Journal of Science and Transport Technology
Journal homepage: https://jstt.vn/index.php/en

Landslide Susceptibility Zoning: Integrating Multiple Intelligent Models with SHAP Analysis

Indra Prakash[1], Dam Duc Nguyen[2*], Nguyen Thanh Tuan[2], Tran Van Phong[3], Le Van Hiep[2]

[1]DDG(R) Geological Survey of India, Gandhinagar, Gujarat, 382010, India

[2]Department of Geotechnical Engineering, University of Transport Technology, 54 Trieu Khuc, Thanh Xuan, Ha Noi, Viet Nam

[3]Institute of Geological Sciences, Vietnam Academy of Science and Technology, 84 Chua Lang Street, Dong Da, Ha Noi, 100000, Viet Nam

Article info
Type of article:
Original research paper

DOI:
https://doi.org/10.58845/jstt.utt.2024.en.4.1.23-41

***Corresponding author:**
E-mail address:
damnd@utt.edu.vn

Received: 27/02/2024
Revised: 28/03/2024
Accepted: 29/03/2024

Abstract: In this study, we aim to delineate landslide susceptibility zones within Dien Bien province, Vietnam, leveraging the capabilities of various machine learning models including Light Gradient Boosting Machine (LGBM), K-Nearest Neighbors (KNN), and Gradient Boosting (GB). Harnessing a dataset comprising 665 data points and encompassing 14 influential factors such as slope, aspect, curvature, elevation, geological composition, Normalized Difference Vegetation Index (NDVI), and proximity to geological features like faults, rivers, and roads, a comprehensive database for landslide modeling was constructed. The analysis entailed rigorous evaluation and comparison of model accuracy employing established statistical metrics, notably Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC).

The findings underscore the efficacy of the Light Gradient Boosting Machine model, exhibiting superior performance with an AUC score of 0.85, surpassing both the Gradient Boosting model (AUC = 0.81) and the K-Nearest Neighbors model (AUC = 0.79). Notably, the Light Gradient Boosting Machine model emerges as a promising tool for precise landslide prediction within the study area, offering significant potential for the creation of a robust landslide susceptibility map. The resulting spatial forecast map for Dien Bien province holds considerable utility for informing land use planning initiatives aimed at mitigating the impact of landslide disasters in the region.

Moreover, the application of SHAP (Shapley Additive explanation) values to quantify the contribution of each factor to landslide susceptibility prediction, offering novel insights into model interpretation and feature importance. The resulting spatial forecast map holds significant implications for land use planning and disaster mitigation efforts in Dien Bien province, showcasing the potential of advanced machine learning techniques in enhancing landslide risk management strategies.

Key words: LGBM, GB, KNN; GIS; Landslide; Dien Bien, Viet Nam.

1. Introduction

Landslides pose significant threats as natural disasters, particularly in mountainous regions [1], where their occurrence can result in substantial damage to both natural landscapes and built environments, often leading to loss of life and

substantial economic repercussions [1]-[3]. Consequently, identifying areas prone to landslides becomes paramount for effective disaster prevention and management. Landslide susceptibility mapping (LSM) offers a means to gauge the likelihood of landslide occurrence within a given area under specific geo-environmental conditions [4], thereby furnishing decision-makers with valuable insights to preempt and mitigate landslide events.

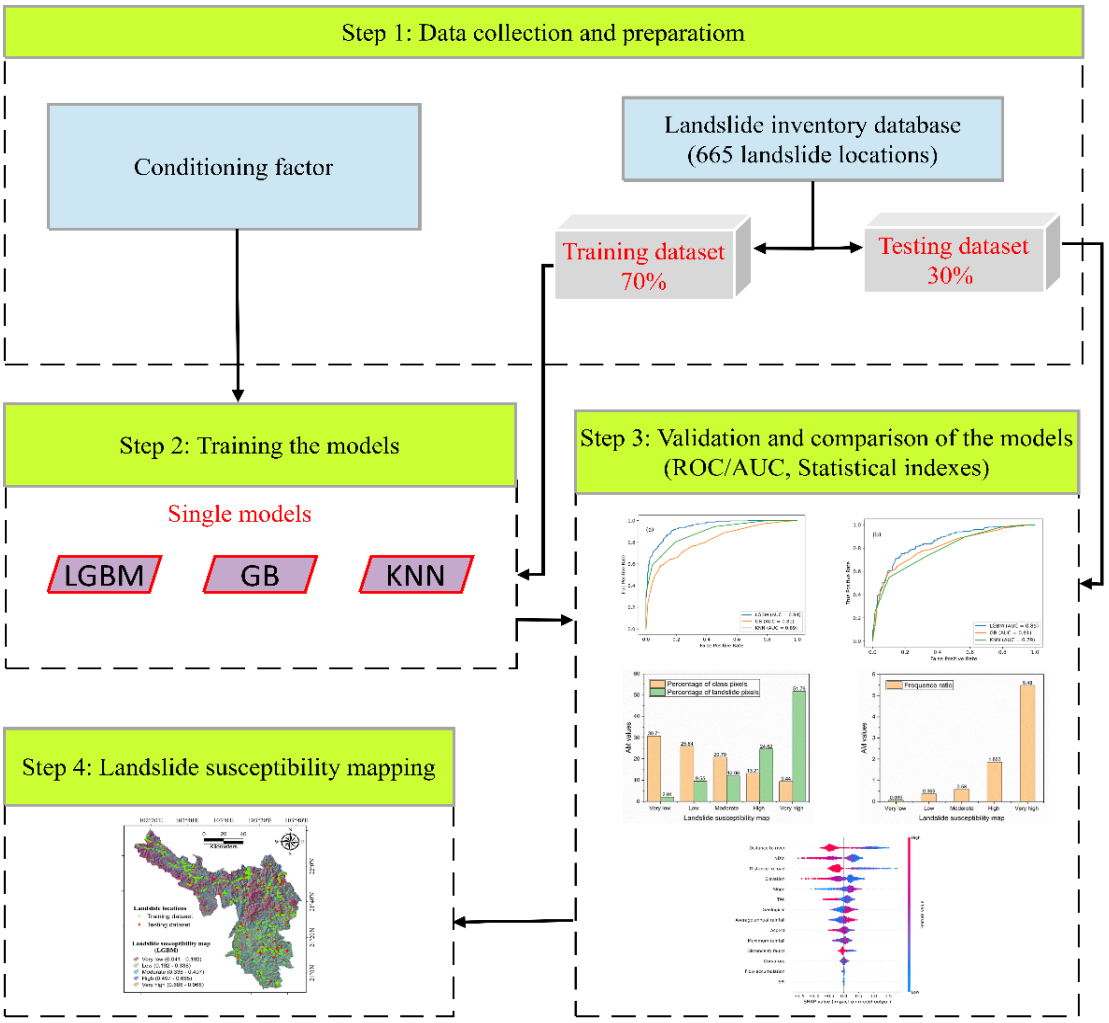


Fig 1. Flow chart of different stages of the present study

Advancements in geographic information systems (GIS) and remote sensing have ushered in a plethora of methodologies for LSM [5], broadly categorized into qualitative and quantitative approaches. Qualitative methods rely heavily on expert knowledge and historical data, such as analytic hierarchy processes and weighted linear combinations [6],[7]. While straightforward to implement, qualitative methods are susceptible to subjective biases. In contrast, quantitative methods, encompassing deterministic and data-driven models, play a pivotal role in LSM. Deterministic models, rooted in physical principles, offer precise estimations but demand extensive geotechnical and hydrogeological data, often impractical for large-scale applications [8]-[10].

In recent years, statistical models leveraging machine learning have gained prominence [10], offering robust solutions to landslide hazard mapping. Traditional statistical techniques, including weight of evidence and logistic regression, while simple, struggle to capture intricate relationships between landslide dynamics and influencing factors [11]. Machine learning algorithms, renowned for their adeptness in handling nonlinear relationships, have emerged as

formidable tools for LSM [12],[13], spawning a myriad of approaches such as support vector machines, k-nearest neighbors, gradient boosting, decision trees, and deep learning neural networks [14],[15].

This study aims to elucidate and compare the performance of various models employed in landslide susceptibility mapping. Specifically, the efficacy of three models—LGBM, GB, and KNN—is explored through empirical research conducted in Dien Bien province, Vietnam, to generate accurate LSM maps. Leveraging techniques like ROC curve analysis and quantitative evaluation metrics, assessment and comparison of forecasting accuracy are conducted. Validation and comparison of model performance are based on relative operating characteristics, sensitivity, specificity, and overall accuracy metrics.

2. Materials and methods

In this study, to forecast landslide zoning in Dien Biên Province, Vietnam, the following four main steps were employed:

(i) Collecting landslide-sensitive points of the study area that have occurred in the past and identification and preparation of layers of effective factors on the occurrence of landslides in the study area,

(ii) Landslide sensitive zoning using machine learning algorithms.

(iii) Evaluating and selecting the most reliable landslide susceptibility map using evaluation criteria such as AUC/ROC and statistical indexes.

(iv) Select the most reliable landslide susceptibility map based on the model's study.

3. Methods used

3.1. LGBM Classifier

In this study, the light gradient boosting machine (LGBM) algorithm was used to build a landslide susceptibility model. LGBM is a machine learning algorithm based on gradient boosting decision trees (GBDT). LGBM's scalability and parallel computing enable it to process large-scale data sets with a relatively small memory footprint [16]. This feature makes LGBM particularly suitable for landslide susceptibility studies, as studies often involve large amounts of remote sensing data and environmental variables. Compared to algorithms such as: RF, SVM and GBDT, LGBM uses a graph-based decision tree algorithm instead of the traditional binary decision tree, which reduces the possibility of overfitting the model. Additionally, LGBM supports L1 and L2 regularization and adopts a leaf growth strategy, which can limit model complexity and reduce the risk of overfitting while maintaining accuracy of the model [17]. The equation of LGBM aims to minimize the loss function, which is usually expressed by the following Equation 1:

$$(y,F)=\sum l(yi, F(xi))+\sum \Omega(ft) \quad (1)$$

where l(yi,F(xi)) represents the loss function, measuring the discrepancy between the predicted value F(xi) and the actual value yi. Ω(ft) denotes the regularization term used to restrict model complexity and prevent overfitting. The summation symbol (∑) indicates the summation of all data samples or base learners.

3.2. Gradient Boosting

Gradient boosting (GB) is one of the typical artificial intelligence methods used to develop classification and regression models to optimize the model learning process to solve non-linear problems [18]. GB is more widely known as decision trees or regression trees. The GB is trained and built by adding new learners in a gradual sequential manner thereby grouping weak prediction models, i.e., decision trees, through the nodes and leaves of the decision tree, and the final prediction result is determined based on the decision nodes [19]. Individual decision trees are weak models, but when viewed as a set (GB), their accuracy is much improved [20]. Therefore, the populations are built gradually in an incremental manner such that every population corrects errors in the previous population, thereby improving the accuracy during model training.

3.3. Kneighbors Classifier

The K-nearest neighbor (KNN) algorithm is a supervised learning algorithm utilized for classification and expectation. It works based on the nearness guideline , which recommends that information focuses with comparative highlights are near to each other [21]. The KNN calculation allots a course to a point based on the closest neighbors in its classification setup. It calculates the separate between the target point and its closest neighbors to decide the course task utilizing remove measurements such as the Euclidean or Manhattan remove. By considering the chosen number of neighbors (K), it chooses the lesson with the most elevated number of votes from these neighbors. The KNN calculation utilizes a subordinate thickness approach and a decision run the show to gather comparative pixels within the include space [22]. This implies that pixels located nearby in this space are considered part of the same class. This model finds applications in different areas such as design acknowledgment, picture preparing, and information investigation. It empowers the classification or expectation of information focuses based on the characteristics of their closest neighbors, leveraging the concepts of nearness and closeness [23].

3.4. Validation methods

3.4.1. Receiver operating characteristic curve (ROC)

The receiver operating characteristic curve (ROC) is a common method used to evaluate the performance of binary classification models [24]. The ROC bend visualizes the classifier s expectation comes about by plotting the genuine positive rate TP on the vertical pivot and the untrue positive rate FP on the flat pivot. When assessing the execution of avalanche vulnerability expectation, the ROC bend can be utilized to survey the classification capacity of the demonstrate for avalanche and non landslide tests [25]. Using landslide samples as positive examples and non-landslide samples as negative examples, the TP and FP values are calculated at different thresholds based on the model's predictions, and the ROC curve is then constructed. The closer the ROC's distance to the top left corner, the more effective the model is at predicting. The quality of the model can be determined by measuring the area under the ROC curve (AUC), which has a range of [0, 1]. The closer the AUC value is to 1, the more accurate the model's predictions are. When the AUC value is greater than 0.8, it suggests that the model has a superior capacity to differentiate between landslide and non-landslide samples, and it can accurately predict and categorize the susceptibility of landslides [26]. The calculation is as the following Equation 2:

$$AUC=\frac{(\sum TP+\sum TN)}{(P+N)} \quad (2)$$

where, TP is the number of landslides that is correctly classified, TN is the number of incorrectly classified landslides, P is the total number of landslides and N is the total number of non-landslides.

3.4.2. Statistical Indexes

In this study, we employed a variety of metrics to assess the performance of models during both the training and validation phases. These metrics encompassed true positives (TP), true negatives (TN), false positives (FP), false negatives (FN), positive predictive value (PPV), negative predictive value (NPV), sensitivity or true positive rate (SST), specificity or true negative rate (SPF), accuracy (ACC), Kappa, root mean square error (RMSE). Below, we will briefly explain each metric.

Positive Predictive Value (PPV): Also known as precision, it is the percentage of positive predictions that are actually positive.

Negative Predictive Value (NPV): It is the percentage of negative predictions that are actually negative.

Sensitivity (SST): Also known as recall or true positive rate, it is the percentage of actual positive instances that are correctly predicted.

Specificity (SPF): Also known as true

negative rate, it is the percentage of actual negative instances that are correctly predicted.

Accuracy (ACC): It is the percentage of total instances that are correctly predicted. It is calculated as:

$$PPV = \frac{TP}{TP+FP} \quad (3)$$

$$NPV = \frac{TN}{TN+FN} \quad (4)$$

$$SST = \frac{TP}{TP+FN} \quad (5)$$

$$SPF = \frac{TN}{TN+FP} \quad (6)$$

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

Where TP (True Positive) and TN (True Negative) are the numbers of correctly classified landslides, FP (false positive) and FN (False negative) are the numbers of landslides incorrectly classified.

Kappa (K): It is a statistical measure that calculates the agreement of prediction with the actual outcomes, taking into account the possibility of agreement occurring by chance.

$$Kappa = \frac{P_o - P_e}{1 - P_e} \quad (8)$$

where P_o is the relative observed agreement among raters (identical to accuracy), and P_e is the hypothetical probability of chance agreement.

A set of quantitative analysis including mean absolute error (MAE), root mean square error (RMSE), were estimated to measure the accuracy of the landslide susceptibility models. The following formulas are accepted for these statistical measures [27]:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Xpred - Xact)^2} \quad (9)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Xpred - Xact| \quad (10)$$

where $Xpred$ is the observed value, $Xact$ is the predicted value, and n is the number of observations.

3.4.3. SHAP

The core idea of SHAP (Shapley Additive exPlanation) is derived from the cooperative game theory, which was proposed by Lundberg and Lee [28], in order to quantify the contribution of players to collaborative games in the early stage [29]. The framework SHAP combines multiple existing approaches in order to create a theoretically sound and intuitive way to interpreting the predictions of any machine learning model. It has been a significant advancement in the field of machine learning model interpretation. The SHAP value determines the magnitude and orientation (positive or negative) of the influence of features on the prediction.

This study uses SHAP values to quantify the contribution of each factor to landslide susceptibility prediction results. SHAP interprets the Shapley value as an additive method of feature mapping, interpreting the model's predicted value as the sum of the attribute values of each input feature [28]:

$$g(x') = \phi_0 + \sum_{m}^{i=1} \phi_i \quad (11)$$

where $g(x')$ is the value of the model, and ϕ_0 is the constant that explains the model, that is, the predicted mean of all training samples. Φ_i is the value of each attribute (Shapley value) as sociated with it.

4. Data used

4.1. Study area

Dien Bien is a mountainous border province in the Northwest region of the country, with geographical coordinates 20°54' - 22°33' North latitude and 102°10' - 103°36' East longitude. Located 504 km west of Hanoi capital, the east and northeast borders Son La province, the north borders Lai Chau province, the northwest borders Yunnan province (China), the west and southwest borders Lao. The area, tectonically active and dissected by shears and geological faults, exhibits a complex terrain in Dien Bien City, dominated by

steep, rugged, and serrated mountains. The region comprises mountains aligned in a northwest-southeast direction, with elevations varying from 200 meters to over 1,800 meters. The terrain gradually slopes from north to south and from west to east. In the North, there are high points of 1.085 m, 1.162 m and 1.856 m (Muong Nhe district), the highest is Pu Den Dinh peak at 1.886 m. In the West, there are high points of 1.127 m, 1.649 m, 1.860 m and the Muong Phang high point range extending down to Tuan Giao. The region features towering mountains interspersed with narrow, steep valleys, rivers, and streams. Notably, the Muong Thanh valley, spanning over 150 square kilometers, stands out as the largest and most renowned expanse in the province and the entire Northwest region. The mountains exhibit significant erosion, giving rise to large plateaus such as the A Pa Chai plateau (Muong Nhe) and the Ta Phinh plateau (Tua Chua). Additionally, various terrain types including valleys, rivers, streams, alluvial terraces, volcanic cones, slopes, and caverns are widely dispersed throughout the area, albeit occupying relatively small portions.

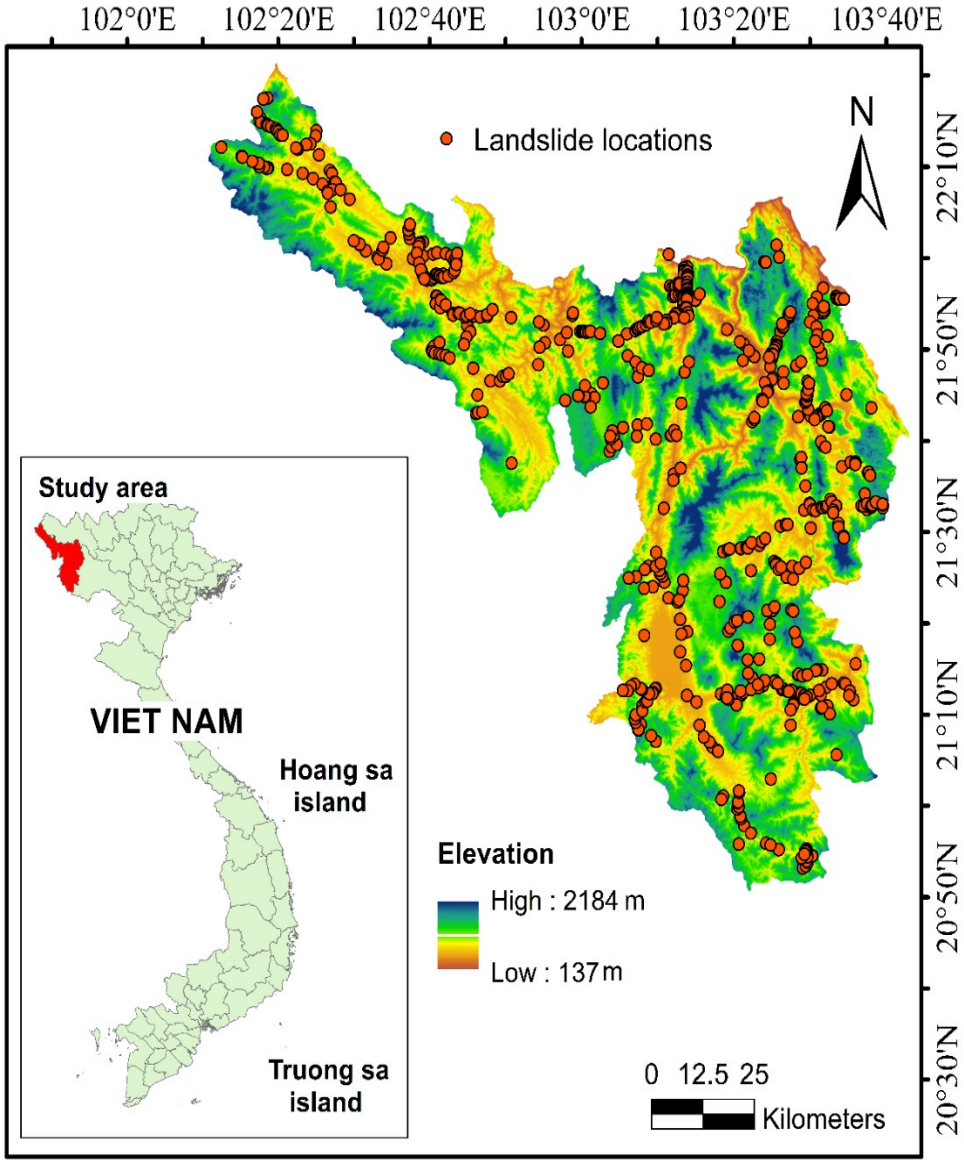


Fig 2. Location of the landslide study area

4.2. Landside inventory

The landslide inventory plays a pivotal role in the assessment of landslide susceptibility, serving as a comprehensive record of historical landslides

within the research area [30]. This initial stage holds paramount importance as it furnishes essential data necessary for refining models that influence the accuracy of Landslide Susceptibility Mapping (LSM) [31]. Therefore, the precision of the landslide inventory directly impacts the effectiveness of LSM models [32]. Higher quality and accuracy in the landslide inventory correspond to enhanced predictive capabilities and execution of LSM models, emphasizing the critical need for meticulous data collection and analysis [32]. Landslide inventory of this study includes 665 past landslide locations which were identified from Google Earth Images and a few field investigations. Most of landslides occur in this area are classified as shallow landslides (Fig 3). Out of these, 70% of landslide data were randomly selected for generating the training dataset and 30% remaining were selected randomly for generating testing dataset. It should be mentioned that, in order to ensure the reliability of the of the modelling applied in this study, another sample with a number of 665 non-landslide locations was generated. This sample was split into 70% of non-landslide data (464 locations) included in the training dataset and 30% (201 locations) included in testing dataset [33]. The produced landslide inventory maps are then associated with contributing geo-environmental parameters such as land cover, topography, geology, geomorphology, and other factors to assess the likelihood of terrain causing a landslide allocated to a susceptibility level [34]-[37].

4.3. Landslide influencing factors

The spatial distribution of landslides is influenced by both triggering factors and modulating variables, which are chosen based on various characteristics such as morphology, geology, hydrology, and human activities within the area [38]. It's important to distinguish between these factors, with modulating variables including features such as roads, faults, geology, slope angle, and land use, while triggering factors encompass phenomena like rainfall and earthquakes. The selection of independent factors for Landslide Susceptibility Mapping (LSM) does not adhere to universal criteria, but rather relies on factors that are non-redundant, consistent, actionable, and measurable. Utilizing tools like ArcGIS, essential modulating variables can be extracted from digital elevation models, including elevation, slope, aspect, curvature, Normalized Difference Vegetation Index (NDVI), geological attributes, distance to faults, flow accumulation, Stream Power Index (SPI), Topographic Wetness Index (TWI), distance to rivers, maximum rainfall, distance to roads, and average annual rainfall.

These fourteen landslide factors can be categorized as follows: Slope angle is one of the important factors affecting the occurrence of landslides [39]. Landslides often occur on slopes with slope angles from 15° to 54° and rarely occur on slopes with small slopes of 0° – 10°. This map is built with different layers including 5 layers: 0 - 11.102188, 11.102188 - 19.088648, 19.088648 - 26.649187, 26.649187 - 35.679176, 35.679176 - 76.4725 (Fig 4.a).

Aspect is an important factor affecting the occurrence of landslides because it affects the moisture content of the materials forming the slope [40]. In this study, the slope direction map is extracted from the 30 m DEM from USGS source (http://earthexplorer.usgs.gov/) digital terrain model with different layers including: Flat, North, Northeast, East, Southeast, South, Southwest, West, Northwest (Fig 4.b).

Curvature of the terrain surface affects the occurrence of landslides because water flow and surface water accumulation depend significantly on the surface shape of the terrain. Landslides often occur in areas with concave terrain surfaces than areas with flat terrain and convex terrain because surface water often accumulates in more concave terrain [41]. In this study, the topographic surface shape map is extracted from the 30m DEM digital elevation model with 3 layers such as concave (< -

0.05), plan (-0.05 –0.05) and convex (> 0.05) (Fig 4.c).

Elevation affects the process of landslides because at different terrain elevations, the weathering level of soil types on slopes is different [42]. In this study, the terrain elevation map is extracted from the 30m DEM digital elevation model and divided into 9 layers: 137 - 443, 443 - 607, 607 - 742, 743 - 874, 874 - 1006, 1006 - 1145, 1145 - 1305, 1305 - 1515, 1515 - 2184 (Fig 4.d).

The influence of geological and tectonic conditions is considered a fundamental factor causing the landslide process, especially lithological composition is one of the most important factors affecting slope stability [43]. Rocks with low durability tend to weather into less durable materials. Geological maps are collected from national data sources at a scale of 1:200.000. Regional geological layers include: Cretaceous system, Devonian system, Neoproterozioc system, Paleogene system, Permian system, Quantermary system, Silurian system, Triassic system (Fig 4.e).

Landslides are closely related to vegetation cover. Areas with low vegetation cover will cause larger landslides than areas with high vegetation cover [44]. Vegetation cover map (NDVI) is taken from satellite data images with and divided into 6 classes: -0.05806 – 0.13206, 0.13206 – 0.181656, 0.181656 – 0.222987, 0.222987 – 0.264317, 0.264317 – 0.313914, 0.313914 - 1 (Fig 4.f).

Distance to faults are products of tectonic movements that cause discontinuity in soil and rock on the slope, thus affecting the process of landslides [45]. In this study, the distance to the faults was selected as a causal factor affecting the landslide process. The fault system is extracted from the geological map at a scale of 1:200.000. The distance map to the faults is established with 6 classes: 0 - 100, 100 - 200, 200 - 300, 300 - 400, 400 - 500, > 500 (Fig 4.g).

Stream power index (SPI) is a metric related to the velocity of flow and the erosion it causes in rivers and streams. The greater the flow power, the greater the erosion rate and impact. The flow power map was established using ArcGIS software and divided into 6 layers: 0 – 619.188, 619.188 – 3715.130, 3715.130 – 9287.825, 9287.825 – 18575.651, 18575.651 – 35912.924, 35912. 924 - 158512.210 (Fig 4.h).

SPI value is calculated by the equation [46]:

$$SPI=A_s.\tan_\beta \quad (12)$$

In which, As is the area of water collection area, β is the terrain slope in degrees.

Topographic wetness index (TWI) is a metric based on the ratio or catchment area to slope angle. It provides a measure of soil moisture that is positively associated with landslide occurrence. The flow power map was established using ArcGIS software and divided into 6 layers: 2.033716 - 4.755845, 4.755845 - 5.728034, 5.728034 - 6.894661, 6.894661 - 8.450164, 8.450164 - 10.459354, 10.459354 - 18.625742 (Fig 4.i).

The TWI value can be constructed as follows [47]:

$$TWI = \ln\frac{A_s}{\tan_\beta} \quad (13)$$

In which, As is the area of the water collection area, β is the terrain slope in degrees

Flow accumulation is considered by some researchers to be an important moderating factor for landslide hazard mapping [48]. It is used to determine the flow or potential flow of rivers and streams. The flow accumulation map was established using ArcGIS software and divided into 6 layers: 0 - 110, 110 - 426, 426 - 1005, 1005 - 2024, 2024 - 4019, 4019 - 15360 (Fig 4.j).

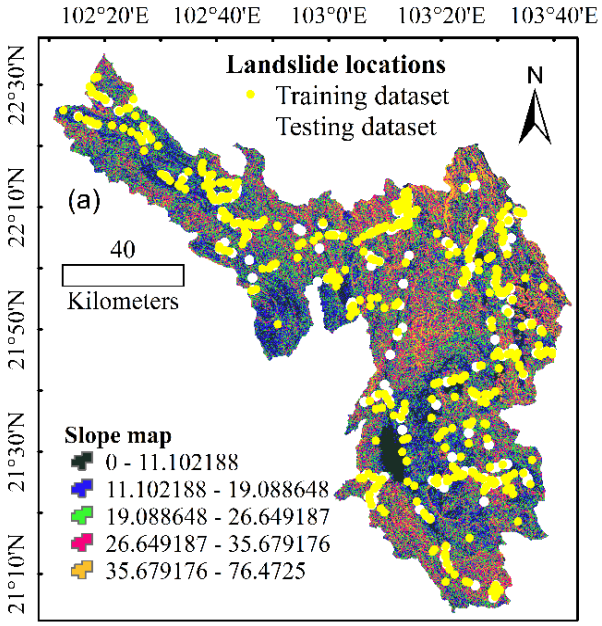
Distance to rivers was selected to analyze the relationship with the landslide occurrence process. Rivers and streams affect the occurrence of landslides because slopes near rivers and streams often have higher humidity than other areas [49]. In addition, water flows in areas with rivers and streams have a direct mechanical impact on the soil and rock of the slope. The river and stream system is extracted from a 1:50.000 scale topographic map. The distance map to the

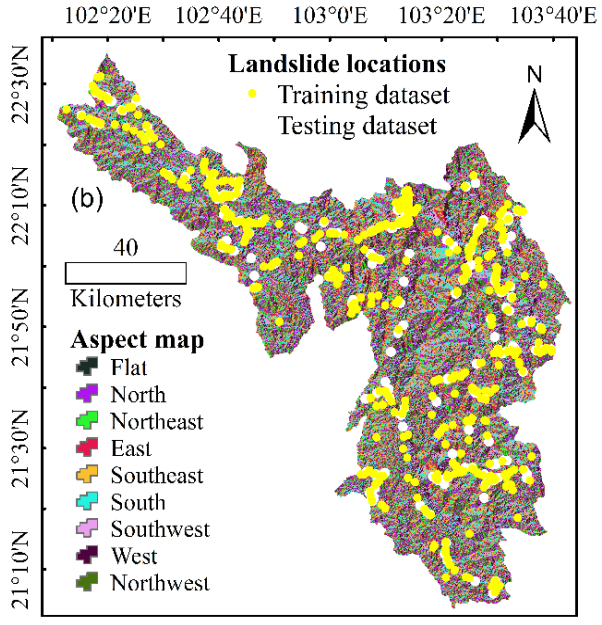
rivers and streams is built into 6 layers: 0 - 100, 100 - 200, 200 - 300, 300 - 400, 400 - 500, > 500 (Fig.4 k).

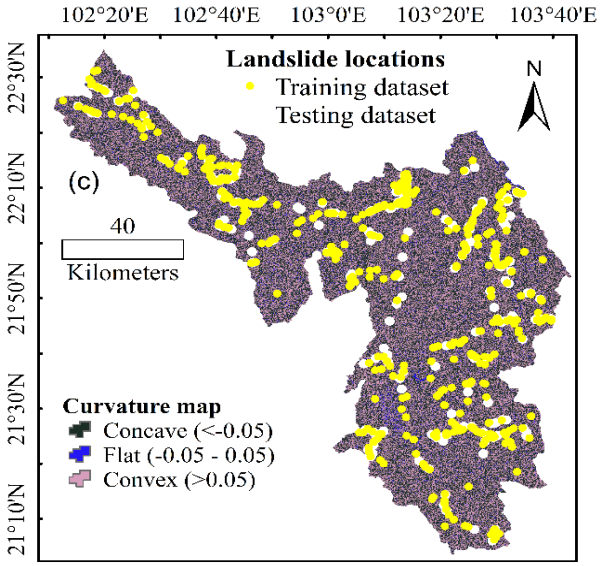
Distance to roads was selected to evaluate the influence of roads on the occurrence of landslides. The process of cutting slopes to build traffic routes often directly impacts the slope, causing loss of continuity of soil and rock on the slope, creating an area of water accumulation that reduces the strength of soil and rock on the slope [50]. Slope affects the occurrence of landslides. The road system is extracted from a 1:50.000 scale topographic map. The distance map to roads is built into 6 layers: 0 - 100, 100 - 200, 200 - 300, 300 - 400, 400 - 500, > 500 (Fig 4.m).

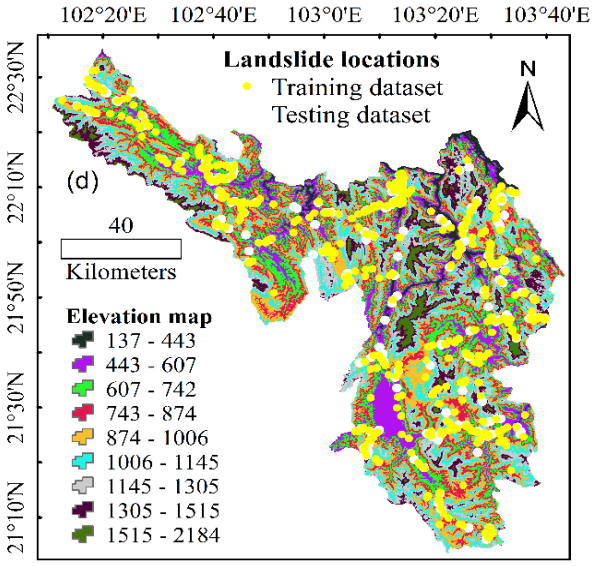
Maximum daily rainfall and Average annual rainfall is the most affecting factor for landslide occurrence [51]. Maximum rainfall map and Average annual rainfall map was constructed from the meteorological data. The Maximum rainfall is divided into five classes including: 62.60 – 86.90, 86.90 – 91.91, 91.91 – 94.54, 94.54 – 98.13, 98.13 – 105.06 (Fig 3.n). Similarly, the average annual rainfall is also divided into five classes including: 1726 – 1830, 1830 – 1900, 1900 – 1970, 1970 – 2050 and 2050 – 2147 (Fig 4.n and Fig 4.o).

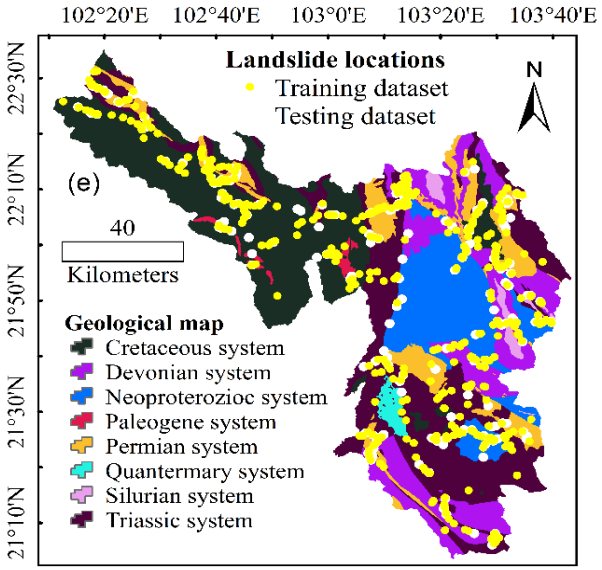
Fig 3. Landslide photos of the study area (Sourcehttps://dienbientv.vn/)

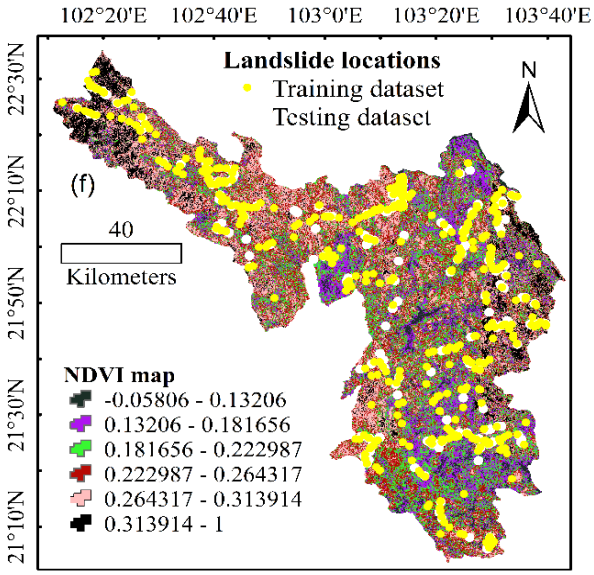


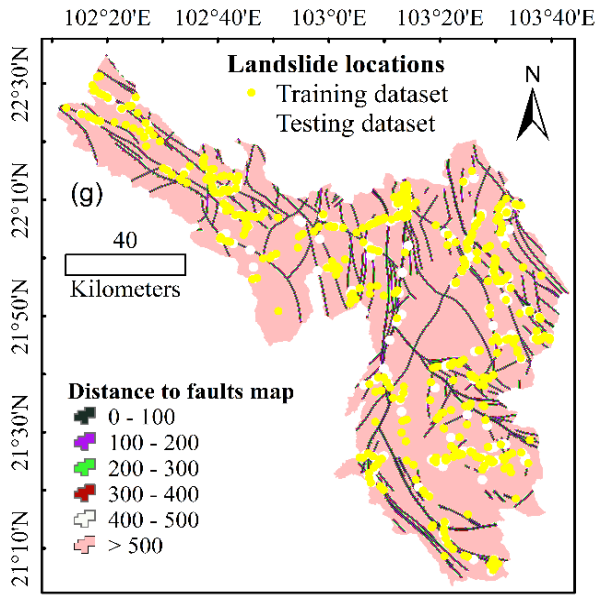


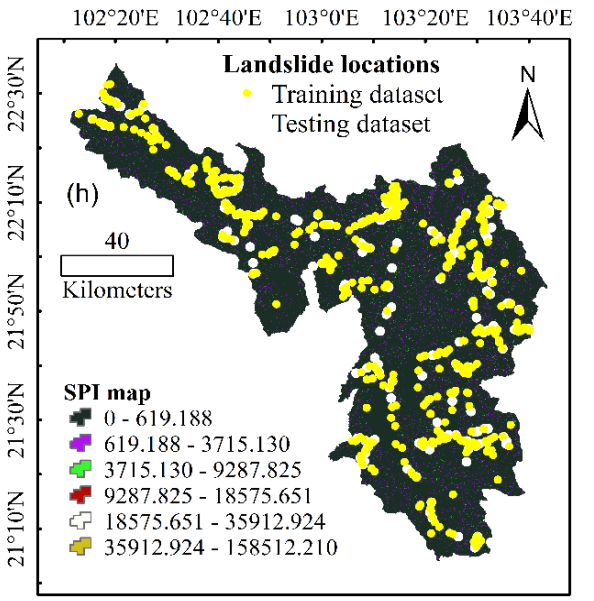

102°20'E
102°40'E
103°0'E
103°20'E
103°40'E
22°30'N
22°10'N
21°50'N
21°30'N
21°10'N
Landslide locations
Training dataset
Testing dataset
N
(c)
40
Kilometers
Curvature map
Concave (<-0.05)
Flat (-0.05 - 0.05)
Convex (>0.05)


102°20'E
102°40'E
103°0'E
103°20'E
103°40'E
22°30'N
22°10'N
21°50'N
21°30'N
21°10'N
Landslide locations
Training dataset
Testing dataset
N
(d)
40
Kilometers
Elevation map
137 - 443
443 - 607
607 - 742
743 - 874
874 - 1006
1006 - 1145
1145 - 1305
1305 - 1515
1515 - 2184


102°20'E
102°40'E
103°0'E
103°20'E
103°40'E
22°30'N
22°10'N
21°50'N
21°30'N
21°10'N
Landslide locations
Training dataset
Testing dataset
N
(e)
40
Kilometers
Geological map
Cretaceous system
Devonian system
Neoproterozioc system
Paleogene system
Permian system
Quantermary system
Silurian system
Triassic system


102°20'E
102°40'E
103°0'E
103°20'E
103°40'E
22°30'N
22°10'N
21°50'N
21°30'N
21°10'N
Landslide locations
Training dataset
Testing dataset
N
(f)
40
Kilometers
NDVI map
-0.05806 - 0.13206
0.13206 - 0.181656
0.181656 - 0.222987
0.222987 - 0.264317
0.264317 - 0.313914
0.313914 - 1


102°20'E
102°40'E
103°0'E
103°20'E
103°40'E
22°30'N
22°10'N
21°50'N
21°30'N
21°10'N
Landslide locations
Training dataset
Testing dataset
N
(g)
40
Kilometers
Distance to faults map
0 - 100
100 - 200
200 - 300
300 - 400
400 - 500
> 500


102°20'E
102°40'E
103°0'E
103°20'E
103°40'E
22°30'N
22°10'N
21°50'N
21°30'N
21°10'N
Landslide locations
Training dataset
Testing dataset
N
(h)
40
Kilometers
SPI map
0 - 619.188
619.188 - 3715.130
3715.130 - 9287.825
9287.825 - 18575.651
18575.651 - 35912.924
35912.924 - 158512.210

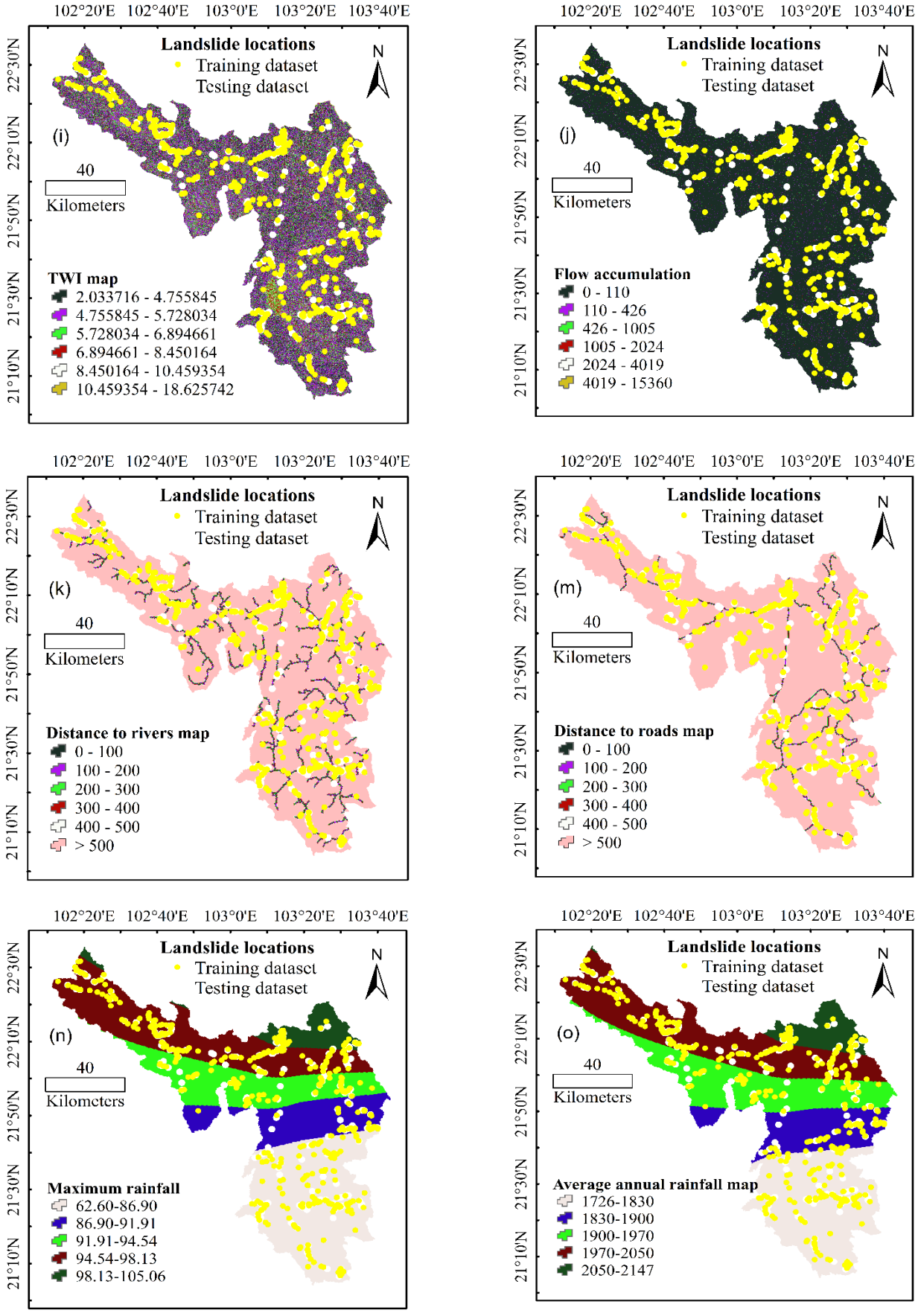


Fig 4. Thematic maps of the study area

5. Results and discussion

5.1. Validation of the models

The spatial landslide prediction model using LGBM, GB and KNN techniques is built on the training data set and verified on the validation data set, and the results of the models' forecasting capacity are shown in Fig 5, Fig 6, Fig 8 and Table 1. The forecasting results using the ROC curve technique (Fig 8) show that the AUC values of all three models LGBM, GB and KNN are high for the whole set. training and validation data. Specifically, the AUC values of the LGBM, GB and KNN models for the training data set are 0.94, 0.81 and 0.89 respectively, while those for the validation data set are 0.85, 0.81 and 0.79 respectively. However, the AUC value of the LGBM model is higher than the GB and KNN models for both training and validation data sets. The forecasting results of the three models using other statistical indicators are shown in Table 1. The values of the statistical indicators of the LGBM model are PPV = 88.65%, NPV = 81.97%, SST = 83.13%, SPF = 87.82%, ACC = 85.32% and K = 0.71 using the training dataset and PPV = 78.89%, NPV = 76.12%, SST = 76.59%, SPF = 78.46%, ACC = 77.50% and K = 0.55 using the validation data set. The values of the statistical indices of the GB model are PPV = 81.37%, NPV = 65.30%, SST = 70.24%, SPF = 77.69%, ACC = 73.36% and K = 0.47, respectively, using the training data set and PPV = 84.92%, NPV = 63.68%, SST = 69.83%, SPF = 81.01%, ACC = 74.25% and K = 0.49 using the validation data set. The values of the statistical indices of the KNN model are PPV = 80.51%, NPV = 80.39%, SST = 80.51%, SPF = 80.39%, ACC = 80.45% and K = 0.61 using the training data set and PPV = 67.34%, NPV = 72.64%, SST = 70.90%, SPF = 69.19%, ACC = 70.00% and K = 0.39 using the validation data set. Fig 5 and Fig 6 show the distribution of root mean square error (RMSE) values of the LGBM, GB and KNN models using the training dataset and validation dataset. Fig 7 depicts the utilization of SHAP to elucidate the features of LGBM, GB, and KNN models. The features are arranged in descending order of importance based on their absolute average SHAP values, which signify the magnitude of their impact on the model's output. Positive SHAP values indicate positive effects, whereas negative SHAP values signify negative effects. In Fig 7, the color spectrum ranges from red to blue, with redder shades indicating larger eigenvalues and bluer shades indicating smaller eigenvalues. A broader spectrum of colors corresponds to a more pronounced effect of the feature, indicating its greater importance in influencing the model's predictions.

Table 1. Accuracy analysis of the models

No	Parameters	Training			Testing		
		LGBM	GB	KNN	LGBM	GB	KNN
1	TP	414	380	376	157	169	134
2	TN	382	303	373	153	128	146
3	FP	53	87	91	42	30	65
4	FN	84	161	91	48	73	55
5	PPV (%)	88.65	81.37	80.51	78.89	84.92	67.34
6	NPV (%)	81.97	65.30	80.39	76.12	63.68	72.64
7	SST (%)	83.13	70.24	80.51	76.59	69.83	70.90
8	SPF (%)	87.82	77.69	80.39	78.46	81.01	69.19
9	ACC (%)	85.32	73.36	80.45	77.50	74.25	70
10	K	0.71	0.47	0.61	0.55	0.49	0.39
11	MAE	0.15	0.27	0.20	0.23	0.26	0.30
12	RMSE	0.38	0.52	0.44	0.47	0.51	0.55

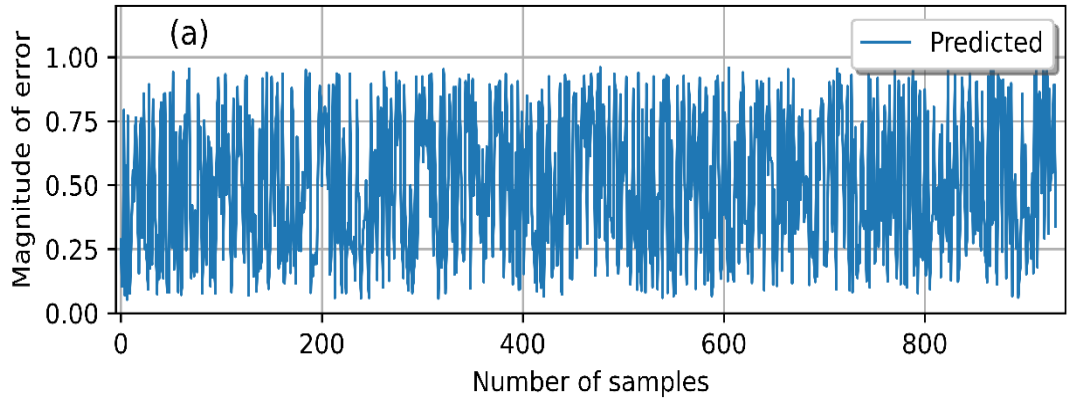


Fig 5. RMSE analysis of the models using training dataset

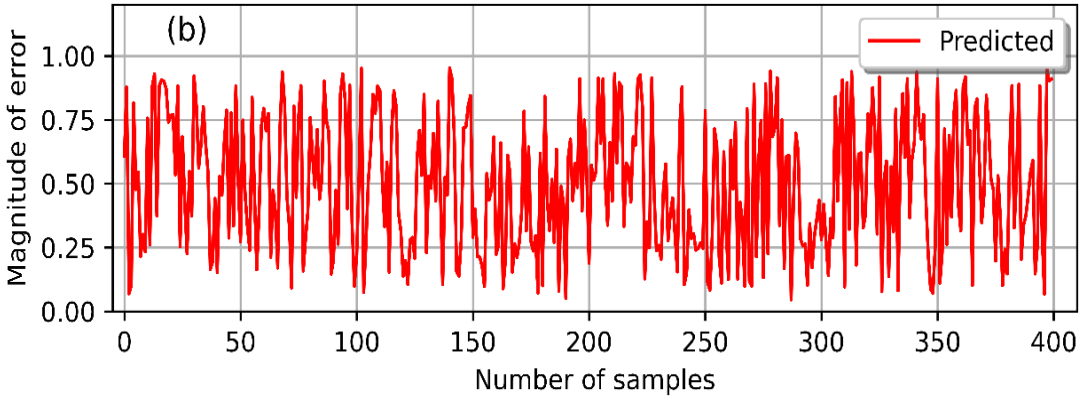


Fig 6. RMSE analysis of the models using validating dataset

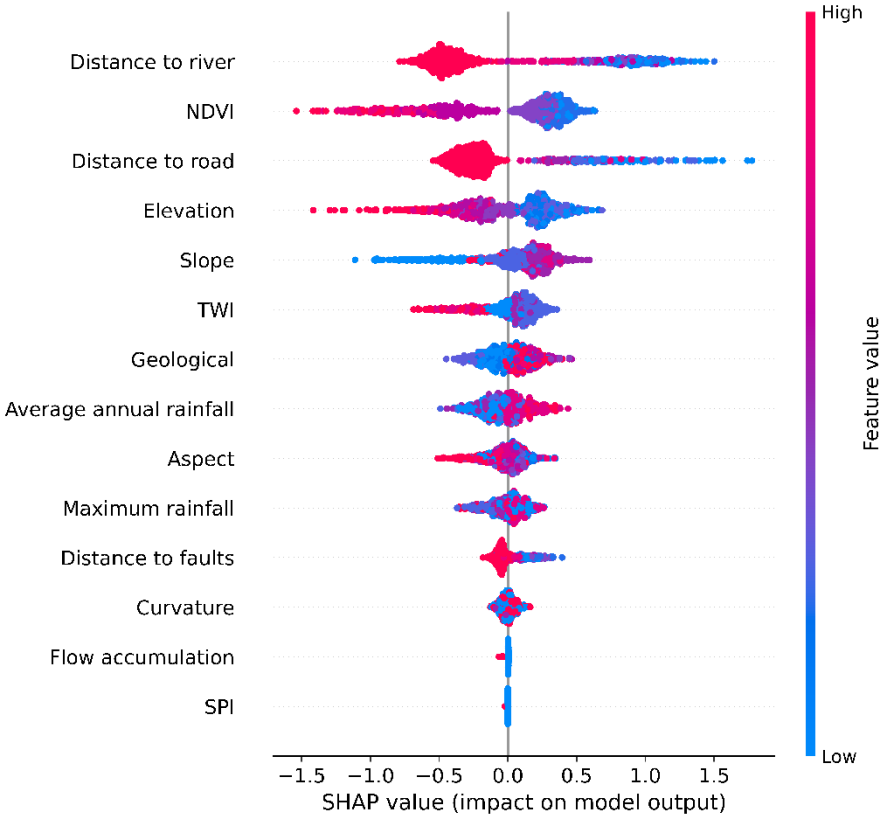


Fig 7. SHAP factor importance

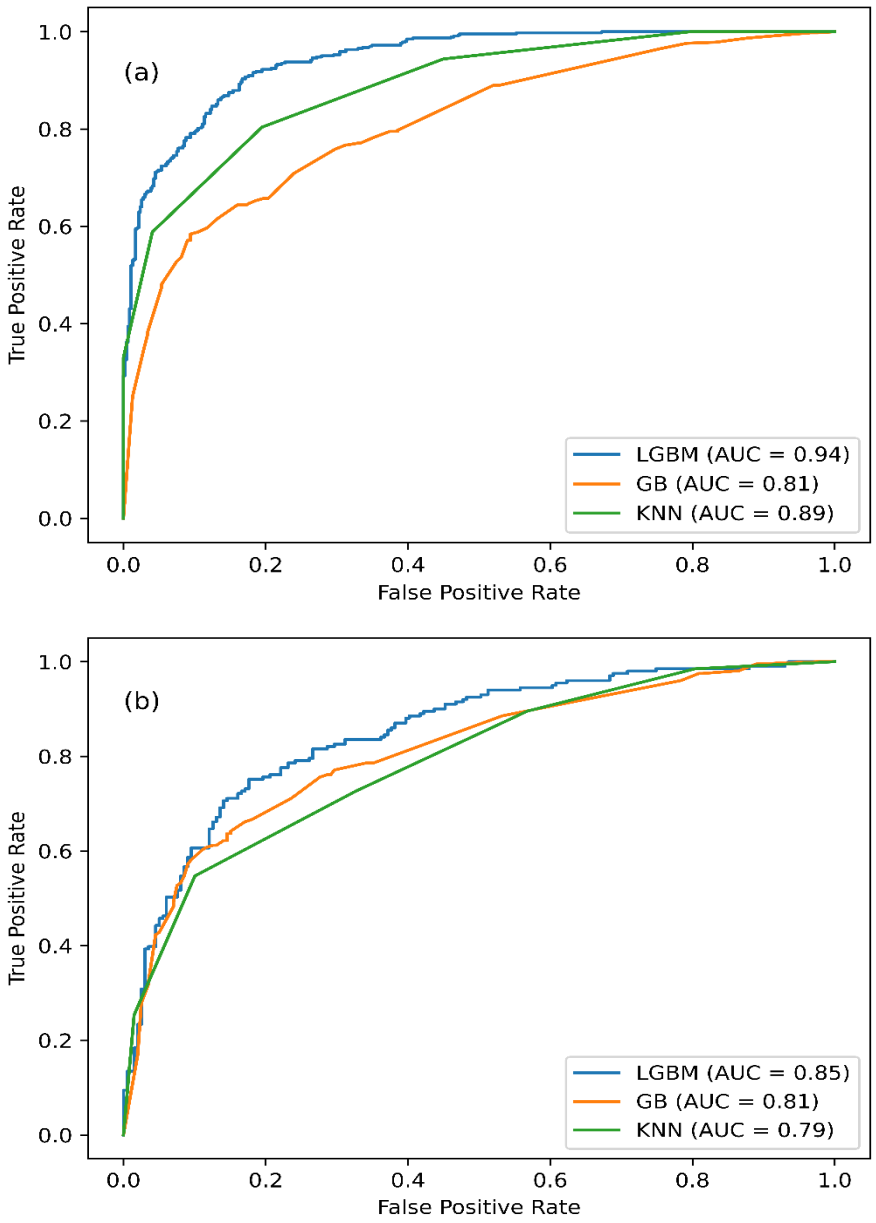


Fig 8. AUC analysis of the models using (a) training dataset and (b) validating dataset

5.2. Construction of landslide susceptibility maps

The landslide susceptibility zoning map was built using the training results of the LGBM model and is shown in Fig 9. Specifically, the probability value of landslide occurrence for pixels in the study area determined through the process of training the LGBM model. These values are then classified into five classes: very high, high, moderate, low and very low using the natural breakpoint classification method built into the ArcGIS application.

Fig 10.b shows the distribution of past landslides across the zoning layers of the landslide hazard zoning map. To evaluate the accuracy of the prediction map, landslides in the validation data are overlapped with the layers of the zoning map and the frequency ratio of occurrence is determined, the results are shown in Fig 10.a. Evaluation results show that most landslides in the past occurred in the very high and high probability classes with the highest frequency ratio values: Very high (5.48), high (1.863). This proves that the spatial landslide prediction map built from the results of the LGBM model is highly accurate and can be used to help minimize the impact caused by landslides.

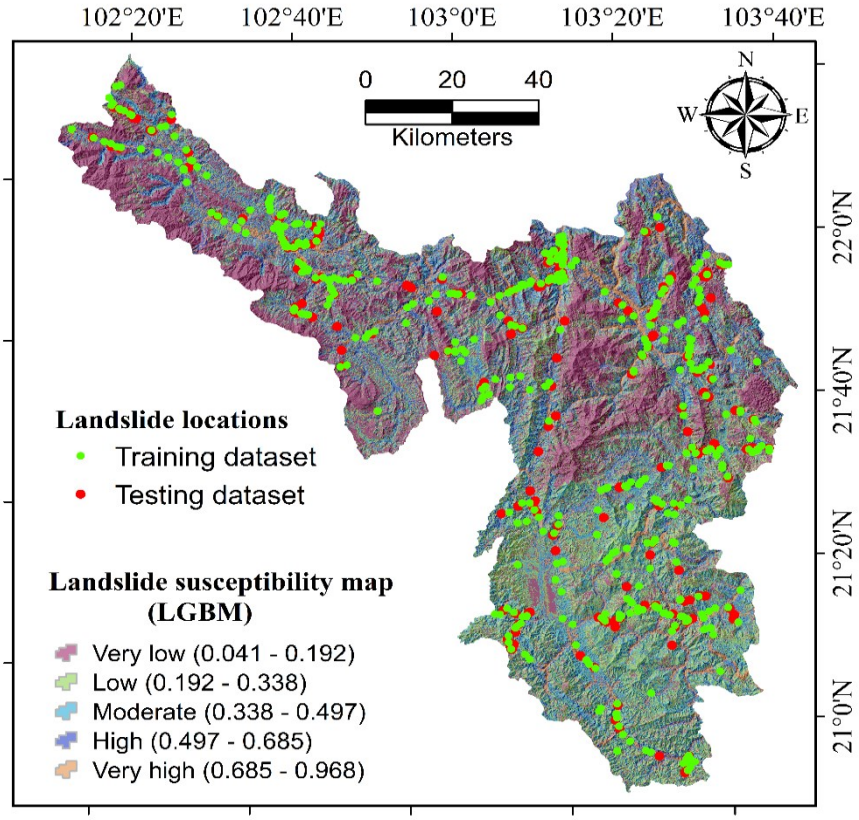


Fig 9. Landslide susceptibility map produced using the LGBM model

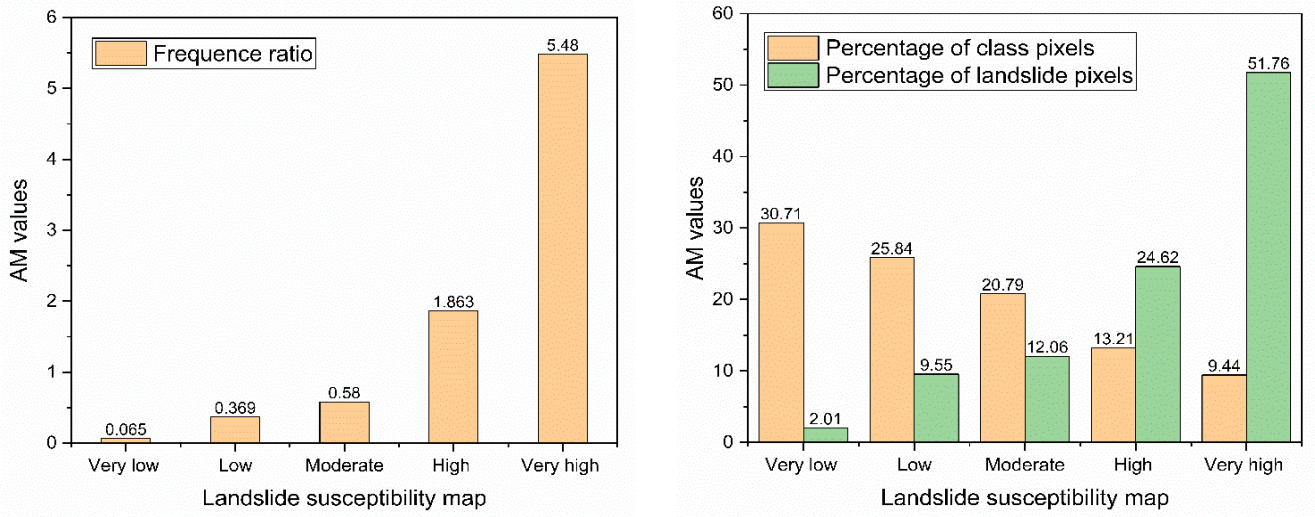


Fig 10. Analysis of landslide density on the susceptibility maps using the models

6. Conclusions

Landslide susceptibility zoning maps are useful tools for effective land use planning to minimize impacts caused by landslide disasters. The article uses advanced artificial intelligence techniques: LGBM, GB and KNN to build a spatial forecast map of landslides in Dien Bien province. A map of the current landslide situation has been built with a total of 665 landslides in the past. A total of 14 landslide cause parameters were selected to build a database used for the prediction model.

The SHAP analysis presented offers valuable insights into feature importance across LGBM, GB, and KNN models. The arrangement of features based on absolute average SHAP values provides a clear understanding of their impact on model output, with positive and negative SHAP values indicating directionality. The color spectrum

in the visualization serves as a visual cue, with varying intensities highlighting the significance of features in influencing model predictions.

Quantitative evaluation techniques such as ROC curves were used to Evaluate and compare the accuracy of models. The results of the study show that all three models LGBM, GB and KNN have high accuracy in building landslide susceptibility zoning maps; However, the LGBM model has higher accuracy than the GB and KNN models. Therefore, the LGBM model can be used as a potential tool in building landslide susceptibility zoning maps. The landslide susceptibility zoning map in Dien Bien province was built with high accuracy and can be used in land use planning and decision making related to landslide disaster management. The proven LGBM technique can be applied to other regions considering the uniqueness and characteristics of each region.

Conflict of Interest: The authors declare that there is no conflict of interest.

Acknowledgment: This research is funded by University of Transport Technology (UTT) under grant number DTTD2022-16.

Funding: This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 105.08-2019.03

Tài liệu tham khảo

[1]. M. Basharat, M.T. Riaz, M.Q. Jan, C. Xu and S. Riaz. (2021). A review of landslides related to the 2005 Kashmir Earthquake: implication and future challenges. *Natural Hazards*, 108, 1-30.

[2]. J.-H. Lee, M.I. Sameen, B. Pradhan and H.-J. Park. (2018). Modeling landslide susceptibility in data-scarce environments using optimized data mining and statistical methods. *Geomorphology*, 303, 284-298.

[3]. G. Berhane, M. Kebede and N. Alfarrah. (2021). Landslide susceptibility mapping and rock slope stability assessment using frequency ratio and kinematic analysis in the mountains of Mgulat area, Northern Ethiopia. *Bulletin of Engineering Geology and the Environment*, 80, 285-301.

[4]. F. Huang, Z. Ye, S.-H. Jiang, J. Huang, Z. Chang and J. Chen. (2021). Uncertainty study of landslide susceptibility prediction considering the different attribute interval numbers of environmental factors and different data-based models. *Catena*, 202, 105250.

[5]. Z. Shirvani. (2020). A holistic analysis for landslide susceptibility mapping applying geographic object-based random forest: A comparison between protected and non-protected forests. *Remote Sensing*, 12(3), 434.

[6]. A. Aditian, T. Kubota and Y. Shinohara. (2018). Comparison of GIS-based landslide susceptibility models using frequency ratio, logistic regression, and artificial neural network in a tertiary region of Ambon, Indonesia. *Geomorphology*, 318, 101-111.

[7]. S. Baharvand, J. Rahnamarad, S. Soori and N. Saadatkhah. (2020). Landslide susceptibility zoning in a catchment of Zagros Mountains using fuzzy logic and GIS. *Environmental Earth Sciences*, 79, 1-10.

[8]. B.T. Pham, D.T. Bui, I. Prakash and M. Dholakia. (2017). Hybrid integration of Multilayer Perceptron Neural Networks and machine learning ensembles for landslide susceptibility assessment at Himalayan area (India) using GIS. *Catena*, 149, 52-63.

[9]. Y. Wang, L. Feng, S. Li, F. Ren and Q. Du. (2020). A hybrid model considering spatial heterogeneity for landslide susceptibility mapping in Zhejiang Province, China. *Catena*, 188, 104425.

[10]. H. Akinci and M. Zeybek. (2021). Comparing classical statistic and machine learning models in landslide susceptibility mapping in Ardanuc (Artvin), Turkey. *Natural Hazards*, 108(2), 1515-1543.

[11]. V.K. Sarda and D.D. Pandey. (2019). Landslide Susceptibility Mapping Using Information Value Method. *Jordan Journal of*

Civil Engineering, 13(2), 335.

[12]. J.M. Habumugisha, N. Chen, M. Rahman, M.M. Islam, H. Ahmad, A. Elbeltagi, G. Sharma, S.N. Liza and A. Dewan. (2022). Landslide susceptibility mapping with deep learning algorithms. *Sustainability*, 14(3), 1734.

[13]. J. Jacinth Jennifer and S. Saravanan. (2022). Artificial neural network and sensitivity analysis in the landslide susceptibility mapping of Idukki district, India. *Geocarto International*, 37(19), 5693-5715.

[14]. L. Bragagnolo, R. Da Silva and J. Grzybowski. (2020). Artificial neural network ensembles applied to the mapping of landslide susceptibility. *Catena*, 184, 104240.

[15]. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones and A. Gomez. (2017). Advances in neural information processing systems. The Neural Information Processing Systems, 30.

[16]. R.M. Aziz, M.F. Baluch, S. Patel and A.H. Ganie. (2022). LGBM: a machine learning approach for Ethereum fraud detection. *International Journal of Information Technology*, 14(7), 3321-3331.

[17]. G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye and T.-Y. Liu. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 3149 - 3157.

[18]. N. Chakrabarty, T. Kundu, S. Dandapat, A. Sarkar and D.K. Kole. (2019). In Flight arrival delay prediction using gradient boosting classifier. *Emerging Technologies in Data Mining and Information Security: Proceedings of IEMIS 2018, Volume 2, Springer*, pp 651-659.

[19]. M.S.I. Khan, N. Islam, J. Uddin, S. Islam and M.K. Nasir. (2022). Water quality prediction and classification based on principal component regression and gradient boosting classifier approach. *Journal of King Saud University-Computer and Information Sciences*, 34(8), 4773-4781.

[20]. L. Lusa. (2017). Gradient boosting for high-dimensional prediction of rare events. *Computational Statistics & Data Analysis*, 113, 19-37.

[21]. E.Y. Boateng, J. Otoo and D.A. Abaye. (2020). Basic tenets of classification algorithms K-nearest-neighbor, support vector machine, random forest and neural network: a review. *Journal of Data Analysis and Information Processing*, 8(4), 341-357.

[22]. K. Balasubramanian and N. Ananthamoorthy. (2022). Correlation-based feature selection using bio-inspired algorithms and optimized KELM classifier for glaucoma diagnosis. *Applied Soft Computing*, 128, 109432.

[23]. A. Jari, A. Khaddari, S. Hajaj, E.M. Bachaoui, S. Mohammedi, A. Jellouli, H. Mosaid, A. El Harti and A. Barakat. (2023). Landslide susceptibility mapping using multi-criteria decision-making (MCDM), statistical, and machine learning models in the Aube Department, France. *Earth*, 4(3), 698-713.

[24]. A. Merghadi, A.P. Yunus, J. Dou, J. Whiteley, B.T. Pham, D.T. Bui, R. Avtar and B. Abderrahmane. (2020). Machine learning methods for landslide susceptibility studies: A comparative overview of algorithm performance. *Earth-Science Reviews*, 207, 103225.

[25]. C.X. Ling, J. Huang and H. Zhang. (2003). In AUC: a better measure than accuracy in comparing learning algorithms. *Advances in Artificial Intelligence: 16th Conference of the Canadian Society for Computational Studies of Intelligence, AI 2003, Halifax, Canada, Proceedings 16, Springer*, pp 329-341.

[26]. A.P. Bradley. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7), 1145-1159.

[27]. K. Mandal, S. Saha and S. Mandal. (2021). Applying deep learning and benchmark machine learning algorithms for landslide

susceptibility modelling in Rorachu river basin of Sikkim Himalaya, India. *Geoscience Frontiers*, 12(5), 101203.

[28]. S.M. Lundberg and S.-I. Lee. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 4768 - 4777.

[29]. W.E. Marcílio and D.M. Eler. (2020). From explanations to feature selection: assessing SHAP values as feature selection mechanism. *2020 33rd SIBGRAPI conference on Graphics, Patterns and Images (SIBGRAPI), IEEE*, pp 340-347.

[30]. A.D. Regmi, K.C. Devkota, K. Yoshida, B. Pradhan, H.R. Pourghasemi, T. Kumamoto and A. Akgun. (2014). Application of frequency ratio, statistical index, and weights-of-evidence models and their comparison in landslide susceptibility mapping in Central Nepal Himalaya. *Arabian Journal of Geosciences*, 7, 725-742.

[31]. G. Paliaga, F. Luino, L. Turconi and F. Faccini. (2019). Inventory of geo-hydrological phenomena in Genova municipality (NW Italy). *Journal of Maps*, 15(2), 28-37.

[32]. F. Guzzetti, A.C. Mondini, M. Cardinali, F. Fiorucci, M. Santangelo and K.-T. Chang. (2012). Landslide inventory maps: New tools for an old problem. *Earth-Science Reviews*, 112(1-2), 42-66.

[33]. M. Shafique, M. van der Meijde and M.A. Khan. (2016). A review of the 2005 Kashmir earthquake-induced landslides; from a remote sensing prospective. J*ournal of Asian Earth Sciences*, 118, 68-80.

[34]. P. Reichenbach, M. Rossi, B.D. Malamud, M. Mihir and F. Guzzetti. (2018). A review of statistically-based landslide susceptibility models. *Earth-science reviews*, 180, 60-91.

[35]. H. Khan, M. Shafique, M.A. Khan, M.A. Bacha, S.U. Shah and C. Calligaris. (2019). Landslide susceptibility assessment using Frequency Ratio, a case study of northern Pakistan. *The Egyptian Journal of Remote Sensing and Space Science*, 22(1), 11-24.

[36]. I. Ilia and P. Tsangaratos. (2016). Applying weight of evidence method and sensitivity analysis to produce a landslide susceptibility map. *Landslides*, 13, 379-397.

[37]. W. Chen, H.R. Pourghasemi, M. Panahi, A. Kornejady, J. Wang, X. Xie and S. Cao. (2017). Spatial prediction of landslide susceptibility using an adaptive neuro-fuzzy inference system combined with frequency ratio, generalized additive model, and support vector machine techniques. *Geomorphology*, 297, 69-85.

[38]. M.A. Hussain, Z. Chen, Y. Zheng, M. Shoaib, S.U. Shah, N. AliandZ. Afzal. (2022). Landslide susceptibility mapping using machine learning algorithm validated by persistent scatterer In-SAR technique. *Sensors*, 22(9), 3119.

[39]. S. Çellek. (2020). Effect of the slope angle and its classification on landslide. *Natural Hazards and Earth System Sciences Discussions*, 1-23.

[40]. T. Arumugam, S. Kinattinkara, S. Velusamy, M. Shanmugamoorthy and S. Murugan. (2023). GIS based landslide susceptibility mapping and assessment using weighted overlay method in Wayanad: A part of Western Ghats, Kerala. *Urban Climate*, 49, 101508.

[41]. K.K. Fatah, Y.T. Mustafa and I.O. Hassan. (2024). Geoinformatics-based frequency ratio, analytic hierarchy process and hybrid models for landslide susceptibility zonation in Kurdistan Region, Northern Iraq. *Environment, Development and Sustainability*, 26(3), 6977-7014.

[42]. P.T.K. Sari, I.B. Mochtar and S. Chaiyaput. (2023). Effectiveness of Horizontal Sub-drain for Slope Stability on Crack Soil Using Numerical Model. *Geotechnical and Geological Engineering*, 41(8), 4821-4844.

[43]. A. Khosronezhad, M. Pourkermani, M. Almasiyan, S. Bouzari and A. Uromeihy. (2024). Landslide susceptibility zoning based on Rock

Engineering System application to the Tehran case study. *Terra Nova*, 36(2), 97-111.

[44]. H. Asada and T. Minagawa. (2023). Impact of Vegetation Differences on Shallow Landslides: A Case Study in Aso, Japan. *Water*, 15(18), 3193.

[45]. S. Ma, H. Qiu, Y. Zhu, D. Yang, B. Tang, D. Wang, L. Wang and M. Cao. (2023). Topographic changes, surface deformation and movement process before, during and after a rotational landslide. *Remote Sensing*, 15(3), 662.

[46]. W. Chen, H. Chai, Z. Zhao, Q. Wang and H. Hong. (2016). Landslide susceptibility mapping based on GIS and support vector machine models for the Qianyang County, China. *Environmental Earth Sciences*, 75, 1-13.

[47]. S. Ma, X. Shao and C. Xu. (2022). Characterizing the distribution pattern and a physically based susceptibility assessment of shallow landslides triggered by the 2019 heavy rainfall event in Longchuan County, Guangdong Province, China. *Remote Sensing*, 14(17), 4257.

[48]. S.L. Cobos-Mora, V. Rodriguez-Galiano and A. Lima. (2023). Analysis of landslide explicative factors and susceptibility mapping in an andean context: The case of Azuay province (Ecuador). *Heliyon*, 9(9), e20170.

[49]. L. Liu, Y. Wu, M. Yin, X. Ma, X. Yu, X. Guo, N. Du, F. Eller and W. Guo. (2023). Soil salinity, not plant genotype or geographical distance, shapes soil microbial community of a reed wetland at a fine scale in the Yellow River Delta. *Science of The Total Environment*, 856, 159136.

[50]. Z. Li, R. Wu, T. Hu, S. Xiao, L. Zhang and D. Zhang. (2023). Stability analysis of an unstable slope in chongqing based on multiple analysis methods. *Processes*, 11(7), 2178.

[51]. H.G. Smith, A.J. Neverman, H. Betts and R. Spiekermann. (2023). The influence of spatial patterns in rainfall on shallow landslides. *Geomorphology*, 437, 108795.