*sustainability*

MDPI

*Article*

# Predicting Sheet and Rill Erosion of Shihmen Reservoir Watershed in Taiwan Using Machine Learning

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

[1] Department of Civil Engineering, National Taipei University of Technology, Taipei 10608, Taiwan
[2] Disaster Prevention Technology Research Center, Sinotech Engineering Consultants, Taipei 11494, Taiwan
[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.)

Received: 7 May 2019; Accepted: 22 June 2019; Published: 1 July 2019


**Abstract:** Shihmen Reservoir watershed is vital to the water supply in Northern Taiwan but the reservoir has been heavily impacted by sedimentation and soil erosion since 1964. The purpose of this study was to explore the capability of machine learning algorithms, such as decision tree and random forest, to predict soil erosion (sheet and rill erosion) depths in the Shihmen reservoir watershed. The accuracy of the models was evaluated using the *RMSE* (Root Mean Squared Error), *MAE* (Mean Absolute Error), and $R^2$. Moreover, the models were verified against the multiple regression analysis, which is commonly used in statistical analysis. The predictors of these models were 14 environmental factors which influence soil erosion, whereas the target was 550 erosion pins installed at 55 locations (on 55 slopes) and monitored over a period of approximately three years. The data sets for the models were separated into 70% for the training data and 30% for the testing data, using the simple random sampling and stratified random sampling methods. The results show that the random forest algorithm performed the best of the three methods. Moreover, the stratified random sampling method had better results among the two sampling methods, as anticipated. The average error (*RMSE* relative to 1:1 line) of the stratified random sampling method of the random forest algorithm is 0.93 mm/yr in the training data and 1.75 mm/yr in the testing data, respectively. Finally, the random forest algorithm predicted that type of slope, slope direction, and sub-watershed are the three most important factors of the 14 environmental factors collected and used in this study for splits in the trees and thus they are the three most important factors affecting the depth of sheet and rill erosion in the Shihmen Reservoir watershed. The results of this study can be employed by decision-makers to improve soil conservation planning and watershed remediation.

**Keywords:** soil erosion; sheet and rill erosion; erosion pins; random forest; decision tree; multiple regression; Shihmen reservoir; machine learning

## 1. Introduction

Modern society is at a crucial stage in the growth of science and technology that is parallel to increases in environmental issues and climate change influenced events, such as floods, droughts and large-scale soil degradation. This has led to a policy shift toward sustainable development to join development, social and environmental solutions. Research has shown that the rate of soil erosion is increasing in many countries of the world through a combination of agricultural practices, soil degradation and increasing intensity and frequency of rainstorms [1]. Moreover, human activities (deforestation, unscientific agriculture, and overgrazing) have become the leading cause for increasing

erosion [2], while natural extreme events (storms, typhoons, floods, and earthquakes) also contribute as large-scale sources of soil erosion.

In Taiwan, rainfall is unevenly distributed throughout the year and is frequently concentrated between the months of May and August, the monsoon season [3]. Moreover, the topography of Taiwan is generally steeply sloping. Water is one of the most difficult natural resources to manage in Taiwan and reservoirs function to harmonize the supply and demand for water for various uses in both the wet and dry seasons. Thus, the conservation of reservoir watersheds is very important [4]. Shihmen reservoir is the third largest reservoir in Taiwan. Increases in urban development, sedimentation and water usage have negatively impacted the lifespan of the reservoir. This has encouraged funding and support by the government of Taiwan to develop watershed conservation and recovery plans in order to decrease soil loss in the watershed. Shihmen reservoir is an important source of potable water to Northern Taiwan, and the reservoir has been drastically affected by sedimentation in the past leading to economic losses and cascading issues. This reinforces the importance of research that evaluates the contributory factors (predictors) of soil erosion in the watershed.

In the Shihmen Reservoir watershed, Chen et al. [5] used the Universal Soil Loss Equation (USLE) model to compare soil erosion under four different Digital Elevation Models (DEMs). The result shows that the amounts of soil erosion were different from four different DEMs. It varies between 61.0 t/ha/year and 95.5 t/ha/year. Apart from this, using erosion pins is a simple and effective method to monitor erosion in gullies and hill slopes by assessing the soil loss (erosion depth) temporally and spatially [6,7]. Edeso et al. [8] evaluated the average soil loss in northern Spain using 29 erosion pins. The pins were installed from July 1993 to September 1994 in areas such as conventional harvesting management plots, harvested plots, and down-slope deep plowing plots. From erosion depth results, it was shown that the soil loss across all types of plots increased. Erosion pins and 137Cs (Caesium-137, a radioisotope used as field evidence to quantify eroded or deposited soil) were deployed to quantify soil erosion and sedimentation rates at Grafton, New South Wales [9]. Results showed that among the 94 pins, 46.8% exhibited signs of erosion, 48.9% of deposition, and 4.3% did not exhibit any soil changes. In Taiwan, Lin et al. [10] used erosion pins to evaluate and compare the effect of rainfall on soil erosion in two sub-watersheds of the Shenmu watershed. The result showed that the soil erosion depth quickly increases when the accumulated rainfall is higher than 200 mm.

In other research, Liu et al. [11] employed erosion pins as a reference to evaluate soil erosion between grid cell analysis and slope unit analysis using GIS in the Shihmen reservoir watershed. The results indicated that the prediction of the grid cell method (GIS average of the entire watershed = 6.9 mm/yr) is very close to the average erosion depth measured by the erosion pins (average = 6.5 mm/yr). Chen and Chen [12] used the Mann-Whitney U-test to evaluate the soil erosion between each sub-watershed of the Shihmen reservoir watershed, and the results indicated that (with 95% confidence) the soil erosion depths of different sub-watersheds are different. This indicates that soil erosion depth is closely related to the sub-watershed (i.e., erosion depth varies spatially and shows significant difference depending on the sub-watershed).

Machine learning is a special application of artificial intelligence and in recent years, machine learning models, such as classification and regression tree (CART), and random forest (RF), have seen increased utility for building regional and local susceptibility maps for landslides. For example, Youssef et al. [13] employed four machine learning models, namely random forest, boosted regression tree (BRT), classification and regression tree, and general linear model (GLM) to build landslide susceptibility maps in Wadi Tayyah Basin, Asir Region, Saudi Arabia. The result showed that the models had high accuracy in landslide susceptibility mapping. On the other hand, Chen et al. [14] developed a landslide susceptibility map for Hanyuan County in China by logistic model tree (LMT), random forest, and classification and regression tree models. The research exhibited that the accuracy of all models was satisfactory, and the random forest model attained the highest accuracy of 0.837 and 0.781 in the training and testing data, respectively. Kuhnert et al. [15] predicted gully erosion density by using a random forest model in Berdekin catchment, Australia. Kheir et al. [16] used the

decision tree model to predict soil and bedrock distribution from gully erosion in Lebanon and the model attained an 87% accuracy for soil and bedrock map. However, to the best of our knowledge, no work has focused on the problem of predicting erosion pin measurements using machine learning algorithms. Therefore, the objectives of this study were to predict the soil erosion depth of erosion pin measurements in the Shihmen Reservoir watershed by using machine learning-based approaches, and to determine the contributory factors to soil erosion in the watershed.

We will address two issues: (1) Which is the best machine learning model to predict erosion pin measurements? and (2) What are the important factors that influence the results of the prediction? It is worth noting that this study is not a conventional soil erosion modeling study, which usually starts from an available model and then proceeds to model parameter calibration, model prediction, and model validation with field data (such as sediment amounts or erosion pin measurements). Instead, this study took the opposite approach by going from erosion pin measurements to erosion depth prediction using machine learning algorithms only. In the process, the erosion pin data from 55 slopes that were distributed throughout a vast watershed (76,000 hectares) were integrated and used collectively to reveal possible patterns of erosion behavior. The purpose is to provide a broad-scale analysis and an erosion overview of the study watershed. It should be noted that the erosion pins used in this study were installed on slopes with no signs of landslide or gully erosion. Therefore, this research only addresses sheet and rill erosion. Landslide and gully erosion are not included.

## 2. Research Methods

The research area, the research method, and the environmental factors considered in this study are introduced in the following subsections.

### *2.1. Study Area*

The Shihmen Reservoir watershed is found in northern Taiwan along the Tahan River and covers 759.53 km$^2$ of mountainous area (Figure 1). The elevation fluctuates from 220 m (Shihmen reservoir dam site) to 3527 m, and the topography contains steep slopes that rise from south to north. The slope gradient is greater than 55% for more than 60% of the watershed [17]. The annual average temperature is 19 °C, and the humidity is 82% and higher in the monsoon season. The average annual precipitation is about 2500 mm/year concentrated between May and August. Predominantly, in the monsoon season, there are frequent tropical storms which inundate Taiwan with heavy rainfall and some storms mature into typhoons, which lead to severe sediment-related disasters annually.

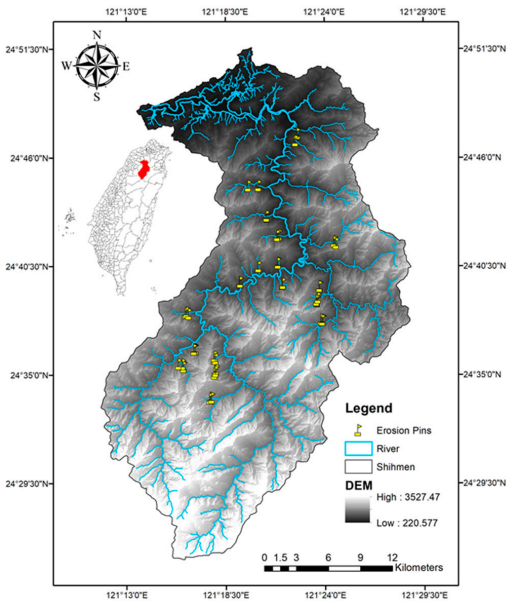



**Figure 1.** Shihmen reservoir watershed and locations of erosion pins on a topographic map.

## 2.2. Research Framework

The predominant examples of applying machine learning to landslide research employed Boolean algebraic data, where the model's target had only two outcomes; Yes (landslide) or No (no landslide), or alternatively, True or False. However, the adaptation of machine learning algorithms to soil erosion research poses an additional layer of difficulty as the target variables are numerical instead of Boolean. Therefore, the model must be able to predict a range of continuous values.

This study represents the first application (to the best of our knowledge) of machine learning algorithms on soil erosion based on erosion pin measurements. Erosion pins were used as soil erosion references to validate the model's results in previous research, however, in this study, the measurements from erosion pins were used as the target of the models to predict soil erosion depths for the entire Shihmen reservoir watershed. Figure 2 shows the flowchart of the process of establishing the decision tree, the random forest and the multiple regression models for this study. The models for this study are developed using the steps described in the following sections.

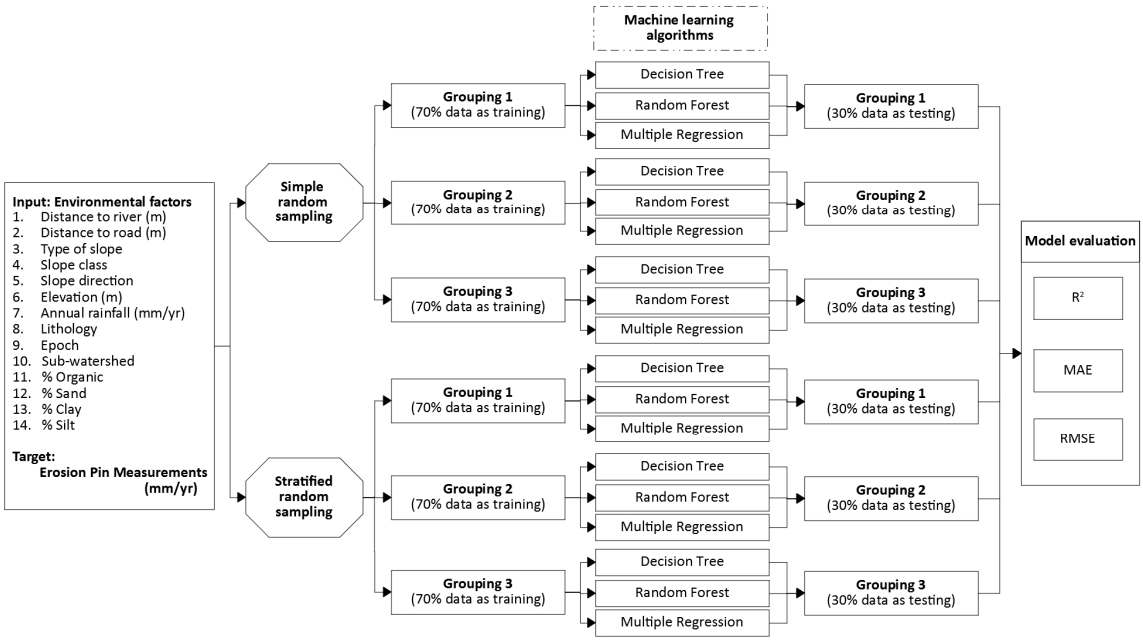



**Figure 2.** Research framework and procedures.

## 2.3. Data Collection

A machine learning model or other statistical inference model requires a target and predictors. The target is the output variable whose values are to be modeled and predicted by the other variables. It is similar to the dependent variable. There must be one target variable in a decision tree analysis. The predictor is a variable which predicts the value of the target variable. It is comparable to the independent variable. There is at least one and usually many predictor variables.

### 2.3.1. Erosion Pin Measurements

The erosion pin measurement, which is the target of the algorithms (the dependent variable), is based on steel bars installed in the watershed. This study employed 550 pins installed at 55 locations (on 55 slopes with 10 pins per slope) over the whole of the Shihmen reservoir watershed (Figure 3) and collected data from 8 September 2008 to 10 October 2011. Lin et al. [10] describes the installation of pins in the field. The average annual erosion depth was calculated for each slope. The measured average erosion depth ranged from 2.17 mm/yr to 13.03 mm/yr.

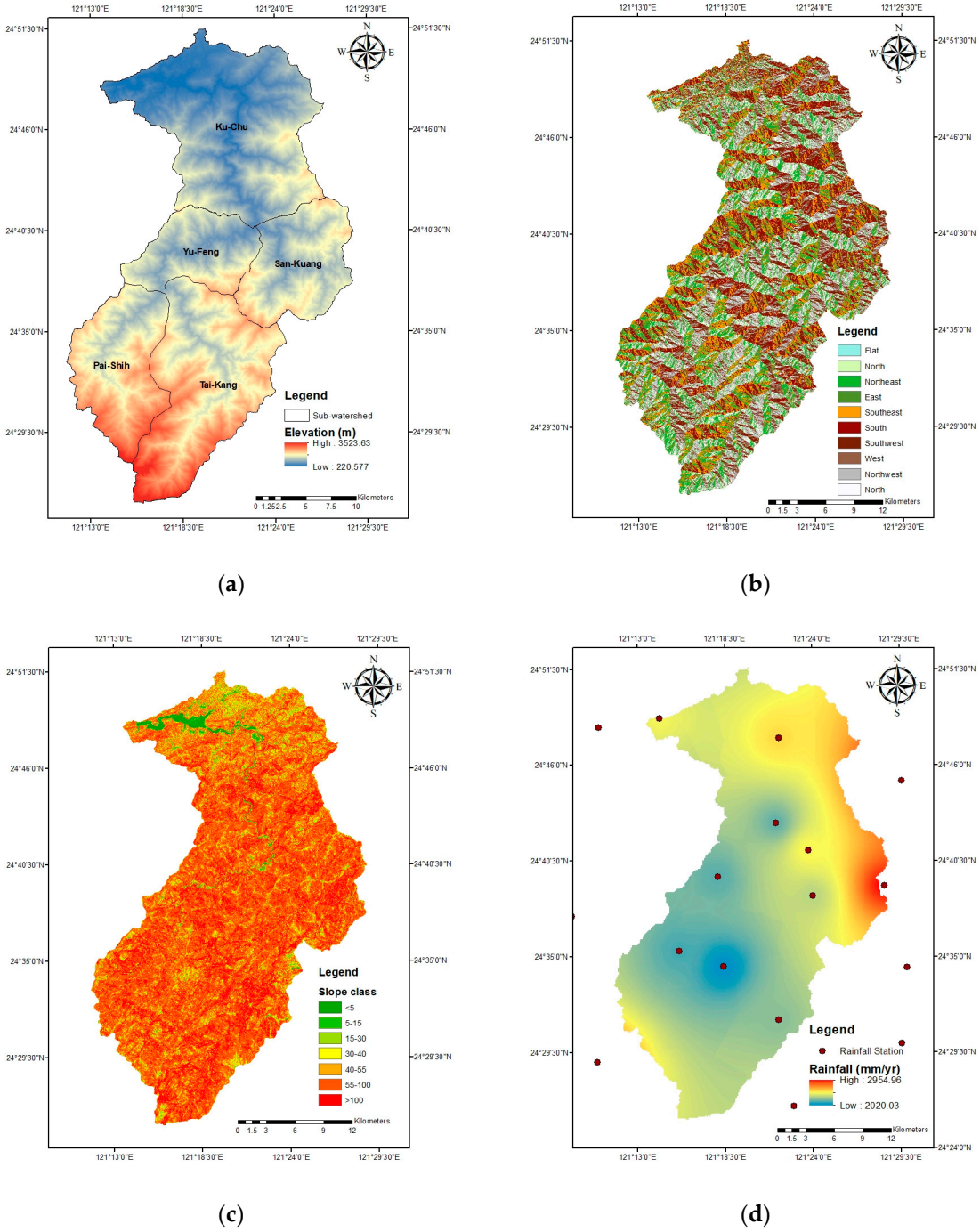


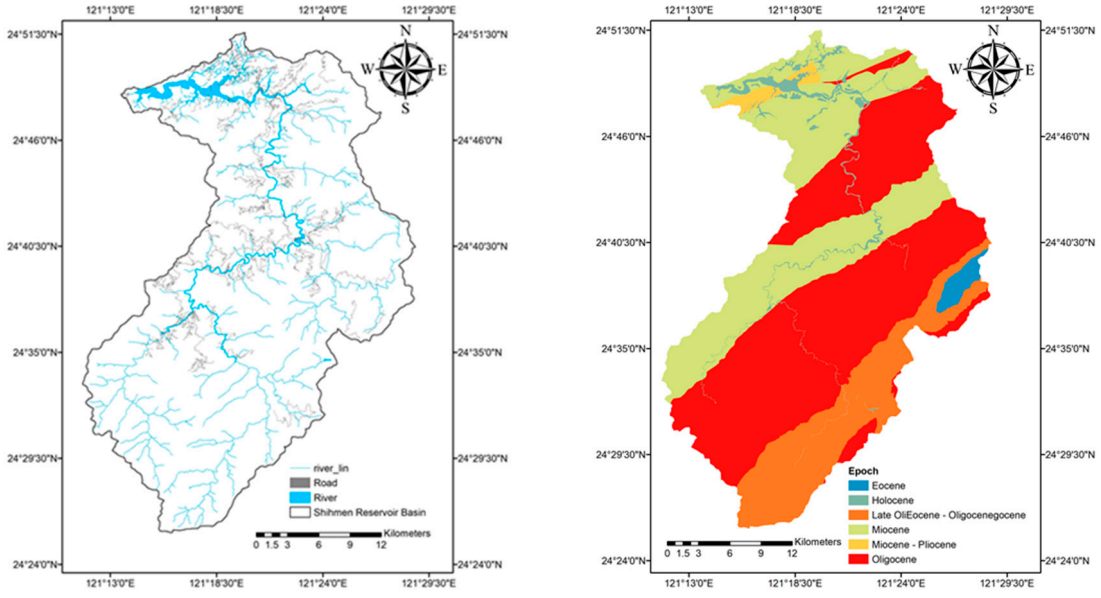

**Figure 3.** Environmental factors: (**a**) Sub-watershed and Elevation, (**b**) Slope direction, (**c**) Slope class, (**d**) Rainfall station and average annual rainfall distribution.

### 2.3.2. Environmental Factors

There were 14 environmental factors which were considered for their impact on soil erosion rates in the Shihmen reservoir watershed and used as the predictors (independent variables) in the models. These factors include sub-watershed, type of slope, slope class, slope direction, elevation, average annual rainfall, distance to river, distance to road, lithology, epoch, and the percentages in the soil of sand, silt, organic content, and clay. The factors were collected from various sources and developed into a geospatial database.

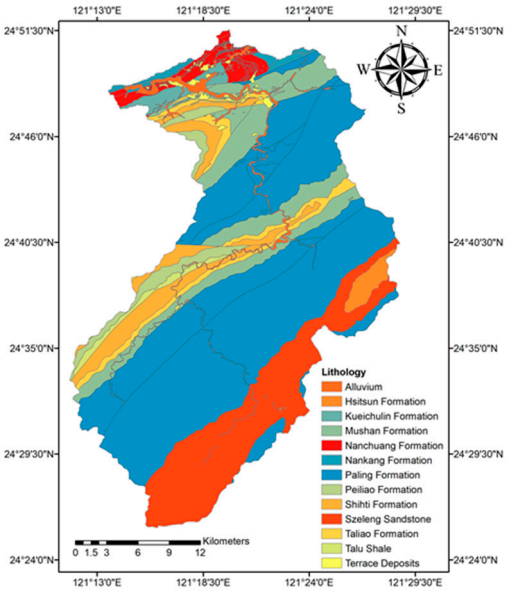
The 10 m × 10 m DEM employed in this study was produced by Taiwan Central Geological Survey in 2013, and the slope class, slope direction, and elevation maps were derived from the DEM using the ArcGIS tool box. Previous soil erosion research has determined that slope and elevation are influential factors in the processes and the quantities of soil erosion [18,19]. The elevation in the Shihmen reservoir watershed ranges from 220 m to 3527 m. The average is 1404 m, and it was divided into five sub-watersheds, namely the Ku-Chu, San-Kuang, Tai-Kang, Yu-Feng, and Pai-Shih sub-watersheds (Figure 3a). Slope direction can be thought of as the aspect which is measured clockwise from 0 to 360 degree. It shows the compass direction that the surface of slope faces at that location (Figure 3b). According to technical regulations from the Soil and Water Conservation Bureau of Taiwan, the slope class data are divided into seven classes by slope percent (<5%, 5–15%, 15–30%, 30–40%, 40–55%, 55–00%, >100%) as shown in Figure 3c. In the entire watershed there are 10 rainfall stations established by the government. However, this study also considers data from 12 rainfall stations which are situated outside and near the border of the watershed. The rainfall data of a total of 22 stations were collected from the Northern Region Water Resources Office, Water Resources Agency, Ministry of Economic Affairs. Rainfall data were collected from 2003 to 2015 and interpolated by IDW tool in ArcGIS (Figure 3d).

The river system data were provided by the Water Resources Agency in 2000, and the road system data by the Industrial Technology Research Institute from 2006 to 2008 on a scale of 1:5000 (Figure 4a). Using the Near tool in ArcGIS, the perpendicular distances between the slopes monitored by erosion pins and any rivers or roads were calculated. Lithostratigraphic units and Epoch were collected from the Central Geological Survey on a scale of 1:50,000 (Figure 4b,c). Finally, the percentages of sand, silt, clay, and organic content in the soil were collected from samples in the Shihmen reservoir watershed [20]. The three types of slope vegetation phases in the watershed are natural, bare, and remediated. These data were collected by field surveys.




(**a**) (**b**)

**Figure 4.** *Cont.*




(**c**)

**Figure 4.** Environmental factors: (**a**) Road and River, (**b**) Epoch, (**c**) Lithostratigraphic units.

### *2.4. Data Separation*

When developing the models, the data is normally divided into two subsets as training data and testing data. The foremost purpose of this division is to use untrained data to check the accuracy of the models and to approve the model predictions. In this study, 70% of the data was for the training stage and 30% of the data for model verification [14,21–23]. There were 55 erosion pin locations (55 slopes), with 38 samples (70%) to build the models and the rest of the samples were reserved as the testing data to verify the models (30% = 17 samples). To evaluate the data sensitivity (to sampling) and to obtain an average result, this step was repeated three times to create three different divisions, namely Grouping #1, Grouping #2, and Grouping #3. Note that even though three groupings were used, they were all based on the same data set. All data sets contain the same numbers for the training and testing data (70%/30%).

Normally, the simple random sampling method is employed to determine a random selection of data from an entire population under 70% of training and 30% of testing. However, this study also applied the stratified random sampling method. In contrast to the simple random sampling method, the stratified random sampling method divides the entire data into small strata based on their characteristics and performs simple random sampling within each stratum to avoid missing representative data in the training or testing data set. For this study, the attribute sub-watershed was chosen as the stratum. This is supported by Chen and Chen [12], who indicate that erosion depth shows significant differences between sub-watersheds. Then, each stratum (sub-watershed) uses simple random sampling to obtain 70% training and 30% testing data.

### *2.5. Model Construction*

Two machine learning algorithms, namely the decision tree and random forest models, were evaluated as models of soil erosion depths in this study, while the multiple regression analysis is used to evaluate the performance of these two algorithms. The algorithms are described below.

#### 2.5.1. Decision Tree

Decision tree is a supervised learning algorithm that comprises of a sequence of "If-Then" rules used to classify data in a model by the training data. Decision trees are used for both classification and

regression purposes depending on if the target variable is categorical or continuous, respectively [24]. The advantage of a decision tree is the ease of constructing the model and the results of the model can be easily interpreted. The main disadvantage of a decision tree is that an overfitting of data could occur. In this study, the classification and regression tree (CART) algorithm [25] in the "rpart" package of R software was used for decision tree modelling.

#### 2.5.2. Random Forest

Random forest is an ensemble learning algorithm which uses multiple decision trees to improve its predictive performance. It avoids the issue of over-fitting exhibited by decision tree algorithms. Random forest contains many trees like a forest. Each tree is built based on classification and regression tree (CART) with a random subset in the training data of each node [25]. The final result of random forest is the average value of all regression trees and it displays the most accurate division between all classification trees [26]. The disadvantage of random forest is that it does not show the final tree of the model. Mean Decrease Gini index (IncNodePurty) is a measure of variable significance based on the Gini impurity index used for the calculation of splits during training. This is effectively a measure of how important a variable is for estimating the trees to create the forest [27]. A higher Mean Decrease Gini index points out a higher variable importance. In this study, the random forest algorithm was run based on the "random forest" package of R software, and the number of trees was set to 1000.

#### 2.5.3. Multiple Regression

Multiple regression is a type of simple linear regression. It explains the relationship between multiple predictor variables and the target [20]. The multiple regression relationship can be expressed as follows:

$$Y = B + A_1X_1 + A_2X_2 + A_3X_3 + \ldots + A_iX_i \quad (1)$$

where $Y$ is the target, $X_i$ are the predictors, $B$ is the value of $Y$ intercept when $X = 0$, and $A_i$ are coefficients.

Multiple regression is a fundamental statistical method. Arnaez et al. [28] used multiple regression to estimate soil erosion and compare it to field experiments and with the USLE-M model. For this study, multiple regression is included in the analysis as a baseline model to evaluate the improvement of accuracy in implementing a machine learning algorithm over a statistically derived regression model, and it was implemented using the "lm" package of the R software.

### *2.6. Model Performance Evaluation*

Three statistical indexes, Coefficient of determination ($R^2$), Root Mean Squared Error (*RMSE*), and Mean Absolute Error (*MAE*), were used to verify the accuracy of models in this study. Coefficient of determination is a statistical measure that indicates the association between the measured and predicted values following the best fit line (Equation (2)). $R^2$ is always between 0% and 100%, and higher values are better.

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum(Y - Y_1)^2}{\sum\left(Y - \overline{Y}\right)^2} \quad (2)$$

where *SSE* is the Sum of Squares of Error, *SST* is the Total Sum of Squares, $\overline{Y}$ is the average of predicted values, and $Y$ and $Y_1$ are defined in Figure 5.

*RMSE* and *MAE* are both commonly used to evaluate the error of models between the observed and predicted data based on the regression line or the 1:1 line, and both show the average model prediction error in units of the variable. The value can range from 0 to infinity, with lower values being better. Figure 5 shows the difference between *RMSE* relative to the regression line and the 1:1 line. It can be seen that the *RMSE* relative to the 1:1 line is the prediction error between the predicted value from the model and the observed value (Equations (3) and (4)). On the other hand, the *RMSE*

relative to the regression line is the prediction error between the predicted value and the value from the regression line (Equations (5) and (6)). Note that it is the 1:1 line that is important in the evaluation of machine learning models because the 1:1 line indicates that the predicted value is the same as the observed value.

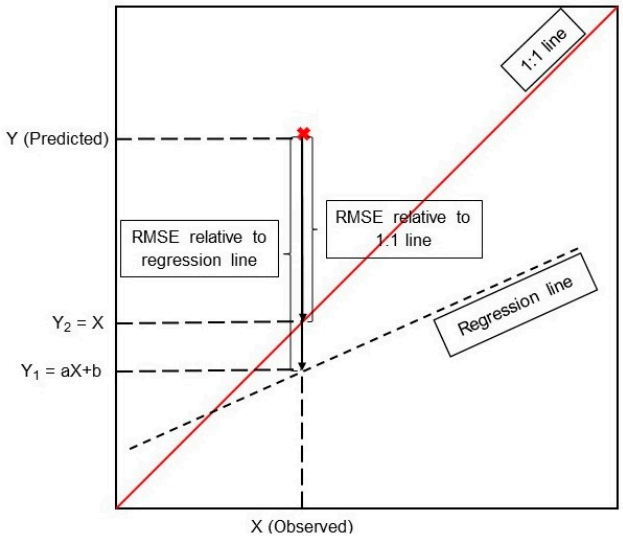



**Figure 5.** Different Root Mean Squared Errors (RMSEs) relative to the regression line and the 1:1 line.

*RMSE* and *MAE* relative to the 1:1 line:

$$RMSE = \sqrt{\frac{\sum (Y - Y_2)^2}{n}} \tag{3}$$

$$MAE = \frac{1}{n} \sum |Y - Y_2| \tag{4}$$

*RMSE* and *MAE* relative to the regression line:

$$RMSE = \sqrt{\frac{\sum (Y - Y_1)^2}{n}} \tag{5}$$

$$MAE = \frac{1}{n} \sum |Y - Y_1| \tag{6}$$

where $Y$ and $X$ are the predicted and observed erosion pin measurements, respectively; $Y_1$ is the value from regression; $Y_2$ is defined in Figure 5; and n is the sample size.

## 3. Results and Discussion

This study has 55 erosion pin measurements (representing the average soil erosion depths of the 55 slopes) and 14 environment factors. They are divided into the training and testing data. To evaluate machine learning models, the accuracy of both the training and testing data needs to be taken into account. Two methods were employed to divide the data, namely the simple random sampling and the stratified random sampling methods. The difference between the two methods is that the stratified random sampling divides the data into small strata first and then samples proportionally within the strata. This allows the stratified random sampling to obtain a proper representation of the components of the data. As a result, the stratified random sampling provides better outcomes than the simple random sampling as researchers can control the sampling to make sure that all types of data are represented in the grouping of data. The following are the important findings of this research.

### 3.1. Comparison between Machine Learning Models and Multiple Regression

Tables 1 and 2 show the statistics of decision tree, random forest, and multiple regression under the two sampling methods, simple random sampling and stratified random sampling, respectively.

**Table 1.** Results of decision tree, random forest, and multiple regression using simple random sampling.

| Simple Random Sampling | Decision Tree | | | | Random Forest | | | | Multiple Regression | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Training | | Testing | | Training | | Testing | | Training | | Testing | |
| **Grouping #1** | **1:1** | **Reg.** | **1:1** | **Reg.** | **1:1** | **Reg.** | **1:1** | **Reg.** | **1:1** | **Reg.** | **1:1** | **Reg.** |
| *MAE* (mm/yr) | 1.42 | 0.94 | 1.72 | 1.25 | 0.79 | 0.31 | 1.33 | 0.35 | 1.21 | 0.94 | 1.48 | 1.21 |
| *RMSE* (mm/yr) | 1.76 | 1.25 | 2.16 | 1.39 | 0.95 | 0.41 | 1.70 | 0.48 | 1.47 | 1.19 | 1.81 | 1.47 |
| $R^2$ | - | 0.50 * | - | 0.05 | - | 0.94 * | - | 0.38 * | - | 0.65 * | - | 0.31 |
| **Grouping #2** | **1:1** | **Reg.** | **1:1** | **Reg.** | **1:1** | **Reg.** | **1:1** | **Reg.** | **1:1** | **Reg.** | **1:1** | **Reg.** |
| *MAE* (mm/yr) | 1.38 | 0.92 | 2.15 | 1.02 | 0.79 | 0.28 | 1.59 | 0.42 | 1.22 | 0.88 | 2.03 | 1.57 |
| *RMSE* (mm/yr) | 1.75 | 1.16 | 2.55 | 1.31 | 0.99 | 0.38 | 1.78 | 0.57 | 1.43 | 1.13 | 2.33 | 1.84 |
| $R^2$ | - | 0.44 | - | 0.12 | - | 0.94 * | - | 0.64 * | - | 0.63 * | - | 0.33 |
| **Grouping #3** | **1:1** | **Reg.** | **1:1** | **Reg.** | **1:1** | **Reg.** | **1:1** | **Reg.** | **1:1** | **Reg.** | **1:1** | **Reg.** |
| *MAE* (mm/yr) | 1.37 | 1.04 | 2.32 | 1.61 | 0.73 | 0.30 | 1.61 | 0.60 | 1.16 | 0.89 | 1.83 | 1.28 |
| *RMSE* (mm/yr) | 1.80 | 1.22 | 2.66 | 1.66 | 0.93 | 0.37 | 1.81 | 0.85 | 1.43 | 1.16 | 2.16 | 1.68 |
| $R^2$ | - | 0.46 * | - | 0.01 | - | 0.95 * | - | 0.33 | - | 0.66 * | - | 0.12 |
| **Average** | **1:1** | **Reg.** | **1:1** | **Reg.** | **1:1** | **Reg.** | **1:1** | **Reg.** | **1:1** | **Reg.** | **1:1** | **Reg.** |
| *MAE* (mm/yr) | 1.39 | 0.97 | 2.06 | 1.29 | 0.77 | 0.30 | 1.51 | 0.46 | 1.20 | 0.90 | 1.78 | 1.35 |
| *RMSE* (mm/yr) | 1.77 | 1.21 | 2.46 | 1.45 | 0.96 | 0.39 | 1.76 | 0.63 | 1.44 | 1.16 | 2.10 | 1.66 |
| $R^2$ | - | 0.47 * | - | 0.06 | - | 0.94 * | - | 0.45 * | - | 0.65 * | - | 0.25 |

(*) Denotes statistically significant results ($R^2 > 0.36$); Reg.: Regression line.

**Table 2.** Results of decision tree, random forest, and multiple regression using stratified random sampling.

| Stratified Random Sampling | Decision Tree | | | | Random Forest | | | | Multiple Regression | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Training | | Testing | | Training | | Testing | | Training | | Testing | |
| **Grouping #1** | **1:1** | **Reg.** | **1:1** | **Reg.** | **1:1** | **Reg.** | **1:1** | **Reg.** | **1:1** | **Reg.** | **1:1** | **Reg.** |
| *MAE* (mm/yr) | 1.63 | 0.83 | 1.87 | 0.96 | 0.76 | 0.24 | 1.44 | 0.84 | 0.95 | 0.77 | 2.51 | 2.33 |
| *RMSE* (mm/yr) | 2.00 | 0.93 | 2.53 | 1.01 | 0.93 | 0.30 | 1.68 | 1.06 | 1.22 | 1.03 | 3.19 | 2.97 |
| $R^2$ | - | 0.22 | - | 0.08 | - | 0.95 * | - | 0.59 * | - | 0.71 * | - | 0.19 |
| **Grouping #2** | **1:1** | **Reg.** | **1:1** | **Reg.** | **1:1** | **Reg.** | **1:1** | **Reg.** | **1:1** | **Reg.** | **1:1** | **Reg.** |
| *MAE* (mm/yr) | 1.19 | 0.87 | 2.01 | 1.03 | 0.68 | 0.24 | 1.51 | 0.97 | 0.95 | 0.84 | 2.90 | 1.84 |
| *RMSE* (mm/yr) | 1.58 | 1.08 | 2.52 | 1.27 | 0.89 | 0.29 | 1.77 | 1.24 | 1.25 | 1.02 | 3.89 | 2.57 |
| $R^2$ | - | 0.47 * | - | 0.32 | - | 0.95 * | - | 0.64 * | - | 0.67 * | - | 0.01 |
| **Grouping #3** | **1:1** | **Reg.** | **1:1** | **Reg.** | **1:1** | **Reg.** | **1:1** | **Reg.** | **1:1** | **Reg.** | **1:1** | **Reg.** |
| *MAE* (mm/yr) | 1.27 | 0.90 | 1.73 | 1.31 | 0.78 | 0.25 | 1.36 | 0.64 | 1.13 | 0.91 | 2.26 | 2.01 |
| *RMSE* (mm/yr) | 1.61 | 1.16 | 2.31 | 1.49 | 0.97 | 0.35 | 1.79 | 0.82 | 1.29 | 1.07 | 3.33 | 2.97 |
| $R^2$ | - | 0.52 * | - | 0.18 | - | 0.94 * | - | 0.51 * | - | 0.69 * | - | 0.14 |
| **Average** | **1:1** | **Reg.** | **1:1** | **Reg.** | **1:1** | **Reg.** | **1:1** | **Reg.** | **1:1** | **Reg.** | **1:1** | **Reg.** |
| *MAE* (mm/yr) | 1.36 | 0.87 | 1.87 | 1.10 | 0.74 | 0.24 | 1.44 | 0.82 | 1.01 | 0.84 | 2.56 | 2.06 |
| *RMSE* (mm/yr) | 1.73 | 1.06 | 2.45 | 1.26 | 0.93 | 0.31 | 1.75 | 1.04 | 1.25 | 1.04 | 3.47 | 2.84 |
| $R^2$ | - | 0.40 * | - | 0.19 | - | 0.95 * | - | 0.58 * | - | 0.69 * | | 0.11 |

(*) Denotes statistically significant results ($R^2 > 0.36$); Reg.: Regression line.

Multiple regression is the baseline model to evaluate the other two machine learning models. Multiple regression, which is commonly employed in various analytical studies, is a statistical model that predicts the erosion depth ($Y$) established by the values of the 14 predictor variables ($X_i$) and is used to evaluate the other two models by comparison of results. The average errors of the models were represented by the *RMSE* relative to the 1:1 line, which is the square root of the average squared differences between the predicted and observed values. The average *RMSE* of the multiple regression analysis in the training data is 1.44 mm/yr, but the average *RMSE* of the testing data is increased to 2.10 mm/yr. Further, the average *RMSE* of the decision tree algorithm in the training data is 1.77 mm/yr, and the average *RMSE* increases to 2.46 mm/yr in the testing data. As an outcome, the error of the decision tree model is higher than the error of the multiple regression analysis for not only the training data but also the testing data. Therefore, the decision tree algorithm is not a good algorithm to predict erosion pin measurements. On the other hand, the random forest algorithm has lower RMSEs of 0.96 mm/yr in the training data and 1.76 mm/yr in the testing data. Both of them are lower than those of the multiple regression analysis, and therefore are the lowest among the three models. From this result, we can conclude that the random forest algorithm is the best model under simple random sampling.

For comparison purposes, we have also included the *RMSE* values relative to the regression line in Table 1. It should be noted that the value of *RMSE* relative to the regression line is always less than the value of *RMSE* relative to the 1:1 line, no matter what the model is and what the data are. This is because the regression line is always the best fit line as is evident from Figure 5. For the training data set, the average *RMSE* of the decision tree algorithm is 1.21 mm/yr, which is higher than the 1.16 mm/yr of the multiple regression analysis. For the testing data set, however, the average *RMSE* of the decision tree algorithm is lower than that of the multiple regression analysis. Their values are 1.45 mm/yr and 1.66 mm/yr, respectively. Lastly, the average *RMSE* of the random forest algorithm is again the lowest among the three models in both the training data (0.39 mm/yr) and the testing data (0.63 mm/yr).

The $R^2$ value measures the statistical association between the predicted and the measured values following the best fit line. In the training data, the average $R^2$ obtained from the three models show a permissible accuracy ($R^2 > 0.36$), which means that all three models can be used with the training data with good results. However, when the trained models are applied to the testing data, only the random forest algorithm maintains a permissible accuracy ($R^2 = 0.45 > 0.36$). The other two models have a much lower average $R^2$ of 0.25 and 0.06, respectively. They are considered insignificant.

As can be seen from Table 1, similar conclusions can be drawn from the *MAE* values as well.

In the case of stratified random sampling where sampling is based on strata defined by sub-watersheds, the results are shown in Table 2. The average *RMSE* relative to the 1:1 line of the multiple regression analysis in the training data is 1.25 mm/yr. However, the average *RMSE* increased to 3.47 mm/yr in the testing data. Additionally, the average value of *RMSE* relative to the 1:1 line of the decision tree algorithm is 1.73 mm/yr in the training data, and the average *RMSE* rises to 2.45 mm/yr in the testing data. As a result, the average *RMSE* of multiple regression is lower than that of the decision tree algorithm in the training data, but the result is reversed in the testing data. Finally, the random forest algorithm shows an average value of *RMSE* of 0.93 mm/yr and 1.75 mm/yr in the training and testing data, respectively. Both are the lowest among the three models. Therefore, we can conclude that the random forest algorithm is the best model under stratified random sampling.

Similar to Table 1, Table 2 also shows that the value of *RMSE* relative to the regression line is always lower than the *RMSE* relative to the 1:1 line no matter what the model is and what the data are. This is because the regression line is always the best fit line (as shown in Figure 5). In the training data set, the average *RMSE* of the decision tree algorithm is 1.06 mm/yr which is higher than the multiple regression analysis of 1.04 mm/yr. In the testing data set, the average *RMSE* of the decision tree algorithm is 1.26 mm/yr, lower than the 2.84 mm/yr of the multiple regression analysis. Again, the *RMSE* of the random forest algorithm is the lowest among the three models in both the training data (0.31 mm/yr) and the testing data (1.04 mm/yr).

For stratified random sampling, the average $R^2$ obtained from the three models also show a permissible accuracy ($R^2 > 0.36$), which means that all three models can be used with the training data with good results. However, when the trained models are applied to the testing data, only the random forest algorithm maintains a permissible accuracy ($R^2 = 0.58 > 0.36$). The other two models have much lower average $R^2$ of 0.11 and 0.19, respectively. They are considered insignificant.

Combining the results of Tables 1 and 2, it can be seen that the random forest algorithm is better than the decision tree algorithm and the multiple regression analysis. Overall, the average value of *RMSE* of the random forest is lower than those of the decision tree algorithm and the multiple regression analysis for both the 1:1 line and the regression line. Furthermore, the average *MAE* value of the random forest algorithm is also the lowest, which also indicates that the random forest algorithm is the best model. In addition, only the average $R^2$ of the random forest algorithm maintains a statistically significant value (>0.36) when the models trained by the training data are applied to the testing data. Based on these three results, it is evident that the random forest algorithm is the best model for predicting erosion pin measurements in the Shihmen reservoir watershed.

In addition, the comparison of Tables 1 and 2 also shows that the stratified random sampling method is better than the simple random sampling method when one is attempting to build a prediction model of erosion pins. In the training data, the average *RMSE* of the random forest algorithm is 0.93 mm/yr under stratified random sampling, which is lower than the 0.96 mm/yr under simple random sampling. Similarly, in the testing data, the average *RMSE* of the random forest algorithm is 1.75 mm/yr under stratified random sampling, which is also lower than the 1.76 mm/yr under simple random sampling. Consequently, it can be concluded that the stratified random sampling method is preferred in erosion pin modeling in the Shihmen reservoir watershed.

To further illustrate the difference between the decision tree, random forest, and multiple regression methods, we took Grouping #2 as an example and plotted the results in Figure 6. The sampling method used here is the stratified random sampling. The figures on the left-hand side of Figure 6 are based on the training data set, whereas the figures on the right-hand side are based on the testing data. It can be seen from Figure 6a,b that the decision tree algorithm summarizes the results into four prediction values within the tree. Therefore, the result of the model is not very encouraging. The multiple regression analysis is better (Figure 6e). However, the model loses its validity and gives poor prediction when it is applied to the testing data (Figure 6f). Only the random forest algorithm performs the best. The model improves its predictive performance by averaging the values from a multitude of decision trees, therefore, its accuracy is higher than a single decision tree and it also outperforms the multiple regression analysis (Figure 6c). When the random forest algorithm is applied to the testing data, a satisfactory result is obtained (Figure 6d).

Since three groupings and two sampling methods were used in this study, they each produced a different multiple regression equation, a different decision tree, and a different random forest model. Without listing all of them here, we again take Grouping #2 of the stratified random sampling as an example and display its multiple regression equation in Equation (7) and its decision tree in Figure 7.

$$\begin{aligned} Y = {} & -613.1 + 0.01168X_1 - 0.0003118X_2 + 1.087X_3 + 0.2413X_4 \\ & -0.157X_5 - 0.0005013X_6 + 0.002939X_7 - 0.06576X_8 - 2.868X_9 \\ & -0.005446X_{10} + 0.6744X_{11} + 6.107X_{12} + 6.387X_{13} + 5.755X_{14} \end{aligned} \quad (7)$$

where $Y$ is the erosion depth, $X_1$ is the distance to road, $X_2$ is the distance to river, $X_3$ is the type of slope, $X_4$ is the slope class, $X_5$ is the slope direction, $X_6$ is the elevation, $X_7$ is the average annual rainfall, $X_8$ is the lithology, $X_9$ is the epoch, $X_{10}$ is the sub-watershed, $X_{11}$ is the percentage of organic material, $X_{12}$ is the percentage of sand, $X_{13}$ is the percentage of clay, and $X_{14}$ is the percentage of silt.

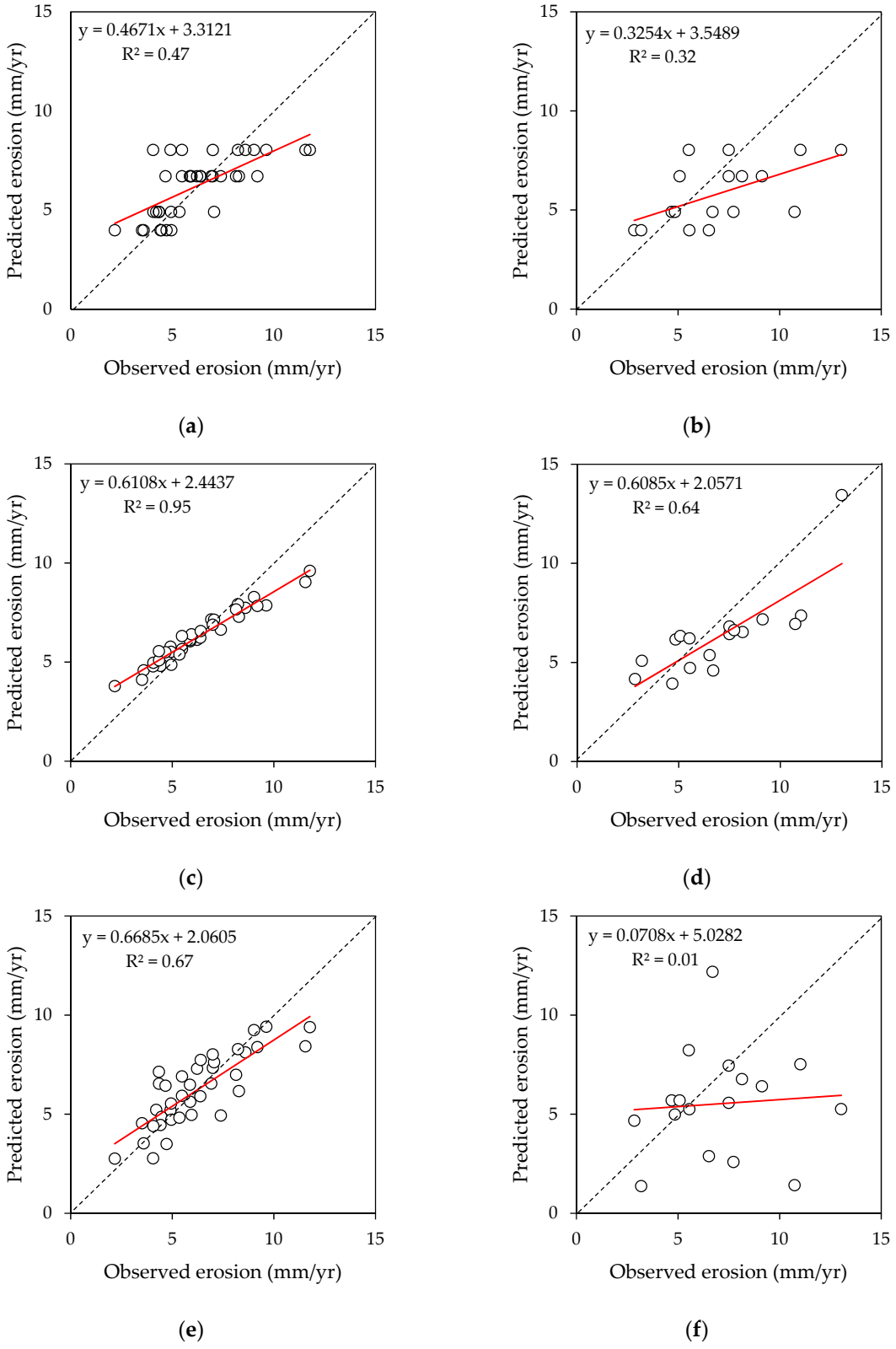



**Figure 6.** Comparison between the predicted values and observed values (using stratified random sampling and Grouping #2 as an example) of (**a**) the training data using decision tree, (**b**) the testing data using decision tree, (**c**) the training data using random forest, (**d**) the testing data using random forest, (**e**) the training data using multiple regression, (**f**) the testing data using multiple regression.

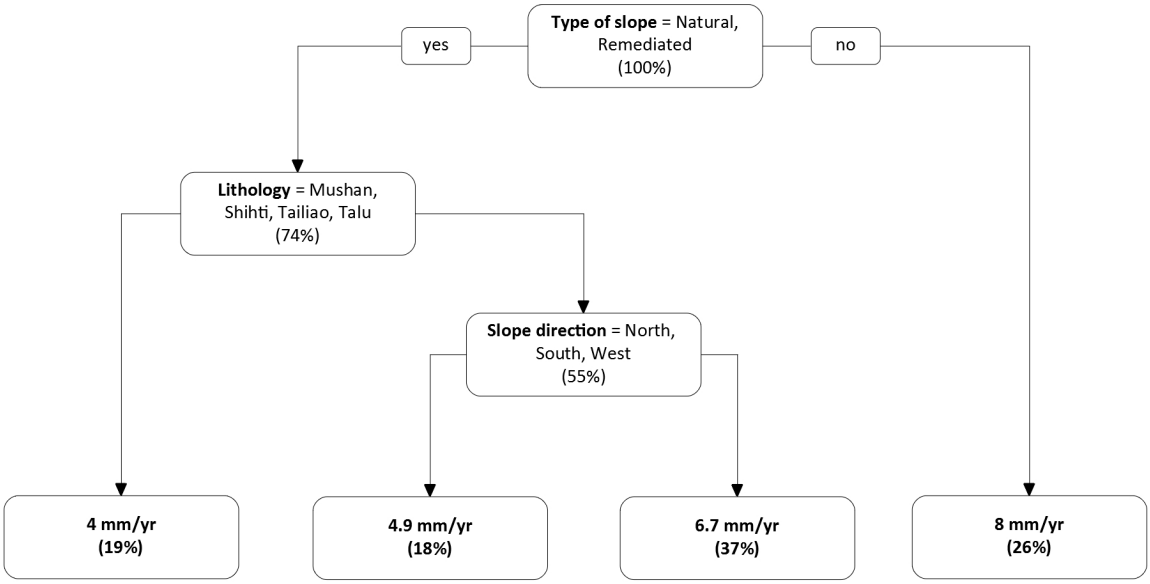



**Figure 7.** The resulting decision tree of Grouping #2 of the stratified random sampling.

### 3.2. Identification of Important Factors

Since the random forest algorithm is the best performing model by all three statistical criteria (*RMSE*, *MAE*, and $R^2$), the Mean Decrease Gini value is used to determine the top three most important factors among the 14 environmental factors, as shown in Figure 8. As can be seen from Figure 8, the three most important factors are type of slope, slope direction, and sub-watershed.

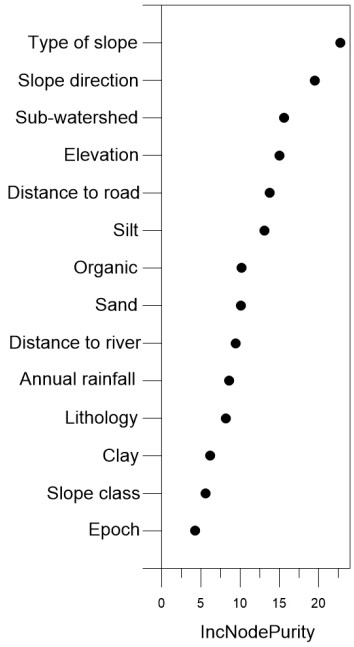



**Figure 8.** Order of important factors of the random forest algorithm using stratified random sampling.

The first important factor, type of slope, represents the different condition that a slope is in (natural, bare, or remediated). If we group the erosion pins according to the type of slopes, it can be seen from Table 3 that they are close to evenly represented in the data. The bare slopes, with the condition that represents soil without vegetation and the absence of protection on the slope, tend to have higher amounts of soil mobilization and transport by water [29]. The average erosion depth is the highest for this group with 8.38 mm/yr. In contrast, the average erosion depth is the lowest in remediated slopes

(5.19 mm/yr), which means that the remediation is successful in minimizing erosion. There are many methods for remediating slopes, such as changing slope lengths, improving drainage, and employing bioengineering stabilization methods. For natural slopes, the average erosion depth is 6.42 mm/yr, a value between the two extremes of bare slopes and remediated slopes. This is consistent with general conception and understanding.

**Table 3.** Average erosion depths of different types of slope.

| | Bare | Natural | Remediated |
|---|---|---|---|
| **Number of pins** | 14 | 21 | 20 |
| **Average (mm/yr)** | 8.38 | 6.42 | 5.19 |

The second important factor, slope direction, has a total of nine compass directions (including flat). The results indicate that the slope direction (slope aspect) can affect soil erosion and this is consistent with previous research. For example, Zhang et al. [30] indicates that slope direction affects the average daily sunlight duration, which in turn impacts the slope. Slopes with high sunlight duration have a higher rate of evaporation of water and this results in less vegetation which can intensify erosion.

For the third important factor, sub-watershed, the erosion pin measurements are significantly different in different sub-watersheds. This is consistent with Chen and Chen [12], who use the Mann-Whitney U-test to prove that soil erosion is statistically different between some of the sub-watersheds of the Shihmen Reservoir watershed with a 95% confidence interval. The important findings of this study are summarized in Table 4.

**Table 4.** Main findings of this study.

| | **Important Findings** |
|---|---|
| Finding #1 | When an insufficient amount of site-specific data is available to apply an empirically or physically based soil erosion model, machine learning-based approaches are shown to provide an alternative method to analyze data from different slopes and predict soil erosion depths in a watershed. |
| Finging #2 | To predict the soil erosion depths of the Shihmen reservoir watershed in Taiwan, the stratified random sampling method is proved to yield better results than the simple random sampling method. |
| Finding #3 | When decision tree and random forest algorithms are compared with multiple regression analysis, the random forest algorithm performed the best of the three methods in the Shihmen reservoir watershed. |
| Finding #4 | The average error (as measured by *RMSE*) of the stratified random sampling method of the random forest algorithm is 0.93 mm/yr in the training data and 1.75 mm/yr in the testing data in the Shihmen reservoir watershed. |

### *3.3. Comparison between Machine Learning Models and Traditional Soil Erosion Models*

It is not the intent of this study to compare conventional soil erosion models with machine learning algorithms. In a typical study of soil erosion modeling, one or several models would be introduced and applied to a study area. Then field measurements (such as sediment amounts and erosion pin measurements) would be used to validate the models. That is the so-called forward approach to analyzing the soil erosion problem. This study is the opposite in that it originated from erosion pin measurements and took a so-called inverse approach. That is, with only erosion pin measurements that are distributed throughout a vast watershed (76,000 hectares) and no applicable soil erosion models (due to the lack of enough site-specific data), how can we interpret the measurements and gain insights from a data science perspective? We consider this to be the fundamental difference between our approach and conventional soil erosion modeling. Other than the lack of site-specific data, we also did not use empirical and process-based soil erosion models because plot-scale studies at 55 locations

(slopes) could not be easily united across the entire watershed to get a total watershed perspective. Given these reasons, we chose machine learning algorithms to provide a broad-scale analysis and an erosion overview of the study watershed, and in the process, we obtained a satisfactory result.

## 4. Conclusions

The multiple regression analysis is a common method of analyzing the influence of environmental factors in statistical studies. In this study, multiple regression was used as a baseline comparison to evaluate the effectiveness of machine learning algorithms, namely the decision tree and the random forest algorithms, in the soil erosion analysis of the Shihmen reservoir watershed. This study utilized 550 erosion pins at 55 different locations (on 55 slopes) within the watershed to evaluate the erosion depth as the prediction target of these models. The predictors of this study were 14 environmental factors established by previous researchers in the literature as influential in soil erosion. The data were divided into training and testing data sets. Repeated sampling was performed three times to compute the average *MAE*, *RMSE*, and $R^2$ for different machine learning algorithms.

The results of this study clearly show that machine learning algorithms can be used to predict erosion pin measurements. Moreover, the average root-mean-squared error of the random forest algorithm is lower than those of the decision tree algorithm and the multiple regression analysis. Therefore, the random forest algorithm is determined to be the best algorithm to predict erosion depths in the Shihmen reservoir watershed. The average *RMSE* is 0.93 mm/yr in the training data, and 1.75 mm/yr in the testing data using the stratified random sampling. This result is very satisfactory. It shows that the prediction error is small, and a random forest consisting of hundreds of decision trees can deliver much better predictions than the other two models. The Mean Decrease Gini index of the random forest algorithm shows that the three most important factors of the 14 environmental factors used in this study for predicting erosion depths are type of slope, slope direction, and sub-watershed.

Finally, the stratified random sampling and simple random sampling methods were applied to divide the data into the training and testing data sets for this study. Although the simple random sampling method is commonly used, it may not be an ideal way to separate the training and testing data because some important subgroups may be underrepresented or missed in the sampling process. On the other hand, the stratified random sampling method separates samples into small strata first before sampling. It allows each stratum to be properly represented in the sampling process. This improves the accuracy of machine learning models over the simple random sampling method. The results of this study support and validate the use of the stratified random sampling method.

**Author Contributions:** Conceptualization, W.C.; Data curation, K.A.N. and B.S.L.; Formal analysis, K.A.N.; Funding acquisition, W.C. and U.S.; Investigation, B.S.L.; Methodology, W.C.; Project administration, W.C.; Resources, W.C., B.S.L. and U.S.; Software, K.A.N.; Supervision, W.C. and U.S.; Validation, K.A.N.; Writing—original draft, K.A.N.; Writing—review & editing, K.A.N., W.C., B.S.L., U.S. and K.T.

**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 numbers MOST 106-2119-M-027-004 and MOST 107-2119-M-027-003).

**Acknowledgments:** The financial assistance in funding this research work is gratefully appreciated.

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

## References

1. Amore, E.; Modica, C.; Nearing, M.A.; Santoro, V.C. Scale effect in USLE and WEPP application for soil erosion computation from three Sicilian basins. *J. Hydrol.* **2004**, *293*, 100–114. [CrossRef]
2. Islam, M.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]
3. Tsai, Z.-X.; You, G.J.Y.; Lee, H.-Y.; Chiu, Y.-J. Use of a total station to monitor post-failure sediment yields in landslide sites of the Shihmen reservoir watershed, Taiwan. *Geomorphology* **2012**, *139*, 438–451. [CrossRef]

4. Chen, Y.-J.; Chang, K.-C. A spatial–temporal analysis of impacts from human development on the Shih-men Reservoir watershed, Taiwan. *Int. J. Remote Sens.* **2011**, *32*, 9473–9496. [CrossRef]
5. 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]
6. 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]
7. Haigh, M.J. The use of erosion pins in the study of slope evolution. *Br. Geomorphol. Res. Group Tech. Bull.* **1977**, *18*, 31–49.
8. Edeso, J.; Merino, A.; Gonzalez, M.; Marauri, P. Soil erosion under different harvesting managements in steep forestlands from northern Spain. *Land Degrad. Dev.* **1999**, *10*, 79–88. [CrossRef]
9. Saynor, M.J.; Loughran, R.J.; Erskine, W.D.; Scott, P. Sediment movement on hillslopes measured by caesium-137 and erosion pins. In *Variability in Stream Erosion and Sediment Transport, Proceedings of the Canberra Symposium*; No. 224; IAHS Publ.: Wallingford, UK, 1994; pp. 87–93.
10. 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]
11. 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]
12. Chen, W.; Chen, A. A statistical test of erosion pin measurements. In Proceedings of the 39th Asian Conference on Remote Sensing (ACRS 2018), Kuala Lumpur, Malaysia, 15–19 October 2018.
13. Youssef, A.M.; Pourghasemi, H.R.; Pourtaghi, Z.S.; Al-Katheeri, M.M. Landslide susceptibility mapping using random forest, boosted regression tree, classification and regression tree, and general linear models and comparison of their performance at Wadi Tayyah Basin, Asir Region, Saudi Arabia. *Landslides* **2016**, *13*, 839–856. [CrossRef]
14. 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]
15. Kuhnert, P.M.; Henderson, A.K.; Bartley, R.; Herr, A. Incorporating uncertainty in gully erosion calculations using the random forests modelling approach. *Environmetrics* **2010**, *21*, 493–509. [CrossRef]
16. Kheir, R.B.; Chorowicz, J.; Abdallah, C.; Dhont, D. Soil and bedrock distribution estimated from gully form and frequency: A GIS-based decision-tree model for Lebanon. *Geomorphology* **2008**, *93*, 482–492. [CrossRef]
17. Tsai, F.; Lai, J.-S.; Chen, W.W.; Lin, T.-H. Analysis of topographic and vegetative factors with data mining for landslide verification. *Ecol. Eng.* **2013**, *61*, 669–677. [CrossRef]
18. Liu, B.; Nearing, M.; Shi, P.; Jia, Z. Slope length effects on soil loss for steep slopes. *Soil Sci. Soc. Am. J.* **2000**, *64*, 1759–1763. [CrossRef]
19. Liu, B.; Nearing, M.A.; Risse, L.M. Slope gradient effects on soil loss for steep slopes. *Trans. ASAE* **1994**, *37*, 1835–1840. [CrossRef]
20. 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]
21. Hong, H.; Panahi, M.; Shirzadi, A.; Ma, T.; Liu, J.; Zhu, A.X.; Kazakis, N. Flood susceptibility assessment in Hengfeng area coupling adaptive neuro-fuzzy inference system with genetic algorithm and differential evolution. *Sci. Total Environ.* **2018**, *621*, 1124–1141. [CrossRef]
22. Lee, S.; Kim, Y.-S.; Oh, H.-J. Application of a weights-of-evidence method and GIS to regional groundwater productivity potential mapping. *J. Environ. Manag.* **2012**, *96*, 91–105. [CrossRef] [PubMed]
23. Riaz, M.T.; Basharat, M.; Hameed, N.; Shafique, M.; Luo, J. A data-driven approach to landslide-susceptibility mapping in mountainous terrain: case study from the Northwest Himalayas, Pakistan. *Nat. Hazards Rev.* **2018**, *19*, 05018007. [CrossRef]
24. Lantz, B. *Machine Learning with R*; Packt Publishing Ltd.: Birmingham, UK, 2013.
25. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]

26. Micheletti, N.; Foresti, L.; Robert, S.; Leuenberger, M.; Pedrazzini, A.; Jaboyedoff, M.; Kanevski, M. Machine learning feature selection methods for landslide susceptibility mapping. *Math. Geosci.* **2014**, *46*, 33–57. [CrossRef]
27. Prasad, A.M.; Iverson, L.R.; Liaw, A. Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. *Ecosystems* **2006**, *9*, 181–199. [CrossRef]
28. Arnaez, J.; Lasanta, T.; Ruiz-Flaño, P.; Ortigosa, L. Factors affecting runoff and erosion under simulated rainfall in Mediterranean vineyards. *Soil Tillage Res.* **2007**, *93*, 324–334. [CrossRef]
29. Bagio, B.; Bertol, I.; Wolschick, N.H.; Schneiders, D.; dos Santos, M.A.N. Water erosion in different slope lengths on bare soil. *Revista Brasileira de Ciência do Solo* **2017**, *41*. [CrossRef]
30. Zhang, W.; Zhou, J.; Feng, G.; Weindorf, D.C.; Hu, G.; Sheng, J. Characteristics of water erosion and conservation practice in arid regions of Central Asia: Xinjiang Province, China as an example. *Int. Soil Water Conserv. Res.* **2015**, *3*, 97–111. [CrossRef]