International Journal of Geo-Information

MDPI

Article

DEM- and GIS-Based Analysis of Soil Erosion Depth Using Machine Learning

Kieu Anh Nguyen and Walter Chen *

Department of Civil Engineering, National Taipei University of Technology, Taipei 10608, Taiwan; t106429401@ntut.edu.tw

* Correspondence: waltchen@ntut.edu.tw; Tel.: +886-2-27712171 (ext. 2628)

Abstract: Soil erosion is a form of land degradation. It is the process of moving surface soil with the action of external forces such as wind or water. Tillage also causes soil erosion. As outlined by the United Nations Sustainable Development Goal (UN SDG) #15, it is a global challenge to "combat desertification, and halt and reverse land degradation and halt biodiversity loss." In order to advance this goal, we studied and modeled the soil erosion depth of a typical watershed in Taiwan using 26 morphometric factors derived from a digital elevation model (DEM) and 10 environmental factors. Feature selection was performed using the Boruta algorithm to determine 15 factors with confirmed importance and one tentative factor. Then, machine learning models, including the random forest (RF) and gradient boosting machine (GBM), were used to create prediction models validated by erosion pin measurements. The results show that GBM, coupled with 15 important factors (confirmed), achieved the best result in the context of root mean square error (RMSE) and Nash–Sutcliffe efficiency (NSE). Finally, we present the maps of soil erosion depth using the two machine learning models. The maps are useful for conservation planning and mitigating future soil erosion.

Keywords: soil erosion; erosion pin; machine learning; morphometric factor; Shihmen Reservoir watershed

Citation: Nguyen, K.A.; Chen, W. DEM- and GIS-Based Analysis of Soil Erosion Depth Using Machine Learning. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 452. https://doi.org/10.3390/ijgi10070452

Academic Editor: Wolfgang Kainz

Received: 18 May 2021
Accepted: 29 June 2021
Published: 1 July 2021

1. Introduction

The United Nations General Assembly adopted 17 sustainable development goals (SDGs) in September 2015, which apply to all countries on the planet. Soil science is intertwined with a number of the SDGs. Among them, soils especially play an essential role in SDGs 2, 3, 6, 7, 12–15 [1].

Soil erosion is a form of land degradation and a severe threat to sustainable development. It is the process of moving surface soil with the action of external forces such as wind or water. Tillage also causes soil erosion. Among them, water erosion is the most tangible form of soil erosion in Taiwan. Soil erosion and sediment movement caused by rainfall and flooding, intense and persistent winds, agricultural activities, grazing, logging, mining, and construction result in significant damage to properties and potentially result in loss of lives, not to mention the livelihood support the land provides for communities. Therefore, it is a global challenge by 2030 to "combat desertification, and halt and reverse land degradation and halt biodiversity loss," as outlined by SDG 15.

Although the soil erosion process may seem to be slow at times, it dramatically impacts soil fertility, agriculture, and the ecosystem. Globally, it is estimated that the average soil erosion from agriculture is 75 billion tons/year ([2,3], as cited in [4]). Other scholars point out that about 85% of the 2 billion hectares of worldwide surface soil degradation stem from wind and water erosion ([5], based on [6,7]). The economic costs of erosion and sedimentation are substantial. For example, the cost of removing sediments alone could be somewhere between USD 7 and USD 68/yard^3 (or USD \$9.16–USD 88.94/m^3) in the US ([8], as cited in [9]). In Iran, the economic costs associated with soil erosion are thought to be around 10 trillion rials or USD 23,750,148 ([10], as cited in [11]). As a result, soil erosion modeling is critical to understanding soil erosion processes and preventing future soil loss.

Recently, the Global Applications of Soil Erosion Modeling Tracker (GASEMT) database was developed using peer-reviewed soil erosion modeling research literature published between 1994 and 2017 and is used to help the UN's global soil erosion assessment. The database contains the most up-to-date information on soil erosion modeling applications from around the world. In total, the GASEMT database contains 435 distinct soil erosion models and model variants. Despite the numerous models available for soil erosion modeling in the GASEMT database, statistics show that entries for watershed-scale applications are the most numerous (59%), and the (revised) universal soil loss equation ((R)USLE) family of models is the most commonly used soil erosion prediction models in the world, at about 41% [12]. Moreover, if USLE-based models such as the water and tillage erosion model/the sediment delivery model (WaTEM/SEDEM), erosion-productivity impact calculator (EPIC), and soil and water assessment tool (SWAT) are included in the same group, then this value could rise to 55%. Since the (R)USLE family is limited to sheet and rill erosion, the great majority of the model applications estimate only sheet and rill erosion amounts. Other types of erosion, notably stream bank erosion, gully erosion, and wind erosion, only account for 3.6% of the model applications combined. Finally, according to the bibliometric analysis based on the enhanced version of the GASEMT database, the (R)USLE model alone also has the largest number of total citations [13].

Understandably, the (R)USLE-family of models are also the most widely used soil erosion prediction models in Taiwan. For example, Chen et al. [14] applied the universal soil loss equation (USLE) model to reduce the order-of-magnitude discrepancy of soil loss estimates in the literature. Liu et al. [15] used two variants of the USLE model (grid cells and slope units) to calculate the soil loss due to sheet and rill erosion.

Beyond the traditional soil erosion models ((R)USLE, EPIC, SWAT, etc.), whether they are classified as empirical, conceptual, or process oriented, a growing alternative is to use machine learning (ML) or multicriteria decision making (MCDM) to improve the modeling ability of soil loss [11,16–19]. However, there are three significant limitations of these ML and ML-like studies. First, many of them only evaluate the presence or absence of soil erosion similar to landslide susceptibility modeling without considering the quantitative amount of soil loss [9,16–18]. Second, they tend to focus on or include gully erosion [16–19]. Third, and most distinctively, some of these studies use subjective evaluation, such as expert opinions, or results from other soil erosion models, as their validation [11,19], thus equivalently training models from subjective judgment and not from objective data such as field measurements.

Nguyen et al. [20] were the first to create machine learning models from field erosion pin measurements, a critical difference from other ML studies on soil erosion. The analysis was improved and expanded to different ML algorithms, including ensemble learning methods [21–23]. However, because some of the environmental factors used in the studies mentioned above were point data, the resulting models could not be directly applied to the entire study area (watershed) without interpolation.

The current study aims to improve the past studies by incorporating more independent variables (factors) derived from the watershed digital elevation model (DEM) and eliminate the dependence on the point data. The purpose is to create a comprehensive ML model that applies to the entire watershed.

2. Materials

Shihmen Reservoir watershed is located in northern Taiwan, which plays a crucial role in the metropolitan and irrigation areas of Taipei and Taoyuan [14]. It is also the third-largest reservoir in Taiwan. Typhoons bring the majority of the annual rainfall of 2350 mm to the Shihmen Reservoir watershed between May and October [24].

2.1. Environmental Factors and Erosion Pin Measurements

The 10 environmental factors (or parameters, or features, or variables, or attributes) examined in this study are main subwatershed, distance to river, distance to road, type

of slope, slope direction, rainfall amount, lithology, epoch, elevation, and slope class. Environmental factors were obtained from various GIS sources such as land use/land cover maps, geological maps, river maps, and road system maps. These factors and four additional factors (% sand, % silt, % clay, and % organic) were previously analyzed in Nguyen et al. [22]. However, the four additional factors were removed from this study because they were point data and could not be directly mapped to the entire study area (watershed). We used morphometric factors to replace the point data.

The erosion pin data used in this study came from field surveys conducted over three years (September 2008 to October 2011). The erosion pins were mounted on 55 slopes in 17 of the 93 subwatersheds of the study area (Figure 1). Each slope had 10 erosion pins mounted, and the average value of the 10 pins represents the slope's erosion depth [25]. The measurements of erosion pins were taken with a caliper, as shown in Figure 2.

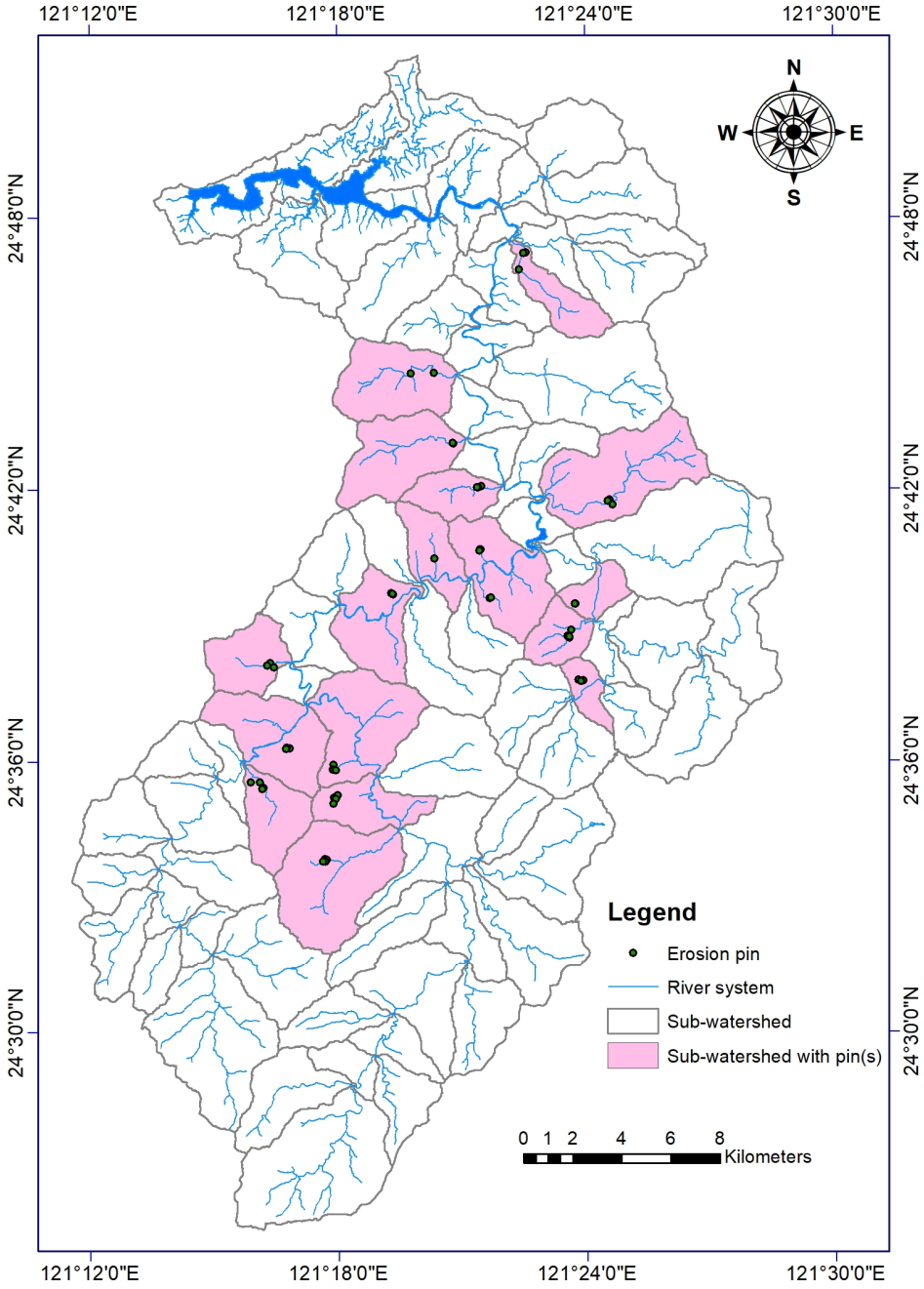


Figure 1. The study area of the Shihmen Reservoir watershed.

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

Figure 2. Illustration of field measurements in the Shihmen Reservoir watershed: (**a**) a natural slope, (**b**) an erosion pin and its label, and (**c**) measuring with a caliper ((**c**) from [22]).

2.2. Morphometric Factors

Morphometric analysis is the "quantitative description and analysis of landforms as practiced in geomorphology that may be applied to a particular kind of landform or to drainage basins and large regions generally" [26]. It is a technique for determining the scale and shape of watersheds, including two types of descriptive numbers: linear scale measurements and dimensionless numbers [27]. This approach can quantify the erosional growth of streams and their drainage watersheds, and compare geomorphic characteristics [28,29].

For this study, the Shihmen Reservoir watershed was divided into 93 subwatersheds to calculate the morphometric factors (or parameters, or features, or variables, or attributes) using the Central Geological Survey (CGS) DEM of Taiwan (10 m resolution) and ArcGIS 10.4.1. First, the DEM was filled in order to create flow paths and flow accumulations. Then, the stream networks were generated based on the flow accumulations of individual cells with a threshold value of 500. Finally, ArcGIS's Stream Link and Watershed functions were used to construct the subwatershed polygons. A total of 26 morphometric factors were calculated and described below (also see Table 1).

Subwatershed area (A) is the total area of a subwatershed. It ranged from 2.88 km^2 to 26.84 km^2 in this study. Research has indicated that total runoff or sediment yield is primarily determined by the subwatershed area [27].

Subwatershed perimeter (P) is the length of the boundary that surrounds a subwatershed. Its value varied between 10.70 and 37.29 km in the study area.

Stream order (U) indicates the complexity of a stream drainage system. The trunk river has the highest stream order and defines the order of a subwatershed [28]. An example of the stream order of a subwatershed is shown in Figure 3.

Number of streams (Nu) is the number of streams of a given stream order in a subwatershed. Figure 3 shows an example of the number of streams. The total number of streams (ΣN_u) is the summation of the number of streams of all orders.

Stream length (Lu) is the total channel length of a given stream order in this study for compatibility with the definition of the number of streams. It is not the cumulative channel length of a given order that includes all lesser orders, as sometimes defined [27]. The total stream length (ΣL_u) is the summation of the stream length of all orders.

Mean subwatershed slope (S) is the average slope of a subwatershed. It is calculated by the Slope function of ArcGIS and characterizes the steepness of a subwatershed.

Mean stream length (Lsm) is defined as the ratio between the stream length and the number of streams of a given stream order in a subwatershed in this study. We computed the average of the mean stream lengths as the characteristic mean stream length of the subwatershed.

Table 1. The morphometric factors used in this study.

	Morphometric Factor	Unit	Formula/Software
1	Subwatershed area (A)	km^2	ArcGIS (Calculate Geometry)
2	Subwatershed perimeter (P)	km	ArcGIS (Calculate Geometry)
3	Stream order (U)	-	ArcGIS (Calculate Geometry)
4	Total number of streams (ΣN_u)	-	ArcGIS (Stream Order)
5	Total stream length (ΣL_u)	km	ArcGIS (Calculate Geometry)
6	Mean subwatershed slope (S)	Degree	ArcGIS (Slope)
7	Mean stream length (Lsm)	km	$Lsm = avg(Lu/Nu)$
8	Subwatershed length (Lb)	km	ArcGIS
9	Stream frequency (Fs)	$1/km^2$	$Fs = \Sigma Nu/A$
10	Drainage density (Dd)	$1/km$	$Dd = \Sigma Lu/A$
11	Constant of channel maintenance (C)	km	$C = 1/Dd$
12	Length of overland flow (Lo)	km	$Lo = 1/(2Dd)$
13	Infiltration number (If)	$1/km^3$	$If = Fs \times Dd$
14	Subwatershed relief (H)	km	$H = h_{max} - h_{min}$
15	Relief ratio (R)	-	$R = H/Lb$
16	Melton index (M)	-	$M = H/\sqrt{A}$
17	Ruggedness number (Rn)	-	$Rn = Dd \times H$
18	Bifurcation ratio (Rb)	-	$Rb = avg(N_u/N_{u+1})$
19	Stream length ratio (Rl)	-	$Rl = avg((L_{u+1}/N_{u+1})/(L_u/N_u))$
20	Ratio Rho (ρ)	-	$\rho = Rl/Rb$
21	Elongation ratio (Re)	-	$Re = (2\sqrt{A/\pi})/Lb$
22	Circularity ratio (Rc)	-	$Rc = 2\sqrt{\pi A}/P$
23	Form factor (Ff)	-	$Ff = A/Lb^2$
24	Shape factor (Bs)	-	$Bs = Lb^2/A$
25	Compactness coefficient (Cc)	-	$Cc = P2\sqrt{\pi A}$
26	Texture ratio (T)	$1/km$	$T = \Sigma Nu/P$

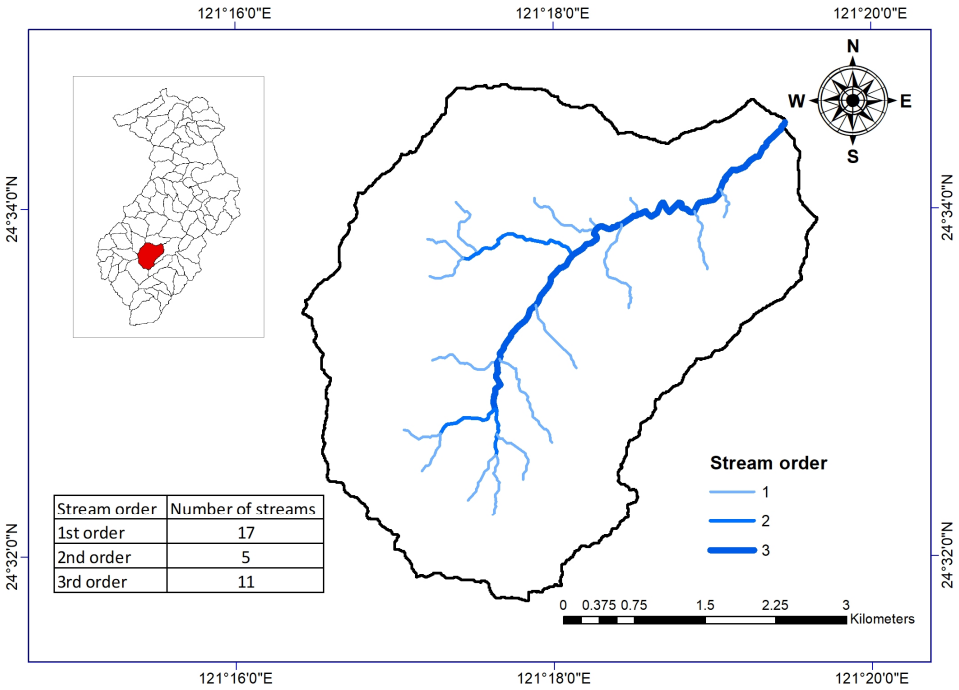


Figure 3. Stream order and the number of streams in a typical subwatershed.

Subwatershed length (Lb) in this study is defined as "the longest dimension of the basin parallel to the principal drainage line," as in the definition of relief ratio below [29]. The length is determined by ArcGIS 10.4.1.

Stream frequency (Fs) is the number of streams per unit area [28]. This value ranged from 0.47 to 2.46 streams/km^2 in this study.

Drainage density (Dd) is defined as the sum of the stream lengths divided by the subwatershed area. It is a crucial indicator of the linear scale of landform elements in a subwatershed [27].

Constant of channel maintenance (C) is defined as the inverse of drainage density. Along with drainage density, this value compares soil's erodibility or other factors influencing surface erosion [29]. Here, metric units were used, and the conversion factor of 5280 (from miles to feet) was ignored.

Length of overland flow (Lo) ranged from 0.32 km to 0.64 km in the study area. It is the length of runoff over the ground surface until it concentrates in definite stream channels and is half the reciprocal of drainage density [28].

Infiltration number (If) is the product of stream frequency and drainage density ([30], as cited in [31]). This value ranged from 0.44 to 2.98 in this study.

Subwatershed relief (H) is the difference in elevations between the lowest (h_{min}) and the highest (h_{max}) points in a subwatershed.

Relief ratio (R) is "the ratio between the total relief of a basin" and "the longest dimension of the basin parallel to the principal drainage line" [29]. For the study area, the relief ratio varied from 0.07 to 0.57.

Melton index (M), or the ruggedness of a subwatershed, is characterized by the dimensionless ratio between the subwatershed relief and the square root of the subwatershed area [32].

Ruggedness number (Rn) is known as the dimensionless product of drainage density and relief. As a result, high drainage density and low relief areas are just as rugged as low drainage density and high relief areas ([33], as cited in [34]).

Bifurcation ratio (Rb) is the average number of branchings or bifurcations of streams. It is defined as the number of streams of a given stream order to that of streams of the next higher order [28]. For a subwatershed, there are different bifurcation ratios for different stream orders. Following the example of Jothimani et al. [35], we computed the average of the bifurcation ratios as the characteristic bifurcation ratio of the subwatershed. For the 93 subwatersheds in the study area, the bifurcation ratio ranged from 0.50 to 8.00.

Stream length ratio (Rl) is defined by the average length of streams of a stream order to the next lower order [28]. Various stream length ratios exist for various stream orders. Therefore, we computed the average of the stream length ratios as the characteristic stream length ratio of the subwatershed, similar to Jothimani et al. [35]. For the 93 subwatersheds in the study area, the stream length ratio ranged from 0.46 to 5.86.

Ratio Rho (ρ) is the stream length ratio divided by the bifurcation ratio [28].

Elongation ratio (Re) is the ratio between the diameter of a circle with the same area as the subwatershed and the longest dimension of the subwatershed parallel to the main drainage line [29], as determined for the relief ratio.

Circularity ratio (Rc) is the circumference of a circle with the same area as the subwatershed divided by the subwatershed perimeter [29].

Form factor (Ff) is the ratio of the width to the length of a subwatershed and is defined as the subwatershed area divided by the square of the length of the subwatershed [36]. The subwatershed length is "measured from a point on the watershed-line opposite the head of the main stream" [36]. Here, we used subwatershed length (Lb) to be the length of the subwatershed.

Shape factor (Bs) is defined as the square of the length of a subwatershed divided by the area of the subwatershed, although other definitions have also been proposed ([37], as cited in [38]). The length of a subwatershed is defined as "the longest dimension from the mouth to the opposite side." Here, we used the subwatershed length (Lb) to represent the length of the subwatershed.

Compactness or compactness coefficient (Cc) is the ratio of the perimeter of the subwatershed to that of a circle with an equal area [36].

Texture ratio is the ratio of the number of crenulations on the contour with the maximum number of crenulations within the subwatershed to the length of the perimeter of the subwatershed [39]. Crenulations are chosen because they indicate streams too small to be shown on a topographic map [27]. The ratio is a measure of channel spacing closeness and thus is related to drainage density. For ease of computation, we used the total number of streams to replace the crenulations in this study. The texture ratio ranged from 0.16 to 1.27.

3. Methods

This study had five objectives: first, to identify and collect morphometric factors and environmental factors that affect soil erosion; second, to use feature selection to identify critical factors that can be used to model soil erosion depths; third, to apply machine learning algorithms to create models that can be used to predict soil erosion depth in the study area; fourth, to assess the validity of the models using statistical indices and threefold cross-validation; fifth and finally, to produce prediction maps of soil erosion depth for the study area.

3.1. Research Framework

Figure 4 depicts the five research steps of this analysis. First, we created an input dataset of 36 independent factors by combining 26 morphometric and 10 environmental factors. Second, we divided the dataset into three folds of roughly the same size based on the main subwatershed attribute to balance the class distribution from the five main subwatersheds [40]. We also used the erosion pin measurement as the target variable. Each time one of the three folds was held as the test data for testing the model, the remaining two folds were used as the training data. The whole process was repeated three times. Third, we applied the random forest (RF) and gradient boosting machine (GBM) to create erosion models based on the training data. Fourth, we assessed the models with the test data. In the process, we eliminated the unessential factors and kept the best models. Finally, we created the spatial distribution maps of soil erosion depth of the study area using the machine learning models.

3.2. Feature Selection

In order to identify the key factors that will generate the most credible soil erosion models, we used feature selection to rank the 36 morphometric and environmental factors in the study. Specifically, the Boruta algorithm was used to select the subsets of factors (predictors) for ML model building.

Boruta is a feature selection algorithm and feature ranking tool based on the RF algorithm and introduced by Kursa et al. [41]. It works by creating a randomized copy of the input dataset, merging it with the original dataset, and constructing the expanded system's classifier. Then, Boruta compares the importance of the factors in the original dataset to those of the randomized factors to identify the key factors. Only factors with greater importance than the randomized factors are considered essential. The advantage of Boruta is that it allows researchers to choose the most significant factors that influence the outcome. For this study, the Boruta package in the R software was used, and the maximum number of times the algorithm was run (maxRun) was set to the default value of 1000.

3.3. Machine Learning Models

In this analysis, two machine learning methods were used. They are the random forest and the gradient boosting machine.

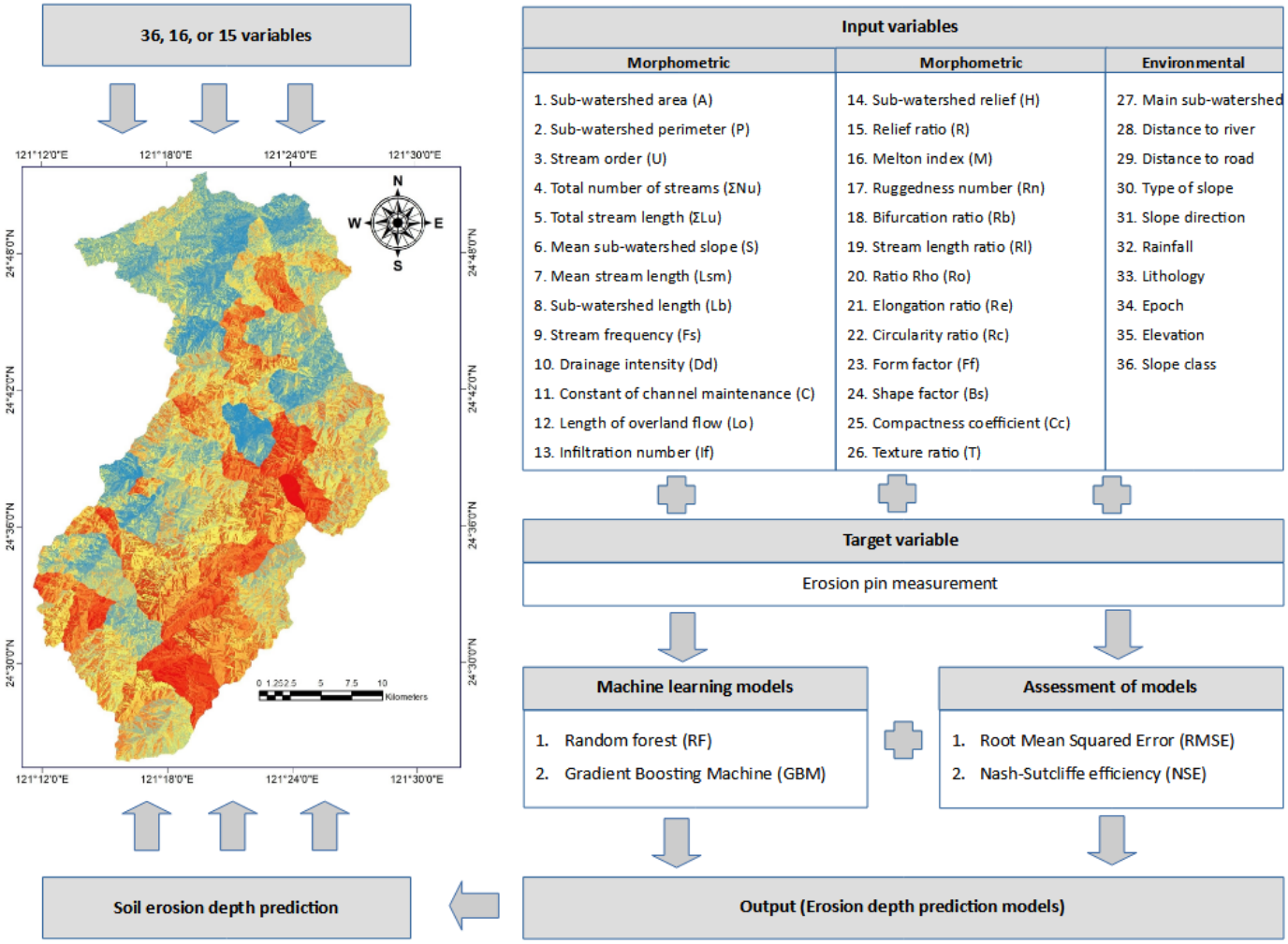


Figure 4. Research flowchart of this study.

Random forest (RF) was proposed by Breiman [42]. It is a supervised ML method that combines all tree-based results into the most appropriate model for the application. The RF algorithm runs many iterations and divides the training dataset (in terms of data and attributes) into many subsets at random to create many trees and produce better results than individual decision trees. The randomForest() package in the R software was used to implement random forest in this analysis, which uses the Gini index to separate data in order to minimize impurity at each node. Tsai et al. [23] provided a more detailed overview of the Gini index and random forest.

Friedman [43] proposed the gradient boosting machine as a simple and highly flexible machine learning tool. It is a widely used machine learning algorithm that has been shown to be effective in a variety of applications [44–46]. The basic idea behind GBM is to build a prediction model using a set of poor learning algorithms, most commonly decision trees. Unlike RF, which produces an ensemble of individual trees in parallel, GBM creates a sequenced tree ensemble. The knowledge gained from previously grown trees is used to grow new trees in a sequential manner. The GBM model was once used to model soil erosion [22]. It was implemented in this study using R software's "gbm" package.

3.4. Assessment of Models

In this study, the ML model performance was evaluated using two statistical indices. As shown in Equations (1) and (2), they are the root mean square error (RMSE) and the Nash–Sutcliffe efficiency (NSE).

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

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

where P is the predicted value, O is the observed value, and $\overline{O}$ is the mean observed value.

RMSE was used to compare the difference between the expected values (model outputs) and the observed values (erosion pin measurements) in the two indices, while NSE was used to determine the effectiveness of the model against the average observed value [20–22].

4. Results

In this analysis, we used R version 4.0.5. In order to assess soil erosion in the Shihmen Reservoir watershed, this study employed two machine learning models, RF and GBM. To substitute four factors that were only point data, 26 morphometric factors were added to the original dataset of 14 environmental factors. In total, 36 variables were examined for their relationship with soil erosion depth (erosion pin measurement). The training data (used to create the ML models) made up two folds of the dataset, while the remaining fold was used to evaluate the models based on RMSE and NSE. Finally, through spatial mapping, machine learning models were used to predict the soil erosion depth for the entire Shihmen Reservoir watershed.

4.1. Feature Selection

Boruta was used as a feature selection tool to assess the relative importance of variables that influence soil erosion. Table 2 and Figure 5 depict the findings. It can be seen that Table 2 was divided into three categories based on decisions: rejected, tentative, and confirmed. They are also ranked by median importance. In total, 15 factors were identified as important, which includes texture ratio, subwatershed length, epoch, elongation ratio, lithology, subwatershed perimeter, form factor, relief ratio, total stream length, Melton index, the total number of streams, elevation, shape factor, subwatershed area, and type of slope. One factor was considered tentative, i.e., the main subwatershed. Moreover, 20 variables were ruled out, which consist of distance to river, mean stream length, ruggedness number, slope direction, ratio Rho, circularity ratio, distance to road, stream length ratio, stream frequency, rainfall, compactness coefficient, stream order, constant of channel maintenance, drainage density, length of overland flow, infiltration number, slope class, subwatershed slope, bifurcation ratio, and subwatershed relief. They should play no important role in the prediction of soil erosion. According to the Boruta analysis, the type of slope, subwatershed area, and shape factor are the three most significant variables among the factors that are shown to be important.

Boruta generates a corresponding "shadow" factor for each factor, whose values were obtained by shuffling the original factor's values across objects. The system then classifies these using all of the extended system's factors and calculates the importance of each factor [47]. Green is used to color the 15 factors listed as important in Figure 5. The 20 rejected factors are colored red, while the one tentative factor is colored yellow. To differentiate the variables, Figure 5 also shows the minimum, mean, and maximum of shadow factors. In general, factors ranked higher than the shadow maximum have been tested to be more significant than chance.

Table 2. Variable importance using Boruta feature selection.

	meanImp	medianImp	minImp	maxImp	Decision
Type of slope	8.162273	8.243055	3.952560	10.425924	Confirmed
Subwatershed area	5.019874	5.052494	2.343560	7.383705	Confirmed
Shape factor	4.976153	5.010055	2.159306	7.532354	Confirmed
Elevation	4.988711	5.000990	2.069170	7.789953	Confirmed
Total number of streams	4.907324	4.938526	2.276170	7.150461	Confirmed
Melton index	4.409082	4.469957	1.627233	7.206727	Confirmed
Total stream length	4.447973	4.438060	1.872146	6.785887	Confirmed
Relief ratio	4.348566	4.412615	0.893647	6.730493	Confirmed
Form factor	4.209047	4.230464	1.780136	6.776284	Confirmed
Subwatershed perimeter	4.064281	4.095632	0.925774	6.298489	Confirmed
Lithology	3.898596	3.915257	0.934244	6.630224	Confirmed
Elongation ratio	3.788302	3.794716	1.129358	5.898901	Confirmed
Epoch	3.517687	3.528155	0.622912	6.683233	Confirmed
Subwatershed length	3.421194	3.444065	0.453943	6.157326	Confirmed
Texture ratio	3.420789	3.421103	0.778195	5.719445	Confirmed
Subwatershed	2.952157	2.983119	−0.229199	5.487332	Tentative
Subwatershed relief	2.355625	2.315015	−0.156997	4.560039	Rejected
Bifurcation ratio	2.122521	2.217065	0.894857	3.185164	Rejected
Mean subwatershed slope	1.990054	2.069925	0.119557	3.807072	Rejected
Slope class	1.999188	2.035054	−0.586225	4.019329	Rejected
Infiltration number	1.761100	1.736976	−0.027713	3.775921	Rejected
Length of overland flow	1.757534	1.690870	−0.044952	3.617237	Rejected
Drainage density	1.674382	1.671754	0.212803	2.613970	Rejected
Constant of channel maintenance	1.412327	1.563865	−1.125096	2.720770	Rejected
Stream order	1.460165	1.537092	−0.199118	2.561430	Rejected
Compactness coefficient	1.492243	1.527261	−0.460479	2.733053	Rejected
Rainfall	1.601451	1.490486	0.124088	3.313065	Rejected
Stream frequency	1.359127	1.464677	−0.542830	2.702161	Rejected
Stream length ratio	1.090233	1.124972	−1.164121	2.851477	Rejected
Distance to road	0.998904	0.815334	−0.609436	2.881991	Rejected
Circularity ratio	0.975395	0.711048	−0.379865	2.311318	Rejected
Ratio Rho	0.518910	0.549485	−0.944962	2.122957	Rejected
Slope direction	0.692622	0.536967	−0.820490	2.884595	Rejected
Ruggedness number	0.183690	0.291990	−2.038411	1.565565	Rejected
Mean stream length	0.486127	0.288427	−1.123192	1.989272	Rejected
Distance to river	−0.937049	−0.925128	−2.557384	0.647234	Rejected

Among the green (important) factors, 4 are environmental factors, while the remaining 11 are morphometric factors. The percentage of the environmental factors in the confirmed group (4/15 = 27%) is slightly less than the overall percentage of the environmental factors in the dataset (10/36 = 28%). On the other hand, the environmental factors account for 100% of the tentative factor (1/1) and 25% (5/20) of the rejected factors. Furthermore, the environmental factors selected in the confirmed group are the type of slope, elevation, lithology, and epoch. Compared to the study by Nguyen et al. [22], which also reported the relative importance of environmental factors, we can see some similarities. The top four factors from Nguyen et al. [22] were slope direction, type of slope, % organic, and elevation. Two (type of slope and elevation) were also selected for this study, while one (slope direction) was not, and the other (% organic) was not included in this study because it is a point data. It is worth noting that Nguyen et al. [22] used 70% training and 30% test data, while this study used threefold cross-validation.

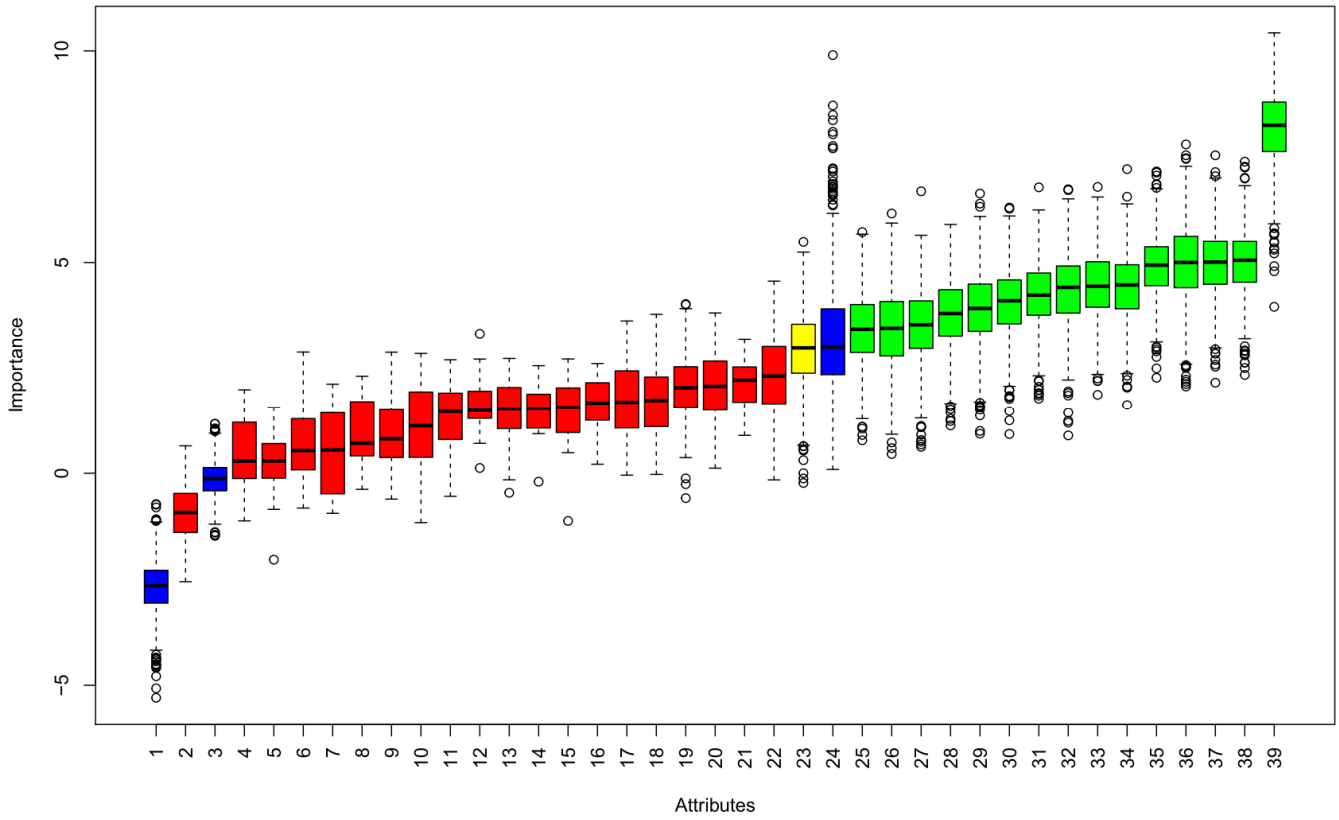


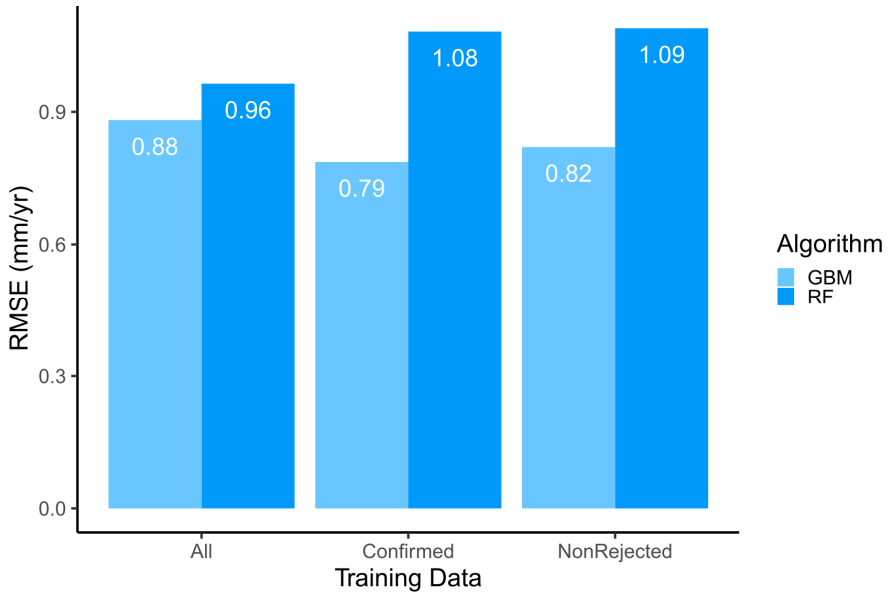
Figure 5. Feature selection with Boruta: 1: shadow min, 2: distance to the river, 3: shadow mean, 4: mean stream length, 5: ruggedness number, 6: slope direction, 7: ratio Rho, 8: circularity ratio, 9: distance to road, 10: stream length ratio, 11: stream frequency, 12: rainfall, 13: compactness coefficient, 14: stream order, 15: constant of channel maintenance, 16: drainage density, 17: length of overland flow, 18: infiltration number, 19: slope class, 20: subwatershed slope, 21: bifurcation ratio, 22: subwatershed relief, 23: main subwatershed, 24: shadow max, 25: texture ratio, 26: subwatershed length, 27: epoch, 28: elongation ratio, 29: lithology, 30: subwatershed perimeter, 31: form factor, 32: relief ratio, 33: total stream length, 34: Melton index, 35: total number of streams, 36: elevation, 37: shape factor, 38: subwatershed area, 39: type of slope.

4.2. Machine Learning

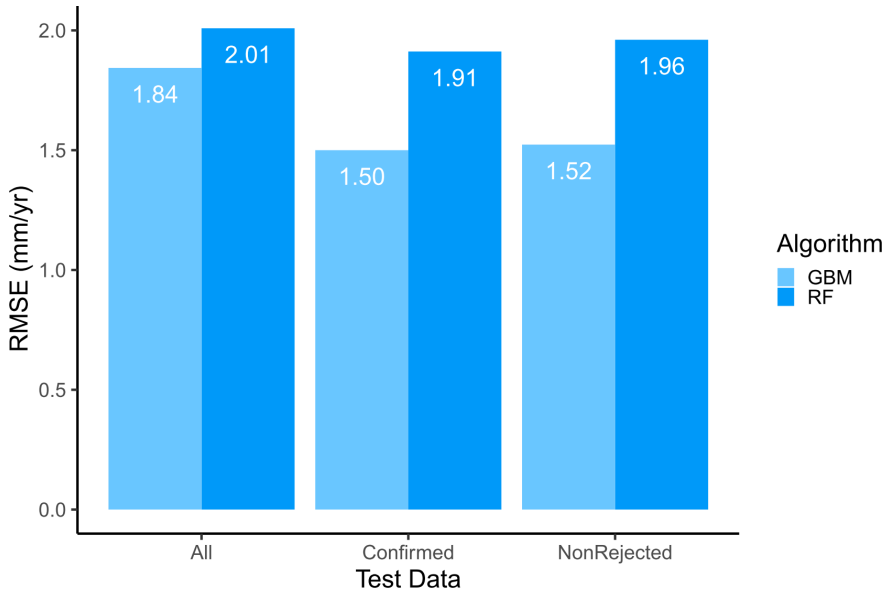
Based on the results of feature selection, we performed machine learning on three sets of factors separately: (1) all 36 factors, (2) 15 confirmed factors, and (3) 15 confirmed factors plus 1 tentative factor. Using threefold cross-validation in each set of factors, the dataset was divided into three, roughly equal folds. Then, two folds were used as the training data, and the other fold was used as the test data. The process was repeated three times so that every fold was used as the test data in the analysis. Both RF and GBM were used to analyze the same data. Finally, the results (RMSE and NSE) of three attempts were averaged. They are shown in Table 3 and Figure 6.

Table 3. Performance comparison of machine learning models using threefold cross-validation.

Model and Factors	No. of Factors	Average RMSE (mm/yr)		Average NSE	
		Training	Test	Training	Test
RF (all)	36	0.96	2.01	0.83	0.25
GBM (all)	36	0.88	1.84	0.84	0.39
RF (confirmed)	15	1.08	1.91	0.79	0.31
GBM (confirmed)	15	0.79	1.50	0.88	0.59
RF (nonrejected)	16	1.09	1.96	0.79	0.27
GBM (nonrejected)	16	0.82	1.52	0.87	0.57



(**a**)



(**b**)

Figure 6. Comparison between parameter selections: (**a**) training data and (**b**) test data (lower is better).

The findings (Table 3) reveal that the ML models delivered good results. Both the average values of RMSE and NSE in Table 3 exhibit the same trend. The smaller the RMSE and the higher the NSE were, the better the model was. As shown in Figure 6, GBM consistently outperforms RF in both training data and test data. GBM also edges out RF in all three datasets that used different factors (all, confirmed, and nonrejected). For the training data, the best RF model result was obtained with the all-factor group, followed by the confirmed group and then the nonrejected group. However, for the test data, the confirmed group is the best, followed by the nonrejected group and then the all-factor group. This shows that the RF models were overfitted with more factors, and that feature selection indeed contributes to improving the ML models when facing unknown data.

On the other hand, the GBM model does not exhibit an overfit bias. For both the training and test data, the confirmed group is the best, followed by the nonrejected group and then the all-factor group.

Overall, the best test result obtained in this study is 1.50 mm/yr (GBM) and 1.91 mm/yr (RF). Both of them are from the confirmed group. Compared to the previous study [22], which used a 70/30 split and only 14 environmental factors, the results are mixed. In terms of RF, the Nguyen et al. [22] result was 1.75 mm/yr, which is better than the current study (1.91 mm/yr). However, in terms of GBM, the present study (1.50 mm/yr) is better than the previous study (1.72 mm/yr). If we only consider the best model, which is GBM in this case, this study is better than the previous study.

4.3. Model Prediction

Using the RF and GBM models, we predicted the soil erosion depth of the entire study area, as shown in Figure 7. The data of the whole Shihmen Reservoir watershed were investigated and then entered into the R software after the preparation of the machine learning models for predicting the soil erosion depth. The results were transferred to the ArcGIS software to create the soil erosion depth maps. Figure 7 showed the spatial distribution of soil erosion depth (in mm/yr) over the Shihmen Reservoir watershed produced by each model's three sets of factors: all, confirmed, and nonrejected. The red area represents a high erosion depth, whereas the blue area has a low erosion depth. Due to the morphometric factors used in the ML models, it is clear that the individual subwatershed has a significant impact on the soil erosion depth distribution.

Figure 7 shows that the all-factor group's maps (a and b) have more variance within individual subwatersheds than the confirmed group's (c and d) and the nonrejected group's maps (e and f). This is most likely due to the fact that there are more variables used in the mapping of all factors (36). The confirmed and nonrejected maps, on the other hand, appear to be more uniform in color throughout each subwatershed. They both have a similar appearance because they used a similar number of variables (15 and 16).

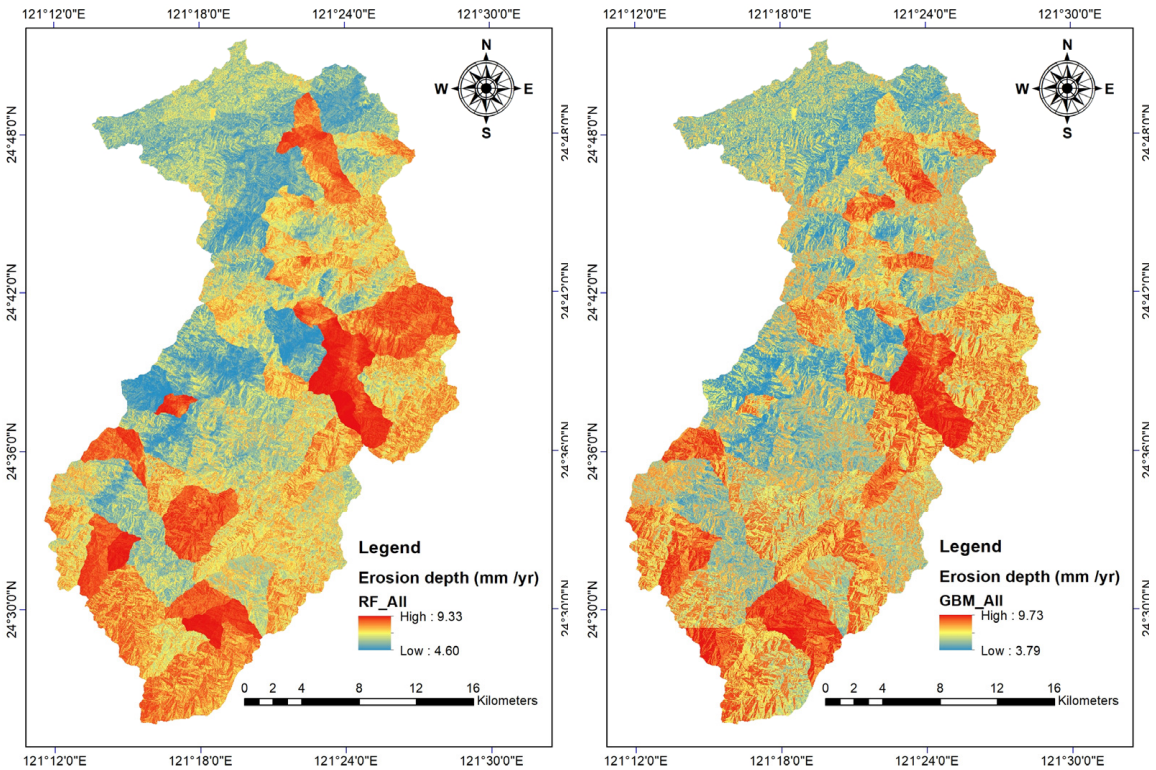


Figure 7. *Cont.*

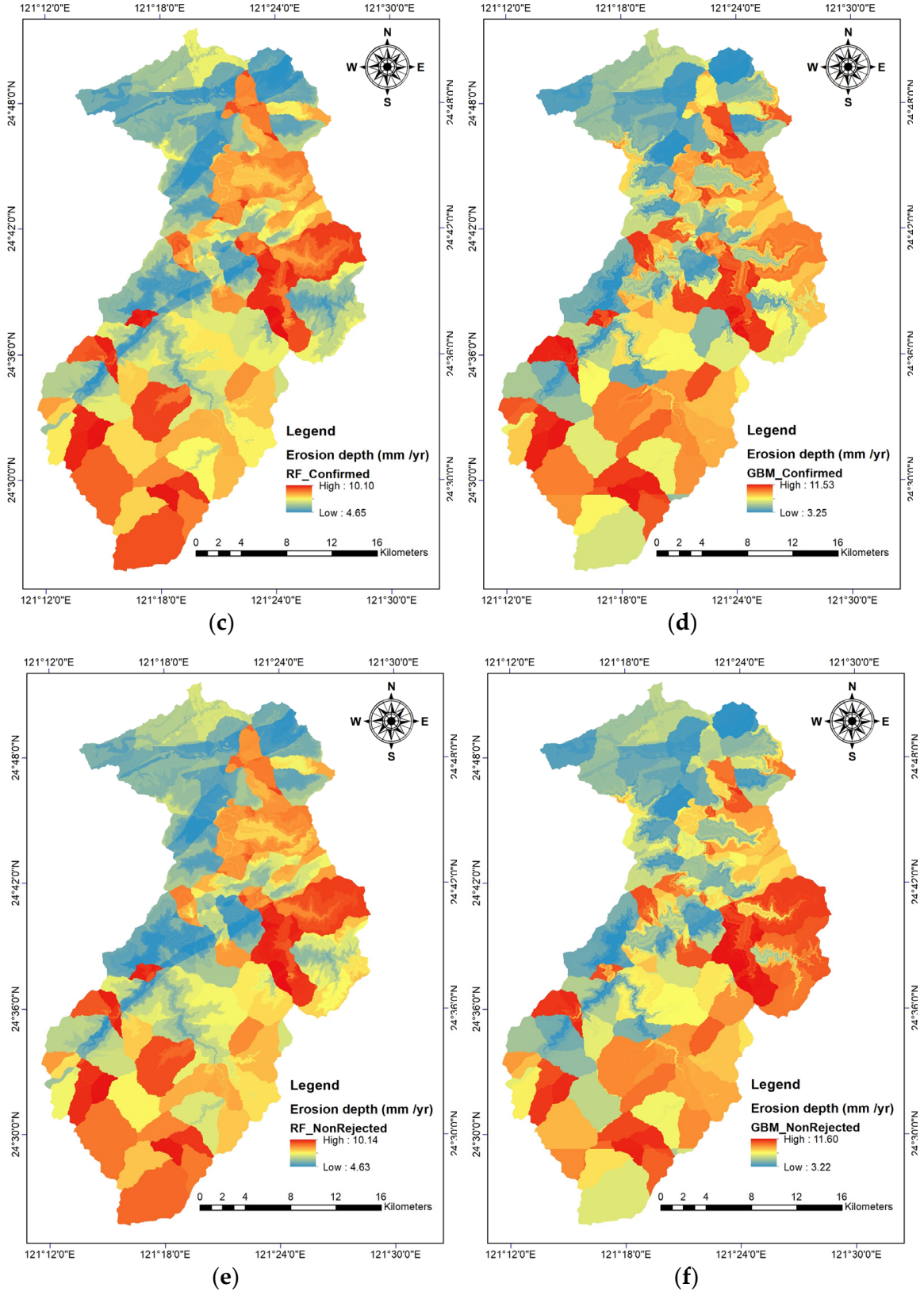


Figure 7. Prediction of soil erosion depth of the entire Shihmen Reservoir watershed using machine learning: (**a**) RF (all), (**b**) GBM (all), (**c**) RF (confirmed), (**d**) GBM (confirmed), (**e**) RF (nonrejected), and (**f**) GBM (nonrejected).

The minimum, mean, and maximum erosion depths expected for the entire Shihmen Reservoir watershed from different model/factor combinations are shown in Table 4. The

table also includes field measurements of erosion pins for comparison. The table shows that the averages of various model/factor combinations are quite similar to the average of erosion pins. However, no model/factor combination accurately forecasts the extreme values of real-world measurements. The predictions are too high for the minimum value and too low for the maximum value.

Table 4. Comparing ML model results with erosion pin measurements.

Erosion Depth (mm/yr)	Min (mm/yr)	Mean (mm/yr)	Max (mm/yr)
RF (all)	4.60	6.77	9.33
GBM (all)	3.79	6.73	9.73
RF (confirmed)	4.65	6.68	10.10
GBM (confirmed)	3.25	6.68	11.53
RF (nonrejected)	4.63	6.67	10.14
GBM (nonrejected)	3.22	6.67	11.60
Erosion Pin measurements	2.17	6.50	13.03

5. Discussion

This study continues to model the soil erosion depth as measured by erosion pins in the Shihmen Reservoir watershed because of the watershed's significance and the degree to which it is affected by soil erosion [20–22]. Since the morphometric features of a watershed influence surface runoff and water erosion, they were included in this research to create a complete picture of the erosion activity in the study region and to improve the ML models. However, due to the overlapping (and sometimes conflicting) nature of some of the morphometric features, the overwhelming number of factors extracted from the morphometric analysis may be a deterrent to further analysis. As a result, feature selection was performed in this study before machine learning modeling. The widely used Boruta algorithm was used to separate the important from the nonimportant factors. In the end, 11 morphometric factors were identified as influential in estimating the soil erosion depth. They are texture ratio, subwatershed length, elongation ratio, subwatershed perimeter, form factor, relief ratio, total stream length, Melton index, total number of streams, shape factor, and subwatershed area. Overall, the morphometric factors were chosen in 42 percent (=11/26) of the cases. On the other hand, only four environmental factors (slope type, elevation, lithology, and epoch) were chosen as important. They account for 40% (=4/10) of the overall environmental factors.

Note that the point data (% sand, % silt, % clay, and % organic) in the original 14 environmental factors had to be removed because they were not available watershed-wide and cannot be used for model prediction of the entire Shihmen Reservoir watershed. Therefore, the lower selection rate of the environmental factors than the morphometric factors in this study could be attributed to the removal of these point data because some of them were shown to be important in the previous study [22].

Another aspect that distinguishes this study from the previous studies [20–22] is the use of threefold cross-validation instead of the 70/30 split with stratified random sampling. The threefold cross-validation divides the dataset into three roughly equal folds with a balanced class distribution. Therefore, each class (stratum) is adequately represented, as with the stratified random sampling. However, in the threefold cross-validation, two folds were used as the training data, and the third fold was used as the test data. The procedure was replicated three times so that the algorithm takes turns using two-thirds of the data as the training data, and each fold was used as the test data only once. The 70/30 split with stratified random sampling, on the other hand, did not rotate the training and test data. To find the average answer, the 70/30 split had to be repeated three times from the beginning using different random seeds.

Regardless of which set of factors was used (all, confirmed, or nonrejected), our analysis shows that GBM consistently outperforms RF in terms of RMSE and NSE. As compared to the previous study [22], the best RMSE value was noticeably reduced from

1.72 mm/yr to 1.50 mm/yr (GBM with confirmed factors). This demonstrates that, despite the elimination of potentially valuable point data, the inclusion of morphometric factors improves the soil erosion modeling.

Additionally, unlike the previous study that used point data [22], this study does not need to interpolate the modeling prediction for the entire research area. Instead, complete maps of the spatial distribution of soil erosion depth can be produced from the ML models directly. The resulting maps (Figure 7) show finer resolution of change with more features in color variation. There is densely packed information not present in the previous maps. It is a huge step forward for soil erosion control and prioritization.

6. Conclusions

To sum up, previous studies built machine learning models for the Shihmen Reservoir watershed using point data that were only available at individual slopes monitored with erosion pins. The current research improved upon past studies by incorporating new independent variables (morphometric factors) derived from the watershed digital elevation model and eliminating the dependence on the point data. A dataset of 36 predictive factors and one target factor was created. Feature selection was performed to remove redundant factors and to avoid the overfitting of models. In the end, 15 important factors were identified that include 4 environmental factors and 11 morphometric factors. Two ML algorithms, RF and GBM, were used in the analysis. Despite the removal of four environmental factors used in previous studies (point data that were not available watershed-wide), the new GBM model in this study shows an improvement in RMSE, which was reduced from 1.72 mm/yr to 1.50 mm/yr. Consequently, we were able to create the most accurate ML model to date of the distribution of soil erosion depth in the study area. This proves the value of adding morphometric factors to soil erosion analysis. Furthermore, the ML models were used to create prediction maps of soil erosion depth of the entire Shihmen Reservoir watershed, which were not possible, and only interpolation approximation was achieved previously (due to the point data issue). The new maps show great details of what needs attention for soil erosion control and prioritization. It is a valuable advancement of our understanding and future study of soil erosion modeling. Since the ML models are data-driven and rely on sufficient monitoring data, it is crucial to improve our data collection methods and use the latest technologies to record information. Solar-powered Internet of Things (IoT) devices that can monitor the change of slope surfaces are currently being experimented with in the Shihmen Reservoir watershed. The inexpensive and large amount of data generated by these devices will likely be the key driver for future research on this topic.

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

Funding: This study was partially supported by 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. Keesstra, S.D.; Bouma, J.; Wallinga, J.; Tittonell, P.; Smith, P.; Cerdà, A.; Montanarella, L.; Quinton, J.N.; Pachepsky, Y.; Van Der Putten, W.H.; et al. The significance of soils and soil science towards realization of the United Nations sustainable development goals. *Soil* **2016**, *2*, 111–128. [CrossRef]
2. Eswaran, H.; Lal, R.; Reich, P.F. Land Degradation: An overview. In *Response to Land Degradation, Proceedings of the 2nd International Conference on Land Degradation and Desertification, Khon Kaen, Thailand, 25–29 January 1999*; Bridges, E.M., Hannam, I.D., Oldeman, L.R., Pening de Vries, F.W.T., Scherr, S.J., Sompatpanit, S., Eds.; Oxford University Press: New Delhi, India, 2002. Available online: http://soils.usda.gov/use/worldsoils/papers/land-degradation-overview.html (accessed on 1 August 2013).
3. Myers, N. *Gaia: An Atlas of Planet Management*; Anchor/DoubleDay: Garden City, NY, USA, 1993.
4. Pimentel, D.; Burgess, M. Soil erosion threatens food production. *Agriculture* **2013**, *3*, 443–463. [CrossRef]
5. Weil, R.R.; Brady, N.C. *The Nature and Properties of Soils*, 15th ed.; Pearson: Harlow, UK, 2017; 1104p.
6. Oldeman, L.R. The global extent of soil degradation. In *Soil Resilience and Sustainable Land Use*; Greenland, D.J., Szabolcs, I., Eds.; CAB International: Wallingford, UK, 1994; pp. 99–118.
7. Daily, G. Restoring value to the world's degraded lands. *Science* **1997**, *269*, 350–354. [CrossRef]
8. U.S. Environmental Protection Agency. *Comparative Costs of Erosion and Sediment Control Construction Activities*; EPA-430/9-73-016; U.S. Government Printing Office: Washington, DC, USA, 1973.
9. Gray, D.H.; Sotir, R.B. *Biotechnical and Soil Bioengineering Slope Stabilization: A Practical Guide for Erosion Control*; John Wiley & Sons: New York, NY, USA, 1996; 378p.
10. National Geosciences Database. 2017. Available online: www.ngdir.ir (accessed on 12 February 2019).
11. Arabameri, A.; Tiefenbacher, J.P.; Blaschke, T.; Pradhan, B.; Bui, D.T. Morphometric analysis for soil erosion susceptibility mapping using novel gis-based ensemble model. *Remote Sens.* **2020**, *12*, 874. [CrossRef]
12. 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. *Sci. Total Environ.* **2021**, *780*, 146494. [CrossRef] [PubMed]
13. Bezak, N.; Mikoš, M.; Borrelli, P.; Alewell, C.; Alvarez, P.; Ayach Anache, J.A.; Baartman, J.; Ballabio, C.; Biddoccu, M.; Cerdà, A.; et al. Soil erosion modelling: A bibliometric analysis. *Environ. Res.* **2021**, *197*, 111087. [CrossRef]
14. 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]
15. 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]
16. Angileri, S.E.; Conoscenti, C.; Hochschild, V.; Märker, M.; Rotigliano, E.; Agnesi, V. Water erosion susceptibility mapping by applying Stochastic Gradient Treeboost to the Imera Meridionale River Basin (Sicily, Italy). *Geomorphology* **2016**, *262*, 61–76. [CrossRef]
17. Chakrabortty, R.; Pal, S.C.; Sahana, M.; Mondal, A.; Dou, J.; Pham, B.T.; Yunus, A.P. Soil erosion potential hotspot zone identification using machine learning and statistical approaches in eastern India. *Nat. Hazards* **2020**, *104*, 1259–1294. [CrossRef]
18. Pourghasemi, H.R.; Yousefi, S.; Kornejady, A.; Cerdà, A. Performance assessment of individual and ensemble data-mining techniques for gully erosion modeling. *Sci. Total Environ.* **2017**, *609*, 764–775. [CrossRef]
19. Svoray, T.; Michailov, E.; Cohen, A.; Rokah, L.; Sturm, A. Predicting gully initiation: Comparing data mining techniques, analytical hierarchy processes and the topographic threshold. *Earth Surf. Process. Landf.* **2012**, *37*, 607–619. [CrossRef]
20. 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]
21. 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]
22. 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. [CrossRef]
23. Tsai, F.; Lai, J.-S.; Nguyen, K.A.; Chen, W. Determining Cover Management Factor with Remote Sensing and Spatial Analysis for Improving Long-Term Soil Loss Estimation in Watersheds. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 19. [CrossRef]
24. Huang, C.L.; Hsu, N.S.; Wei, C.C. Coupled Heuristic Prediction of Long Lead-Time Accumulated Total Inflow of a Reservoir during Typhoons Using Deterministic Recurrent and Fuzzy Inference-Based Neural Network. *Water* **2015**, *7*, 6516–6550. [CrossRef]
25. 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]
26. Encyclopaedia Britannica. Morphometric Analysis. 2009. Available online: https://www.britannica.com/science/morphometric-analysis (accessed on 4 May 2021).
27. Strahler, A.N. Quantitative analysis of watershed geomorphology. *Eos Trans. Am. Geophys. Union* **1957**, *38*, 913–920. [CrossRef]
28. Horton, R.E. Erosional development of streams and their drainage basins; hydrophysical approach to quantitative morphology. *Bull. Geol. Soc. Am.* **1945**, *56*, 275–370. [CrossRef]
29. Schumm, S.A. Evolution of drainage systems and slopes in badlands at Perth Amboy, New Jersey. *Bull. Geol. Soc. Am.* **1956**, *67*, 597–646. [CrossRef]
30. Faniran, A. The index of drainage intensity: A provisional new drainage factor. *Aust. J. Sci.* **1968**, *31*, 326–330.

31. Arango, M.I.; Aristizábal, E.; Gómez, F. Morphometrical analysis of torrential flows-prone catchments in tropical and mountainous terrain of the Colombian Andes by machine learning techniques. *Nat. Hazards* **2021**, *105*, 983–1012. [CrossRef]
32. Melton, M.A. The geomorphic and paleoclimatic significance of alluvial deposits in southern Arizona. *J. Geol.* **1965**, *73*, 1–38. [CrossRef]
33. Melton, M.A. *An Analysis of the Relation among Elements of Climate, Surface Properties and Geomorphology*; Tech. Rep. II; Office of Navy Research, Department of Geology, Columbia University: New York, NY, USA, 1957; 102p.
34. Patton, P.C.; Baker, V.R. Morphometry and floods in small drainage basins subject to diverse hydrogeomorphic controls. *Water Resour. Res.* **1976**, *12*, 941–952. [CrossRef]
35. Jothimani, M.; Abebe, A.; Dawit, Z. Mapping of soil erosion-prone sub-watersheds through drainage morphometric analysis and weighted sum approach: A case study of the Kulfo River basin, Rift valley, Arba Minch, Southern Ethiopia. *Model. Earth Syst. Environ.* **2020**, *6*, 2377–2389. [CrossRef]
36. Horton, R.E. Drainage-Basin Characteristics. Trans. *Am. Geophys. Union* **1932**, *13*, 350–361. [CrossRef]
37. *Corps of Engineers*; Civil Works Inv., Project CW 153; Department of the Army, Washington District, The Unit Hydrograph Compilations: Washington, DC, USA, 1949.
38. Morisawa, M. Measurement of Drainage-Basin Outline Form. *J. Geol.* **1958**, *66*, 587–591. [CrossRef]
39. Smith, K.G. Standards for grading texture of erosional topography. *Am. J. Sci.* **1950**, *248*, 655–668. [CrossRef]
40. 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; Volume 4, pp. 2439–2443.
41. Kursa, M.B.; Jankowski, A.; Rudnicki, W.R. Boruta—A system for feature selection. *Fundam. Inform.* **2010**, *101*, 271–285. [CrossRef]
42. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
43. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
44. Chen, C.; Yang, D.; Gao, S.; Zhang, Y.; Chen, L.; Wang, B.; Mo, Z.; Yang, Y.; Hei, Z.; Zhou, S. Development and performance assessment of novel machine learning models to predict pneumonia after liver transplantation. *Respir. Res.* **2021**, *22*, 94. [CrossRef] [PubMed]
45. Kim, J.; Park, Y.; Park, S.; Jang, H.; Kim, H.J.; Na, D.L.; Lee, H.; Seo, S.W. Prediction of tau accumulation in prodromal Alzheimer's disease using an ensemble machine learning approach. *Sci. Rep.* **2021**, *11*, 5706. [CrossRef]
46. Mamun, O.; Wenzlick, M.; Hawk, J.; Devanathan, R. A machine learning aided interpretable model for rupture strength prediction in Fe-based martensitic and austenitic alloys. *Sci. Rep.* **2021**, *11*, 5466. [CrossRef]
47. Kursa, M.B.; Rudnicki, W.R. Feature selection with the boruta package. *J. Stat. Softw.* **2010**, *36*, 1–13. [CrossRef]