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

UNIVERSITY OF TRANSPORT TECHNOLOGY

Detection of extreme rainfall-induced landslides in mountainous areas using ensemble machine learning and Sentinel-2 satellite image analysis

Viet Long Doan[1*], Thanh Thien Le[1], Pham Phuoc Thanh[2]

[1]The University of Danang, University of Science and Technology, Danang, Vietnam; dvlong@dut.udn.vn

[2]Geotechnical Engineering and Artificial Intelligence research group, University of Transport Technology, 54 Trieu Khuc, Thanh Liet, Hanoi, Vietnam; thanhphamphuoc82@gmail.com

Article info
Type of article:
Original research paper

DOI:
https://doi.org/10.58845/jstt.utt.2026.en.6.3.239-257

***Corresponding author:**
Email address:
dvlong@dut.udn.vn

Received: 19/05/2026
Received in Revised Form: 18/06/2026
Accepted: 05/07/2026

Abstract: Landslide detection is a fundamental prerequisite for subsequent investigations, including landslide susceptibility mapping and landslide risk assessment. In mountainous regions with complex terrain conditions, this task is particularly challenging due to the limited accessibility of landslide sites for direct field investigation and data acquisition. Recent advances in remote sensing image analysis have enabled efficient detection and extraction of landslide information through the analysis of changes in land cover and topographic characteristics derived from pre- and post-event imagery. Nevertheless, the application of remote sensing techniques still faces several limitations, particularly in discriminating actual landslide areas from vegetation loss caused by anthropogenic activities such as deforestation and infrastructure construction. To overcome these limitations, this study applied ensemble machine learning approaches using satellite imagery acquired from the Sentinel-2 satellite to detect landslide locations. Pre- and post-event satellite images associated with the extreme rainfall-induced landslide events that occurred in late October 2020 in the mountainous region of Phuoc Son were collected and analyzed. In addition to satellite imagery, several conditioning factors, including slope and rainfall, were incorporated into the models to minimize the misclassification of non-landslide areas as landslides. The training and validation results obtained from three models, namely Logistic Regression, Random Forest, and XGBoost (eXtreme Gradient Boosting), demonstrated excellent predictive capability for all models. Among them, the XGBoost model achieved the highest predictive performance, with an AUC value of 0.999 on the validation dataset. Subsequently, the XGBoost model was employed to identify landslide locations across the entire study area. Validation based on field survey data and Google Earth imagery further confirmed the high effectiveness and reliability of the proposed approach.

Keywords: Landslide detection, XGBoost, Random Forest, Logistic Regression, Sentinel 2.

1. Introduction

Landslides are among the most widespread natural hazards worldwide, causing severe damage to infrastructure and substantial loss of human life [1], [2]. The collection of landslide information, including location and occurrence time, enables researchers to identify triggering factors and develop landslide susceptibility, hazard, and risk maps, thereby supporting disaster prevention and mitigation efforts for this highly destructive hazard [3]. Information on landslide locations can be obtained through field surveys [4], aerial image interpretation, or satellite image analysis [5]. Among these approaches, field surveys are considered the most reliable method for directly identifying and mapping landslide locations. However, access to landslide sites in mountainous regions characterized by complex terrain is often extremely difficult [4]. Furthermore, investigating landslides that occurred in the distant past presents additional challenges, as surface conditions may have changed over time due to processes such as vegetation regrowth at the landslide sites. To overcome these limitations, aerial image analysis using unmanned aerial vehicles (UAVs) has been increasingly adopted and has demonstrated high effectiveness [5]. Nevertheless, this approach requires substantial financial resources when applied over large areas. In addition, flight permission regulations may pose further constraints on its practical implementation [5].

For studies aimed at identifying landslide locations over large areas while maintaining cost efficiency, satellite image analysis represents an effective alternative [6]. Currently, two principal types of satellite imagery are widely utilized for this purpose, namely radar imagery and optical imagery [6]. Radar imagery is acquired from Synthetic Aperture Radar (SAR) satellites, including Sentinel-1, ERS, ENVISAT, ALOS PALSAR, and RADARSAT [7]. In contrast, optical imagery can be obtained from satellites such as Landsat and Sentinel-2 [3], [8]. With recent technological advancements, optical imagery offers several advantages over radar imagery, including: (i) intuitive visualization that facilitates the identification of surface features such as vegetation, soil, buildings, infrastructure, and water bodies; (ii) relatively simple analytical procedures requiring fewer preprocessing steps, whereas radar imagery often involves complex processing techniques for noise reduction and terrain correction; (iii) higher spatial resolution, enabling more effective object detection and classification; and (iv) the availability of numerous freely accessible datasets, making optical imagery highly suitable for large-scale investigations and long-term monitoring [9]. Owing to these advantages, optical imagery has been extensively applied in landslide detection studies [10].

To date, numerous studies have investigated landslide detection using optical imagery in combination with various analytical techniques, including Change Vector Analysis (CVA) [11], the Normalized Difference Vegetation Index (NDVI), Principal Component Analysis (PCA), and Independent Component Analysis (ICA) [3]. Several studies have reported that NDVI-based change detection approaches generally achieve higher performance than other analytical methods [3], [12]. NDVI analysis enables effective landslide detection through the identification of substantial variations in NDVI values or vegetation loss between pre-event and post-event images [13]. Nevertheless, this approach still presents several limitations, including the inability to detect landslides occurring in non-vegetated areas and the misclassification of land-cover changes caused by anthropogenic activities such as deforestation and construction works. To overcome these limitations, machine learning techniques incorporating multiple input datasets can be employed to reduce noise and improve the accuracy of landslide detection analyses [14]. Among these approaches, ensemble learning models such as Random Forest and XGBoost have demonstrated superior classification performance compared with conventional standalone machine learning models [15], [16].

Landslides occur in various forms, including slides, falls, topples, and flows, and may be triggered by multiple factors such as rainfall, earthquakes, snowmelt, and human activities [17]. In tropical monsoon countries such as Vietnam, shallow landslides induced by intense rainfall in mountainous regions represent the most common type of landslide occurrence. These events typically result in the removal of existing vegetation cover on hillslopes. Therefore, this study employed NDVI-based change detection techniques to identify vegetation disturbances and detect landslide locations. The landslide events triggered by extreme rainfall in late October 2020 in the mountainous region of Phuoc Son were selected for database development and analysis. In addition to NDVI, several conditioning factors, including slope, rainfall, were incorporated into the analysis to enhance landslide detection accuracy. Ensemble learning techniques based on Random Forest and XGBoost models were applied to support the classification and identification of landslide locations. This study is expected to provide valuable insights into the detection of extreme rainfall-induced landslides in large-scale mountainous and densely forested regions.

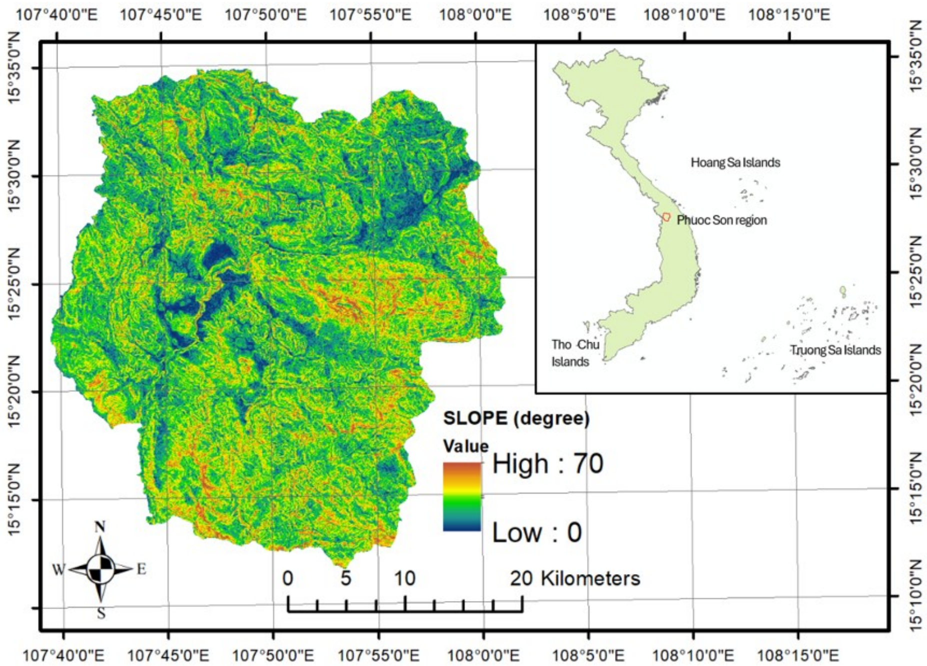


Fig. 1. Location of the Phuoc Son mountainous area, Da Nang City, Vietnam.

Fig. 2. Images of landslides captured by UAVs following the severe landslide event in the mountainous area of Phuoc Son

2. Study area

The mountainous region of Phuoc Son (Fig. 1) is located in the southwestern part of Da Nang, Vietnam (formerly Phuoc Son District, Quang Nam Province, prior to the administrative reorganization). The study area covers approximately 1,142 km². The terrain is predominantly mountainous, with elevations ranging from 30 m to 2,088 m above sea level. The topography is highly dissected, characterized by numerous valleys and a dense river and stream network. The study area is largely covered by forest, accounting for approximately 90% of the total area. Transportation accessibility in this region is extremely limited due to the low density of the road network.

Regarding hydro-meteorological conditions, the area receives an average annual rainfall of approximately 3,000 mm, with precipitation mainly concentrated in October and November. In recent years, several extreme rainfall events have occurred in the region, triggering numerous severe landslides. A representative example is the landslide event that occurred in late October 2020, which caused catastrophic damage throughout the study area. Extreme rainfall recorded at a rain gauge station within the region reached nearly 400 mm/day prior to the landslide event and served as the primary triggering factor for this disaster. An aerial view captured using an unmanned aerial vehicle (UAV) (Fig. 2) illustrates the severe destruction caused by this event. This event can be considered a representative case of rainfall-induced landslides, and the comprehensive collection of landslide inventory data across the entire region is essential for subsequent investigations. Given the large spatial extent and highly complex terrain conditions of the Phuoc Son mountainous area, the application of remote sensing techniques represents a highly feasible and effective solution.

3. Methodology

To achieve the objective of accurately identifying landslide locations triggered by extreme rainfall, this study employed multiple approaches and established a research workflow consisting of the following steps:

(i) Collection of landslide inventory data following the event of October 28, 2020 (polygon format), as well as non-landslide locations, including areas with intact vegetation cover, naturally non-vegetated areas, and areas of vegetation loss caused by deforestation and construction activities.

(ii) Collection of NDVI maps before and after the landslide event, rainfall maps, and slope maps.

(iii) Development and evaluation of landslide location prediction models using ensemble learning approaches, including LR, XGBoost, and Random Forest models.

(iv) Prediction of landslide locations across the entire study area.

(v) Validation of the prediction results at several representative locations.

3.1. Collection of Sentinel-2 imagery data

The Sentinel-2 satellite is a multispectral Earth observation system developed under the Global Monitoring for Environment and Security initiative and jointly coordinated by the European Commission and the European Space Agency [3]. The system consists of two polar-orbiting satellites, namely Sentinel-2A and Sentinel-2B, operating on the same orbit with a 180° phase difference. These satellites are equipped with a Multispectral Instrument (MSI) comprising 13 spectral bands (10 m, 20 m, and 60 m spatial resolutions), ranging from the visible and near-infrared (VNIR) to the shortwave infrared (SWIR) regions. By combining data from both satellites, the temporal resolution can reach 5 days. Level-1C and Level-2A products are freely available to users for land monitoring, emergency management, and risk mapping purposes [18]. This study employed Sentinel-2 Level-2A imagery, which provides Bottom-of-Atmosphere (BOA) surface reflectance data. Cloud-affected pixels were masked using the s2cloudless cloud removal algorithm available in Google Earth Engine, applying a maximum cloud probability threshold of 20%. For the purpose of identifying landslide locations based on vegetation cover changes, this study employed the Normalized Difference Vegetation Index (NDVI), which is capable of characterizing land surface vegetation conditions during the analysis process. This index reflects vegetation quality and facilitates the discrimination between vegetated surfaces and bare soil. NDVI can be derived from Sentinel-2 satellite imagery and is calculated using the following equation [19]:

$$NDVI=\frac{NIR\text{-}Red}{NIR+Red}=\frac{B8\text{-}B4}{B8+B4} \quad (1)$$

where NIR and Red represent the reflectance values of the near-infrared and red wavelength bands, respectively.

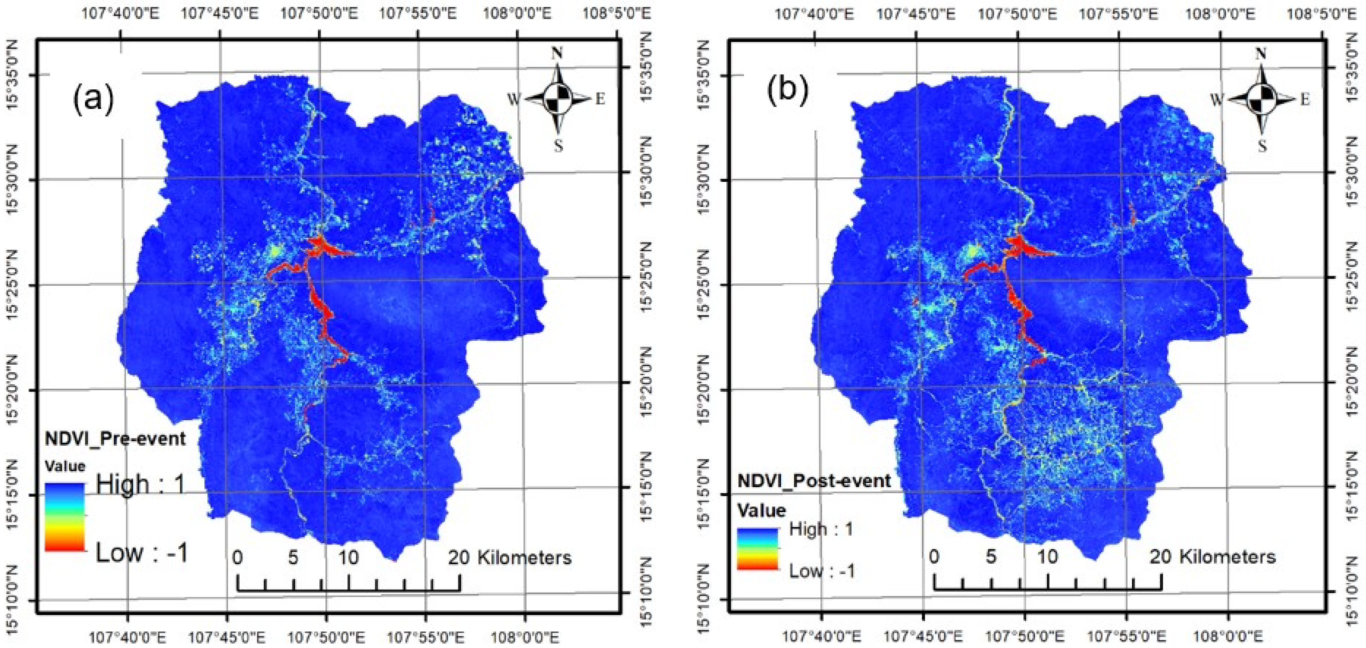


Fig. 3. NDVI derived from Sentinel-2 imagery: (a) Pre-landslide imagery, and (b) Post-landslide imagery

Based on the timing of landslide occurrences in the study area, which were concentrated in late October 2020, this study selected high-quality satellite image series (with cloud cover below 20%) acquired during the pre-landslide period (from 01 June 2020 to 31 August 2020) and the post-landslide period (from 29 October 2020 to 28 February 2021) for analysis. A cloud removal algorithm was applied to minimize the effects of cloud cover on the surface observations of the study area (Fig. 3).

3.2. Landslide inventory

The landslide dataset used for training the landslide location prediction models included both landslide and non-landslide locations, with all data collected in polygon format. Representative landslide locations were identified following the landslide event triggered by extreme rainfall on October 28, 2020 (Fig. 4). Meanwhile, non-landslide locations were collected from various areas, including representative cases such as areas where vegetation cover remained intact after the landslide event, areas where vegetation cover loss resulted from deforestation, and areas affected by the construction of public infrastructure (Fig. 4).

The diversity in the selection of non-landslide locations helps prevent the model from confusing vegetation loss caused by landslides with vegetation loss resulting from human activities. To ensure objectivity in model prediction, the numbers of landslide and non-landslide pixels were balanced. In total, 12 landslide locations (equivalent to 2,033 pixels) and 14 non-landslide locations (equivalent to 2,017 pixels) were used (Table 1). Landslide locations (Table 1) were selected and validated using UAV imagery. For this event, the selected landslide locations were concentrated in areas affected by extreme rainfall, primarily in the southern part of the study area. The slopes of the identified landslide locations generally ranged from 15° to 40°, which is representative of the overall terrain conditions across the study area. For the non-landslide dataset, 14 representative locations were carefully selected (Table 1). These locations included: (i) areas without vegetation cover loss; (ii) areas lacking vegetation cover before the event but exhibiting vegetation cover after the event; (iii) areas with vegetation cover loss caused by deforestation and agricultural activities; (iv) areas with vegetation cover loss resulting from construction activities; and (v) areas without vegetation cover both before and after the landslide event. The diversity of the selected non-landslide locations is expected to improve the ability of machine learning models to accurately distinguish landslide occurrences and reduce the misclassification of areas affected by human activities as landslides.

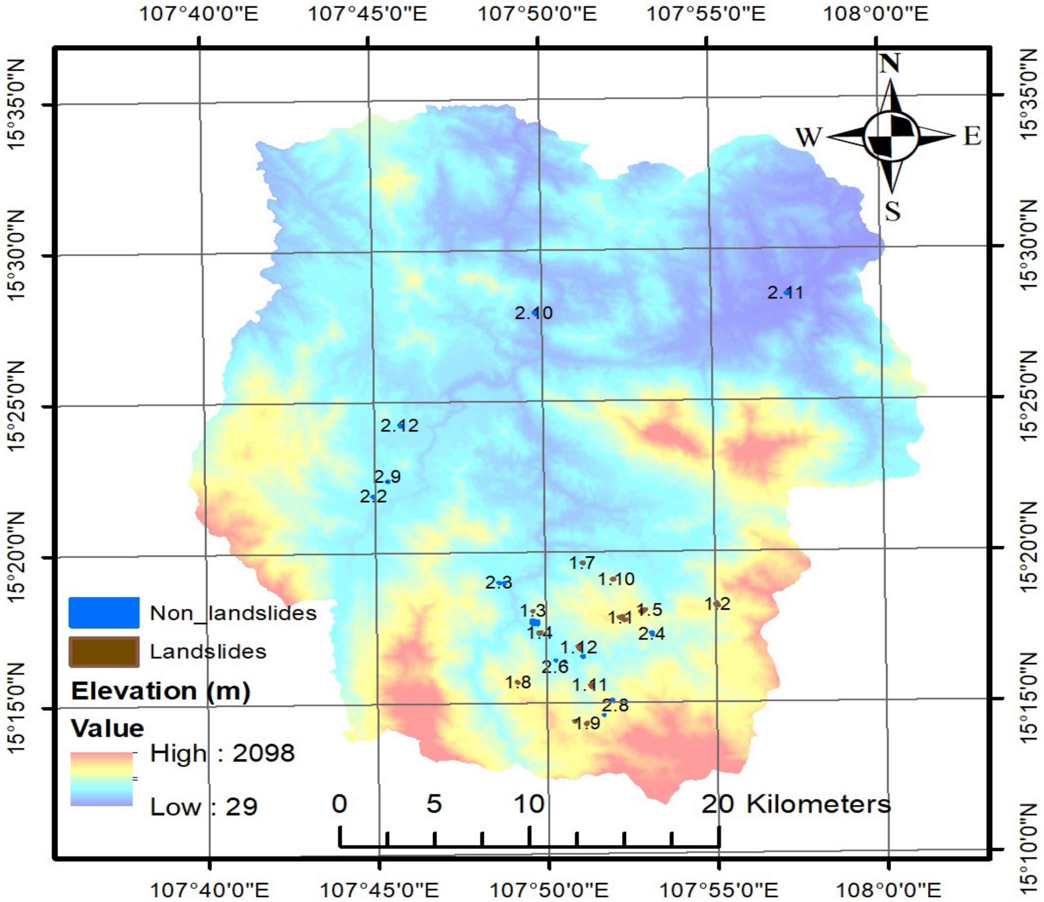


Fig. 4. Representative landslide and non-landslide locations

3.3. Landslide influencing factors mapping

This study identified landslide locations based on vegetation cover changes. To minimize the misclassification of human-induced land cover changes as landslides, additional influencing factors beyond NDVI variation were incorporated into the analysis. The landslide event of 28 October 2020 was triggered by extreme rainfall; therefore, rainfall conditions associated with landslide initiation were selected as a key factor for landslide identification. Furthermore, slope is widely recognized as one of the most important factors influencing landslide occurrence [20]. This factor can help distinguish actual landslide locations from areas where vegetation cover changes are caused by agricultural practices or construction activities. Accordingly, this study incorporated two additional factors, including extreme rainfall triggering landslides and slope to improve the assessment and identification of landslide locations.

Table 1. Landslide inventory statistics

Location	No.	Characteristics	Number of Locations	Number of Pixels (10 m × 10 m)	Notes
Landslide	1.1÷1.12	Landslides triggered by extreme rainfall	12	2033	
Non-landslide	2.1	Areas without vegetation cover loss	1	875	
	2.2÷2.3	Areas without vegetation cover before the event and with vegetation cover after the event	2	159	
	2.4÷2.7	Areas with vegetation cover loss due to deforestation and agricultural activities	4	492	
	2.8÷2.13	Areas with vegetation cover loss due to construction activities	6	457	
	2.14	Areas without vegetation cover (both before and after the landslide event)	1	34	

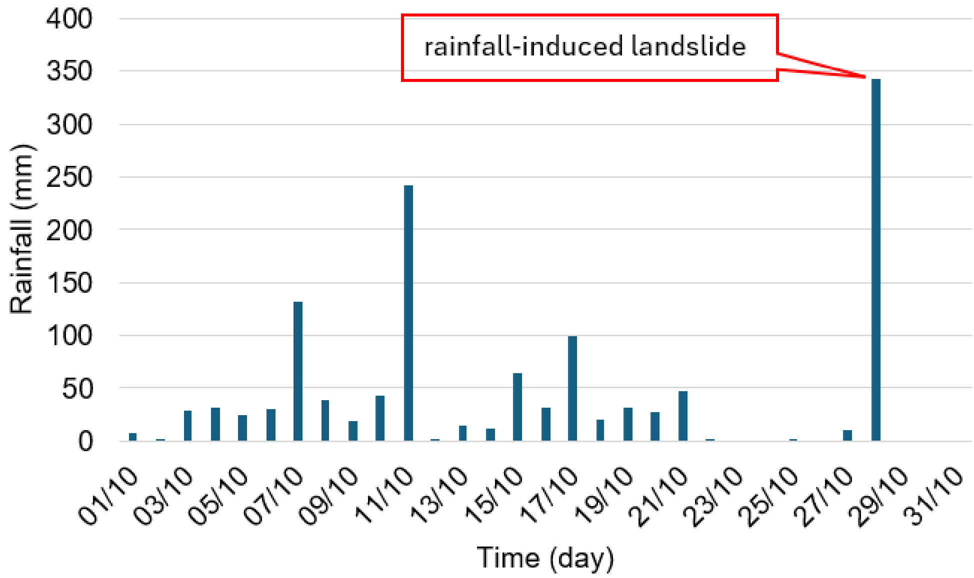


Fig. 5. Daily rainfall variation in the study area before and after the landslide event

a. Landslide-triggering rainfall

This study analyzed the rainfall time series based on the occurrence of the landslide event on October 28, 2020. Analysis of daily rainfall data from the Phuoc Thanh Rainfall Station, located within the landslide-affected area (Fig. 5), revealed the primary triggering factor of the landslide event. Specifically, during the period from October 22 to October 27, 2020, the study area experienced almost no rainfall. Subsequently, an extreme rainfall event occurred on October 28, 2020, with a total daily rainfall of approximately 350 mm, while no significant rainfall was recorded during the remaining days of October 2020. These findings indicate that this extreme rainfall event was the main triggering factor responsible for the occurrence of landslides in the area.

Based on this evidence, the present study proposed the use of rainfall data recorded on October 28, 2020, from rainfall stations within the study area for further analysis. The spatial distribution map of daily rainfall was interpolated from rainfall data collected at five rainfall stations in the region (Fig. 6) using the Inverse Distance Weighting (IDW) interpolation method [21].

b. Slope

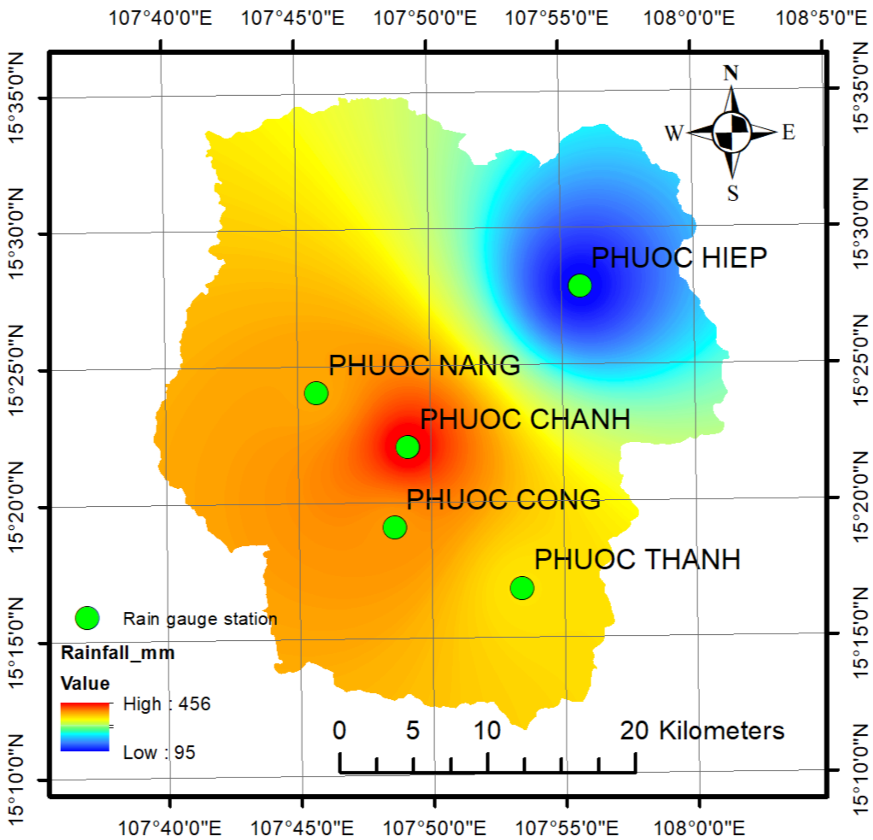


Fig. 6. Spatial distribution map of rainfall on October 28, 2020, and the locations of rainfall stations

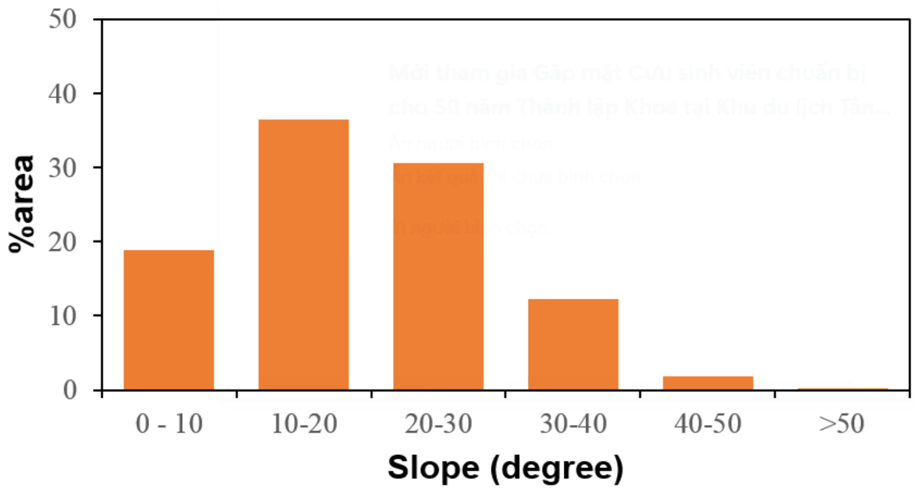


Fig. 7. Distribution chart of the study area by slope classes

Slope is considered one of the most influential factors affecting landslide occurrence. As the slope angle increases, the shear resistance of unconsolidated materials (soil and rock) generally decreases [22]. Slope angle can be regarded as an indicator of slope stability and can be derived from DEM data [23]. In general, gentle slopes tend to exhibit lower landslide susceptibility than steep slopes. In many regions, landslides are more likely to occur on slopes ranging from 15° to 45° [24]. However, landslide analyses indicate that the relationship between slope angle and slope stability is not straightforward, as the steepest slopes do not necessarily correspond to the highest landslide susceptibility. In practice, many steep rocky slopes exhibit greater stability than relatively gentle slopes composed of materials with weak geotechnical properties [25]. Therefore, the slope factor used in this study may also assist the model in better identifying human-induced disturbances that reduce vegetation cover, such as deforestation, agricultural activities, and infrastructure construction.

In this study, the slope map (Fig. 1) was derived from the ALOS PALSAR DEM dataset (https://www.earthdata.nasa.gov/) with a spatial resolution of 12.5 m × 12.5 m. The slope map was classified into several categories, including: <10°, 10–20°, 20–30°, 30–40°, 40–50°, and >50°. The frequency distribution chart (Fig. 7) indicates that approximately 20% of the study area has slopes lower than 10°, which corresponds to areas with very low landslide susceptibility [26].

3.4. Data evaluation

This study employed the Variance Inflation Factor (VIF) to assess multicollinearity among the input variables. The VIF is calculated using the following equation:

$$VIF=\frac{1}{1-R^2} \quad (2)$$

where R represents the multiple correlation coefficient between one influencing factor and the remaining influencing factors in the model.

In several previous studies, variables with VIF values greater than 5 are considered to exhibit multicollinearity issues, while variables with VIF values exceeding 10 should be removed from the model [27].

3.5. Landslide prediction model

a. Logistic regression

Logistic Regression (LR) is a generalized linear regression model that is well suited for multivariate problems. LR was introduced in the late 1960s and early 1970s [28]. Unlike conventional linear regression models, LR constrains the output values to the range of 0 to 1 through a sigmoid activation function, allowing the results to be expressed as probabilities. Although termed a regression method, LR is more commonly applied to classification problems and has been widely used in landslide susceptibility prediction studies [29], [30].

b. Random Forest

Random Forest (RF) is an ensemble learning method based on the Bagging approach, capable of accurately classifying large datasets through the use of multiple decision trees [31]. The main hyperparameters of the RF model include *n_estimators, max_depth, max_features, min_samples_leaf, min_samples_split*. One of the key tasks during model training is to determine the optimal combination of these hyperparameters. The RF model has been widely applied in landslide classification studies and has demonstrated high predictive performance [32].

c. Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) is an ensemble learning method based on the Boosting approach. It is a high-performance machine learning algorithm developed by Chen [33]. XGBoost utilizes multiple Classification and Regression Trees (CART) and integrates them using the Gradient Boosting technique. The main hyperparameters of the XGBoost model include *n_estimators, learning_rate, max.depth, colsample_bytree, min_child_weight, subsample*. XGBoost has been widely applied in classification problems and has demonstrated excellent predictive capability [32].

3.6. Model evaluation

For the LR, RF, and XGBoost classification models, this study employed several evaluation metrics, including Accuracy (ACC), Kappa coefficient (k), Positive Predictive Value (PPV), Negative Predictive Value (NPV), and the Area Under the ROC Curve (AUC) [34]. In general, higher values of ACC, k, and AUC approaching 1 indicate greater predictive accuracy of the model. Meanwhile, PPV and NPV reflect the balance performance of the model. A smaller difference between NPV and PPV values indicates that the model performs well in predicting both landslide and non-landslide locations [19].

4. Result and Discussion

4.1. Data Evaluation

To ensure reliable model performance, the input variables, including NDVI_diff (the difference between the post-event NDVI and the pre-event

NDVI), slope, and rainfall, were evaluated to eliminate potential multicollinearity effects. The multicollinearity assessment results based on the VIF index (Table 2) showed that none of the variables had a VIF value greater than 5. This indicates that the three variables do not exhibit multicollinearity and are therefore suitable for use in the prediction models.

Table 2. VIF values of the factors influencing landslides

Conditioning factor	NDVI_diff	Slope	Rainfall
VIF value	1.44	1.44	1.04

Table 3. The hyperparameter sets of the XGBoost and Random Forest (RF) models

Model	Parameter	Hyperparameter
XGBoost	n_estimators	150
	learning_rate	0.05
	max.depth	5
	colsample_bytree	0.8
	min_child_weight	3
	subsample	0.8
Random Forest	n_estimators	100
	max_depth	10
	max_features	sqrt
	min_samples_leaf	4
	min_samples_split	5

Table 4. Predictive performance of the LR, RF, and XGBoost Models

Evaluation metric	LR model		RF model		XGBoost model	
	Training	Testing	Training	Testing	Training	Testing
True Positive	1517	391	1623	401	1621	401
False Positive	130	26	10	3	5	1
True Negative	1484	377	1604	400	1609	402
False Negative	109	16	3	6	5	6
Accuracy	0.926	0.948	0.996	0.988	0.996	0.991
Kappa	0.852	0.896	0.992	0.977	0.993	0.982
Sensitivity (SST)	0.933	0.960	0.998	0.985	0.996	0.985
Specificity (SPF)	0.919	0.935	0.993	0.992	0.996	0.997
AUC	0.985	0.988	0.998	0.997	0.999	0.999

4.2. Predictive performance of the models

This study used a total of 2,033 landslide pixels and 2,017 non-landslide pixels. The dataset was subsequently divided into two groups: one for model training and the other for model evaluation. The data partitioning process was conducted randomly. However, to ensure objectivity, this study additionally applied the K-fold cross-validation method [35]. Specifically, the dataset was divided into k = 5 folds, of which four folds were used for model training and one-fold was used for model validation. The hyperparameters of the XGBoost and Random Forest ensemble models are presented in Table 3.

The model evaluation results (Table 4) indicate that most models achieved excellent predictive performance. Regarding the ACC metric, the XGBoost model ranked highest with an ACC value of 0.991, followed by the RF model (0.988) and the LR model (0.948). A similar pattern was observed for the Kappa, SST, and SPF indices, where the XGBoost model consistently demonstrated superior predictive capability with the highest statistical values. In addition, the differences between SST and SPF values across all three models were relatively small, indicating that the models provided balanced predictive performance for both landslide and non-landslide locations.

In addition to the statistical metrics, the predictive performance of the models was also evaluated using ROC curves (Fig. 8). A larger area under the ROC curve indicates higher prediction accuracy [36]. Overall, all three models produced excellent prediction results, with very high AUC values. Among them, the XGBoost model exhibited the best predictive performance, with an AUC value of 0.999, followed by the RF model (AUC = 0.997) and the LR model (AUC = 0.988).

Based on the analyses of the statistical evaluation metrics and AUC values presented above, the XGBoost model outperformed the other models in predictive performance. Therefore, the XGBoost model was selected to identify landslide locations across the entire study area.

4.3. Factor Importance

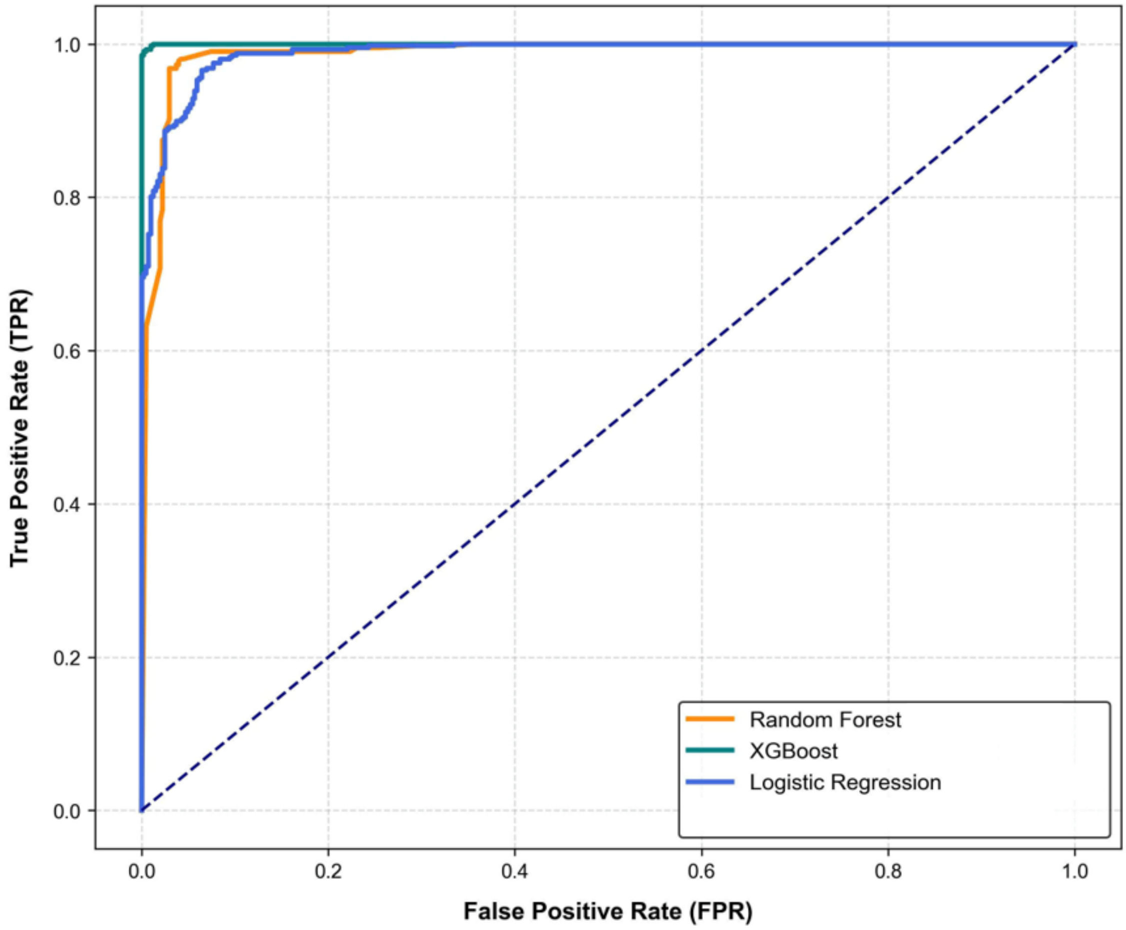
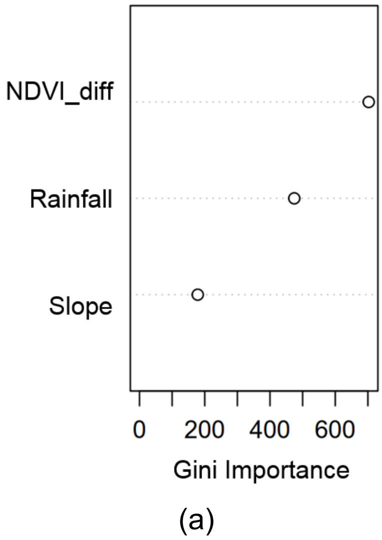


Fig. 8. ROC curve analysis results of the LR, RF, and XGBoost models

The importance index reflects the contribution level of each factor to the prediction model results. In this study, the RF and XGBoost models were employed to evaluate the importance of factors such as NDVI_diff, slope, and rainfall. These models apply different importance metrics: the RF model uses Gini Importance [37], whereas the XGBoost model employs Gain value [38]. Factors with higher importance values indicate a stronger influence on landslide occurrence.

The importance assessment results (Fig. 9) demonstrate the contribution level of each factor according to the respective evaluation methods. Overall, the NDVI_diff exhibited the greatest influence, followed by rainfall and slope. Considering the characteristics of the mountainous Phuoc Son region, where vegetation cover accounts for approximately 90% of the study area, variations in NDVI or the loss of vegetation cover represent clear indicators for identifying landslide locations. This explains why the NDVI_diff showed the highest level of influence.



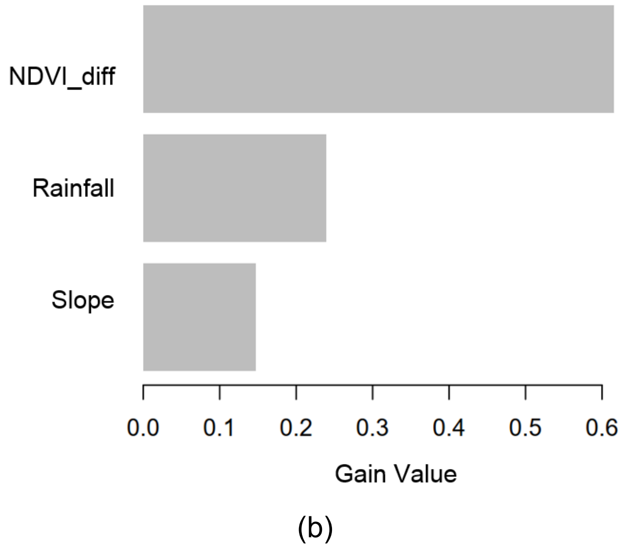


Fig. 9. Importance assessment of influencing factors: (a) Random Forest method, (b) XGBoost method

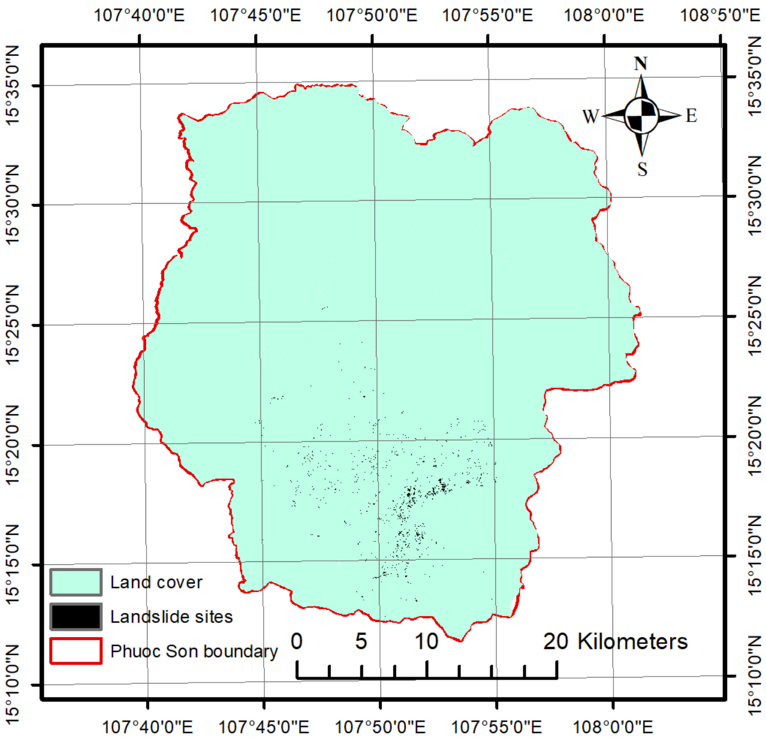
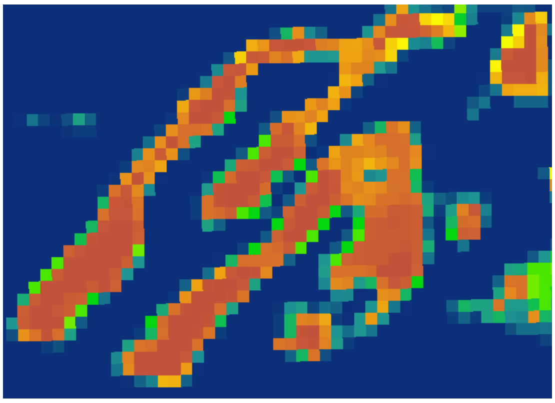


Fig. 10. Landslide inventory map derived from the XGBoost model for the entire Phuoc Son mountainous region

4.4. Prediction of landslide locations in the study area

The XGBoost model, which demonstrated superior predictive performance among the applied models, was selected to predict landslide locations across the entire study area. This task required extensive computational time due to the large dataset consisting of approximately 11,420,000 map pixels covering the entire Phuoc Son region.

The landslide inventory map (Fig. 10) reveals that a large number of landslide sites were identified. A total of 2,328 landslide sites were detected in the mountainous Phuoc Son area after the major landslide event that occurred on 28 October 2020. It can be observed that most landslide locations are concentrated in the southern part of the Phuoc Son mountainous region. This finding is consistent with actual records of landslide occurrences triggered by extreme rainfall in this area. When compared with actual imagery acquired from UAV (Fig. 11), the landslide locations identified by the XGBoost model showed strong agreement with the actual surface conditions displayed in the UAV imagery.

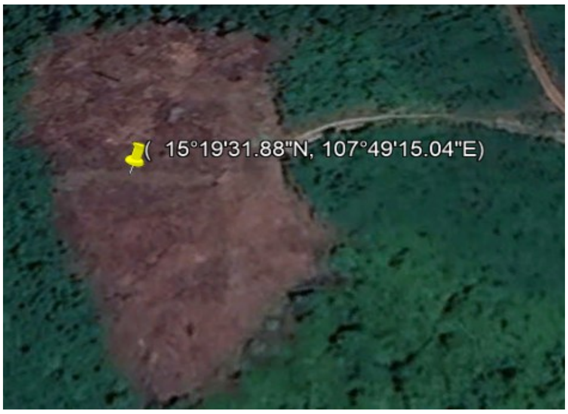


(a)

(b)

Fig. 11. Comparison between landslide locations identified by the XGBoost model (brown areas indicate landslide scars) (a) and actual imagery acquired from UAV (b)

(a)



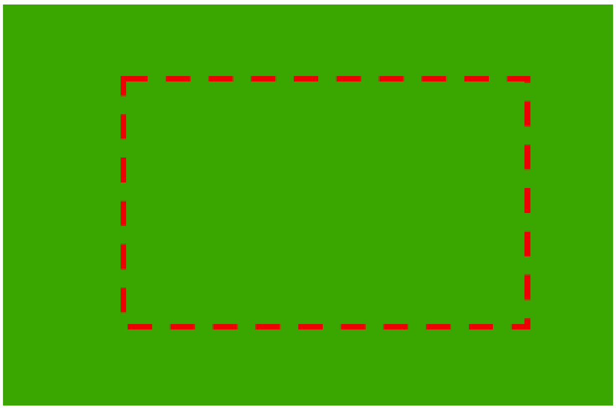
(b)

(c)

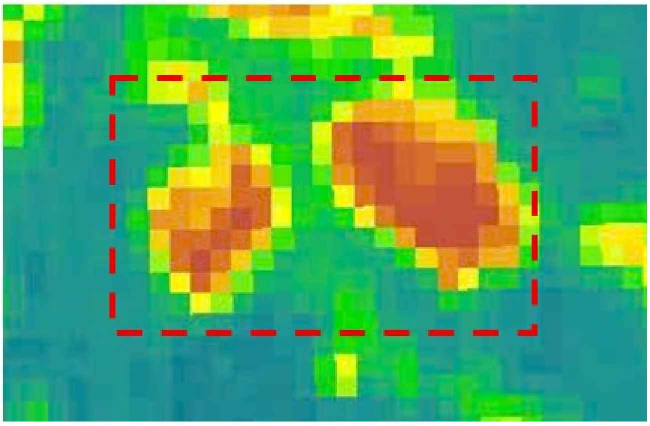
Fig. 12. Google Earth imagery illustrating surface conditions: (a) vegetation loss caused by landslides, (b) vegetation loss caused by deforestation, and (c) vegetation loss caused by construction activities

Fig. 13. Areas (white background) showing post-landslide imagery acquired on October 1, 2021 from Google Earth data

(a)



(b)



(c)

Fig. 14. Validation at areas of vegetation loss caused by deforestation: (a) imagery provided by Google Earth on October 1, 2021, and (b) prediction results generated by the XGBoost model (No landslides were recorded) (c) Post-landslide NDVI image (red areas indicate zones of vegetation cover loss)

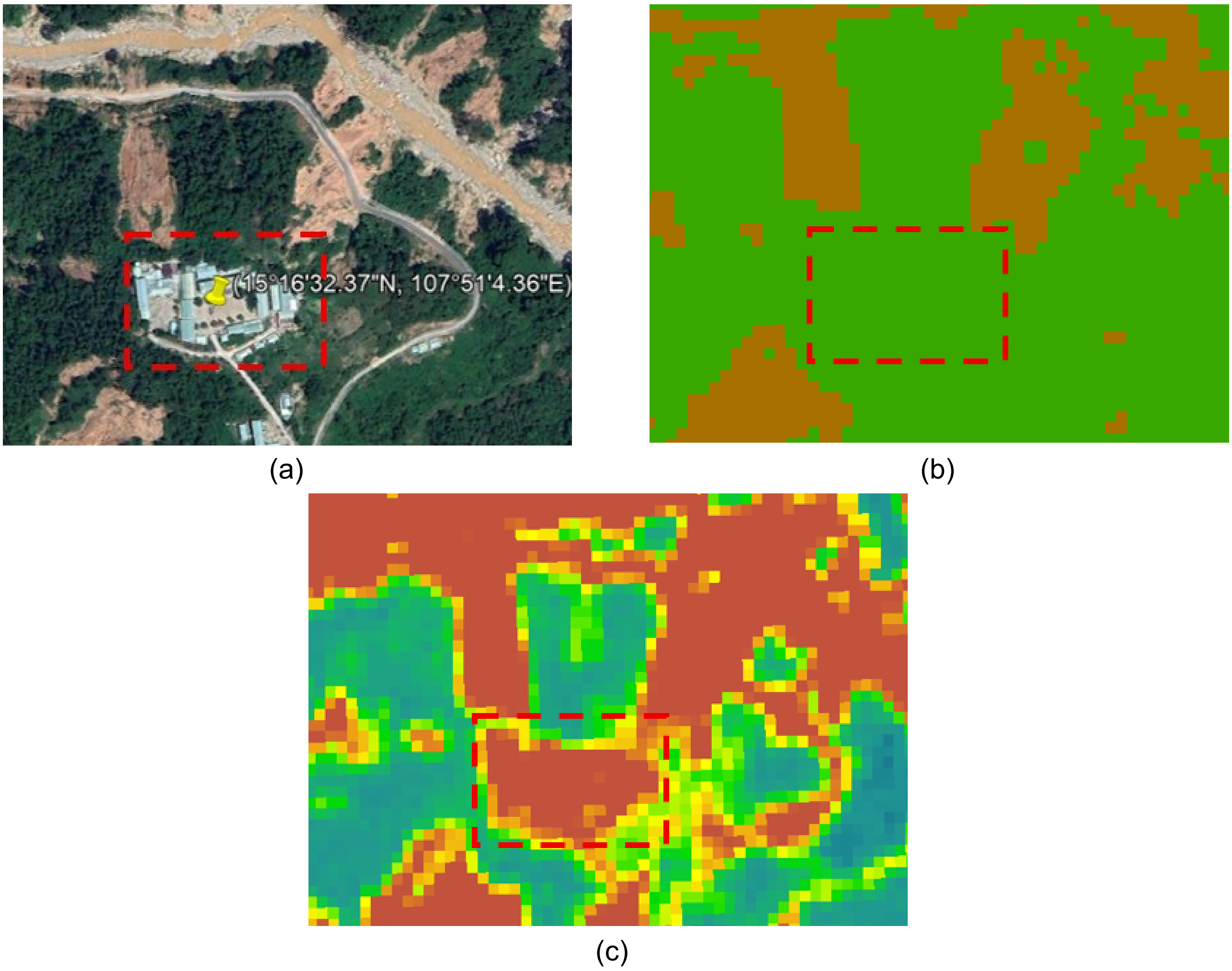


Fig. 15. Validation at areas of vegetation loss caused by construction activities: (a) imagery provided by Google Earth on October 1, 2021, and (b) prediction results generated by the XGBoost model (No landslides were recorded), (c) Post-landslide NDVI image (red areas indicate zones of vegetation cover loss)

In map analysis, imagery provided by Google Earth offers very high spatial resolution (up to 0.5 m), and visual interpretation of these images serves as an effective source of information for analysis [39]. In landslide investigations, Google Earth imagery can support the identification of vegetation loss caused by landslides (Fig. 12a), deforestation (Fig. 12b), and construction activities (Fig. 12c). However, this dataset also presents certain limitations, as the imagery is compiled from multiple scenes acquired at different times and with varying temporal resolutions. Consequently, it is not possible to observe the complete surface conditions of the entire study area at the same point in time. For the representative landslide event triggered by extreme rainfall in late October 2020, the imagery provided by Google Earth on October 1, 2021 (approximately a year after the landslide event) covered only part of the mountainous region of Phuoc Son (the white area in Fig. 13), while the closest pre-event imagery available was acquired in 2019.

It can be observed that these results demonstrate that the identification of landslide locations across the entire Phuoc Son mountainous region immediately after the event represents a significant contribution, helping to overcome the limitations associated with relying solely on Google Earth imagery for analysis.

4.5. Validation of the ability to avoid false identification of landslide locations

In studies aimed at identifying landslide locations based on vegetation-cover changes, there is a high risk of confusion between vegetation loss caused by landslides and vegetation loss resulting from human activities, such as deforestation and construction activities. During the training process of the XGBoost model, this study incorporated similar non-landslide cases to enable the model to learn how to reduce false identification of landslide locations.

After generating the landslide distribution map for the entire study area, the validation of the model's ability to accurately distinguish landslide locations in several representative cases became an essential step for further assessing the reliability of this study. Several representative cases were selected to verify the model's capability to avoid false identification of landslide locations, including: (i) evaluation at areas of vegetation loss caused by deforestation (Fig. 14), and (ii) evaluation at areas affected by construction activities (Fig. 15).

The validation results presented in Fig. 14 and Fig. 15 further confirm the effectiveness of the XGBoost model trained using the approach proposed in this study. In addition to accurately identifying landslide locations, the model also demonstrated the capability to exclude non-landslide areas associated with human activities. This represents a highly promising outcome for landslide inventory data collection, which is considered a critical task in landslide research.

5. Conclusion

Landslide detection plays a critical role in landslide-related studies. However, this task becomes particularly challenging in mountainous regions characterized by complex topography and limited accessibility. The landslide events triggered by extreme rainfall in the mountainous area of Phuoc Son in late October 2020 represent a typical case, involving a large number of landslide occurrences requiring rapid identification and mapping. In this study, remote sensing techniques integrated with ensemble learning models based on bagging and boosting approaches were applied to detect these landslide locations. A comprehensive database was constructed from representative landslide and non-landslide samples within the study area. Notably, the non-landslide dataset encompassed various land-cover conditions, including areas with minor vegetation changes as well as areas affected by anthropogenic activities such as deforestation and construction works. Furthermore, several conditioning factors, including the NDVI_diff derived from pre-event and post-event imagery, slope, and rainfall associated with landslide initiation, were incorporated as input variables for the prediction models.

The integration of remote sensing techniques, ensemble learning algorithms, and an appropriate data selection strategy substantially enhanced model performance. Validation results demonstrated that the XGBoost model achieved outstanding predictive accuracy, with an AUC value reaching 0.999. The model exhibited excellent capability in identifying landslide locations in the mountainous region of Phuoc Son following the extreme rainfall event. Verification at several representative landslide sites further confirmed that the model was not only effective in detecting landslide occurrences but also capable of distinguishing vegetation loss caused by landslides from vegetation disturbances induced by human activities.

The proposed approach can be effectively applied to other regions with similar environmental conditions, thereby reducing the time, effort, and cost associated with landslide inventory mapping. Nevertheless, several limitations remain, including (i) The proposed method is primarily effective for shallow landslides associated with vegetation loss, while the spatial resolution of 10 m × 10 m is still insufficient for accurately detecting small-scale landslides; (ii) Landslides occurring on non-vegetated surfaces could not be detected by the proposed approach; (iii) Due to adverse weather conditions, high-quality Sentinel-2 images with low cloud coverage immediately after the landslide event were unavailable for analysis; and (iv)

Landslide validation relied exclusively on UAV imagery and Google Earth observations, while detailed field-based verification at the landslide sites was not conducted. These limitations also highlight important directions for future research aimed at improving landslide detection capability through advanced remote sensing technologies.

Declaration of Generative AI

During the preparation of this work, the authors do not use Generative AI and AI-assisted technologies tools in order to generate the content of the manuscript.

References

[1] M. Dilley, R.S. Chen, U. Deichmann, A.L. Lerner-Lam, M. Arnold. (2005). Natural disaster hotspots: a global risk analysis. *The World Bank. ISBN 0-8213-5930-4*

[2] O. Kjekstad, L. Highland. (2009). Economic and Social Impacts of Landslides. Landslides–Disaster Risk Reduction, pp. 573–587. *Springer.* https://doi.org/10.1007/978-3-540-69970-5_30

[3] Y. Qin, P. Lu, Z. Li. (2018). Landslide Inventory Mapping From Bitemporal 10 M Sentinel-2 Images Using Change Detection Based Markov Random Field. International Archives of the Photogrammetry. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-3, 1447–1452. https://doi.org/10.5194/isprs-archives-XLII-3-1447-2018

[4] R. Pellicani, G. Spilotro. (2015). Evaluating the quality of landslide inventory maps: comparison between archive and surveyed inventories for the Daunia region (Apulia, Southern Italy). *Bulletin of Engineering Geology and the Environment*, 74, pp. 357–367. https://doi.org/10.1007/s10064-014-0639-z

[5] J. Sun, G. Yuan, L. Song, H. Zhang. (2024). Unmanned aerial vehicles (UAVs) in landslide investigation and monitoring: A review. *Drones*, 8(1), 30. https://doi.org/10.3390/drones8010030

[6] N. Casagli, E. Intrieri, V. Tofani, G. Gigli, F. Raspini. (2023). Landslide detection, monitoring and prediction with remote-sensing techniques. *Nature Reviews Earth & Environment*, 4, 51–64. https://doi.org/10.1038/s43017-022-00373-x

[7] A. Barra, O. Monserrat, P. Mazzanti, C. Esposito, M. Crosetto, G.S. Mugnozza. (2016). First insights on the potential of Sentinel-1 for landslides detection. *Geomatics, Natural Hazards and Risk*, 7(6), 1874–1883. https://doi.org/10.1080/19475705.2016.1171258

[8] M.W. Mwaniki, D.N. Kuria, M.K. Boitt, T.G. Ngigi. (2017). Image enhancements of Landsat 8 (OLI) and SAR data for preliminary landslide identification and mapping applied to the central region of Kenya. *Geomorphology*, 282, 162–175. https://doi.org/10.1016/j.geomorph.2017.01.015

[9] S. Plank, D. Hölbling, C. Eisank, B. Friedl, S. Martinis, A. Twele. (2015). Comparing object-based landslide detection methods based on polarimetric SAR and optical satellite imagery—A case study in Taiwan. *Proceedings of the 7th International Workshop on Science and Applications of SAR Polarimetry and Polarimetric Interferometry, POLinSAR, 2015*.

[10] Q. Huang, C. Wang, Y. Meng, J. Chen, A. Yue. (2020). Landslide monitoring using change detection in multitemporal optical imagery. *IEEE Geoscience and Remote Sensing Letters*, 17(2), 312–316. DOI: 10.1109/LGRS.2019.2918254

[11] Z. Li, W. Shi, P. Lu, L. Yan, Q. Wang, Z. Miao. (2016). Landslide mapping from aerial photographs using change detection-based Markov random field. *Remote Sensing of Environment*, 187, 76–90. https://doi.org/10.1016/j.rse.2016.10.008

[12] P. Lu, Y. Qin, Z. Li, A.C. Mondini, N. Casagli. (2019). Landslide mapping from multi-sensor data through improved change detection-based Markov random field. *Remote Sensing of*

Environment, 231, 111235. https://doi.org/10.1016/j.rse.2019.111235

[13] Q. Guo, L. Zhu, A. Li, L. Gu. (2022). Landslide identification method based on NDVI change detection. *Remote Sensing Technology and Application*, 37(1), 17–23. https://doi.org/10.11873/j.issn.1004-0323.2022.1.0017

[14] A. Mohan, A.K. Singh, B. Kumar, R. Dwivedi. (2021). Review on remote sensing methods for landslide detection using machine and deep learning. *Transactions on Emerging Telecommunications Technologies*, 32(7), e3998. https://doi.org/10.1002/ett.3998

[15] E.K. Sahin. (2022). Comparative analysis of gradient boosting algorithms for landslide susceptibility mapping. *Geocarto International*, 37(9), 2441–2465. https://doi.org/10.1080/10106049.2020.1831623

[16] T. Kavzoglu, A. Teke. (2022). Predictive performances of ensemble machine learning algorithms in landslide susceptibility mapping using random forest, extreme gradient boosting (XGBoost) and natural gradient boosting (NGBoost). *Arabian Journal for Science and Engineering*, 47, 7367–7385. https://doi.org/10.1007/s13369-022-06560-8

[17] L. Highland. (2004). Landslide types and processes. DOI:10.3133/FS20043072

[18] D. Phiri, M. Simwanda, S. Salekin, V.R. Nyirenda, Y. Murayama, M. Ranagalage. (2020). Sentinel-2 data for land cover/use mapping: A review. *Remote Sensing*, 12(14), 2291. https://doi.org/10.3390/rs12142291

[19] V.L. Doan, B.-Q.-V. Nguyen, H.T. Pham, C.C. Nguyen, C.T. Nguyen. (2023). Effect of time-variant NDVI on landside susceptibility: A case study in Quang Ngai province, Vietnam. *Open Geosciences*, 15, 20220550. https://doi.org/10.1515/geo-2022-0550

[20] P. Reichenbach, M. Rossi, B.D. Malamud, M. Mihir, F. Guzzetti. (2018). A review of statistically-based landslide susceptibility models. *Earth-Science Reviews*, 180, 60–91. https://doi.org/10.1016/j.earscirev.2018.03.001

[21] N.D. Vo, P. Gourbesville. (2016). Application of deterministic distributed hydrological model for large catchment: a case study at Vu Gia Thu Bon catchment, Vietnam. *Journal of Hydroinformatics*, 18(5), 885–904. https://doi.org/10.2166/hydro.2016.138

[22] S. Lee, K. Min. (2001). Statistical analysis of landslide susceptibility at Yongin, Korea. *Environmental Geology*, 40, 1095–1113. https://doi.org/10.1007/s002540100310

[23] C.J. van Westen, N. Rengers, R. Soeters. (2003). Use of geomorphological information in indirect landslide susceptibility assessment. *Natural Hazards*, 30, 399–419. https://doi.org/10.1023/B:NHAZ.0000007097.42735.9e

[24] A.K. Saha, R.P. Gupta, I. Sarkar, M.K. Arora, E. Csaplovics. (2005). An approach for GIS-based statistical landslide susceptibility zonation—with a case study in the Himalayas. *Landslides*, 2, 61–69. https://doi.org/10.1007/s10346-004-0039-8

[25] D.J. Varnes and the International Association of Engineering Geology. Commission on Landslides and Other Mass Movements on Slopes. (1984). Landslide Hazard Zonation: a review of principles and practice. *The UNESCO Press*.

[26] V.L. Doan, B.-Q.-V. Nguyen, C.C. Nguyen, C.T. Nguyen. (2024). Effect of time-variant rainfall on landslide susceptibility: A case study in Quang Ngai Province, Vietnam. *Vietnam Journal of Earth Sciences*, 46(2), 202–220. https://doi.org/10.15625/2615-9783/20065

[27] B. Pradhan, M.I. Sameen. (2017). Landslide susceptibility modeling: optimization and factor effect analysis. *Laser Scanning Applications in Landslide Assessment, pp. 115–132. Springer*. https://doi.org/10.1007/978-3-319-55342-9_6

[28] A.F. Cabrera. (1994). Logistic regression analysis in higher education: An applied

perspective. Higher education: Handbook of theory and research, vol. 10, pp. 225–256. *Agathon Press*.

[29] B.T. Pham, D.T. Bui, I. Prakash, L.H. Nguyen, M.B. Dholakia. (2017). A comparative study of sequential minimal optimization-based support vector machines, vote feature intervals, and logistic regression in landslide susceptibility assessment using GIS. *Environmental Earth Sciences*, 76, 371. https://doi.org/10.1007/s12665-017-6689-3

[30] B. Kalantar, B. Pradhan, S.A. Naghibi, A. Motevalli, S. Mansor. (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, Natural Hazards and Risk*, 9(1), 49–69. https://doi.org/10.1080/19475705.2017.1407368

[31] L. Breiman. (2001). Random Forests. Machine Learning, 45, 5–32. https://doi.org/10.1023/A:1010933404324

[32] E.K. Sahin. (2020). Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGBoost, gradient boosting machine, and random forest. *SN Applied Sciences*, 2, 1308. https://doi.org/10.1007/s42452-020-3060-1

[33] T. Chen, C. Guestrin. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *2016, pp. 785–794*. https://doi.org/10.1145/2939672.2939785

[34] P. Frattini, G. Crosta, A. Carrara. (2010). Techniques for evaluating the performance of landslide susceptibility models. *Engineering Geology*, 111(1–4), 62–72. https://doi.org/10.1016/j.enggeo.2009.12.004

[35] P. Jiang, J. Chen. (2016). Displacement prediction of landslide based on generalized regression neural networks with K-fold cross-validation. *Neurocomputing*, 198, 40–47. https://doi.org/10.1016/j.neucom.2015.08.118

[36] M. Kantardzic. (2011). Data mining: concepts, models, methods, and algorithms. *John Wiley & Sons*. DOI:10.1002/9781118029145

[37] B.H. Menze, B.M. Kelm, R. Masuch, U. Himmelreich, P. Bachert, W. Petrich, F.A. Hamprecht. (2009). A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics*, 10, 213. https://doi.org/10.1186/1471-2105-10-213

[38] H. Shin. (2022). XGBoost regression of the most significant photoplethysmogram features for assessing vascular aging. *IEEE Journal of Biomedical and Health Informatics*, 26(7), 3354–3361. DOI: 10.1109/JBHI.2022.3151091

[39] G.B. Fisher, C.B. Amos, B. Bookhagen, D.W. Burbank, V. Godard, S.J. Whitmeyer. (2012). Channel widths, landslides, faults, and beyond: The new world order of high-spatial resolution Google Earth imagery in the study of earth surface processes. *Special Paper of the Geological Society of America*, 492, 1–22. DOI 10.1130/2012.2492(01)