

Optimizing Machine Learning Models for Landslide Susceptibility Mapping in Yen Bai Province, Vietnam

Tran, T. L.,^{1*} Nemoto, T.,¹ Dhar, S.,¹ Truong, X. Q.² and Raghavan, V.¹

¹Department of Geosciences, Osaka Metropolitan University, Japan,

E-mail: sp228721@st.omu.ac.jp*, t.tunglam.research@gmail.com*

²Vietnam National University, Hanoi School of Interdisciplinary Sciences and Arts, Hanoi, Vietnam

* Corresponding Author

DOI: <https://doi.org/10.52939/ijg.v22i9.5196>

Abstract

Landslide hazard assessment is crucial for mitigating landslide risks. Landslide Susceptibility Mapping (LSM) provides a practical and cost-effective tool for zoning areas prone to landslides. LSM expresses in the form of a probability of landslide occurrence in each pixel. This study implements Machine Learning (ML) models to generate Susceptibility Maps in Van Yen (VY). Landslides were mapped after landslide events occurred. The datasets were created by combining balanced numbers of landslide and non-landslide points with 17 contributing factors, including topographic, geological, hydrological, anthropogenic, and vegetation factors. To ensure all factors have the same range of values, landslide areas standardized using the Frequency Ratio (FR). The effectiveness of the method will be evaluated on the same dataset before and after applying FR. Four ML models Random Forest (RF), Support Vector Machine (SVM), Logistic Regression (LR), and Extreme Gradient Boosting (XGBoost) were trained in VY area, then validated on the dataset in Mu Cang Chai (MCC). Evaluation process was conducted using Accuracy Evaluation and Efficient Global Optimization (EGO). The results indicate that RF and XGBoost achieved the highest and most consistent performance along with the highest learning capacity. However, Geologic factors have showed to be low on impact values towards the model's performance, suggesting the need to include other critical geologic factors. The study can contribute to future Python-based frameworks for susceptibility analysis applicable to diverse topographic regions.

Keywords: Bayesian Optimization, Landslide, Machine Learning, Open-Source, Susceptibility

1. Introduction

Landslides are ubiquitous in any terrestrial environment with slopes, driven by tectonic. A single landslide event may include more than one slope failure triggered by the same phenomenon [1]. They can result from various factors such as high accumulation of rainfall, earthquakes, or volcanic activity [2]. The magnitude of the impact depends on the number of exposed elements and their associated vulnerabilities, the consequences of the impacts and the intensity of the landslide event [1]. Additionally, lithology, morphology, and external triggers such as prolonged heavy rain and seismic activity play a crucial role in landslide occurrences [2]. Vietnam's northern region, especially the Mu Cang Chai (MCC) and the Van Yen (VY) districts in Yen Bai province is highly susceptible to landslides due to rugged terrain, high rainfall, and deforestation. According to a study conducted by the Ministry of Resources and Environment of Vietnam, over 70% of communes in Yen Bai province face high or very high landslide risk, particularly in MCC and VY districts [3]. The

landslides may not be stopped or controlled; however, the losses can be reduced by establishing a decision support system to predict possible landslides or identifying landslide prone areas for management [4]. Landslide Susceptibility Map (LSM) is a crucial tool for assessing the likelihood of landslides in each pixel, based on local terrain conditions [2]. By categorizing regions from low to high susceptibility, LSMs assist planners in making informed land-use decisions and implementing hazard mitigation strategies. These maps consider factors like slope, soil type, and water flow impacts, aiming to pinpoint areas predisposed to landslides [5]. The mapping process involves analyzing the extent and triggers of landslides, such as topographic features, urbanization and vegetation cover, and documenting both historical and recent occurrences. Landslide susceptibility is mathematically defined as the probability of slope failure occurring under specific geo-environmental conditions, assuming that future landslides will occur under conditions

similar to past events [6]. By assuming that future landslides will occur under similar conditions as past events, susceptibility assessments help predict future landslide locations.

Machine Learning (ML) techniques have increasingly been used to solve scientific and engineering problems, notably in producing landslide susceptibility maps [7]. Among these techniques, decision-based tree ensemble methods like Random Forest (RF) and Extreme Gradient Boosting (XGBoost) have shown superior generalized performance compared to other models [8]. Landslide susceptibility assessments typically use a landslide inventory and various contributing factors. To increase prediction accuracy, it is crucial to identify the most important factors and eliminate less relevant ones, although this remains challenging without a universally accepted method [9]. Principal Component Analysis (PCA) is frequently used to reduce the dimensionality of contributing factors while preserving essential information, thus facilitating a clearer understanding of their relationships [10]. Additionally, the Pearson Correlation Coefficient helps identify the factors most associated with landslides and pinpoint redundant factors [9]. In the context of the research, it is used to evaluate how each landslide influencing factor correlates with each other and filter out the one with the least impact towards the prediction of the ML models.

Landslide influencing factors have been standardized using Frequency Ratio (FR) method. The FR method characterizes the spatial relationship between landslide distribution and environmental factors before ML modeling [11]. The FR serves to quantitatively represent the distinction between areas where landslides have occurred versus those where they have not, relative to a given factor's attribute. The method was used to determine the level of correlation between the location of the landslides and the eight landslide-related factors chosen for the study [12]. ML models coupled with FR preprocessing achieve higher prediction accuracy than models without considering these nonlinear connections [11]. The effectiveness of the method will be evaluated on the same dataset before and after applying FR.

Integrating statistical analysis with ML models offers a more effective approach to landslide susceptibility mapping (LSM). This method ensures that ML models are trained on a diverse set of different landslide-influencing factors while retaining only those with the most significant impact on predictions. As a result, both the accuracy and reliability of LSMs are improved. In summary, this research aims to achieve the following targets:

- Assess the performance of various ML models for landslide susceptibility mapping in northern Vietnam's mountainous regions, where human activities significantly influence topography.
- PCA and correlation analysis techniques to identify the most significant landslide-contributing factors within each category.
- Mitigate common biases in ML models that could result in the misvaluation of critical factors, enhancing prediction robustness and generalizability.
- Develop a complete ML framework for susceptibility mapping by integrating binary classification with statistical analysis.

2. Study area and Materials

2.1 Study Area

Van Yen (VY) district is a mountainous district located in the northern part of Yen Bai province, with geographical coordinates ranging from 104°20'17" to 104°47'38" East longitude and 21°39'57" to 22°12'12" North latitude (Figure 1). It is one of the most affected areas indicating high and very high potential landslide occurrences in the province [13]. This region covering an area of about 1389 km², features complex terrain characterized by hills and mountains within the Red River valley. The Con Voi Mountain range stands in the northeast of the district, with its highest peak reaching 1450 m. The district has diverse topographies, including high mountains, interlaced valleys, and a dense network of rivers and streams. In the northwest of the district, there are several moderately high mountains with rugged terrain and high slopes [14]. The VY district is influenced by a tropical monsoon climate with an annual average rainfall ranging from 2200 to 2400 mm, particularly concentrated during the rainy season from June to September annually [13].

2.2 Validation Area

The MCC district (Lat. 21°39'00"N–21°50'00"N; Long. 103°56'00"E–104°23'00"E) is selected as the study area. It is located in the northwest of Yen Bai province of Viet Nam covering an area of about 1196.47 km², as seen in Figure 2. The topography of the area is hilly with deep intervening valleys intervening. Mountains in the region occupy steep slopes up to 77 degrees at places where elevation of the area [14]. Its elevation varies from 315 to 2844 m with the average elevation of about 1515 m (above standard sea level). The area is occupied by igneous, sedimentary and metamorphic rocks. The volcanic extrusive rocks of Tu Le and Ngoi Thia and Tram Tau formation are predominant in the area.

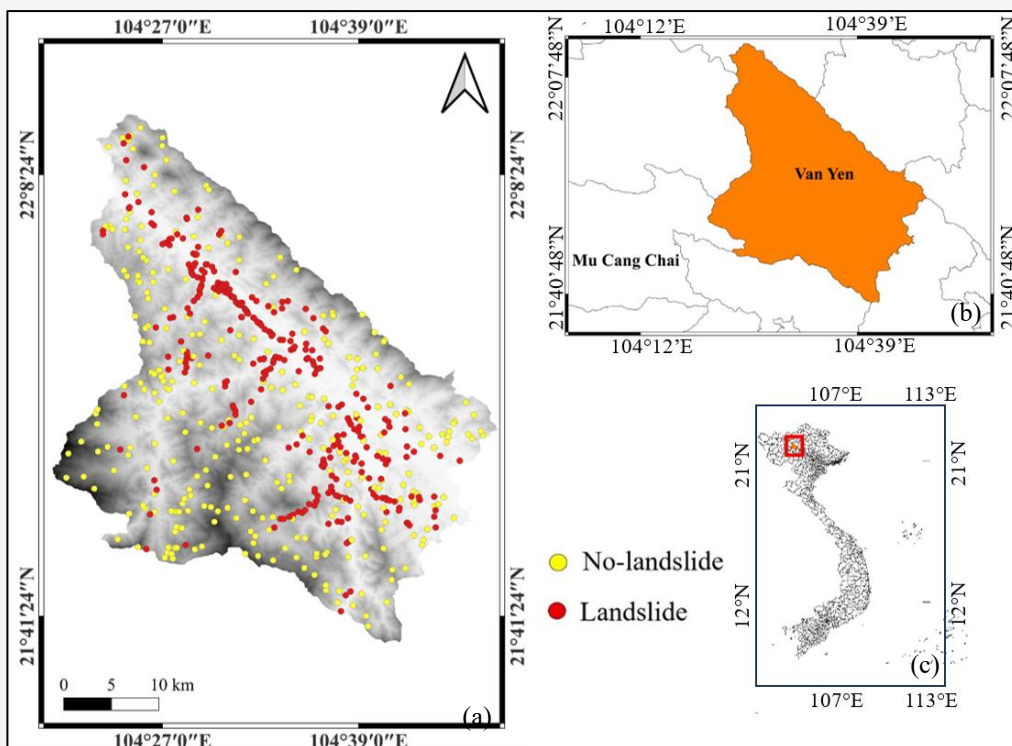


Figure 1: The selected study area: (a) landslide inventory, (b) Van Yen district, Yen Bai province, Viet Nam, and (c) the study area in northern Vietnam

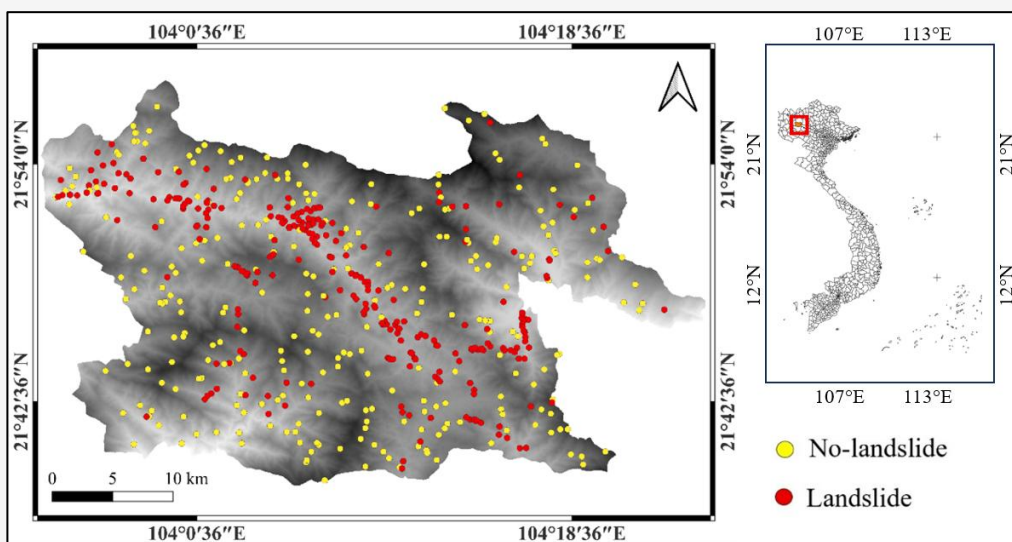


Figure 2: Validation area in Mu Cang Chai district

The area is situated in a tropical monsoon region having heavy rainfall (average 3700–5490 mm) during the months May– October. The annual mean humidity in this area is about 81 % and mean temperature 14.3 °C. Around 61.76 % of the study region is covered by forests [3].

2.3 Landslide Inventory

The landslide inventory used in Vietnam is produced by the project “Artificial Intelligence and Earth Observation Technologies for Landslide Studying in the Northern Mountain of Vietnam” by Hanoi University of Natural Resources and Environment [13] and [15].

The locations of landslides were recorded through conducted field surveys and satellite images. However, it met many challenges due to the high cloud cover rate in the area throughout the year, making it difficult to get a clear image of the landslide scarps. A map of landslides in VY and MCC was created, where sampling points are produced by drawing a point at the location of the identified landslides. Each landslide event is counted as a single landslide point on the inventory regardless of their size [16]. By treating all landslides equally, this selection aims to highlight the thresholds of at which a landslide is triggered rather than distinguishing different types of landslides [16]. The inventory includes information on the timing of landslide occurrences, though it does not specify their current activity status. The study utilizes landslide data from previous research, in VY (Study area) and MCC (Validation area), as seen in Figure 1 and Figure 2. The VY district was selected to validate the ML models due to its high landslide incidence and similar topographic condition to MCC district. Figure 1 shows the spatial database for the VY area, which includes 308 landslide points and 308 no-landslide points. Similarly, Figure 2 shows the dataset for MCC area, comprising 286 landslide points and 286 no-landslide points. Landslide points represent locations where landslides have occurred, and any area without a recorded landslide can be considered as a no-landslide point. Including both landslide points and no-landslide points is crucial for ML models to distinguish between the two classes [17]. Maintaining an equal number between landslide and no-landslide samples makes it easier to interpret landslide susceptibility without needing to adjust the data [16]. This balanced approach also prevents issues that can arise when one class is overrepresented, such as the model incorrectly identifying landslide-prone areas as safe because there are too many no-landslide points [17].

2.4 Contributing Factors

To evaluate landslide susceptibility, beside Topography being the main triggering factor, it is necessary to take account of other variables. Even though there is no universal standard for selecting landslide influencing factors, it is common practice to put these impact factors into internal and external factors [18]. Internal conditioning factors include topographic and geologic variables while external factors include meteorological factors and human's influence in the area [17]. Table 1 shows 17 contributing impact factors that were considered in this study, encompassing topographic, geologic,

hydrologic, anthropogenic, and vegetational data. These factors were derived from open data sources, existing databases, and previous research conducted in the area. Open-source Geographic Information System (GIS) software including QGIS (Quantum Geographic Information System) and GRASS GIS (Geographic Resources Analysis Support System). The topographic factors were generated from Digital Elevation Models (DEMs) obtained from the Advanced Land Observing Satellite-Phased Array L-band Synthetic Aperture Radar (ALOS-PALSAR) data.

2.4.1 Topographic factors

Digital Elevation Models (DEMs) are crucial for landslide events to trigger, along with Slope. Elevation has always been the first variable to be considered. Furthermore, DEM models are also used to calculate other contributing factors particularly topographic data. Many studies on landslide susceptibility assessments make use of Open-Source DEMs available for public access [19]. Changes in Elevation create drastic change in topography and in geomorphology, vegetation, stream system. It is the first triggering factor that alter landslide susceptibility value [20]. The DEM used in Vietnam was achieved from Advanced Land Observing Satellite-Phased Array L-band Synthetic Aperture Radar (ALOS-PALSAR) project, with the resolution of 12.5m. These DEMs are used to correct geometric and radiometric distortions in the data, making it more accurate and useful for scientific application [21]. Slope is a critical factor, with steeper slopes increasing shear stress leading to landslide susceptibility [19]. This is because steeper slopes demand greater resistance to maintain stability compared to gentler slopes, due to the balancing forces acting on the slope [2].

The aspect of terrain affects the direction of precipitation, sunlight, and wind. Therefore, it has effects on the growth of vegetation, rate of erosion, and thickness of soil. It is measured clockwise from 0° (due north) to 360° (again due north), coming full circle [19]. Curvature measures how fast a slope changes over time along specific directional planes relative to water flow and gravity. Profile Curvature measures the rate of change of slope along the downhill path of water flow. While Planform Curvature measures the rate of change of aspect perpendicular to the downhill direction. Negative curvature indicates divergent flow conditions and accelerated water flow, affecting landslide susceptibility [19] and [22].

Table 1: Overview of the landslide impact factors

Category	Factors/Variables	Type	Spatial Resolution (m)	Source
Topographic	Elevation	Continuous	12.5	ALOS-PALSAR
	Slope	Continuous	12.5	ALOS-PALSAR
	Aspect	Continuous	12.5	ALOS-PALSAR
	Topographic Wetness Index (TWI)	Continuous	12.5	ALOS-PALSAR
	Profile Curvature	Continuous	12.5	ALOS-PALSAR
	Planform Curvature	Continuous	12.5	ALOS-PALSAR
	Terrain Ruggedness Index (TRI)	Continuous	12.5	ALOS-PALSAR
	Roughness	Continuous	12.5	ALOS-PALSAR
Geologic	Faults	Categorical	12.5	[13]
	Lithology	Categorical	12.5	[13]
	Lithologic Boundaries	Categorical	12.5	[13]
Hydrologic	Stream Power Index (SPI)	Continuous	12.5	ALOS-PALSAR
	Distance from rivers	Continuous	12.5	[13]
Anthropogenic	Distance from roads	Continuous	12.5	[13]
	Land Use Land Cover (LULC)	Continuous	12.5	[13]
Vegetation	NDVI	Continuous	10.0	Sentinel-2
	NDWI	Continuous	10.0	Sentinel-2

The Terrain Ruggedness Index (TRI) is the measurement of surface irregularity by calculating the root-mean-square elevation difference between a center pixel and all 8 surrounding neighboring cells. High TRI values indicate highly dissected terrain where sharp slope and rapid water runoff increase landslide susceptibility [22]. Similarly, Roughness indicates local elevation contrast, calculated strictly as the absolute difference between the maximum and minimum elevation measured in a group of 3x3 pixels [23]. High roughness values mean steep slope breaks [22]. The Topographic Wetness Index (TWI) is used to quantify the spatial pattern of surface water accumulation and soil moisture on terrain geometry, therefore, it is inversely related to landslide susceptibility [19], as defined in Equation 1:

$$TWI = \ln\left(\frac{A}{\tan \alpha}\right)$$

Equation 1

Where: A = Area of a specific catchment and α = Slope gradient of the specific area.

2.4.2 Hydrologic factors

The Stream Power Index (SPI) represents the erosive force of water flows (rivers and streams) across terrain. SPI value increases the steeper the slope and is low alluvial plain, as calculated in Equation 2 [24]:

$$SPI = \ln(A \cdot \tan \alpha)$$

Equation 2

Where: A = Area of a specific catchment, and α = Slope gradient of the specific area.

Distance from rivers (River buffer) is a key hydrological and geomorphological conditioning factor used in landslide susceptibility mapping. It measures how close a specific piece of land is to the nearest stream or river channel, since water erosion is one of the main factors that trigger landslide in this area [25].

2.4.3 Geological factors

Northern Vietnam has undergone significant deformation due to the collision of the Indo-China block with the South China block and the Indian subcontinent. The geology of MCC is dominated by fine-grained igneous rocks such as rhyolite and porphyritic rhyolite, while VY features a mix of igneous and sedimentary rocks, including quartz-sericite schist, quartz-mica schist, and quartzitic sandstone. Landslides in VY frequently occur in unconsolidated Quaternary deposits and sedimentary rocks, particularly near the boundaries of harder formations like metamorphic rocks and schists. The geological map was utilized from previous research about landslide on the study area [13]. Due to the complex geological characteristics of the study area, the geological map was classified into different geologic groups (Each contains different geologic formations).

Geological units that covered less than 10% of the study area were excluded, as their contribution to landslide risk was considered negligible, and this allowed for a clearer visualization of the significant geological features. Upon inspection, several landslides also occurred near geological boundaries, in MCC, around 27% of landslides are within 500 meters of a geological boundary and for VY area, it's approximately 29%.

It is further noted that the scale of landslides tends to decrease with distance from these boundaries, with small to medium-scale landslides occurring approximately 2-3 km away. Hence why the buffer map for geological boundary was constructed for the analysis. Fault line is a structural geologic factor that measures how close a given area is to tectonic fault. It evaluates the proximity to tectonic disturbance that can influence slope instability [26]. Additionally, a fault buffer map was created from geological data to assess the impact of geological faults on landslide occurrences.

2.4.4 Geological Factors

Distance from the road (Road buffer) is an anthropogenic impact factor that evaluates the how close any given pixel in the area to the nearest road network [18]. Distance from road networks is one of the most critical factors as constructing roads and facilities in steeped mountainous terrain requires cutting through existing slopes, therefore, heavily alter the terrain topography [18] and [25]. Land Use/Land Cover (LULC) is also used in this study. The form of land use is an important index in inconstancy of the slopes that affect the earth's traits and results in changes in its action [26].

2.4.5 Vegetation factors

Generally, the probability of the occurrence of the landslide on the naturally vegetated surface is lower than the bare lands [19]. The spectral indices are computed using the Visible band (RED, GREEN, BLUE) or RGB images and Near infrared (NIR) band

image from Sentinel-2. Equation 3 shows the Normalized Difference Vegetation Index (NDVI), which quantifies vegetation greenness and ground stability, as vegetation acts as a natural stabilizer against slope failure. It utilizes radiances or reflectance from a red channel around 0.66 μm and a NIR channel around 0.86 μm [27]:

$$NDVI = \frac{NIR - RED}{NIR + RED} \quad \text{Equation 3}$$

Normalized Difference Water Index (NDWI) is used to highlight open water features in a satellite image, allowing a water body to "stand out" against the soil and vegetation. This index utilizes radiances or reflectance from a green channel around 0.55 μm and a NIR channel around 0.86 μm , as seen in Equation 4 [28]:

$$NDWI = \frac{GREEN - NIR}{GREEN + NIR} \quad \text{Equation 4}$$

3. Methodology

Figure 3 shows the processing framework of this study. The methodology begins with defining the problem as a binary classification task (Landslide/No-landslide) [18]. After the landslide contributing factors are selected, these factors are combined with the landslide inventory to form the sampling points suitable for binary classification task.

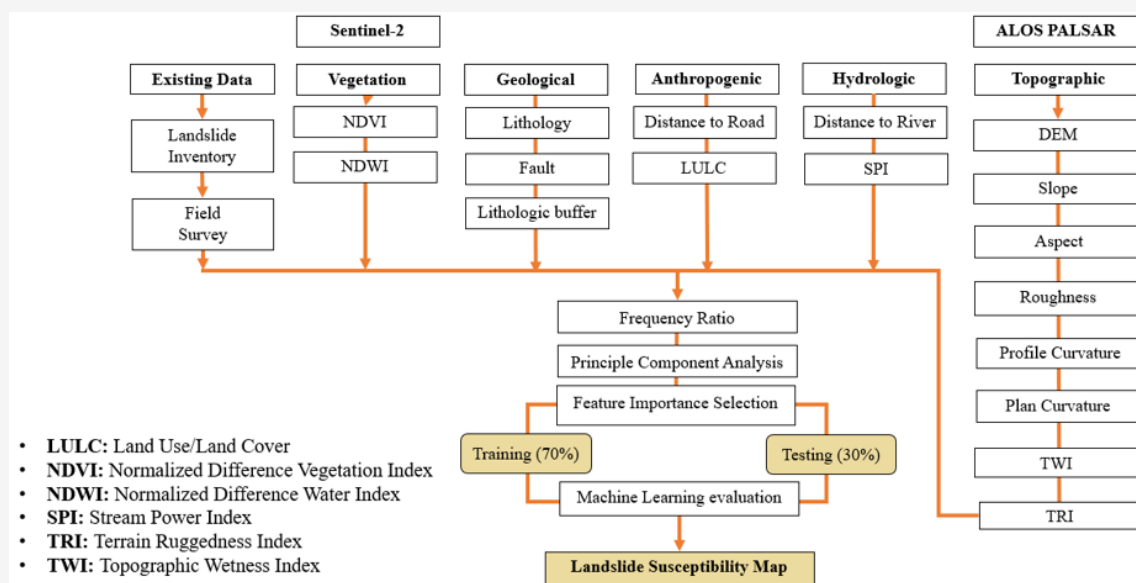


Figure 3: Machine Learning framework

3.1 PyGILE-PLUS

The Machine Learning tasks were performed on Python GeoInformatics Lab Environment-Plus (PyGILE-Plus) environment. It is a comprehensive, headless Docker environment for geospatial research with algorithms across multiple GIS platforms plus a complete Python geospatial stack [29]. It comes with pre-configured Jupyter notebooks and libraries for geospatial data analysis.

3.2 Frequency Ratio

In general, to predict landslides, it is necessary to assume that landslide occurrence is determined by landslide-related factors, and that future landslides will occur under the same conditions as past landslides. On this basis, the relationship between areas where a landslide has occurred and landslide-related factors can be distinguished from the relationship between areas without past landslides and landslide-related factors [12]. The frequency ratio is the ratio of the area where landslides occurred in the total study area, and also, is the ratio of the probability of a landslide occurrence to a non-occurrence for a given attribute [12]. The FR value is calculated using Equation 5 [30]:

$$FR = \frac{\sum \frac{Npix(1)}{Npix(2)}}{\sum \frac{Npix(3)}{Npix(4)}} \quad \text{Equation 5}$$

Where:

$Npix(1)$ is the number of landslide pixels of the factor class

$Npix(2)$ is the total number of pixels of the sub-class over the entire study area,

$Npix(3)$ is the total number of landslide pixels of the study area, and

$Npix(4)$ is the total number of pixels of the study area.

17 factors were standardized using FR method as in Table 2, Table 3 and Table 4. Table 2 contains topographic factors that were derived from DEM. Table 3 includes vegetation cover layers and geologic data. Table 4 contains Anthropogenic factors such as LULC and a buffer map of the road system in the area.

3.3 Factor Analysis

3.3.1 principal component analysis

PCA is an unsupervised multivariate technique used to analyze datasets composed of inter-correlated quantitative variables [31]. It transforms these original variables into a reduced set of linearly

uncorrelated variables known as principal components (PCs) [32]. Since the variables do not have the same units and scales, all the variables were first standardized, prior to analysis by centering to a zero mean of 0 and scaling to unit variance of 1 [18] and [32]. The chart in Figure 4 was plotted using the first two dimensions PC1 and PC2. The PCA resulted in 17 principal components, collectively accounting for 100% of the total variance in the dataset. The first three components PC1, PC2, and PC3 accounted for approximately 20%, 16%, and 10% of the data variability respectively, and so on. As shown in Figure 4, the first 12 principal components are already sufficient to cover 91% of the dataset's variability

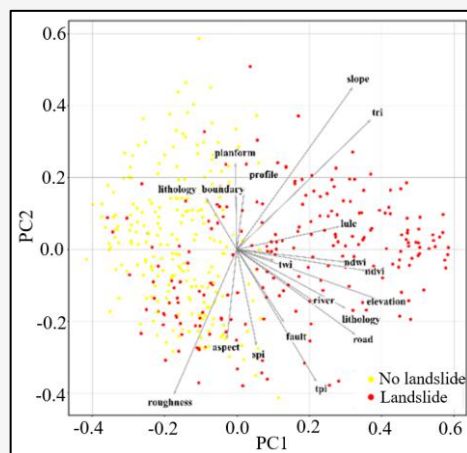


Figure 4: PCA loading plot shows how much each factor affects a PC

3.3.2 Feature Selection

To reduce the dimensionality of the dataset and improve model performance, PCA and Pearson Correlation Coefficients were used to evaluate the weight of each contributing factor and eliminate correlated ones. The correlated factor with the lowest weight value was removed from the dataset without affecting the performance of the classification models performed later. Factors with correlations exceeding ± 0.75 were excluded from the dataset. In Figure 5, TRI and Slope have the correlation value of 0.88, therefore, TRI was removed from the dataset, as Slope is the core factor influencing landslide activities [13]. Feature Importance analysis, a function of the RF model in the Python environment, was deployed to evaluate the impact each of the remaining factors in the new dataset has towards the prediction of RF model (Figure 6). The graph showed Road factor (road buffer) has the highest influence towards landslide. This is aligned with The PCA result which showed the road factor has high impact on both PC1 and PC2 (Figure 4).

Table 2: Topographic impacts factors

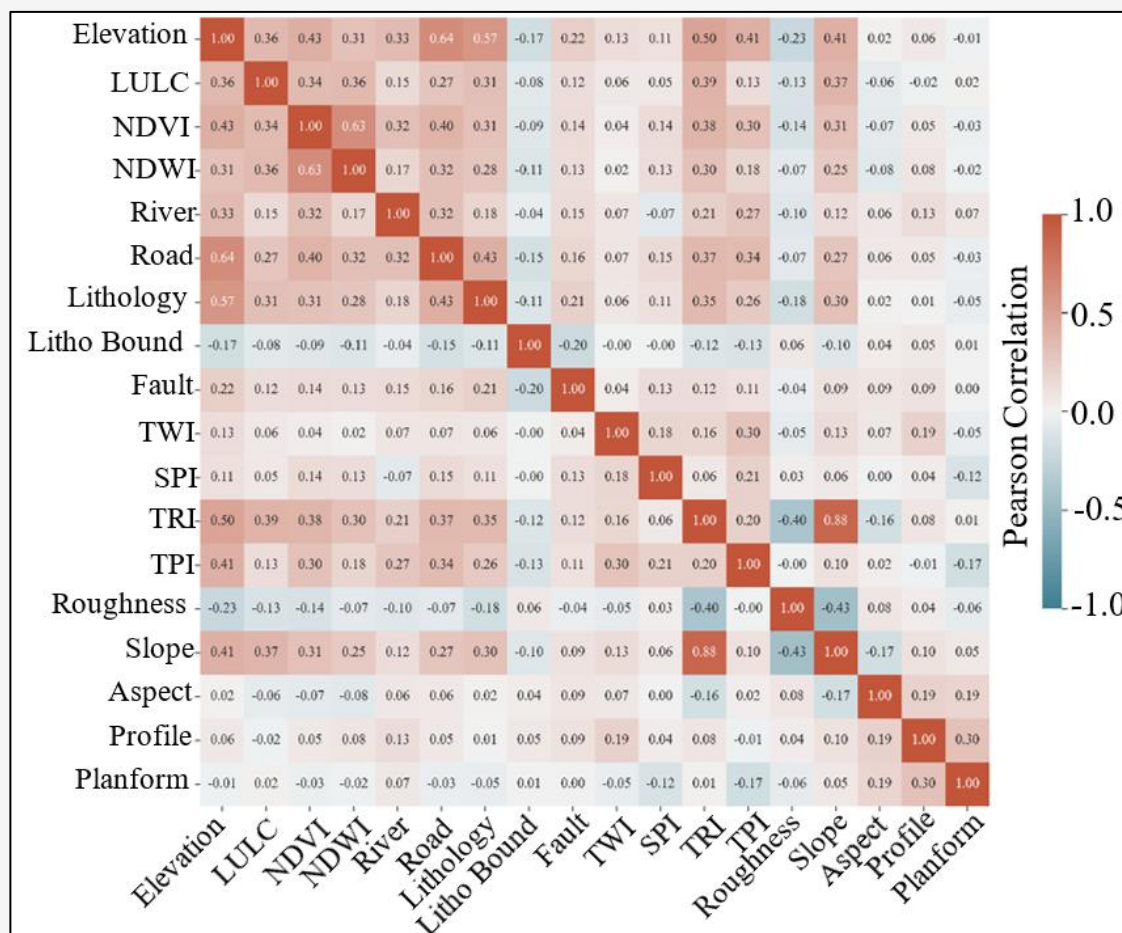
Factor	Class	LS point	Landslide (%)	Class pixel	Class pixel (%)	FR
Elevation (m)	-2 to 130	197	63.96	1514,047	16.95	3.77
	130 to 350	76	24.68	2380,916	26.65	0.93
	350 to 600	21	6.82	2442,084	27.33	0.25
	> 600	14	4.55	2597,028	29.07	0.16
Aspect (Degree)	0.71 to 72	79	25.65	1,809,162	20.52	1.25
	72 to 143	67	21.75	1,707,399	19.37	1.12
	143 to 215	66	21.43	1,789,605	20.30	1.06
	216 to 287	44	14.29	1,815,487	20.59	0.69
	287 to 360	52	16.88	1,693,688	19.21	0.88
Slope (Degree)	0 to 10	123	39.94	1,069,835	11.97	3.33
	10 to 20	111	36.04	2,050,280	22.95	1.57
	20 to 30	57	18.51	4,285,861	47.97	0.39
	30 to 80	17	5.52	1,528,099	17.10	0.32
Roughness	0 to 5	100	32.47	1,960,234	25.75	1.26
	5 to 10	100	32.47	4,062,669	53.36	0.61
	10 to 20	91	29.55	1,383,385	18.17	1.63
	20 to 181	17	5.52	207,683	2.73	2.02
Planform Curvature	-1.60 to -0.01	98	31.82	2,532,245	28.35	1.12
	-0.01 to 0.01	88	28.57	3,066,319	34.33	0.83
	0.01 to 0.05	122	39.61	3,332,837	37.32	1.06
Profile Curvature	-0.13 to -0.01	73	23.70	1,989,539	22.28	1.06
	-0.01 to 0.01	174	56.49	5,166,288	57.84	0.98
	0.005 to 0.18	61	19.81	1,775,574	19.88	1.00
TRI	0 to 5	110	35.71	1,171,868	13.12	2.72
	5 to 10	102	33.12	2,248,751	25.17	1.32
	10 to 20	85	27.60	4,273,203	47.83	0.58
	20 to 163.15	11	3.57	1,240,252	13.88	0.26
TWI	3.71 to 7.00	35	11.36	1,605,644	17.97	0.63
	7.00 to 9.00	217	70.45	5,561,051	62.25	1.13
	9.00 to 20.66	56	18.18	1,767,380	19.78	0.92
SPI	0 to 400	173	56.17	4,563,506	51.50	1.09
	400 to 1,500	122	39.61	3,083,248	34.80	1.14
	> 1,500	13	4.22	1,213,845	13.70	0.31

Table 3: Natural impact factors

Factor	Class	LS point	Landslide (%)	Class pixel	Class pixel (%)	FR
NDVI	-0.34 to 0.25	99	32.14	1,199,747	13.43	2.39
	0.25 to 0.5	192	62.34	5,229,073	58.53	1.07
	0.5 to 0.6	17	5.52	2,505,254	28.04	0.20
NDWI	-1.0 to -0.8	53	17.21	3,354,996	37.55	0.46
	-0.8 to -0.6	126	40.91	3,570,459	39.96	1.02
	-0.6	129	41.88	2,008,619	22.48	1.86
Rivers distance (m)	0 to 50	72	23.38	1,664,553	18.63	1.25
	50 to 100	79	25.65	1,623,769	18.18	1.41
	100 to 300	151	49.03	4,353,760	48.73	1.01
	300 to 1,036	6	1.95	1,291,991	14.46	0.13
Faults distance (m)	0 to 100	34	11.04	718,241	8.04	1.37
	100 to 300	64	20.78	1,100,654	12.32	1.69
	300 to 500	46	14.94	1,009,538	11.30	1.32
	500 to 7,761.98	164	53.25	6,105,641	68.34	0.78
Lithology boundary (m)	0 to 50	7.14	396,733	4.44	1.61	7.14
	50 to 100	5.84	362,711	46.84	0.12	5.84
	100 to 200	10.39	694,689	8.37	1.24	10.39
	>300	76.62	7,479,942	4.44	17.25	76.62
Lithology	Quaternary	20	6.49	607,264	10.80	0.60
	Cam Duong Formation	18	5.84	293,871	5.23	1.12
	Co Phuc Formation	3	0.97	213,544	3.80	0.26
	Tu Le – Ngoi Thia Formation	11	3.57	600,675	10.68	0.33
	Xom Giau Formation	1	0.32	153,919	2.74	0.12
	Sa Pa Formation	1	0.32	129,307	2.30	0.14
	Ngoi Chi Formation	37	12.01	1,311,869	23.33	0.51
	Nam Thap Formation	90	29.22	509,817	9.07	3.22
	Sinh Quyen Formation	76	24.68	3,375,646	60.03	0.41
	Suoi Bang Formation	3	0.97	606,472	10.78	0.09
	Tram Tau Formation	48	15.58	1,131,728	20.12	0.77

Table 4: Anthropogenic impact factors

Factor	Class	LS point	Landslide (%)	Class pixel	Class pixel (%)	FR
LULC	Water	10	3.25	187,679	2.10	1.55
	Build-up	21	6.82	136,303	1.53	4.47
	Paddy-field/ Terrace	74	24.03	745,923	8.35	2.88
	Non-forest vegetation	81	26.30	2,931,550	32.81	0.80
	Forest	80	25.97	4,184,314	46.84	0.55
	Bareland/Others	42	13.64	748,044	8.37	1.63
Roads distance (m)	0 to 100	173	56.17	1,407,751	15.76	3.56
	100 to 300	117	37.99	1,796,557	20.11	1.89
	300 to 500	8	2.60	1,161,427	13.00	0.20
	500 to 5,863.15	10	3.25	4,568,339	51.13	0.06

**Figure 1:** Pearson correlation Matrix between landslide influence factors

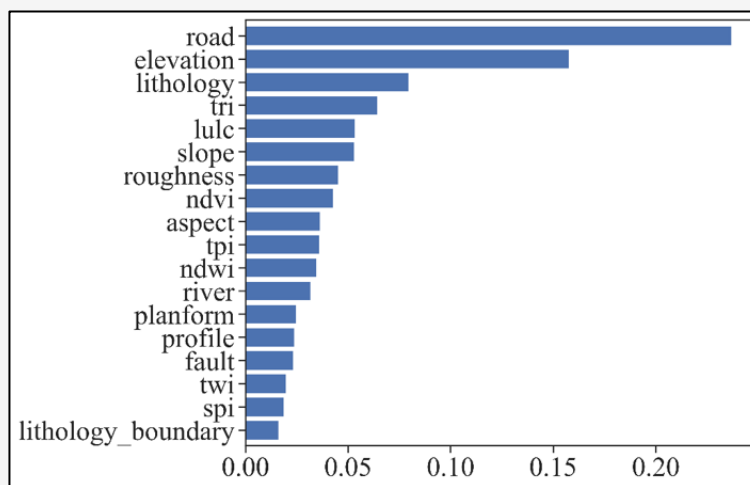


Figure 6: Feature Importance histogram

3.4 Tuning Machine Learning Models

3.4.1 Hyperparameter tuning and 10-fold validation

Before the ML models can be trained, they need to be tuned. Hyperparameters are the external settings of a machine learning algorithm that you configure before training the model. The K-Fold Validation splits the training dataset into a number of subsets (k subsets), iteratively, some of them are used to learn the model, while the others are exploited to assess its performance [18]. This calculation is part of the “Sklearn.model_selection” function from the Scikit-learn library. After the training steps are completed, the now trained ML models can apply what they have learnt on the validation area without the need to repeat the same processes again. K-Fold Validation was applied to the training dataset, comprising 70% of the total dataset. The training dataset is split into 10 equal subsets of data (10-fold validation) to evaluate if they perform well to unseen data before starting training the ML models. In each iteration, 9 subsets were used for training, while the remaining subsets were chosen for validation. This process was repeated subsequently across 10 folds, ensuring each subset was used as the validation set only once. The best-performing hyperparameter set from the 10-fold validation was then utilized to train the ML models using full training data set. The hyperparameters were then saved in the Grid Search. Table 5 showed the most optimal parameters for each ML model.

3.4.2 Random Forest

The RF model is a supervised learning ensemble classifier that constructs multiple decision trees to form a “Decision Forest”. The mode is well known for its accuracy in binary classification tasks. Each tree votes on the predicted outcome, and the majority vote determines the final result [33]. RF is computationally efficient, with shorter training times

compared to other supervised methods, making it an attractive option. Its strong predictive performance has proven particularly effective for LSM [7] and [34].

3.4.3 Support vector machine

Support Vector Machine (SVM) is a well-established supervised learning algorithm that excels across various applications. It learns from training data and predicts outcomes by separating data into categories using hyperplanes. SVM maximizes the margin between the hyperplane and the nearest training data points on either side [32]. SVM was selected, as its kernel tricks provide a unique solution for complex problems, though the parameter selection can be intensive in terms of computation [16].

3.4.4 Logistic regression

Logistic Regression (LR) is a widely used classification algorithm for binary data, where data points belong to one of two classes. Although it is fundamentally a linear regression, LR employs the sigmoid (logistic) function as its cost function. This algorithm is one of the most used methods for landslide susceptibility modeling at the regional scale [16] and [35].

3.4.5 Extreme gradient boosting

Gradient Boosting involves creating a sequence of weak models that address each other's weaknesses [8]. Each model focuses on correcting errors from the previous ones. Key steps include selecting a suitable differentiable loss function for the problem. An advantage of Gradient Boosting is its flexibility to use various loss functions without needing new algorithms [14]. XGBoost is an advanced supervisory algorithm optimized based on Gradient Boosting Decision Tree [36].

Table 5: Hyperparameters used in the ML models

Model	Parameter	Value
Random Forest	Number of estimators	30
	Maximum Features	Square root
Support Vector Machine	Gamma	0.01
	C (Cost)	10
	Kernel	RBF
Logistic Regression	C (Cost)	0.001
	Penalty	L2
	Solver	Liblinear
Extreme Gradient Boosting	Learning Rate	0.01
	Maximum depth	7.00
	Number of estimators	100.00
	Gamma	0.50
	Subsample	1.00
	Colsample by Tree	1.00
	Regularization Alpha	0.00
	Regularization Lambda	5.00
	Scale Positive Weight	1.00

Table 6: Accuracy metrics of all models in study area

Models	Accuracy	Kappa	Sensitivity	Specificity
Before standardization				
RF	0.83	0.66	0.76	0.90
SVM	0.88	0.76	0.84	0.92
LR	0.81	0.62	0.72	0.94
XGBoost	0.84	0.68	0.75	0.92
After standardization				
RF	0.83	0.66	0.79	0.87
SVM	0.82	0.63	0.74	0.89
LR	0.82	0.63	0.70	0.94
XGBoost	0.84	0.69	0.87	0.82

3.5 Model training

All four ML models (RF, XGBoost, SVM, LR) used in this study are implemented using the scikit-learn library's "sklearn.model_selection" module, under the names "RandomForestClassifier," "XGBClassifier," "SVC" and "LogisticRegression". The four models are all classifier, while LR is classifier that uses direct regression coefficient through the sigmoid function. While the codes for RF, SVM, and LR models are open for public access from [18], the code for XGBoost model is developed specifically for this research. The code is available at [37] along with the modified and original code from the author [18]. An accuracy assessment was performed to evaluate the predictive quality and classification effectiveness of each model. Table 6 shows the performance of four ML models before and after standardization, which were evaluated using the Confusion Matrix and the accuracy metrics in VY. The accuracy metrics comprise of 2 major indexes: Accuracy Score and Kappa Score, which can be plotted using the

"Sklearn.metrics" function from the Scikit-learn library.

3.5.1 Accuracy Evaluation

One of the methods of conducting accuracy assessments used in this study is by producing a confusion matrix in order to compare the ground truth with the results from the classification [38]. In this context, a True Positive (TP) occurs when a model correctly identifies an actual landslide, while a True Negative (TN) occurs when the model correctly identifies a no-landslide area. A False Positive (FP) happens when a no-landslide area is incorrectly labeled as a landslide by the model, and a False Negative (FN) occurs when the model fails to detect an actual landslide. "In natural hazard evaluation, False Negatives (FNs) represent the most critical error (the 'worst-case scenario'), as unstable or landslide-prone slopes that are incorrectly classified as safe. To calculate how many errors of this type that the model makes, an error matrix (confusion matrix) was deployed to derive standard

accuracy metrics, including Overall Accuracy (OA), showing the proportion of all the samples that were correctly classified, as seen in Equation 6 [38]:

$$OA = \frac{TP + FP}{TP + TN + FP + FN} \quad \text{Equation 6}$$

The second metric is Cohen's Kappa (κ), it is a measure of the percentage of right answers made purely by chance or by predicting the most common category. A high kappa score means there is a high amount correct predictions that is not a result of coincidence. Kappa gives you a score between 0 and 1, but the value can it be negative suggesting there is systematic disagreement. Equation 7 shows Cohen's Kappa formula [38]:

$$\kappa = \frac{OA - CA}{1 - CA} \quad \text{Equation 7}$$

Where:

OA is the overall accuracy of the model
 CA is the chance agreement

The Receiver Operating Characteristic (ROC) curve measures the performance of classification problems at different threshold settings. The ROC is a probability curve for binary classification problems, and the Area Under the Curve (AUC) indicates ability to distinguish between landslide and non-landslide instances. Higher AUC values mean the model better is at predicting the correct classes [39]. An AUC of 1.0 representing perfect classification and 0.5 indicating performance equivalent to random chance. The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR). In Equation 8, the TPR, also called sensitivity, measures the proportion of correctly identified positive (true positives) in relation to all actual positive events (true positives and false negatives) [39]:

$$Sensitivity = \frac{TP}{TP + FN} \quad \text{Equation 8}$$

Specificity, or the True Negative Rate (TNR), reflects the proportion of actual negative instances specifically, stable non-landslide locations that are correctly identified by the model by using Equation 9 (True Negatives and False Positives):

$$Specificity = \frac{TN}{TN + FP} \quad \text{Equation 9}$$

The FPR used to plot the ROC curve is calculated using Equation 10:

$$FPR = 1 - Specificity \quad \text{Equation 10}$$

3.5.2 Convergence rates of efficient global optimization

The Convergence Rate of Efficient Global Optimization (EGO) defines the mathematically guaranteed speed at which an EGO strategy approaches the true minimum of the unknown function. In another way, it is the mathematical measure of how fast an EGO model can achieve the most optimal results as more combinations are performed. The most popular methods for solving the efficient global optimization problem are expected-improvement algorithms. This approach is often called Bayesian Optimization (BO). When plotted, it compares how fast and how well each model achieves its best ROC-AUC score during global (Bayesian) optimization, in short, its learning capacity. Efficient Global Optimization places a Gaussian process minimizing a continuous, as shown in Equation 11 [40]:

$$f : X \rightarrow R^d \quad \text{Equation 11}$$

Where: X is a compact subset of R^d

The Gaussian process provides a Bayesian posterior probability distribution that describes potential values for $f(x)$ at a candidate point x [40]. The core selection formula in Bayesian Optimization refers to the acquisition function, which determines where to sample next by balancing exploration and exploitation. The most common acquisition function is Expected Improvement (EI). It maximizes the expected improvement. Equation 12 defines it as the expected positive difference between the function value $f(x)$ and f_*^n [41]:

$$EI_n(x) = E_n \left[\left(f_*^n - f(x) \right)^+ \right] \quad \text{Equation 12}$$

The expected improvement algorithm selects the next design point by maximizing the EI. This algorithm was calculated using Equation 13 [42]:

$$x_{n+1} = \arg \max EI_x(x) \quad \text{Equation 13}$$

Where:

EI is the Expected Improvement
 f_*^n is the best value so far

$F(x)$ is the value the function might take at the candidate point x

$(f_*^n - f(x))^+$ is positive improvement:

If $f(x)$ is better (e.g., lower for minimization) than f_*^n , we get a positive number.

4. Results and Discussion

4.1 Model Performance

Table 6 compared the performance of all four ML models before and after applying FR standardization on the dataset. All models were evaluated using confusion matrix and accuracy metrics. The average accuracy score of all models was 0.83. Before standardization, SVM achieved the highest accuracy score at 0.88, however, it was reduced to 0.82 after the model was standardized. Following standardization, XGBoost emerged at the top with the highest accuracy score of 0.84, with RF slightly behind, maintaining a consistent accuracy score of 0.83 across both scenarios. Kappa score on average is within the range of 0.61–0.80, which can be classified as having substantial performance beyond coincidence, indicating generally good classification performance.

Combining with high accuracy, it shows that four ML models have learned the patterns of both classes well. The range of Sensitivity values from 0.70 to 0.87 for all models indicating the models are effective at classifying positive class (landslide points) both before and after standardization. Specificity values range from 0.82 to 0.94 across all models, showing a good ability to recognize negatives accurately, however, there are slight decreases after standardization with RF and SVM, while XGBoost lost from 0.92 to 0.82. No model showed trade-off values (e.g., very high sensitivity with very low specificity, or high accuracy with low Kappa) indicating overfitting did not happen. The trained models were validated on the MCC area to ensure their generalizability. The validation results

showed low accuracy assessment. This result happened due to the ML models being trained in another area. Even though the two areas have similar topography, it was proven that this was not enough to prevent bias in the ML model. Accuracy score assessment following a similar pattern similar with XGBoost and RF model achieving the highest score.

Before standardization, the models performed poorly. Table 7 showed the assessment of four ML models in the validation area. The average accuracy score was 0.55. Furthermore, average Kappa scores for all ML models were below 0.20 indicating the many of the correctly classified results were randomly guessed. Most of the accuracy points are from over generalization of No-landslide class.

XGBoost experiences the most significant boost, achieving the highest Accuracy (0.74) and Kappa Score (0.48), along with a strong Sensitivity (0.86). RF and LR drastically improved classification stability. Their initial accuracy points jumped from 0.57 and 0.54 to 0.67, and Kappa scores jumped to around 0.34. However, SVM model experienced a slight decrease in performance. Before standardization, SVM had high Specificity value at 0.99 due to over generalization of the No-landslide class. After standardization, it started doing the same thing to the Landslide class (Sensitivity = 0.98), further decreasing Kappa score to 0.03. Among all models being validated with MCC data, SVM by far had the lowest performance. The overall accuracy and Kappa score of the validation on the external area indicates that the SVM and LR models do not generalize well on the unseen data. However, FR method did show improvements across all metrics, and increased stability and consistent performance between different models on the same dataset. The ROC value measures how well the model distinguishes between the two classes. Before applying FR, there was a slight bump in the performance of SVM. This also happened with AUC value, as seen in Figure 7(a).

Table 7: Accuracy metrics of all models in validation area

Models	Accuracy	Kappa	Sensitivity	Specificity
Before standardization				
RF	0.57	0.14	0.24	0.89
SVM	0.53	0.06	0.06	0.99
LR	0.54	0.08	0.10	0.98
XGBoost	0.57	0.14	0.25	0.89
After standardization				
RF	0.67	0.34	0.54	0.80
SVM	0.52	0.03	0.98	0.05
LR	0.67	0.35	0.54	0.80
XGBoost	0.74	0.48	0.86	0.62

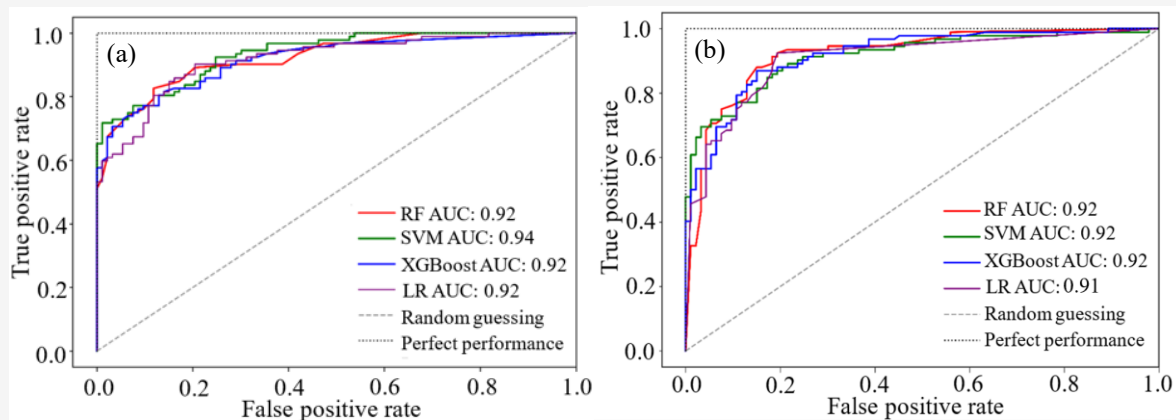


Figure 7: ROC curves for the ML models (a) before applying standardization with FR, and (b) after applying standardization with FR

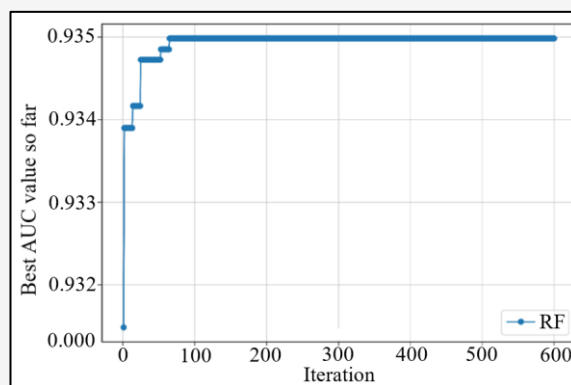


Figure 8: Convergence rate of RF model using Bayesian Optimization

After the standardization, the AUC value of SVM reduced to 0.92, the same as RF and XGBoost, while LR slightly jumped down to 0.91 from 0.92. The differences in accuracy before and after standardization are minor, indicating stable model behavior across all models under both conditions with RF and XGBoost performed comparatively well, as seen in Figure 7(b). The RF model converged at an AUC of 0.934 after 700 iterations (Figure 8). Beyond this point, additional trees produced little improvement in AUC, indicating that 700 trees were sufficient to achieve stable predictive performance.

4.3 Landslide Susceptibility Mapping

The susceptibility maps are visualized using equal interval classes to ensure comparability across all maps [44]. These maps represent the probability of each pixel experiencing a landslide, with susceptibility categorized into five risk levels: Very Low, Low, Medium, High, and Very High. A color gradient (blue, green, yellow, orange, and red) is used to illustrate these levels, following expert judgment criteria established in previous research [13]. To estimate probabilities, the "Predict_proba" function from the scikit-learn library is applied, determining

the likelihood of each pixel being classified as either "Landslide" or "No-landslide." These predictions are based on learned relationships between landslide factors (e.g., topography, hydrology, vegetation) and past occurrences. Each ML model generates a distinct LSM, where color-coded pixels indicate varying susceptibility levels. The resulting LSMs for the study area and validation area are plotted in the Jupyter Notebook environment, with Figures 9 and 10 displaying the study area maps. A gradient from blue (Very Low) to red (Very High) visually represents the increasing landslide risk. As seen in Figure 9, the LSM of RF and XGBoost both show similarity with concentrated red patterns and mostly blue backgrounds with clear boundaries, while RF expressed the transition between different risk levels, XGBoost drew sharp boundaries between these areas. RF model tends to highlight more localized high-susceptibility patches, reflecting its strong fit to complex pattern. While in XGBoost, each area is well defined and less likely to get mixed. In Figure 10, the LSM of SVM also emphasizes areas with high risk, but the orange to red (high to very high) zones are narrower.

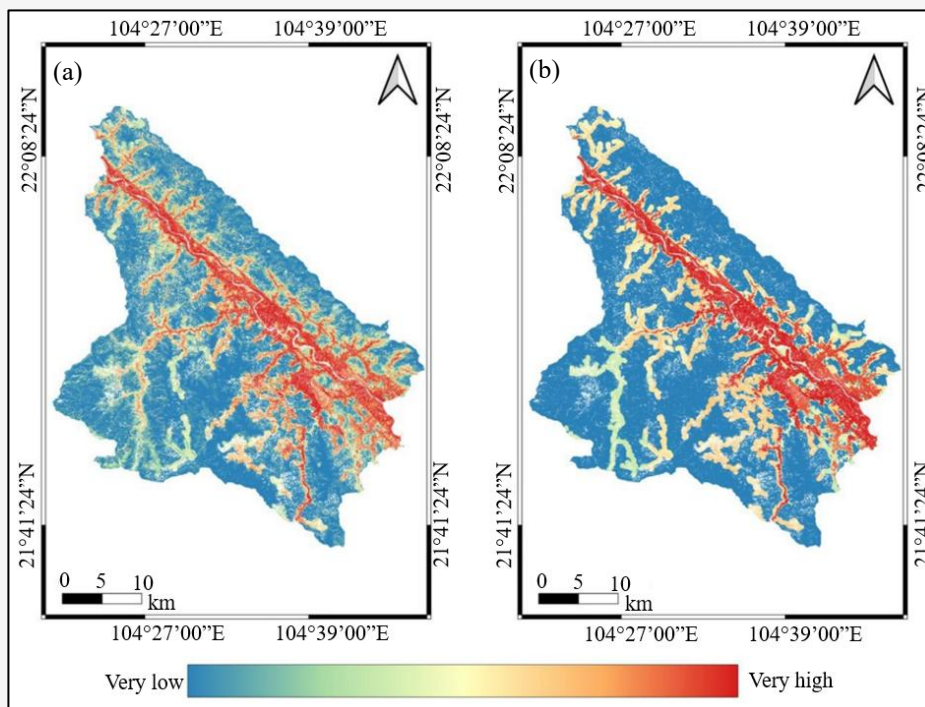


Figure 9: LSM of VY area generated (a) RF, and (b) XGBoost

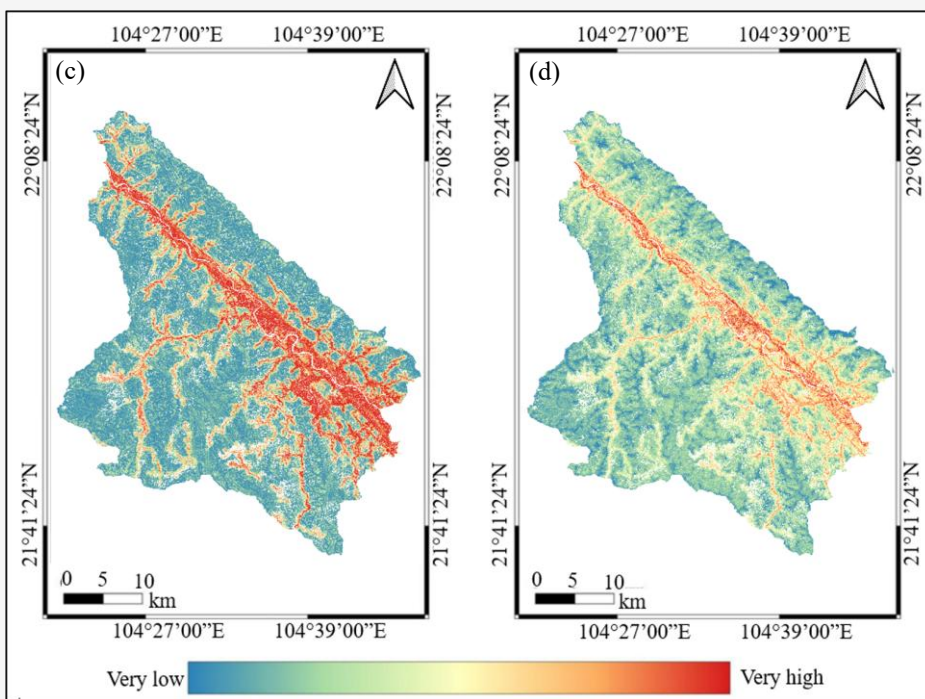


Figure 10: LSM of VY area generated (c) SVM, and (d) LR

The background areas of the two LSM maps produced by SVM and LR models showed larger green zone (low landslide risk). SVM's map highlights sharper separation between high risk corridors and low risk surroundings. The LR showed

the smoothest pattern, drawing the average risk level closer to the center (medium). Yellow/orange (medium–high) susceptibility spreads over larger areas, contrasts between safe and unsafe areas are less visible. Through LR, the area has moderate risk.

5. Conclusion

The study demonstrates that RF and XGBoost models achieved the highest of the four models tested. Though RF is only slightly behind, it continues to be a reliable ML model while XGBoost shows high potential for its robustness. Kappa Score, Sensitivity and Specificity remained very important metrics when assessing ML models, indicating the accuracy score of the classification task is not by chance, as shown in validation results. The models have poor generalizability when perform validation, since the differences in topography are much bigger than anticipated. The validation result showed high Specificity value, which measures TNR, but with low Sensitivity to landslide. The problem with all 4 models when validation is that models get confused with landslide causative factors. This might be a problem in the training dataset when all landslides are too centralized in certain regions. Furthermore, external validation brings a lot of bias from the training data into a new data set. Therefore, the method should be re-evaluated before being deployed in future research. The convergence analysis using BO indicated that the trained ML models converged quickly within few tries with RF and XGBoost achieving very high ROC-AUC scores early on. While accuracy metrics are good for evaluating classification results, EGO models can be used to see the learning capability of ML models on a dataset. FR was proved to be a useful method to standardize the dataset, which composes different factors across different scales and types, preventing situations when ML models have inconsistent performance. Finding a means to standardize dataset on the same scale is a very important step in ML analysis. It is crucial that the distribution of training points represents a wide range of landslide-influencing factors across the entire study area to ensure robust model performance. The application of the ML workflow in this study suggests that similar approaches can be used for other landslide-prone regions. From the results of the analysis, it can be concluded that the landslide inventories still contain bias from the method the landslide is collected, InSAR method is proposed to be incorporated to build a more accurate landslide inventory. High-resolution DEM will also be utilized for more precise analysis of small and medium-scale landslides. With enhanced landslide inventories, advanced deep learning models will be explored to achieve more accurate and robust landslide classification. Furthermore, future research will also focus on setting threshold between slope failures and landslides when identifying landslides, to prevent confusing the models. In order to cover a wider range of influence, climate factors (e.g., rain precipitations,

wind and humidity...) will be included. The scope of the study will be expanded to other regions, including Japan, to assess the effectiveness of ML models across diverse geological and topographic settings. Many landslides happened on ingenious, metamorphic rocks and limestone. This indicates there are moderate - thick layers of weathered soil on top of these rocks. Future research will integrate weathered layer thickness on to further increase the impact value of geologic factors when using ML evaluation.

Acknowledgments

This research was supported by SPRING (Support for Pioneering Research Initiated by the Next Generation) under the auspices of Osaka Metropolitan University (OMU), Osaka, Japan. My sincere gratitude goes to the Laboratory of Geoinformatics, Department of Geosciences, OMU for providing me with the opportunity and to my supervisors for their mentorship.

References

- [1] Froude, M., J. and Petley, D. N., (2018). Global Fatal Landslide Occurrence From 2004 to 2016. *Natural Hazards and Earth System Sciences*, Vol. 18(8); 2161–2181. <https://doi.org/10.5194/nhess-18-2161-2018>.
- [2] Reichenbach, P., Rossi, M., Malamud, B. D., Mihir, M. and Guzzetti, F., (2018). A Review of Statistically-Based Landslide Susceptibility Models. *Earth-Science Reviews*, Vol. 180; 60–91. <https://doi.org/10.1016/j.earscirev.2018.03.001>.
- [3] Pham, B., T., Bui, D., T., Pham, H., V., Le, H., Q., Prakash, I. and Dholakia, M. B., (2016). Landslide Hazard Assessment Using Random SubSpace Fuzzy Rules Based Classifier Ensemble and Probability Analysis of Rainfall Data: A Case Study at Mu Cang Chai District, Yen Bai Province (Viet Nam). *Journal of the Indian Society of Remote Sensing*, Vol. 45(4); 673–683. <https://doi.org/10.1007/s12524-016-0620-3>.
- [4] Pham, B., T., Shirzadi, A., Shahabi, H., Omidvar, E., Singh, S., K., Sahana, M., Asl, D., T., Ahmad, B., B., Quoc, N., K. and Lee, S., (2019). Landslide Susceptibility Assessment by Novel Hybrid Machine Learning Algorithms. *Sustainability*, Vol. 11(16). <https://doi.org/10.3390/su11164386>.
- [5] Landslide Susceptibility Mapping. Website name: Geological Survey Ireland. [Online]. Available: <https://www.gsi.ie/en-ie/programmesandprojects/geohazards/projects/Pages/Lan>

- dslide-Susceptibility-Mapping.aspx [Accessed: Aug. 12, 2024].
- [6] Guzzetti, F., Reichenbach, P., Cardinali, M., Galli, M. and Ardizzone, F., (2005). Probabilistic Landslide Hazard Assessment at The Basin Scale. *Geomorphology*, Vol. 72(1-4); 272–299. <https://doi.org/10.1016/j.geomorph.2005.06.002>.
 - [7] Goetz, J., N., Brenning, A., Petschko, H. and Leopold, P., (2015). Evaluating Machine Learning and Statistical Prediction Techniques for Landslide Susceptibility Modeling. *Computers & Geosciences*, Vol. 81; 1–11. <https://doi.org/10.1016/j.cageo.2015.04.007>.
 - [8] Merghadi, A., Yunus, A., P., Dou, J., Whiteley, J., Thai Pham, B., Bui, D., T., Avtar, R. and Abderrahmane, B., (2020). Machine Learning Methods for Landslide Susceptibility Studies: A Comparative Overview of Algorithm Performance. *Earth-Science Reviews*, Vol. 207. <https://doi.org/10.1016/j.earscirev.2020.103225>.
 - [9] Chen, C. and Fan, L., (2023). Selection of Contributing Factors for Predicting Landslide Susceptibility Using Machine Learning and Deep Learning Models. *Stochastic Environmental Research and Risk Assessment*, Vol. 37(9); 3411–3429. <https://doi.org/10.1007/s00477-023-02556-4>.
 - [10] Xu, Z., Che, A. and Zhou, H., (2024). Seismic Landslide Susceptibility Assessment using Principal Component Analysis and Support Vector Machine. *Scientific Reports*, Vol. 14. <https://doi.org/10.1038/s41598-023-48196-0>.
 - [11] Sheng, M., Zhou, J., Chen, X., Teng, Y., Hong, A. and Liu, G., (2022). Landslide Susceptibility Prediction Based on Frequency Ratio Method and C5.0 Decision Tree Model. *Frontiers in Earth Science*, Vol. 10. <https://doi.org/10.3389/feart.2022.918386>.
 - [12] Lee, S. and Talib, J. A., (2005). Probabilistic Landslide Susceptibility and Factor Effect Analysis. *Environmental Geology*, Vol. 47(7); 982–990. <https://doi.org/10.1007/s00254-005-1228-z>.
 - [13] Truong, X. Q., Tran, N. D., Dang, N. H. D., Hang, T. D., Nguyen, Q. D., Yordanov, V., Brovelli, M. A., Duong, A. Q. and Khuc, T. D., (2023(a)). WebGIS and Random Forest Model for Assessing the Impact of Landslides in Van Yen District, Yen Bai Province, Vietnam. *Environmental Science and Engineering*, Springer, Heidelberg, 445–464. <https://doi.org/10.1007/978-3-031-17808-527>.
 - [14] Tran, V., A., Thanh, D., K., Quang, T., X., Binh, N., A. and Thanh, P., T., (2024). Application of Potential Machine Learning Models in Landslide Susceptibility Assessment: A Case Study of Van Yen District, Yen Bai Province, Vietnam. *Quaternary Science Advances*, Vol. 14. <https://doi.org/10.1016/j.qsa.2024.100181>.
 - [15] Truong, X. Q., Dang, N. H. D., Hang, T. D., Tran, N. D., Nga, T. T. D., Tran, V. A., Yordanov, V., Brovelli, M. A. and Khuc, T. D., (2023(b)). Random Forest Analysis of Land Use and Land Cover Change Using Sentinel-2 Data in Van Yen, Yen Bai Province, Vietnam. *Environmental Science and Engineering*, Springer, Heidelberg, 429–445. https://doi.org/10.1007/978-3-031-20463-0_27.
 - [16] Brock, J., Schratz, P., Petschko, H., Muenchow, J., Micu, M. and Brenning, A., (2020). The Performance of Landslide Susceptibility Models Critically Depends on the Quality of Digital Elevation Models. *Geomatics Natural Hazards and Risk*, Vol. 11(1); 1075–1092. <https://doi.org/10.1080/19475705.2020.1776403>.
 - [17] Ma, Z., Mei, G. and Piccialli, F., (2020). Machine Learning for Landslides Prevention: A Survey. *Neural Computing and Applications*, Vol. 33(17); 10881–10907. <https://doi.org/10.1007/s00521-020-05529-8>.
 - [18] Ageenko, A., Hansen, L. C., Lyng, K. L., Bodum, L. and Arsanjani, J. J., (2022). Land Slide Susceptibility Mapping Using Machine Learning: A Danish Case Study. *ISPRS International Journal of Geo-Information*, Vol. 11(6). <https://doi.org/10.3390/ijgi11060324>.
 - [19] Rabby, Y. W., Ishtiaque, A. and Rahman, M. S., (2020). Evaluating the Effects of Digital Elevation Models in Landslide Susceptibility Mapping in Rangamati District, Bangladesh. *Remote Sensing*, Vol. 12(17). <https://doi.org/10.3390/rs12172718>.
 - [20] Thang, N. H., Suzuki, S., Hayashi, K., Dinh, N. Q., Van, T. N. and Abe, S., (2024). Geologic Controls and Rainfall Thresholds for Landslide Occurrence in Northern Vietnam. *Journal of the Japan Landslide Society*, Vol. 61(5); 175–185. <https://doi.org/10.3313/jls.61.175>.
 - [21] ALOS PALSAR – Radiometric Terrain Correction. *Alaska Satellite Facility*. [Online]. Available: <https://asf.alaska.edu/datasets/daac/alos-palsar-radiometric-terrain-correction/>. [Accessed: Aug. 11, 2026].

- [22] Saleem, N., Huq, M., E., Yaw, N., Javed, A. and Sajjad, A., (2019). Parameters Derived from and/or Used with Digital Elevation Models (DEMs) for Landslide Susceptibility Mapping and Landslide Risk Assessment: A Review. *ISPRS International Journal of Geo-Information*, Vol. 8(12). <https://doi.org/10.3390/ijgi8120545>.
- [23] Chaikaew, N., Kumsap, C., Wutthisakkaroorn, C., and Pimmasarn, S. (2023). Landslide Susceptibility Mapping Using LiDAR Data: A Case Study of Khao Yai National Park, Thailand. *International Journal of Geoinformatics*, Vol. 19(3), 1–12. <https://doi.org/10.52939/ijg.v19i3.2597..>
- [24] Zhang, S., Li, R., and Wang, F., (2019). Characteristics of Landslides Triggered by the 2018 Hokkaido Eastern Ibari Earthquake, Northern Japan. *Landslides*, Vol. 16; 1691–1708. <https://doi.org/10.1007/s10346-019-01207-6>.
- [25] Kanwal, S., Atif, S. and Shafiq, M., (2016). GIS Based Landslide Susceptibility Mapping of Northern Areas of Pakistan, A Case Study of Shigar and Shyok Basins. *Geomatics, Natural Hazards and Risk*, Vol. 8(2); 348–366. <https://doi.org/10.1080/19475705.2016.1220023>.
- [26] Abedini, M. and Tulabi, S., (2018). Assessing LNRF, FR, and AHP Models in Landslide Susceptibility Mapping Index: A Comparative Study of Nojjan Watershed in Lorestan Province, Iran. *Environmental Earth Sciences*, Vol. 77. <https://doi.org/10.1007/s12665-018-7524-1>.
- [27] Rouse, J. W., Haas, R. H., Schell, J. A. and Deering, D. W., (1974). Monitoring Vegetation Systems in the Great Plains with ERTS. *Third Earth Resources Technology Satellite Symposium*, Vol. 1; 309–317. <https://ntrs.nasa.gov/citations/19740022614>.
- [28] McFeeters, S. K., (1996). The Use of The Normalized Difference Water Index (NDWI) in the Delineation of Open Water Features. *International Journal of Remote Sensing*, Vol. 17(7); 1425–1432. <https://doi.org/10.1080/01431169608948714>.
- [29] Awasthi, B., Ninsawat, S., Raghavan, V. and Nemoto, T., (2025). PyGILE-Plus: Python GeoInformatics Lab Environment-Plus. *Zenodo*. [Computer software]. Available: <https://doi.org/10.5281/zenodo.16146572>.
- [30] Khuc, T. D., Truong, X. Q., Tran, V. A., Bui, D. Q., Bui, D. P., Ha, H., Tran, T. H. M., Pham, T. T. T. and Yordanov, V., (2023). Comparison of Multi-Criteria Decision Making, Statistics, and Machine Learning Models for Landslide Susceptibility Mapping in Van Yen District, Yen Bai Province, Vietnam. *International Journal of Geoinformatics*, Vol. 19(7); 33–45. <https://doi.org/10.52939/ijg.v19i7.2743>.
- [31] Abdi, H. and Williams, L. J., (2010). Principal Component Analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, Vol. 2(4); 433–459. <https://doi.org/10.1002/wics.101>.
- [32] Kuhn, M. and Johnson, K., (2013). *Applied Predictive Modeling*. 1st ed., Springer, New York. <https://doi.org/10.1007/978-1-4614-6849-3>.
- [33] Breiman, L., (2001). Random Forests. *Machine Learning*, Vol. 45(1); 5–32. <https://doi.org/10.1023/A:1010933404324>.
- [34] Dou, J., Yunus, A. P., Bui, D. T., Merghadi, A., Sahana, M., Zhu, Z., Chen, C., Khosravi, K., Yang, Y., and Pham, B. T., (2019). Assessment of Advanced Random Forest and Decision Tree Algorithms for Modeling Rainfall-Induced Landslide Susceptibility in The Izu-Oshima Volcanic Island, Japan. *The Science of the Total Environment*, Vol. 662; 332–346. <https://doi.org/10.1016/j.scitotenv.2019.01.221>.
- [35] Brenning, A., (2005). Spatial Prediction Models for Landslide Hazards: Review, Comparison and Evaluation. *Natural Hazards and Earth System Sciences*, Vol. 5(6); 853–862. <https://doi.org/10.5194/nhess-5-853-2005>.
- [36] Dao, M., Phan, T., and Nguyen, Q. (2026). Comparing Five Machine Learning Approaches for Local-Scale Landslide Susceptibility Mapping in the Huoi Reng Watershed, Central Vietnam. *International Journal of Geoinformatics*, Vol. 22(8), 23–41. <https://doi.org/10.52939/ijg.v22i8.5139>.
- [37] Tran, T. L., (2025). Landslide-Susceptibility-Mapping. *GitHub*. [Computer software]. Available: <https://github.com/T-tunglam/Landslide-susceptibility-mapping>. [Accessed: October 10, 2026].
- [38] Lillesand, T. M., Kiefer, R. W. and Chipman, J. W., (2008). *Remote Sensing and Image Interpretation*. 6th ed., John Wiley & Sons, Hoboken.
- [39] Kotu, V. and Deshpande, B., (2015). *Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner*. 1st ed., Elsevier, Morgan Kaufmann.

- [40] Bull, A. D., (2011). Convergence Rates of Efficient Global Optimization Algorithms. *Journal of Machine Learning Research*, Vol. 12; 2879–2904. <https://doi.org/10.48550/arXiv.1101.3501>.
- [41] Frazier, P. I., (2018). A Tutorial on Bayesian Optimization. *arXiv preprint arXiv:1807.02811*, Vol. 1(1); 1–22. <https://doi.org/10.48550/arXiv.1807.02811>.
- [42] Jones, D. R., Schonlau, M. and Welch, W. J., (1998). Efficient Global Optimization of Expensive Black-Box Functions. *Journal of Global Optimization*, Vol. 13(4); 455–492. <https://doi.org/10.1023/A:1008306431147>.
- [43] Alibrahim, H. and Ludwig, S. A., (2021). Hyperparameter Optimization: Comparing Genetic Algorithm against Grid Search and Bayesian Optimization. 2021 *IEEE Congress on Evolutionary Computation (CEC)*, Kraków, Poland, 1551–1559. <https://doi.org/10.1109/CEC45853.2021.9504761>.
- [44] Chung, C. F. and Fabbri, A. G., (2003). Validation of Spatial Prediction Models for Landslide Hazard Mapping. *Natural Hazards*, Vol. 30; 451–472. <https://doi.org/10.1023/b:naz.0000007172.62651>.