soil systems

Article

Robustness of Optimized Decision Tree-Based Machine Learning Models to Map Gully Erosion Vulnerability

Hasna Eloudi [1], Mohammed Hssaisoune [1,2,3,*], Hanane Reddad [4], Mustapha Namous [5], Maryem Ismaili [5], Samira Krimissa [5], Mustapha Ouayah [5] and Lhoussaine Bouchaou [1,3]

1 Applied Geology and Geo-Environment Laboratory, Faculty of Sciences, Ibn Zohr University, Agadir 80000, Morocco
2 Faculty of Applied Sciences, Ibn Zohr University, Ait Melloul 86150, Morocco
3 International Water Research Institute, Mohammed VI Polytechnic University, Ben Guerir 43150, Morocco
4 Laboratoire d'Ingénierie & de Technologies Appliquées (LITA), École Supérieure de Technologie de Beni Mellal, Sultan Moulay Slimane University, Beni-Mellal 23000, Morocco
5 Data Science for Sustainable Earth Laboratory (Data4Earth), Sultan Moulay Slimane University, Beni-Mellal 23000, Morocco
* Correspondence: m.hssaisoune@uiz.ac.ma; Tel.: +212-670929680

Abstract: Gully erosion is a worldwide threat with numerous environmental, social, and economic impacts. The purpose of this research is to evaluate the performance and robustness of six machine learning ensemble models based on the decision tree principle: Random Forest (RF), C5.0, XGBoost, treebag, Gradient Boosting Machines (GBMs) and Adaboost, in order to map and predict gully erosion-prone areas in a semi-arid mountain context. The first step was to prepare the inventory data, which consisted of 217 gully points. This database was then randomly subdivided into five percentages of Train/Test (50/50, 60/40, 70/30, 80/20, and 90/10) to assess the stability and robustness of the models. Furthermore, 17 geo-environmental variables were used as potential controlling factors, and several metrics were examined to evaluate the performance of the six models. The results revealed that all of the models used performed well in terms of predicting vulnerability to gully erosion. The C5.0 and RF models had the best prediction performance (AUC = 90.8 and AUC = 90.1, respectively). However, according to the random subdivisions of the database, these models exhibit small but noticeable instability, with high performance for the 80/20% and 70/30% subdivisions. This demonstrates the significance of database refining and the need to test various splitting data in order to ensure efficient and reliable output results.

Keywords: soil erosion; inventory data; performance; robustness; spatial prediction

Citation: Eloudi, H.; Hssaisoune, M.; Reddad, H.; Namous, M.; Ismaili, M.; Krimissa, S.; Ouayah, M.; Bouchaou, L. Robustness of Optimized Decision Tree-Based Machine Learning Models to Map Gully Erosion Vulnerability. *Soil Syst.* **2023**, *7*, 50. https://doi.org/10.3390/soilsystems7020050

Academic Editor: Luis Eduardo Akiyoshi Sanches Suzuki

Received: 10 February 2023
Revised: 12 May 2023
Accepted: 14 May 2023
Published: 16 May 2023

1. Introduction

Soil erosion is known as a loosening of sediment from the uplands to the valley floor induced by runoff [1]. This phenomenon is described as a catastrophic global issue with extensive environmental, social, and economic repercussions [2]. Soil erosion endangers water and soil resources, both of which are vital to human existence and the environmental equilibrium. There are several types of soil erosion, the most notable of which is gully erosion (GE) [3]. This type contributes to landscape shaping while also causing significant damage such as the degradation of arable land fertility, damage to water infrastructure and shortening of its life span, and the disruption of countries' economic and societal circumstances [4]. This phenomenon has affected one-third of the world's arable land in the last few decades [5]. According to the literature, soil erosion affects more than 10 million hectares of agricultural land each year, with annual global loss rates of approximately 43 Pg [6]. According to FAO [7], soil losses due to soil erosion are estimated to result in a $1 billion economic loss. Soil erosion affects 40% of Moroccan territory, with annual loss rates ranging from 23 to 55 t/ha/yr on average, and extreme values reaching

524 t/ha/yr in some places [8]. Furthermore, agriculture is the main source of income for the people who live in the mountainous areas of Morocco. However, these areas are heavily affected by soil erosion, which decreases the amount of fertile land, reduces the quality and quantity of water, and has other serious economic and social effects [9]. In this respect, the Lakhdar watershed in the Moroccan High Atlas is one of the regions impacted by significant soil degradation as a result of its very complicated physical features, such as a very high topography and a steep slope occupied by rocks with differing properties. In connection with these factors, the study of gully erosion vulnerability may be a crucial tool for understanding erosive processes in comparable environments. As a result, identifying areas prone to soil erosion is an important step toward good natural resource management and long-term protection and a deeper comprehension of the erosive processes and the factors that influence this phenomenon under current climatic conditions.

Since the 1930s, numerous models have been developed to estimate soil loss rates and qualitatively assess soil erosion sensitivity. Currently, combining remote sensing with geographic information systems makes this task easier and more efficient [10]. According to the literature, there seem to be two distinct methods of soil erosion analysis: Qualitative and quantitative approaches. To assess medium- and long-term soil loss rates, empirical models such as the Universal Soil Loss Equation (USLE), Modified Universal Soil Loss Equation (MUSLE), and Revised Universal Soil Loss Equation (RUSLE) have been employed [11,12]. There are also physical models that can be used to quantify soil loss averages, such as the Water Erosion Prediction Project Model (WEPP) [13], the Chemical Runoff and Erosion for Agricultural Management System (CREAMS) [14], and other models such as The Erosion-Productivity Impact Calculator (EPIC) [15] and the Limburg Soil Erosion Model (LSEM). These models, however, cannot predict gully erosion susceptibility because this type of erosion is controlled by several factors not completely taken into account by their formulas, including topographic, hydraulic, climatic, soil conditions, and morphometric characteristics [16]. Furthermore, quantitative models necessitate calibration and are subject to significant uncertainty in terms of differences between predicted and measured loss rates [17]. In this regard, other bivariate and multivariate statistical models have been developed [18]. Furthermore, hierarchical process analysis (HPA) methods [19], logistic regression models [20], Weight-of-Evidence (WoE) models, and entropy indexes are used to evaluate the sensitivity to gully erosion [21,22]. Moreover, machine learning techniques have proven to be an effective tool for assessing and mapping gully erosion [23,24]. These methods are a subset of the artificial intelligence field that is based on the hypothesis that computer programs can learn from inventory and model input data without the need for human intervention [25]. Presently, the use of machine learning-based approaches has become popular, particularly in the mapping and monitoring of natural hazards, because they produce high-accuracy results in data processing, classification, and prediction [26]. In addition, numerous researchers have tested the high performance of deep learning models in the spatial prediction of vulnerable zones to gullying phenomena [27]. On the basis of this previous investigation, we aim to fill a gap in the analysis of inventory data by investigating various possible subdivisions and proposing to researchers and decision-makers a simple, less expensive, and effective method for predicting soil erosion vulnerability.

The objectives of this investigation are to identify the factors that cause gully erosion and to test several "decision tree" models to develop a gully erosion susceptibility detection and prediction model suitable for mountainous and semi-arid areas. It is therefore essential to evaluate the stability of these models against the variation in training and testing percentages. Because of this, we will test how well these models perform with five different splitting of data on Training and Testing: 90/10%, 80/20%, 70/30%, 60/40%, and 50/50%. The advantage of this kind type of decision tree model is that it determines the relationships between the explanatory variables, the dependent factors, and the occurrence of the phenomenon in a simple tree structure. This makes these models more comprehensive compared to mathematical formulas or correspondence tables. According to the literature, numerous studies have used these models to assess and monitor gully

erosion vulnerability [24,28]. Despite this, the use of these combined approaches to predict areas susceptible to gully erosion on the one hand, and their tests under different quantities of input data subdivision on the other hand, remains very limited. Additionally, the combination of different types of decision trees has never been tested in Morocco, lending originality to this research. Finally, the development of these advanced methods to map gully erosion-vulnerable areas is critical because it will support decision-making in terms of planning and implementing sustainable policies and strategies for land management of water and soil resources.

2. Materials and Methods

2.1. Study Area

The Lakhdar watershed is one of the Oum Er-Rbia sub-basins (Figure 1) and is located in the Atlas Mountains axial zone and covers approximately 1600 km^2. The study area is divided into three geomorphological distinct units: High mountains with altitudes of up to 4000 m, plateaus, and valleys carved deeply by soil erosion. The region has a major hydraulic structure of critical importance in terms of drinking water supply for Marrakech city as well as irrigation of the Haouz plain downstream. Geologically, the Lakhdar watershed is composed of an amalgam of lithologies with a dominance of Jurassic limestones and Permo-Triassic sandstones; its upstream part is primarily characterized by detrital deposits represented essentially by clays, marls, and alluvial deposits of the Quaternary period (Figure 2 and Table 1). The study area is classified as a semi-arid zone with hot summers (from June to August) and cold winters (since December to February). The aridity primarily affects the downstream portion of the watershed area; however, the upstream portion is controlled by high altitudes, resulting in significant spatial differences in rainfall amounts. In general, the average annual rainfall is approximately 450 mm, with maximum values of 600 mm recorded in the upstream portion and minimum values of 300 mm recorded primarily in the downstream areas. The area under investigation has a deteriorated vegetative cover, which is exacerbated by the dynamics and anthropic activities that invade the area. This is supported by a 36% reduction in forest area over the last few decades. As a result, the watershed area serves as a test bed for studying soil erosion processes and comprehending the erosive processes that occur in a semi-arid mountainous area.

Table 1. Description of lithological units in Lakhdar watershed, Morocco.

Class	Description
1	Silurian: Graptolitic shales
2	Stephano-Triassic: Sandstones and red conglomerates
3	Permian-Triassic: Basalt
4	Lower Lias: Limestones, and red clays
5	Lower Lias: Limestones and marls
6	Middle Lias: Limestones
7	Upper Lias: Conglomerates, sandstones, and clays
8	Dogger: Marls and limestones
9	Quaternary: Alluvial and Rockfull

2.2. Methodology

The current study's approach includes several major steps illustrated by the flowchart presented in Figure 3. In addition, the same figure presents an overview of the approach that was developed for probabilistic gully erosion susceptibility using decision tree models (C5.0, XGBoost, treebag, GBM, and Adaboost) to produce accurate Gully Erosion Susceptibility Maps (GESMs).

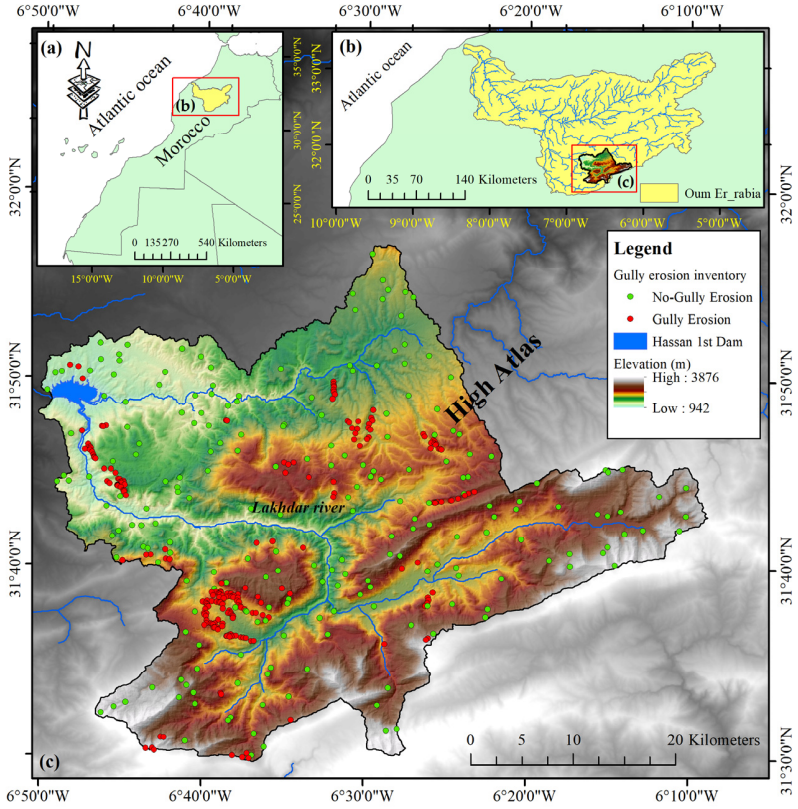


Figure 1. Location maps of the study area (**a**) at Moroccan scale, (**b**) at Oum Er-Rbia watershed scale, and (**c**) DEM of Lakhdar watershed.

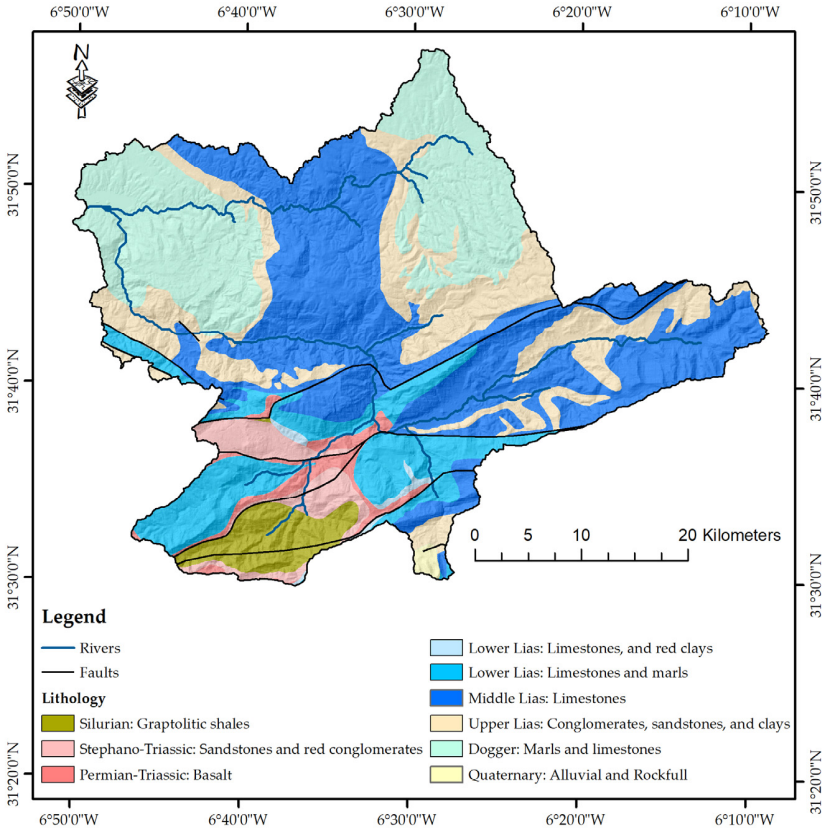


Figure 2. Geological map of the study area.

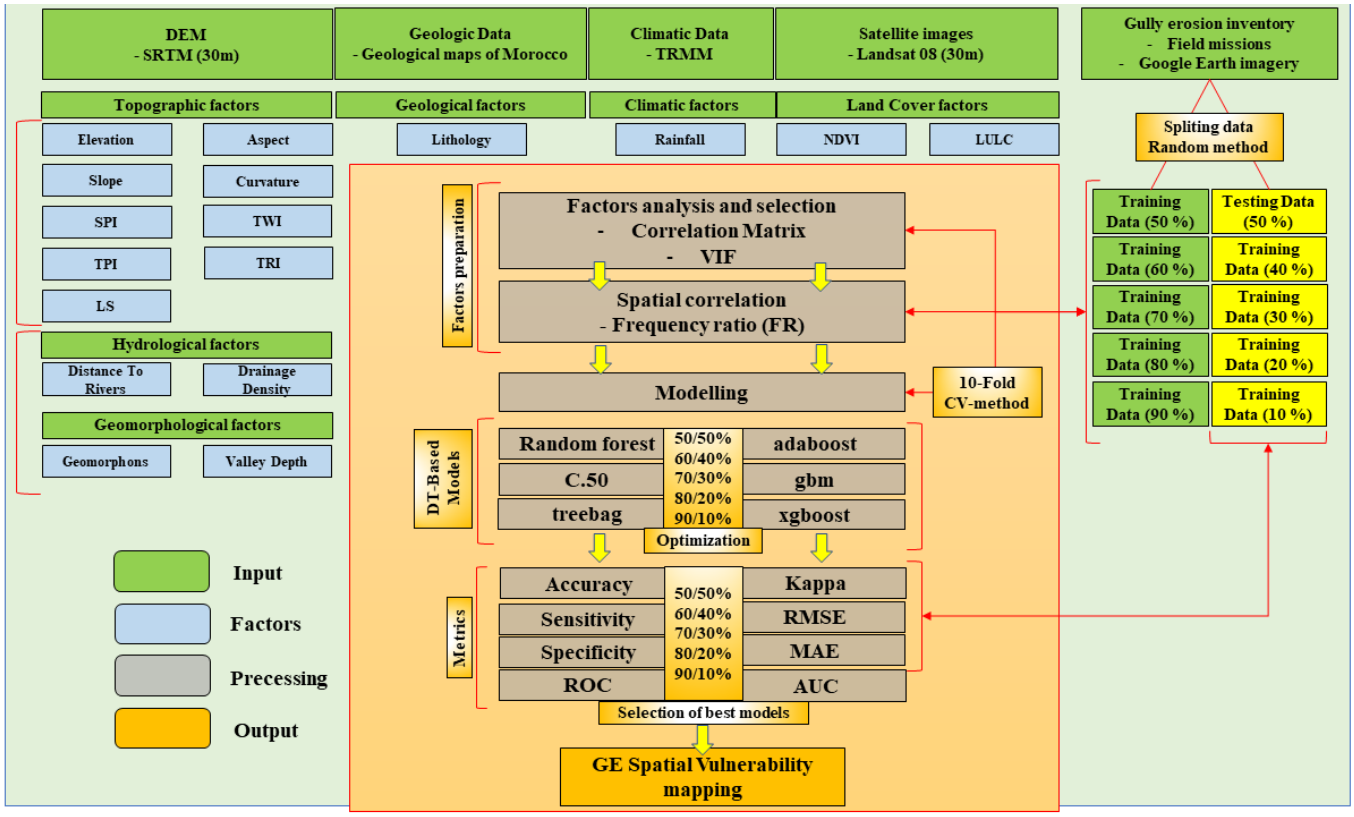


Figure 3. A detailed flowchart shows research methods.

2.2.1. Gully Erosion Inventory Mapping

One of the most important indicators for assessing gully erosion susceptibility is the gully erosion inventory map (GEIM), which presents the spatial locations of the gullies. The distribution of traditional and present gully locations can be used to estimate the potential probability of gully erosion in a region. As a result, it is important to create a gully erosion inventory map in order to estimate the optimal future gully erosion [29]. The GEIM is required for the preparation of GESMs by various predictive models [30] and was used as the dependent variable in this study. For GEIM preparation, gully locations were identified by conducting fieldwork in the study region combined with google earth image analysis. Gully locations were determined using a handheld GPS device. In the study area, 217 gullies were collected (Figure 1). This database was then randomly subdivided into five quantities (50/50%, 60/40%, 70/30%, 80/20%, and 90/10%) to assess the performance, robustness, and stability of the models (Figure 2).

2.2.2. Dataset Preparation for Spatial Modelling

The selection of Gully Erosion Conditioning Factors (GCFs) is a crucial stage in the development of GESMs using several techniques [31]. In this study, 17 geo-environmental variables were used for spatial modelling of gully erosion, including elevation, slope, aspect, rainfall, LandUse-LandCover (LULC), Normalized Difference Vegetation Index (NDVI), distance to rivers, Drainage Density, Valley Depth, Curvature, Lithology, Geomorphons, Topographic Position Index (TPI), Topographic Wetness Index (TWI), Topographic Roughness Index (TRI), Slope Length (LS), and Stream Power Index (SPI) (Table 2), while taking previous literature and multicollinearity into account. Note that for all quantitative factors, the classification is based on the Natural break technique, as suggested by the majority of researchers [32].

Table 2. Data sources used in this study.

Factors	Data Layers	Data Source
Topographic factors	Elevation Slope (°) Stream Power Index (SPI) Topographic Position Index (TPI) Slope Length (LS) Aspect Curvature Topographic Wetness Index (TWI) Topographic Roughness Index (TRI)	SRTM-DEM (Digital Elevation Model) were downloaded from the website of United States Geological Survey (USGS) (http://gdex.cr.usgs.gov/gdex/ (accessed on 2 August 2022)); Pixel size of 30 m × 30 m.
Hydrological factors	Distance To Rivers Drainage Density	
Geomorphological factors	Valley depth Geomorphons	
Geological factors	Lithology	Geologic map of Ouaouizghte-Dades 1/200,000 Bourcart et al., 1942 [32] Geologic map of Demnate-Telouate 1/200,000 Termier, 1941 [32]
Climatic factors	Rainfall (mm)	TRMM data
LAND cover factors	Normalized Difference Vegetation Index (NDVI) LandUse-LandCover (LULC)	LANDSAT-8 OLI TIRS satellite image

The elevation data layer was created using the digital elevation model (DEM) obtained from the USGS (Figure 4a). The study area's altitude was separated into five groups: 942–1513 m, 1504–1947 m, 1937–2379 m, 2381–2866 m, and 2860–3876 m (Figure 4a). The slope has a big effect on how gullies form [33]. The slope map was created in GIS using a DEM and was divided into five groups: 0–9, 10–18, 19–26, 27–36, and 37–71° (Figure 4a). The aspect map, similar to that of the slope map, was created from the DEM and divided into nine classes: Flat, north, northeast, east, southeast, south, southwest, west, and northwest. The curvature is also mapped from DEM using GIS and divided into five classes −24.9 to −2.4, −2.3 to −0.9, −0.8 to −0.4, 0.5 to 2.1, and 2.2 to 30.2 (Figure 4a). The sediment power index (SPI) reveals the discharge, carrying potential, and water erosion energy, which influences the sensitivity to gully erosion [34]. The following Equation (1) was used to obtain the SPI from the DEM:

$$SPI = As \times \tan\beta, \quad (1)$$

where As is the upstream drainage area and β is the slope degree. The SPI was classified into the five sub-categories of 0–443, 444–959, 960–1587, 1588–2547, and 2548–9410 (Figure 4a). The topographic wetness index (TWI) is regarded as a key gully erosion conditioning factor. Using the following Equation (2), the TWI was obtained from DEM data:

$$TWI = \mathrm{In}(As/\tan\beta), \quad (2)$$

where As is the upstream drainage area and β is the slope degree. The TWI was categorized into five classes: 2–6, 7–8, 9–11, 12–16, and 17–25 (Figure 4a). The slope length (LS) factor was calculated also from the DEM by means of Equation (3).

$$LS = (m + 1) \times [As/22] \times [\sin\beta/0.0896], \quad (3)$$

where As is the upstream drainage area and β is the slope degree. The LS was categorized into five classes: 0–4.16, 4.16–9.02, 9.02–14.58, 14.58–27.76, and 27.76–177 (Figure 4b). The Terrain ruggedness index (TRI) indicates the elevation difference between the surrounding

cells of a DEM [35]. The TRI was classified into five classes: 0–4.49, 4.49–8.66, 8.66–13.80, 13.80–22.46, and 22.49–81.83 (Figure 4b). The topographic position index (TPI) is also calculated using DEM; TPI is a terrain classification method in which the altitude of each data point is compared to its neighbors. In a nutshell, we calculate the height difference between each data point, or pixel in a raster DEM, and its immediate surroundings. The TPI was classified into five classes: −13–152, 152–298, 298–459, 459–632, and 632–966 (Figure 4b). Drainage density factors were also used and categorized into five classes: 0.14–0.46, 0.47–0.64, 0.65–0.79, 0.8–0.93, and 0.94–1.3 (Figure 4b). The distance from the river map was prepared by applying the Euclidian distance buffer (EDB) tool in GIS (Figure 4a). It was classed into five sub-classes, namely 0–185 m, 186–419 m, 420–668 m, 669–966 m, and 967–2052 m (Figure 4a).

Despite the fact that gully erosion is highly dependent on the lithology qualities of the exposed material near the earth's surface, lithology indicators play an essential function in assessing gully erosion vulnerability [33]. The lithological map was generated from the available geological data of Morocco and was classified into nine classes numbered one through nine (Figure 4a and Table 1). The NDVI was calculated using the Landsat 8 imagery in a GIS environment following this Equation (4).

$$\mathrm{NDVI} = (\mathrm{NIR} - \mathrm{R})/(\mathrm{NIR} + \mathrm{R}), \quad (4)$$

where NIR is the near-infrared spectrum and R is the red spectrum. The map was categorized into five classes: −0.12 to 0.1, 0.11 to 0.14, 0.15 to 0.2, 0.21 to 0.31, and 0.32 to 0.58 (Figure 4a). The Land Use Land Cover (LULC) map was obtained from Landsat 8 imagery based on the supervised classification process in the GIS environment. Water bodies, soil bare, sparse vegetation, agricultural land, and forest are the LULC classes (Figure 4b).

The geomorphological factors used are Valley depth and Geomorphons. The first was classified into five classes: −13–152, 152–298, 298–459, 459–632, and 632–966. The second was classified into ten classes: Flat, summit, Ridge, shoulder, spur, slope, hollow, footslope, and Valley depression (Figure 4b). Rainfall is a major factor that directly contributes to gully erosion, and annual precipitation data were obtained from the Tropical Rainfall Measuring Mission (TRMM). According to the rainfall map, the annual average rainfall in the study area ranges between 390 and 610 mm.year^{-1}. The most significant values are found in the south, while precipitation decreases sharply in the north (Figure 4a). The rainfall map was subdivided into five classes: 330–395, 395–450, 450–505, 505–552, and 552–610.

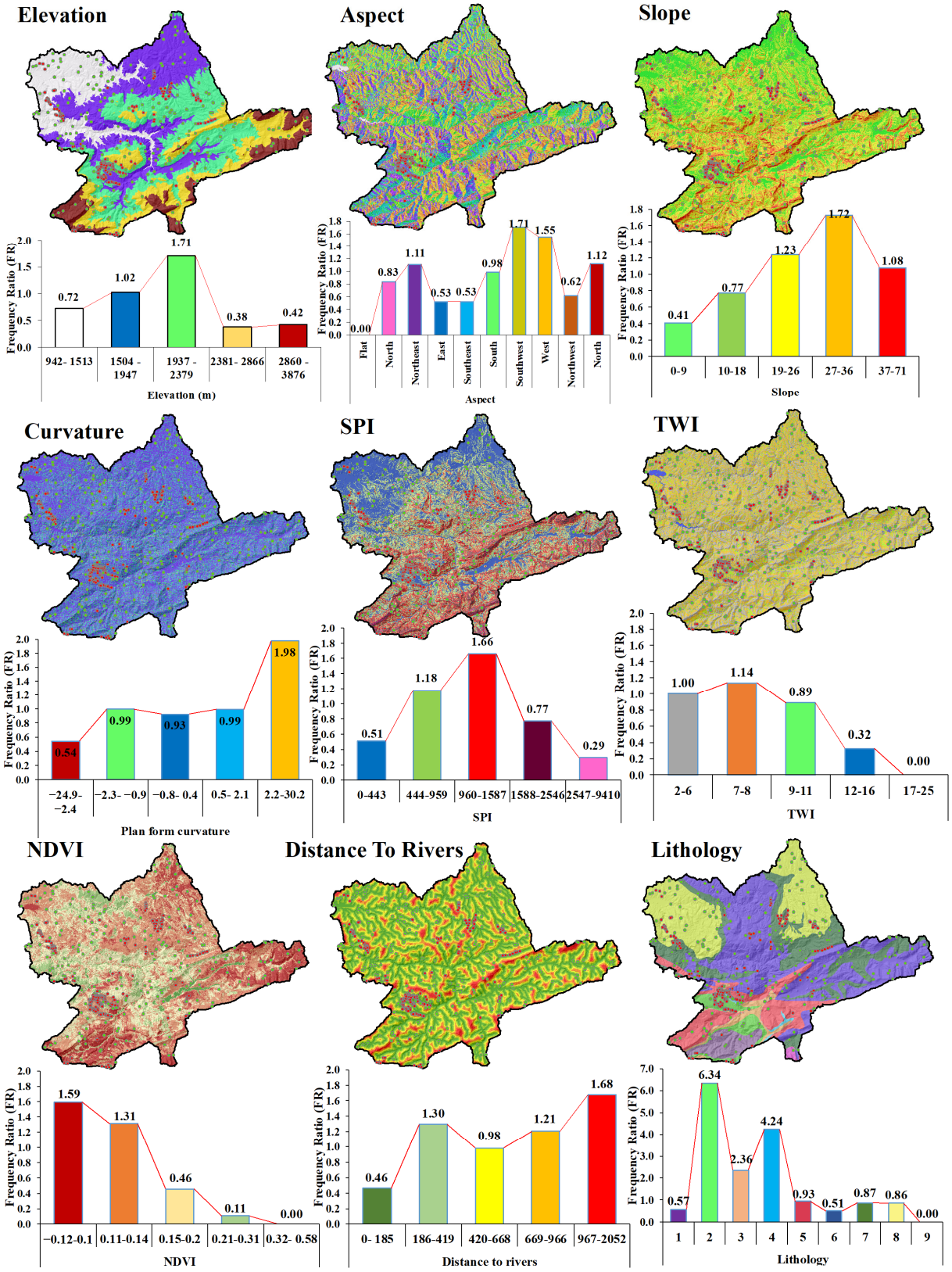
2.2.3. Multicollinearity Analysis

The multicollinearity test is an important approach to measure the linear dependency among the specified independent parameters in statistical modelling. This method needs to be applied to machine learning models in order to improve their performance [31]. This study used the correlation matrix and variable inflation factor (VIF) methods to determine the multicollinearity of the Gully erosion factors. Using the correlation between predictor pairs alone has limitations, whether small or large [36].

2.2.4. Decision Tree-Based Approaches

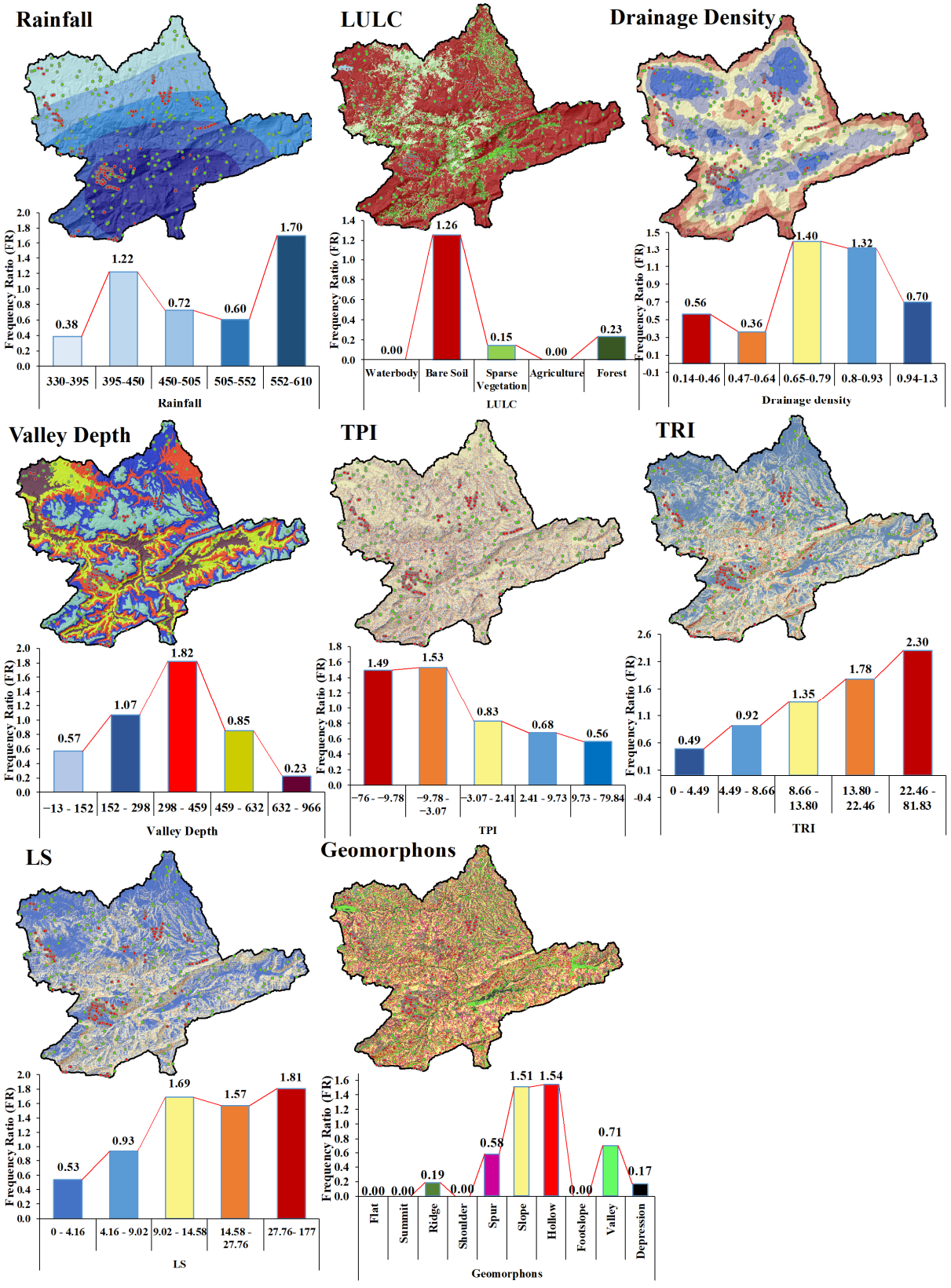
Random Forest (RF)

The random forest (RF) algorithm is a statistical technique for controlling a large number of connected variables [37]. In 2001, Breiman [38] developed the technique as a binary tree decision-making system [39]. RF may also assess dynamic trends and understand nonlinear connections between explanatory and dependent variables. It will also merge multiple data formats due to the lack of a uniform distribution of the data used. RF is ideally suited to geographical studies and is often employed in land movement sensitivity mapping [40]. This method combines many decision trees, with many bootstrap samples obtained from the data and a range of input variables arbitrarily added to each tree. Furthermore, the RF approach categorizes elements according to their relevance. The weights are determined by taking the average decline in forecast accuracy.



(**a**)

Figure 4. *Cont.*



(**b**)

Figure 4. (**a**,**b**). Conditioning factor maps and spatial correlation between factors and Gully erosion location using FR method.

C5.0

C5.0 is a decision tree technique that works by first testing the classifier to classify unseen data and then using the final decision. Pandya and Pandya [41] demonstrate decisively that C5.0 is an improvement over C4.5 in terms of processing time, memory consumption efficiency, error, and, ultimately, classification accuracy. When compared to more advanced and complicated machine learning models (e.g., neural networks and support vector machines), the C5.0 algorithm decision trees perform almost as well but are considerably easier to understand and use [42].

Adaboost

Adaboost is a method for reducing the error of a weak learning algorithm. In theory, the weak learning algorithm can be any that can generate classifiers that are only marginally better than random guessing [43]. There are two primary issues with boosting: Determining how to modify the training set so that the weak classifier can train using it and how to combine the weak classifiers gained during training to form a strong one. Previous authors [44] developed the Adaboost (adaptive boosting) method, which adjusts the weight without requiring prior information on learner learning. Adaboost has been employed in ensembles to increase prediction performance, most notably in neural networks [45], support vector machines [46], and decision trees [47]. The classifier uses an adaptive resampling strategy to select training samples, which means that a misclassified dataset generated by a prior classifier is chosen more frequently than correctly classified ones, allowing a new classifier to perform well in a fresh dataset. Each iteration gives the dataset a weight so that the following integration concentrates on reweighted datasets that were previously misclassified. In the final classifier, the ensemble predictions are weighted. The Adaboost algorithm can be applied to two-class problems, multi-class single-label issues, multi-class multi-label problems, single-label problem categories, and regression problems [47].

Treebag

Bagging or bootstrap aggregation is an ensemble method developed [48] that involves repeatedly training the same algorithm using different subsets of the training data. After that, the final output forecast is averaged over all sub-model projections. Bagging, in general, increases classification accuracy by lowering the variation of classification incertitude [49]. Freund and Schapire [48] claim that bagging can considerably enhance accuracy if changing the learning set creates a major change in the predictor built. The ensemble's majority vote is used to forecast a test sample [50]. Bagging attempts to reduce the error level owing to the variation of the base classifier by voting on the predictions of each classifier because each ensemble member is trained with a separate set of data [48].

Gradient Boosting Machine (GBM)

The Gradient Boosting Machine (GBM) is a forward-learning ensemble approach developed by [51] that is commonly used in machine learning. It is an effective method for developing predictive models for regression and classification tasks. GBM assists us in obtaining a predictive model in the form of an ensemble of weak prediction models such as decision trees [52]. When a decision tree performs poorly as a learner, the resulting algorithm is known as gradient-boosted trees [30]. Most supervised learning algorithms, in general, rely on a unique predictive model, such as decision trees and regression models. However, some supervised ML algorithms rely on the ensemble, which is a combination of various models. In other words, when multiple base models contribute predictions, boosting algorithms adapt to an average of all predictions. GBM is made up of three components, which are as follows: Weak learners, a loss function, and an additive model.

Extreme Gradient Boosting (Boost)

The gradient boosting theory is the basis for the XGBoost model, which combines a set of weak learners' predictions to create a robust learner through an additive training

strategy [53]. The XGBoost model requires a number of parameter selections to predict the model, but the performance is always dependent on the selection of the optimal parameters. Thus, in the modelling process, the user needs to select three key parameters: colsample by tree (the portion of the variables to be used in each tree), subsample (the subsample ratio for the data to be considered in each tree), and nrounds (the maximum number of boosting iterations).

2.2.5. Models' Optimization

Cross-validation is an extremely effective tool in advanced and powerful machine learning models [54]. It allows us to make better use of our data and provides us with much more information about the performance of our algorithms. In this research, we used two approaches: K-fold cross-validation and tuning hyperparameters. For the first approach, the K-fold cross-validation method splits the input dataset into K groups of identical-size samples. The name given to these samples is folds. The prediction process uses k-1 folds for the separate training data and the remaining folds are used for the testing data. This is a popular CV approach because it is simple to understand and produces fewer biased results compared to other techniques. For the second approach, the process of tuning the parameters present as item sets while building ML models is known as hyperparameter tuning. These parameters are defined by us and can be manipulated as desired by the scientist in order for the model to perform well.

2.2.6. Validation and Accuracy Assessment

To assess the robustness of the used ML DT-based models used in the GE modelling process, we employed a number of statistics-based metrics, including sensitivity and specificity. This enables us to assess how the gully modelling predictive skill is employed to classify gully locations; specificity denotes the non-gully areas, while sensitivity denotes the gully area. These methods are relevant to predicting gully and non-gully areas. In addition, the kappa approach is utilized to assess the reliability of a gully erosion model. The values fall within the interval of -1 to 1, with 1 representing the best results. In addition, we used the accuracy, RMSE, and MAE values to assess the performance of the models tested for each data subdivision. A high value of accuracy and lower values of RMSE and MAE indicates better results of gully erosion modelling. Finally, the receiver operating characteristic curve (ROC) is regarded as a standard metric for evaluating the results of using ML models. To evaluate the performance of the modelling process, we use four types of possible metrics: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). All of the equations used to calculate these parameters are mentioned below:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{5}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}}, \tag{6}$$

$$\text{Accuracy} = \frac{\text{TN} + \text{TP}}{\text{TP} + \text{FP} + \text{TN} + \text{TP}}, \tag{7}$$

$$\text{Kappa} = \frac{\text{Accuracy} - \text{B}}{1 - \text{B}}, \tag{8}$$

$$\text{Where B} = \frac{(\text{TP} + \text{FN})(\text{TP} + \text{FP}) + (\text{FP} + \text{TN})(\text{FN} + \text{TN})}{\sqrt{\text{TP} + \text{TN} + \text{FN} + \text{FP}}}, \tag{9}$$

$$\text{RMSE} = \sqrt{\frac{1}{\text{n}} \sum_{\text{i}=1}^{\text{n}} (\text{X}_\text{P} - \text{X}_\text{A})^2}, \tag{10}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|(X_P - X_A)^2\right|, \quad (11)$$

2.2.7. Variable Importance Analysis

We adopted two methods based on the RF model to generate a classification of factors according to their importance. The first is the mean decrease in accuracy and the second is the mean decrease in Gini. The mean decrease in accuracy shows how much the model accuracy loses when a factor is left out. The more the accuracy decreases, the greater the significance of the variable for effective results. The mean decrease in Gini is a measure of variable importance based on the principle that whenever a node is split on variable m, the Gini impurity criterion for the two descendent nodes is lesser compared to the parent node. Adding the Gini reductions for each variable across all trees offers a rapid measure of variable importance [55].

3. Results

3.1. Preliminary Data Analysis

The correlation matrix and the variance inflation factor (VIF) were used to examine the collinearity between the explanatory factors (Figure 5 and Table 3). The correlation matrix shows a high value of 0.8 between the LS factor and the TRI factor, while the collinearity of the remaining factors remains acceptable. The VIF shows a tolerance level with values less than 5: Curvature (1.069), TWI (1.075), and Distance to Rivers have the lowest VIF values (1.088), and the highest value is related to the LS-factor (3.396). As a result of the collinearity test, the LS factor has been omitted from the analysis.

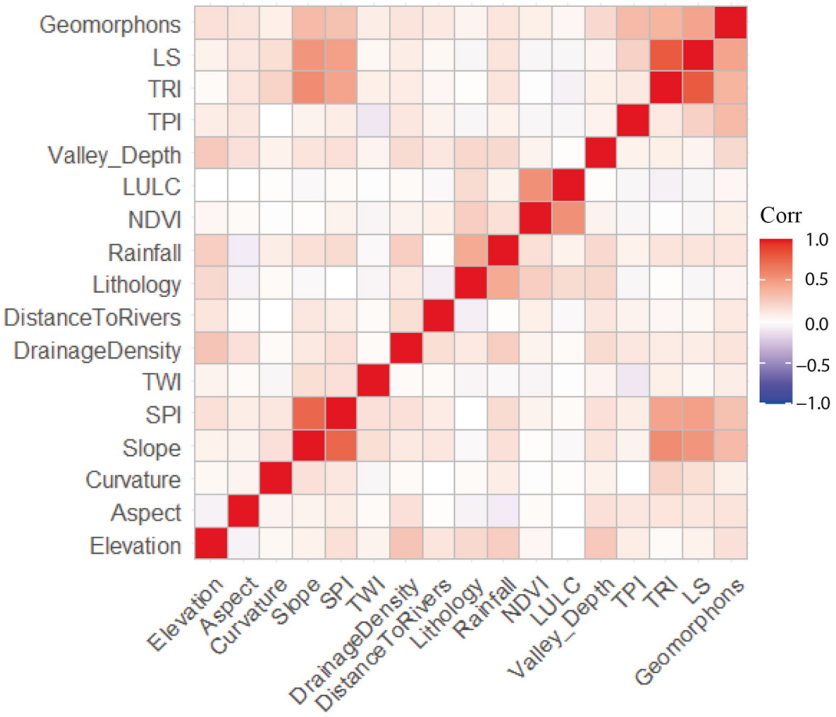


Figure 5. Correlation matrix results between conditioning factors. LandUse-LandCover (LULC), Normalized Difference Vegetation Index (NDVI), Topographic Position Index (TPI), Topographic Wetness Index (TWI), Topographic Roughness Index (TRI), Slope Length (LS), and Stream Power Index (SPI).

Table 3. Variance inflation factor (VIF) and Tolerance (TOL) results.

Factors	VIF	TOL
Elevation	1.264	0.791
Aspect	1.118	0.895
Curvature	1.069	0.936
Slope	2.805	0.356
SPI	2.461	0.406
TWI	1.075	0.930
Drainage Density	1.214	0.824
Distance To Rivers	1.088	0.919
Lithology	1.366	0.732
Rainfall	1.406	0.711
NDVI	1.537	0.650
LULC	1.464	0.683
Valley Depth	1.207	0.828
TPI	1.217	0.822
TRI	3.301	0.303
LS	3.396	0.295
Geomorphons	1.494	0.669

3.2. Spatial Relationship between Gully Locations and Effective Factors

A bi-variate statistical approach based on the frequency ratio (FR) was used to correlate causative factors with the spatial distribution of gullies (Figure 4a,b). For a given factor, this ratio determines the likelihood of gully occurrence versus non-occurrence [56]. The highest value of FR is 6.34 represented by the lithology class occupied by sandstones and red conglomerates followed by the class of limestones and red clays, which had an FR value of 4.24, and lastly, the Basalts class with an FR of 2.36. The TRI factor and curvature represent a strong spatial correlation with the gullies with an FR value of 2.30 (class 22.46–81.83) and 1.98 (class 2.2–30.2), respectively. The topographic factors also showed a high spatial correlation with an FR value of 1.82 for valleys ranging in depth from 298 m to 459 m followed by the highest class of the LS factor with an FR of 1.81, and then the slope class (27–36°) with an FR of 1.72 and the elevation class, which ranges from 1937–2379 m with an FR value of 1.71. The majority of gullies developed on the southwest-facing slopes, which is represented by the high value of the Aspect factor (FR = 1.71). Rainfall also has a strong concordance with gully development where the highest value of FR (1.70) is given to the maximum rainfall class (552 and 610 mm). Compared with the rest of the factors, the majority of gullies developed in areas where the distance to rivers was more than 552 m (FR = 1.68), areas classified as the moderate SPI class (FR = 1.66), bare soil areas with an FR = 1.26 for the LULC factor, and areas in which the NDVI class ranged between −0.12 and 0.1 (FR = 1.59). Furthermore, the majority of gullies form on slopes and cavity areas, as indicated by the geomorphic factor, in which FR for these classes is 1.54. The TWI naturally correlates with gully formation areas; in the current study, this index has an FR = 1.14 represented by classes 7–8 of the TWI factor.

3.3. Variable Importance Analysis

Two measures were considered to identify the importance of the predictive factors of gully erosion: The average decrease in accuracy and the average decrease in Gini, which is based on the RF model with four subdivisions of the input database (50%, 60%, 70%, 80%, and 90%) (Figure 6). In general, the results of these two measures show that all variables play a role in gully formation. However, some factors were more important in predicting the spatial distribution of GE based on the average decrease in the accuracy index. The results show that the factors lithology, geomorphons, elevation, and LULC are the most important in terms of controlling gully formation. This can be explained by the study area's mountainous and geomorphological characteristics, as well as its continuous active tectonic aspect. These findings also demonstrate the effect of anthropogenic action on gully erosion sensitivity and the role of vegetation cover protection in combating this phenomenon. Furthermore, the average decrease in the Gini index results is in perfect agreement with the previous results, confirming the importance of lithology, geomorphological unit, and vegetation cover protection in the formation of gullies.

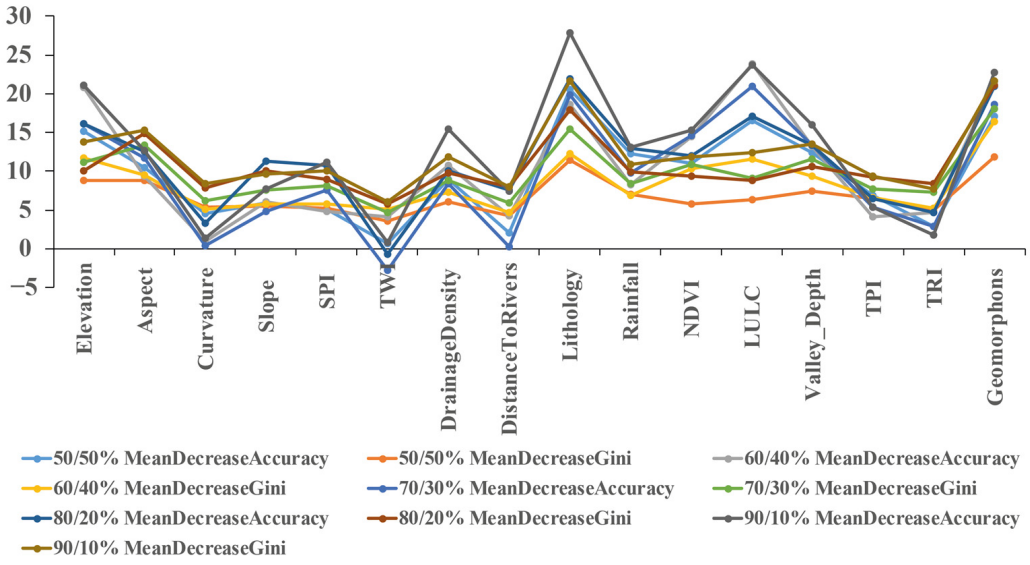


Figure 6. Conditioning factors importance assessment using RF algorithm.

3.4. Gully Erosion Susceptibility Mapping

The gully erosion susceptibility maps (GESMs) allow us to visualize the spatial distribution of gullies and identify the areas vulnerable to gully formation. GESMs were produced using the R interface and reclassified using the natural break method in GIS software (Figure 7). The percentages of the areas occupied by each gully erosion sensitivity class in relation to each model are shown in Figure 8. According to these results, the higher sensitivity classes account for 24% of the total area for the RF and XGBoost models, 23% of the study area for the C5.0 and GMB models, 25% of the area for the treebag model, and 28% of the total area for the Adaboost model. However, the areas with moderate gullying susceptibility range in percentage of the area from 28% for Adaboost to 24% for the RF model, 22% for the C5.0, and 19% for the treebag and GBM models. These findings are consistent with field observations, as the majority of mapped gullies are classified as having high or very high gully erosion sensitivity. Furthermore, all gully erosion sensitivity maps show increasing spatial variation from the very low gully erosion class to the very high gully erosion class, demonstrating the effectiveness of the models used and the reliability of the results.

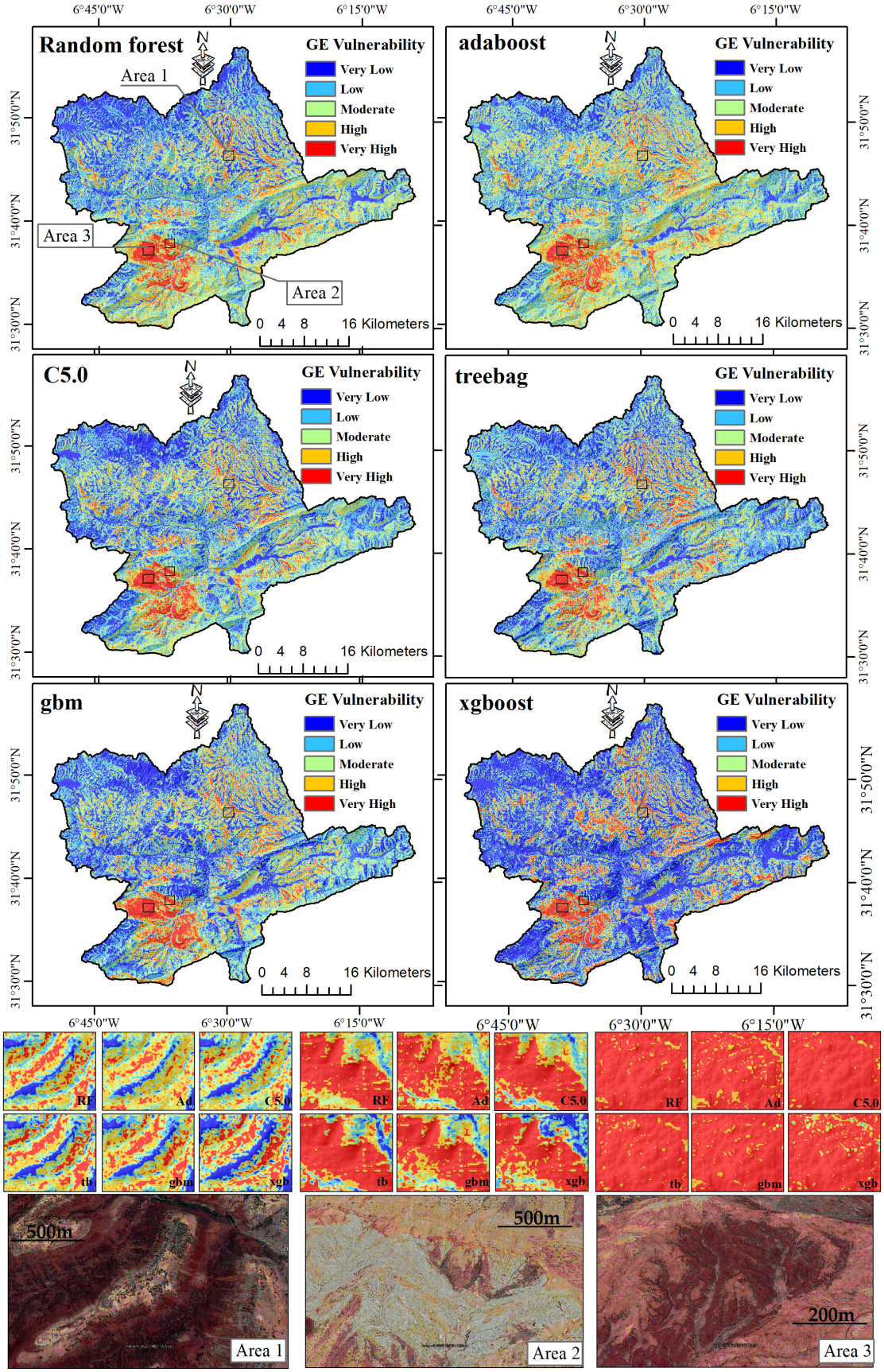


Figure 7. Gully erosion sensitivity maps using Dt-based models.

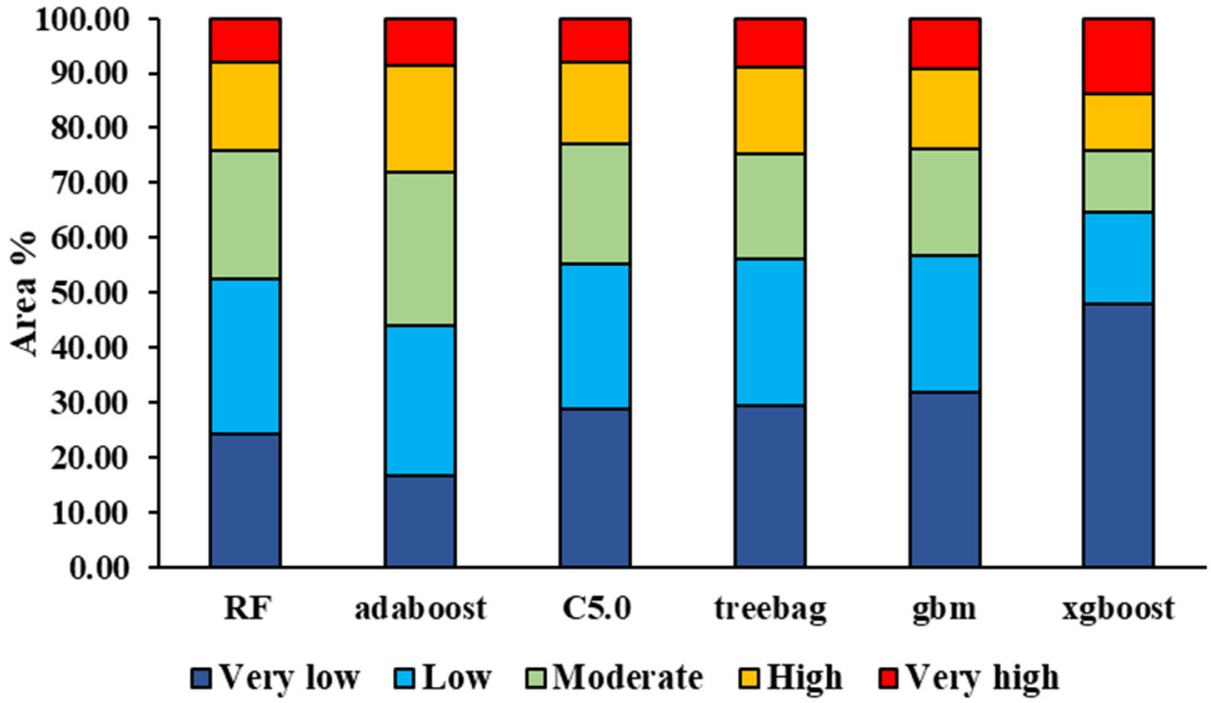


Figure 8. Distribution of gully erosion (GE) susceptibility classes (area %).

3.5. Model Accuracy and Validation Results

In this regard, six models were used, the input data were evaluated by cross-validation ten times for each model, and the accuracy was calculated on the basis of five subdivisions by four parameters, namely Kappa, ROC-AUC, RMSE, and MAE. As a result, we were able to identify the degree of discrimination and reliability, reflecting the performance of the chosen models. Comparing the results obtained (Figures 9–11), the C5.0 model shows a better performance, especially for the 70/30% subdivision, with an AUC value of 90.80% followed by the RF model with an AUC value equal to 90.10% for the 80/20% subdivision, then XGBoost and Adaboost models with an AUC of 90% for the 70/30% subdivision, then the GBM model with an AUC of 88.20% for the 90/10% subdivision, and finally, the treebag model with an AUC of 87.7% for the 70/30% subdivision. This demonstrates that the entire accuracy of the used models is high, particularly at the 70/30% and 80/20% subdivisions for the majority of these models. The average Kappa index values for the RF, C5.0, Adaboost, GBM, treebag, and XGBoost models are 0.58, 0.56, 0.59, 0.55, 0.57, and 0.54, respectively. These results are classified as acceptable to moderate. The average RMSE values range between 0.45 for the RF and Adaboost models, 0.46 for the C5.0 and treebag models, and 0.47 for the GBM and XGBoost models, indicating that the output results are of high quality and reliability. In the 10-fold cross-validation analysis, the prediction models used demonstrated robustness and stability for the calibration and validation datasets. These models also had a high accuracy, which exceeded 80% for the set of random subdivisions used.

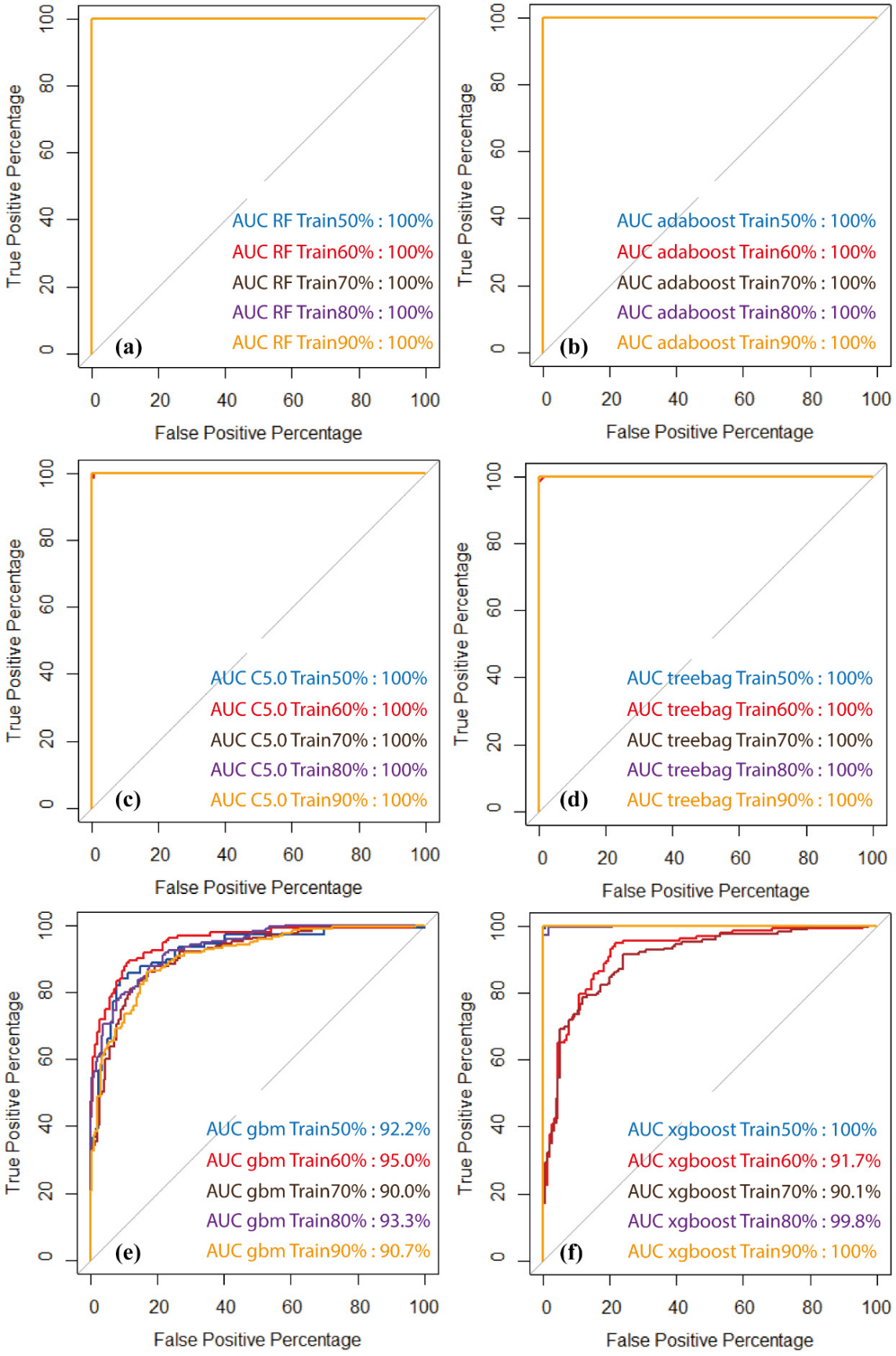


Figure 9. Success rate curve and the area under curve values of the DT-based models using the training dataset: RF (**a**), Adaboost (**b**), C5.0 (**c**), treebag (**d**), (**e**) GBM, and XGBoost (**f**).

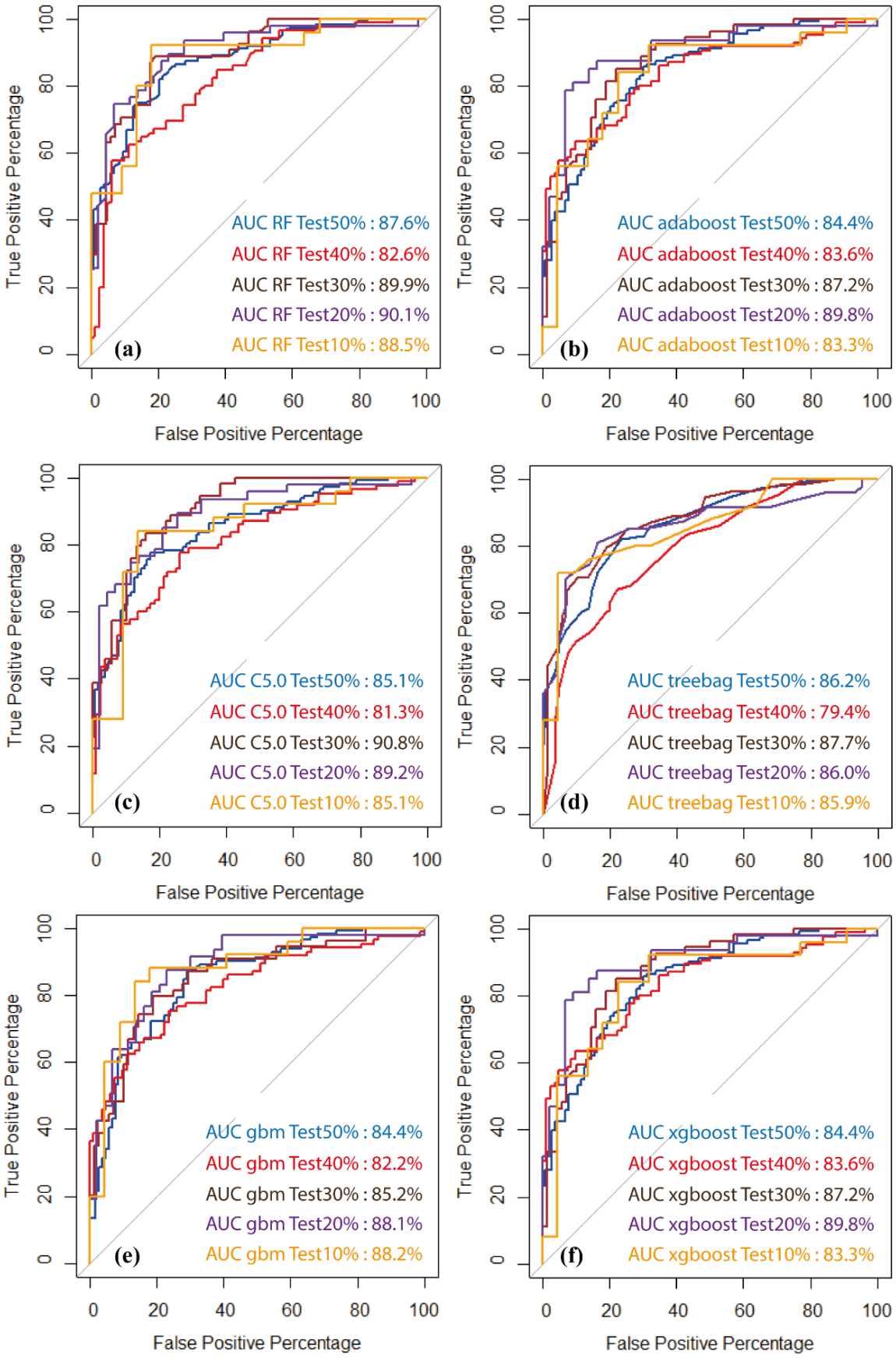


Figure 10. Prediction rate curve and the area under curve of the DT-based models using the testing dataset: RF (**a**), Adaboost (**b**), C5.0 (**c**), treebag (**d**), (**e**) GBM, and XGBoost (**f**).

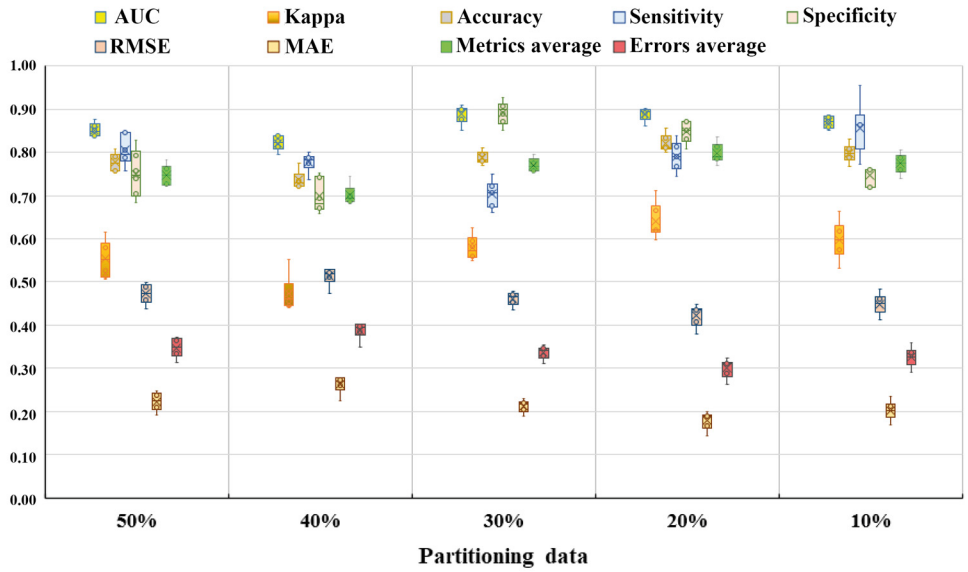


Figure 11. Validation results and metric parameters calculation of all models in each splitting quantity using testing data.

4. Discussion

In this section, the results are discussed in three parts: (i) The analysis of the models' performances; (ii) the investigation of the importance of each geo-environmental factor in the modelling of gully erosion; and (iii) the analysis of gully erosion vulnerability mapping results.

4.1. Accuracy Assessment and Comparison

The performance of the modelling is based on two fundamental aspects: Discrimination and reliability. In this respect, the evaluation of the performance of the GE sensitivity models was carried out according to five random subdivisions (50/50%, 60/40%, 70/30%, 80/20%, and 90/10%) by 10-fold cross-validation through several statistical metrics, namely the Kappa index, AUC, RMSE, and MAE.

In terms of prediction accuracy, the C5.0-70/30% model performed the best (AUC = 90.8), followed by the RF-80/20% model (AUC = 90.1), the Adaboost-70/30% model (AUC = 90), the XGBoost-80/20% model (AUC = 89.8), the GBM-80/20% model (AUC = 88.2), and the treebag-70/30% model (AUC = 87.7). These precision values indicate that all of the models utilized demonstrated a high level of performance and robustness, making them applicable to a variety of study domains and the monitoring and evaluation of natural hazards such as soil erosion, landslides, floods, and others [25,57,58]. To confirm this performance, however, the use of a single accuracy indicator may increase the margin of error, leading to potentially inaccurate results [59]. In this regard, the determination of additional accuracy indices such as RMSE, MAE, and the Kappa index can bolster the validity of the employed models [60]. Although the values of AUC and Kappa in terms of the discrimination index are greater in the present study, the values of RMSE and MAE are lower, indicating that the majority of gully inventory points were recognized on the final gully erosion sensitivity maps, reflecting the accuracy of the used models.

Moreover, despite the fact that these decision tree models are highly intuitive and do not necessitate a great deal of work in the preparation and processing of the database, they do require some effort. The obtained results indicate that the accuracy measures have high sensitivity to random database partitioning. Nonetheless, the majority of models perform better at the 70/30% and 80/20% subdivision levels, indicating that one of the disadvantages of decision tree-based models is that a simple change in the database can result in a change in the general structure of the decision tree and, as a result, model

instability. For this reason, it is necessary to test multiple subdivisions in conjunction with 10-fold cross-validation to select a more accurate prediction model.

In the case of managing natural hazards such as gully erosion, the primary goal of the manager is to identify high-risk regions. However, the cost and time required to accomplish this goal are extremely significant. Consequently, the adoption of predictive models can be advantageous in terms of costs and resources mobilized to solve such an issue, since these models enable managers to concentrate on management priorities, thereby enhancing the efficiency of decision making.

4.2. Geoenvironmental Variable Importance Analysis

Several studies have highlighted that a large database is necessary for obtaining accurate results and a more accurate prediction of gully-vulnerable locations [61]. For this purpose, in this work, 17 factors were utilized to build GESMs, including topographical, hydrological, geomorphological, climatological, and soil-property-associated factors. The integration of these parameters with inventory data facilitates the identification of regions with a high risk of gully erosion.

According to five random subdivisions (50, 60, 70, 80, and 90%) of the model training database and using two measures (the average decrease in accuracy and the average decrease in precision), the RF technique determined the importance of the factors. The overall examination of these results revealed that all influencing variables contribute to gully formation. Furthermore, lithology, elevation, geomorphic factors, and LULC are the most significant contributors. This is consistent with the mountainous character of the study area and also demonstrates the visible influence of human interference on natural ecosystems on the acceleration of soil erosion. This is because the combination of highly friable lithologies such as clays and marls, high altitudes, and degraded vegetation cover facilitates gully development, particularly on steep slopes and in places with damaged vegetation cover [62]. Multiple investigations in comparable circumstances have confirmed that these variables effectively regulate the degree of soil particle detachment and gully formation vulnerability [33]. Furthermore, the LULC factor refers to human activities and natural land surface changes. In addition, the lack of a viable alternative economic sector for the local population, other than forest exploitation, significantly exacerbates soil erosion (wood, pasture, etc.). Therefore, people strive to make a living by clearing, overgrazing, and overexploiting firewood in order to satisfy the significant rise in demand for arable land [63].

In other words, areas covered by friable lithologies such as clays and marls are the most susceptible to soil particle detachment [64]; therefore, vegetation cover protects the soil, and its degradation increases the likelihood of gully formation [62]. In addition, research on the effect of topographic parameters on gully formation in arid and semi-arid contexts has revealed the existence of direct and indirect impacts of topographic circumstances on the evolution of vegetation cover, rainfall, and runoff kinetic energy [65–68]. In reality, topographical features influence the local climate, which is characterized by geographically and temporally localized rainfall events, therefore places with steep slopes, such as hillsides, are characterized by high runoff velocities. This results in soil saturation, a substantial separation of soil particles, and the creation of ravines. The geomorphic element, which is also of major importance, verified this. This feature, which enables the mapping of slope units [69] and demonstrates that the majority of gullies are related to slopes and depressions, validates the effect of topography on the expression of erosive processes in mountainous regions.

Finally, all factors demonstrated significance in predicting and identifying regions with a high vulnerability to gully erosion; however, only the LS component was excluded because it was inconsistent with the other topographic variables.

4.3. Gully Erosion VulnerAbility Maps

Taking into consideration the subdivision where each model performs best, various models were used to develop vulnerability maps. The findings reveal that certain factors

influence the spatial variability in vulnerability more strongly than others. Thus, the rise in the proportion of the most susceptible regions from upstream to downstream of a basin is directly attributable to the topographical impact. This is consistent with the substantial geographical link between these sites and classes with slopes above 27 degrees, TRI > 22, as well as slopes and Hollow units in Geomorhons' factor. Moreover, precipitation and LULC seem to be significant elements in regulating gully development, which is why all models anticipate that gully formation will be greatest in regions with high precipitation and degraded vegetation cover. These conclusions are comparable to those of earlier studies conducted in specific localities of the High Atlas, Morocco [70,71].

Comparing the maps generated by the various models, it is evident that the Adaboost model predicts more susceptible regions than the other models, especially in comparison to the XGBoost model, which predicts the fewest vulnerable areas. In general, the differences between the predictions of the models are limited; this is evident in Figure 7, where the areas highly susceptible to gully formation were predicted almost identically by all six models (Figure 7—Areas 2 and 3); however, for the low-vulnerability areas, only minor differences between XGBoost and the other models can be observed (Figure 7—Area 1). In general, the results of this study using RF, C5.0, Adaboost, XGBoost, treebag, and GBM models demonstrates that machine learning methods are capable of producing GESMs with great precision. This can be viewed as a fundamental tool to aid planners and managers in ensuring the sustainable and effective management of soil erosion-affected areas in a semi-arid mountain setting.

5. Conclusions

Gully erosion is a phenomenon of great complexity. To ensure appropriate management of this phenomenon, it is vital to comprehend the geographical distribution of gullies and detect regions with a high possibility of gully formation. Six decision tree models based on machine learning algorithms (Random Forest (RF), C5.0, XGBoost, 18 treebag, Gradient Boosting Machines (GBMs), and Adaboost) were tested to determine the role of 17 parameters in gully formation in a semi-arid environment with a hilly character and to test their stability in response to the changing splitting quantities in input data. The outcome was six erosion vulnerability maps for gullies. The examination of these results demonstrates that all the utilized models are robust and extremely reliable at predicting and identifying the sensitivity to gully erosion and that the most influential factors are Lithology, LandUse-LandCover (LULC), Geomorphons, and Elevation factors. In addition, the analysis of factors and their effects on gully formation and soil degradation revealed that topographical factors, such as geomorphological units and valley depths, play a significant role in the formation of gullies in this mountain environment. The validation of these results is likewise satisfactory, as they demonstrate congruence between the regions predicted by the ML models and the inventory points recovered from the real field data. This substantiates the accuracy of the predicted gullies' future results. The results also confirmed the need to test the performance of the models under many subdivisions of the input data in order to build a more accurate and stable model in terms of prediction. In this semi-arid highland context, the vulnerability maps generated have been shown to be a valuable tool for the sustainable management and planning of gully-erosion-affected areas.

Author Contributions: Conceptualization, H.E., M.H. and H.R.; methodology, H.E., M.H., M.N. and M.I.; software, H.E., M.I., M.O. and S.K.; validation, M.H., H.R. and L.B.; formal analysis, H.E.; investigation, H.E., M.N., S.K. and M.O.; resources, H.E., M.H., H.R. and L.B.; data curation, H.E., M.I., S.K. and M.O.; writing—original draft preparation, H.E.; writing—review and editing, H.E., M.H., H.R. and L.B.; visualization, H.E. and M.N.; supervision, M.H.; project administration, L.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: This work was carried out within the CHARISMA Project with the assistance of the Hassan II Academy of Science and Technology. This project is supported by GeanTech project funded by OCP Foundation. We acknowledge special support from IRD representation in Morocco.

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

References

1. Poesen, J.; Nachtergaele, J.; Verstraeten, G.; Valentin, C. Gully erosion and environmental change: Importance and research needs. *Catena* **2003**, *50*, 91–133. [CrossRef]
2. Roy, P.; Chandra Pal, S.; Arabameri, A.; Chakrabortty, R.; Pradhan, B.; Chowdhuri, I.; Tien Bui, D. Novel ensemble of multivariate adaptive regression spline with spatial logistic regression and boosted regression tree for gully erosion susceptibility. *Remote Sens.* **2020**, *12*, 3284. [CrossRef]
3. Li, Z.; Fang, H. Impacts of climate change on water erosion: A review. *Earth-Sci. Rev.* **2016**, *163*, 94–117. [CrossRef]
4. Zabihi, M.; Pourghasemi, H.R.; Motevalli, A.; Zakeri, M.A. Gully erosion modeling using GIS-based data mining techniques in Northern Iran: A comparison between boosted regression tree and multivariate adaptive regression spline. In *Natural Hazards GIS-Based Spatial Modeling Using Data Mining Techniques*; Springer: Cham, Switzerland, 2019; pp. 1–26. [CrossRef]
5. Gupta, G.S. Land degradation and challenges of food security. *Rev. Eur. Stud.* **2019**, *11*, 63. [CrossRef]
6. Borrelli, P.; Robinson, D.A.; Panagos, P.; Lugato, E.; Yang, J.E.; Alewell, C.; Wuepper, D.; Montanarella, L.; Ballabio, C. Land use and climate change impacts on global soil erosion by water (2015–2070). *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 21994–22001. [CrossRef]
7. FAO. *Global Soil Status, Processes and Trends. Status of the World's Soil Resources (SWSR)—Main Report of the Food and Agriculture Organization*; FAO: New York, NY, USA, 2015.
8. Acharki, S.; El Qorchi, F.; Arjdal, Y.; Amharref, M.; Bernoussi, A.S.; Aissa, H.B. Soil erosion assessment in Northwestern Morocco. *Remote Sens. Appl. Soc. Environ.* **2022**, *25*, 100663. [CrossRef]
9. Markhi, A.; Laftouhi, N.; Grusson, Y.; Soulaimani, A. Assessment of potential soil erosion and sediment yield in the semi-arid N′ fis basin (High Atlas, Morocco) using the SWAT model. *Acta Geophys.* **2019**, *67*, 263–272. [CrossRef]
10. 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.* **2013**, *46*, 33–57. [CrossRef]
11. Smith, S.J.; Williams, J.R.; Menzel, R.G.; Coleman, G.A. Prediction of sediment yield from southern plains grasslandds with the modified universal soil loss equation. *J. Range Manag.* **1984**, *37*, 295–297. [CrossRef]
12. Renard, K.G.; Foster, G.R.; Weesies, G.A.; Porter, J.P. RUSLE, revised universal soil loss equation. *J. Soil Water Conserv.* **1991**, *46*, 30–33.
13. Flanagan, D.C.; Nearing, M.A. *USDA-Water Erosion Prediction Project: Hill Slope and Watershed Model Documentation. NSERI Report No. 10*; USDA-ARS National Soil Erosion Research Laboratory: West Lafayette, IN, USA, 1995.
14. Wischmeier, W.H.; Smith, D.D. *Predicting Rainfall Erosion Losses: A Guide to Conservation Planning. Agriculture Handbook. 282*; USDA-ARS: Beltsville, MA, USA, 1978.
15. Williams, J.R.; Jones, C.A.; Dyke, P.T. *The EPIC Model. United States Department of Agriculture (USDA) Teachnical Bulletin No. 1768*; United States Department of Agriculture: Washington, DC, USA, 1990.
16. Gayen, A.; Saha, S. Application of weights-of-evidence (WoE) and evidential belief function (EBF) models for the delineation of soil erosion vulnerable zones: A study on Pathro river basin, Jharkhand, India. *Model. Earth Syst. Environ.* **2017**, *3*, 1123–1139. [CrossRef]
17. Alewell, C.; Borrelli, P.; Meusburger, K.; Panagos, P. Using the USLE: Chances, challenges and limitations of soil erosion modelling. *Int. Soil Water Conserv. Res.* **2019**, *7*, 203–225. [CrossRef]
18. Luca, F.; Conforti, M.; Robustelli, G. Comparison of GIS-based gullying susceptibility mapping using bivariate and multivariate statistics: Northern Calabria, South Italy. *Geomorphology* **2011**, *134*, 297–308. [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.* **1991**, *37*, 607–619. [CrossRef]
20. Conoscenti, C.; Angileri, S.; Cappadonia, C.; Rotigliano, E.; Agnesi, V.; Marker, M. Gully erosion susceptibility assessment by means of GIS-based logistic regression: A case of Sicily (Italy). *Geomorphology* **2014**, *204*, 399–411. [CrossRef]
21. Dube, F.; Nhapi, I.; Murwira, A.; Gumindoga, W.; Goldin, J.; Mashauri, D.A. Potential of weight of evidence modelling for gully erosion hazard assessment in Mbire District—Zimbabwe. *Phys. Chem. Earth* **2014**, *67*, 145–152. [CrossRef]
22. Zakerinejad, R.; Maerker, M. An integrated assessment of soil erosion dynamics with special emphasis on gully erosion in the Mazayjan basin, southwestern Iran. *Nat. Hazards* **2015**, *79*, 25–50. [CrossRef]
23. Du Plessis, C.; Van Zijl, G.; Van Tol, J.; Manyevere, A. Machine learning digital soil mapping to inform gully erosion mitigation measures in the Eastern Cape, South Africa. *Geoderma* **2020**, *368*, 114287. [CrossRef]
24. Zhao, X.; Chen, W. Gis-based evaluation of landslide susceptibility models using certainty factors and functional trees-based ensemble techniques. *Appl. Sci.* **2020**, *10*, 16. [CrossRef]

25. Sahour, H.; Gholami, V.; Vazifedan, M. A comparative analysis of statistical and machine learning techniques for mapping the spatial distribution of groundwater salinity in a coastal aquifer. *J. Hydrol.* **2020**, *591*, 125321. [CrossRef]
26. Marjanović, M.; Kovačević, M.; Bajat, B.; Voženílek, V. Landslide susceptibility assessment using SVM machine learning algorithm. *Eng. Geol.* **2011**, *123*, 225–234. [CrossRef]
27. Chen, W.; Lei, X.; Chakrabortty, R.; Pal, S.C.; Sahana, M.; Janizadeh, S. Evaluation of different boosting ensemble machine learning models and novel deep learning and boosting framework for head-cut gully erosion susceptibility. *J. Environ. Manag.* **2021**, *284*, 112015. [CrossRef]
28. Alaboz, P.; Dengiz, O.; Demir, S.; Şenol, H. Digital mapping of soil erodibility factors based on decision tree using geostatistical approaches in terrestrial ecosystem. *Catena* **2021**, *207*, 105634. [CrossRef]
29. Pal, S.C.; Chakrabortty, R.; Arabameri, A.; Santosh, M.; Saha, A.; Chowdhuri, I.; Roy, P.; Shit, M. Chemical weathering and gully erosion causing land degradation in a complex river basin of Eastern India: An integrated field, analytical and artificial intelligence approach. *Nat. Hazards* **2022**, *110*, 847–879. [CrossRef]
30. Saha, S.; Roy, J.; Arabameri, A.; Blaschke, T.; Tien Bui, D. Machine Learning-Based Gully Erosion Susceptibility Mapping: A Case Study of Eastern India. *Sensors* **2020**, *20*, 1313. [CrossRef]
31. Pourghasemi, H.R.; Sadhasivam, N.; Kariminejad, N.; Collins, A.L. Gully erosion spatial modelling: Role of machine learning algorithms in selection of the best controlling factors and modelling process. *Geosci. Front.* **2020**, *11*, 2207–2219. [CrossRef]
32. Tiwari, A.; Arun, G.; Vishwakarma, B.D. Parameter importance assessment improves efficacy of machine learning methods for predicting snow avalanche sites in Leh-Manali Highway, India. *Sci. Total Environ.* **2021**, *794*, 148738. [CrossRef]
33. Rahmati, O.; Tahmasebipour, N.; Haghizadeh, A.; Pourghasemi, H.R.; Feizizadeh, B. Evaluation of different machine learning models for predicting and mapping the susceptibility of gully erosion. *Geomorphology* **2017**, *298*, 118–137. [CrossRef]
34. Conforti, M.; Aucelli, P.; Robustelli, G.; Scarciglia, F. Geomorphology and GIS analysis for mapping gully erosion susceptibility in the Turbolo Stream catchment (Northern Calabria, Italy). *Nat. Hazards* **2011**, *56*, 881–898. [CrossRef]
35. Sharma, M.; Garg, R.D.; Badenko, V.; Fedotov, A.; Min, L.; Yao, A. Potential of airborne LiDAR data for terrain parameters extraction. *Quat. Int.* **2021**, *575*, 317–327. [CrossRef]
36. Holloway, J.; Rudy, A.; Lamoureux, S.; Treitz, P. Determining the terrain characteristics related to the surface expression of subsurface water pressurization in permafrost landscapes using susceptibility modelling. *Cryosphere* **2017**, *11*, 1403–1415. [CrossRef]
37. Gutiérrez, Á.G.; Schnabel, S.; Contador, F.L. Gully erosion, land use and topographical thresholds during the last 60 years in a small rangeland catchment in SW Spain. *Land Degrad. Dev.* **2009**, *20*, 535–550. [CrossRef]
38. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
39. Ravì, D.; Bober, M.; Farinella, G.M.; Guarnera, M.; Battiato, S. Semantic segmentation of images exploiting DCT based features and random forest. *Pattern Recognit.* **2016**, *52*, 260–273. [CrossRef]
40. Zhang, G.; Cai, Y.; Zheng, Z.; Zhen, J.; Liu, Y.; Huang, K. Integration of the Statistical Index Method and the Analytic Hierarchy Process technique for the assessment of landslide susceptibility in Huizhou, China. *Catena* **2016**, *142*, 233–244. [CrossRef]
41. Pandya, R.; Pandya, J. C5. 0 algorithm to improved decision tree with feature selection and reduced error pruning. *Int. J. Comput. Appl.* **2015**, *117*, 18–21. [CrossRef]
42. Putra, F.; Sitanggang, I. Classification model of air quality in Jakarta using decision tree algorithm based on air pollutant standard index. *IOP Conf. Ser. Earth Environ. Sci.* **2020**, *528*, 012053. [CrossRef]
43. Pham, B.T.; Nguyen, M.D.; Nguyen-Thoi, T.; Ho, L.S.; Koopialipoor, M.; Kim Quoc, N.; Armaghani, D.J.; Le, H.V. A novel approach for classification of soils based on laboratory tests using Adaboost, Tree and ANN modeling. *Transp. Geotech.* **2021**, *27*, 100508. [CrossRef]
44. Freund, Y.; Schapire, R.E. Experiments with a new boosting algorithm. *ICML* **1996**, *96*, 148–156.
45. West, D.; Dellana, S.; Qian, J. Neural network ensemble strategies for financial decision applications. *Comput. Oper. Res. Appl. Neural Netw.* **2005**, *32*, 2543–2559. [CrossRef]
46. Wang, S.; Mathew, A.; Chen, Y.; Xi, L.; Ma, L.; Lee, J. Empirical analysis of support vector machine ensemble classifiers. *Expert Syst. Appl.* **2009**, *36*, 6466–6476. [CrossRef]
47. Hong, H.; Liu, J.; Bui, D.T.; Pradhan, B.; Acharya, T.D.; Pham, B.T.; Zhu, A.-X.; Chen, W.; Ahmad, B.B. Landslide susceptibility mapping using J48 Decision Tree with AdaBoost, Bagging and Rotation Forest ensembles in the Guangchang area (China). *Catena* **2018**, *163*, 399–413. [CrossRef]
48. Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123–140. [CrossRef]
49. Chan, J.C.-W.; Paelinckx, D. Evaluation of Random Forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. *Remote Sens. Environ.* **2008**, *112*, 2999–3011. [CrossRef]
50. Banfield, R.E. *Learning on Complex Simulations*; University of South Florida: Tampa, FL, USA, 2007.
51. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29*, 1189–1232. Available online: https://www.jstor.org/stable/2699986 (accessed on 10 May 2023). [CrossRef]
52. Sahin, E.K. Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGBoost, gradient boosting machine, and random forest. *SN Appl. Sci.* **2020**, *2*, 1308. [CrossRef]

53. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, San Francisco, CA, USA, 13–17 August 2016; Association for Computing Machinery: New York, NY, USA; pp. 785–794. [CrossRef]
54. Ramezan, C.A.; Warner, T.A.; Maxwell, A.E. Evaluation of sampling and cross-validation tuning strategies for regional-scale machine learning classification. *Remote Sens.* **2019**, *11*, 185. [CrossRef]
55. Breiman, L.; Cutler, A. A deterministic algorithm for global optimization. *Math. Program.* **1993**, *58*, 179–199. [CrossRef]
56. Lee, S.; Pradhan, B. Landslide hazard mapping at Selangor, Malaysia using frequency ratio and logistic regression models. *Landslides* **2006**, *4*, 33–41. [CrossRef]
57. Guo, Z.; Shi, Y.; Huang, F.; Fan, X.; Huang, J. Landslide susceptibility zonation method based on C5. 0 decision tree and K-means cluster algorithms to improve the efficiency of risk management. *Geosci. Front.* **2021**, *12*, 101249. [CrossRef]
58. Masselink, R.H.; Temme, A.J.A.M.; Giménez Díaz, R.; Casalí Sarasíbar, J.; Keesstra, S.D. Assessing hillslope-channel connectivity in an agricultural catchment using rare-earth oxide tracers and random forests models. *Cuad. Investig. Geográfica* **2017**, *43*, 19–39. [CrossRef]
59. Tehrany, M.S.; Pradhan, B.; Mansor, S.; Ahmad, N. Flood susceptibility assessment using GIS-based support vector machine model with different kernel types. *Catena* **2015**, *125*, 91–101. [CrossRef]
60. Pham, B.T.; Prakash, I.; Singh, S.K.; Shirzadi, A.; Shahabi, H.; Tran, T.-T.T.; Tien Bui, D. Landslide susceptibility modeling using reduced error pruning trees and different ensemble techniques: Hybrid machine learning approaches. *Catena* **2019**, *175*, 203–218. [CrossRef]
61. Romer, C.; Ferentinou, M. Shallow landslide susceptibility assessment in a semiarid environment—A quaternary catchment of KwaZulu-Natal, South Africa. *Eng. Geol.* **2016**, *201*, 29–44. [CrossRef]
62. Arabameri, A.; Tiefenbacher, J.P.; Blaschke, T.; Pradhan, B.; Tien Bui, D. Morphometric analysis for soil erosion susceptibility mapping using novel gis-based ensemble model. *Remote Sens.* **2020**, *12*, 874. [CrossRef]
63. Bouzekraoui, H.; El Khalki, Y.; Mouaddine, A.; Lhissou, R.; El Youssi, M.; Barakat, A. Characterization and dynamics of agroforestry landscape using geospatial techniques and field survey: A case study in central High-Atlas (Morocco). *Agrofor. Syst.* **2016**, *90*, 965–978. [CrossRef]
64. Azareh, A.; Rahmati, O.; Rafiei-Sardooi, E.; Sankey, J.B.; Lee, S.; Shahabi, H.; Ahmad, B.B. Modelling gully-erosion susceptibility in a semi-arid region, Iran: Investigation of applicability of certainty factor and maximum entropy models. *Sci. Total Environ.* **2019**, *655*, 684–696. [CrossRef]
65. Nazari Samani, A.; Ahmadi, H.; Jafari, M.; Ghoddousi, J. Geomorphic threshold conditions for gully erosion in Southwestern Iran (Boushehr-Samal watershed). *J. Asian Earth Sci.* **2009**, *35*, 180–189. [CrossRef]
66. Bochet, E.; García-Fayos, P. Factors controlling vegetation establishment and water erosion on motorway slopes in Valencia, Spain. *Restor. Ecol.* **2004**, *12*, 166–174. [CrossRef]
67. Wang, L.; Wei, S.; Horton, R.; Shao, M.A. Effects of vegetation and slope aspect on water budget in the hill and gully region of the Loess Plateau of China. *Catena* **2011**, *87*, 90–100. [CrossRef]
68. Beullens, J.; Van de Velde, D.; Nyssen, J. Impact of slope aspect on hydrological rainfall and on the magnitude of rill erosion in Belgium and northern France. *Catena* **2014**, *114*, 129–139. [CrossRef]
69. Luo, W.; Liu, C.C. Innovative landslide susceptibility mapping supported by geomorphon and geographical detector methods. *Landslides* **2018**, *15*, 465–474. [CrossRef]
70. Barakat, A.; Rafai, M.; Mosaid, H.; Islam, M.S.; Saeed, S. Mapping of Water-Induced Soil Erosion Using Machine Learning Models: A Case Study of Oum Er Rbia Basin (Morocco). *Earth Syst. Environ.* **2022**, *7*, 151–170. [CrossRef]
71. Meliho, M.; Khattabi, A.; Mhammdi, N. A GIS-based approach for gully erosion susceptibility modelling using bivariate statistics methods in the Ourika watershed, Morocco. *Environ. Earth Sci.* **2018**, *77*, 655. [CrossRef]