Assessing gully erosion susceptibility using topographic derived attributes, multi-criteria decision-making, and machine learning classifiers

Ahmed J. Al-Bawi, Alaa M. Al-Abadi, Biswajeet Pradhan & Abdullah M. Alamri

Published online: 25 Oct 2021.

Submit your article to this journal ↗

Article views: 9

View related articles ↗

View Crossmark data ↗

Taylor & Francis
Taylor & Francis Group

Assessing gully erosion susceptibility using topographic derived attributes, multi-criteria decision-making, and machine learning classifiers

Ahmed J. Al-Bawi[a], Alaa M. Al-Abadi[a], Biswajeet Pradhan[b,c,d] 🆔 and Abdullah M. Alamri[e]

[a]Department of Geology, College of Science, University of Basrah, Basrah, Iraq; [b]The Centre for Advanced Modelling and Geospatial Information Systems, Faculty of Engineering and IT, University of Technology Sydney, NSW, Australia; [c]Department of Energy and Mineral Resources Engineering, Sejong University, Gwangjin-gu, Seoul, Korea; [d]Earth Observation Center, Institute of Climate Change, Universiti Kebangsaan Malaysia, Bangi, Selangor, Malaysia; [e]Department of Geology and Geophysics, College of Science, King Saud University, Riyadh, Saudi Arabia

ABSTRACT

Gully erosion is an erosive process that contributes considerably to the shape of the earth's surface and is a major contributor to land degradation and soil loss. This study applied a methodology for mapping gully erosion susceptibility using only topographic related attributes derived from a medium-resolution digital elevation model (DEM) and a hybrid analytical hierarchy process (AHP) and the technique for an order of preference by similarity to ideal solutions (TOPSIS) and compare the results with naïve Bayes (NB) and support vector machine learning (SVM) algorithms. A transboundary sub-basin in an arid area of southern Iraq was selected as a case study. The performance of the developed models was compared using the receiver operating characteristic curve (ROC). Results showed that the areas under the ROC were 0.933, 0.936, and 0.955 for AHP-TOPSIS, NB, and SVM with radial basis function, respectively, which indicated that the performance of simply derived AHP-TOPSIS model is similar to sophisticated NB and SVM models. Findings indicated that a medium resolution DEM and AHP-TOPSIS are a promising tool for mapping of gully erosion susceptibility.

1. Introduction

Gully erosion is a major source of soil deterioration in many regions of the world. Gully erosion is now generally recognized as a global indication of desertification and land degradation, as it is viewed as a potentially destructive process that poses a threat to life and property. The evaluation of this process and the elements that

CONTACT Biswajeet Pradhan ✉ Biswajeet24@gmail.com; Biswajeet.Pradhan@uts.edu.au; Alaa M. Al-Abadi ✉ alaa.atiaa@uobasrah.edu.iq

Table 1. Topography related attributes and their importance in gully erosion process.

Topographic factor	Importance
Elevation	Effect potential energy, vegetation, and climate (Moore et al. 1991). Elevation has been identified as one of the important factors in gully erosion susceptibility. The gully occurrence and development mechanism is closely related to the different elevation of the terrain (Zabihi et al. 2018)
Slope	Controls velocity of overland flow and runoff generation. It is also considered as one of the important factors in the gully erosion process (Lucà et al. 2011). The gentle slope is expected to be more exposure to gully development as a result of surface runoff (Agnesi et al. 2011)
Aspect	Controls solar insolation, rainfall intensity, soil moisture, evapotranspiration, flora and fauna distribution and abundance and thus has an indirect effect on the gully erosion process (Carrara et al. 1991; Moore et al. 1991; Conforti et al. 2011)
Profile curvature	Plays an important role in controlling flow acceleration, rates of erosion and deposition (Shary et al. 2002; Jiang et al. 2021)
Plan Curvature	Effects the terrain instability because it controls converging/diverging flow, soil water content and soil characteristics (Moore et al. 1991; Lucà et al. 2011; Sujatha & Sridhar 2019)
TWI	Effects the location and size of saturated source areas of runoff generation (Lucà et al. 2011) and thus plays a major role in control gully initiation and development (Arabameri et al. 2019).
SPI	Reflects the erosive power of water flow based on the assumption that flow rate is proportional to specific catchment area (Moore et al. 1991; Pourghasemi et al. 2013)
LS	Is a factor that was used to quantify soil erosion rate using the RUSLE equation and basically reflects the effect of topography on erosion (Renard et al. 1997; Arabameri et al. 2019).

contribute to it are becoming increasingly essential. Gully formation is a complicated natural process that is influenced by a variety of factors, including topography, soil, land use/land cover (LULC), and climate. The growing interest in gully erosion research stems from a desire to better understand its impacts and, as a result, try to mitigate their negative consequences (Al-Abadi and Al-Ali 2018). To control and avoid the negative impacts of gully erosion, the spatial modeling of the gullies, as well as the variables that influence their development, should be assessed. Studying gully erosion susceptibility using conventional methods is a difficult task that requires a lot of money and effort. With the advancement of geographic information systems (GIS), remote sensing, and the evolution of computational and intelligent solution algorithms, geospatial modeling has rendered the process of studying susceptibility to gully erosion simpler. The developed techniques enabled the researchers to study the gully erosion susceptibility at large scale and draw maps to indicate which areas are more vulnerable to erosion to protect them in the future. In this context, statistical bivariate, multivariate, and the machine learning model proved to be very efficient to predict the gully erosion susceptibility (Dube et al. 2014; Angileri et al. 2016; Conoscenti et al. 2017; Al-Abadi and Al-Ali 2018).

In general, the related primary and secondary topographic attributes, lithology, soil properties, land use/land cover, and parameters related to climatic conditions are used to build the gully erosion susceptibility models (Gómez-Gutiérrez et al. 2015). Topography is well known to play the most important role in the development of gullies, as it controls the overland flow's erosive intensity. It is well known that topography plays the most important role in developing gullies, as it determines the

erosive power of the overland flow. The topography is also closely related to other variables that affect gully development, such as lithology, climate, and vegetation cover (Table 1). In addition, the availability of data with sufficient resolution in many cases, especially in transboundary basins, is limited to topography. Therefore, the development of a model for mapping gully erosion with only topographic features is of great importance in gully susceptibility studies (Gómez-Gutiérrez et al. 2015).

One of the disadvantages of statistical and machine learning models is their need for training samples to detect patterns in the input data. Such training samples are sometimes not available, few, or insufficient to train models. Therefore, the use of expert knowledge-based models such as multi-criteria decision-making (MCDM) is preferred. The MCDM is a sub-discipline of operations research that explicitly evaluates multiple conflicting criteria in decision-making problems. The MCDM is a prominent decision-making method as it strengthens decision-makers' ability to take the decision by simultaneously evaluating all the criteria and objectives (Al-Abadi et al. 2018). Despite their use in many scientific, engineering, social and economic fields, the use of MCDM in mapping gully erosion is still limited (Kheir et al. 2007).

The northeast Maysan Governorate, bordering Iran, is often heavily affected by water erosion that caused many issues including the losses of rich soil that sustains the agricultural activities, the issuance of high costs, reduced agricultural potential, and migration of people in the villages of this region. Although there are many solutions to mitigate the effects of this erosion and loss of soils, they are not relevant in all areas vulnerable to erosion due to limited human and financial resources. Therefore, identification of the area most vulnerable to erosion is often given a top priority.

The aim of this research is to investigate the possibility of mapping gully erosion susceptibility with adequate precision using only topographical attributes derived from a digital elevation model, knowledge-based MCDM, and Machine learning models in the framework of GIS platform. Although this study used the same modeling techniques (MCDM and machine learning models) that have been used before in many published studies (Arabameri et al. 2019), the novelty of this manuscript is not in the using of these models in the gully erosion susceptibility but the possibility of using only the topographical related factors derived from the medium resolution DEM for accurate modeling of gully erosion susceptibility and compare the outcomes of MCDM with advanced machine learning classifiers. The second difference between this study and other similar studies is the process of representing gullies in geographic information system (GIS), where most of the previous studies use the "point feature" to represent individual gully and create "gully inventory map" regardless of the size and length of these gullies. Due to the fact that most of the recognized gullies in the study area are longitudinal and extend to long distances, the area effected by gullies is represented in a more realistic way by choosing a large number of points to create the "inventory map" and this representation approach is used for the first time in gully erosion susceptibility as many previous studies represent single gully by a point regardless of the size of this gully and with other gullies the "gully inventory map" is created.

To attain this objective, an MCDM model was developed through the hybridization of analytical hierarchy process and the technique for an order of preference by similarity to ideal solutions (AHP-TOPSIS) to map gully erosion susceptibility using only topographic factors. The AHP-TOPSIS model's reliability was compared with two highly efficient machine-learning classifiers: support vector machine (SVM) and Naïve Bayes (NB) to demonstrate the efficacy of APH-TOPSIS in mapping gully erosion susceptibility. We selected SVM and NB here, because these techniques are master machine learning algorithms and known to be accurate and easy to implement, especially when the training data is sufficient. The major advantages of these techniques include easy to design and build, do not require iterative parameter estimation, robust to noise data, can handle different types of data (numeric, nominal, category, etc.), have good generalization capabilities which prevent it from over-fitting, and relatively memory efficient (Fiesler & Beale 2020).

The results of this study contribute to providing useful insights into ways to protect soil and establish land-use projects, which contribute to avoiding soil loss and ensuring agricultural production.

2. Material and methods

2.1. Study area

A sub-watershed situated in the northeastern part of the Maysan Governorate, Iraq (Figure 1) was regarded as the area of case study. The sub-watershed is 205 km^2 in area and has elevation ranges varying from 8 to 272 m. In the territories of Iraq, the largest part of the study area is situated, while a small portion is expanded to the land of Iran. Most of the study region is flat and without features except small hills on the eastern border between Iraq and Iran. The climate of the area is dry and hot in summer and relatively cold and wet during winter. Precipitation (most of them are rainfall) dominates during winter, spring, and autumn. Temperature over 50 °C commonly occurs in summer, whereas the temperature below 0 °C is rarely recorded. The Quaternary sediments cover the whole of the study area (Figure 2a). The main lithological units are Quaternary alluvium and alluvial fan deposits. These sediments comprise mainly gravel, sand, and silty sand. In terms of the tectonic framework, the majority of considered sub-watershed is situated in the Mesopotamian zone (Tigris subzone). The Mesopotamia zone is situated on the eastern side of the stable shelf, and the Hamrin-Makhul area is restricted to the east (Jassim and Goff 2006). Two types of soils found in the basin: aridisols and inceptisols (Figure 2b) which cover 163 km^2 (79%) and 17 km^2 (17%) area, respectively excluding the area outside of the Iraqi territory (25 km^2, 12%). Two major types of land covers can be recognized in the sub-basin through the analysis of remote sensing data using maximum likelihood supervised classification supported by ground truths. The recognized classes are shrubland distributes over an area of 35 km^2 (32%) and bare land covers 73 km^2 (68%) (Figure 2c).
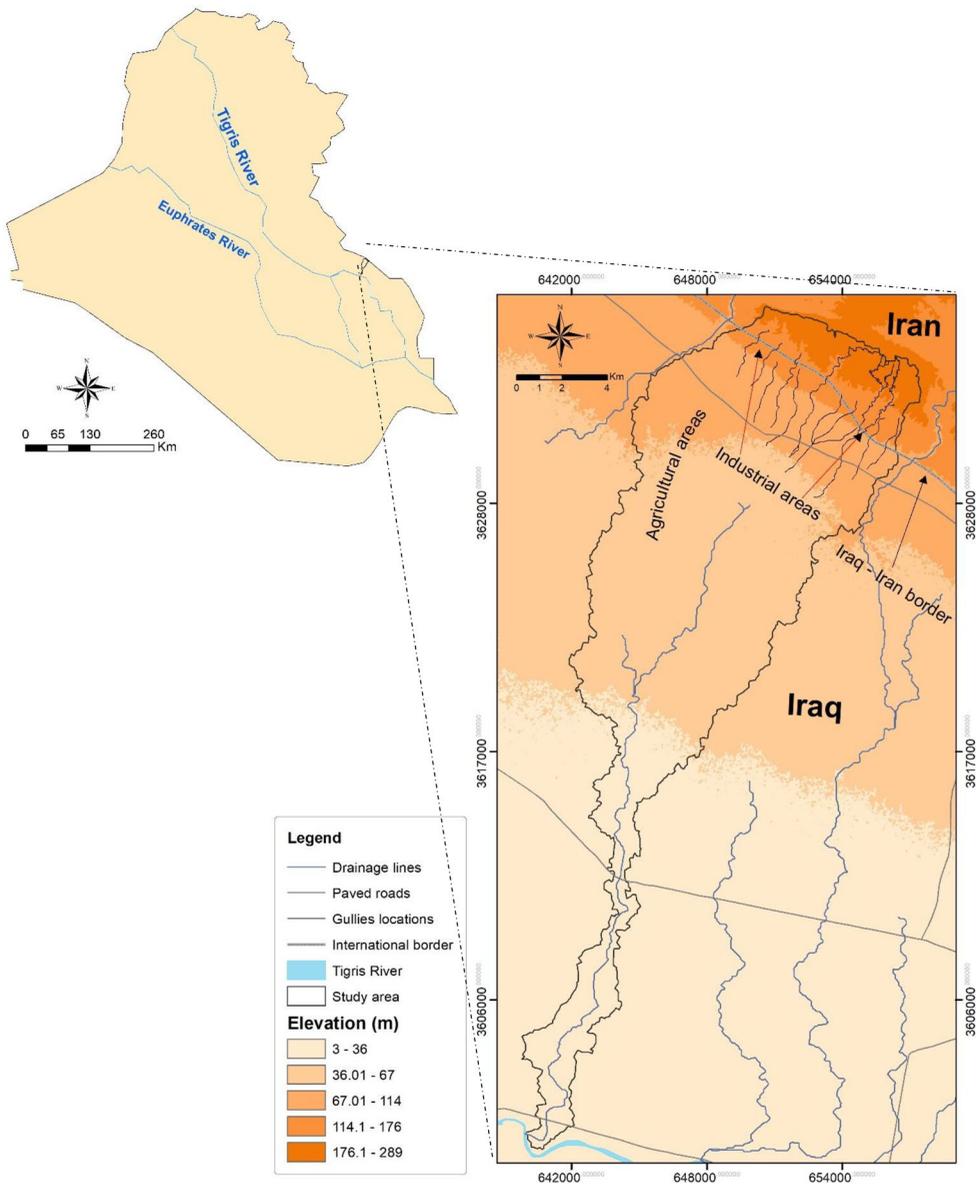
Figure 1. Location of the study area in Iraq.

2.2. Methodology

The adopted methodology in this study consists of five steps (Figure 3): (1) preparation of the gully inventory map; (2) generation of topographical attributes from DEM; (3) carry out feature selection for identifying the relevance of the factors in the building the susceptibility model; (4) application of models and mapping gully erosion susceptible zones; and (5) validation model results and compares the performances of MCDM and ML to choose which one the best for modelling gully erosion in the study area. A detail description of these steps is given as follows.
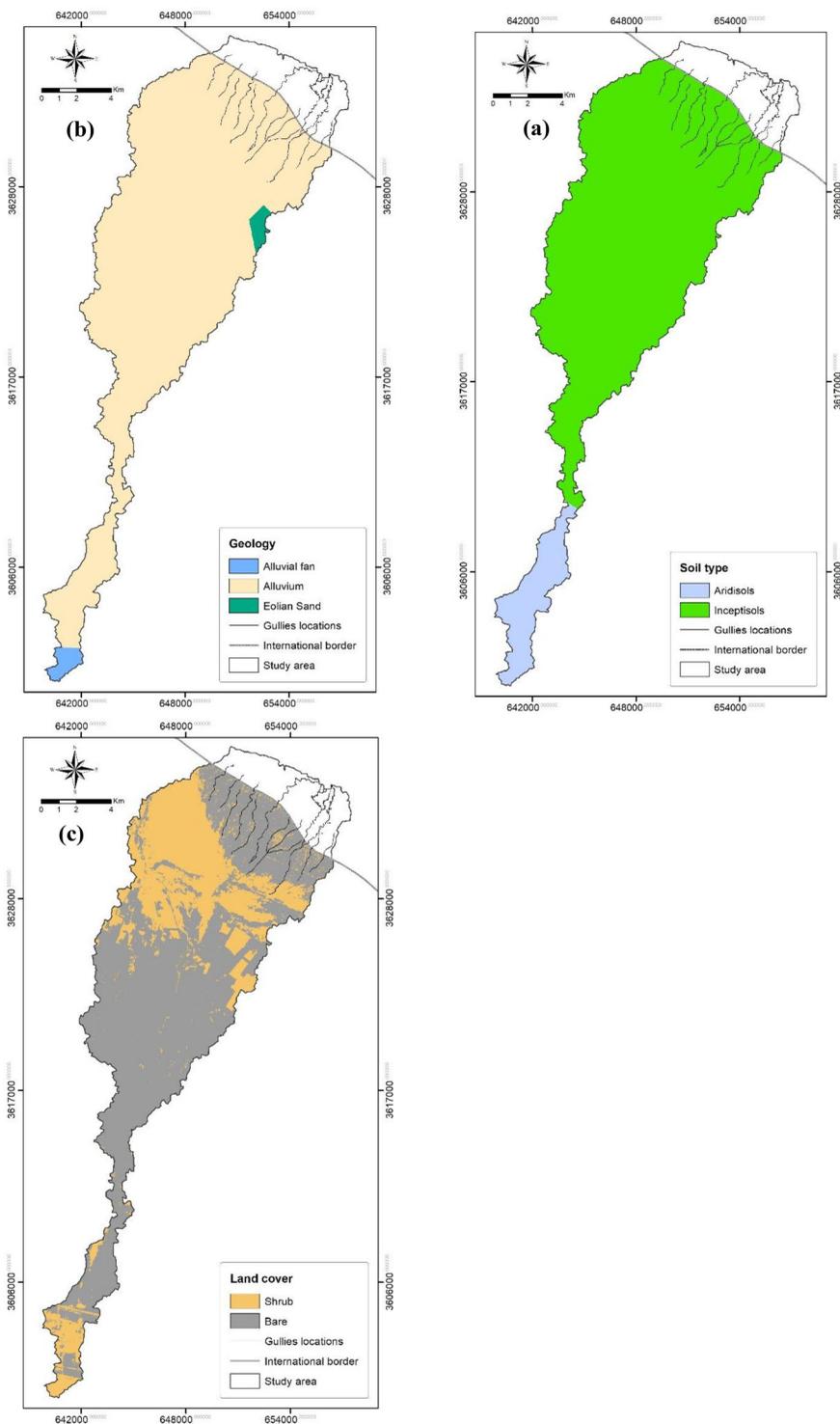
Figure 2. Layers of (a) geology (b) soil, and (c) land cover.

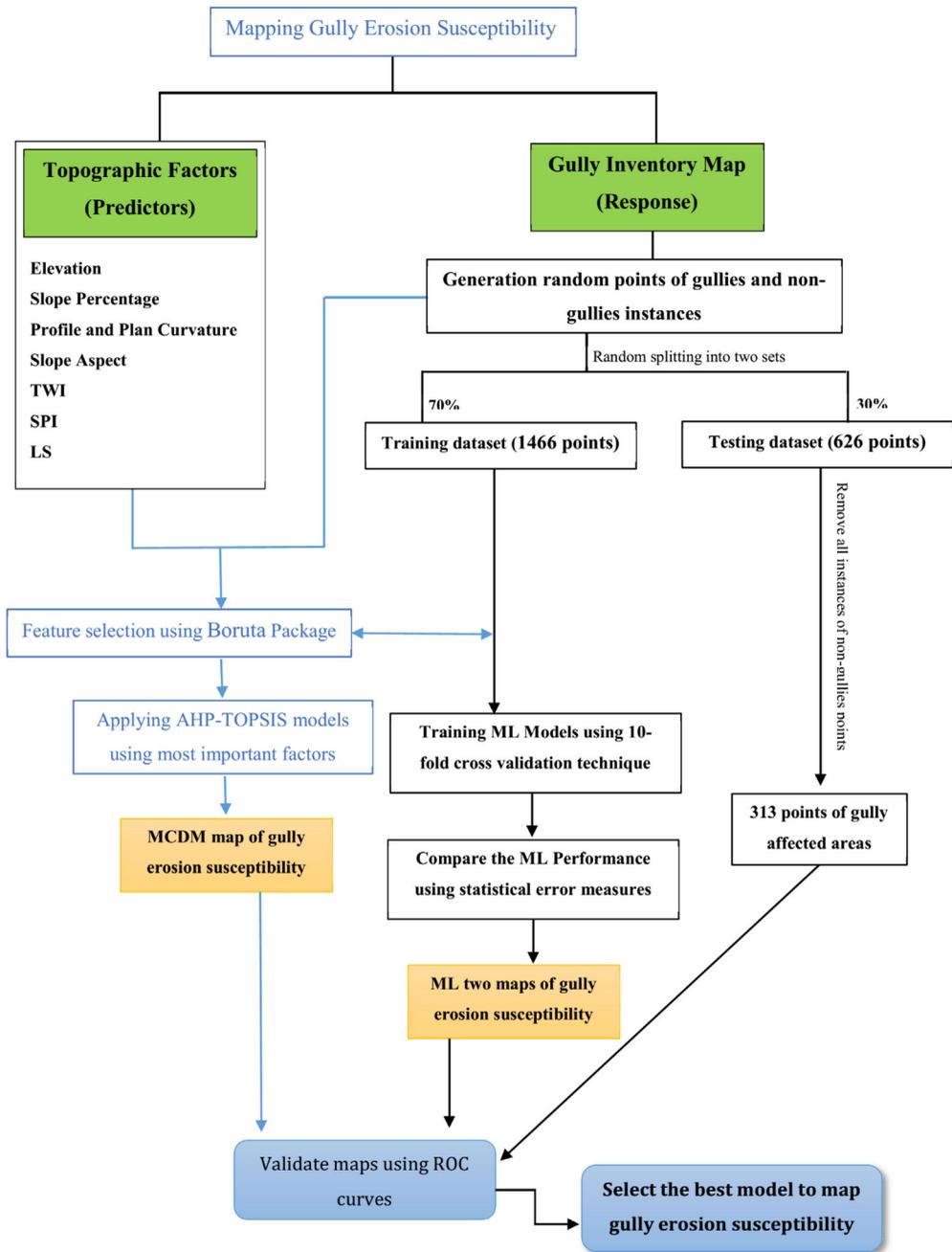**Figure 3.** Flow chart showing the steps used for the implementation of the gully erosion suscepti-bility model.

2.2.1. Gullies inventory map

The gully inventory map was prepared through comprehensive analyses of Google Earth images, verified by extensive field survey. A total of 10 gullies were identified (Figure 4). The gullies in the region were found to have linear plan shape with a
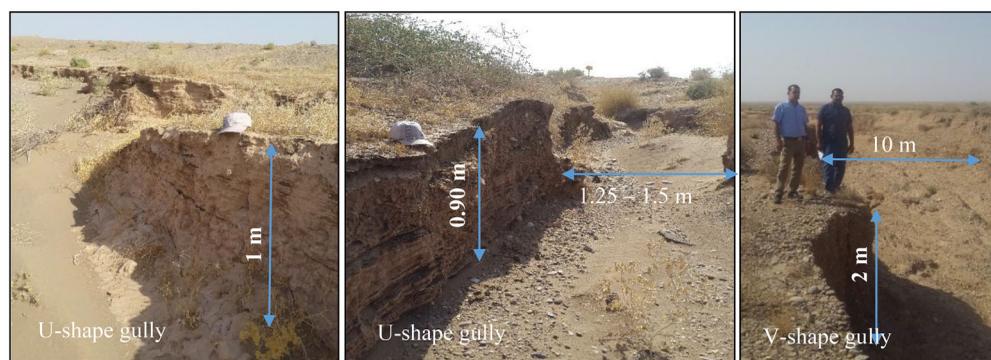
Figure 4. Photos showing the gully features in the study area.

length range from a few meters to hundreds of meters (Al-Abadi and Ali 2018). The gullies depth ranges from 0.5 m to over 2 m. The width of the gullies ranges 1.25 − 20 m. Most of the gullies have Type U cross sections, but often V-shaped gullies have also been found. Concentrated runoff created the majority of the gullies in the study region. Gully head cuts are most common on slopes of 15–30%, where surface runoff accumulates. In most cases, gullies are linked to the drainage lines network, allowing material eroded from upland regions to be evacuated more easily. Almost all gullies are formed on slopes underlain by alluvial fan sediments having a considerable amount of clay. Sometimes gullies appear close to the paved roads and industrial places, especially the sand and gravel quarries.

2.2.2. Preparation of topographic related factors

Recently, obtaining digital elevation model (DEM) with appropriate spatial resolution has become easy, especially after providing easy access to DEM data at free of charge by many governmental organizations and international agencies. The use of high-resolution DEM files does not necessarily produce better results (Gómez-Gutiérrez et al. 2015). In general, high-resolution data requires more time and cost for processing. Besides, the high-resolution DEM is not available for free of charge for the whole world. For these reasons, a DEM of SRTM (*Shuttle Radar Topography Mission*) was used in the study. The choice of this type of DEM is due to several reasons: (1) it has a fairly high resolution (1 arc-second or around 30 m); (2) it has a near-global coverage (from 56°S to 60°N); and (3) it is available in the public domain. After pre-processing of SRTM tiles (mosaic tiles, re-projection, and fill voids), the DEM was used to generate a number of primary and secondary topographic factors including: elevation, slope percent, slope aspect, profile and plan curvatures, topographic wetness index (TWI), stream power index (SPI), and slope-length (LS). From a review of the existing references on the mapping of gully erosion susceptibility, which are in the tens, it can be said that there is no study that does not use these factors in the development of gully erosion maps and almost in all types of terrain. For this reason, these factors are relied upon here to map the gully erosion of the study area. The elevation was derived from pre-processed DEM and was categorized into 10 classes (Figure 5a). Slope layer was also directly derived from DEM and classified into five classes: <2%,
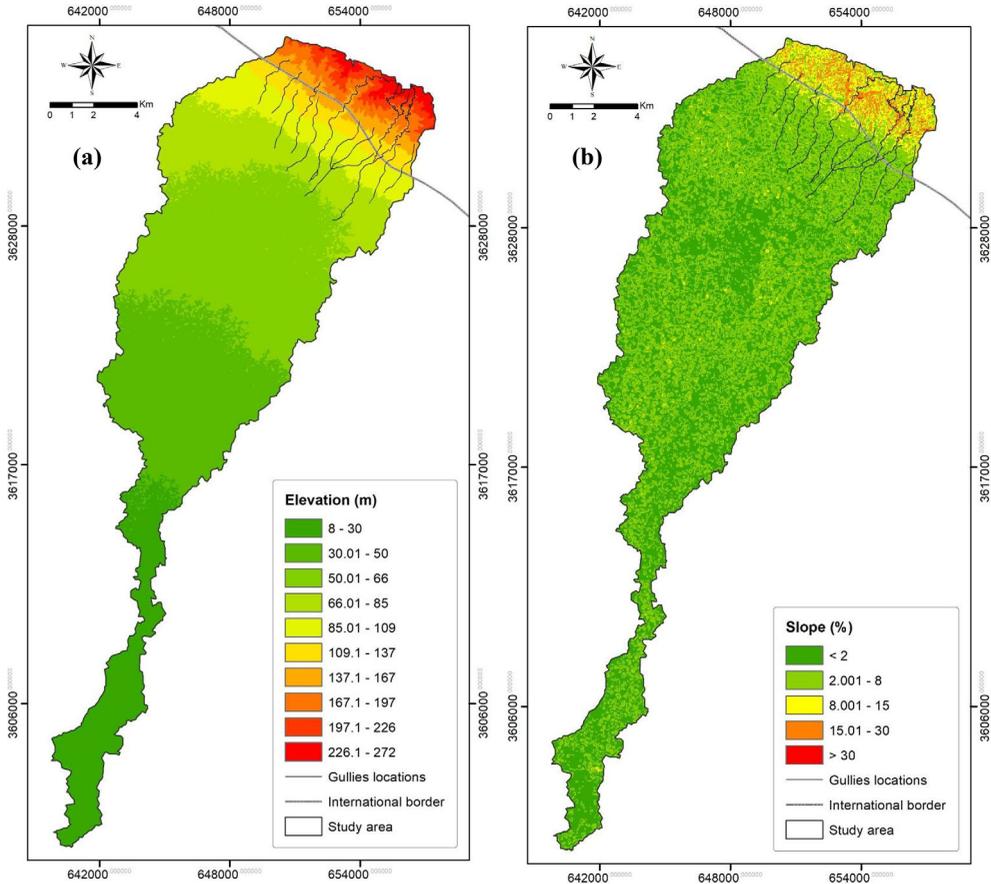
Figure 5. Topographic related factors: (a) Elevation, (b) Slope percentage, (c) Profile curvature, (d) Plan curvature, (e) Aspect, (f) TWI, (g) SPI, and (h) LS.

2–8%, 8–15%, 15–30%, and >30% (de Winner 2007) (Figure 5b). The profile and plan curvatures were represented by three classes: concave, flat, and convex (Figure 5c, d). Aspect was classified into nine classes (Figure 5e). The TWI was calculated as:

$$\text{TWI} = \ln(A_s/\tan\beta) \tag{1}$$

where, A_s is the specific catchment area and β is the slope in degree. The SPI was computed using the following equation (Moore et al. 1991):

$$\text{SPI} = A_s\tan\beta \tag{2}$$

Finally, the *LS* factor was calcualted using the equation developed by Moore and Burch(1986):

$$\text{LS} = (\text{fa} \times \text{rastercellsize}/22.13)^{0.4} \times (\sin\beta/22.13)^{1.3} \tag{3}$$

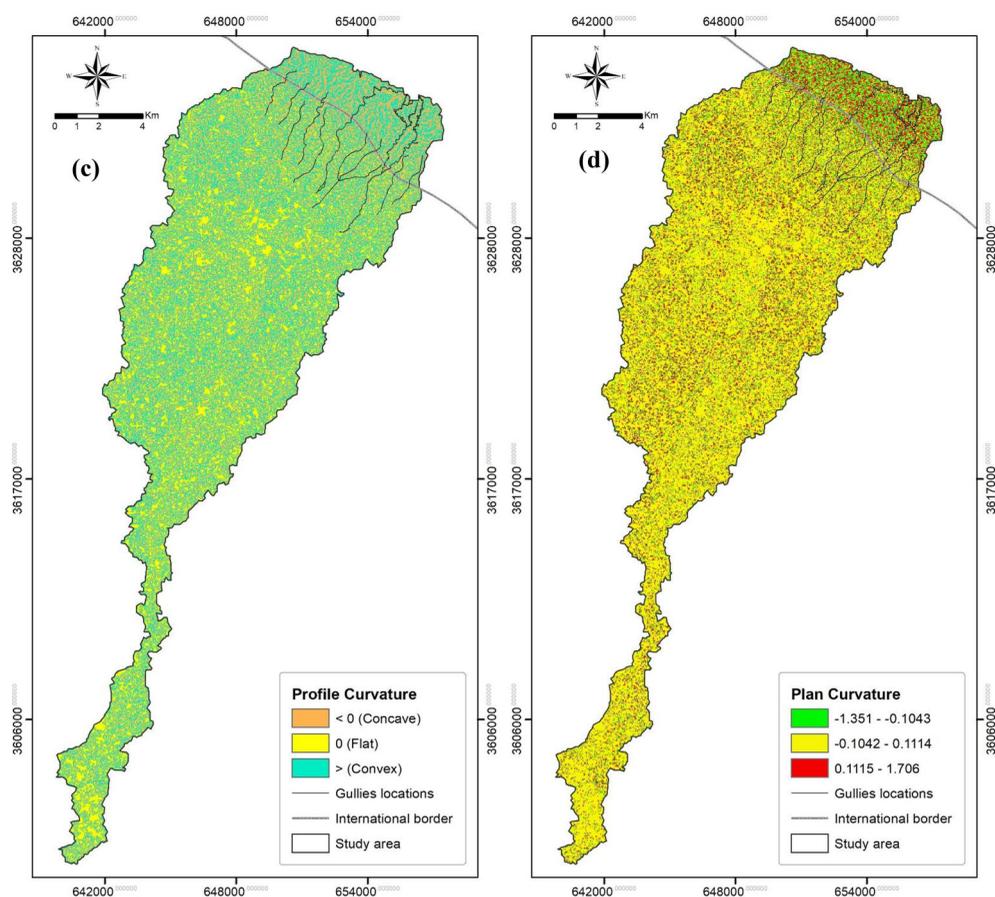where, fa is flow accumulation (Sujatha & Sridhar 2018).

Figure 5. (Continued).

The TWI, SPI, and LS factors were classified into five classes (Figure 5f–h). The raster map of each factor was prepared with 590 columns and 1245 rows, and thus, the total number of cells was 734,550.

2.2.3. Preparation of data for the development of models

The gully erosion susceptibility maps in this study were prepared using different paradigms, and therefore, the pre-processing of data was important in order to compare the results. The difference between this study and other similar studies is the process of representing gullies in geographic information system (GIS), where most of the previous studies use the "point feature" to represent individual gully and create "gully inventory map" regardless of the size and length of these gullies. In this study, most of the recognized gullies in the study area are longitudinal and extend to long distances, the gullies are represented here in a more realistic way by choosing a large number of points from the polygons that represent the spatial extent of these gullies. For this reason, the number of points used here is relatively large (1046 points); this representation approach is used for the first time in gully erosion susceptibility studies; let's take an example to explain that. Figure 6 shows a part of the study area with a longitudinal gully with different map scale. It is clear from Figure 1, it is impossible
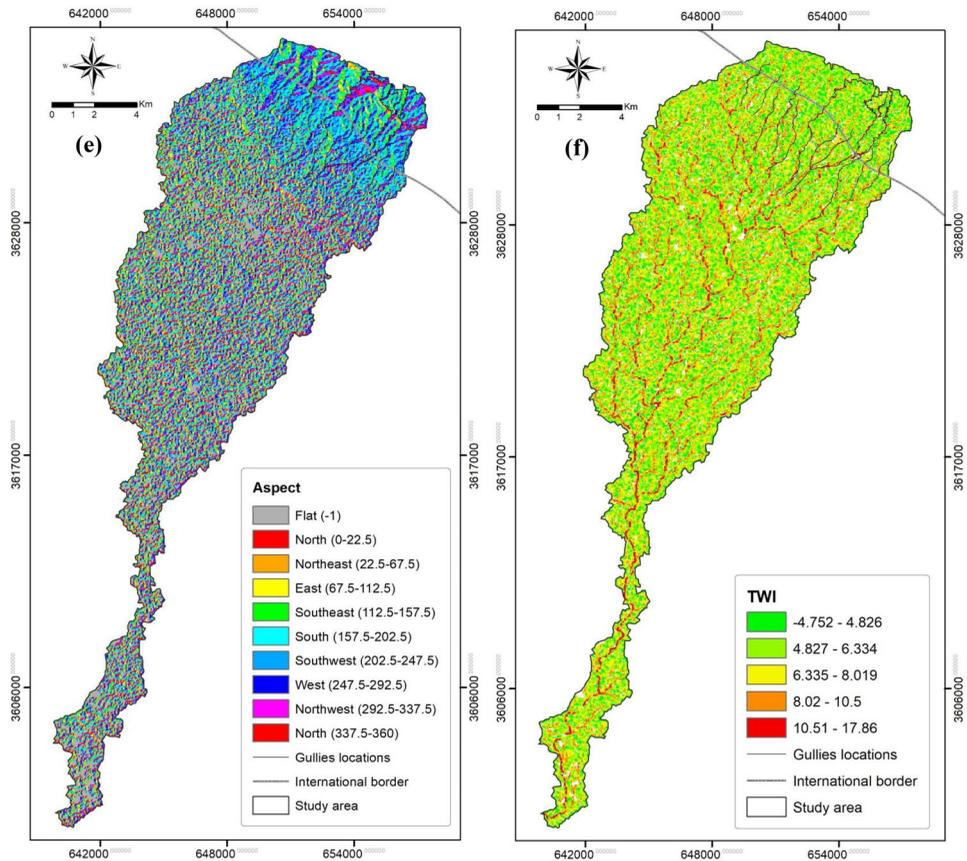
to represent this gully by only one-point feature in the GIS to refer to the area affected by this gully. Therefore, we developed gully inventory map to create a large number of points that refer to the areas affected by gullies instead of representing each individual gully with only one point and this is the only time this representation method used in gully susceptibility analysis.

Numerous points were created to cover the gullies in the entire area (1046 points). The same number of non-gully points was also created to represent areas not affected by gully erosion. Hence, the total number of created points was 2092. Using the random sampling process available in caret package of R statistical software, the dataset was divided with a ratio of 70/30. 70% of data (1466 points) was used for building and training the ML models (the MCDM model do not need sample training data). The remaining data 30% (626 points) were retained for testing the models using the relative operating characteristic (ROC) technique. For the 626 points, all the instances are not affected by gullies (313 points) were removed and the remaining points (gullies) were kept to compare the performance of the MCDM and ML models.
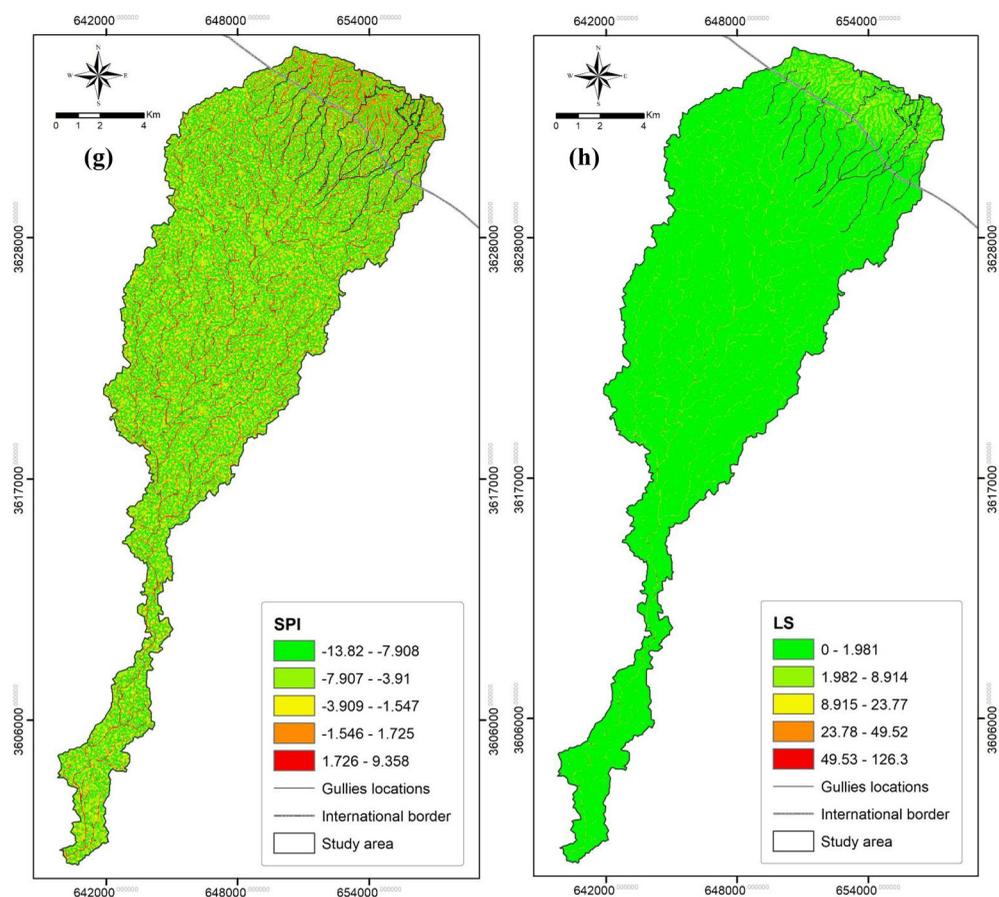
Figure 5. (Continued).

2.2.4. Modeling techniques

2.2.4.1 Feature selection. Feature selection is a process of selecting a subset of relevant features for the construction of a predictive model. The process helps in (1) simplifying the predictive models so that they can be easily developed and interpreted; (2) shorten the time of model training; and (3) enhance generalization by reducing model over-fitting (James et al. 2014). In general, there are three types of feature selection approaches, namely, filter, wrapper, and embedded. Each of these methods has advantages and disadvantages. In the present study, a wrapper algorithm, namely the Boruta Algorithm (BA) was used to investigate the features and order their importance in predicting gully erosion. BA is built based on random forest algorithm and has a capability to capture all the important features that are uncorrelated and non-redundant in a dataset with respect to an outcome (dependent) variable. A detail description of this algorithm can be found in Kursa and Rudnicki (2010).

2.2.4.2 Ahp and TOPSIS. AHP is an MCDA technique pioneered by Saaty (1980). In this technique, weights are allocated in a process of pair-wise relative comparison according to expert opinions (Bathrellos et al. 2017). In AHP, the problem being
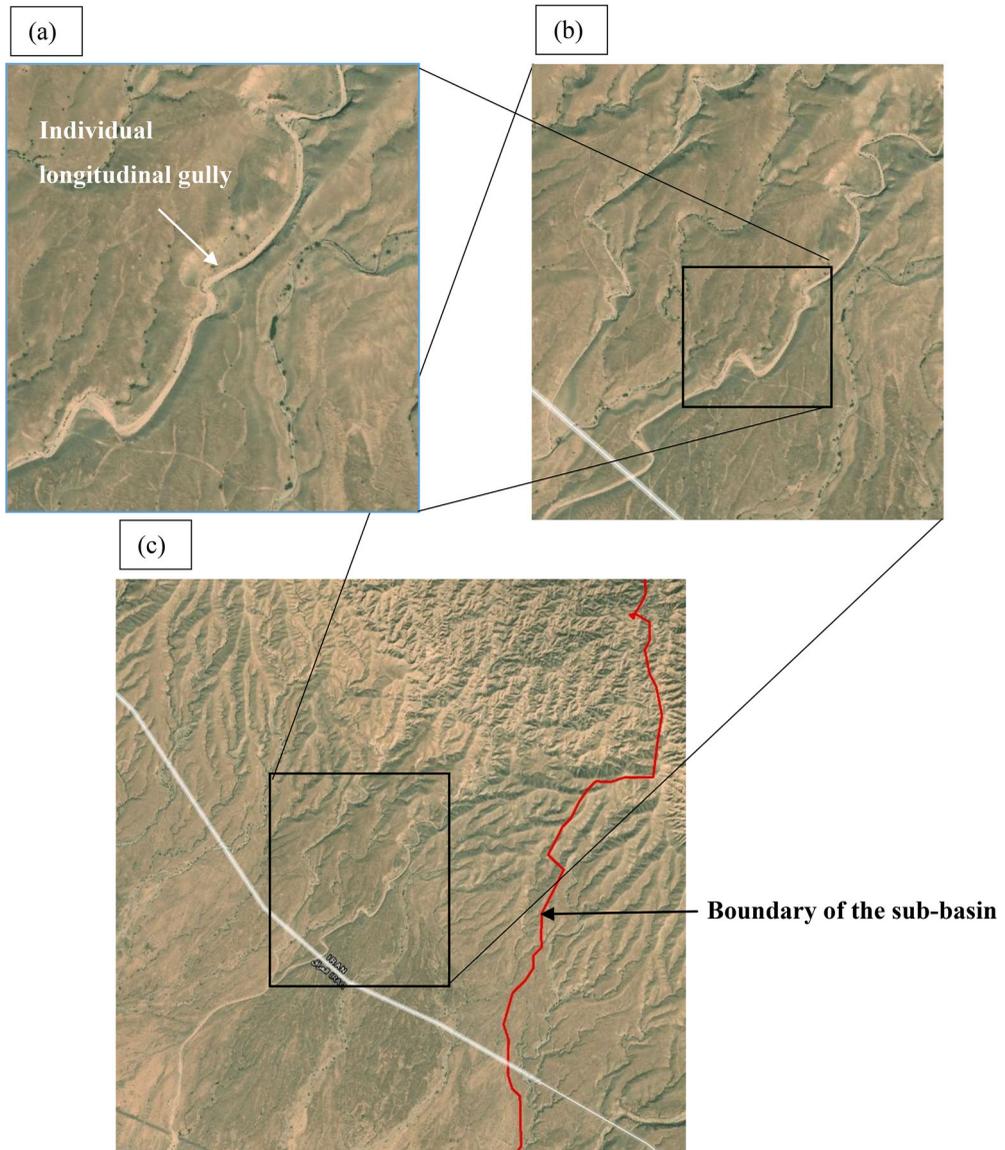
Figure 6. Individual gully in different map scale: (a) 1: 5,000, (b) 1:10,000, and (c) 1:25,000.

question is formulated as a hierarchical model. The AHP model in a simple form consists of three levels: goal, criteria, and alternative. The pair-wise comparison is implemented at each level of the hierarchy to obtain the judgmental matrix using Saaty's nine-point scale (Saaty 2008). Essentially, the pair-wise comparison contributes to evaluating each factor's contribution independently (Rezaei-Moghaddam and Karami 2008). The eigenvector of the judgmental matrix yields the weights of the elements in the hierarchy.

If $C = \{C_j | j = 1, 2, ..., n\}$ be the set of criteria, the decision matrix A resulting from pair-wise comparison can be written as:

$$A = \begin{bmatrix} a_{11} & a_{12} & ... & a_{1n} \\ a_{21} & a_{22} & ... & a_{2n} \\ ... & ... & ... & ... \\ a_{n1} & a_{n2} & ... & a_{nn} \end{bmatrix}, \tag{4}$$

where, a_{ij} is the pairwise comparison between i and j of a level with the upper level. The eigenvalue method was used to assign matrix elements weights involved in the analysis. The vector of weights is calculated by taking the principal eigenvector (w) on matrix A.

$$Aw = \lambda_{\max} w. \tag{5}$$

The decision matrix is said to be consistent if the following rule is satisfied

$$a_{ij} = a_{ik} \times a_{kj}; \ \forall i, \ j, \ k. \tag{6}$$

To assess the consistency judgment, Saaty (1980) used the consistency ratio CI:

$$CI = \frac{\lambda_{\max} - n}{n - 1}. \tag{7}$$

The TOPSIS is another MCDM method (Hwang & Yoon 1981), which was developed based on the concept that the chosen alternative should have the shortest geometric distance from the positive ideal solution (PIS) and the longest geometric distance from the negative ideal solution (NIS). The PIS is a solution that maximizes the benefit criteria and minimizes the cost criteria, while the NIS do the reverse (maximizes the cost criteria and minimizes the benefit criteria) (Assari et al. 2012). The steps used to implement TOPSIS begin with normalization of the decision matrix (Hwang and Yoon 1981):

$$r_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{m} x_{ij}}} \tag{8}$$

where, $i = 1, 2, \ldots, m$; $j = 1, 2, \ldots, n$; x_{ij} and r_{ij} are the original and normalized scores of the decision matrix, respectively.

The weighted normalized decision matrix is estimated by multiply weights w_i of the criteria by r_{ij} obtained using eq. (8):

$$v_{ij} = w_i r_{ij}. \tag{9}$$

The PIS and NIS are calculated as:

$$PIS = \left(\max v_{ij} | j \in J\right), \ \left(\min v_{ij} | j \in 'J\right)$$

$$NIS = \left(\min v_{ij} | j \in J\right), \ \left(\max v_{ij} | j \in 'J\right) \tag{10}$$

where, v_{ij} is the weighted and normalized x_{ij}, and J and $'J$ are the benefit and cost criteria, respectively. The separations of each alternative from the PIS and NIS are calculated as:

$$S_j^+ = \sqrt{\sum_{i=1}^{n} \left(v_{ij} - v_i^+\right)^2} \qquad j = 1, \ 2, \ ..., n$$

$$S_j^- = \sqrt{\sum_{i=1}^{n} \left(v_{ij} - v_i^-\right)^2} \qquad j = 1, \ 2, \ ..., n \tag{11}$$

The relative closeness coefficient to the ideal solution of each alternative is computed as:

$$C_j = \frac{S_j^-}{S_j^+ + S_j^-} \qquad j = 1, 2, \ ..., n \tag{12}$$

By comparing the C_j, the ranking of alternatives is determined.

2.2.4.3 Support vector machine (SVM). SVM is a ML technique designed to solve both classification and regression problems. It seeks to find a hyperplane that best divides a dataset into distinguished classes. A hyperplane is a line that linearly separates a dataset. Support vectors, on the other hand, are the data points located close to the hyperplane. They are critical elements of the SVM as removing these points dramatically changing the position of the dividing hyperplane. The margin is the distance between the hyperplane and the support vectors (Figure 7a). The SVM algorithm chooses a hyperplane with the greatest possible margin between the hyperplane and any point within the training dataset, and thus, the new data could be classified correctly. In case of difficulty to define a clear hyperplane, the two-dimensional data is converted to a three-dimension through kernelling concept (Figure 7b) and the hyperplane turns into a plane. The dataset is continually mapped into higher and higher dimensions until a hyperplane can be formed for optimal segregation.

Given a set of training examples $\{x_n\}_{n=1}^{N}$ $(x \in \mathbb{R}^L)$ and its corresponding output $\{y_n\}_{n=1}^{N}$ with $y_n \in \{-1, 1\}$, the linear separating hyper-plane can be described as:

$$w.x_i + b = 0 \tag{13}$$

where, w is a weight vector that defines the orientation of the hyperplane in the feature space, and b is the offset of the hyperplane from the origin (Cortes & Vapnik 1995). Using the Lagrangian multipliers, the following optimization problem is solved to get the optimal hyperplane:

$$\text{Minimize} \ \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i x_j) \tag{14}$$
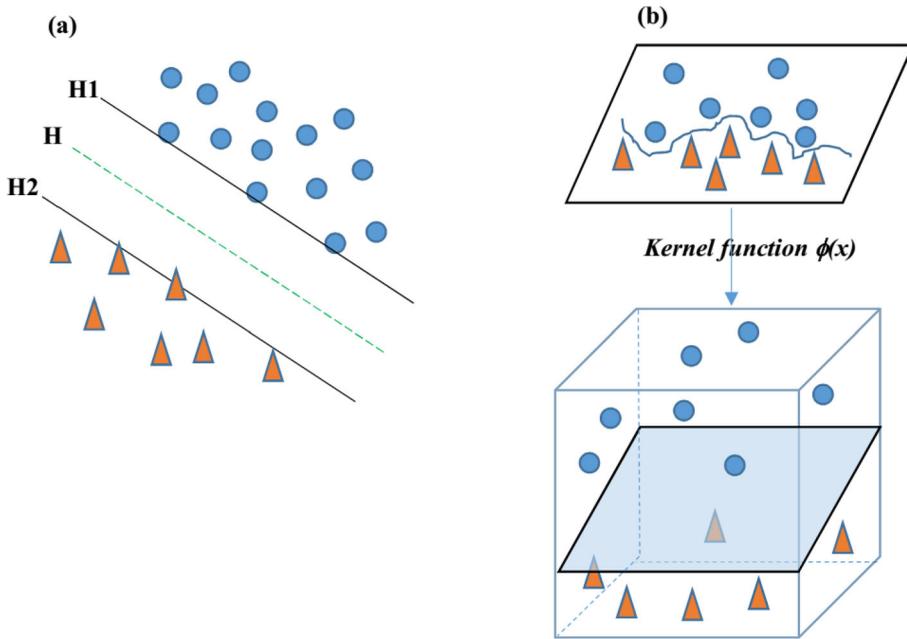
Figure 7. The basic of SVM: (a) the soft margin – separating hyperplane; (b) the kernel function (modified after Yao et al. (2008)).

$$\text{Subject to} \sum_{i=1}^{n} \alpha_i y_j = 0, \ 0 \le \alpha_i \le C, \tag{15}$$

where, α_i are lagrange multipliers and C is the cost of the penalty.

When the data can't be separated linearly, SVM introduces kernel function (Wilk-Kolodziejczyk et al. 2016) which maps the original data space into a high-dimension feature space. The idea is that when the data are transformed into a higher dimension, it can be easily separated (Nanda et al. 2018).

The classification decision function is then presented as,

$$g(x) = \text{sign}\left(\sum_{i-1}^{n} y_i \alpha_j K(x_i, x_j) + b\right), \tag{16}$$

where, $K(x_i, x_j)$ is the kernel function. The successful application of SVM depends on the right choice of kernel type and its parameters (Damaŝeviĉius 2010).

2.2.4.4 Naïve bayes (NB). The NB is a family of simple probabilistic algorithms based on Baye's theorem with strong independence assumptions between the features (Shmueli et al. 2018). NB has many advantageous including easy to design and build, does not require complex iterative parameter estimation, and robust to irrelevant feature and noise Soria et al. (2011).
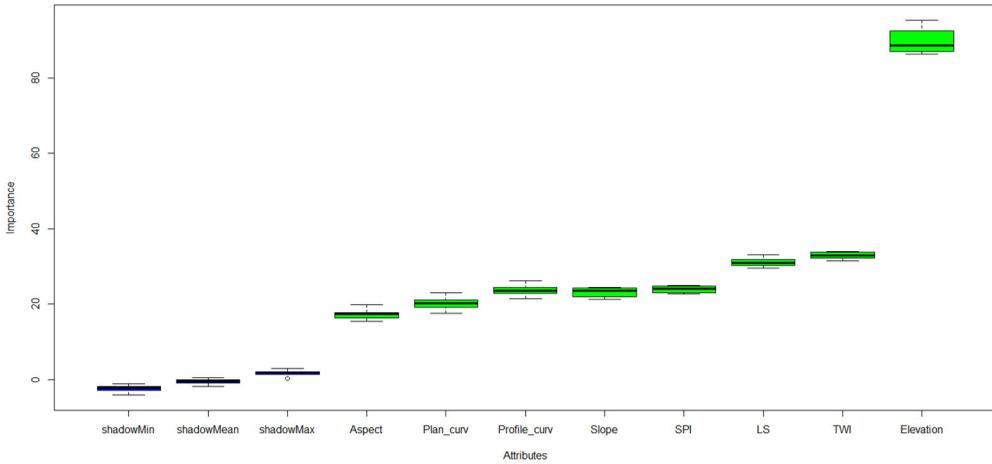
Figure 8. Result of feature selection analysis (Boruta package).

Given a class variable y and dependent feature vector x $(x_1, x_2, ..., x_n)$, Baye's theorem states the following relationship (Zhang & Su 2004):

$$P(y|x_1, x_2, ..., x_n) = \frac{P(y)P(x_1, x_2, ..., x_n|y)}{P(x_1, x_2, ..., x_n)} \tag{17}$$

With naïve conditional independence assumption, eq. (13) turns into

$$P(x_i|y, x_1, ..., x_{i-1}, x_{i+1}, ..., x_n) = P(x_i|y) \tag{18}$$

For all i, eq. (14) can be simplified as,

$$P(y| x_1, x_2, ..., x_n) = \frac{P(y) \prod_{i=1}^{n} P(x_i|y)}{P(x_1, x_2, ..., x_n)} \tag{18}$$

As $P(x_1, x_2, ..., x_n)$ is constant to the given input, the following classification rules can be used:

$$P(y|x_1, x_2, ..., x_n) \propto P(y) \prod_{i=1}^{n} P(x_i|y)$$

$$\Downarrow$$

$$\hat{y} = \arg \max_{y} P(y) \prod_{i=1}^{n} P(x_i|y) \tag{19}$$

The main weakness of NB lies in the assumption that the factors involved in model construction must be independent of each other; however, if there are factors dependent on each other, a large number of incorrect classifications may occur.

2.2.4.5 Performance evaluation metrics. To investigate the performance capability of the ML models, five statistical based confusion matrix measures were calculated: accuracy, sensitivity, specificity, precision, and F1 score. A more detail how these measures calculate and how they are used for measuring the model predictive ability is found in Appendix I.

In this analysis, the ROC curve was used to compare the performance of a hybrid MCDM model built and to select the best machine learning models. The ROC is a graphical plot which clarifies a predictive model ability when its threshold for discrimination is varied. It is created by plotting the true positive rate (sensitivity) against the false positive rate (1 – specificity) for various threshold settings. The area under the ROC (AUC) detects the quality of a classifier by explaining the model's ability to forecast correctly the occurrence and non-occurrence of pre-defined events (Mason & Graham 2002). The AUC value ranges from 0.5 (random prediction represented by a diagonal straight line) to 1 (perfect prediction).

2.2.4.6 Software used and the procedure adapted. Different software were used in this study for mapping the gully erosion susceptibility. The online version of AHP calculator (https://bpmsg.com/ahp/ahp-calc.php) was used for applying AHP technique and determine the weights of topographic related factors used in this analysis. To run the TOPSIS algorithm, the "topsis" package (https://cran.r-project.org/web/packages/topsis/index.html) in R statistical software was used. The weights derived using AHP were passed to the algorithm and the cost type normalization technique (larger the better) was used to normalize all eight factors. This type of normalization was used after studying the relationship between gully locations and the topographic factors affecting their development. The caret (Kuhn 2008) and e1071 (https://cran.r-project.org/web/packages/e1071/index.html) packages of R statistical software were used to develop the SVM and NB machine learning models. A k-fold cross validation procedure was used to train the models. This approach is a re-sampling technique used to evaluate the performance of predictive models through randomly partitioning the original data set into k groups (folds) of approximately equal size (James et al. 2014). One sub-sample is used in the verification test, and the remaining sub-samples are used for model training. The re-sampling process is repeated *k* times (the folds) and the results from the folds averaged to produce a single estimate (Chen et al. 2018). For SVM models, four kernel functions were used: linear, radial, polynomial, and sigmoid. The raw data was pre-processed using centre-scaling function and the parameters of each kernel were tuned using the grid-search approach. The grid-search works by looking for the combination of parameters in a given region's length and then emitting the best parameters through the minimum classification error to construct the classifier (Nanda et al. 2018). The tuned SVM models were implemented using the optimized parameters, and the result were compared using the six statistical errors to choose the best one. The NB model was developed in two stages using e1071 and caret packages. In the first stage, the train data was pre-proceed using the Box Cox method (normalization), standardized using centre-scaling, and reduction via principal component procedure (PCA). After pre-processing, NB classifier was trained with 10-fold cross validation without tuning the hyper-parameters of the classifier. In the next stage, the hyper-parameters of the NB model were tuned

Table 2 Results obtained using AHP technique.

Factor	Elevation	LS	TWI	SPI	Aspect	Slope	Plan curvature	Profile curvature
Priority	33.1%	23.1%	15.7%	10.6	7.1	4.8%	3.3%	2.4%
Rank	1	2	3	4	5	6	7	8

Consistency ratio = 2.9%

using grid-search approach. Finally, the ROC curves were produced by IDRISI Selva software.

3. Results

3.1. Feature selection

The results obtained using BA for the selection of features were summarized in Figure 8. All topographical related factors significantly contribute to gully erosion in the sub-basin (average merit > 0). The most important factor was elevation, followed by TWI, LS, SPI, profile curvature, slope, plane curvature, and aspect. Therefore, all the factors were included in the modelling of gully erosion susceptibility.

3.2. Application of AHP-TOPSIS model

The weights of the factors according to their importance in mapping gully erosion were calculated using the AHP method and depending on the feature selection results (Section 3.1), where the highest weight was assigned to the elevation and the lowest weight was assigned to the profile curvature. The pair-wise comparison and the consistency ratio were used to judge the consistency (Table 2). It can be concluded from Table 2 that the judgment was consistent (the consistency ratio was smaller than 0.1 (0.029)). The highest weight was given to elevation (33.1%), followed by LS (23.1%), TWI (15.7%), SPI (10.6), aspect (7.1%), slope (4.8%), plan curvature (3.3%), and profile curvature (2.4%).

The ranked values obtained using TOPSIS were mapped using ArcGIS 10.2 and classified using quantile classification scheme into three categories: very low – low, moderate, high – very high (Figure 9a). Areas occupied by each category are given in Table 3. Moderate and high – very high susceptibility zones were found to occupy 61% of the study area (about $124 \, km^2$), while very low – low susceptibility zones were found to encompass the remaining part of the sub-basin (39%; $81 \, km^2$). The high susceptibility zones were distributed in the upper part of the sub-basin and in some areas in the centre. The moderate zone covers the central part of the study area and the low susceptible zone was found to cover the lower parts of the basin beside some parts in the centre. Validation of this model using ROC curves (Figure 10) revealed a value of AUC equal to 0.933, which indicates an excellent prediction by the APH-TOPSIS model.

3.3. Application of SVM and NB models

Table 4 shows the results of SVM models using the optimized hyperparameters. In terms of accuracy, all the SVM models performed well. The best SVM model was
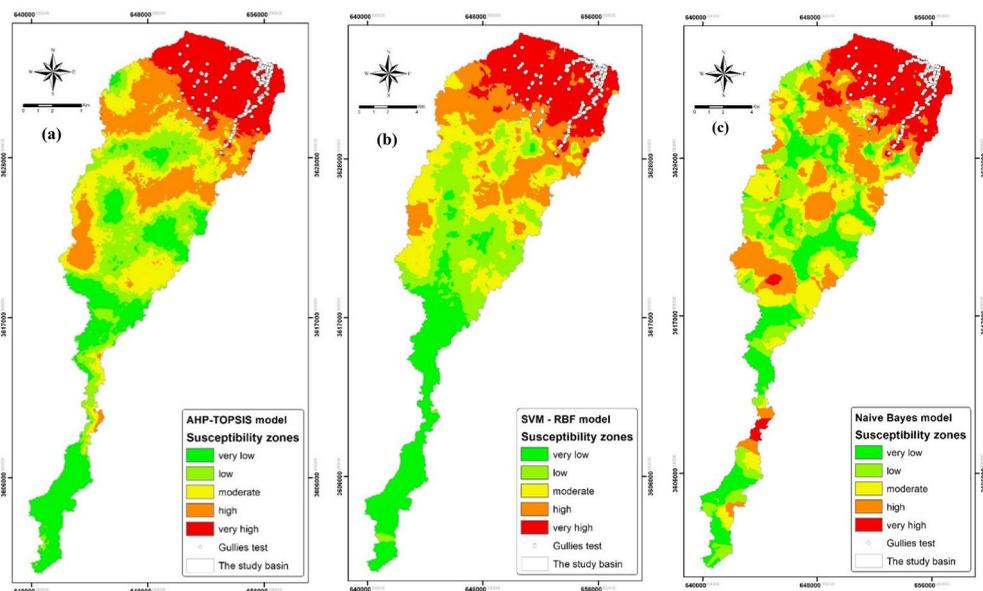
Figure 9. Gully erosion susceptibility maps obtained using: (a) hybrid AHP-TOPSIS; and (b) SVM-Radial, (c) Naïve Bayes.

Table 3. Areas covered by different categories of gully erosion susceptibility.

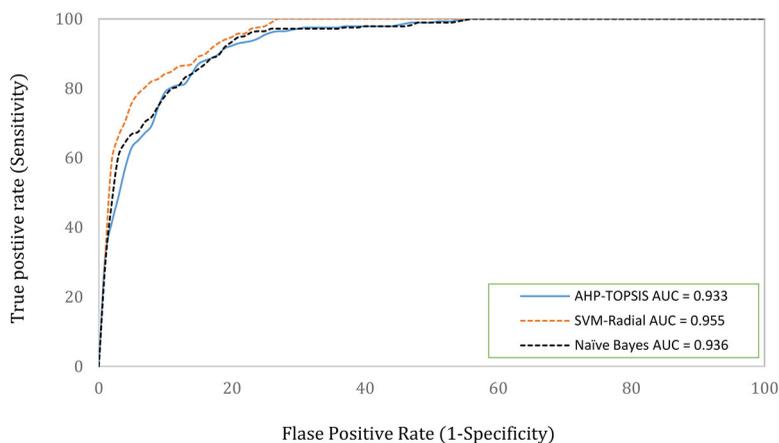| Model	very low – low	Moderate	high – very high
	Gully hazard zones		
AHP-TOPSIS	39% (81 km²)	22% (44 km²)	39% (80 km²)
SVM – Radial	38% (79 km²)	22% (44 km²)	40% (82 km²)
Naïve Bayes	37% (75 km²)	22% (44 km²)	42% (86 km²)

Figure 10. Validation of the models using ROC curve.

found for the RBF kernel (Accuracy = 87.6%). In terms of sensitivity, the best SVM model for classifying the gullies was also SVM with RBF function (Sensitivity = 90.4%), followed by SVM with polynomial kernel type (Sensitivity = 70.5%). On the

Table 4. The evaluation statistical error measures for the SVM with different kernel functions.

| Evaluation measure	Kernel type (10-fold cross validation training mode)			
	Linear	Polynomial	RBF	Sigmoid
Accuracy	0.848	0.852	0.876	0.782
Sensitivity	0.695	0.705	0.904	0.563
Specificity	0.881	0.916	0.847	0.789
Precision	0.808	0.789	0.847	0.774
F1 score	0.842	0.843	0.872	0.780

other hand, the best model for classifying non-gullies cases was SVM with polynomial kernel (Specificity = 91.6%), followed by SVM with linear, RBF and Sigmoid kernel (Specificity = 88.1%, 84.7%, and 78.9%, respectively). With respect to the "Precision" in measurement, the SVM with RBF kernel performed the best (84.7%), followed by SVM with a linear kernel (80.8%), SVM with polynomial kernel (84.7%) and SVM with Sigmoid kernel (77.4%). Examining the predictive capability of the SVM model using the F1 score revealed that the SVM with the RBF was the best model (F1 = 87.2%), followed by SVM with polynomial kernel (84.3%), SVM with linear kernel (84.2%) and SVM with sigmoid kernel (78%). Based on these results, the best SVM model was SVM with RBF kernel. This model was then chosen to map gully erosion susceptibility to compare the results of AHP-TOPSIS. The probability values of the SVM with RBF were interpolated using the ordinary kriging interpolation technique in ArcGIS 10.3. The interpolated values were categorized into five gully erosion susceptibility zones using quantile classification scheme (Figure 9b): very low, low, moderate, high, and very high. Areas occupied by these zones are presented in Table 3. The values were found to be closely similar to that derived using the AHP-TOPSIS model. Examining the SVM model performance using the ROC curve for the testing data revealed a value of AUC equal to 0.955 (excellent model). The gully erosion hazard zones were found distributed in the study area in the same manner as that generated by AHP-TOPSIS model. In general, the very-high zones were found to encompass the northern parts, a moderate zone in the central part, whereas very low – low zones were found to cover the southern part.

The high susceptibility zones were found to distribute in the upper part of the sub-basin and in some areas in the central part, while the moderate zone was found mainly concentrated in the central part and the low erosion susceptible zone was found to cover the lower parts of the area beside some parts in the centre. Validation of the model using ROC curves (Figure 10) revealed a value of AUC equal to 0.955, which indicates an excellent performance model.

Results of the NB model for the first stage (without tuning the algorithm hyper-parameters) are presented in Table 5. The overall accuracy of the model in this stage was 0.83. The predictive capability of the NB model was improved after tuning the hyper-parameters (stage two). The accuracy of the NB model in the second stage (tuning the hyperparameters) was 86.4% (Table 5). The probability of the tuned NB model was exported to the ArcGIS 10.3 and interpolated to reveal the levels of gully erosion susceptibility in the study area. The probability values were found in the range 0.02–1.00. The values were classified using a quantile classification scheme into five zones in a similar fashion adopted for other models (AHP-TOPSIS and SVM-

Table 5. Performance of NB model in term of statistical measures.

Evaluation measure	Without tuning process	With tuning and preprocess
Accuracy	0.839	0.864
Sensitivity	0.864	0.877
Specificity	0.800	0.830
Precision	0.801	0.829
F1 score	0.833	0.844
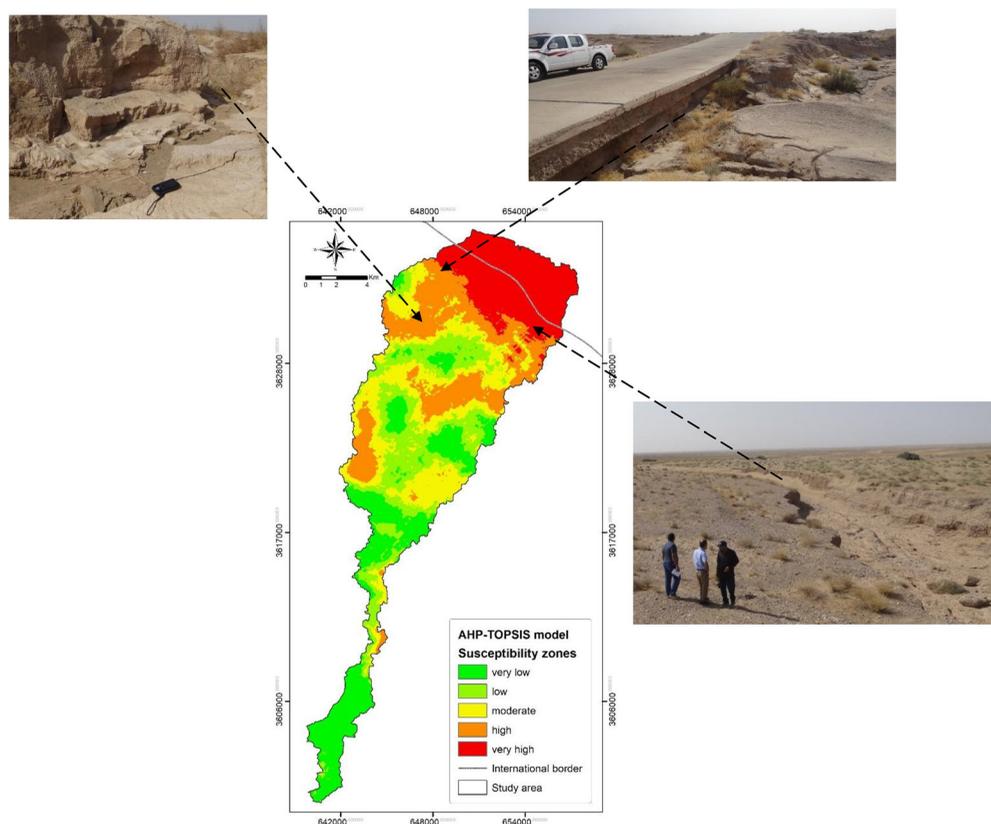
Figure 11. Ground truth check and field validation of the gully erosion susceptibility map.

RBF). Table 3 summarizes the areas covered by those five zones. Examination of the performance of the model for the test data using the ROC curve showed 0.936 AUC value, which was slightly better value than AHP-TOPSIS model (Figure 10).

The locations of three distinct gullies that were not utilized in the training and validation test were overlaid over the gully erosion map produced by AHP-TOPSIS to further evaluate the susceptibility of the gully erosion map produced by AHP-TOPSIS, Figure 11. Most of these gullies are found to be distribute in the moderate, high, and very high susceptibility zones. This is convincing evidence that the maps generated here may be used to determine where heavy soil erosion is likely to occur, and so make a major contribution to soil protection and land use management.

4. Discussion

Gully erosion is a complicated geomorphic hazard that results in significant economic losses as well as substantial infrastructural and natural environment degradation. However, modelling of soil erosion susceptibility with high accuracy is still difficult to predict, especially in the region where the available data is poor (Arabameri et al. 2019). Topography is widely recognized for playing a key part in the formation of gullies, since it influences the erosive strength of the overland flow. Other factors that influence gully development, including as lithology, climate, and plant cover, are all influenced by topography. As a result, in gully susceptibility research, developing a model for mapping gully erosion using just topographic characteristics is critical. For this regard, in this study, a methodology is presented to develop gully erosion susceptibility map using a hybrid AHP-TOPSIS using only topographic related factors derive from a medium resolution DEM. The topographic factors used in the study include elevation, slope, profile and plan curvatures, aspect, TWI, SPI, and LS. The benefit of adopting this model is that it is an expert knowledge-driven model that may be constructed without the need for a training sample dataset. Such training samples (the locations of gullies or areas affected by a specific gully) are sometimes not available, few or insufficient to complete the training phase of the data-driven models, such as statistical and machine learning models. The findings of AHP-TOPSIS are compared to those of two well-known accurate machine learning classifiers, SVM and NB, to demonstrate the benefits of employing such a simple model.

The contribution of factors used in the analysis of gully erosion models is of considerable interest, and it has received a lot of attention in the literature. The importance of these factors in a specific region, is depended on the nature of the study area; i.e., the soil type, LULC, exposed lithological units, geomorphology, and meteorological conditions, and thus a factor that has a high contribution in a specific area may have a lower contribution to other areas with different conditions. The other important things affect the contribution of these factors are the modelling techniques used and evaluation metric employed (Arabameri et al. 2019). The finding of this research confirm that the elevation, TWI, SPI, and LS are the most contribution factors in developing soil erosion. The less contribution factors are Profile curvature, Slope, Plane curvature, and Aspect. The northern section of the research region, as well as the slope of the hills into Iraqi territory to the southeast, are dominated by higher elevation. The combination of heavy rains on Iranian soil and the rough terrain generally results in a high concentration of runoff. The concentrated runoff leads to soil erosion and transports soil material deep into the sub-watershed under investigation. As a result, the gully head cuts are generally in high places, which, in addition to surface runoff and slope, contributes to produce these gullies, as does the composition of the soil and Quaternary sediments, which primarily consist of sand, gravel, silt, and clay that is simple to carry. This may explain why the elevation is the most important factors in the developing gullies in the study area. TWI and SPI also had a substantial impact on gully head erosion, which aids in the identification of locations with higher soil moisture content, water accumulation content, and erosive power (Zabihi et al. 2018). The LS, on the other hand, is a

sediment transport capacity index that is used to calculate the impact of surfaces runoff velocity on erosion. Gully erosion becomes more common when the LS, TWI, and SPI rise (Rong et al. 2019).

The performance of the hybrid knowledge-driven AHP-TOPSIS and data-driven SVM and NB machine learning classifiers to map gully erosion susceptibility zone on the study area were evaluated and compared. Validation of the results using the ROC indicated that the SVM (AUC = 0.955) with radial kernel function outperformed NB (AUC = 0.936) and AHP-TOPSIS (AUC = 0.933) models. Although AHP-TOPSIS did not outperform ML models, it is simple to build and does not require training data, which are both advantages of simple models. The convergence of the results of the hybrid AHP-TOPSIS model with advantageous ML classifiers (SVM and NB) is probably due the fact that the derivation of weights for this model by AHP method did not depend on the opinions of experts which are usually misleading and bear a lot of uncertainty. Instead, the AHP weights were built based on the feature selection by Boruta package, which depends on the random forest algorithm known to be very accurate in discovering a pattern in input data. Arabameri et al. (2019) found similar findings when they compared the COPRAS MCDM model ensembles to three machine learning models: logistic regression, boosted regression tree, and random forest algorithms. They found that combining MCDM with ML methods to produce factor weights, which are required to apply MCDM models, significantly improves the model. It is almost the only study with which the results of this study can be compared, despite the difference in the method used and the factors used in the gully erosion modelling.

Mapping of gully erosion susceptibility in the study area using the three models used revealed three hazards: very low-low, moderate, high-very high. The areas occupy by these zones are approximately the same for the three models and distribute as 78, 45, and $83\,km^2$ for the very low-low, moderate, high-very high, respectively. The high-very high zones mainly encompass the northern parts of the study area and correlate with the high elevation range, high slope, north–northwest, east, and west aspects, and high values of TWI, SPI, and LS. The moderate and very low–low zones mainly occupy the middle and south of the considered area and without specified pattern.

5. Concluding remarks

This study contributes to a systematic comparison and evaluation of two machine learning models and a hybrid MCDM model for modelling gully erosion susceptibility and using only derived topographic factors from a medium resolution DEM. The main conclusions of this study are: (i) the most effective factors controlling the gully erosion development in the study area are elevation, TWI, SPI, and LS. (ii) The performance of the hybrid AHP-TOPSIS model is substantially improved and approaches the performance of the advanced machine learning models when the feature selection that depend on random forest algorithm contribute to assign weights using AHP and rank the important factors instead of using expert opinions having a lot of uncertainty. (iii) The SVM (AUC = 0.955) with radial kernel function beat the NB (AUC

= 0.936) and AHP-TOPSIS (AUC = 0.933) models in terms of ROC validation. Although AHP-TOPSIS did not beat ML models, it is easy to construct and does not require training data, both of which are benefits of simple models. (v) Modelling of gully erosion susceptibility in the considered area indicated that high susceptibility zones mostly cover the northern parts of the study area and is associated with a high elevation range, steep slope, north–northwest, east, and west aspect, and high TWI, SPI, and LS values. The low and moderately susceptibility zones are mostly found in the centre and south of the study region, with no discernible pattern. (vi) Using just topographical data generated from the medium resolution DEM, the susceptibility to gully erosion may be modelled with high accuracy using a hybrid AHP-TOPSIS model supported by the Boruta package to rank the relevance of factors and calculate suitable factor weights.

Disclosure statement

The authors declare no conflict of interest.

Funding

Data availability

Raw data were generated at University of Basrah, Iraq. Derived data supporting the findings of this study are available from the corresponding author [Biswajeet Pradhan] on request.

References

Agnesi V, Angileri S, Cappadonia C, Conoscenti C, Rotigliano E. 2011. Multi parametric GIS analysis to assess gully erosion susceptibility: a test in southern sicily, Italy. Landform Anal. 17:15–20.

Al-Abadi AM, Al-Ali AK. 2018. Susceptibility mapping of gully erosion using GIS-based statistical bivariate models: a case study from Ali Al-Gharbi District, Maysan Governorate, southern Iraq. Environ Earth Sci. 77(6):249.

Al-Abadi AM, Al-Kubaisi QY, Ma A-G. 2018. Mapping groundwater zones contaminated by hydrocarbons in the Dammam aquifer in the Karbala–Najaf plateau, Iraq. Environ. Earth Sci. 77(633). https://doi.org/10.1007/s12665-018-7827-2.

Angileri SE, Conoscenti C, Hochschild V, Märker M, Rotigliano E, Agnesi V. 2016. Water erosion susceptibility mapping by applying Stochastic Gradient Treeboost to the Imera Meridionale River Basin (Sicily, Italy). Geomorphology. 262(1):61–76.

Arabameri A, Pradhan B, Rezaei K, Conoscenti C. 2019. Gully erosion susceptibility mapping using GIS-based multi-criteria decision analysis techniques. Catena. 180:282–297.

Assari A, Maheshand TM, Assari E. 2012. Role of public participation in sustainability of historical city: usage of TOPSIS method. IJST. 5(3):1–2294.

Bathrellos GD, Skilodimou HD, Chousianitis K, Youssef AM, Pradhan B. 2017. Suitability estimation for urban development using multi-hazard assessment map. Sci Total Environ. 575: 119–134. https://doi.org/10.1016/j.scitotenv.2016.10.025.

Carrara A, Cardinali M, Detti R, Guzzetti F, Pasqui V, Reichenbach P. 1991. GIS techniques and statistical models in evaluating landslide hazard. Earth Surf Process Landforms. 16(5): 427–445.

Chen W, Zhang S, Li R, Shahabi H. 2018. Performance evaluation of the GIS-based data mining techniques of best-first decision tree, random forest, and naïve Bayes tree for landslide susceptibility modeling. Sci Total Environ. 644:1006–1018. 2018.06.389.

Conforti M, Aucelli P, Robustelli G, Scarciglia F. 2011. Geomorphology and GIS analysis for mapping gully erosion susceptibility in the Turbolo Stream catchment (Northern Calabria, Italy). Nat Hazards. 56(3):881–898. https://doi.org/10.1007/s11069-010-9598-2.

Conoscenti C, Agnesi V, Cama M, Caraballo-Arias NA, Rotigliano E. 2017. Assessment of gully erosion susceptibility using multivariate adaptive regression splines and accounting for terrain connectivity. Land Degrad Dev. 29(3):724–736. https://doi.org/10.1002/ldr.2772.

Cortes C, Vapnik V. 1995. Support-vector networks. Mach Learn. 20(3):273–297.

Damaŝeviĉius R. 2010. Structural analysis of regulatory DNA sequences using grammar inference and Support Vector Machine. Neurocomputing. 73(4–6):633–638. https://doi.org/10.1016/j.neucom.2009.09.018

Dube F, Nhapi I, Murwira A, Gumindoga W, Goldin J, Mashauri DA. 2014. Potential of weight of evidence modelling for gully erosion hazard assessment in Mbire District—Zimbabwe. Phys Chem Earth. 67–69:145–152. https://doi.org/10.1016/j.pce.2014.02.002.

Fiesler E, Beale R. 2020. Handbook of neural computation. Routledge: CRC Press.

Gómez-Gutiérrez Á, Conoscenti C, Angileri SE, Rotigliano E, Schnabel S. 2015. Using topographical attributes to evaluate gully erosion proneness (susceptibility) in two Mediterranean basins: advantages and limitations. Nat Hazards. 79(S1):291–314. https://doi.org/10.1007/s11069-015-1703-0.

Hwang C-L, Yoon K. 1981. Methods for multiple attribute decision making. In: Multiple attribute decision making. Lecture notes in economics and mathematical systems, vol 186. Springer, Berlin, Heidelberg.

James G, Daniela W, Hastie T, Robert T. 2014. An introduction to statistical learning: with applications in R. New York: Springer Publishing Company, p. 426.

Jassim SZ, Goff JC. 2006. Geology of Iraq. Brno: Dolin, Prague and Moravian Museum, p. 431.

Jiang C, Fan W, Yu N, Liu E. 2021. Spatial modeling of gully head erosion on the Loess Plateau using a certainty factor and random forest model. Sci Total Environ. 783:147040. https://www.sciencedirect.com/science/article/pii/S0048969721021100.

Kheir RB, Wilson J, Deng Y. 2007. Use of terrain variables for mapping gully erosion susceptibility in Lebanon. Earth Surf Process Landforms. 32(12):1770–1782. https://doi.org/10.1002/esp.1501.

Kuhn M. 2008. Caret package. J Stat Soft. 28(5):1–26.

Kursa MB, Rudnicki WR. 2010. Feature selection with the Boruta package. J Stat Softw. 36(11):1–13.

Lucà F, Conforti M, Robustelli G. 2011. Comparison of GIS-based gullying susceptibility mapping using bivariate and multivariate statistics: Northern Calabria, South Italy. Geomorphology. 134 (3–4):297–308. https://doi.org/10.1016/j.geomorph.2011.07.006.

Mason SJ, Graham NE. 2002. Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. Q J R Meteorol Soc. 128(584):2145–2166. https://doi.org/10.1256/ 00359000 2320603584.

Moore ID, Burch GJ. 1986. Physical basis of the length slope factor in the Universal Soil Loss Equation. Soil Sci Soc Am J. 50(5):1294–1298.

Moore ID, Grayson RB, Ladson AR. 1991. Digital terrain modeling: a review of hydrological, geomorphological and biological applications. Hydrol Process. 5(1):3–30. https://doi.org/10.1002/hyp.3360050103.

Nanda MA, Seminar KB, Nandika D, Maddu A. 2018. A comparison study of kernel functions in the support vector machine and its application for termite detection. Information. 9(1):5. https://doi.org/10.3390/info9010005.

Pourghasemi HR, Pradhan B, Gokceoglu C, Moezzi KD. 2013. A comparative assessment of prediction capabilities of Dempster-Shafer and Weights-of-evidence models in landslide susceptibility mapping using GIS. Geo Nat Haz Risk. 4(2):93–118. https://doi.org/10.1080/19475705.2012.662915.

Rahmati O, Tahmasebipour N, Haghizadeh A, Pourghasemi HR, Feizizadeh B. 2017. Evaluating the influence of geo-environmental factors on gully erosion in a semi-arid region of Iran: an integrated framework. Sci Total Environ. 579:913–927. https://doi.org/10.1016/j.scitotenv.2016.10.176.

Renard KG, Foster GR, Weesies GA. (Coordinators), 1997. Predicting soil erosion by water: a guide to conservation planning with the revised universal soil loss equation (RUSLE). USDA Agricultural Handbook No. 703, p. 404. Washington D.C.

Rezaei-Moghaddam K, Karami E. 2008. A multiple criteria evaluation of sustainable agricultural development models using AHP. Environ Dev Sustain. 10(4):407–426. https://doi.org/10.1007/s10668-006-9072-1.

Rong L, Duan X, Zhang G, Gu Z, Feng D. 2019. Impacts of tillage practices on ephemeral gully erosion in a dry-hot valley region in southwestern China. Soil Tillage Res. 187:72–84. https://doi.org/10.1016/j.still.2018.11.012.

Saaty TL. 1980. The analytic hierarchy process. New York: McGraw-Hill.

Saaty TL. 2008. Decision making with the analytical hierarchy process. IJSSCI. 1(1):83–98.

Shary PA, Sharaya LS, Mitusov AV. 2002. Fundamental quantitative methods of land surface analysis. Geoderma. 107(1–2):1–32. https://doi.org/10.1016/S0016-7061(01)00136-7.

Shmueli G, Bruce P, Yahav I, Patel N, Lichtendahl K. 2018. Data mining for business analytics: concepts, techniques, and applications in R. John Wiley& Sin Inc, Washinton D.C.

Soria D, Garibaldi JM, Ambrogi F, Biganzoli EM, Ellis IO. 2011. A 'non-parametric' version of the naive Bayes classifier. Knowledge Based Syst. 24(6):775–784.

Sujatha ER, Sridhar V. 2018. Spatial Prediction of Erosion Risk of a Small Mountainous Watershed Using RUSLE: a case-study of the palar sub-watershed in Kodaikanal, South India. Water. 10(11):1608. https://doi.org/10.3390/w10111608.

Sujatha ER, Sridhar V. 2019. Mapping of erosion susceptibility using a weighted linear combination model: a case study of a hill sub-watershed in Kodaikkanal, Western Ghats, South India. Remote Sens Appl Soc Environ. 14:34–45. https://doi.org/10.1016/j.rsase.2019.02.004.

Wilk-Kolodziejczyk D, Regulski K, Gumienny G. 2016. Comparative analysis of the properties of the nodular cast iron with carbides and the austempered ductile iron with use of the machine learning and the support vector machine. G. Int J Adv Manuf Technol. 87 (1–4):1077–1093. https://doi.org/10.1007/s00170-016-8510-y.

Yao X, Tham LG, Dai FC. 2008. Landslide susceptibility mapping based on support vector machine: a case study on natural slopes of Hong Kong, China. Geomorphology. 101(4):572–582. https://doi.org/10.1016/j.geomorph.2008.02.011.

Zabihi M, Mirchooli F, Motevalli A, Darvishan AK, Pourghasemi HR, Zakeri MA, Sadighi F. 2018. Spatial modelling of gully erosion in Mazandaran Province, northern Iran. Catena. 161:1–13.

Zhang H, 2004. and, Su J. Naive Bayesian classifiers for ranking. Boulicaut JF, Esposito F, Giannotti F, Pedreschi D (eds) Machine learning: ECML 2004. ECML 2004. Lecture notes in computer science, vol 3201. Berlin, Heidelberg: Springer.

Appendix 1

Confusion matrix evaluation measures used in this study

In this appendix, the definition and mathematical formulas use for ML performance evaluation are summarized.

Accuracy is the proportion of instances (gullies and non-gullies) that are predicted correctly by the model. It is calculated via the following formula: $\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$; where TP is the number of cases that predicted correctly as gullies, FN is the number of instances that

A. J. AL-BAWI ET AL.

predicted incorrectly as gullies, TN is the number of instances that predicted correctly as non-gullies, and FP is the number of instances that predicted incorrectly as non-gullies. Sensitivity (named as Recall too) is the proportion of positive instances that are predicted correctly and is calculated as: Sensitivity $= \frac{TP}{TP+FN}$. Specificity is the proportion of negative instances that are predicted correctly: Specificity $= \frac{TN}{TN+FP}$. Precision is defined as the ratio of correctly predicted instances to the total predicted positive instances: Precision $= \frac{TP}{TP+FP}$. Finally, the F1 score is the weighted average of Precision and Sensitivity and is computed as: F1 score $= \frac{2\times(\text{Sensitivity}\times\text{Precision})}{\text{Sensitivity}+\text{Precision}}$. For all the above evaluation measures, the model is considered performed well when the value is near to unity.