Natural Hazards (2024) 120:7211–7244
https://doi.org/10.1007/s11069-024-06481-9

ORIGINAL PAPER

# Predictive machine learning for gully susceptibility modeling with geo-environmental covariates: main drivers, model performance, and computational efficiency

**Kwanele Phinzi**[1] · **Szilárd Szabó**[2]

Received: 7 February 2023 / Accepted: 29 January 2024 / Published online: 12 March 2024


## Abstract

Currently, machine learning (ML) based gully susceptibility prediction is a rapidly expanding research area. However, when assessing the predictive performance of ML models, previous research frequently overlooked the critical component of computational efficiency in favor of accuracy. This study aimed to evaluate and compare the predictive performance of six commonly used algorithms in gully susceptibility modeling. Artificial neural networks (ANN), partial least squares, regularized discriminant analysis, random forest (RF), stochastic gradient boosting, and support vector machine (SVM) were applied. The comparison was conducted under three scenarios of input feature set sizes: small (six features), medium (twelve features), and large (sixteen features). Results indicated that SVM was the most efficient algorithm with a medium-sized feature set, outperforming other algorithms across all overall accuracy (OA) metrics (OA = 0.898, $F1$-score = 0.897) and required a relatively short computation time (< 1 min). Conversely, ensemble-based algorithms, mainly RF, required a larger feature set to reach optimal accuracy and were computationally demanding, taking about 15 min to compute. ANN also showed sensitivity to the number of input features, but unlike RF, its accuracy consistently decreased with larger feature sets. Among geo-environmental covariates, NDVI, followed by elevation, TWI, population density, SPI, and LULC, were critical for gully susceptibility modeling. Therefore, using SVM and involving these covariates in gully susceptibility modeling in similar environmental settings is strongly suggested to ensure higher accuracy and minimal computation time.

**Keywords** Gully erosion · Machine learning · Predictive modeling · Accuracy · Computational efficiency · Geo-environmental predictors

✉ Kwanele Phinzi
kwanelep48634@gmail.com

[1] Department of Geography and Environmental Studies, University of Zululand, KwaDlangezwa 3886, South Africa

[2] Department of Physical Geography and Geoinformatics, Faculty of Sciences and Technology, University of Debrecen, Debrecen 4032, Hungary

## 1 Introduction

Gully erosion is the most pressing environmental issue driving landscape and soil degradation worldwide (Poesen et al. 2003; Valentin et al. 2005; Mararakanye and Le Roux 2012; Bennett and Wells 2019). Gullies not only reduce soil fertility (Pimentel et al. 1995), potentially leading to food insecurity, but also pose a significant challenge to the sustainable management of ecosystems across the globe (Poesen et al. 2003; Magliulo 2012; Gayen and Pourghasemi 2019; Hitouri et al. 2022). The global costs associated with soil erosion and subsequent sedimentation deposits, of which gullies account for the most, amount to $400 billion per year (Pimentel et al. 1995; Pimentel 2006). At the catchment scale, gullies contribute about 10–95% of total sediment losses, although they occupy a small proportion (less than 5%) of the landscape (Poesen et al. 2003; Roberts et al. 2022). Thus, predicting areas susceptible to gully erosion is required for targeted interventions to reduce the negative impacts of soil erosion at the catchment scale.

Although gully erosion research spans over a century (Castillo and Gómez 2016; Liu et al. 2021), gully modeling is still in its infancy compared to rill and sheet erosion (Roberts et al. 2022), with well-established and widely adopted models worldwide, such as the universal soil loss equation (USLE) (Wischmeier and Smith 1978) and its derivatives, the revised USLE (RUSLE) (Renard et al. 1997) and modified USLE (MUSLE) (Williams and Berndt 1977). Besides, existing gully erosion models designed to simulate sediment yield, gully development, morphology, and head-cut retreat rates have not been widely used globally due to their geographic specificity, the complexity of their physical processes, and challenges in obtaining observation data (Balogh and Novák 2020; Roberts et al. 2022). Traditional methods for assessing gully erosion involve the use of various devices, such as microtopographic profilers, rulers, total stations, tapes, and poles (Capra and Scicolone 2002). These methods determine gully volume, cross-sectional areas, and length of gully reaches with high accuracy (Castillo et al. 2012) but are often expensive and limited to accessible locations. This hampers a precise understanding of gully susceptibility at catchment scales (Vrieling et al. 2010). Similarly, manual digitization and interpretation of gullies from aerial or satellite images is time-consuming and potentially subjective, which impedes repeatability. High-resolution digital surface models derived from terrestrial Light Detection and Ranging (LiDAR) technology hold the potential for in-depth gully assessment (Goodwin et al. 2017). However, the associated survey or data collection costs make them unsuitable for large-scale catchment studies. A reasonable alternative overcoming these limitations is the use of data-driven techniques such as machine learning (ML).

Although ML was introduced in the 1990s, it became popular for gully erosion assessment only after 2010 (Svoray et al. 2012; Lana et al. 2022), and its applications to gully erosion have since increased, even more so in the last three years. A literature search in Scopus (in February 2023) using keywords like “gully” and “machine learning” indicated that 294 research papers were published between 2012 and 2022, of which more than 50% were published in the last 3 years (2020–2022). This significant increase in the use of ML in gully erosion can be attributed to the increased availability of free geospatial data, a prerequisite for building predictive ML models.

In gully erosion, ML models are primarily used to automatically extract or detect gullies from airborne and space-borne remotely-sensed data using either pixel or object-based methods (Shruthi et al. 2011; Mararakanye and Nethengwe 2012; Gafurov and Yermolayev 2020; Phinzi et al. 2020, 2021). With additional spatial detail, high-resolution sensors permit accurate extraction of small and narrow gully features, although their high acquisition

costs inhibit large-scale mapping. More recently, ML has been increasingly applied to gully susceptibility modeling at catchment scales (Dewitte et al. 2015; Pourghasemi et al. 2017; Arabameri et al. 2019; Huang et al. 2022; Lana et al. 2022; Kulimushi et al. 2023). This approach concerns mapping areas with varying degrees (i.e., low, medium, high, and very high) of gully vulnerability and analyzing the relationship between gully occurrence and spatial variability of numerous geo-environmental predictors (Conoscenti and Rotigliano 2020).

Various ML methods are applied and evaluated to select the best model with the highest predictive accuracy for gully susceptibility mapping. For example, Gayen et al. (2019) mapped gully susceptibility using four ML methods: multivariate additive regression spline (MARS), flexible discriminant analysis (FDA), random forest (RF), and support vector machines (SVM) and reported the highest prediction accuracy for RF based on the area under the curve (AUC = 96.2%). Using artificial neural network (ANN), convolutional neural network (CNN), and deep neural network (DNN), Chowdhuri et al. (2021) mapped gully susceptibility, and the results showed that DNN outperformed other methods with 95.8% AUC. Hosseinalizadeh et al. (2019) predicted gully headcut susceptibility with functional trees (FT), naïve bayes (NBTree), and RF models and found RF to be the most efficient model with an AUC of 96.5%. Huang et al. (2022) applied RF, SVM, ANN, and a generalized linear model (GLM) to gully susceptibility, and RF yielded the best predictive accuracy (AUC = 90.5%). Most of these studies, however, focus on the accuracy and not computational efficiency, another critical aspect of performance evaluation in predictive ML. A necessary step during the model building process is to fine-tune relevant hyperparameters for optimal model performance. A typical approach involves exhaustively searching for the appropriate hyperparameter combination through cross-validation, a computationally demanding procedure, especially when fine-tuning multiple parameters. For this reason, an efficient ML model should maximize accuracy and minimize computation time (Gislason et al. 2006; Belgiu and Drăgu 2016). So far, more emphasis is being placed on the latter when assessing the performance of ML models. Furthermore, most gully susceptibility studies often use a fixed set of predictor variables (Conoscenti et al. 2013; Bernini et al. 2021; Chowdhuri et al. 2021; Lana et al. 2022), hampering insights into how different ML models might perform (in terms of accuracy and computation time) when smaller, medium, and larger subsets of predictors are used, which is a significant research gap.

We hypothesize that predictor subsets of varying sizes (e.g.,, small, medium, and large) influence the prediction accuracy and computation time of different ML models. To test this hypothesis, six different commonly used ML models, including ANN, partial least squares (PLS), regularized discriminant analysis (RDA), RF, stochastic gradient boosting (SGB), and SVM, were compared. The objectives of this study are threefold: (i) to select geo-environmental variables with the greatest predictive power to model gully susceptibility, (ii) to evaluate the performance of ML models under different scenarios of input feature set sizes, and (iii) to develop an optimal model that is fast with superior predictive performance.

## 2 Study area

The study area is confined to a tertiary catchment (T12) in the Eastern Cape Province of South Africa. The catchment lies between latitude 31° 29′ 38″ S and 32° 02′ 24″ S and longitude 27° 25′ 01″ E and 28° 29′ 34″ E, covering a surface area of about 2145 km$^2$ with

an elevation range of 534–1772 m above sea level (Fig. 1). It is drained by the Mgwali, Qumanco, and Tora Rivers, which are the tributaries of the Mbhashe River system, one of the major rivers in the country. The climate is semi-arid, and most rainfall occurs in summer, with a mean annual precipitation of approximately 800 mm (DWA 2010). Such climatic conditions favor the growth of natural vegetation and support agricultural activities. Grassland, consisting of indigenous grass species, is the dominant natural vegetation type, with some forest patches scattered throughout the catchment. Agriculture characterizes land use activity and mainly includes subsistence crops (e.g.,, maize) and livestock (e.g.,, cattle) farming, although there are few commercial farms. Mudstones and sandstones of the Tarkastad Formation and Molteno Formation underlie the area. Due to the easily weathered parent material of these geological rock types, soils in this area are inherently susceptible to erosion (DWA 2010), which explains widespread soil erosion, particularly gully erosion. Classical gullies, also referred to as permanent gullies exhibiting V-shape and U-shape characteristics, are the prevalent type of erosion and are commonly located in the foot slopes across the entire catchment, where unconsolidated soil material from the hillslope accumulates. Although most gullies are typically found in gently sloping grasslands and abandoned agricultural lands, certain gullies are observed in elevated hillslope areas characterized by poor vegetation cover. Examples of gullies in the study area captured during the field survey are depicted in Fig. 2a and b, while Fig. 2c and d provide an aerial view from Google Earth images, depicting some of the most extensive gully systems. Chromic Luvisols, whose top layer has a high silt content (ISRIC 2002), are extensive in gently sloping and nearly flat areas, covering about 87% of the catchment. Eutric Planosols (6%) and Solodic Planasols (7%), commonly found in hilly areas, cover the rest of the catchment. Planosols are highly unstable soils, making them vulnerable to gully erosion (Du Plessis et al. 2020).

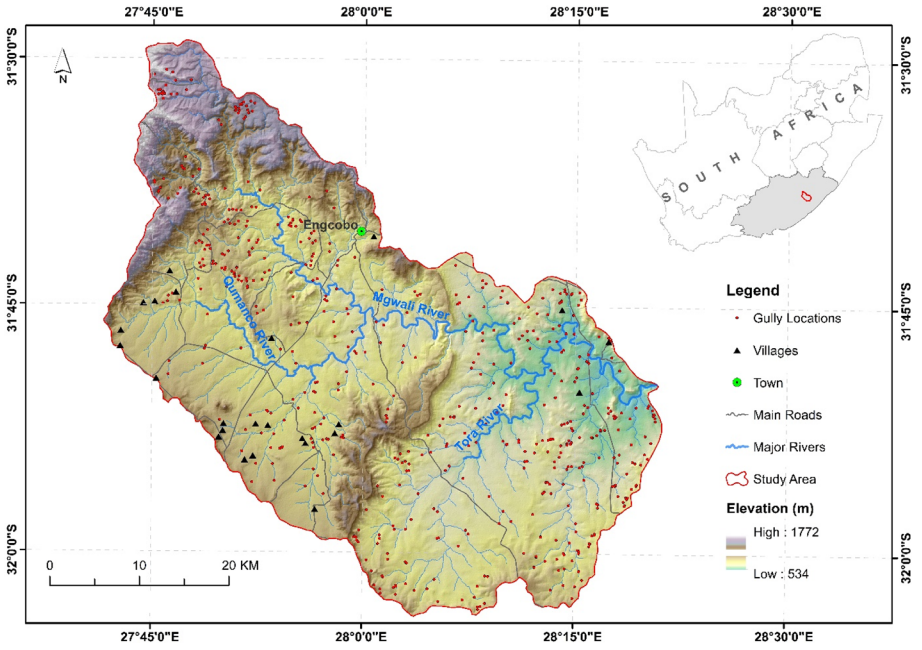



**Fig. 1** Study area map showing the distribution of gully locations (data source: shuttle radar topography mission—SRTM DEM)

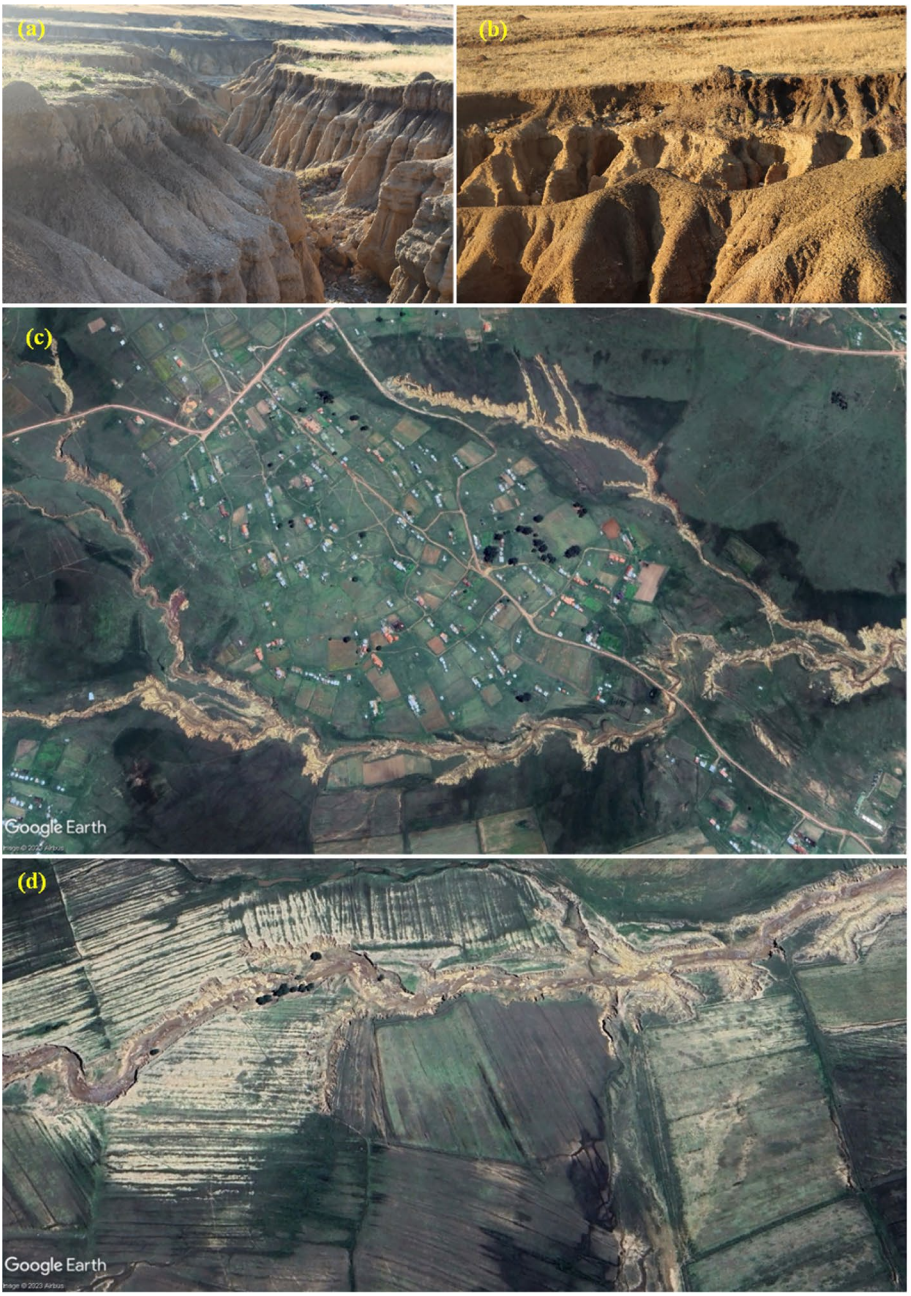



**Fig. 2** Some examples of gullies observed in the study area: **a** and **b** are field photos captured in August 2021, and **c** and **d** are Google earth images

## 3 Materials and methods

This section details the methods and data used to achieve the objectives of this study. Figure 3 summarizes the procedure followed for gully susceptibility modeling. All ML-related experiments and analyses were executed in R 4.2.2 software (R Core Team 2021), while ArcGIS 10.4 (ESRI 2022) was used for mapping and computing geo-environmental

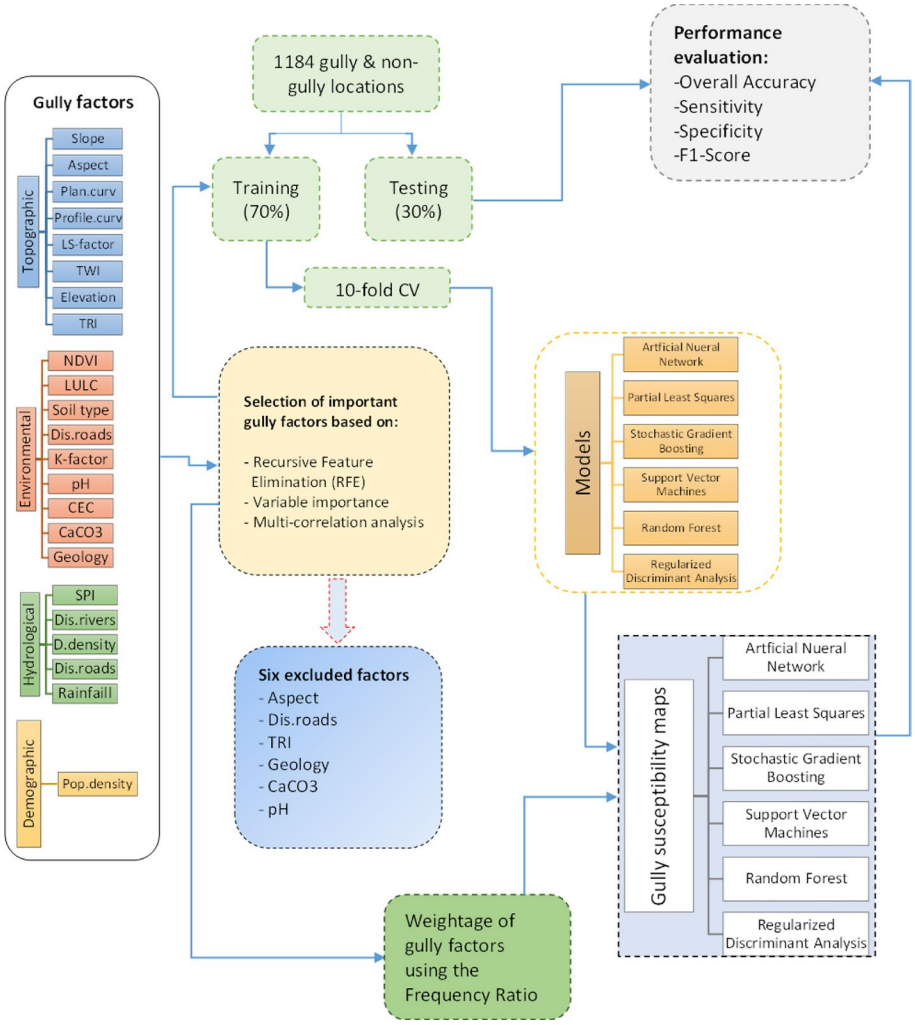



**Fig. 3** Flowchart summarizing the methodology utilized in this study (VIF = variable inflation factor, CV = cross-validation, TRI = topographic ruggedness index, Plan.curv = plan curvature, Profile.curv = profile curvature, LS = slope length and steepness, TWI = topographic wetness index, NDVI = normalized difference index, LULC = land use/land cover, Dis.roads = distance from roads, CEC = cation exchange capacity, SPI = stream power index, Dis.rivers = distance from rivers, D.density = drainage density, Pop.density = population density)

predictors. All the spatial data were projected to the Universal Traverse Mercator (UTM) zone 35 South, referenced to the world datum (WGS84). A standard spatial resolution of 30 m was used for raster data.

### 3.1 Geo-environmental predictors

One of the objectives of this study was to identify the most important geo-environmental variables for gully susceptibility modeling. To achieve this, we initially compiled a set of 22 variables encompassing topographic, environmental, hydrological, and demographic factors (Table 1). The spatial distributions of these geo-environmental variables are shown in the Appendix (Fig. 10). The selection of these variables was based on their potential influence, data availability, and their utilization in previous studies (Valentin et al. 2005; Le Roux and Sumner 2012; Conoscenti et al. 2014; Phinzi and Ngetar 2019b; Ghaedi and Shojaian 2020; Ebhuoma et al. 2022). Elevation data for this study was sourced from the void-filled Shuttle Radar Topography Mission (SRTM) DEM, downloaded from the United States Geological Survey (USGS). In addition, slope, aspect, plan curvature, profile curvature, slope length and steepness (LS-factor) (Moore and Burch 1986), topographic wetness index (TWI), and terrain ruggedness index (TRI), stream power index (SPI) (Moore et al. 1991), distance from rivers, and drainage density were also derived from the SRTM DEM.

Soil data consisting of soil physical (organic matter content, sand, silt, and clay content) and chemical properties such as cation exchange capacity (CEC), calcium carbonate ($CaCO_3$), and pH were obtained from the digital soil map of the world (FAO 2003). Soil erodibility, represented by the *K*-factor, was computed from the soil's physical properties using the empirical relation of Williams (1995). The normalized difference vegetation index (NDVI), a proxy for vegetation, was computed from a single date, cloud-free Landsat-9 Operational Land Imager (OLI) acquired on 08 February 2022. The Landsat image was downloaded from the USGS website. The land use/land cover (LULC) map for the study area was extracted from the South African National Land Cover (SANLC) dataset, available on the Department of Forestry, Fisheries, and Environment website. The study has eight LULC classes, including built-up land, barren land, cultivated land, forest, grassland, mines and quarries, water bodies, and wetlands.

The land type map was prepared from the South African Land Type Survey database (Land Type Survey Staff). The map comprises nine broad land types grouped according to the prevailing climate, terrain, and dominant soil types found within the land type (Van Zijl et al. 2013; Du Plessis et al. 2020). Distance from roads was computed from the road network data available in vector format. Geology for this study consisting of Mudstones and sandstones of the Tarkastad Formation and Molteno Formation was downloaded from the South African National Space Agency (SANSA). Long-term (1981–2021) annual gridded rainfall data from the Climate Hazards Group Infrared Precipitation (CHIRPS) were used (Funk et al. 2015). The CHIRPS product provides quasi-global rainfall estimates at a spatial resolution of 0.05° and was resampled to 30 m. Finally, the 2020 population density dataset was downloaded from the WorldPop database in Geotiff format at a spatial resolution of 100 m.

**Table 1** Geo-environmental predictors considered for gully susceptibility modeling

| Geo-environmental predictors | | Class range | Spatial resolution* | Data source |
|---|---|---|---|---|
| Topographical | Slope | 0–67.07° | 30 m | SRTM DEM (https://earthexplorer.usgs.gov/) |
| | Aspect | 9 classes | 30 m | SRTM DEM (https://earthexplorer.usgs.gov/) |
| | Plan curvature | − 9–8 | 30 m | SRTM DEM (https://earthexplorer.usgs.gov/) |
| | Profile curvature | − 10–1 | 30 m | SRTM DEM (https://earthexplorer.usgs.gov/) |
| | LS-factor | 0–3228.69 | 30 m | SRTM DEM (https://earthexplorer.usgs.gov/) |
| | TWI | 2.73–24.99 | 30 m | SRTM DEM (https://earthexplorer.usgs.gov/) |
| | Elevation | 534–1772 m | 30 m | SRTM DEM (https://earthexplorer.usgs.gov/) |
| | TRI | 0.89–0.11 | 30 m | SRTM DEM (https://earthexplorer.usgs.gov/) |
| Environmental | NDVI | − 0.23–0.72 | 30 m | Landsat-9 (https://earthexplorer.usgs.gov/) |
| | LULC | 8 classes | 20 m | Department of forestry, fisheries and environment (https://egis.environment.gov.za) |
| | Land type | 9 classes | – | Land type survey staff (1972–2006) |
| | Distance from roads | 0–12647.2 m | – | Roads shapefile (https://www.hotosm.org/) |
| | *K*-factor | 0.06–0.12 | – | FAO world soil database |
| | Soil pH | 6.2–6.9 | – | FAO world soil database |
| | CEC | 8.4–13.1 cmol/kg | – | FAO world soil database |
| | CaCO3 | 0–1 | – | FAO world soil database |
| | Geology | 2 classes | – | South African National Space Agency (http://atlas.sansa.org.za/atlas-geology.html) |
| Hydrological | SPI | − 13.82–13.35 | 30 m | SRTM DEM (https://earthexplorer.usgs.gov/) |
| | Distance from rivers | 0–3079.5 m | 30 m | SRTM DEM (https://earthexplorer.usgs.gov/) |
| | Drainage density | 0–4.51 km/km$^2$ | 30 m | SRTM DEM (https://earthexplorer.usgs.gov/) |
| | Rainfall | 651.94–766.19 mm | 0.05° | CHIRPS (Funk et al. 2015) |
| Demographic | Population density | 0–85 people/km$^2$ | 100 m | WorldPop database (www.worldpop.org) |

*SRTM* shuttle radar topography mission, *DEM* digital elevation model, *TWI* topographic wetness index, *LS* slope length and steepness, *TRI* terrain ruggedness index, *NDVI* normalized difference vegetation index, *LULC* land use/land cover, *CEC* cation exchange capacity, $CaCO_3$ calcium carbonate, *SPI* stream power index

*The dashed (–) line indicates that the data were initially collected in polygon form and subsequently converted into raster format. All datasets underwent resampling to ensure a common spatial resolution of 30 m

### 3.2 Variable selection

Selecting an optimal subset of predictors is a critical step in the modeling process required to avoid overfitting and multicollinearity, thereby improving predictive performance. Consequently, the recursive feature elimination (RFE) algorithm was used to remove non-informative predictors from the original dataset through iteration. A random forest-based RFE was applied with a tenfold cross-validation, repeated five times. Initially, the algorithm considers all predictors and eliminates non-informative predictors (one at a time) in each iteration until an optimal subset of predictors is attained (Csatáriné Szabó et al. 2020; Varga et al. 2021). Although RFE generates an optimal feature set, it does not show bivariate correlations of predictors. For this reason, the correlation matrix, tolerance ($1-r^2$), and variance inflation factor (VIF = $1/(1-r^2)$) were also computed to detect multicollinearity. Multicollinearity was considered to exist if bivariate correlations among predictors exceeded 0.75, tolerance was < 0.1, and VIF was > 10 (Mason and Perreault Jr 1991; Gareth et al. 2013; Kuhn and Johnson 2013; Vatcheva et al. 2016). If collinearity was detected, the variables concerned were removed, resulting in an optimal feature set of sixteen predictors. The predictors were then ranked in their relative importance in gully prediction using RF-based variable importance. We generated three feature sets (large set = 16 features, medium set = 12 features, and small set = 6 features) from this optimal feature set to investigate the performance of ML models when the number of predictors is varied. The small feature set comprised the top six most important predictors, the medium set comprised the top twelve important predictors (i.e., small set + six additional predictors), and the large set comprised the complete set of predictors (i.e., all sixteen variables). For convenience, we refer to these feature sets as large, medium, and small sets throughout the manuscript. Partial dependence plots (PDPs) were employed to further examine the variables identified as most important in the modeling of gully susceptibility. Friedman (2001) introduced PDPs to depict the marginal effect of one or two predictor variables on the predicted outcome. A positive *y*-axis value indicates that for that specific value of a predictor variable, the model is likely to predict the correct class for that observation, while a negative value suggests a negative impact on predicting the correct class (Friedman 2001). The analysis was carried out to determine whether the considered variables had a positive or negative impact on the prediction of gully erosion. PDPs were executed in the R software with the "pdp" (Greenwell 2017) and "RandomForest" (Liaw and Wiener 2002) packages.

## 4 Reference data collection

Reference data comprising training and testing sets are central to supervised ML. Following a field survey of gully erosion hotspots and visual interpretation of high-resolution Google Earth images, 592 gully points were randomly collected from the gullies that were large enough and distinguishable. Gullies within the study area are extensive and complex, forming narrow linear and dendritic patterns with a dense network of smaller gullies. Due to this complexity, gully locations were mainly collected at gully heads and in the middle part of a gully in the case of long narrow linear gullies with limited branching or visible gully heads. Most gullies in the area are permanent and exhibit varying morphological characteristics such as lengths (30–274 m), depths (1.22–6.90 m), and widths (4.66–15 m), with primarily V-shaped and U-shaped cross-sections (Phinzi et al. 2021). Despite this

complexity, their relatively large size and lack of vegetation cover made it easy to locate gullies in the field and on Google Earth.

Similarly, 592 non-gully points were collected in locations where a different land use/cover (e.g.,, forest, built-up area, roads, bare soil, and grassland) existed other than a gully. Finally, gully and non-gully points from Google Earth were converted into a vector shapefile, which was exported to the R software for extracting pixel values from each geo-environmental predictor. In essence, reference data had 1184 observations with 16 predictors (large set) and a response variable (e.g.,, gully occurrence), which is binary (g = gully and ng = non-gully). Reference data were randomly partitioned into training (70%) and testing (30%) data. This procedure was repeated for the small and medium sets.

### 4.1 Gully susceptibility modeling

A two-step process was followed for gully susceptibility modeling using six ML techniques (ANN, PLS, RDA, RF, SGB, and SVM) and a bivariate statistical method (frequency ratio). First, we used the caret package (Kuhn 2008) to train all ML algorithm through a tenfold cross-validation process with five repetitions, resulting in 50 candidate models for each ML algorithm. In this phase, model optimization and selection were carried out based on a combination of hyper-parameters that maximized model performance, as illustrated in the Appendix (Fig. 11). Next, a final model with the highest accuracy was selected and applied to predict gully erosion at a pixel level, resulting in a binary output (g and ng pixels) for each ML method. Second, the frequency ratio (FR) was used to assign weights to each geo-environmental predictor and reclassify the binary output of each ML algorithm into four gully susceptibility classes (e.g.,, low, moderate, high, and very high). The resulting gully susceptibility maps were generated with a spatial resolution of 30 m. The following subsections briefly describe each ML method, including relevant hyper-parameters that were tuned and how the FR was utilized. To delve deeper into the specifics of each ML algorithm, the reader is encouraged to consult the references provided for each respective algorithm.

#### 4.1.1 Artificial neural network (ANN)

ANN consists of layers with artificial neurons that mimic the way biological neurons function in the human brain. An input layer of neurons (also called nodes or units), one or more hidden layers of neurons, and a final layer of output neurons define a typical architecture of an ANN. Neurons in each layer are connected to other neurons in the next layer, and each connection is associated with a specific weight (Wang 2003). These weights determine the importance of each geo-environmental predictor in gully prediction, where predictors with larger weights contribute considerably to gully prediction. ANN can implicitly recognize complex nonlinear relationships between the response variable and predictors, simultaneously identifying all possible interactions among the predictor variables (Tu 1996). Several packages are available in R software to implement ANN. In this study, ANN was performed using the "nnet" (Venables and Ripley 2013) function in the caret package (Kuhn 2008). A grid search of two hyper-parameters, including the number of neurons in the hidden layer (size = 5, 10, 15) and a regularization parameter to avoid over-fitting (decay = 0.001, 0.01, 0.1), was defined to find the best combination of values that maximizes the predictive performance.

#### 4.1.2 Partial least squares (PLS)

Initially designed for dimension reduction (Wold 1966), PLS is becoming popular for solving classification problems, including gully susceptibility (Pham et al. 2020; Pourghasemi et al. 2020). PLS intends to form components that capture most of the information in the explanatory variables helpful in predicting the response variable (Garthwaite 1994). It achieves this by constructing linear combinations (components) of the original predictors from which a set of latent variables with the best predictive power is extracted (Abdi 2003), then regressing the response variable on these latent variables (Chung and Keles 2010). PLS only has one hyper-parameter, the number of components. A tune length of 15 was selected to find the number of components with the highest accuracy. PLS was executed with the "pls" (Wehrens and Mevik 2007) function in the caret package.

#### 4.1.3 Regularized discriminant analysis (RDA)

Discriminant analysis (DA) is a generative ML approach that uses Bayes' theorem to model the conditional distribution of the predictors $\boldsymbol{X}$ in each of the predefined response classes $\boldsymbol{Y}=1$ (i.e., gully) and $\boldsymbol{Y}=0$ (i.e., non-gully), then estimates the probability of $\boldsymbol{Y}$ given the value of $\boldsymbol{X}$ (Welch 1939; Friedman 1989; Gareth et al. 2013). DA operates by finding one or more linear combinations of predictors that best discriminate or separate the response classes (Alkarkhi and Alqaraghuli 2018). New observations can be predicted and assigned to predefined response classes using linear or quadratic discriminant functions. RDA is a regularization method derived from linear DA (LDA) and quadratic DA (QDA). Thus, one would expect the RDA to perform better than its predecessors (Wu et al. 1996), although they all rely on the same assumption about data distribution (e.g.,, assume a multivariate normal distribution of data). Furthermore, an advantage of RDA over its predecessors is that it has tuning parameters, including the balance between LDA and QDA (gamma) and the correlation of predictors (lambda), making it a robust classifier. We fine-tuned these parameters using a tune length of 5. The "rda" (Friedman 1989) function in the caret package was used to execute RDA.

#### 4.1.4 Random forest (RF)

Introduced by Breiman (2001), RF is a popular ML method that addresses the shortcomings of decision trees by ensembling randomly numerous decision trees to improve predictive performance. RF accomplishes this by bootstrap or begging aggregation, where multiple predictors are generated using classification trees. The predictions from all these trees are then averaged to form a final prediction. The bootstrap component induces randomness, ensuring that the individual trees are different (Irizarry 2019). First, the algorithm generates small subsets of data (called bootstrap samples) by randomly sampling several observations from the training data with replacement. These bootstrap samples are used to train many random decision trees separately. The remaining observations, known as "out-of-bag" samples that have not been fed as training data to these decision trees, are used for evaluation (Abdi 2020). Second, the algorithm randomly selects a subset of features for each decision tree, minimizing the risk of overfitting while improving predictive accuracy. These features make RF a robust classifier and a good candidate for solving classification problems like gully susceptibility, where several predictors are involved. Additionally, RF offers several advantages (Rodriguez-Galiano et al. 2012): including its non-parametric

nature, high accuracy in classification, and the ability to assess variable importance. It is also flexible to conduct various types of data analyses, encompassing regression, supervised and unsupervised classifications (Rodriguez-Galiano et al. 2012). The "rf" function from the caret package (Kuhn 2008) was used to perform RF classification. The grid search method was used for tuning the mtry parameter, the number of trees at each split (mtry = 20).

#### 4.1.5 Stochastic gradient boosting (SGB)

Boosting was traditionally developed for classification problems (Valiant 1984) and concerned with combining several weak classifiers to form a robust classifier (Kuhn and Johnson 2013). There are many variants of boosting algorithms, and SGB (Friedman 2002), also known as gradient boosting machines (GBM), is among the most recent and popular algorithms. Much like RF, SGB uses bagging, a technique where a set of random decision trees are generated, and each tree is trained on a random subset of the training data. The main difference is that SGB generates an ensemble of several shallow trees sequentially, where each tree learns and improves on the previous one, while RF generates an ensemble of deep independent trees (Boehmke and Greenwell 2019). SGB uses a sequential ensemble approach in which boosting starts with a weak model and sequentially boosts its performance by building a new tree at each iteration from a random subsample of the training set, improving the model's prediction accuracy (Moisen et al. 2006; Boehmke and Greenwell 2019). SGB was implemented through the caret package using the "gbm" (Ridgeway 2007) function. Hyper-parameters include the depth of each tree (interaction depth = 1, 5, 9), the number of trees ($n$ trees = 1500), the learning rate of the algorithm (shrinkage = 0.1), and the minimum number of observations for the trees terminal node ($n$ minobsinode = 20). A grid search method was used to find a combination of these parameters with the highest accuracy.

#### 4.1.6 Support vector machines (SVM)

SVM is a binary classifier that originates in statistical learning theory (Vapnik 1999), capable of solving two-class and multiclass classification problems. The objective of SVM is to find a hyperplane (a decision boundary) that best separates the two classes of data points (called support vectors) in a feature space. It achieves this objective using the kernel trick, which expands the feature space to the point (margin) where the classes are perfectly separable. Furthermore, SVM uses various kernel functions, making it highly flexible for finding even complex non-linear boundaries (Boehmke and Greenwell 2019). In addition, SVM can simplify complex problems and achieve effective classification performance when applied to real-world scenarios (Hearst et al. 1998). The algorithm has two important hyper-parameters: the cost (C) parameter that penalizes misclassified data points (i.e., samples that lie on the wrong side of the decision boundary) and the sigma parameter that controls the SVM decision boundary. SVM was performed with the "svmRadial" function in the caret package (Kuhn 2008), which had the highest accuracy compared to other kernel functions (linear, polynomial, and sigmoid). A grid search of C (2, 4, 6, 8, 10) and sigma (0.01, 0.02, 0.04, 0.06, 0.08, 0.10) was generated to find the optimal combination of these parameter values.

#### 4.1.7 Frequency ratio (FR)

FR is a bivariate statistical technique successfully applied to various natural hazards such as flooding (Rahmati et al. 2016; Shafapour Tehrany et al. 2019), landslides (Lee and Sambath 2006; Anbalagan et al. 2015), and, more recently, gully erosion (Roy and Saha 2019; Amare et al. 2021; Azedou et al. 2021; Lana et al. 2022). It expresses the ratio between the occurrence and non-occurrence of a natural hazard (in this case, gully erosion) based on its spatial relationship with associated influencing factors (i.e., gully predictors) (Lee and Pradhan 2007). Similar to conditional probability, an FR ratio of $>1$ represents a strong relationship, while a ratio of $<1$ represents a weak relationship between gullies and predictor classes (Anbalagan et al. 2015). In this study, the FR was utilized to assign weights to each geo-environmental predictor and reclassify ML-derived gully maps into four gully susceptibility classes (e.g.,, low, moderate, high, and very high).

### 4.2 Model performance evaluation

The confusion matrix was used to evaluate the performance of ML algorithms based on testing data. It cross-tabulates predictions against actual values using four possible outcomes (Irizarry 2019): true positive (TP), false negative (FN), false positive (FP), and true negatives (TN). Various statistical metrics were quantified from these outcomes (Table 2), including the overall proportion of correctly classified cases (overall accuracy), the proportion of actual positives (sensitivity, also called recall), the proportion of negatives (specificity), the proportion of positive outcomes called positive that are positive (precision), and the harmonic average of precision and recall ($F1$-score).

## 5 Results

### 5.1 Variable selection and multicollinearity analysis

Variables were first ranked in their relative importance in the modeling process. Results indicated that NDVI (importance = 100%), elevation (26%), TWI (25%), population density (23%), SPI (22%), and LULC (19%) had considerable predictive power for gully susceptibility modeling (Fig. 4a). The least important variables, including aspect (0%), geology (0.20%), TRI (0.27%), and distance from roads (3.83%), were removed by the RFE algorithm, retaining 18 variables (out of 22). The applied algorithm showed no improvement in accuracy after the 18th variable (Fig. 4b). This implied that the four predictors following the 18th variable did not exert any influence on predictive accuracy. Although calcium

**Table 2** Statistical metrics derived from the confusion matrix (TP = true positive, FN = false negative, FP = false positive, and FN = false negative)

| Metric | Formula |
|---|---|
| Overall accuracy (OA) | $\frac{TP+TN}{TP+TN+FN}$ |
| Sensitivity (recall) | $\frac{TP}{TP+FN}$ |
| Specificity | $\frac{TN}{TN+FP}$ |
| Precision | $\frac{TP}{TP+FP}$ |
| $F1$-score | $2 \times \frac{precision \times recall}{precision+recall}$ |

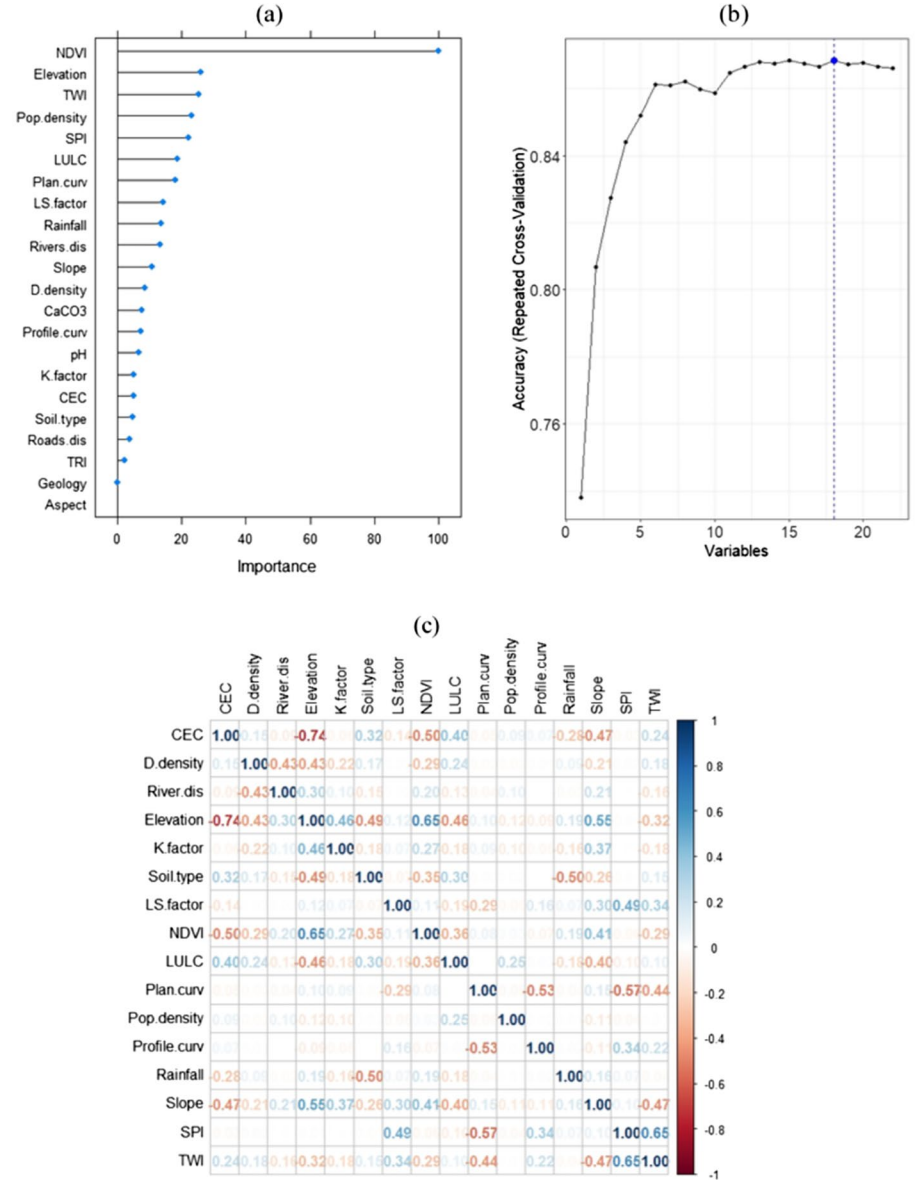



**Fig. 4** Ranking and selection of important predictors based on: **a** variable importance, **b** recursive feature elimination (the blue-dashed line marks the eighteenth variable), and **c** correlation matrix. *Note* CEC = cation exchange capacity; D.density = drainage density; River.dis = distance from rivers; Plan.curv = plan curvature; Pop.density = population density; Profile.curv = profile curvature; SPI = stream power index; TWI = topographic wetness index

carbonate ($CaCO_3$) and pH were part of the retained predictors, they had a high correlation ($r^2 > 0.75$) with other predictors and hence were also removed to avoid multicollinearity, resulting in 16 predictors (Fig. 4c). These predictors were grouped into three feature sets (small, medium, and large sets) based on their relative importance, as detailed in Sect. 3.2.

**Table 3** Tolerance (Tol) and variance inflation factor (VIF) for geo-environmental predictors in three feature sets (large, medium, and small sets)

| Large set | | | Medium set | | | Small set | | |
|---|---|---|---|---|---|---|---|---|
| Variables | Tol | VIF | Variables | Tol | VIF | Variables | Tol | VIF |
| Population density | 0.87 | 1.16 | Rainfall | 0.90 | 1.2 | Population density | 0.92 | 1.09 |
| Distance from rivers | 0.75 | 1.33 | Population density | 0.88 | 1.13 | LULC | 0.73 | 1.37 |
| Profile curvature | 0.70 | 1.42 | Distance from rivers | 0.79 | 1.27 | NDVI | 0.55 | 1.80 |
| LULC | 0.67 | 1.49 | LULC | 0.68 | 1.47 | SPI | 0.52 | 1.92 |
| LS factor | 0.63 | 1.59 | Drainage density | 0.68 | 1.47 | Elevation | 0.48 | 2.07 |
| Rainfall | 0.63 | 1.59 | LS factor | 0.66 | 1.52 | TWI | 0.46 | 2.17 |
| Drainage density | 0.62 | 1.60 | Plan curvature | 0.59 | 1.69 | | | |
| Land type | 0.57 | 1.76 | NDVI | 0.52 | 1.92 | | | |
| K factor | 0.54 | 1.86 | Elevation | 0.42 | 2.37 | | | |
| NDVI | 0.51 | 1.94 | Slope | 0.31 | 3.21 | | | |
| Plan curvature | 0.50 | 1.98 | SPI | 0.29 | 3.50 | | | |
| CEC | 0.32 | 3.14 | TWI | 0.25 | 3.96 | | | |
| Slope | 0.30 | 3.39 | | | | | | |
| SPI | 0.26 | 3.86 | | | | | | |
| TWI | 0.23 | 4.38 | | | | | | |
| Elevation | 0.18 | 5.65 | | | | | | |

*TWI* topographic wetness index, *LS* slope length and steepness, *NDVI* normalized difference vegetation index, *LULC* land use/land cover, *CEC* cation exchange capacity, *SPI* stream power index

Other multicollinearity diagnostic tools, including tolerance and VIF, were used to check if multicollinearity exists among predictors in each feature set. Almost all predictors had relatively high tolerance ($>0.22$) and low VIF ($<4.5$) values which indicate the non-existence of multicollinearity (Table 3). On the contrary, elevation yielded the lowest tolerance (0.18) and highest VIF (5.65) in a larger set. Nevertheless, these values also fall within the acceptable VIF threshold, considering a long-standing rule of thumb (i.e., VIF $<10$) for the non-existence of multicollinearity (Gareth et al. 2013; Vatcheva et al. 2016). Thus, all geo-environmental predictors in each feature set met these criteria and were used in the modeling process.

## 5.2 Model performance evaluation

The performance (i.e., accuracy and processing time) of the six algorithms (ANN, PLS, RDA, RF, SGB, and SVM) varied with different feature subsets (Fig. 5). SVM was more efficient with medium and small sets than large feature sets across all evaluation metrics. The algorithm obtained the highest $F1$-score (0.897), OA (0.898), and specificity (0.908) with the medium feature set, taking the shortest computation time ($<2$ min). SGB closely followed SVM in most evaluation metrics, but it was more efficient with larger feature sets and took more time ($>5$ min) to compute than SVM. An exception was in sensitivity, where SGB produced the highest accuracy (0.893) with a medium feature set and was computationally efficient ($<3.5$ min). Following SVM and SGB was RF, the most computationally expensive algorithm. RF was sensitive to the number of input features and

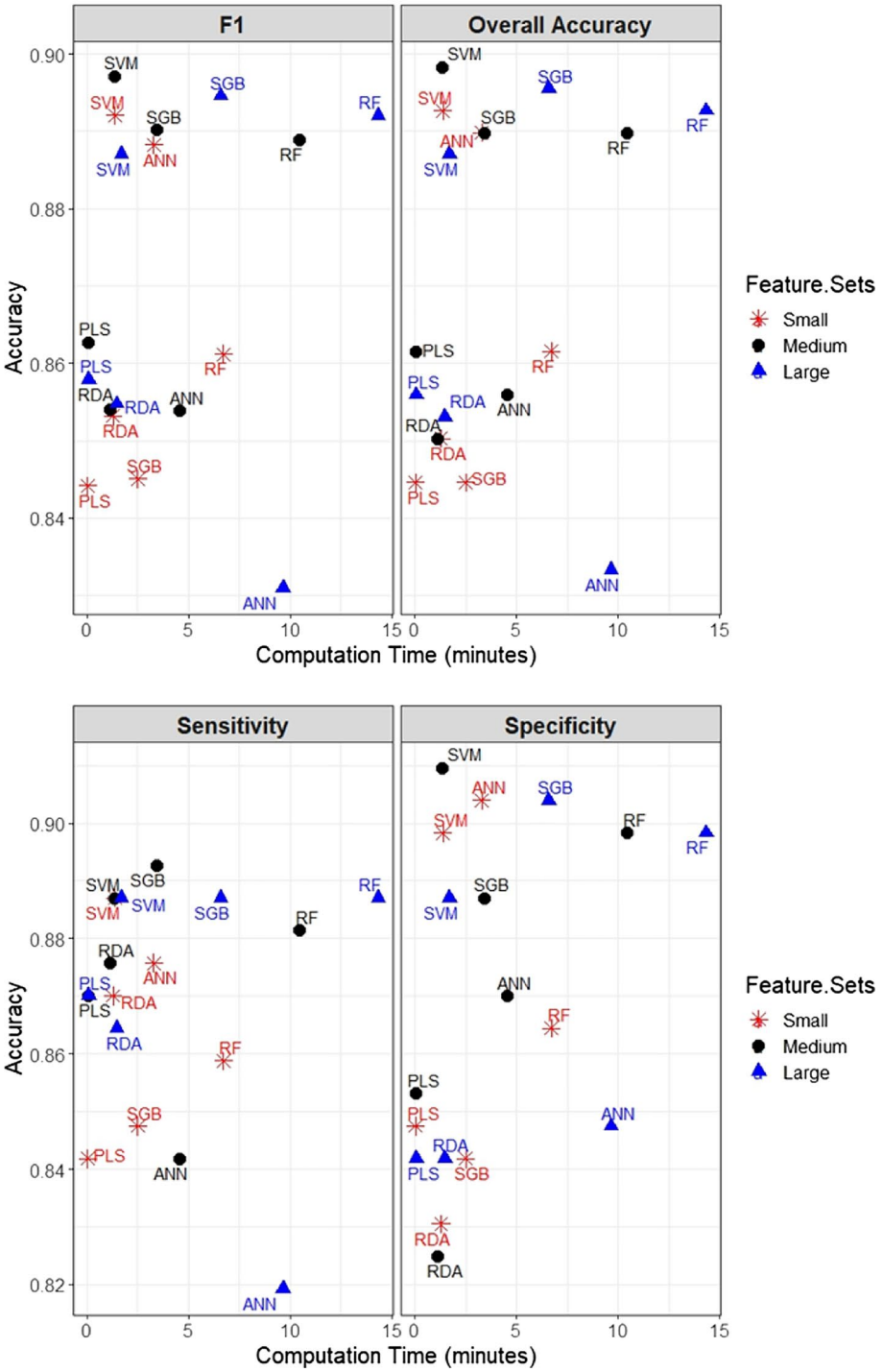
**Fig. 5** Predictive performance of ML models (ANN = artificial neural network, PLS = partial least squares, RDA = regularized discriminant analysis, RF = random forest, SGB = stochastic gradient boosting, SVM = support vector machines) using smaller (Sf), medium (Mf), and larger (Lf) feature sets

performed poorly with small and medium sets compared to large feature sets. Although the algorithm consistently yielded high accuracies with a large feature set across all evaluation metrics, it required more computation time (about 15 min) than any other algorithm in this study. On the contrary, PLS was the fastest algorithm, taking less than a minute to compute, and the number of features did not affect this computation time. Besides, it produced relatively high accuracies (0.85–0.87) with a medium feature set compared to other sets. Like PLS, RDA also took a relatively short computation time (< 1.5 min), and the accuracy values varied from 0.82 to 0.87, with most high values belonging to a large set. Much like RF, ANN was also sensitive to the number of input features and tended to take several minutes (about 10 min) of computation when a larger feature set was used. However, the algorithm consistently achieved relatively high accuracies (i.e., $F1$ score, OA, and specificity > 0.88) with the small feature set, outperforming PLS and RDA (except in sensitivity) across most evaluation metrics.

Regarding susceptibility maps, all six algorithms produced comparable outputs, depicting reasonable gully susceptibility, with the exception of RDA, which exhibited a different pattern (Fig. 6). Class-wise metrics, particularly specificity (< 0.85), indicate that RDA highly misclassified gully susceptibility. Notably, areas characterized by low gully susceptibility were predominantly classified as having medium to high gully susceptibility. This misclassification pattern can also be observed with a small and medium feature set, where RDA produced the lowest specificity values (≤ 0.83). SVM exhibited superior performance in its gully susceptibility map, aligning with the previously mentioned highest accuracy results. RF and SGB also yielded high-quality maps comparable to SVM, despite necessitating a larger feature set. In contrast, even though ANN utilized the smallest feature set, leading to a shorter computation time, it proved efficient and generated high-quality maps on par with other algorithms with superior predictive capabilities. Similarly, PLS produced an impressive gully susceptibility map within the shortest possible computation time (e.g.,, < 1 min).

### 5.3 Gully susceptibility and key controlling geo-environmental variables

All six algorithms indicated that a considerable portion (71–84%) of the catchment is of low gully susceptibility, whereas only 0.2–2.6% is under high gully susceptibility (Fig. 7). Furthermore, ranging from 4.5 to 6.6%, the proportion of very high gully susceptibility was almost the same across all algorithms. A very high susceptibility class corresponded to severely gullied areas, suggesting that these algorithms can detect individual gullies. In particular, the SVM-derived map is a good example illustrating areas with varying degrees of gully susceptibility (Fig. 8). Accordingly, SVM was used to produce this final map due to its superior predictive performance (Sect. 4.2). Medium to very high gully susceptibility is primarily confined to low-lying areas with gentle to flat slopes throughout the catchment (Fig. 8c–f), although some elevated and hilly parts (i.e., Fig. 8a) were under these gully susceptibility classes. Extensive gully systems were remarkable in the central (Fig. 8c), eastern (Fig. 8e), and southeastern (Fig. 8f) parts of the catchment. These areas are characterized by grasslands typically utilized for livestock grazing under subsistence farming. Abandoned agricultural lands with extensive gullies were also observed in the area. In


F1
Overall Accuracy
Sensitivity
Specificity
Accuracy
Computation Time (minutes)
Feature.Sets
Small
Medium
Large

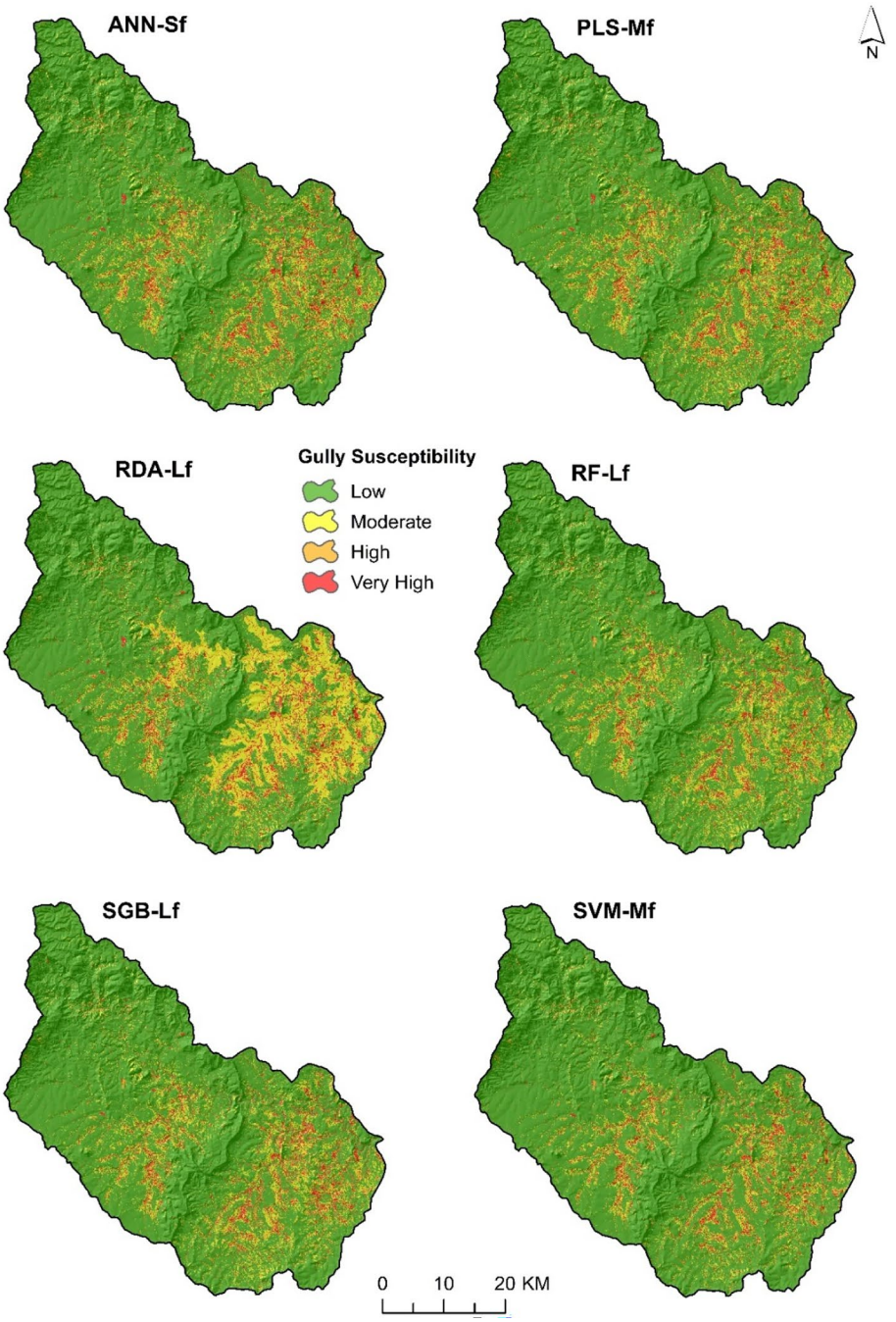



**Fig. 6** Gully susceptibility maps for each algorithm computed with smaller (Sf), medium (Mf) or larger (Lf) feature sets (ANN = artificial neural network, PLS = partial least squares, RDA = regularized discriminant analysis, RF = random forest, SGB = stochastic gradient boosting, SVM = support vector machines)

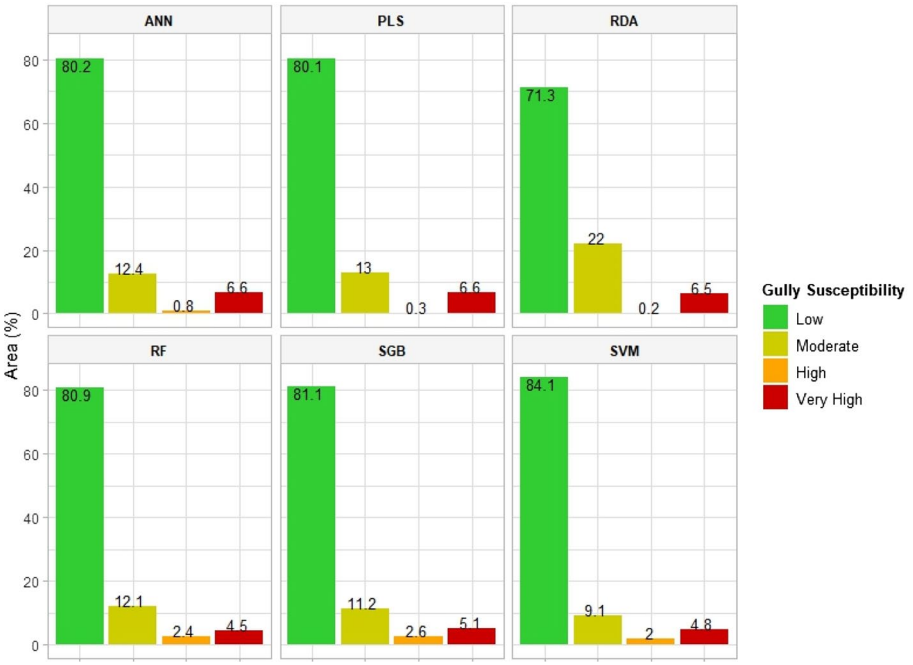



**Fig. 7** The proportion of area under different levels of gully susceptibility (ANN = artificial neural network, PLS = partial least squares, RDA = regularized discriminant analysis, RF = random forest, SGB = stochastic gradient boosting, SVM = support vector machines)

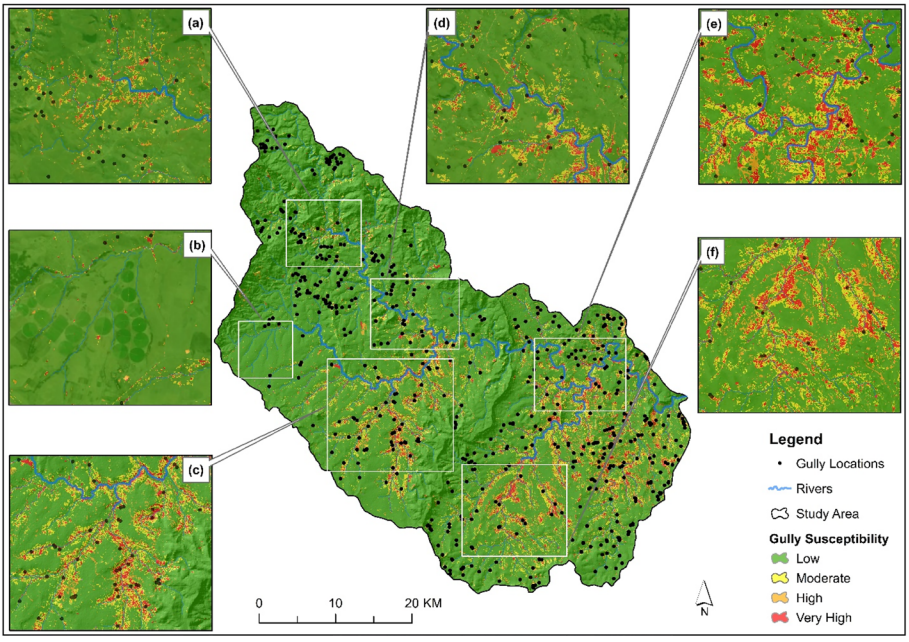



**Fig. 8** Distribution of gully locations across different parts (**a**–**f**) of the catchment with varying degrees of gully susceptibility

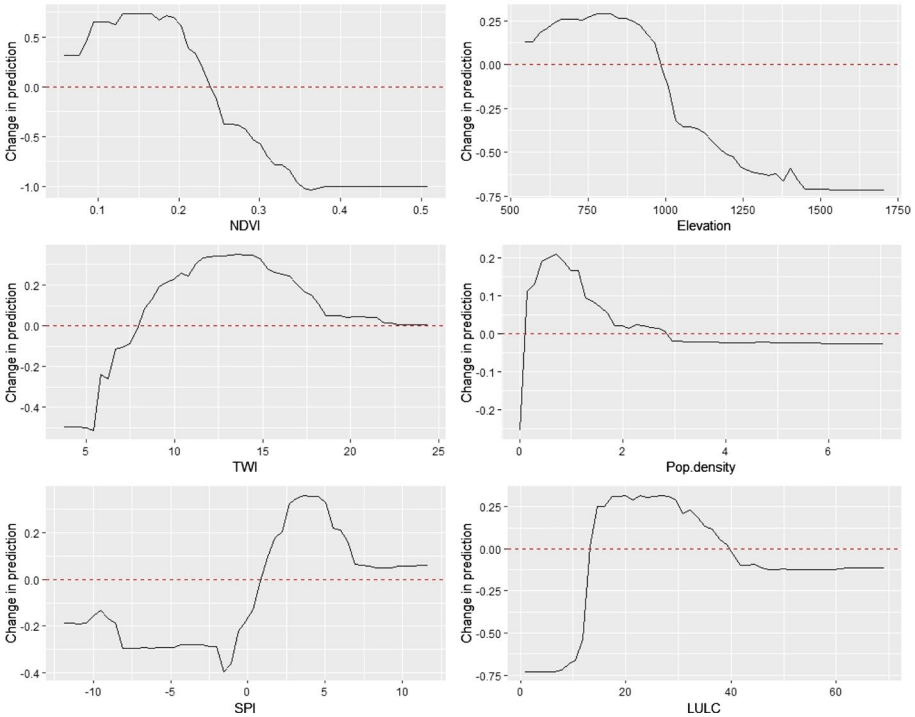



**Fig. 9** Partial dependence plots (PDP) of the six most important predictor variables (NDVI = normalized difference vegetation index, TWI = topographic wetness index, Pop.density = population density, SPI = stream power index)

contrast, the western section of the catchment, particularly those areas engaged in commercial farming, exhibited lower susceptibility to gully erosion (Fig. 8b).

Based on the variable importance analysis (see Sect. 4.1), NDVI, elevation, TWI, population density, SPI, and LULC were the six most influential variables. Each predictor variable exhibited a distinct PDP (Fig. 9), highlighting their varied influences on gully prediction. For instance, when NDVI values were low, there was a positive change in gully prediction, whereas high NDVI values were associated with a negative change in prediction. This implied that lower NDVI values positively affected gully susceptibility prediction, while higher values had a negative impact on the accuracy of the prediction outcome. This observation aligns with the fact that over 96% of gully pixels were found within a low NDVI range of − 0.23–0.35 (Table 4), suggesting that areas with little or no vegetative cover are highly susceptible to gullying.

Elevation showed a similar trend to NDVI in terms of its impact on gully prediction. Lower elevation values (< 500 m) were associated with a positive effect on gully prediction, but as the elevation increased to around 1000 m, gully prediction was negatively affected. In essence, the model showed a higher likelihood of accurately predicting gully susceptibility in areas with lower elevation compared to those at higher elevations. Notably, the majority (85%) of gully pixels in the catchment were concentrated in regions with low elevation (538–893 m), reinforcing the notion that these areas are more susceptible to gully formation than elevated areas (> 1000 m). These low-lying areas exhibited varying

**Table 4** Distribution of gully pixels across different value ranges for each of the six most critical geo-environmental predictors

| Gully factor | Class | Gully pixels | Gully pixels (%) |
|---|---|---|---|
| NDVI | − 0.23–0.18 | 132725 | 18.068 |
| | 0.19–0.24 | 196895 | 26.803 |
| | 0.25–0.29 | 167825 | 22.846 |
| | 0.30–0.35 | 133145 | 18.125 |
| | 0.36–0.72 | 24277 | 3.305 |
| Elevation (m) | 538–765 | 276441 | 37.625 |
| | 766–893 | 342962 | 46.679 |
| | 894–1039 | 24540 | 3.340 |
| | 1040–1255 | 10286 | 1.400 |
| | 1256–1772 | 748 | 0.102 |
| TWI | < 6.13 | 287932 | 39.189 |
| | 6.14–7.53 | 295262 | 40.187 |
| | 7.54–9.36 | 62294 | 8.479 |
| | 9.37–11.89 | 9437 | 1.284 |
| | 11.90–24.99 | 52 | 0.007 |
| Population density (people/km$^2$) | 0–9 | 654515 | 89.143 |
| | 10–85 | 0 | 0 |
| SPI | < − 6.25 | 156181 | 21.257 |
| | − 6.24–− 1.56 | 117122 | 15.941 |
| | − 1.55–0.36 | 166513 | 22.663 |
| | 0.37–3.23 | 169113 | 23.017 |
| | 3.24–13.35 | 46048 | 6.267 |
| LULC | Forested land | 8126 | 1.106 |
| | Grassland | 318769 | 43.389 |
| | Waterbodies | 2703 | 0.368 |
| | Wetlands | 1073 | 0.146 |
| | Barren land | 58494 | 7.962 |
| | Cultivated | 239029 | 32.535 |
| | Built-up | 26123 | 3.556 |
| | Mines and quarries | 606 | 0.082 |

*NDVI* normalized difference vegetation index, *TWI* topographic wetness index, *SPI* stream power index, *LULC* land use/land cover

degrees of water accumulation, as evident in TWI values ranging from < 6.13 to 7.53, with about 80% of gully pixels falling within this range. The PDP revealed that very low TWI values negatively impact gully prediction, but as the values increase, TWI begins to positively influence gully prediction. However, at some point, higher TWI values did not have any effect on gully prediction, leading to a constant PDP line around zero. Low TWI values indicate reduced wetness or water accumulation, while higher values indicate areas where surface water flow is more likely to occur (Sörensen et al. 2006), heightening vulnerability to gully erosion. The increased susceptibility of these areas to gullying can also be attributed to the increased erosive power of the flow, as indicated by high SPI values. The PDP corroborated this observation, revealing that lower SPI values negatively impacted gully prediction, while high values had a positive effect on prediction. However, at a certain point, the PDP line remained constant just above zero, indicating reduced variation in

prediction with higher SPI values. This suggests that the highest proportion of gully pixels was not associated with either very low (< 6.25) or extremely high (13.35) SPI values (Table 4). Instead, the majority (about 60%) of gully pixels were located within the intermediate SPI range (> − 6.25 to 3.23).

As indicated by the PDP, a lower population density had a positive impact on gully prediction, whereas higher population density had a negative effect on the prediction. When the population density was low, there was a more pronounced positive change in prediction, suggesting that low population density significantly influences gully modeling. Conversely, as population density increased, the PDP line fell below zero, and there was a reduced variation in prediction, indicating a lower likelihood of accurately predicting gully susceptibility in areas with higher population density. This observation aligns with expectations, considering that nearly 90% of the gully pixels were found in areas with lower population density. Regarding LULC, the PDP illustrated that various LULC classes have distinct effects on gully susceptibility prediction. Although the plot does not explicitly indicate which LULC classes have positive or negative effects, it is reasonable to infer that grassland and cultivated land positively influenced gully susceptibility modeling. This assumption is supported by the fact that grassland had the highest proportion of gully pixels (43%), followed by cultivated lands (33%), suggesting that these two classes are more susceptible to gully erosion compared to other LULC types in the study area.

## 6 Discussion

### 6.1 Model performance and computational efficiency

Our findings revealed that SVM was efficient in terms of both predictive accuracy and computation time when using a medium-sized feature set, outperforming other algorithms. While an increase in the number of input features had a minimal impact on computation time, it considerably affected the accuracy, with the largest feature set resulting in decreased accuracy across all metrics. This finding highlights that SVM's predictive performance can be negatively affected when numerous redundant variables are incorporated into model building, which aligns with previous studies (Hastie et al. 2009; Zhang et al. 2016). For instance, Fan and Fan (2008) demonstrated that SVM classification utilizing all features performs as poorly as random guessing. They attributed such a poor performance to the accumulation of noise in a high-dimensional feature space, where target classes overlap due to features having very similar information. In this study, an optimal subset for SVM comprised 12 predictor variables.

Partly, the remarkable success of SVM in this study can be also attributed to the class number of the target variable. Given the binary nature of the problem (gully vs. non-gully), the impressive results obtained by SVM in this study are expected, since SVM is inherently a binary classifier. When intuitively reflecting on this observation, the results are logical, as it is simpler for the model to distinguish between data points of only two classes (gully and non-gully) in comparison to dealing with data points from multiple classes. Additionally, the choice of an appropriate kernel function plays a crucial role in determining the SVM classification outcome by identifying an optimal hyperplane that separates these data points (support vectors). In this study, the application of the radial basis function (RBF), a non-linear kernel, ensured the perfect separation of gully and non-gully support vectors in the feature space, resulting in enhanced accuracy. Similarly, Rahmati et al. (2017) reported

that SVM applied with the RBF outperformed other kernel functions in accurately predicting gully susceptibility.

Although ensemble-based algorithms such as RF and SGB closely followed SVM in terms of predictive performance, they required larger input feature sets to achieve their maximum predictive capability, resulting in extended computation times. Such a finding demonstrates that these algorithms benefit from a larger feature set during their model building process. For instance, RF constructs an ensemble model of decision trees from random feature subsets (Nguyen and Huang 2015). The use of a larger feature set for each decision tree enhances the individual trees' robustness, reducing the risk of overfitting and enhancing predictive accuracy. In this study, RF performed optimally with the largest feature set, comprising 16 features. In the context of image classification, RF attained its highest average accuracy when the number of features falls within the range of 10–100 (Sheykhmousa et al. 2020). Indeed, various gully prediction studies employing RF utilized predictor variables within this range. For instance, Jiang et al. (2021) utilized 11 features, Gayen et al. (2019) employed 12 features, and Pham et al. (2020) used 14 features. In these studies, the RF algorithm estimated gully susceptibility with an AUC ranging from 88 to 96%, which is comparable to the accuracy (OA > 88%) achieved in our study. Several gully susceptibility studies indicated that RF outperformed SVM (Garosi et al. 2019; Pourghasemi et al. 2020; Hitouri et al. 2022; Huang et al. 2022). However, in these studies, the model comparison was based on a fixed set of predictor variables, while our study revealed that feature sets of varying sizes affect the performance of these algorithms differently. For example, when using a larger feature set, RF models outperformed SVM models using the same feature set size across all accuracy metrics. Nevertheless, with a medium-sized feature set, SVM demonstrated superior performance compared to all RF models, irrespective of a feature set size.

Similar to RF, the impressive performance of SGB with a larger feature set can be attributed to the inherent functionality of the algorithm. SGB relies on the ensemble of weak learners to capture complex relationships in the data (Zhang and Haghani 2015; Boehmke and Greenwell 2019; González et al. 2020). When the feature set is limited, the model may lack the necessary complexity to accurately represent underlying patterns. Therefore, it is reasonable to suggest that a larger feature set allowed the model to capture more meaningful patterns, leading to an improvement in predictive accuracy. In contrast, the behavior of ANN deviated from this trend. Specifically, ANN consistently achieved optimal accuracy results when working with a smaller feature set. Conversely, when utilizing a larger feature set, the accuracy of the ANN was at its lowest, recording extended computation time. In this context, with the increased number of features, there is a higher likelihood of incorporating irrelevant or redundant features. Such features, lacking meaningful contributions, can introduce noise, resulting in diminished performance, as indicated by the lowest accuracy and prolonged computation times.

In contrast to other algorithms, the size of the feature set had minimal impact on the computation time of PLS. While PLS produced gully susceptibility maps comparable to advanced algorithms such as SVM and ensemble-based methods, its accuracy was relatively lower. One possible explanation for this disparity is that PLS involved tuning only one hyperparameter (number of components), in contrast to other algorithms that had more than one hyperparameter. A combination of different hyperparameters is crucial for optimal model performance (Van Rijn and Hutter 2018; Yang and Shami 2020). Additionally, PLS relies on a single predictive model, unlike RF and SGB, which operate by combining multiple models into an ensemble for improved performance. Like PLS, RDA exhibited a relatively short computation time regardless of the feature set size. However, RDA

demonstrated the poorest performance among all algorithms in predicting gully susceptibility, generating maps with a distinct pattern suggestive of misclassification. The poor performance of RDA compared to other algorithms can be attributed to its assumption of a multivariate normal distribution in the data. In contrast, the other algorithms employed in this study are non-parametric and do not make such assumptions about data distribution, making them flexible and more effective in handling diverse distributions.

### 6.2 Key geo-environmental variables influencing gully erosion

Gully erosion is the result of both natural factors, such as vegetation, topography, and climate, and human-induced factors like LULC and population density (Valentin et al. 2005). Unsurprisingly, geo-environmental variables like NDVI, elevation, TWI, population density, SPI, and LULC, representing these gully erosion factors, were identified as the most important variables influencing gully susceptibility in this study. This finding aligns with previous research (Chowdhuri et al. 2021; Han et al. 2022; Jaafari et al. 2022; Roy and Saha 2022), affirming the crucial role of these variables in predicting gully susceptibility. For instance, Bernini et al. (2021) reported NDVI as the most crucial factor influencing formation of U-shaped gully systems. Recently, Barakat et al. (2022) also indicated the prominence of this variable in their erosion study. Typically, areas exhibiting lower NDVI values, which is indicative of poor or no vegetation cover, are more prone to gully erosion than densely vegetated areas. Indeed, inadequate vegetation cover has been identified as the primary factor driving gully erosion (Muñoz-Robles et al. 2010; Jahantigh and Pessarakli 2011). This observation aligns with our findings, as the majority of gully occurrences (pixels) were concentrated in areas characterized by low NDVI values (< 0.29). Phinzi and Ngetar (2017) also noted the presence of gullies within an NDVI range of 0.15–0.25, highlighting a correlation between gully formation and lower NDVI values. However, it is worth noting that different types of gully systems may exhibit different NDVI values. Bernini et al. (2021) identified low NDVI values (< 0.1) for V-shaped gullies and higher values (> 0.3) for U-shaped gullies. In our study, consistent with the PDP results, low NDVI values positively impacted gully susceptibility prediction, while high NDVI values had a negative predictive effect. This reinforces the notion that gullies are more prevalent in areas with deficient vegetative cover in the study area. Le Roux and Sumner (2012) also discovered a correlation between gully locations and poor vegetative cover in their research.

We observed that sparsely vegetated areas or bare surfaces were located in low-lying terrains with elevations ranging from 538 to 893 m, encompassing the majority of gully pixels (85%). Similar findings were reported by Haung et al. (2022) and Chowdhuri et al. (2021), emphasizing elevation as the primary factor influencing gully erosion. We also observed that the same areas of low elevation were predominantly associated with TWI values ranging from 6.13 to 7.53, indicating areas prone to wetness. These saturated areas create favorable conditions for gully formation, as the strength of surface soils diminishes when they become wet (Le Roux and Sumner 2012). Indeed, more than 70% of gully pixels fell within this TWI range, aligning with the study of Arabameri et al. (2019a), where a comparable proportion of gully pixels fell within a similar TWI range (< 4.41–6.84). In our study, the susceptibility of low-lying areas to gully formation could also be attributed to increased water surface runoff from the hillslope to the footslope. The intensified stream flow power, evident in high SPI values, also facilitates gullying in saturated, less vegetated areas characterized by low elevation. These findings underscore the complex interplay between topographic (elevation and TWI), hydrologic (SPI), and vegetation (NDVI) attributes in facilitating the development of gullies.

In addition to natural factors, human-related factors such as LULC and population density played a crucial role in gully erosion. Our findings indicated that grassland and cultivated lands collectively accounted for the highest proportion of gully pixels (76%), making them the most susceptible LULC types to gully erosion. This finding is expected, considering that cultivated areas and grasslands often experience soil disturbance, creating conditions favorable for the development of gullies (Le Roux and Sumner 2012). Our study, supported by field observations, indicated a higher impact of gully erosion in communal areas compared to nearby commercial areas, especially in the western parts of the catchment. This discrepancy arises because commercial areas generally have gentler slopes and lower rainfall-runoff ratios compared to communal areas (Meadows and Hoffman 2002). Le Roux et al. (2008) projected grassland to be the most vulnerable to erosion in South Africa due to rapid population and agricultural intensification. Meadows and Hoffman (2002) highlighted that the severity of soil degradation, including gully erosion, was most pronounced in communal areas where grazing is the dominant land use form. Recently, Olivier et al. (2022) reported approximately 40% of erosion in communal areas resulting from overgrazing and cattle tracks, while only 8.4% was linked to population pressure. Numerous studies conducted in South Africa concur that these anthropogenic activities have exacerbated soil erosion in communal areas (Beckedahl and de Villiers 2000; Kakembo and Rowntree 2003; Mhangara et al. 2012; Phinzi and Ngetar 2019a).

In contrast to earlier research (Reich et al. 1999; Meadows and Hoffman 2002), our study revealed that regions with lower population density exhibit a higher susceptibility to gully formation compared to those with higher population density. This observation aligns with findings from studies conducted in different parts of Africa. For instance, in Zimbabwe, Mambo and Archer (2007) reported that areas most susceptible to land degradation had the lowest population density, and a similar trend was observed in Tanzania by Achten et al. (2008), where gully occurrence was inversely related to high population density. Our findings, supported by field observations, illustrated that a considerable proportion of gullies (89% of gully pixels) was found in sparsely populated rural areas, particularly in grasslands and abandoned agricultural fields. Extensive gully systems were more prominent in those rural areas characterized by a relatively lower population density ($< 10$ persons/km$^2$) compared to densely populated areas (85 persons/km$^2$), particularly in urban zones around the small town of Engcobo. The results from the PDP further confirmed this trend, showing that lower population density positively affected gully erosion prediction, while higher population density had a negative predictive impact.

### 6.3 Limitations and future prospects

The predictor variables utilized in this study are not exhaustive, and future investigations could explore additional variables not included in our analysis. Geology demonstrated relatively low importance (0.2%) in explaining gully susceptibility. This warrants further investigation, particularly in the South African context, given that the geological rocks (specifically mudstones and sandstones) underlying our study area are known to contribute to gully formation, primarily due to highly erodible duplex soils derived from these rocks (Laker 2004; Le Roux and Sumner 2012). A more in-depth machine learning study on gully susceptibility incorporating geology and related variables such as soil type, *K*-factor, and chemical soil properties is recommended at the regional scale. At the catchment scale, unfortunately, there was less variation in the values of these predictors, which likely explains their lower importance in predicting gully formation at this scale. Especially

geology, with only two classes, did not contribute useful information to the model due to its near-zero variance, which is why it was excluded from the modeling.

The soil properties obtained from the world soil map are not deemed suitable for analysis due to their limited spatial heterogeneity at this scale. Unfortunately, these were the only soil data accessible for use in this study. Nevertheless, given the global availability of such soil data, future research could explore the significance of diverse soil properties in influencing gully formation on smaller scales (larger geographic areas). NDVI, identified as having the utmost importance in gully susceptibility, was derived from freely accessible Landsat data with the longest global image archive. This presents an opportunity to incorporate this variable in appraising the temporal dynamics of gully susceptibility at the catchment level.

This study proved comparing algorithms based on a fixed set of input features will invariably disadvantage or advantage one algorithm over another. This discrepancy arises because specific algorithms excel with specific feature set sizes. Therefore, caution should be exercised in identifying an appropriate subset of predictor variables, a consideration that may vary from one study to another, depending on the study area size, dataset characteristics, and data availability. Hence, for ensemble-based algorithms, it is advisable to incorporate as many features as possible, particularly those with significant predictive power, even though this approach can be computationally demanding (Conoscenti et al. 2013). Conversely, SVM and ANN exhibited exceptional performance with fewer input feature sets and demonstrated computational efficiency. Therefore, using these algorithms, particularly SVM, is recommended to achieve higher accuracy while minimizing computation time.

## 7 Conclusions

This study evaluated and compared the predictive performance of six commonly used algorithms in gully susceptibility under different scenarios of feature set sizes. SVM was the most efficient algorithm with a medium-sized feature set regarding accuracy (OA = 0.898, $F1$-score = 0.897) and computation time. Conversely, ensemble-based algorithms, mainly RF, required a larger set of predictors to reach maximum accuracy, taking several minutes of computation. Like RF, ANN was also sensitive to the number of input features, although its predictive performance was inversely related to the number of input features, reaching its maximum predictive accuracy with the smallest feature set. PLS and RDA were computationally efficient but the least-performing algorithms (OA and $F1$-score < 0.87). This suggests that predictor subsets of varying sizes influence both the prediction accuracy and computational efficiency of different ML algorithms. Despite varying accuracies, all algorithms produced reasonable and interpretable gully susceptibility maps, except RDA, which exhibited a completely different pattern. Geo-environmental covariates such as NDVI, elevation, TWI, population density, SPI, and LULC were identified as crucial for gully susceptibility modeling in this study. Accordingly, we recommend their inclusion in gully susceptibility assessments in similar environmental contexts. The study concluded that SVM is the most optimal algorithm, and its map was adopted for further visual analysis of gully susceptibility, including influencing gully factors in the catchment.

## Appendix

See Figs. 10 and 11.

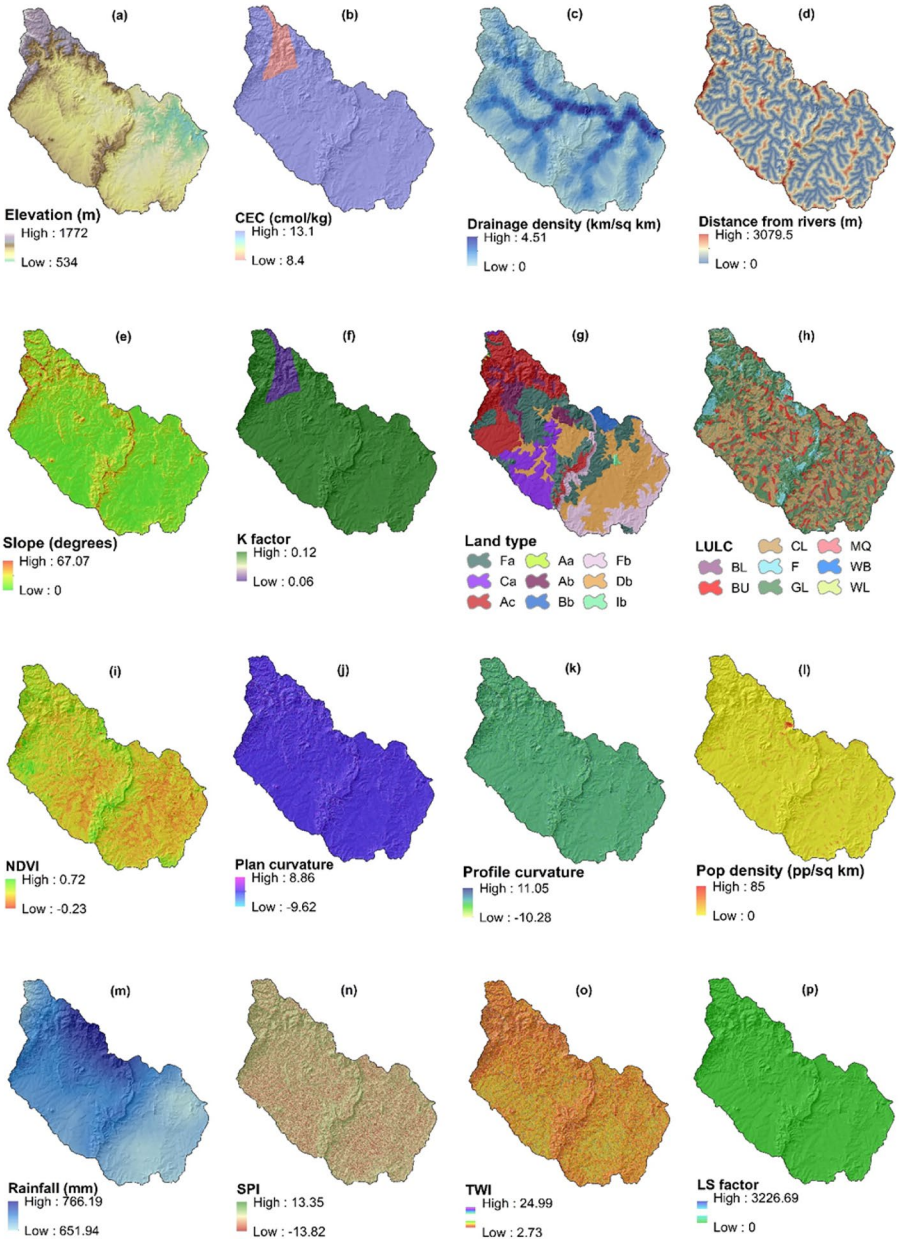



**Fig. 10** Geo-environmental covariates: **a** elevation, **b** cation exchange capacity (CEC), **c** drainage density, **d** distance from rivers, **e** slope, **f** *K* factor, **g** land type, **h** land use/land cover (LULC), **i** NDVI, **j** plan curvature, **k** profile curvature, **l** population density, **m** rainfall, **n** stream power index (SPI), **o** topographic wetness index (TWI), **p** slope length and steepness (LS), **q** aspect, **r** $CaCO_3$, **s** distance from roads, **t** geology, **u** terrain ruggedness index (TRI), and **v** soil pH

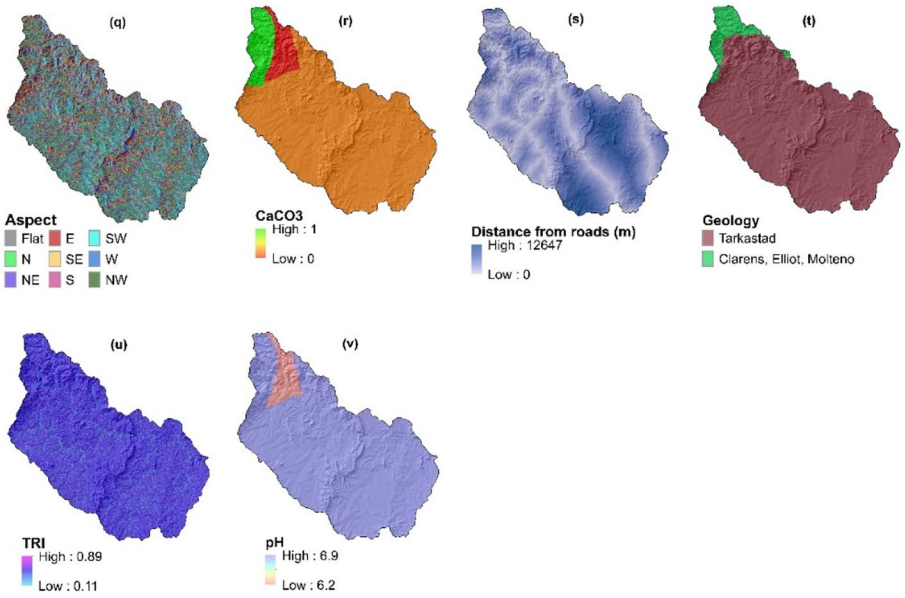



**Fig. 10** (continued)

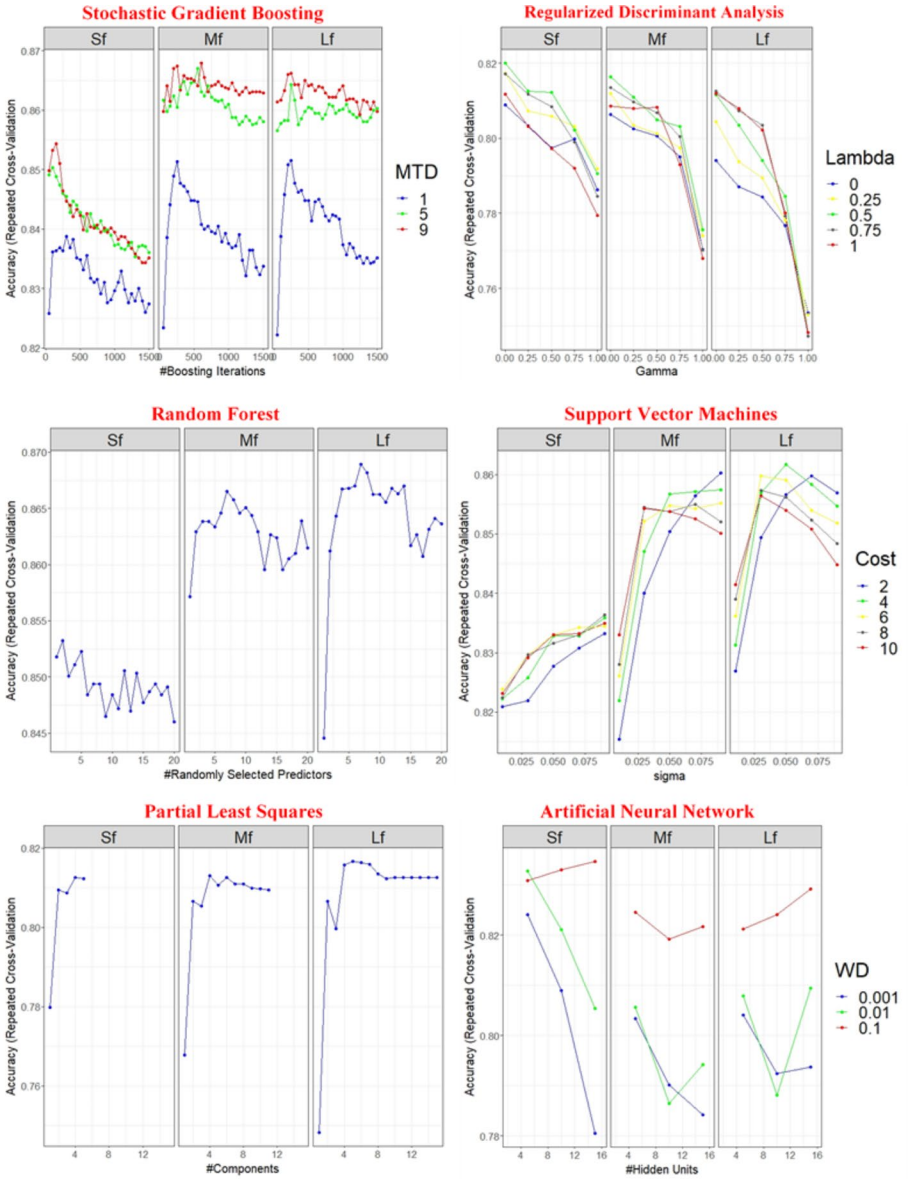



**Fig. 11** Hyperparameter tuning and the optimal combination of values selected for the final models in each ML algorithm across three scenarios of feature set sizes (Sf: small feature, Mf: medium feature, and Lf: large feature sets)

**Acknowledgements** We express our gratitude to the two anonymous reviewers for their valuable comments, which enhanced the quality of this manuscript. The first author is immensely grateful to the Tempus Public Foundation for funding his Ph.D. studies through the Stipendium Hungaricum Scholarship Program. The author is equally grateful to the Department of Higher Education and Training (DHET) of South Africa for the additional support.

**Author contributions** KP conceptualized, designed the study, performed analysis, wrote the first draft, and revised the manuscript. SS supervised the research, reviewed, edited, and revised the manuscript. Both authors read and approved the final manuscript.

**Funding** Open access funding provided by University of Zululand. This research was funded by the NKFI K 138079 and the KKP144068 projects.

## Declarations

**Conflict of interest** The authors have no relevant financial or non-financial interests to disclose.



## References

Abdi H (2003) Partial least square regression (PLS regression). Encycl Res Methods Soc Sci 6:792–795

Abdi AM (2020) Land cover and land use classification performance of machine learning algorithms in a boreal landscape using Sentinel-2 data. Gisci Remote Sens 57:1–20. https://doi.org/10.1080/15481603.2019.1650447

Achten WMJ, Dondeyne S, Mugogo S et al (2008) Gully erosion in south eastern Tanzania: spatial distribution and topographic thresholds. Zeitschrift Fur Geomorphologie 52:225–236

Alkarkhi AFM, Alqaraghuli WAA (2018) Discriminant analysis and classification. In: Alkarkhi AFM, Alqaraghuli WAA (eds) Easy statistics for food science with R. Academic Press, London, p 213

Amare S, Langendoen E, Keesstra S et al (2021) Susceptibility to gully erosion: applying random forest (RF) and frequency ratio (FR) approaches to a small catchment in Ethiopia. Water (basel) 13:216

Anbalagan R, Kumar R, Lakshmanan K et al (2015) Landslide hazard zonation mapping using frequency ratio and fuzzy logic approach, a case study of Lachung Valley, Sikkim. Geoenvironmental Disasters 2:1–17

Arabameri A, Chen W, Loche M et al (2019) Comparison of machine learning models for gully erosion susceptibility mapping. Geosci Front. https://doi.org/10.1016/j.gsf.2019.11.009

Azedou A, Lahssini S, Khattabi A et al (2021) A methodological comparison of three models for gully erosion susceptibility mapping in the rural municipality of El Faid (Morocco). Sustainability 13:682

Balogh S, Novák TJ (2020) Trends and hotspots in landscape transformation based on anthropogenic impacts on soil in Hungary, 1990–2018. Hungarian Geographical Bulletin 69:349–361

Barakat A, Rafai M, Mosaid H et al (2022) Mapping of water-induced soil erosion using machine learning models: a case study of Oum Er Rbia basin (Morocco). Earth Syst Environ 7:1–20

Beckedahl HR, de Villiers AB (2000) Accelerated erosion by piping in the eastern Cape province, South Africa. S Afr Geogr J 82:157–162. https://doi.org/10.1080/03736245.2000.9713709

Belgiu M, Drăgu L (2016) Random forest in remote sensing: a review of applications and future directions. ISPRS J Photogramm Remote Sens 114:24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011

Bennett SJ, Wells RR (2019) Gully erosion processes, disciplinary fragmentation, and technological innovation. Earth Surf Process Landf 44:46–53

Bernini A, Bosino A, Botha GA, Maerker M (2021) Evaluation of gully erosion susceptibility using a maximum entropy model in the upper Mkhomazi river basin in South Africa. ISPRS Int J Geoinf 10:729

Boehmke B, Greenwell BM (2019) Hands-on machine learning with R. CRC Press, Boca Raton

Breiman L (2001) Random forests. Mach Learn 45:5–32

Capra A, Scicolone B (2002) SW—soil and water: ephemeral gully erosion in a wheat-cultivated area in Sicily (Italy). Biosyst Eng 83:119–126

Castillo C, Gómez JA (2016) A century of gully erosion research: urgency, complexity and study approaches. Earth Sci Rev 160:300–319
Castillo C, Pérez R, James MR et al (2012) Comparing the accuracy of several field methods for measuring gully erosion. Soil Sci Soc Am J 76:1319–1332. https://doi.org/10.2136/sssaj2011.0390
Chowdhuri I, Pal SC, Saha A et al (2021) Evaluation of different DEMs for gully erosion susceptibility mapping using in-situ field measurement and validation. Ecol Inform 65:101425
Chung D, Keles S (2010) Sparse partial least squares classification for high dimensional data. Stat Appl Genet Mol Biol. https://doi.org/10.2202/1544-6115.1492
Conoscenti C, Rotigliano E (2020) Predicting gully occurrence at watershed scale: comparing topographic indices and multivariate statistical models. Geomorphology 359:107123
Conoscenti C, Agnesi V, Angileri S et al (2013) A GIS-based approach for gully erosion susceptibility modeling: a test in Sicily, Italy. Environ Earth Sci 70:1179–1195
Conoscenti C, Angileri S, Cappadonia C et al (2014) Gully erosion susceptibility assessment by means of GIS-based logistic regression: a case of Sicily (Italy). Geomorphology 204:399–411. https://doi.org/10.1016/j.geomorph.2013.08.021
Csatáriné Szabó Z, Mikita T, Négyesi G et al (2020) Uncertainty and overfitting in fluvial landform classification using laser scanned data and machine learning: a comparison of pixel and object-based approaches. Remote Sens (basel) 12:3652
Land Type Survey Staff Land Type Survey Database. Pretoria: ARC–Institute for Soil, Climate and Water. Pretoria
Dewitte O, Daoudi M, Bosco C, Van Den Eeckhaut M (2015) Predicting the susceptibility to gully initiation in data-poor regions. Geomorphology 228:101–115
Du Plessis C, Van Zijl G, Van Tol J, Manyevere A (2020) Machine learning digital soil mapping to inform gully erosion mitigation measures in the Eastern Cape. South Africa Geoderma 368:114287
DWA (2010) Mbashe River trends report (2007–2010) Department of Water Affairs river health Program eastern Cape
Ebhuoma O, Gebreslasie M, Ngetar NS et al (2022) Soil erosion vulnerability mapping in selected rural communities of Uthukela catchment, South Africa, using the analytic hierarchy process. Earth Systems and Environment 6:1–14
ESRI (2022) ArcGIS Desktop (Version 10.4)
Fan J, Fan Y (2008) High dimensional classification using features annealed independence rules. Ann Stat 36:2605
FAO (2003) The digital soil map of the world, land and water development division. FAO, Rome
Friedman JH (1989) Regularized discriminant analysis. J Am Stat Assoc 84:165–175
Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29:1189–1232
Friedman JH (2002) Stochastic gradient boosting. Comput Stat Data Anal 38:367–378
Funk C, Peterson P, Landsfeld M et al (2015) The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. Sci Data 2:1–21
Gafurov AM, Yermolayev OP (2020) Automatic gully detection: neural networks and computer vision. Remote Sens (basel) 12:1743
Gareth J, Daniela W, Trevor H, Robert T (2013) An introduction to statistical learning: with applications in R. Spinger, New York
Garosi Y, Sheklabadi M, Conoscenti C et al (2019) Assessing the performance of GIS-based machine learning models with different accuracy measures for determining susceptibility to gully erosion. Sci Total Environ 664:1117–1132
Garthwaite PH (1994) An interpretation of partial least squares. J Am Stat Assoc 89:122–127
Gayen A, Pourghasemi HR (2019) Spatial modeling of gully erosion: a new ensemble of CART and GLM data-mining algorithms. Spatial modeling in GIS and R for earth and environmental sciences. Elsevier, Netherlands, pp 653–669
Gayen A, Pourghasemi HR, Saha S et al (2019) Gully erosion susceptibility assessment and management of hazard-prone areas in India using different machine learning algorithms. Sci Total Environ 668:124–138
Ghaedi S, Shojaian A (2020) Spatial and temporal variability of precipitation concentration in Iran. Geogr Pannon 24:241–251
Gislason PO, Benediktsson JA, Sveinsson JR (2006) Random forests for land cover classification. Pattern Recognit Lett 27:294–300
González S, García S, Del Ser J et al (2020) A practical tutorial on bagging and boosting based ensembles for machine learning: algorithms, software tools, performance study, practical perspectives and opportunities. Information Fusion 64:205–237

Goodwin NR, Armston JD, Muir J, Stiller I (2017) Monitoring gully change: a comparison of airborne and terrestrial laser scanning using a case study from Aratula, Queensland. Geomorphology 282:195–208
Greenwell BM (2017) pdp: An R package for constructing partial dependence plots. R J 9:421–436
Han J, Guzman JA, Chu ML (2022) Gully erosion susceptibility considering spatiotemporal environmental variables: midwest US region. J Hydrol Reg Stud 43:101196
Hastie T, Tibshirani R, Friedman JH (2009) The elements of statistical learning: data mining, inference, and prediction, 2nd edn. Springer, New York
Hearst MA, Dumais ST, Osuna E et al (1998) Support vector machines. IEEE Intell Syst Appl 13:18–28
Hitouri S, Varasano A, Mohajane M et al (2022) Hybrid machine learning approach for gully erosion mapping susceptibility at a watershed scale. ISPRS Int J Geoinf 11:401
Hosseinalizadeh M, Kariminejad N, Chen W et al (2019) Gully headcut susceptibility modeling using functional trees, naïve Bayes tree, and random forest models. Geoderma 342:1–11
Huang D, Su L, Zhou L et al (2022) Assessment of gully erosion susceptibility using different DEM-derived topographic factors in the black soil region of northeast China. Int Soil Water Conserv Res 11(1):97–111
Irizarry RA (2019) Introduction to data science: data analysis and prediction algorithms with R. CRC Press, Boca Raton
ISRIC (2002) Luvisols (lv)
Jaafari A, Janizadeh S, Abdo HG et al (2022) Understanding land degradation induced by gully erosion from the perspective of different geoenvironmental factors. J Environ Manag 315:115181. https://doi.org/10.1016/j.jenvman.2022.115181
Jahantigh M, Pessarakli M (2011) Causes and effects of gully erosion on agricultural lands and the environment. Commun Soil Sci Plant Anal 42:2250–2255
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
Kakembo V, Rowntree KM (2003) The relationship between land use and soil erosion in the communal lands near Peddie town, eastern Cape, South Africa. Land Degrad Dev 14:39–49. https://doi.org/10.1002/ldr.509
Kuhn M (2008) Building predictive models in R using the caret package. J Stat Softw 28:1–26
Kuhn M, Johnson K (2013) Applied predictive modeling. Springer
Kulimushi LC, Bashagaluke JB, Prasad P et al (2023) Soil erosion susceptibility mapping using ensemble machine learning models: a case study of upper Congo river sub-basin. Catena (amst) 222:106858
Laker MC (2004) Advances in soil erosion, soil conservation, land suitability evaluation and land use planning research in South Africa, 1978–2003. South Afr J Plant Soil 21:345–368
Lana JC, de Castro PTA, Lana CE (2022) Assessing gully erosion susceptibility and its conditioning factors in southeastern Brazil using machine learning algorithms and bivariate statistical methods: a regional approach. Geomorphology 402:108159
Le Roux JJ, Sumner PD (2012) Factors controlling gully development: Comparing continuous and discontinuous gullies. Land Degrad Dev 23:440–449. https://doi.org/10.1002/ldr.1083
Le Roux JJ, Morgenthal TL, Malherbe J et al (2008) Water erosion prediction at a national scale for South Africa. Water Sa 34:305–314
Lee S, Pradhan B (2007) Landslide hazard mapping at Selangor, Malaysia using frequency ratio and logistic regression models. Landslides 4:33–41
Lee S, Sambath T (2006) Landslide susceptibility mapping in the Damrei Romel area, Cambodia using frequency ratio and logistic regression models. Environ Geol 50:847–855
Liaw A, Wiener M (2002) Classification and Regression by randomForest. R News 2:18–22
Liu G, Zheng F, Wilson GV et al (2021) Three decades of ephemeral gully erosion studies. Soil Tillage Res 212:105046
Magliulo P (2012) Assessing the susceptibility to water-induced soil erosion using a geomorphological, bivariate statistics-based approach. Environ Earth Sci 67:1801–1820
Mambo J, Archer E (2007) An assessment of land degradation in the save catchment of Zimbabwe. Area 39:380–391
Mararakanye N, Le Roux JJ (2012) Gully location mapping at a national scale for South Africa. S Afr Geogr J 94:208–218. https://doi.org/10.1080/03736245.2012.742786
Mararakanye N, Nethengwe NS (2012) Gully erosion mapping using remote sensing techniques. South Afr J Geomat 1:109–118
Mason CH, Perreault WD Jr (1991) Collinearity, power, and interpretation of multiple regression analysis. J Mark Res 28:268–280

Meadows ME, Hoffman MT (2002) The nature, extent and causes of land degradation in south Africa: legacy of the past, lessons for the future? Area 34:428–437. https://doi.org/10.1111/1475-4762.00100
Mhangara P, Kakembo V, Lim KJ (2012) Soil erosion risk assessment of the Keiskamma catchment, south Africa using GIS and remote sensing. Environ Earth Sci 65:2087–2102. https://doi.org/10.1007/s12665-011-1190-x
Moisen GG, Freeman EA, Blackard JA et al (2006) Predicting tree species presence and basal area in Utah: a comparison of stochastic gradient boosting, generalized additive models, and tree-based methods. Ecol Modell 199:176–187
Moore ID, Burch GJ (1986) Physical basis of the length-slope factor in the universal soil loss equation. Soil Sci Soc Am J 50:1294–1298
Moore ID, Grayson RB, Ladson AR (1991) Digital terrain modeling: a review of hydrological, geomorphological, and biological applications. Hydrol Process 5:3–30
Muñoz-Robles C, Reid N, Frazier P et al (2010) Factors related to gully erosion in woody encroachment in south–eastern Australia. Catena (amst) 83:148–157
Nguyen T-T, Huang JZ, Nguyen TT (2015) Unbiased feature selection in learning random forests for high-dimensional data. Sci World J. https://doi.org/10.1155/2015/471371
Olivier G, Van De Wiel MJ, De Clercq WP (2022) Intersecting views of gully erosion in South Africa. Earth Surf Process Landf 48(1):119–142
Pham QB, Mukherjee K, Norouzi A et al (2020) Head-cut gully erosion susceptibility modeling based on ensemble random forest with oblique decision trees in Fareghan watershed, Iran. Geomat, Nat Hazards Risk 11:2385–2410
Phinzi K, Ngetar NS (2017) Mapping soil erosion in a quaternary catchment in eastern Cape using geographic information system and remote sensing. South Afr J Geomat 6:11. https://doi.org/10.4314/sajg.v6i1.2
Phinzi K, Ngetar NS (2019a) Land use/land cover dynamics and soil erosion in the Umzintlava catchment (T32E), eastern Cape, South Africa. Transactions of the Royal Society of South Africa 74:223–237. https://doi.org/10.1080/0035919X.2019.1634652
Phinzi K, Ngetar NS (2019b) The assessment of water-borne erosion at catchment level using GIS-based RUSLE and remote sensing: a review. Int Soil Water Conserv Res 7:27–46. https://doi.org/10.1016/j.iswcr.2018.12.002
Phinzi K, Abriha D, Bertalan L et al (2020) Machine learning for gully feature extraction based on a pan-sharpened multispectral image: multiclass vs Binary approach. ISPRS Int J Geoinf. https://doi.org/10.3390/ijgi9040252
Phinzi K, Holb I, Szabó S (2021) Mapping permanent gullies in an agricultural area using satellite images: efficacy of machine learning algorithms. Agronomy 11:333
Pimentel D (2006) Soil erosion: a food and environmental threat. Environ Dev Sustain 8:119–137
Pimentel D, Harvey C, Resosudarmo P et al (1995) Environmental and economic costs of soil erosion and conservation benefits. Science 267:1117–1123
Poesen J, Nachtergaele J, Verstraeten G, Valentin C (2003) Gully erosion and environmental change: importance and research needs. Catena (amst) 50:91–133. https://doi.org/10.1016/S0341-8162(02)00143-1
Pourghasemi HR, Yousefi S, Kornejady A, Cerdà A (2017) Performance assessment of individual and ensemble data-mining techniques for gully erosion modeling. Sci Total Environ 609:764–775. https://doi.org/10.1016/j.scitotenv.2017.07.198
Pourghasemi HR, Sadhasivam N, Kariminejad N, Collins AL (2020) Gully erosion spatial modeling: role of machine learning algorithms in selection of the best controlling factors and modeling process. Geosci Front 11:2207–2219
R Core Team (2021) R: a language and environment for statistical computing. R Foundation for statistical computing, Vienna
Rahmati O, Pourghasemi HR, Zeinivand H (2016) Flood susceptibility mapping using frequency ratio and weights-of-evidence models in the Golastan province. Iran Geocarto Int 31:42–70
Rahmati O, Tahmasebipour N, Haghizadeh A et al (2017) Evaluation of different machine learning models for predicting and mapping the susceptibility of gully erosion. Geomorphology 298:118–137
Reich P, Eswaran H, Beinroth F (1999) Global dimensions of vulnerability to water and wind erosion. In: Scott DE, Mohtar RH, Steinhardt GC (eds) Sustaining the Global Farm. Proceeding from the 10th International Soil Conservation Organization Meeting. Purdue University and USDA-ARS National Soil Erosion Research Laboratory, pp 838–846
Renard KG, Foster GR, Weesies GA, et al (1997) Predicting soil erosion by water: A guide to conservation planning with the Revised Universal Soil Loss Equation (RUSLE). U.S. Department of Agriculture, Agricultural Research Service, Washington DC
Ridgeway G (2007) Generalized boosted models: a guide to the gbm package. Update 1:2007

Van Rijn JN, Hutter F (2018) Hyperparameter importance across datasets. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining. pp 2367–2376
Roberts ME, Burrows RM, Thwaites RN, Hamilton DP (2022) modeling classical gullies—a review. Geomorphology 407:108216
Rodriguez-Galiano VF, Ghimire B, Rogan J et al (2012) An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS J Photogramm Remote Sens 67:93–104
Roy J, Saha S (2019) 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 15:100247
Roy J, Saha S (2022) Ensemble hybrid machine learning methods for gully erosion susceptibility mapping: K-fold cross validation approach. Artif Intell Geosci 3:28–45
Shafapour Tehrany M, Kumar L, Neamah Jebur M, Shabani F (2019) Evaluating the application of the statistical index method in flood susceptibility mapping and its comparison with frequency ratio and logistic regression methods. Geomat Nat Haz Risk 10:79–101
Sheykhmousa M, Mahdianpari M, Ghanbari H et al (2020) Support vector machine versus random forest for remote sensing image classification: a meta-analysis and systematic review. IEEE J Sel Top Appl Earth Obs Remote Sens 13:6308–6325. https://doi.org/10.1109/JSTARS.2020.3026724
Shruthi RBV, Kerle N, Jetten V (2011) Object-based gully feature extraction using high spatial resolution imagery. Geomorphology 134:260–268. https://doi.org/10.1016/j.geomorph.2011.07.003
Sörensen R, Zinko U, Seibert J (2006) On the calculation of the topographic wetness index: evaluation of different methods based on field observations. Hydrol Earth Syst Sci 10:101–112
Svoray T, Michailov E, Cohen A et al (2012) Predicting gully initiation: comparing data mining techniques, analytical hierarchy processes and the topographic threshold. Earth Surf Process Landf 37:607–619
Tu JV (1996) Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. J Clin Epidemiol 49:1225–1231
Valentin C, Poesen J, Li Y (2005) Gully erosion: Impacts, factors and control. Catena (amst) 63:132–153. https://doi.org/10.1016/j.catena.2005.06.001
Valiant LG (1984) A theory of the learnable. Commun ACM 27:1134–1142
Van Zijl GM, Le Roux PAL, Turner DP (2013) Disaggregation of land types using terrain analysis, expert knowledge and GIS methods. South Afr J Plant Soil 30:123–129
Varga OG, Kovács Z, Bekő L et al (2021) Validation of visually interpreted corine land cover classes with spectral values of satellite images and machine learning. Remote Sens (basel) 13:857
Vatcheva KP, Lee M, McCormick JB, Rahbar MH (2016) Multicollinearity in regression analyses conducted in epidemiologic studies. Epidemiology (Sunnyvale). https://doi.org/10.4172/2161-1165.1000227
Venables WN, Ripley BD (2013) Modern applied statistics with S-PLUS. Springer Science & Business Media, New York
Vrieling A, Sterk G, de Jong SM (2010) Satellite-based estimation of rainfall erosivity for Africa. J Hydrol (amst) 395:235–241
Wang S-C (2003) Artificial neural network. Interdisciplinary computing in java programming. Springer, New York, pp 81–100
Wehrens R, Mevik B-H (2007) The pls package: principal component and partial least squares regression in R. J Stat Softw 18:1–24
Welch BL (1939) Note on discriminant functions. Biometrika 31:218–220
Williams JR (1995) The EPIC model. Computer models of watershed hydrology. Resources Publications, Highlands Ranch, pp 909–1000
Williams JR, Berndt HD (1977) Sediment yield prediction based on watershed hydrology. Trans ASAE 20:1100–1104
Wischmeier WH, Smith DD (1978) Predicting rainfall erosion losses: a guide to conservation planning. Department of Agriculture, Science and Education Administration
Wu W, Mallet Y, Walczak B et al (1996) Comparison of regularized discriminant analysis linear discriminant analysis and quadratic discriminant analysis applied to NIR data. Anal Chim Acta 329:257–265
Yang L, Shami A (2020) On hyperparameter optimization of machine learning algorithms: theory and practice. Neurocomputing 415:295–316
Zhang Y, Haghani A (2015) A gradient boosting method to improve travel time prediction. Transp Res Part C Emerg Technol 58:308–324
Zhang X, Wu Y, Wang L, Li R (2016) Variable selection for support vector machines in moderately high dimensions. J R Stat Soc Series B Stat Methodol 78:53–76