Available Online at http://www.recentscientific.com

CODEN: IJRSFP (USA)

International Journal of Recent Scientific Research
Vol. 15, Issue, 05, pp.4704-4713, May, 2024

ISSN: 0976-3031

International Journal of Recent Scientific Research

DOI: 10.24327/IJRSR

Research Article

PREDICTION OF GULLY EROSION SUSCEPTIBILITY MAPPING USING XGBOOSTING MACHINE LEARNING ALGORITHM

Md Hasanuzzaman[1,2] and Pravat Kumar Shit[2]*

[1]Research Centre in Natural and Applied Science, Raja N. L. Khan Women’s College (Autonomous),
Vidyasagar University, Midnapore-721102, West Bengal, India
[2]PG Department of Geography, Raja N. L. Khan Women’s College (Autonomous), Gope Palace,
Midnapore-721102, West Bengal, India

DOI: http://dx.doi.org/10.24327/ijrsr.20241505.0879

ARTICLE INFO

Article History:

Received 9nd April, 2024
Received in revised form 24th April, 2024
Accepted 16th May, 2024
Published online 28th May, 2024

Keywords:

Gully erosion; Geo-environmental factors; Machine learning algorithm; Receiver Operating Characteristic (ROC); Sita Nala small watershed.

ABSTRACT

Gully erosion presents a significant threat to the environment, putting agriculture, wildlife habitats, human safety, infrastructure, and soil health at risk. Mapping areas vulnerable to gully erosion accurately demands selecting the right machine learning model, given the varied environmental factors influencing gully formation. In this study, we utilized machine learning algorithms based on extreme gradient boosting (XGB) to craft a highly precise gully erosion susceptibility map (GESM) for the Sita Nala small watershed, a tributary located on the right bank of the Subarnarekha River in West Bengal, India. Our investigation involved an in-depth analysis of gully erosion mapping with twenty-four variables and scrutiny of a dataset comprising 200 sample points, equally representing gullies and non-gullies. To assess multicollinearity, we utilized Information Gain Ratio (IGR) and Variance Inflation Factors (VIF) tests. The results revealed that drainage density (0.77), elevation (0.74), geomorphology (0.72), Land Use/Land Cover (LULC) (0.72), and Normalized Difference Vegetation Index (NDVI) (0.68) are the most critical factors influencing GESM. Employing a quantile classification approach, we generated three distinct categories of GESMs, ranging from areas with no gully erosion to those with moderate gully susceptibility area and high gully susceptibility area. Approximately 13.49% of the basin area was identified as being dominated by gully erosion, highlighting the urgent need for targeted management strategies in these regions. We evaluated the XGB model's performance on both training and testing data using various statistical tests, including Root Mean Square Error (RMSE), Kappa index, Mean Absolute Error (MAE), Accuracy (ACC), Receiver Operating Characteristic (ROC), and R^2. While both models produced satisfactory results, the XGB model exhibited strong performance, achieving an ROC value of 84.2%. However, the present study underscores that machine learning can accurately identify areas prone to gully erosion, providing valuable insights for policymakers to implement sustainable management practices.

Copyright© The author(s) 2024,This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited.

INTRODUCTION

Gully erosion stands as a pressing environmental concern worldwide, affecting the functionality of various soil and land systems. Its presence serves as tangible evidence of historical instances of intense soil erosion, reflecting shifts in the environment influenced by interactions among geomorphological features, changes in land use, and extreme weather events [1]. Despite their relatively small footprint within catchment areas, gully channels significantly contribute to sediment discharge, exacerbating runoff and sediment connectivity across landscapes [2]. This increased connectivity amplifies the risk of flooding and sediment buildup in reservoirs, emphasizing the crucial necessity for environmental scientists, land managers, and policymakers to comprehend the intricate relationship between environmental change and land degradation [3].

The magnitude of gully erosion's impact is notable globally, with its accounting for a considerable 55% of land degradation worldwide, affecting approximately two billion hectares of land [4]. The repercussions extend to soil depletion, habitat destruction, water contamination, sediment accumulation in water bodies, and heightened vulnerability to flooding. Additionally, gully erosion has detrimental effects on agricultural productivity, infrastructure integrity, and public safety [5].

Corresponding author:* **Pravat Kumar Shit
PG Department of Geography, Raja N. L. Khan Women’s College (Autonomous), Gope Palace,
Midnapore-721102, West Bengal, India

Changes in land use land cover (LULC) directly and indirectly influence the extent of gully erosion [6]. In India, the problem of water-induced rill and gully erosion is particularly severe, with an alarming annual soil loss rate of 16.4 tons per hectare, resulting in the loss of an estimated five billion metric tons of soil annually [7]. Given these implications, accurately mapping susceptibility to gully erosion is crucial for effectively mitigating its adverse effects [8].

Various statistical models have been utilized for mapping gully erosion susceptibility, including machine learning, multi-criteria decision-making with Analytical Hierarchy Process (AHP), and bivariate & multivariate statistical models [1; 6; 9; 10; 11]. Machine learning techniques have significantly advanced gully erosion prediction, outperforming traditional methods by discerning intricate patterns, analyzing vast datasets, uncovering hidden correlations, and mitigating human bias [12]. These algorithms continually enhance accuracy, particularly adept at identifying complex changes and unforeseen scenarios, even in data-limited contexts [13]. Moreover, machine learning-based models excel in assessing the impact of climate change-induced runoff on gully erosion compared to alternative techniques [14]. Selecting an appropriate machine learning model is crucial in developing an accurate gully erosion susceptibility mapping (GESM), given the variability in model performance across different environmental risks [15].

However, the present study, we utilized the machine learning model based extreme gradient boosting (XGB) algorithm for assessing the gully erosion susceptibility mapping of Sita Nala small-watershed area in the part of Chhota Nagpur plateau fringe region, India. The machine learning model based XGB model have emerged as powerful tools for predicting gully erosion and its effectiveness in addressing slope-related geo-environmental hazards. The XGB model's scalability and efficiency make it indispensable for memory-restricted environments, while its ability to manage sparse data and vast datasets ensures precise estimates.

MATERIALS AND METHODS

Study area

The present study focused on the Sita Nala small-watershed, a tributary situated on the right bank of the Subarnarekha River in West Bengal, India, spanning from 22° 05´ 24.54´´ N to 22° 09´ 13.47´´ N latitude and 87° 01´ 40.33´´ E to 87° 01´ 53.06´´ E longitude with an area of 38.606 km² (Fig. 1). The study area, located in the Chhotanagpur plateau fringe region, is characterized by significant gully erosion due to its undulating terrain and varied landforms [2]. With a tropical monsoon climate, the region experiences concentrated rainfall from June to September, averaging 1500 mm annually, with over 80% occurring during the monsoon season, leading to intense downpours and significant surface runoff [2]. The surface runoff coefficient ranges from 0.4 to 0.7 due to steep terrain, shallow soils, and limited vegetation cover, accelerating soil erosion and surface water runoff during monsoon periods [16]. Human activities such as deforestation and improper agricultural practices exacerbate gully erosion [2], posing significant threats to soil fertility, agriculture, water quality, and environmental stability. Topsoil loss contributes to reduced land productivity and downstream sedimentation, impacting aquatic ecosystems. Efforts to manage this issue include soil conservation, sustainable land use, and community awareness initiatives aimed at mitigating gully erosion's effects and ensuring ecological balance [2].

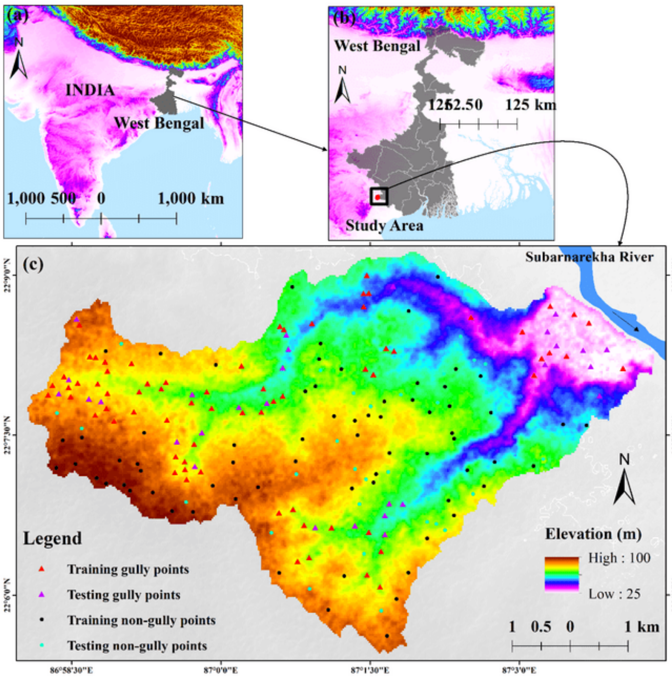

Fig.1 Location of the study area of Sita Nala small-watershed, a tributary situated on the right bank of the Subarnarekha River in West Bengal, India

Data based

As evidenced by previous research, the interplay of various geo-environmental factors such as topographical and hydrological characteristics, soil properties, land use/land cover (LULC), and a human activity contributes to gully formation in susceptible areas [17]. Drawing from a literature review of relevant studies [17; 18; 19], as well as considering the physiographic characteristics of the study area, multicollinearity tests, data availability, and research scale, this study identifies 24 conditioning factors influencing gully erosion (Figure 2).

Inventory of gully erosion locations

In the initial stages of developing gully erosion susceptibility models (GESMs), an essential step involves the creation of a gully inventory map [5]. To achieve this, we employed Digital Elevation Model (DEM) data with a resolution of 12.5 meters and utilized Google Earth Pro software for delineating gullies and generating the inventory maps. Prior to model execution, the accuracy of the gully inventory map was validated through field surveys conducted between October and December 2023, using a Garmin GPS etrex10 device for ground verification. A total of 100 gullies were randomly selected within the study area, with depths ranging from 1.98 meters to 8.17 meters and lengths varying between 0.033 and 0.413 km. These gullies were digitized as polygons along with their converted point features for subsequent model integration. Additionally, to meet model prerequisites, 100 non-gully locations were randomly chosen for comparison [8]. Thus, the study incorporated 200 random sample points, comprising 100 non-gully points and 100 gully points. Following the assignment of binary values, with gully points labeled as 1 and non-gully points as 0, the dataset was divided into training and testing subsets [20] (Fig. 1). Of the total sample set, 30% (60 points) were allocated for the testing dataset, while the remaining 70% (140 points) were utilized for model training. Finally, the study concluded by processing the datasets and executing the models using ArcGIS and R software (version 4.2.0).

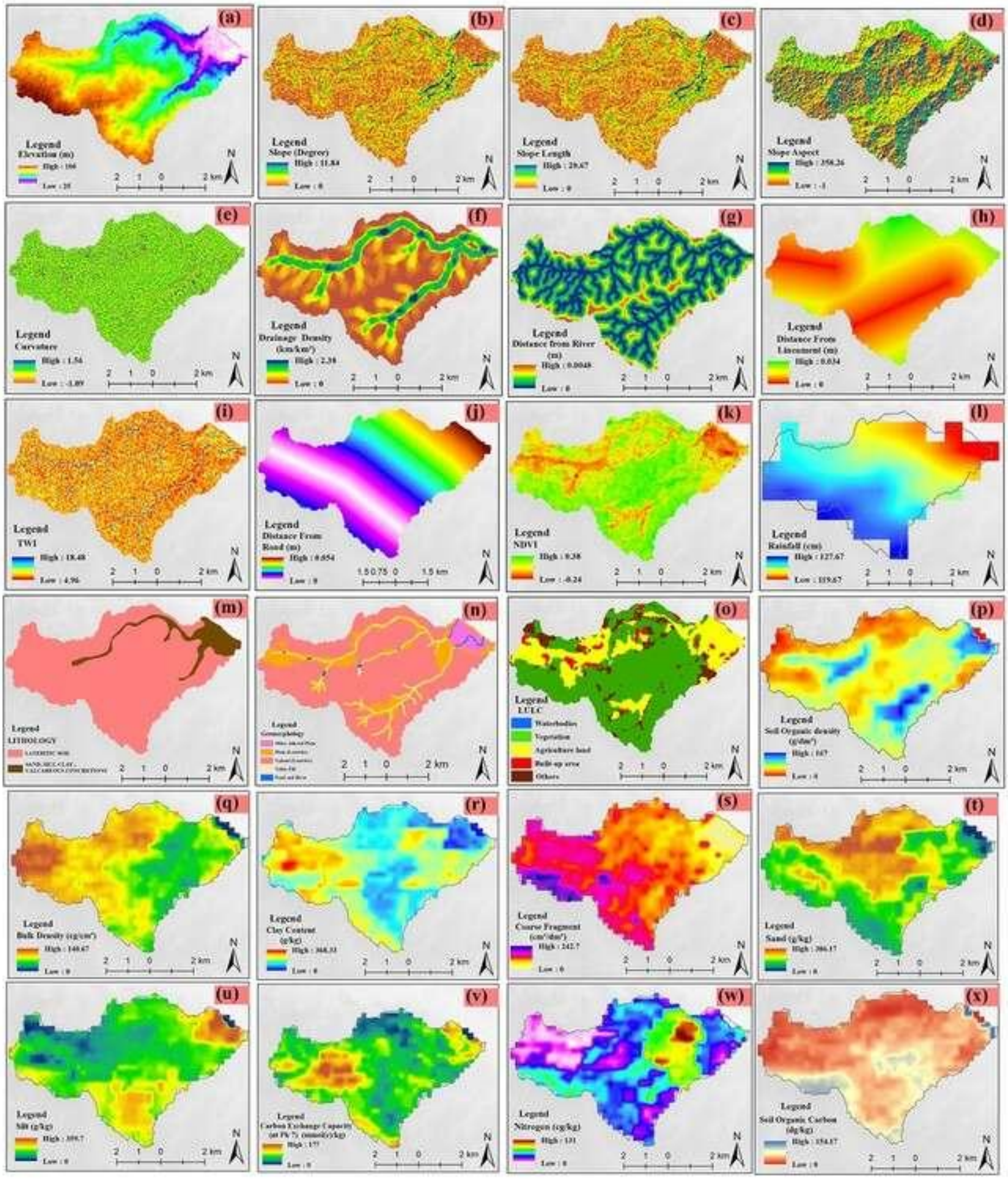

Fig. 2 Twenty four variables used in this study (a) elevation, (b) slope, (c) slope length, (d) slope aspect, (e) curvature, (f) drainage density, (g) distance from river, (h) distance from lineament, (i) TWI (j) distance to road, (k) NDVI, (l) rainfall, (m) lithology, (n) geomorphology, (o) LULC, (p) soil organic density, (q) bulk density, (r) clay content, (s) course fragment, (t) sand, (u) silt, (v) carbon exchange capacity, (w) Nitrogen, and (x) soil organic carbon.

Gully erosion conditioning factors

The formation of gullies is influenced by a multitude of environmental factors, and accurately identifying erosion-prone areas depends on carefully selecting the relevant contributing factors. In this study, 24 parameters governing gully erosion were identified based on available data, extensive field surveys, and previous research [8; 15; 20; 21; 22] to compute the Gully Erosion Susceptibility Models (GESMs) (Table 1). The factors chosen for this investigation include elevation, slope, slope length, slope aspect, curvature, drainage density, distance from river, distance from lineament, topographic wetness index (TWI), distance to road, normalized difference vegetation index (NDVI), rainfall, lithology, geomorphology, LULC, soil organic density, bulk density, clay content, course fragment, sand, silt, carbon exchange capacity, Nitrogen, and soil organic carbon. These relevant factors are illustrated in Figure 2 and detailed in Table 1.

Table 1 The current study details the data utilized, its respective sources, and the spatial resolution employed

Sl. No.	Category	Data source	Resolution
1.	Alos Palsar Dem (Elevation)	https://search.asf.alaska.edu	12.5 × 12.5 m
2.	Slope	Extracted from DEM	12.5 × 12.5 m
3.	Slope length	Extracted from DEM	12.5 × 12.5 m
4.	Slope aspect	Extracted from DEM	12.5 × 12.5 m
5.	Curvature	Extracted from DEM	12.5 × 12.5 m
6.	Drainage density (DD)	Extracted from DEM	12.5 × 12.5 m
7.	Distance from the river (DFR)	Extracted from DEM	12.5 × 12.5 m
8.	Distance from the lineament (DFL)	Extracted from DEM	12.5 × 12.5 m
9.	Topographic Weightiness Index (TWI)	Extracted from DEM	12.5 × 12.5 m
10.	Rainfall	WorldClim website	885.67×885.67 m
11.	NDVI	Satellite image (USGS website)	30 × 30 m
12.	Land Use and Land Cover (LULC)	Satellite image (USGS website)	30 × 30 m
13.	Distance from the road (DR)	(https://www.openstreetmap.org	30 × 30 m
14.	Lithologic	Survey of India (bhukosh.gsi.gov.in)	30 × 30 m
15.	Geomorphology	Survey of India (bhukosh.gsi.gov.in)	30 × 30 m
16.	Soil organic density (SOD)	https://soilgrids.org	250×250 m
17.	Bulk density	https://soilgrids.org	250×250 m
18.	Clay Content in Soil (SC)	https://soilgrids.org	250×250 m
19.	Coarse fragments	https://soilgrids.org	250×250 m
20.	Sand	https://soilgrids.org	250×250 m
21.	Silt	https://soilgrids.org	250×250 m
22.	Carbon exchange capacity (CEC)	https://soilgrids.org	250×250 m
23.	Nitrogen	https://soilgrids.org	250×250 m
24.	Soil organic carbon (SOC)	https://soilgrids.org	250×250 m

Multicollinearity assessment

To ensure the robustness of our model, we conducted tests to detect and mitigate multicollinearity among the 24 selected gully erosion factors. Multicollinearity arises when factors exhibit high correlation, potentially leading to inaccuracies in modeling [5. We employed two methodologies: Information Gain Ratio (IGR) and Variance Inflation Factors (VIF). VIF values exceeding 10 or falling below 0.1 indicate multicollinearity issues [23]. Conversely, the IGR method assesses the relative importance of each factor in predicting gully formation likelihood. A higher IGR value, determined by Average Merit (AM), indicates greater significance [23]. By utilizing a combination of VIF, Pearson's correlation coefficients, tolerance criteria, and IGR analysis, we diagnosed and mitigated multicollinearity, ensuring the appropriateness of the chosen factors for our gully erosion model.

*Extreme gradient boosting (*XGB*)*

The XGB algorithm, introduced by [24], was chosen for this study due to its status as a cutting-edge tool within the machine learning community. This algorithm is built upon classification trees [25] and the gradient boosting framework [26]. XGB, an extensively used machine learning system, enhances the performance of classification trees [27]. A classification tree establishes rules to categorize each instance of gully erosion based on predisposing factors within a graph structure. In this framework, a single tree is constructed, with leaves assigned scores indicating the likelihood of a gully falling into a specific factor class, whether categorical (e.g., lithotypes) or ordinal (e.g., reclassified slope steepness).

Within the XGB framework, the loss function used to train the ensemble model is augmented with regularization, penalizing the complexity of trees. This regularization technique can improve the performance of the gully erosion model by mitigating overfitting. Overfitting occurs when a model performs well on the training data but struggles with new datasets, limiting its predictive ability [28]. Regularization helps mitigate overfitting and enhances the flexibility of the gully erosion prediction model. XGB combines the outcomes of various tree models by averaging their weighted results [29]. An iterative process is employed, using weak prediction models to refine the overall prediction model at each step by correcting misclassifications from the previous iteration. The XGB model is constructed by optimizing a specific objective function.

$$OF\left(\theta\right)=\sum_{i=1}^{n} l\left(y_{i,}\ \bar{y}_i\right)a+\sum_{i=1}^{k}\Omega\left(f_k\right) \qquad (1)$$

Here, $\sum_{i=1}^{n} l\left(y_{i,}\ \bar{y}_i\right)$ represents the loss function of root mean square error utilized for model fitting on the training data. $\sum_{i=1}^{k}\Omega\left(f_k\right)$ refers to the regularization term, aiding in preventing overfitting. K denotes the number of individual trees, and f_k represents a tree within the ensemble. $y_{i,}$ and $\bar{y}_i$ denote the actual and predicted class outputs, respectively.

The training procedure of XGB is centered around minimizing the objective function mentioned earlier. This is achieved by iteratively adding weak learners to the ensemble. Importantly, as the learning of trees advances, the complexity of the model increases; however, the regularization term acts to mitigate overfitting issues by controlling the number of leaf nodes in the tree [29]. Further details regarding the construction phases of the XGBoost model can be found in previous studies by [24; 29].

In this study, we adhered to recommended parameter configurations for XGB tuning, as outlined by [28]. These include specifying the maximum number of training rounds (iterations) around 200, capping tree depth at 6, and fine-tuning other parameters such as learning rate, regularization, variable selection, and minimum child weight. Furthermore, we utilized ArcGIS software (version 10.8) to delineate the basin's gully-dominant region by overlaying the final gully susceptibility maps generated by XGB models, with a specific focus on areas classified as having a very high susceptibility class.

Model validation

Validation and accuracy assessment play a pivotal role in evaluating models for management studies [30]. Without proper validation, the interpretation of machine learning model outputs holds less real-world significance. In this study, various statistical indices, including classification accuracy Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Kappa Index (K), R², Receiver Operating Characteristic (ROC) curve, and Accuracy (ACC), among others, were employed for validating and assessing the accuracy of the model results.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

$$Kappa(k) = \frac{P_c - P_{cxp}}{1 - P_{cxp}} \quad (3)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{i=n}(X_{ei} - X_{oi})^2} \quad (4)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{i=n}\left|X_{ei} - X_{oi}\right| \quad (5)$$

Where, P_c refers to a number of pixels to be classified correctly as gully eroded or non-gully eroded pixels; P_{cxp} denotes expected results. X_{oi} and X_{ei} are the i^{th} observed and model estimation values, respectively, and n is the number of data points [23].

RESULTS

Influence of key factors on GESM

We examined the potential multicollinearity (redundancy) among the 24 selected geo-environmental factors through two techniques: Information Gain Ratio (IGR) and Variance Inflation Factors (VIF). The results presented in Table 2 indicate no concerns regarding multicollinearity, as all VIF values range between 0.1 and 10. According to the statistical tests, there exists a low level of collinearity among these factors, indicating their relative independence. Additionally, Table 2 illustrates the relative influence of each factor on GESM within the Sita Nala River basin. The table displays the Average Merit (AM) for each factor, representing its significance in the GESM. Factors with the highest average merit values, such as drainage density (0.77), elevation (0.74), geomorphology (0.72), LULC (0.72), and NDVI (0.68), exert the greatest influence on an area's susceptibility to gully erosion. Conversely, the distance from bulk density (0.21) exhibits the least influence. Subsequent to these primary factors in descending order of influence are Carbon exchange capacity (0.23), Silt (0.25), Distance from the lineament (0.25), and curvature (0.29). These findings provide crucial insights into the primary drivers of gully erosion in this region, facilitating the development of more effective management strategies.

Table 2 Multicollinearity diagnosis using VIF tests and Information Gain Ratio (Average Merit)

No.	Influencing factors	VIF	Information gain ratio
1.	DEM (Elevation)	3.85	0.74
2.	Slope	2.11	0.42
3.	Slope length	1.19	0.33
4.	Slope aspect	1.33	0.36
5.	Curvature	1.88	0.29
6.	Drainage density (DD)	3.97	0.77
7.	Distance from the river (DFR)	2.41	0.55
8.	Distance from the lineament (DFL)	1.39	0.25
9.	Topographic Weightiness Index (TWI)	1.57	0.30
10.	Rainfall	2.46	0.49
11.	NDVI	3.22	0.68
12.	Land Use and Land Cover (LULC)	3.48	0.70
13.	Distance from the road (DR)	1.89	0.34
14.	Lithology	1.79	0.48
15.	Geomorphology	3.74	0.72
16.	Soil organic density (SOD)	2.21	0.57
17.	Bulk density	1.59	0.21
18.	Clay Content in Soil (SC)	2.99	0.63
19.	Coarse fragments	1.91	0.61
20.	Sand	2.34	0.38
21.	Silt	2.11	0.25
22.	Carbon exchange capacity (CEC)	1.89	0.23
23.	Nitrogen	1.78	0.28
24.	Soil organic carbon (SOC)	1.39	0.31

Gully erosion susceptibility mapping

GESMs serve as crucial instruments for environmental protection and sustainable development, aiding in the comprehension of erosion patterns and the formulation of mitigation strategies. This study utilized 24 factors influencing gully formation to develop a GESM employing the XGB model (Fig. 5). To delineate clear risk zones, the quantile classification method within ArcGIS software was utilized, enabling the classification of susceptibility levels into three zones: areas with no gully erosion, those with high susceptibility, and those with very high susceptibility [31]. The examination of each susceptibility zone offered a comprehensive understanding of the vulnerability of different areas within the study region to gully erosion. Figure 3 illustrates the GESMs generated by the XGB models, showcasing the predicted susceptibility levels across the Sita Nala River basin. Monitoring revealed slight disparities in the spatial distribution of risk zones, with the XGB model classifying a larger portion of the study area into the high gully susceptibility area and moderate gully susceptibility area at 13.49% and 11.24% respectively (Table 3, and Fig. 3). Conversely, an area was classified as non-gully erosion area (75.24%) by the XGB model. Encouragingly, the model exhibited good classification abilities and effectively identified the prominent gully-prone areas within the basin. Ultimately, the XGB modeling techniques proved effective in constructing GESMs that delineate distinct zones of varying susceptibility levels throughout the river basin under investigation. The model demonstrated satisfactory performance, accurately reflecting and aligning with the observed patterns of gully prevalence across the region. By pinpointing areas susceptible to gully erosion, these GESMs offer valuable tools for prioritizing conservation efforts and devising sustainable land management practices, thereby contributing to environmental protection and the long-term health of the basin.

Validation of model

To assess the effectiveness of the XGB model, we conducted a battery of statistical tests on both the training and testing datasets. These tests included ACC, MAE, Kappa index, R², and RMSE. Analysis of the training dataset revealed that the XGB model exhibited superior performance across all metrics (Table 4). It achieved an impressive accuracy rate of 88.1%, a Kappa index of 0.84 (indicating strong agreement), a remarkably high R² of 0.85, along with the lowest RMSE (0.16) and MAE (0.12) values.

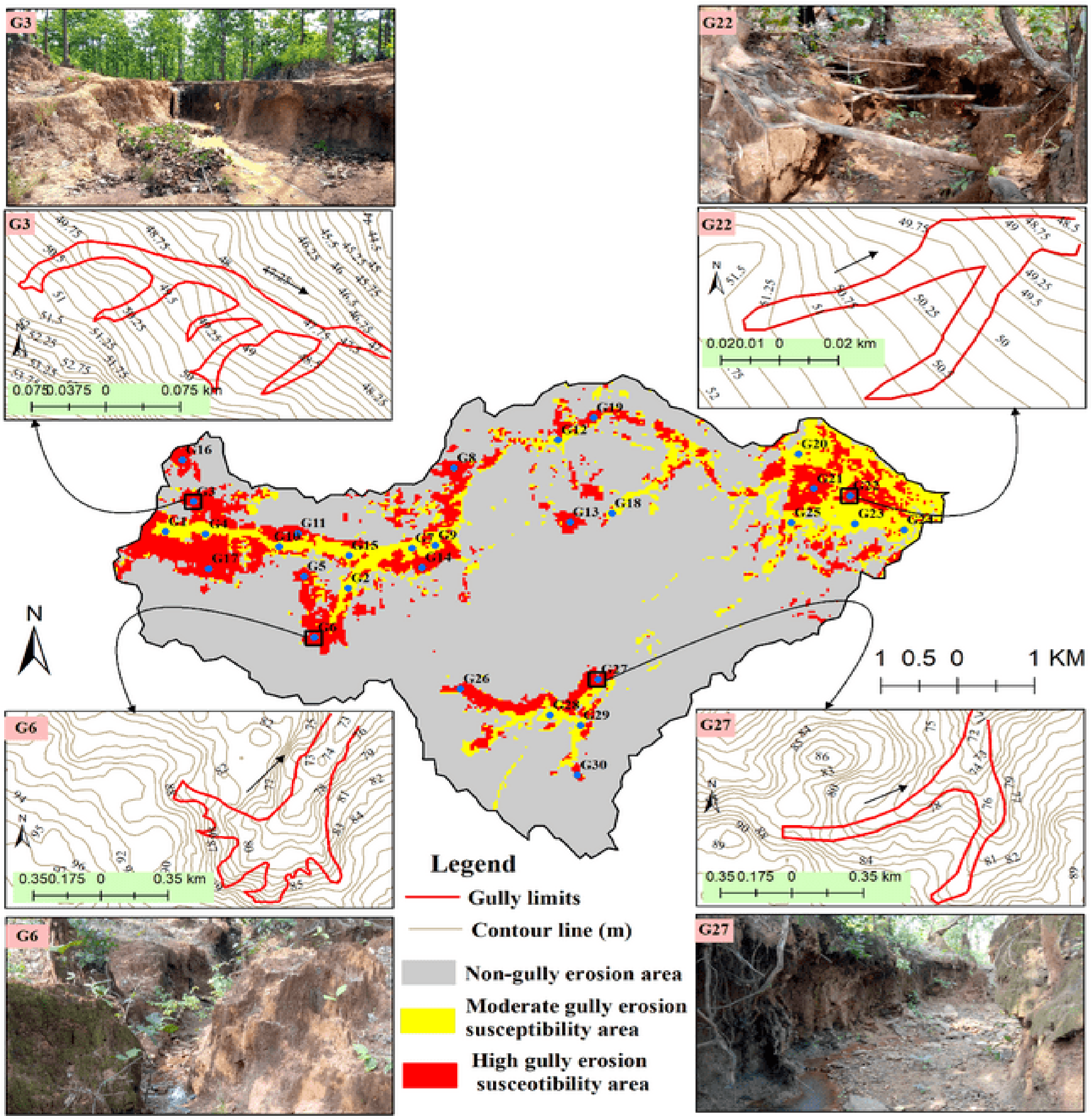

Fig. 3 Prediction of gully erosion susceptibility mapping with ground truth

Table 3 Gully erosion susceptibility mapping of the study area using the XGB algorithms

Classes	XGB	
	Area in km²	Area in %
High gully susceptibility area (0.75-1)	5.21	13.49
Moderate gully susceptibility area (0.5-0.75)	4.34	11.24
Non-gully erosion area (less than 0.5)	29.05	75.24
Total	38.60	100

Table 4 Prediction performances of the proposed XG Boost model using the training and testing dataset.

Statistical index	XGB	
	Training	Testing
Accuracy (%)	88.1	87.7
Kappa index (K)	0.84	0.85
MAE	0.12	0.15
RMSE	0.16	0.14
R²	0.85	0.88

These findings underscore the XGB model's exceptional capability to accurately forecast gully erosion. Furthermore, the evaluation results underscored the consistent and robust performance of the XGB models, as they consistently demonstrated comparable levels of predictive accuracy across both the testing and training datasets when evaluated using the same set of statistical metrics (refer to Table 4). This suggests that the models generalize well and can be confidently applied to unseen data.

For a more assessment of the model performance, we utilized ROC curve analysis, depicted in Figure 4. The ROC curve visually illustrates the balance between the false positive rate (the proportion of non-gully areas erroneously classified as gullies) and the true positive rate (the proportion of actual gully areas correctly identified) across various probability thresholds. Both models demonstrated highly satisfactory predictive capabilities in delineating gully erosion susceptibility zones, as indicated by the Area Under the Curve (AUC) values, which stood at 84.2% for the XGB model. This substantial AUC value, with a slight advantage observed for the XGB model, signifies robust model performance, where higher values correspond to greater accuracy in distinguishing between gully-prone and non-gully areas.

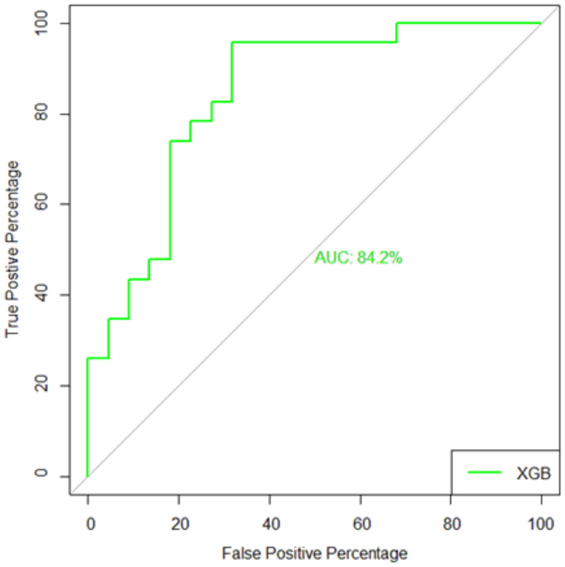

Fig.4 ROC of the prediction model

DISCUSSION

Gully erosion poses a grave environmental threat, impacting not only the landscape but also human safety and agricultural productivity. It contributes to habitat destruction, soil loss, water contamination, infrastructure damage, and increased flood risks. Conventional statistical approaches to analyzing gully erosion may be constrained by issues like overfitting and uncertainty [32]. This study presents a robust solution by harnessing machine learning techniques. We have developed a precise Gully Erosion Susceptibility Model (GESM) tailored for the Sita Nala River basin. This model integrates 24 crucial factors influencing gully formation and employs two advanced machine learning algorithms, notably the XGB model. Our research stands out for its comprehensive methodology. Unlike many studies on gully erosion, we incorporate a diverse array of data types, employ rigorous validation methods, and consider a comprehensive set of controlling factors. Additionally, we utilized Information Gain Ratio (IGR) and Variance Inflation Factor (VIF) techniques to evaluate the 24 factors and ensure they are not overly correlated (multicollinearity). Analysis of the Average Merit (AM) highlighted drainage density (0.77), elevation (0.74), geomorphology (0.72), LULC (0.72), and NDVI (0.68) as the most influential factors shaping the GESM for the Sita Nala River basin. Conversely, bulk density emerged as the least influential factor. Our findings resonate with numerous studies conducted across various geomorphological and climatic regions. Studies by Huang *et al.* [33] in China, Saha *et al.* [34] in India, Setargie *et al.* [35] in Ethiopia, and Garosi *et al.* [21] in Iran consistently underscore the critical role of diverse factors in gully erosion development, aligning with the outcomes of our investigation.

To evaluate the XGB model's performance thoroughly, we conducted an extensive array of statistical tests on both the training and testing datasets. These tests encompassed ROC analysis, RMSE, MAE, R², ACC, and Kappa index analysis. While both models demonstrated robust performance and suitability for this study, the XGB model emerged as particularly adept at predicting GESM. This finding contrasts with prior studies by Saha *et al.* [34], Avand *et al.* [36], and Hosseinalizadeh *et al.* [37], which favored the RF model for its accuracy. Monitoring revealed slight disparities in the spatial distribution of risk zones, with the XGB model classifying a larger portion of the study area into the High and moderate gully erosion susceptibility categories, accounting for 13.49% and 11.24%, respectively (refer to Table 3). Conversely, the XGB model identified 75.24% of the area as non-gully erosion areas. To validate the accuracy of our models, we conducted follow-up field surveys in January and March 2024, utilizing GPS technology to obtain precise location data. This field validation process is pivotal as it illustrates how well the model predictions align with real-world conditions. The findings of this research unequivocally demonstrate the suitability of the XGB machine learning technique for GESM. By facilitating targeted interventions, this approach transcends mere management strategies, fostering sustainable land use practices by integrating community involvement and promoting ecological benefits. Ultimately, this integrated approach aims to mitigate the adverse impacts of gully erosion on the environment, soil, and land, thereby contributing to a more sustainable future.

CONCLUSIONS

This study successfully developed a precise GESM for the Sita Nala River basin, employing machine learning algorithms, specifically the XGB algorithm, and incorporating 24 essential factors influencing gully formation. To ensure these factors were not overly correlated, we employed IGR and VIF techniques, which revealed no multicollinearity issues. Our analysis within the basin identified drainage density (0.77), elevation (0.74), geomorphology (0.72), Land Use/Land Cover (LULC) (0.72), and Normalized Difference Vegetation Index (NDVI) (0.68) as the most significant factors controlling susceptibility to gully erosion, while distance from the lineament had the least impact. We conducted rigorous statistical tests (RMSE, MAE, Kappa index, R^2, ACC, and ROC) on both training and testing data to evaluate the model's performance. The XGB model demonstrated commendable performance and proved well-suited for predicting gully erosion susceptibility zones within the GESM. The model identified approximately 13.49% of the basin area as highly susceptible to gully erosion, forming what can be termed as gully-dominated zones. This critical finding emphasizes the urgent need for targeted management techniques in these vulnerable areas. Our findings suggest that the machine learning based on XGB model effectively identified regions with actively developing gullies. By focusing on these vulnerable areas, decision-makers can implement tailored and sustainable programs and policies to mitigate the future impacts of gully erosion on residents.

Acknowledgement

The authors show their kind acknowledgment to the Dept. of Geography, Raja N. L. Khan Women's College (Autonomous), West Bengal, India for their laboratory facilities and kind encouragement.

Contributions

Md. Hasanuzzaman - conceptualized and planned the study, conducted the survey, analyzed the data, and interpreted the results. P.K.Shit - supervised the study and reviewed and edited the manuscript. All authors have read and approved the final manuscript.

Ethic declarations

Competing interests

The authors declare that they have no competing interests.

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Data availability

The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

References

1. Azareh, A., Rahmati, O., Rafiei-Sardooi, E., Sankey, J.B., Lee, S., Shahabi, H. and Ahmad, B.B., 2019. Modelling gully-erosion susceptibility in a semi-arid region, Iran: Investigation of applicability of certainty factor and maximum entropy models. *Science of the Total Environment*, *655*, pp.684-696.
2. Shit, P.K., Pourghasemi, H.R. and Bhunia, G.S. eds., 2019. *Gully erosion studies from India and surrounding regions*. Springer Nature.
3. Poesen J, Nachtergaele J, Verstraeten G, Valentin C. 2003. Gully erosion and environmentalbchange: importance and research needs. CATENA. 50(2–4):91–133.
4. Hassen, G., & Bantider, A. (2020). Assessment of drivers and dynamics of gully erosion in case of Tabota Koromo and Koromo Danshe watersheds, South Central Ethiopia. *Geoenvironmental Disasters*, *7*(1), 5.
5. Chen, W., Lei, X., Chakrabortty, R., Pal, S.C., Sahana, M. and Janizadeh, S., 2021. Evaluation of different boosting ensemble machine learning models and novel deep learning and boosting framework for head-cut gully erosion susceptibility. *Journal of Environmental Management*, *284*, p.112015.DOI: 10.1016/j.jenvman.2021.112015
6. Igwe, O., John, U.I., Solomon, O. and Obinna, O., 2020. GIS-based gully erosion susceptibility modeling, adapting bivariate statistical method and AHP approach in Gombe town and environs Northeast Nigeria. *Geoenvironmental Disasters*, *7*, pp.1-16.
7. Majhi, A., Nyssen, J. and Verdoodt, A., 2021. What is the best technique to estimate topographic thresholds of gully erosion? Insights from a case study on the permanent gullies of Rarh plain, India. *Geomorphology*, *375*, p.107547. DOI: 10.1016/j.geomorph.2020.107547
8. Rahmati, O., Tahmasebipour, N., Haghizadeh, A., Pourghasemi, H.R. and Feizizadeh, B., 2017. Evaluation of different machine learning models for predicting and mapping the susceptibility of gully erosion. *Geomorphology*, *298*, pp.118-137.
9. Arabameri, A., Rezaei, K., Pourghasemi, H.R., Lee, S. and Yamani, M., 2018. GIS-based gully erosion susceptibility mapping: a comparison among three data-driven models and AHP knowledge-based technique. *Environmental earth sciences*, *77*, pp.1-22.
10. Razavi-Termeh, S.V., Sadeghi-Niaraki, A. and Choi, S.M., 2020. Gully erosion susceptibility mapping

using artificial intelligence and statistical models. *Geomatics, Natural Hazards and Risk*, *11*(1), pp.821-844.

11. Mehmood, Q., Qing, W., Chen, J., Yan, J., Ammar, M. and Rahman, G., 2021. Susceptibility assessment of single gully debris flow based on AHP and extension method. *Civil Engineering Journal*, *7*(06).
12. Mohebzadeh, H., Biswas, A., Rudra, R. and Daggupati, P., 2022. Machine learning techniques for gully erosion susceptibility mapping: a review. *Geosciences*, *12*(12), p.429.
13. Ghorbanzadeh, O., Shahabi, H., Mirchooli, F., Valizadeh Kamran, K., Lim, S., Aryal, J., Jarihani, B. and Blaschke, T., 2020. Gully erosion susceptibility mapping (GESM) using machine learning methods optimized by the multi-collinearity analysis and K-fold cross-validation. *Geomatics, Natural Hazards and Risk*, *11*(1), pp.1653-1678.
14. Liu, G., Arabameri, A., Santosh, M. and Nalivan, O.A., 2023. Optimizing machine learning algorithms for spatial prediction of gully erosion susceptibility with four training scenarios. *Environmental Science and Pollution Research*, *30*(16), pp.46979-46996.
15. Mosavi, A., Golshan, M., Janizadeh, S., Choubin, B., Melesse, A.M. and Dineva, A.A., 2022. Ensemble models of GLM, FDA, MARS, and RF for flood and erosion susceptibility mapping: a priority assessment of sub-basins. *Geocarto International*, *37*(9), pp.2541-2560.
16. Samanta RK, Bhunia GS, Shit PK (2016) Spatial modelling of soil erosion susceptibility mapping in lower basin of Subarnarekha River (India) based on geospatial techniques. Modeling Earth Systems and Environment. 2(2):99.doi.org/10.1007/ s40808-016-0170-2
17. El Maaoui MA, Sfar Felfoul M, Boussema MR, Snane MH. 2012. Sediment yield from irregularly shaped gullies located on the Fortuna lithologic formation in semi-arid area of Tunisia. CATENA. 93:97–104.
18. Conoscenti C, Agnesi V, Cama M, Caraballo-Arias NA, Rotigliano E. 2018. Assessment of gully erosion susceptibility using multivariate adaptive regression splines and accounting for terrain connectivity. Land Degrad Dev. 29(3):724–736.
19. Tien Bui D, Shirzadi A, Shahabi H, Chapi K, Omidavr E, Pham BT, Talebpour Asl D, Khaledian H, Pradhan B, Panahi M, *et al*. 2019. A novel ensemble artificial intelligence approach for gully erosion mapping in a semi-arid watershed (Iran). Sensors. 19(11):2444.
20. Hitouri, S., Meriame, M., Ajim, A.S., Pacheco, Q.R., Nguyen-Huy, T., Bao, P.Q., ElKhrachy, I. and Varasano, A., 2024. Gully erosion mapping susceptibility in a Mediterranean environment: A hybrid decision-making model. *International Soil and Water Conservation Research*, *12*(2), pp.279-297.
21. Garosi, Y., Sheklabadi, M., Pourghasemi, H.R., Besalatpour, A.A., Conoscenti, C. and Van Oost, K., 2018. Comparison of differences in resolution and sources of controlling factors for gully erosion susceptibility mapping. *Geoderma*, *330*, pp.65-78. DOI: 10.1016/j.geoderma.2018.05.027
22. Hasanuzzaman, M., Shit, P.K., Bera, B. and Islam, A., 2023. Characterizing recurrent flood hazards in the Himalayan foothill region through data-driven modelling. *Advances in Space Research*, *71*(12), pp.5311-5326.
23. Khosravi, K., Shahabi, H., Pham, B.T., Adamowski, J., Shirzadi, A., Pradhan, B., Dou, J., Ly, H.-B., Gróf, G., Ho, H.L., 2019. A comparative assessment of flood susceptibility modeling using Multi-Criteria Decision-Making Analysis and Machine Learning Methods. Journal of Hydrology 573, 311–323.
24. Chen, T., Guestrin, C., 2016. Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, pp. 785–794.
25. Breiman L, Friedman J, Stone CJ, Olshen RA. 1984. Classification and regression trees. Taylor & Francis, Milton Park, Abingdon-on-Thames, Oxfordshire United Kingdom.
26. Friedman JH. 2001. Greedy function approximation: a gradient boosting machine. Ann Stat. 29(5):1189–1232.
27. Li P. 2010. Robust logitboost and adaptive base class (ABC) logitboost. Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence, UAI 2010 [Internet]. p. 302–311. [accessed 2020 Oct 13]; https://www.researchwithrutgers.com/en/publications/robust-logitboost-and-adaptive-base-class-abc-logitboost. Jul 8-11, Catalina Island, California
28. Boehmke, B., Greenwell, B.M., 2019. Hands-on machine learning with R. CRC Press.
29. Gusain K, Gupta A, Popli B. 2018. Transition-aware human activity recognition using extreme gradient boosted decision trees. In: Choudhary RK, Mandal JK, Bhattacharyya D, editors. Advanced computing and communication technologies. Singapore: Springer; p. 41-49.
30. Wang H, Moayedi H, Kok Foong L. 2020. Genetic algorithm hybridized with multilayer perceptron to have an economical slope stability design. Eng Comput.. 28:1-2.
31. Tehrany MS, Pradhan B, Mansor S, Ahmad N. 2015. Flood susceptibility assessment using GIS-based support vector machine model with different kernel types. CATENA. 125:91–101.
32. Telikani, A., Tahmassebi, A., Banzhaf, W. and Gandomi, A.H., 2021. Evolutionary machine learning: A survey. *ACM Computing Surveys (CSUR)*, *54*(8), pp.1-35.
33. Huang, D., Su, L., Zhou, L., Tian, Y. and Fan, H., 2023. Assessment of gully erosion susceptibility using different DEM-derived topographic factors in the black soil region of Northeast China. *International Soil and Water Conservation Research*, *11*(1), pp.97-111.
34. 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. https://doi.org/10.3390/ s20051313.
35. Setargie, T.A., Tsunekawa, A., Haregeweyn, N., Tsubo, M., Fenta, A.A., Berihun, M.L., Sultan, D., Yibeltal, M., Ebabu, K., Nzioki, B. and Meshesha, T.M., 2023. Random Forest–based gully erosion susceptibility assessment across different agro-ecologies of the Upper Blue Nile basin,

Ethiopia. *Geomorphology*, p.108671.DOI: 10.1016/j.geomorph.2023.108671.

36. Avand, M., Janizadeh, S., Naghibi, S.A., Pourghasemi, H.R., Khosrobeigi Bozchaloei, S. and Blaschke, T., 2019. A comparative assessment of random forest and k-nearest neighbor classifiers for gully erosion susceptibility mapping. *Water*, *11*(10), p.2076.
37. Hosseinalizadeh, M., Kariminejad, N., Chen, W., Pourghasemi, H.R., Alinejad, M., Behbahani, A.M. and Tiefenbacher, J.P., 2019. Gully headcut susceptibility modeling using functional trees, naïve Bayes tree, and random forest models. *Ge oderma*, *342*, pp.1-11.

How to cite this article:

Md Hasanuzzaman and Pravat Kumar Shit.(2024). Prediction of gully erosion susceptibility mapping using xgboosting machine learning algorithm. *Int J Recent Sci Res.*15(05), pp.4704-4713.

****