land

MDPI

Article

Spatial Prediction and Mapping of Gully Erosion Susceptibility Using Machine Learning Techniques in a Degraded Semi-Arid Region of Kenya

Kennedy Were [1,*], Syphyline Kebeney [2], Harrison Churu [2], James Mumo Mutio [2], Ruth Njoroge [2], Denis Mugaa [2], Boniface Alkamoi [2], Wilson Ng'etich [2] and Bal Ram Singh [3,*]

[1] Kenya Agricultural and Livestock Research Organization, Kenya Soil Survey, P.O. Box 14733, Nairobi 00800, Kenya
[2] School of Agriculture and Biotechnology, University of Eldoret, P.O. Box 1125, Eldoret 30100, Kenya; syphyline.kebeney@uoeld.ac.ke (S.K.); harrison.churu@uoeld.ac.ke (H.C.); mumomutio@gmail.com (J.M.M.); ruthnjoroge@uoeld.ac.ke (R.N.); sagrsosm001@uoeld.ac.ke (D.M.); alkamoi.boniface@uoeld.ac.ke (B.A.); w.ngetich@physics.org (W.N.)
[3] Faculty of Environmental Sciences and Natural Resource Management, Norwegian University of Life Sciences, P.O. Box 5003, 1432 Ås, Norway
* Correspondence: kenwerez@yahoo.com or kennedy.were@kalro.org (K.W.); balram.singh@nmbu.no (B.R.S.)

Abstract: This study aimed at (i) developing, evaluating and comparing the performance of support vector machines (SVM), boosted regression trees (BRT), random forest (RF) and logistic regression (LR) models in mapping gully erosion susceptibility, and (ii) determining the important gully erosion conditioning factors (GECFs) in a Kenyan semi-arid landscape. A total of 431 geo-referenced gully erosion points were gathered through a field survey and visual interpretation of high-resolution satellite imagery on Google Earth, while 24 raster-based GECFs were retrieved from the existing geodatabases for spatial modeling and prediction. The resultant models exhibited excellent performance, although the machine learners outperformed the benchmark LR technique. Specifically, the RF and BRT models returned the highest area under the receiver operating characteristic curve (AUC = 0.89 each) and overall accuracy (OA = 80.2%; 79.7%, respectively), followed by the SVM and LR models (AUC = 0.86; 0.85 & OA = 79.1%; 79.6%, respectively). In addition, the importance of the GECFs varied among the models. The best-performing RF model ranked the distance to a stream, drainage density and valley depth as the three most important GECFs in the region. The output gully erosion susceptibility maps can support the efficient allocation of resources for sustainable land management in the area.

Keywords: soil erosion; land degradation; sustainable land management; landscape restoration; spatial prediction; machine learning

Citation: 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. https://doi.org/10.3390/land12040890

Academic Editors: Vesna Zupanc, Nejc Bezak and Carla Ferreira

Received: 5 March 2023
Revised: 4 April 2023
Accepted: 13 April 2023
Published: 15 April 2023

1. Introduction

Soil is a precious and irreplaceable natural resource that offers vital ecosystem services and performs multiple ecological functions to support life on Earth. It plays a role in the production of food, fodder and timber, purification and storage of water, cycling of nutrients, filtration of toxic substances, and preservation of biodiversity, among others [1]. Despite this, a third of the world's soils are degraded, with 25–40 billion tons of soil being lost annually due to erosion [2,3]. This has implications for the productivity, resilience and sustainability of both agricultural and ecological systems [4]. Although soil erosion is global, Africa is the worst-hit continent, with the most destructive process being gully erosion [5]. Gully erosion is a complex and destructive geomorphic phenomenon driven by various environmental factors, including soil type, lithology, topography, climate, land use and vegetation [6,7]. For instance, in parts of East Africa, it has been attributed to increased overland flow owing to the declining and low vegetation cover [8]. Gully erosion

is of great concern to scientists because runoff concentrates and flows detach and transport large amounts of soil particles, carving out wide (>0.3 m), deep (>0.6 m) and long channels across the landscape, causing many problems downstream [9,10]. The channels, which are difficult to eliminate, can stem from mechanisms, such as incision, piping, fluting, mass wasting, subsurface flows, and the development of rills and cracks in the soils [11,12].

The environmental and socio-economic repercussions of gully erosion are manifold. They range from desertification, flooding and stagnation of agricultural productivity to the degradation of water quality in rivers, depletion of essential soil nutrients, infrastructural damage and loss of soil biodiversity and agro-pastoral lands [4,5,11–16]. To prevent, halt and reverse gully formation and expansion, geographic targeting and uptake of sustainable land management technologies, innovations and management practices, including soil and water conservation measures, is indispensable. This calls for spatially-explicit information on the risk of gully erosion, which can support the delineation and prioritization of hotspots for intervention.

The need for spatially-explicit information has stimulated several studies, which have supplied sufficient evidence on the applicability of different models in mapping the predisposition of various landscapes to gully erosion. Such studies have also been boosted by the rapid advances in machine learning and geocomputing capabilities, as well as by the increased availability of reliable and open-access geospatial data. Thus far, the capabilities of several modeling techniques in mapping gully erosion susceptibility have been demonstrated, including knowledge-based models, such as the analytical hierarchy process [14,17] and statistical models, such as binary logistic regression, frequency ratio, the weight of evidence and multivariate adaptive regression splines [13,18–22], and machine learning and deep learning neural networks models, such as support vector machines, random forest, boosted regression trees, artificial neural networks, maximum entropy and convolutional neural networks [10,23–36].

Despite the multiplicity of studies that have contributed knowledge on the capabilities of different modeling techniques in gully erosion susceptibility mapping, very few have focused on Africa's arid and semi-arid lands, where the severity of land degradation is disturbing [37]. Recently, Busch et al. [9] reported satisfactory classification accuracy after applying the random forest technique to model the gully erosion susceptibility across a semi-arid environment in Ethiopia. Similarly, Igwe et al. [17] obtained promising results when they applied the frequency ratio and analytical hierarchy process in a semi-arid region of northeastern Nigeria; however, these studies never attempted to assess the performance of different machine learning techniques comparatively. It would be interesting, for instance, to know whether different modeling techniques would yield uniform results when applied in the same geographic setting or even when a single technique is implemented in different arid and semi-arid landscapes within the continent. In brief, knowledge about the dynamics of gully erosion and performance of the existing modeling techniques, especially the state-of-the-art machine learning methods, in Africa's less-studied and data-scarce arid and semi-arid environments is still insufficient.

In order to bridge this gap in knowledge, this study was carried out in West Pokot, a semi-arid region in northwestern Kenya, where deep and severely eroded gullies, carved by water and intensified by recurrent droughts, flash floods, overgrazing, deforestation and inappropriate agricultural practices, have ravaged the landscape [38]. This gullied landscape has instigated the loss of animal and human life, increased the physical separation of neighbors, aggravated water scarcity during the dry season by lowering the water table, and reduced the amount of cultivable, habitable and grazing land, threatening the agro-pastoral system. This study aimed to (i) develop logistic regression, support vector machines, boosted regression trees and random forest models and compare their performance in predicting the spatial patterns of gully erosion susceptibility; and (ii) establish the comparative importance of the environmental factors that determine the gullying process. The logistic regression model provided a basis for comparing the machine learning models. The outputs will provide evidence of gully-prone sites to support the prioritization and

spatial targeting of technological solutions, interventions and programs for sustainable soil management, restoration of the degraded lands and realization of a land degradation-neutral environment in accord with Sustainable Development Goal 15.3 [39]. Moreover, the outputs will yield new evidence about the performance of different machine learning methods in the arid and semi-arid regions of Africa, which will ensure holistic scientific discussions on the best modeling approach for gully erosion susceptibility mapping.

2. Materials and Methods

Figure 1 summarizes the flow of data and methods that were instrumental in the production of the gully erosion susceptibility maps. The major steps comprised: (a) spatial data acquisition and preparation; (b) exploratory data analysis and variable selection; (c) model development (i.e., fitting, evaluating and comparing models); and (d) spatial prediction and mapping (i.e., the application of the models to generate spatially-distributed gully erosion susceptibility maps).

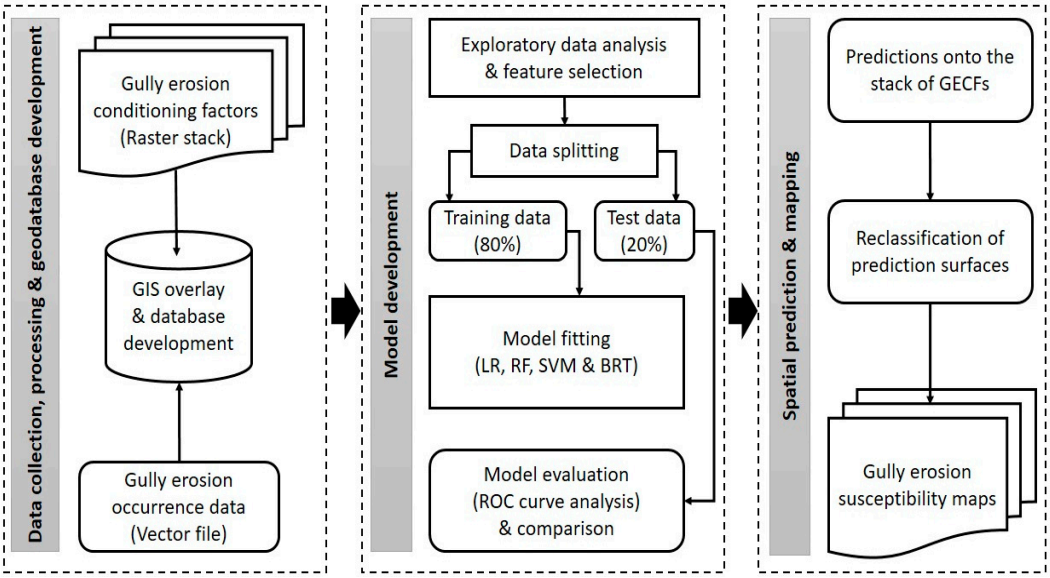


Figure 1. Flowchart of the data and methods used for gully erosion susceptibility mapping.

2.1. Description of the Study Area

This study was conducted in Senetwo location, West Pokot County, Kenya (Figure 2), which covers approximately 49 km^2. It lies between latitudes 1°18′–1°23′ N and longitudes 35°7′–35°12′ E, with the altitudes varying from 1510 to 2180 m above sea level. In terms of climate, the mean annual temperature is about 22 °C, while the mean annual rainfall is about 750 mm. The rainfall regime is bimodal, with the long-rain season starting from March to May and the short-rain season from August to November [40,41]. The soils are well-drained, shallow to moderately deep, sandy loam to sandy clay loam, dark brown to dark reddish-brown (or dark reddish-brown to dark red), the underlying metamorphic rocks of which are mostly gneisses, rich in ferromagnesian minerals [42]. The physiography is characterized by hills and mid-level uplands with gentle and moderately steep slopes covered by grasses and sparse trees. Small-scale agro-pastoralism dominates the land use and economic organization of the residents, who are mainly the Pokot, with food crops, such as maize, beans, sorghum and millet being grown, and animals, such as cattle, goats, sheep and chickens being kept [43].

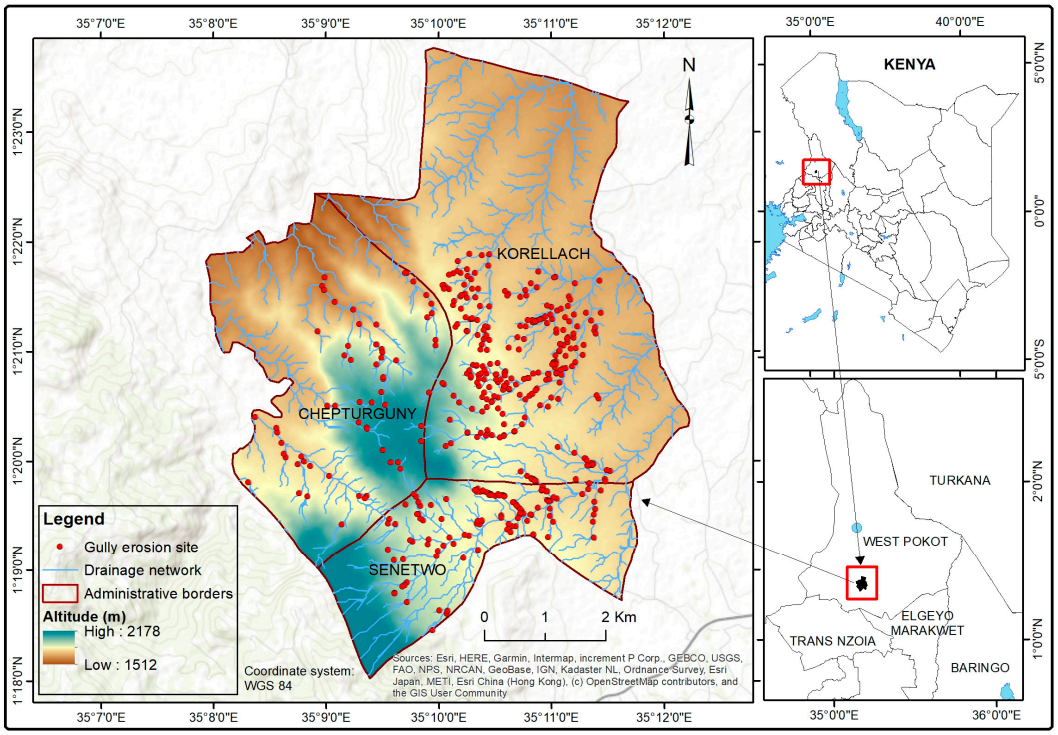


Figure 2. Location of the study area and gully erosion sites.

2.2. Spatial Data Acquisition and Preparation

The data depicting gully erosion occurrence and the conditioning factors were procured from various sources in different formats and preprocessed to build models for predicting and mapping the spatial patterns of gully erosion susceptibility in the study area.

2.2.1. Gully Erosion Occurrence Data

The gully locations were extracted from high-resolution satellite imagery on Google Earth Pro through visual interpretation and were verified through a field visit in June 2021. New gully locations were also added during the field visit (Figure 3). A total of 431 points were gathered and spatially referenced to represent the presence of gullies. For modeling purposes, a similar number of points were randomly selected by using geographical information systems (GIS) tools to represent the absence of gullies [19]. Ultimately, a balanced set of data and an inventory of gullies were created (Figure 2).

Figure 3. Photos illustrating the form of gullies in the study area.

2.2.2. Gully Erosion Conditioning Factors

Twenty-four (24) auxiliary datasets depicting the conditioning factors that could potentially explain and predict the occurrence of gullies were gathered in raster formats

based on a literature review, data availability and expert knowledge. The rainfall, digital elevation model, soil texture (clay and sand content) and atmospherically-corrected Landsat 8 operational land imager data were directly downloaded from existing geodata portals (see Table 1), while the rest of the data were derived from them. That is, 17 terrain parameters, including the slope aspect, slope angle, slope length-steepness factor, topographic wetness index, topographic position index, topographic ruggedness index, stream power index, flow accumulation, flow direction, plan curvature, profile curvature, convergence index, convexity, landform, valley depth, geomorphons and texture were extracted from a digital elevation model without sinks using the relevant terrain analysis tools in a GIS [44].

Table 1. Gully erosion conditioning factors.

Factor	Scale	Proxy for/Effects	Source
Elevation (DEM)	30 m	Micro-climate, vegetation, drainage network	https://earthexplorer.usgs.gov (Accessed on 1 August 2021)
Rainfall (1970–2000)	1 km	Soil moisture, volume of surface runoff, sediment transport capacity, slope stability	https://worldclim.org (Accessed on 1 August 2021)
Slope angle (gradient)	30 m	Overland and subsurface flows, erosive energy of overland flow, flow velocity, drainage density, sediment transport capacity, infiltration rate	DEM
Slope length-steepness (LS) factor			
Flow accumulation	30 m	Soil moisture (saturation), surface runoff	DEM
Topographic wetness index			
Slope aspect	30 m	Evapotranspiration, soil moisture, vegetation structure, weathering rate, micro-climate	DEM
Plan curvature	30 m	Concentration of overland flow, flow velocity (rate)	DEM
Profile curvature			
Convexity			
Convergence index			
Terrain ruggedness index			
Topographic position index			
Geomorphons			
Landform			
Texture			
Valley depth			
Stream power index	30 m	Stream incision, slope erosion	DEM
Land use/cover	30 m	Slope stability, evapotranspiration, infiltration, overland flow, surface runoff generation, sediment dynamics	Landsat 8 OLI imagery https://earthexplorer.usgs.gov (Accessed on 1 August 2021)
NDVI			
Drainage density	30 m	Flow magnitude, sediment transport capacity, infiltration, surface runoff	DEM
Distance to stream			
Clay content	0–20 cm depth	Infiltration rate, surface runoff, erosion resistance, subsurface flow and piping	https://isda-africa.com/isdasoil (Accessed on 1 August 2021)
Sand content			

Additionally, the land cover layer was obtained through the supervised classification of Landsat 8 satellite imagery, utilizing the maximum likelihood algorithm [45], while the normalized difference vegetation index layer was derived by manipulating the surface reflectance values from the Landsat 8 bands 4 and 5 using the raster algebra operations [46]. A stream networks layer was first created from the (depressionless) digital elevation model using hydrological tools to process the drainage density and distance to stream layers.

Thereafter, the line density tool was used to create the drainage density layer, whereas the Euclidean distance tool was applied to compute the distance to stream layer.

The 23 thematic raster layers were clipped to the region of interest, resampled to a common spatial resolution (30 m) using the bilinear interpolation method, transformed to UTM WGS84 Zone 36N and stacked. After that, the balanced set of gully erosion occurrence points was overlaid on the stack of conditioning factors to extract the raster values to the points. Finally, a spatial database was built, which served as the input data for gully erosion susceptibility modeling.

2.3. Spatial Modeling, Prediction and Mapping

Before modeling, the input data were randomly split into training and testing datasets in the ratio of 80:20. The former dataset was used to fit models, while the latter was used to validate the fitted models in terms of their predictive capacity.

2.3.1. Exploratory Data Analysis

Multi-collinearity among the gully erosion conditioning factors was detected by computing their variance inflation factors in a regression model and by analyzing their correlation coefficients. From a pair, one conditioning factor was discarded when the correlation coefficient was equal to or more than 0.8 [47]. Moreso, the conditioning factors with variance inflation factors exceeding 10 were excluded from further analysis [48]. This step reduced redundancy and ensured the accurate estimates of model parameters and measures of statistical significance.

2.3.2. Model Development

Four predictive models were built using logistic regression, support vector machines, boosted regression trees and random forest techniques.

- Logistic regression

The target variable had only two possible outcomes: the presence (1) and absence (0) of gullies. Hence, the binary logistic regression technique was appropriate for explaining the relationship between the target variable and the associated conditioning factors, as well as for predicting the probability of a gully event. The binary logistic regression technique is based on logit transformation [48,49], which aims at linearizing the S-shaped logistic response function as defined in Equation (1):

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \propto + \sum_{i=1}^{n} \beta_i x_i \quad (1)$$

where $\hat{\pi}(x)$ is the probability that an event will occur, $\propto$ is the constant and β_i is the coefficient of the explanatory variables x_i. The ratio $\frac{\pi(x)}{1-\pi(x)}$ is known as the odds, or likelihood ratio, while $\log\left(\frac{\pi(x)}{1-\pi(x)}\right)$ is referred to as log odds, or the logit transformation of $\hat{\pi}(x)$. After transformation, Equation (2) is applied to convert the log odds into conditional probabilities and force the values to lie between 0 and 1. Values close to 1 imply a higher chance of an occurrence.

$$\hat{\pi}(x) = \frac{e^{\propto + \sum_{i=1}^{n} \beta_i x_i}}{1 + e^{\propto + \sum_{i=1}^{n} \beta_i x_i}} = \frac{1}{1 + e^{-(\propto + \sum_{i=1}^{n} \beta_i x_i)}} \quad (2)$$

First, a full logistic regression model was fitted using the maximum likelihood method to estimate the model parameters and, thereafter, was reduced by a stepwise regression algorithm. The reduced model was selected on the basis of the Akaike information criterion, while its goodness-of-fit, significance and parameter estimates were tested using the Hosmer–Lemeshow, likelihood ratio and Wald statistics [48,49], respectively.

- Random forest

This technique uses recursive binary splitting to grow several uncorrelated decision trees, which it ultimately combines to make a classification. The classifier draws random samples with replacements from the training data and grows one tree for each sample. Two-thirds of the sample is used to grow a tree, while the out-of-bag sample is used to assess its predictive accuracy and the importance of the environmental covariates. In growing a tree, the random forest algorithm splits the feature space at each tree node into two by using a random subset of environmental covariates and then groups the target variable at the two descendant nodes to minimize dissimilarities [50]. Binary splitting continues recursively until the sample size goes below a certain threshold at a terminal node. Three hyperparameters are defined in the process, including the (i) number of trees to grow (*ntree*), (ii) smallest sample size at each terminal node (*nodesize*), and (iii) the number of environmental covariates to consider at each node for splitting (*mtry*). In this study, the *mtry* was determined through a special tuning algorithm, whereas the *ntree* and the *node size* were set to 1000 and 5 (default values), respectively.

For the predictions, the random forest algorithm takes new data through each tree and classifies them based on majority voting [51,52]. It also computes the probability of each class membership from the fraction of votes received. The mean decrease accuracy is calculated to rank the covariates according to their importance. Firstly, the prediction error for each tree is estimated using the out-of-bag sample. Thereafter, the values of the covariates in the out-of-bag samples are randomly permutated in turn, and the prediction errors are re-calculated using the modified out-of-bag data. Finally, the differences in the average out-of-bag errors before and after permutation indicate variable importance.

- Support vector machines

This is a supervised learning technique that aims to minimize the structural or empirical risk as it separates classes [53]. By using kernel functions, this technique transforms the original input data from a low dimensional space, where classes are linearly inseparable, into a feature space of much higher dimensionality, where it fits an optimal separating hyperplane, which maximizes the margins of the boundaries of two classes with minimal errors and complexity [30,53,54]. The fitted nonlinear hyperplane easily classifies (or predicts) new data. Here, we used the Gaussian radial basis function kernel (Equation (3)) to convert the original input data into a higher dimension.

$$K(x_i, x_j) = \exp^{-\frac{||x_i, x_j||^2}{2\sigma^2}} \quad (3)$$

where K is the kernel function, x is the input vector, and σ is the bandwidth parameter (*sigma*), which controls the degree of nonlinearity in the hyperplane [55]. Sigma and the regularization (*cost*) parameter had to be specified. The latter governs the tradeoff between the complexity of the model and empirical errors, which also controls overfitting. Optimal values for these two parameters were chosen through the grid search method with 10-fold cross-validation [56].

- Boosted regression trees

Similar to random forest, the boosted regression trees algorithm constructs a set of trees but merges them by using a boosting technique to derive a final model with an improved predictive ability [57]. That is, it bootstraps the training data and constructs the first regression tree by applying equal weights to the data points. Thereafter, the boosted regression trees algorithm fits the second regression tree by giving higher weights to the observations that were poorly predicted in the first step. This process continues iteratively until a model with a low prediction error is obtained. The final model is a summation of the regression trees fitted in the entire iterative process [57,58]. During model building, five hyperparameters are tuned, including the (i) *learning rate* to determine each tree's contribution to the model, (ii) *tree complexity* to control the depth of the variable

interactions or the number of splits in each tree, (iii) *bag fraction* to specify the proportion of the bootstrapped sample for each boosting iteration, (iv) *number of trees* to indicate the number of trees to be fitted, considering the learning rate and tree complexity, and (v) *terminal node size* to define the minimal sample size at the terminal nodes [59]. In this study, a few combinations of the five hyperparameters were tested in a grid search process to establish the values that yielded the minimum prediction error.

2.3.3. Model Evaluation and Comparison

The predictive power of the models was determined by computing the area under the receiver operating characteristic curve (i.e., AUC), utilizing the testing data. The curve plotted sensitivity (true positives) as a function of 1—the specificity (false positives) for all the possible cut-off values that could be taken to interpret the predicted probabilities as gully erosion events [60,61]. The curve depicted the tradeoff between the rate of making true predictions of gully erosion events and that of making false predictions [62]. The computed AUC values ranged between 0.5 and 1, with 0.5 implying a random, 0.6–0.7 a good, 0.7–0.8 an acceptable, 0.8–0.9 an excellent, and 0.9–1.0 an outstanding performance [63]. Apart from the AUC, the specificity, sensitivity and overall accuracy were also calculated for each model using Equations (4)–(6) after cross-classifying the observed and the predicted gully and non-gully events in a 2 × 2 contingency table (Table 2) [49].

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

where TP = true positive and FN = false negative

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{5}$$

where TN = true negative and FP = false positive

$$\text{Overall accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{6}$$

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

Table 2. Cross-tabulation of the observed and predicted gully erosion events.

		Predicted	
Observed		Presence (1)	Absence (0)
	Presence (1)	TP (1 | 1)	FN (1 | 0)
	Absence (0)	FP (0 | 1)	TN (0 | 0)

Note: TP = true positive, TN = true negative, FP = false positive, and FN = false negative.

2.3.4. Model Application

The parameters of each model were applied to the stack of gully erosion controlling factors to create prediction raster surfaces, which were subsequently reclassified into five categories (natural groupings) based on the similarities in the data values using Jenks' natural breaks method. The five classes reflected areas of very low, low, moderate, high and very high susceptibility to gully erosion. Finally, the gully erosion susceptibility maps were produced.

2.4. Software

Data preparation, geocomputation, spatial prediction and mapping were executed on ArcGIS, QGIS, SAGA GIS and R Studio platforms [64].

3. Results

3.1. Exploratory Data Analysis

Correlation analysis revealed that the rainfall and elevation (r = 0.84), sand content and clay content (r = 0.87), and landform and slope gradient (r = 0.92) were highly correlated. Thus, rainfall, clay content and landform were eliminated from subsequent analysis to break up the near-linear dependency in the gully erosion conditioning factors and reduce redundancy. The absence of multi-collinearity in the remaining 21 predisposing factors was also confirmed by their respective variance inflation factors, which did not exceed 10 (Table 3). Otherwise, multi-collinearity would have biased the estimation of the parameters and measures of statistical significance, leading to prediction inaccuracies, especially in logistic regression modeling.

Table 3. Results of the multi-collinearity tests.

Factor	VIF	Factor	VIF	Factor	VIF
Aspect	1.28	LS Factor	5.51	Flow direction	1.18
Convergence index	2.49	NDVI	2.86	Flow accumulation	1.23
Convexity	1.46	Sand content	3.49	Geomorphons	2.41
Plan curvature	3.57	Slope gradient	2.25	Land cover	2.25
Profile curvature	1.85	Stream power index	1.62	Topographic position index	1.97
Drainage density	1.54	Distance to stream	1.61	Topographic wetness index	1.75
Elevation	3.07	Texture (SAGA)	1.25	Valley depth	2.58

Note: VIF = variance inflation factor; NDVI = normalized difference vegetation index; LS Factor = slope length and steepness factor.

3.2. Models of Gully Erosion Susceptibility and Relative Importance of the Conditioning Factors

The remaining 21 conditioning factors were used to build the logistic regression, support vector machines, boosted regression trees and random forest models that were subsequently applied to predict gully erosion occurrence. According to the p values of the likelihood ratio, Hosmer–Lemeshow and the Wald statistics presented in Table 4, the resultant logistic regression model was statistically significant, fitted the data adequately, and eight out of the twenty-one conditioning factors had significant effects on gully erosion occurrence. In particular, the drainage density, sand content, valley depth, elevation and stream power index had increasing effects on the odds of gully erosion occurrence; that is, the odds multiplied by 1.22, 1.21, 1.02, 1.01 and 1.00, respectively, for every one-unit increase in their values. By contrast, the distance to stream, slope gradient and plan curvature had decreasing effects on the odds of gully erosion occurrence; that is, the chances were reduced by about 1%, 95% and 100%, respectively, for every one-unit change in their values. Although the effects of the valley depth, elevation, stream power index and distance to stream were statistically significant, their magnitudes were rather inconsiderable. In addition, the odds ratios of these three factors were close to 1, implying that increasing them by one unit only slightly changed the likelihood of gully formation.

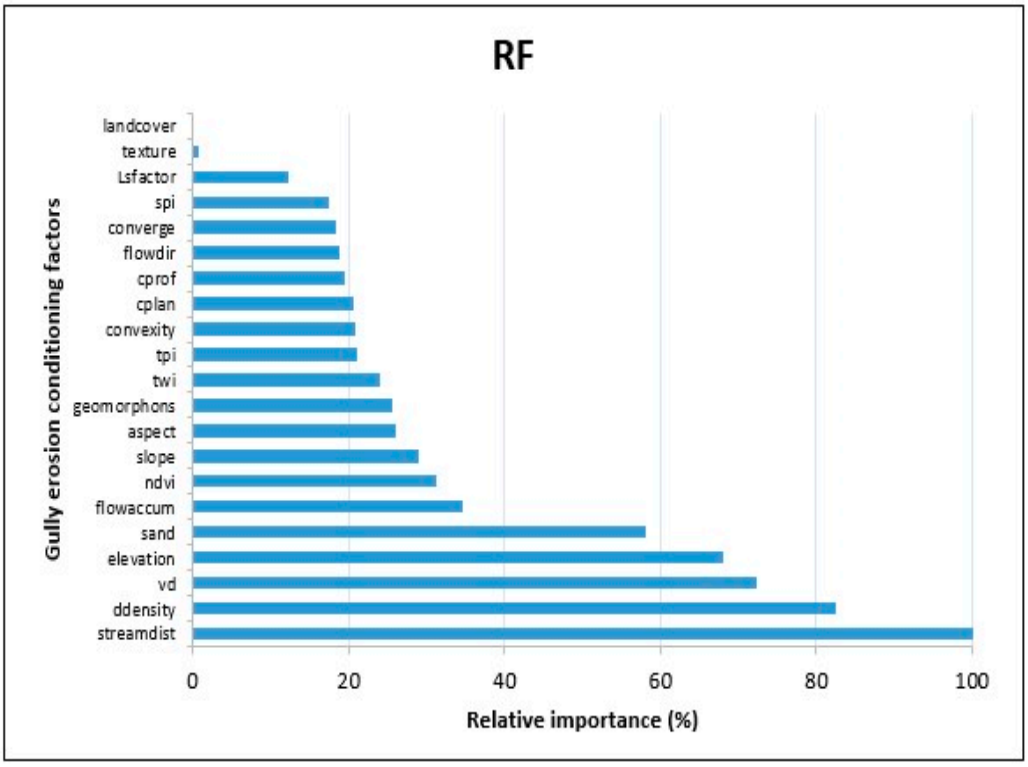
Unlike the logit model, which estimated the parameters that quantified the effects of the conditioning factors on the likelihood of gully erosion occurrence, the three machine learners ranked their significance based on their mean-decrease-in-accuracy scores (Figure 4a–c). These scores were subsequently converted into percentages. The results showed that the boosted regression trees, support vector machines and random forest models ranked the conditioning factors somewhat differently. According to the support vector machines and random forest models, distance to stream was the most important determinant of gully erosion susceptibility, followed by the drainage density and valley depth (Figure 4a,c), whereas the boosted regression trees model ranked sand content, elevation and stream density as the three most important factors that control gullying (Figure 4b). Furthermore, the results showed that land cover had the least influence in the boosted regression trees and random forest models, while the stream power index had the lowest importance in the support vector machines model. It is also notable that the order

of importance of the 21 conditioning factors varied depending on the machine learning algorithm used for model development.

Table 4. Summary statistics for the reduced logistic regression model.

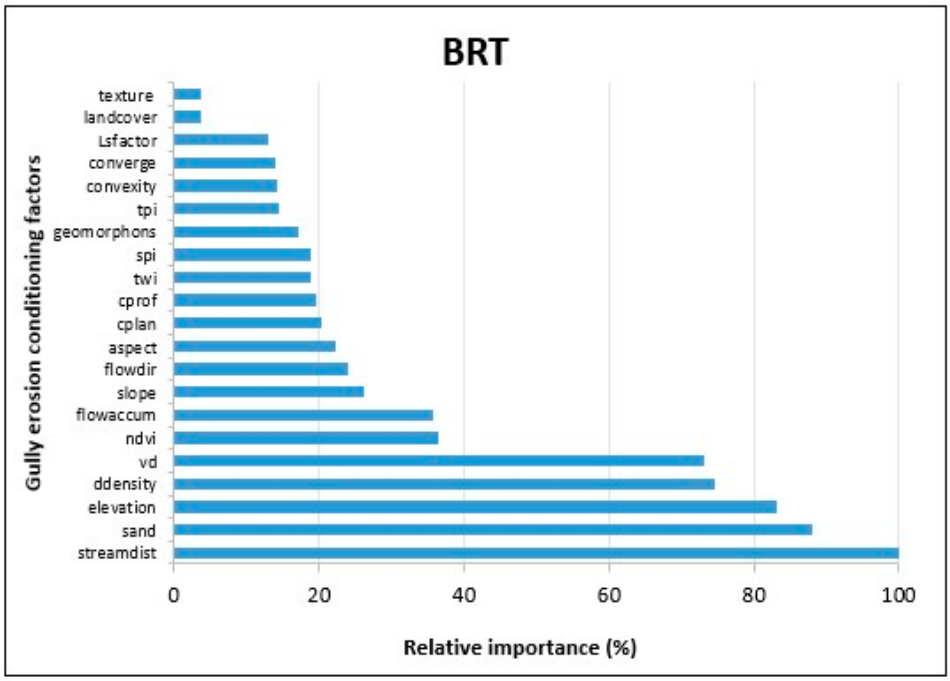
Parameter	Estimate	Std. Error	Odds Ratio	*p* Value
(Intercept)	−36.1502	5.0454	0.0000	0.0000
Plan curvature	−148.9885	42.8495	0.0000	0.0005
Drainage density	0.1991	0.0554	1.2203	0.0003
Sand content	0.1879	0.0276	1.2067	0.0000
Elevation	0.0137	0.0022	1.0138	0.0000
Valley depth	0.0165	0.0036	1.0167	0.0000
Distance to stream	−0.0095	0.0029	0.9906	0.0009
Slope gradient	−2.9885	1.3371	0.0504	0.0254
Stream power index	0.0001	0.0000	1.0001	0.0227
Pr > LRo χ^2	0.0000			
Pr > HL χ^2	0.7072			

Note: LR_o = likelihood ratio; HL = Hosmer–Lemeshow.

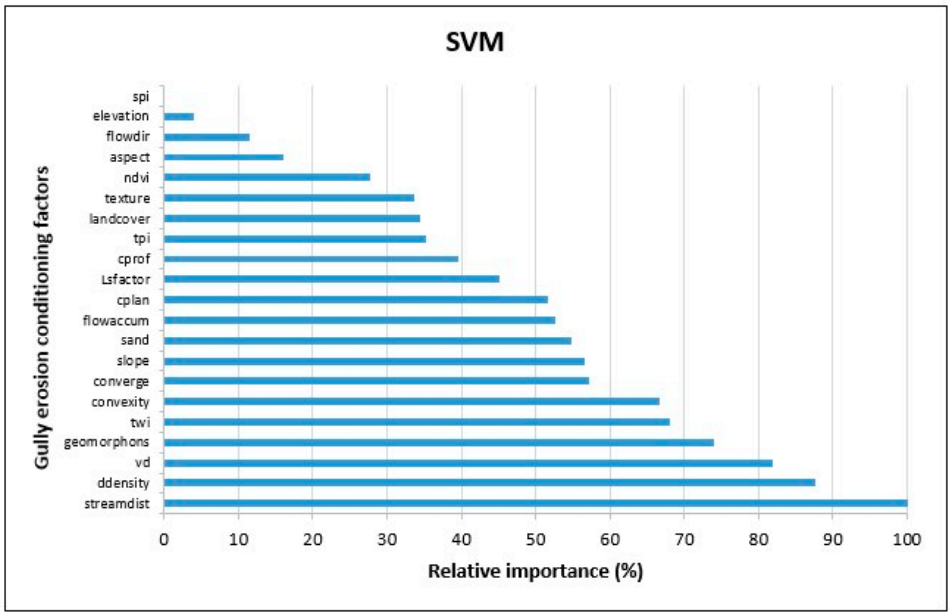


(**a**)

Figure 4. *Cont.*



(**b**)



(**c**)

Figure 4. (**a**) Variable importance for the random forest model. Streamdist = distance to stream; ddensity = drainage density; flowaccum = flow accumulation; ndvi = normalized difference vegetation index; twi = topographic wetness index; spi = stream power index; cprof = profile curvature; cplan = plan curvature; flowdir = flow direction; tpi = topographic position index; converge = convergence index; vd = valley depth; LSfactor = length-slope factor. (**b**) Variable importance for the boosted regression trees model. (**c**) Variable importance for the support vector machines model.

3.3. Model Evaluation and Comparison

The high AUC values (>0.80) and receiver operating characteristic curves, shown in Figure 5, indicate that the fitted logistic regression, boosted regression trees, support vector machines and random forest models had excellent capacity to predict new gully erosion occurrence and that each model provided a greater true positivity rate for any given false positivity rate. Boosted regression trees and random forest models displayed

similar predictive capabilities, having the highest AUC (0.89), followed by support vector machines (AUC = 0.86) and logistic regression (AUC = 0.85). Evidently, the differences in the AUC values of the four models were marginal. The overall prediction accuracy, sensitivity and specificity values were also high (Table 5), affirming the good performance of the models. Specifically, the values of these three performance indicators exceeded 70% in all the models except for the specificity of the logistic regression model (66%), indicating that the benchmark logit model had a slightly lower ability to predict the absence of gullies relative to the machine learning models.

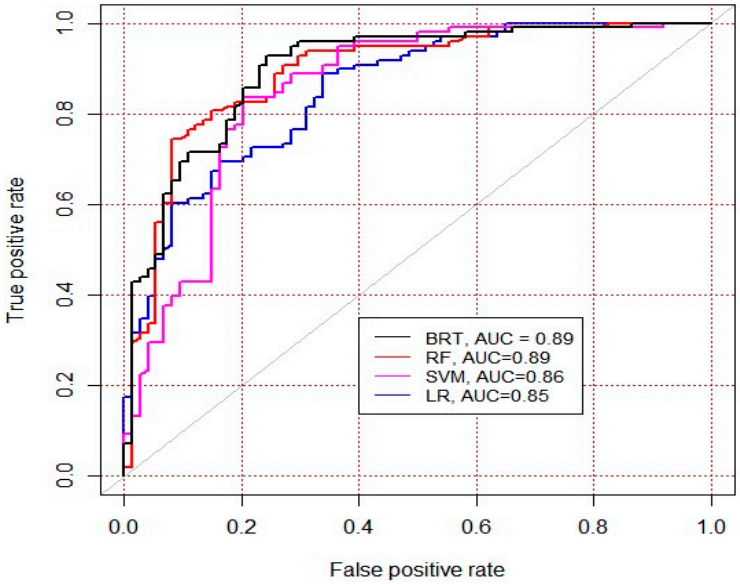


Figure 5. Evaluation of the boosted regression trees, support vector machines and logistic regression models using receiver operating characteristic curves.

Table 5. Evaluation of the models using proportions in 2 × 2 contingency tables.

Model	Predicted			
	Observed	**Presence**	**Absence**	**% Correct**
LR	Presence	80	18	81.6 a
	Absence	25	49	66.2 b
		Overall accuracy (%)		79.6
SVM	Presence	79	19	80.6 a
	Absence	16	58	78.4 b
		Overall accuracy (%)		79.1
BRT	Presence	77	21	78.6 a
	Absence	14	60	81.1 b
		Overall accuracy (%)		79.7
RF	Presence	80	18	81.6 a
	Absence	16	58	78.4 b
		Overall accuracy (%)		80.2

Note: a = sensitivity; b = specificity.

3.4. Spatial Patterns of Gully Erosion Susceptibility

Visually, the mapping results of the boosted regression trees, support vector machines and logistic regression models had some similarities in terms of the spatial distribution of the probability of gully erosion occurrence in the area (Figure 6). Zones of high to very high erosion susceptibility mostly appeared in the eastern and southeastern parts, whereas those of very low susceptibility mostly occurred on the western side in the four resultant maps.

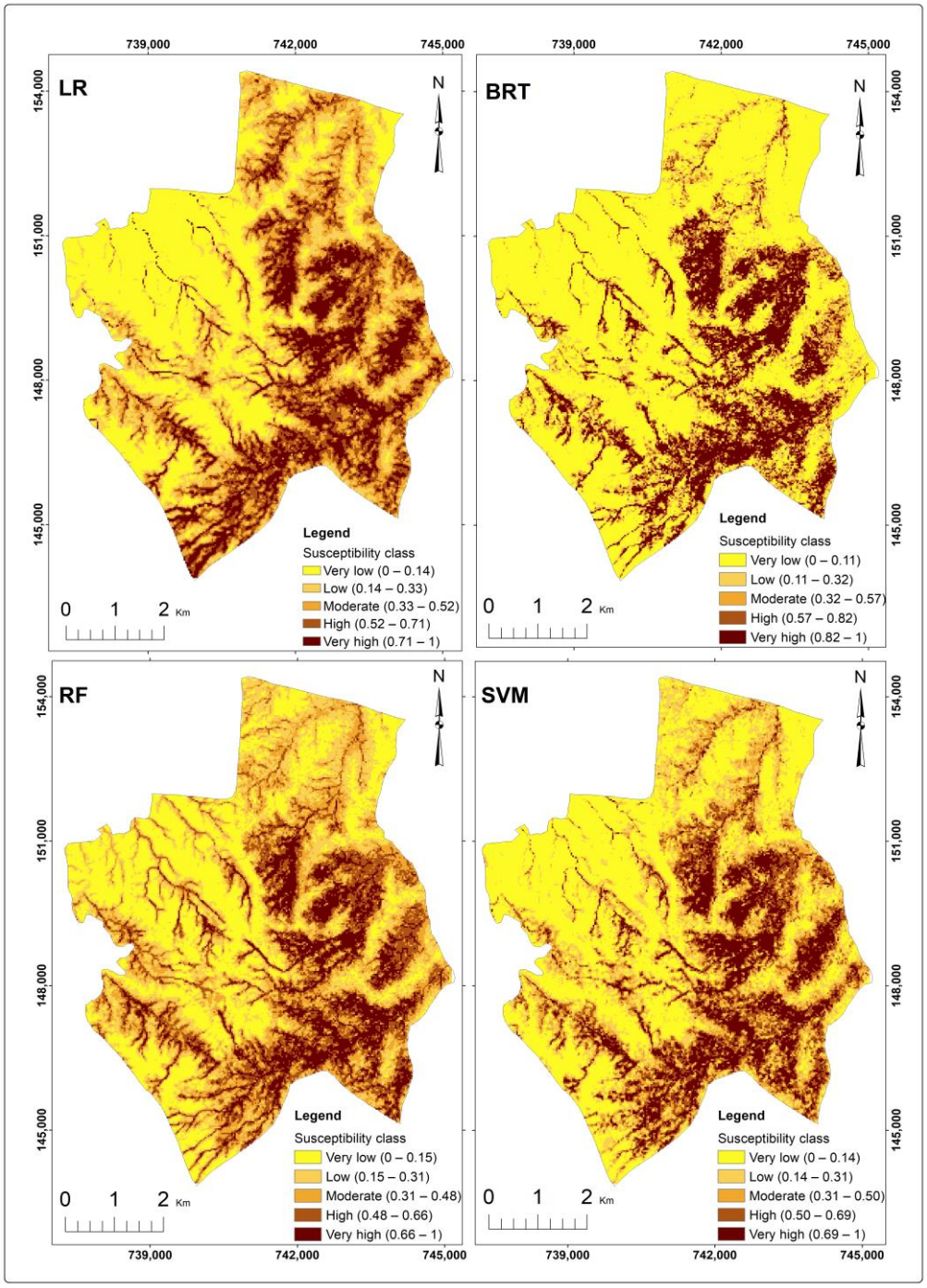


Figure 6. Spatial patterns of the susceptibility to gully erosion.

However, quantitative analysis of the mapping results revealed the less obvious areal differentiations occasioned by the four modeling approaches (Table 6). The random forest-based map showed that 51% of the landscape had low (24%) to very low (27%) gully erosion susceptibility, 21% had moderate susceptibility, and the rest had high (18%) to very high (12%) susceptibility. Unlike random forest and the support vector machines, the boosted regression trees model predictions assigned a large portion of the study area to the very low (56%) and a small part to the moderate (9%) and high (9%) susceptibility classes.

Table 6. Areal extents and proportions of the susceptibility classes.

Class	LR		BRT		RF		SVM	
	Area (km^2)	%	Area (km^2)	%	Area (km^2)	%	Area (km^2)	%
Very low	14.81	30.41	27.06	55.56	13.12	26.95	18.07	37.10
Low	9.66	19.84	5.80	11.91	11.56	23.73	9.59	19.69
Moderate	8.90	18.28	4.18	8.59	9.77	20.06	7.46	15.32
High	8.79	18.04	4.15	8.51	8.59	17.63	7.12	14.61
Very high	6.54	13.43	7.52	15.43	5.66	11.63	6.47	13.28
Total	48.71	100.00	48.71	100.00	48.71	100.00	48.71	100.00

4. Discussion

4.1. Relative Importance of the Gully Erosion Conditioning Factors

In this study, the support vector machines, boosted regression trees and random forest techniques were applied to map gully erosion susceptibility and establish the importance of its conditioning factors. Despite being applied in the same environmental setting, the three machine learning techniques yielded inconsistent environmental variable-importance results. The support vector machines and random forest models ranked distance to stream, drainage density and valley depth, while the boosted regression trees model ranked elevation, sand content and distance to stream as the three most influential conditioning factors. Conflicting variable-importance results have been widely observed and reported. For example, Chuma et al. [65] conducted a similar study in the Democratic Republic of Congo using artificial neural network, random forest, boosted regression trees and maximum entropy algorithms and noted a variation in the environmental determinants of gully erosion from one model to another. Arabameri et al. [66] ascribed such conflicts to the complexity of factors involved in gully development and the inherent structural differences in machine learning techniques. That is, each machine learning technique manipulates the input data differently to unravel the complex linkages between the gully erosion events and predisposing factors. The structural differences might also explain the variations in the proportion of land under each susceptibility class for the four modeling approaches (Table 6).

Nonetheless, the ranking of the most important conditioning factors seemed plausible because, for instance, it was evident from the resultant maps and field observations that most of the areas with high to very high gully erosion susceptibility were concentrated around the drainage channels and on the eastern and southeastern parts where the elevation was relatively low. Tien Bui et al. [25], Nhu et al. [29] and Pourghasemi et al. [31] also found that gully erosion evolved as the distance to streams decreased, drainage density increased and elevation decreased. This can be attributed to the effects of these factors on the flow magnitude, velocity and concentration, as well as on the sediment transport capacity, vegetal cover and mechanisms, such as incision, seepage and piping [24,67].

Lastly, it is worth noting that our variable-importance results differed from those reported by Amare et al. [68] and Busch et al. [9], who also conducted their studies in semi-arid contexts in the region. They found land use, drainage density and elevation to be the most important factors that control gully erosion. However, such discrepancies are unsurprising because, as Vanmaercke et al. [12] and Nhu et al. [29] argued, gully erosion conditioning factors tend to be context-specific and vary with spatial scales and spatial data attributes, including scale, quality, type and quantity. This implies that the most important conditioning factors reported here should not be carelessly extrapolated to other semi-arid regions; instead, appropriate investigations would be needed to unravel the important environmental drivers of gully erosion in each setting.

4.2. Model Evaluation and Comparison

Again, the three machine learning methods yielded slightly different performance results in the same environment and a small improvement in the accuracy of predicting gully erosion. Boosted regression trees and random forest models returned the highest AUC (0.89), followed by support vector machines (AUC = 0.86) and logistic regression (AUC = 0.85). The better performance of the random forest model compared to the support vector machines and logistic regression models coincides with the findings of comparable studies conducted in other regions [32,33,69–71]. However, the similarity of performance exhibited by the boosted regression trees and random forest models differs from the results of Arabameri et al. [23], Chuma et al. [65], Hembram et al. [72], Amiri et al. [73], Rahmati et al. [74] and Wang et al. [75], who reported that random forest outperformed the boosted regression trees in modeling gully erosion susceptibility. The superiority of random forest has been proven in many environmental modeling studies, such as landslide susceptibility mapping, because, as Youssef et al. [76] pointed out, it generates stable results, avoids overfitting and deals with missing data, outliers and multi-collinearity relatively well.

Interestingly, some past studies achieved higher predictive accuracies than this one. For instance, within the region, Busch et al. [9] and Amare et al. [68] realized AUC values of 0.99 and 0.95, respectively, while outside of the region, Pourghasemi et al. [31], Gayen et al. [72] and Saha et al. [77] attained 0.96, 0.96 and 0.99, respectively, using the random forest algorithm. Such variations in the performance statistics can be attributed to several factors, ranging from the optimization (tuning) of hyperparameters and site characteristics to the quality and quantity of the auxiliary datasets. In terms of data quality and quantity, some of the auxiliary data used in this study, including the sand content and digital elevation model, were by themselves products of spatial prediction. The digital elevation model was further used to derive other gully erosion conditioning factors, such as slope and aspect; hence, the inherent errors may have been propagated in the subsequent procedures and biased the model results. Nevertheless, these were the best and most readily available data at the time. In addition, the effects of other important soil-related conditioning factors, such as the exchangeable sodium, sodium adsorption ratio, electrical conductivity, soil depth, water-dispersible clays and lithology, were not explicitly accounted for due to lack of data. This could have also impacted model performances. Thus, in future research, the models should be re-evaluated and refined to account for the missing soil-related variables as the data become available.

5. Conclusions and Outlook

This study aimed to develop, evaluate and compare the performance of boosted regression trees, support vector machines, random forest and logistic regression models in mapping gully erosion risk, as well as to determine the most important conditioning factors in a semi-arid landscape of northwestern Kenya. The resulting models exhibited excellent predictive capabilities, although the machine learning models showed their superiority over the benchmark logistic regression model. Differences in the performance of the models were rather small; thus, we conclude that each technique is promising and has the potential to generate reliable maps, information and tools for gully erosion risk management in similar semi-arid environments globally. However, the performance of the models and the importance of the gully erosion conditioning factors might be variable. It would always be worthwhile to compare a set of machine learning models and select the best-performing model to generate a gully erosion susceptibility map in each context. Regarding the relative influence of the conditioning factors, random forest and support vector machine analyses revealed that the distance to stream, drainage density and valley depth were the top three environmental factors that predisposed the region to gully erosion.

Besides its contribution to the state of knowledge and literature on the performance of state-of-the-art machine learning techniques in the less-studied and data-scarce arid and semi-arid lands of Africa, the findings also improve our general understanding of

the spatial dynamics and determinants of gully erosion. Such an understanding, coupled with the resultant gully erosion susceptibility maps, can support agro-environmental experts and other stakeholders in strategic planning, formulation of strategies, and efficient allocation of resources to prevent, halt and reverse land degradation in order to achieve land degradation-neutral landscapes in compliance with Sustainable Development Goal 15.3. For example, the output maps from this study suggest to potential agro-environmental projects that intensive monitoring and investments in sustainable land management technologies, innovations and practices, such as terracing, revegetation and check dams, should be targeted at the zones with high to very high likelihood of gully erosion on the eastern and southeastern parts of the study area. Future research should re-evaluate and refine the models to account for other soil-related gully erosion conditioning factors, such as the exchangeable sodium, sodium adsorption ratio, electrical conductivity, soil depth and water-dispersible clays as these data become available. Moreover, studies attempting to unravel the future evolution of gully erosion in the study area and similar environments would be of great interest.

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

Funding: This research was developed within the framework of the GIS Support Project (Project no: 17-229) and the Drylands Farmers Research Network Project (Project no: 19-135), funded by the McKnight Foundation (USA).

Data Availability Statement: The data presented in this study are available on request from the corresponding author (KW).

Acknowledgments: The authors thank the County government of West Pokot, Thomas Lokoriong'or (local community representative) and Francis Musungu (driver) for their invaluable support and cooperation during the field campaigns.

Conflicts of Interest: The authors declare no conflict 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. Lorenz, K.; Lal, R.; Ehle, K. Soil organic carbon stock as an indicator for monitoring land and soil degradation in relation to United Nations' Sustainable Development Goals. *Land Degrad. Dev.* **2019**, *30*, 824–838. [CrossRef]
2. FAO and ITPS. *Status of the World's Soil Resources—Main Report*; Food and Agriculture Organization of the United Nations and Intergovernmental Technical Panel on Soils: Rome, Italy, 2021. Available online: https://www.fao.org/documents/card/en/c/c6814873-efc3-41db-b7d3-2081a10ede50/ (accessed on 25 August 2021).
3. UNCCD. *The Global Land Outlook*, 1st ed.; United Nations Conventions to Combat Desertification: Bonn, Germany, 2017. Available online: https://www.unccd.int/sites/default/files/documents/2017-09/GLO_Full_Report_low_res.pdf (accessed on 10 April 2023).
4. Rahmati, O.; Tahmasebipour, N.; Haghizadeh, A.; Pourghasemi, H.R.; Feizizadeh, B. Evaluating the influence of geo-environmental factors on gully erosion in a semi-arid region of Iran: An integrated framework. *Sci. Total Environ.* **2017**, *579*, 913–927. [CrossRef]
5. 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]
6. Conforti, M.; Aucelli, P.P.C.; 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]
7. Conoscenti, C.; Agnesi, V.; Angileri, S.; Cappadonia, C.; Rotigliano, E.; Märker, M. A GIS-based approach for gully erosion susceptibility modelling: A test in Sicily, Italy. *Env. Earth Sci.* **2013**, *70*, 1179–1195. [CrossRef]
8. Ireri, C.; Krhoda, G.O.; Mukhovi, M.S. Bivariate-based susceptibility mapping for gully erosion in Wanjoga River catchment Upper Tana Basin, Kenya. *East Afr. J. Sci. Technol. Innov.* **2021**, *2*, 1–15. [CrossRef]

9. Busch, R.; Hardt, J.; Nir, N.; Schütt, B. Modeling gully erosion susceptibility to evaluate human impact on a local landscape system in Tigray, Ethiopia. *Remote Sens.* **2021**, *13*, 2009. [CrossRef]
10. Roy, J.; Saha, S. Ensemble hybrid machine learning methods for gully erosion susceptibility mapping: K-fold cross validation approach. *Artif. Intell. Geosci.* **2022**, *3*, 28–45. [CrossRef]
11. Zabihi, M.; Mirchooli, F.; Motevalli, A.; Khaledi, D.A.; Pourghasemi, H.R.; Zakeri, M.A.; Sadighi, F. Spatial modelling of gully erosion in Mazandaran Province, northern Iran. *Catena* **2018**, *161*, 1–13. [CrossRef]
12. Vanmaercke, M.; Panagos, P.; Vanwalleghem, T.; Hayas, A.; Foerster, S.; Borrelli, P.; Rossi, M.; Torri, D.; Casali, J.; Borselli, L.; et al. Measuring, modelling and managing gully erosion at large scales: A state of the art. *Earth-Sci. Rev.* **2021**, *218*, 103637. [CrossRef]
13. Dewitte, O.; Daoudi, M.; Bosco, C.; Eeckhaut, M.V.D. Predicting the susceptibility to gully initiation in data-poor regions. *Geomorphology* **2015**, *228*, 101–115. [CrossRef]
14. Arabameri, A.; Rezaei, K.; Pourghasemi, H.R.; Lee, S.; Yamani, M. GIS-based gully erosion susceptibility mapping: A comparison among three data-driven models and AHP knowledge-based technique. *Environ. Earth Sci.* **2018**, *77*, 628. [CrossRef]
15. 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]
16. Yazie, T.; Mekonnen, M.; Derebe, A. Gully erosion and its impacts on soil loss and crop yield in three decades, northwest Ethiopia. *Model. Earth Syst. Environ.* **2021**, *7*, 2491–2500. [CrossRef]
17. Igwe, O.; Ikechukwu, J.U.; Onwuka, S.; Ozioko, O. GIS-based gully erosion susceptibility modeling, adapting bivariate statistical method and AHP approach in Gombe town and environs Northeast Nigeria. *Geoenviron. Disasters* **2020**, *7*, 32. [CrossRef]
18. Conoscenti, C.; Angileri, S.; Cappadonia, C.; Rotigliano, E.; Agnesi, V.; Märker, M. Gully erosion susceptibility assessment by means of GIS-based logistic regression: A case of Sicily (Italy). *Geomorphology* **2014**, *204*, 399–411. [CrossRef]
19. Conoscenti, C.; Agnesi, V.; Cama, M.; Caraballo-Arias, N.A.; Rotigliano, E. Assessment of gully erosion susceptibility using multivariate adaptive regression splines and accounting for hydrological connectivity. *Land Degrad. Dev.* **2018**, *29*, 724–736. [CrossRef]
20. Rahmati, O.; Haghizadeh, A.; Pourghasemi, H.R.; Noormohamadi, F. Gully erosion susceptibility mapping: The role of GIS-based bivariate statistical models and their comparison. *Nat. Hazards* **2016**, *82*, 1231–1258. [CrossRef]
21. Javidan, N.; Kavian, A.; Pourghasemi, H.R.; Conoscenti, C.; Jafarian, Z. Gully erosion susceptibility mapping using multivariate adaptive regression splines-replications and sample size scenarios. *Water* **2019**, *11*, 2319. [CrossRef]
22. Razavi-Termeh, S.V.; Sadeghi-Niaraki, A.; Choi, S. Gully erosion susceptibility mapping using artificial intelligence and statistical models. *Geomat. Nat. Hazards Risk* **2020**, *11*, 821–844. [CrossRef]
23. Arabameri, A.; Pradhan, B.; Pourghasemi, H.R.; Rezaei, K.; Kerle, N. Spatial modelling of gully erosion using GIS and R programing: A comparison among three data mining algorithms. *Appl. Sci.* **2018**, *8*, 1369. [CrossRef]
24. Avand, M.; Janizadeh, S.; Naghibi, S.A.; Pourghasemi, H.R.; Bozchaloei, S.K.; Blaschke, T. A comparative assessment of random forest and k-nearest neighbor classifiers for gully erosion susceptibility mapping. *Water* **2019**, *11*, 2076. [CrossRef]
25. Tien Bui, D.; Shirzadi, A.; Shahabi, H.; Chapi, K.; Omidavr, E.; Thai Pham, B.; Asl, D.T.; Khaledian, H.; Pradhan, B.; Panahi, M.; et al. A novel ensemble artificial intelligence approach for gully erosion mapping in a semi-arid watershed (Iran). *Sensors* **2019**, *19*, 2444. [CrossRef]
26. Band, S.S.; Janizadeh, S.; Pal, S.C.; Saha, A.; Chakrabortty, R.; Shokri, M.; Mosavi, A. Novel ensemble approach of Deep Learning Neural Network (DLNN) model and Particle Swarm Optimization (PSO) algorithm for prediction of gully erosion susceptibility. *Sensors* **2020**, *20*, 5609. [CrossRef] [PubMed]
27. Ghorbanzadeh, O.; Shahabi, H.; Mirchooli, F.; Kamran, K.V.; Lim, S.; Aryal, J.; Jarihani, B.; Blaschke, T. Gully erosion susceptibility mapping (GESM) using machine learning methods optimized by the multi-collinearity analysis and K-fold cross-validation. *Geomat. Nat. Hazards Risk* **2020**, *11*, 1653–1678. [CrossRef]
28. Lei, X.; Chen, W.; Avand, M.; Janizadeh, S.; Kariminejad, N.; Shahabi, H.; Costache, R.; Shahabi, H.; Shirzadi, A.; Mosavi, A. GIS-based machine learning algorithms for gully erosion susceptibility mapping in a semi-arid region of Iran. *Remote Sens.* **2020**, *12*, 2478. [CrossRef]
29. Nhu, V.; Janizadeh, S.; Avand, M.; Chen, W.; Farzin, M.; Omidvar, E.; Shirzadi, A.; Shahabi, H.; Clague, J.J.; Jaafari, A.; et al. GIS-Based Gully erosion susceptibility mapping: A comparison of computational ensemble data mining models. *Appl. Sci.* **2020**, *10*, 2039. [CrossRef]
30. Pourghasemi, H.R.; Yousefi, S.; Kornejady, A.; Cerdà, A. Performance assessment of individual and ensemble data-mining techniques for gully erosion modeling. *Sci. Total Environ.* **2017**, *609*, 764–775. [CrossRef]
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. Ahmadpour, H.; Bazrafshan, O.; Rafiei-Sardooi, E.; Zamani, H.; Panagopoulos, T. Gully erosion susceptibility assessment in the Kondoran watershed using machine learning algorithms and the Boruta feature selection. *Sustainability* **2021**, *13*, 10110. [CrossRef]
33. Arabameri, A.; Pal, S.C.; Costache, R.; Saha, A.; Rezaie, F.; Danesh, A.S.; Pradhan, B.; Lee, S.; Hoang, N. Prediction of gully erosion susceptibility mapping using novel ensemble machine learning algorithms. *Geomat. Nat. Hazards Risk* **2021**, *12*, 469–498. [CrossRef]

34. Saha, S.; Sarkar, R.; Thapa, G.; Roy, J. Modeling gully erosion susceptibility in Phuentsholing, Bhutan using deep learning and basic machine learning algorithms. *Environ. Earth Sci.* **2021**, *80*, 295. [CrossRef]
35. Yang, A.; Wang, C.; Pang, G.; Long, Y.; Wang, L.; Cruse, R.M.; Yang, Q. Gully erosion susceptibility mapping in highly complex terrain using machine learning models. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 680. [CrossRef]
36. Jaafari, A.; Janizadeh, S.; Abdo, H.G.; Mafi-Gholami, D.; Adeli, B. Understanding land degradation induced by gully erosion from the perspective of different geoenvironmental factors. *J. Environ. Manag.* **2022**, *315*, 115181. [CrossRef] [PubMed]
37. Vanmaercke, M.; Chen, Y.; Haregeweyn, N.; De Geeter, S.; Campforts, B.; Heyndrickx, W.; Tsunekawa, A.; Poesen, J. Predicting gully densities at sub-continental scales: A case study for the Horn of Africa. *Earth Surf. Process. Landf.* **2020**, *45*, 3763–3779. [CrossRef]
38. County Government of West Pokot. County Integrated Development Plan (2018–2022). 256p. Available online: https://www.devolution.go.ke/wp-content/uploads/2020/02/Westpokot-CIDP-2018-2022.pdf (accessed on 25 August 2021).
39. Reith, J.; Ghazaryan, G.; Muthoni, F.; Dubovyk, O. Assessment of land degradation in semiarid Tanzania—Using multiscale remote sensing datasets to support sustainable development goal 15.3. *Remote Sens.* **2021**, *13*, 1754. [CrossRef]
40. Wairore, J.N.; Mureithi, S.M.; Wasonga, O.V.; Nyberg, G. Characterization of enclosure management regimes and factors influencing their choice among agro-pastoralists in north-western Kenya. *Pastor. Res. Policy Pract.* **2015**, *5*, 14. [CrossRef]
41. Wairore, J.N.; Mureithi, S.M.; Wasonga, O.V.; Nyberg, G. Benefits derived from rehabilitating a degraded semi-arid rangeland in private enclosures in West Pokot County, Kenya. *Land Degrad. Dev.* **2015**, *27*, 532–541. [CrossRef]
42. Touber, L. Landforms and Soils of West Pokot District, Kenya: A Site Evaluation for Range-Land Use. Wageningen (The Netherlands), The Winand Staring Centre for Integrated Land, Soil and Water Research. Report No. 50. Available online: https://esdac.jrc.ec.europa.eu/content/landforms-and-soils-west-pokot-district-site-evaluation-rangeland-use-report-no-p50 (accessed on 25 August 2021).
43. Arukulem, Y.E.; Makindi, S.M.; Obwoyere, G.O. Climate variability and the associated impacts on smallholder agriculture in Senetwo location, Kenya. *Int. J. Sci. Res.* **2015**, *4*, 845–850.
44. Wilson, J.P.; Gallant, J.C. *Terrain Analysis: Principles and Applications*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2000.
45. Campbell, J.B. *Introduction to Remote Sensing*, 3rd ed.; Taylor & Francis: London, UK, 2002.
46. Lillesand, T.M.; Kiefer, R.W.; Chipman, J.W. *Remote Sensing and Image Interpretation*, 6th ed.; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2008.
47. Menard, S. *Applied Logistic Regression Analysis, Quantitative Applications in the Social Sciences*; No. 106; Sage: London, UK, 2002.
48. Montgomery, D.C.; Peck, E.A.; Vining, G.G. *Introduction to Linear Regression Analysis*; Wiley: Hoboken, NJ, USA, 2006.
49. Agresti, A. *An Introduction to Categorical Data Analysis*; Wiley: Hoboken, NJ, USA, 2007.
50. Cutler, A.; Cutler, D.R.; Stevens, J.R. Random Forests. In *Ensemble Machine Learning: Methods and Applications*; Zhang, C., Ma, Y., Eds.; Springer: New York, NY, USA, 2012; pp. 157–175.
51. Taalab, K.; Cheng, T.; Zhang, Y. Mapping landslide susceptibility and types using random forest. *Big Earth Data* **2018**, *2*, 159–178. [CrossRef]
52. Arabameri, A.; Nalivan, O.A.; Saha, S.; Roy, J.; Pradhan, B.; Tiefenbacher, J.P.; Ngo, P.T.T. Novel ensemble approaches of machine learning techniques in modeling the gully erosion susceptibility. *Remote Sens.* **2020**, *12*, 1890. [CrossRef]
53. Hong, H.; Pradhan, B.; Jebur, M.N.; Tien Bui, D.; Xu, C.; Akgun, A. Spatial prediction of landslide hazard at the Luxi area (China) using support vector machines. *Environ. Earth Sci.* **2016**, *75*, 40. [CrossRef]
54. Tien Bui, D.; Pradhan, B.; Lofman, O.; Revhaug, I. Landslide susceptibility assessment in Vietnam using support vector machine, decision tree and Naïve Bayes models. *Math. Prob. Eng.* **2012**, 974638. [CrossRef]
55. Pradhan, B. A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS. *Comput. Geosci.* **2013**, *51*, 350–365. [CrossRef]
56. Kavzoglu, T.; Colkesen, I. A kernel function analysis for support vector machines for land cover classification. *Int. J. Appl. Earth Observ. Geoinfomat.* **2009**, *11*, 352–359. [CrossRef]
57. Elith, J.; Leathwick, J.R.; Hastie, T.A. Working guide to boosted regression trees. *J. Anim. Ecol.* **2008**, *77*, 802–813. [CrossRef]
58. Ließ, M.; Schmidt, J.; Glaser, B. Improving the spatial prediction of soil organic carbon stocks in a complex tropical mountain landscape by methodological specifications in machine learning approaches. *PLoS ONE* **2016**, *11*, e0153673. [CrossRef] [PubMed]
59. Yang, R.; Zhang, G.; Liu, F.; Lu, Y.; Yang, F.; Yang, F.; Yang, M.; Zhao, Y.; Li, D. Comparison of boosted regression tree and random forest models for mapping topsoil organic carbon concentration in an alpine ecosystem. *Ecol. Indic.* **2016**, *60*, 870–878. [CrossRef]
60. Pontius, R.G.; Schneider, L.C. Land cover change model validation by an ROC method for the Ipswich watershed, Massachusetts, USA. *Agric. Ecosyst. Environ.* **2001**, *85*, 239–248. [CrossRef]
61. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. [CrossRef]
62. Metz, C.E. Basic principles of ROC analysis. *Seminars in Nuclear Science III* **1978**, *8*, 283–298. [CrossRef] [PubMed]
63. Hosmer, D.W.; Lemeshow, S. *Applied Logistic Regression*; Wiley Series in Probability and Statistics; Wiley: Hoboken, NJ, USA, 2000. [CrossRef]
64. R Development Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2020.

65. Chuma, G.B.; Mugumaarhahama, Y.; Mond, J.M.; Bagula, E.M.; Ndeko, A.B.; Lucungu, P.B.; Karume, 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* **2023**, *129*, 103295. [CrossRef]
66. Arabameri, A.; Chen, W.; Loche, M.; Zhao, X.; Li, Y.; Lombardo, L.; Cerda, A.; Pradhan, B.; Tien Bui, D. Comparison of machine learning models for gully erosion susceptibility mapping. *Geosci. Front.* **2020**, *11*, 1609–1620. [CrossRef]
67. Arabameri, A.; Blaschke, T.; Pradhan, B.; Pourghasemi, H.R.; Tiefenbacher, J.P.; Tien Bui, D. Evaluation of recent advanced soft computing techniques for gully erosion susceptibility mapping: A comparative study. *Sensors* **2020**, *20*, 335. [CrossRef]
68. Amare, S.; Langendoen, E.; Keesstra, S.; van der Ploeg, M.; Gelagay, H.; Lemma, H.; van der Zee, S.E.A.T.M. Susceptibility to gully erosion: Applying random forest (RF) and frequency ratio (FR) approaches to a small catchment in Ethiopia. *Water* **2021**, *13*, 216. [CrossRef]
69. Bouramtane, T.; Hilal, H.; Rezende-Filho, A.T.; Bouramtane, K.; Barbiero, L.; Abraham, S.; Valles, V.; Kacimi, I.; Sanhaji, H.; Torres-Rondon, L.; et al. Mapping gully erosion variability and susceptibility using remote sensing, multivariate statistical analysis, and machine learning in South Mato Grosso, Brazil. *Geosciences* **2022**, *12*, 235. [CrossRef]
70. Garosi, Y.; Sheklabadi, M.; Conoscenti, C.; Pourghasemi, H.R.; van Oostef, K. Assessing the performance of GIS-based machine learning models with different accuracy measures for determining susceptibility to gully erosion. *Sci. Total Environ.* **2019**, *664*, 1117–1132. [CrossRef] [PubMed]
71. Gayen, A.; Pourghasemi, H.R.; Saha, S.; Keesstra, S.; Bai, S. Gully erosion susceptibility assessment and management of hazard-prone areas in India using different machine learning algorithms. *Sci. Total Environ.* **2019**, *668*, 124–138. [CrossRef]
72. Hembram, T.K.; Saha, S.; Pradhan, B.; Maulud, K.N.A.; Alamri, A.M. Robustness analysis of machine learning classifiers in predicting spatial gully erosion susceptibility with altered training samples. *Geomat. Nat. Hazards Risk* **2021**, *12*, 794–828. [CrossRef]
73. Amiri, M.; Pourghasemi, H.R.; Ghanbarian, G.A.; Afzali, S.F. Assessment of the importance of gully erosion effective factors using Boruta algorithm and its spatial modeling and mapping using three machine learning algorithms. *Geoderma* **2019**, *340*, 55–69. [CrossRef]
74. Rahmati, O.; Tahmasebipour, N.; Hamid, A.H.; 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]
75. Wang, F.; Sahana, M.; Pahlevanzadeh, B.; Pal, S.C.; Shit, P.K.; Piran, M.J.; Janizadeh, S.; Band, S.S.; Mosavi, A. Applying different resampling strategies in machine learning models to predict head-cut gully erosion susceptibility: Predict head-cut gully erosion susceptibility. *Alex. Eng. J.* **2021**, *60*, 5813–5829. [CrossRef]
76. Youssef, A.M.; Pourghasemi, H.R. Landslide susceptibility mapping using machine learning algorithms and comparison of their performance at Abha Basin, Asir Region, Saudi Arabia. *Geosci. Front.* **2021**, *12*, 639–655. [CrossRef]
77. Saha, S.; Roy, J.; Arabameri, A.; Blaschke, T.; Tien Bui, D. Machine learning-based gully erosion susceptibility mapping: A case studyof eastern india. *Sensors* **2020**, *20*, 1313. [CrossRef] [PubMed]