Journal of Hydroinformatics

Journal of Hydroinformatics Vol 00 No 0, 1 doi: 10.2166/hydro.2023.327

Mapping of soil erosion susceptibility using advanced machine learning models at Nghe An, Vietnam

Chien Quyet Nguyen[a], Tuyen Thi Tran[b], Trang Thanh Thi Nguyen[b], Thuy Ha Thi Nguyen[c,d], T. S. Astarkhanova[d], Luong Van Vu[c], Khac Tai Dau[c], Hieu Ngoc Nguyen[e], Giang Hương Pham[f], Duc Dam Nguyen[g], Indra Prakash[h] and Binh Pham[g,*]

[a] Faculty of Geography, Hanoi National University of Education, Vietnam 136 Xuan Thuy Str., Cau Giay District, Hanoi, Vietnam
[b] Faculty of Geography, School of Education, Vinh university
[c] School of Agriculture and Resources, Vinh University, Nghe An, Vietnam
[d] Peoples' Friendship University of Russia, Moscow 117198, Russia
[e] Nghe An University of Economics, Nghe An, Vietnam
[f] Faculty of Geography, Thai Nguyen University of Education
[g] University of Transport Technology, Hanoi 100000, Vietnam
[h] DDG (R) Geological Survey of India, Gandhinagar 382010, India
*Corresponding author. E-mail: binhthaiphamvn@gmail.com

ABSTRACT

Soil Erosion Susceptibility Mapping (SESM) is one of the practical approaches for managing and mitigating soil erosion. This study applied four Machine Learning (ML) models namely the Multilayer Perceptron (MLP) classifier, AdaBoost, Ridge classifier, and Gradient Boosting classifier to perform SESM in a region of Nghe An province, Vietnam. The development of these models incorporated seven factors influencing soil erosion: slope degree, slope aspect, curvature, elevation, Normalized Difference Vegetation Index (NDVI), rainfall, and soil type. These factors were determined based on 685 identified soil erosion locations. According to SHapley Additive exPlanations (SHAP) analysis, soil type emerged as the most significant factor influencing soil erosion. Among all the developed models, the Gradient Boosting classifier demonstrated the highest prediction power, followed by the MLP classifier, Ridge classifier, and AdaBoost, respectively. Therefore, the Gradient Boosting classifier is recommended for accurate SESM in other regions too, taking into account the local geo-environmental factors.

Key words: gradient boosting classifier, machine learning, grid search, soil erosion, Vietnam

HIGHLIGHTS

- Soil erosion has been modeled and a soil erosion susceptibility map was generated.
- Several ML models, including the MLP classifier, Ada Boost, Ridge classifier, and Gradient Boosting classifier were implemented.
- Developed models were tuned using the Grid search CV technique.
- The Gradient Boosting classifier performed the best.
- About 33% of the study area has a high and very high susceptibility to soil erosion occurrence.

1. INTRODUCTION

Soil erosion is globally acknowledged as a significant challenge that poses a threat not only to agricultural productivity but also to socio-economic advancement and the sustainability of economies.

Rill, inter-rill, sheet, gully, badland, and landslide are the different kinds of soil erosion forms causing the degradation of land. Landslide occurrence, as a sub-category of mass movement, is one of the main forms of soil erosion and soil degradation in Eastern Asian countries such as Nepal, Malaysia, and Vietnam, due to their unique topography, geology river system, land use pattern, and location in tropical monsoon regions. Rainfall is one of the main causes of soil erosion. Soil erosion has several negative impacts on the region's environment and economy. It can lead to reduced soil fertility, decreased agricultural productivity, and increased sedimentation in waterways, which can affect fish populations and water quality. Soil erosion can also lead to increased costs for farmers due to the need for soil conservation measures and increased use of inputs such as

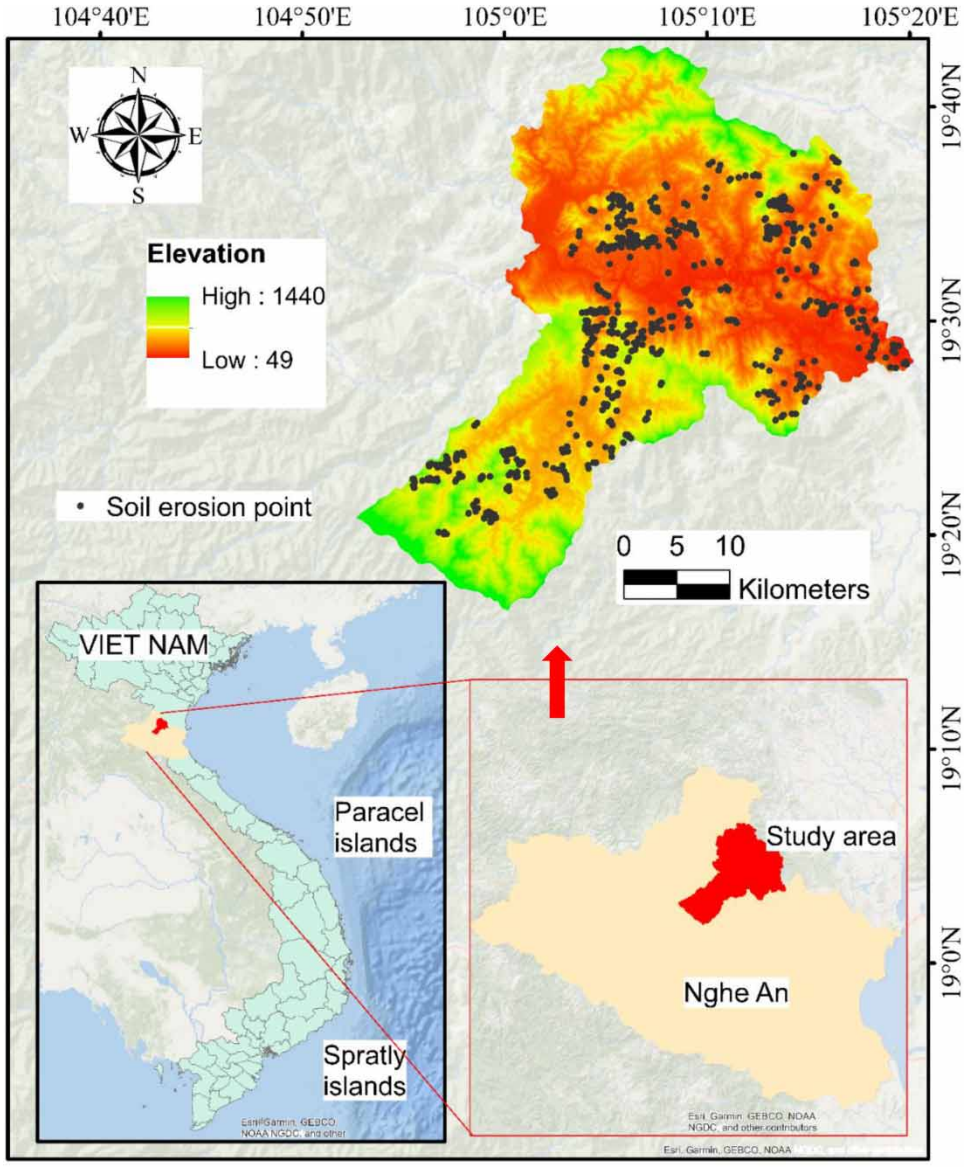

Figure 1 | Location map of the study area showing the location of soil erosion sites.

fertilizers and pesticides. Soil erosion also affects morphology of the area and cause serious problem to water quality, aquatic systems and their habitats, and reduction of dam reservoir storage (Khosravi *et al.* 2022).

The issue of soil erosion is escalating day by day due to human interference in altering land use patterns for agriculture and development, deforestation, and also due to the effects of climate change (Yang *et al.* 2003). For sustainable development, it is required to prevent soil erosion and degradation of soil mass. Identifying areas susceptible to soil erosion is a key step in implementing effective remedial measures (Yesuph & Dagnew 2019). These measures aim to stabilize soil in various areas including ground surfaces, hill slopes, and catchments. Planning for such measures, which may include strategic plantation and construction of structures to curb river bank erosion, is essential in maintaining the integrity of these landscapes. Areas vulnerable to soil erosion can be identified by systematic soil erosion susceptibility mapping (SESM) to mitigate the problems of soil erosion (Saha *et al.* 2019). In many natural hazard problems, it has been found that future hazards occur under conditions similar to those of the past (Van Westen *et al.* 2006). Hence, the creation of an inventory, documenting past erosion locations' inventory plays an important role in SESM. The traditional ways of SESM include the use of models such as the Revised Universal Soil Loss Equation (RUSLE), which predicts long-term, average-annual erosion by water for a broad range

of farming (Prasannakumar *et al.* 2012). However, this model can be implemented based on expert opinion, and therefore, it is associated with high uncertainty. Other traditional methods include the use of Weighted Overlay Analysis (WOA) to identify soil erosion high-risk areas (Ulain *et al.* 2022).

In recent years, advanced Machine Learning (ML) models have been increasingly utilized in fields such as hydrology and geotechnical engineering (Nearing *et al.* 2021; Zhang *et al.* 2021). These models are particularly effective in addressing classification and prediction challenges, including those found in SESM. Golkarian *et al.* (2023) compared the prediction accuracy of various ML models namely Convolutional Neural Networks (CNNs) and CNN optimized by gray wolf optimizer (CNN-GWO), Support Vector Machine (SVM) and SVM optimized by GWO (SVM-GWO), Group Method of Data Handling (GMDH) and GMDH optimized by GWO (GMDH-GWO), and Extreme Gradient Boosting (XGBoost) with the RUSLE model for soil erosion in Iran, and ML models are significantly better than the traditional RUSLE model. Sajedi-Hosseini *et al.* (2018) applied the fuzzy analytical network process (ANP) for soil erosion modeling in Iran, and stated that the fuzzy FANP model has reasonable prediction power. Mosavi *et al.* (2020) applied a new kind of random forest model named weighted subspace random forest (WSRF), and compared the results with the Gaussian Process and Naive Bayes (NB) model for SESM at Noor-Rood watershed, Iran, and reported that WSRF outperformed the other studied models. Sahour *et al.* (2021) implemented three different ML/linear models, including deep learning (DL), boosted regression trees (BRT), and multiple linear regression (MLR) for soil erosion in Iran, and reported that the BRT model outperformed the other studied models. Khosravi *et al.* (2023) applied three different kinds of DL models namely CNN, RNN, and LSTM for SESM in the North of Iran, and suggested that RNN has the highest prediction accuracy compared with the other studied models.

In general, literature reviews indicate that ML models are superior to SESM. This is because ML methods offer several advantages over traditional methods. They can handle large amounts of data and complex relationships between variables, which can lead to more accurate predictions (Vishnu & Rajput 2020). Moreover, model development and selection of the best model is a continuous process for solving classification and prediction problems. In the present study, therefore, we have used four ML classifier algorithms namely Multilayer Perceptron (MLP), Ada Boost, Ridge, and Gradient Boosting for the development of SESM. In addition, soil erosion is a significant challenge for Nghe An province, Vietnam. Efforts to mitigate its impacts are essential for the region's sustainable development. Therefore, we have selected a part of this province for soil erosion study and development of SES models. The novelty of the current work lies in its unique approach to SESM. Specifically, this study is one of the first to assess and compare the predictive power of the Gradient Boosting classifier (GBC) model with other established models such as the MLP, Ada Boost, and Ridge classifier. While Extreme Gradient Boosting has been utilized for natural hazard assessments, the application of the GBC model in the geoscience field is relatively rare. This study, therefore, not only fills a significant gap in the literature but also contributes to our understanding of the potential and effectiveness of different ML models in geoscience and SESM. By exploring these proposed models, this research could pave the way for more innovative and effective approaches to SESM in the future. Python software was used for data analysis and plotting while ArcGIS software was used for data processing and mapping.

2. STUDY AREA

The study area is a part of Nghe An province, Vietnam, located in the upper middle part of the Lam River basin, covering an area of about 1,074 sq. km. This province is characterized by a diverse topography, comprising mountains, hills, plains, and coastal areas. The western part of the province is dominated by the Annamite Range, which runs from north to south and forms a natural border with Laos. The eastern part of the province consists of coastal plains and sandy beaches.

Nghe An province has a complex geology, with a variety of rocks and mineral deposits. The province is rich in minerals such as coal, iron, lead, zinc, copper, gold, and silver. The western part of the province is also known for its karst landscapes and cave systems. Nghe An province has a diverse range of crops, including rice, maize, cassava, sugarcane, and coffee. This region has a tropical monsoon climate, with two distinct seasons. The average temperature varies between 23° and 25 °C, and the annual rainfall is around 1,600–1,800 mm.

Soil erosion in Nghe An province is a significant problem that affects the region's agricultural productivity, environmental health, and economic development. The province experiences high levels of soil erosion due to a combination of natural and human factors. Natural factors contributing to soil erosion include the region's topography, which is characterized by steep slopes that increase the risk of soil erosion. The province also experiences high levels of rainfall, which can cause soil erosion

during heavy downpours. Human factors contributing to soil erosion include intensive agricultural practices such as excessive tillage, use of chemical fertilizers and pesticides, as well as overgrazing by livestock. These practices can result in soil degradation, nutrient depletion, and ultimately, soil erosion. In addition, deforestation is another significant contributor to soil erosion in the province. Deforestation leads to the loss of vegetation cover, which can cause soil erosion and soil nutrient depletion. Climate change is also a contributing factor as it can increase the frequency and severity of rainfall events, which can exacerbate soil erosion in the province.

3. MATERIALS AND METHODS

3.1. Data used

3.1.1. Soil erosion inventory

An inventory of soil erosion was prepared for 685 locations within the Lam River basin of Nghe An province that have been affected by soil erosion (Figure 2). This inventory was used to evaluate factors that could potentially influence future soil erosion in the study area. Historical soil erosion data locations were identified using Google Earth images, documentary sources, satellite data, and field surveys with the help of the Global Positioning System (GPS) tool. For modeling, non-soil erosion and soil erosion areas were delineated and extracted from the total study area by drawing buffers around soil erosion individual sites. For the ML model study, 685 non-soil erosion locations (i.e. equal to the number of identified erosion sites) were randomly selected using GIS software. Both types of soil erosion and non-soil erosion data were separated into a 70:30 ratio for training and testing models, respectively (Nguyen *et al.* 2021b).

The entire historical soil dataset, including training and testing datasets, was converted to a raster format, and soil erosion pixels were allocated '1' and non-soil erosion pixels '0'. In the next step, all assigned pixels were overlaid with all seven considered soil erosion geo-environmental related conditioning factors to extract the attribute values including historical soil erosion data which are located within the soil erosion conditioning factors. In the subsequent step, extracted values were converted into an Excel file and exported to the GIS and Weka software for soil erosion modeling and mapping.

3.1.2. Soil erosion influencing factors

Identification of the relevant potential soil erosion influencing factors is crucial for developing a robust model for SESM. Redundant and irrelevant factors/ features are to be removed from the models using feature selection methods to overcome the problem of dimensionality and to reduce overfitting problems in models. Soil erosion in any area is influenced by a multitude of factors, including topography, geo-environment, geology, hydrology, anthropogenic activities, climate conditions, and rainfall. In this study, we have incorporated seven of these factors into our model's development, based on historical records and past experiences. In this study, a digital elevation model (DEM) of 30 m resolution (https://asterweb.jpl.nasa.gov/gdem.asp) was used for the development of slope degree, slope aspect, curvature, and elevation maps. A Normalized Difference Vegetation Index (NDVI) map was generated using Landsat 7 satellite images collected from USGS (https://

Figure 2 | Example of soil erosion site photo of the study area.

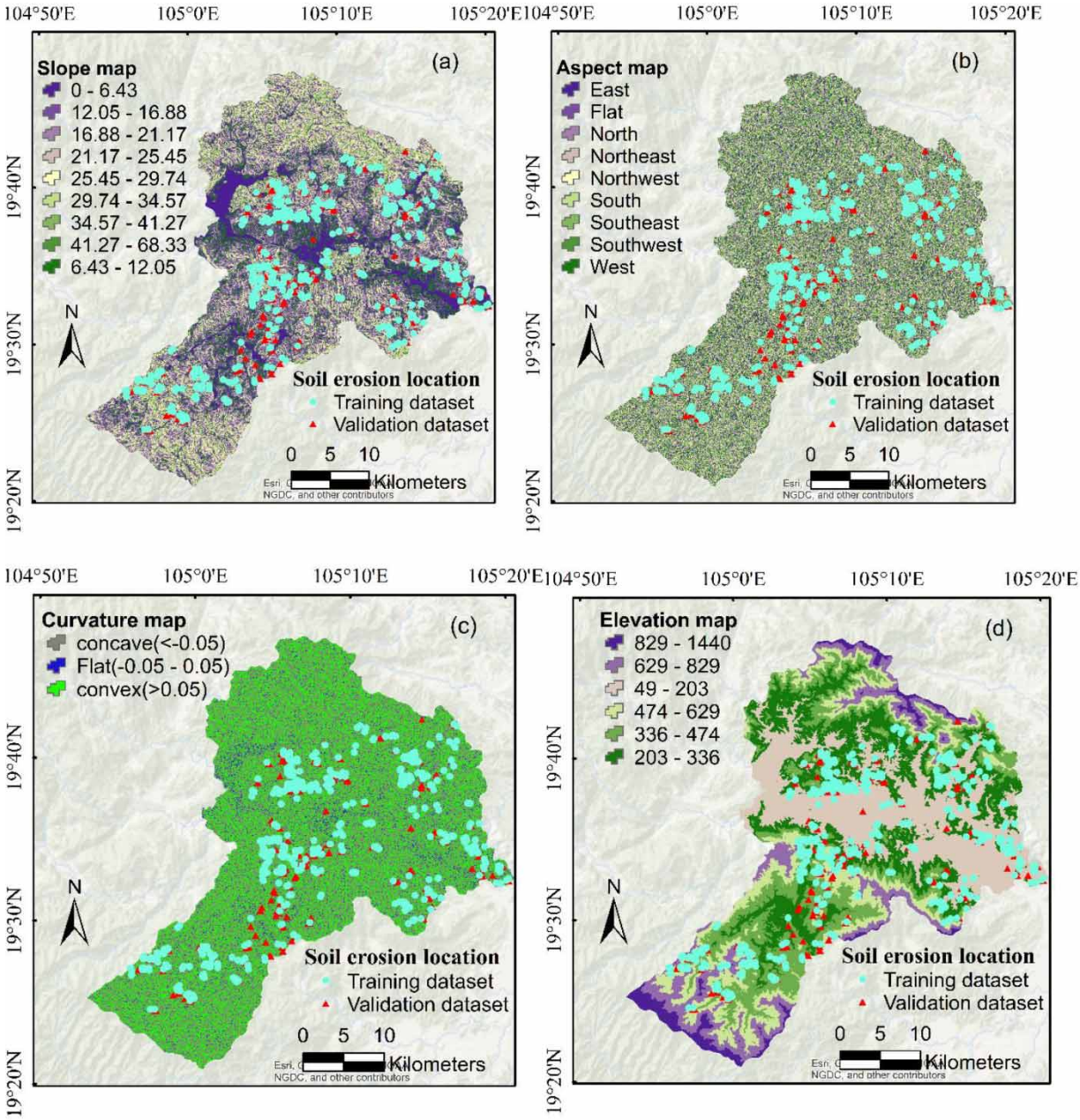

Figure 3 | Generated maps of the soil erosion influencing factor of the study area: (a) slope, (b) aspect, (c) curvature, (d) elevation, (e) NDVI, (f) rainfall, and (g) soil type. (*continued.*).

earthexplorer.usgs.gov/). Rainfall and soil maps were extracted from the available published data and maps collected from authorized agencies. A brief description of the soil erosion affecting factors is given below:

The slope angle significantly contributes to soil erosion due to landslides and direct removal of soil in the event of rain (Bradford & Foster 1996; Wu *et al.* 2017). Soil erosion in the area is more on the slope angle, which is more than 10 degrees. The higher the slope angle, the greater the possibility of soil erosion. For the study area, a slope angle map was generated and divided into nine classes (Figure 3(a)).

The slope aspect significantly influences the direction and intensity of solar radiation, moisture, and rainfall, all of which contribute to soil erosion (van Breda Weaver 1991; Sadeghi *et al.* 2012). Additionally, the wind flow, which is also dependent on the slope aspect, can further erode soil from sloping surfaces. To better understand these effects, a slope aspect map of the

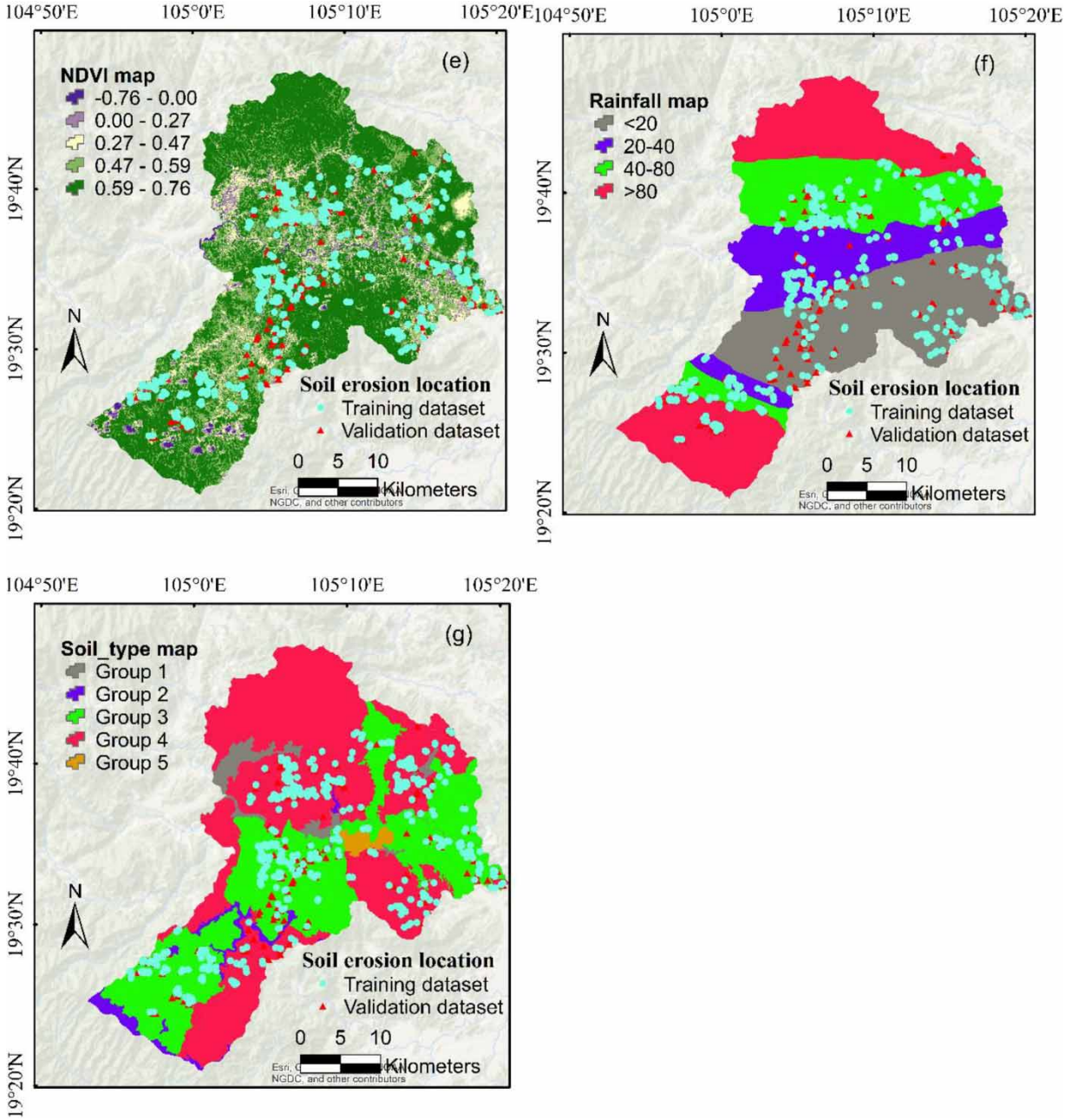

Figure 3 | Continued.

study area was created. This map was divided into nine categories: flat, north, northeast, east, southeast, south, southwest, west, and northwest (Figure 3(b)).

The curvature of the surface significantly affects the runoff in the area, leading to increased soil erosion on the convex surface (D'souza & Morgan 1976; Stefano *et al.* 2000). This is in contrast to concave and flat surfaces, where the accumulation and infiltration of surface water are more pronounced, thereby reducing the likelihood of erosion. A curvature map was prepared from the DEM and classified into three classes: convex, flat, and concave (Figure 3(c)).

Elevation plays a crucial role in the occurrence of soil erosion. It directly influences factors such as the volume, intensity, and duration of rainfall, the type of vegetation cover, and soil depth (Aslam *et al.* 2021). These effects are particularly pronounced in hilly and mountainous regions where variations in elevation are significant. An elevation map of the study

area was derived from the DEM. This map was then categorized into six distinct classes for a more detailed study (Figure 3(d)).

NDVI affects soil erosion as this factor evaluates the vegetation cover (Karaburun 2010). In this study, this index was derived using Landsat 7 satellite images. NDVI quantifies the amount of green vegetation using the formula: NDVI = (NIR – Red)/(NIR + Red), where NIR represents Near Infrared spectral reflectance (Figure 3(e)).

Rainfall is the primary contributor to soil erosion, accounting for the majority of total erosion, especially in tropical areas (Martınez-Casasnovas *et al.* 2002). It leads to the disintegration of soil particles, breakdown of soil aggregates, and migration of eroded soil materials or sediment. Furthermore, rainfall is a key factor triggering landslides (Varnes 1978). A rainfall map of the study area was constructed using the mean annual rainfall data collected over a 30-year period (1991–2021) from the Meteorological and Hydrological Station of North Central Vietnam (Figure 3(f)).

Soil type plays a significant role in soil erosion. The susceptibility of soil to erosion depends on its physico-mechanical and chemical properties (Mekonnen & Melesse 2011). Factors such as the mineral composition, size, and texture of the soil are crucial in determining its erodibility. For instance, loose, sandy, cohesionless soil is more prone to erosion than sticky, clayey soil. The soil map of the study area, derived from the existing soil map of Nghe An province, was reclassified into five distinct groups. These groups are illustrated in Figure 3 and detailed in Table 1.

3.2. Methods used

3.2.1. MLP classifier

MLP classifier algorithm, which is a sub-category of the feed forward artificial neural network (ANN) model, relies on underlying Neural Networks (NNs) to implement the classification scheme (Sahana *et al.* 2022). At its core, the MLP classifier is composed of multiple interconnected layers of artificial neurons or nodes. These layers include an input layer, one or more hidden layers, and an output layer (Yao *et al.* 2021). Each neuron in a layer is connected to every neuron in the subsequent layer, forming a dense and layered network. This architecture allows the MLP classifier to capture intricate patterns and relationships within the data. One of the distinguishing features of the MLP classifier is its capability to handle non-linear and high-dimensional data (Guragai *et al.* 2020). Training an MLP classifier involves a process known as backpropagation, where the model iteratively adjusts its internal parameters (weights and biases) to minimize a specified loss function (Ding *et al.* 2023). This optimization process aims to reduce the discrepancy between the predicted outputs and the true labels in the training data. MLP classifiers can be applied to both binary and multiclass classification problems. In binary classification, the output layer typically consists of a single neuron with a sigmoid activation function, while in multiclass classification, the output layer can have multiple neurons, each representing a class and employing activation functions.

3.2.2. Adaboost

AdaBoost is a kind of statistical classification model, which is one of the most successful and efficient algorithms of the ensemble learning technique, especially to reduce the sensitivity of noisy data through hybridization (Nhu *et al.* 2020). Therefore, the AdaBoost model can be integrated with many other types of algorithms to enhance the performance of base algorithms (Yang *et al.* 2021). Although the AdaBoost model is developed for binary classification, it can be for implementation to multiple classes or bounded intervals as well (Sun *et al.* 2021). In the first step, a subset is created from the training

Table 1 | Soil type group's description

No	Group	Description
1	Red-yellow ferralite soil formed on sandstone	The soil mechanical composition is light, the proportion of clay is about <20%, limited water holding capacity.
2	Red-yellow ferralite formed on acidic igneous rocks	The mechanical component is light. The clay content in the soil is < 30%.
3	Red-yellow soil is formed on metamorphic rocks	The mechanical composition of the soil is quite heavy, the proportion of clay is about 25–35%,
4	Humus on the mountain	Humus on the mountain is formed on shale, sandstone, at an altitude of over 700 m. The soil has a granular structure, the proportion of clay in the soil is quite high, 35–40%.
5	Alluvial soil in the valley	The mechanical component is light, mostly fine grain. The clay content in the soil is < 30%.

data, and therefore, one early classifier algorithm is developed; then, the early model is implemented to predict all instances in the training dataset; next, the misclassified cases get higher weights, but the weights of the correctly classified cases remain the same; and finally, the weights of all cases in the training dataset are scaled and a new subset is then randomly created to construct a next classifier-based model. This process continues until pre-defined stopping criteria are reached. The final model is calculated based on a weighted sum of all classifier-based models.

3.2.3. Ridge classifier

A ridge classifier is a powerful ML algorithm designed for solving classification problems (Elgeldawi *et al.* 2021). This algorithm is an extension of the well-known Ridge regression technique used for regression tasks, and it brings the benefits of regularization to the realm of classification. At its core, the Ridge classifier is a linear classification model (Nouman *et al.* 2023). It operates by establishing a linear decision boundary within the feature space to distinguish between different classes. One of the notable strengths of the Ridge classifier is its versatility in handling both binary and multiclass classification tasks. In binary classification, it discerns between two classes, while in multiclass classification, it extends its capabilities to classify data into multiple classes through strategies like one-vs-all or one-vs-one. Additionally, the Ridge classifier offers the ability to provide probability estimates for class predictions, which can be invaluable in applications where understanding the model's confidence is essential. However, like many ML algorithms, fine-tuning hyperparameters is crucial to achieving optimal performance. Among these hyperparameters, the regularization strength, often denoted as alpha or lambda, is pivotal in controlling the level of regularization applied to the model.

3.2.4. Gradient boosting classifier

A Gradient Boosting classifier is an effective, flexible, and robust ML model that has gained immense popularity for solving classification problems (Khan *et al.* 2022). It belongs to the family of boosting algorithms and is known for its exceptional predictive accuracy and ability to handle complex data relationships (Bentéjac *et al.* 2021). At its core, Gradient Boosting is an ensemble learning method that combines the predictions of multiple weak learners, typically decision trees, to create a robust and accurate predictive model. In this model, each weak learner is trained to correct the errors made by the previous ones. This sequential training process leads to a powerful ensemble model that continually improves its predictive performance. One of the notable advantages of Gradient Boosting is its high predictive accuracy. By focusing on correcting errors made by the earlier models, it gradually refines its predictions, resulting in a strong and accurate model. Gradient Boosting is also robust to overfitting, a common issue in ML. The sequential nature of the algorithm and the regularization techniques applied during training, such as shrinkage and depth constraints on decision trees, help to prevent overfitting and improve the model's generalization to unseen data. While it may require careful hyperparameter tuning and can be computationally expensive, the superior performance it offers often justifies the effort, making the Gradient Boosting classifier a valuable tool for data scientists and ML practitioners in various domains.

3.2.5. SHAP analysis

SHapley Additive exPlanations (SHAP) is a tool used in ML for explainability and interpretability (Mangalathu *et al.* 2020). It computes the contribution of each feature to the prediction, making the outcomes of ML models more understandable. Thus, it provides a way to understand and interpret complex ML models, making them more accessible and useful in practical applications. In SHAP's plot, the x-axis shows the SHAP value while the y-axis indicates the variable name (commonly in the order of importance from top to bottom) (Park *et al.* 2022). A SHAP value shows how much is the change in log-odds, and can be used to interpret the probability of success. In the variable name axis, the value next to them is the mean SHAP value. In addition, feature importance can be extracted based on SHAP values. Positive SHAP values indicate that a feature contributes positively to the prediction, causing the model to predict a higher output value. On the other hand, negative SHAP values suggest that a feature contributes negatively to the prediction, causing the model to predict a lower output value. According to the color bar, red color shows higher values while lower values are presented in blue. In this study, the SHAP technique was used to feature the importance of the input factors used in soil erosion modeling.

3.3. Validation methods

Evaluation and comparison of the model were conducted using a Receiver Operating Characteristic curve (ROC) approach and several statistical indices, including the positive predictive value (PPV), negative predictive value (NPV), sensitivity (SST), specificity (SPF), accuracy (ACC), Kappa (K), root mean square error (RMSE), and mean absolute error (MAE).

The ROC curve is a plot that displays the diagnostic ability of a binary classifier system—on the y-axis, a true positive rate, whereas on the x-axis, a false positive rate (Chakrabortty *et al.* 2020). A true positive rate or recall, or sensitivity or probability of detection is defined as the rate of correctly determined soil erosion locations. A false positive rate or probability of false alarm is defined as the rate of wrongly determined soil erosion locations. The ROC curve quantitatively measures model performance through the Area Under the ROC Curve (AUC) (Bien *et al.* 2022, 2023). The AUC value ranges from 0.5 to 1. An ideal model has an AUC equal to 1, indicating a perfect classification. Therefore, the larger the AUC value, the higher the performance of the model.

With standard statistical indices (Chakrabortty *et al.* 2022; Chakrabortty & Pal 2023), PPV is defined as the likelihood that pixels in the study area are correctly classified as soil erosion data; NPV is considered as the probability of pixels being correctly classified as non-soil erosion data; ACC is the proportion of soil erosion and non-soil erosion data pixels that are correctly classified; SST is the proportion of soil erosion data pixels that are correctly identified; SPF is the proportion of non-soil erosion data pixels that are correctly identified; and Kappa (K) statistic is used to evaluate the reliability of the SESM. Additionally, RMSE and MAE were applied for model evaluation and comparison through the error analysis of the models. Generally, higher values of SST, SPF, ACC, K, RMSE, MAE, NPV, and PPV indicate better predictive capability of the models. In contrast, lower values of RMSE and MAE indicate better predictive capability of the models.

4. RESULTS AND DISCUSSION

4.1. Importance of the input factors in the soil erosion modeling

In this study, sensitivity analysis was carried out by applying SHAP analysis to evaluate the importance of the input factors in the modeling of soil erosion. The SHAP technique's plot is composed of thousands of distinct points, with higher values represented in more intense red and lower values in deeper blue. This color scheme corresponds to the feature values. If the points on one side of the central line progressively shift from red to blue, it indicates that increasing or decreasing values, respectively, push the predicted soil erosion in that direction (Figures 4 and 5). It can be seen from SHAP analysis that while all considered factors are important, soil type is the most influential factor on soil erosion occurrence, with a mean SHAP value of 0.63. This is followed by slope degree (mean SHAP: 0.42), rainfall (mean SHAP: 0.41), elevation (mean SHAP: 0.33), NDVI (mean SHAP: 0.28), aspect (mean SHAP: 0.18), and curvature (mean SHAP: 0.08), respectively. This aligns with the natural process of soil erosion, as soil erodibility depends on the particle size, compactness, cohesiveness, and mineral composition of soil (Egbueri *et al.* 2021). Rainfall or runoff is the primary causative factor for soil erosion in the study area (Martínez-Mena *et al.* 2020). Heavy rainfall in Nghe An province during typhoons causes significant soil erosion and deep landslides. In addition, it is important to note that other parameters such as NDVI and land use patterns can

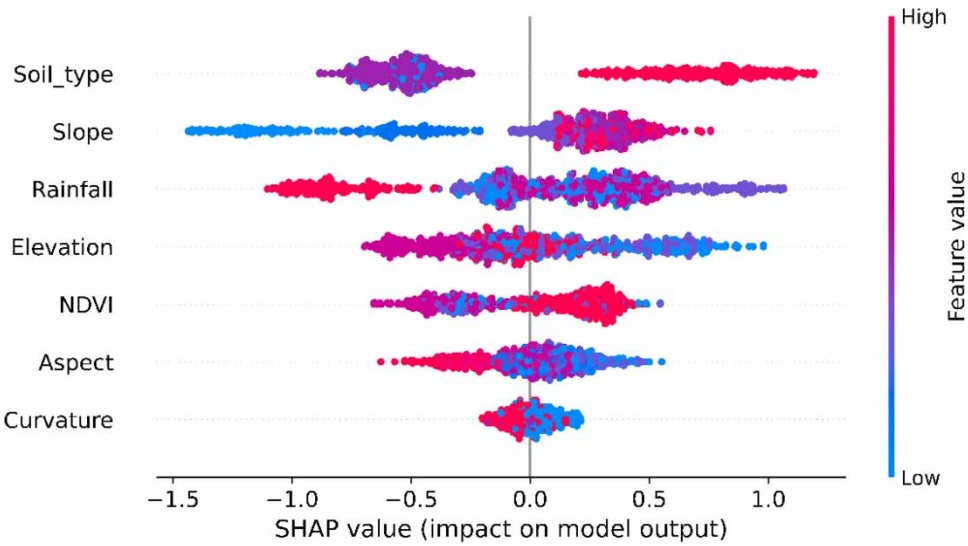

Figure 4 | SHAP values of the soil erosion affecting factors used in this study.

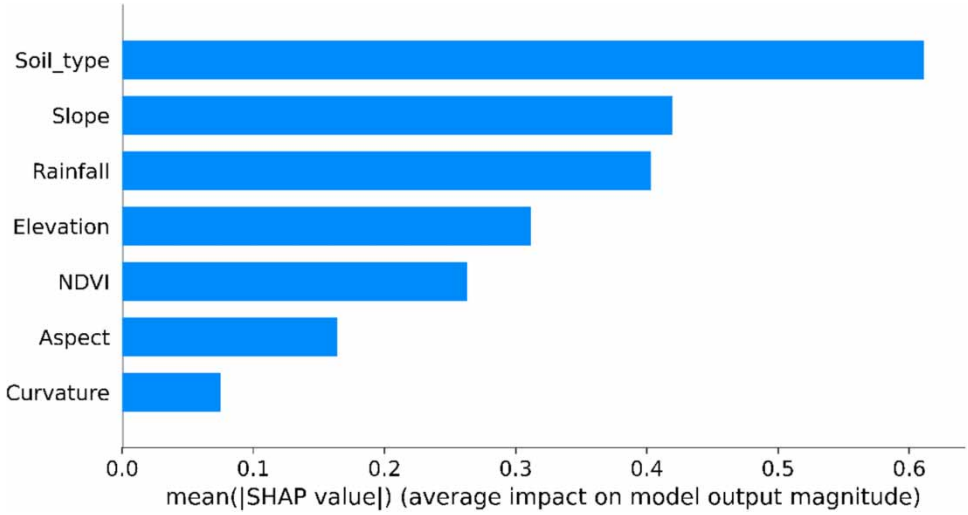

Figure 5 | Mean absolute SHAP values of the soil erosion affecting factors used in this study.

also significantly influence soil erosion (Ayalew *et al.* 2020). Therefore, the factors influencing soil erosion can vary from region to region, depending on ground conditions and geo-environmental factors.

4.2. Hyperparameter tuning for training the models

Hyperparameter tuning involves finding a set of optimal hyperparameter values for the learning algorithm to control the learning process of the ML model. This technique helps in maximizing the performance of the model. In this study, we have used the Grid Search (GS) technique for hyperparameter tuning as it has certain advantages such as (Hossain & Timmer 2021): (i) Exhaustive search: the GS conducts a thorough exploration of the search space, ensuring that no potential combination of hyperparameters is overlooked, and (ii) Interpretability: by evaluating all possible combinations, the GS provides a comprehensive understanding of how each hyperparameter influences the model's performance.

Basically, the GS implements through various parameters that enter into the parameter grid and generates the best integration of parameters, based on a scoring metric of researcher choice. By applying the GS technique, the optimal values of hyperparameters of each model are obtained and presented in Table 2.

Table 2 | Best values of tuning hyperparameters of the models using the GS technique

		Models			
No	**Parameters**	**MLP classifier**	**AdaBoost**	**Ridge classifier**	**Gradient boosting classifier**
1	Activation	logistic	–	–	–
2	Alpha	1	–	290	–
3	Hidden layers	100	–	–	–
4	Learning rate	–	0.1	–	–
5	Number of estimators	–	200	–	200
6	Fit intercept	–	–	true	
7	Max depth	–	–	–	4
8	Max features	–	–	–	0.3
9	Min samples leaf	–	–	–	100
10	Cross validation	10	10	10	10

4.3. Validation of the models

Validation results for the four developed models, which were implemented using a set of quantitative statistical indices including PPV, NPV, SST, SPF, ACC, K, MAE, and RMSE, are summarized in Table 3. It indicated that the performance of all models is good, but that of the Gradient Boosting classifier is the best on both training and validation data, followed by the MLP classifier, Ridge classifier, and AdaBoost, respectively. The Gradient Boosting classifier model has the lowest value of error index (MAE = 0.238 and RMSE = 0.488) in the validation phase. The NPV classifier is also high with an 80% probability of correctly classifying pixels of the soil erosion class; in terms of SST, it had a 74.32% probability of incorrectly classifying soil erosion pixels into the soil erosion classes; in terms of SPF, 78.3% of non-soil erosion pixels were correctly classified as a non-soil erosion location. In terms of NPV, it had a 72.2% probability that correctly classified non-soil erosion pixels as a non-soil erosion location. The Gradient Boosting classifier with an ACC value of 76.15% and a K value of 0.523 outperformed other models in the validation phase, followed by theMLP (ACC:74.2% and K:0.484), AdaBoost (ACC:70.80% and K:0.416), and Ridge classifiers (ACC:70.56% and K:0.411).

Based on the finding from the ROC curve, the MLP classifier model with an AUC of 0.89 for the training phase has the highest prediction power among all developed models but it has a lower performance than the Gradient Boosting classifier model during the validation phase. It shows that the MLP classifier model has a lower generalization power, while the Gradient Boosting classifier model has the highest generalization power. Overall, the Gradient Boosting classifier model with an AUC of 0.83 for validation phases, has better prediction power among all developed models followed by the MLP classifier (AUC:0.81), Ridge classifier (AUC: 0.71), and AdaBoost (AUC:0.79), respectively (Figure 6(a) and 6(b)).

Based on the above analysis of the validation results of the models, it can be concluded that among the four ML classifiers applied in the study area, the Gradient Boosting classifier demonstrated the best performance in accurately predicting soil erosion susceptibility. This was followed by the MLP classifier, Ridge classifier, and AdaBoost, respectively. The superior predictive capability of the Gradient Boosting classifier can be attributed to its high flexibility and ability to optimize on different loss functions. In addition, this model has other advantages such as: (i) it is less prone to overfitting as it builds trees sequentially, and each tree focuses on correcting the errors of the previous ones, (ii) it can effectively handle features with different scales, and (iii) it can more effectively handle outliers of the data used. The findings of this work are also in line with the other published works (Nguyen *et al.* 2021a; Saha *et al.* 2022).

4.4. Construction of soil erosion susceptibility maps

A soil erosion susceptibility map of the study area was developed using the best ML model, namely the Gradient Boosting classifier, as shown in Figure 7. In the first step, soil erosion susceptibility indices of the whole study area were extracted by training the model for all pixels of the study area. Thereafter, these indices were classified into various categories based on the natural break classification method embedded in the ArcGIS application (Amiri *et al.* 2019). Five categories namely very low, low, moderate, high, and very high soil erosion susceptibility were identified to construct soil erosion susceptibility maps of the study area. It can be seen that the north part of the study area has a higher susceptibility to soil erosion

Table 3 | Accuracy analysis of the models

		Training dataset				Validation dataset			
No	Parameters	MLP classifier	AdaBoost	Ridge classifier	Gradient Boosting classifier	MLP classifier	AdaBoost	Ridge classifier	Gradient Boosting classifier
1	PPV (%)	73.695	60.752	65.136	79.958	68.447	60.194	68.932	80.097
2	NPV (%)	79.167	77.500	71.042	76.875	80.000	81.463	72.195	72.195
3	SST (%)	77.925	72.932	69.180	77.530	77.473	76.543	71.357	74.324
4	SPF (%)	75.099	66.429	67.126	79.355	71.616	67.068	69.811	78.307
5	ACC (%)	76.434	69.135	68.092	78.415	74.209	70.803	70.560	76.156
6	K	0.529	0.383	0.362	0.568	0.484	0.416	0.411	0.523
7	MAE	0.236	0.309	0.319	0.215	0.258	0.292	0.294	0.238
8	RMSE	0.485	0.556	0.565	0.464	0.508	0.540	0.543	0.488

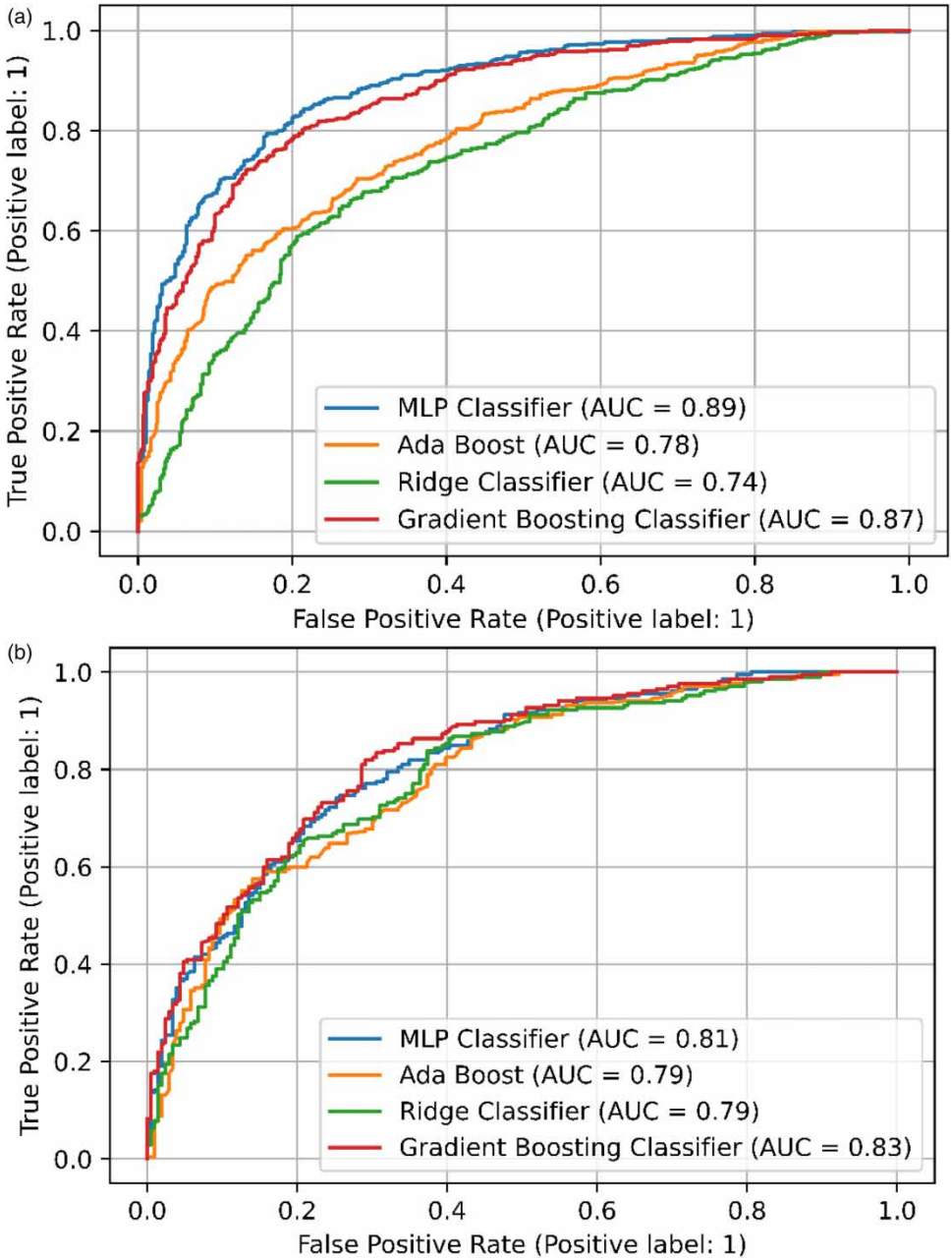

Figure 6 | Model performance evaluation using ROC approach for (a) training and (b) validating datasets.

in comparison to the other parts. Frequency ratio analysis, which is a ratio of the percentage of soil erosion pixels and class pixels on each susceptibility class of the map (Gayen *et al.* 2019), was then used to validate the performance of the soil erosion susceptibility map generated. It can be observed that the frequency ratio values have an increasing trend from very low susceptibility (0.38) classes to very high susceptibility (1.7); thus, it can be stated that the achieved map has a high degree of accuracy (Figure 8).

5. CONCLUDING REMARKS

In this study, four ML models namely the MLP classifier, AdaBoost, Ridge classifier, and Gradient Boosting classifier were applied for SESM in a part of Nghe An province, Vietnam. Seven soil erosion influencing factors were considered for model

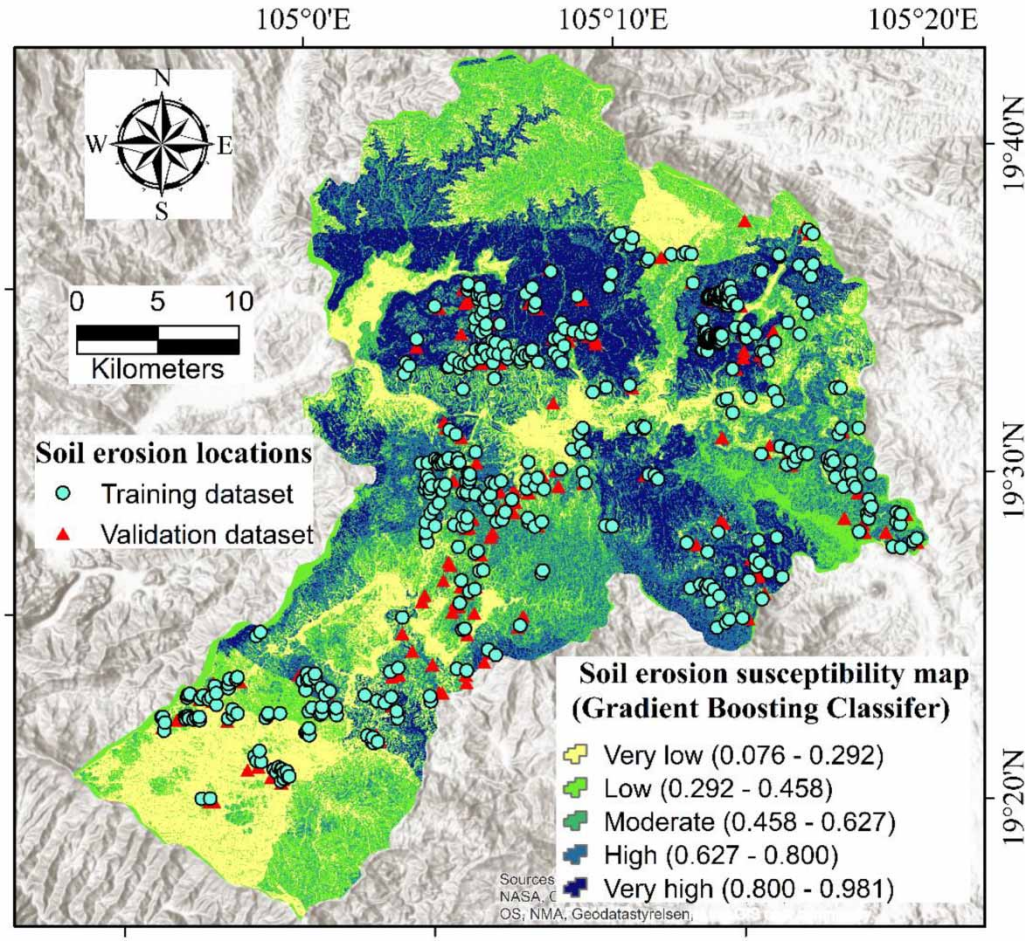

Figure 7 | Soil erosion susceptibility map generated from Gradient Boosting classifier.

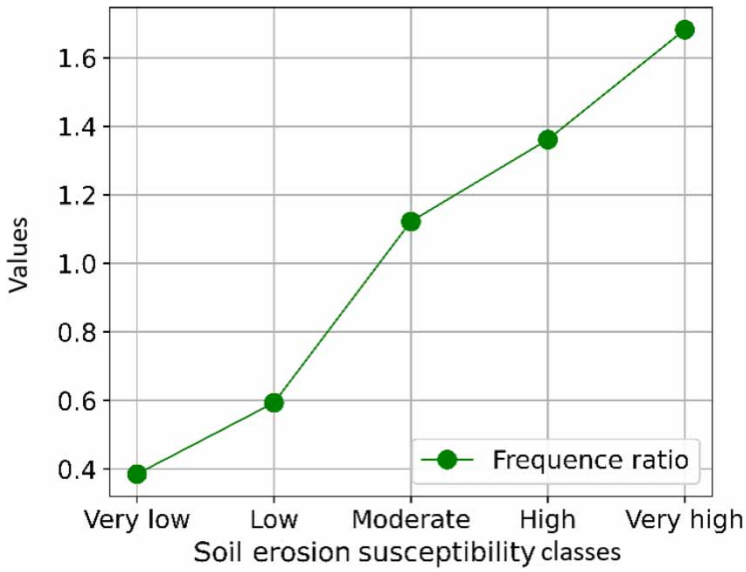

Figure 8 | Frequency ratio analysis of the susceptibility map generated from Gradient Boosting classifier.

development. These factors include slope degree, slope aspect, curvature, elevation, NDVI, rainfall, and soil type. Feature selection methods, including SHAP, were used to evaluate the importance of the relevant soil erosion influencing factors. All these factors were found to be relevant as input parameters in modeling. However, SHAP analysis indicated that soil type is the most important factor, followed by slope degree, rainfall, elevation, NDVI, aspect, and curvature, respectively.

The developed models were tuned using the GS technique, which proved successful in hyperparameter tuning. The performance of the models was evaluated using standard statistical measures including the ROC curve. Among all four models, the Gradient Boosting classifier performed the best, followed by the MLP classifier, Ridge classifier, and AdaBoost, respectively. Therefore, it is recommended that the Gradient Boosting classifier is an effective tool for SESM in the study area and could potentially be applied in other areas depending on ground conditions and local geo-environmental conditions.

It is important to note that factors such as glacier-induced erosion, solifluction, and wind erosion were not considered in this study as they were not relevant to the study area. In addition, it is important to note that in this work, the GS technique was used for turning the hyperparameters of the models; however, other methods such as random search and Bayesian optimization also have their own strengths for hyperparameter tuning. For instance, random search can be more efficient than GS when dealing with a large number of hyperparameters. Bayesian optimization, on the other hand, uses probability to find the optimal set of hyperparameters, which can be more effective in certain situations. Therefore, the choice of method depends on the specific requirements and constraints of the ML task.

DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Amiri, M., Pourghasemi, H. R., Ghanbarian, G. A. & Afzali, S. F. 2019 Assessment of the importance of gully erosion effective factors using Boruta algorithm and its spatial modeling and mapping using three machine learning algorithms. *Geoderma* **340**, 55–69.

Aslam, B. *et al.* 2021 Soil erosion susceptibility mapping using a GIS-based multi-criteria decision approach: Case of district Chitral, Pakistan. *Ain Shams Engineering Journal* **12**, 1637–1649.

Ayalew, D. A., Deumlich, D., Šarapatka, B. & Doktor, D. 2020 Quantifying the sensitivity of NDVI-based C factor estimation and potential soil erosion prediction using Spaceborne earth observation data. *Remote Sensing* **12**, 1136.

Bentéjac, C., Csörgő, A. & Martínez-Muñoz, G. 2021 A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review* **54**, 1937–1967.

Bien, T. X. *et al.* 2022 Landslide susceptibility mapping at sin Ho, Lai Chau province, Vietnam using ensemble models based on fuzzy unordered rules induction algorithm. *Geocarto International* **37**, 17777–17798.

Bien, T. X., Jaafari, A., Van Phong, T., Trinh, P. T. & Pham, B. T. 2023 Groundwater potential mapping in the Central Highlands of Vietnam using spatially explicit machine learning. *Earth Science Informatics* **16**, 131–146.

Bradford, J. & Foster, G. 1996 Interrill soil erosion and slope steepness factors. *Soil Science Society of America Journal* **60**, 909–915.

Chakrabortty, R. & Pal, S. C. 2023 Modeling soil erosion susceptibility using GIS-based different machine learning algorithms in monsoon dominated diversified landscape in India. *Modeling Earth Systems and Environment* **9**, 2927–2942.

Chakrabortty, R. *et al.* 2020 Soil erosion potential hotspot zone identification using machine learning and statistical approaches in eastern India. *Natural Hazards* **104**, 1259–1294.

Chakrabortty, R., Pal, S. C., Santosh, M., Roy, P. & Chowdhuri, I. 2022 Gully erosion and climate induced chemical weathering for vulnerability assessment in sub-tropical environment. *Geomorphology* **398**, 108027.

Ding, C. *et al.* 2023 Performance prediction for a fuel cell air compressor based on the combination of backpropagation neural network optimized by genetic algorithm (GA-BP) and support vector machine (SVM) algorithms. *Thermal Science and Engineering Progress* **44**, 102070.

D'souza, V. & Morgan, R. 1976 A laboratory study of the effect of slope steepness and curvature on soil erosion. *Journal of Agricultural Engineering Research* **21**, 21–31.

Egbueri, J. C., Igwe, O. & Unigwe, C. O. 2021 Gully slope distribution characteristics and stability analysis for soil erosion risk ranking in parts of southeastern Nigeria: A case study. *Environmental Earth Sciences* **80**, 292.

Elgeldawi, E., Sayed, A., Galal, A. R. & Zaki, A. M. 2021 Hyperparameter tuning for machine learning algorithms used for arabic sentiment analysis, informatics. *MDPI* 79.

Gayen, A., Pourghasemi, H. R., Saha, S., Keesstra, S. & Bai, S. 2019 Gully erosion susceptibility assessment and management of hazard-prone areas in India using different machine learning algorithms. *Science of the Total Environment* **668**, 124–138.

Golkarian, A., Khosravi, K., Panahi, M. & Clague, J. J. 2023 Spatial variability of soil water erosion: Comparing empirical and intelligent techniques. *Geoscience Frontiers* **14**, 101456.

Guragai, B., AlShorman, O., Masadeh, M. & Heyat, M. B. B. 2020 A survey on deep learning classification algorithms for motor imagery. In: *2020 32nd International Conference on Microelectronics (ICM)*. IEEE, pp. 1–4.

Hossain, R. & Timmer, D. 2021 Machine learning model optimization with hyper parameter tuning approach. *Glob. J. Comput. Sci. Technol. D Neural Artif. Intell* **21**.

Karaburun, A. 2010 Estimation of C factor for soil erosion modeling using NDVI in Buyukcekmece watershed. *Ozean Journal of Applied Sciences* **3**, 77–85.

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

Khosravi, K., Golkarian, A., Melesse, A. M. & Deo, R. C. 2022 Suspended sediment load modeling using advanced hybrid rotation forest based elastic network approach. *Journal of Hydrology* **610**, 127963.

Khosravi, K. *et al.* 2023 Soil water erosion susceptibility assessment using deep learning algorithms. *Journal of Hydrology* **618**, 129229.

Mangalathu, S., Hwang, S.-H. & Jeon, J.-S. 2020 Failure mode and effects analysis of RC members based on machine-learning-based SHapley Additive exPlanations (SHAP) approach. *Engineering Structures* **219**, 110927.

Martınez-Casasnovas, J., Ramos, M. & Ribes-Dasi, M. 2002 Soil erosion caused by extreme rainfall events: Mapping and quantification in agricultural plots from very detailed digital elevation models. *Geoderma* **105**, 125–140.

Martínez-Mena, M. *et al.* 2020 Long-term effectiveness of sustainable land management practices to control runoff, soil erosion, and nutrient loss and the role of rainfall intensity in Mediterranean rainfed agroecosystems. *Catena* **187**, 104352.

Mekonnen, M. & Melesse, A. M. 2011 Soil erosion mapping and hotspot area identification using GIS and remote sensing in northwest Ethiopian highlands, near Lake Tana. Nile River basin: hydrology, climate and water use, 207–224.

Mosavi, A. *et al.* 2020 Susceptibility mapping of soil water erosion using machine learning models. *Water* **12**, 1995.

Nearing, G. S. *et al.* 2021 What role does hydrological science play in the age of machine learning? *Water Resources Research* **57**, e2020WR028091.

Nguyen, K. A., Chen, W., Lin, B.-S. & Seeboonruang, U. 2021a Comparison of ensemble machine learning methods for soil erosion pin measurements. *ISPRS International Journal of Geo-Information* **10**, 42.

Nguyen, Q. H. *et al.* 2021b Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. *Mathematical Problems in Engineering* **2021**, 1–15.

Nhu, V.-H. *et al.* 2020 Landslide susceptibility mapping using machine learning algorithms and remote sensing data in a tropical environment. *International Journal of Environmental Research and Public Health* **17**, 4933.

Nouman, M. *et al.* 2023 Malicious node detection using machine learning and distributed data storage using blockchain in WSNs. *IEEE Access* **11**, 6106–6121.

Park, J. *et al.* 2022 Interpretation of ensemble learning to predict water quality using explainable artificial intelligence. *Science of the Total Environment* **832**, 155070.

Prasannakumar, V., Vijith, H., Abinod, S. & Geetha, N. 2012 Estimation of soil erosion risk within a small mountainous sub-watershed in Kerala, India, using Revised Universal Soil Loss Equation (RUSLE) and geo-information technology. *Geoscience Frontiers* **3**, 209–215.

Sadeghi, S. H., Moosavi, V., Karami, A. & Behnia, N. 2012 Soil erosion assessment and prioritization of affecting factors at plot scale using the Taguchi method. *Journal of Hydrology* **448**, 174–180.

Saha, S., Gayen, A., Pourghasemi, H. R. & Tiefenbacher, J. P. 2019 Identification of soil erosion-susceptible areas using fuzzy logic and analytical hierarchy process modeling in an agricultural watershed of Burdwan district, India. *Environmental Earth Sciences* **78**, 1–18.

Saha, A. *et al.* 2022 Threats of soil erosion under CMIP6 SSPs scenarios: An integrated data mining techniques and geospatial approaches. *Geocarto International* **37**, 17307–17339.

Sahana, M. *et al.* 2022 Rainfall induced landslide susceptibility mapping using novel hybrid soft computing methods based on multi-layer perceptron neural network classifier. *Geocarto International* **37**, 2747–2771.

Sahour, H., Gholami, V., Vazifedan, M. & Saeedi, S. 2021 Machine learning applications for water-induced soil erosion modeling and mapping. *Soil and Tillage Research* **211**, 105032.

Sajedi-Hosseini, F., Choubin, B., Solaimani, K., Cerdà, A. & Kavian, A. 2018 Spatial prediction of soil erosion susceptibility using a fuzzy analytical network process: Application of the fuzzy decision making trial and evaluation laboratory approach. *Land Degradation & Development* **29**, 3092–3103.

Stefano, C. D., Ferro, V., Porto, P. & Tusa, G. 2000 Slope curvature influence on soil erosion and deposition processes. *Water Resources Research* **36**, 607–617.

Sun, J., Fujita, H., Zheng, Y. & Ai, W. 2021 Multi-class financial distress prediction based on support vector machines integrated with the decomposition and fusion methods. *Information Sciences* **559**, 153–170.

Ulain, Q. *et al.* 2022 Identification of soil erosion-Based degraded land areas by employing a geographic information system – A case study of Pakistan for 1990–2020. *Sustainability* **14**, 11888.

van Breda Weaver, A. 1991 The distribution of soil erosion as a function of slope aspect and parent material in Ciskei, Southern Africa. *GeoJournal* **23**, 29–34.

Van Westen, C., Van Asch, T. W. & Soeters, R. 2006 Landslide hazard and risk zonation – why is it still so difficult? *Bulletin of Engineering Geology and the Environment* **65**, 167–184.

Vishnu, V. K. & Rajput, D. S. 2020 A review on the significance of machine learning for data analysis in big data. *Jordanian Journal of Computers and Information Technology* **6**.

Wu, S., Yu, M. & Chen, L. 2017 Nonmonotonic and spatial-temporal dynamic slope effects on soil erosion during rainfall-runoff processes. *Water Resources Research* **53**, 1369–1389.

Yang, D., Kanae, S., Oki, T., Koike, T. & Musiake, K. 2003 Global potential soil erosion with reference to land use and climate changes. *Hydrological Processes* **17**, 2913–2928.

Yang, J., Ma, H., Dou, J. & Guo, R. 2021 Harmonic characteristics data-driven THD prediction method for LEDs using MEA-GRNN and improved-AdaBoost algorithm. *IEEE Access* **9**, 31297–31308.

Yao, Y., Yang, X., Lai, S. H. & Chin, R. J. 2021 Predicting tsunami-like solitary wave run-up over fringing reefs using the multi-layer perceptron neural network. *Natural Hazards* **107**, 601–616.

Yesuph, A. Y. & Dagnew, A. B. 2019 Soil erosion mapping and severity analysis based on RUSLE model and local perception in the Beshillo Catchment of the Blue Nile Basin, Ethiopia. *Environmental Systems Research* **8**, 1–21.

Zhang, W. *et al.* 2021 Application of deep learning algorithms in geotechnical engineering: A short critical review. *Artificial Intelligence Review* 1–41.

First received 22 May 2023; accepted in revised form 2 November 2023. Available online 15 December 2023