Point and pixel inclusive machine learning models for exploring gully erosion susceptibility

Swades Pal*, Sandipta Debanshi* and Satyajit Paul

*Department of Geography, University of Gour Banga, Malda, West Bengal, India; \( ^{b} \)Department of Geography, Gour Mahavidyalaya, Malda, West Bengal, India

ABSTRACT
Sample point based spatial model derived from Machine Learning (ML) algorithms often generalizes the spatial pattern of an event. The present study has tried to highlight how far it is acceptable and can it replace with pixel based modeling? The present study has presented a comparative view of pixel and sample point based modeling of gully erosion susceptibility of the upper Mayurakshi basin to assess the predictabilities. Random forest (RF), Support Vector Machine (SVM), and (ADB) have been applied for developing pixel and point based models in Python and WEKA software environments respectively. The models show that 14–25% area located mainly in the upper parts of the study unit is very highly susceptible to gully erosion. Based on the accuracy and performance level using area under curve (AUC) of Receiver operating curve, sensitivity, precision, F1 score, MCC, pixel based ensemble models are superior to point based modeling. The point-based models often have an inferior agreement with training and testing data. So, pixel based models could not be replaced with point based models. RF is found as the best representative model. The study, therefore, recommends using pixel based modeling for this or a similar purpose. Since the models have figured out the gully erosion susceptible areas, it would be a useful tool for related planning processes.

1. Introduction
Machine Learning (ML) algorithms based on spatial models coupled with and Geographical Information System (GIS) has been used to address different issues in a wide array of disciplines, including epidemiology (Auchincloss et al. 2012), climate change (Mansfield et al. 2020), natural resource (Frey 2020), environmental vulnerability (Pal and Debanshi 2021), environmental hazards (Wang et al. 2021). Gully erosion is one of the major issues that have become a global concern with the increasing coverage of degraded land worldwide (Secretariat of the UNCCD 2013). The global coverage of degraded land is about 30% (Nkonya et al. 2016) and about 44% in India (Mythili and Goedecke 2016). Soil erosion by water, especially in the gully, is the most common soil degradation process...
Gully erosion featuring deep channels eroded by running water causes that degradation locally by forming badlands over lateritic terrain (Tang et al. 2013). Gully erosion is capable of causing acute problems by causing the formation of badlands (Cánovas et al. 2017; Wuepper et al. 2021), eroding fertile soil of agricultural land (Han et al. 2018), clogging sediment in a reservoir (Dutta 2016), depleting of groundwater table ( Tilahun et al. 2016). Thus, sustainable practices for soil conservation should be adopted to secure future development ( Vanmaercke et al. 2016; Arabameri et al. 2020). The present study area, the Mayurakshi river basin, suffers from the severe problem of gully erosion due to its geo-environmental location over the degraded eastern fringe of the Chottanagpur plateau (Bhattacharyya 2014; Debanshi and Pal 2020). Gully erosion susceptibility mapping is essential for initiating such conservation practice (Zabihi et al. 2018; Debanshi and Pal 2020). A gully erosion susceptibility mapping can be done by detecting the relationship between the existing gully occurrences and geo-environmental gully conditioning factors (Nicu 2021; Rahmati et al. 2017).

ML algorithms are becoming increasingly applied tools for assessing gully erosion and are considered robust over the traditional statistical tool for their accurate predictability (Azareh et al. 2019; Arabameri et al. 2020; Saha et al. 2020). Several software and programming languages like R studio ( Arabameri et al. 2018; Hosseinalizadeh et al. 2019a), Waikato Environment for Knowledge Analysis (WEKA) ( Hosseinalizadeh et al. 2019b), Python ( Wang et al. 2020; Can et al. 2021) are used for implementing the ML algorithms. Several studies reported the robustness of ML models over the multi-criteria decision approach models, especially when feature space is complex and the target variable is involved with the input dataset in a non-linear relationship ( Rodriguez-Galiano et al. 2012; Churpek et al. 2016; Xie et al. 2018; Jung and Lee 2019). ML algorithms have the potential to identify the complex non-linear relationship ( Mosavi et al. 2018) and eventually produce the spatial model of gully erosion susceptibility based on associated predisposing factors of gully erosion (Saha et al. 2020; Borrelli et al. 2022). Integrating remote sensing and GIS with ML and ensemble algorithms has made mapping any environmental issue, including gully erosion, easier and more precise (Pal and Paul 2020; Saha et al. 2020; Rafiee Sardooi et al. 2021). For example, Arabameri et al. ( 2018) and Hosseinalizadeh et al. (2019a) used R programming language and GIS technique to assess gully erosion susceptibility and reported greater performance accuracy and preciseness than WEKA is a software that provides a set of ML and ensemble algorithms and leverages data mining and model building (Ahmadpour et al. 2021). Many studies have used the WEKA tool and GIS for developing spatial models (Oh et al. 2019; Hosseinalizadeh et al. 2019b; Pham et al. 2020). Since the software provides a set of already programmed algorithms, the parameters can be optimized to achieve the best performance (Pal et al. 2020; Pham et al. 2020). Recently, the python programming language and associated libraries are also used for developing ML-based predictive spatial models (Wang et al. 2020; Arabameri et al. 2021; Can et al. 2021).

In this context, in recent times, the generalization of the spatial predictive model developed based on ML and Deep Learning (DL) has been in discussion ( Maxwell et al. 2016; Neyshabur et al. 2017; Kawaguchi et al. 2017; Maxwell et al. 2020a, 2020b). This generalization was sometimes the different geographical regions from which the model has been trained and applied (Maxwell et al. 2020b) or sometimes the disparate data set ( Hoese and Kuenzer 2020; Hoese et al. 2020). However, in most cases, the spatial models integrated with ML algorithms and GIS limit the type of analysis performed (Hogland and Anderson 2017). The process of such model building involves multiple sequential steps viz. (1) building dataset using GIS, (2) importing the dataset in ML algorithm performing software, (3)
training of the model using a set of samples of that imported dataset where the target variable is known and define a relationship between target variable and explanatory variable, (4) applying the model on the rest of the dataset, and then (5) producing a spatially explicit surface in GIS environment using the output of the predictive model. Producing final outputs in this manner is often challenged by multiple issues like learning multiple software, importing and exporting a large dataset, long processing time, and the requirement for more extensive storage space (Hogland and Anderson 2014, 2015). While producing spatial models of a larger geographical area, these limitations intensify, mainly because of the issue of storage space. Since the remote sensing datasets are becoming more refined day by day, a large dataset is needed to be handled to take all the pixels of a large geographical region into account (Hogland and Anderson 2017). Model building using a previously programmed set of algorithms like the WEKA tool suffers from limitations due to its lesser storage capacity (Naudts et al. 2004). To overcome this issue, sample point based model building is practised. This method has a fast model building benefit but often gets subjected to generalizing or smearing the nuances of the place located further from the chosen sample sites or points. This problem reduces the precision of the mapping, and eventually, the initiation of the conservation practice is compromised. This is an acute problem and could be resolved by taking all the pixels of the study area into consideration while producing any model. Python libraries give such an opportunity to handle large datasets containing all the pixels of a larger geographical region. It is an open space, and the algorithms need to be programmed (Brownlee 2016), making it difficult for non-expert users. In addition, machine learning needs hyper-parameter optimization to increase the prediction accuracy, which requires additional computation (Kotthoff et al. 2019). Hyper-parameters are the parameters that are set before the training process, and optimization can be achieved by manual search or automatic search method (Pradhan et al. 2021). The auto-optimization function of the WEKA tool, to some extent, eases the hyper-parameter optimization process (Kotthoff et al. 2019), but it has limited scope, and it optimizes the hyper-parameters up to 70% (Brownlee 2019). Python library enables to overcome this limitation with a broader spectrum of hyper-parameters optimization (Brownlee 2016). Therefore, python libraries are considered superior for optimizing the hyper-parameters than WEKA (Mitrpanont et al. 2017). Here the research gap could be pointed out whether the time and expertise expensive laborious work of pixel inclusive modelling are essential to replace the representative sample point based modelling for spatial analysis. If yes, then to what extent the precision of the pixel inclusive models greater than the time effective and fast approach of representative sample point based modeling. To answer this query, the pros and cons of sample point based modeling using WEKA tool and pixel inclusive modeling using Python libraries have been decided. Considering this issue and addressing the aforementioned research gap, the present study has intended to examine the better credential of Sample Point Based Modeling (SPBM) and Pixel Based Modeling (PBM) to explore gully erosion susceptibility.

2. Study area

The Mayurakshi river basin is located over the Chottanagpur plateau fringe region of eastern India (Figure 1a) between 23°15’N to 24°34’15”N latitude and 86°58’E to 88°20’30”E longitude and covers more than 5403 km² (Figure 1b and 1c). The entire basin consists of the Rarh tract dominated by secondary lateritic formation. The rivers that originated from the Chottanagpur Plateau took part in such formation (Jha and Kapat 2003, 2009; Pal 2016). Geologically (Figure 1b), Granitic gneiss is a dominant rock type that belongs to the Pleistocene age that covers most of the upper and middle catchment of the
basin (Figure 1e and 1f). Granet-biotite gneiss is also found over a considerable area in the north-eastern part of the basin. Geomorphologically (2c), the upper and middle catchment of the basin consists of a denudational origin pediment-pedeplain complex, and in a few patches, moderately dissected hills and valleys originating from the denudational process are found. On the other hand, lower catchment mostly belongs to flood plain geomorphologically comprised of fluvial origin older alluvial plain, older flood plain, and active flood plain. Geologically also, this part of the basin belongs to newer and older alluvium. Eventually, this part of the basin was not subjected to severe soil erosion;

Figure 1. Geographical setup of the study area. (a) India, (b) geomorphological divisions, (c) geological division, (d) area of interest with training sites, (e) elevation of the study area, (f) hillshade of the study area.
instead, an alluvial deposition by the western tributaries took part behind the formation of this part (Bandyopadhyay et al. 2014). Apart from that, the confluence part of the eastern side of the basin is merged with the Ganges delta. Almost all the gully occurrences are found in the upper and middle part of the basin. In many cases these gullies appear as 1st and 2nd order streams. Araujo and Pejon (2015) and Joshi et al. (2016) reported the same in their study area also. In this part the numbers of 1st and 2nd order streams are respectively 4827 and 1091 (extracted from SOI topographical maps). Geologically this part belongs to Granitic gneiss formation and Granet-biotite gneiss formation and geomorphologically belongs to denudational origin pediment-pedeplain complex and denudational origin moderately dissected hill and valleys. Therefore, these geological and geomorphological divisions have been taken as an area of interest in this study, and the lower part of the basin has been excluded from the investigation (Figure 1b–d). This area of interest covers more than 3500 km² which is about 65% of the entire basin area. Figure 2 presents the channel network of some gully erosion prone sites.

Figure 2. Channel network of some gully erosion prone sites, (a–l) extracted from Google earth imagery and (j–m) extracted from SOI topographical maps.
3. Materials and methods

3.1. Materials

For this study, 18 data layers have been prepared using various data sources like USGS Landsat images with 30 m spatial resolution (path/row: 139/43 and 139/44) to prepare land use/land cover (LULC), Ferrous mineral index, Normalized Difference Vegetation Index (NDVI), Bare Soil Index (BSI), Water Supplying Vegetation Index (WSVI) and monthly fluctuation of a temperature map of the basin area. The map of geological divisions has been taken from the Geological Survey of India, whereas BHUVAN (an Indian web-based utility) has been subscribed for extracting the geomorphological map of the study area. Mouza-wise soil texture map has been derived from National Informatics Centre (NIC) to prepare a soil erosivity map. The average monthly and annual rainfall map of the Department of Science and Technology (DST) has been employed in this study for rainfall information and its further derivations. The spatial data layers of the slope, aspect, and stream power index (SPI) have been prepared using SRTM DEMs with 30 m spatial resolution. Survey of India (SOI) toposheets of 1972 (scale 1:50,000) have been used for developing the base map of the study, along with preparing the maps of the drainage network and water covered area and extracting the location of gully head bundh (Figure 3). Recent Google Earth imagery (22/10/2021) has also been subscribed for extracting the location of check dams. All Landsat imageries and SRTM DEMs have been preprocessed before final layer preparation.

3.2. Software taken for PBM and SPBM

Generally, SPBM is considered a cost-effective and quick approach for environmental studies (Chabuk et al. 2020). Representation of the surface dynamics through these sample points minimizes the variability between two subsequent points and generates specifiable continuity in the surface (Zhu et al. 2020). PBM is another approach where all the pixels are considered for the generation of the final output (Riyadi et al. 2020). In pixel based modelling, the model has been applied to a standard dataset (70% for training and 30% for validation), and the final trained model has been applied to the entire dataset or pixel (Zhu et al. 2020). There are different platforms available for sample point-based modelling (SPBM) and pixel based modelling (PBM), such as Google Cloud ML Engine, WEKA, Amazon Machine Learning (AML), Shogun, Apache Spark MLlib, Pytorch, Keras, TensorFlow, R-studio (González et al. 2020). These platforms run various programming languages like Python, Java, C++, and R to run different machine learning algorithms. Recently, Python and R have been used to model various environmental phenomena for efficiency and big data handling capability. Whereas software like WEKA and KNIME is used to run relatively small sample point datasets. For this study, we have used WEKA for sample point based modelling (SPBM), and Python has been used for pixel based modelling (PBM).

3.3. Preparing data layers for contributing factors of gully erosion susceptibility

Total eighteen factors under four-factor clusters like (1) erodibility (soil texture, LULC, geological aspect, BSI, ferrous mineral, geomorphology) (2) erosivity (soil erosivity, yearly average temperature, monthly fluctuation of temperature, WSVI) (3) resistance (NDVI, distance from GHB, water coverage, distance from check dams) and (4) topographical factors (slope, curvature, SPI, distance form 1st and 2nd order stream) have been taken for
Figure 3. Raster data layers for the respective factors, (a) soil texture, (b) LULC, (c) geological aspect, (d) BSI, (e) ferrous mineral, (f) geomorphology, (g) soil erodivity, (h) yearly average rainfall, (i) monthly fluctuation of surface temperature, (j) WSVI, (k) NDVI, (l) distance from GHB, (m) water coverage, (n) distance from check dams, (o) curvature, (p) SPI, (q) distance from first- and second-order stream, (r) slope.
gully erosion susceptibility mapping. These factors have been selected based on their importance in determining the gully erosion of the study region and the taking experience from the existing literature like Debanshi and Pal (2020) and Saha et al. (2020). Spatial data layers have been prepared for each factor. After preprocessing the SRTM, DEM has been employed to derive the data layers of slope, curvature, and SPI. Several studies (Meliho et al. 2018; Shit et al. 2020) have used these parameters for measuring soil and gully erosion. Slope, SPI directly influences the erosion rate and also, to some extent, determines the distribution of gullies (Kertész and Gergely 2011). Gullies expand rapidly over the steep slope compared to flat land (Marden et al. 2012). The SPI indicates the erosive power of water, considering the discharge proportional to the catchment area (Debanshi and Pal 2020). The surface analysis tool of ArcGIS software (v-10.2) has been used to facilitate slope and curvature derivation (Figures 3r and 2o), while the raster calculator tool has been used to calculate SPI (Figure 3p) of the basin area with the help of Eq. (1) (Conforti et al. 2011). Similarly, satellite imagery of the Landsat 8 series has been used to derive the data layers of LULC, BSI, ferrous mineral, monthly fluctuation of surface temperature, NDVI, and WSVI. LULC of the region can influence the gully development in many ways (Gelagay and Minale 2016). The bare ground is more prone to gully development than a dense forest (Feizizadeh et al. 2013). Supervised classification based on maximum likelihood classifier has been carried out in ERDAS Imagine software (v-9.2) for LULC identification and mapping (Figure 3b). Since the bare ground positively contributes to gully development of its lower resistance rate (Jahantigh and Pessarakli 2011). Equation (2) has been implemented in the raster calculator for deriving BSI (Elfadaly et al. 2017) (Figure 3d). The inbuilt function of the ERDAS Imagine software has been used to produce the data layer of ferrous content (Figure 3e). It has been recognized as a predictor of gully erosion because of the capability of ferrous minerals to initiate chemical weathering in the soil. Studies like (Jha and Kapat 2003, 2009) recognized this factor for gully development in this region. The temperature fluctuation is the key factor for the mechanical disintegration of rock (Bandfield et al. 2011), and it could be an important predictor of gully development. Monthly Land Surface Temperature (LST) has been calculated (Eq. (3)) following the guideline provided by Landsat Project Science Office (2002), and its Coefficient of Variation (CV) has been calculated for the spatial measurement and preparation of monthly fluctuation of surface temperature data layer (Figure 3i). Keeping the ability of denser forests to protect the soil surface from being eroded, the NDVI has been frequently incorporated (Arabameri et al. 2018; Roy et al. 2020; Zhou et al. 2021) in gully erosion studies. Calculation of NDVI (Eq. (4)) (Townshend and Justice 1986) and spatial mapping (Figure 3k) have been performed in the raster calculator of ArcGIS. The spatial data layers of LST and NDVI have been used to calculate the WSVI (Eq. (5)), which has been incorporated in this study as an indicator of drought severity (Figure 3j) (Alshaikh 2015). 1st and 2nd order streams often behave like gullies and accelerate soil and gully erosion (Araujo and Pejon 2015; Joshi et al. 2016). To arrest the erosion of soil and its deposition in the reservoir of the Massanjore dam which was constructed for hydro-electric generation, total of 242 smaller reservoirs have been installed at different parts of the basin. These reservoirs feature concrete barrier (locally called bundh) in front of it mostly constructed at the head of the gully and therefore, called Gully Head Bundh (GHB). These GHBs and check dams are capable of arresting soil erosion promoted by 1st and 2nd order streams (Pal and Debanshi 2018). The Euclidian distance tool of Arc GIS software has been employed to compute the spatial distance of the basin area from 1st and 2nd order streams, GHBs, check dams, and generate the spatial data layers at a 30 m × 30 m pixel scale (Figure 3q, 3l, and 3n). 1st and 2nd
order streams have been digitized; the locations of GHBs have been extracted from the
SOI toposheets, while the locations of check dams have been extracted from the Google
earth imagery. The frequency of existing gullies per unit area of each geological division
(scale—1:200,000) has been calculated, and a raster layer with a 30 m × 30 m cell size has
been generated for each division, considering its gully frequency for facilitating the input
of spatial model development. The same procedure has been repeated for geomorpho-
logical divisions also. For producing the spatial data layer of soil texture (Figure 3a), the
proportion of coarser sand in the top soil has been considered since coarse-textured is
very prone to gully formation and erosion (Pal 2016). The proportion of coarser sand in
the top soil layer of the upper catchment has been measured by extracting soil samples
from 43 sites and testing with the help of a digital sieve shaker. The village level soil tex-
ture maps (1:100,000) of NIC have also been subscribed to get the soil texture data. The
percentage of coarser sand in the soil surface has been identiﬁed based on USDA (1999)
soil classiﬁcation system. The DST-provided rainfall map has been used for preparing the
data layer of average yearly rainfall (Figure 3h). Soil erosivity has also been incorporated
in this study to consider the impact of rainfall intensity on the soil. Since the concen-
trated rainfall in fewer months causes higher erosion than the uniform rainfall for the
entire year (Pal and Debanshi 2018), the erosivity of the rainfall has been considered as a
parameter in this study. To fulﬁll this purpose, the average rainfall data of last 30 years of
two stations (Suri, Dumka) have been taken and district resource maps and the Ph.D. the-
esis of Bhattacharyya (2014) have been consulted. For producing the data layers (Figure
3g), soil erosivity has been calculated using the modiﬁed Fournier index (Eq. (6)) (1960).
The parameter of water coverage has been considered in this study because of the pres-
ence of large reservoir river projects like Masanjore or Tilpara in the area. The data layer
(Figure 3m) has been prepared by categorizing the basin area into water logged and non-
water-logged areas and demarcating the reservoir area from the SOI toposheet. Variance
inflation factor (VIF) statistics have been computed for this. Vörösmarty and Dobos
(2020) reported value of <5 does mean there is no co-linearity problem. The result shows
that the VIF score is within permissible limits, therefore, no signiﬁcant multi-co-linearity
exists among selected factors.

\[ S P I = A s \times \tan \sigma \]  

(1)

where \( As = \) speciﬁc catchment area in meters, \( \sigma = \) slope gradient in degrees

\[ B S I = \frac{(b_{SWIR} + b_R) - (b_{NIR} + b_B)}{(b_{SWIR} + b_R) + (b_{NIR} + b_B)} \]  

(2)

where \( b_{SWIR} = \) Short Wave Infra-Red band brightness value, \( b_R = \) Red band brightness value, \( b_{NIR} = \) Near Infra-Red band brightness value, \( b_B = \) Blue band brightness value

\[ L S T = \frac{T_B}{1 + \left\{ (\lambda \times T_B / \rho) \times \ln \varepsilon \right\}} \]  

(3)

where \( T_B = \) At satellite temperature, \( \lambda = \) wavelength of emitted radiance in meters, \( \rho = h \times c / r, h = \) Planck’s constant, \( r = \) Boltzmann constant, and \( c = \) velocity of light \( \varepsilon = \) Land surface emissivity.

\[ NDVI = \frac{(b_{IR} - b_R)}{(b_{IR} + b_R)} \]  

(4)

where \( b_{IR} = \) Infra-Red band brightness value, \( b_R = \) Red band brightness value

\[ W S V I = \frac{NDVI}{LST} \]  

(5)
\[ FI = \sum_{i=1}^{12} \frac{P_i^2}{P} \]  

where \( P_i \) = Precipitation of the month \( I \), and \( P \) = Mean annual Precipitation.

### 3.4. Modelling gully erosion susceptibility

#### 3.4.1. Selection of training sites for PBM and SPBM

For selecting the training sites for the gully erosion susceptibility cluster modelling, the Gully affected sites are considered ‘1’, and non-affected sites are considered ‘0’. 3658 points from both the gully and non-gully erosion points, 70% (2560 points) of data have been used for modeling, and 30% (1098 points) were kept for validation. The training sites for gully affected and non-affected points have been recognized using SOI toposheet with 1:50,000 scale, high-resolution Google Earth imagery with 2.62 m. spatial resolution, and field investigation. Since the gully-affected sites were mostly concentrated in the upper catchment of the study area, both the training and testing sites have been taken from these regions. After delineating affected regions with the help of SOI toposheets, Google earth imagery and field investigation, training sites have been randomly extracted as point feature from those demarcated gully erosion prone areas. Here distance from gully was not considered. In this context, studies like Süzen and Doyuran (2004) suggested considering equal numbers of vulnerable and non-vulnerable points for better performance of the employed models. Following this principle, equal training sites were extracted from gully affected and non-gully affected areas. The broken ground and erosion prone surface areas have been kept aside while training sites extraction for non-gully affected areas.

#### 3.4.2. Factor cluster modelling technique

After the Selection of the sample dataset, three ML classification algorithms, random forest (RF), AdaBoosting (ADB), and support vector machine (SVM), have been applied to the training dataset. During this training process, 1 has been assigned for gully affected training sites, and 0 has been assigned for non-gully affected training sites. In the sample point based modelling (SPBM), the output of each cluster was mapped with the help of interpolation technique using inverse distance weighting (IDW) in the ArcGIS environment. Whereas, in PBM, the classification algorithm has been run on the entire dataset, including all pixels of the study area. Although the ML techniques have run upon the basic regression techniques, the ML techniques have superiority over traditional statistical methods for delineating the database’s complex behavior, especially the geospatial phenomena (Chowdhuri et al. 2021; Panahi et al. 2021). Therefore, we have employed the three ensemble machine learning techniques: Random Forest (RF), AdaBoosting (ADB), and Support Vector Machine (SVM) to explore the gully erosion susceptible zones for the present study area. Zhou et al. (2020), Pal and Paul (2021), and Pal and Debanshi (2021) have successfully employed these ML techniques to inspect and map various geospatial phenomena in their studies. A detailed methodological analysis of the SPBM and PBM has been discussed below.

In the PBM, the prediction of all the pixels of the basin area was in binary expression i.e. 1 for gully erosion susceptible and 0 for non-susceptible, and eventually, the factor cluster wise gully erosion prediction mapping has been done in binary expression. But in the case of SPBM, the interpolated map, the susceptibility value varied 0 to 1, where 0 indicated non-susceptible and 1 showed susceptible. So the maps of gully erosion
prediction derived from each factor cluster have been converted to a binary map considering 0.5 as the threshold. In this way after producing the factor cluster wise gully erosion prediction maps for both PBM and SPBM approach the maps have been spatially integrated in order to yield gully erosion susceptibility. Since there were four maps containing 0 and 1 values, the values of gully erosion susceptibility maps varied from 0 to 4. In this map 0 represents those areas where all four clusters negated gully erosion and 4 represents those parts where all four clusters predicted gully erosion. Based on the agreement of the cluster on gully erosion prediction, the susceptibility has been categories into five classes, viz. extremely susceptible, highly susceptible, moderately susceptible, least susceptible, and relatively safe. The susceptibility value 4 has been categorized as extremely susceptible followed by susceptibility value 3 which denotes prediction agreement of 3 clusters has been categorized as highly susceptible. The susceptibility values 2 and 1 denote the agreement of two and one clusters and have been categorized respectively as moderately susceptible and least susceptible. The 0 susceptibility denotes the agreement of no cluster regarding the prediction of gully erosion and therefore, has been labeled relatively safe.

### 3.4.3. Random Forest (RF)

In unsupervised ensemble learning, the random forest has proven to be a reliable and robust algorithm for classification and regression (Schonlau and Zou 2020; Meshram et al. 2021). The random forest algorithm uses the bootstrap aggregation technique while growing multiple decision trees (Syam and Kaul 2021). Therefore, each tree can predict and ‘vote’ independently at the time of supply of the input data for the corresponding classes \( \hat{a}, Y^f_{rf} = \text{majorityvote}\{X^c(a)^+\}^Y \), where \( X^c(a)^+ \) the class’s prediction of the \( b \)th random forest (Hartini et al. 2021). The voting capability of RF reduces the noise and improves the robustness and prediction capability of the input data (Zhou et al. 2020). RF reduces the dimensionality in the database by using a built-in feature selection system that can regulate multiple parameters without removing some of them (Tella et al. 2021). By estimating the increase in prediction error in the dataset, RF can compute the variable importance scores of each constituent tree and also the entire database (Pal and Paul 2020; Elavarasan and Vincent 2021). Studies like Chan and Paelinckx (2008) and Pal and Mather (2003) reported that RF (uses bagging) is more sensitive to noise than the algorithms based on boosting technique. Finally, RF uses proximity between pairs of variables to locate the outliers (Norouzi and Moghaddam 2020). To calculate the proximity between two samples, RF calculates the frequency of appearance of the sample at the same terminal node (Wangchuk and Bolch 2020). After the computation of proximity for each pair of causes and the formation of trees, the variables are normalized by dividing with numbers of trees (Nahkala et al. 2021). For the present study, a random forest algorithm from the Scikit-learn ML library from python 3.6 has been applied for PBM, and the Random Forest tool from WEKA has been used for SPBM.

### 3.4.4. AdaBoosting (ADB) classification

AdaBoosting, or adaptive boosting (ADB), is a tree-based ensemble algorithm that uses an iterative process to generate weight at each sample for both classification and regression (Mehmood and Asghar 2021; Pal and Paul 2021). ADB consecutively grows multiple interdependent classifiers and generates multiple classifiers to ‘vote’ them (Xu et al. 2021). ADB generates equal weights for each dataset. In the iteration process, the weights of each unclassified dataset have been increased, whereas the weight for correctly classified samples is gradually decreased (Asif et al. 2021). These weights indicate the classifier’s
accuracy and are a function of the overall weights of the correctly classified samples (Nishat et al. 2021). The algorithm’s accuracy depends on the given weight of correctly classified samples. Unlike RF, ADB is sensitive to noise in the training data; therefore, it is necessary to preprocess the training data to obtain optimum output (He et al. 2020). Python-based AdaBoost Classifier from ensemble learning tool from Scikit-learn ML library is used to classify factor clusters in this present study, whereas AdaBoostM1 tool in WEKA is used for SPBM classification.

3.4.5. Support vector machine (SVM)

SVM is a popular robust supervised classification technique initially developed for binary classification problems (Cortes and Vapnik 1995). In this classification, several classes can be determined as an arrangement of \( n\)-class or \( n(n-1)/2\) or binary classification techniques by applying different selection schemes that lead to a final decision (Ul Din and Mak 2021). SVM generates an optimum hyperplane in a given training set that elucidates a complete separation of linearly different clusters in a hyper-surface (Talukdar et al. 2021). During the training, SVM searches for a hyper-plane that best separates the dataset into two classes (binary), such as 1 and \(-1\). These instances can be identified as \( h_1 \) and \( h_{-1} \) or support vectors (Deiss et al. 2020). The hyperplane between \( h_1 \) and \( h_{-1} \) is a 0 or non-linear convex programming problem. The pairs of parameters of the separating hyperplane determination can be determined by solving the following optimization problem:

\[
\min_{w, b, \xi} \frac{1}{2} \|w^T w\| + C \sum^{I}_{i=1} \xi_i
\]

In respect to \( 1 - \xi_i - y_i (w \cdot x + b) \leq 0, -\xi_i \leq 0, i = 1, \ldots, n \)

where penalty parameter (\( C \)) and margin of tolerance (\( \xi \)) needed to be tuned to accuracy and efficiency of the model, for this study, the SMOreg tool from WEKA has been used for SPBM, and SVC algorithm from Python has been used for PBM classification in Python. GridsearchCV tool from the Scikit-learn ML library has been used for hyperparameter selection.

3.4.6. Hyperparameter optimization

The selection of a proper set of hyperparameters is necessary to extract maximum performance and accuracy from the model and is considered an integral part of machine learning-based modelling (Kotthoff et al. 2019). Studies like Khalid et al. (2020) and Victoria et al. (2021) reported that high accuracy and performance could be achieved by tuning hyperparameters. There are various techniques to tune hyperparameters like randomized search, grid search, Bayes optimization, halving grid and random search, hyper Opt-Sklearn, kNN, or IBK. For this, in the present study, kNN or IBK techniques have been used to tune the hyperparameter in WEKA for SPBM, whereas grid search optimization and the K-fold cross-validation technique have been used to adjust the hyperparameter and validate the predicted dataset. Each cluster model has run a 10-fold iterative process for every 340 candidates totaling 2400 fits. In Python, the GridSearchCV tool from the Sk-learn ML library is employed for this present work.

The final output layers of pixel based modelling (PBM) and sample point based modelling (SPBM) have been classified into five equidistant susceptible zones: relative safe, least susceptible, moderately susceptible, highly susceptible, and extremely susceptible classes. The areas emerged as low to very low prone to gully erosion, categorized as the least
susceptible and relative safe zones. In contrast, areas with increasing gully erosion effects are moderately, highly, and extremely susceptible zones.

3.5. Accuracy assessment of the PBM and SPBM

To affirm the performance of the employed models, six statistical matrices, namely percentage of correctly classified data, receiver operating characteristics (ROC) curve derived area under curve (AUC), Precision, Sensitivity, F1-score, and Matthew’s correlation coefficient (MCC), have been calculated based on 1098 sample points. The validation techniques have been calculated based on four matrices which are true positive (TP), false negative (FN), true negative (TN), and false-positive (FP). Based on these matrices, the specificity and sensitivity of the models have been calculated. The ROC shows the distinguishing capability of the model between classes and is constructed by plotting sensitivity (Y-axis) against 1-specificity (X-axis), where AUC denotes the degree of separability (Narkhede 2018). Sensitivity or recall has been calculated using TP and FN, and specificity has been calculated using TN and FP. In MCC, all four matrices have been used for better reliability and performance (Chicco et al. 2021). The MCC value ranges from −1 to 1,
where $-1$ indicates the least accuracy in the prediction and vice-versa. The F1-score has been calculated against the value of precision and sensitivity. Therefore, it gives a better result, even better in uneven class distribution (Chicco and Jurman 2020). The value of ROC-AUC, Precision, and F1-score, ranges between 0–1, where absolute 0 or near to 0 indicates less reliability on the model and 1 indicates maximum reliability on the employed models (Yao and Shepperd 2020). All four matrices have some superiority and flaws against another; therefore, a score of any single model cannot be taken as a good indicator for model validation. The calculation for the stated matrices is as follows (Eqs. (9)–(12)).

$$\text{recall} = \text{sensitivity} = \frac{Tp}{Tp + Fn} \quad (9)$$

$$\text{precision} = \frac{Tp}{(Tp + Fp)} \quad (10)$$

$$\text{MCC} = \frac{Tp \times Tn - Fp \times Fn}{\sqrt{(Tp + Fp)(Tp + Fn)(Tn + Fp)(Tn + Fn)}} \quad (11)$$

$$F1-score = \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (12)$$

4. Results

4.1. Gully erosion susceptible models PBM and SPBM

Figure 4 displays the comparative views of pixels and sample points based on ensemble ML models for gully erosion susceptibility in the upper Mayurakshi river basin. Each map has been classified into five gully erosion susceptible zones, and the area under each class is shown in Table 1. About 15–20% of the area has been detected as a very highly susceptible gully erosion zone as per pixel based ML models. On the other hand, this ranges from 1–26% in the case of sample point based ML models. From these statistics, it can be stated that sample point based modelling reflected a greater degree of prediction uncertainty. The gap between pixel and point based very highly susceptible area have been found very high in the ADB model and found low in the case of the RF model. Such inconsistency in prediction is caused for the question of the reliability of sample point based models. Geographically, the upper parts of the study region are high to very highly susceptible to gully erosion. Highly erodible coarse-textured laterite soil, relatively sloppy Chottanagpur plateau fringe, and highly dense 1st and 2nd orders streams are mainly responsible for greater gully erosion susceptibility (Pal 2016; Pal and Debanshi 2018;
Debanshi and Pal (2020). Areal patches of very highly erosion susceptibility look relatively continuous in the case of point-based modelling, and this geolocation is deliberately different from pixel based ML model output. So, justification of the suitability of the model’s application for such work is essential. Moreover, it is also necessary to justify the suitability of the pixel and point based ML models. For that accuracy, a test is highly crucial.

Table 2 shows the accuracy and performance level of the applied models with the help of some statistical techniques like AUC-ROC, precision, sensitivity, F1 score, and MCC. In the case of sample point based ML models, AUC and MCC range from 0.53–0.74 and 0.48–0.69, whereas it ranges from 0.81–0.91 and 0.78–0.91, respectively, in the case of pixel based ML models. Sensitivity, precision, and F1 score also represent the same trend (Table 2). Based on these statistics, it could be inferred that pixel based gully erosion susceptibility models are superior to sample point based models. Obviously, in the case of sample point based models, the sample size may influence the accuracy level; however, at any sample size, it will not be as precise as the pixel based models. Among the tree applied ensemble ML models, RF was the best representative since the accuracy level, as per all the test statistics, was high in pixel and point-based models. AUC-ROC value shows a very good to excellent agreement between training and test data set in case of all the models applied for predicting pixel based susceptibility. However, RF has been recognized as the best suited. The agreement between the test and training dataset was unacceptable to fairly good in point based ML models. To assess the stability of the validation result, we have performed the validation technique with 11 different sample sites. However, the result did not vary significantly, and the best performance result has been mentioned here (Table 2).

Field data regarding extension of the gully, widening, and deepening of the gully has been collected from thirty-seven sites to validate the ML based gully erosion, susceptible models. Gully erosion rate have been correlated with the degree of gully erosion susceptibility. Correlation coefficients between the degree of susceptibility and head ward extension of the gully, gully widening and deepening were 0.65–0.77, 0.58–0.62, and 0.73–0.81, respectively, in case of pixel based models, and all these values are statistically significant at 0.05 to 0.1 level of significance. The highest correlation has been found in case of the RF model. In the case of sample point based ML models, correlation coefficient values range from 0.12–0.35, 0.13–0.29, and 0.17–0.34, and the significance level is <0.05.

### 4.2. Factor cluster based models

Figure 5 displays the comparative views of pixel and point based gully erosion susceptibility modeling of four-factor clusters like (1) erodibility, (2) erosivity, (3) resistance, and (4)
The nature of susceptible areas has been determined based on the considered factors under each factor cluster. Point based models show continuous patches of susceptible areas, whereas its areal extension is highly fragmented. All the factor clusters based model has recognized comprehensive parts of the upper catchment area under gully erosion susceptibility. This proportion is susceptible high in the case of topographical factor cluster. Both pixel and point based factor cluster models have recognized upper parts of the study region is under soil erosion susceptibility; however, in terms of geolocation, these are not always coinciding. So, it is necessary to justify whether the pixel based model is superior to the sample point based model. Here one thing is clear if the factor cluster based model is validated with the testing data set, it may show relatively poor agreement in most cases. However, considering the relative difference in statistical values of the applied measures, it could be easy to infer the suitability of the models. Table 3 depicts each factor cluster model’s validation and model performance statistics.

Table 3. Factor cluster based models (1) Pixel based and (2) Sample point based approaches.

Figure 5. Factor cluster based models (1) Pixel based and (2) Sample point based approaches.
agreement between training and testing datasets. Here also RF model has appeared the best suited than other models.

5. Discussion

The study exhibited that about 20% area at the upper part of the study area is highly susceptible to gully erosion. The predictive ability of the RF model is the most credible since its agreement between training and the testing dataset is very high. The predictive ability of pixel based models is far better than sample point based models. Why is a higher degree of gully erosion susceptibility found in the upper part of the study region? Why predictive ability of pixel based models is higher than sample size based models. May sample size be an issue in improving the accuracy of the models?

The upper reach of every river basin is characterized by higher 1st and 2nd order stream frequency and density and is prone to gully erosion with different intensities (Das 2014). New gully formation and extension of the gully are very general mechanisms in such an area. But in a river basin where highly erodible coarse texture soil is available, such a rate is very high. In the Mayurakshi river basin, the lateritic composition of soil enriched with silica, aluminium, iron etc., is highly prone to chemical weathering and consequent erosion (Ghosh et al. 2015). A field study also revealed that gully formation and extension are very prominent even in the forested lateritic region. In the region where hard lateritic duricrust is there, the rate of gully erosion is relatively less. In the present case study, a few lateritic patches are where hard cap lateritic mass is armoured by the fragile laterite soil. Poor undergrowth vegetation in the Sal (Shorea robusta) forest available in a wider part of the upper catchment also insists intensive gully erosion. The growing rate of stone collection from the hilly region, settlement growth in the forested

| ML techniques | Clusters | Modeling software | Correctly classified (%) | ROC (AUC) | Precision | Sensitivity | F1-score | MCC | Support |
|---------------|----------|-------------------|--------------------------|-----------|-----------|------------|----------|-----|---------|
| RF            | Erodibility | PBM (Python)      | 79.220                   | 0.788     | 0.783     | 0.782      | 0.782    | 0.765 | 1098    |
|               |           | SPBM (WEKA)       | 48.910                   | 0.425     | 0.418     | 0.412      | 0.415    | 0.415 | 1098    |
|               | Erosivity | PBM (Python)      | 79.100                   | 0.789     | 0.781     | 0.780      | 0.780    | 0.698 | 1098    |
|               |           | SPBM (WEKA)       | 44.210                   | 0.438     | 0.426     | 0.411      | 0.418    | 0.405 | 1098    |
|               | Resistance| PBM (Python)      | 78.960                   | 0.784     | 0.771     | 0.771      | 0.771    | 0.765 | 1098    |
|               |           | SPBM (WEKA)       | 43.420                   | 0.429     | 0.424     | 0.417      | 0.420    | 0.408 | 1098    |
|               | Topographical| PBM (Python)    | 76.890                   | 0.756     | 0.751     | 0.749      | 0.749    | 0.734 | 1098    |
|               |           | SPBM (WEKA)       | 41.320                   | 0.409     | 0.398     | 0.391      | 0.394    | 0.381 | 1098    |

| ADB           | Erodibility | PBM (Python)      | 77.100                   | 0.775     | 0.778     | 0.765      | 0.771    | 0.758 | 1098    |
|               |           | SPBM (WEKA)       | 35.760                   | 0.344     | 0.335     | 0.331      | 0.333    | 0.328 | 1098    |
|               | Erosivity | PBM (Python)      | 76.330                   | 0.758     | 0.754     | 0.751      | 0.752    | 0.740 | 1098    |
|               |           | SPBM (WEKA)       | 32.570                   | 0.318     | 0.311     | 0.310      | 0.310    | 0.307 | 1098    |
|               | Resistance| PBM (Python)      | 75.400                   | 0.745     | 0.736     | 0.731      | 0.733    | 0.718 | 1098    |
|               |           | SPBM (WEKA)       | 31.630                   | 0.306     | 0.299     | 0.297      | 0.298    | 0.287 | 1098    |
|               | Topographical| PBM (Python)    | 72.800                   | 0.709     | 0.695     | 0.690      | 0.692    | 0.663 | 1098    |
|               |           | SPBM (WEKA)       | 30.100                   | 0.296     | 0.291     | 0.287      | 0.289    | 0.275 | 1098    |

| SVM           | Erodibility | PBM (Python)      | 75.180                   | 0.745     | 0.731     | 0.727      | 0.729    | 0.721 | 1098    |
|               |           | SPBM (WEKA)       | 44.570                   | 0.431     | 0.425     | 0.432      | 0.428    | 0.426 | 1098    |
|               | Erosivity | PBM (Python)      | 74.890                   | 0.748     | 0.736     | 0.731      | 0.733    | 0.729 | 1098    |
|               |           | SPBM (WEKA)       | 42.150                   | 0.410     | 0.398     | 0.391      | 0.394    | 0.387 | 1098    |
|               | Resistance| PBM (Python)      | 73.110                   | 0.721     | 0.723     | 0.718      | 0.720    | 0.712 | 1098    |
|               |           | SPBM (WEKA)       | 41.510                   | 0.397     | 0.388     | 0.381      | 0.384    | 0.369 | 1098    |
|               | Topographical| PBM (Python)    | 71.980                   | 0.719     | 0.711     | 0.700      | 0.705    | 0.684 | 1098    |
|               |           | SPBM (WEKA)       | 39.470                   | 0.374     | 0.368     | 0.361      | 0.364    | 0.350 | 1098    |
region, and illegal deforestation are all responsible for creating an ambient condition of highly intensive soil erosion. Some forested areas and their peripheral part are also prone to high gully erosion susceptibility. Higher drainage frequency and density in its upper part coupling with a relatively steeper slope and fragile laterite soil firmly control the gully erosion. Pal and Debanshi (2018) also condemned the fragility of laterite soil as a dominant factor for the high rate of gully erosion in this area.

Pixel based modeling in Python software provided more accurate results than sample point based modeling in this study. There is no need for interpolation while every pixel has been considered for analysis. So, here total pixel is the total sample points. But in the case of sample point based modeling, unknown sites have been predicted using the interpolation method based on known values of the training sites. Interpolation often generalizes the trend of value gradient. It may not always represent reality properly (Abdulmanov et al. 2021). During interpolation, several issues are associated with making the map more precise. Sample point size for interpolation is the first and foremost issue that can dictate the precision of the interpolated map. The sample size is indirectly related to the degree of accuracy of the interpolated map (Simpson and Wu 2014). In this study, 3658 sample sites have been taken for analysis.

Ensemble machine learning models have been applied for predicting gully erosion susceptibility considering its robustness and high precision level. In the present study, all models have been found acceptable, but precision is high in the case of the RF model. In the RF algorithm, a built-in features selection system is used to reduce dimensionality without removing any data in the database. Therefore, there is very little chance of data loss or outlier intervention in the dataset, making this algorithm more reliable and error-free than the other algorithm (Zhou et al. 2020). The RF (which uses bagging) is more sensitive to noise than the algorithms based on boosting technique; therefore, it generates more accurate predictions by controlling the noise in the dataset (Pal and Mather 2003; Chan and Paelinckx 2008). The multi-model approach has been adopted here to enhance the reliability of the comparison. Multiple models with the same aim cross-validate themselves and, therefore, are more reliable in the modeling approach (Arabameri et al. 2020). Moreover, such an approach helps justify the best representative model of all. Considering these advantages, in recent times so, many scholars like Kadavi et al. (2018), Mosavi et al. (2018), Hong et al. (2018), Liu et al. (2017), and Jamali (2019) have taken this line of thinking.

The identification of extreme gully erosion susceptible zones has its implication from the management point of view. These areas deserve special care for resisting gully erosion and soil loss. There may be some ways to check gully erosion in this region theoretically; however, check damming is an effective measure. For the last seven decades, this engineering mechanism has been highly relied on. Small scale check damming can break down the slope and store the regional soil. A good number of check dams have been installed in the catchment of the Massanjore dam over the Mayurakshi river to check soil erosion and thereby lengthen the longevity of the dam (Pal 2016; Debanshi and Pal 2020). So, this could be a good way. Renovation of existing dams and installing a new dam in the deserving sites may check the rate of soil erosion. Further research is necessary to determine the deserving sites for new check dam sites. As far as the soil type is concerned, it is pretty challenging to arrest such erosion but check damming may slow down such rate. Since the natural vegetation of this region is tall with lesser canopy density, that cannot contribute much to protecting the soil from splash erosion. So vegetation area should be increased, emphasizing dwarf varieties of plants having greater canopy density, which will reduce fall velocity and provide good resistance to gully erosion.
The present study has compared the pixel and point based ensemble ML models for gully erosion susceptibility. This approach is very new, and the finding of this may encourage scholars to do with software environments like Python, where every pixel is a sample. It also proved that sample point-based modeling, which is done frequently, has considerable noise, and such output may mislead spatial planning. It is further to be mentioned that here only the output of particular sample size has been generated; if instead of this, the same modeling process was run using different sample sizes and accuracy was assessed, the result was more easily convincing. So the analysis of sample size effect on model output is a good scope of work.

6. Conclusion

The present study inferred that pixel based ensemble models’ predictability is higher than sample point based models. The RF model shows the highest precision for predicting gully erosion susceptibility in pixel and point based modeling. Coarser laterite soil dominated fertile soil in the upper part of the study region is more susceptible than its counterparts. So, the study recommends using pixel based ensemble modeling to predict susceptibility instead of sample point-based modeling. The conventional practice of modeling with sample point data using interpolation technique generalizes the spatial character and thereby reduces the precision of the model. An increase in the undergrowth and check damming may be some fruitful measures to slow down the gully erosion rate. In such a case, implementing management strategies with limited resources require greater accuracy and a higher degree of spatial differentiation for prioritizing the area of action. Since the study demonstrated the gully erosion susceptible region, this information will be useful to the decision-makers for protecting gully erosion and soil resources.

Authors’ contributions

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Swades Pal, Sandipta Debanshi and Satyajit Paul. The first draft of the manuscript was written by Swades Pal and Satyajit Paul and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Disclosure statement

The authors have no relevant financial or non-financial interests to disclose.

References

Abdulmanov R, Miftakhov I, Ishbulatov M, Galeev E, Shafeeva E. 2021. Comparison of the effectiveness of GIS-based interpolation methods for estimating the spatial distribution of agrochemical soil properties. Environ Technol Innov. 24:101970.
Ahmadpour H, Bazarfshan O, Rafaei-Sardooi E, Zamani H, Panagopoulos T. 2021. Gully erosion susceptibility assessment in the Kondoran watershed using machine learning algorithms and the Boruta feature selection. Sustainability. 13(18):10110.
Alshaikh AY. 2015. Space applications for drought assessment in Wadi-Dama (West Tabouk), KSA. Egypt J Remote Sens Space Sci. 18(1):S43–S53.
Arabameri A, Chandra Pal S, Costache R, Saha A, Rezaie F, Seyed Danesh A, Pradhan B, Lee S, Hoang N-D. 2021. Prediction of gully erosion susceptibility mapping using novel ensemble machine learning algorithms. Geomat Nat Hazards Risk. 12(1):469–498.
Arabameri A, Chen W, Loche M, Zhao X, Li Y, Lombardo L, Cerda A, Pradhan B, Bui DT. 2020. Comparison of machine learning models for gully erosion susceptibility mapping. Geosci Front. 11(5):1609–1620.

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

Araujo TP, Pejon OJ. 2015. Topographic threshold to trigger gully erosion in a Tropical region—Brazil. In: Engineering geology for society and territory. Vol. 3. Cham: Springer. p. 627–630.

Asif M, Nishat MM, Faisal F, Dip RR, Udoy MH, Shikder M, Ahsan R. 2021. Performance evaluation and comparative analysis of different machine learning algorithms in predicting cardiovascular disease. Eng Lett. 29(2).

Auchincloss AH, Gebreab SY, Mair C, Diez Roux AV. 2012. A review of spatial methods in epidemiology, 2000–2010. Annu Rev Public Health. 33:107–122.

Azareh A, Rahmati O, Rafie-Sardooi E, Sankey JB, Lee S, Shahabi H, Ahmad BB. 2019. Modelling gully-erosion susceptibility in a semi-arid region, Iran: investigation of applicability of certainty factor and maximum entropy models. Sci Total Environ. 655:684–696.

Bandfield JL, Ghent RR, Vasavada AR, Lawrence SJ, Robinson MS. 2011. Lunar surface rock abundance and regolith fines temperatures derived from LRO Diviner Radiometer data. J Geophys Res. 116(E12):1–18.

Bandyopadhyay S, Kar NS, Das S, Sen J. 2014. River systems and water resources of West Bengal: a review. Vol. 3. Bengaluru: Geological Society of India Special Publication; p. 63–84.

Bhattacharyya A. 2014. Evolution of hydro genomic characteristics of flood in the Mayurakshi river basin of eastern India [Ph.D. thesis]. Department of Geography, Viswa-Bharati University.

Borrelli P, Poesen J, Vanmaercke M, Ballabio C, Hernandes J, Maerker M, Scarpa S, Panagos P. 2022. Monitoring gully erosion in the European Union: a novel approach based on the Land Use/Cover Area frame Survey (LUCAS). Int Soil Water Conser Res. 10(1):17–28.

Brownlee J. 2016. Machine learning mastery with Python: understand your data, create accurate models, and work projects end-to-end. San Francisco: Machine Learning Mastery Pty Ltd; p. 100–120.

Brownlee J. 2019. Machine learning mastery with Weka. Ebook ed., Vol. 1. p. 4.

Can R, Kocaman S, Gokceoglu C. 2021. A comprehensive assessment of XGBoost algorithm for landslide susceptibility mapping in the upper basin of Ataturk Dam, Turkey. Applied Sciences. 11(11):4993.

Canovas JB, Stoffel M, Martín-Duque JF, Corona C, Lucía A, Bodoque JM, Montgomery DR. 2017. Gully evolution and geomorphic adjustments of badlands to reforestation. Sci Rep. 7(1):1–8.

Chabuk A, Al-Madhlom Q, Al-Maliki A, Al-Ansari N, Hussain HM, Laue J. 2020. Water quality assessment along Tigris River (Iraq) using water quality index (WQI) and GIS software. Arab J Geosci. 13(14):1–23.

Chan JCW, Paelinckx D. 2008. Evaluation of Random Forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. Remote Sens Environ. 112(6):2999–3011.

Chicco D, Jurman G. 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 21(1):1–13.

Chicco D, Tötsch N, Jurman G. 2021. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. BioData Min. 14(1):1–22.

Chowdhiari I, Pal SC, Chakrabortty R, Malik S, Das B, Roy P. 2021. Torrential rainfall-induced landslide susceptibility assessment using machine learning and statistical methods of eastern Himalaya. Nat Hazards. 107(1):697–722.

Churpek MM, Yuen TC, Winslow C, Meltzer DO, Kattan MW, Edelson DP. 2016. Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. Crit Care Med. 44(2):368.

Conforti M, Aucelli PP, Robustelli G, Scarciglia F. 2011. Geomorphology and GIS analysis for mapping gully erosion susceptibility in the Turbolo stream catchment (Northern Calabria, Italy). Nat Hazards. 56(3):881–898.

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

Das D. 2014. Identification of erosion prone areas by morphometric analysis using GIS. J Inst Eng India Ser A. 95(1):61–74.

Debanshi S, Pal S. 2020. Assessing gully erosion susceptibility in Mayurakshi river basin of eastern India. Environ Dev Sustain. 22(2):883–914.

Deiss L, Margenot AJ, Culman SW, Demyan MS. 2020. Tuning support vector machines regression models improves prediction accuracy of soil properties in MIR spectroscopy. Geoderma. 365:114227.
Dutta S. 2016. Soil erosion, sediment yield and sedimentation of reservoir: a review. Model Earth Syst Environ. 2(3):1–18.

Elavarasan D, Vincent PDR. 2021. A reinforced random forest model for enhanced crop yield prediction by integrating agrarian parameters. J Ambient Intell Hum Comput. 12:1–14.

Elfadaly A, Wafa O, Abouarab MA, Guida A, Spanu PG, Lasaponara R. 2017. Geo-environmental estimation of land use changes and its effects on Egyptian Temples at Luxor City. IJGI. 6(11):378.

Feizzadeh B, Blaschke T, Nazmfar H, Rezaei Moghaddam MH. 2013. Landslide susceptibility mapping for the Urmia Lake basin, Iran: a multi-criteria evaluation approach using GIS. Int J Environ Res. 7(2):319–336.

Frey U. 2020. Putting machine learning to use in natural resource management—improving model performance. E&S. 25(4):1–33.

Gelagay HS, Minale AS. 2016. Soil loss estimation using GIS and Remote sensing techniques: a case of Koga watershed, Northwestern Ethiopia. Int Soil Water Conserv Res. 4(2):126–136.

Ghosh S, Guchhait SK, Hu XF. 2015. Characterization and evolution of primary and secondary laterites in northwestern Bengal Basin, West Bengal, India. J Palaeogeogr. 4(2):203–230.

González S, García S, Del Ser J, Rokach L, Herrera F. 2020. A practical tutorial on bagging and boosting based ensembles for machine learning: algorithms, software tools, performance study, practical perspectives and opportunities. Inf Fusion. 64:205–237.

Han X, Lv P, Zhao S, Sun Y, Yan S, Wang M, Han X, Wang X. 2018. The Effect of the Gully Land consolidation project on soil erosion and crop production on a typical watershed in the loess plateau. Land. 7(4):113.

Hartini S, Rustam Z, Saragih GS, Vargas MJS. 2021. Estimating probability of banking crises using random forest. IAES Int J Artif Intell. 10(2):407.

He H, Khoshelham K, Fraser C. 2020. A multiclass TrAdaBoost transfer learning algorithm for the classification of mobile lidar data. ISPRS J Photogramm Remote Sens. 166:118–127.

Hoeser T, Kuenzer C. 2020. Object detection and image segmentation with deep learning on earth observation data: a review - part I: evolution and recent trends. Remote Sens. 12(10):1667.

Hoeser T, Bachofer F, Kuenzer C. 2020. Object detection and image segmentation with deep learning on Earth observation data: a review—part II: applications. Remote Sens. 12(18):3053.

Hogland J, Anderson NM. 2014. Improved analyses using function datasets and statistical modeling. In: Proceedings of the 2014 ESRI Users Conference, July 14–18, 2014, San Diego, CA. Redlands, CA: Environmental Systems Research Institute. p. 166–182. Online: http://proceedings.esri.com/library/user-conf/proc14/papers/166_182.pdf.

Hogland JS, Anderson NM. 2015. Estimating FIA plot characteristics using NAIP imagery, function modeling, and the RMRS Raster Utility coding library. In: Stanton SM, Christensen GA, comp. Pushing boundaries: new directions in inventory techniques and applications: Forest Inventory and Analysis (FIA) symposium 2015. 2015 December 8–10, Portland, Oregon. Gen. Tech. Rep. PNW-GTR-931, Vol. 931. Portland, OR: US Department of Agriculture, Forest Service, Pacific Northwest Research Station, p. 340–344.

Hogland J, Anderson N. 2017. Function modeling improves the efficiency of spatial modeling using big data from remote sensing. BDCC. 1(1):3.

Hong H, Liu J, Bui DT, Pradhan B, Acharya TD, Pham BT, Zhu A-X, Chen W, Ahmad BB. 2018. Landslide susceptibility mapping using J48 Decision Tree with AdaBoost, Bagging and Rotation Forest ensembles in the Guangchang area (China). Catena. 163:399–413.

Hosseinalizadeh M, Kariminejad N, Chen W, Pourghasemi HR, Alinejad M, Behbahani AM, Tiefenbacher JP. 2019a. Spatial modelling of gully headcuts using UAV data and four best-first decision classifier ensembles (BFTree, Bag-BFTree, RS-BFTree, and RF-BFTree). Geomorphology. 329:184–193.

Hosseinalizadeh M, Kariminejad N, Chen W, Pourghasemi HR, Alinejad M, Behbahani AM, Tiefenbacher JP. 2019b. Gully headcut susceptibility modeling using functional trees, naïve Bayes tree, and random forest models. Geoderma. 342:1–11.

Jahantigh M, Pessarakli M. 2011. Causes and effects of gully erosion on agricultural lands and the environment. Commun Soil Sci Plant Anal. 42(18):2250–2255.

Jamali A. 2019. Evaluation and comparison of eight machine learning models in land use/land cover mapping using Landsat 8 OLI: a case study of the northern region of Iran. SN Appl Sci. 1(11):1448.

Jha V. C, Kapat S. 2003. Gully erosion and its implications on land use, a case study. In: Land degradation and desertification. Jaipur and New Delhi: Publ; p. 156–178.

Jha VC, Kapat S. 2009. Rill and gully erosion risk of lateritic terrain in South-Western Birbhum District, West Bengal, India. Soc Nat. 21(2):141–158.
Joshi V, Susware N, Sinha D. 2016. Estimating soil loss from a watershed in Western Deccan, India, using revised universal soil loss equation. Landsc Environ. 10(1):13–25.

Jung HS, Lee S, editors. 2019. Machine learning techniques applied to geoscience information system and remote sensing. Basel: MDPI.

Kadavi P, Lee CW, Lee S. 2018. Application of ensemble-based machine learning models to landslide susceptibility mapping. Remote Sens. 10(8):1252.

Kawaguchi K, Kaelbling LP, Bengio Y. 2017. Generalization in deep learning. arXiv Preprint arXiv: 1710.05468; p. 1–28.

Kertész A, Gergely J. 2011. Gully erosion in Hungary, review and case study. Procedia-Soc Behav Sci. 19: 693–701.

Khalid R, Javaid N. 2020. A survey on hyperparameters optimization algorithms of forecasting models in smart grid. Sustain Cities Soc. 61: 102275.

Kotthoff L, Thornton C, Hoos HH, Hutter F, Leyton-Brown K. 2019. Auto-WEKA: automatic model selection and hyperparameter optimization in WEKA. In: Automated machine learning. Cham: Springer. p. 81–95.

Liu Y, Yang Y, Jing W, Yue X. 2017. Comparison of different machine learning approaches for monthly satellite-based soil moisture downscaling over Northeast China. Remote Sens. 10(2):31.

Mansfield LA, Nowack PJ, Kasoar M, Everitt RG, Collins WJ, Voulgarakis A. 2020. Predicting global patterns of long-term climate change from short-term simulations using machine learning. NPJ Clim Atmos Sci. 3(1):1–9.

Marden M, Arnold G, Seymour A, Hambling R. 2012. History and distribution of steepland gullies in response to land use change, East Coast Region, North Island, New Zealand. Geomorphology. 153: 81–90.

Maxwell AE, Bester M, Guillen L, Ramezan C, Carpinello D, Fan Y, Hartley F, Maynard S, Pyron J. 2020a. Semantic segmentation deep learning for extracting surface mine extents from historic topographic maps. Remote Sens. 12(24):4145.

Maxwell AE, Pourmohammadi P, Poyner JD. 2020b. Mapping the topographic features of mining-related valley fills using mask R-CNN deep learning and digital elevation data. Remote Sens. 12(3):547.

Maxwell AE, Warner TA, Strager MP. 2016. Predicting palustrine wetland probability using random forest machine learning and digital elevation data-derived terrain variables. Photogramm Eng Remote Sens. 82(6):437–447.

Mehmood Z, Asghar S. 2021. Customizing SVM as a base learner with AdaBoost ensemble to learn from multi-class problems: a hybrid approach AdaBoost-MSVM. Knowl-Based Syst. 217:106845.

Meliho M, Khattabi A, Mhammdi N. 2018. A GIS-based approach for gully erosion susceptibility modelling using bivariate statistics methods in the Ourika watershed, Morocco. Environ Earth Sci. 77(18): 1–14.

Meshram SG, Safari MJS, Khosravi K, Meshram C. 2021. Iterative classifier optimizer-based pace regression and random forest hybrid models for suspended sediment load prediction. Environ Sci Pollut Res Int. 28(9):11637–11649.

Mitrpanont J, Sawangphol W, Vithantirawat T, Paengkaew S, Suwannasing P, Daramas A, Chen YC. 2017. A study on using Python vs Weka on dialysis data analysis. In: 2017 2nd International Conference on Information Technology (INCIT), November. IEEE. p. 1–6.

Mosavi A, Ozturk P, Chau KW. 2018. Flood prediction using machine learning models: literature review. Water. 10(11):1536.

Mythili G, Goedecke J. 2016. Economics of land degradation in India. In: Economics of land degradation and improvement–a global assessment for sustainable development. Cham: Springer. p. 431–469.

Nahkala BA, Kaleita AL, Soupir ML. 2021. Empirical tool development for prairie pothole management using AnnAGNPS and random forest. Environ Model Softw. 147:105241.

Narkhede S. 2018. Understanding AUC-ROC curve. Towards Data Science, Vol. 26, no. 1. p. 220–227.

Naudts J, Mestre G, Hoffman G. 2004. Improvement of Weka, a datamining tool. Universiteit Gent.

Neyshabur B, Bhojanapalli S, McAllester D, Srebro N. 2017. Exploring generalization in deep learning. 203:1–16. arXiv Preprint arXiv:1706.08947.

Nicu IC. 2021. Is digital shoreline analysis system “fit” for gully erosion assessment? Catena. 203:1–16. https://doi.org/10.1016/j.catena.2021.105307

Nishat MM, Hasan T, Nasrullah SM, Faisal F, Asif MAAR, Hoque MA. 2021. Detection of Parkinson’s disease by employing boosting algorithms. In: 2021 Joint 10th International Conference on Informatics, Electronics & Vision (ICIEV) and 2021 5th International Conference on Imaging, Vision & Pattern Recognition (iCIVPR), August. IEEE. p. 1–7.
Nkonya E, Mirzabaev A, Von Braun J. 2016. Economics of land degradation and improvement—A global assessment for sustainable development. The Hague: Springer Nature; p. 686.

Norouzi H, Moghaddam AA. 2020. Groundwater quality assessment using random forest method based on groundwater quality indices (case study: Miandoab plain aquifer, NW of Iran). Arab J Geosci. 13(18):1–13.

Oh HJ, Syifa M, Lee CW, Lee S. 2019. Land subsidence susceptibility mapping using bayesian, functional, and meta-ensemble machine learning models. Appl Sci. 9(6):1248.

Pal M, Mather PM. 2003. An assessment of the effectiveness of decision tree methods for land cover classification. Remote Sens Environ. 86(4):554–565.

Pal S. 2016. Identification of soil erosion vulnerable areas in Chandrabhaga river basin: a multi-criteria decision approach. Model Earth Syst Environ. 2(1):1–11.

Pal S, Debanshi S. 2018. Influences of soil erosion susceptibility toward overloading vulnerability of the gully head bundhs in Mayurakshi River basin of eastern Chottanagpur Plateau. Environ Dev Sustain. 20(4):1739–1775.

Pal S, Debanshi S. 2021. Machine learning models for wetland habitat vulnerability in mature Ganges delta. Environ Sci Pollut Res. 28(15):19121–19146.

Pal S, Paul S. 2020. Assessing wetland habitat vulnerability in moribund Ganges delta using bivariate models and machine learning algorithms. Ecol Indic. 119:106866.

Pal S, Paul S. 2021. Linking hydrological security and landscape insecurity in the moribund deltaic wetland of India using tree-based hybrid ensemble method in Python. Ecol Inf. 65:101422.

Pal S, Talukdar S, Ghosh R. 2020. Damming effect on habitat quality of riparian corridor. Ecol Indic. 114:106300.

Panagos P, Ballabio C, Himics M, Scarpa S, Matthews F, Bogonos M, Poesen J, Borrelli P. 2021. Projections of soil loss by water erosion in Europe by 2050. Environ Sci Policy. 124:380–392.

Panahi M, Dodangeh E, Rezaie F, Khosravi K, Van Le H, Lee M-J, Lee S, Thai Pham B. 2021. Flood spatial prediction modeling using a hybrid of meta-optimization and support vector regression modeling. Catena. 199:105114.

Pham BT, Avand M, Janizadeh S, Phong TV, Al-Ansari N, Ho LS, Das S, Le HV, Amini A, Bozchaloei SK, et al. 2020. GIS based hybrid computational approaches for flash flood susceptibility assessment. Water. 12(3):683.

Pradhan B, Sameen MI, Al-Najjar HA, Sheng D, Alamri AM, Park HJ. 2021. A meta-learning approach of optimisation for spatial prediction of landslides. Remote Sens. 13(22):4521.

Rafiei Sardooi E, Azareh A, Mesbahzadeh T, Soleimani Sardoo F, Parteli EJ, Pradhan B. 2021. A hybrid model using data mining and multi-criteria decision-making methods for landslide risk mapping at Golestan Province, Iran. Environ Earth Sci. 80(15):1–25.

Rahmati O, Tahmasebipour N, Haghizadeh A, Pourghasemi HR, Feizizadeh B. 2017. Evaluation of different machine learning models for predicting and mapping the susceptibility of gully erosion. Geomorphology. 298:118–137.

Riyadi S, Mayanti S, Damariati C, Puspita S. 2020. Deep learning for pixel-based edge models classification of tertiary dentine images. In: Proceedings of the 2020 the 3rd International Conference on Information Science and System, March. p. 57–60.

Rodriguez-Galiano VF, Ghimire B, Rogan J, Chica-Olmo M, Rigol-Sanchez JP. 2012. An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS J Photogramm Remote Sens. 67:93–104. https://doi.org/10.1016/j.isprsjprs.2011.11.002

Roy P, Chakrabortty R, Chowdhuri I, Malik S, Das B, Pal SC. 2020. Development of different machine learning ensemble classifier for gully erosion susceptibility in Gandheswari Watershed of West Bengal, India. In: Machine learning for intelligent decision science. Switzerland: Springer Nature; p. 1–26.

Saha S, Roy J, Arabameri A, Blaschke T, Tien Bui D. 2020. Machine learning-based gully erosion susceptibility mapping: a case study of Eastern India. Sensors. 20(5):1313.

Schonlau M, Zou RY. 2020. The random forest algorithm for statistical learning. Stata J. 20(1):3–29.

Secretariat of the UNCCD. 2013. A stronger UNCCD for a land-degradation neutral world. Issue Brief, Bonn, Germany. p. 33.

Shit PK, Bhunia GS, Pourghasemi HR. 2020. Gully erosion susceptibility mapping based on bayesian weight of evidence. In: Gully erosion studies from India and surrounding regions. Cham: Springer. p. 133–146.

Simpson G, Wu YH. 2014. Accuracy and effort of interpolation and sampling: can GIS help lower field costs? IJGI. 3(4):1317–1333.
Süzen ML, Doymuran V. 2004. Data driven bivariate landslide susceptibility assessment using geographical information systems: a method and application to Asarsuyu catchment, Turkey. Eng. Geol. 71(3-4): 303–321. https://doi.org/10.1016/S0013-7952(03)00143-1.

Syam N, Kaul R. 2021. Random forest, bagging, and boosting of decision trees. In: Machine learning and artificial intelligence in marketing and sales. Bingley: Emerald Publishing Limited; p. 139–182. https://doi.org/10.1108/978-1-80043-880-4-20211006

Talukdar S, Mankotia S, Shamimuzzaman M, Mahato S. 2021. Wetland-inundated area modeling and monitoring using supervised and machine learning classifiers. In: Advances in remote sensing for natural resource monitoring. Hoboken, NJ: Wiley Online; p. 346–365.

Tang WJ, Liu HH, Liu BY. 2013. Effects of gully erosion and gully filling on soil degradation in the black soil region of Northeast China. J Mt Sci. 10(5):913–922.

Tella A, Balogun AL, Adebisi N, Abdullah S. 2021. Spatial assessment of PM10 hotspots using Random Forest, K-Nearest Neighbour and Naïve Bayes. Atmos Pollut Res. 12(10):101202.

Tilahun SA, Ayana EK, Guzman CD, Dagnew DC, Zegeye AD, Tebebu TY, Yitaferu B, Steenhuis TS. 2016. Revisiting storm runoff processes in the upper Blue Nile basin: the Debre Mawi watershed. Catena. 143:47–56.

Townshend JR, Justice CO. 1986. Analysis of the dynamics of African vegetation using the normalized difference vegetation index. Int J Remote Sens. 7(11):1435–1445.

Ul Din S, Mak HWL. 2016. Retrieval of Land-Use/Land Cover Change (LUCC) Maps and Urban Expansion Dynamics of Hyderabad, Pakistan via Landsat Datasets and Support Vector Machine Framework. Remote Sens. 13(16):3337. https://doi.org/10.3390/rs13163337.

USDA. 1999. Soil taxonomy; a basic system of soil classification for making and interpreting soil surveys, 2nd ed. Washington, DC: USDA, Natural Resources Conservation Service.

Vanmaercke M, Poesen J, Van Mele B, Demuzere M, Bruynseels A, Golosov V, Bezerra JFR, Bolysov S, Dvinskikh A, Frankl A, et al. 2016. How fast do gully headcuts retreat? Earth Sci Rev. 154:336–355.

Victoria AH, Maragatham G. 2021. Automatic tuning of hyperparameters using Bayesian optimization. Evol Syst. 12(1):217–223.

Vörösmarty G, Dobos I. 2020. Green purchasing frameworks considering firm size: a multicollinearity analysis using variance inflation factor. Supply Chain Forum. 21(4):290–301.

Wang H, Zhang L, Yin K, Luo H, Li J. 2021. Landslide identification using machine learning. Geosci Front. 12(1):351–364.

Wang Z, Liu Q, Liu Y. 2020. Mapping landslide susceptibility using machine learning algorithms and GIS: a case study in Shexian County, Anhui Province, China. Symmetry. 12(12):1954.

Wangchuk S, Bolch T. 2020. Mapping of glacial lakes using Sentinel-1 and Sentinel-2 data and a random forest classifier: strengths and challenges. Sci Remote Sens. 2:100008.

Wuepper D, Borrelli P, Panagos P, Lauber T, Crowther T, Thomas A, Robinson DA. 2021. A ‘debt’ based approach to land degradation as an indicator of global change. Glob Chang Biol. 27(21):5407–5410.

Xie J, Song Z, Li Y, Zhang Y, Yu H, Zhan J, Ma Z, Qiao Y, Zhang J, Guo J. 2018. A survey on machine learning-based mobile big data analysis: challenges and applications. Wireless Commun Mobile Comput. 2018:1–19.

Xu Y, Cong K, Zhu Q, He Y. 2021. A novel AdaBoost ensemble model based on the reconstruction of local tangent space alignment and its application to multiple faults recognition. J Process Control. 104:158–167.

Yao J, Shepperd M. 2020. Assessing software defection prediction performance: why using the Matthews correlation coefficient matters. In: Proceedings of the Evaluation and Assessment in Software Engineering. pp. 120–129.

Zabihi M, Mirchooili F, Motevalli AR, Khaledi Darvishan A, Pourghasemi HR, Zakeri MA, Sadighi F. 2018. GIS-based gully erosion spatial modeling in northern Iran. Catena. 161:1–13.

Zhou X, Lu P, Zheng Z, Tolliver D, Keramati A. 2020. Accident prediction accuracy assessment for highway-rail grade crossings using random forest algorithm compared with decision tree. Reliab Eng Syst Saf. 200:106931.

Zhou Y, Zhang B, Qin W, Deng Q, Luo J, Liu H, Yang D, Wang H, Zhao Y. 2021. Primary environmental factors controlling gully distribution at the local and regional scale: an example from Northeastern China. Int Soil Water Conserv Res. 9(1):58–68.

Zhu H, Gowen A, Feng H, Yu K, Xu JL. 2020. Deep spectral-spatial features of near infrared hyperspectral images for pixel-wise classification of food products. Sensors. 20(18):5322.