Quaternary Science Advances 14 (2024) 100181

Contents lists available at ScienceDirect

Quaternary Science Advances

journal homepage: www.sciencedirect.com/journal/quaternary-science-advances

Application of potential machine learning models in landslide susceptibility assessment: A case study of Van Yen district, Yen Bai province, Vietnam

Van Anh Tran [a,f], Thanh Dong Khuc [b,*], Xuan Quang Truong [c], An Binh Nguyen [d], Truong Thanh Phi [e]

[a] Faculty of Geomatics and Land Administration, Hanoi University of Mining and Geology, Hanoi, Viet Nam
[b] Faculty of Bridge and Roads, Hanoi University of Civil Engineering, Hanoi, Viet Nam
[c] Faculty of Information Technology, Hanoi University of Natural Resources and Environment, Hanoi, Viet Nam
[d] Ho Chi Minh City Institute of Resources Geography, Vietnam Academy of Science and Technology, Ho Chi Minh City, Viet Nam
[e] Faculty of Geology, Hanoi University of Natural Resources and Environment, Hanoi, Viet Nam
[f] Innovations for Sustainable and Responsible Mining (ISRM) Research Group, Hanoi University of Mining and Geology, Hanoi, Viet Nam

ARTICLE INFO

Keywords:
Landslides
Random forest
Support vector machine
Gradient boosting
Van yen

ABSTRACT

Landslides are natural hazards that cause significant damage to both property and human lives. This study employs potential machine learning models such as Random Forest (RF), Gradient Boosting (GB), and Support Vector Machine (SVM) to assess landslide susceptibility in Van Yen District, Yen Bai Province, Vietnam, that experiences a higher frequency of landslides compared to other localities in the region. The study incorporates thirteen input variables, including elevation, slope angle, aspect, plan curvature, profile curvature, Topographic Wetness Index (TWI), distance to faults, lithology, distance to roads, distance to rivers, land cover, rainfall, and Normalized Difference Vegetation Index (NDVI). To construct the models, landslide statistics reports were utilized, consisting of 302 landslide points collected through field surveys and 52 landslide points determined using Radar Sentinel-1 images. The Google Earth Engine cloud computing platform is utilized for constructing the landslide susceptibility models. The outcome of the research is a landslide susceptibility map with five levels: very low, low, moderate, high, and very high. The Area Under the Curve (AUC) is used as a metric to evaluate the performance of all three models. The findings indicate that, besides similarities observed in landslide susceptibility maps for previously occurred landslides, the Random Forest model demonstrates a favorable performance compared to the other models, with an AUC of 0.883.

1. Introduction

Landslides are a highly dangerous form of natural hazards, causing numerous fatalities and extensive property damage each year (Froude and Petley, 2018). They can be triggered by various factors such as high accumulation of rainfall, earthquakes, or volcanic activity. Additionally, lithology, morphology, and triggers such as prolonged heavy rain and earthquakes contribute to landslide occurrences (Reichenbach et al., 2018). In Southeast Asia, the occurrence of landslides is often linked to rapid urbanization, including activities like road construction, deforestation, economic development, population growth, land use changes, and more recently, climate change (Petley, 2010). In recent times, landslides have been increasing in intensity, frequency, and scale, resulting in severe impacts on human lives, infrastructure, and economic development within these countries (He et al., 2021).

As a result, numerous studies have focused on landslide susceptibility mapping. However, selecting an appropriate method for each study area and ensuring high accuracy poses a challenge. Among the commonly used algorithms, machine learning algorithms have emerged as the most prominent due to their superior accuracy compared to conventional statistical methods. These machine-learning algorithms can be categorized based on their complexity. Classical methods such as Logistic Regression (LR) and Bayesian models are simple machine learning algorithms rooted in statistics. Meanwhile, Support Vector Machines (SVM) and Decision Trees (DT) represent examples of the first simple models. Noteworthy studies include Saito et al. (2009), who applied the decision tree DT, and Heckmann et al. (2013), who employed LR for landslide susceptibility determination. Furthermore, there have been

* Corresponding author.
E-mail address: dongkt@huce.edu.vn (T.D. Khuc).

https://doi.org/10.1016/j.qsa.2024.100181
Received 29 September 2023; Received in revised form 7 March 2024; Accepted 9 March 2024
Available online 16 March 2024

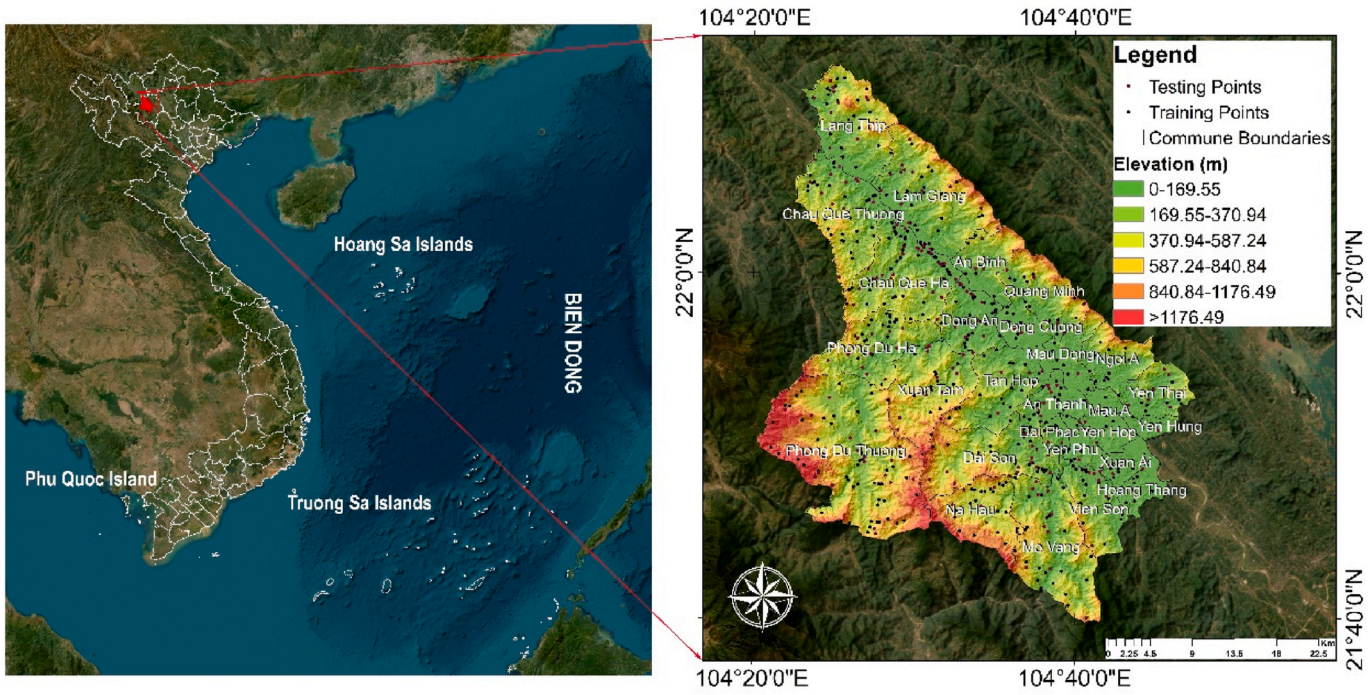


Fig. 1. Location of the study area and the landslide inventory map.

studies comparing different simple machine-learning algorithms. For instance, Kalantar et al. (2018) compared Artificial Neural Networks (ANN), LR, and SVM in one of their studies.

In a study conducted by Chang et al. (2023), Random Forest (RF) and multi-layer perceptron (MLP) machine learning models were applied to construct landslide hazard maps in Chongyi County, China. These models demonstrated promising results in landslide mapping applications.

The first neural network applied to landslides was the Artificial Neural Network (ANN), as studied by Lee et al. (2003) in their work on landslide identification in Boun, Korea. Subsequently, improvements were made to neural networks, resulting in nested structures with increased complexity, such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), also referred to as deep neural networks or deep learning (LeCun et al., 2015; Shin et al., 2016).

In a case study conducted by Chen et al. (2017) in Shaanxi, China, Logical Model Trees (LMT), Random Forest (RF), and Classification and Regression Trees (CART) were compared using a standard dataset of 171 landslides. The study found that RF outperformed the other models.

Ensemble Learning Methods (ELM) involve combining multiple individual machine learning algorithms. Pham et al. (2019) combined different ELM algorithms to create four new hybrid prediction models, which were then evaluated over a test site of approximately 250 square kilometers in the Indian Himalayas. The results were positive, with the algorithm named Bagging based Reduced Error Pruning Trees (BREPT) demonstrating the highest reliability. Additionally, Fang et al. (2021) implemented four ELMs that consisted of components like CNN and RNN, combined with Support Vector Machines (SVM) and Logistic Regression (LR). They also analyzed the correlation between various geomorphological parameters and landslide susceptibility. The results indicated that Ensemble Machine Learning (ELM) outperformed individual algorithms, as confirmed by Yordanov et al. (2021).

In this study, our objective is to utilize three machine learning methods, namely Support Vector Machine, Random Forest, and Gradient Boosting, to predict landslide susceptibility in the Van Yen, Yen Bai province. This research area experiences a higher annual frequency of landslides compared to other regions in the mountainous provinces of Northern Vietnam. The study selected representative models from simple to advanced, to assess their applicability and identify the best performing model for application in the research area. While previous studies have indicated that ensemble machine learning models tend to exhibit superior performance, the results still reveal varied assessments across regions. This study contributes to elucidating the suitability of these models for areas with high landslide frequency, such as Van Yen district, Yen Bai province, Vietnam. Additionally, we intend to leverage the processing tools and available data sources on the Google Earth Engine cloud computing platform to develop the fastest and most effective predictive model. This platform proves convenient for data mining and supports various machine-learning algorithms for constructing prediction models.

Due to the high annual frequency of landslides in the study area, the locations of the landslide survey sites were mainly taken along the roads, so in this study, we also propose to use additional landslides points made from SAR time series images by Persistent Scattering Interferometric Synthetic Aperture Radar (PSInSAR) method. These supplementary points will enhance the modeling process, while the field survey points will serve as a means to assess accuracy. Furthermore, the study area is characterized by complex terrain and notable variations in land cover, which are the primary factors contributing to landslides. Our research comprehensively evaluates the process of selecting an effective landslide prediction model to support land use management and reduce landslide risks.

2. Study area and data used

2.1. Study area

Van Yen 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. The eastern border shares with Luc Yen and Yen Binh districts, the western border with Van Chan district, the southern border with Tran Yen district, and the northern border with Van Ban and Bao Yen district of Lao Cai province (Thanh Thi Pham et al., 2020).

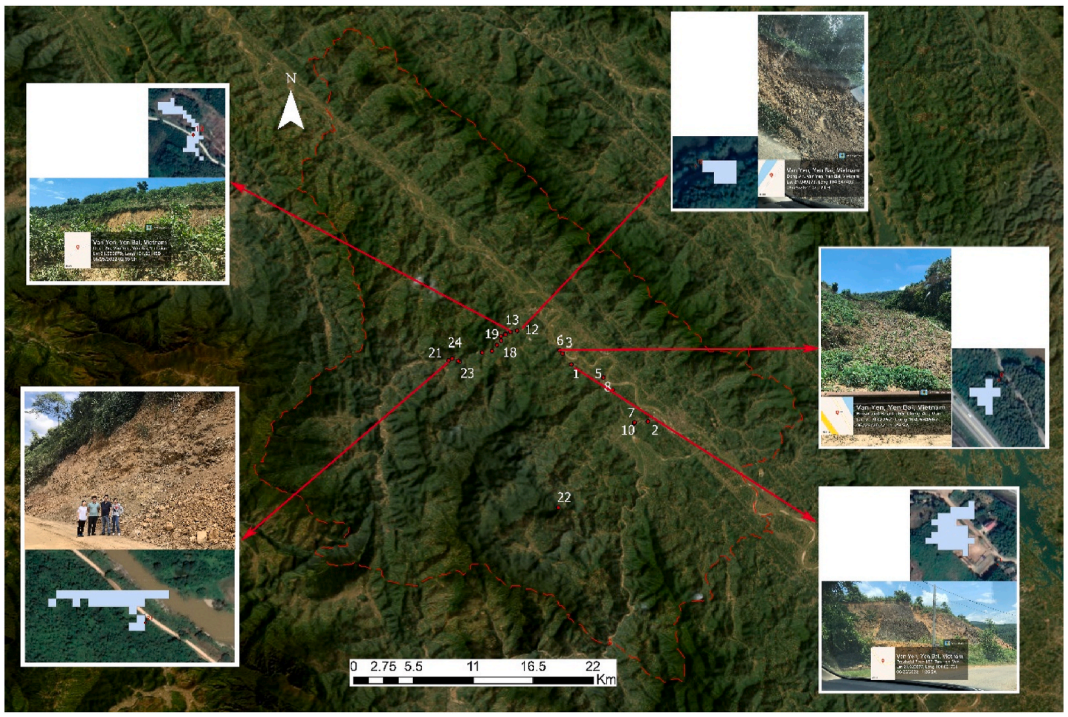


Fig. 2. Locations of several landslide points identified through PSInSAR and the validation process (photo by authors).

This region 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 (Tran et al., 2021).

Van Yen 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. This district experiences the highest frequency of landslides compared to the northern mountainous region and Yen Bai province, resulting in significant losses to human lives, infrastructure, and property (Ha et al., 2018).

2.2. Determination of the inventory points of landslide

Landslide inventory data plays a critical role in predictive modeling, together with landslide controlling factor layers throughout the landslide modeling process. As mentioned above, in order to increase the number of landslide inventory points, we processed a series of Radar Sentinel-1 images. Details of this image processing have been presented in the article of Tran V.A. et al. (2021). Here we only want to describe briefly the method of determining these landslide points.

The PSInSAR method, proposed by Ferretti et al. (2001), relies on using a series of multitemporal SAR images taken at the same location to identify permanent scattering points. These points are instrumental in detecting terrain deformation. The PSInSAR method has been employed for landslide monitoring by several researchers, including Colesanti et al. (2003), Colesanti and Wasowski (2006), Oliveira et al. (2014), Ciampalini et al. (2016), and Yazici and Tunc Gormus (2022). The availability of archived SAR images, along with highly frequent acquisitions, provides PSInSAR with the capability to measure and monitor terrain changes both in the past and the present. In our study, we utilized Sentinel-1 images acquired from January 1, 2019, to January 1, 2021 (Tran et al., 2021). Landslide points are recognized as PS (persistant scattering points), with over 50 thousand PS points having been detected in the region. Despite their abundance, we specifically choose PS points situated on slopes exceeding 10° and possessing negative values.

Field surveys were conducted in 2013, 2015, and 2017, resulting in a total of 302 landslide points. Due to some difficult areas remaining unexplored during the field survey, 52 landslide points were selected from the processed Radar Sentinel-1 satellite images, representing 52 landslide events synthesized from PS points. These points were verified through fieldwork and cross-checked on Google Earth before being incorporated into the inventory dataset (Fig. 2). Consequently, the current number of identified landslide points stands at 354. (Fig. 1).

2.3. Landslide susceptibility model input data

The input data for the landslide susceptibility model consists of various variables and parameters that are used to assess and predict the susceptibility of landslides. These data elements play a crucial role in the accuracy and effectiveness of the prediction model (Chen et al., 2017). The vector data layers are rasterized using ArcMap 10.8 software, followed by standardizing their resolution to 12.5 m to facilitate the data integration process for the model. The key input data typically include:

- Topographic data: This includes information such as elevation, slope angle, aspect, and curvature of the terrain. These factors influence the stability of slopes and are essential in evaluating landslide susceptibility (Pham et al., 2019). In this study, we used a Digital Elevation Model (DEM) derived from ALOS PALSAR with 12.5 m spatial resolution in 2020, which was used to create layers of elevation (Fig. 1), slope angle (Fig. 3(a)), aspect (Fig. 3(b)), plan curvature (Fig. 3(c)), and profile curvature (Fig. 3(d)).
- Topographic Wetness Index (TWI): TWI is an index that describes the process of hydrological flow based on the basin area and specific topographic slope characteristics (Bui et al., 2023). The TWI is used to assess the spatial distribution of soil moisture, which can be transformed from a digital elevation model according to the equation (Fig. 3(e)):

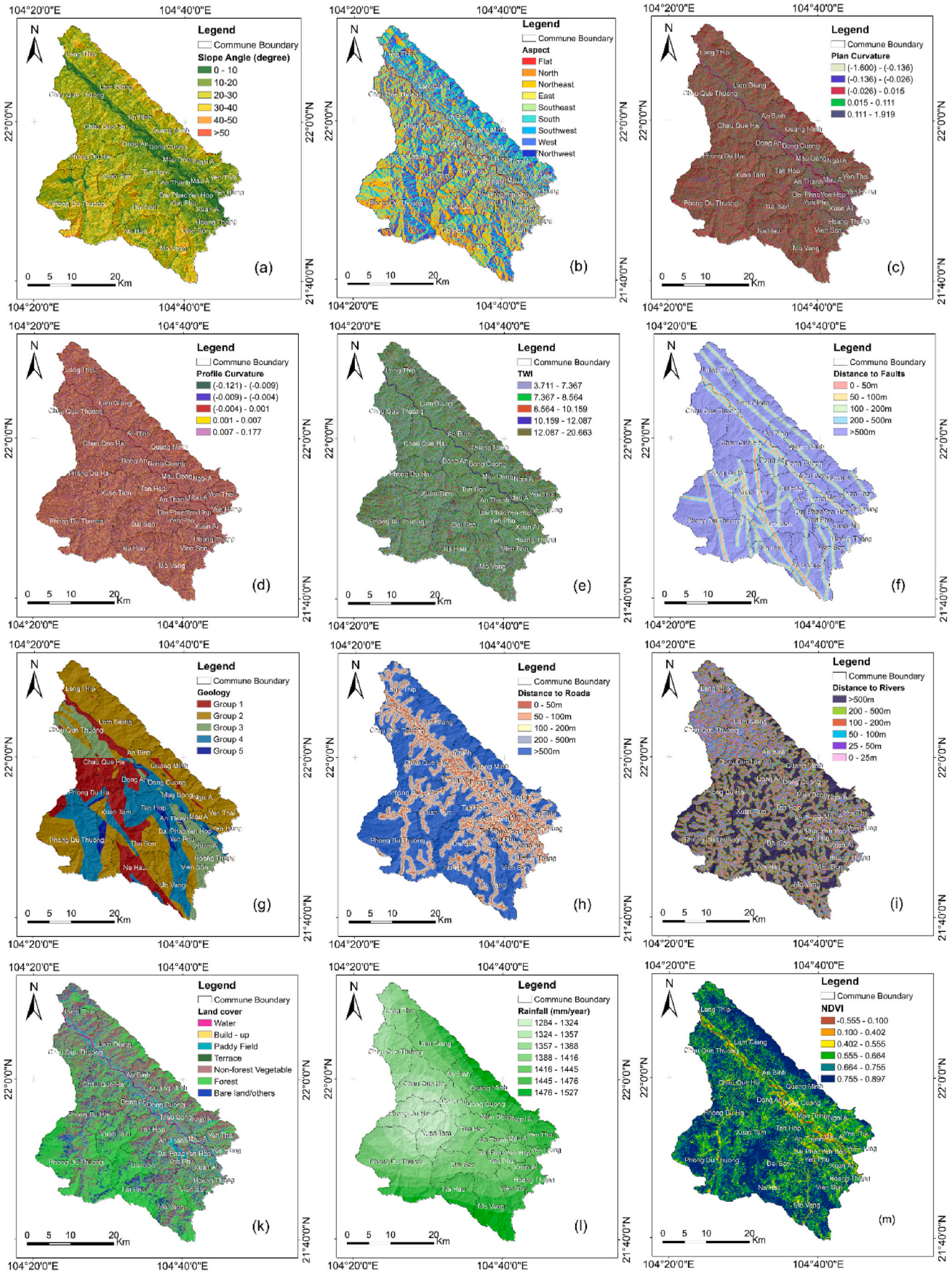


Fig. 3. Conditioning factors of the study area. (a)Slope Angle; (b)Aspect; (c)Plan Curvature; (d)Profile Curvature; (e)TWI; (f)Distance to Faults; (g)Lithology; (h) Distance to Road; (i)Distance to River; (k)Land Cover; (l)Rainfall; (m)NDVI.

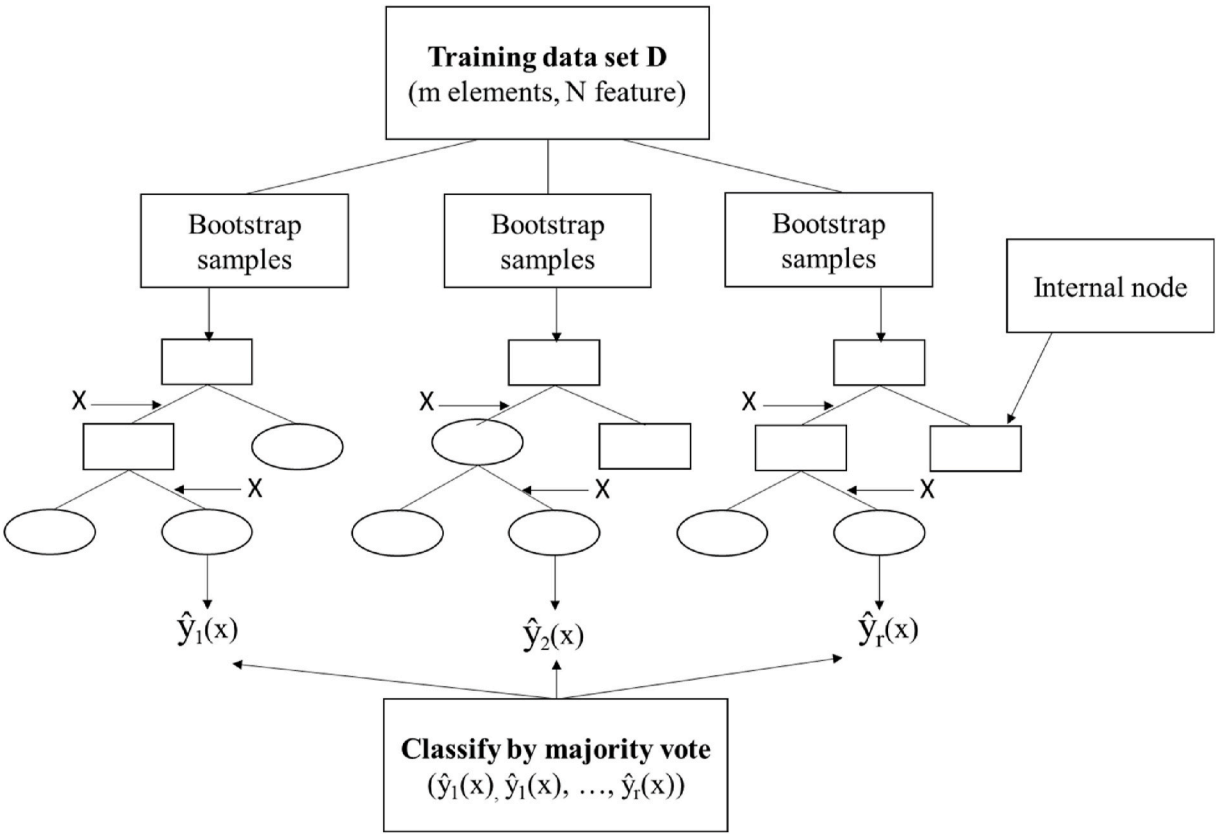


Fig. 4. Random Forest formation.

$$\mathrm{TWI} = \frac{A_s}{\tan \beta} \tag{1}$$

where A_s is the specific basin area (m^2/m) and β is topographic slope angle.

- Geological data: The geological map (Fig. 3(g)) of Van Yen District, along with distances to faults (Fig. 3(f)), were both established using a 1:200,000 scale geological map and fault data provided by the Vietnam Institute of Geosciences and Mineral Resources in 2018. The geological map includes five lithology groups, ranging from group 1 to group 5, based on increasing hardness.
- Distance to roads and distance to rivers: The distance to roads and distance to rivers maps were constructed based on the road and river map of the OpenStreetMap source in 2022. Roads can affect landslide occurrence in multiple ways. Construction and excavation activities associated with road development can alter the stability of slopes and increase the likelihood of landslides (Ha et al., 2022). Moreover, roads can act as a conduit for water, channeling rainfall and increasing water infiltration into the soil, which can contribute to slope instability. The proximity of slopes to roads may also indicate the level of human activity and potential disturbance, further influencing landslide susceptibility. For the study area, the distance to roads (Fig. 3(h)) is buffered with distances of 50 m, 100 m, 200 m and greater than 500 m. Rivers and water bodies can significantly impact landslides. Water flowing along rivers can erode and undermine slopes, reducing their stability and triggering landslides (Abedini et al., 2019). The proximity of slopes to rivers may indicate areas susceptible to erosion and increased water pressure, which can weaken the soil and contribute to slope failures (Ha et al., 2022). River valleys often have specific geological and geomorphological characteristics that can increase landslide susceptibility. In the case of Van Yen District, the distance to rivers (Fig. 3(i)) is buffered with distances of 25 m, 50 m, 100 m, 200 m, 500 m, and greater than 500 m.
- Rainfall: Precipitation plays a significant role in triggering landslides by saturating the soil, increasing pore water pressure, and reducing soil strength. The rainfall map (Fig. 3(l)) was interpolated using Kriging method from the average data of 10 years (2012–2022) from six stations located within and surrounding the study area (Bui et al., 2023). The rainfall data from these stations is provided by National Centre for Hydrometeorological Forecasting under the Ministry of Natural Resources and Environment of Vietnam.
- Land cover and vegetation density: The type of land cover, such as forests or urban areas, and the density of vegetation can influence slope stability. Vegetation plays a crucial role in enhancing slope stability by reinforcing soil cohesion and reducing the risk of landslides. In this research, a land cover map (Fig. 3(k)) was generated by classifying seven sub-classes using a classification method based on Sentinel-2 satellite imagery with a spatial resolution of 10 m. Additionally, a Normalized Difference Vegetation Index (NDVI) (Fig. 3(m)) data layer represents vegetation density with a resolution of 10 m was created from this image.

3. Research methods

3.1. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a machine learning algorithm commonly used for binary classification tasks. SVM constructs an optimal hyperplane to separate different samples by maximizing the distance between them. It can handle both linearly separable and non-linearly separable data (Cortes and Vapnik, 1995; Drucker et al., 2003).

In the case of linearly separable data, SVM seeks a hyperplane that maximizes the distance between two groups of samples and the closest samples from each group. However, in practice, data is often not linearly

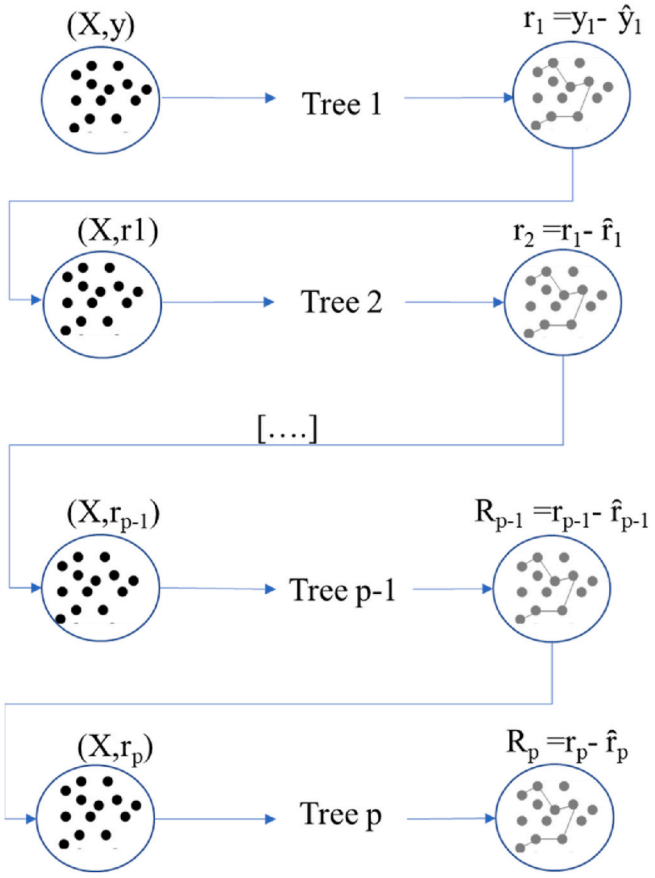


Fig. 5. Diagram of the Gradient Boosting model.

separable, so SVM is used for nonlinear data. In this scenario, relaxation variables are introduced to allow for some misclassifications, and penalty factors are used to balance the trade-off between maximizing the margin and minimizing misclassification errors (Abedini et al., 2019; Huang and Zhao, 2018; Kamran et al., 2021).

To solve the optimization problem, we use the method of Lagrange multipliers, leading to the dual formulation of the problem with a kernel function that enables SVM to operate in high-dimensional feature spaces. The commonly used kernel functions include linear, polynomial, Gaussian radial basis, and sigmoid kernels. Among these kernels, the Gaussian radial basis kernel is popular due to its ability to optimize model parameters and perform well with both large and small datasets.

3.2. Random Forest

Random Forest is an algorithm that consists of numerous individual decision trees functioning collectively (Breiman, 2001). Each decision tree in the Random Forest predicts the class, and the class with the highest number of votes is chosen as the model's final prediction (Fig. 4). The Random Forest model is highly effective for image classification because it utilizes a multitude of smaller models with diverse rules to make the ultimate decision. Although each sub-model may differ and be weak, the classification result is more accurate compared to using a single model due to the "wisdom of the crowd" principle. Random Forest employs a technique known as bagging, which means that during each split of the tree, only a limited subset of features is considered rather than all the features of the model.

As its name suggests, Random Forest (RF) is based on two key concepts: (1) Randomness, indicating the utilization of randomness, and (2) Forest, representing the presence of multiple decision trees (Sun et al., 2020; Zhou et al., 2021).

Table 1
Landslide conditioning factors and their classification.

Factor	Classification	Classification method
Elevation (m)	(1) [0–169.55], (2) [169.55–370.94], (3) [370.94–587.24], (4) [587.24–840.84], (5) [840.84–1176.49], (6) >1176.49	Natural Breaks
Slope angle (degree)	(1) 0–10, (2) 10–20, (3) 20–30, (4) 30–40, (5) 40–50, (6) > 50	Manual Classification
Aspect	(1) Flat (−1), (2) North (0–22.5), (3) Northeast (22.5–67.5), (4) East (67.5–112.5), (5) Southeast (112.5–157.5), (6) South (157.5–202.5), (7) Southwest (202.5–247.5), (8) West (247.5–292.5), (9) Northwest (292.5–337.5), (2) North (337.5–360)	Azimuth
Plan Curvature	(1) [(-1.6) - (−0.136)], (2) [(-0.136) - (0.026)], (3) [(-0.026) - 0.015)], (4) [(0.015–0.111], (5) [0.111–1.919]	Natural Breaks
Profile Curvature	(1) [(-1.121) - (−0.009)], (2) [(-0.009) - (−0.004)], (3) [(-0.009) - 0.002)], (4) [(0.002–0.007], (5) [0.007–0.177]	Natural Breaks
TWI	(1) [3.711–7.367], (2) [7.367–8.564], (3) [8.564–10.159], (4) [10.159–12.087], (5) [12.087–20.663]	Natural Breaks
Distance to faults	(1) [0–50], (2) [50–100], (3) [100–200], (4) [200–500], (5) >500	Manual Classification
Lithology	(1) Group 1, (2) Group 2, (3) Group 3, (4) Group 4, (5) Group 5	Lithology categories
Distance to road	(1) [0–50], (2) [50–100], (3) [100–200], (4) [200–500], (5) >500	Manual Classification
Distance to river	(1) [0–25], (2) [25–50], (3) [50–100], (4) [100–200], (5) [200–500], (6) >500	Manual Classification
Land cover	(1) Water, (2) Build-up, (3) Paddy field, (4) Terrace, (5) Non-forest vegetable, (6) Forest, (7) Bare land/others	Land cover categories
Rainfall	(1) [1284–1324], (2) [1324–1357], (3) [1357–1388], (4) [1388–1416], (5) [1416–1445], (6) [1445–1476], (7) [1476–1527]	Natural Breaks
NDVI	(1) [(-0.555)-0.100], (2) [0.100–0.402], (3) [0.402–0.555], (4) [0.555–0.664], (5) [0.664–0.755], (6) [0.755–0.897]	Natural Breaks

3.3. Gradient boosting

Gradient Boosting (GB) is based on the idea of creating a sequence of weak models that complement each other's shortcomings (Friedman et al., 2000). In Boosting, subsequent models strive to minimize the errors made by the previous models. The concept of gradient boosting involves three key steps. Firstly, a suitable differentiable loss function is chosen based on the problem at hand. One advantage of the gradient boosting model is that it can accommodate different loss functions without requiring new algorithms. It only requires selecting an appropriate loss function and incorporating it into the gradient boosting framework. Secondly, a weak learner is constructed to make predictions. In gradient boosting, decision trees are commonly used as weak learners. Specifically, regression trees are employed, generating continuous output for splits that can be aggregated. This allows the combination of outputs from different models, leading to improved predictions by progressively refining the residuals. The trees are built in a greedy manner, often subject to certain constraints to ensure they remain weak learners and can still be constructed using a greedy approach (Fig. 5). Finally, an additive model is created by aggregating the predictions of the weak learners to minimize the loss function. This process of adding trees occurs incrementally, one tree at a time. The output generated by each new tree is added to the outputs of the previously constructed sequence of trees, enhancing the final model's output. This iterative process continues until the optimized value for the loss function is achieved, signaling the completion of the gradient boosting process (Iban and Bilgilioglu, 2023; Sahin, 2022).

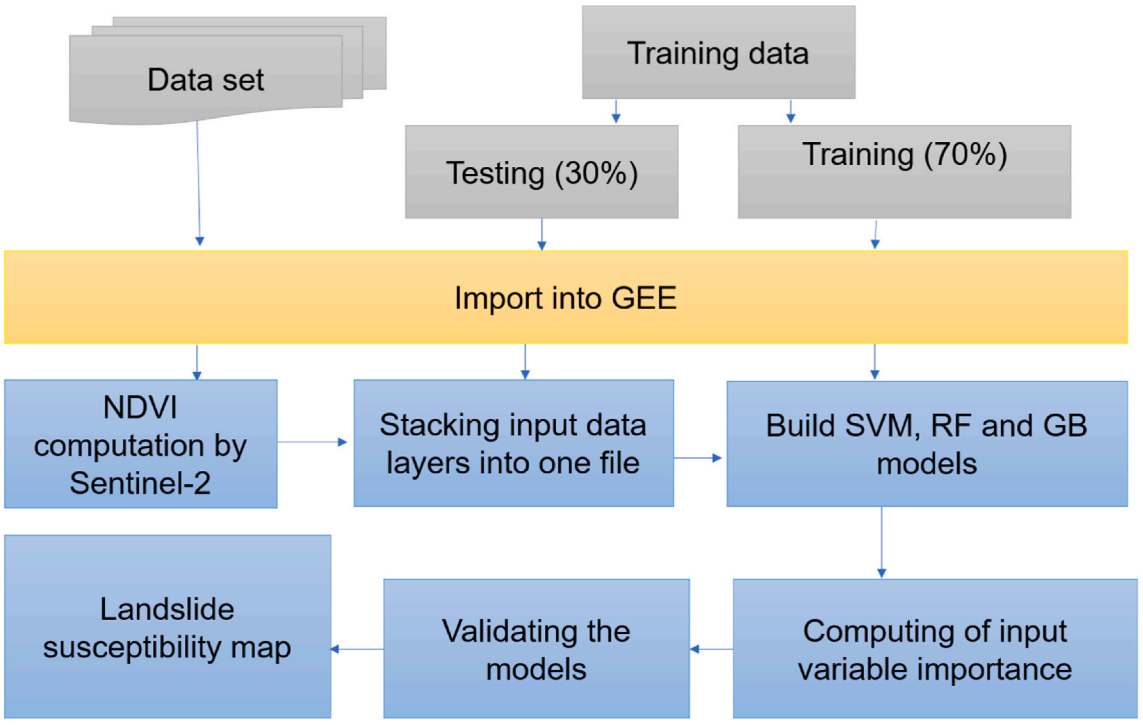


Fig. 6. Flow chart of image processing and building predictive models by methods SVM, RF and GB.

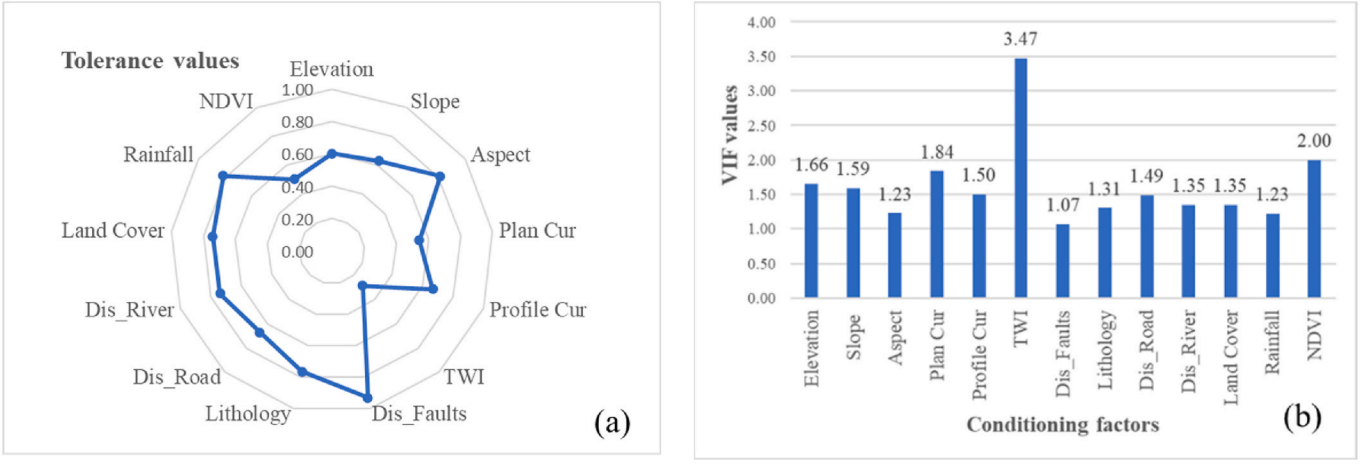


Fig. 7. Multicollinearity assessment results based on VIF and Tolerance Values.

$$min_{C_n} = 1:\mathrm{N}, w_n = 1:N^{L\left(\left(y,\sum_{n=1}^{N} C_n W_n\right)\right)} \quad (2)$$

where:

- L: loss function value
- y: label
- c_n: is the weight of the n weak learner
- w_n: n weak learner

3.4. Tool for processing

The Google Earth Engine (GEE) cloud computing platform is utilized as the tool for constructing the predictive model. GEE combines numerous satellite images and geospatial data from diverse sources to serve as input data for the model. This approach allows for reducing desktop data preparation, thereby reducing data preparation time for model input. Table 1 and Fig. 3 provide a summary of the datasets, including those prepared on the desktop and obtained from the cloud.

The training set consists of 248 landslide points and 248 no landslide points, from which values for elevation, slope, aspect, NDVI, LULC, TWI, rainfall, plan curvature, profile curvature, lithology, distance to roads, distance to rivers, distance to fault, and landslide location are extracted. The values are labeled as either 1 (slides) or 0 (no landslide). The flowchart in Fig. 6 illustrates the process of image processing and building the predictive model using three methods: SVM, RF, and GB.

4. Modelling prediction and performance

The performance of a predictive model is evaluated using the ROC curve and the area under the ROC curve (AUC). The ROC curve illustrates the relationship between Sensitivity and (1-Speccificity) for both landslide and non-landslide positions. The AUC serves as a comprehensive performance metric for land subsidence prediction models. A

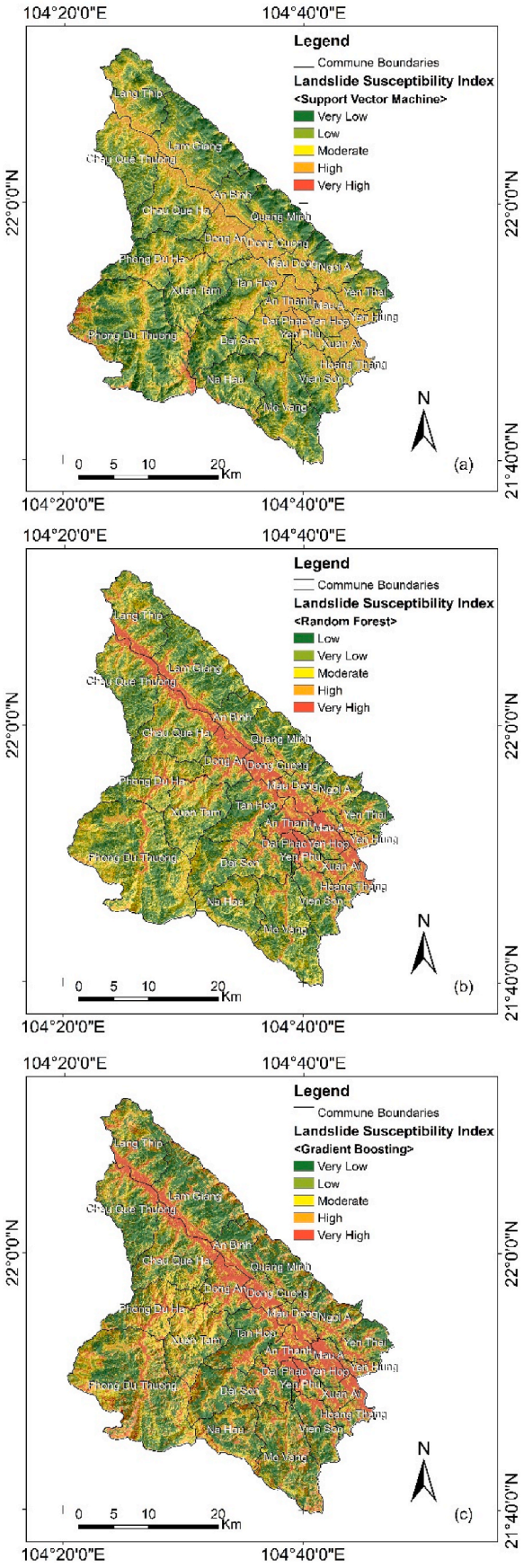


Fig. 8. Landslide susceptibility map predicted by models: (a) Support Vector Machine, (b) Random Forest, (c) Gradient Boosting.

higher AUC value indicates a more effective model.

The sensitivity is computed using the following formula:

$$\text{Sensitivity} = \frac{TP}{(TP + FN)} \tag{3}$$

where TP represents the count of True Positives, and FN denotes the number of False Negatives.

Specificity can be computed using the formula below:

$$\text{Specificity} = \frac{TN}{(TN + FP)} \tag{4}$$

where TN corresponds to the true negatives, and FP represents the count of false positives.

The overall accuracy (OA) is calculated based on all four indices: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). The formula for computing the overall accuracy (OA) is given as follows:

$$\text{OA} = \frac{TP + TN}{(TP + TN + FP + FN)} \tag{5}$$

The ROC, AUC, and OA curves are valuable tools for evaluating landslide models. The ROC curve enables the assessment of model performance based on sensitivity and specificity. Sensitivity represents the proportion of predicted landslides that are correctly identified as landslides, while specificity denotes the proportion of non-landslide areas that are accurately classified as such. Moreover, OA provides a measure of the model's overall accuracy by comparing correct scores with the total number of testing points.

These assessment methods not only help verify the reliability of results but also facilitate model comparison and selection for landslide prediction purposes. Researchers can use ROC and AUC curves to gain insights into model performance and make informed decisions when choosing the most suitable model for their specific landslide prediction needs.

5. Results and discussions

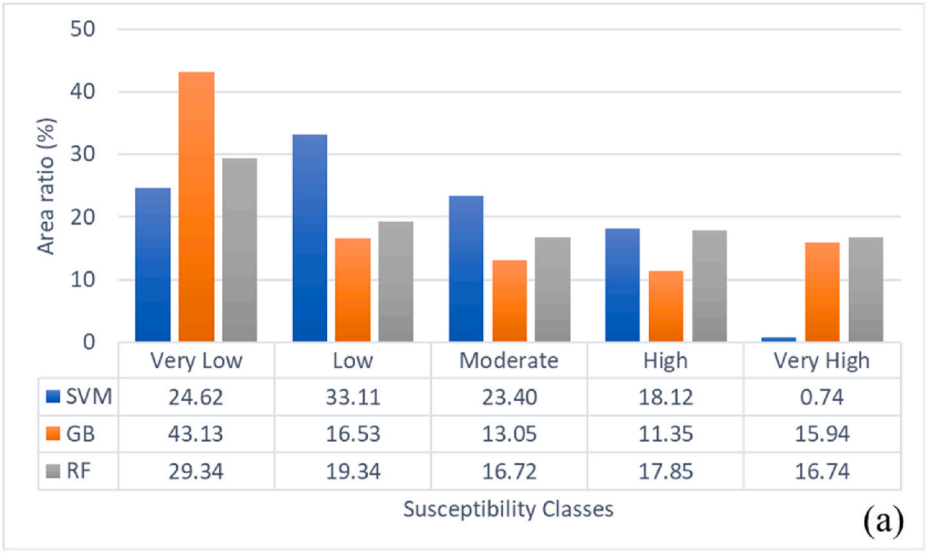
5.1. Multicollinearity analysis

Multicollinearity is a phenomenon in which factors within a model are correlated and influence each other, leading to a reduction in the explanatory power of variables. In statistical research, multicollinearity is considered to occur when the Variance Inflation Factor (VIF) exceeds 10 or when the tolerance is less than 0.1. Fig. 7 illustrates the values of Tolerance and VIF in assessing multicollinearity for 13 factors in the landslide model within the study area (Daviran et al., 2023).

The Tolerance values indicate a low degree of correlation among the factors, as all of them are greater than 0.1 (Fig. 7(a)). Conversely, the VIF values for the factors range from 1.07 (Distance to Faults) to 3.47 (Topographic Wetness Index), ensuring that they are all less than 10 (Fig. 7(b)). This suggests the absence of severe multicollinearity among the factors in the model. Consequently, all the landslide conditioning factors in this case are retained to construct the landslide susceptibility prediction model.

5.2. Landslide susceptibility models

The landslide susceptibility zoning map demonstrates consistency and similarity in classifying the landslide sensitivity levels among all three models. Particularly, areas with high and very high landslide sensitivity tend to be concentrated along transportation routes. Based on Fig. 8, it can be observed that areas with high and very high landslide susceptibility tend to be in close proximity to transportation routes, such as along the national road CT05 and the provincial road DT163. Additionally, significant landslide-prone areas are identified in Chau Que Ha,



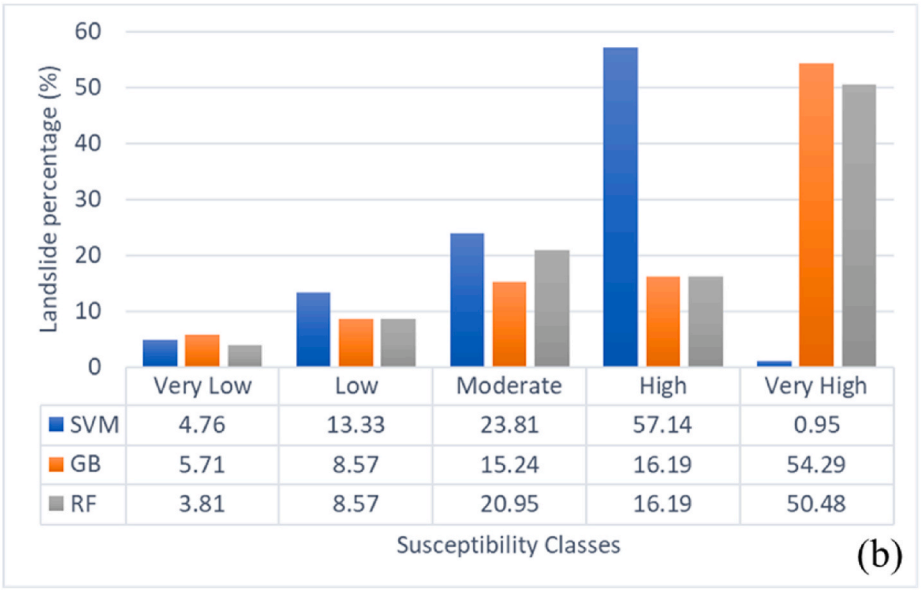


Fig. 9. Distribution charts (a) percentage of susceptibility level classes, (b) percentage of landslide survey points in the testing dataset by three models.

Phong Du Ha, and Phong Du Thuong, which are characterized by high slopes beside transportation road.

Fig. 9(a) illustrates the percentage distribution of susceptibility level classes for all three models: Support Vector Machine (SVM), Gradient Boosting (GB), and Random Forest (RF). The results show that the models tend to distribute the susceptibility levels differently. Specifically, for the "Very Low" class, the percentages are 24.62% for SVM, 43.13% for GB, and 29.34% for RF. Conversely, for the "Very High" class, the percentages are 0.74% for SVM, 15.94% for GB, and 16.74% for RF. It can be observed that SVM has the highest percentage of area for the "Very Low" susceptibility level and the lowest percentage for the "Very High" susceptibility level. GB demonstrates a higher capability of predicting samples in the "Very Low" susceptibility level compared to SVM, but this decreases as it predicts samples in the "Very High" susceptibility level. On the other hand, RF shows a higher capability of predicting samples in the "High" susceptibility level compared to GB, with percentages of 17.85% and 11.35%, respectively.

Fig. 9(b) displays the percentage distribution of landslide survey points in the testing dataset across susceptibility level classes for all three models: Support Vector Machine (SVM), Gradient Boosting (GB), and Random Forest (RF). The results present the percentages of survey points allocated to each susceptibility class, such as "Very Low," "Low," "Moderate," "High," and "Very High." The findings clearly indicate differences in the distribution of points among the models. SVM tends to predict samples into the "Low," "Moderate," and "High" classes the most, with distribution percentages of 13.33%, 23.81%, and 57.14%, respectively. GB exhibits the highest percentage of points allocated to the "Very High" class, with a distribution rate of 54.29%, while the RF model also demonstrates a similar tendency, with 50.48% of points distributed to the "Very High" class.

The distribution percentages also highlight distinctions between the models, as evidenced by SVM having the lowest percentage in the "Very High" class at 0.95%, whereas GB and RF both have higher percentages at 54.29% and 50.48%, respectively, in the same class. These distribution patterns indicate the forecasting performance of the models for landslide occurrences. Landslide points are more likely to occur in areas with "High" susceptibility levels when predicted using SVM, whereas they tend to occur in the "Very High" susceptibility class when predicted with GB and RF models.

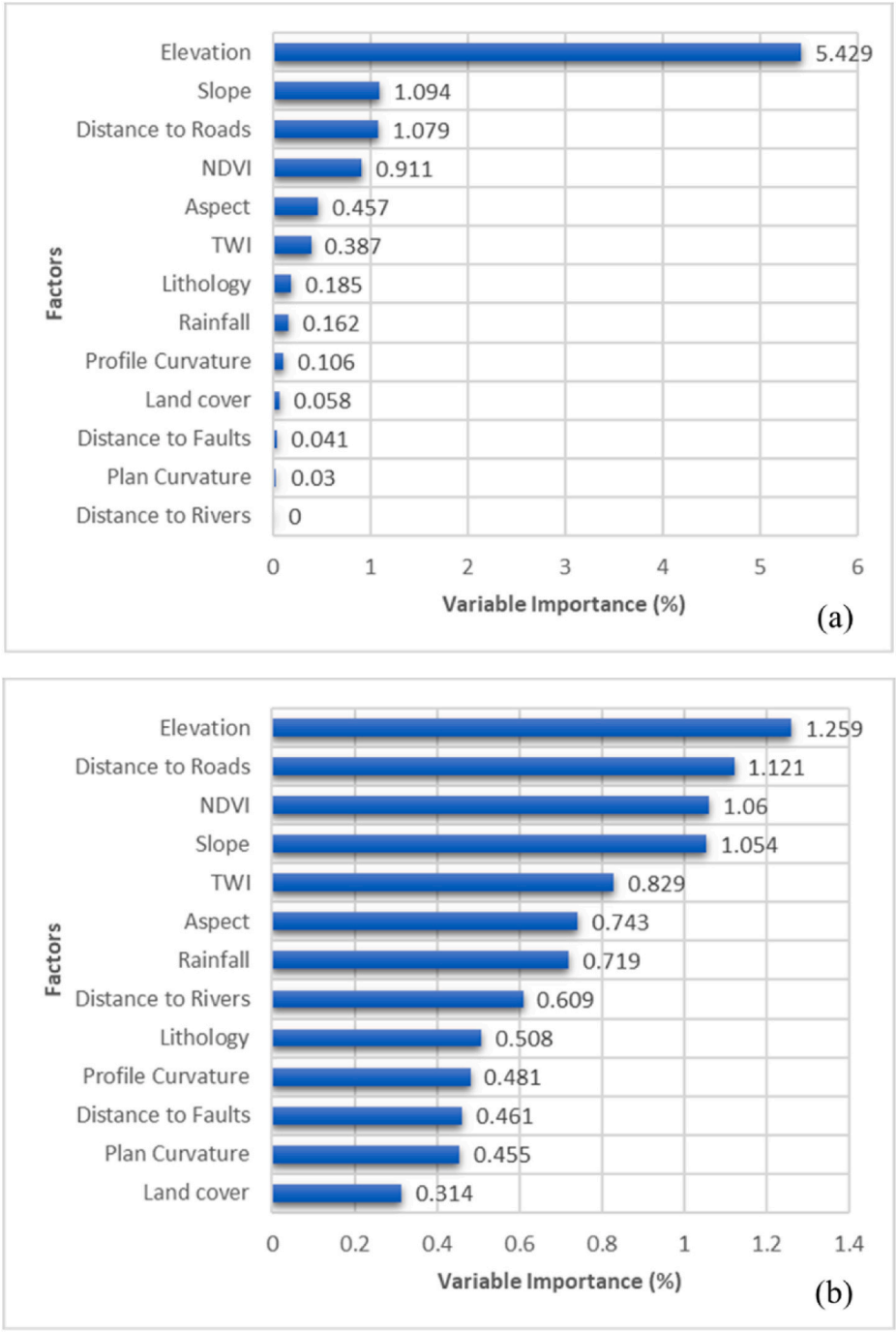


Fig. 10. Variable importance with models: (a) Random Forest, (b) Gradient Boosting.

Table 2
Model performance assessment parameters.

	Gradient Boosting	Random Forest	Support Vector Machine
TP	82	75	64
TN	82	93	92
FP	24	13	14
FN	23	30	41
Overall Accuracy	0.777	0.796	0.739
Sensitivity	0.781	0.714	0.610
Specificity	0.774	0.877	0.868
AUC	0.861	0.883	0.815

5.3. Variables contribution analysis

The results from both models Gradient Boosting and Random Forest provide insights into the importance of factors in predicting landslide occurrences. The Gradient Boosting model reveals that Elevation and Slope are the two most significant factors, with importance values of 5.429 and 1.094, respectively (Fig. 10(a)). This highlights the strong influence of terrain elevation and slope on the likelihood of landslides. Additionally, factors such as Distance to Roads, NDVI, Aspect, and TWI also contribute significantly, with importance values ranging from 0.387 to 1.079. In contrast, Plan Curvature and Distance to Rivers have the lowest importance values and contribute minimally to the prediction model.

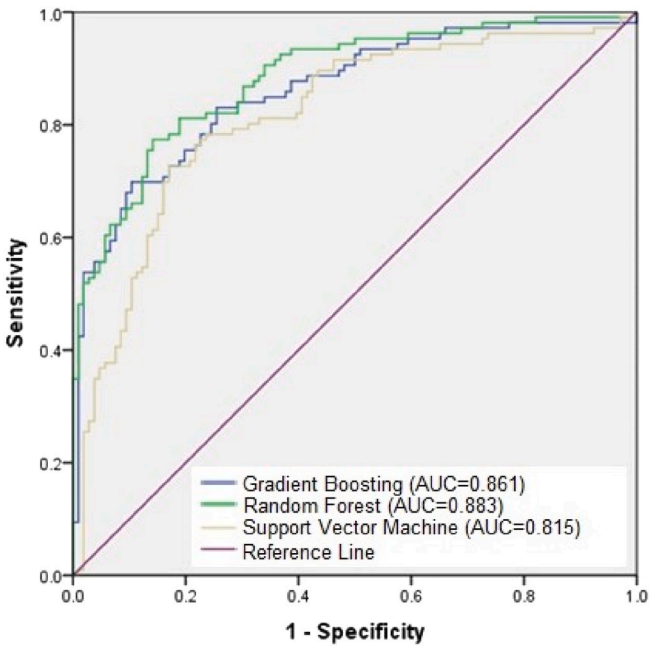


Fig. 11. ROC and AUC of models.

Random Forest model identifies Elevation and Distance to Roads as the two most important factors, with corresponding importance values of 1.259 and 1.121 (Fig. 10(b)). Overall, the importance values of Random Forest tend to be lower compared to Gradient Boosting. This difference might arise from how Random Forest operates by creating multiple independent decision trees and combining their results. Factors such as NDVI, Slope, TWI, Aspect, and Rainfall make substantial contributions, displaying importance values within the range of 0.719–1.060. Among these factors, Plan Curvature and Land Cover exhibit the lowest levels of importance.

Although there are variations in the order and degree of importance of factors between the two models, both models agree that Elevation and Slope are crucial in evaluating landslide sensitivity. These discrepancies may be attributed to the different ways the two models process data and determine the importance of factors. This emphasizes the necessity of employing multiple models in assessments to arrive at reliable conclusions and support effective decision-making in landslide risk management and prediction.

5.4. Model quality assessment

In this study, three machine learning models, namely Support Vector Machine (SVM), Gradient Boosting, and Random Forest, were compared for predicting landslides using the Overall Accuracy, Sensitivity, Specificity, and Area Under the Curve (AUC) for testing dataset (Table 2). The results indicate that Gradient Boosting and Random Forest outperformed SVM in prediction accuracy. The overall accuracy of SVM was 0.777, while that of Gradient Boosting and Random Forest was 0.739 and 0.796, respectively. This demonstrates that both Gradient Boosting and Random Forest achieved higher prediction accuracy compared to SVM.

Table 2 shows that the Random Forest model has a sensitivity of 0.714, which is lower than Gradient Boosting (0.781) but higher than SVM (0.610). However, all three models showed good performance in detecting landslides. The specificity of the Random Forest model was 0.877, higher than both Gradient Boosting (0.774) and SVM (0.868). This indicates that Random Forest had a better ability to detect non-landslide cases compared to the other two models.

Finally, to assess the overall model performance, we used the Area Under Curve (AUC). The results showed that Random Forest had the highest AUC of 0.883, followed by Gradient Boosting with an AUC of 0.861, and SVM had the lowest AUC of 0.815 (Fig. 11). This indicates that Random Forest performed best among the three tested models in terms of prediction accuracy.

6. Conclusion

The study utilized data from 13 classes of factors influencing landslides, along with field survey data on landslides combined with PS-InSAR analysis from Sentinel-1 images to create landslide susceptibility maps using Support Vector Machine (SVM), Random Forest (RF), and Gradient Boosting (GB) machine learning models. The research revealed that high landslide susceptibility occurred in areas along major transportation routes and roads in Chau Que Ha, Phong Du Ha, and Phong Du Thuong communes. The results demonstrated that the application of combined models yielded more optimized effectiveness in generating landslide susceptibility maps, with the Random Forest model showing the best performance compared to the Gradient Boosting and the Support Vector Machine, with AUC values of 0.883, 0.861, and 0.815, respectively. Furthermore, the GEE cloud-based tool proves highly effective by harnessing numerous available cloud-based data sources and leveraging highly efficient machine learning algorithms for constructing predictive models.

CRediT authorship contribution statement

Van Anh Tran: Writing – review & editing, Writing – original draft, Visualization, Software, Conceptualization. **Thanh Dong Khuc:** Writing – original draft, Visualization, Software, Data curation, Conceptualization. **Xuan Quang Truong:** Writing – review & editing, Software. **An Binh Nguyen:** Software. **Truong Thanh Phi:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgment

This study is conducted based on the individual contributions of the authors, including data collection, data processing, building, and revising manuscripts.

References

Abedini, M., Ghasemian, B., Shirzadi, A., Bui, D.T., 2019. A comparative study of support vector machine and logistic model tree classifiers for shallow landslide susceptibility modeling. Environ. Earth Sci. 78, 560. https://doi.org/10.1007/s12665-019-8562-z.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Bui, Q.D., Ha, H., Khuc, D.T., Nguyen, D.Q., von Meding, J., Nguyen, L.P., Luu, C., 2023. Landslide susceptibility prediction mapping with advanced ensemble models: son La province, Vietnam. Nat. Hazards 116, 2283–2309. https://doi.org/10.1007/s11069-022-05764-3.

Chang, Z., Catani, F., Huang, F., Liu, G., Meena, S.R., Huang, J., Zhou, C., 2023. Landslide susceptibility prediction using slope unit-based machine learning models considering the heterogeneity of conditioning factors. J. Rock Mech. Geotech. Eng. 15, 1127–1143. https://doi.org/10.1016/j.jrmge.2022.07.009.

Chen, W., Xie, X., Wang, J., Pradhan, B., Hong, H., Bui, D.T., Duan, Z., Ma, J., 2017. A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility. Catena 151, 147–160. https://doi.org/10.1016/j.catena.2016.11.032.

Ciampalini, A., Raspini, F., Frodella, W., Bardi, F., Bianchini, S., Moretti, S., 2016. The effectiveness of high-resolution LiDAR data combined with PSInSAR data in

landslide study. Landslides 13, 399–410. https://doi.org/10.1007/s10346-015-0663-5.
Colesanti, C., Ferretti, A., Prati, C., Rocca, F., 2003. Monitoring landslides and tectonic motions with the permanent scatterers technique. Eng. Geol. 68, 3–14. https://doi.org/10.1016/S0013-7952(02)00195-3.
Colesanti, C., Wasowski, J., 2006. Investigating landslides with space-borne synthetic aperture radar (SAR) interferometry. Eng. Geol. 88, 173–199. https://doi.org/10.1016/j.enggeo.2006.09.013.
Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20, 273–297. https://doi.org/10.1007/BF00994018.
Daviran, M., Shamekhi, M., Ghezelbash, R., Maghsoudi, A., 2023. Landslide susceptibility prediction using artificial neural networks, SVMs and random forest: hyperparameters tuning by genetic optimization algorithm. Int. J. Environ. Sci. Technol. 20, 259–276. https://doi.org/10.1007/s13762-022-04491-3.
Drucker, H., C, C., Kaufman, L., Smola, A., Vapnik, V., 2003. Support vector regression machines. Adv. Neural Inf. Process. Syst. 9.
Fang, Z., Wang, Y., Peng, L., Hong, H., 2021. A comparative study of heterogeneous ensemble-learning techniques for landslide susceptibility mapping. Int. J. Geogr. Inf. Sci. 35, 321–347. https://doi.org/10.1080/13658816.2020.1808897.
Friedman, J., Hastie, T., Tibshirani, R., 2000. Additive logistic regression: a statistical view of boosting. Ann. Stat. 28, 337–407. https://doi.org/10.1214/aos/1016218223.
Froude, M.J., Petley, D.N., 2018. Global fatal landslide occurrence from 2004 to 2016. Nat. Hazards Earth Syst. Sci. 18, 2161–2181. https://doi.org/10.5194/nhess-18-2161-2018.
Ha, D.T., Trang, B.T.T., Thanh, N.K., 2018. Study on mapping landslide susceptibility for Van Yen district, Yen Bai province. Vietnam Sci. Technol. J. 61.
Ha, H., Bui, Q.D., Khuc, T.D., Tran, D.T., Pham, B.T., Mai, S.H., Nguyen, L.P., Luu, C., 2022. A machine learning approach in spatial predicting of landslides and flash flood susceptible zones for a road network. Model. Earth Syst. Environ. 8, 4341–4357. https://doi.org/10.1007/s40808-022-01384-9.
He, Q., Jiang, Z., Wang, M., Liu, K., 2021. Landslide and wildfire susceptibility assessment in Southeast Asia using ensemble machine learning methods. Rem. Sens. https://doi.org/10.3390/rs13081572.
Heckmann, T., Gegg, K., Gegg, A., Becht, M., 2013. Sample size matters: investigating the effect of sample size on a logistic regression debris flow susceptibility model. Nat. Hazards Earth Syst. Sci. Discuss. 1, 2731–2779. https://doi.org/10.5194/nhessd-1-2731-2013.
Huang, Y., Zhao, L., 2018. Review on landslide susceptibility mapping using support vector machines. Catena 165, 520–529. https://doi.org/10.1016/j.catena.2018.03.003.
Iban, M.C., Bilgilioglu, S.S., 2023. Snow avalanche susceptibility mapping using novel tree-based machine learning algorithms (XGBoost, NGBoost, and LightGBM) with eXplainable Artificial Intelligence (XAI) approach. Stoch. Environ. Res. Risk Assess. 37, 2243–2270. https://doi.org/10.1007/s00477-023-02392-6.
Kalantar, B., Pradhan, B., Naghibi, A., Motevalli, A., Mansor, S., 2018. Assessment of the effects of training data selection on the landslide susceptibility mapping: a comparison between support vector machine (SVM), logistic regression (LR) and artificial neural networks (ANN). Geomatics. Nat. Hazards Risk. https://doi.org/10.1080/19475705.2017.1407368.
Kamran, K.V., Feizizadeh, B., Khorrami, B., Ebadi, Y., 2021. A comparative approach of support vector machine kernel functions for GIS-based landslide susceptibility mapping. Appl. Geomatics 13, 837–851. https://doi.org/10.1007/s12518-021-00393-0.
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444. https://doi.org/10.1038/nature14539.
Lee, S., Ryu, J.-H., Lee, M.-J., Won, J.-S., 2003. Use of an artificial neural network for analysis of the susceptibility to landslides at Boun, Korea. Environ. Geol. 44, 820–833. https://doi.org/10.1007/s00254-003-0825-y.
Oliveira, S., Zêzere, J., Catalao, J., Nico, G., 2014. The contribution of PSInSAR interferometry to landslide hazard in weak rock-dominated areas. Landslides 12. https://doi.org/10.1007/s10346-014-0522-9.
Petley, D., 2010. On the impact of climate change and population growth on the occurrence of fatal landslides in South, East and SE Asia. Q. J. Eng. Geol. Hydrogeol. 43, 487–496. https://doi.org/10.1144/1470-9236/09-001.
Pham, B.T., Prakash, I., Singh, S.K., Shirzadi, A., Shahabi, H., Tran, T.-T.-T., Bui, D.T., 2019. Landslide susceptibility modeling using Reduced Error Pruning Trees and different ensemble techniques: hybrid machine learning approaches. Catena 175, 203–218. https://doi.org/10.1016/j.catena.2018.12.018.
Reichenbach, P., Rossi, M., Malamud, B.D., Mihir, M., Guzzetti, F., 2018. A review of statistically-based landslide susceptibility models. Earth Sci. Rev. 180, 60–91. https://doi.org/10.1016/j.earscirev.2018.03.001.
Sahin, E.K., 2022. Comparative analysis of gradient boosting algorithms for landslide susceptibility mapping. Geocarto Int. 37, 2441–2465. https://doi.org/10.1080/10106049.2020.1831623.
Saito, H., Nakayama, D., Matsuyama, H., 2009. Comparison of landslide susceptibility based on a decision-tree model and actual landslide occurrence: the Akaishi Mountains, Japan. Geomorphology 109, 108–121. https://doi.org/10.1016/j.geomorph.2009.02.026.
Shin, H., Roth, H., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., Summers, R., 2016. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans. Med. Imag. 35 https://doi.org/10.1109/TMI.2016.2528162.
Sun, D., Wen, H., Wang, D., Xu, J., 2020. A random forest model of landslide susceptibility mapping based on hyperparameter optimization using Bayes algorithm. Geomorphology 362, 107201. https://doi.org/10.1016/j.geomorph.2020.107201.
Thanh Thi Pham, N., Nong, D., Raghavan Sathyan, A., Garschagen, M., 2020. Vulnerability assessment of households to flash floods and landslides in the poor upland regions of Vietnam. Clim. Risk Manag. 28, 100215 https://doi.org/10.1016/j.crm.2020.100215.
Tran, V.A., Truong, X.Q., Nguyen, D.A., Longoni, L., Yordanov, V., 2021. Landslides monitoring with time series of sentinel-1 imagery in yen Bai province-vietnam. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. - ISPRS Arch. 46, 197–203. https://doi.org/10.5194/isprs-archives-XLVI-4-W2-2021-197-2021.
Yazici, B.V., Tunc Gormus, E., 2022. Investigating persistent scatterer InSAR (PSInSAR) technique efficiency for landslides mapping: a case study in Artvin dam area, in Turkey. Geocarto Int. 37, 2293–2311. https://doi.org/10.1080/10106049.2020.1818854.
Yordanov, V., Biagi, L., Truong, X., Tran, V., Brovelli, M., 2021. An overview of geoinformatics state-of-the-art techniques for landslide monitoring and mapping. Int. Arch. Photogram. Rem. Sens. Spatial Inf. Sci. 205–212. https://doi.org/10.5194/isprs-archives-XLVI-4-W2-2021-205-2021.
Zhou, X., Wen, H., Zhang, Y., Xu, J., Zhang, W., 2021. Landslide susceptibility mapping using hybrid random forest with GeoDetector and RFE for factor optimization. Geosci. Front. 12, 101211 https://doi.org/10.1016/j.gsf.2021.101211.