Gully erosion susceptibility prediction in Mollisols using machine learning models

Y. Wang, Y. Zhang, and H. Chen

Abstract: In recent years, gully erosion has caused soil loss, land degradation, and a large sediment yield in the Mollisols in northeastern China, threatening agricultural development and national food security. Moreover, the prediction of gully erosion remains a great challenge owing to the difficulty of determining suitable environmental indicators and identifying the best models for predicting gully erosion prone areas. Therefore, the objective of this study was to quantify the contributions of the main factors controlling gully erosion and to identify the best model for predicting areas susceptible to gully erosion in Hailun City, northeastern China. Initially, the spatial distribution of the gully erosion was investigated through visual interpretation of GaoFen-1 satellite images. The analyzed gullies were evenly distributed in the study region, and we selected 70% of the gullies as the training data set and the remaining 30% as the validation data set. Subsequently, 12 variables, including the elevation, slope, aspect, plan curvature, profile curvature, topographic wetness index (TWI), soil type, land use, normalized difference vegetation index (NDVI), precipitation, distance from rivers, and distance from existing gullies, were selected as the indicators of gully erosion. Then, multicollinearity analysis was conducted to determine the main indicators without linearity. Finally, the contributions of the indicators and the areas susceptible to gully erosion were determined using machine learning models, including support vector machine (SVM), multilayer perceptron neural network (MLPNN), random forest (RF), and extreme gradient boosting (XGBoost) models. The results revealed that there was no multicollinearity among the 12 indicators, so they were all employed in the machine learning models for the gully erosion susceptibility prediction. The XGBoost model had the highest R² and lowest root mean square error (RMSE) values in the model validation stage (0.81 and 0.60, respectively), followed by the RF (0.78 and 0.61, respectively), MLPNN (0.65 and 0.70, respectively), and SVM (0.62 and 0.70, respectively). The gully distance had the largest relative importance score (>35%) for gully erosion, followed by the profile curvature, plan curvature, land use, elevation, and soil type, which had relative importance scores of 10% to 15%. The gully erosion susceptibility map revealed that the central part of the study area was more susceptible to gully erosion than the other regions. These results can help managers to identify the regions that are prone to gully erosion and to design soil conservation practices to slow down the erosion process.

Key words: controlling factors—gully erosion susceptibility—machine learning—Mollisols—relative importance

Soil erosion, an important cause of soil loss, land degradation, and sediment yield, threatens agricultural development and the ecological environment (Zhao et al. 2016; Poesen 2018). The Mollisol region in northeastern China is an important agricultural production base and is subject to severe soil erosion. The average erosion rate in this region was approximately 15 Mg ha⁻¹ y⁻¹ or more during the last century (Nearing et al. 2017). Gully erosion is a common type of soil erosion in the Mollisol region in northeastern China. Gully erosion is defined as linear soil erosion over a short time period, even after an intense rainfall event. At present, there are nearly 300,000 gullies in the Mollisols region in northeastern China, and 61% of them are distributed in cultivated land. Moreover, the area of cultivated land has decreased by 0.5% to 3% (Ministry of Water Resources 2010). Therefore, to prevent gully erosion, it is necessary to determine the main factors controlling gully erosion and to identify the spatial pattern of the areas prone to gully erosion.

Based on previous studies, gully erosion is influenced by the climate, topography, hydrology, geology, and environmental conditions (Pourghasemi et al. 2017a; Gayen et al. 2019; Mokarram and Zarei 2021). In recent years, a variety of studies have been conducted to investigate the controlling factors in different areas. For example, Mokarram and Zarei (2021) determined the areas prone to gully erosion in the southern region of Fars Province and found that the normalized different vegetation index (NDVI), slope, topographic wetness index (TWI), altitude, terrain ruggedness index (TRI), lithology, and land use were the most important controlling factors. Chen (2021) reported that altitude, land use, distance from road, and soil characteristics were the key factors controlling gully erosion in the Meinard watershed. Amiri et al. (2019) revealed that three factors, including the land use, distance from river, and clay content, had the most important impacts on gully erosion, and they quantified the contribution of each factor. However, many factors influencing the distribution of gully erosion prone areas have been identified, and it is too costly and time-consuming to employ them all in the determination of gully erosion prone areas. In addition, because the Mollisol region in northeastern China is a rolling and hilly area, with concentrated rainfall in summer, alternating freezing and thawing in winter and spring, ridged tillage, and long-term
high-intensity utilization, the distribution of the erosion gullies is complicated, and the controlling factors have large uncertainties (Zhang et al. 2018; Wen et al. 2021). At present, the main factors influencing the gully erosion in the Mollisol region in northeastern China and their contributions remain unclear.

Owing to the limited data available, the prediction of gully erosion remains a great challenge worldwide, especially at the regional scale and larger scales (e.g., basin scale, province scale, and country scale). Gully susceptibility prediction models can be categorized into four types, including primal inventory-based models, heuristic models, data-driven statistical models, and physically-based deterministic models (Soleimanpour et al. 2021). Among these categories, data-driven statistical models are considered to be more reliable than the other types of models that rely on subjective and heuristic methods (Chen et al. 2021). As data-driven statistical models, machine learning methods have been applied in prediction systems, especially for the determination of highly complex relationships (Kubat 2017). Machine learning models such as the random forest (RF) (Roy et al. 2020), boosted regression tree (BRT) (Buston and Elith 2011), extreme gradient boosting (XGBoost) (Yang et al. 2021), support vector machine (SVM) (Roy et al. 2020), and artificial neural network (ANN) (Pourghasemi et al. 2017b) models have been applied for gully erosion susceptibility (GES) prediction. Previous research has focused on the development of ANN models to predict areas susceptible to gully erosion; however, the multilayer perceptron neural network (MLPNN) also has a great potential for computing nonlinear relationships between inputs and outputs, and it has been seldom used in gully prediction studies (Pourghasemi et al. 2017b; Oludolapo et al. 2012; de Oliveira et al. 2021). In previous studies, various machine learning models, such as supervised learning, neural network, and tree-based models, have been reported to be helpful in GES mapping. In addition, the contributions of the key controlling factors could also be obtained using machine learning models in this region. However, there is still debate as to the most effective and accurate model for gully erosion prediction because of the models’ complex structures.

Therefore, the objectives of this study were (1) to select suitable indicators of gully erosion and assess the contributions of the key factors controlling gully erosion; (2) to compare the performances of different machine learning models (SVM, MLPNN, RF, and XGBoost) and identify the best model for the prediction of GES; and (3) to obtain a GES map of the Mollisol region in northeastern China. Although the four models used in this study have been employed individually in previous studies, their application and comparison in a new, unstudied region contributes to our knowledge of the reliability of machine learning methods in predicting areas prone to gully erosion in different regions.

Materials and Methods

Study Area. Hailun (46°58′ to 47°52′ N, 126°14′ to 127°45′ E) is located in the central Mollisol region in Heilongjiang Province, China, with elevations ranging from 144 to 484 m (figure 1). It is an important national commodity grain base in China, and the large difference in elevation and long-term extensive usage of the land have led to seri-
ous soil erosion in Hailun City. This area is plain land, with an area of 4,667 km². Hailun is located in the north temperate continental monsoon climate zone, with a maximum temperature of 37°C and a minimum temperature of −39.5°C during the year. The main soil types include black soil, meadow soil, swamp soil, and small amounts of paddy soil, dark-brown soil, and albic soil (Tang et al. 2021).

Mapping of Gully Erosion Situation. Since gully erosion prone areas are closely linked to the distribution pattern of existing gullies, gully erosion investigation is essential for GES prediction. As figure 2 shows, the gullies photographed using an unmanned aerial vehicle were captured in the central part of Hailun, with a viewing azimuth of ~45°. There were 4,662 existing gullies distributed in Hailun based on visual interpretation of GaoFen-1 (GF-1) images with a 2 m resolution. GaoFen is the first civilian optical remote sensing satellite that was independently developed by China. This GF series satellite is a high-definition Earth observation system (HDEOS) developed by the China Academy of Space Technology (CAST) (Sun et al. 2018). Of the 4,662 gullies identified, 70% (3,263 gullies, evenly distributed throughout the entire study area) were selected and used to calibrate the prediction model; the other 30% (1,399 gullies) were used for the model validation (figure 1).

Selection of Indicators. The indicators were selected based on the following criteria: data availability, previous experiences and reports in the literature, data connection and heterogeneity, and local geo-environmental characteristics (Arabameri et al. 2018; Teimouri and Kornejady 2019; Zabihi et al. 2019; Gayen et al. 2019; Chen et al. 2021). The data connection was taken into account because the factors should be connected to each other and work together to influence the gully erosion process. In addition, the data heterogeneity was considered to ensure that the factors with a similar pattern across the study region were excluded while more factors with different properties were involved in the modeling process (Soleimanpour et al. 2021). It is worth noting that some of the environmental indicators (e.g., soil erodibility and runoff) with high-resolution data were not available for the study area, so such indicators were not used in this study. In addition, Hailun is mainly a water erosion area, so wind erosion related factors such as the wind direction and velocity were not considered in this study. Finally, 12 indicators were selected: elevation, slope, aspect, plan curvature, profile curvature, TWI, soil type, land use, NDVI, precipitation, distance from rivers, and distance from existing gullies (table 1). We found that half of the indicators were topography-related variables. This occurred because Hailun is a rolling and hilly area, so the topographic features have more impact on gully erosion in this region. Moreover, the topographic factors were associated with other environmental variables (e.g., precip-
Table 1
Details of the controlling factors used in gully erosion prediction.

<table>
<thead>
<tr>
<th>Controlling factor</th>
<th>Range or type</th>
<th>Spatial resolution (m)</th>
<th>Data source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elevation (m)</td>
<td>144 to 484</td>
<td>12.5</td>
<td>Derived from DEM [<a href="http://glovis.usgs.gov/">http://glovis.usgs.gov/</a>]</td>
</tr>
<tr>
<td>Slope (degree)</td>
<td>0 to 11.94</td>
<td>12.5</td>
<td>Derived from DEM</td>
</tr>
<tr>
<td>Aspect</td>
<td>Flat, north, northeast, east, southeast, south, southwest, west, northwest</td>
<td>12.5</td>
<td>Derived from DEM</td>
</tr>
<tr>
<td>Plan curvature</td>
<td>0 to 66.65</td>
<td>12.5</td>
<td>Derived from DEM</td>
</tr>
<tr>
<td>Profile curvature</td>
<td>0 to 3.17</td>
<td>12.5</td>
<td>Derived from DEM</td>
</tr>
<tr>
<td>TWI</td>
<td>1.58 to 21.96</td>
<td>12.5</td>
<td>Derived from DEM</td>
</tr>
<tr>
<td>Soil type</td>
<td>Black soil, dark-brown soil, fluvo-aquic soil, peat soil</td>
<td>–</td>
<td>The Second National Soil Survey of China</td>
</tr>
<tr>
<td>Land use</td>
<td>Traffic land, resident land, dry land, reservoir land, paddy rice, river, lake, grassland, wetland, meadow, coniferous forest, broad leaved forest, canal, mixed broadleaf-conifer forest</td>
<td>30</td>
<td>Interpreted from Landsat 8 OLI image [<a href="http://glovis.usgs.gov/">http://glovis.usgs.gov/</a>]</td>
</tr>
<tr>
<td>NDVI</td>
<td>–0.30 to 0.94</td>
<td>250</td>
<td>MODIS products (MOD13) [<a href="https://ladsweb.nascom.nasa.gov/">https://ladsweb.nascom.nasa.gov/</a>]</td>
</tr>
<tr>
<td>Annual mean precipitation (mm)</td>
<td>405 to 596</td>
<td>–</td>
<td>China Meteorological Data Service Centre [<a href="http://www.nmic.cn/">http://www.nmic.cn/</a>]</td>
</tr>
<tr>
<td>Distance from rivers (m)</td>
<td>367 to 24,492</td>
<td>–</td>
<td>Interpreted from Landsat 8 OLI image</td>
</tr>
<tr>
<td>Distance from existed gullies (m)</td>
<td>0 to 5,000</td>
<td>–</td>
<td>Interpreted from Landsat 8 OLI image</td>
</tr>
</tbody>
</table>

Notes: DEM = digital elevation model. TWI = topographic wetness index. NDVI = normalized difference vegetation index.

Topographic attributes play an important role in the expansion and control of gully erosion by indirectly affecting the vegetation and precipitation characteristics (Nhu et al. 2020). In this study, the digital elevation model (DEM) data were obtained from the Advanced Land Observing Satellite (ALOS) phased array type L-band synthetic aperture radar (PALSAR) sensor, with a resolution of 12.5 m. The slope and aspect can indirectly affect the erosion process by controlling the solar exposure, vegetation type, and soil moisture (Zabibi et al. 2018). The plan and profile curvature affect the divergence or convergence of water during its descent (Avand et al. 2019). The TWI also has an impact on the morphometric and hydrological characteristics. Thus, slope, aspect, plan curvature, profile curvature, and TWI maps were derived from the DEM data using the ArcGIS 10.6 software. The TWI is expressed as equation 1 (Moore et al. 1991):

\[ TWI = \ln(As / \tan \theta) \]  

where \( As \) is the cumulative catchment area \( (m^2 \cdot m^{-1}) \) of a point, and \( \theta \) is the slope at this point. The soil properties affect the flow of the current subsurface and the occurrence of piping erosion, which can cause gully erosion (Rahmati et al. 2016). In this study, a soil type map of the study area was obtained based on the Second National Soil Survey of China. Land use is an important factor in the occurrence of gully erosion due to its impact on the vegetation conditions. In this study, a land use map was interpreted from GF-1 images acquired in 2021 (figure 1) using the object-based classification method in the eCognition 9.0 software. There were 14 land use types in the Second National Soil Survey of China, defined as equations 2 and 3:

\[ Tolerance = 1 - R^2_M \]  

\[ VIF = 1 / \text{Tolerance} \]
where $R^2_j$ is the regression coefficient of the determination of factor $J$ on all of the other indicators. Tolerance > 0.10 and VIF > 5 indicate a multicollinearity problem (Pourghasemi et al. 2017b).

**Machine Learning Models for Gully Erosion Susceptibility Prediction.** The structures of machine learning models are different and somewhat complex, and in general, these models have various accuracies and transparencies. For example, the SVM has the potential to solve the overfitting problem when using high-dimensional data (Chilingaryan et al. 2018). Neural networks have a good ability to model complex nonlinear relationships between dependent and independent variables (Cui et al. 2018). The RF and XGBoost models are both tree-based models, and they can obtain the relative importance of the predictors (Zhang et al. 2021). The goal of this study was to identify a more widely applicable type of model for GES prediction. Thus, we considered and compared the currently widely used types of machine learning models, including supervised learning, multilayer perceptrons, and tree-based methods, for predicting areas susceptible to gully erosion. The most well-known models of these types are the SVM, MLPNN, RF, and XGBoost algorithms.

**Support Vector Machine.** SVM regression is a supervised, nonparametric statistical learning method, and it uses structural risk minimization (SRM) to obtain an optimal overall response (Vapnik 1995). The SVM model can handle a high-dimensional multivariate space (Karatzoglou et al. 2006). It includes four types of classification functions (i.e., linear, polynomial, radial, and sigmoid) and each kernel has its own optimization parameters (Joachims 1997). There are two main features that need to be optimized: one is the selection of the kernel function, and the other one is the noise tolerance of each kernel. The information regarding the specific SVM parameters used for the GES prediction has not been explored as deeply.

**Multilayer Perceptron Neural Network.** An ANN can build the complex and functional relationships between different variables without setting any null hypothesis (Yang et al. 2017). The MLPNN is one of the most commonly used neural models for prediction, and it uses the back-propagation technique for learning. It is interconnected by connecting forces represented by the synaptic weights (de Oliveira et al. 2021). The MLPNN contains three layers: a first input layer, a middle hidden layer, and a last output layer. After the data set is input into the model, the training process is repeated until the error is reduced to an acceptable value.

**Random Forest.** RF regression is a tree modeling decision-making method, which combines the anticipation of several single algorithm using rules (Breiman 2001). Three user-defined parameters, including the number of trees (ntree), the number of variables used as predictors for each tree (mtry), and the minimum size in each terminal mode (node size), are contained in the RF model (Friedman and Meulman 2003). In this study, the ntree was set as 500 to 2,000, and the optimal ntree was selected to stabilize the error. In general, the mtry is set to one-third of the number of selected indicators (Zhang et al. 2019).

**Extreme Gradient Boosting.** The XGBoost model is also a decision tree-based model, and it is a modification of the gradient boosting model. Compared with the RF model, the XGBoost model grows each tree based on the residuals of the previous tree, and some studies have found that it outperformed RF in regression models (Stojic et al. 2019; Huang et al. 2020). This algorithm is generally applied in classification, credit rating, and prediction studies (Cherif and Kortebi 2019). Among the existing machine learning methods, this algorithm has gained popularity because of its generalizability, low risk of overfitting, and high interpretability (Liu et al. 2021). The XGBoost model outperforms other machine learning models in terms of its predictive capability, and it has also been used to accurately estimate the aboveground biomass of maize (Zea mays L.) (Zhang et al. 2021).

The contributions of the indicators of gully erosion can also be calculated quantitatively using the RF and XGBoost models. In this study, the importance of the factors in the gully erosion prediction was determined according to the relative importance, which was calculated through comprehensive evaluation (equation 4) (Niu et al. 2021).

$$w_i = 0.5 \times A_i / \sum_{i=1}^{n} A_i + 0.5 \times B_i / \sum_{i=1}^{n} B_i , \quad (4)$$

where $w_i$ is the relative importance of factor $i$, $A_i$ and $B_i$ are the %IncMSE and IncNodePurity values of factor $i$, respectively; and $n$ is the total number of factors. %IncMSE is the increase in the percentage of the mean squared error, $i.e.$, the value by which the accuracy of the prediction decreases after the variable is removed. IncNodePurity is the increase in the node purity. A larger node purity indicates a more important variable (Angermueller et al. 2016).

In general, machine learning models depend on databases composed of point samples, which represent the occurrence of gullies and the nonoccurrence of gullies. The sample points containing gullies were derived from the interpreted GF-1 imagery. The sample points without gullies were obtained using the random sampling tool in ArcGIS 10.6.

**Model Validation and Accuracy Assessment.** To compare the reliabilities of the predictions of the different models, the coefficient of determination ($R^2$), root mean square error (RMSE), and mean absolute error (MAE) were selected to compare the predictions of the different models (Paul et al. 2019; Pourghasemi et al. 2020). The $R^2$ value is usually used as an index of the goodness-of-fit and is computed to ensure the fit of the predicted outcomes (Harel 2009). The RMSE can be used as a benchmark for crosschecking the inaccuracies of models (Tien Bui et al. 2016). However, when the model contains large values and outliers, the RMSE becomes susceptible to imprecision (Chai and Draxler 2014). Thus, the MAE was also employed in this study. In this study, these three parameters were used to identify the best models for gully erosion prediction. They were calculated as equations 5 through 7:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (z_i - z_i^*)^2}{\sum_{i=1}^{n} (z_i - \bar{z})^2} , \quad (5)$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (z_i - z_i^*)^2}{n}} , \quad \text{and}$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |z_i - z_i^*| , \quad (7)$$

where $z_i$ and $z_i^*$ are the observed values and predicted values at site $i$, $\bar{z}$ is the average of the observed values, and $n$ is the number of samples. Then, the best model was selected to conduct the GES prediction in the study area. In addition, the performances of the four models were assessed based on the receiver operating characteristic (ROC) curve, which is a typical measure for assessing the outcomes of the application of machine learning models. The area under the curve (AUC) represents the model's performance.
Values close to 1 indicate a perfect model, while values close to 0.5 indicate an inaccurate model (Chung and Fabbri 2003).

Results and Discussion

Extraction of Indicators. To prepare the GES map for Hailun, we employed 12 indicators. Based on the above methods, these indicators were extracted. The results are presented in figure 3. In figure 3, most of the subfigures exhibit different spatial patterns, providing environmental indicators for gully erosion prediction from multiple aspects.

Multicollinearity Test. The multicollinearity assessment was conducted to exclude biased variables and select appropriate indicators for the modeling. The multicollinearity analysis results for the 12 indicators are presented in table 2. The multicollinearity of the selected variables was analyzed based on the VIF and tolerance limit. The lowest VIF represents the highest tolerance levels. The lowest and highest VIF were related to the distance from existing gullies (1.046) and slope (2.020). The 12 indicators all had VIFs of less than 5 (the VIF limit). Therefore, there was no multicollinearity among the 12 factors.

Comparison of the Prediction Models. By comparing the four models (SVM, MLPNN, RF, and XGBoost) for GES prediction, it was found that the XGBoost model was the best model for predicting GES because it had the lowest RMSE (0.60) and MAE (0.50) values and the highest R² value (0.81) (table 3). In addition, the RF model also had a good performance, with R² = 0.78, RMSE = 0.61, and MAE = 0.50. The SVM model had the poorest performance, with R² = 0.62, RMSE = 0.70, and MAE = 0.50. The performances of the models were also evaluated based on their ROC curves (figure 4). Figure 4 clearly shows that the SVM, MLPNN, RF, and XGBoost models all had AUROC values of 0.832, 0.867, 0.966, and 0.985, respectively. The XGBoost and RF models outperformed the other two models, achieving the best results in terms of their abilities to predict the occurrence and absence of gullies. Therefore, we selected the XGBoost and RF models as the final prediction models for the GES mapping.

Relative Importance of Indicators. The relative importance can reflect the influence of each controlling factor on the GES prediction. In this study, both the RF and XGBoost models had good performances, so only these two models were used to calculate the contribution of each factor. Thus, the relative importance of the indicators of gully erosion were obtained based on the combination of the importance scores calculated using the RF and XGBoost models (figure 5).

As figure 5 shows, the distance from existing gullies was the most important of the 12 indicators, and its relative importance was greater than 35%. The profile curvature, plan curvature, land use, elevation, and soil type were the second most important factors controlling gully erosion, with contributions of 10% to 15%. The other six indicators, including slope, river distance, precipitation, aspect, NDVI, and TWI, had the smallest impacts on gully erosion, with contributions of less than 5%.

Mapping of Areas Susceptible to Gully Erosion. The areas susceptible to gully erosion were predicted using the four different machine learning methods, but only the two prediction results obtained from the models with good performances are presented. Figure 6 presents the GES maps for Hailun City obtained using the XGBoost and RF models. The GES in the study area was classified into four levels: very high, high, medium, and low level. The spatial heterogeneities of the GES presented by the two GES maps obtained using the XGBoost and RF models are consistent. These GES maps indicate that the central part of Hailun was more susceptible to gully erosion than the northeastern and southwestern parts. In addition, the areas and percentages of the four levels for GES are shown in table 4. Most of the study area was predicted to have a medium susceptibility to gully erosion (53.82% to 61.64%), while the areas with low (8.12% to 11.59%) and very high (5.32% to 5.78%) susceptibility levels were small. The area with a high level of susceptibility was fairly large (24.92% to 28.81%).

Discussion. The intensive gully erosion in the Mollisol region in northeastern China, which is an important region of crop production in China, has caused serious soil loss and cropland degradation (Gao et al. 2015). Thus, it is essential to identify the areas susceptible to gully erosion and to slow down the erosion process in this region. In general, the selection of effective indicators of gully erosion is a crucial step in identifying the areas susceptible to gully erosion. However, there are no fixed rules for choosing suitable indicators for GES prediction (Gutiérrez et al. 2009; Chen et al. 2021). Based on the above-mentioned criteria and multicollinearity test results (table 2), all 12 indicators were retained and employed in the model for predicting areas susceptible to gully erosion.

In recent years, different researchers have tried to predict areas susceptible to gully erosion by incorporating different types of algorithms, generally including traditional statistical and novel machine learning methods (Garosi et al. 2018; Soleimanpour et al. 2021; Chen et al. 2021; Mokarram and Zarei 2021). However, the use of machine learning methods has been reported to be a reliable approach for GES prediction (Bunker and Thabtah 2019; Chen et al. 2021). For example, Soleimanpour et al. (2021) revealed that the quick, unbiased, efficient statistical tree (QUEST) model outperformed other prediction models, such as the frequency ratio (FR) model and the evidential belief function (EBF) model, and this model achieved the highest area under the receiver operating characteristic (AUROC) curve and true skill statistic (TSS) values of 83.2% and 0.63, respectively. Chen et al. (2021) compared different machine learning models for predicting areas susceptible to gully erosion and found that the deep boost (DB) model had a significantly higher accuracy than other models, including the boosted tree (BT) model, boosted generalized linear model (BGLM), and BRT model.

As the results in this study show, the XGBoost model outperformed the other three algorithms (SVM, MLPNN, and RF), and its ability to predict areas susceptible to gully erosion in the study was demonstrated (table 3 and figure 4). In addition, based on the comparison of the four gully erosion prediction models, the tree-based machine learning models (such as the RF and XGBoost) had better GES prediction performances than the other models. The SVM model only has three hyperparameters for tuning, which leads to its weakness in fitting the complex relationships between predictors and indicators. Thus, preanalysis has been highly recommended, which can be regarded as a baseline for comparison with other models (Tang and Na 2021). The MLPNN has a higher model generalization ability due to its multilayered architecture, but as the number of hidden layers increases, the estimation behind the network becomes more complex and opaque (Mittendorf et al. 2022). The nonparametric tree-based models have the advantages of resistance to
Figure 3
Indicators extracted in this study: (a) elevation, (b) slope, (c) aspect, (d) plan curvature, (e) profile curvature, (f) topographic wetness index, (g) soil type, (h) land use, (i) normalized difference vegetation index, (j) precipitation, (k) distance from rivers, and (l) distance from gullies.
height, distance from a fault, slope, and index of connectivity were the most important factors in gully erosion prediction. Pourghasemi et al. (2020) showed that the distance from rivers was one of the most effective factors in determining areas prone to gully erosion. Chen et al. (2021) reported that elevation was the most important factor in GES modeling. There are some consistencies between the results of this study and those of other studies on GES prediction. For example, in this study, the topography-related factors were also identified as important factors for GES prediction. Nevertheless, the distance to existing gullies has seldom been selected as an indicator in GES prediction studies. However, in this study, it was found that this indicator played the most significant role (relative importance of >35%) in the GES prediction. Existing gullies play an important role in changing the surface hydrological processes and also provide a platform for accumulated precipitation to be transformed into runoff. Therefore, it was concluded that obtaining the spatial distribution of the existing gullies is essential for GES prediction. The distribution of the existing gullies provided a base map for the gully erosion conditions in the study area, and the areas susceptible to gully erosion were identified based on the existing gullies. However, this indicator is not suitable for modeling gully erosion prediction in other areas without data for existing gullies, and thus, it would not necessarily be the most essential indicator in gully prediction in other study areas.

Indeed, the gully erosion was also affected by other environmental factors. For example, the spatial distribution of the accumulated precipitation is modified by the existing gullies, rivers, and road networks in a specific region. In this study, the land use type also had a significant impact on the gully erosion prediction, which is consistent with the results of similar studies. Mokarram and Zarei (2021) reported that land use intensified the gully erosion process. Chen et al. (2021) showed that land use had the greatest impact on the GES. In particular, densely vegetated regions are less prone to soil erosion, while barren and sparsely vegetated regions are more prone to soil erosion. The soil properties play a significant role in gully erosion prediction because different soil types have various soil and water retention abilities.

### Summary and Conclusions

In this study, four different machine learning algorithms were applied to investigate the factors controlling gully erosion and to predict the areas susceptible to gully erosion in Hailun in the Mollisol region in northeastern China. Accordingly, 12 factors, including topography, soil, land use, vegetation, and climate-related variables, were selected to investigate the main factors controlling gully erosion. The contributions of the controlling factors and the areas susceptible to gully erosion in Hailun City were determined using four machine learning algorithms, including the SVM, MLPNN, RF, and XGBoost models. In this study, the areas susceptible to gully erosion were predicted with an adequate accuracy using the different machine learning models. Through comparison of the four prediction models, it was found that the XGBoost and RF models both achieved good performances in predicting the areas susceptible to gully erosion.

In addition, the selection of the most relevant environmental indicators for predicting areas susceptible to gully erosion is challenging due to the anthropogenic alterations of the land resources. In this study, it was found that the most important factors controlling gully erosion were the gully distance, profile curvature, plan curvature, land use, elevation, and soil type. The central part of the study area was more susceptible to gully erosion than the other areas, which is consistent with the visual interpretation results.

In future studies, the indicators (e.g., DEM and NDVI data) with a high spatial resolution will be used to improve the accuracy of the prediction of the areas susceptible to gully erosion. In addition to the 12 indicators selected in this study, other variables such as the soil texture, electrical conductivity, and run-off speed will be employed to identify the areas susceptible to gully erosion. Owing to the complex relationships between gully erosion and the controlling factors, other types of machine learning algorithms, especially deep learning models (e.g., convolutional neural networks), should be applied in GES prediction.

### Acknowledgements

The work was supported by the National Key R&D Program of China (2021YFD1500800), the National Natural Science Foundation of China (U19A2061), the Science and Technology Project for Black Soil Granary (XDA28080500), and the National Earth System Science Data Center. We thank LetPub (www.letpub.com) for its linguistic assistance during the preparation of this manuscript.

### References


#### Table 2

<table>
<thead>
<tr>
<th>Factors</th>
<th>Tolerance</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elevation</td>
<td>0.638</td>
<td>1.567</td>
</tr>
<tr>
<td>Slope</td>
<td>0.495</td>
<td>2.020</td>
</tr>
<tr>
<td>Aspect</td>
<td>0.949</td>
<td>1.054</td>
</tr>
<tr>
<td>Plan curvature</td>
<td>0.707</td>
<td>1.414</td>
</tr>
<tr>
<td>Profile curvature</td>
<td>0.645</td>
<td>1.550</td>
</tr>
<tr>
<td>TWI</td>
<td>0.810</td>
<td>1.235</td>
</tr>
<tr>
<td>Soil type</td>
<td>0.918</td>
<td>1.089</td>
</tr>
<tr>
<td>Land use</td>
<td>0.691</td>
<td>1.447</td>
</tr>
<tr>
<td>NDVI</td>
<td>0.886</td>
<td>1.129</td>
</tr>
<tr>
<td>Precipitation</td>
<td>0.868</td>
<td>1.152</td>
</tr>
<tr>
<td>Distance from river</td>
<td>0.870</td>
<td>1.149</td>
</tr>
<tr>
<td>Distance from existing gullies</td>
<td>0.956</td>
<td>1.046</td>
</tr>
</tbody>
</table>

Notes: VIF = variance inflation factor. TWI = topographic wetness index. NDVI = normalized difference vegetation index.

#### Table 3

<table>
<thead>
<tr>
<th>Model</th>
<th>R²</th>
<th>RMSE</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>0.62</td>
<td>0.70</td>
<td>0.50</td>
</tr>
<tr>
<td>MLPNN</td>
<td>0.65</td>
<td>0.70</td>
<td>0.49</td>
</tr>
<tr>
<td>RF</td>
<td>0.78</td>
<td>0.61</td>
<td>0.50</td>
</tr>
<tr>
<td>XGBoost</td>
<td>0.81</td>
<td>0.60</td>
<td>0.50</td>
</tr>
</tbody>
</table>

Notes: RMSE = root mean square error. MAE = mean absolute error. SVM = support vector machine. MLPNN = multilayer perceptron neural network. RF = random forest. XGBoost = extreme gradient boosting.


Figure 4
Receiver operator curve (ROC) and area under the curve (AUC) values for the four models ([a] support vector machine [SVM], [b] multilayer perceptron neural network [MLPNN], [c] random forest [RF], and [d] extreme gradient boosting [XGBoost]).

(a) SVM = 0.832
(b) MLPNN = 0.867
(c) RF = 0.966
(d) XGBoost = 0.985
Figure 5
Relative importance of the indicators for gully erosion prediction.

Indicators

- Gully distance
- Profile curvature
- Plan curvature
- Land use
- Elevation
- Soil type
- Slope
- River distance
- Precipitation
- Aspect
- NDVI
- TWI


Mokarram, M., and A.R. Zarei. 2021. Determining prone areas to gully erosion and the impact of land use change on it by using multiple-criteria decision-making algorithm in arid and semi-arid regions.


Figure 6
Gully erosion susceptibility (GES) maps obtained using the (a) extreme gradient boosting (XGBoost) and (b) random forest (RF) models.

Table 4
Area of each gully susceptibility levels in extreme gradient boosting (XGBoost) and random forest (RF) model.

<table>
<thead>
<tr>
<th>Model</th>
<th>Class</th>
<th>Area (km²)</th>
<th>Percentage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>XGBoost</td>
<td>Very high</td>
<td>248.28</td>
<td>5.32</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>1,163.02</td>
<td>24.92</td>
</tr>
<tr>
<td></td>
<td>Medium</td>
<td>2,876.74</td>
<td>61.64</td>
</tr>
<tr>
<td></td>
<td>Low</td>
<td>378.96</td>
<td>8.12</td>
</tr>
<tr>
<td>RF</td>
<td>Very high</td>
<td>269.75</td>
<td>5.78</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>1,344.56</td>
<td>28.81</td>
</tr>
<tr>
<td></td>
<td>Medium</td>
<td>2,511.78</td>
<td>53.82</td>
</tr>
<tr>
<td></td>
<td>Low</td>
<td>540.91</td>
<td>11.59</td>
</tr>
</tbody>
</table>


