



Identifying sensitivity of factor cluster based gully erosion susceptibility models

Swades Pal¹ · Satyajit Paul¹ · Sandipta Debanshi¹

Received: 10 March 2022 / Accepted: 13 July 2022 / Published online: 26 July 2022
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2022

Abstract

The present study has attempted to address the issue of sensitivity of different clusters of factors towards gully erosion in the Mayurakshi river basin. Firstly, the gully erosion susceptibility of the basin area has been mapped by integrating using 18 parameters divided into four factor-cluster, viz. erodibility, erosivity, resistance, and topographical cluster, with the help of four machine learning (ML) models such as random forest (RF), gradient boost (GBM), extreme gradient boost (XGB), and support vector machine (SVM). Results show that almost 20% and 25% of the upper catchment of the basin belongs to extreme and high gully erosion susceptibility. Among the applied algorithms, RF is appeared as the best performing model. The spatial association of factor cluster-based models with the final susceptibility model is found the highest for the erosivity cluster, followed by the erodibility cluster. From the sensitivity analysis, it becomes clear that geology and soil texture are dominant contributing factors to gully erosion susceptibility. The geological formation of unclassified granite gneiss and geomorphological formation of denudational origin pediment-pediplain complex is dominant over the entire upper catchment of the basin, and therefore, can be considered regional factors of importance. Since the study has figured out the different grades of susceptible areas with dominant factors and factor cluster, it would be useful for devising planning for gully erosion check measures. From economic particularly food security purpose, it is very essential since it is concerned with precious soil loss and negative effects on agriculture.

Keywords Gully erosion susceptibility · Factor clustering · Machine learning algorithms · Sensitivity analysis

Introduction

Although gully erosion creates narrow and deep channels that occupy tiny parts of a catchment (Amare et al. 2021) but poses a significant impact on the regional geo-environment and economic prosperity (Roy and Saha 2021). It causes Badlands formation (Cánovas et al. 2017), removal of fertile topsoil (Han et al. 2018), reservoir sedimentation (Dutta 2016), and lowering of the groundwater table (Tilahun et al.

2016), etc. Therefore, gully erosion should be checked along with sustainable management practices in order to ensure future development (Arabameri et al. 2020a). Due to its severe impact, gully erosion has attracted plenty of research interest (Arabameri et al. 2020a; Saha et al. 2020; Busch et al. 2021; Yang et al. 2021a, b; Sidorchuk 2021). Gully erosion susceptibility mapping is considered to be the first step of implementing sustainable management practices (Debanshi and Pal 2020). Such mapping is possible by detecting the relationship between gully erosion and gully conditioning factors (Rahmati et al. 2017). The conditioning factors of gully erosion belong to different groups or clusters like topographical, erodibility, erosivity, etc. (Conforti et al. 2011). Sometimes, different resistance like vegetation cover and installation of check dams may play crucial positive role in controlling soil erosion, and those can be grouped as resistance cluster (Debanshi and Pal 2020). Though gully erosion susceptibility mapping has been paid enough attention, factor cluster-specific mapping and assessing their contribution to overall erosion susceptibility are not adequately explored.

Communicated by Marcus Schulz.

✉ Sandipta Debanshi
debanshi.sandipta93@gmail.com

Swades Pal
swadespal2017@gmail.com

Satyajit Paul
spaulofficial.geo@gmail.com

¹ Department of Geography, University of Gour Banga, Malda, West Bengal, India

Moreover, most of the existing studies have focused on the contribution of individual factors to susceptibility based on some statistical analysis. However, how inclusion or exclusion of one single factor can bring changes in the spatial pattern of susceptibility is almost absent. However, this analysis can exhibit regional priority factors of importance towards susceptibility.

The application of Geographical Information System (GIS) with the modern day's advanced modelling techniques has made spatial research on environmental events or natural hazards easier and more robust (Reichstein et al., 2019). However, the drawbacks of physical models based on statistical techniques, as mentioned by (Mosavi et al. 2018), promote the adaption of advanced data-driven algorithm-based models like machine learning (ML) in spatial research. The continuous advancement of the ML-based modelling approach over the last two decades showed its suitability for spatial prediction with an acceptable rate of outperforming the conventional models (Mosavi et al. 2017). In the case of spatially dense estimation, tree-based algorithms like decision tree (DT) or RF are considered efficient ML algorithms. Ortiz-García et al. (2014) described how ML algorithms efficiently model complex data structures of the input variable and produce output. Many algorithms like neuro-fuzzy (Dineva et al. 2014), artificial neural networks (ANNs) (Kim et al. 2016), support vector regression (SVR) (Taherei Ghazvinei et al. 2018), support vector machine (SVM) (Mosavi et al. 2017), etc. are reportedly capable of making both short and long-term prediction. Apart from that, there are several forms of principal monotonicity inference (PMI) like discrete principal monotonicity inference (DPMI), multivariate principal monotonicity inference (MPMI) which are used to handle irregular, nonlinear, uncertain, and multi-variate dependent data in the field of hydro-system (Cheng et al., 2016), climate classification (Cheng et al. 2017). In the present context, various ML algorithms have been explored and effectively used to predict and compare gully erosion. Along with predicting environmental events like gully erosion, Pourghasemi et al. (2020) presented the role of ML in the selection of its controlling factors.

Sometimes, the spatial models developed with the help of ML or DL algorithms provide a generalized result (Maxwell et al., 2016; Neyshabur et al. 2017; Kawaguchi et al. 2017; Maxwell et al., 2020a, b; 2020a). The root of generalization is geographical diversity, error in machine learning through training datasets (Maxwell et al., 2020a, b), or the disparate dataset and limited training sites (Hoeser and Kuenzer 2020; Hoeser et al., 2020). Limited training sites based on spatial model building and interpolation are often done by scholars. This process is faster but makes output coarser in resolution since it creates a map based on the interpolation method. For the sake of precision in output, instead of this approach, the pixel inclusive approach is essential for such modelling.

Python libraries have the potential to include all the pixels over a larger geographical region. It is open-source software, and algorithms can be programmed as per need (Brownlee 2019). Moreover, the optimization of hyperparameters is an integral part of ML-based modelling to enhance the accuracy of the prediction (Kotthoff et al. 2019). This pixel inclusive work is highly time-consuming. However, it can yield a more precise result which is of utmost necessity in the planning process.

Hyper-parameters are those parameters that need to be set before training the machine, and those can be achieved either by manual searches or by the auto-optimization programme (Pradhan et al. 2021). Manual searches require previous knowledge, expertise, and professional skill; therefore, it becomes difficult to set effective hyper-parameters (Bergstra and Bengio 2012). On the other hand, the auto-optimization process allows for overcoming these difficulties (Pradhan et al. 2021). Python libraries have a broader scope of hyper-parameter optimization (Brownlee 2019) with robust auto-optimization functions (Mitrpanont et al. 2017). Therefore, using the python libraries, the present study has attempted to fill the aforementioned gap and aimed to produce pixel inclusive spatial model of gully erosion susceptibility based on the factor cluster model and assessed to explain the role of individual factor clusters in overall susceptibility. In addition, this study also has attempted to explore the role of the individual factors at pixel scale to determine the spatial pattern of the susceptible areas. This may be a novel approach to investigating gully erosion which is capable of identifying the sensitivity of each factor cluster towards the occurrence of gully erosion and, in turn, may also help determining the local factor of gully erosion.

Study area

The present study area is the Mayurakshi river basin which is located over the Chottanagpur plateau fringe region of eastern India. The basin area spreads over more than 5400 km² between 23°15'N to 24°34'15"N to 86°58'E to 88°20'30"E over two states of India, namely Jharkhand and West Bengal (Fig. 1a). The entire basin area is characterized by the Rarh tract, and the Granitic gneiss is the dominant rock type that covers most of the upper and middle catchment of the basin (Fig. 1b). In the north-eastern part of the basin, over a considerable area, granet-biotite gneiss is also found. Geomorphologically (Fig. 1c), these parts of the basin belong to the denudational origin pediment-pedepain complex and dissected hills and valleys. On the other hand, the lower part of the basin consists of older and newer alluvium flood plain of fluvial origin. The confluence part of the basin on the eastern side is merged with the Ganges delta. The lower catchment of the basin is not subjected to considerable gully erosion.

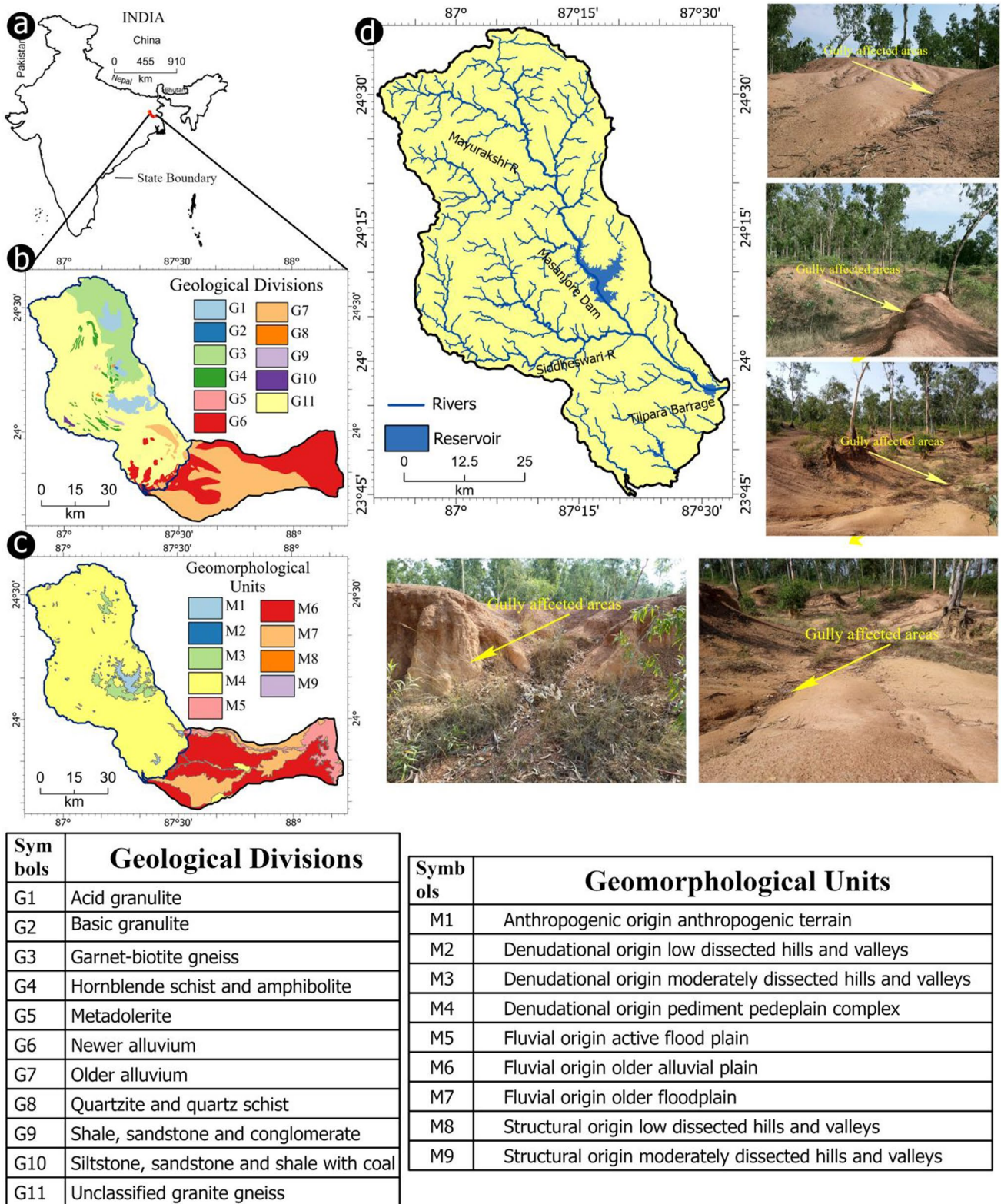


Fig. 1 Location of the study area with geographical setup. **a** India, **b** geological division, **c** geomorphological divisions, **d** area of interest

On the contrary, almost all the gully formation is witnessed in the upper and middle catchment of the basin. Field investigation and topographical maps show, in this basin area, the geological formations of granitic gneiss formation and granet-biotite gneiss coupled with the geomorphological formation of denudational origin pediment-pediplain complex and denudational origin moderately dissected hill and valleys are mostly subjected to gully formation and erosion. Therefore, for minutely focusing on the most gully erosion effected area, the lower catchment of the basin has been excluded from the investigation, and most affected geological and geomorphological divisions have been taken as an area of interest in this study (Fig. 1d).

This region comes under tropical monsoon climate type with hot-humid summer and cold-dry winter. The average annual temperature of the region is 24–28 °C, with a minimum of 15–17 °C and a maximum of 35–38 °C. Average yearly rainfall is about 1650–1700 mm, and more than 80% is experienced in the monsoon season, covering the months of June to September. Such a climatic profile corresponding to the aluminium and ferrous mineral of the soil promotes lateralization (Ghosh et al. 2015). These lateritic tracts are highly prone to gully formation. Especially the high summer temperature and mild cold in the winter highly favour laterite formation. This lateralization leads to greater solubility of the silica and increases the chance of desertification. Moreover, <20 °C average temperature intensifies the mechanical weathering and a considerably high amount of rainfall in the monsoon season results in leaching, where fluctuation of

groundwater in response to the wet and dry season conducts groundwater laterite formation (Sarkar et al. 2020).

Materials and methods

Materials

Eighteen data layers have been used for creating a gully erosion susceptibility map in the present study region. All the selected parameters and the data sources used for preparing the spatial data layers are mentioned in Table 1. All Landsat images and SRTM DEMs have been pre-processed before final layer preparation.

Selection of contributing factors, factor clustering, and data layer preparation

The parameter for gully erosion susceptibility mapping has been divided into four clusters, namely erodibility, erosivity, resistance, and topographical factors (Fig. 2). These clusters are formed based on their role in determining the gully erosion. Studies like Conforti et al. (2011) Debanshi and Pal (2020), and Saha et al. (2020) reported similar factor clustering in their studies. The factors under different clusters are mentioned in Table 1. Factors have been selected considering their influence on gully initiation and expansion. Few topographical configurations like slope, curvature, stream power index (SPI), distance from existing gullies, etc. play a

Table 1 Factors under different clusters, their data sources, and VIF scores

| Clusters | Factors | Data source | VIF |
|-------------|---------------------------------|---|------|
| Topographic | Slope | SRTM DEMs (spatial resolution — 30 m) | 1.55 |
| | Curvature | | 1.41 |
| | SPI | | 1.47 |
| | Distance from stream | | 1.49 |
| Erodibility | Geology | 1st and 2nd order stream extracted from SOI toposheets | 1.88 |
| | Geomorphology | Geological map provided by GSI | 1.48 |
| | Soil texture | Geomorphological map downloaded from BHUVAN (an Indian web absed utility) | 1.82 |
| | BSI | Field data and NIC provided village level texture data | 3.29 |
| | Ferrous content | Landsat satellite images (spatial resolution 30 m, path/row- 139/43) | 3.46 |
| | LULC | | 1.17 |
| Erosivity | Average monsoon rainfall | DST provided annual rainfall map | 3.51 |
| | Soil erosivity | District resource map of rain fall distribution | 3.10 |
| | Monthly temperature fluctuation | Landsat satellite images (spatial resolution 30 m, path/row- 139/43) | 2.54 |
| | WSVI | | 2.42 |
| Resistance | NDVI | | 3.46 |
| | Distance from GHB | Location of GHBs extracted from SOI toposheets | 1.76 |
| | Distance from check dams | Location of check dams extracted from Google earth imagery | 1.42 |
| | Surface water coverage | Coverage of reservoir extracted from SOI toposheets | 2.47 |

significant role behind gully erosion. Several previous studies (Meliho et al., 2018; Shit et al. 2020; Azedou et al. 2021) have used these factors for assessing gully erosion in a similar environment. These factors have been selected as the topographical factor cluster in this study.

Slope, SPI directly influence the erosion rate and also, to some extent, determine 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). 1st and 2nd-order streams often behave like gullies and tend to expand and pose a risk of gully erosion in the nearby areas (Araujo and Pejon 2015; Joshi et al. 2016). The straight-line distance from these streams has been computed by extracting the 1st and 2nd order streams from the SOI toposheets. SRTM DEM has been employed to derive the data layers of the slope, curvature, and SPI. The surface analysis tool of ArcGIS software (v-10.2) has been used to facilitate slope and curvature derivation (Fig. 2r and o), while the raster calculator tool has been used to calculate SPI (Eq. 1) (Fig. 2p).

Apart from the topographical configuration, there are a few characteristics of soil and land surface which influence the propensity of gully erosion. These characteristics are geological division, geomorphological division, the texture of the surface soil, ferrous content in the surface soil, the bareness of the land surface, LULC features of the region, etc. All these have been included in this study as factors under the erodibility cluster. Geological characteristics, including rock types and lithological alignments, are very closely related to gully erosion and determine the gully development process according to composition characteristics and mechanical properties (Conforti et al., 2011). The frequency of existing gullies per unit area of each geological division has been calculated, and a raster layer has been produced (Fig. 2c) to facilitate the input of spatial model development. Similarly, the geomorphological environment determines the nature and intensity of erosion (Evelpidou et al. 2018), and the same procedure has been followed for geomorphological divisions as well (Fig. 2f). Soil texture is strongly correlated to soil erosion, where a greater proportion of coarser sand in the top layer of soil makes favourable conditions for gully initiation (Pal 2016; Pal and Debanshi, 2018). For producing the spatial data layer of soil texture (Fig. 2a), the proportion of coarser sand in the topsoil layer of the upper catchment was measured by extracting soil samples from 43 sites and testing them with the help of a digital sieve shaker. The village-level soil texture maps of the National Informatics Centre (NIC) have also been taken to get the data regarding the soil texture. The percentage of coarser sand in the soil surface has been identified based on USDA (1999) soil classification system. The inbuilt function

of the ERDAS Imagine software (v-9.2) has been used to produce the data layer of ferrous content (Fig. 2e). 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) identified this factor important for gully development in this region. Since the bare ground positively contributes to gully development due to its lower resistance (Jahantigh and Pessarakli, 2011), Eq. 2 has been implemented in the raster calculator for deriving Bare Soil Index (BSI) (Elfadaly et al. 2017) and incorporating bareness of the ground as an input factor (Fig. 2d). LULC of a region can influence the gully development in many ways (Gelagay, and Minale, 2016). Supervised classification based on maximum likelihood classifier has been performed in ERDAS Imagine software (v-9.2) for LULC identification and mapping (Fig. 2b).

Besides inherent characteristics of the soil surface, climatic factors like rainfall intensity, rainfall erosivity, fluctuation of surface temperature, dryness, etc. are also capable of significantly contributing to the soil and gully erosion. These factors have been considered under erosivity cluster. The DST-provided rainfall map has been used for preparing the data layer of average yearly rainfall (Fig. 2h). Soil erosivity has also been incorporated in this study to take the impact of rainfall intensity on soil into consideration. The concentrated rainfall in fewer months causes higher erosion than the uniform rainfall for the entire year (Pal and Debanshi, 2018). For producing the data layer (Fig. 2g), soil erosivity has been calculated using the modified Fournier index (Eq. 3) (1960). 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. 4) 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 (Fig. 2i). Sometimes, dryness of the surface soil may promote the severity of erosion; therefore Water Supplying Vegetation Index (WSVI) has been calculated (Eq. 5) (Fig. 2j) as an indicator of soil moisture.

The resistance to the gully erosion in the form of vegetation coverage, water coverage, and implementation of gully checking measures is also important for assessing gully erosion susceptibility. Considering the ability of the denser forest 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. In the present area, multiple patches of semi-denser forest areas are noticed. Therefore, calculation of NDVI (Eq. 6) (Townshend and Justice 1986) and spatial mapping (Fig. 2k) have been performed. Some gully controlling measures like the construction of Gully Head Bundhs (GHBs) and check

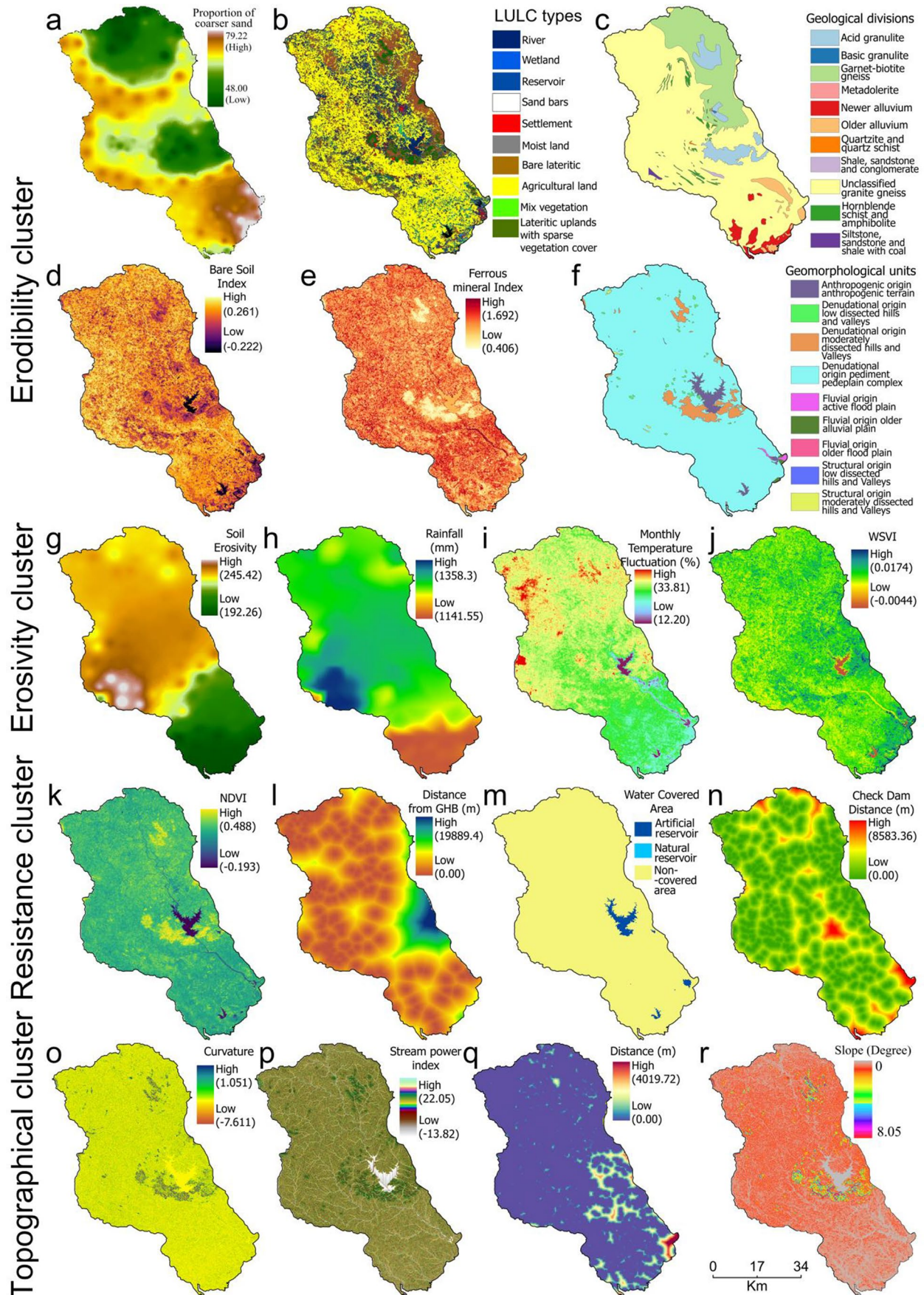


Fig. 2 Raster data layers for the respective factors, **a** soli texture, **b** LULC, **c** geological aspect, **d** BSI, **e** ferrous mineral, **f** geomorphology, **g** soil erosivity, **h** monthly average rainfall, **i** monthly fluctuation of surface temperature, **j** WSVI, **k** NDVI, **l** distance from gully headed bundh, **m** water coverage, **n** distance from check dams, **o** curvature, **p** SPI, **q** distance from first- and second- order stream, **r** slope

dams were constructed in this region. The straight-line distance from the GHB and check dams has been measured, and maps have been prepared (Fig. 2l and o). The parameter of water coverage has been considered in this study because of the presence of large reservoir river projects like Massanjore or Tilpara in the area. The data layer (Fig. 2m) has been prepared by categorizing the basin area into the waterlogged and non-water logged areas and demarcating the reservoir area from SOI toposheet. All these gully erosion resisting factors have been considered under the resistance cluster.

$$SPI = A_s * \tan \sigma \quad (1)$$

where, A_s is the specific catchment area in metres, σ is the slope gradient in degrees.

$$BSI = \frac{(b_{SWIR} + b_R) - (b_{NIR} + b_B)}{(b_{SWIR} + b_R) + (b_{NIR} + b_B)} \quad (2)$$

where, b_{SWIR} is the short wave infra-red band brightness value, b_R is the red band brightness value, b_{NIR} is the near infra-red band brightness value, b_B is the blue band brightness value.

$$FI = \sum_{I=1}^{12} \frac{P_i^2}{P} \quad (3)$$

where,

Where, P_i is the precipitation of month I , and P is the mean annual precipitation.

$$LST = T_B / [1 + \{(\lambda * T_B / \rho) * \ln \varepsilon\}] \quad (4)$$

where, T_B is the at satellite temperature, λ is the wavelength of emitted radiance in metres, $\rho = h * c / r$, h is the Planck's constant, r is the Boltzmann constant, and c is the velocity of light, ε is the land surface emissivity.

$$NDVI = \frac{(b_{IR} - b_R)}{(b_{IR} + b_R)} \quad (5)$$

where, b_{IR} is the infra-red band brightness value, b_R is the red band brightness value.

Assessing multi-collinearity of the datalayers

While proceeding with multiple factors, the existence of inter-correlated factors, which is termed multi-collinearity,

may reduce the accuracy of the model output considerably (Kalantar et al., 2020). It arises when in multi-variate modelling, multiple factors considered an independent variable have equal prediction capability (Jaafari et al. 2017). In this study, the conditioning factors of gully erosion are subjected to multi-collinearity, and therefore, a multi-collinearity test using the variance inflation factor (VIF) test (Eq. 7) has been done on employed factors under different factor cluster (Arora et al., 2019). The VIF generally measures the disagreement within a model with multiple relations using the variance of a model with the target variable alone (Hong et al. 2019). The entire procedure involves Eqs. 7 and 8, where n different VIFs have been calculated. Parse, in the case of a linear model with n variable, an ordinary least square regression is conducted first with Y and all the explanatory variables (Eq. 7).

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + e \quad (6)$$

where, α is the constant, β is the slope, and e is the error. The VIF of i th factor is calculated using the following formula (Eq. 8).

$$VIF_i = \frac{1}{1 - R_i^2} \quad (7)$$

where, R_i^2 is the coefficient of determination of i th factor in the above equation.

The entire calculation has been carried out in the SPSS (v-22) environment in the present study. The value of the co-linearity test in this study ranges from 1.169 to 3.512, which is less than the maximum threshold value for co-linearity tolerance ($VIF \geq 5$) (Lin and Billa 2021; Maiti et al. 2021) (Table 1). Therefore, the selected factor is non-collinear to each other and quite suitable for the present study.

Modelling gully erosion susceptibility

Selection of gully and non-gully effected regions

To identify the gully affected and non-gully affected points, high-resolution Google earth imagery with 2.62 m spatial resolution, field investigation, and SOI toposheet with a 1:50,000 scale have been used. A total of 3658 points from both the gully and non-gully erosion points have been taken, from which 80% (2926 points) data has been used for modelling and 20% (732 points) kept for validation purposes. For gully erosion cluster modelling, the entire dataset of the gully and non-gully affected zones have been convert into binary classes, where '1' is considered gully affected sites and '0' is considered non-affected sites.

Modelling the factor clusters

Four machine learning (ML) classification algorithms, namely random forest (RF), gradient boosting (GDB), XGBoosting (XGB), and support vector machine (SVM), have been used for four factor cluster binary classification in the python environment. The classified maps are combined in the ArcGIS environment in order to prepare the final gully susceptibility map (Fig. 3). Details of methodological analysis of the factor clusters mapping are discussed in subsequent section.

Random forest (RF)

The random forest (RF) is a robust ensemble ML algorithm used for both classification and regression for unsupervised learning (Schonlau and Zou 2020). The RF algorithm is widely used in spatial modelling (Meshram et al. 2021).

RF uses DT's as the core element to reduce the estimate of classes. Boots trap aggregation is an integral part of RF-based estimation and training the dataset (Syam and Kaul 2021; Meshram et al. 2021). The Python-based *Scikit-learn* ML library has been used for cluster modelling in this present study. To optimize the RF algorithm, the selected eighteen parameters have been considered input variable for the number of trees. To reach the best possible output, 5- and tenfold cross-validation has been applied in this study. Since the factor cluster are varying to each other, the 'vote' technique of RF has been used to reduce the noise or outlier in the gully erosion controlling factors as input variables. The four gully erosion factor clusters are pretty distinctive from each other such as erodibility cluster differs from topographical, erodibility, and resistant clusters, yet they influence the overall condition of the gully. Such dimensionality and distinctiveness of the pre-cursor variable can affect the model's overall performance. Apart from clustering the variable according to their nature of influence on gully, the

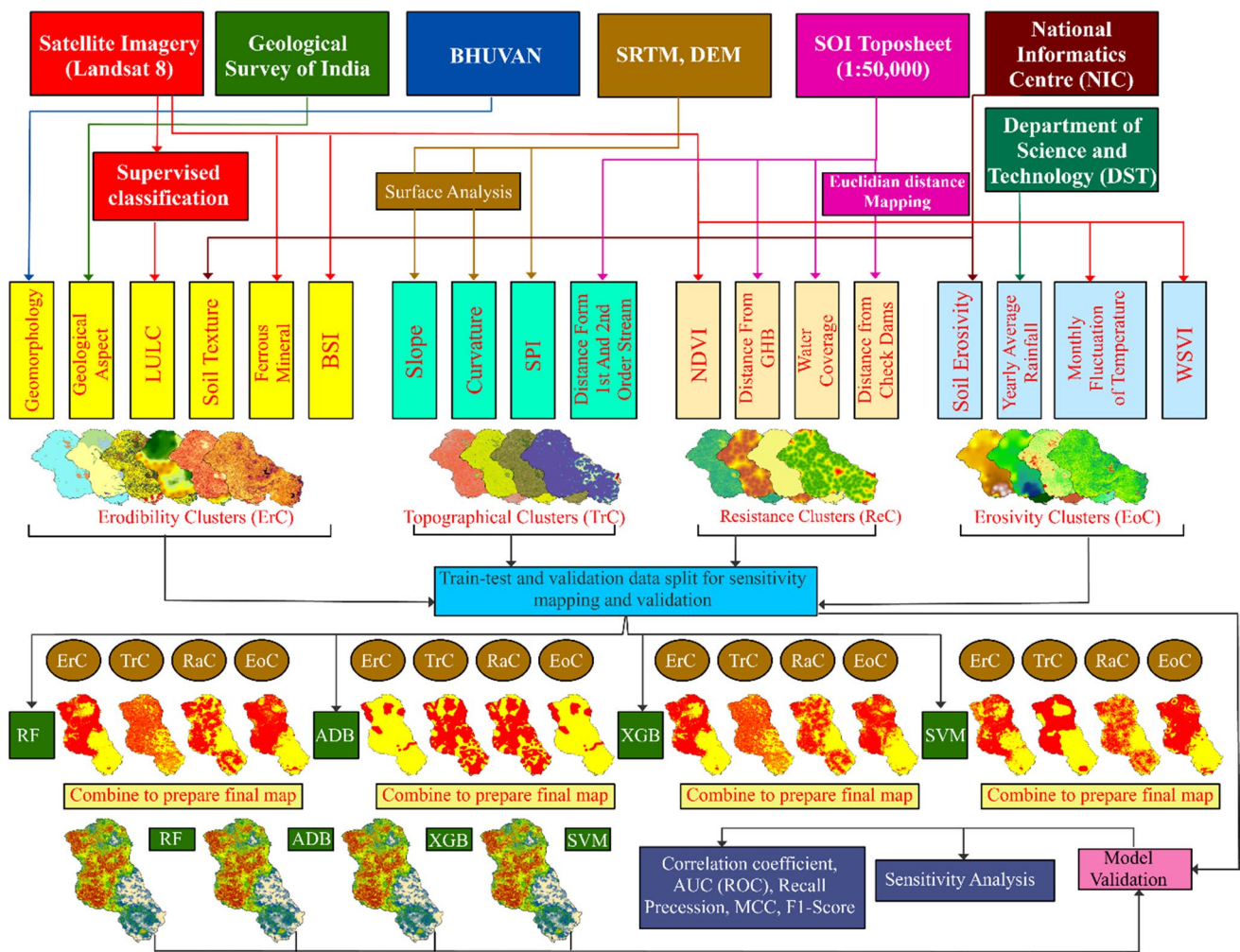


Fig. 3 Methodological proceedings of the study

feature section system of RF has been used, which can also improve the accuracy of the model. Tella et al. (2021) used similar feature section of RF in their study. A proximity algorithm is used to locate the outliers in terminal nodes (Norouzi and Moghaddam 2020).

Gradient boosting model (GBM)

Gradient boosting (GBM) is another tree-based ensemble machine learning technique used for supervised learning (Zhang et al. 2021). Although the RF technique is a step ahead of its predecessor technique (DTree's model) by using 'voting', feature section technique, and proximity algorithm, it lacks in determining weaklings in the database (Yang et al. 2021a, b). The GBM uses the classification and regression tree (CART) technique to determine the weak learners among predictor variables (Handoko et al. 2020). Studies by Jun (2021) and Yang et al. (2021a, b) showed that GBM has strong predictive power. This technique is beneficial for the present study, where four-factor clusters incorporating the gully controlling factors are distinctive and influence each other in many ways in the occurrence of gully erosion. Such weak predictors can disrupt the model's performance; therefore, GBM-based 'best-fit' optimization has been used to improve the performance of and accuracy of the employed model. Although the GBM has a higher predictive capability than RF, in the case of noisy data, it sometimes leads to overfitting (Yang et al. 2021a, b). Therefore, the 5 and tenfold cross-validation technique has been applied to optimize the model's performance in this study to avoid this overfitting problem of GBM. From the *Scikit-learn* ensemble package of Python, *GradientBoosting* classifier has been used for this study to run the algorithm.

XGBoost (XGB)

Among the different boosting techniques, the XGBoost or Xtreme gradient boosting technique is quite an efficient and popular method and it is widely used for its excellent predictability and efficiency. It is quite recent algorithm, proposed by Chen and Guestrin (2016). XGBoost is an extreme variation of the gradient boosting model (GBM) (Osman et al. 2021; Ke et al. 2017). XGB uses regularization boosting and parallel processing techniques to overcome the overfitting problem (Gui et al. 2020; Fauzan and Murfi 2018). Unlike other GBM, XGB is powerful and effective in noisy datasets (Naghibi et al. 2020; Liu et al. 2020). Therefore, the predictor selection method in a backward manner has been applied to structure the model for better performance (Sahin 2020; Li et al. 2020). With such optimization, the model takes more time but gives better accuracy by using

appropriate predictors for our dataset. This ensemble boosting approach can fit non-linear relationships; therefore, this technique can fit the non-linear relationship between the predictors of the four factor clusters. Where factor clusters are the independent variable and the 3658 sample points from both the gully and non-gully erosion sites are the target variables. The 80% of the dataset has been used to train the dataset and find out the set of predictors with maximum predictive capability using 5- and tenfold cross validation technique and grid search method. After that, the set of predictors has been run on the entire dataset to prepare the final cluster-wise susceptibility maps. The *XGBClassifier* from the *SK-learnxgboost* ensemble package has been used to predict the clusters in this present study.

Support vector machine (SVM)

Along with ensemble machine learning and generic programming algorithms, SVM represents a newer and advanced generation of ML algorithms (Du et al. 2020). In simple terms, SVM generates an optimum hyperplane clusters in a hyper-surface using a binary classifier (Acortes, and Vapnik 1995). In generic machine learning, training of the algorithm is done to minimize the empirical training error, which leads to an overfitting problem (Deiss et al., 2020). SVM maximizes the boundary between hyperplane and data to minimize the generalization error in the predicted data (Talukdar et al., 2021).

During model training, SVM searches hyper-plane that best separates the dataset into 1 and -1 binary classes, which can be identified as h_1 and h_{-1} or support vectors (Deiss et al., 2020). In this study, the gully susceptible and non-susceptible areas can be distinguished by using the hyperplane separation function of SVM where gully susceptible pixels are considered 1 and non-susceptible as 0. The separating margin or hyperplane between h_1 and h_{-1} is referred as 0 or non-linear convex programming problem. The non-linear relationship among the factor cluster parameters can be determined using such a function. The pairs of parameters of the separating hyperplane determination can be determined by solving the following optimization problem (Eq. 9):

$$\min_{w,b,\xi} \frac{1}{2} \| (w^T \cdot w) + C \left(\sum_{i=1}^l \xi_i \right) \quad (8)$$

$$\text{In respect to } 1 - \xi_i - y_i(w \cdot x + b) \leq 0, -\xi_i \leq 0, i = 1, \dots, n \quad (9)$$

where, penalty parameter (C) and margin of tolerance (ξ) needed to be tuned in order to accuracy and efficiency of the model.

To find out the susceptible and non-susceptible zone over the study area among the four factor clusters, the gamma function along with radial basis function (RBF) has been used where gamma value ranges from 1 to 0.0001 for different clusters and ‘rbf’ kernel used to distinguish the hyperplane among the different clusters. The grid search CV technique with 5- and tenfold CV have used to obtain maximum performance of the employed models. Parameter combination with maximum score has been used generate final spatial layers. Python-based SVC algorithm has been used to built the model.

Hyperparameter optimisation

At any machine learning, a selected or default set of parameters cannot perform equally on different sets of databases; therefore, for optimum performance of the given input data, we have employed the Grid-Search hyperparameters optimization technique for this present study. A tenfold K cross-validation technique is also used along with Grid-Search for better data optimization (Gulzat et al. 2020). Studies like Daviran et al. (2021), Abdi (2020), and Kaur et al. (2020) reported that K -fold cross-validation along with Grid-Search has a better capability for parameter selection and also optimization. Table 2 shows that values for different hyperparameters of each model have been given. After the selection of parameter values, the model has tested on the sample

dataset as well as the entire dataset, and final predicted values were obtained.

Accuracy assessment

The present work has employed two different types of validation techniques, namely accuracy matrices and sensitivity analysis to assess the accuracy of the different models.

Performance assessment of the models

Performance of the employed ML techniques has been evaluated through six calculated matrices considering ground truth data extracted from 20% of the gully and non-gully samples sites, based on high-resolution Google Earth imagery. The six employed matrices are the percentage of correctly classified data, the area under receiver operating characteristics (AUROC) or AUC(ROC), precision, sensitivity, F1-score, and Matthew’s correlation coefficient (MCC) (Eqs. 9–12). The validation metrics have been calculated based on four metrics such as true positive, false negative, true negative, and false positive. Percentage of correctly classified data calculated against true positive, true negative, and total size of the sample, whereas sensitivity and precision are based on true positive, false negative, and false positive values of the predicted data (Harimoorthy and Thangavelu 2021). In Matthew’s correlation coefficient (MCC), all four metrics have been used for better reliability and performance

Table 2 Hyperparameter optimization outputs using GridSearch

| Models | RF | GBM | XGB | SVM |
|-------------|--|--|---|---------------------------------|
| Erodibility | Bootstrap: true, max depth: 10, max features: 3, min samplesleaf: 3, min samplesplit: 10, no. of estimators: 100 | Learning rate: 0.001, max depth: 80, no. of estimators: 200, subsample: 0.5 | Colsample by tree: 0.7, gamma: 0.3, learning rate: 0.1, max depth: 50, min child weight: 1 | C: 100, gamma:0.1, kernel: rbf |
| Erosivity | Bootstrap: true, max depth: 10, max features: 2, min samplesleaf: 3, min samplesplit: 8, no. of estimators: 300 | Learning rate: 0.01, max depth: 100, no. of estimators: 1000, subsample: 0.7 | Colsample by tree: 0.5, gamma: 0.4, learning rate: 0.3, max depth: 50, minchild weight: 1 | C: 10, gamma:0.1, kernel: rbf |
| Resistance | Bootstrap: true, max depth: 80, max features: 3, min samplesleaf: 3, min samplesplit: 8,no. of estimators: 300 | Learning rate: 0.01, max depth: 100, no. of estimators: 300, subsample: 0.5 | Colsample by tree: 0.5, gamma: 0.2, learning rate: 0.05, max depth: 15, min-child weight: 7 | C: 1000, gamma: 1, kernel: rbf |
| Topographic | Bootstrap: true, max depth: 10, max features: 2, min samplesleaf: 3, min samplesplit: 8,no. of estimators: 200 | Learning rate: 0.1, max depth: 100, no. of estimators: 100, subsample: 0.5 | Colsample by tree: 0.5, gamma: 0.0, learning rate: 0.01, maxdepth: 15, mini-child weight: 3 | C: 10, gamma:0.001, kernel: rbf |

(Chicco et al. 2021). The MCC value ranges from -1 to 1 , where -1 indicates the least accuracy in the prediction and vice-versa. AUROC is calculated against the true and false positive value, of the calculated against the predicted data, and finally, F1-score is calculated based on the precision and sensitiveness of the dataset. AUROC, F1-score, and precision range between 0 and 1 , where a value near 1 indicates more reliability and accuracy of models. Equations for the stated metrics are expressed in Eqs. 11–14.

$$\text{recall} = \text{sensitivity} = \frac{Tp}{Tp + Fn} \quad (10)$$

$$\text{precision} = \frac{Tp}{(Tp + Fp)} \quad (11)$$

$$\text{MCC} = \frac{Tp \times Tn - Fp \times Fn}{\sqrt{(Tp + Fp)(Tp + Fn)(Tn + Fp)(Tn + Fn)}} \quad (12)$$

$$\text{F1 - score} = \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (13)$$

Sensitivity analysis

To measure the input parameter's, cluster's-impact on susceptibility at spatial scale, sensitivity analysis has been done. In this process, six parameters, namely NDVI, slope, stream distance, BSI, soil erosivity, soil texture, and geology, have been selected according to their importance and contribution in the cluster (importance value derived from cluster wise ML algorithms). The parameters have been subtracted one by one to generate the sensitivity model and see their individual impact at local, regional levels. AUROC, precision, sensitivity, F1-score, and MCC matrices have been calculated to measure the degree of sensitivity of the parameters. In order to detect the role of each cluster towards gully erosion susceptibility, correlation coefficients for each factor-based model and final model of gully erosion susceptibility have been calculated. For doing so, including all the pixels of the study area, PCA tool from ArcGIS software has been employed.

Results

Gully erosion susceptible models and their accuracies

Figure 4 shows the gully erosion susceptible models developed using ensemble models like (a) random forest, (b) gradient boost model, (c) XGBoost, and (d) SVM. Each model has

been classified into five susceptibility classes, from extremely susceptible to relatively safe. All the models show that 15–20% of the study region, particularly in the upper parts of the study region dominated by Granite gneiss geological formation, is extremely susceptible to gully erosion (Table 3). All the models have pointed out almost the same geographical area belonging to unclassified granite gneiss as gully erosion susceptible. This tract is overlaid with thick laterite soil of granite and gneissic origin. Oxidation is a dominant process of the weathering process, which encourages the consequent lateralization process in this area (Ghosh and Guchhait 2015). High silica and ferrous mineral content increase the erodibility of soil. Erosivity triggered by highly skewed rainfall during monsoon time enhances the mechanism of gullying and extension of the gully (Ghosh et al. 2015). Headward erosion of gully, gully widening, and deepening are the chief mechanism of gully induce soil erosion (Pal 2016; Arabameri et al., 2020b). Toe cutting and failure of gully banks are also found in some deep gullies. In the plateau fringe area, where a relatively steeper slope exists, this rate of gullying is more prominent. A wide part of the hilly region and lateritic tract of the study region is composed of sal (*Shorea robusta*) forest, but lack of undergrowth insists gully erosion even in the forested region. Deforestation is also a major issue in this region, and it also stimulates gully erosion activities (Pal and Debanshi, 2018). In all the susceptible classes, the areal extent differs marginally, and therefore, all the models could provide quite a similar kind of accuracy. However, to assess the most suitable one, an accuracy assessment is highly essential.

Table 4 depicts the accuracy level of the applied model and their performances. AUC value ranges from 0.78 to 0.91 , MCC value ranges from 0.76 to 0.91 , precision, sensitivity, and F1 score range from 0.75 to 0.94 , 0.81 to 0.92 , and 0.78 to 0.93 respectively. All these values have clearly certified that the applied models are in good to extremely good agreement with training and testing data. Among the applied models, RF appears as the best representative since the accuracy and performance levels are the highest, followed by GBM, XGBoost, and SVM (Table 4).

Field evidence-based study also validated the same. For instance, the area between Amgachhipahar and Jhar-napahar in the Masalia community development block (at the north-eastern side of Masanjore dam) is a highly gully occurrence and susceptible site, which was recognized by the RF model as an extremely susceptible area, whereas other models identified the same area mostly as moderate to highly susceptible.

Factor cluster-based gully erosion susceptible models

Figure 5 portrays four factor cluster-based gully erosion susceptibility models depicting the areas of susceptibility and

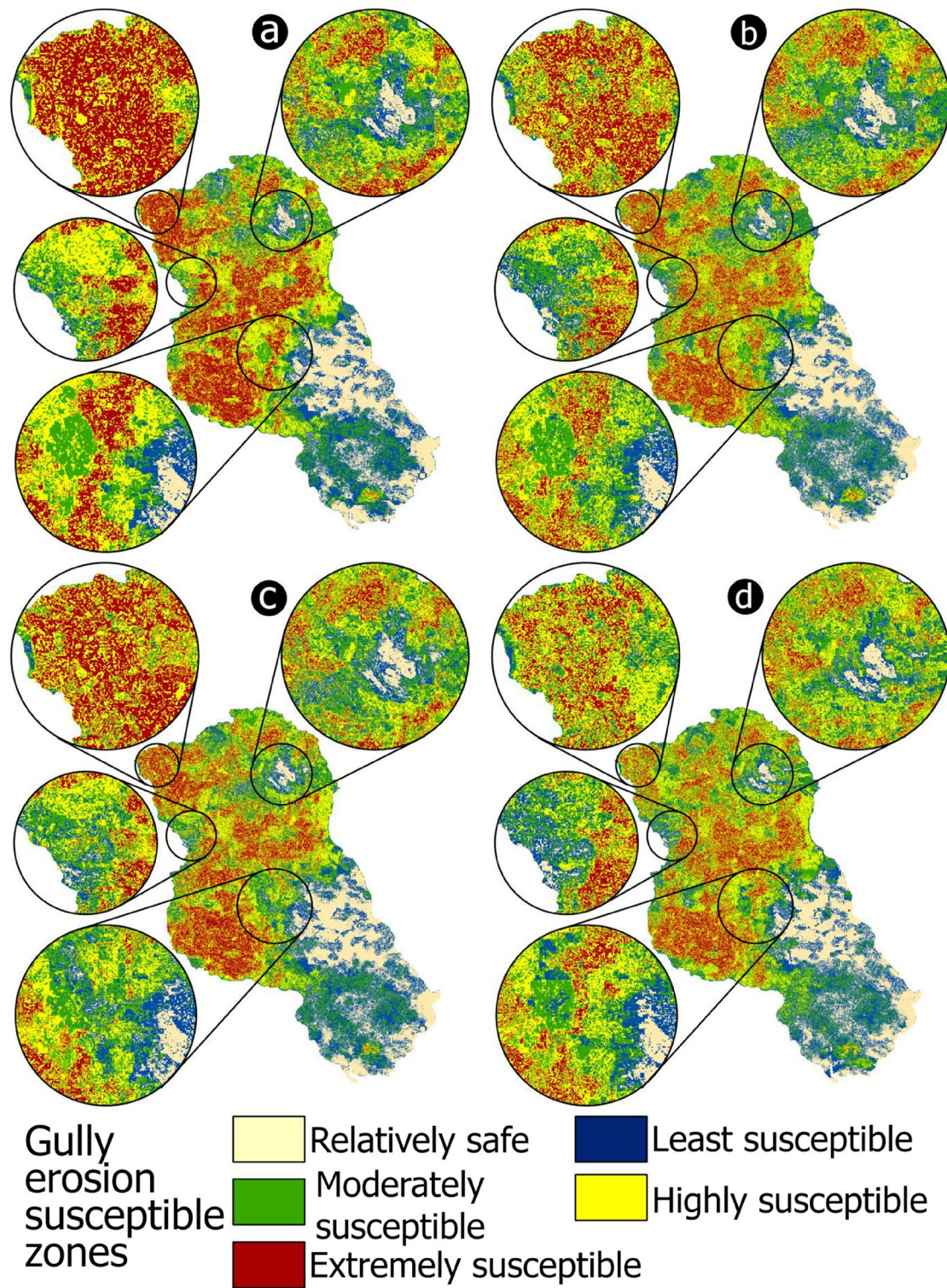


Fig. 4 Gully erosion susceptible zones **a** random forest, **b** gradient boost model, **c** XGBoost, and **d** SVM

non-susceptibility. Erodibility and erosivity factor cluster-based models depict a wider part of the upper catchment in the continuous stretch is susceptible to gully erosion. Since the geological and geomorphological divisions and soil

texture are primary factors of erodibility cluster, the concentration of gully erosion-prone divisions of geology and geomorphology in the upper catchment provided such result. Soil texture is another important erodibility factor, and a

Table 3 Percentage of area under different gully erosion susceptible zones

| Models | RF | GBM | XGB | SVM |
|------------------------------------|--------------------------|--------------------------|--------------------------|--------------------------|
| Gully erosion susceptibility zones | Percentage to total area | Percentage to total area | Percentage to total area | Percentage to total area |
| Relatively safe | 16.62 | 17.66 | 16.96 | 16.06 |
| Less susceptible | 18.04 | 19.30 | 19.61 | 19.13 |
| Moderate susceptible | 19.55 | 21.82 | 21.42 | 23.37 |
| High susceptible | 25.64 | 25.58 | 25.13 | 26.84 |
| Very high susceptible | 20.14 | 15.64 | 16.89 | 14.60 |

Table 4 Statistical accuracy assessment of different ML models

| Classifiers | K-fold | Correctly classified (%) | ROC (AUC) | Precision | Sensitivity | F1-score | MCC | Support |
|-------------|--------|--------------------------|-----------|-----------|-------------|----------|-------|---------|
| RF | 5 | 97.25 | 0.91 | 0.941 | 0.921 | 0.931 | 0.912 | 1098 |
| | 10 | 95.29 | 0.88 | 0.931 | 0.933 | 0.932 | 0.929 | 1098 |
| GBM | 5 | 94.23 | 0.87 | 0.892 | 0.910 | 0.901 | 0.917 | 1098 |
| | 10 | 93.45 | 0.84 | 0.874 | 0.903 | 0.888 | 0.890 | 1098 |
| XGBoost | 5 | 94.81 | 0.85 | 0.895 | 0.901 | 0.898 | 0.894 | 1098 |
| | 10 | 93.31 | 0.83 | 0.834 | 0.912 | 0.871 | 0.886 | 1098 |
| SVM | 5 | 91.01 | 0.81 | 0.781 | 0.792 | 0.786 | 0.779 | 1098 |
| | 10 | 89.21 | 0.78 | 0.750 | 0.814 | 0.781 | 0.765 | 1098 |

higher proportion of coarser sand in the lateritic tract of the upper catchment makes the upper catchment susceptible in terms of erodibility. On the other hand, resistance and topographical factor cluster-based models have identified some similar areas to previous two factor cluster models. However, some new areas, even in the lower parts of the study unit, have been figured out. In the case of topographical factor cluster, the distribution of the susceptible areas is less continuous but covers a wider part of the study region. In the case of the resistance cluster, most of the forest patches are situated in the upper catchment, which protects to some extent. The lack of such considerable forest patches in the lower increases the exposure to gully erosion compared to the upper catchment. When testing data is overlapped with each factor cluster model to obtain the model's accuracy, matching accuracy is found between fairly good and good. All the statistical measures applied envisage the same truth. Among the used models, in the case of RF models, the accuracy level is found high, followed by XGBoost, GBM, and SVM. The accuracy value is found to be higher in the case of the erodibility factor cluster, followed by the erosivity factor cluster (Table 5).

The correlation coefficient between the factor cluster and the final model output of the respective applied models also reveals the same trend (Table 6). Correlation coefficients between erodibility and erosivity factor cluster model of final RF model of gully erosion susceptibility are respectively

0.759 and 0.776. Correlation values are quite less in the case of topographical and resistance factor clusters in reference to all the models. RF model-based correlation has revealed the stronger correlation followed by GBM, XGB, and SVM. All the values of the correlation coefficient are statistically significant at <0.01 level of significance. This again justifies that RF model output is more acceptable.

Sensitivity analysis

Sensitivity analysis at a spatial scale has been done, excluding the selected factors. Figure 6 shows the RF model output after excluding the chosen factors one by one. Change of accuracy level due to the exclusion of one factor reflects the importance of one factor. Similarly, spatial scale sensitivity modelling also helps to recognize exclusion of one factor brings significant changes in which parts of the study area, and it thus helps to identify the factors of regional importance. From Table 7, it is very evident that the departure value of all the accuracy statistics is high after the exclusion of geology factor followed by soil texture, erosivity, etc. (Table 7). These determinants are very decisive in bringing a significant change in gully erosion susceptibility in the unclassified granite and gneissic part and, therefore, could be considered regional factors of importance. BSI is an important determinant of the patches around the Masanjore dam.

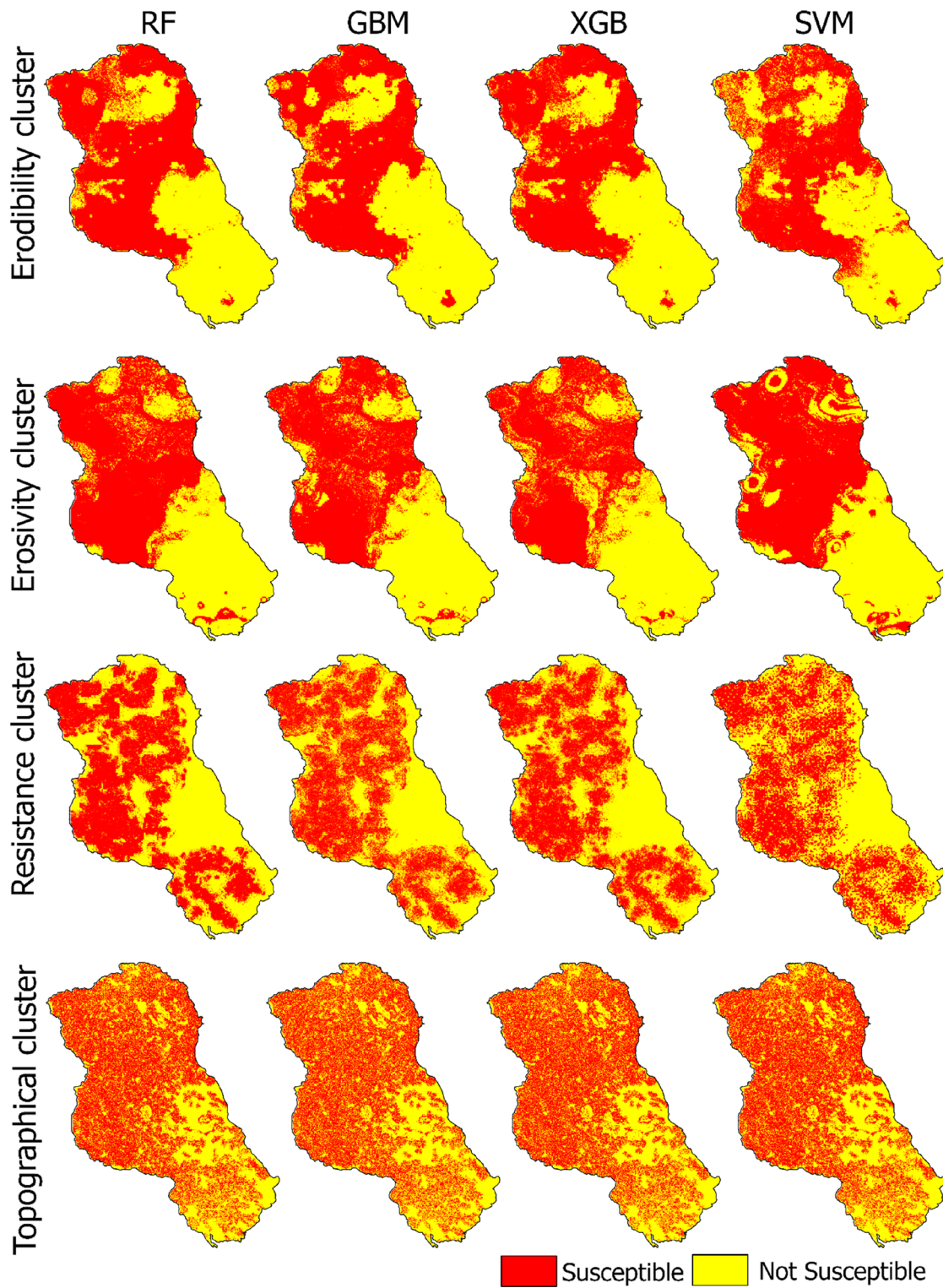


Fig. 5 Cluster specific gully erosion susceptible zones based on different ML models

Discussion

ML model output has demonstrated that 14–20% of the study area mainly in the upper catchment is highly

susceptible to gully erosion. All the applied models have good acceptability in reference to their accuracy level. However, the RF model is found to be the best representative. The erodibility factor cluster is the most determining

Table 5 Statistical accuracy of the factor cluster models

| Models | Clusters | Correctly classified (%) | ROC (AUC) | Precision | Sensitivity | F1-score | MCC | Support |
|--------|-----------------------|--------------------------|-----------|-----------|-------------|----------|-------|---------|
| RF | Erodibility cluster | 79.22 | 0.788 | 0.783 | 0.781 | 0.782 | 0.765 | 1098 |
| | Erosivity cluster | 79.10 | 0.709 | 0.701 | 0.700 | 0.700 | 0.698 | 1098 |
| | Resistance cluster | 78.96 | 0.784 | 0.771 | 0.771 | 0.771 | 0.765 | 1098 |
| | Topographical cluster | 76.89 | 0.756 | 0.751 | 0.748 | 0.749 | 0.734 | 1098 |
| GBM | Erodibility cluster | 72.63 | 0.716 | 0.711 | 0.706 | 0.708 | 0.691 | 1098 |
| | Erosivity cluster | 71.28 | 0.703 | 0.700 | 0.693 | 0.696 | 0.691 | 1098 |
| | Resistance cluster | 69.72 | 0.685 | 0.681 | 0.681 | 0.681 | 0.678 | 1098 |
| | Topographical cluster | 68.98 | 0.681 | 0.678 | 0.670 | 0.674 | 0.667 | 1098 |
| XGB | Erodibility cluster | 75.33 | 0.742 | 0.736 | 0.709 | 0.722 | 0.697 | 1098 |
| | Erosivity cluster | 75.11 | 0.738 | 0.725 | 0.692 | 0.708 | 0.680 | 1098 |
| | Resistance cluster | 73.83 | 0.721 | 0.703 | 0.680 | 0.691 | 0.665 | 1098 |
| | Topographical cluster | 72.31 | 0.709 | 0.697 | 0.678 | 0.687 | 0.659 | 1098 |
| SVM | Erodibility cluster | 75.18 | 0.745 | 0.731 | 0.727 | 0.729 | 0.721 | 1098 |
| | Erosivity cluster | 74.89 | 0.748 | 0.736 | 0.731 | 0.733 | 0.729 | 1098 |
| | Resistance cluster | 73.11 | 0.721 | 0.723 | 0.718 | 0.720 | 0.712 | 1098 |
| | Topographical cluster | 71.98 | 0.719 | 0.711 | 0.700 | 0.705 | 0.684 | 1098 |

Table 6 Correlation coefficient between factor cluster ML and final ML models

| Gully erosion models | Erodibility cluster | Erosivity cluster | Resistance cluster | Topographical cluster |
|----------------------|---------------------|-------------------|--------------------|-----------------------|
| RF | 0.759 | 0.776 | 0.603 | 0.620 |
| GBM | 0.749 | 0.762 | 0.529 | 0.544 |
| XGB | 0.741 | 0.746 | 0.587 | 0.610 |
| SVM | 0.691 | 0.724 | 0.411 | 0.381 |

factor cluster for measuring gully erosion susceptibility. Geological factor, soil texture, association of ferrous mineral, etc. are the major factors under this factor cluster with high importance towards gully erosion. Sensitivity analysis also has reported that geology, soil texture, and erosivity are the major contributing factors enhancing gully erosion susceptibility.

Multi-model ML approach has been taken for modelling gully erosion susceptibility with the aim to validate the models by themselves. This is the advantage of the multi-model approach. ML models based on training sites provide credible output in the case of prediction work (Benedetto et al. 2020; Xenochristou and Kapelan 2020). Benedetto et al. (2020), Xenochristou and Kapelan (2020), and Fayaz et al. (2020) have successfully applied RF, SVM, GBM, and XGB models and reported excellent credibility in predictive modelling (Pal and Paul 2021). In their works, always RF was not found as the best representative model. However, in many cases, it has been seen as the best representative. For example, Gianinetto

et al. (2020) and Chakraborty et al. (2020) for modelling soil erosion susceptibility, Al-Najjar and Pradhan (2021) for modelling landslide susceptibility, Islam et al. (2021) for modelling flood susceptibility, etc., the RF algorithm provides a built-in feature selection system that reduces dimensionality without removing any data. Such a function makes the chances of data loss very little, and the algorithm becomes relatively more reliable and error-free (Zhou et al. 2020). In addition, since the RF uses bagging, it is more sensitive to the noise and capable of controlling it than the boosting technique-based algorithms and resulting in greater prediction accuracy (Chan and Paelinckx 2008; Pal and Mather 2003).

This work can be compared to the previous similar studies (Avand et al. 2019; Saha et al. 2020; Pham et al. 2020; Amare et al. 2021) regarding model accuracy. The accuracy (ROC-AUC) of the previous studies was observed to range between 0.87 and 0.99. Among the applied algorithms, all these previous studies reported well accuracy of the RF algorithm. In the present study, the same algorithm has achieved quite an identical accuracy level. Since this work is entirely based on all pixels inclusive, generalization effect as found in case of point-based work has been minimized. Often, it is found that model building parameters are ranked based on their importance using statistical measures like information gain ratio (Costache et al. 2020; Bui et al. 2020). But, in reality, some less important model building parameters may play crucial role in determining gully erosion susceptibility. So, pixel scale sensitivity analysis can resolve such problems. Since the present study has taken this approach, factors of regional and local importance along with their

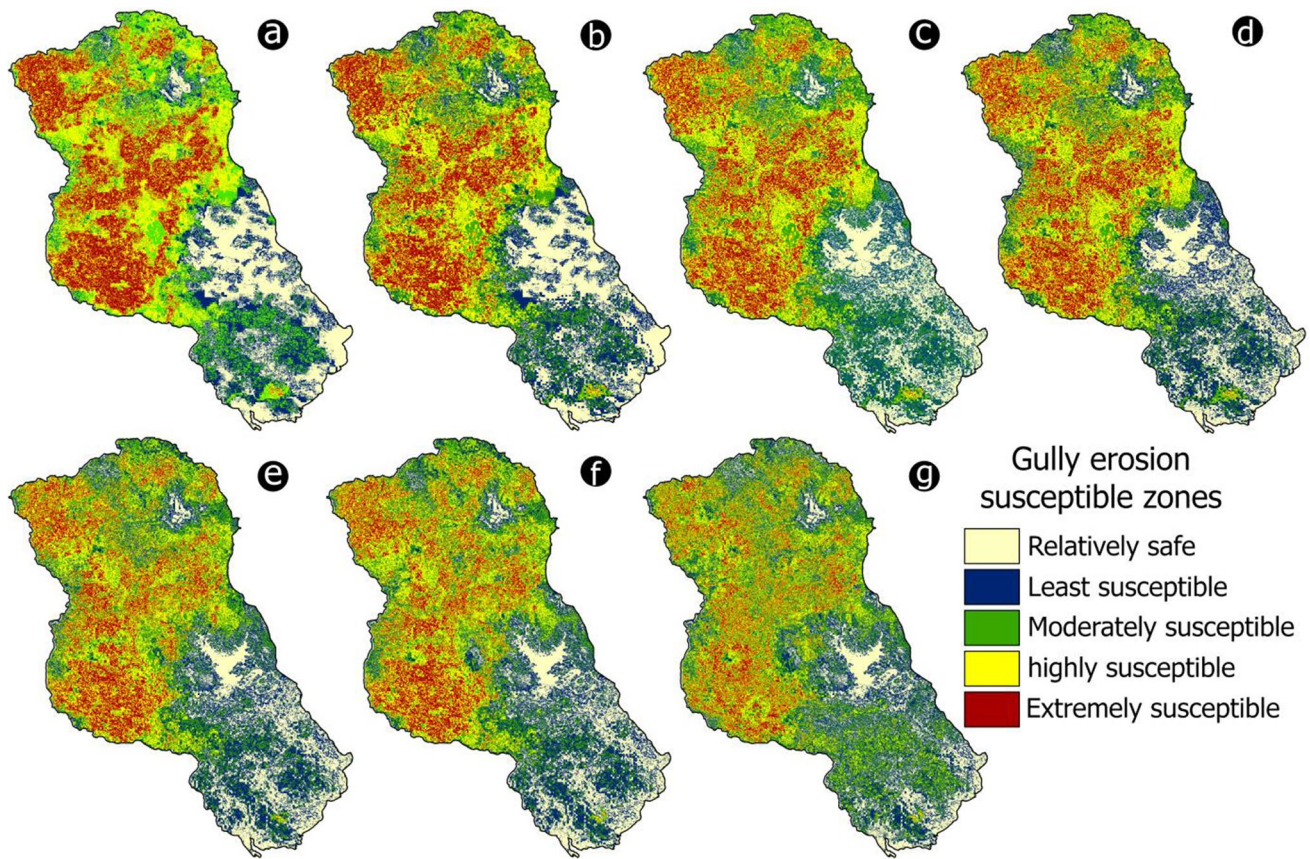


Fig. 6 Sensitivity models excluding the following factors one by one **a** excluding NDVI; **b** NDVI and slope; **c** NDVI, slope, and stream distance; **d** NDVI, slope, stream distance, and BSI; **e** NDVI, slope, stream distance, BSI, and soil erosivity; **f** NDVI, slope, stream dis-

tance, BSI, soil erosivity, and soil texture; **g** NDVI, slope, stream distance, BSI, soil erosivity, soil texture, and geology based on RF model

Table 7 Statistical test of the sensitivity models done in reference to RF model (departure values after excluding individual parameter)

| Classifiers | Parameters | K-fold | Correctly classified (%) | ROC (AUC) | Precision | Sensitivity | F1-score | MCC | Support |
|-------------|-----------------|--------|--------------------------|-----------|-----------|-------------|----------|-------|---------|
| RF | NDVI | 10 | 4.08 | 0.006 | 0.049 | 0.054 | 0.052 | 0.058 | 1098 |
| | Slope | | 13.20 | 0.077 | 0.120 | 0.133 | 0.127 | 0.128 | 1098 |
| | Stream distance | | 14.95 | 0.092 | 0.140 | 0.132 | 0.136 | 0.146 | 1098 |
| | BSI | | 20.16 | 0.149 | 0.209 | 0.198 | 0.204 | 0.218 | 1098 |
| | Soil erosivity | | 27.06 | 0.211 | 0.260 | 0.247 | 0.254 | 0.275 | 1098 |
| | Soil texture | | 31.86 | 0.267 | 0.331 | 0.321 | 0.326 | 0.339 | 1098 |
| | Geology | | 33.11 | 0.286 | 0.345 | 0.342 | 0.344 | 0.361 | 1098 |

average rank have been explored. Sensitivity analysis has the capability to do this. The present study has used this approach for spatial prediction of change which may occur if one factor is excluded from the analysis. Gully erosion susceptibility mapping using bivariate, multi-variate statistics, and ML models is very common, but sensitivity analysis at spatial scale is rarely found, but this has profound importance. Moreover, factor cluster-based modelling and its role

in the prediction process is also very important but paid less attention to. This approach can help recognize a set of factors as a factor cluster and its role in gully erosion susceptibility mapping.

Spatially figuring out the gully erosion susceptible areas with varying intensity and set of responsible factors based on sensitivity at spatial level can help to develop region specific plan of gully erosion check in order to arrest the valuable

soil resources. When sensitivity is computed only at numerical scale including entire study unit as a whole, it may not reflect the regionally sensitive factors. Without this, application of suitable planning will be in vain. Since the study clearly mapped the degree of susceptibility and concerned factors of it, this work has enough planning implication. Moreover, to cater the agriculturally dependent growing population in this region, rilling, gullying, and consequent soil loss have become a major barrier. The upper catchment of the study region majorly dominated with Chottanagpur plateau fringe secondary laterite soil with poor cohesion, fertility, and high erosivity. The ambience of gullying is very suitable in this region. Resilience of the problem is the major alternative to cohabit with the situation. Without proper planning, food security of the region would be in front of a big question. In this standpoint, findings of this work have its societal implication apart from its needs in natural resource conservation. The issue which is addressed is widely found across the world, but its background, causative factors may not be similar. Therefore, the regional findings are regionally more important, but the approach of study and sensitivity analysis for finding out factors of regional importance could be applied universally.

Conclusion

The study has explored 14–20% study area mainly in the upper catchment as the gully erosion susceptible using ensemble ML models. Unclassified granitic and the gneissic composed area are found to be highly susceptible to gully erosion. Factor cluster-based modelling has reported that the erodibility factor cluster is the best representative, followed by the erosivity factor cluster. Geology and soil texture have been found as the dominant contributing factors to gully erosion, predicted through sensitivity analysis. Among the applied models, the RF model is found as the best representative for predicting gully erosion susceptibility. Factor cluster-based modelling and spatial scale sensitivity analysis for identifying factors of regional and local importance are two innovative parts of this present work. Credible model output with a highly acceptable accuracy level encourages use of ensemble ML models. Sensitivity analysis has clearly imaged the factor of regional importance, and therefore, it is also recommended to use this approach while dealing with such or similar work.

However, this study does not focus on providing any quantification of gully erosion in this region which is a major limitation of this study. To overcome this limitation, few case studies can be conducted on different gully erosion susceptible areas. Since the present work has figured out the gully erosion susceptible area and identified dominant factors of regional and local importance, this study would

be instrumental to planners for devising local to regional level planning for gully erosion check and conserving soil erosion. Soil resource is a major stay of agriculture and this economic activity is the basis of economic fate of the majority of the people of this region. So, the loss of soil is almost synonymous with the loss of agricultural production as well as food security. From soil preservation and food security standpoints, the findings of the work would be very effective.

Acknowledgements For this study, the authors would like to convey their gratitude to USGS for providing Landsat imageries. They are also thankful to Indrajit Mandal for his assistance during field survey and software handling. No fund was received for this study.

Author contribution All the authors contributed to the study conception and design. Conceptualization, supervision, editing, and reviewing were performed by Swades Pal. Data curation, methodology designing, investigation, and validation were performed by Satyajit Paul. Formal analysis and writing of original draft were performed by Sandipta Debanshi. All the authors read and approved the final manuscript.

Data availability The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Competing interests The authors declare no competing interests.

References

- Abdi AM (2020) Land cover and land use classification performance of machine learning algorithms in a boreal landscape using Sentinel-2 data. *GIScience & Remote Sensing* 57(1):1–20. <https://doi.org/10.1080/15481603.2019.1650447>
- Acortes C, Vapnik V (1995) Support vector networks. *Machine Learning* 20(1):273–297
- Al-Najjar HH, Pradhan B (2021) Spatial landslide susceptibility assessment using machine learning techniques assisted by additional data created with generative adversarial networks. *Geosci Front* 12(2):625–637
- Amare S, Langendoen E, Keesstra S, Ploeg MVD, Gelagay H, Lemma H, van der Zee SE (2021) Susceptibility to gully erosion: applying random forest (RF) and frequency ratio (FR) approaches to a small catchment in Ethiopia. *Water* 13(2):216
- Arabameri A, Pradhan B, Pourghasemi HR, Rezaei K, Kerle N (2018) Spatial modelling of gully erosion using GIS and R programming: a comparison among three data mining algorithms. *Appl Sci* 8(8):1369
- Arabameri A, Chen W, Blaschke T, Tiefenbacher JP, Pradhan B, Tien Bui D (2020) Gully head-cut distribution modeling using machine learning methods—a case study of nwiran. *Water* 12(1):16

- Arabameri A, Chen W, Loche M, Zhao X, Li Y, Lombardo L, Bui DT (2020) Comparison of machine learning models for gully erosion susceptibility mapping. *Geosci Front* 11(5):1609–1620
- Araujo TP, Pejon OJ (2015) Topographic threshold to trigger gully erosion in a Tropical region—Brazil. In *Engineering Geology for Society and Territory* 3(627):630 (**Springer, Cham**)
- Arora, A., Pandey, M., Siddiqui, M. A., Hong, H., & Mishra, V. N. (2019). Spatial flood susceptibility prediction in Middle Ganga Plain: comparison of frequency ratio and Shannon's entropy models. *Geocarto International*, 1–32.
- Avand M, Janizadeh S, Naghibi SA, Pourghasemi HR, Khosrobeigi Bozchaloei S, Blaschke T (2019) A comparative assessment of random forest and k-nearest neighbor classifiers for gully erosion susceptibility mapping. *Water* 11(10):2076
- Azedou A, Lahssini S, Khattabi A, Meliho M, Rifai N (2021) A methodological comparison of three models for gully erosion susceptibility mapping in the rural municipality of El Faid (Morocco). *Sustainability* 13(2):682
- Bandfield, J. L., Ghent, R. R., Vasavada, A. R., Paige, D. A., Lawrence, S. J., & Robinson, M. S. (2011). Lunar surface rock abundance and regolith fines temperatures derived from LRO Diviner Radiometer data. *Journal of Geophysical Research: Planets*, 116(E12).
- Benedetto, U., Dimagli, A., Sinha, S., Cocomello, L., Gibbison, B., Caputo, M., & Angelini, G. D. (2020). Machine learning improves mortality risk prediction after cardiac surgery: systematic review and meta-analysis. *The Journal of Thoracic and Cardiovascular Surgery*.
- Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2).
- Brownlee J (2019) Machine learning mastery with Weka. Ebook Edition 1:4
- Bui DT, Hoang ND, Martínez-Álvarez F, Ngo PTT, Hoa PV, Pham TD, Costache R (2020) A novel deep learning neural network approach for predicting flash flood susceptibility: a case study at a high frequency tropical storm area. *Sci Total Environ* 701:134413
- Busch R, Hardt J, Nir N, Schütt B (2021) Modeling gully erosion susceptibility to evaluate human impact on a local landscape system in Tigray. *Ethiopia Remote Sensing* 13(10):2009
- Cánovas 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
- Chakraborty R, Pal SC, Sahana M, Mondal A, Dou J, Pham BT, Yunus AP (2020) Soil erosion potential hotspot zone identification using machine learning and statistical approaches in eastern India. *Nat Hazards* 104(2):1259–1294
- 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
- Chen, T. and Guestrin, C., 2016, August. Xgboost: a scalable tree boosting system. In *Proceedings of the 22nd acmsigkdd international conference on knowledge discovery and data mining* (pp. 785–794).
- Cheng G, Dong C, Huang G, Baetz BW, Han J (2016) Discrete principal-monotonicity inference for hydro-system analysis under irregular nonlinearities, data uncertainties, and multivariate dependencies Part i: Methodology Development. *Hydrolo Process* 30(23):4255–4272
- Cheng G et al (2017) Climate classification through recursive multivariate statistical inferences: a case study of the Athabasca River Basin, Canada. *International Journal of Climatology* 37:1001–1012
- 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 Mining* 14(1):1–22
- Conforti M, Auceili 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
- Costache R, Pham QB, Sharifi E, Linh NTT, Abba SI, Vojtek M, Khoi DN (2020) Flash-flood susceptibility assessment using multi-criteria decision making and machine learning supported by remote sensing and GIS techniques. *Remote Sensing* 12(1):106
- Daviran M, Maghsoudi A, Ghezelbash R, Pradhan B (2021) A new strategy for spatial predictive mapping of mineral prospectivity: automated hyperparameter tuning of random forest approach. *Comput Geosci* 148:104688
- Debanshi S, Pal S (2020) Assessing gully erosion susceptibility in Mayurakshi river basin of eastern India. *Environ Dev Sustain* 22(2):883–914
- Dineva, A., Várkonyi-Kóczy, A. R., & Tar, J. K. (2014). Fuzzy expert system for automatic wavelet shrinkage procedure selection for noise suppression. In *IEEE 18th International Conference on Intelligent Engineering Systems INES 2014* (pp. 163–168). IEEE.
- Du P, Bai X, Tan K, Xue Z, Samat A, Xia J, Liu W (2020) Advances of four machine learning methods for spatial data handling: a review. *Journal of Geovisualization and Spatial Analysis* 4:1–25
- Dutta S (2016) Soil erosion, sediment yield and sedimentation of reservoir: a review. *Modeling Earth Syst Environ* 2(3):1–18
- 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. *ISPRS Int J Geo Inf* 6(11):378
- Evelpidou, N., Kampolis, I., & Karkani, A. (2018). Geomorphic features associated with erosion. In *Natural Hazards* (pp. 205–232). CRC Press.
- Fauzan, M.A. and Murfi, H., 2018. The accuracy of XGBoost for insurance claim prediction. *Int. J. Adv. Soft Comput. Appl*, 10(2).
- Fayaz, M., Khan, A., Rahman, J. U., Alharbi, A., Uddin, M. I., & Alouffi, B. (2020). Ensemble machine learning model for classification of spam product reviews. *Complexity*, 2020.
- Gelagay HS, Minale AS (2016) Soil loss estimation using GIS and remote sensing techniques: a case of Koga watershed, Northwestern Ethiopia. *Int Soil and Water Conserv Res* 4(2):126–136
- Ghosh S, Guchhait SK (2015) Characterization and evolution of laterites in West Bengal: implication on the geology of northwest Bengal Basin. *Transactions* 37(1):93–119
- Ghosh S, Guchhait SK, Xiu-Fang Hu (2015) Characterization and evolution of primary and secondary laterites in northwestern Bengal Basin, West Bengal, India. *J Palaeogeogr* 4(2):203–230
- Gianinetto M, Aiello M, Vezzoli R, Polinelli FN, Rulli MC, Chiarelli DD, Soncini A (2020) Future scenarios of soil erosion in the Alps under climate change and land cover transformations simulated with automatic machine learning. *Climate* 8(2):28
- Gui K, Che H, Zeng Z, Wang Y, Zhai S, Wang Z, Luo M, Zhang L, Liao T, Li H, Zhao L (2020) Construction of a virtual PM2.5 observation network in China based on high-density surface meteorological observations using the extreme gradient boosting model. *Environ Int* 141:105801
- Gulzat T, Lyazat N, Siladi V, Gulbakyt S, Maksatbek S (2020) Research on predictive model based on classification with parameters of optimization. *Neural Network World* 30(5):295
- Han X, Lv P, Zhao S, Sun Y, Yan S, Wang M, 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

- Handoko J, Hendryli DE, Herwindiati J (2020) November. Gradient boosting tree for land use change detection using Landsat 7 and 8 imageries: a case study of Bogor area as water buffer zone of Jakarta. In IOP Conf Series: Earth and Environ Sci 581(1):012045
- Harimoorthy K, Thangavelu M (2021) Multi-disease prediction model using improved SVM-radial bias technique in healthcare monitoring system. *J Ambient Intell Humaniz Comput* 12(3):3715–3723
- Hoester 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 Sensing* 12(10):1667
- Hoester T, Bachofer F, Kuenzer C (2020) Object detection and image segmentation with deep learning on Earth observation data: a review—Part II: Applications. *Remote Sensing* 12(18):3053
- Hong H, Jaafari A, Zenner EK (2019) Predicting spatial patterns of wildfire susceptibility in the Huichang County, China: an integrated model to analysis of landscape indicators. *Ecol Ind* 101:878–891
- Islam ARMT, Talukdar S, Mahato S, Kundu S, Eibek KU, Pham QB, Linh NTT (2021) Flood susceptibility modelling using advanced ensemble machine learning models. *Geosci Front* 12(3):101075
- Jaafari A, Gholami DM, Zenner EK (2017) A Bayesian modeling of wildfire probability in the Zagros Mountains. *Iran Ecological Informatics* 39:32–44
- 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
- Jha VC, Kapat S (2003) Gully erosion and its implications on land use, a case study. *Land degradation and desertification*. Publ, Jaipur and New Delhi, pp 156–178
- Jha VC, Kapat S (2009) Rill and gully erosion risk of lateritic terrain in South-Western Birbhum District, West Bengal. *India Sociedade & Natureza* 21(2):141–158
- Joshi V, Suswara N, Sinha D (2016) Estimating soil loss from a watershed in Western Deccan, India, using revised universal soil loss equation. *Landscape & Environment* 10(1):13–25
- Jun, M.J., 2021. A comparison of a gradient boosting decision tree, random forests, and artificial neural networks to model urban land use changes: the case of the Seoul metropolitan area. *International Journal of Geographical Information Science*, pp.1–19. <https://doi.org/10.1080/13658816.2021.1887490>
- Kalantar B, Ueda N, Idrees MO, Janizadeh S, Ahmadi K, Shabani F (2020) Forest fire susceptibility prediction based on machine learning models with resampling algorithms on remote sensing data. *Remote Sensing* 12(22):3682
- Kaur S, Aggarwal H, Rani R (2020) Hyper-parameter optimization of deep learning model for prediction of Parkinson's disease. *Mach vis Appl* 31(5):1–15
- Kawaguchi, K., Kaelbling, L. P., & Bengio, Y. (2017). Generalization in deep learning. *arXiv preprint arXiv:1710.05468*.
- Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu TY (2017) Lightgbm: a highly efficient gradient boosting decision tree. *Adv Neural Inf Process Syst* 30:3146–3154
- Kertész Á, Gergely J (2011) Gully erosion in Hungary, review and case study. *Procedia Soc Behav Sci* 19:693–701
- Kim S, Matsumi Y, Pan S, Mase H (2016) A real-time forecast model using artificial neural network for after-runner storm surges on the Tottori coast, Japan. *Ocean Eng* 122:44–53
- Kotthoff, L., Thornton, C., Hoos, H. H., Hutter, F., & Leyton-Brown, K. (2019). Auto-WEKA: automatic model selection and hyper-parameter optimization in WEKA. In *Automated Machine Learning* (pp. 81–95). Springer, Cham.
- Li, R., Cui, L., Fu, H., Meng, Y., Li, J. and Guo, J., 2020. Estimating high-resolution PM1 concentration from Himawari-8 combining extreme gradient boosting-geographically and temporally weighted regression (XGBoost-GTWR). <https://doi.org/10.1016/j.atmosenv.2020.117434>
- Lin JM, Billa L (2021) Spatial prediction of flood-prone areas using geographically weighted regression. *Environmental Advances* 6:100118
- Liu K, Chen W, Lin H (2020) XG-PseU: an eXtreme gradient boosting based method for identifying pseudouridine sites. *Mol Genet Genomics* 295(1):13–21. <https://doi.org/10.1007/s00438-019-01600-9>
- Maiti A, Zhang Q, Sannigrahi S, Pramanik S, Chakraborti S, Cerda A, Pilla F (2021) Exploring spatiotemporal effects of the driving factors on COVID-19 incidences in the contiguous United States. *Sustain Cities Soc* 68:102784
- 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, 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
- Maxwell AE, Bester MS, Guillen LA, Ramezan CA, Carpinello DJ, Fan Y, Pyron JL (2020) Semantic segmentation deep learning for extracting surface mine extents from historic topographic maps. *Remote Sensing* 12(24):4145
- Maxwell AE, Pourmohammadi P, Poyner JD (2020) Mapping the topographic features of mining-related valley fills using mask R-CNN deep learning and digital elevation data. *Remote Sensing* 12(3):547
- 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* 28(9):11637–11649
- Mitranont, J., Sawangphol, W., Vithantirawat, T., Paengkaew, S., Suwannasing, P., Daramas, A., & Chen, Y. C. (2017, November). A study on using Python vs Weka on dialysis data analysis. In *2017 2nd International Conference on Information Technology (INCIT)* (pp. 1–6). IEEE.
- Mosavi A, Ozturk P, Chau KW (2018) Flood prediction using machine learning models: literature review. *Water* 10(11):1536
- Mosavi, A., Rabczuk, T., & Varkonyi-Koczy, A. R. (2017, September). Reviewing the novel machine learning tools for materials design. In *International Conference on Global Research and Education* (pp. 50–58). Springer, Cham.
- Naghibi SA, Hashemi H, Berndtsson R, Lee S (2020) Application of extreme gradient boosting and parallel random forest algorithms for assessing groundwater spring potential using DEM-derived factors. *J Hydrol* 589:125197. <https://doi.org/10.1016/j.jhydrol.2020.125197>
- Neyshabur, B., Bhojanapalli, S., McAllester, D., & Srebro, N. (2017). Exploring generalization in deep learning. *arXiv preprint arXiv:1706.08947*.
- 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
- Ortiz-García EG, Salcedo-Sanz S, Casanova-Mateo C (2014) Accurate precipitation prediction with support vector classifiers: a study including novel predictive variables and observational data. *Atmos Res* 139:128–136
- Osman, A.I.A., Ahmed, A.N., Chow, M.F., Huang, Y.F. and El-Shafie, A., 2021. Extreme gradient boosting (Xgboost) model to predict

- the groundwater levels in Selangor Malaysia. *Ain Shams Engineering Journal* <https://doi.org/10.1016/j.asej.2020.11.011>
- 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 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, Paul S (2021) Linking hydrological security and landscape insecurity in the moribund deltaic wetland of India using tree-based hybrid ensemble method in python. *Eco Inform* 65:101422
- Pham QB, Mukherjee K, Norouzi A, Linh NTT, Janizadeh S, Ahmadi K, Anh DT (2020) Head-cut gully erosion susceptibility modelling based on ensemble Random Forest with oblique decision trees in Fareghan watershed. *Iran Geomatics, Natural Hazards and Risk* 11(1):2385–2410
- 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 Sensing* 13(22):4521
- 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
- Roy, P., Chakraborty, R., Chowdhuri, I., Malik, S., Das, B., & Pal, S. C. (2020). Development of different machine learning ensemble classifier for gully erosion susceptibility in Gandheswari Watershed of West Bengal, India. *Machine learning for intelligent decision science*, 1–26.
- Roy J, Saha S (2021) Integration of artificial intelligence with meta classifiers for the gully erosion susceptibility assessment in Hinglo river basin. *Eastern India Advances in Space Research* 67(1):316–333
- 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
- Sahin EK (2020) Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGBoost, gradient boosting machine, and random forest. *SN Applied Sciences* 2(7):1–17. <https://doi.org/10.1007/s42452-020-3060-1>
- Sarkar T, Mishra M, Chatterjee S (2020) On detailed field-based observations of laterite and laterization: a study in the Paschim Medinipur lateritic upland of India. *J Sediment Environ* 5(2):219–245
- Schonlau M, Zou RY (2020) The random forest algorithm for statistical learning. *Stand Genomic Sci* 20(1):3–29
- Shit, P. K., Bhunia, G. S., & Pourghasemi, H. R. (2020). Gully erosion susceptibility mapping based on bayesian weight of evidence. In *Gully Erosion Studies from India and Surrounding Regions* (pp. 133–146). Springer, Cham.
- Sidorchuk A (2021) Models of gully erosion by water. *Water* 13(22):3293
- Syam, N., & Kaul, R. (2021). Random forest, bagging, and boosting of decision trees. In *Machine Learning and Artificial Intelligence in Marketing and Sales*. Emerald Publishing Limited.
- Taherei Ghazvinei P, Hassanpour Darvishi H, Mosavi A, Yusof KBW, Alizamir M, Shamshirband S, Chau KW (2018) Sugarcane growth prediction based on meteorological parameters using extreme learning machine and artificial neural network. *Engineering Applications of Computational Fluid Mechanics* 12(1):738–749
- 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, Dagne DC, Zegeye AD, Tebebu TY, 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
- USDA. (1999). Natural resources conservation service, soil taxonomy a basic system of soil classification for making and interpreting soil surveys, second edition.
- Xenochristou M, Kapelan Z (2020) An ensemble stacked model with bias correction for improved water demand forecasting. *Urban Water Journal* 17(3):212–223
- Yang A, Wang C, Pang G, Long Y, Wang L, Cruse RM, Yang Q (2021) Gully erosion susceptibility mapping in highly complex terrain using machine learning models. *ISPRS Int J Geo Inf* 10(10):680
- Yang, Y., Chung, H. and Kim, J.S., 2021b. Local or neighborhood? Examining the relationship between traffic accidents and land use using a gradient boosting machine learning method: the case of Suzhou Industrial Park, China. *Journal of Advanced Transportation*, 2021b.
- Zhang T, He W, Zheng H, Cui Y, Song H, Fu S (2021) Satellite-based ground PM2.5 estimation using a gradient boosting decision tree. *Chemosphere* 268:128801
- 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, Zhao Y (2021) Primary environmental factors controlling gully distribution at the local and regional scale: an example from Northeastern China. *International Soil and Water Conservation Research* 9(1):58–68

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.