Applying different resampling strategies in machine learning models to predict head-cut gully erosion susceptibility

DOI:
10.1016/j.aej.2021.04.026

Citation for published version (APA):
Wang, F., Sahana, M., Pahlevanzadeh, B., Pal, S. C., Shit, P. K., Piran, M. J., Janizadeh, S., Band, S. S., & Mosavi, A. (2021). Applying different resampling strategies in machine learning models to predict head-cut gully erosion susceptibility: predict head-cut gully erosion susceptibility. Alexandria Engineering Journal, 60(6), 5813-5829. https://doi.org/10.1016/j.aej.2021.04.026

Published in:
Alexandria Engineering Journal

Citing this paper
Please note that where the full-text provided on Manchester Research Explorer is the Author Accepted Manuscript or Proof version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version.

General rights
Copyright and moral rights for the publications made accessible in the Research Explorer are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Takedown policy
If you believe that this document breaches copyright please refer to the University of Manchester’s Takedown Procedures [http://man.ac.uk/04Y6Bo] or contact uml.scholarlycommunications@manchester.ac.uk providing relevant details, so we can investigate your claim.
Applying different resampling strategies in machine learning models to predict head-cut gully erosion susceptibility

Fengjie Wang a, Mehebub Sahana b, Bahareh Pahlevanzadeh c,*, Subodh Chandra Pal d, Pravat Kumar Shit e, Md. Jalil Piran f,*, Saeid Janizadeh g, Shahab S. Band h,*, Amir Mosavi i,j,*

a College of Science, Donghua University, Shanghai 201620, China
b School of Environment, Education & Development, University of Manchester, UK
c Research Department of Design and System Operations, Regional Information Center for Science and Technology (RICeST), Shiraz, Fars, Iran
d Department of Geography, The University of Burdwan, Burdhaman, West Bengal Pin: 713104, India
e Department of Geography and Environment Management, Medinipur, India
f Department of Computer Science and Engineering, Sejong University, Seoul 05006, South Korea
g Department of Watershed Management Engineering and Sciences, Faculty in Natural Resources and Marine Science, Tarbiat Modares University, Tehran 14115-111, Iran
h Future Technology Research Center, College of Future, National Yunlin University of Science and Technology, 123 University Road, Section 3, Douliou, Yunlin 64002, Taiwan, ROC
i School of the Built Environment, Oxford Brookes University, Oxford OX3 0BP, UK
j John von Neumann Faculty of Informatics, Obuda University, 1034 Budapest, Hungary

Received 27 December 2020; revised 25 March 2021; accepted 3 April 2021

KEYWORDS
Head-cut gully erosion; Resampling algorithms; Bootstrap; K-fold cross validation; Machine learning; Boosted regression tree; Support vector machine; Random forest

Abstract Gully erosion is one of the advanced forms of water erosion. Identifying the effective factors and gully erosion predicting is one of the important tools to control and manage such phenomenon. The main purpose of this study is to evaluate the effect of four different resampling algorithms including cross-validation (5-fold and 10-fold) and bootstrapping (Bootstrap and Optimism bootstrap) on boosted regression tree (BRT), support vector machine (SVM), and random forest (RF) models in spatial modeling and evaluation of head-cut gully erosion in Konduran watershed. For this purpose, based on an extensive field survey, the points of the head-cut of the gully erosion were identified first, and a map of the distribution of head-cut gully erosion in the study area was prepared. Then 18 variable identify and prepare as factors affecting the occurrence of head-cut gully erosion. To assess the efficiency of the models, receiver operating characteristics (ROC) and area under the curve (AUC) were used. Through the assessment result we indicate...
The results of the assessment indicated that the use of resampling algorithms increases the efficiency of the models. The integrated optimism-bootstrap-BRT, optimism-bootstrap-SVM, and Optimism-Bootstrap-RF models with AUC 0.85, 0.823 and 0.89 respectively, outperformed the cross-validation 5fold (BRT, SVM, RF), Cross-validation 10fold (BRT, SVM, RF) and Bootstrap (BRT, SVM, RF) integrated algorithms.

© 2021 THE AUTHORS. Published by Elsevier BV on behalf of Faculty of Engineering, Alexandria University. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Gullies are part of the most important processes that result in land degradation. Gullies cause the loss of 10% to 90% of soil erosion [1]. Gullies form the most fragile land systems and deteriorate the physio-chemical properties of the soils. They are hot spots of soil erosion and an enormous amount of soil is eroded from these lands and transported to low land areas and increasing the risk of sedimentation and flooding in a catchment region [2,3].

Gully erosion is a frequent event in the arid and semi-arid climatic region and in a dry-wet climatic situation, leads to a large amount of sediment yield [4]. Iran experiences approximately 1 mm/yr reduction in soil thickness across the whole country due to water erosion, and 75.8% of the geographical area is exposed [5]. The combination of the prolonged dry periods and short wet seasons leads to a large amount of discharge, promoting surface runoff and sedimentation [6]. A large rate of soil erosion through head-cut gullies in agriculture is unsustainable and requires a way out in the form of a management strategy. However, studying gully erosion and predicting head-cut gully erosion is difficult in complex environments [2,7].

Many researchers have handled gully erosion susceptibility mapping (GESM) using RS and GIS techniques. They have used various traditional data mining approaches including weights of evidence (WoE) [5,8–10], conditional analysis (CA) [11,12], logistic regression (LR) [12,13], index or entropy (IOE) [14,15], evidential belief function (EBF) [9], certainty factor (CF), frequency ratio (FR) [16], analytical hierarchy process (AHP) [14,17]. But Regional GESM techniques needed factors affecting gully erosion data variables from different sources at spatial scales, which may contain ambiguity and uncertainties. And traditional data mining approaches could not establish the relationship between geo-environmental factors and gully erosion processes. So, new modeling techniques are required that move away from traditional data mining techniques, could solve ongoing concerns, and intensify the

![Fig. 1](image_url) Location of the study area in Iran.
model performance and precision to predict head-cut gully susceptibility mapping.

Recently, machine learning (ML) models have become more popular than traditional data mining techniques. ML is one of the artificial intelligence (AI) techniques that use soft computer algorithms to explore and predict information by learning from training datasets [18–20]. ML algorithms that have been used for GESM techniques include boosted regression tree (BRT) [21], random forest (RF) [14,22], classification and regression trees (CART) [14], support vector machine (SVM) [23], artificial neural networks (ANN) [24], maximum entropy (ME) [17,25], linear discriminant analysis (LDA) [26], stochastic gradient tree-boost (SGT), and multivariate adaptive regression splines (MARS) [27], and logistic model tree (LMT) [9].

Previous studies verify that there are more faces that may be investigated and that a large number of potentially valuable approaches have not yet been totally performed to assess gully erosion susceptibility mapping. However, the present study to establish the relationship between gully conditioning factors and head-cut gully occurrence using three ML algorithms, includes BRT, SVM, and RF models and evaluation of the performance of different resampling algorithms in ML models to predict head-cut gully erosion susceptibility.

2. Materials and methods

2.1. Description of the study area

Konduran watershed is located 60 km west of Bandar Lengeh, Hormozgan province, Iran between latitudes 51° 26' and 51° 28' north and longitudes 26° 44' and 26° 51' east, located in a coastal watershed (Fig. 1). The highest point in the northwest of the study area is 418 m above sea level and the lowest point in the southwest is 1 m above sea level. The study area has an area of 75.95 Km². Annual rainfall in this area is 165 mm, the maximum annual rainfall is 392.7 and the minimum is 43 mm. The physiographic type of the region is flood plain. The floodplain area is relatively flat with low elevation and gentle slope, which is caused by fine sedimentary deposition over long periods. The soil is deep to very deep with medium to the light texture (silt, silt loam, and loam sand). The predominant geological formations in the area include Gachsaran formations consisting of gypsum and indrite layers, Mishan formation consisting mainly of marine, calcareous, Aghajari formation outcrops consisting of rock and marl sand and salt domes. The closest village to the gully area is Konduran village, which is 500 m from the gully above. So, the development of gully erosion in the village is of great concern to the villagers (Administration of the natural resource of Hormozgan Province).

2.2. Methodology

For this purpose, using extensive field surveys, locations of areas sensitive to head-cut gully erosion were collected with the global positioning system (GPS), Garmin (76 CSX Garmin), and entered into the ArcGIS software. The distribution of head-cut gully erosion was prepared with a total of 103 head-cut gullies in the case study and 103 points were randomly selected as non-head-cut gully points. The data was divided into two stages: training and validation, so that 70% of the data were randomly selected as training data and the other 30% were handled for validation [16,23].

In the present study, for spatial modeling of head-cut gully erosion susceptibility, the effect of different resampling algorithms on different methods of ML algorithms were evaluated. The flowchart of the research steps is shown in Fig. 2.

Fig. 2 Methodological flow chart.
Fig. 3  Head-cuts Gully erosion conditioning factors: (a) altitude, (b) slope, (c) aspect, (d) distance from river, (e) drainage density, (f) plan curvature, (g) profile curvature, (h) distance from road, (i) NDVI, (j) land use, (k) lithology, (l) soil, (m) clay, (n) silt, (o) sand, (p) TPI, (q) SPI, (r) TWI.
2.3. Dataset preparation

The choice of the type of factors affecting the occurrence of natural hazard phenomena varies in order to the different features of the region. Accordingly, researchers have considered a set of different topographic, hydrological, and lithological factors in the research of gully erosion zoning [23,28,29]. Therefore, in the present study, 18 variables including altitude, slope, aspect, distance from rivers, drainage density, plan curvature, profile curvature, distance from roads, normalized Difference vegetation index (NDVI), land use, lithology, soil type, clay, silt, sand, Topographic position index (TPI), stream...
power index (SPI) and Topographic wetness index (TWI) were considered as effective variables in the incidence of head-cut gully erosion.

The digital elevation model with a resolution of 12.5°×12.5 m was obtained from the phased array type L-band synthetic aperture radar (PALSAR) (https://search.asf.alaska.edu/#/) and morphometric factors such as altitude, slope, aspect, curvature were provided in ArcGIS 10.5.

Topographic characteristics such as altitude play an important role mainly in the expansion and control of gully erosion, by indirectly affecting the characteristics of vegetation and precipitation [14,29]. According to the altitude map of Konduran watershed, the minimum altitude is 1 and the maximum is 418 m above sea level (Fig. 3a). Low-slope areas have a high potential for surface runoff concentration and exposure to the process of gully erosion [5,12]. In this research, a slope map was prepared from DEM (Fig. 3b). The slope aspect can indirectly affect the erosion processes by controlling sun exposure, vegetation type, soil moisture content [15,30]. The aspect variable was provided in ArcGIS 10.5 based on the DEM and was classified into 9 categories (Fig. 3c). In general, the role of the plan and profile curvature on the occurrence of gully erosion is in the divergence or convergence of water during the descent [22,28]. Therefore, the plan and profile curvature layers are selected according to their effect on the creation and emancipation of gully erosion. Using ArcGIS 10.5, plan and profile curvature maps were prepared (Fig. 3d and Fig. 3e). Drainage density reflects the geological nature, soil characteristics, and vegetation conditions of each region and, by affecting the amount of runoff, causes and spreads erosion [5,29]. In order to prepare a drainage density map, line density extension in ArcGIS 10.5 was used (Fig. 3f). In many areas, the stream network is effective in developing gully erosion, and areas near the river are more susceptible to erosion. In this study, the distance from rivers layer was prepared based on the Euclidean distance extension in ArcGIS 10.5 software [9,22] (Fig. 3g). Roads create intermittent penetration levels and soil disturbance, interrupting and concentrating the surface flow and discharging it in the downstream slopes, and increasing the process of gully erosion [31]. Distance from road map was prepared based on the Euclidean distance extension in ArcGIS 10.5 software (Fig. 3h). Land-use is one of the important factors in the appearance of gully phenomena due to the impact on land use and vegetation. In general, regions with low vegetation are sensitive to erosion than regions with high vegetation because vegetation causes soil resistance to erosion [5,16]. Clay, silt and sand percentage maps were prepared from the soil grid site with a pixel size of 250°×250 m (Fig. 3i, Fig. 3m, and Fig. 3n).

The NDVI shows the vegetation status in the area. Considering that the vegetation status in hydrological and soil studies is of great importance, in this study, NDVI was used. NDVI varies between −1 and +1, and the value of −1 indicates areas without vegetation coverage and +1 shows areas with dense vegetation coverage [33,34]. In order to prepare the NDVI map, bands 5 (near-infrared) and 4 (red) from OLI imagery Landsat satellite taken on 21/6/2019 were used (Fig. 4o).

Due to the effect of TPI, SPI, and TWI on morphometric and hydrological characteristics [16,35,36], TPI, SPI, and TWI have been considered as important variables in causing gully erosion. SPI, TPI, and TWI factors were provided in SAGA GIS 2.6 (Fig. 3p, Fig. 3q, and Fig. 3s).

### 2.4. Multi-collinearity

Multi-collinearity is a condition that shows that an independent factor is a linear function of other independent factors [37]. If linearity is high, it means that there is a high correlation between factors and it can reduce the accuracy of modeling [38]. In order to survey multi-collinearity, the variance inflation factor (VIF) is used. Variables with VIF less than 5 are considered as variables with low and suitable linearity for modeling [25,39].

Tolerance is a ratio of the relative scatter to the scatter of that variable, which is not justified by the linear relationships of that variable with other independent variables in the model [37]. Given that tolerance is a ratio, its value varies between zero and one. A value close to one means that in an independent variable a small part of its scattering is justified by other independent variables. A value close to zero means that a variable is almost a linear combination of other independent variables, and the data has multiple common linear relationships [40].

| Table 1 | Evaluate multi-collinearity in independent variables. |
|---------|--------------------------------------------------------|
| Variables | VIF    | Tolerance |
| Altitude  | 4.07   | 0.24      |
| Slope     | 1.96   | 0.51      |
| Aspect    | 1.12   | 0.89      |
| Distance from river | 3.33 | 0.3     |
| Drainage density | 2.36 | 0.42     |
| Plan curvature | 2.09 | 0.48     |
| Profile curvature | 1.89 | 0.53     |
| Distance from road | 1.96 | 0.51     |
| NDVI      | 2.61   | 0.38      |
| Land use  | 2.43   | 0.41      |
| Lithology | 1.43   | 0.69      |
| Soil      | 2.88   | 0.35      |
| Clay      | 4.85   | 0.21      |
| Silt      | 3.06   | 0.33      |
| Sand      | 4.46   | 0.22      |
| TPI       | 1.65   | 0.6       |
| SPI       | 1.22   | 0.81      |
| TWI       | 1.41   | 0.71      |
2.5. ML algorithms

ML approaches are used to analysis that and build a full automated analytical models. Such approaches learn from data, identify patterns, and accordingly make rational decisions [54]. Supervised learning (SL) is one category of ML approaches, the goal is to classify the input data into some classes based on some pre-defined labels according to their features. The main tasks in SL are classification, regression, etc. The most well-known approaches in SL are linear regression, random forest, decision tree, SVM, and K-nearest neighbors (KNN).

2.5.1. Boosted regression tree (BRT)

The boosted regression tree (BRT) model schemed as a ML algorithms based on classification and regression trees along with the Boosting algorithm [4,41]. This model, unlike algorithms that predict the average, uses a stage-wise progressive method (analysis and matching between variables and data of the training group) and more non-parametric statistics. This model can be used to predict quantitative (regression tree) or classification outcomes (classification tree) [41]. In order to implement this model, rpart package in R software was used.

2.5.2. Support vector Machine (SVM)

The SVM is used for data classification and regression analysis. In other words, after identifying the input data of the model and the target data, the SVM model divides the data into separate groups after analysis between the independent and dependent variables [42]. SVM is an algorithm that finds a specific type of linear model that produces the maximum margin of a hyperplane. Maximizing the hyperplane margin

Fig. 5  Select best interaction depth in BRT model based on four resampling algorithms a) 5 fold CV, b) 10 fold CV, c) Bootstrap and d) OB.
leads to maximizing segregation between classes. The closest training points to the maximum of the hyperplane margin are called support vectors, and these functions are used to obtain the boundary between classes. SVM could be thought of as an algorithm in which two classes are hyperplane by a separating superscript that is defined on the training dataset [43]. Choosing the optimal kernel type is an important step in the SVM model, as it can greatly increase the accuracy of the model. The kernel functions used in SVM are usually divided into four groups: linear kernels, polynomial kernels, radial base kernels, and circular kernels [42]. In this study, e1071 package was used in R software for modeling based on SVM with a radial kernel function.

2.5.3. Random forest (RF)

RF is an effective tool in predicting target variables or pattern classification is the decision tree. RF is an ensemble method that mixes multi algorithms to produce a repeated prediction of each phenomenon [44]. In general, the single decision tree is prone to over-fitting and has little generalizability. When a decision tree is formed, small changes in learning patterns can cause fundamental changes in the structure of that tree. RF could learn complex types and remark nonlinear relationships between independent factors and response factors. It could also constitute various sorts of data due to the lack of normal distribution of the data used [44]. This method is a combination of several decision trees, in the structure of which, several samples of Bootstrap are extracted from the data and a number of input variables are randomly involved in the structure of each tree. In the sampling stage, near 1-third of data is not sampled, and as an example out of the bag (OOB), this data is used to determine important variables as well as to estimate the error. Then a tree is spread on each sample of Bootstrap. In the stage of structuring a tree in each branch, from all independent M variables, the m variable is randomly selected for the division. After making the whole tree, the validation
data is introduced to the tree and the number of trees for the input vector of output is obtained [45]. In this study, random forest model calculations were performed in R software and randomForest package.

2.6. Resampling

Resampling techniques are a manner of frequently carrying samples from a data set and refitting an assigned model on each sample with the purpose of learning more about the adapted model. Resampling techniques could be time-consuming and costly, as they often need the same statistical methods to be performed under different data sets [46]. To receive extra information about the fitted model, return the resampling techniques of the model of attention to the samples created from the training set. There are various resampling algorithms such as cross-validation (CV), leave-one-out cross-validation (LOOCV), K-fold cross-validation and Bootstrap [46,47].

2.6.1. K-fold cross validation (K-Fold CV)

K-fold CV functions are based on randomly dividing the set of observations into K sections or folds of approximately similar size. Any of the K folds is handled as the validation set while the other K - 1 folds are applied as the validation set to create K calculates the validation error [48]. The cross-validation process, which is repeated K times, is then used exactly once as validation data with each of the K samples. The results K can then be averaged to produce a single estimate. The advantage of this method over repetitive random sampling is that all observations are used for both training and validation, and each observation is used accurately for validation. The K-fold CV evaluated validation error occurs from the mean of these calculations. Standard values for K are 5 or 10 since

---

**Fig. 7** Select best number of m try in RF model based on four resampling algorithms a) 5-fold CV, b) 10-fold CV, c) Bootstrap and d) OB.
these values demand less calculation than when K is equal to n [48].

When we have a small size of observations, dividing it into training and validation sets may lead to a very small test set [49]. By chance, we can get almost any performance in this set. This is even worse when we are faced with a multi-class problem. If we use cross-validation in this case, we create different K models, so we can predict all our data. For each instance, we predict with a model that does not see this instance. When we create different models using our K learning algorithm and test it on different sets of K tests, we can have more confidence in the performance of our algorithm. When we do a single evaluation on our test suite, we get only one result.

2.6.2. Bootstrap

The bootstrap is a generally appropriate algorithm that could be handled to quantify the uncertainty affiliated with an addressed estimator or statistical learning method, such as those for which it is tough to achieve a size of variability [50]. The bootstrap produces different data sets by frequently sampling observations from the fundamental data set. These created data sets could be applied to determine variability in lieu of sampling independent data sets from the entire group. The bootstrap produces different data sets by frequently sampling observations from the fundamental data set. These created data sets could be applied to determine variability in lieu of sampling independent data sets from the entire group. The sampling hired by the bootstrap implicates randomly choosing n observations with substitution, which intends some observations could be elected duplicated times while other observations are not involved at all [50]. This procedure is iterated N times to generate N bootstrap data sets, M*1, M*2, ..., M * N, which could be handled to determine other quantities include standard error (Fig. 4).

2.7. Validation models

In this research, the parameters of the receiver operating characteristics (ROC) were applied to the assessment efficiency of the models. The parameters in this curve include sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV), and area under the curve (AUC) [16,26].
AUC is equal to the probability of distinguishing between dependent variables and independent variables by a model. Different values of the AUC are between 0.5 and 1. If AUC is 0.5, it indicates that the model is random, and if this value is equal to 1, the model can best distinguish between occurrence and non-occurrence points. AUC of 0.7 to 0.8 indicates a good efficiency model, between 0.8 and 0.9 excellent, and more than 0.9 indicates the excellent detection power of the model [51].

\[
TPR = \frac{TP}{TP + FN},
\]

(1)

\[
TNR = \frac{TN}{TN + FP},
\]

(2)

\[
PPV = \frac{\text{Number of positives}}{\text{Number of positives} + \text{Number of false positives}},
\]

(3)

\[
NPV = \frac{\text{Number of true negatives}}{\text{Number of true negatives} + \text{Number of false negatives}},
\]

(4)

\[
AUC = \frac{\sum TP + \sum TN}{P + N},
\]

(5)

where TPR is sensitivity, TNR is Specificity, TP is indicated true positive of gully occurrence, and TN indicated True negative of gully non-occurrence.

3. Results

3.1. Multi-collinearity analysis

Multi-collinearity data can interfere with the accuracy and interpretation of findings where every linear statistical procedure has been used. Multi-collinearity is described as a linear relationship in a dataset linking two or more independent variables [23,35]. The TOL and VIF values are either < 0.1 or ≥ 10, suggesting strong multi-collinearity [52,53]. The multi-collinearity test of 18 Geo-environmental gully headcut erosion susceptibility parameters showed that the highest and lowest values of variance inflation factor (VIF) are 4.07 and 1.12 whereas the highest and lowest tolerance values are 0.89 and 0.21 respectively (Table 1). The results of multi-collinearity showed that all independent variables have VIF < 5, these variables are independent and there is not a strong correlation between them, therefore, all 18 variables were selected for modeling HGES modeling.

3.2. Determine best parameters

The performance and selection of the best resampling method from tune parameters for the BRT, SVM, and RF ML models are shown in Fig. 5, Fig. 6, and Fig. 7 respectively.

3.3. Assessment efficiency of models

The accuracy of the maps developed by OB-BRT, OB-SVM, OB-RF models has been checked by the ROC curves. The validation of the outcomes clearly showed that the Optimism Bootstrap BRT, Optimism Bootstrap SVM, and Optimism
Bootstrap RF models with ROC values of 0.85, 0.823, and 0.89 respectively had exceptional levels of precision than rest of the ensemble and standalone ML models (Fig. 8 and Fig. 9). Results obtained using the positions of 103 recorded gully and non-gully points, were validated. The graphical outcomes of the suitable output tests revealed that the ensemble ML models (Optimism Bootstrap BRT, Optimism Bootstrap SVM, and Optimism Bootstrap RF) had a decent performance compared to others (Fig. 10).

3.4. Head-cut gully erosion susceptibility modeling

ML models such as BRT, SVM, RF along with four resampling methods K-fold (5 and 10) CV and Bootstrapping (Bootstrap (B) and Optimism Bootstrap (OB)) were applied to predicting the gully erosion areas. Resampling techniques have been utilized since resampling is a methodology that will support all the data used. After determining the accuracy of ensemble models based on resampling algorithms, the best ensemble models of each group were selected to predict head-cut gully erosion maps (Fig. 11).

The map of gully head-cut susceptibility using OB-BRT model (Fig. 11) shows the highest level of precision in comparison with the standalone BRT, 5-fold CV BRT, 10-fold CV BRT, and B-BRT respectively. According to the gully head-cut susceptibility map of OB-BRT model, the maximum part of the region is engaged by very low (20.03%) to low (28.66%) classes, while high (17.55%), moderate (25.60%), and very high (8.16%) classes covered other parts of the region respectively (Table 2).

The map of the susceptibility to gully head-cut using the SVM Optimism Bootstrap model (Fig. 11) indicates the high level of accuracy compared to the standalone SVM, 5-fold CV SVM, and B-SVM respectively. According to the gully head-cut susceptibility map of OB-SVM model, the maximum part of the region is engaged by very low (14.08%) to low (28.73%) classes, while high (17.70%), moderate (29.80%) and very high (9.70%) susceptibility classes occupied rest of the studied region respectively (Table 2).

The Optimism Bootstrap RF Model (Fig. 11) map displays the maximum degree of accuracy compared to the rest of the standalone and ensemble models such as 5-fold CV RF, 10-
Fig. 11  Gully head-cut erosion susceptibility using the best resampling algorithms in a) BART, b) SVM, and c) RF models.
fold CV RF, and B-RF. According to the gully head-cut susceptibility map of OB-RF model, the maximum parts of the region are engaged by very low (36.68%) to low (21.80%) classes, while high (14.14%), moderate (18.64%) and very high (8.73%) susceptibility classes covered rest of the studied region respectively (Table 2).

**Table 2** Gully head-cut erosion susceptibility classes area.

| Susceptibility class | Models         | OB-BRT | OB-SVM | OB-RF |
|----------------------|----------------|--------|--------|-------|
|                      | Km²            | %      | Km²    | %     | Km²   | %     |
| Very low             | 15.21          | 20.03  | 10.69  | 14.08 | 27.86 | 36.68 |
| Low                  | 21.77          | 28.66  | 21.82  | 28.73 | 16.56 | 21.80 |
| Moderate             | 19.44          | 25.60  | 22.63  | 29.80 | 14.16 | 18.64 |
| High                 | 13.33          | 17.55  | 13.44  | 17.70 | 10.74 | 14.14 |
| Very high            | 6.2            | 8.16   | 7.33   | 9.70  | 6.63  | 8.73  |

**Fig. 12** Result of importance value based on OB-RF model.

**Fig. 13** Partial plot for importance variable a) altitude, distance from road, b) distance from river and NDVI.
3.5. Importance value

The results indicate that altitude, distance from the river, distance from the road, land use, NDVI, and percentage of clay are of maximum importance, while the rest of the other variables, such as TWI, SPI, TPI, aspect, lithology, and slope are of less importance (Fig. 12). In this region, the causes of gully head-cut formation are closely linked to land use, NDVI, and distance from the road, with the latter being the most influential variable. These agents also play a crucial function in aggravating the degradation of gullies. Most gully head-cuts' shapes manifest the impact of land use activities induced surface runoff on the progress and expansion of this kind of gully head-cut development. The partial plots for important variables are also showing the same importance level (Fig. 13).

4. Discussion

In recent years, with the advancement of computer science and modeling, the use of these methods to solve various problems, including natural disasters, has increased. Therefore, comparing the performance of different models and identifying the best model is one of the necessities in these studies. Although comparing the performance of different models with common criteria used in various studies, the results may be misleading. That is why it is very important to find the best method based on their abilities [51]. Therefore, in this study, a field-based methodology was explored by implementing three ML models and four resampling procedures for the development of a gully head-cut susceptibility map, taking into account subsurface piping erosion. While modeling gully head-cut susceptibility, it is necessary to calculate the projecting capacity and multicollinearity of the eighteen selected variables and therefore, the importance values of each conditioning variable have been calculated [35]. Furthermore, few studies have focused on gully head-cut erosion to establish and project the correlation between gully head-cuts and their absence, considering a number of individual variables using models [6]. As a result, in the present study, we have calculated the importance of each parameter behind the appearance of gully head-cut positions in Konduran watershed. The findings are associated with the statement that the intensity of gully head-cut erosion depends on certain variables such as the volume of the runoff, altitude, distance from the road, distance from the river, percentage of clay, NDVI, and land use.

A large number of ML models are available, but efficient and advanced methods for spatial modeling of the occurrences of gully head-cut erosion susceptibility are needed. Here, ML models such as BRT, SVM, RF as well as resampling approaches were applied to evaluate the performance of ensemble ML models by comparing the performance of individual ML models (BRT, SVM and RF) to predict the precise gully head-cut erosion susceptibility map and gully head-cut locations based on eighteen selected significant parameters. Methods of resampling have been utilized as resampling is a method that can help other applied analysts and solve the unbalancing in response variables [51]. While they do not solve all inferential issues, unlike resampling analytical sampling distributions, CV is still used; it’s used to ensure the adequacy of a statistical model. The explanation that cross-validation and bootstrap are likely to be less reliable is that they have smaller efficient goodness-of-fit tests relative to the Optimism Bootstrap which is always the chosen process. Evaluation of modeling performance confirms that the best method is a set of models (OB-BRT, OB-SVM, and OB-RF models) with exceptional precision at 85%, 82.3%, and 89%, respectively, in the prediction of head-cut locations compared to the rest of the ensemble and standalone ML methods. In general, the Bootstrap BRT, Bootstrap SVM and Bootstrap RF set enhances the generalization of base predictors for gully head-cut locations to find the Optimism Bootstrap set of combined models with improved accuracy. In other fields and different ML models also the ability of Optimism Bootstrap resampling method have showed [51,52].

A larger number of gully head-cut occurrences were observed in and around the residential area of the region than the other categories of land use, which reveals the obvious effect of land use on gully head-cut. Inadequate management practices and land use reforms have been noted to play a significant impact in gully head cutting, as subsurface piping and gully head-cut encounters have contributed to gully formation and increase. Our findings are consistent with those to which it is inferred that gully head-cuts are mainly generated by runoff, land use, percentage of clay, and altitude [25]. In addition, the resampling-based ensemble ML models (OB-BRT, OB-SVM and OB-RF models) have identified these gully head-cut places more accurately on a regional scale than other tested ML ensembles and standalone models.

Gully head-cut and subsurface piping erosion are one of the most significant causes of soil erosion in the research region and are the major sources of sediment supply in the lower area. Such adverse effects have created huge damages to the economy in the region. Therefore, appropriate management measures should be implemented keeping in view, the health and sustainability of the local environment of this region.

5. Conclusion

In the present study, we applied different resampling algorithms on three ML approaches to investigate the spatial modeling efficiency of gully head-cut erosion susceptibility prediction in the Konduran watershed. Variable importance of explanatory factors showed that altitude, distance from the road, distance from the river, percentage of clay, NDVI and land use, had the most significant impact on the occurrence of gully head-cut erosion in the Konduran watershed. Validation of the models was rendered on the basis of the Receiver Operating Characteristic (ROC) curve and area under the curve. The results showed that Optimism bootstrap in comparison to other resampling methods, has better performance to increase the accuracy of ML models. Validation of the outcomes clearly showed that the OB-BRT, OB-SVM, and OB-RF models had exceptional levels of precision depending upon selected significant variables. The outcome of this gully head-cut susceptibility map can be useful for land use planning, soil and water conservation, and, as a result, for sustainable development of the region.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.
References

[1] J. Poesen, J. Nachtergaele, G. Verstraeten, C. Valentin, Gully erosion and environmental change: importance and research needs, Catena 50 (2003) 91–133.

[2] P.K. Shit, R. Paira, G. Bhunia, R. Maiti, Modeling of potential gully erosion hazard using geo-spatial technology at Garbhetu block, West Bengal in India, Model. Earth Syst. Environ. 1 (2015) 2.

[3] A. Malik, A. Kumar, Comparison of soft-computing and statistical techniques in simulating daily river flow: a case study in India, 2018, https://doi.org/10.5938/2455-7145.2018.00029.2.

[4] A. Arabameri, A. Cerda, J. Rodrigo-Comino, B. Pradhan, M. Sohrabi, T. Blaschke, D. Tien Bui, Proposing a novel predictive technique for gully erosion susceptibility mapping in arid and semi-arid regions (Iran), Remote Sens. 11 (2019) 2577.

[5] O. Rahmati, A. Haghizadeh, H.R. Pourghasemi, F. Noormohamadi, Gully erosion susceptibility mapping: the role of GIS-based bivariate statistical models and their comparison, Nat. Hazards. 82 (2016) 1231–1258, https://doi.org/10.1007/s11069-016-2239-7.

[6] A. Arabameri, A. Cerda, B. Pradhan, J.P. Tiefenbacher, L. Lombardo, D.T. Bui, A methodological comparison of head-cut based gully erosion susceptibility models: Combined use of statistical and artificial intelligence, Geomorphology 107136 (2020).

[7] A. Bernatek-Jakiel, J. Poesen, Subsurface erosion by soil piping: Significance and research needs, Earth-Sci. Rev. 185 (2018) 1107–1128.

[8] F. Dube, I. Nhapi, A. Murwira, W. Gumindoga, J. Goldin, D. Mashaauri, Potential of weight of evidence modelling for gully erosion hazard assessment in Mbire District – Zimbabwe, J. Phys. Chem. E.A.R. (2014), https://doi.org/10.1016/j.

[9] A. Arabameri, B. Pradhan, K. Rezaei, M. Yamani, H.R. Pourghasemi, L. Lombardo, Spatial modelling of gully erosion using evidential belief function, logistic regression, and a new ensemble of evidential belief function-logistic regression algorithm, L. Degrad. Dev. 29 (2018) 4035–4040.

[10] P.K. Shit, G.S. Bhunia, H.R. Pourghasemi, Gully Erosion Susceptibility Mapping Based on Bayesian Weight of Evidence, in: Gully Eros. Stud. from India Surround. Reg. Reg., Springer, 2020: pp. 133–146.

[11] G. Magliulo, V. Capozzi, R. Ramasco, Seismic performance of R/C frames with overstress discontinuities in elevation, Bull. Earthq. Eng. 10 (2012) 679–694.

[12] C. Conoscenti, S. Angileri, C. Cappadonia, E. Rotigliano, V. Agnesi, M. Märker, Gully erosion susceptibility assessment by means of GIS-based logistic regression: a case of Sicily (Italy), Geomorphology 204 (2014) 399–411, https://doi.org/10.1016/j.geomorph.2013.08.021.

[13] T. Vanwalleghem, M. Van Den Eeckhaut, J. Poesen, G. Govers, J. Deckers, Spatial analysis of factors controlling the presence of closed depressions and gullies under forest: Application of rare event logistic regression, Geomorphology 95 (2008) 504–517.

[14] A. Arabameri, K. Rezaei, H.R. Pourghasemi, S. Lee, M. Yamani, GIS-based gully erosion susceptibility mapping: a comparison among three data-driven models and AHP knowledge-based technique, Environ. Earth Sci. 77 (2018) 1–22, https://doi.org/10.1007/s12665-018-7808-5.

[15] M. Zabihi, F. Mirchooli, A. Motevali, A.K. Darvishan, H.R. Pourghasemi, M.A. Zakeri, F. Sadighi, Spatial modelling of gully erosion in Mazandaran Province, northern Iran, Catena. 161 (2018) 1–13.

[16] A. Azarch, O. Rahmati, E. Rafiei-Sardooi, J.B. Sankey, S. Lee, H. Shahabi, B. Bin Ahmad, Modelling gully-erosion susceptibility in a semi-arid region, Iran: Investigation of applicability of certainty factor and maximum entropy models, Sci. Total Environ. 655 (2019) 684–696.

[17] R. Zakerinejad, M. Maerker, An integrated assessment of soil erosion dynamics with special emphasis on gully erosion in the Mazayyan basin, southwestern Iran, Nat. Hazards. 79 (2019) 25–50, https://doi.org/10.1007/s11069-019-1700-3.

[18] O.A. Arqub, N. Shawagfeh, Application of reproducing kernel algorithm for solving Dirichlet time-fractional diffusion-Gordon types equations in porous media, J. Porous Media. 22 (2019).

[19] O.A. Arqub, Numerical solutions of systems of first-order, two-point BVPs based on the reproducing kernel algorithm, Calcolo. 55 (2018) 1–28.

[20] O. Abu Arqub, Numerical algorithm for the solutions of fractional order systems of Dirichlet function types with comparative analysis, Fundam. Informaticae. 166 (2019) 111–137.

[21] A. Arabameri, B. Pradhan, H.R. Pourghasemi, K. Rezaei, N. Kerle, Spatial modelling of gully erosion using GIS and R programming: A comparison among three data mining algorithms, Appl. Sci. 8 (2018) 1369.

[22] M. Avand, S. Janizadeh, S.A. Naghibi, H.R. Pourghasemi, S. Khoosrozbeigi Bozchaleei, T. Blaschke, A comparative assessment of random forest and k-nearest neighbor classifiers for gully erosion susceptibility mapping, Water 11 (2019) 2076.

[23] O. Rahmati, N. Tahmasebipour, A. Haghizadeh, H.R. Pourghasemi, B. Feizizadeh, Evaluation of different machine learning models for predicting and mapping the susceptibility of gully erosion, Geomorphology 298 (2017) 118–137.

[24] H.R. Pourghasemi, S. Yousefi, A. Kornejady, A. Cerda, Performance assessment of individual and ensemble data-mining techniques for gully erosion modeling, Sci. Total Environ. 609 (2017) 764–775.

[25] H. Pourghasemi, S. Youse, A. Kornejady, A. Cerda, Science of the Total Environment Performance assessment of individual and ensemble data-mining techniques for gully erosion modeling, 609 (2017) 764–775, https://doi.org/10.1016/j.

[26] A. Arabameri, H.R. Pourghasemi, Spatial modeling of gully erosion using linear and quadratic discriminant analyses in GIS and R, in: Spat. Model. GIS R Earth Environ. Sci., Elsevier, 2019: pp. 299–321.

[27] M. Amiri, H.R. Pourghasemi, Mapping and Preparing a Susceptibility Map of Gully Erosion Using the MARS Model, in: Gully Eros. Stud. from India Surround. Reg. Reg., Springer, 2020: pp. 405–413.

[28] V.-H. Nhu, S. Janizadeh, M. Avand, W. Chen, M. Farzin, E. Omidvar, A. Shirzadi, H. Shahabi, J.B. Clage, A. Jaafari, others, Gis-based gully erosion susceptibility mapping: A comparison of computational ensemble data mining models, Appl. Sci. 10 (2020) 2039.

[29] A. Gayen, H.R. Pourghasemi, S. Saha, S. Keesstra, S. Bai, Gully erosion susceptibility assessment and management of hazard-prone areas in India using different machine learning algorithms, Sci. Total Environ. 668 (2019) 124–138.

[30] Y. Garosi, M. Sheklabadi, C. Conoscenti, H.R. Pourghasemi, K. Van Oost, Assessing the performance of GIS-based machine learning models with different accuracy measures for determining susceptibility to gully erosion, Sci. Total Environ. 664 (2019) 1117–1132.

[31] B. Choubin, O. Rahmati, N. Tahmasebipour, B. Feizizadeh, H. R. Pourghasemi, Application of Fuzzy Analytical Network Process Model for Analyzing the Gully Erosion Susceptibility, in: Nat. Hazards GIS-Based Spat. Model. Using Data Min. Tech., Springer, 2019: pp. 105–125.

[32] Á. Gómez-Gutierrez, C. Conoscenti, S.E. Angileri, E. Rotigliano, S. Schnabel, Using topographical attributes to evaluate gully erosion proneness (susceptibility) in two
mediterranean basins: advantages and limitations, Nat. Hazards 79 (2015) 291–314, https://doi.org/10.1007/s11069-015-1703-0.

[33] A. Arabameri, B. Pradhan, K. Rezaei, M. Sohrabi, Z. Kalantari, GIS-based landslide susceptibility mapping using numerical risk factor bivariate model and its ensemble with linear multivariate regression and boosted regression tree algorithms, J. Mt. Sci. 16 (2019) 595–618, https://doi.org/10.1007/s11629-018-5168-y.

[34] A. Frankl, M. Vanmaerecke, J. Nyssen, J. Poesen, Gully prevention and rehabilitation: a review, in: 8th Int. Symp. Gully Eros., 2019: p. 67.

[35] J. Roy, S. Saha, GIS-based gully erosion susceptibility evaluation using frequency ratio, cosine amplitude and logistic regression ensembled with fuzzy logic in Hinglo River Basin, India, Remote Sens. Appl. Soc. Environ. 15 (2019) 100247.

[36] A. Arabameri, S. Saha, J. Roy, W. Chen, T. Blaschke, D. Tien Bui, Landslide susceptibility evaluation and management using different machine learning methods in the Gallicash River Watershed, Iran, Remote Sens. 12 (2020) 475.

[37] W.J. Wilson, TREATING MULTICOLLINEARITY WITH SAS, (n.d.).

[38] M.H. Graham, Confronting multicollinearity in ecological multiple regression, Ecology 84 (2003) 2809–2815.

[39] M. Amir, H.R. Pourghasemi, G.A. Ghanbarian, S.F. Afzali, Assessment of the importance of gully erosion effective factors using Boruta algorithm and its spatial modeling and mapping using three machine learning algorithms, Geoderma 340 (2019) 55–69.

[40] S.S. Kumari, Multicollinearity: Estimation and elimination, J. Contemp. Res. Manag. 3 (2008) 87–95.

[41] S.A. Naghibi, H.R. Pourghasemi, B. Dixon, GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran, Environ. Monit. Assess. 188 (2016) 44.

[42] V. Jakkula, Tutorial on support vector machine (svm), Sch. EECS, Washingt. State Univ. 37 (2006).

[43] H.-C. Kim, S. Pang, H.-M. Je, D. Kim, S.Y. Bang, Constructing support vector machine ensemble, Pattern Recognit. 36 (2003) 2757–2767.

[44] A. Liaw, M. Wiener, others, Classification and regression by randomForest, R News. 2 (2002) 18–22.

[45] A.A. Monillio, Random forests, Lect. Stat. Found. Data Anal, 2009.

[46] P.I. Good, Resampling methods, Springer, 2006.

[47] S.N. Lahiri, Resampling methods for dependent data, Springer Science & Business Media, 2013.

[48] T. Fushiki, Estimation of prediction error by using K-fold cross-validation, Stat. Comput. 21 (2011) 137–146.

[49] R. Kohavi et al, A study of cross-validation and bootstrap for accuracy estimation and model selection, Ijcai (1995) 1137–1145.

[50] B. Efron, R.J. Tibshirani, An introduction to the bootstrap, CRC Press, 1994.

[51] E. Yesilnacar, T. Topal, Landslide susceptibility mapping: a comparison of logistic regression and neural networks methods in a medium scale study, Hendek region (Turkey), Eng. Geol. 79 (2005) 251–266.

[52] J.F. Hair, W.C. Black, B.J. Babin, R.E. Anderson, R.L. Tatham, Multivariate data analysis (Vol. 5, No. 3, pp. 207-219), (1998).

[53] I. Chowdhuri, S.C. Pal, R. Chakrabortty, Flood susceptibility mapping by ensemble evidential belief function and binomial logistic regression model on river basin of eastern India, Adv. Sp. Res. 65 (2020) 1466–1489.

[54] U. Butt, M. Mehmood, S. Shah, R. Amin, M. Shaukat, S. Raza, D.Y. Suh, Md. Jalil Piran, A review of machine learning algorithms for cloud computing security, Electronics 9 (9) (2020) 1379.