technology, Hanoi, Vietnam
[2]University of Transport Technology, Hanoi 100000, Vietnam
[3]Department of Science & Technology, Bhaskarcharya Institute for Space Applications and Geo-Informatics (BISAG), Government of Gujarat, Gandhinagar 382002, India
[4]Center for Information and Archives and Journal of Geology (CIAJG)

Received 09 March 2020; Received in revised form 21 April 2020; Accepted 25 June 2020

ABSTRACT

Landslide susceptibility mapping is a helpful tool for assessment and management of landslides of an area. In this study, we have applied first time Forest by Penalizing Attributes (FPA) algorithm-based Machine Learning (ML) approach for mapping of landslide susceptibility at Muong Lay district (Vietnam). For this aim, 217 historical landslides locations were identified and analyzed for the development of FPA model and generation of susceptibility map. Nine landslide topographical and geo-environmental conditioning factors (curvature, geology/lithology, aspect, distance from faults, rivers and roads, weathering crust, slope, and deep division) were utilized to construct the training and validating datasets for landslide modeling. Different quantitative statistical indices including Area Under the Receiver Operating Characteristic (ROC) curve (AUC) were used to evaluate the performance of the model. The results indicate that the predictive capability of the FPA is very good for landslide susceptibility mapping on both training (AUC = 0.935) and validating (AUC = 0.882) datasets. Thus, the novel FPA based ML model can be utilized for the development of accurate landslide susceptibility map of the study area and this approach can also be applied in other landslide prone areas.

Keywords: Landslide susceptibility mapping; machine learning; AUC; ROC; GIS; Vietnam.

1. Introduction

Landslides are one of the most disastrous geo-hazards affecting life and property of inhabitants in hilly areas (Zhong et al., 2020). Nowadays, landslides events are occurring more frequently due to change in the land use pattern with increasing population and climate change effect (Shirzadi et al., 2012; Shirzadi et al., 2017). It requires more attention of governments and hazard managers to find a better way for controlling and preventing this natural phenomenon (Zhang et al., 2016). Mapping of landslide susceptibility is one of the important and effective tools for assessment and management of landslides (Dou et al., 2020; Ghasemain et al., 2020).

[*]Corresponding author, Email: hai banglh@utt.edu.vn

Such maps help in better land use planning by decision makers for reducing the damages induced by landslides, especially in hilly areas (Nguyen et al., 2019).

In recent decades, various statistical approaches have been proposed and used for mapping of landslide susceptibility, which includes expert's opinion techniques, weighted techniques, and machine learning (ML) techniques (Nohani et al., 2019; Zhou et al., 2018). Out of these techniques, ML is considered as more accurate and advanced approaches for better performance of models in generating landslide susceptibility map (Zhou et al., 2018). Achour and Pourghasemi (2019) evaluated some of the ML methods such as Support Vector Machine (SVM), Random Forest (RF), and Boosted Regression Tree (BRT) which could improve the accuracy of landslide susceptibility maps. Chang et al. (2019) applied and compared different ML models such as RF, SVM, and Logistic Regression (LR) for mapping landslide susceptibility, and proved that these ML models performed well for generation of landslide susceptibility maps. Hu et al. (2020) compared several ML models such as Naïve Bayes (NB) and SVM with Fractal Theory (FT) model. Other popular machine learning techniques used in mapping of landslide susceptibility include Decision trees (DT) (Pham et al., 2016) and Artificial Neural Networks (ANN) (Harmouzi et al., 2019), and Adaptive Neuro-Fuzzy Inference System (ANFIS) (Aghdam et al., 2016).

Recently, Adnan and Islam (2017) proposed a novel decision forest algorithm namely Forest by Penalizing Attributes (FPA) which builds a set of highly accurate decision trees using the strength of all non-class attributes available in a data set. The proposed algorithm promotes imposes penalties (disadvantageous weights) and strong diversity to those attributes which participated in the latest tree to create the subsequent trees. The analysis results indicated that FPA algorithm is good in creating more balanced and highly accurate decision forests in comparison with other prominent decision forest algorithms, thus it is very good technique in the domain of expert and intelligent systems. In view of this, in this research, the FPA algorithm-based ML model has been used for mapping of landslide susceptibility at Muong Lay district (Vietnam) where landslides often occur every year. Various quantitative statistical indices including Area Under the ROC curve (AUC) were used to evaluate the predictive capability of the model. Weka and GIS software were used for modeling and data processing, respectively.

2. Characteristic of study area

District of Muong Lay, which is lied in the northwest of Vietnam, was chosen as the study area covering about 114.03 km^2 area (Fig. 1). The district is described by rugged topography, with the elevation ranges from 125 to 1778m with slope gradients up to 73%. Average temperature in the area is 22–23°C and rainfall 1483 mm/year. In this area the Dien Bien Phu Fault is prominent active fault. Sheared weathered sedimentary and metamorphic rocks and Quaternary sediments form vulnerable zones for landslides, which occur usually during monsoon period.

3. Materials and Methods

3.1. Geospatial data

For landslide modeling, an inventory map of landslides is constructed using past and present landslide data from available records, satellite and Google Earth images (Nhu et al., 2020a) (Fig. 1). Mostly landslides occur along the Vietnam Highways 6 and 12 are of rotational, translational, debris, rock fall and mixed types. In total 217 landslide events were recorded in the Muong Lay district. Out of these, 70% landslide events (locations) were utilized to generate training dataset for the proposed model, whereas 30% remaining event locations were used to generate testing dataset for the evaluation of the model.

On the base of the local topographical and geo-environmental conditions, and literature review (Phong et al., 2019), nine landslide factors: curvature, deep division, aspect, geology/lithology, weathering crust, slope, and distance from faults, rivers and roads were used in the model study. Detail description of these factors (Fig. 1) and their frequency ratio analysis (Fig. 2) is also presented in Phong et al. (2019).

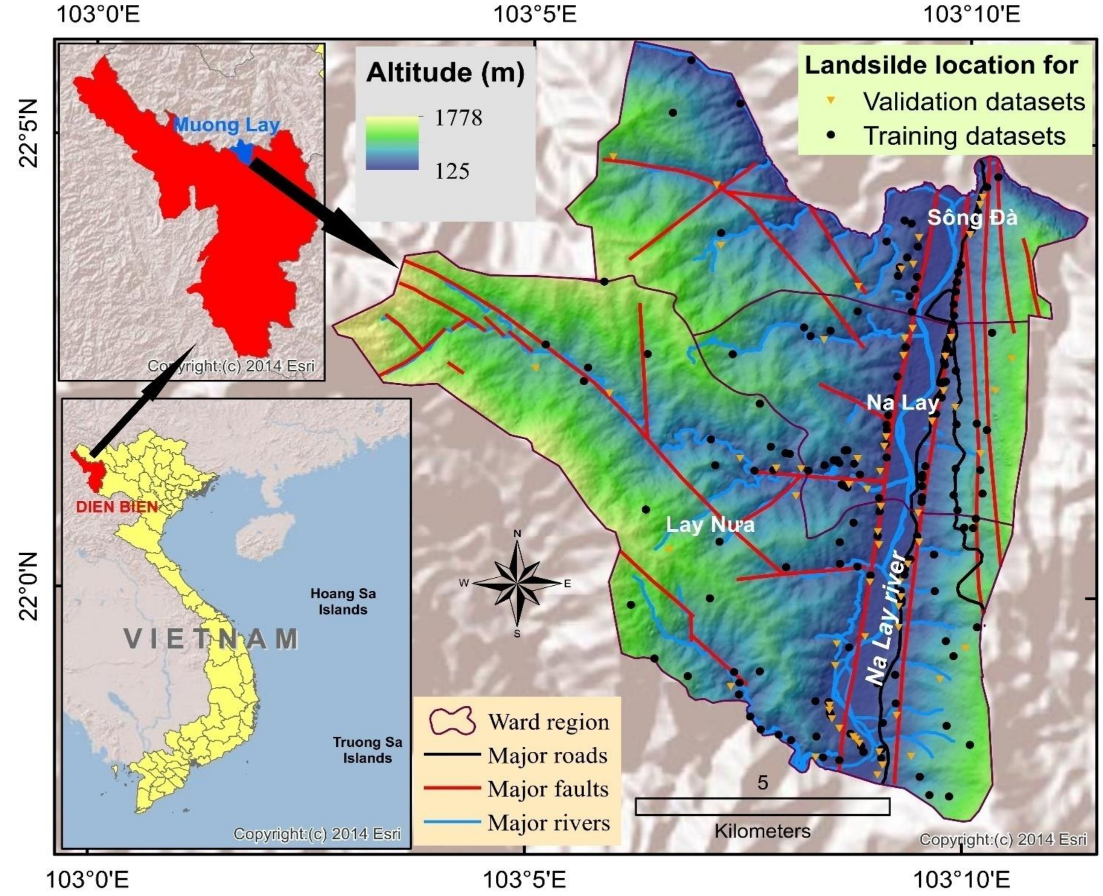


Figure 1. Location map of Muong Lay district, Vietnam showing landslide event points

3.2. Methods used

3.2.1. Forest by Penalizing Attributes (FPA) algorithm

FPA is a recently developed algorithm, which avoids several drawbacks of Random Forest (Adnan and Islam, 2017; Hong et al., 2020). In the FPA algorithm, entire attribute set was used to construct the next decision trees and penalties was simultaneously imposed to the participated attributes in the latest decision tree (Hong et al., 2020; Samat et al., 2019). Each level in the tree possessed a certain weight-range and weights were randomly distributed to the participating attributes (Adnan and Islam, 2017; Hong et al., 2020). With classical decision tree algorithms, more than one tree might have a chance to be the identical if the training datasets used the same distribution of weights on attributes (Hong et al., 2020; Samat et al., 2019). In order to avoid the construction of similar trees, the mechanism of FPA was different and could be summarized in four steps as below (Adnan and Islam, 2017):

Step 1: Generate a bootstrap sample from the training data set.

Step 2: Using the weights of the attributes, a decision tree is generated based on the previously built bootstrap sample.

Step 3: Weight values and gradual weight increment values are updated for all the attributes from the latest tree.

Step 4: Choosing the corresponding weight increment values, which is not in the latest tree, and used to update the weights of the respective attributes.

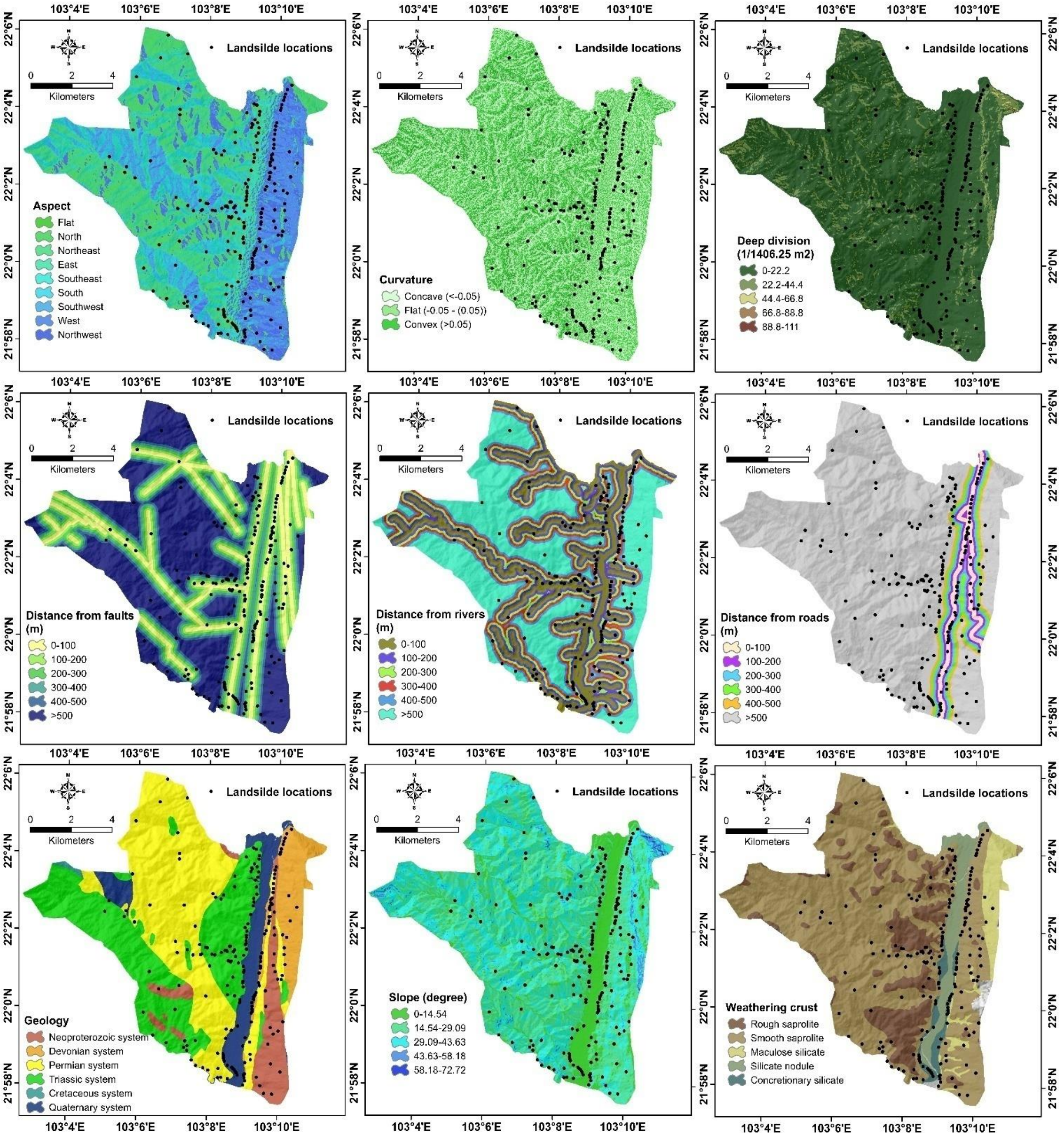


Figure 2. Landslide conditioning factors: (a) aspect, (b) curvature, (c) deep division, (d) distance from faults, (e) distance from rivers, (f) distance from roads, (g) geology, (h) slope, and (i) weathering crust

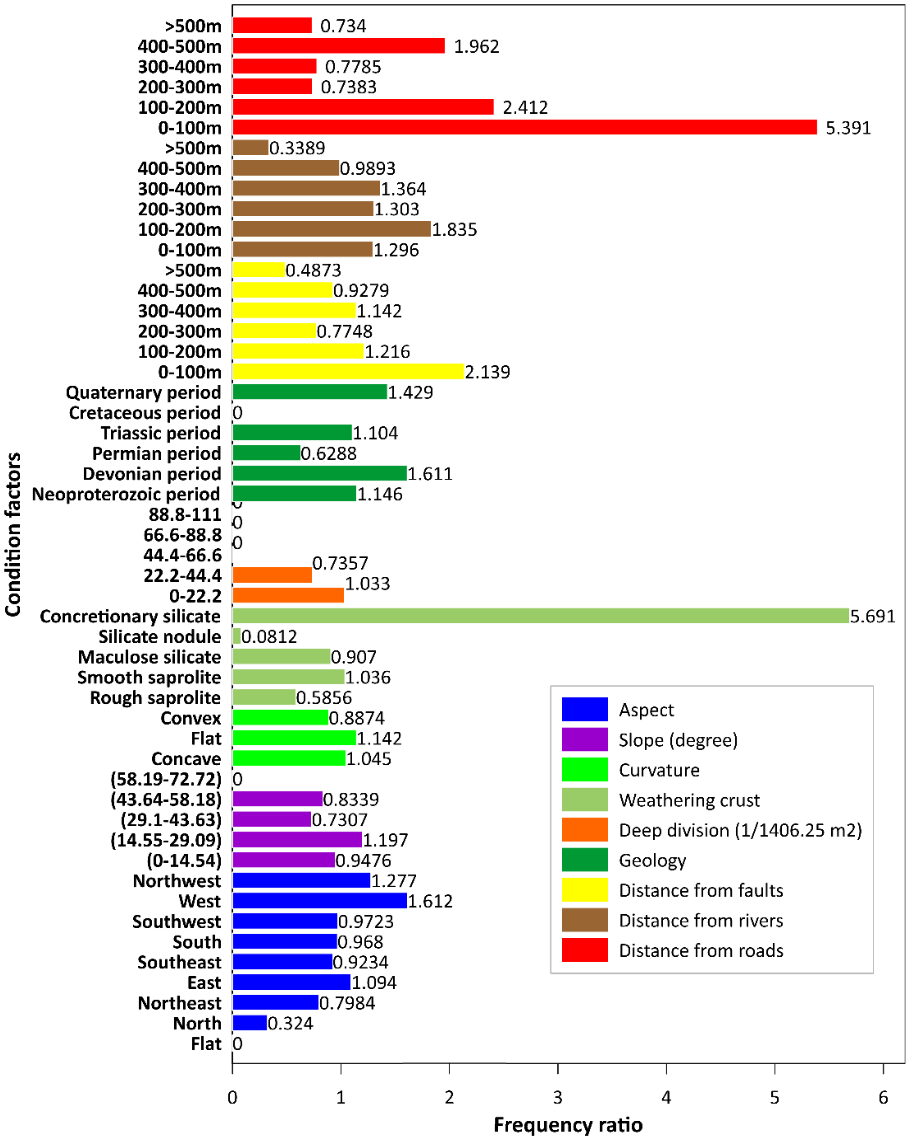


Figure 3. Frequency ratio analysis of landslide conditioning factors

3.2.2. Validation methods

In this study, popular quantitative statistical indexes such as Negative Predictive Value (NPV), Positive Predictive Value (PPV), Kappa index (k), Accuracy (ACC), Root Mean Square Error (RSME), Specificity (SPF), and Sensitivity (SST) were used to assess predictive capability of the FPA model (Dao et al., 2020; Nguyen et al., 2020; Van Dao et al., 2020). Quantitatively, smaller RMSE values represent better accuracy for landslide models. Whereas, higher values of NPV, PPV, SPE, k, SST, and ACC show better accuracy (Nguyen et al., 2019).

In addition, Area Under the ROC curve (AUC) was computed to validate predictive capability of the model. This is the standard performance metric for evaluating classification problems (Pourghasemi et al., 2020). The AUC represents probability that a landslide model will rank for a randomly chosen positive landslide sample higher than a randomly chosen non-landslide. The value of AUC is based on the specificity and sensitivity values on the ROC curve (Nhu et al., 2020b). The highest value of AUC = 1 is best for the prediction of any model.

4. Results and discussion

4.1. Validation of landslide susceptibility model

Landslide susceptibility model using FPA was validated on both training and testing datasets (Table 1 and Figs. 4, 5). In term of training dataset, the FPA has a good performance as the values of PPV, NPV, ACC, SPF, SST, and Kappa are 88.08%, 85.43%, 86.75%, 87.76%, 85.81% and 0.735, respectively. With testing dataset, performance of the FPA is also good as the values of PPV, NPV, ACC, SPF, SST, and Kappa are 83.33%, 78.46%, 80.92%, 82.26%, 79.71%, and 0.615, respectively (Table 1). Figure 4 shows that the error of the models is small on both training (RMSE = 0.322) and testing (RMSE = 0.373) datasets. Based on the ROC curve analysis, the FPA model performs well for both training (AUC = 0.935) and testing (AUC = 0.882) datasets.

In general, it is reasonable to state that the FPA model has a good performance for mapping of landslide susceptibility at the study area. Compared with other published studies using same datasets, it can be seen that the FPA is much better models than SVM (AUC = 0.87), LR (AUC = 0.863), ANN (AUC = 0.865), and REPT (AUC = 0.851) (Phong et al., 2019). This is reasonable as FPA performance is similar to Bagging or Random Forest, where bootstrap samples are used to construct the decision trees (Adnan and Islam, 2017). Therefore, the trees are different if they are constructed from different bootstrap samples. Moreover, once the first tree is constructed, the second tree will be different from the previously constructed tree due to mechanism of building trees in FPA. Last but not least, the strategy of the weight increment of FPA algorithm will generate different distribution of weight to different trees. Another advantage of the FPA is that it avoids any hyper-parameters during the learning process (Adnan and Islam, 2017). With these advantages, performance of the FPA is a robust and reliable model for mapping of landslide susceptibility as observed in this study.

Table 1. Model performance using various validation criteria

No.	Parameters	Training model	Validation model
1	TP	133	55
2	TN	129	51
3	FP	18	11
4	FN	22	14
5	PPV (%)	88.08	83.33
6	NPV (%)	85.43	78.46
7	SST (%)	85.81	79.71
8	SPF (%)	87.76	82.26
9	ACC (%)	86.75	80.92
10	Kappa	0.735	0.615

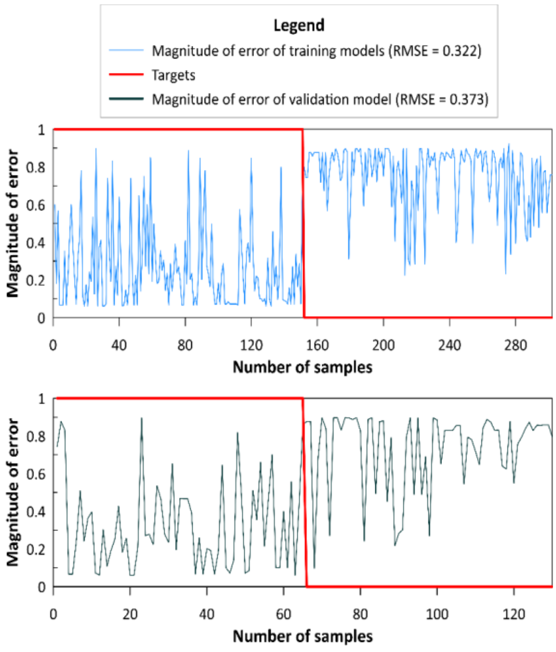


Figure 4. Error analysis of the models

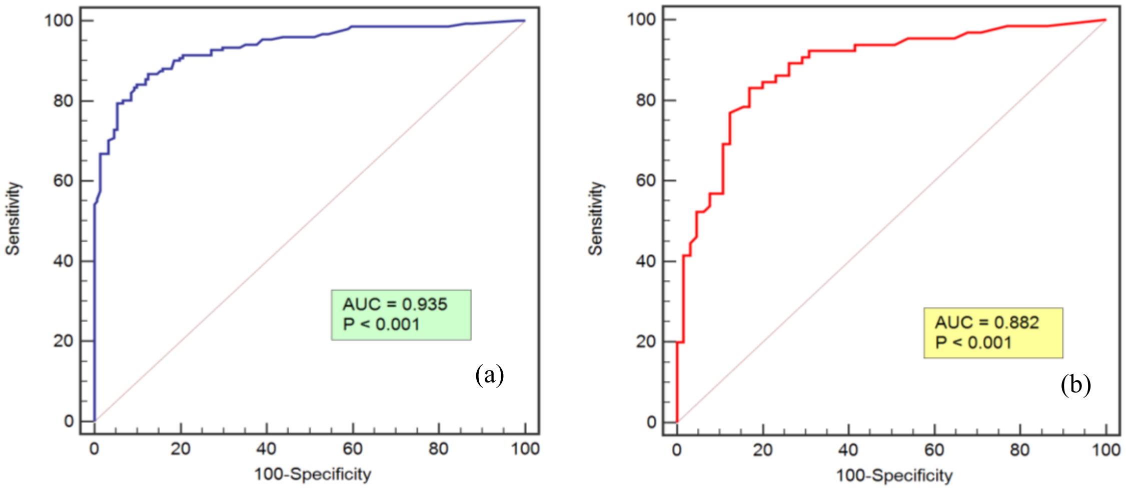


Figure 5. Evaluation of model using ROC curve: (a) training dataset (b) validating dataset

4.2. Construction and validation of landslide susceptibility map

Map of landslide susceptibility was finally constructed using the FPA model (Fig. 6). This process was implemented in two main steps. Firstly, indexes of landslide susceptibility were extracted from the training process of FPA. These indexes were then assigned for all pixels of the area. Secondly, these indexes were finally classified into five classes (very high, high, moderate, low and very low) using the natural break classification method in GIS application to construct the final landslide susceptibility map (Fig. 6).

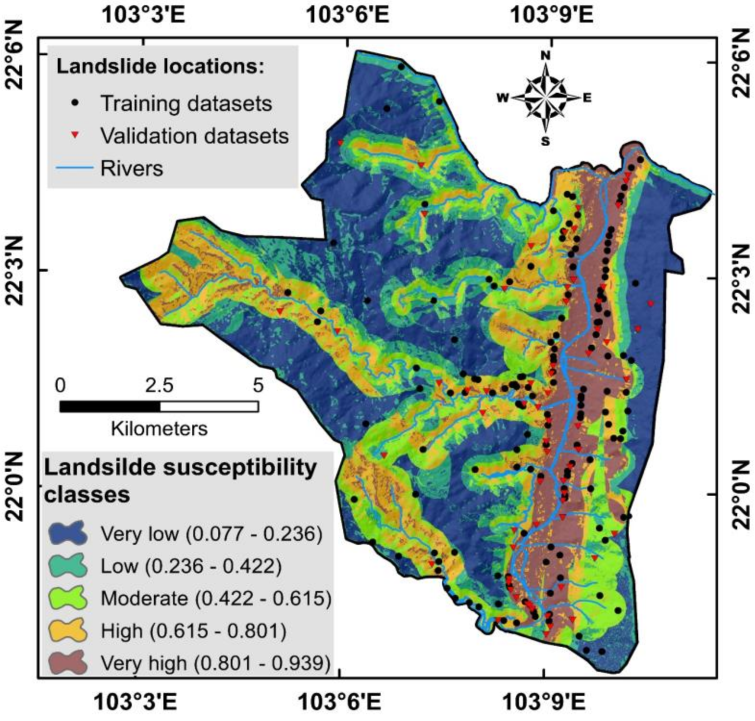


Figure 6. Landslide susceptibility maps produced by FPA model

Validation of the constructed landslide susceptibility map was also done using frequency ratio analysis, which indicated that 15.6% of the study area falls into very high susceptibility zone, 18.9 % in high susceptibility zone, 16.76% in moderate susceptibility zone, 16.8% in low susceptibility zone, and 31.8% in very low susceptibility zone (Fig. 7). Validation of the susceptibility map was also done using frequency ratio analysis (Fig. 7). Results show that very high and high susceptibility classes have the highest values of frequency ratio (2.668 for very high class and 1.917 for high class) which indicates that the performance of landslide susceptibility map constructed is good and reliable for practical application in better land use planning and hazard management at the study area.

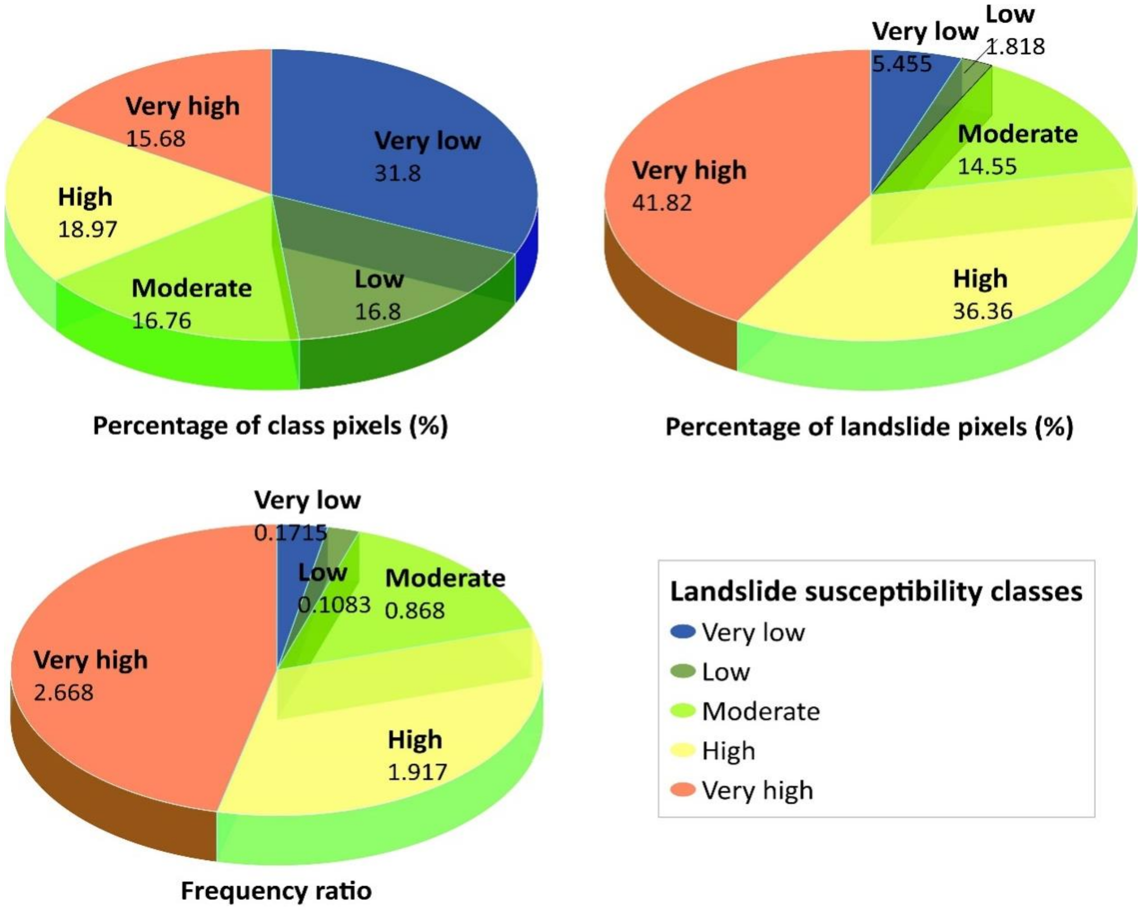


Figure 7. Validation of landslide susceptibility map

5. Concluding remarks

In this study, we have applied first time FPA algorithm-based ML approach for landslide susceptibility mapping at Muong Lay district (Vietnam). Results of the statistical analysis show that the performance of the FPA is very good for landslide susceptibility mapping on both training (AUC: 0.935) and testing (AUC: 0.882) datasets. Analysis of the model performance based on other statistical methods such as NPV, PPV, ACC, SPF, SST, and Kappa also show very good results. RMSE Value of the model was also small on both training (RMSE: 0.322) and testing (RMSE: 0.373) datasets. Performance evaluation of FPA model shows that this model is a robust and reliable for mapping of landslide susceptibility. Thus, the proposed FPA based ML model can be utilized for the construction of accurate landslide susceptibility map and for better land use planning and hazard management not only of the study area but also of other landslide prone areas depending on the local geo-environmental factors.

The FPA algorithm is relatively new which has been applied in landslide study in the present paper. Results indicated that it is effective in generating highly accurate results, thus it is a promising ML algorithm for landslide studies. Its further application in other areas is required to be explored for proper landslide management.

Acknowledgments

We thank the support from the project "Geological hazards assessment of Dien Bien - Lai Chau fault zone base on application machine learning, artificial intelligence", VAST05.05/20-21.

Conflict of interest: None

References

Achour Y., Pourghasemi H.R., 2019. How do machine learning techniques help in increasing accuracy of landslide susceptibility maps? Geoscience Frontiers.

Adnan M.N., Islam M.Z., 2017. Forest PA: Constructing a decision forest by penalizing attributes used in previous trees. Expert Systems with Applications, 89, 389–403.

Aghdam I.N., Varzandeh M.H.M., Pradhan B., 2016. Landslide susceptibility mapping using an ensemble statistical index (Wi) and adaptive neuro-fuzzy inference system (ANFIS) model at Alborz Mountains (Iran). Environmental Earth Sciences, 75(7), 553.

Chang K.-T., Merghadi A., Yunus A.P., Pham B.T., Dou J., 2019. Evaluating scale effects of topographic variables in landslide susceptibility models using GIS-based machine learning techniques. Scientific reports, 9(1), 1–21.

Dao D.V., Ly H.-B., Vu H.-L.T., Le T.-T., Pham B.T., 2020. Investigation and Optimization of the C-ANN Structure in Predicting the Compressive Strength of Foamed Concrete. Materials, 13(5), 1072.

Dou J., Yunus A.P., Merghadi A., Shirzadi A., Nguyen H., Hussain Y., Avtar R., Chen Y., Pham B.T., Yamagishi H., 2020. Different sampling strategies for predicting landslide susceptibilities are deemed less consequential with deep learning. Science of The Total Environment, 720, 137320.

Ghasemain B., Asl D.T., Pham B.T., Avand M., Nguyen H.D., Janizadeh S., 2020. Shallow landslide susceptibility mapping: A comparison between classification and regression tree and reduced error pruning tree algorithms. Vietnam Journal of Earth Sciences. 42(3), 208–227. Doi: 10.15625/0866-7187/42/3/14952.

Harmouzi H., Nefeslioglu H.A., Rouai M., Sezer E.A., Dekayir A., Gokceoglu C., 2019. Landslide susceptibility mapping of the Mediterranean coastal zone of Morocco between Oued Laou and El Jebha using artificial neural networks (ANN). Arabian Journal of Geosciences, 12(22), 696.

Hong H., Liu J., Zhu A.-X., 2020. Modeling landslide susceptibility using LogitBoost alternating decision trees and forest by penalizing attributes with the bagging ensemble. Science of the total environment, 718, 137231.

Hu Q., Zhou Y., Wang S., Wang F., 2020. Machine learning and fractal theory models for landslide susceptibility mapping: Case study from the Jinsha River Basin. Geomorphology, 351, 106975.

Nguyen P.T., Ha D.H., Jaafari A., Nguyen H.D., Van Phong T., Al-Ansari N., Prakash I., Le H.V., Pham B.T., 2020. Groundwater Potential Mapping Combining Artificial Neural Network and Real AdaBoost Ensemble Technique: The Dak Nong Province Case-study, Vietnam. International Journal of Environmental Research and Public Health, 17(7), 2473.

Nguyen V.-T., Tran T.H., Ha N.A., Ngo V.L., Nadhir A.-A., Tran V.P., Duy Nguyen H., Malek M.A., Amini A., Prakash I., 2019. GIS Based Novel

Hybrid Computational Intelligence Models for Mapping Landslide Susceptibility: A Case Study at Da Lat City, Vietnam. Sustainability, 11(24), 7118.

Nhu V.-H., Shirzadi A., Shahabi H., Chen W., Clague J.J., Geertsema M., Jaafari A., Avand M., Miraki S., Asl D.T., 2020a. Shallow Landslide Susceptibility Mapping by Random Forest Base Classifier and its Ensembles in a Semi-Arid Region of Iran. Forests, 11(4), 421.

Nhu V.-H., Shirzadi A., Shahabi H., Singh S.K., Al-Ansari N., Clague J.J., Jaafari A., Chen W., Miraki S., Dou J., 2020b. Shallow Landslide Susceptibility Mapping: A Comparison between Logistic Model Tree, Logistic Regression, Naïve Bayes Tree, Artificial Neural Network, and Support Vector Machine Algorithms. International Journal of Environmental Research and Public Health, 17(8), 2749.

Nohani E., Moharrami M., Sharafi S., Khosravi K., Pradhan B., Pham B.T., Lee S., M. Melesse A., 2019. Landslide susceptibility mapping using different GIS-based bivariate models. Water, 11(7), 1402.

Pham B.T., Bui D.T., Dholakia M., Prakash I., Pham H.V., 2016. A comparative study of least square support vector machines and multiclass alternating decision trees for spatial prediction of rainfall-induced landslides in a tropical cyclones area. Geotechnical and Geological Engineering, 34(6), 1807–1824.

Phong T.V., Phan T.T., Prakash I., Singh S.K., Shirzadi A., Chapi K., Ly H.-B., Ho L.S., Quoc N.K., Pham B.T., 2019. Landslide susceptibility modeling using different artificial intelligence methods: A case study at Muong Lay district, Vietnam. Geocarto International, 1–24.

Pourghasemi H.R., Kornejady A., Kerle N., Shabani F., 2020. Investigating the effects of different landslide positioning techniques, landslide partitioning approaches, and presence-absence balances on landslide susceptibility mapping. Catena, 187, 104364.

Samat A., Liu S., Persello C., Li E., Miao Z., Abuduwaili J., 2019. Evaluation of ForestPA for VHR RS image classification using spectral and superpixel-guided morphological profiles. European journal of remote sensing, 52(1), 107–121.

Shirzadi A., Saro L., Joo O.H., Chapi K., 2012. A GIS-based logistic regression model in rock-fall susceptibility mapping along a mountainous road: Salavat Abad case study, Kurdistan, Iran. Natural hazards, 64(2), 1639–656.

Shirzadi A., Shahabi H., Chapi K., Bui D.T., Pham B.T., Shahedi K., Ahmad B.B., 2017. A comparative study between popular statistical and machine learning methods for simulating volume of landslides. Catena, 157, 213–226.

Van Dao D., Jaafari A., Bayat M., Mafi-Gholami D., Qi C., Moayedi H., Van Phong T., Ly H.-B., Le T.-T., Trinh P.T., 2020. A spatially explicit deep learning neural network model for the prediction of landslide susceptibility. Catena, 188, 104451.

Zhang G., Cai Y., Zheng Z., Zhen J., Liu Y., Huang K., 2016. Integration of the statistical index method and the analytic hierarchy process technique for the assessment of landslide susceptibility in Huizhou, China. Catena, 142, 233–244.

Zhong C., Liu Y., Gao P., Chen W., Li H., Hou Y., Nuremanguli T., Ma H., 2020. Landslide mapping with remote sensing: challenges and opportunities. International Journal of Remote Sensing, 41(4), 1555–1581.

Zhou C., Yin K., Cao Y., Ahmed B., Li Y., Catani F., Pourghasemi H.R., 2018. Landslide susceptibility modeling applying machine learning methods: A case study from Longju in the Three Gorges Reservoir area, China. Computers & Geosciences, 112, 23–37.