International Journal of Geo-Information

MDPI

Article

Comparison of Ensemble Machine Learning Methods for Soil Erosion Pin Measurements

Kieu Anh Nguyen [1], Walter Chen [1,*], Bor-Shiun Lin [2] and Uma Seeboonruang [3,*]

[1] Department of Civil Engineering, National Taipei University of Technology, Taipei 10608, Taiwan; t106429401@ntut.edu.tw
[2] Disaster Prevention Technology Research Center, Sinotech Engineering Consultants, Taipei 11494, Taiwan; bosch.lin@sinotech.org.tw
[3] Faculty of Engineering, King Mongkut's Institute of Technology Ladkrabang, Bangkok 10520, Thailand
* Correspondence: waltchen@ntut.edu.tw (W.C.); uma.se@kmitl.ac.th (U.S.); Tel.: +886-2-27712171 (ext. 2628) (W.C.); +66-2329-8334 (U.S.)

Abstract: Although machine learning has been extensively used in various fields, it has only recently been applied to soil erosion pin modeling. To improve upon previous methods of quantifying soil erosion based on erosion pin measurements, this study explored the possible application of ensemble machine learning algorithms to the Shihmen Reservoir watershed in northern Taiwan. Three categories of ensemble methods were considered in this study: (a) Bagging, (b) boosting, and (c) stacking. The bagging method in this study refers to bagged multivariate adaptive regression splines (bagged MARS) and random forest (RF), and the boosting method includes Cubist and gradient boosting machine (GBM). Finally, the stacking method is an ensemble method that uses a meta-model to combine the predictions of base models. This study used RF and GBM as the meta-models, decision tree, linear regression, artificial neural network, and support vector machine as the base models. The dataset used in this study was sampled using stratified random sampling to achieve a 70/30 split for the training and test data, and the process was repeated three times. The performance of six ensemble methods in three categories was analyzed based on the average of three attempts. It was found that GBM performed the best among the ensemble models with the lowest root-mean-square error (RMSE = 1.72 mm/year), the highest Nash-Sutcliffe efficiency (NSE = 0.54), and the highest index of agreement (d = 0.81). This result was confirmed by the spatial comparison of the absolute differences (errors) between model predictions and observations using GBM and RF in the study area. In summary, the results show that as a group, the bagging method and the boosting method performed equally well, and the stacking method was third for the erosion pin dataset considered in this study.

Keywords: soil erosion; erosion pin; ensemble machine learning; Shihmen Reservoir watershed; bagging; boosting; stacking

Citation: Nguyen, K.A.; Chen, W.; Lin, B.-S.; Seeboonruang, U. Comparison of Ensemble Machine Learning Methods for Soil Erosion Pin Measurements. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 42. https://doi.org/10.3390/ijgi10010042

Received: 6 December 2020
Accepted: 16 January 2021
Published: 19 January 2021

1. Introduction

Soil erosion is a severe global issue affecting farming and the environment in tropical and subtropical areas. In particular, soil erosion leads to environmental damage such as soil nutrient loss, pollution by sedimentation, and the increased possibility of flooding. The rate of soil erosion depends on soil characteristics, climate, slope steepness [1], land use, and protective vegetation [2]. In addition, eroded soils lose 75% to 80% of carbon content [3], which results in a deficit of terrestrial carbon budget. Anthropogenic activity is a major cause of soil erosion [4]. Soil degradation rapidly has intensified with the rising population in the 20th century [5] and beyond.

Many theoretical/empirical models can be used to study soil erosion. According to Borrelli et al. [6], 435 distinct models and model variants were used to analyze soil erosion from 1994 to 2017 in 1697 scientific articles in the Scopus database. The top five

most applied models are RUSLE (Revised Universal Soil Loss Equation), USLE (Universal Soil Loss Equation), WEPP (Water Erosion Prediction Project), SWAT (Soil and Water Assessment Tool), and WaTEM/SEDEM (Water and Tillage Erosion Model and Sediment Delivery Model). More than half of these studies (58% of 1697 articles) included validation of the research results in the form of expert knowledge, measured sediment yield, measured erosion rates, or comparison with other models.

Taiwan has numerous natural hazards, and typhoons and earthquakes occur every year. The average annual precipitation is 2500 mm, which is concentrated from May to September [7]. For this reason, a large amount of soil is eroded and transported away by flowing water. In recent years, many studies have been conducted to model soil erosion in Taiwan. Traditionally, a physically or empirically based soil erosion model is needed for such studies. For example, Fan and Wu [8] developed equations to evaluate the relationship between slope steepness, soil properties, rainfall intensity, and interrill soil erosion rate in Taiwan. Lo [9] used the agricultural nonpoint source pollution model (AGNPS) to quantify soil erosion in the Bajun River basin and Tsengwen Reservoir watershed. Alternatively, Chiu et al. [10] measured the 137Cs concentrations at 60 sampling sites in the Shihmen Reservoir watershed to estimate soil erosion. Chen et al. [11] applied the USLE model to the same watershed to study the effect of the digital elevation model (DEM) on soil erosion. Finally, Liu et al. [12] followed up by applying slope units to soil erosion modeling in the same watershed.

In contrast to the traditional soil erosion modeling approach, in which a physically or empirically based model is needed to make a prediction and field data are collected to verify the model's correctness, a machine learning (ML) based approach does not require an a priori model. The field measurements (such as those of erosion pins) are used directly to formulate rules and make generalizations from the data (i.e., predictions). Although ML-based approaches have been extensively used in relevant fields such as landslides susceptibility mapping [13,14], soil thickness prediction [15], digital soil mapping [16], and biomass retrieval [17], it has only recently been applied to soil erosion pin study [18,19]. To improve upon previous methods of quantifying soil erosion, the study explored the possible application of ensemble machine learning algorithms to the Shihmen Reservoir watershed in northern Taiwan. Three categories of ensemble methods were considered in this study: (a) Bagging, (b) boosting, and (c) stacking. The results were compared among the ensemble methods in terms of three statistical indices: (1) Root-mean-square error (RMSE), (2) Nash-Sutcliffe efficiency (NSE), and (3) index of agreement (d).

2. Methods

An erosion pin is a wooden or metal rod inserted into the ground for measuring the change of the ground surface. The pin referred to in this study is made of metal with a diameter of 15.9 mm and a length of 300 mm. About 270 mm of the pin is embedded in the ground with the exposed part painted red. The erosion pin is one of the simplest and most effective methods for monitoring ground surface variation due to soil erosion and sediment deposition [20]. It has been used to monitor sheet erosion, gully erosion, landslides, and stream bank erosion [21–23]. Lin et al. [24] has documented the procedures for installing erosion pins in Taiwan. Measurements of erosion pins have been collected from various watersheds. The Shihmen Reservoir watershed data show that the average soil erosion was 90.6 t per hectare per year [12].

Ensemble machine learning is a technique that combines several base ML models (either homogeneous or heterogeneous) to make better predictions. Note that the word "model" has a different meaning in machine learning than for soil erosion. As explained earlier, a machine learning (ML) based approach does not require an a priori (soil erosion) model to work. In addition, ensemble methods gain performance over individual ML models [25]. Ensemble methods have been applied to many diverse fields such as banking [26], big data security [27], and breast cancer diagnosis [28]. In the environment and related fields, Pham et al. [29] applied several ensemble methods (AdaBoost, Bagging, Dagging,

MultiBoost, Rotation Forest, and Random Subspace) to evaluate the landslide susceptibility in the Himalayan India. The results showed that the area under the curve (AUC) of the receiver operating characteristic (ROC) curve were all higher than 0.876. Similarly, Tehrany et al. [30] used an ensemble support vector machine (SVM) and the weights-of-evidence method to conduct the flood susceptibility mapping in Terengganu of Malaysia. Their results improved flood modeling by 29%.

An ensemble method is viewed as a compound model. The purpose of such a model is to achieve better predictive performance by reducing the noise or error between observed and predicted data. Ensemble methods are usually grouped into bootstrap aggregating (bagging), boosting, and stacking methods. All three categories attempt to tune their predictions to the observations by decreasing model variance, bias, or both simultaneously. The main difference is that bagging and boosting usually work with homogeneous models, whereas stacking excels in combining heterogeneous models. Their respective approaches are described in the following sub-sections. In this study, we will select two methods from each category for analysis and comparison.

2.1. Bagging

Bagging is a technique that builds multiple homogeneous models from different subsamples of the same training dataset to obtain more accurate predictions than its individual models. It is an application of the bootstrap procedure to high-variance machine learning problems. For example, random forest (RF) is the bagging of decision trees (DT). Using CART (classification and regression trees) as an example, RF randomly samples the training dataset multiple times (with replacement) to obtain many subsamples. Then, a decision tree is built for each subsample using CART. Finally, RF issues its prediction by combing the results of all decision trees either by voting (classification) or averaging (regression). RF is a very effective machine learning tool, and it has been applied to various research problems, including soil erosion [18]. Therefore, we chose RF and bagged MARS (multivariate adaptive regression splines) in this study to compare with other ensemble algorithms.

MARS was first introduced by Friedman and Roosen [31]. It explores the relationship between the dependent and independent variables in a way very similar to least-squares regression [32]. The advantages of MARS include its computational efficiency, its ability to yield easy to interpret models, and its function to quantify the contribution of predictor variables. However, its lack of accurate prediction is one of the most significant drawbacks [32]. To address this issue, bagging was introduced to MARS to become bagged MARS to improve the classification accuracy. The bagged MARS model was implemented by the "earth" package in R, and the RF model was implemented by the "randomForest" package.

2.2. Boosting

Boosting refers to a group of algorithms that utilize weighted averages to make weak learning algorithms stronger learning algorithms. Unlike bagging that relies on each model running independently and then aggregated at the end, boosting runs sequentially by using later models to fix the prediction errors of the predecessor models in the sequence. For this study, we have selected Cubist and gradient boosting machine (GBM) for comparison with other ensemble methods.

Cubist is a prediction-oriented regression model proposed by Quinlan [33,34]. The general idea of the Cubist regression model is briefly described here. During the tree growth stage, many branches and leaves are grown. Linear regression models are added to the leaves of the tree. The Cubist method creates a series of "if-then" rules. Each rule has an associated multivariate linear model [35]. The corresponding model is used to calculate the predicted value if a set of variables satisfies the conditions of the rule. Rules are eliminated via pruning and/or combined for simplification. The main advantage of the Cubist method

is the addition of multiple training committees to balance case weights. In this study, the Cubist model was implemented by the "caret" and "Cubist" packages in R.

GBM was proposed by Friedman [36,37] as a simple and highly adaptive method for machine learning [38]. It is an improved boosting algorithm for regression and classification problems. The basic theory of GBM is to produce a prediction model constructed by a group of weak learning algorithms, typically decision trees. Each tree is grown sequentially by using the information from previously grown trees [39]. In this study, the GBM model was implemented by the "gbm" package in R.

2.3. Stacking

Stacking, sometimes called stacked generalization, is an ensemble machine learning method that combines multiple heterogeneous base or component models via a meta-model. The base model is trained on the complete training data, and then the meta-model is trained on the predictions of the base models. The advantage of stacking is the ability to explore the solution space with different models in the same problem. In this study, two stacking ensembles were chosen. They are (1) RF + DT + LM + Artificial Neural Networks (ANN) + SVM, and (2) GBM + DT + LM + ANN + SVM. For implementation, the "caretEnsemble" package in R was used.

2.4. Model Assessment

In order to evaluate the performance of ensemble models, three statistical indices were used as the evaluation criteria: (1) Root-mean-square error (RMSE), (2) Nash–Sutcliffe efficiency (NSE), and (3) index of agreement (d). These statistical indices have been frequently used in many studies [40,41].

RMSE is a good indicator for evaluating the model performance for continuous variables. In this study, RMSE represents the differences between erosion pin measurements and ensemble model predictions. It can be written as follows:

$$\mathrm{RMSE} = \sqrt{\frac{\sum(\mathrm{P}-\mathrm{O})^2}{\mathrm{n}}} \tag{1}$$

where P and O are the predicted and observed values, respectively.

NSE defines the relative magnitude between the "noise" and "information" [42]. Its value ranges from $-\infty$ to 1. The closer the value of NSE to 1, the more efficient the model is. In the case that NSE is negative, the model is considered poor because the observed mean serves as a better prediction than the model. NSE is defined as follows:

$$\mathrm{NSE} = 1 - \frac{\sum(\mathrm{P}-\mathrm{O})^2}{\sum(\mathrm{O}-\overline{\mathrm{O}})^2} \tag{2}$$

where $\overline{\mathrm{O}}$ is the mean observed value.

Finally, the index of agreement (d) is often used to represent the model performance [43]. Its value ranges from 0 to 1, with higher values indicating better agreement between the predictions and observations. The d index is defined as follows:

$$\mathrm{d} = 1 - \frac{\sum(\mathrm{P}-\mathrm{O})^2}{\sum\left(|\mathrm{P}-\overline{\mathrm{O}}| + |\mathrm{O}-\overline{\mathrm{O}}|\right)^2} \tag{3}$$

Note that we do not use R^2 (coefficient of determination) as a statistical index in the model evaluation. As pointed out by Nguyen et al. [18,19], R^2 evaluates the fit to the regression line only. High R^2 does not mean small differences between predictions and observations. This point is illustrated in Figure 1, in which a poor model has a perfect R^2 but only mediocre (even poor) values of other statistical indices. Good predictions should fall on the 45° line instead.

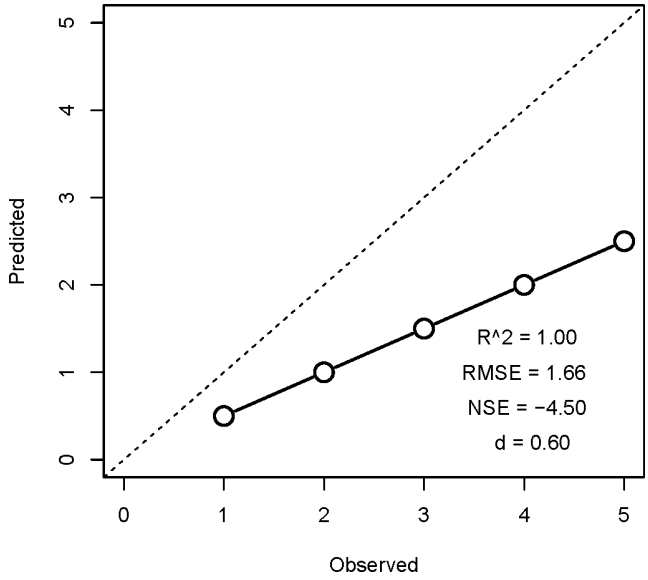


Figure 1. The comparison of R^2, root-mean-square error (RMSE), Nash–Sutcliffe efficiency (NSE), and d of a poor predictive model.

2.5. Study Area

In this study, the Shihmen Reservoir watershed in northern Taiwan was selected as the research area (Figure 2). Shihmen Reservoir watershed covers 76,340 ha of land area with a maximum elevation of 3527 m. The rainy season of the watershed coincides with the typhoon months. Therefore, heavy rainfalls are common [44]. In northern Taiwan, Shihmen Reservoir plays an essential role in providing drinking water for domestic use, irrigation for agriculture, and flood control for typhoon-related disasters [45]. Figure 2 also shows a photo of a metal erosion pin with the exposed part painted red and a picture of measurement being taken by a micrometer.

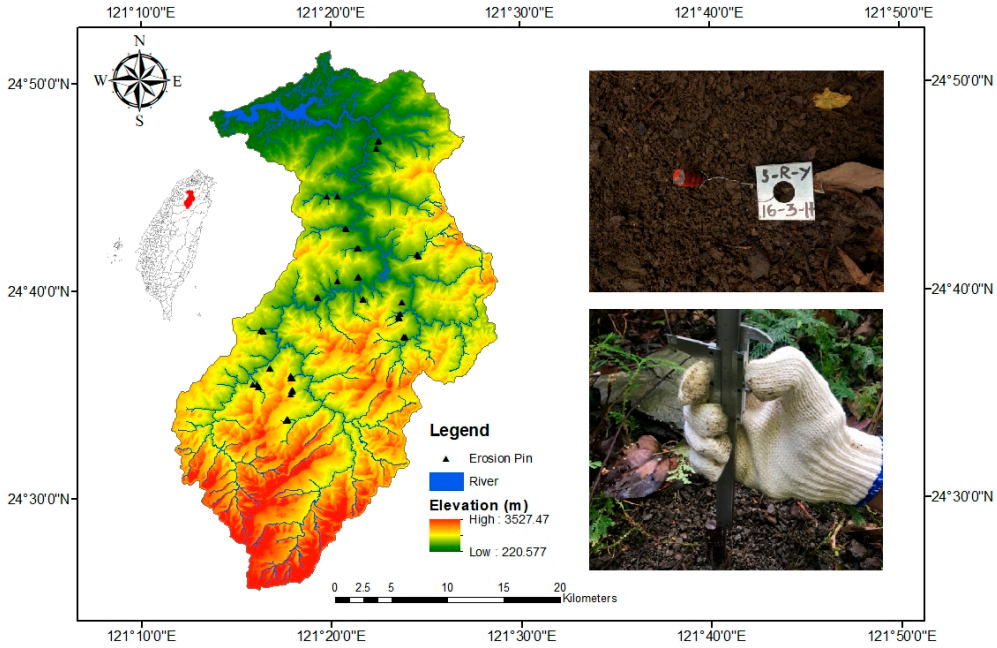


Figure 2. The study area and locations of erosion pins.

The workflow of this research is shown in Figure 3, and includes three parts: (1) Data collection, (2) data preparation, and (3) soil erosion pin analysis. During the first stage, the target variable and predictive variables were compiled into a dataset. Then, the dataset was

split into training and test data using a stratified random sampling method. Finally, the training data were fed to the ensemble methods to create prediction models. The models were tested using the test data, and the statistical indices were computed.

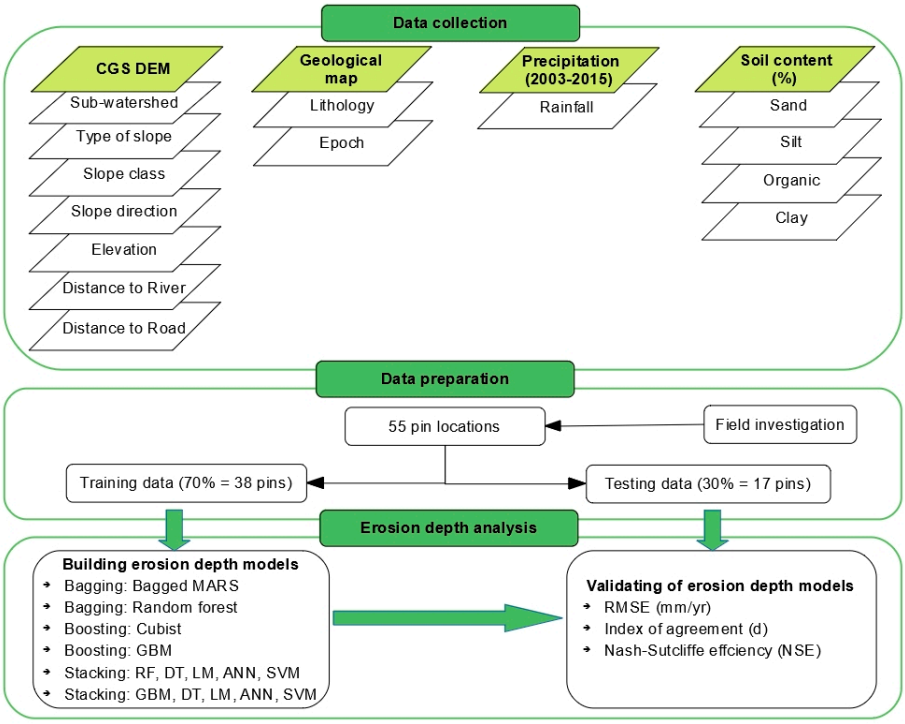


Figure 3. The soil erosion pin modeling flow chart using ensemble methods.

Fourteen attributes in four categories were used as the predictive variables, as shown in Figure 3. Five of the attributes were point data that were not available watershed-wide but only available at the erosion pins' locations. The dataset was separated into two groups using a 70/30 split, which is the common ratio used by many studies [46–48]. The entire process was repeated three times to determine the average result.

The target variable is the erosion pin measurement. A total of 550 pins were installed on 55 slopes (10 pins per slope) in the Shihmen Reservoir watershed (Figure 2). Lin et al. [24] documented the installation procedures. The erosion depth measurements were collected from 8 September 2008 to 10 October 2011. The measurements were averaged by slopes.

To derive topography-related attributes, such as elevation, sub-watershed, slope class, slope direction, distance to river, and distance to road, we used the Central Geological Survey (CGS) DEM created from an airborne LiDAR (light detection and ranging) survey, which has a spatial resolution of 10 m (Table 1). The DEM data were created in 2013 [11].

The average annual rainfall of the study area was calculated from 22 rainfall stations from 2003 to 2015. The slopes were classified into seven classes: (1) <5%, (2) 5–15%, (3) 15–30%, (4) 30–40%, (5) 40–55%, (6) 55–100%, and (7) >100%, based on the classification system of the Soil and Water Conservation Bureau. Slope direction is the aspect of a slope, which can be flat or facing north, northeast, east, southeast, south, southwest, west, or northwest. The distance to river was calculated based on the river network map. Similarly, the distance to road was found by the road network map with a scale of 1:5000. Both distances were calculated as the shortest distance between the individual erosion pin and the river or road network using ArcGIS 10.2 software.

Finally, lithology and epoch were from the geological maps of the Central Geological Survey. The scales were both 1:50,000. The soil contents were the percent of sand, silt, organic, and clay. The data were provided by Lin et al. [49].

Table 1. The predictive factors used in the modeling of soil erosion pin of the study area.

Type	Factor/Attribute	Source of Data	Scale or Resolution
CGS DEM	Sub-watershed	Central Geological Survey, Ministry of Economic Affairs of Taiwan	10 m × 10 m
	Type of slope		
	Slope class		
	Slope direction		
	Elevation		
	Distance to river		
	Distance to road		
Geological map	Lithology	Central Geological Survey	1:50,000
	Epoch		
Precipitation	Average annual rainfall	Northern Region Water Resources Office, Water Resources Agency, Ministry of Economic Affairs (2003–2015, 22 stations)	10 m × 10 m
Soil content	% sand	Lin et al. (2019)	10 m × 10 m
	% silt		
	% organic		
	% clay		

3. Results and Discussion

The results of three types of ensemble learning (that is, bagging, boosting, and stacking) and their comparisons are presented below.

3.1. Bagging

In the bagging category, we selected bagged MARS and random forest (also used by [18]) as the representative ensemble learning methods. Three repeated samplings (groupings) of the same dataset were made to create three different 70/30 splits using the stratified random sampling method. After fitting the model, the resulting statistical metrics, including the RMSE, NSE, and d values were calculated and these are listed in Table 2. Among the three indices of bagged MARS, RMSE ranges from 0.92 to 1.83 mm/year for the training data and from 1.70 to 2.18 mm/year for the test data. In addition, NSE ($-\infty$ to 1) varies from 0.38 to 0.83 for the training data and from 0.19 to 0.60 for the test data. Finally, d (0 to 1) goes from 0.64 to 0.94 for the training data and from 0.60 to 0.85 for the test data. As expected, all three indicators are better for the training data than for the test data with no exception. The same observation can be made on the RF metrics as well.

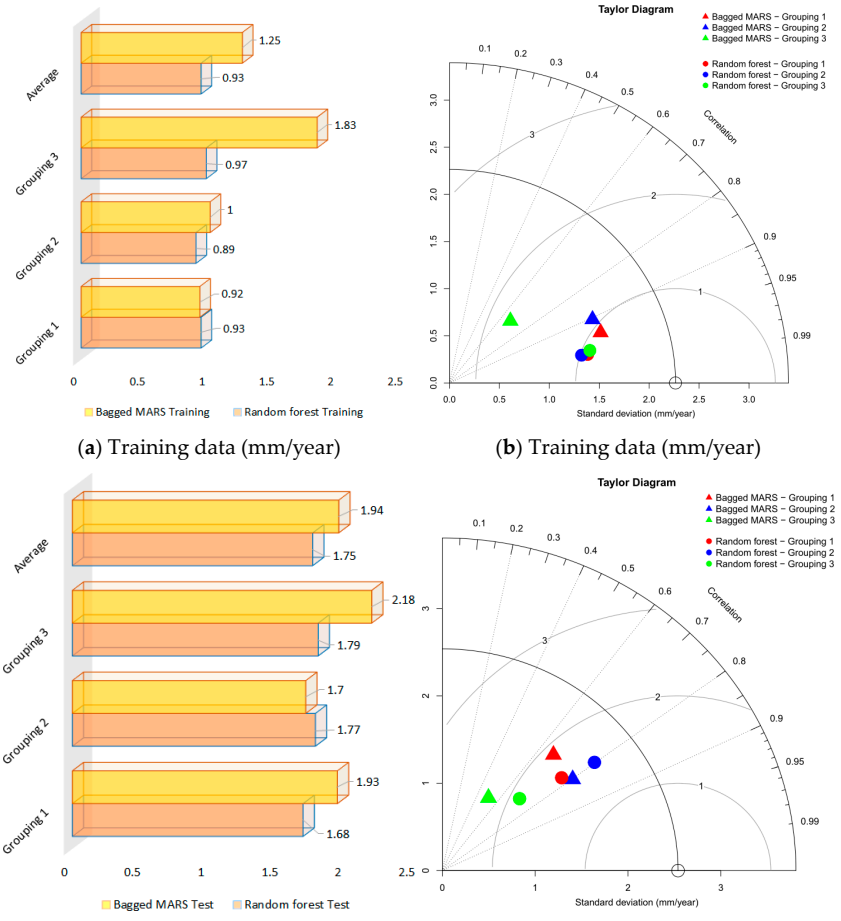
To evaluate the relative appropriateness of bagged MARS and RF as tools for ensemble learning, we compared the RMSE values of the two ensemble methods. The results were mixed, as shown in Figure 4. RF out-performed bagged MARS two out of three times in both the training data and the test data, as shown in Figure 4a,c and was superior overall. The same conclusion was also confirmed by the lower average RMSE values of RF than bagged MARS.

If we plot the results on a Taylor diagram, we can observe further differences between RF and bagged MARS. As can be seen from Figure 4b, all three sampling (grouping) results of the RF cluster together with very similar RMSE, correlation, and standard deviation. This indicates consistent training results across different samples. By contrast, only two of the sampling results of bagged MARS cluster together. The third is very far away with a substantially smaller correlation and standard deviation and much larger RMSE than the

other two. This shows that bagged MARS did not work equally well with different training data. In this case, it had difficulties fitting a model to grouping (sampling) #3. Hence, it lost to RF and fell short in the statistical comparison. Figure 4d shows a Taylor diagram comparison between the test datasets. A larger spread is observed in the data for both bagged MARS and RF. However, notice the closeness of data points of the same color. This shows that bagged MARS and RF generated similar predictions on the same test dataset. For instance, although the green triangle and green circle are very far apart in Figure 4b, they are relatively close to each other in Figure 4d.

Table 2. Statistical metrics of bagging ensemble models (bagged multivariate adaptive regression splines (MARS) and random forest (RF)).

Model	Statistical Index	Grouping 1		Grouping 2		Grouping 3		Average	
		Training	Test	Training	Test	Training	Test	Training	Test
bagged MARS	RMSE	0.92	1.93	1.00	1.70	1.83	2.18	1.25	1.94
	NSE	0.83	0.42	0.79	0.60	0.38	0.19	0.67	0.40
	d	0.94	0.78	0.93	0.85	0.64	0.60	0.84	0.74
RF	RMSE	0.93	1.68	0.89	1.77	0.97	1.79	0.93	1.75
	NSE	0.83	0.56	0.83	0.57	0.82	0.45	0.83	0.53
	d	0.94	0.82	0.93	0.85	0.93	0.72	0.93	0.80



(**a**) Training data (mm/year) (**b**) Training data (mm/year)

(**c**) Test data (mm/year) (**d**) Test data (mm/year)

Figure 4. The comparison of RMSE (mm/year), standard deviation, and correlation between bagged MARS and random forest.

3.2. Boosting

Two boosting ensemble methods were explored in this study. They were the GBM and Cubist models. Table 3 shows the performance metrics (RMSE, NSE, and d) under three different samplings (groupings) of the erosion pin data. For groupings #1 and #2, GBM won against Cubist in every category (lower RMSE, higher NSE, and higher d values). However, for grouping #3, the result is mixed. On the one hand, GBM lost to Cubist in all three indices concerning the training data. On the other hand, GBM beat Cubist in the test data by winning two of the three indices. The RMSE of GBM is lower than that of Cubist (1.71 vs. 1.98). Similarly, the NSE of GBM is higher than that of Cubist (0.50 vs. 0.33). However, GBM's d value is lower than Cubist's (0.75 vs. 0.78). This is one of the rare instances where NSE and d do not agree with each other. Putting everything together, we can conclude that GBM is still superior to Cubist.

Table 3. Statistical metrics of boosting ensemble models (gradient boosting machine (GBM) and Cubist).

Model	Statistical Index	Grouping 1		Grouping 2		Grouping 3		Average	
		Training	Test	Training	Test	Training	Test	Training	Test
GBM	RMSE	0.19	1.86	0.53	1.59	1.11	1.71	0.61	1.72
	NSE	0.99	0.46	0.94	0.65	0.77	0.50	0.90	0.54
	d	1.00	0.81	0.98	0.87	0.91	0.75	0.96	0.81
Cubist	RMSE	0.69	2.08	0.94	1.78	0.88	1.98	0.84	1.95
	NSE	0.91	0.33	0.81	0.56	0.86	0.33	0.86	0.41
	d	0.97	0.70	0.93	0.83	0.95	0.78	0.95	0.77

Figure 5 visually compares the average RMSE (mm/year) values between GBM and Cubist. For the training data, the RMSE of Cubist is 0.84 mm/year, inferior to the 0.61 mm/year of GBM. For the test data, a similar result is obtained. The RMSE of Cubist is 1.95 mm/year, not as good as the 1.72 mm/year of GBM. As a result, it can be confirmed that GBM is the better performing model in the category of boosting ensemble methods.

Figure 5. Comparison of the average RMSE (mm/year) values between Cubist and GBM.

3.3. Stacking

Stacking is the third and final category of ensemble methods examined in this study. The basic idea of stacking is to combine several weak models together to use their predictions as attributes in an overall meta-model. The meta-model is trained to yield better

predictions than the base models (component models). Using the same six ML models studied by Nguyen et al. [18,19], we created Table 4 and classified the ML models into four categories: Tree model, neural network model, hyperplane model, and linear regression model. The average RMSE values of these models are also shown in Table 4. Based on Table 4, we picked the four weakest models (one from each category) and used them as the base models in the stacking ensemble learning. They are DT, ANN, SVM, and LM. Furthermore, RF and GBM, one from bagging and the other from boosting, were chosen as the meta-models to form two sets of stacking models: (1) RF + DT + LM + ANN + SVM and (2) GBM + DT + LM + ANN + SVM. All of the statistical indices were computed in R in this study.

Table 4. The comparison of four types of machine learning (ML) models by the average RMSE (mm/year) values.

Type of Model	Model	Average RMSE (mm/year)	
		Training	Test
Tree model	Decision tree	1.73	2.45
	Random forest	0.93	1.75
Neural network model	ANN	1.23	2.36
	ANFIS	0.01	2.05
Hyperplane model	SVM	1.43	2.61
Linear regression model	LM	1.25	3.47

The results of RF-stacking (RF + DT + LM + ANN + SVM) and GBM-stacking (GBM + DT + LM + ANN + SVM) are shown in Table 5. For the training datasets, GBM-stacking outperformed RF-stacking in groupings #1 and #3 in all three statistical indices (RMSE, NSE, and d). In grouping #2, GBM-stacking also won against RF-stacking in terms of RMSE (1.45 vs. 1.46), but lost to RF-stacking in terms of d values (0.79 vs. 0.83). This is a rare case that an inconsistency between RMSE and d is observed.

Table 5. Statistical metrics of stacking ensemble models (RF + DT + LM + ANN + SVM and GBM + DT + LM + ANN + SVM).

Model	Statistical Index	Grouping 1		Grouping 2		Grouping 3		Average	
		Training	Test	Training	Test	Training	Test	Training	Test
RF + DT + LM + ANN + SVM	RMSE	1.38	2.04	1.46	2.48	1.70	1.94	1.51	2.15
	NSE	0.63	0.35	0.55	0.15	0.46	0.35	0.55	0.29
	d	0.87	0.74	0.83	0.65	0.72	0.69	0.81	0.69
GBM + DT + LM + ANN + SVM	RMSE	1.34	2.12	1.45	2.67	1.62	1.97	1.47	2.26
	NSE	0.65	0.30	0.55	0.02	0.51	0.33	0.57	0.22
	d	0.87	0.71	0.79	0.53	0.76	0.66	0.81	0.64

As for the test datasets, the best-performing model is reversed. In all three different groupings (samplings), RF-stacking out-performed GBM-stacking in all statistical indices with no exception. In summary, although GBM-stacking performed best with the training data, it lost the test data. Since the predictive performance of a model is based on unseen test data, we conclude that RF-stacking is the better stacking model of the two. The visual comparison between RF-stacking and GBM-stacking is shown in Figure 6, where all three indices were plotted (RMSE, NSE, and d). The numbers in the figure represent the average values of three different groupings (partitioning).

Figure 6. The comparison of RMSE (mm/year), NSE, and d between RF-stacking and GBM-stacking.

3.4. Comparison of Ensemble Models

So far, we have compared six ensemble methods in three categories and determined the best-performing model for each category. The next question is how they compare with one another in a six-way comparison. The results were compiled in Figure 7 using the RMSE values. Among the six ensemble models examined in this study, if we only consider the training data, the overall best-performing model is GBM (0.61 mm/year), followed by Cubist (0.84 mm/year). Both of them belong to the boosting category. Their average d values are as high as 0.96 and 0.95, respectively. However, if we only focus on the test data, although the overall winning model is still GBM (1.72 mm/year), RF (1.75 mm/year) will replace Cubist (1.95 mm/year) as the second best-performing model. If NSE and d are considered instead of RMSE, the conclusion does not change. In those cases, GBM and RF remain the two best ensemble models. Based on Figure 7, we rank the models from the best to the worst as follows:

1. Training: GBM > Cubist > RF > bagged MARS > Stacking (GBM) > Stacking (RF)
2. Test: GBM > RF > bagged MARS > Cubist > Stacking (RF) > Stacking (GBM)

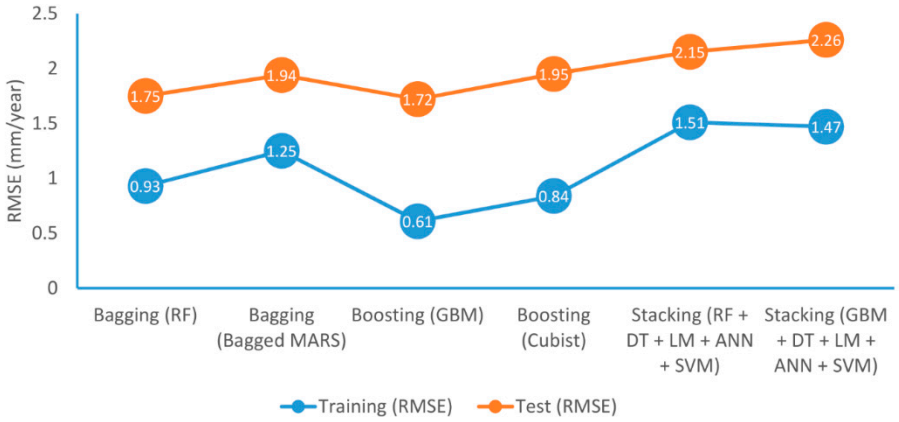


Figure 7. The comparison of ensemble models by the average RMSE (mm/year) values.

For machine learning, it is more appropriate to judge ML models' performance by the unseen test data. Therefore, for the test data, GBM (boosting) is the overall winner, followed by RF (bagging). The next best-performing models are bagged MARS (bagging) and Cubist (boosting). Clearly, the top four places were split evenly between the bagging method and the boosting method. Hence, these two types of ensemble models, bagging

and boosting, rival each other in performance and are equally good in predicting soil erosion depths measured by erosion pins. By contrast, the stacking ensemble methods (RF + DT + LM + ANN + SVM and GBM + DT + LM + ANN + SVM) seem to lag behind both the bagging and boosting methods. We found this result very intriguing. Despite GBM and RF being the top-two performing models, they do not perform as well when they are used in the stacking approach to combine other weaker models. Perhaps this is because GBM and RF worked with 14 attributes directly when used individually as single models. When GBM and RF were used as meta-models in the stacking approach, they were only trained on the predictions of the base models (DT, LM, ANN, and SVM). Not being able to work with the underlying 14 attributes directly seems to have undermined the ability of GBM and RF to make better predictions.

3.5. Model Predictions and Factor Importance

Because GBM and RF are the two best ensemble models in a class of their own (RMSE = 1.72 and 1.75 mm/year, respectively), we only display their respective ensemble learning results in Figures 8 and 9. As shown in the figures, the size and color of the circles represent the absolute differences (errors) between model predictions and observations ($|Obs - Pre|$). Proportional symbols were used. Therefore, the larger the circle, the bigger the difference. Similarly, the redder the dot color, the more significant the error. The contrasting results produced by GBM and RF are evident. For GBM, it is clear from Figure 8 that most points have low error except for those in the eastern part of the study area. By contrast, Figure 9 shows that RF has large errors in both the eastern and southern parts of the study area. To better visualize the error distribution, we further mapped the spatially interpolated values (absolute errors) throughout the study area. As shown in Figure 8, the resulting watershed is mostly green (associated with low error) for GBM, except in the eastern part, where large errors due to the steep topography and undesirable slope directions bring a return to red colors. However, according to Figure 9, the watershed is only in green in the northern region for RF. The rest of the watershed is colored red or yellow.

Furthermore, we created two vertical profiles on the map, one in an east-west direction and the other in a north-south direction, to compare the model predictions and observations for both GBM and RF. As shown in Figure 8, the black line is the observation, and the blue line is the GBM prediction. Both lines show a similar trend and move in the same direction: If the observation increases so does the prediction, and if the observation decreases so does the prediction. However, it seems evident that GBM tends to under-estimate the high values of observation and over-estimate the low values of observation in both the north-south and the east-west profiles. Furthermore, if we plot the same profiles in Figure 9 and use a red line to represent the RF prediction, we can see a similar result. The RF prediction also moves in the same direction as the observation. The RF model also underestimates the high values of observation and overestimates the low values of observation in the north–south and the east–west profiles. However, the difference is that RF errors more in comparison with GBM.

In summary, both GBM and RF perform better in the watershed's northern region, where the reservoir is located. Since the watershed slopes from south to north (as shown previously in Figure 2), it implies that the models more adequately captured the erosion behavior (measured by erosion pins) at the lower elevations than at the higher elevations. The eastern part is problematic. Neither of the two models works well here. In general, GBM outperformed RF because GBM matches the available measurements better than RF in the west and south. This makes GBM the overall best model, which is consistent with the RMSE-based results.

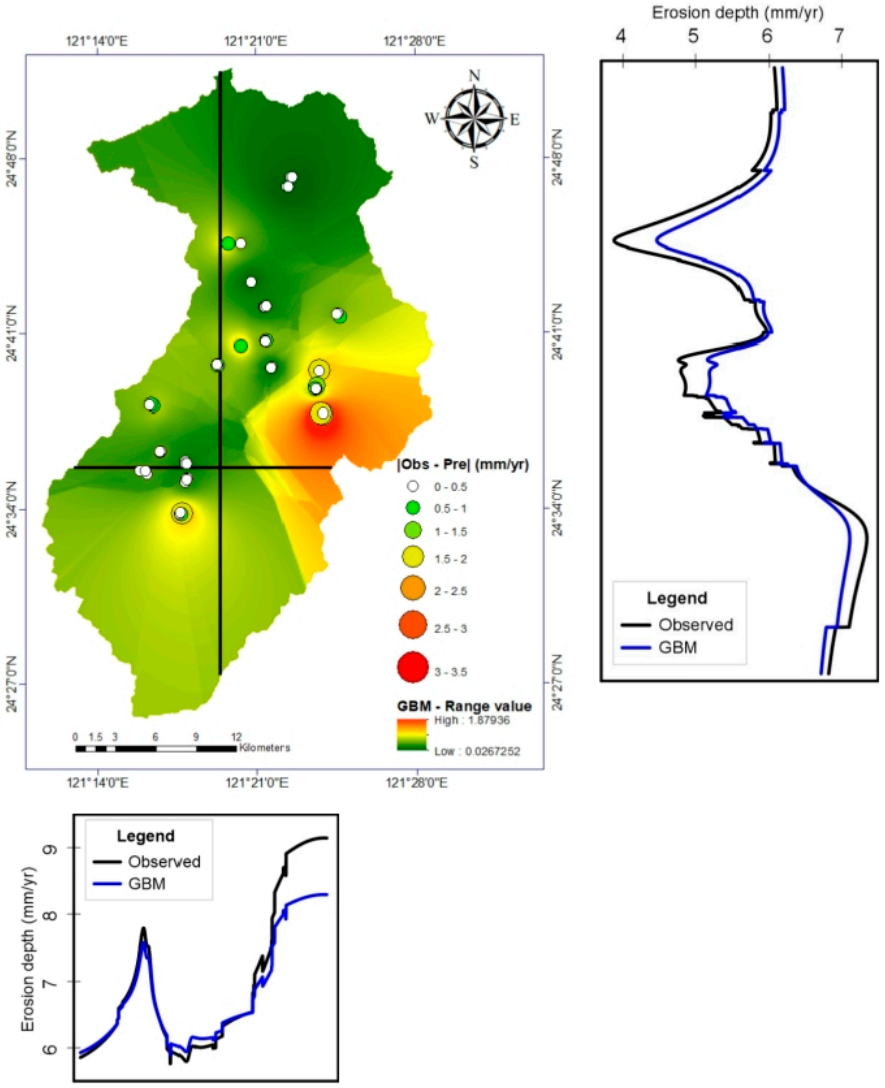


Figure 8. Map and profile of the study area showing the absolute difference between model prediction (GBM) and observation (mm/year).

Both GBM and RF generate a rank of importance for the 14 attributes used in this study. The six simulations (three of GBM and three of RF) are combined in a boxplot to illustrate the range of ranks (factor importance) as shown in Figure 10. Several points in this figure merit a closer look. First, the grey box in the figure shows the range of ranks between the first and third quartiles of the results. It is evident that each attribute has a variable range of ranks. Second, the black line in the box shows the median rank (1 being the most important and 14 being the least important). The median rank and the gray box can be used to compare the relative importance of attributes. Therefore, it can be seen from Figure 10 that A (slope direction) and B (type of slope) are the two overall most important attributes in the GBM and RF models. They consistently rank higher than the other 12 attributes. Attribute D (elevation) is interesting. It has the fourth-lowest median rank in the comparison. At the same time, it also has the longest grey box and the broadest range than any of the other attributes. This means that the elevation has changeable importance depending on the model and dataset. In some instances, the elevation is very important, whereas in others, it is not.

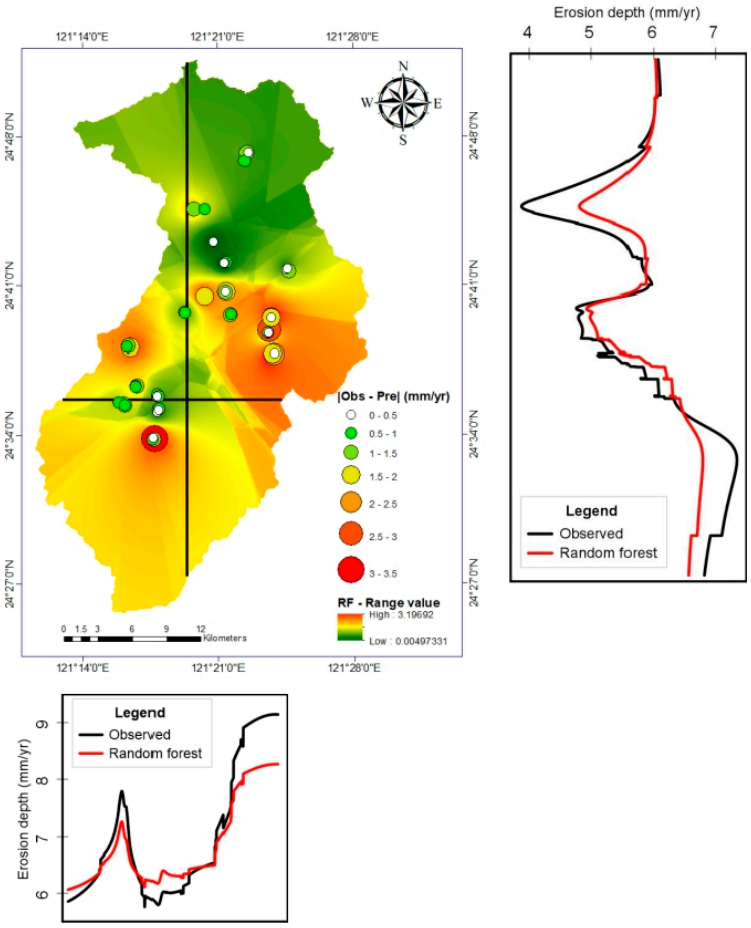


Figure 9. Map and profile of the study area showing the absolute difference between model prediction (RF) and observation (mm/year).

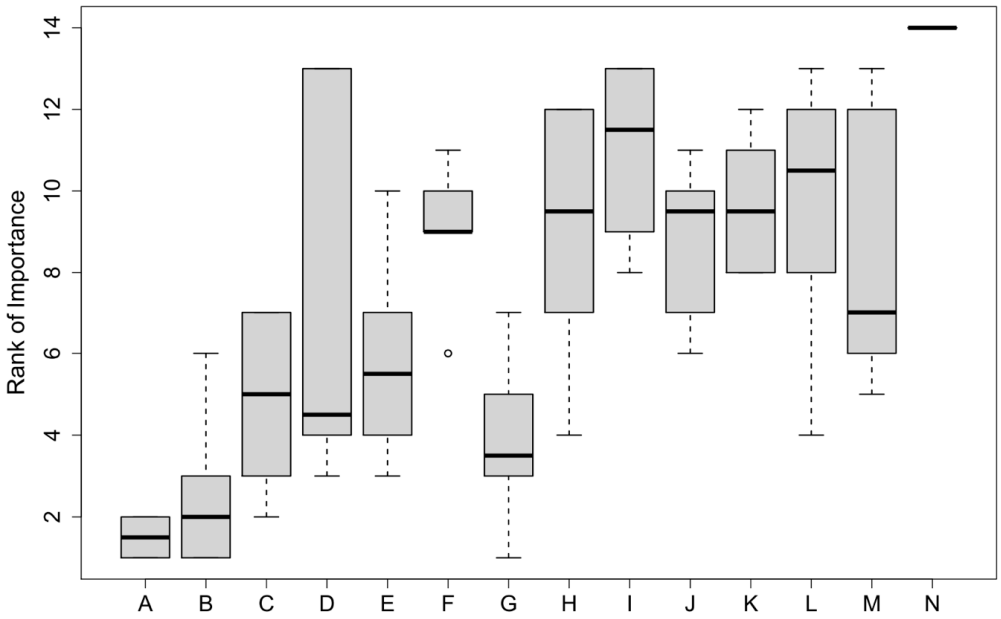


Figure 10. The relative importance of the 14 attributes used in this study. The lower the number, the higher the importance of the attribute. (A: Slope direction, B: Type of slope, C: % Organic, D: Elevation, E: Distance to road, F: % Sand, G: Subwatershed, H: % Clay, I: Slope class, J: Lithology, K: Distance to river, L: Average annual rainfall, M: % Silt, N: Epoch).

4. Conclusions

Soil erosion is a significant threat to the environment and the livelihood of the region, and should be taken seriously to mitigate the disastrous consequences. Hence, a precise spatial prediction of soil erosion is a critical need. In this study, we applied six ensemble learning methods (bagged MARS, RF, GBM, Cubist, RF-stacking, and GBM-stacking) of three categories (bagging, boosting, and stacking) to model the measurements of erosion pins in the Shihmen Reservoir watershed of Taiwan. The purpose was to improve the modeling accuracy and forecasting capability related to soil erosion. In the process of assessing the performance of the three types of ensemble approaches, we have learned how effective the ensemble methods are with unseen test data.

The study results show that the ensemble methods improve the prediction accuracy as measured by three statistical indices—RMSE, NSE, and d. Among the three categories of ensemble methods, bagging and boosting work equally well on the unseen test data. Stacking is the least favorable approach, with its RMSE trailing behind other types of ensemble algorithms. Individually speaking, we found GBM to be the best fitting model. Its RMSE, NSE, and d values are 1.72 mm/year, 0.54, and 0.81, respectively. The second-best model is RF. We used both GBM and RF to map the absolute differences (errors) between model predictions and observations ($|Obs - Pre|$) in the study area. The results show that both models perform well in the watershed's northern region (where the reservoir is located) and perform relatively poorly in the watershed's eastern part due to the steep topography and undesirable slope directions. What makes GBM superior to RF is that GBM also works well in the western and southern parts of the study area while RF does not. This conclusion is consistent with the RMSE-based results.

Finally, as an additional discovery in the study, we noticed two cases of inconsistent statistical indices during the model comparison. One of them happened when we compared GBM with Cubist (test data of grouping #3). The other occurred when we examined the performance of the two stacking models (training data of grouping #2). In both of these cases, RMSE and NSE favored one model, but d preferred the other. Therefore, potentially contradictory conclusions could be made if we only rely on one single index. This danger exemplifies the need to present multiple indices in such studies.

Author Contributions: Conceptualization, Walter Chen; data curation, Bor-Shiun Lin; formal analysis, Kieu Anh Nguyen and Walter Chen; funding acquisition, Walter Chen and Uma Seeboonruang; investigation, Kieu Anh Nguyen and Walter Chen; methodology, Walter Chen; project administration, Walter Chen and Uma Seeboonruang; resources, Walter Chen and Bor-Shiun Lin; software, Kieu Anh Nguyen; supervision, Walter Chen and Uma Seeboonruang; visualization, Kieu Anh Nguyen; writing—original draft, Kieu Anh Nguyen and Walter Chen; writing—review and editing, Walter Chen, Bor-Shiun Lin, and Uma Seeboonruang. All authors have read and agreed to the published version of the manuscript.

Funding: This study was partially supported by the National Taipei University of Technology-King Mongkut's Institute of Technology Ladkrabang Joint Research Program (Grant Numbers NTUT-KMITL-106-01, NTUT-KMITL-107-02, and NTUT-KMITL-108-01) and the Ministry of Science and Technology (Taiwan) Research Project (Grant Number MOST 109-2121-M-027-001).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ruiz-Sinoga, J.D.; Martínez-Murillo, J.F. Hydrological response of abandoned agricultural soils along a climatological gradient on metamorphic parent material in southern Spain. *Earth Surf. Process. Landf.* **2009**, *34*, 2047–2056. [CrossRef]
2. García-Ruiz, J.M. The effects of land uses on soil erosion in Spain: A review. *Catena* **2010**, *81*, 1–11. [CrossRef]
3. Morgan, R.P.C. *Soil Erosion and Conservation*; John Wiley & Sons: Hoboken, NJ, USA, 2009.

4. Islam, R.; Jaafar, W.Z.W.; Hin, L.S.; Osman, N.; Hossain, A.; Mohd, N.S. Development of an intelligent system based on ANFIS model for predicting soil erosion. *Environ. Earth Sci.* **2018**, *77*, 186. [CrossRef]
5. Lal, R. Soil degradation by erosion. *Land Degrad. Dev.* **2001**, *12*, 519–539. [CrossRef]
6. Borrelli, P.; Alewell, C.; Alvarez, P.; Anache, J.A.A.; Baartman, J.; Ballabio, C.; Bezak, N.; Biddoccu, M.; Cerdà, A.; Chalise, D.; et al. Soil erosion modelling: A global review and statistical analysis. *EarthArxiv* **2020**. [CrossRef]
7. Yeh, S.-C.; Wang, C.-A.; Yu, H.-C. Simulation of soil erosion and nutrient impact using an integrated system dynamics model in a watershed in Taiwan. *Environ. Model. Softw.* **2006**, *21*, 937–948. [CrossRef]
8. Fan, J.-C.; Wu, M.-F. Effects of soil strength, texture, slope steepness and rainfall intensity on interrill erosion of some soils in Taiwan. In Proceedings of the 10th International Soil Conservation Organization Meeting, Purdue University, USDA-ARS National Soil Erosion Research Laboratory, W. Lafayette, IN, USA, 24–29 May 1999.
9. Lo, K.F.A. Erosion assessment of large watersheds in Taiwan. *J. Soil Water Conserv.* **1995**, *50*, 180–183.
10. Chiu, Y.-J.; Chang, K.-T.; Chen, Y.-C.; Chao, J.-H.; Lee, H.-Y. Estimation of soil erosion rates in a subtropical mountain watershed using 137Cs radionuclide. *Nat. Hazards* **2011**, *59*, 271–284. [CrossRef]
11. Chen, W.; Li, D.-H.; Yang, K.-J.; Tsai, F.; Seeboonruang, U. Identifying and comparing relatively high soil erosion sites with four DEMs. *Ecol. Eng.* **2018**, *120*, 449–463. [CrossRef]
12. Liu, Y.-H.; Li, D.-H.; Chen, W.; Lin, B.-S.; Seeboonruang, U.; Tsai, F. Soil Erosion Modeling and Comparison Using Slope Units and Grid Cells in Shihmen Reservoir Watershed in Northern Taiwan. *Water* **2018**, *10*, 1387. [CrossRef]
13. Huang, Y.; Zhao, L. Review on landslide susceptibility mapping using support vector machines. *Catena* **2018**, *165*, 520–529. [CrossRef]
14. Reichenbach, P.; Rossi, M.; Malamud, B.D.; Mihir, M.; Guzzetti, F. A review of statistically-based landslide susceptibility models. *Earth-Sci. Rev.* **2018**, *180*, 60–91. [CrossRef]
15. Lagomarsino, D.; Tofani, V.; Segoni, S.; Catani, F.; Casagli, N. A Tool for Classification and Regression Using Random Forest Methodology: Applications to Landslide Susceptibility Mapping and Soil Thickness Modeling. *Environ. Model. Assess.* **2017**, *22*, 201–214. [CrossRef]
16. Heung, B.; Ho, H.C.; Zhang, J.; Knudby, A.; Bulmer, C.E.; Schmidt, M.G. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. *Geoderma* **2016**, *265*, 62–77. [CrossRef]
17. Ali, I.; Greifeneder, F.; Stamenkovic, J.; Neumann, M.; Notarnicola, C. Review of Machine Learning Approaches for Biomass and Soil Moisture Retrievals from Remote Sensing Data. *Remote Sens.* **2015**, *7*, 16398–16421. [CrossRef]
18. Nguyen, K.A.; Chen, W.; Lin, B.-S.; Seeboonruang, U.; Thomas, K. Predicting Sheet and Rill Erosion of Shihmen Reservoir Watershed in Taiwan Using Machine Learning. *Sustainability* **2019**, *11*, 3615. [CrossRef]
19. Nguyen, K.A.; Chen, W.; Lin, B.-S.; Seeboonruang, U. Using Machine Learning-Based Algorithms to Analyze Erosion Rates of a Watershed in Northern Taiwan. *Sustainability* **2020**, *12*, 2022. [CrossRef]
20. Haigh, M.J. The use of erosion pins in the study of slope evolution. *Br. Geomorphol. Res. Group Tech. Bull.* **1977**, *18*, 31–49.
21. Ghimire, S.K.; Higaki, D.; Bhattarai, T.P. Estimation of Soil Erosion Rates and Eroded Sediment in a Degraded Catchment of the Siwalik Hills, Nepal. *Land* **2013**, *2*, 370–391. [CrossRef]
22. Couper, P.; Stott, T.; Maddock, I. Insights into river bank erosion processes derived from analysis of negative erosion-pin recordings: Observations from three recent UK studies. *Earth Surf. Process. Landf. J. Br. Geomorphol. Res. Group* **2002**, *27*, 59–79. [CrossRef]
23. Lawler, D.M.; Couperthwaite, J.; Bull, L.J.; Harris, N.M. Bank erosion events and processes in the Upper Severn basin. *Hydrol. Earth Syst. Sci.* **1997**, *1*, 523–534. [CrossRef]
24. Lin, B.-S.; Thomas, K.; Chen, C.-K.; Ho, H.-C. Evaluation of soil erosion risk for watershed management in Shenmu watershed, central Taiwan using USLE model parameters. *Paddy Water Environ.* **2016**, *14*, 19–43. [CrossRef]
25. Dietterich, T.G. Ensemble methods in machine learning. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). In *International Workshop on Multiple Classifier Systems*; Springer: Berlin/Heidelberg, Germany, 2000; Volume 1857 LNCS, pp. 1–15.
26. Erdal, H.; Karahanoğlu, İ. Bagging ensemble models for bank profitability: An empirical research on Turkish development and investment banks. *Appl. Soft Comput.* **2016**, *49*, 861–867. [CrossRef]
27. Abawajy, J.; Kelarev, A.; Chowdhurry, M.U. Large Iterative Multitier Ensemble Classifiers for Security of Big Data. *IEEE Trans. Emerg. Top. Comput.* **2014**, *2*, 352–363. [CrossRef]
28. Hsieh, S.-L.; Hsieh, S.-H.; Cheng, P.-H.; Chen, C.-H.; Hsu, K.-P.; Lee, I.-S.; Wang, Z.; Lai, F. Design Ensemble Machine Learning Model for Breast Cancer Diagnosis. *J. Med. Syst.* **2012**, *36*, 2841–2847. [CrossRef]
29. Pham, B.T.; Bui, D.T.; Prakash, I.; Dholakia, M. Hybrid integration of Multilayer Perceptron Neural Networks and machine learning ensembles for landslide susceptibility assessment at Himalayan area (India) using GIS. *Catena* **2017**, *149*, 52–63. [CrossRef]
30. Tehrany, M.S.; Pradhan, B.; Jebur, M.N. Flood susceptibility mapping using a novel ensemble weights-of-evidence and support vector machine models in GIS. *J. Hydrol.* **2014**, *512*, 332–343. [CrossRef]
31. Friedman, J.H.; Roosen, C.B. An introduction to multivariate adaptive regression splines. *Stat. Methods Med. Res.* **1995**, *4*, 197–217. [CrossRef]
32. Otok, B.W.; Akbar, M.S.; Guritno, S.; Subanar, S. Ordinal Regression Model using Bootstrap Approach. *J. ILMU DASAR* **2007**, *8*, 54–67.

33. Quinlan, J.R. Learning with continuous classes. In Proceedings of the 5th Australian Joint Conference on Artificial Intelligence, Hobart, Australia, 16–18 November 1992.
34. Quinlan, J.R. Combining instance-based and model-based learning. In Proceedings of the 10th International Conference on Machine Learning, Amherst, MA, USA, 27–29 June 1993.
35. Zhou, J.; Li, E.; Wei, H.; Li, C.; Qiao, Q.; Armaghani, D.J. Random Forests and Cubist Algorithms for Predicting Shear Strengths of Rockfill Materials. *Appl. Sci.* **2019**, *9*, 1621. [CrossRef]
36. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
37. Friedman, J.H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* **2002**, *38*, 367–378. [CrossRef]
38. Zhou, J.; Li, E.; Yang, S.; Wang, M.; Shi, X.; Yao, S.; Mitri, H.S. Slope stability prediction for circular mode failure using gradient boosting machine approach based on an updated database of case histories. *Saf. Sci.* **2019**, *118*, 505–518. [CrossRef]
39. Ridgeway, G. Generalized Boosted Models: A guide to the GBM package. *Update* **2007**, 1–15.
40. Acharya, G.; Cochrane, T.; Davies, T.R.H.; Bowman, E. Quantifying and modeling post-failure sediment yields from laboratory-scale soil erosion and shallow landslide experiments with silty loess. *Geomorphology* **2011**, *129*, 49–58. [CrossRef]
41. Du, S.; Zhang, J.; Deng, Z.; Li, J. A New Approach of Geological Disasters Forecasting using Meteorological Factors based on Genetic Algorithm Optimized BP Neural Network. *Elektron. Elektrotech.* **2014**, *20*, 57–62. [CrossRef]
42. Nash, J.E.; Sutcliffe, J.V. River flow forecasting through conceptual models part I—A discussion of principles. *J. Hydrol.* **1970**, *10*, 282–290. [CrossRef]
43. Willmott, C.J. On the validation of models. *Phys. Geogr.* **1981**, *2*, 184–194. [CrossRef]
44. Chen, C.-S.; Chen, Y.-L. The Rainfall Characteristics of Taiwan. *Mon. Weather Rev.* **2003**, *131*, 1323–1341. [CrossRef]
45. Chang, F.-J.; Chang, Y.-T. Adaptive neuro-fuzzy inference system for prediction of water level in reservoir. *Adv. Water Resour.* **2006**, *29*, 1–10. [CrossRef]
46. Chen, W.; Panahi, M.; Pourghasemi, H.R. Performance evaluation of GIS-based new ensemble data mining techniques of adaptive neuro-fuzzy inference system (ANFIS) with genetic algorithm (GA), differential evolution (DE), and particle swarm optimization (PSO) for landslide spatial modelling. *Catena* **2017**, *157*, 310–324. [CrossRef]
47. Rogan, J.; Franklin, J.; Stow, D.; Miller, J.; Woodcock, C.; Roberts, D. Mapping land-cover modifications over large areas: A comparison of machine learning algorithms. *Remote Sens. Environ.* **2008**, *112*, 2272–2283. [CrossRef]
48. Ramos-Pollán, R.; Guevara-López, M.Á.; Oliveira, E. Introducing ROC curves as error measure functions: A new approach to train ANN-based biomedical data classifiers. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). In *Iberoamerican Congress on Pattern Recognition*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 517–524.
49. Lin, B.-S.; Chen, C.-K.; Thomas, K.; Hsu, C.-K.; Ho, H.-C. Improvement of the K-Factor of USLE and Soil Erosion Estimation in Shihmen Reservoir Watershed. *Sustainability* **2019**, *11*, 355. [CrossRef]