Hindawi
Applied and Environmental Soil Science
Volume 2021, Article ID 6665485, 14 pages
https://doi.org/10.1155/2021/6665485

Research Article

Evaluation of Different Machine Learning Models for Predicting Soil Erosion in Tropical Sloping Lands of Northeast Vietnam

Tuan Vu Dinh,[1,2] **Nhat-Duc Hoang**,[2,3] **and Xuan-Linh Tran**[2,3]

[1]*VNU University of Science, Vietnam National University, Hanoi, Vietnam*
[2]*Institute of Research and Development, Duy Tan University, Da Nang 550000, Vietnam*
[3]*Duy Tan University, Da Nang 550000, Vietnam*

Correspondence should be addressed to Tuan Vu Dinh; vudinhtuan@vnu.edu.vn

Received 12 October 2020; Revised 11 March 2021; Accepted 22 March 2021; Published 5 April 2021

Academic Editor: Maman Turjaman

Soil erosion induced by rainfall under prevailing conditions is a prominent problem to farmers in tropical sloping lands of Northeast Vietnam. This study evaluates possibility of predicting erosion status by machine learning models, including fuzzy k-nearest neighbor (FKNN), artificial neural network (ANN), support vector machine (SVM), least squares support vector machine (LSSVM), and relevance vector machine (RVM). Model evaluation employed a historical dataset consisting of ten explanatory variables and soil erosion featured four different land use managements on hillslopes in Northwest Vietnam. All 236 data samples representing soil erosion/nonerosion events were randomly prepared (80% for training and 20% for testing) to assess the robustness of the five models. This subsampling process was repeatedly carried out by 30 rounds to eliminate the issue of randomness in data selection. Classification accuracy rate (CAR) and area under receiver operating characteristic (AUC) were used to evaluate performance of the five models. Significant difference between different algorithms was verified by the Wilcoxon test. Results of the study showed that RVM model achieves the best outcomes in both training (CAR = 92.22% and AUC = 0.98) and testing phases (CAR = 91.94% and AUC = 0.97). Four other learning algorithms also demonstrated good performance as indicated by their CAR values surpassing 80% and AUC values greater than 0.9. Hence, these results strongly confirm the efficacy of applying machine learning models for soil erosion prediction.

1. Introduction

Water erosion often causes loss of soil from the field, breakdown of soil structure, and decline of organic matter and nutrients [1]. Erosion leads to reduction of cultivable soil depth and decline in soil fertility, eventually reducing production. Furthermore, sedimentation downstream reduces the capacity of rivers, reservoirs, and drainage ditches, which shortens their designs' life. It also enhances the risk of flooding and blocks irrigation channels [2]. The soil erosion severity is highly variable depending on site's climate, soil, topography, cropping, and land management [3]. Particularly, soil erosion potential in tropical areas is high due to heavy rainfall coupled with land management such as mono cropping in the uplands of Northwest Vietnam [4, 5]. Accelerated erosion is often observed at the beginning of the cropping season when heavy rains coincide with poor ground cover [6]. Climatic condition, soil characteristic, land form, and land management significantly contribute to soil erosion in different weights that need to be investigated.

Soil loss studies at the plot scale have been of crucial importance to identify the mechanism of the processes. The erosion plot experiments can help to introduce new erosion prevention technologies as it provides access to reliable and consistent erosion measurements and large numbers of data necessary to test new models [7]. Most recent empirical models employed data from plot studies such as USLE/RUSLE based on Universal Soil Loss Equation $A = RKLSCP$, where A is computed soil loss, R is the rainfall-runoff erosivity factor, K is a soil erodibility factor, L is the slope length factor, S is the slope steepness factor, C is a cover management factor, and P is a supporting practices factor

[8, 9], SWAT [10], a physically based model Water Erosion Prediction Project (WEPP) [11], and Tradeoffs (InVEST) Sediment Delivery Ratio (SDR) model [12], etc.

Machine learning approaches could provide a helpful alternative to deal with the multivariate and complex nature of problems in soil science and geoscience [13–16]. Artificial neural network (ANN) generally predicts soil loss at acceptable results [17, 18] or even better than that of WEPP model (2011) [19]. Kohonen Neural Networks (KNN), multivariate adaptive regression splines [20], and support vector classification coupled with metaheuristic [21] being used for runoff-erosion modeling had shown a superior result to the conventional multiple linear regression model [22]. Soil erosion prediction is a complex and dynamic process, requiring comparison of various advanced machine learning algorithms. Machine learning has demonstrated great potentiality and effectiveness for solving complex soil science problems. This modern method can construct data-driven models from historical datasets and establish prediction models used for predicting various complex phenomena including soil erosion [23–25].

This study elucidates potential application of five competent machine learning models to predict soil erosion: artificial neural network (ANN), support vector machine (SVM), least squares support vector machine (LSSVM), relevance vector machine (RVM), and fuzzy k-nearest neighbor (FKNN) using a dataset containing ten explanatory variables, collected from fields in Northwest Vietnam. The ANN method is inspired from the actual neural systems of human brain; this method possesses the universal approximating capability and can accurately approximate any nonlinear function [26, 27]. SVM is a robust machine learning model which is based by the structural risk minimization [28]; therefore, SVM is less susceptible to overfitting than ANN. LSSVM and RVM can be considered as variants of the original SVM. The first reformulates the model training procedure of SVM so that it is only required to solve a linear system instead of a constrained nonlinear programming problem in SVM [29]. The latter model of RVM takes advantage of Bayesian framework to construct more robust and sparse models which may result in less numbers of support vectors than the standard SVM [30]. The sparseness property of a RVM means that this approach can be resilient to noise and less susceptible to noisy data samples [31, 32]. In addition, the FKNN [33] is an extension of the standard k-nearest neighbor (KNN) algorithm; this model incorporates the fuzzy theory into the KNN model structure to enhance the flexibility of data modeling and better constructs the class decision boundary. Due to such characteristics and advantages, these five models are selected to be employed in this study.

2. Research Methodology

2.1. The Dataset. The erosion dataset was collected from two experiments that featured four different land use managements in Northwest Vietnam during three years (2009–2911). Details of the experiments have been described in [34]. In brief, erosion plots were arranged in a randomized complete block design with four treatments, three replicates. The employed treatments represent conventional local farmers' maize cropping practice based on slashing, burning, and ploughing with fertilization (1), and soil conservation practices such as grass barrier (*Panicum maximum*) (2), minimum tillage with cover crop (*Arachis pintoi*) (3), or relay cropping with Adzuki beans (*Phaseolus calcaratus*). The minimum tillage and/or cover crop option provided better land cover and less disturbed soil condition, hence lowering soil loss. Each plot is sized 72 m^2 (4 m wide and 18 m slope length), laid on slopes within 24.8–34.8 degrees. A system of buckets was installed to collect the deposited sediment subjected to soil erosion from the above plots. Erosion data were recorded on storm basic in the three years: 2009–2011.

2.2. Description of Soil Erosion Data. Climate, soil, topography, and land use factors affect rill and inter-rill soil erosion caused by raindrop impact and surface runoff. More precisely, soil erosion depends on the erosivity caused by the amount and intensity of rainfall and runoff, and the resistance of the soil surface or the degree of erodibility caused by intrinsic soil properties, adopting land use practices, and the topography of the landscape as described by slope length and steepness. To represent these factors, a set of ten explanatory variables has been chosen as described in Table 1. Data distributions are shown in the histograms (Figure 1). In this study, we classify the dependent variables either as "erosion" or "nonerosion." When soil loss measured in the field is greater than 3 tons per hectare, it is considered as a significant erosion in tropical regions [36]; otherwise, the loss is negligible. A total of 236 data samples had been collected, within which 118 records were classified as "erosion."

$$E = 1099[1 - 0.72\ \exp(-1.27i)]. \quad (1)$$

OC denotes organic matter.

3. Machine Learning Methods for Soil Erosion Status Prediction

3.1. Artificial Neural Network (ANN). ANN is a widely employed machine learning method inspired by biological neural networks. This method simulates the knowledge acquisition and reasoning processes occurring the human brain [37–41]. Given the learning task is to train a function $f: X \in R^D \longrightarrow Y \in R^1$, where D denotes the number of input attributes, an ANN model employed to learn the function f typically includes the input, hidden, and output layers.

Via a training process, the knowledge learnt by an ANN model is adapted and stored in the form of matrices of connection weights. Generally, the parameters of an ANN model are trained via a process that employs the framework of error backpropagation [42, 43]. Overall, an ANN-based soil erosion classification model can be expressed as follows:

Table 1: Statistical descriptions of soil erosion influencing factors [20].

Influencing factors	Notation	Min	Max	Mean	Std.
EI30 (MJ*mm/ha∗hr)	X_1	0.000	3008.930	573.642	814.696
Slope (degree)	X_2	24.830	34.770	29.049	2.324
OC topsoil (%)	X_3	0.890	2.790	1.747	0.584
pH topsoil	X_4	5.130	7.060	5.866	0.581
Topsoil bulk density g/cm^3	X_5	1.230	1.580	1.398	0.080
Topsoil porosity (%)	X_6	46.340	59.480	52.762	3.016
Topsoil texture (silt fraction %)	X_7	31.350	37.710	33.902	1.486
Topsoil texture (clay fraction %)	X_8	18.610	38.350	29.138	4.807
Topsoil texture (sand fraction %)	X_9	29.660	46.510	36.954	4.375
Soil cover (%)	X_{10}	1.050	97.640	44.276	26.741

Note: *EI30* denotes the kinetic rainfall energy which is the product of total storm energy (E) times the maximum 30 min intensity (I30). Storm energy E is adapted for tropical condition [35]:

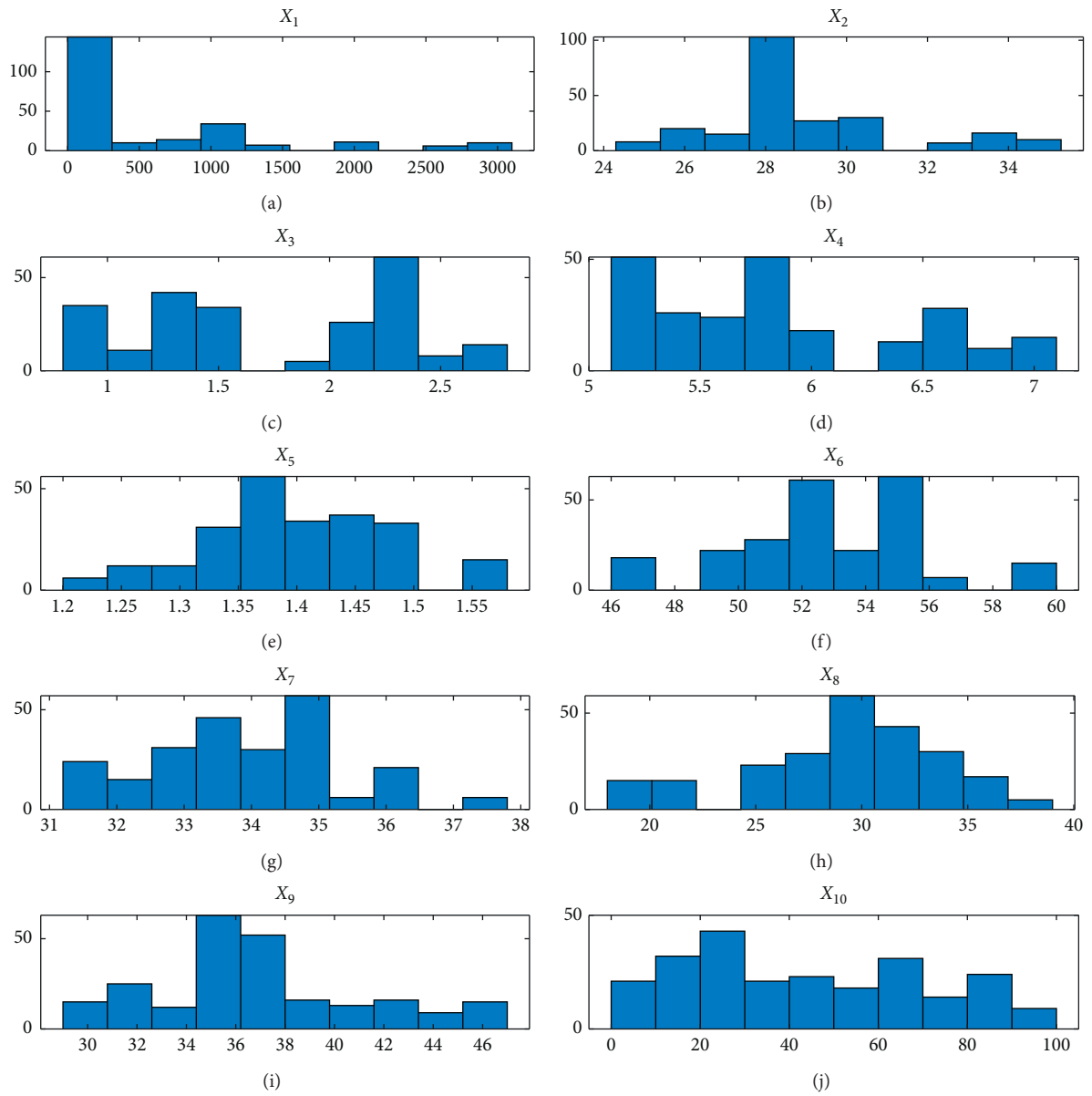


Figure 1: Data distribution.

$$f(X) = \mathrm{SM}[b_1 + W_{L1L2} \times (f_A(b_0 + W_{L0L1} \times X))], \quad (2)$$

where b_0 and b_1 denote the two bias vectors of the input and hidden layers, respectively, f_A represents the activation function. SM is the softmax activation function [44, 45], W_{L0L1} is the matrix of connection weights between the input and hidden layer, and W_{L1L2} denotes that between the hidden and the softmax layer.

The softmax activation function used to compute the class probability is expressed as follows:

$$\delta(z) = \frac{\exp(z_i)}{\sum_{i=0}^{CN-1} \exp(z_i)}, \quad (3)$$

where CN represents the number of output classes.

3.2. Support Vector Machine (SVM). Proposed by [28], the SVM algorithm was a powerful method for linear binary classification. The algorithm aims at constructing a hyperplane to separate positive and negative samples with the margin as large as possible. The SVM models are highly suitable for medium-size datasets and are less susceptible to overfitting than ANN models [46–48]. Given a training dataset $\{x_k, y_k\}_{k=1}^{N}$ with input data $x_k \in R^n$ and corresponding class labels $y_k \in \{-1, +1\}$, the SVM algorithm establishes a decision boundary so that the gap between classes is as large as possible. Moreover, SVM relies on the kernel trick to cope with nonlinear classification problems [49–51]. The formulation of the SVM training process can be described as the following optimization problem:

$$\begin{aligned} &\text{minimize } J_p(w, e) = \frac{1}{2} w^T w + c \frac{1}{2} \sum_{k=1}^{N} e_k^2, \\ &\text{subjected to } y_k\left(w^T \varphi(x_k) + b\right) \geq 1 - e_k, \quad k = 1, \ldots, N, \quad e_k \geq 0, \end{aligned} \quad (4)$$

where $w \in R_n$ denotes a normal vector to the classification hyperplane and $b \in R$ represents the model bias; $e_k > 0$ denotes slack variables; c is a penalty constant; and $\varphi(x)$ represents a nonlinear mapping from the input space to the high-dimensional feature space.

By solving the above constrained optimization problem, the final SVM model used for pattern classification is expressed as follows [52]:

$$y(x_l) = \mathrm{sign}\left(\sum_{k=1}^{SV} \alpha_k y_k K(x_k, x_l) + b\right), \quad (5)$$

where α_k is the solution of the dual form of the optimization described in equation (2), SV represents the number of support vectors (the number of $\alpha_k > 0$), and $K(.)$ denotes the radial basis function (RBF) kernel [52]:

$$K(x_k, x_l) = \exp\left(-\frac{\|x_k - x_l\|^2}{2\sigma^2}\right), \quad (6)$$

where σ denotes the RBF parameter.

3.3. Least Squares Support Vector Machine (LSSVM). LSSVM is a least squares version of the standard SVM within which the model structure is identified by solving a set of linear system instead of a nonlinear optimization problem [53, 54]. Similar to the standard SVM, the LSSVM relies on kernel functions to deal with complex and nonlinear datasets [55–57]. The LSSVM formulation for pattern classification can be stated as follows [58]:

$$\begin{aligned} &\text{minimize } J_p(w, e) = \frac{1}{2} w^T w + \gamma \frac{1}{2} \sum_{k=1}^{N} e_k^2, \\ &\text{subjected to } y_k\left(w^T \varphi(x_k) + b\right) = 1 - e_k, \quad k = 1, \ldots, N, \end{aligned} \quad (7)$$

where $w \in R^n$ is the normal vector to the classification hyperplane, and $b \in R$ is the bias; $e_k \in R$ represents error variables; and $\gamma > 0$ denotes a regularization constant.

By solving the above optimization problem, the LSSVM classification model can be expressed as follows:

$$y(x) = \mathrm{sign}\left(\sum_{k=1}^{N} \alpha_k y_i K(x_k, x_l) + b\right), \quad (8)$$

where α_k and b are the solution of the systems stated in equation (4). $K(.)$ also denotes the RBF kernel function [54].

3.4. Relevance Vector Machine (RVM). RVM, proposed by [59], is a Bayesian inference-based method that can be employed for solving classification problems. The functional form of RVM is similar to that of the support vector machine. Furthermore, an expectation maximization based method is utilized to construct the RVM prediction model [60].

Compared to the aforementioned SVM and LSSVM, the Bayesian-based RVM requires fewer tuning parameter; hence, the model construction phase of the RVM can be fast to accomplish [61, 62]. Furthermore, a RVM model often results in good predictive performance thanks to its sparseness property. It is because a RVM model relies on a small number of relevant vectors extracted from the training samples to construct the classification model [31].

The RVM-based classification model is presented compactly as follows [30]:

$$y(x, w) = \sum_{m=1}^{M} w_m \varphi_m(x) + w_o = w \cdot \varphi, \quad (9)$$

where $w = [w_0, w_1, \ldots, w_M]$ represents a vector of the model weights and $\varphi = [1, \varphi_1(x_t), \varphi_2(x_t), \ldots, \varphi_M(x_t)]$ denotes a vector of Gaussian basis functions.

The Gaussian basis function basis is given as follows : $\varphi_m(x)$

$$= \exp\left(\frac{-\|x - x_m\|^2}{2 \times b^2}\right), \quad (10)$$

where b represents the width of the Gaussian basis function.

3.5. Fuzzy k-Nearest Neighbor (FKNN). Proposed by [33], the FKNN algorithm is an extension of the original k-nearest neighbor [63]. One major advantage of the FKNN is that it takes into account the distances among samples. The FKNN utilizes the concept of fuzzy logic to express the membership strength of data instances in each class. The membership degree of a data instance in a class is computed as a function of distance to its nearest neighbors [64].

The FKNN classifier computes a fuzzy partition matrix $U = [u_{ij}]$ as follows [64]:

$$u_{ij}(x) = u_i\left(x_j\right) = \begin{cases} 0.51 + \left(\frac{n_i}{k}\right) \times 0.49, & if\ c\left(x_j\right) = i \\ \left(\frac{n_i}{k}\right) \times 0.49, & if\ c\left(x_j\right) \neq i \end{cases}, \tag{11}$$

where n_i denotes the number of neighbors of the data instance x_j that is actually in the $i\ th$ class and $c(x_j)$ represents the class label of x_j.

Based on the matrix U, the fuzzy memberships of a new sample x in different classes can be obtained and the class label having the largest membership degree is selected as the output for a new input data x. The fuzzy memberships of x are computed as follows:

$$u_i(x) = \frac{\sum_{j=1}^{k} u_{ij}\left(1/\left(\left\|x - x_j\right\|^{(2/(m-1))}\right)\right)}{\sum_{j=1}^{k}\left(1/\left(\left\|x - x_j\right\|^{(2/(m-1))}\right)\right)}, \tag{12}$$

where $i = 1,2,\ldots, C$, and $j = 1,2,.., k$. Moreover, k is the number of nearest neighbors. The parameter m is called the fuzzy strength coefficient.

3.6. Performance Metrics. The classification accuracy rate (CAR) is employed to measure and compare the performance of classifiers. CAR is the percentage of correct classified cases calculated by the following equation:

$$\mathrm{CAR} = \frac{Nc}{Na} \times 100\%, \tag{13}$$

where Nc and Na represent the numbers of correctly classified instances and the total number of instances, respectively.

In addition to CAR, true positive rate (TPR) (the percentage of positive instances correctly classified), false positive rate (FPR) (the percentage of negative instances misclassified), false negative rate (FNR) (the percentage of positive instances misclassified), and true negative rate (TNR) (the percentage of negative instances correctly classified) are also utilized to quantify the performance of classifier [65]. The formulation for calculating the above four metrics is as follows:

$$\begin{aligned} \mathrm{TPR} &= \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \\ \mathrm{FPR} &= \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}, \\ \mathrm{FNR} &= \frac{\mathrm{FN}}{\mathrm{TP} + \mathrm{FN}}, \\ \mathrm{TNR} &= \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}, \end{aligned} \tag{14}$$

where TP, TN, FP, and FN represent the numbers of true positive, true negative, false positive, and false negative, respectively.

Receiver operating characteristic, a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied, can be applied to summarized TPR and FPR [65]. The area under the ROC curve, or AUC for short, can be calculated to quantitatively exhibit the classification performance of a model [66].

3.7. Selection of Model Parameters. All values representing soil erosion/nonerosion are randomly divided into two subsets: a training dataset (80%) was used for model establishment and a test dataset (20%) was used to measure the model generalization capability. The data were normalized into the range between 0 and 1 with the Z-score normalization by the following equation:

$$X_n = \frac{X_o - \mu_x}{s_x}, \tag{15}$$

where X_n and X_o denote the normalized and the original input variables, respectively, and μ_x and s_x are the mean and the standard deviation of the variable X_o, respectively.

Five machine learning algorithms, ANN, SVM, LSSVM, RVM, and FKNN, were used to establish the soil erosion status prediction models. The FKNN model is coded in MATLAB environment by the authors. The ANN and SVM models are implemented in MATLAB environment with the Statistics and Machine Learning Toolbox [67]. The LSSVM and RVM models are established via the toolboxes developed by [68, 69], respectively.

A fivefold cross-validation procedure coupled with grid search was carried out to identify appropriate free parameters for model performance. The model training and prediction was repeatedly carried out five times on five mutual exclusive groups being separated from the whole dataset. Model selection was based on a set of free parameters that leads to the highest average CAR. Moreover, the grid search procedure employed for a model with two free parameters is described in Algorithm 1.

4. Results and Discussion

4.1. A Preliminary Analysis on the Relevancy of Input Factors with Mutual Information. A preliminary analysis on feature relevancy was evaluated prior to model training and prediction phases. This analysis may help to identify irrelevant

```
Establishing Training Subset 1, Training Subset 1
Establishing parameter pools P1, P2
PM = {∅}// Performance matrix
For i = 1: NHP1//NHP1 = number of available hyperparameters in P1
   p1 = P1(i)
  For j = 1: NHP2//NHP2 = number of available hyperparameters in P2
    p2 = P2(j)
    For f = 1:5//5 is the number of data folds
    Train the prediction model with the training set (80%)
      Model prediction with the testing subset (20%)
  CarFold(f) = CAR (%) of the testing subset
    End For
    PM(i,j) = CarFol d(f)/5
  End For
End For
Finding the best set of p1 and p2 based on PM
```

Algorithm 1: A typical grid search procedure for parameter selection.

input variables. In this study, the mutual information method [70] was utilized to compute the independence relationship of each conditioning factor to the class label (erosion/nonerosion). It is proper to note that large mutual information indicates a strong relevancy between the conditioning factor and the class label. The analysis result is shown in Figure 2 that provides the mutual information values of all input variables. It is clearly shown that the input factor X_8 (topsoil texture-clay) obtains the highest mutual information value, followed by the input factor X_1 (EI30), X_9 (topsoil texture-sand), X_{10} (soil cover), and X_4 (pH topsoil). The factors X_3 (OC top soil), X_2 (slope), and X_5 (topsoil bulk density) receive comparatively low mutual information values. Since all mutual information values are not null, the subsequent model establishment phase should take into account all of the ten factors. This study shows that mutual information of slope (factor 2) is lower than pH (factor 4), which may be explained by change in pH causing physical properties of the soil to change in the clayey soil rich in Al, Ca, and Mg [34] leading to the development of soil water erosion [71].

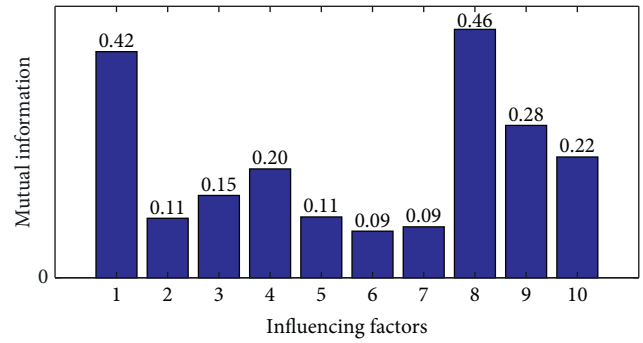


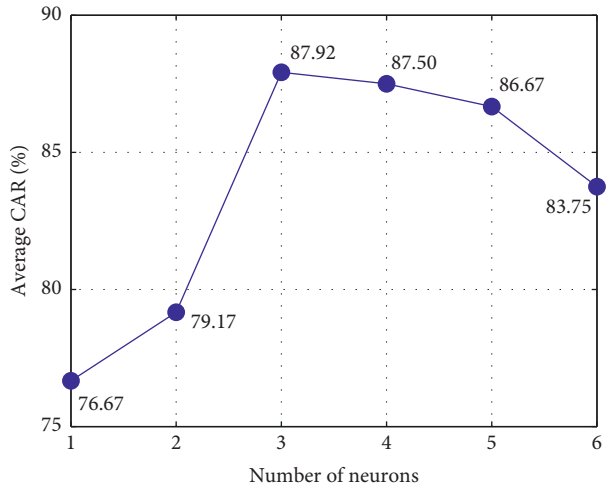
Figure 2: Preliminary analysis of feature relevancy with mutual information.

4.2. Model Calibration. The ANN model requires the selection of the number of neurons in the hidden layers and the learning rate. In this experiment, we study the number of neurons within the range of 5 to 30 and the learning rate parameter within the set of [0.001, 0.01, 0.1, 1]. The model performances of ANN with different number of neurons are reported in Figure 3(a). The best ANN model (CAR = 87.92%) corresponds to a model consisting of 15 neurons and the learning rate = 0.01.

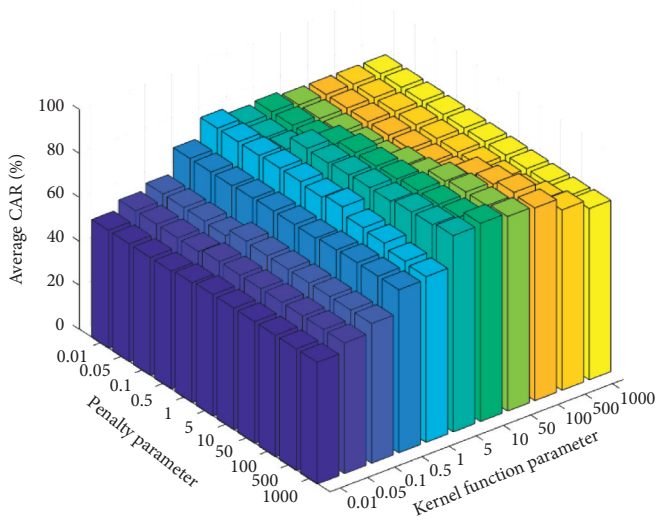
For the case of SVM, the model performance corresponding to different sets of the penalty parameter c and the kernel function parameter σ is investigated. It is worth noticing that the parameter c influences the model complexity and the parameter σ affects the smoothness of the classification boundary of SVM. These two parameters of an SVM model are allowed to be varied within the range of 0.01 and 1000. The SVM model performance with each pair of c and σ is illustrated in Figure 3(b). The best values of the penalty parameter c and the kernel function parameter σ are 1000 and 10 (with CAR = 90.11%), respectively.

Figure 3(c) reports the model selection of LSSVM in which the regularization (γ) and the kernel function (σ) parameters are studied. The best LSSVM corresponds to $\gamma = 10$ and $\sigma = 5$ with CAR = 88.50%. In the case of FKNN, the highest model accuracy obtained from the fivefold cross-validation is accompanied with the value of nearest neighbors $k = 3$ and the fuzzy strength $m = 2$ (see Figure 3(d)). In addition, the classification accuracy of RVM corresponding to different values of the Gaussian bandwidth (b) is provided in Figure 3(e) in which $b = 0.015$ is the most suitable value that leads to an average CAR = 91.45%.

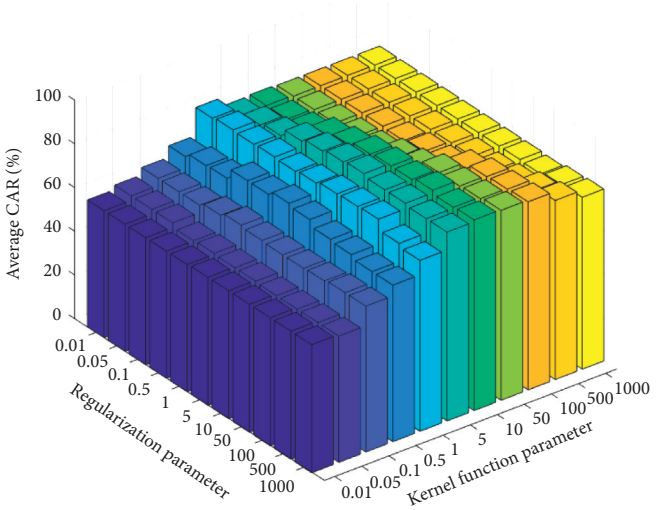
4.3. Water Erosion Prediction Modeling. It is noted that a single run of experiment cannot reliably exhibit the capability of the soil erosion status prediction model due to the issue of randomness in data selection. Thus, a repeated subsampling process consisting of 30 runs was carried out. After 30 runs, the performance metrics of the five employed models are summarized in Table 2. Figure 4 illustrates the



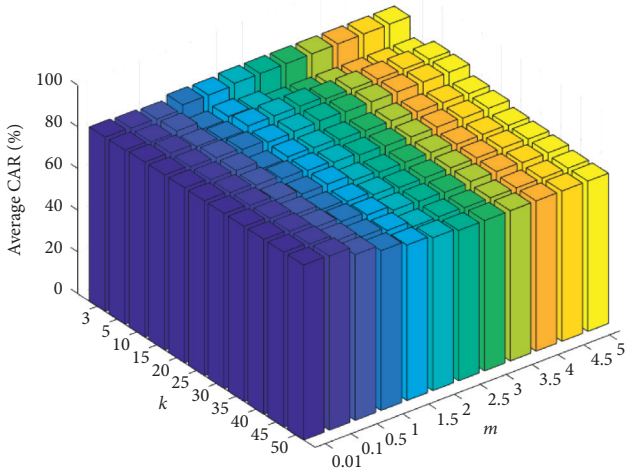
(a)



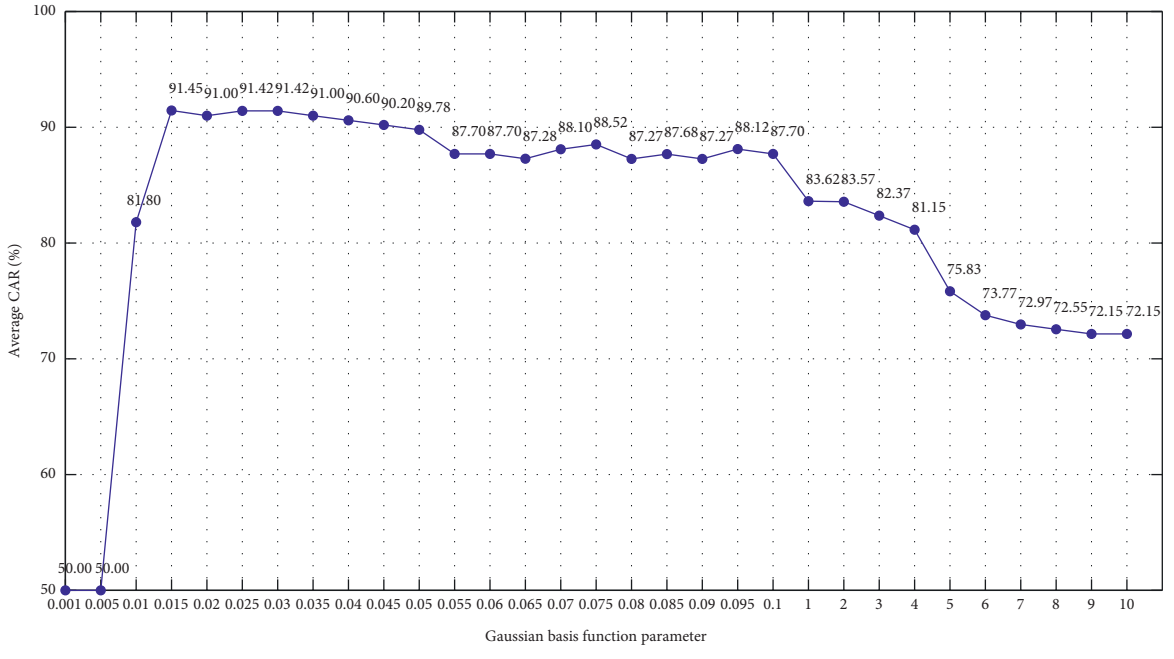
(b)



(c)



(d)

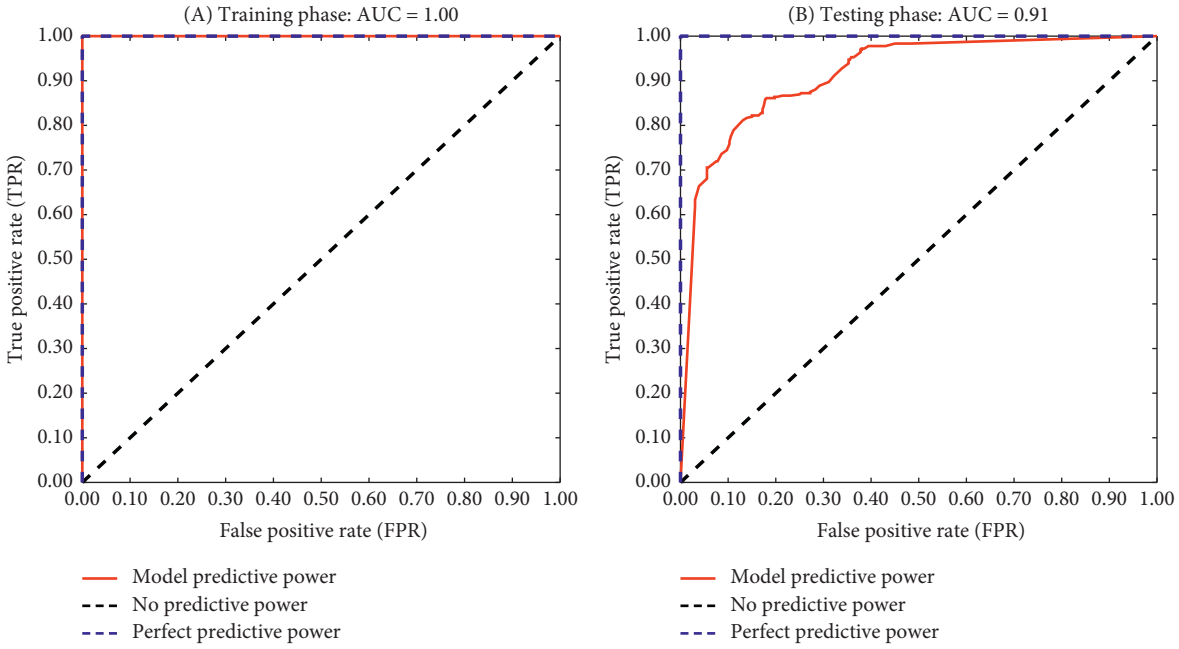


(e)

Figure 3: Model performance analysis: (a) ANN, (b) SVM, (c) LSSVM, (d) FKNN, and (e) RVM.

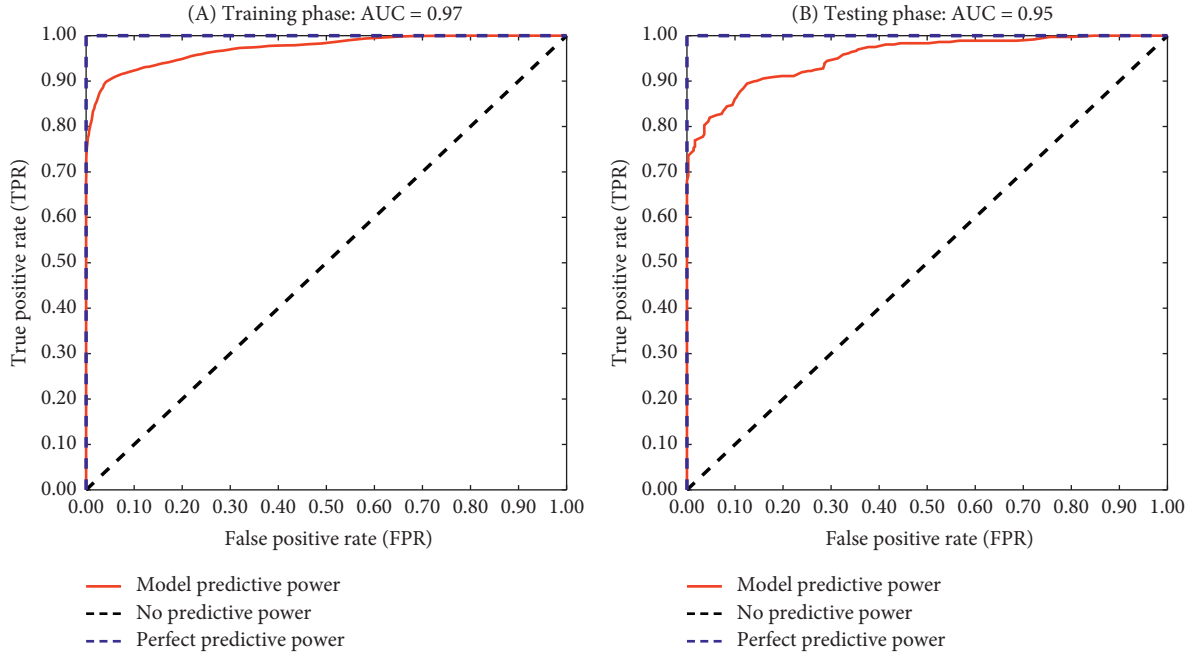
Table 2: Performance metrics of five soil erosion prediction models.

Metrics	FKNN		ANN		SVM		LSSVM		RVM	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Training phase										
CAR (%)	100.00	0.00	92.72	0.95	96.21	0.71	93.02	0.69	92.22	0.86
AUC	1.00	0.85	0.97	0.00	0.97	0.01	0.94	0.01	0.98	0.00
TPR	1.00	0.00	0.90	0.01	0.95	0.01	0.88	0.01	0.90	0.02
FPR	0.00	0.00	0.05	0.01	0.03	0.01	0.02	0.01	0.05	0.01
FNR	0.00	0.00	0.10	0.01	0.05	0.01	0.12	0.01	0.10	0.02
TNR	1.00	0.00	0.95	0.01	0.97	0.01	0.98	0.01	0.95	0.01
Testing phase										
CAR (%)	83.19	6.33	88.33	5.62	85.97	7.62	88.61	5.12	91.94	5.99
AUC	0.91	0.05	0.95	0.03	0.86	0.09	0.91	0.05	0.97	0.03
TPR	0.86	0.09	0.88	0.08	0.89	0.07	0.83	0.09	0.91	0.08
FPR	0.20	0.11	0.11	0.10	0.17	0.13	0.06	0.08	0.07	0.09
FNR	0.14	0.09	0.13	0.08	0.11	0.07	0.17	0.09	0.09	0.08
TNR	0.80	0.11	0.89	0.10	0.83	0.13	0.94	0.08	0.93	0.09

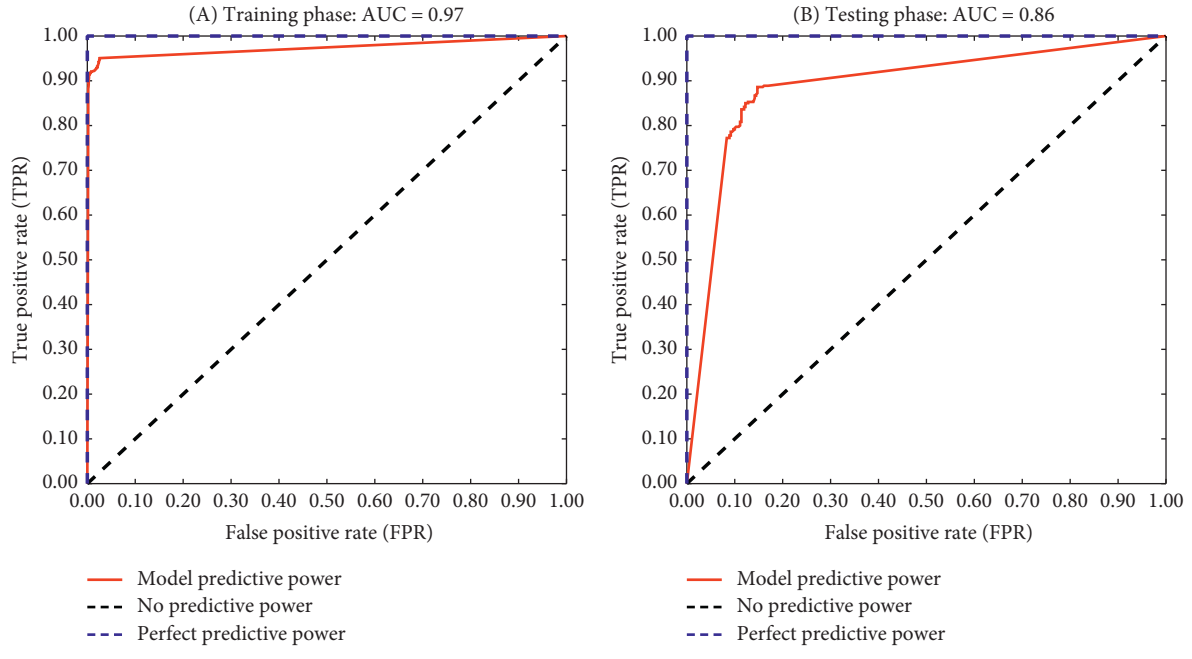


(a)

Figure 4: Continued.



(b)



(c)

Figure 4: Continued.

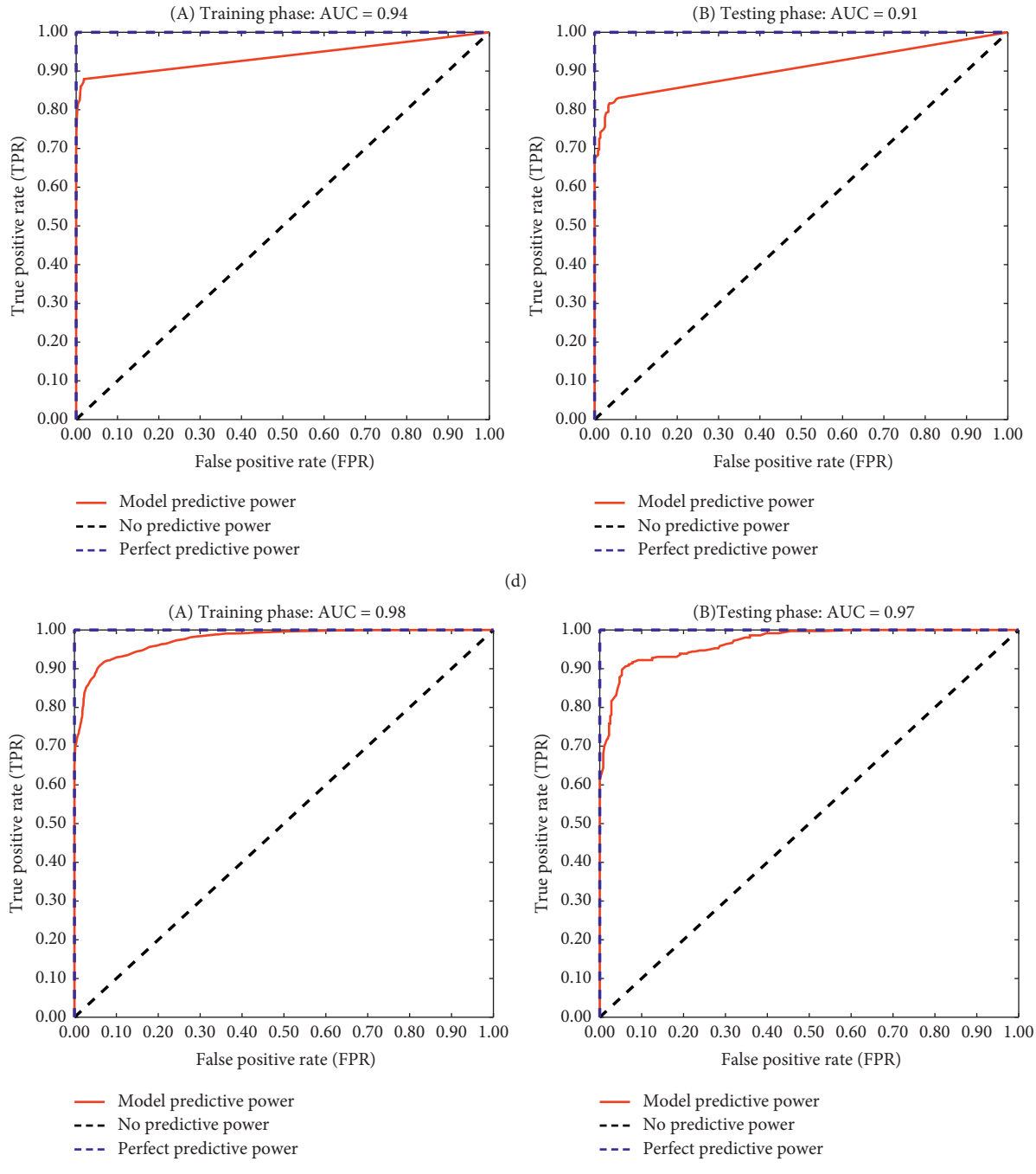


Figure 4: Receiver operating characteristic curves of models: (a) FKNN, (b) ANN, (c) SVM, (d) LSSVM, and (e) RVM.

average ROC curves with each model in both training and testing phases.

The RVM model has obtained the most desirable prediction accuracy in the testing phase (CAR = 91.74% and AUC = 0.97) (Table 2). The LSSVM is the second best model (CAR = 88.61% and AUC = 0.91); ANN ranks as the third model (CAR = 88.33% and AUC = 0.95), followed by SVM (CAR = 85.97% and AUC = 0.86) and FKNN (CAR = 83.19% and AUC = 0.91). The results also point out exceptionally high values of TPR (0.90) and TNR (0.94) yielded by RVM. The box plot shown in Figure 5 summarizes the CAR and AUC results of the five models obtained from 30 runs.

In addition, the Wilcoxon signed-rank test [72] was employed to investigate whether the prediction performances of each pair of methods were statistically different. This is a nonparametric hypothesis test used for model comparison. The significance level of the test (p value) was set to be 0.05. Based on the threshold p value = 0.05, if the p value of the test was lower than 0.05, we could reject the null hypothesis that the performances of the two models of

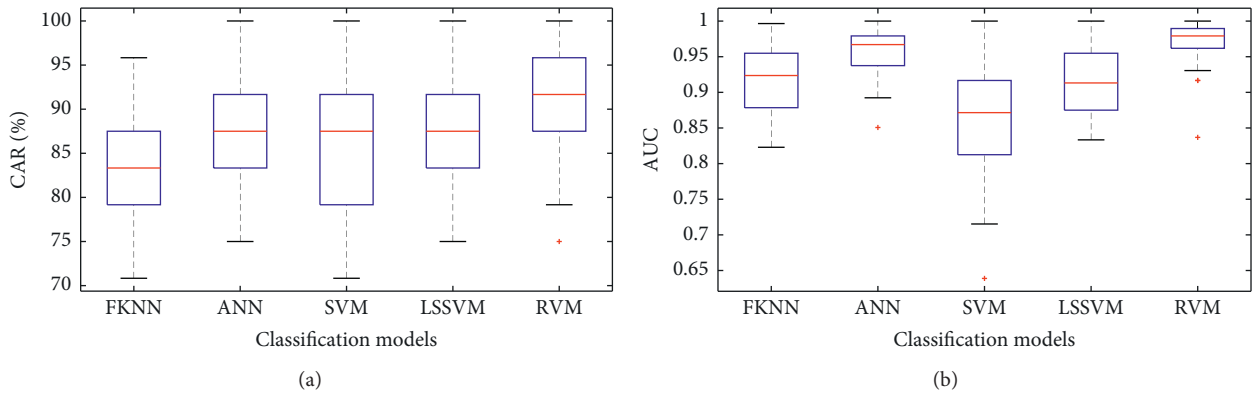


Figure 5: Performance of soil erosion prediction models: (a) CAR and (b) AUC.

Test result					
Models	FKNN	ANN	SVM	LSSVM	RVM
FKNN	x	−−	−	−−	−−
ANN	++	x	+	−	−−
SVM	++	−	x	−	−−
LSSVM	++	+	+	x	−−
RVM	++	++	++	++	x

p-values					
Models	FKNN	ANN	SVM	LSSVM	RVM
FKNN	x	0.0054	0.0823	0.0056	0.0002
ANN	0.0054	x	0.1725	0.8599	0.0337
SVM	0.0823	0.1725	x	0.2616	0.0043
LSSVM	0.0056	0.8599	0.2616	x	0.0125
RVM	0.0002	0.0337	0.0043	0.0125	x

Figure 6: Performance comparison of the five soil erosion prediction models using Wilcoxon signed-rank test.

interest are statistically indifferent. Comparison of each pair of models is presented in Figure 6. In this table, the symbols "++," "+," "−−," and "−" stand for a significant win, a win, a significant loss, and a loss, respectively. Observably, RVM attains four significant wins over other benchmark models. LSSVM, as the second best approach, obtains a significant win over FKNN, and two wins over ANN and SVM. FKNN receives three significant losses in the duals with ANN, LSSVM, and RVM, and one loss in the dual with SVM.

Based on the experimental results supported by the employed statistical test, it can be stated with confidence that the RVM is the best suited method for the current dataset. The outstanding performance of this machine learning approach can be explained by its advantages including the ease of model establishment and improved generalization. The first advantage of the RVM may stem from the fact that this model only requires one hyperparameter which is the width of the Gaussian basis function. The second advantage of the model is based on the model sparseness; the RVM only selects a small portion of the training samples as crucial data points to construct the classification model. Therefore, this advanced machine learning model is less susceptible to noisy data points than other employed machine learning approaches. Based on these findings, the RVM is strongly recommended for soil erosion prediction problems under this tropical prevailing condition. Broader spectrum of data collected in wider conditions is required for a more comprehensive prediction in the future.

It is proper to note that most conventional erosion prediction models based on physical or empirical or both face difficulties in model development and in predictive accuracy. Moreover, model parameters often need to be calibrated against observed data, creating problems with model identification and the physical interpretability of model parameter [73]. Developing concepts for erosion processes requires a considerable length of time due to the natural complexity of the systems where erosion occurs.

Furthermore, the appropriateness of erosion concepts commonly employed in model structures is still questionable [74]. Despite the fact that physical processes of detachment, transport, and deposition in overland flow are well recognized and have been widely incorporated within erosion models, the experimental procedures to test conditions when processes are occurring concurrently have only recently been developed [75]. Our initiative of using machine learning approaches therefore proves to be a promising alternative for erosion prediction in which it overcomes obstacles in parameterization, calibration, and validation processes that are often considered to be the main difficulties while applying conventional models.

5. Conclusion

This study evaluated performances of five machine learning algorithms, namely, FKNN, ANN, SVM, LSSVM, and RVM, using a historical dataset collected in tropical slopping fields featured by ten soil erosion conditioning factors. Experimental results supported by the Wilcoxon signed-rank test pointed out that RVM was deemed best suited for the problem at hand. The RVM model achieved the best

performance in the testing phase (CAR = 91.94% and AUC = 0.97). Four other learning algorithms also demonstrated good performance as indicated by their CAR values surpassing 80% and AUC values greater than 0.9. Thus, these results strongly confirm the efficacy of applying machine intelligence for solving the problem of interest. Furthermore, RVM can be a very promising tool to assist landowners and managers to quickly identify potential soil erosion areas and develop preventive measures. The reasons for the good performance of RVM may lie in the fact that this model utilizes Bayesian inference to obtain parsimonious solutions the soil erosion prediction problem in this study which is modeled as a pattern classification task. The employed Bayesian inference of RVM can help to result in a robust classification model which features a small number of support vectors. Therefore, the decision boundary constructed by such support vectors has good generalization property and resilience to noise. These facts explain why predictive accuracy of RVM is better than those obtained from other machine learning models.

Future extensions of the current works may include the following:

(i) Investigation of the capabilities of other advanced machine learning models (such as tree ensemble, functional tree, gradient boosted regression tree, stochastic gradient tree boost, alternating decision tree, logistic model tree, boosted regression trees, random forest, and naive Bayes variants) in soil erosion prediction

(ii) Collection of more data samples to increase the current data size and therefore enhance the generalization as well as applicability of the current data-driven models

(iii) Investigation of other influencing factors of soil erosion to ameliorate the explicability of the current study

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was funded by the Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 105.08-2017.302.

References

[1] M. Spekken, S. De Bruin, J. P. Molin, and G. Sparovek, "Planning machine paths and row crop patterns on steep surfaces to minimize soil erosion," *Computers and Electronics in Agriculture*, vol. 124, pp. 194–210, 2016.

[2] R. P. C. Morgan, *Soil Erosion and Conservation*, Blackwell Science Ltd, Oxford, England, 3rd edition, 2005.

[3] S. A. El-Swaify, "Factors affecting soil erosion hazards and conservation needs for tropical steeplands," *Soil Technology*, vol. 11, pp. 3–16, 1997.

[4] G. Clemens, S. Fiedler, N. D. Cong, N. Van Dung, U. Schuler, and K. Stahr, "Soil fertility affected by land use history, relief position, and parent material under a tropical climate in NW-Vietnam," *Catena*, vol. 81, no. 2, pp. 87–96, 2010.

[5] P. Schmitter, H. L. Fröhlich, G. Dercon et al., "Redistribution of carbon and nitrogen through irrigation in intensively cultivated tropical mountainous watersheds," *Biogeochemistry*, vol. 109, no. 1-3, pp. 133–150, 2012.

[6] Y. Song, Q. Li, D. Feng, J. J. Zou, and W. Cai, "Texture image classification with discriminative neural networks," *Computational Visual Media*, vol. 2, pp. 367–377, 2016.

[7] M. A. Nearing, G. Govers, and L. D. Norton, "Variability in soil erosion data from replicated plots," *Soil Science Society of America Journal*, vol. 63, no. 6, pp. 1829–1835, 1999.

[8] W. A. Bagwan and R. S. Gavali, "Delineating changes in soil erosion risk zones using RUSLE model based on confusion matrix for the Urmodi River Watershed, Maharashtra, India," *Modeling Earth Systems and Environment*, 2020.

[9] K. G. Renard and J. R. Freimund, "Using monthly precipitation data to estimate the R-factor in the revised USLE," *Journal of Hydrology*, vol. 157, no. 1-4, pp. 287–306, 1994.

[10] Z. Kliment, J. Kadlec, and J. Langhammer, "Evaluation of suspended load changes using Ann AGNPS and SWAT semi-empirical erosion models," *Catena*, vol. 73, no. 3, pp. 286–299, 2008.

[11] J. M. Laflen, L. J. Lane, and G. R. Foster, "WEPP: a new generation of erosion prediction technology," *Journal of Soil and Water Conservation*, vol. 46, pp. 34–38, 1991.

[12] I. D. U. H. Piyathilake, R. G. I. Sumudumali, E. P. N. Udayakumara, L. V. Ranaweera, J. M. C. K. Jayawardana, and S. K. Gunatilake, "Modeling predictive assessment of soil erosion related hazards at the Uva Province in Sri Lanka," *Modeling Earth Systems and Environment*, 2020.

[13] M. Abedini, B. Ghasemian, A. Shirzadi et al., "A novel hybrid approach of bayesian logistic regression and its ensembles for landslide susceptibility assessment," *Geocarto International*, vol. 34, no. 13, pp. 1427–1457, 2018.

[14] R. Mohanty, S. Suman, and S. K. Das, "Prediction of vertical pile capacity of driven pile in cohesionless soil using artificial intelligence techniques," *International Journal of Geotechnical Engineering*, vol. 12, no. 2, pp. 209–216, 2018.

[15] Z. Mosaffaei and A. Jahani, "Modeling of ash (Fraxinus Excelsior) bark thickness in urban forests using Artificial Neural Network (ANN) and regression models," *Modeling Earth Systems and Environment*, 2020.

[16] M. A. Shahin, "State-of-the-art review of some artificial intelligence applications in pile foundations," *Geoscience Frontiers*, vol. 7, no. 1, pp. 33–44, 2016.

[17] M. Kim and J. E. Gilley, "Artificial Neural Network estimation of soil erosion and nutrient concentrations in runoff from land application areas," *Computers and Electronics in Agriculture*, vol. 64, no. 2, pp. 268–275, 2008.

[18] P. Licznar and M. A. Nearing, "Artificial neural networks of soil erosion and runoff prediction at the plot scale," *Catena*, vol. 51, no. 2, pp. 89–114, 2003.

[19] M. F. Yusof, H. M. Azamathulla, and R. Abdullah, "Prediction of soil erodibility factor for Peninsular Malaysia soil series using ANN," *Neural Computing and Applications*, vol. 24, no. 2, pp. 383–389, 2014.

[20] V. D. Tuan, X.-L. Tran, M.-T. Cao, T. C. Tran, and N.-D. Hoang, "Machine learning based soil erosion susceptibility prediction using social spider algorithm optimized multivariate adaptive regression spline," *Measurement*, vol. 164, p. 108066, 2020.

[21] T. V. Dinh, H. Nguyen, X.-L. Tran, and N.-D. Hoang, "Predicting rainfall-induced soil erosion based on a hybridization of adaptive differential evolution and support vector machine classification," *Mathematical Problems in Engineering*, vol. 2021, Article ID 6647829, 1 page, 2021.

[22] C. A. S. De Farias and C. A. G. Santos, "The use of Kohonen neural networks for runoff-erosion modeling," *Journal of Soils and Sediments*, vol. 14, no. 7, pp. 1242–1250, 2014.

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

[24] A. Arabameri, S. Chandra Pal, R. Costache et al., "Prediction of gully erosion susceptibility mapping using novel ensemble machine learning algorithms," *Geomatics, Natural Hazards and Risk*, vol. 12, no. 1, pp. 469–498, 2021.

[25] C. Conoscenti, V. Agnesi, M. Cama, N. A. Caraballo-Arias, and E. Rotigliano, "Assessment of gully erosion susceptibility using multivariate adaptive regression splines and accounting for terrain connectivity," *Land Degradation & Development*, vol. 29, no. 3, pp. 724–736, 2018.

[26] C. C. Aggarwal, *Neural Networks and Deep Learning*, Springer, Berlin, Germany, 2018.

[27] S. O. Haykin, *Neural Networks and Learning Machines*, Pearson, London, UK, 2008.

[28] V. N. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, Inc., Hoboken, NJ, USA, 1998.

[29] J. A. K. Suykens and J. Vandewalle, *Neural Processing Letters*, vol. 9, no. 3, pp. 293–300, 1999.

[30] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, vol. 1, pp. 211–244, 2001.

[31] K.-W. Liao, N.-D. Hoang, and S.-C. Chang, "Estimating landslide occurrence via small watershed method with relevance vector machine," *Earth Science Informatics*, vol. 13, no. 2, pp. 249–260, 2020.

[32] D. T. Bui, H. Shahabi, A. Shirzadi et al., "A novel integrated approach of relevance vector machine optimized by imperialist competitive algorithm for spatial modeling of shallow landslides," *Remote Sensing*, vol. 10, no. 10, p. 1538, 2018.

[33] J. M. Keller, M. R. Gray, and J. A. Givens, "A fuzzy K-nearest neighbor algorithm," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-15, no. 4, pp. 580–585, 1985.

[34] V. D. Tuan, T. Hilger, L. MacDonald et al., "Mitigation potential of soil conservation in maize cropping on steep slopes," *Field Crops Research*, vol. 156, pp. 91–102, 2014.

[35] K. C. McGregor, R. L. Bingner, A. J. Bowie, and G. R. Foster, "Erosivity index values for Northern Mississippi," *Transactions of the ASAE*, vol. 38, no. 4, pp. 1039–1047, 1995.

[36] C. Valentin, F. Agus, R. Alamban et al., "Runoff and sediment losses from 27 upland catchments in Southeast Asia: impact of rapid land use changes and conservation practices," *Agriculture, Ecosystems & Environment*, vol. 128, no. 4, pp. 225–238, 2008.

[37] F. Abdollahi, S. Hosseini, M. Sabet, S. H. Esmaeili-Faraj, and F. Amiri, "A novel study of the gas lift process using an integrated production/injection system using artificial neural network approach," *Modeling Earth Systems and Environment*, 2020.

[38] M. T. Hagan, H. B. Demuth, M. H. Beale, and O. D. Jesús, *Neural Network Design*, Paperback, Oklahoma City, OK, USA, 2nd edition, 2014.

[39] H. Moayedi, D. Tien Bui, A. Dounis, L. Kok Foong, and B. Kalantar, "Novel nature-inspired hybrids of neural computing for estimating soil shear strength," *Applied Sciences*, vol. 9, no. 21, p. 4643, 2019.

[40] J. I. Mwaura and B. K. Kenduiywo, "County level maize yield estimation using artificial neural network," *Modeling Earth Systems and Environment*, 2020.

[41] T.-H. Tran and N.-D. Hoang, "Predicting colonization growth of algae on mortar surface with artificial neural network," *Journal of Computing in Civil Engineering*, vol. 30, no. 6, 2016.

[42] M. T. Hagan and M. B. Menhaj, "Training feedforward networks with the Marquardt algorithm," *IEEE Transactions on Neural Networks*, vol. 5, no. 6, pp. 989–993, 1994.

[43] Ł. Sadowski and J. Hoła, "ANN modeling of pull-off adhesion of concrete layers," *Advances in Engineering Software*, vol. 89, pp. 17–27, 2015.

[44] G. Hinton, N. Srivastava, and K. Swersky, "Neural networks for machine learning lecture notes," 2012, https://wwwcstorontoedu/~hinton/coursera/lecture4/lec4pdf>.

[45] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[46] M.-Y. Cheng and N.-D. Hoang, "Typhoon-induced slope collapse assessment using a novel bee colony optimized support vector classifier," *Natural Hazards*, vol. 78, no. 3, pp. 1961–1978, 2015.

[47] R. Gupta, M. A. Alam, and P. Agarwal, "Modified support vector machine for detecting stress level using EEG signals," *Computational Intelligence and Neuroscience*, vol. 2020, Article ID 8860841, 2020.

[48] G. M. Hadjidemetriou, P. A. Vela, and S. E. Christodoulou, "Automated pavement patch detection and quantification using support vector machines," *Journal of Computing in Civil Engineering*, vol. 32, no. 1, 2018.

[49] J.-S. Chou, K.-H. Yang, and J.-Y. Lin, "Peak shear strength of discrete fiber-reinforced soils computed by machine learning and metaensemble Methods," *Journal of Computing in Civil Engineering*, vol. 30, no. 6, 2016.

[50] Q. He, Q. Zhang, H. Wang, and C. Zhang, "Local similarity-based fuzzy multiple kernel one-class support vector machine," *Complexity*, vol. 2020, Article ID 8853277, 2020.

[51] D. T. Bui, A. T. Tran, H. Klempe, B. Pradhan, and I. Revhaug, "Spatial prediction models for shallow landslide hazards: a comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree," *Landslides*, vol. 13, pp. 361–378, 2016.

[52] L. H. Hamel, *Knowledge Discovery with Support Vector Machines*, John Wiley & Sons, Inc., Hoboken, NJ, USA, 2009.

[53] M.-Y. Cheng, D. Prayogo, and Y.-W. Wu, "Prediction of permanent deformation in asphalt pavements using a novel symbiotic organisms search–least squares support vector regression," *Neural Computing and Applications*, vol. 31, pp. 6261–6273, 2019.

[54] N.-D. Hoang and D. Tien-Bui, "A novel relevance vector machine classifier with cuckoo search optimization for spatial prediction of landslides," *Journal of Computing in Civil Engineering*, vol. 30, no. 5, 2016.

[55] H. Han, X. Cui, Y. Fan, and H. Qing, "Least squares support vector machine (LS-SVM)-based chiller fault diagnosis using

fault indicative features," *Applied Thermal Engineering*, vol. 154, pp. 540–547, 2019.

[56] Y.-H. Wu and H. Shen, "Grey-related least squares support vector machine optimization model and its application in predicting natural gas consumption demand," *Journal of Computational and Applied Mathematics*, vol. 338, pp. 212–220, 2018.

[57] P.-P. Xi, Y.-P. Zhao, P.-X. Wang, Z.-Q. Li, Y.-T. Pan, and F.-Q. Song, "Least squares support vector machine for class imbalance learning and their applications to fault detection of aircraft engine," *Aerospace Science and Technology*, vol. 84, pp. 56–74, 2019.

[58] J. Suykens, J. V. Gestel, J. D. Brabanter, B. D. Moor, and J. Vandewalle, *Least Square Support Vector Machines*, World Scientific Publishing Co Pte Ltd, Singapore, 2002.

[59] M. E. Tipping, "The relevance vector machine," *Advances in Neural Information Processing Systems*, vol. 12, pp. 652–658, 2000.

[60] N.-D. Hoang and D. T. Bui, "Predicting earthquake-induced soil liquefaction based on a hybridization of kernel Fisher discriminant analysis and a least squares support vector machine: a multi-dataset study," *Bulletin of Engineering Geology and the Environment*, vol. 77, no. 1, pp. 191–204, 2018.

[61] H. Abbas and J. Tezcan, "Relevance vector machines modeling of nonstationary ground motion coherency," *Soil Dynamics and Earthquake Engineering*, vol. 120, pp. 262–272, 2019.

[62] A. K. Samantaray, G. Singh, and M. Ramadas, "Application of the relevance vector machine to drought monitoring," in *Soft Computing for Problem Solving*, pp. 891–898, Springer, Singapore, 2019.

[63] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[64] M.-Y. Cheng and N.-D. Hoang, "Evaluating contractor financial status using a hybrid fuzzy instance based classifier: case study in the construction industry," *IEEE Transactions on Engineering Management*, vol. 62, no. 2, pp. 184–192, 2015.

[65] R. C. Prati, G. E. A. P. A. Batista, and D. F. Silva, "Class imbalance revisited: a new experimental setup to assess the performance of treatment methods," *Knowledge and Information Systems*, vol. 45, no. 1, pp. 247–270, 2015.

[66] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, "An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics," *Information Sciences*, vol. 250, pp. 113–141, 2013.

[67] Mathworks, *Statistics and Machine Learning Toolbox*, The MathWorks, Inc, Bengaluru, Karnataka, India, 2016.

[68] K. De Brabanter, P. Karsmakers, F. Ojeda, and C. Alzate, *LS-SVMlab Toolbox User's Guide Version 1.8 Internal Report 10-146*, ESAT-SISTA, KULeuven (Leuven, Belgium), 2010.

[69] M. E. Tipping, "Sparse Bayesian models (and the RVM)," 2009, http://wwwmiketippingcom/sparsebayeshtm.

[70] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[71] S. Matsumoto, S. Ogata, H. Shimada, T. Sasaoka, A. Hamanaka, and G. Kusuma, "Effects of pH-induced changes in soil physical characteristics on the development of soil water erosion," *Geosciences*, vol. 8, no. 4, p. 134, 2018.

[72] S. Sidney, *Non-parametric Statistics for the Behavioral Sciences*, McGraw-Hill, New York, NY, USA,, 1988.

[73] W. S. Merritt, R. A. Letcher, and A. J. Jakeman, "A review of erosion and sediment transport models," *Environmental Modelling & Software*, vol. 18, no. 8-9, pp. 761–799, 2003.

[74] C. Huang, L. K. Wells, and L. D. Norton, "Sediment transport capacity and erosion processes: model concepts and reality," *Earth Surface Processes and Landforms*, vol. 24, no. 6, pp. 503–516, 1999.

[75] I. Albaradeyia, A. Hani, and I. Shahrour, "WEPP and ANN models for simulating soil loss and runoff in a semi-arid Mediterranean region," *Environmental Monitoring and Assessment*, vol. 180, no. 1-4, pp. 537–556, 2011.