LANDSLIDE HAZARD MAPPING USING ENSEMBLE MACHINE LEARNING ALGORITHM IN BA BE LAKE BASIN

Le Van Thin, Dao Anh Tuan, Nguyen Dang Giap
Key Laboratory of River and Coastal Engineering
Vietnam Academy for Water Resources
Nguyen Thi Thuy Quynh
Academy of Finance

Abstract: *The paper presents landslide hazard mapping for Ba Be lake basin using four machine learning models namely: Logistic Regression (LR), Random Forest (RF), Gradient Boosting (GBC) and Xgboost (XGB). Based on field surveys, literature review and available data, ten indicators were used to derive the landslide hazard, including: slope, soil, plan curvature, normalized difference vegetation index (NDVI), topography, geomorphons, distance from roads, distance from rivers, density of streams and rainfall accumulation. These indicators were arranged in grid cells. The Receiver Operating Curves (ROC) and Area Under Curve (AUC) were used to validate the modes. The results of the analysis showed that the RF and GBC models had the highest predictive ability (AUC = 0.88), followed by the XGB models with AUC = 0.86 and the last one is LR with AUC = 0.78. The results could be useful for planners in general land use planning and management.*
Keywords: *Landslide, hazard mapping, machine learning*

1. INTRODUCTION

One of the most serious geological hazards in the world is landslide, especially in mountainous areas. The landslide events are causing widespread destruction of infrastructure, the loss of life and properties [1]. Mountains cover about 65% of land in Ba Be lake basin where is prone to landslide hazard. According to the results of landslide investigation, the area of Ba Be lake and the surrounding areas has more than 800 landslide sites in the period from 2000 to 2014 [2].

Ba Be lake basin is small with an area of 464.68 km² located in the west of Bac Kan province and adjacent to Tuyen Quang province. Ba Be lake basin is contributed by three major tributaries, namely Bo Lu, Cho Leng and Ta Han.

Although the main research object is Ba Be lake basin, in this study area only 36 landslide sites have been collected, so an expanded area is considered to enhance landslide data and rainfall stations surrounds Ba Be lake area with part of Tuyen Quang province as follows:

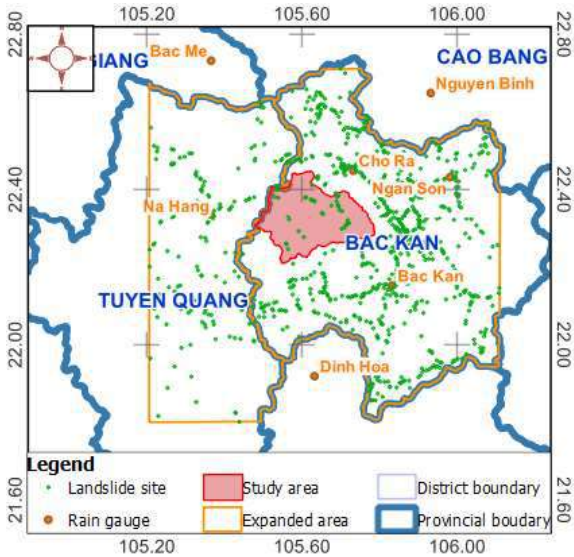


Figure 1: Study area

Receipt Date: August 15th, 2022
Review Approval Date: September 16th, 2022
Publish Approval Date: October 03th, 2022

2. METHODS

Qualitative approach to develop the landslide susceptibility mapping and modelling have been adopted by several researchers [3] [4] [5]. Recently, machine learning techniques such as artificial neural network are applied to landslide susceptibility analysis to remove subjectivity in qualitative analysis [6] [7] [8].

2.1. Data collection and preparation

To carry out this study, the required data were collected and sorted from different data sources such as DEM from Ministry of Natural Resources and Environment (MONRE) with 1/10.000 scale, satellite image from Google Earth Engine (GEE), soil from Department of Natural Resources and Environment Bac Kan (DONRE), road from Open Street Map (OSM), landslide sites from MONRE [2] and field survey, daily rainfall from MONRE.

Table 1: Sources of data

No.	Data	Sources	Note/Details
	DEM	MONRE	1/10.000 scale
	Landsat images	GEE	Period from 2000 to 2020
	Soil map	DONRE	Collected in 2020
	Road map	OSM	Downloaded in April 2021
	Landslide sites	DARD and field survey	Vietnam Institute of Geosciences and Mineral Resources (VIGMR); Bac Kan Provincial Commanding Committee of Natural Disaster Prevention and Control, Search and Rescue (PCDPSR); Field trip in 2020.
	Daily rainfall	MONRE	From 1986 to 2020 for 7 gauges (Bac Me, Bao Lac, Ngan Son, Nguyen Binh, Cho Ra, Na Hang, Dinh Hoa) and Bac Kan for 1996-2020 period.

Terrain analyses include slope, plan curvature, geomorphons, density of streams. NDVI is extracted satellite images from 2000 ÷ 2020 using GEE tool. The resolution of factors analysis is 20m.

The whole factor analysis is done in python language: GIS spatial analysis using GDAL, and GRASS libraries or QGIS toolbox, data series analysis using Pandas and Numpy libraries, charts using Matplotlib and Seaborn libraries. The maps is presented by QGIS.

2.2. Spatial database

1. Landslide inventory database

This study have collected data on landslides from 2 sources: (1) from PCDPSR with several landslide locations collected in recent period (with unspecified timeline); and (2) The landslide investigation project which was implemented in Bac Kan by the Vietnam Institute of Geosciences and Mineral Resources in 2014 [2].

The collected landslide sites are reprojected to VN-2000 and removed duplicate points at the two data sources, besides, some unsuitable landslide sites (after review and evaluation) were also removed. In the Ba Be lake basin and the surrounding area, 808 landslide sites were selected in both Bac Kan and Tuyen Quang provinces in this study.

2. Landslide indicators

- Slope

The force acting on the steep slope is the main cause of landslides [9], in many studies, the slope is considered as a major factor causing landslides [10] [11] [12]. Depending on the characteristics of each area (in terms of geological structure, landcover and so forth), there is a different hazard level of landslide.

There are some minor adjustments in the slope used in this study compared to the slope of the grid cells. It is assume that a grid cell located on a steep shoulder also has a very high risk of landslide while the slope of that grid cell is very low. Therefore, the considered slope of a grid cell in this landslide assessment as the average slope of the three grid cells in the direction flow which originates from that grid cell. The newly formed slope map is a combination of the slope map and the flow direction map.

A new tool was developed in python to generate the new slope map (called slope down).

- *Plan curvature*

Plan curvature is the curvature of a contour line at a given point on the topographic surface and it is generated from a digital elevation model (DEM). This factor has an unlimited range that can take either positive or negative value. A positive value of the index indicates flow divergence while a negative plan curvature value indicates flow convergence [13]. The plan curvature is calculated by SAGA tool in QGIS.

- *Distances from the river and road*

In the study area, most of the landslide sites are located near roads and rivers. The proximity tool in the GDAL library is utilized to calculate the distances of each pixel to rivers and roads.

- *Normalized Difference Vegetation Index (NDVI)*

Although in many landslide studies, land use and land cover (LULC) maps are used, in this study, the NDVI is considered as the preferred index. Because landslides are collected over a 20-year period, it is not possible to determine the LULC corresponding to each period. Some studies have shown that higher NDVI refers to lower probability of hazard of landslides. However, in the process of data collection and analysis, it is found that even in areas with high vegetation index, many landslides appeared. Landsat images in Google Earth Engine were collected, processed and calculated the NDVI for the entire study area.

- *Elevation*

Topographic elevation is considered as an impact on the distribution of rainfall and humidity in the area. DEM data is obtained from MONRE at the scale of 1/10,000

- *Soil*

Different soil types have different cohesive characteristics, which have direct effects on landslides. The land map was collected from DONRE to build a soil map for the study area.

- *Geomorphons*

Geomorphons is a factor that is rarely considered in landslide studies. However, it is found that hilly areas have a greater hazard of landslides than other areas, so this index is considered to assess the impact on landslide hazard.

Geomorphons were described by Jasiewicz as a calculation method in 2013 [14] to indicate 10 common types of landforms including: flat; peak; ridge; shoulder; hollow; slope; spur; footslope; valley; and pit as figure below:

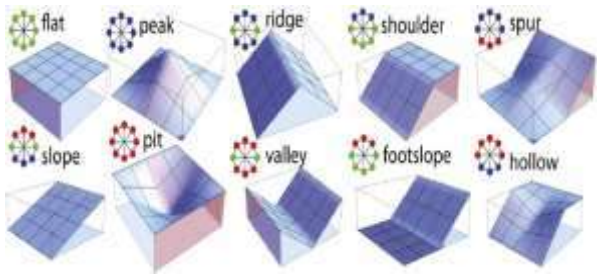


Figure 2: Types of landforms

The r.geomorphons tool in GRASS library were used to build this map.

- *Density of streams*

High density of streams increases soil saturation and reduces cohesion, which are indirect cause of the landslide process. The density of rivers and streams was calculated based on a self-developed algorithm in python language.

- *Precipitation*

The characteristics of the largest accumulated rain by days at 8 rainfall stations are shown as follows:

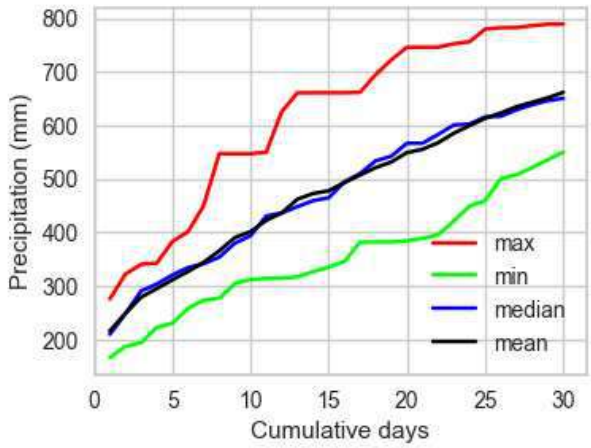


Figure 3: Characteristics of accumulated rainfall

There have been a number of studies that have attempted to determine a correlation between the time of precipitation accumulation and the landslide event, but have failed to show a specific way [15]. At each different time of accumulation of precipitation, landslide events also have different responses.

The figure 3 show that the accumulation period of 9÷11 days, the total maximum and minimum rainfall accumulation was almost unchanged. In this study, 10-day cumulative rainfall was selected as the rainfall to calculate landslide hazard.

Since it is not possible to determine the time of landslide occurrence for each collected landslide site (2000 ÷ 2020 period), this study is assumed that there was a rain event over a 20-year period as the landslide-induced precipitation threshold for the collected landslide sites. Based on this assumption, precipitation with a return period of 20 years for 10-day cumulative precipitation is calculated.

2.3. Modeling landslide

1. Logistic regression (LR)

The LR model has become one of the most popular methods in machine learning algorithms. This is a multivariate statistical method that is generally used to deal with binary classification issues. An outstanding feature of LR is that it calculates the weights for each conditioning factor.

2. Random Forest

Random forest is an ensemble learning method, first proposed by Breiman [16] and Cutler [17] that builds multiple decision trees through different subsets of data and votes the results of many decisions tree to get the results of the random forest.

A large number of available studies have shown that random forests have a significant tolerance to extraneous factors and noise, unlikely to over-fit, and of high prediction accuracy and stability [18].

3. Gradient boosting classification

Similar to RF, gradient boosting classification (GBC) is another technique useed to perform supervised machine learning. GBC has generated a predictive model in the form of an ensemble of weak prediction models as decision trees [19].

4. Extreme gradient boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) is a kind of efficient and optimized gradient tree boosting algorithm and has recently gained wide popularity, especially due to its exceptional performance in Kaggle competitions.

The core achievement behind the XGBoost algorithm is its scalability in all scenarios and fast processing execution by providing bagging-bootstrap aggregation and randomization, XGBoost both prevents overfitting problems while taking into account the variance trade-off. Due to such special advantages, XGBoost has recently become one of the most popular ML algorithms in landslide susceptibility mapping studies [20] [21]

3. RESULT AND DISCUSSION

3.1. Data distributions

The distribution of data greatly affects the results of machine learning models. Big data bias causes inaccurate results. Therefore, input data need to renormalize the in the value range between 0 and 1 as shown in the tables 2. In addition, soil and geomorphons are not numeric format, the frequency ratio method was applied to convert categories to numbers for machine learning model as table 3:

Table 2: Actual and normalized data

No.	Landslide indicator	Actual value range	Normalized value
1	Slope	0 ÷ 77.2 º	Divide by 100
2	Plan Curvature	-124.98 ÷ 51.77	0 ÷ 1
3	Distance from road	0 ÷ 7,955	0 ÷ 1
4	Distance from streams	0 ÷ 1,627	0 ÷ 1
5	NDVI	-1 ÷ 1	-1 ÷ 1
6	Elevation	5.96 ÷ 1914.84 m	0 ÷ 1
7	Soil	Separate value	Frequency ratio → 0 ÷ 1
8	Geomorphons	Separate value	Frequency ratio → 0 ÷ 1
9	Density of streams	0 ÷ 1	0 ÷ 1
10	10 days accumulated rainfall	281.34 ÷ 433.45 mm	Divide by 1000

Table 3: Frequency ratio for soil and geomorphons

No.	Class	Area (km^2)	Area percentage (%)	Number of landslides	Percentage of landslides (%)	FR	LSI
Geomorphons							
1	Peak	182.53	2.50%	6	0.70%	0.3	-1.2
2	Ridge	953	12.90%	150	18.60%	1.44	0.366
3	Hollow	1,082.91	14.60%	155	19.20%	1.31	0.271
4	Slope	1,663.56	22.50%	197	24.40%	1.09	0.082
5	Spur	1,128.88	15.20%	139	17.20%	1.13	0.121
6	Valley	1,666.93	22.50%	112	13.90%	0.62	-0.485
7	Pit	727.44	9.80%	49	6.10%	0.62	-0.482
Total		7,405.25	100.00%	808	100.00%		
Soil							
1	Rocky mountain	648.8	8.80%	31	3.80%	0.44	-0.826
2	Rhodic Ferrasols	187.25	2.50%	28	3.50%	1.37	0.315
3	Xanthic Ferrasols	300.34	4.10%	26	3.20%	0.79	-0.231
4	Ferralic Acrisols	6,041.62	81.60%	712	88.10%	1.08	0.077
5	Humic Acrisols	228.07	3.10%	11	1.40%	0.44	-0.816
Total		7,406.08	100.00%	808	100.00%		

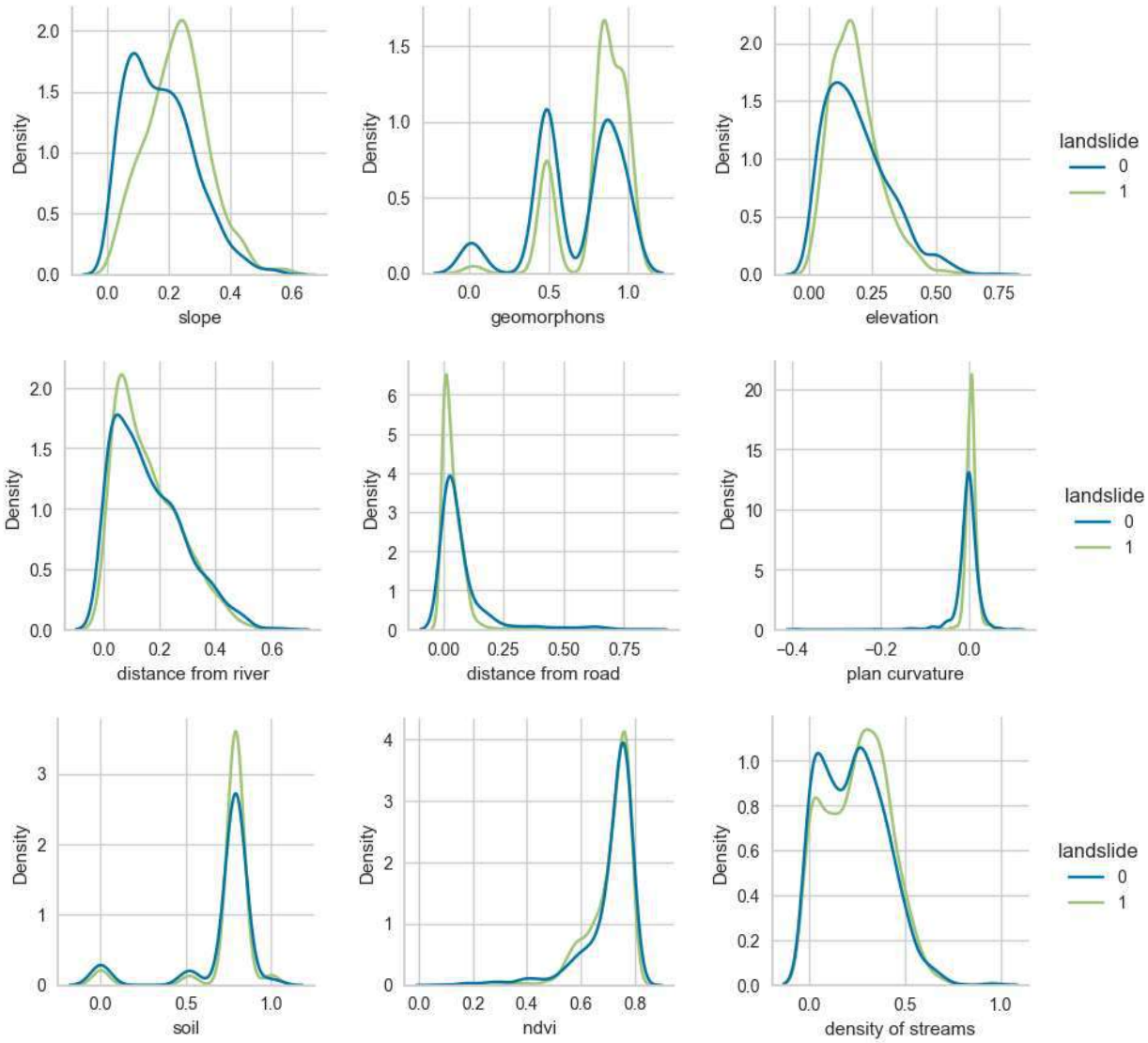


Figure 4: Graphs of distributions of the variables (after normalization)

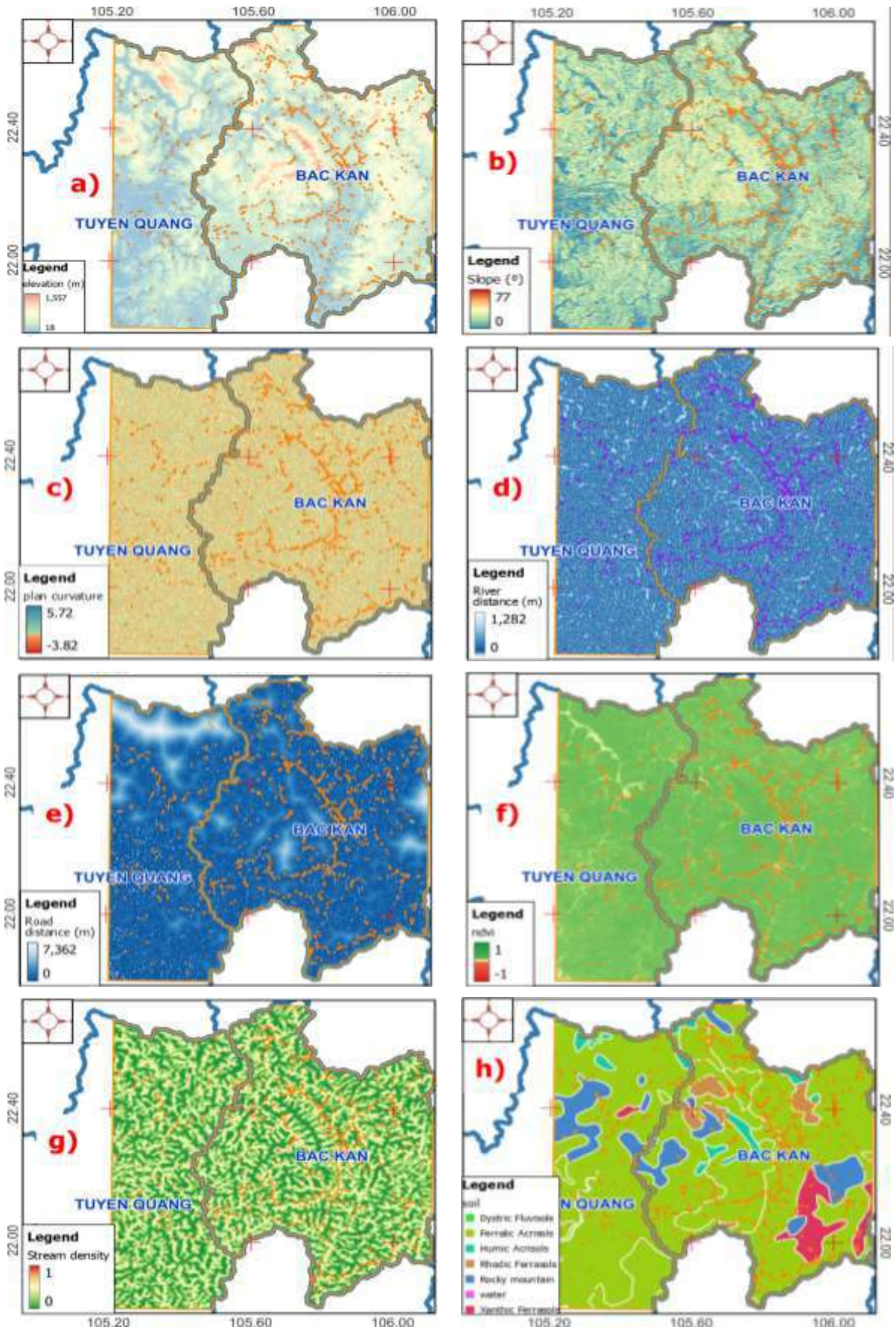
The distribution of data shows that the landslide sites have a higher slope than the non-slide sites. Meanwhile, areas with hollow, spur and slope landform tend to have more landslides than others.

At the elevation from 100÷500m, there are the most number of landslide points (615/808 points), especially from 250÷500m, there are 398 landslide points (accounting for nearly 50% of the total number of landslides).

A large number of landslide points also occurred in areas with normalized difference vegetation index ranging from 0.4÷0.6, some places with high NDVI (≥ 0.6) still appeared landslide spots.

According to the plan curvature, the slightly curved areas have more landslides than other areas. Beside, the area near rivers, streams and roads also has a higher density of landslides than other areas.

Furthermore, landslide sites occur more frequently at locations with slopes from 25π to 40π, located on the mid-mountain area.


105.20
105.60
106.00
22.40
22.00
a)
BAC KAN
TUYEN QUANG
Legend
elevation (m)
1,557
18
b)
Legend
Slope (°)
77
0
c)
Legend
plan curvature
5.72
-3.82
d)
Legend
River distance (m)
1,282
0
e)
Legend
Road distance (m)
7,362
0
f)
Legend
ndvi
1
-1
g)
Legend
Stream density
1
0
h)
Legend
soil
Dystric Fluvisols
Ferralic Acrisols
Humic Acrisols
Rhodic Ferralsols
Rocky mountain
water

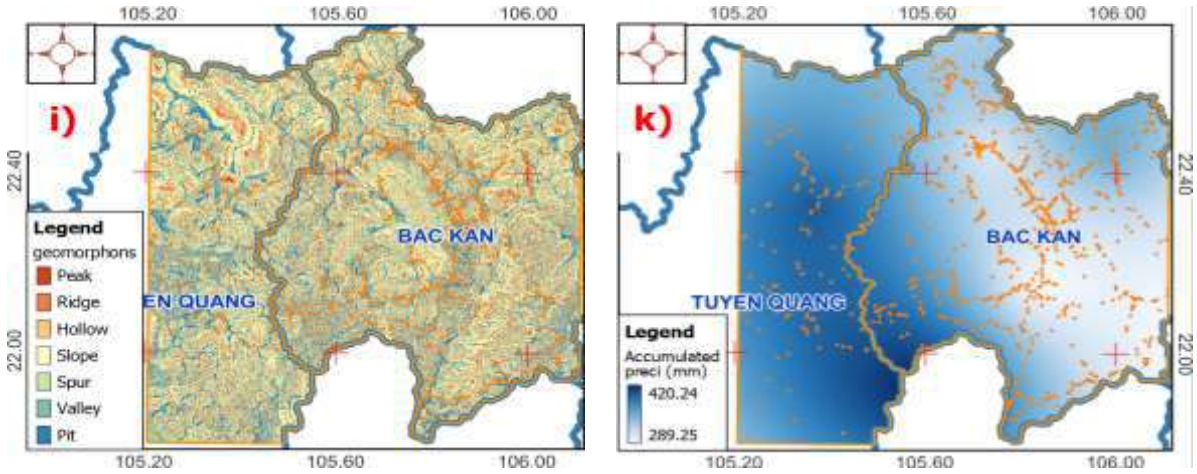


Figure 5: a) Digital Elevation Model; b) Slope; c) Plan Curvature; d) Distance from river; e) Distance from road; f) Normalized Difference Vegetation Index; g) Stream density; h) Soil; i) Geomorphons; k) Precipitation (20 year-return period in 10 day-induced rainfall)

Sites near streams and roads are more prone to landslides than distant points, while spots with slight or no curvature tend to be more prone to landslides than areas with extreme curvature.

Rhodic Ferrasols are soils susceptible to landslides, followed by Ferralic Acrisols, in addition, the density of rivers and streams at about 10÷40% more susceptible than other areas.

3.2. Models

1. ROC and AUC

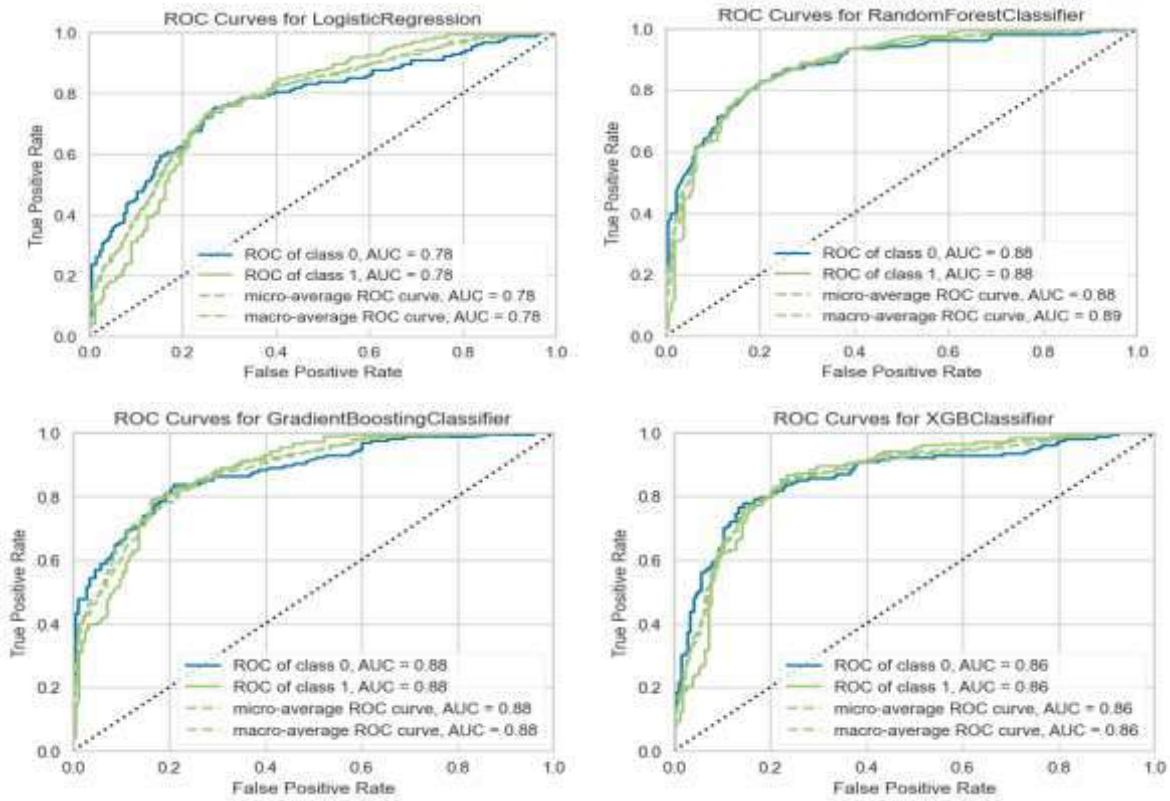


Figure 6: ROC curves and AUC value for each model of machine learning

Simulation results show that the RF model gives the best simulation ability (AUC ≈ 0.88) and has the highest accuracy (accuracy = 0.81), followed by two models, GBC and XGB, give as equivalent. The LR model gives the lowest results with AUC = 0.78 and accuracy = 0.73.

2. Factor importance analysis

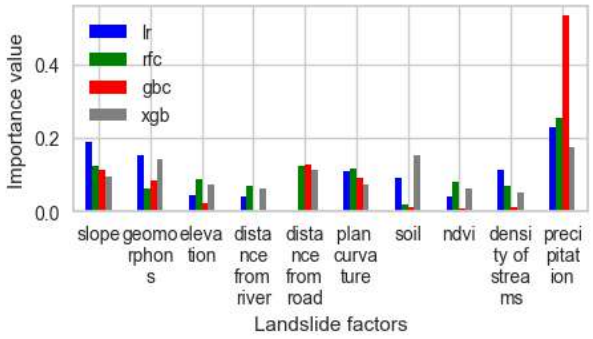


Figure 7: The importance value of each factor by machine learning model

Each machine learning model has a different assessment of the importance of each factor (figure 7). However, coefficient of precipitation in all models is highest, followed by slope. The GBC model has a bias in the precipitation coefficient, while the XGB model assumes that the precipitation coefficient, soil type and geomorphons factor are also quite important in landslide hazard assessment.

The RF model assesses rainfall, slope, distance to roads and plan curvature as important factors in landslide hazard assessment. Meanwhile, the LR model suggests that in addition to rainfall, slope, geomorphology and plan curvature have a great influence.

Although there are different views on the importance of each factor, the models give very good evaluation results, which enriches the views in the assessment of landslide hazard. In the final result, the average combination of all models will be used.

3.3. Result of landslide hazard maps

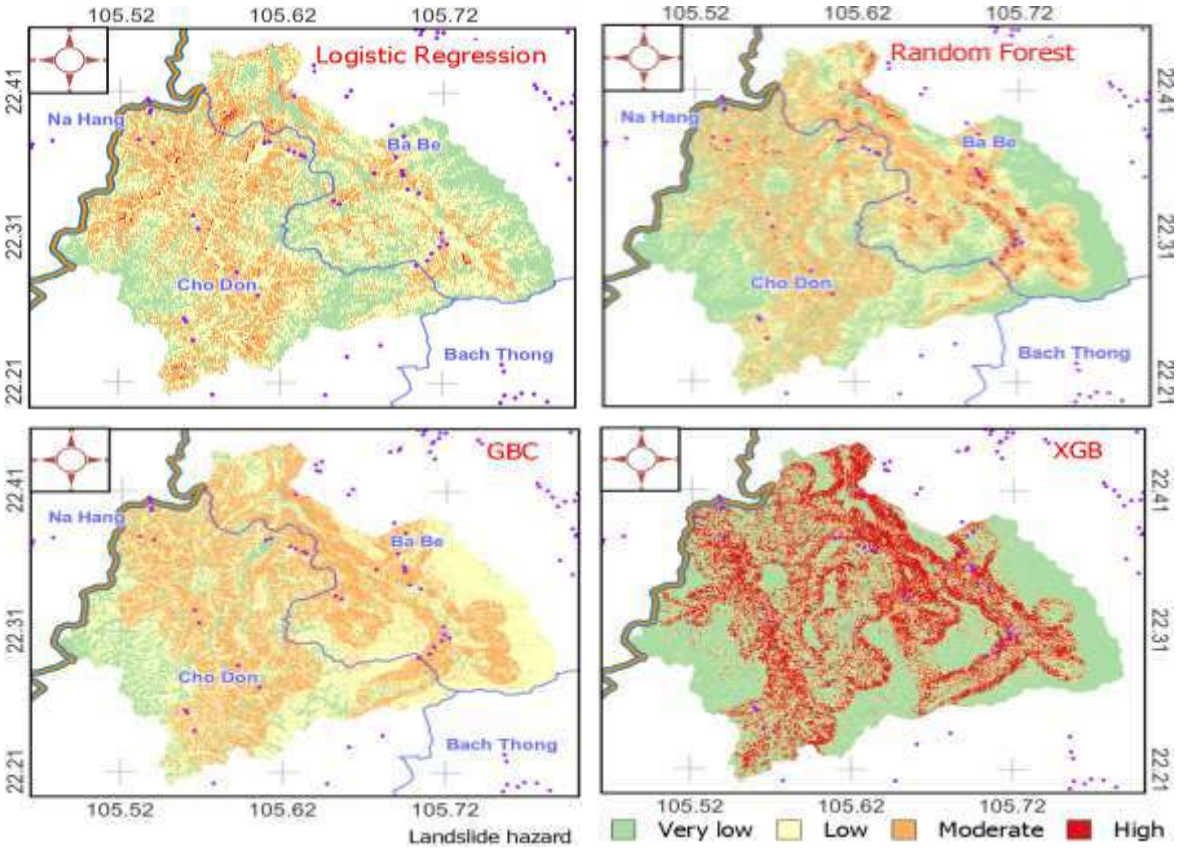


Figure 8: Landslide hazard mapping for each machine learning model

Different views on the importance of these factors have resulted in different landslide hazard maps (figure 8). LR and RF models suggest that sites with high precipitation and slope are the main causes of landslides, these sites are distributed in the mountainous slopes. In addition, the points considered to be high hazard for XGB and GBC models are distributed near the traffic roads where there is steep slope too.

Although all have good accuracy, 2 models LR and RF show a better suitable in landslide hazard distribution than the other 2 models. Although the majority of landslide sites are located next to roads, this factor should still be only a small factor in the formation of landslide hazard. The bias in road distance in the collected data has caused the unbiased recognition of these two models. An average combined map of 4 machine learning models is the final product that this study purpose and the landslide hazard map is presented in figure 9 as below.

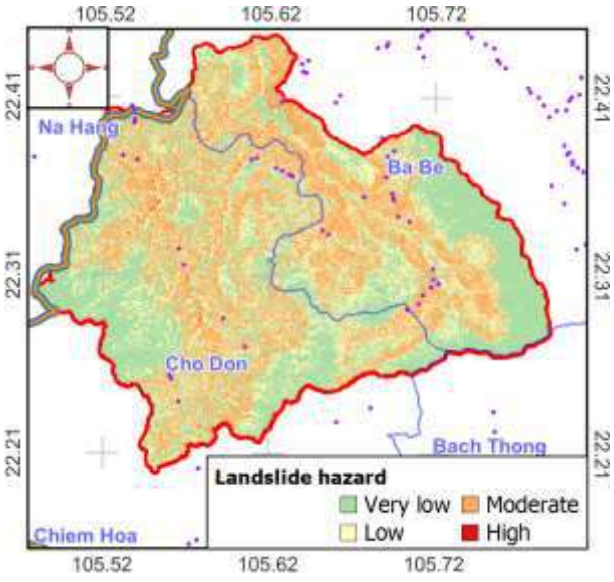


Figure 9: Landslide hazard mapping

The points with high hazard of landslides account for a very small part of the scattered distribution at the foot of the mountain, where there is a steep slope. This combination is similar to the principle of the RF model, which is a decision based on the votes of each model.

4. CONCLUSION

To build a landslide hazard map for Ba Be lake basin, 10 factors were utilized combining 4 machine learning models. The simulation results show that precipitation and slope are two decisive factors in the process of landslide hazard formation.

Each machine learning model has a different view of the importance of the factors, which creates different landslide hazard map products. Although the models all have good evaluation indicators, the LR and RF models are our two favorite models due to their agreement with the evaluation of many previous scientists.

This product could be useful to managers and scientists, contributing to the richness of the perception of landslide hazard on a large scale. Additionally, they could be useful to planners in land use planning and management.

ACKNOWLEDGEMENT

This is the original product of an independent national project: "Research on solutions to some major natural disasters causing natural instability in Ba Be lake area to serve the socio-economic development of the locality". We would like to thank the Ministry of Science and Technology as well as Bac Kan province for funding this research.

REFERENCES

[1] Parise, M. and Jibson, R.W., "A seismic landslide susceptibility rating of geologic units based on analysis of characterstics of landslides triggered by the 17 January, 1994 Northridge, California earthquake," *Engineering geology*, vol. 58, no. 3-4, p. 251–270, 2000, DOI: 10.1016/S0013-7952(00)00038-7.

[2] Le Quoc Hung, "The final report of the landslide status 1:50,000 scale in the mountainous area of Bac Kan province (in Vietnamese)," Monre, Hanoi, 2014.

[3] Aleotti P and Chowdhury R, "Landslide hazard assessment: summary review and new perspectives," *Bulletin of Engineering Geology and the Environment*, vol. 58, pp. 21-44, 1999, DOI:10.1007/s100640050066.

[4] J. Corominas, C. van Westen, P. Frattini, L. Cascini, J.-P. Malet, S. Fotopoulou, F. Catani, M. Van Den Eeckhaut, O. Mavrouli, F. Agliardi, K. Pitilakis, M. G. Winter, M. Pastor, S. Ferlisi, V. Tofani, J. Hervás and J. T. Smith, "Recommendations for the quantitative analysis of landslide risk," *Bulletin of Engineering Geology and the Environment*, vol. 73, pp. 209-263, 2014, DOI:10.1007/s10064-013-0538-8.

[5] Innocent C.Nnorom, UgochukwuEwuzie and Sunday O.Eze, "Multivariate statistical approach and water quality assessment of natural springs and other drinking water sources in Southeastern Nigeria," *Heliyon*, vol. 5, no. 1, p. e01123, 2019, DOI:10.1016/j.heliyon.2019.e01123.

[6] Elisa Arnone, Antonio Francipane, Leonardo V. Noto, Antonino Scarbaci and Goffredo La Loggia, "Strategies investigation in using artificial neural network for landslide susceptibility mapping: application to a Sicilian catchment," *Journal of Hydroinformatics*, vol. 16, no. 2, pp. 502-515, 2014, DOI:10.2166/hydro.2013.191.

[7] L.Bragagnolo, R.V. daSilva and J.M.V.Grzybowski, "Artificial neural network ensembles applied to the mapping of landslide susceptibility," *Catena*, vol. 184, no. 104240, 2020, DOI:10.1016/j.catena.2019.104240.

[8] Bin Li, Nianqin Wang and Jing Chen, "GIS-Based Landslide Susceptibility Mapping Using Information, Frequency Ratio, and Artificial Neural Network Methods in Qinghai Province, Northwestern China," *Advances in Civil Engineering*, vol. 2021, no. 4758062, 2021, DOI:10.1155/2021/4758062.

[9] "Landslides 101," USGS, [Online]. Available: https://www.usgs.gov/programs/landslide-hazards/landslides-101. [Accessed 15 09 2021].

[10] F.C Dai, C.F Lee and X.H Zhang, "GIS-based geo-environmental evaluation for urban land-use planning: a case study," *Engineering geology*, vol. 61, no. 4, pp. 257-271, 2001, DOI:10.1016/S0013-7952(01)00028-X.

[11] Keith Forbes and Jeremy Broadhead, "Forest and landslides: The role of trees and forests in the prevention of landslides and rehabilitation of landslide-affected areas in Asia," Food and Agriculture Organization of the United Nations Regional Office for Asia and the Pacific, Bangkok, 2013.

[12] Asmelash Abay, Giulio Barbieri and Kifle Woldearegay, "GIS-based Landslide Susceptibility Evaluation Using Analytical Hierarchy Process (AHP) Approach: The Case of Tarmaber District, Ethiopia," *Momona Ethiopian Journal of Science,* vol. 11, no. 1, pp. 14-36, 2019, DOI: 10.4314/mejs.v11i1.2.

[13] "Curvature function," ESRI, [Online]. Available: https://pro.arcgis.com/en/pro-app/latest/help/analysis/raster-functions/curvature-function.htm. [Accessed 13 01 2022].

[14] JarosławJasiewicz and Tomasz F.Stepinski, "Geomorphons — a pattern recognition approach to classification and mapping of landforms," *Elsevier B.V,* vol. 182, no. 15, pp. 147-156, 2013, DOI:10.1016/j.geomorph.2012.11.005.

[15] Samuele Segoni, Stefano Luigi Gariano and Ascanio Rosi, Rainfall Thresholds and Other Approaches for Landslide Prediction and Early Warning, Basel, Switzerland: MDPI, ISBN 978-3-0365-0931-0.

[16] Leo Breiman, "Bagging predictors," *Machine Learning,* vol. 24, no. 2, p. 123–140, 1996, DOI:10.1007/BF00058655.

[17] D. Richard Cutler, Thomas C. Edwards Jr, Karen H. Beard, Adele Cutler, Kyle T. Hess, Jacob Gibson and Joshua J. Lawler, "RANDOM FORESTS FOR CLASSIFICATION IN ECOLOGY," *Ecological Society of America,* vol. 88, no. 11, pp. 2783-2792, 2007, DOI:10.1890/07-0539.1.

[18] LI Z G, "Several Research on Random Forest Improvement," *Xiamen University,* 2013.

[19] Jeffrey Strickland, Data Analytics Using Open-Source tools, Lulu.com; First edition, 2016.

[20] Recep Can, Sultan Kocaman and Candan Gokceoglu, "A Comprehensive Assessment of XGBoost Algorithm for Landslide Susceptibility Mapping in the Upper Basin of Ataturk Dam, Turkey," *Applied Sciences,* vol. 11, no. 4493, 2021, DOI:10.3390/app11114993.

[21] Emrehan Kutlug Sahin, "Comparative analysis of gradient boosting algorithms for landslide susceptibility mapping," *Geocarto International,* vol. 37, no. 9, 2022, DOI:10.1080/10106049.2020.1831623.