land

MDPI

Article

Machine Learning Models for the Spatial Prediction of Gully Erosion Susceptibility in the Piraí Drainage Basin, Paraíba Do Sul Middle Valley, Southeast Brazil

Jorge da Paixão Marques Filho [1,*], Antônio José Teixeira Guerra [1], Carla Bernadete Madureira Cruz [1], Maria do Carmo Oliveira Jorge [1] and Colin A. Booth [2,*]

[1] Departament of Geography, Federal University of Rio de Janeiro, Rio de Janeiro 21910-240, Brazil; antonio.guerra@igeo.ufrj.br (A.J.T.G.); carlamad@igeo.ufrj.br (C.B.M.C.); maria.jorgerj@ccmn.ufrj.br (M.d.C.O.J.)

[2] School of Engineering, University of the West of England, Bristol BS16 1QY, UK

[*] Correspondence: marquesfilho.j.p@ufrj.br (J.d.P.M.F.); colin.booth@uwe.ac.uk (C.A.B.)

Abstract: Soil erosion is a global issue—with gully erosion recognized as one of the most important forms of land degradation. The purpose of this study is to compare and contrast the outcomes of four machine learning models, Classification and Regression (CART), eXtreme Gradient Boosting (XGBoost), Random Forest (RF), and Support Vector Machine (SVM), used for mapping susceptibility to soil gully erosion. The controlling factors of gully erosion in the Piraí Drainage Basin, Paraíba do Sul Middle Valley were analysed by image interpretation in Google Earth and gully erosion samples (n = 159) were used for modelling and spatial prediction. The XGBoost and RF models achieved identical results for the area under the receiver operating characteristic curve (AUROC = 88.50%), followed by the SVM and CART models, respectively (AUROC = 86.17%; AUROC = 85.11%). In all models analysed, the importance of the main controlling factors predominated among Lineaments, Land Use and Cover, Slope, Elevation and Rainfall, highlighting the need to understand the landscape. The XGBoost model, considering a smaller number of false negatives in spatial prediction, was considered the most appropriate, compared to the Random Forest model. It is noteworthy that the XGBoost model made it possible to validate the hypothesis of the study area, for susceptibility to gully erosion and identifying that 9.47% of the Piraí Drainage Basin is susceptible to gully erosion. Furthermore, replicable methodologies are evidenced by their rapid applicability at different scales.

Keywords: gully erosion susceptibility; land degradation; machine learning; spatial modelling

check for updates

Citation: Filho, J.d.P.M.; Guerra, A.J.T.; Cruz, C.B.M.; Jorge, M.d.C.O.; Booth, C.A. Machine Learning Models for the Spatial Prediction of Gully Erosion Susceptibility in the Piraí Drainage Basin, Paraíba Do Sul Middle Valley, Southeast Brazil. *Land* **2024**, *13*, 1665. https://doi.org/10.3390/land13101665

Academic Editors: Nejc Bezak, Carla Ferreira, Vesna Zupanc and Hanoch Lavee

Received: 31 August 2024
Revised: 30 September 2024
Accepted: 9 October 2024
Published: 13 October 2024

1. Introduction

Soil erosion processes are influenced by precipitation (intensity and amount), slope angle, land use and management, and soil properties. The clearance of vegetation, including deforestation, and agricultural activities are determining factors for the occurrence of soil erosion and surface runoff, induced by the rainfall regime [1]. Depending on the specificity of the controlling factors, erosion can be categorized as sheet, rill and gully erosion. However, in this research work we will only address gully erosion.

A gully can be defined as an erosive incision, in unconsolidated materials, resulting from the concentration of water flows, following intense rains or ice melting. Furthermore, gullies are erosive features with depths > 0.5 m and cannot be obliterated by agricultural machinery [2]. Gully erosion has been recognized worldwide as one of the most important forms of land degradation. It can be found in rural or urban areas.

In this sense, several national and international authors have drawn attention to this type of erosive process, which in addition to causing effects in the place where they occur, also, through surface and subsurface runoff, causes silting of water bodies and areas located further downstream [3–5].

The process in which a given landscape is established is dynamic, unstable, unrepeatable and occurs in a portion of the geographic space [6]. In this sense, understanding natural laws, the portion of geographic space and the history of formation of environmental conditions are ways of understanding the landscape.

Agriculture and pasture are activities that can result in different forms of land degradation, especially when these activities do not consider the limits imposed by the environments, with regard to gully erosion. As a consequence, in addition to the loss of land for rural activities, there is an increase in river and reservoir silting [7,8].

Due to the complexity inherent in the occurrence of gullies, methods such as multicriteria decision analysis by Geographic Information Systems (GIS-MCDA) [9] and machine learning [10] have become recurrent. GIS-MCDA can be understood as an approach that contributes to decision-making, combining data and geospatial considerations, according to their importance in understanding this issue [11]. However, the subjectivity implicit in GIS-MCDA occurs mainly due to the attribution of weights to the analysis factors and results that are often less accurate, in relation to the machine learning method [12]. Furthermore, empirical models, usually due to the difficulty of dealing with multiple controlling factors, cannot correctly estimate areas susceptible to erosion [10].

Process-based, or deterministic, models do not allow for their application on large scales, especially due to their parameterization, which is fundamental as input data for the established equation [13,14]. In turn, stochastic models model a given natural system, through the distribution of probabilities across several variables and n values, producing several probable solutions and enabling the assessment of uncertainty, carried out by the modeler [14].

In this sense, due to the limitations for spatial modelling of models based on multicriteria decision analysis, due to their subjectivity and process-based or deterministic models, due to their difficulty in applying on a large scale, we opted for the use of stochastic models, for machine learning.

Machine learning is an empirical method that uses regression, or classification, and is recommended for problem solving, where theoretical knowledge is not yet consolidated [15] and for data analysis. In mapping gully erosion susceptibility, the machine learning method provides satisfactory results, such as prediction and metrics for model evaluation/validation [10,16].

Despite the importance of gully erosion, there has been minimal effort to develop reliable models for its formation and evolution. Therefore, using measurements of width, depth and length of gullies in the study area, together with the monitoring of several geomorphological features, this research work aims at mapping the susceptibility to erosion by gullies in Piraí drainage basin. In this sense, the purpose of the study was to compare and contrast four machine learning models, Classification and Regression (CART), eXtreme Gradient Boosting (XGBoost), Random Forest (RF), and Support Vector Machine (SVM) for gully erosion susceptibility mapping in the Piraí Drainage Basin, Paraíba do Sul Middle Valley.

2. Materials and Methods

2.1. Study Area Characterization

The Piraí drainage basin, located in Rio de Janeiro State (Figure 1), intersects the municipalities of Barra do Piraí, Engenheiro Paulo Frontin, Mendes, Piraí and Vassouras, totalling 1019.87 km^2. Furthermore, it is part of the context of the Depression of the Middle Valley Paraíba do Sul, which plays an important role as a regional base level. In Paraíba do Sul-Embu tectonic terrain, the Quirino complex stands out in this compartment, the largest outcropping area, especially in the Shear Zone of Paraíba do Sul River, with NE–SW direction. For example, reflected in the lineaments, concentrated downstream in the Piraí drainage basin.

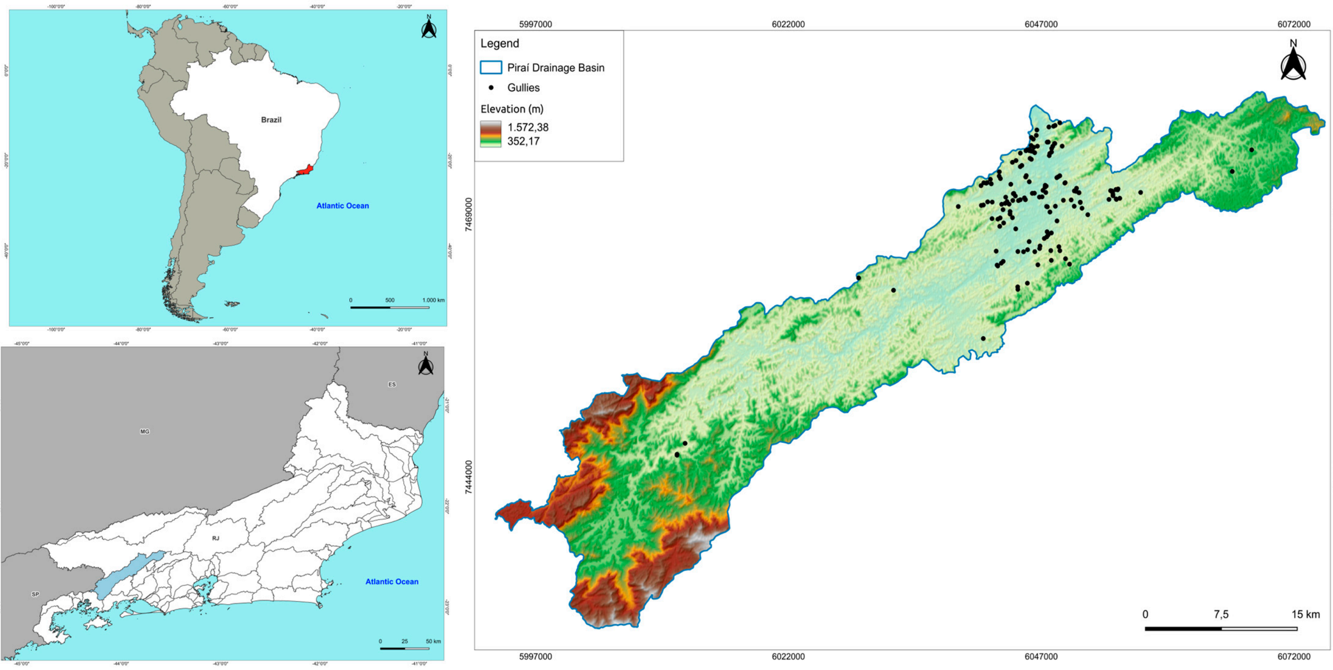

Figure 1. Study area localization and gully erosion sites.

This condition has caused a depressed surface delimited by Mantiqueira and Mar Mountain Ranges, due to the Cenozoic tectonics that originated the Southwest Brazil Continental Rift [17,18].

A depression of the Middle Vale Paraíba do Sul is a hemigraben, close to the Mantiqueira mountain alignments (Front) and Mar Mountain reverse. It is characterized by landforms of hilly domains, which may include low hills and landforms of hills and low hills and gentle to medium slopes [19]. On the hilly domain, in the middle Paraíba do Sul valley, as convex or convex–concave strands [18], active geomorphological processes occur, which are observed by intense rills and gullies, relief inversion, drainage captures, and structural concavities [20,21]. For the erosive or hydro-erosive process to occur, a slope angle is necessary 3° [1].

In relation to the depression of Paraíba do Sul Middle Valley, where the Piraí Drainage Basin is contained, Red–Yellow Oxisols, Red–Yellow Argisols, Yellow Oxisols, and Yellow Argisols dominate on gentle slopes [22]. The occurrence of sub-humid tropical climate domains, with annual rainfall averages between 1200 mm and 1800 mm [20] enabling the development of humid soils [18,22]. Furthermore, the Atlantic Forest biome is predominant and in the study area, isolated forest fragments refer to secondary formations, arising from the abandonment of agricultural areas [23].

The settlement of Paraíba do Sul middle valley has been characterized by several economic cycles, such as coffee growing and its subsequent replacement by dairy farming in the XIXth century, promoting soil depletion and accelerated erosion on the slopes, due to changes in the regional hydroclimatic dynamics [24]. Concavity's structural control, associated with the subsurface hydrological dynamics and soil use and management, favour the occurrence of erosion processes [19].

2.2. Methodological Procedures

2.2.1. Gully Inventory Data

Soil samples (n = 234) from gullies (Figure 2) were acquired by the Geological Survey of Brazil [25]. Although there are several classifications for gully measurements, it was decided to adopt the following criteria: >0.5 m for depth and >0.5 m for width, usually over 50 m long, with steep walls.

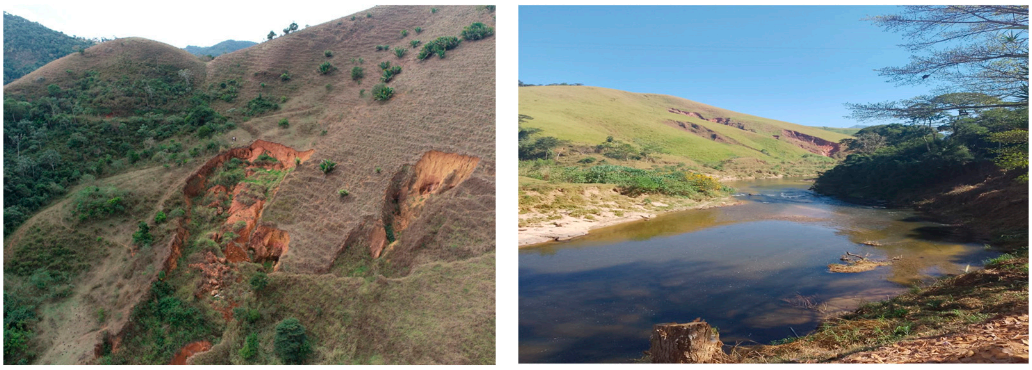

Figure 2. Photos illustrating the occurrence of gully erosion in the study area.

The main reason is that this classification can be used for both tropical and temperate environments, as several authors have highlighted in their articles [1,4,5,7]. Thus, using Google Earth, width and depth were measured, using the criteria mentioned above. Among the 234 samples evaluated, 75 had georeferencing problems and, as such, were excluded, with only 159 samples being used, representative of the total amount of the sample setting, without the data exclusion. Subsequently, further samples were generated without the occurrence of gullies, resulting in 159 samples randomly [26] to balance the previous sample set, totaling 318 samples.

In the present article, the samples referring to the non-occurrence of gullies were called non-eroded (0) and with occurrence, eroded (1) and divided between training (70%) and test (30%) samples, according to [10,16,26,27], for the predictive model of gully susceptibility.

Cross-validation or K-fold was adopted to increase the robustness of the models, especially regarding the randomness of the training data [28] by arbitrarily dividing the training and/or validation data into K classes [29]. In this sense, the training data were stratified into 4 K-Folds to enable cross-validation of the models for susceptibility to erosion by gullies and can be subdivided, according to the research needs and adopted in several studies on natural risks [30].

2.2.2. Multicollinearity in Controlling Factors for Gully Erosion

This section is subdivided—it provides a concise and precise description of the experimental results, their interpretation, as well as the experimental conclusions that can be drawn. In studies for susceptibility mapping, it is necessary to consider multicollinearity in the controlling factors, as it may possibly influence the predictive model [10]. In the present research study, the following thresholds were adopted: Variable Inflation Factors (VIF) < 10 and TOL > 0.1, adopted by [16,27]. VIF can be understood as the inverse of TOL, which is the residue of the given variables in a given multivariate regression [10].

To prepare the controlling factors, referring to geomorphological and hydrological dynamics, the Forest and Building Removed Digital Elevation Model [31] was used, with a spatial resolution of 30 m. For hydrological conditioning in the digital elevation model, in the calculation of land surface parameters, it was decided to apply the filling method to remove spurious or erroneous depressions [32]. In the analysis carried out by [33], digital elevation models, which undergo pre-processing, present more satisfactory results.

Geomorphometry is the science that aims to quantify and to analyse the earth's surface [34]. In the calculation of land surface parameters, such as dissection index, LS factor, profile curvature, plan curvature, slope angle, topographic ruggedness index, topographic wetness index [10], stream power index [16], and specific contribution area, the Whitebox package in R was used.

Elevation (z) is a land surface parameter or fundamental raw material for calculating various terrain attributes [35].

$$z = \mathrm{f}(\mathrm{x},\mathrm{y}) \tag{1}$$

where, z is elevation.

Dissection index is a modification referring to [36], which presupposes understanding the relationship between Absolute and Local Relief to assess the degree of denudation of landforms.

$$\text{Dissection Index} = \frac{(\mathrm{z} - \mathrm{z_{min}})}{(\mathrm{z_{max}} - \mathrm{z_{min}})} \tag{2}$$

where, z is elevation, z_{min} is the minimum elevation in a given 3 × 3 moving window and z_{max} is the maximum elevation in a given 3 × 3 moving window.

Local relief is the difference between the maximum and minimum elevations in a given portion, and is a metric that can indicate the denudation of landforms [37,38].

$$\text{Local Relief} = \mathrm{z_{max}} - \mathrm{z_{min}} \tag{3}$$

where, z is elevation, z_{min} is the minimum elevation in a given 3 × 3 moving window and z_{max} is the maximum elevation in a given 3 × 3 moving window.

Plan Curvature is the measurement between the convergence and divergence of flows on the surface and inside the soil, in which values < 0 are convergent and >0 are divergent [39,40].

$$\text{Plan Curvature} = \frac{\mathrm{q2r + 2pqs + p2t}}{\mathrm{(p2 + q1)}\sqrt{1 + \mathrm{p}^2 + \mathrm{q}^2}} \tag{4}$$

where, p is the derivative of z on the x–axis, q is the derivative of z on the y–axis, and s is the second derivative of z on the x– and y–axes.

Profile Curvature is the measurement between the acceleration and deceleration of flows on the surface and inside the soil, in which values < 0 are decelerated and >0, accelerated [39,40].

$$\text{Profile Curvature} = \frac{\mathrm{p2r + 2pqs + q2t}}{\mathrm{(p2 + q1)}\sqrt{\left(1 + \mathrm{p}^2 + \mathrm{q}^2\right)3}} \tag{5}$$

where, p is the derivative of z on the x–axis, q is the derivative of z on the y–axis, and s is the second derivative of z on the x– and y–axes.

The relationship between the length (l) of the slope and its slope angle(s) is estimated, as assumed by [41]. In which, in this equation, the conditions of erodibility and stability are considered, in order to estimate such conditions in a more appropriate way.

$$\text{Sediment Transport Index} = \left(\frac{\mathrm{a.}}{22.13}\right)^{0.4} \left(\frac{\mathrm{s}}{0.0896}\right)^{1.3} \tag{6}$$

where, a is the contribution area and s is the slope angle.

Slope angle or gradient of the terrain can be understood with the rate of change in elevation, in the axes or co-ordinates of X and Y [42].

$$\text{Slope Angle} = \arctan\sqrt{\mathrm{p}^2 + \mathrm{q}^2} \tag{7}$$

where, p is the derivative of z on the x–axis and q is the derivative of z on the y–axis.

Stream Power Index is the proportion or relationship between the contributing area and the slope angle of the land. It refers to the potential for erosive flow and processes in the landscape [40,43], where negative values correspond to areas with sediment accumulation and positive values, and steeper slopes with a risk of erosion [44].

$$\text{Stream Power Index} = \mathrm{In}\ (1 + \mathrm{a}\ .\ \tan(\mathrm{s})) \tag{8}$$

where, a is the specific contribution area and s is the slope radians.

Specific Catchment Area is the ratio of the contributing area over a portion of the length of the orientation of a given slope [40].

Terrain Roughness Index, proposed by [45], aims to quantify the heterogeneity of the terrain or relief units

$$\text{Topographic Ruggedness Index} = \mathrm{Y}\left[\sum(\mathrm{x_{ij}} - \mathrm{x_{00}})^2\right]^{\frac{1}{2}} \tag{9}$$

where, *p* is the maximum elevation in a given 3 × 3 moving window and q is the minimum elevation in a given moving window 3 × 3 and *s*.

The Topographic Wetness Index [46] presupposes understanding the spatial distribution of the effects of the Contribution Area, such as the hydrological response on the soils and, therefore, on the landforms, and in which positive values estimate humid or saturated areas and negative non-humid or unsaturated areas [47].

$$\text{Topographic Wetness Index} = \text{In}\left[\frac{\mathrm{a}}{\tan(\mathrm{s})}\right] \tag{10}$$

where, *a* is the specific contribution area and *s* is the slope radians.

The literature, regarding gully susceptibility mapping, controlling factors (such as lithology, rainfall, distance to highways, distance to rivers [10], land use and cover [26], and soils [45]) have been addressed [10,48]. In this sense, for the present research work, a geological and lineaments map was adopted at 1:400,000 scale [49], a land use and cover map at 1:100,000 scale [50], and a soils map at 1:250,000 scale [22].

To calculate the distance to highways and the distance to rivers, vector data were used, respectively from road sections and hydrography, at 1:25,000 scale [51]. In terms of rainfall, WorldClim climate data [52], version 2.1, were used, which have 1 km^2 spatial resolution, with a historical series between 1970–2000, and the average rainfall was subsequently calculated using map algebra. Subsequently, the data were resampled to a spatial resolution of 30 metres, and the data in vector format were also converted to matrix format at the same resolution. These procedures were carried out to make data resolutions compatible in the predictive model.

It is important to emphasise that for replicability and reproducibility purposes, in the dataset used, 11 controlling factors, such as the digital elevation model, which can calculate the parameters of the earth's surface and climate time series data (rainfall), are available free of charge across the entire globe. More specific data can be replaced by radiometric indices, as in vegetation and soil dynamics, therefore enabling the use of this methodology for different contexts of landscape structures.

In studies for susceptibility mapping, it is necessary to consider multicollinearity in the controlling factors, as it may possibly influence the predictive model [10]. In the present research study, the following thresholds were adopted, VIF < 10 and TOL > 0.1, adopted by [16,27], using the USDM package in R programming.

2.2.3. Thresholding in Controlling Factors for Gully Erosion

In order to structure the understanding of the specificities of the controlling factors in the study area, descriptive statistics metrics were applied, such as mean, standard deviation, minimum and maximum for continuous data and mode for discrete data, to identify the thresholds controlling factors (geomorphological, hydrological, and others) and locations, according to discrete and thematic data.

2.2.4. Machine Learning Models

Several studies such as [10,27,53–57] compare machine learning models for susceptibility to erosion by gullies, aiming to identify the most appropriate one. However, there is no consensus in the literature regarding in which situations a given model can be superimposed on another.

In this sense, the choice of the eXtreme Gradient Boosting, Random Forest, and Support Vector Machine models occur because they present the most satisfactory results in the analyses. Although, the Classification and Regression Tree model, predecessor and precursor of the Random Forest algorithm, known for some similarities, is rarely analysed or compared, hence its choice.

Classification and Regression Tree

The Classification and Regression Tree is a decision tree algorithm, its main difference being the possibility of using uncategorized variables, that is, not labelled but numerical and is based on decision-making logic (If-else) and regression analysis [58,59]. The functionality of this model is the creation of a given decision tree, with successive divisions in the training data, that is recursive and with a predefined limit, which after this process performs the appropriate labelling. Furthermore, the Classification and Regression Tree model is highly sensitive to training data [59].

The cf hyperparameter refers to the complexity of the model. In optimizing the single hyperparameter (cf), a sequence of values was used, using the grid search with four-fold-cross-validation method. The choice to use the Classification and Regression Tree model is due to a few comparative studies, with the exception of [53], which identified good performance, compared to the General Linear Model (GLM).

eXtreme Gradient Boosting

eXtreme Gradient Boosting (XGBoost) is a machine learning model based on gradient tree boosting that provides scalability and makes it possible to model complex relationships [60]. Typically, a decision tree for classification establishes rules to affect the separability and labelling of each instance, in this article for gully erosion, based on the predisposing (controlling) factors in a decision structure [54]. However, eXtreme Gradient Boosting produces a sequence of decision trees gradually and for each new tree constructed, the correction of previous errors is sought, and thus progressively more accurate predictions [54,55].

The hyperparameters, nrounds, the increase in iterations, max depth, the maximum depth of the decision tree, eta, the model learning rate, gamma, the minimum loss reduction, colsample by tree, the proportion of column subsamples, min child weight, the minimum sum of the instance weight and the subsample, and the percentage of the sub-sample were optimized using a sequence of values through the grid search with a four-fold-cross-validation method. Moreover, in several recent studies, the XGBoost model has demonstrated excellent performance for gully erosion susceptibility [27,54–56], thus highlighting its choice here.

Random Forest

The decision tree is a driving force for several applications, quickly, which presents significantly accurate results [61]. The Random Forest machine learning algorithm is a non-parametric and random classifier, which consists of classifiers structured in decision trees. Attributes are randomly distributed and each decision tree sends its unit's "vote" to differentiate a class [62].

The Random Forest model uses the premise of classification and regression trees (CART) to define each tree from a given sample and repeats k times to define the trees through a random subgroup of variables, in this case the controlling factors for susceptibility and erosion by gullies at each node [13]. The Random Forest model was chosen as it presents the most robust performance in gully erosion susceptibility models [10,53,56,57]. When using this model, the following hyperparameters were considered: mTry (2) which is the number of predictors sampled for splitting at each node. In optimizing the hyperparameter, a sequence of values was used, using the grid search with a four-fold-cross-validation method.

Hyperparameters such as ntree, referring to the number of decision trees and nodesize, and the minimum size of terminal nodes were defined automatically, based on the definition of mTry. Furthermore, it is possible to identify the most significant contributions of the most significant factors or variables in the model prediction. The Mean Decrease Gini calculates the impurity of the data, the probability of occurrence in classification, and the labelling of classes. Therefore, the Mean Decrease Gini is fundamental for identifying the importance of variables and the higher the value obtained by the factor, the greater its contribution to the predictive model and vice versa [62].

Support Vector Machine

The Support Vector Machine's functionality is to identify the well-established limit between two classes for subsequent classification [63]. Furthermore, it is a linear and generalized classifier, and is a popularly adopted method for regression problems and can be used for applications in data classifications [64].

In this model, the input data for training, labelled with a given label, fit into an optimal separating hyperplane, maximizing the margins of the boundaries between the two classes [27–64]. Among the hyperparameters of this model, sigma, which controls the non-linearity of the hyperplane, and the regularization (cost) hyperparameter, are fundamental to optimize the model and control overfitting [56]. To optimize the values of the hyperparameters, a sequence of values was used, using the grid search with a four-fold-cross-validation method. The Support Vector Machine model is promising for susceptibility to erosion by gullies [27], as well as an indicator by [56] and therefore adopted in the present study.

Machine Learning Models Implementation

The application of the Classification and Regression Tree, Random Forest, Support Vector Machine, and eXtreme Gradient Boosting machine learning models were done in R with the rpart, randomForest, e1071, and xgboost packages with the caret package.

For classification between the Non-Eroded (0) and Eroded (1) labels in the Machine Learning Models, and after classification, a probability matrix was generated for the recognition of spatial patterns of susceptibility to gully in R with terra package and after using the Natural Jenks method. Natural breaking is an approach that groups similar data and accentuates the differentiation between them [65].

2.2.5. Model Evaluations

Metrics for evaluating the predictive machine learning model, using a binary confusion matrix, consider the relationship between true and false positives and negatives [66]. According to [66], defining true positives or negatives corresponds to the identical labelling between the real and predicted classes. False positives correspond to the erroneous or incorrect classification of the predicted classes, while in false negatives a class is wrongly assigned for subsequent classification.

Among the most commonly used metrics in gully erosion susceptibility mapping, accuracy, precision, recall, F1-Score, and ROC Curve stand out. Accuracy corresponds to the degree of measurement between modelling and reality, precision refers to the proportion between true positives and sums that were predicted as positive in the model. Recall is similar to precision, however it only considers the sum of truly positive data, and the F1-Score is the harmonic mean between precision and recall, that is, if the F1-Score has higher values, the more robust the relationship is between the two metrics [67].

$$\text{Accuracy} = \frac{\text{TP} - \text{FN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{11}$$

where, TP = true positive, TN = true negative, FP = false positive, and FN = false negative

$$\text{Precision} = \frac{\text{TP}}{\text{TP} - \text{FP}} \tag{12}$$

where, TP = true positive and FP = false positive

$$\text{Recall} = \frac{\text{TP}}{\text{TP} - \text{FN}} \quad (13)$$

where, TP = true positive and FN = false negative

$$\text{F1-Score} = \frac{\text{TP}}{\text{TP} - \frac{1}{2}(\text{FP} - \text{FN})} \quad (14)$$

where, TP = true positive, FP = false positive, and FN = false negative

The ROC Curve is defined as the relationship between sensitivity and specificity, where the first is the proportion of positives classified correctly, and the second the proportion of negatives classified correctly [67]. Metrics to evaluate its performance were done in R with caret and pROC packages. For specificity, spatial patterns in gully occurrence in susceptibility mapping addressed the Natural Jenks approach in ArcGIS 10.8 software.

2.2.6. Minimum Mapping Unit

Due to the multiplicity of data scales of the controlling factors and ensuring the reliability of the proposed mapping scale, the procedure was carried out to identify the most appropriate minimum mappable area. The minimum mapping unit is the smallest feature that can be captured when imaging by remote sensing [68].

For this, the matrix cartographic generalization proposed by [69] was used, which suggests the calculation between the ratio of the minimum mappable area (40 ha) referring to the mapping scale, 1:250,000, and the size of the pixel area, 900 m^2. We enabled the appropriate choice of the movable window, in this case 5 × 5, as it is close to the result of the number of pixels relative to the previously mentioned ratio, 27.7 pixels.

3. Results

3.1. Multicollinearity in Controlling Factors for Gully Erosion

In Table 1, analyses of the VIF and TOL statistical tests to identify correlations between the controlling factors were performed. Among the 18 controlling factors, Dissection Index, Local Relief, Topographic Ruggedness Index, and Sediment Transport Index presented collinearity and were excluded. According to the data below on the controlling factors, no significant correlation or connections were identified in the variables used to gully erosion susceptibility mapping.

Table 1. Results of multicollinearity tests in controlling factors.

Controlling Factors	VIF	TOL
Elevation	4.7	0.21
Distance to Rivers	1.20	0.83
Distance to Roads	1.26	0.79
Land Use and Land Cover	1.17	0.85
Lineaments	1.3	0.77
Lithology	1.09	0.92
Plan Curvature	1.08	0.92
Profile Curvature	1.20	0.83
Rainfall	3.21	0.31
Slope	4.32	0.23
Specific Contributing Area	3.21	0.31
Stream Power Index	4.17	0.24
Soils	1.72	0.58
Topographic Wetness Index	2.74	0.36

3.2. Thresholding in Controlling Factors for Gully Erosion

In Table 2, it is possible to understand how the predominance of gully erosion occurs in the Piraí Drainage Basin, Paraíba do Sul Middle Valley, where, in the prerogative of geomorphological and hydrological thresholds, the gullies stand out in a range with low elevation, compared to the altimetric extension of 1572.88 m. In relation, the density of lineaments has a considerable concentration of faults and fractures.

Table 2. Exploratory statistics of susceptibility thresholds for gully erosion in controlling factors.

Controlling Factors	Mean	Min	Max	Mode
Elevation	448.15 m	383.71 m	592.14 m	-
Distance to Rivers	-	-	-	0–100 m
Distance to Roads	-	-	-	0–100 m
Land Use and Land Cover	-	-	-	Pasture
Lineaments	164,242.59 px/km^2	0 px/km^2	200,902.26 px/km^2	-
Lithology	-	-	-	Rio Turvo Suite
Plan Curvature	0.0022 m^{-1}	−0.0226 m^{-1}	0.0255 m^{-1}	-
Profile Curvature	0.0004 m^{-1}	−0.0059°/m	0.0043°/m	-
Rainfall	1205.48 mm	1172.38 mm	1272.87 mm	-
Slope	22.63°	6.09°	33.07°	-
Specific Contributing Area	45.05 m^2/m	29.12 m^2/m	232.96 m^2/m	-
Stream Power Index	18.46	3.11	108.22	-
Soils	-	-	-	Red–Yellow Argisols
Topographic Wetness Index	4.53	3.80	6.62	-

With the curvatures in profile and plan, it is possible to conceive that the predominance of convex-convex or divergent slope shapes, according to equations 4 and 5 and the slope, occurs in a range between low to medium slope and reference. In the flow power index, it is clear that, due to the predominance of values, as indicated in the literature, they indicate a greater risk to erosion, in this case susceptibility to erosion by gullying.

Through the specific contribution area that is directly related to the topographic humidity index, it is possible to identify through the positive values of the index that are located in humid or saturated areas.

It is possible to understand the occurrence of gullies, as approximately every 100 m from a given gully there is a drainage or a highway. Furthermore, pasture predominates as the main land use and cover, in favour of gullies. In relation to lithology and soils, the Rio Turvo suite with Granitoids and Orthogneisses and Red–Yellow Argisols predominates.

3.3. Performance of Machine Learning Models

Table 2 and Figure 3 show the results regarding the metrics used to evaluate the four predictive models for mapping susceptibility to gully erosion. To evaluate the performance of the models, we chose to use five performance metrics—Accuracy, Precision, Recall, F1-Score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC).

The Classification and Regression Tree model obtained the least satisfactory result in Accuracy (0.851) and AUC-ROC (85.11%) and in the other performance metrics compared, and in the two classes Non-Eroded and Eroded (1). In this model, in the Eroded class, it is Precision (0.836) and lower than Recall (0.873), indicating the model's underestimation of the occurrence of gullies and in the Non-Eroded class, the opposite situation occurs, respectively Precision (0.866) and Recall (0.829), identifying the overestimation of the non-occurrence of gullies in the study area.

The eXtreme Gradient Boosting (XGBoost) and Random Forest models achieved the most robust results, referring to the five performance metrics adopted, however, each of these models presents a peculiarity that is reflected in the underestimation or overestimation of the occurrence of gullies, with identical values in Accuracy (0.882) and AUC-ROC (88.30%), representing an excellent predictive capacity. By evaluating the metrics, it is possible to understand the performance of the model and it is clear that Precision in both models, there is an underestimation of the occurrence for the emergence of gullies (Eroded), respectively, XGBoost (0.860) and RF (0.875), and Recall XGBoost (0.914) and RF (0.893).

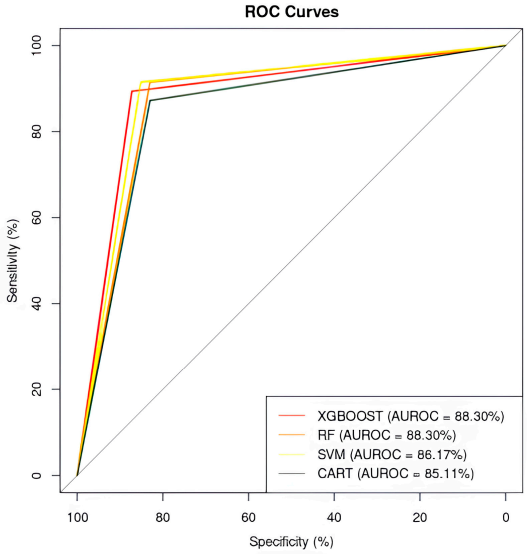

Figure 3. Comparison between CART: classification and regression tree model; XGBoost: extreme gradient boosting model; RF: random forest model and SVM: support vector machine model using receiver operating characteristic curves (AUROC).

These two values allow us to identify that, in the XGBoost model, the largest overestimation occurs in the Non-Eroded class, referring to Precision, XGBoost (0.909), RF (0.891), Recall XGBoost (0.851), and RF (0.872). In the Support Vector Machine model, with Accuracy (0.861) and AUC-ROC (86.17%), it presents the same behavior regarding the underestimation of the occurrence of gully (Eroded), with Precision (0.826) and Recall (0.914) of overestimation (Non-Eroded), and Precision (0.904) and Recall (0.808), being the model with the highest underestimation and overestimation for the gully erosion susceptibility model (Table 3).

Table 3. Comparison between performance metrics of machine learning models.

		Performance Metrics			
Model		**Accuracy**	**Precision**	**Recall**	**F1-Score**
CART	Non-Eroded Eroded	0.851	0.866 0.836	0.829 0.873	0.847 0.854
XGBoost	Non-Eroded Eroded	0.882	0.909 0.860	0.851 0.914	0.879 0.886
RF	Non-Eroded Eroded	0.882	0.891 0.875	0.872 0.893	0.881 0.884
SVM	Non-Eroded Eroded	0.861	0.904 0.826	0.808 0.914	0.853 0.868

3.4. Variables Importance

The contribution of the controlling factors to the Classification and Regression Tree Model (Figure 4A) shows that, among 14 selected variables, only 11 were considered in the predictive model, excluding Specific Contributing Area and Distance to Rivers and Soils. In this case, the controlling factor Lineaments (64.22) was the most significant variable in the model, while Land Use and Cover (46.24), Slope (41.10), Elevation (40.98), and Rainfall (35.53), although significant, contributed moderately to highly in the modelling.

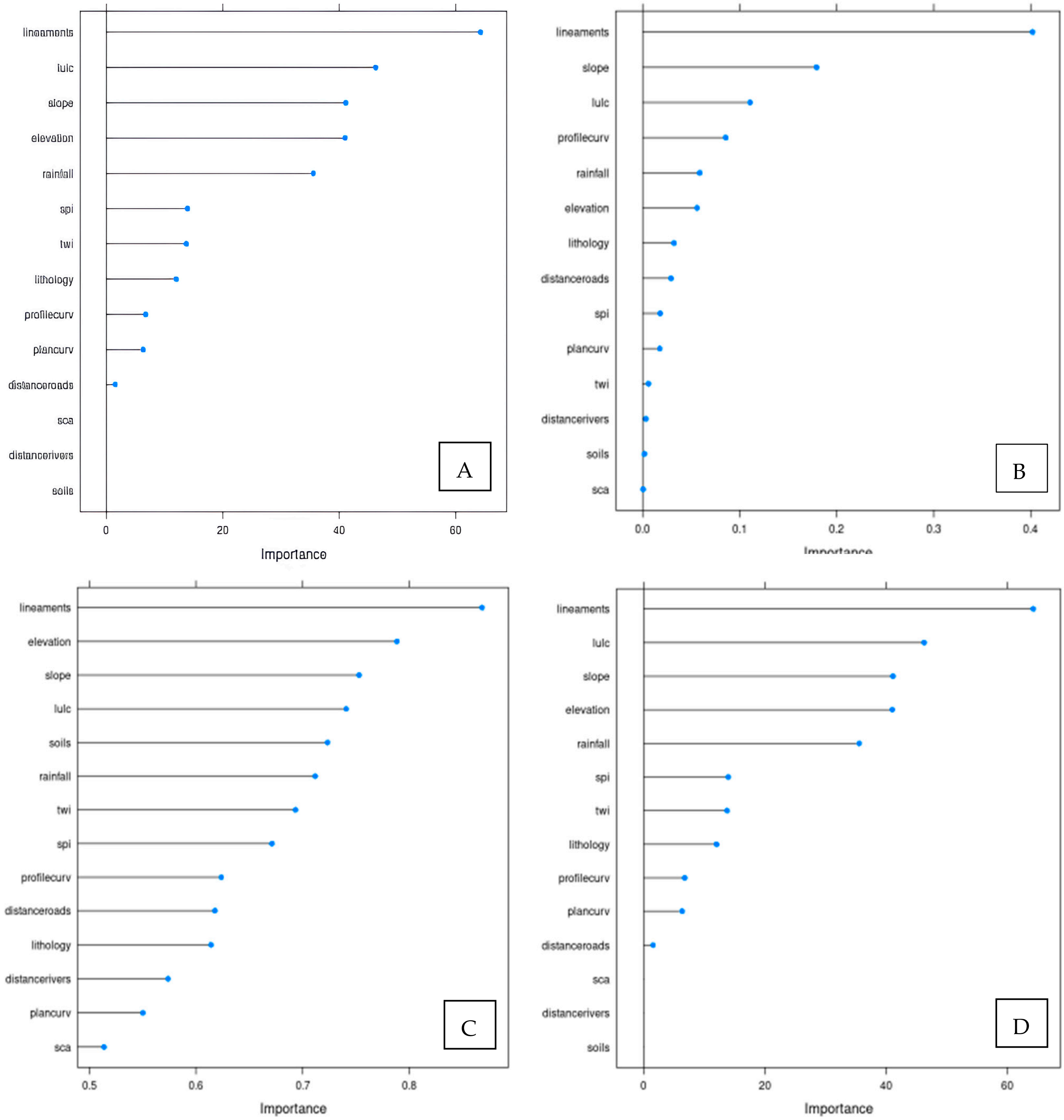

Figure 4. (**A**) Variables' importance for classification and regression tree model, respectively: lineaments = lineaments; lulc = land use and cover; slope = slope; elevation = elevation; rainfall = rainfall; spi = stream power index; twi = topographic wetness index; lithology = lithology; profilecurv = profile curvature; plancurv = plan curvature; distanceroads = distance to roads; sca = specific contributing area; distancerivers = distance to rivers; soils = soils. (**B**) Variables importance for eXtreme gradient boosting model; (**C**) Variables importance for random forest model and (**D**) Variables importance for support vector machine model.

The controlling factors such as Stream Power Index (13.90), Topographic Wetness Index (13.734), and Lithology (12.00) do not present significant values in contribution, but they differ from the other controlling factors, showing a moderate to low contribution. Profile Curvature (6.74), Plan Curvature (6.33), and Distance to Roads (1.52) were the least significant controlling factors in the model.

Regarding the eXtreme Gradient Boosting Model (Figure 4B), 14 selected variables were considered in the predictive model and the controlling factor Lineaments (40.18) demonstrated the greatest contribution in the modelling. Slope (17.92), Land Use and Land Cover (11.08), Profile Curvature (8.56), Rainfall (5.87), and Elevation (5.60) contributed moderately to highly in the modelling. Necessarily Lithology (3.21) and Distance to Roads (2.91) highlight a moderate to low contribution, considering the situation of this model and Stream Power Index (1.79), Plan Curvature (1.76), Topographic Wetness Index (0.59), Distance to Rivers (0.31), Soils (0.18), and Specific Contributing Area (0.04) were the least significant controlling factors in this predictive modelling.

The result regarding the contribution of the controlling factors to the Random Forest prediction (Figure 4C) shows that the 14 selected variables were considered in the predictive model. In this case, the controlling factor Lineaments was, again, the most relevant controlling factor for the model (22.13). Factors such as Elevation (13.33), Rainfall (11.61), Slope (11.26), and Land Use and Cover (11.16) in this model were considered as moderate to high contributions.

Stream Power Index (8.08), Topographic Wetness Index (6.83), and Profile Curvature (6.30) show a moderate to low contribution. Plan Curvature (5.38), Soils (4.25), Lithology (4.08), Distance to Roads (2.98), Distance to Rivers (2.29), and Specific Contributing Area were the least significant controlling factors in the predictive model.

In the Support Vector Machine Model (Figure 4D), the 14 selected variables were considered in the predictive model and, again, the lineaments controlling factor made the largest contribution (86.81), while Elevation (78.83), Slope (75.29), Land Use and Land Cover (74.09), Soils (72.33), and Precipitation (71.18) were considered as moderate to high contributions.

The controlling factors Topographic Wetness Index (69.34), Stream Power Index (67.12), Profile Curvature (62.38), Distance to Roads (61.78), and Lithology (61.42) were considered as moderate to low contributions. Distance to Rivers (57.38), Plan Curvature (55.02), and Specific Contributing Area (51.38) were considered as moderate to low contributions.

Unlike the Classification and Regression Tree, eXtreme Gradient Boosting, and Random Forest models, the importance or contribution of variables, in this case controlling factors, is performed using the ROC Curve to identify which attribute is most significant when modelling, contrary to the models previously mentioned, based on decision trees that use the Accuracy metric.

Through this comparison, it is possible to identify that the Lineaments controlling factor is the most significant attribute or variable in the four predictive models for susceptibility to gully erosion. Furthermore, Specific Contributing Area and then Distance to Rivers were the least significant controlling factors in the modelling process, showing a similarity in this aspect.

The controlling factors with moderate to high contribution were concentrated between Land Use and Cover, Slope, Elevation, and Rainfall, highlighting the importance in models of susceptibility to erosion and gullies, except for eXtreme Gradient Boosting adding the Profile Curvature factor and Support Vector Machine, the controlling factor soils. In the moderate to low controlling factors, Stream Power Index and Topographic Index were the attributes that made this contribution the most. The other controlling factors such as Distance to Roads, Lithology, Profile Curvature, Plan Curvature, and Soils did not obtain a clear pattern, depending on the model, occupying any extract for the contribution.

3.5. Gully Erosion Susceptibility Mapping

In recognizing the geospatial patterns of susceptibility classes (Figure 5), the Natural Jenks method [65] was used to identify those areas most prone to gullying (Table 4). In the very low and low susceptibility classes, the Classification and Regression Tree and Random Forest models achieved similar results and diverged from the Support Vector Machine and especially eXtreme Gradient results.

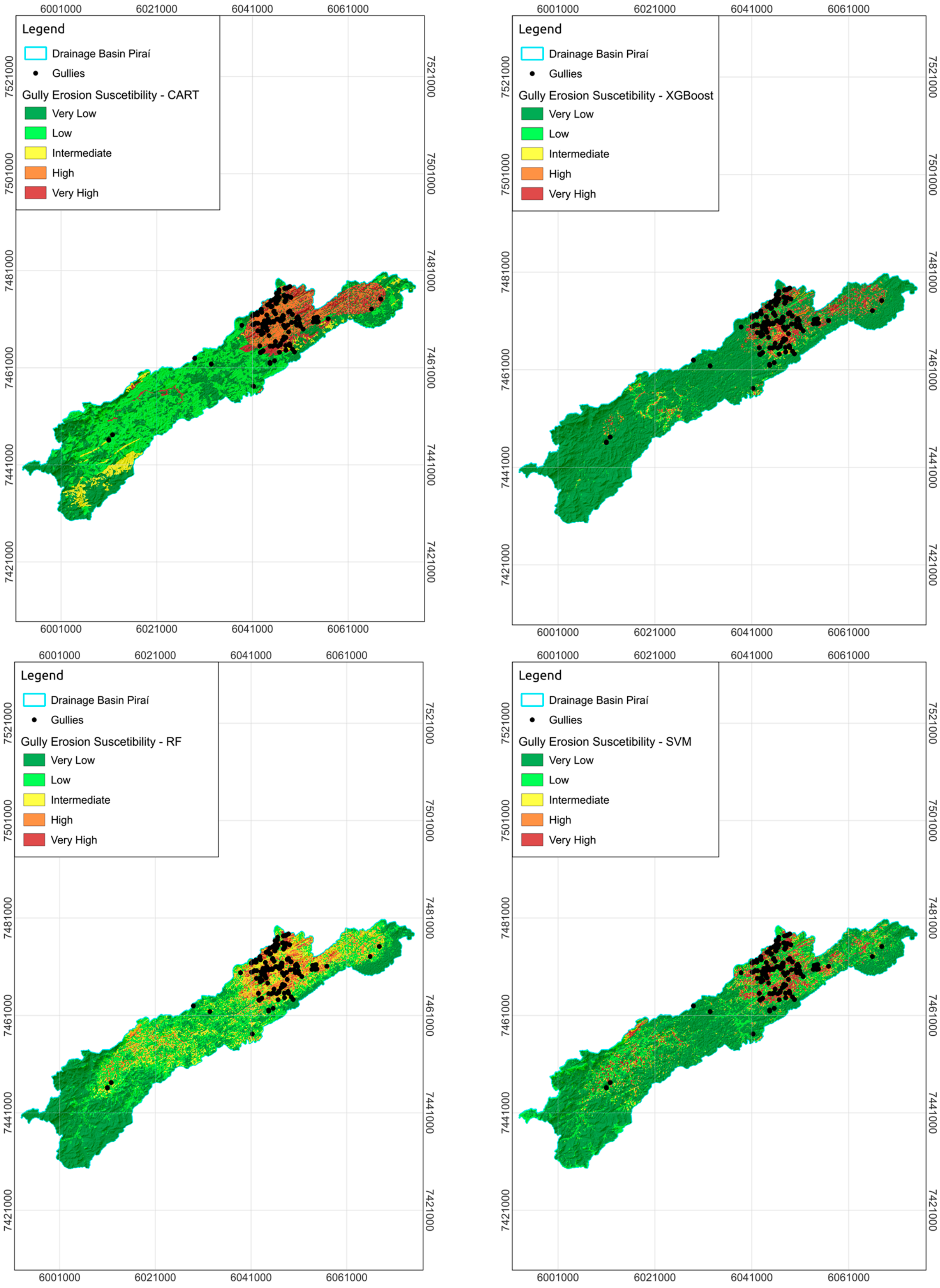

Figure 5. Different spatial patterns of susceptibility to gully erosion, according to each machine learning model, respectively. CART = classification and regression tree; XGBoost = extreme gradient boosting; RF = random forest and SVM = support vector machine.

Table 4. Area size and percentage of susceptibility to gully erosion.

Model		Gully Erosion Susceptibility				
		Very Low	Low	Intermediate	High	Very High
CART	Percent (%)	43.31	35.03	4.89	8.11	9.16
	Area (Km^2)	441.68	357.23	49.92	82.68	93.47
XGBoost	Percent (%)	83.08	5.89	2.07	2.36	7.11
	Area (Km^2)	847.34	60.04	21.11	24.03	72.47
RF	Percent (%)	41.02	30.23	17.23	7.88	3.63
	Area (Km^2)	418.40	308.25	180.82	80.39	37.02
SVM	Percent (%)	61.83	22.13	4.62	3.71	8.21
	Area (Km^2)	630.58	225.72	47.11	37.83	83.75

In the classes of susceptibility to erosion by intermediate gullies, running the Random Forest model, percentages were insignificant. The results referring to the high class again indicate similarity between the Classification and Regression Tree and Random Forest models, as well as with the eXtreme Gradient Boosting models. Considering the very high susceptibility classes to gully erosion, except in the Random Forest model, they assumed similar percentages.

4. Discussion

4.1. Multicollinearity in Controlling Factors for Gully Erosion

Among 18 controlling factors, Dissection Index, Local Relief, Topographic Ruggedness Index, and Sediment Transport Index were the factors that showed the greatest multicollinearity, considering VIF < 10 and TOL > 0.1 [16,27]. It is possible to understand that both the Topographic Ruggedness Index and Sediment Transport Index, in their formulations, consider the slope, respectively, with a moving window and related to the contribution area. Although Dissection Index and Local Relief do not use slope directly or indirectly, spatially they have similar results. In this sense, the use of these controlling factors could bias predictive modelling.

4.2. Performance of Machine Learning Models

Considering the comparison between the four models for susceptibility to gully erosion, the Classification and Regression Tree model was the least satisfactory model, and this is due to its sensitivity to the training data [59]. Although, the Support Vector Machine model achieved formidable performance, it is not superior to eXtreme Gradient Boosting and Random Forest, as indicated in studies by [27,56]. This analysis corroborates [10], showing that the Classification and Regression Tree model, as it is more sensitive, results in predictions with reduced overall accuracy. Although the Support Vector Machine model makes it possible to analyse complex and non-linear relationships [53], in general, it performs more robust results with smaller datasets [59] and more sensitive to noisy data, compared to the Random Forest model [62] and more interesting for susceptibility to landslides and/or floods [10,53].

eXtreme Gradient Boosting and Random Forest model in relation to Accuracy and AUC-ROC achieved identical, more robust and satisfactory results for susceptibility to gully erosion, as identified by [27,54–56] for eXtreme Gradient Boosting and [10,53,56,57] for Random Forest. In this sense, the present study is like the result obtained by [56] since, based on k-fold cross validation (cv), the performances were identical between the Bagging and Boosting strategies. The eXtreme Gradient Boosting and Random Forest models, unlike the other models analysed, can deal with a large volume of data and complex relationships between the data [10,53,60].

Based on the nature of the Precision and Recall metrics, the eXtreme Gradient Boosting model is the most satisfactory model, although the Precision metric considers in its analysis the occurrence of false positives (fp), indicating that there is an overestimation for the prediction in the eXtreme Gradient Boosting model, compared to the Random Forest model. However, the Recall metric identifies that false negatives (fn) are lower in the eXtreme Gradient Boosting model and higher in the Random Forest model, indicating that there is a higher rate of true positives (tp) in the eXtreme Gradient Boosting model in relation to the model Random Forest, analysing the Eroded class, that is, for real occurrences of susceptibility to erosion by gullies.

For the Non-Eroded class, this scenario is the opposite, since Precision obtained superior results in the eXtreme Gradient Boosting model and inferior results in the Random Forest model, indicating the occurrence of a higher rate of false positives in the second model mentioned. However, considering the Recall metric, the Random Forest model presented a more robust performance to identify the non-occurrence of susceptibility to gully erosion. In this sense, it is understood that, although the models have identical performance metrics, as mentioned previously, the Random Forest model demonstrated greater efficiency for identifying the Non-Eroded class and the eXtreme Gradient Boosting model for the Eroded class.

Precision and recall are inversely proportional metrics [68], necessarily for susceptibility mapping, as in this case for erosion by gullies, it is understood that the identification of the true positive rate or recall is fundamental. This makes it possible to identify real occurrences of susceptibility. For these reasons, it is understood that the eXtreme Gradient Boosting model was the most appropriate model to identify real occurrences of susceptibility to erosion by gullies.

Furthermore, it is important to highlight that, due to the multiplicity of scales between the different data used for the controlling factors and the diverse nature of geographic data, they can cause uncertainty in the modelling, for example, digital elevation models, climate time series data, and drainage network data. In this sense, the use of performance metrics from machine learning models is essential to understand these limitations and uncertainties, as well as choosing the Precision metrics and more specifically Recall, to identify the most suitable model to be used.

As well as the sample set (n = 159), they may cause bias in the analyses. However, the use of machine learning models such as eXtreme Gradient Boosting [60] and Random Forest [61] make it possible to reduce bias and to provide more robust results, as indicated in this research.

4.3. Thresholding in Controlling Factors for Gully Erosion

Several studies involving the mapping and/or models of susceptibility to erosion by gullies have not been concerned with understanding the thresholds of the controlling factors for erosion by gullies, rather only the importance of the controlling factors in predictive modelling.

In relation to Elevation, although difficult to interpret, as it is absolute, it is understood that due to the occurrence of gullies in the lower portions, it can be related to the hilly domain [20,21]. In the profile and plan convex–convex curvatures, it is similar [19], which presupposes a predominant convex slope form in profile and diverges considering the concave plan form, thus, as a slope it is average, without establishing any certain threshold.

In the flow power index, specific contribution area, and topographic wetness index, positive values represent humid areas or saturated areas, in which surface runoff predominates. In relation to the distance to rivers and highways, it is not possible to infer whether there is a positive correlation between such distances and the occurrence of gullies. In pasture, as the predominant use for the emergence of gullies, it refers to the various changes in land use and cover over two centuries [24]. Concerning lineaments, lithology, and soils, the Rio Turvo suite with Granitoids and Orthogneisses and Red–Yellow Argisols can facilitate the occurrence of mechanical discontinuities, based on this conjunction of factors.

4.4. Variables Importance

The four models for susceptibility to gully erosion, generally, present a predominance over certain controlling factors, such as Lineaments, Land Use and Cover, Slope, Elevation, and Rainfall, which are consistent with [24] on the acceleration of processes erosion in the study area. This statement can be verified based on the presence of controlling factors such as Land Use and Cover and Precipitation, relating hydroclimatic changes to the various changes in land use and cover. Thus, Slope and Elevation, according to [20,21], highlight the nature of the occurrence of gullies in hilly domains in the study area.

Among the four models for susceptibility to gully erosion, eXtreme Gradient Boosting not only agrees with [21,24], but it is similar to the hypothesis developed by [19,20]. Since the aforementioned model considers not only the structural aspect (Lineaments), such as faults and fractures, like the other models, but also the identification of slope shapes (Profile Curvature), which may or may not favour the emergence of gullies, with greater importance in the configuration of the gully erosion susceptibility model. In this regard, it is a more explanatory model and consistent with the literature on susceptibility to erosion by gullies in the Piraí Basin Drainage, Paraíba do Sul Middle Valley.

Highlighting the robustness of the XGBoost model in relation to modelling, in relation to complexity [60] and the importance of understanding natural laws, the portion of geographic space and the history of formation of environmental conditions are ways of understanding the landscape [6] to choose the controlling factors consistent with the dynamicity of the landscape.

4.5. Gully Erosion Susceptibility Mapping

It is understood that the similarity noted between the Classification and Regression Tree and Random Forest models for susceptibility to gully erosion in the very low, low, and high classes arises from the Random Forest model using the very high gully assumption, except the Random model Forest assumed similar percentages as the Classification and Regression Tree model [13].

Susceptibility to gully erosion is predominant in the very low susceptibility class in the four models analysed, while considering the combination of the high and very high classes, it does not exceed 18%. Considering the eXtreme Gradient Boosting model as the most appropriate, it is identified that 9.47% or 96.50 km^2, depending on the high to very high susceptibility classes.

5. Conclusions

eXtreme Gradient Boosting was the more appropriate, robust, and satisfactory model to identify susceptibility to gully erosion in the Drainage Piraí Basin, Paraíba do Sul Middle Valley, and this was identified as 9.47% or 96.50 km^2, depending on the high to very high susceptibility classes.

Considering the performance metrics for machine learning models, as well as the importance variables, it was the one that most resembled the current hypothesis and was consistent with the literature. Unlike physical models for the occurrence of gullies, they can be implemented on a large scale. It is understood that studies for susceptibility to erosion by gullies should not only be concerned with performance or performance metrics for reliability, as well as to enable validation or not, by conceptual, empirical, physical, or stochastic models.

In this sense, it is recommended that public policies be enacted to enable the construction of geospatial data for the constant monitoring of gully occurrences to enable even more robust models, as well as discussions regarding areas that correspond to the percentage of high to very high susceptibility for the prevention and recovery of these areas.

The premise of the methodology in this article makes it possible to understand the spatial pattern of susceptibility to gully erosion, regardless of the landscape structure. In this sense, with the possibility of accelerating erosion processes due to global climate change, truly replicable methodologies, as in this study, are evidenced by their rapid applicability at different scales from the local to the global context.

In this sense, mapping susceptibility to erosion by gullies makes it possible to carry out adequate land use management, anticipating problems related to management. Due to the complexity and multiplicity of controlling factors, investigation into more specific aspects of their dynamics can contribute to predictive modelling, such as soils' physical and chemical properties, and the inclusion of temporal variability in land use and land cover.

Still, due to the inconclusiveness of the literature on the most appropriate machine learning model, further investigations and comparisons with hybrid and deep learning models are needed, which have shown robust results.

Author Contributions: Conceptualization, J.d.P.M.F. and A.J.T.G.; validation, J.d.P.M.F. and C.B.M.C.; formal analysis, J.d.P.M.F.; investigation, J.d.P.M.F., A.J.T.G. and C.B.M.C.; resources, J.d.P.M.F.; writing—original draft preparation, J.d.P.M.F.; writing—review and editing, A.J.T.G., C.B.M.C., M.d.C.O.J. and C.A.B.; visualization, A.J.T.G., C.B.M.C., M.d.C.O.J. and C.A.B.; supervision, A.J.T.G. and C.B.M.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received a grant from the National Council for Scientific and Technological Development of Brazil—CNPq (in Portuguese: Conselho Nacional de Desenvolvimento Científico e Tecnológico), process no. 140241/2021-3.

Data Availability Statement: The original data presented in the study are openly available in GitHub at 10.5281/zenodo.13924282 [insert reason here].

Acknowledgments: We would like to thank the Post-Graduate Program in Geography (UFRJ), financial support from CNPq (Brazilian National Council), and Ana Karolina Tomé and Thainara Cordeiro Pereira, undergraduates at the Geography Department, Federal University of Rio de Janeiro, for measuring the gullies on Google Earth.

Conflicts of Interest: The authors declare no conflicts of interest. The funder had no role in the design of the study, in the collection, analyses, or interpretation of data, in the writing of the manuscript, or in the decision to publish the results.

References

1. Guerra, A.J.T.; Fullen, M.A.; Jorge, M.C.O.; Bezerra, J.F.R.; Shokr, M.S. Slope processes, mass movement and soil erosion: A review. *Pedosphere* **2017**, *27*, 27–41. [CrossRef]
2. Soil Science Society of America. *Glossary of Soil Science Terms*, 2nd ed.; Soil Science Society of America: Madison, WI, USA, 2008.
3. Fullen, M.A.; Catt, J.A. *Soil Management—Problems and Solutions*, 1st ed.; Oxford University Press: Oxford, UK, 2004.
4. Poesen, J.; Torri, D.; Vanwalleghem, T. Gully erosion: Procedures to adopt when modelling soil erosion in landscapes affected by gullying. In *Handbook of Erosion Modelling*, 1st ed.; Morgan, R., Nearing, M., Eds.; Blackwell Publishing Ltd.: Hoboken, NJ, USA, 2010; pp. 360–386.
5. Barbosa, W.C.S.; Guerra, A.J.T.; Valladares, G.S. Soil erosion modeling using the revised universal soil loss equation and a geographic information system in a watershed in the northeastern Brazilian Cerrado. *Geosciences* **2024**, *14*, 78. [CrossRef]
6. Phillips, J.D. Laws, place, history and the interpretation of landforms. *Earth Surf. Process. Landf.* **2016**, *42*, 347–354. [CrossRef]
7. Ciccolini, U.; Buffalini, M.; Materazzi, M.; Dramis, F. Gully erosion development in drainage basins: A new morphometric approach. *Land* **2024**, *13*, 792. [CrossRef]
8. Guerra, A.J.T.; Bezerra, J.F.R.R.; Jorge, M.C.O. Recuperação de voçorocas e de áreas degradadas, no Brasil e no mundo—Estudo de caso da voçoroca do Sacavém-São Luís—MA. *Rev. Bras. Geomorfol.* **2023**, *24*, 1–20. [CrossRef]
9. Arabameri, A.; Cerda, A.; Tiefenbacher, J.P. Spatial pattern analysis and prediction of gully erosion using novel hybrid model of entropy-weight of evidence. *Water* **2019**, *11*, 1129. [CrossRef]
10. Pourghasemi, H.R.; Sadhavisam, 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.* **2019**, *11*, 2207–2219. [CrossRef]
11. Malczewski, J. GIS-based multicriteria decision analysis: A survey of the literature. *Int. J. Geogr. Inf. Syst.* **2006**, *20*, 703–726. [CrossRef]
12. Vojtek, M.; Vojteková, J.; Costache, R.; Pham, Q.B.; Lee, S.; Arshad, A.; Sahoo, S.; Linh, N.T.T.; Anh, D.T. Comparison of multi-criteria-analytical hierarchy process and machine learning-boosted tree models for regional flood susceptibility mapping: A case study from Slovakia. *Geomat. Nat. Hazards Risk* **2021**, *12*, 1153–1180. [CrossRef]
13. Mohebzadeh, H.; Biswas, A.; Rudra, R.; Daggupati, P. Machine learning techniques for gully erosion susceptibility mapping: A review. *Geosciences* **2022**, *12*, 429. [CrossRef]

14. Renard, P.; Alcolea, A.; Ginsbourger, D. Stochastic versus Deterministic Approaches. In *Environmental Modelling: Finding Simplicity in Complexity*, 2nd ed.; Wainwright, J., Mulligan, M., Eds.; Wiley: Hoboken, NJ, USA, 2013; pp. 133–150.
15. Lary, D.J.; Alavi, A.H.; Gandomi, A.H.; Walker, A.L. Machine learning in geoscience and remote sensing. *Geosci. Front.* **2016**, *7*, 3–10. [CrossRef]
16. Chuma, G.B.; Mugumaarhahama, Y.; Mond, J.M.; Bagula, E.M.; Ndeko, A.B.; Lucungu, P.B.; Katume, K.; Mushagalusa, G.N.; Schmitz, S. Gully erosion susceptibility mapping using four machine learning methods in Luzinzi watershed, eastern Democratic Republic of Congo. *Phys. Chem. Earth Parts A/B/C* **2023**, *123*, 103295. [CrossRef]
17. Riccomini, C.; Sant'anna, L.G.; Ferrari, A.L. Evolução geológica do Rift continental do sudeste do Brasil. In *Geologia do Continente Sul-Americano: Evolução da Obra de Fernando Flávio Marques de Almeida*, 1st ed.; Mantesso-Neto, V., Bartorelli, A., Carneiro, C.D.R., Brito-neves, B.B., Eds.; Editora Beca: São Paulo, Brazil, 2004; pp. 383–405.
18. Dantas, M.E.; Ferreira, C.E.O.; Shinzato, E. Relevo. In *Solos do Rio de Janeiro—Gênese, Classificação e Limitações ao Uso Agrícola*, 1st ed.; Pereira, M.G., Anjos, L.H.C., Neto, E.C.S., Eds.; Editora Atena: Ponta Grossa, Brazil, 2023; pp. 19–59.
19. Dantas, M.E. Geomorfologia do estado do Rio de Janeiro. In *Projeto Rio de Janeiro*, 1st ed.; CPRM–Serviço Geológico do Brasil: Brasilia, Brazil, 2001; pp. 95–195.
20. Avelar, A.S.; Coelho Netto, A.L. Fraturas e desenvolvimento de unidades geomorfológicas côncavas no Médio Vale do Rio Paraíba do Sul. *Rev. Bras. Geociências* **1992**, *22*, 222–227. [CrossRef]
21. Coelho Netto, A.L. Evolução de cabeceiras de drenagem no médio Vale do Rio Paraíba do Sul (SP/RJ): Bases para um modelo de formação e crescimento da rede de canais sob controle estrutural. *Rev. Bras. Geomorfol.* **2003**, *4*, 69–100. [CrossRef]
22. Carvalho Filho, A.; Lumbreras, J.F.; Wittern, K.P.; Lemos, A.L.; dos Santos, R.D.; Calderano Filho, B.; de Oliveira, R.P.; Aglio, M.L.D.; de Souza, J.S.; Chaffin, C.E.; et al. *Levantamento de Reconhecimento de Baixa Intensidade dos Solos do Estado do Rio de Janeiro*, 1st ed.; Embrapa Solos: Rio de Janeiro, Brazil, 2003.
23. Loureiro, H.A.S. Diagnóstico de Erosão por Voçorocas: Experimentos com Geotecnologias e Solos na Bacia do Alto Rio Piraí—Rio Claro-RJ. Ph.D. Thesis, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil, 2019.
24. Dantas, M.E.; Coelho Netto, A.L. A denudação antropogênica da paisagem: Processos erosivos deposicionais no médio Vale do Rio Paraíba do Sul. In *Geografia Histórica do Café no Vale do Rio Paraíba do Sul*, 1st ed.; Oliveira, R., Lazos, A., Eds.; Editora Puc-Rio: Rio de Janeiro, Brazil, 2018; pp. 107–126.
25. Bitar, O.Y. *Cartas de Suscetibilidade a Movimentos Gravitacionais de Massa e Inundações—1:25,000 Nota Técnica Explicativa*, 1st ed.; IPT–Instituto de Pesquisas Tecnológicas do Estado de São Paulo and CPRM–Serviço Geológico do Brasil: São Paulo, Brazil, 2016.
26. Baiddah, A.; Krimissa, S.; Hajji, S.; Ismaili, M.; Abdelrahman, K.; El Bouzekraoui, M.; Eloudi, H.; Elaloui, A.; Khouz, A.; Badreldin, N.; et al. Head-cut gully erosion susceptibility mapping in semi-arid region using machine learning methods: Insight from the high atlas, Morocco. *Front. Earth Sci.* **2023**, *11*, 1184038. [CrossRef]
27. Bammou, Y.; Benzougagh, B.; Abdessalam, O.; Brahim, I.; Kader, S.; Spalevic, V.; Sestras, P.; Ercisli, S. Machine learning models for gully erosion susceptibility assessment in the Tensift catchment, Haouz Plain, Morocco for sustainable development. *J. Afr. Earth Sci.* **2024**, *213*, 105229. [CrossRef]
28. Ghorbanzadeh, O.; Blaschke, T.; Aryal, J.; Gholaminia, K. A new GIS-based technique using an adaptive neuro-fuzzy inference system for land subsidence susceptibility mapping. *J. Spat. Sci.* **2020**, *65*, 401–418. [CrossRef]
29. Pal, S.C.; Arabameri, A.; Blaschke, T.; Chowdhuri, I.; Saha, A.; Chakrabortty, R.; Lee, S.; Band, S.S. Ensemble of machine-learning methods for predicting gully erosion susceptibility. *Remote Sens.* **2020**, *12*, 3675. [CrossRef]
30. Roy, J.; Saha, S. Ensemble hybrid machine learning methods for gully erosion susceptibility mapping: K-fold cross validation approach. *Artif. Intelli. Geosci* **2022**, *3*, 28–45. [CrossRef]
31. Hawker, L.; Uhe, P.; Paulo, L.; Sosa, J.; Savage, J.; Sampson, C.; Neal, J. A 30 m global map of elevation with forests and buildings removed. *Environ. Res. Lett* **2022**, *17*, 24016. [CrossRef]
32. Lindsay, J.B.; Creed, I.F. Removal of artifact depressions from digital elevation models: Towards a minimum impact approach. *Hydrol. Process.* **2005**, *19*, 3113–3126. [CrossRef]
33. Reuter, H.I.; Hengl, T.; Gessler, P.; Soille, P. Preparation of DEMs for geomorphometric analysis. In *Geomorphometry: Concepts, Software, Applications*, 1st ed.; Hengl, T., Reuter, H.I., Eds.; Elsevier: Amsterdam, The Netherlands, 2009; pp. 87–120.
34. Pike, R.J.; Evans, I.; Hengl, T. Geomorphometry: A brief guide. In *Geomorphometry: Concepts, Software, Applications*, 1st ed.; Hengl, T., Reuter, H.I., Eds.; Elsevier: Amsterdam, The Netherlands, 2009; pp. 3–30.
35. Franklin, S.E. Interpretation and use of geomorphometry in remote sensing: A guide and review of integrated applications. *Int. J. Remote Sens.* **2020**, *41*, 7700–7733. [CrossRef]
36. Evans, I.S. General geomorphometry, derivatives of altitude, and descriptive statistics. In *Spatial Analysis in Geomorphology*, 1st ed.; Chorley, R.J., Ed.; Harper & Row: New York, NY, USA, 1972; pp. 17–90.
37. Ahnert, F. Functional relationships between denudation, relief, and uplift in large, mid-latitude drainage basins. *Am. J. Sci.* **1970**, *268*, 243–263. [CrossRef]
38. Ahnert, F. Local relief and the height limits of mountain ranges. *Am. J. Sci* **1984**, *284*, 1035–1055. [CrossRef]
39. Speight, J.G. A parametric approach to landform regions. In *Progress in Geomorphology*, 1st ed.; Institute of British Geographers: London, UK, 1974; Alden & Mowbray Ltd. at the Alden Press: Oxford, UK, 1994; pp. 213–230.

40. Florinsky, I.V. Topographic Surface and Its Characterization. In *Digital Terrain Analysis in Soil Science and Geology*, 2nd ed.; Florinsky, I.V., Ed.; Elsevier: Amsterdam, The Netherlands, 2016; pp. 7–76.
41. Desmet, P.J.J.; Govers, G. A GIS procedure for automatically calculating the USLE LS Factor on topographically complex landscape units. *J. Soil Water Conserv.* **1996**, *51*, 427–433.
42. Shary, P.A.; Sharaya, L.S.; Mitusov, A.V. Fundamental quantitative methods of land surface analysis. *Geoderma* **2002**, *107*, 1–32. [CrossRef]
43. Moore, I.D.; Grayson, R.B.; Ladson, A.R. Digital terrain modelling: A review of hydrological, geomorphological and biological applications. *Hydrol. Process.* **1991**, *5*, 3–30. [CrossRef]
44. Sanchez, E.F.; Alvarez, C.I. Prioritization of Hydrological Restoration Areas Using AHP and GIS in Dulcepamba River Basin in Bolivar–Ecuador. *Hydrology* **2024**, *11*, 81. [CrossRef]
45. Riley, S.J.; DeGloria, S.D.; Elliot, R. A terrain ruggedness index that quantifies topographic heterogeneity. *Int. J. Sci.* **1999**, *5*, 23–27.
46. Beven, K.J.; Kirkby, M.J. A physically-based variable contributing area model of basin hydrology. *Hydrol. Sci. Bull.* **1979**, *24*, 43–69. [CrossRef]
47. Grabs, T.; Seibert, J.; Bishop, K.; Laudon, H. Modeling spatial patterns of saturated areas: A comparison of the topographic wetness index and a dynamic distributed model. *J. Hydrol.* **2009**, *373*, 15–23. [CrossRef]
48. Al-Bawi, A.J.; Al-Abadi, A.M.; Pradhan, B.; Alamri, A.M. Assessing gully erosion susceptibility using topographic derived attributes, multi-criteria decision-making, and machine learning classifiers. *Geomat. Nat. Hazards Risk* **2021**, *12*, 3035–3062. [CrossRef]
49. Heilbron, M.; Almeida, J.C.H.; Eirado, L.G. *Geologia e Recursos Minerais do Estado do Rio de Janeiro*, 1st ed.; CPRM: Brasilia, Brazil, 2016.
50. Instituto Estadual do Ambiente. Mapeamento de Uso do Solo e Cobertura Vegetal. 2018. Available online: https://visualizador.inde.gov.br/ (accessed on 8 June 2024).
51. Instituto Brasileiro de Geografia e Estatística. Base Comum Vetorial do Estado do Rio de Janeiro Versão 2018 edgv 3.0. Available online: https://www.ibge.gov.br/geociencias/downloads-geociencias.html?caminho=cartas_e_mapas/bases_cartograficas_continuas/bc25/rj/versao2018_edgv_3.0/ (accessed on 8 June 2024).
52. Fick, S.E.; Hijmans, R.J. WorldClim 2: New 1 km spatial resolution climate surfaces for global land areas. *Int. J. Climatol.* **2017**, *37*, 4302–4315. [CrossRef]
53. Gayen, A.; Pourghasemi, H.R.; Saha, S.; Keesstra, S.; Bai, S. Gully erosion susceptibility assessment and management of haz-ard-prone areas in India using different machine learning algorithms. *Sci. Total Environ.* **2019**, *668*, 124–138. [CrossRef]
54. Arabameri, A.; Chandra Pal, S.; Costache, R.; Saha, A.; Rezaie, F.; Seyed Danesh, A.; Pradhan, B.; Lee, S.; Hoang, N.D. Prediction of gully erosion susceptibility mapping using novel ensemble machine learning algorithms. *Geomat. Nat. Hazards Risk* **2021**, *12*, 469–498. [CrossRef]
55. Hasanuzzaman, M.D.; Adhikary, P.P.; Shit, P.K. Gully erosion susceptibility mapping and prioritization of gully-dominant sub-watersheds using machine learning algorithms: Evidence from the Silabati River (tropical river, India). *Adv. Space Res.* **2024**, *73*, 1653–1666. [CrossRef]
56. Were, K.; Kebeney, S.; Churu, H.; Mutio, J.M.; Njoroge, R.; Mugaa, D.; Alkamoi, B.; Ng'etich, W.; Singh, B.R. Spatial Prediction and Mapping of Gully Erosion Susceptibility Using Machine Learning Techniques in a Degraded Semi-Arid Region of Kenya. *Land* **2023**, *12*, 890. [CrossRef]
57. Huang, D.; Su, L.; Zhou, L.; Tian, Y.; Fan, H. Assessment of gully erosion susceptibility using different DEM-derived topographic factors in the black soil region of Northeast China. *Int. Soil Water Conserv. Res.* **2023**, *11*, 97–111. [CrossRef]
58. Krzywinski, M.; Altman, N. Classification and regression trees. *Nat. Methods* **2017**, *14*, 757–758. [CrossRef]
59. Zafar, Z.; Zubair, M.; Zha, Y.; Fahd, S.; Nadeem, A.A. Performance assessment of machine learning algorithms for mapping of land use/land cover using remote sensing data. *Egypt. J. Remote Sens. Space Sci.* **2024**, *27*, 216–226. [CrossRef]
60. Chen, T.Q.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
61. Witten, H.; Frank, E.; Hall, M.A. *Data Mining Practical Machine Learning Tools and Techniques*, 4th ed.; Morgan Kaufmann Publishers: Burlington, MA, USA, 2018.
62. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
63. Pal, M.; Mather, P.M. Support vector machines for classification in remote sensing. *Int. J. Remote Sens.* **2005**, *26*, 1007–1011. [CrossRef]
64. Wang, L.; Yan, J.; Mu, L.; Huang, L. Knowledge discovery from remote sensing images: A review. *WIREs Data. Min. Knowl.* **2020**, *10*, 1–31. [CrossRef]
65. De Smith, M.J.; Goodchild, M.F.; Longley, P.A. Univariate classification schemes. In *Geospatial Analysis—A Comprehensive Guide*, 7th ed.; Winchelsea Press: Winchelsea, East Sussex, England, 2024.
66. Maxwell, A.E.; Warner, T.A.; Guillén, L.A. Accuracy assessment in convolutional neural network-based Deep Learning remote sensing studies—Part 1: Literature review. *Remote Sens.* **2021**, *13*, 2450. [CrossRef]
67. Sammut, C.; Webb, G.I. *Encyclopedia of Machine Learning and Data Mining*, 2nd ed.; Springer: New York, NY, USA, 2017.

68. Buckley, A.R. Minimum Mapping Unit (MMU). In *Encyclopedia of Geographic Information Science*, 1st ed.; Kemp, K.K., Ed.; SAGE Research Methods: London, UK, 2008; pp. 287–288.
69. Spínola, D.N.; Filho, E.I.F.; Portes, R.C.; Resck, B.C. Comparação entre dois métodos de generalização cartográfica semi-automática em ambiente matricial. In Proceedings of the XV Simpósio Brasileiro de Sensoriamento Remoto, Curitiba, Brazil, 30 April–5 May 2011; pp. 2294–2301.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.