Water Resources Management
https://doi.org/10.1007/s11269-024-03760-9

A Comparative Assessment of Decision Tree Algorithms for Index of Sediment Connectivity Modelling

Haniyeh Asadi[1] · Mohammad T. Dastorani[1] · Roy C. Sidle[2] · Afshin Jahanshahi[3]

Received: 2 November 2023 / Accepted: 16 January 2024

Abstract
Assessment of the spatial distribution of potential pathways of sediment transport and the degree of linkage between sediment sources and the channel network within a watershed represents a valuable analysis for informing management decisions on sediment yield and transfer. Given the limitations of conventional methods for determining index of sediment connectivity (IC), there is a need to provide a flexible and efficient approach with the ability to apply different factors. In this regard, five decision tree-based machine learning models: M5 prime (M5P), random tree (RT), random forest (RF), alternating model tree (AMT), and reduced error pruning tree (REPT) were tested using geomorphic and climatic factors. Two databases were constructed with 200 and 1600 classes at 50 watersheds in Queensland, Australia. In these models, IC was assessed as an output parameter and six attributes that affect IC were assigned as input parameters (i.e., elevation, slope, area, length of stream channel, normalized difference vegetation index, and rainfall). Statistical validation and comparison of model predictions with calculated IC values based on the approach of Borselli et al. (Catena 75:268–277, 2008) were performed. Based on the statistical criteria, the RF model produced the most robust estimations of IC compared to other models and performed very well for IC modelling, especially in smaller subsections of watersheds. Accordingly, these findings can play an effective role for implementing watershed management and soil and water resources management measures.

Keywords Sediment connectivity · Machine learning · Decision tree algorithm · Random forest · Geomorphic factors

✉ Mohammad T. Dastorani
dastorani@um.ac.ir

[1] Faculty of Natural Resources and Environment, Ferdowsi University of Mashhad, Mashhad, Iran

[2] Mountain Societies Research Institute, University of Central Asia, Bishkek, Kyrgyzstan

[3] Department of Watershed Management, Sari Agricultural Sciences and Natural Resources University, Sari, Iran

1 Introduction

Connectivity is commonly understood as the degree of linkage among component parts of geomorphic systems (e.g., Cavalli et al. 2013; Heckmann and Schwanghart 2013). Many researchers attempted to formulate connectivity definitions and developed various indices and models to assess connectivity and its spatial and temporal distribution (Cavalli et al. 2013; Fryirs 2013; Bracken et al. 2015; Gay et al. 2016; Masselink et al. 2016). Sediment connectivity as a hydrological linkage includes two distinct concepts (structural/physical and functional/process-based), which assess the spatial distribution of potential pathways of sediment transport and dynamics of hydrogeomorphic processes in different sections of the watershed (between possible sediment sources and potential sinks) (Turnbull et al. 2018; Najafi et al. 2021). Given the need for appropriate low-cost methods to assess sediment connectivity, various approaches have been applied in previous studies, including the digital elevation model (DEM) of difference (DoD) technique (Croke et al. 2013; Heckmann and Vericat 2018), geographical information system (GIS) modelling based on morphometric characteristics (Borselli et al. 2008; Cavalli et al. 2016), spatial network analysis (Phillips et al. 2015; Fressard and Cossart 2019), and intelligent modelling based on physical characteristics of the watershed (Asadi et al. 2023a). The index of sediment connectivity (IC) is a hydrogeomorphic tool that practically assesses sediment connectivity from hillslopes to downstream channels (Martini et al. 2022). The application of IC in the study of hydrologic and geomorphic processes in different watersheds around the world has increased substantially in the past two decades (Fryirs 2013; Parsons et al. 2015; Wohl et al. 2019; Koci et al. 2020). For example, it has been used in implementation of post-fire programs in burned watersheds (López-Vicente et al. 2020), modelling and quantifying the probability of flooding at major road-stream junctions (Kalantari et al. 2019), improving flood hazard assessment and management (Keesstra et al. 2018), and providing practical guidelines for mitigation and restoration of abandoned lands (Marchamalo et al. 2016). In general, the IC can be used to distinguish homogeneous sections with similar potentials for sediment transport that helps to prioritize different parts of watershed for sediment control measures and to facilitate watershed management, especially in regions with high erosion and sediment delivery rates (Asadi et al. 2023a; González-Romero et al. 2021).

In recent years, machine learning (ML) models have been increasingly used as a powerful tool to simulate complex phenomenon (Asadi et al. 2021; Aghamolaei and Hessami-Kermani 2023; Chia et al. 2023; Gelete 2023; Zhao et al. 2023). Decision tree (DT) algorithms are one of the main tools of ML that have been widely applied, including to study issues in water resources, hydrology, climatology, and hydraulics (Bui et al. 2020). Example of uses include prediction of bedload transport rates in gravel-bed rivers (Khosravi et al. 2020), suspended sediment loads (Al-Mukhtar 2019), dissolved oxygen concentrations in rivers (Heddam and Kisi 2018), apparent shear stress (Khozani et al. 2019), water quality indices (Bui et al. 2020), land degradation (Yousefi et al. 2021), and daily river flow (Ghorbani et al. 2020). Most of these studies have compared different ML models with DT-based models and confirmed the superiority of DT models compared with other methods for similar conditions. The better performance of DT-based models is related to the lack of hidden layers and model transparency, handling data from various scales, facilitating the construction of rules for prediction of complex relationships, and statistical analysis without any assumptions of statistical distribution (Tehrany et al. 2013).

Although advances have been achieved related to quantifying the degree of sediment connectivity throughout the world and the predictive power of various DT-based algorithms has been proven for different hydrological phenomena, the performance of these algorithms to predict IC has not been studied. Therefore, the main objectives of this study are: (1) investigating the efficiency of DT-based algorithms, namely M5 prime (M5P), random tree (RT), random forest (RF), alternating model tree (AMT), and reduced error pruning tree (REPT) in estimation of IC; (2) comparing the predictive power of these models; and (3) performing a sensitivity analysis of the effective variables used. On the other hand, there are no studies that have used climatic variables to develop an efficient ML-based approach for estimating IC. Therefore, investigating rainfall as a climatic input in estimation of IC in addition to the geomorphic characteristics of the watersheds is the fourth objective of this study. Finally, the fifth objective is to examine the effect of modelling scale on the model performance.

2 Study Area

To pursue these objectives, 50 watersheds in Queensland, Australia were evaluated (Fig. 1; Table 1) in two case studies: in the first case study, 50 watersheds were categorized into four classes (low, medium-low, medium-high, and high) (Tiranti et al. 2018; Najafi et al. 2021) (e.g., Fig. 2). In the second case study, 50 watersheds were categorized into 32 classes (Asadi et al. 2023a) (e.g., Fig. 3). Classification of watersheds was done based on values of IC and using a natural breaks classification algorithm that, overall, created 200 and 1600 classes in case studies 1 and 2, respectively. It should be noted that the selected watersheds had minimal anthropogenic influences (i.e., without dams or the major abstractions in the upstream reaches). A basic flowchart of the methodology is shown in Fig. 4.

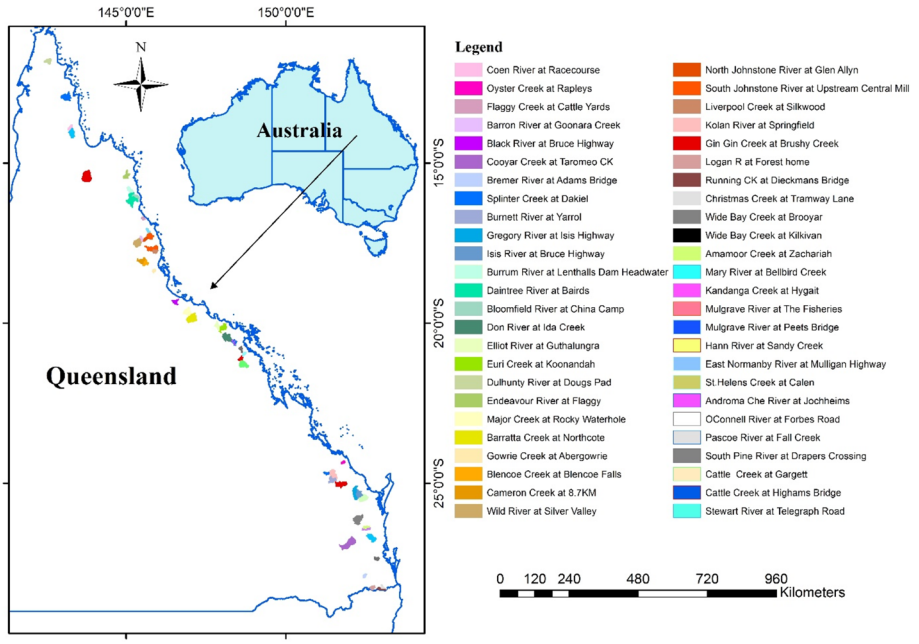

Fig. 1 Location of 50 studied watersheds in Queensland, Australia

Table 1 Name and hydrological characteristics of the studied watersheds

Basin	Watershed	Drainage area (km^2)	Basin	Watershed	Drainage area (km^2)
Archer	Coen River at Racecourse	172	Johnstone	North Johnstone River at Glen Allyn	165
Baffle	Oyster Creek at Rapleys	194	Johnstone	South Johnstone River at Upstream Central Mill	400
Barron	Flaggy Creek at Cattle Yards	150	Johnstone	Liverpool Creek at Silkwood	242
Barron	Barron River at Goonara Creek	127	Kolan	Kolan River at Springfield	551
Black	Black River at Bruce Highway	256	Kolan	Gin Gin Creek at Brushy Creek	531
Brisbane	Cooyar Creek at Taromeo CK	963	Logan-Albert	Logan R at Forest home	175
Brisbane	Bremer River at Adams Bridge	125	Logan-Albert	Running CK at Dieckmans Bridge	128
Burnett	Splinter Creek at Dakiel	139	Logan-Albert	Christmas Creek at Tramway Lane	166
Burnett	Burnett River at Yarrol	370	Mary	Wide Bay Creek at Brooyar	655
Burrum	Gregory River at Isis Highway	454	Mary	Wide Bay Creek at Kilkivan	322
Burrum	Isis River at Bruce Highway	446	Mary	Amamoor Creek at Zachariah	133
Burrum	Burrum River at Lenthalls Dam Headwater	515	Mary	Mary River at Bellbird Creek	486
Daintree	Daintree River at Bairds	911	Mary	Kandanga Creek at Hygait	143
Daintree	Bloomfield River at China Camp	264	Mulgrave-Russell	Mulgrave River at The Fisheries	357
Don	Don River at Ida Creek	604	Mulgrave-Russell	Mulgrave River at Peets Bridge	520
Don	Elliot River at Guthalungra	273	Normanby	Hann River at Sandy Creek	984
Don	Euri Creek at Koonandah	429	Normanby	East Normanby River at Mulligan Highway	297
Ducie	Dulhunty River at Dougs Pad	332	OConnell	St.Helens Creek at Calen	118
Endeavour	Endeavour River at Flaggy	337	OConnell	Androma Che River at Jochheims	230
Haughton	Major Creek at Rocky Waterhole	468	OConnell	OConnell River at Forbes Road	167
Haughton	Barratta Creek at Northcote	753	Olive-Pascoe	Pascoe River at Fall Creek	651
Herbert	Gowrie Creek at Abergowrie	124	Pine	South Pine River at Drapers Crossing	156
Herbert	Blencoe Creek at Blencoe Falls	226	Pioneer	Cattle Creek at Gargett	326
Herbert	Cameron Creek at 8.7KM	360	Pioneer	Cattle Creek at Highams Bridge	198
Herbert	Wild River at Silver Valley	591	Stewart	Stewart River at Telegraph Road	470

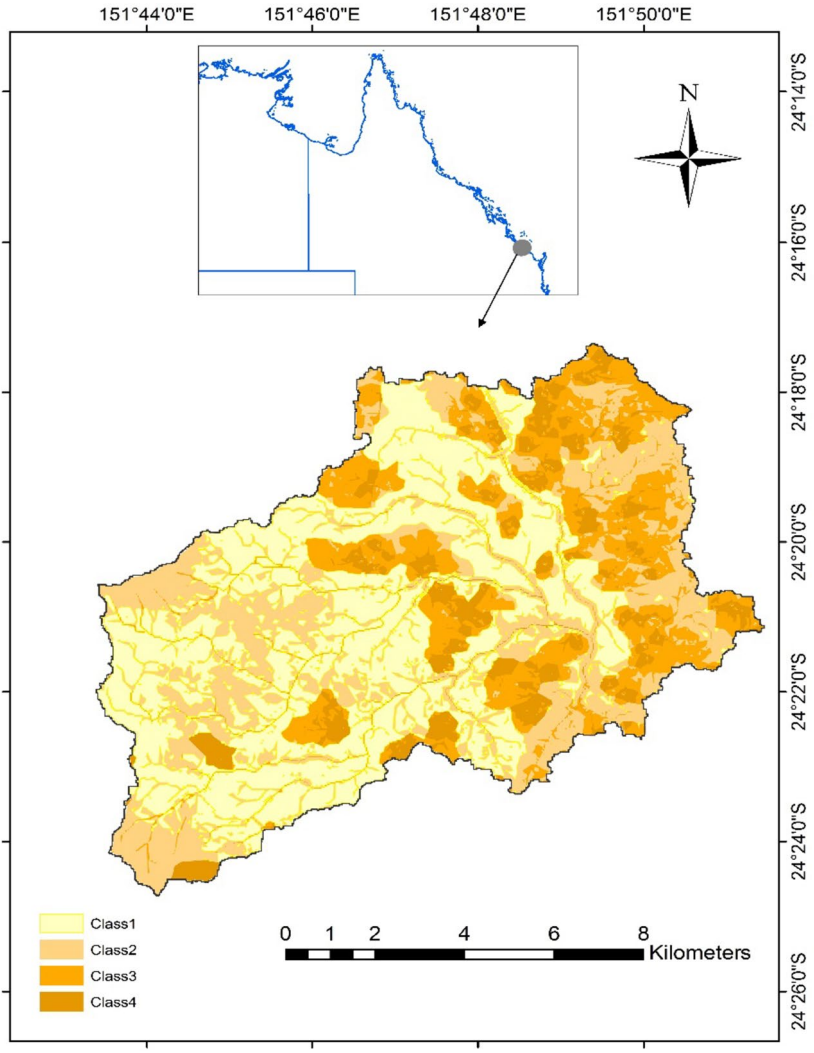

Fig. 2 Map of 4 classes of Oyster Creek watershed in case study 1 as an example

3 Methodology

3.1 Development and Application of DT-Based Algorithms

3.1.1 Sample Size

Overall, 200 and 1600 classes were generated for case studies 1 and 2, respectively. The statistical parameters of input and output variables (elevation, slope, area, length of stream channel, normalized difference vegetation index (NDVI), rainfall, and IC) in these classes were calculated (Table 2). To avoid over-fitting, K-fold cross-validation was used to train and test the

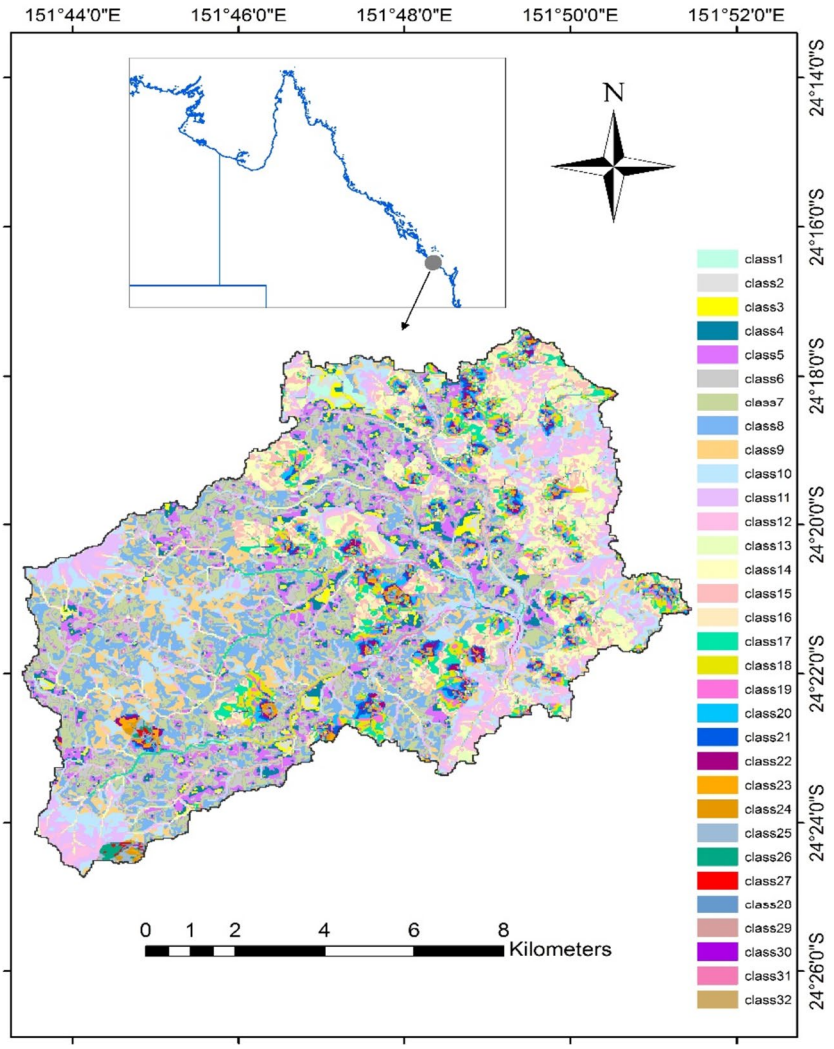

Fig. 3 Map of 32 classes of Oyster Creek watershed in case study 2 as an example

DT-based models (Asadi et al. 2023b). In this approach, all data were randomly partitioned into equal sized subsamples (i.e., 5 and 10 subsamples in case studies 1 and 2, respectively). Of the 5 subsamples in case study 1, 4 subsamples (160 classes) and of the 10 subsamples in case study 2, 9 subsamples (1440 classes) were selected for training; the one remaining subsample in each case study was used for testing. This process in case studies 1 and 2 was repeated 5 and 10 times, respectively and each time one of the subsamples was used as the validation data.

3.1.2 Preparation of Modelling Dataset

IC Factor Among methods to calculate sediment connectivity, the topography-based structural IC introduced by Borselli et al. (2008) is widely used due to the lack of large data requirements and

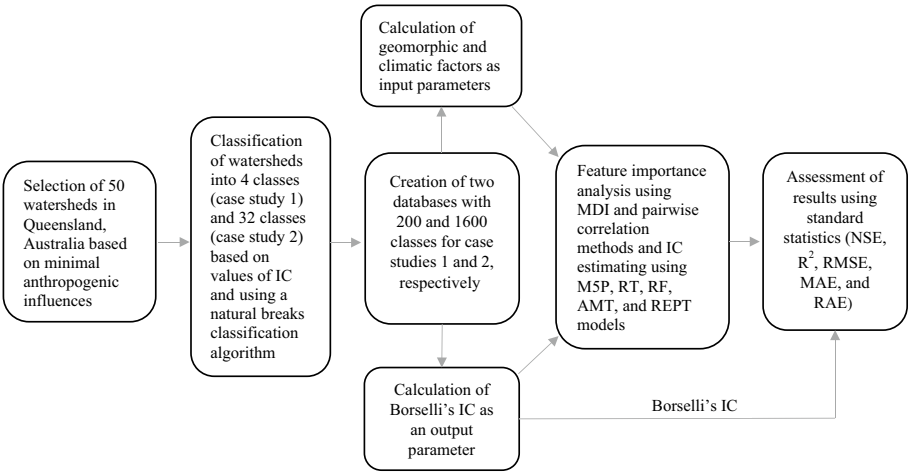

Fig. 4 A basic flowchart of research methodology

Table 2 Statistical parameters of studied variables for the total data in case studies 1 and 2

Case study	Statistical parameter	Variable						
		$ELEV_m$ (m)	SLP_m (%)	Area (km^2)	LS (km)	$NDVI_m$	P_m (mm)	IC_m
1	x_{min}	32.251	2.71	0.095	0.235	0.414	0.233	-2.936
	x_{max}	910.079	42.495	413.371	132.52	0.759	138.75	2.739
	$\bar{x}$	319.006	17.085	89.034	38.859	0.597	45.180	-0.697
	σ_x	207.900	9.131	81.027	23.558	0.098	32.588	1.163
	G_1	1.008	0.393	1.618	1.413	-0.135	0.908	0.338
	β_2	0.726	-0.626	3.021	2.782	-1.291	0.113	-0.483
2	x_{min}	17.863	1.931	0.004	0.00	0.088	-	-3.435
	x_{max}	1093.177	144.298	148.859	285.918	0.857	-	3.828
	$\bar{x}$	318.993	17.895	10.991	5.088	0.595	-	-0.095
	σ_x	209.609	11.791	17.329	9.664	0.102	-	1.723
	G_1	0.992	1.318	2.706	18.434	-0.319	-	0.086
	β_2	0.626	7.961	9.760	490.709	-0.218	-	-1.061

[1]$ELEV_m$ is average elevation, SLP_m is average slope, LS is total length of stream, $NDVI_m$ is average normalized difference vegetation index, P_m is average monthly rainfall, IC_m is average index of sediment connectivity, x_{min} is the minimum value of the data, x_{max} is the maximum value of the data, $\bar{x}$ is the mean of the data, σ_x is the standard deviation, G_1 is the skewness, β_2 is the kurtosis

availability of the required data. Borselli's IC depicts the potential connectivity between different parts of watersheds and is computed in a GIS environment using dynamic (e.g., land use) and static (e.g., topography) attributes. This index is comprised of two components: (1) the upslope component is the downward routing potential of sediment generated from upslope and (2) the downslope component is the flow path length that a particle travels to reach a specified target or sink. Equations for the calculation of this index are expressed in Table 3 (Borselli et al. 2008).

Table 3 Description equations for calculation of IC

No.	Equation	Definition
(1)	$D_{up} = \overline{WS}\sqrt{A}$	D_{up} is the upslope component, $\bar{W}$ is the average weighting factor (dimensionless), $\bar{S}$ is the mean slope gradient (m/m), and A (m^2) is the upslope contributing area.
(2)	$D_{dn} = \sum_{i=1}^{n} \frac{d_i}{W_i S_i}$	D_{dn} is the downslope component, W_i is a dimensionless weighting factor of the i^{th} cell, S_i is the slope gradient of the i^{th} cell (m/m), and d_i (m) is the length of the flow path along the i^{th} cell based on the steepest downslope direction.
(3)	$IC = \log_{10}(\frac{D_{up}}{D_{dn}})$	IC values change in the range of $-\infty$ and $+\infty$, where larger IC values show increasing connectivity.

Construction of IC Factor IC raster maps were prepared via ModelBuilder in an ArcGIS environment (Borselli et al. 2008). The maps used for this model were: (1) raster maps with resolutions of 30 × 30 m of the hydrologically enforced DEM (DEM-H) (Jarihani et al. 2015), and (2) the weighting factor raster maps with resolution 30 × 30 m derived from the cover-management factor of the Universal Soil Loss Equation (USLE)/Revised Universal Soil Loss Equation (RUSLE) (Wischmeier and Smith 1978; Renard 1997). Since the measurement of the cover-management factor by field surveys is difficult, first remotely-sensed land cover maps (i.e., NDVI) were collected and then the cover-management factor maps of RUSLE were developed using the following equation (Durigon et al. 2014):

$$C = \left(\frac{1 - NDVI}{2}\right) \tag{1}$$

where C is the cover-management factor which ranges between 0 and 1; 0 indicates dense vegetation cover and protected soil and 1 indicates unprotected bare soil.

Construction of IC-Influencing Factors The use of relevant and influencing input factors in supervised ML algorithms is important (Asadi et al. 2022). Different factors affect IC, which in this study, elevation, slope, area, length of stream channel, and NDVI were used for model input in both case studies. Also, rainfall was investigated as an input in case study 1. Maps of physical factors, such as slope gradient, elevation, and stream channel length were constructed using SRTM DEM with a resolution 30 m × 30 m using GIS software (ArcGIS 10.6.).

NDVI maps (30 m resolution from 2015 to 2022) were extracted from LANDSAT/LC08/C01/T1_32DAY_NDVI products available at https://explorer.earthengine.google.com/. Monthly gridded rainfall maps (from 2015 to 2022) were collected from SILO (Scientific Information for Land Owners) through the https://silo.longpaddock.qld.gov.au/gridded-data website. In this database, gridded daily climate surfaces derived either from splining or kriging the observational data with a resolution of approximately 5 km × 5 km. Maps of several factors influencing IC, as well as IC for Oyster Creek watershed are represented in Fig. 5.

3.1.3 Features Importance

Feature importance is a widely used analytical method that has been applied in modelling using ML algorithms due to its simplicity and interpretability of feature ranking (Asadi et al. 2023b). One of the effective feature ranking methods is pairwise correlation, which is inspired in the correlation-based feature selection (CFS) method (Jiménez et al. 2021). Also, some of the ML algorithms provide feature importance, for example, the RF algorithm evaluates feature importance based mean decrease in impurity (MDI) (Ali et al. 2021). In our study, two methods (MDI and pairwise correlation) were used for analysis of feature importance in which higher obtained values show higher predictive capability of the factors.

3.1.4 Descriptions of the Models

Five DT-based models were investigated to predict IC in two case studies. These models are briefly introduced as follows:

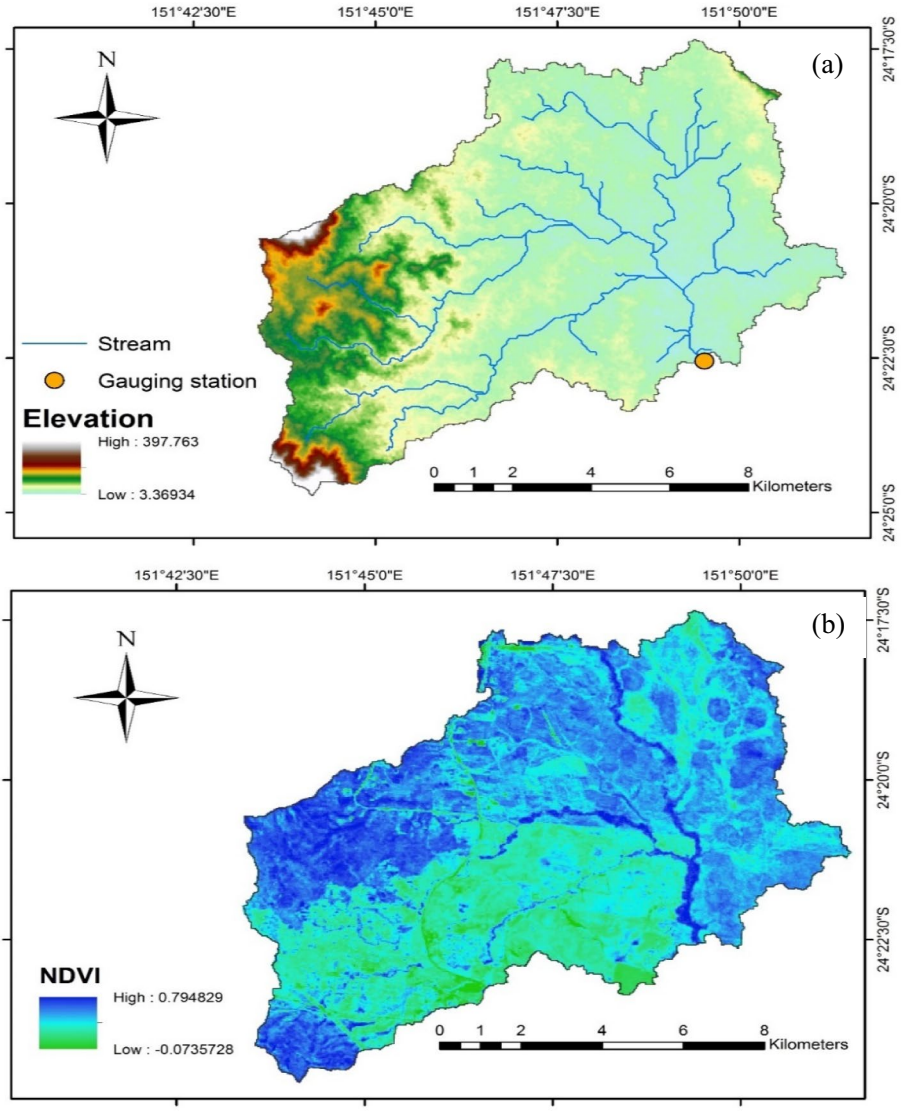

Fig. 5 Maps of some IC-influencing factors, namely elevation and stream length (**a**), NDVI (**b**), and slope (**c**) as well as IC factor map (**d**) for Oyster Creek watershed

Reduced Error Pruning Tree (REPT) The REPT algorithm, which builds a decision or regression tree based on the information gain/variance reduction, is a hybrid of the Reduced Error Pruning (REP) method and the DT method (Quinlan 1987). This algorithm is performed in four steps: (1) creating multiple trees in various iterations (Jayanthi and Sasikala 2013); (2) selecting the best tree from multiple trees; (3) applying the REP technique to avoid over-fitting; and (4) handling missing values using a C4.5 algorithm and sorting the values of numerical attributes. REP is a simple pruning method (Quinlan 1987) that decreases

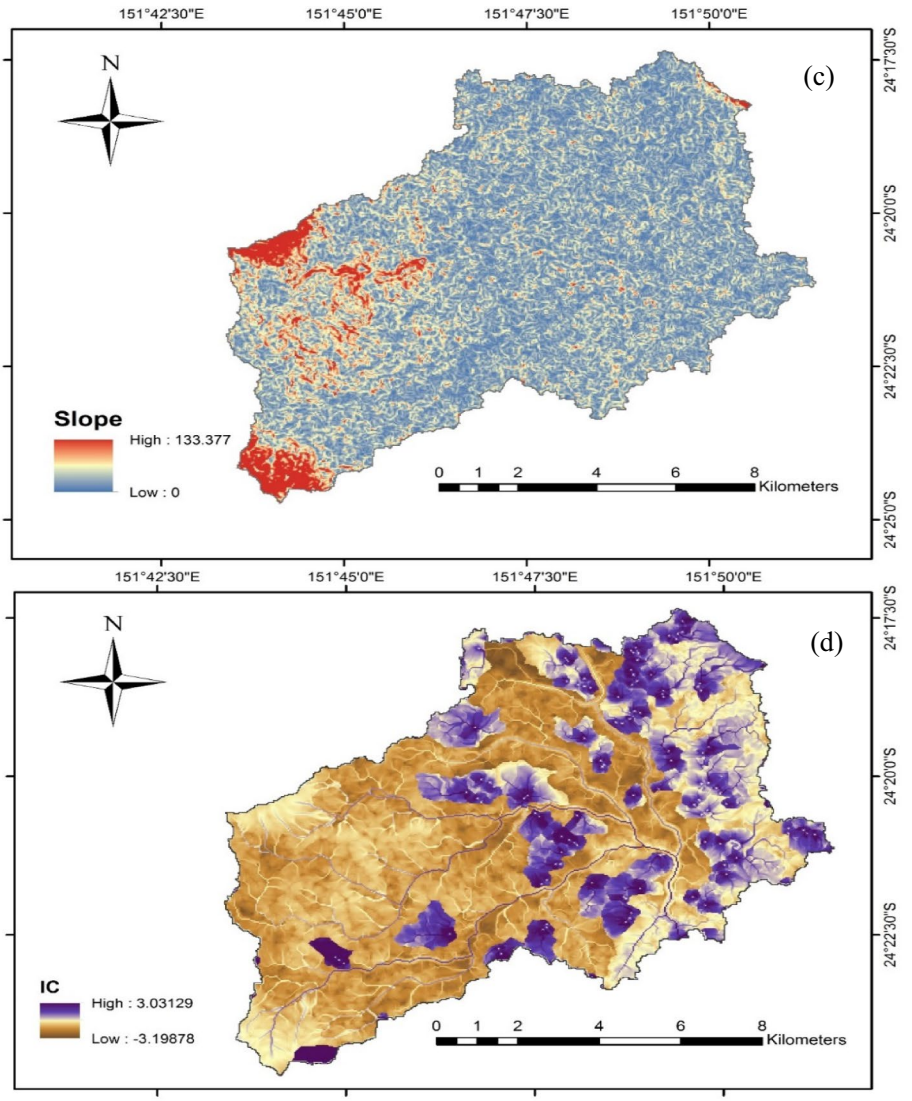

Fig. 5 (continued)

the complexity of the tree structure (Mohamed et al. 2012) by removing some leaves and branches of the tree which provide little power for classification (Galathiya et al. 2012).

Random Forest (RF) RF is a DT-based algorithm (Breiman 2001) that is currently popular and requires few parameters to tune (Senagi and Jouandeau 2022). Several randomized decision trees are combined and a forest of decision trees is produced in which every tree predicts a class and the final decision is achieved by averaging all predictions (Ali et al. 2021). This algorithm is trained in three steps: (1) drawing a bootstrap sample from the

training data; (2) growing a decision tree for each bootstrap sample by selecting the best split among the subset selected randomly from all the features; and (3) repeating these steps until an adequately large number of trees are produced (Mutanga et al. 2012).

M5 Prime (M5P) The M5P algorithm as a DT-based algorithm first introduced by Quinlan (1992). This algorithm has high flexibility (Zhan et al. 2011) and functions in four steps: (1) splitting input spaces and constructing the tree; using the standard deviation reduction (SDR), error reduction is maximized to achieve the best model performance; (2) developing a linear regression model in each of the sub-spaces using the data associated with that sub-space; (3) pruning the tree, starting after the tree is constructed to eliminate undesired sub-trees and the attributes are reduced one by one to minimize estimated error; and (4) smoothing the tree, performed to compensate for sharp discontinuities between adjacent linear models at the leaves of the pruned tree (Wang and Witten 1997).

Random Tree (RT) The RT algorithm (Aldous 1991, 1993) is a fast and flexible learner that uses the decision tree to develop the model and build the decision trees on a random subset of columns (LaValle 1998). The RT is formed by a stochastic process, which has one difference regarding decision trees; only a random subset of attributes is available for each split of the training dataset.

Alternating Model Tree (AMT) AMT (Freund and Mason 1999) is a class of regression trees that contains splitter and prediction nodes (two prediction nodes are generated at each splitter node). This algorithm is performed in three steps: (1) selecting the splitting variable by considering all input variables and splitting on the median value of the data that reaches a particular node; (2) fitting two linear regression models on the subsets resulting from the split and then placing these two models at the prediction nodes attached to the recent splitter node; and (3) to achieve the final prediction, the individual predictions made at each visited prediction node are multiplied by a shrinkage parameter and summed together (the number of splitter nodes and the shrinkage parameter must be set by the user) (Fijani and Khosravi 2023).

3.2 Evaluation Criteria

Six quantitative metrics including coefficient of determination (R^2), root mean squared error (RMSE), mean absolute error (MAE), relative absolute error (RAE), root relative square error (RRSE), and Nash-Sutcliffe efficiency coefficient (NSE) were used for performance analysis of the models in the testing dataset. The equations and the performance classification for these indices are expressed in Table 4.

4 Results and Discussion

4.1 Feature Importance Analysis

The importance of features was estimated using MDI and pairwise correlation methods in both case studies. For a better understanding of feature ranking, the results are visually presented in Fig. 6. In case study 1, feature selection results based on the MDI method indicate that area is the most important factor for IC modelling, followed by slope, length of stream channel, NDVI, elevation, and rainfall, respectively. Based on pairwise correlation,

Table 4 Model evaluation metrics

No.	Equation	Value	Performance classification	References
(1)	$R^2 = \left[\frac{\sum_{i=1}^{n}\left(O_i-\bar{O}\right)\left(E_i-\bar{E}\right)}{\sqrt{\sum_{i=1}^{n}\left(O_i-\bar{O}\right)^2}\sqrt{\sum_{i=1}^{n}\left(E_i-\bar{E}\right)^2}}\right]^2$	$0.70 < R^2 < 1.00$	Very good	Ayele et al. (2017)
		$0.60 < R^2 < 0.70$	Good	
		$0.50 < R^2 < 0.60$	Satisfactory	
		$R^2 < 0.50$	Unsatisfactory	
(2)	$NSE = 1-\frac{\sum_{i=1}^{n}(O_i-E_i)^2}{\sum_{i=1}^{n}\left(O_i-\bar{O}\right)^2}$	$0.75 < NSE \leq 1.00$	Very good	Moriasi et al. (2007)
		$0.65 < NSE \leq 0.75$	Good	
		$0.50 < NSE \leq 0.65$	Satisfactory	
		$NSE \leq 0.50$	Unsatisfactory	
(3)	$MAE = \frac{\sum_{i=1}^{n}\lvert O_i-E_i\rvert}{n}$	MAE, RMSE RAE and RRSE range from 0 to $+\infty$	The lower the RMSE, MAE, RAE and RRSE, the better the model performance	Zounemat-Kermani et al. (2016)
(4)	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(O_i-E_i)^2}$			
(5)	$RAE = \frac{\sum_{i=1}^{n}\lvert O_i-E_i\rvert}{\sum_{i=1}^{n}\lvert O_i-\bar{O}\rvert}$			
(6)	$RRSE = \sqrt{\frac{\sum_{i=1}^{n}(O_i-E_i)^2}{\sum_{i=1}^{n}\left(O_i-\bar{O}\right)^2}}$			

[3] n is the number of data, O_i: is the i^{th} observed IC, E_i is the i^{th} estimated IC, $\bar{O}$ is the average of observed IC, $\bar{E}$ is the average of estimated IC

area is the most important factor, followed by slope, length of stream channel, NDVI, rainfall, and elevation, respectively. In case study 2, feature selection results based on both MDI and pairwise correlation methods show that slope is the most important factor for IC modelling, followed by area, length of stream channel, NDVI, and elevation, respectively.

The two most important features for assessing IC according to the applied techniques are slope and area. The importance of slope is easily justified because with an increase in slope angle, the time for infiltration decreases and runoff increases (Youssef et al. 2015) resulting in higher connectivity. Also, watershed area influences hydrological connectivity (Borselli et al. 2008); an increase in watershed area affects the delivery of sediment to channels and the sediment output from the watershed (De Vente and Poesen 2005). The significance of these factors is consistent with previous results (Asadi et al. 2023a). The length of the stream channel (longitudinal connectivity) ranks third in both case studies because it is an important factor in controlling water flow, connection, and sediment transport. High concentrations of flow into and within stream channels result in more connection and sediment transport (Croke et al. 2005). The next important factor which impacts hydrological processes including sediment connectivity is NDVI (Asadi et al. 2023a). Vegetation density is negatively correlated with runoff generation (Tehrany et al. 2014); areas with high vegetation density and forest cover will reduce surface runoff, and this decrease in runoff causes less internal linkages among sediment sources

Fig. 6 Graphical representation of features importance for case studies 1 and 2

within the watershed. Thus, NDVI affects the resistance of cells against runoff and sediment flow (Cavalli et al. 2013). Our results showed that although elevation had the lowest impact on estimating IC, it does affect runoff as low altitude regions have a higher potential for flow connectivity due to water flowing down from higher elevations. Moreover, watershed altitude often influences the amount and type of precipitation (Garcia-Martino et al. 1996), which affects surface runoff, energy balance, and hydrology of watershed (Ding et al. 2014). Also, rainfall, as a required input in case study 1, was considered since it is highly correlated with runoff generation and is the most important factor for generating runoff. However, the spatiotemporal dynamics of rainfall are complicated (Jahanshahi and Booij 2023) and not easily described in IC modeling. Although runoff occurs when the rainfall intensity exceeds the infiltration capacity of the soil, physical factors such as slope gradient, elevation, soil types, soil moisture, land use patterns, and topography also control the amount of rainfall that can infiltrate into the ground, and hence the amount of rainfall which becomes flow (Jahanshahi et al. 2022).

4.2 Model Evaluation

All the five models were validated using the test dataset in case studies 1 and 2 (Table 5). For case study 1, the RF model had the highest performance based on standard statistical parameters (i.e., NSE, R^2, RMSE, MAE, and RAE), followed by the M5P, AMT, REPT, and RT models, respectively. In case study 2, the RF model also had the highest predictive capability, followed by the REPT, RT, AMT, and M5P models, respectively. A test of model performance based on the NSE metric shows: (1) in case study 1, the RF model has very good performance, the M5P and AMT models have good performance, and the REPT and RT models are unsatisfactory. In case study 2, all techniques have very good performance in determining IC except M5P which has good performance.

The comparison of the results produced by the RF model in both case studies indicated this model performed better in case study 2 compared with case study 1. The RF model accuracy was 20.8% and 19.5% higher based on NSE and R^2, respectively, and the error values were 16.9%, 21.6%, 48.4%, and 43.8% lower based on RMSE, MAE, RAE, and RRSE, respectively, in case study 2 compared to case study 1. Thus, modelling within more homogenous sections of watersheds produces superior results even with fewer factors, consistent with the previous results (Asadi et al. 2023a). To evaluate the performance of the DT-based algorithms, scatter plots of test data are shown for all models in both case studies (Fig. 7).

Table 5 Performance of 5 models for prediction of IC in case studies 1 and 2

Case study	Model	RMSE	MAE	RAE	RRSE	NSE	R^2
1	**RT**	0.814	0.627	0.640	0.696	0.398	0.562
	AMT	0.662	0.500	0.511	0.566	0.667	0.701
	REPT	0.782	0.602	0.614	0.668	0.414	0.567
	M5P	0.593	0.485	0.497	0.508	0.745	0.749
	RF	0.546	0.435	0.446	0.468	0.771	0.780
2	**RT**	0.640	0.466	0.314	0.371	0.862	0.865
	AMT	0.702	0.474	0.319	0.407	0.835	0.835
	REPT	0.639	0.447	0.301	0.371	0.863	0.863
	M5P	0.933	0.724	0.488	0.541	0.710	0.723
	RF	0.454	0.341	0.230	0.263	0.931	0.932

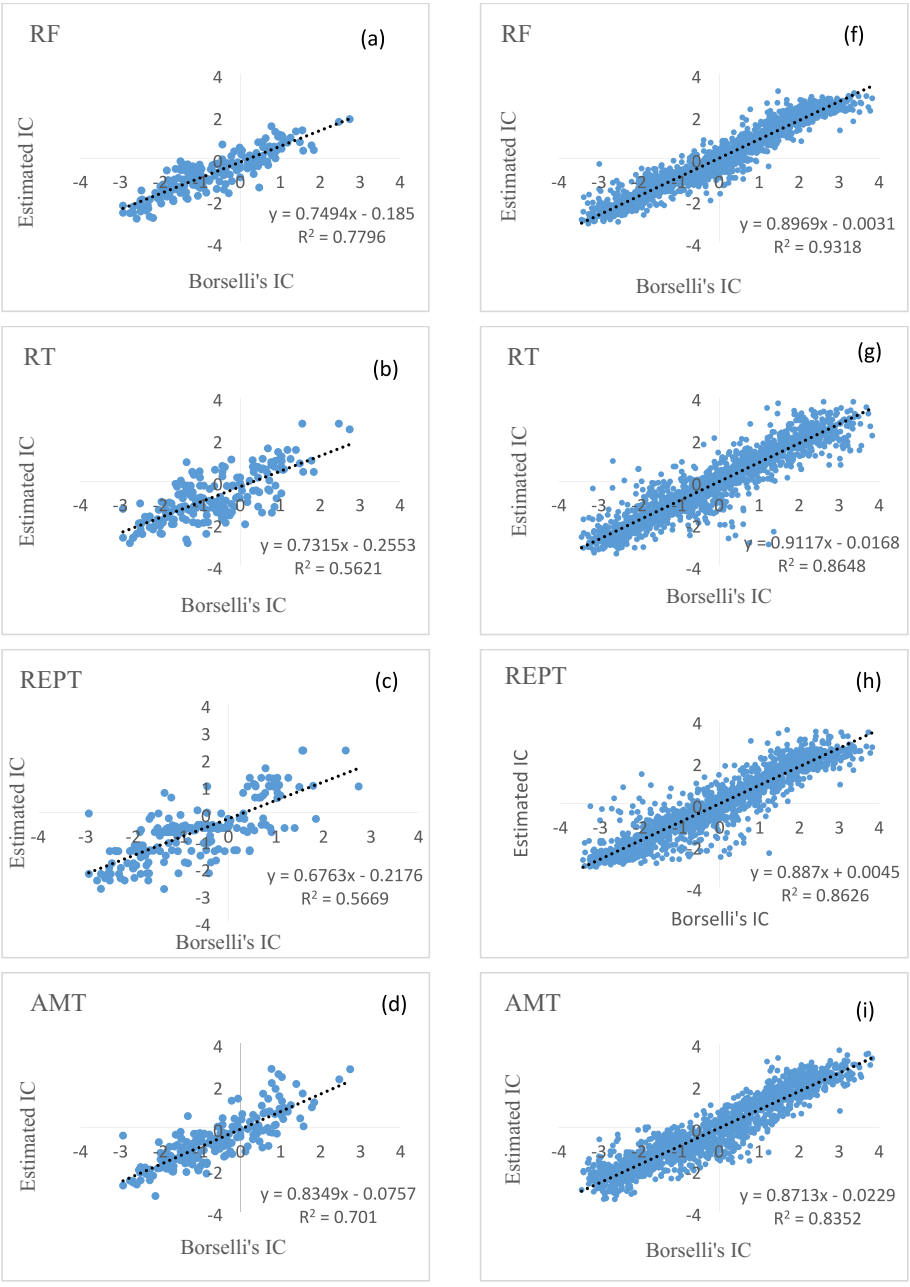

Fig. 7 Scatter plots of the Borselli's IC and estimated IC for all models in case study 1 (a, b, c, d, and e) and case study 2 (f, g, h, i, and j)

In both case studies, comparison of plots indicates that the best agreement between Borselli's IC and estimated IC was obtained using the RF model. Also, estimated IC values (in all models except M5P) were in closer agreement with Borselli's IC values in case study

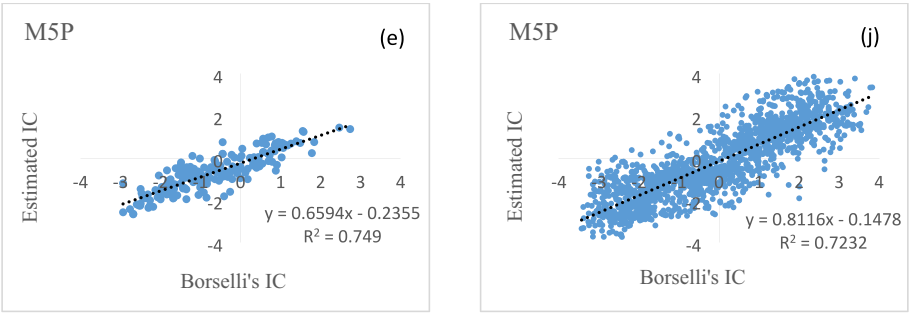

Fig. 7 (continued)

2 compared to case study 1. Therefore, the accuracy of modelling at small measurement scales of watersheds was higher than that at larger scales. Since water and sediment transfers change according to the scale at which they are observed, hence, the concept of connectivity and measurement scales are related (Cammeraat 2002).

Also, our investigation showed that some estimated values were not accurate, which often occurs in modeling (Sichingabula 1998). The reasons may be that the IC can be influenced by different factors including hydrological variables (e.g., soil moisture) and physical characteristics of watershed (e.g., soil types), which were not used in our IC estimators. Additionally, geomorphic and hydrologic perturbations in watershed, including landslides and floods can affect water and sediment fluxes (Ziegler et al. 2014) as well as the degree of connectivity among different parts of a watershed. Moreover, the use of the other ML models, high quality data, and the data with high temporal and spatial resolution may improve results (Asadi et al. 2022).

Results of our models (M5P, RT, RF, AMT and REPT) were compared with the results of Borselli's model. It is noteworthy that calculate values of IC in Borselli's model may not be completely accurate due to limitations and uncertainties in computation (Heckmann and Vericat 2018). Thus, the lack of comparison of model results with geomorphologic and sediment field observations is the main limitation of our study; this issue is the focus of future studies. The results obtained from this study indicate that the RF model can be applied as a reliable method for estimating IC. Advantages of this model are its accuracy, ability to deal with small sample sizes, and relatively few parameters required to tune (Biau and Scornet 2016).

Overall, the inability of the conventional methods (e.g., GIS-based models) to simultaneously apply different influencing factors (e.g., geomorphological, meteorological, and hydrological variables) in estimation of IC is the main limitation of these methods, while in ML-based approaches, the ability to use different factors regardless of physical processes can create a more flexible, comprehensive, and efficient tool for use by management experts and the personnel involved quantitative assessments of sediment connectivity. It is noteworthy that more detailed studies should investigate the potential of this approach with various IC-influencing factors in different watersheds.

5 Conclusion

The accurate estimation of IC is a prerequisite for understanding the linkages amongst various parts of watershed with different land use and topographic features. IC is not only suitable to characterize sediment dynamics within the watershed, but also to help

geomorphological interpretation of the watershed. Accordingly, the availability of easily applicable methods for estimation of IC promotes potential for implementing this concept as a management tool. Due to the non-linear and complex behavior of sediment transport within watersheds, DT-based machine learning algorithms have the potential to accurately estimate IC. Our study tested this potential for the first time by examining the prediction power of M5 prime (M5P), random tree (RT), random forest (RF), alternating model tree (AMT), and reduced error pruning tree (REPT) models. Findings revealed that in case study 1, the RF and M5P models were successful in assessing IC, while in case study 2, all techniques were successful. Among these methods, the RF model had the highest prediction power in both case studies. Moreover, the results indicate that the three most significant geomorphic features according to feature importance and correlation value were slope, area, and length of stream channel in both case studies. Generally, findings provide a relatively inexpensive and efficient ML-based approach for rapid prediction of IC that can be used particularly in developing countries where the lack of adequate technical skills, equipment, and budget are serious constraints.

Author Contributions Haniyeh Asadi: Conceptualization, Methodology, Investigation, Formal analysis, Writing – original draft. Mohammad T. Dastorani: Methodology, Supervision, Writing – review & editing. Roy C. Sidle: Methodology, Writing – review & editing. Afshin Jahanshahi: Formal analysis, Writing – review & editing.

Funding This work was supported by the Ferdowsi University of Mashhad (grant number FUM-64635).

Data Availability The corresponding author can provide the data that back up the study's conclusions upon request.

Declarations

Ethical Approval All authors read and approved the final manuscript.

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

References

Aghamolaei Z, Hessami-Kermani M-R (2023) Developing a new artificial intelligence framework to estimate the thalweg of rivers. Water Resour Manag 1–25. https://doi.org/10.1007/s11269-023-03632-8

Al-Mukhtar M (2019) Random forest, support vector machine, and neural networks to modelling suspended sediment in Tigris River-Baghdad. Environ Monit Assess 191:673

Aldous D (1993) The continuum random tree III. Ann Probab 248–289

Aldous D (1991) The continuum random tree. II. An overview. Stoch Anal 167:23–70

Ali MM, Paul BK, Ahmed K et al (2021) Heart disease prediction using supervised machine learning algorithms: performance analysis and comparison. Comput Biol Med 136

Asadi H, Dastorani MT, Khosravi K, Sidle RC (2022) Applying the C-Factor of the RUSLE Model to improve the prediction of suspended sediment concentration using Smart Data-Driven models. Water 14

Asadi H, Dastorani MT, Sidle RC (2023a) Estimating index of sediment connectivity using a smart data-driven model. J Hydrol 620

Asadi S, Tartibian B, Moni MA (2023b) Determination of optimum intensity and duration of exercise based on the immune system response using a machine-learning model. Sci Rep 13:8207

Asadi H, Dastorani MT, Sidle RC, Shahedi K (2021) Improving Flow Discharge-suspended sediment relations: Intelligent algorithms versus data separation. Water 13

Ayele GT, Teshale EZ, Yu B, Rutherfurd ID, Jeong J (2017) Stream flow and sediment yield prediction for watershed prioritization in the Upper Blue Nile River Basin. Ethiopia. Water 9(782). https://doi.org/10.3390/w9100782

Biau G, Scornet E (2016) A random forest guided tour. TEST 25:197–227
Borselli L, Cassi P, Torri D (2008) Prolegomena to sediment and flow connectivity in the landscape: a GIS and field numerical assessment. Catena 75:268–277
Bracken LJ, Turnbull L, Wainwright J, Bogaart P (2015) Sediment connectivity: a framework for understanding sediment transfer at multiple scales. Earth Surf Process Landforms 40:177–188
Breiman L (2001) Random forests. Mach Learn 45:5–32
Bui DT, Khosravi K, Tiefenbacher J et al (2020) Improving prediction of water quality indices using novel hybrid machine-learning algorithms. Sci Total Environ 721
Cammeraat LH (2002) A review of two strongly contrasting geomorphological systems within the context of scale. Earth Surf Process Landforms 27:1201–1222
Cavalli M, Tarolli P, Dalla Fontana G, Marchi L (2016) Multi-temporal analysis of sediment source areas and sediment connectivity in the Rio Cordon catchment (Dolomites). Rend Online Soc Geol Ital 39:27–30
Cavalli M, Trevisani S, Comiti F, Marchi L (2013) Geomorphometric assessment of spatial sediment connectivity in small Alpine catchments. Geomorphology 188:31–41
Chia MY, Koo CH, Huang YF, Di Chan W, Pang JY (2023) Artificial intelligence generated synthetic datasets as the remedy for data scarcity in water quality index estimation. Water Resour Manag 37(15):6183–6198
Croke J, Fryirs K, Thompson C (2013) Channel–floodplain connectivity during an extreme flood event: implications for sediment erosion, deposition, and delivery. Earth Surf Process Landforms 38:1444–1456
Croke J, Mockler S, Fogarty P, Takken I (2005) Sediment concentration changes in runoff pathways from a forest road network and the resultant spatial pattern of catchment connectivity. Geomorphology 68:257–268
De Vente J, Poesen J (2005) Predicting soil erosion and sediment yield at the basin scale: scale issues and semi-quantitative models. Earth Sci Rev 71:95–125
Ding B, Yang K, Qin J et al (2014) The dependence of precipitation types on surface elevation and meteorological conditions and its parameterization. J Hydrol 513:154–163
Durigon VL, Carvalho DF, Antunes MAH et al (2014) NDVI time series for monitoring RUSLE cover management factor in a tropical watershed. Int J Remote Sens 35:441–453
Fijani E, Khosravi K (2023) Hybrid iterative and tree-based machine learning algorithms for lake water level forecasting. Water Resour Manag 37(14):5431–5457
Freund Y, Mason L (1999) The alternating decision tree learning algorithm. In icml 99:124–133
Fressard M, Cossart E (2019) A graph theory tool for assessing structural sediment connectivity: development and application in the Mercurey vineyards (France). Sci Total Environ 651:2566–2584
Fryirs K (2013) Dis) Connectivity in catchment sediment cascades: a fresh look at the sediment delivery problem. Earth Surf Process Landforms 38:30–46
Galathiya AS, Ganatra AP, Bhensdadia CK (2012) Improved decision tree induction algorithm with feature selection, cross validation, model complexity and reduced error pruning. Int J Comput Sci Inf Technol 3:3427–3431
Garcia-Martino AR, Warner GS, Scatena FN, Civco DL (1996) Rainfall, runoff and elevation relationships in the Luquillo Mountains of Puerto Rico. Caribb J Sci 32:413–424
Gay A, Cerdan O, Mardhel V, Desmet M (2016) Application of an index of sediment connectivity in a lowland area. J Soils Sediments 16:280–293
Gelete G (2023) Hybrid extreme gradient boosting and nonlinear ensemble models for suspended sediment load prediction in an agricultural catchment. Water Resour Manage Please provide complete bibliographic details of this reference. Water Resour Manag 1–29. https://doi.org/10.1007/s11269-023-03629-3
Ghorbani MA, Deo RC, Kim S et al (2020) Development and evaluation of the cascade correlation neural network and the random forest models for river stage and river flow prediction in Australia. Soft Comput 24:12079–12090
González-Romero J, López-Vicente M, Gómez-Sánchez E et al (2021) Post-fire management effects on sediment (dis) connectivity in Mediterranean forest ecosystems: Channel and catchment response. Earth Surf Process Landforms 46:2710–2727
Heckmann T, Schwanghart W (2013) Geomorphic coupling and sediment connectivity in an alpine catchment—exploring sediment cascades using graph theory. Geomorphology 182:89–103
Heckmann T, Vericat D (2018) Computing spatially distributed sediment delivery ratios: inferring functional sediment connectivity from repeat high-resolution digital elevation models. Earth Surf Process Landforms 43:1547–1554
Heddam S, Kisi O (2018) Modelling daily dissolved oxygen concentration using least square support vector machine, multivariate adaptive regression splines and M5 model tree. J Hydrol 559:499–509

Jahanshahi A, Booij MJ (2023) Exploring controls on rainfall–runoff events: spatial dynamics of event runoff coefficients in Iran. Hydrol Sci J 68:954–966
Jahanshahi A, Ghazanchaei Z, Navari M et al (2022) Dependence of rainfall-runoff model transferability on climate conditions in Iran. Hydrol Sci J 67:564–587
Jarihani AA, Callow JN, McVicar TR et al (2015) Satellite-derived Digital Elevation Model (DEM) selection, preparation and correction for hydrodynamic modelling in large, low-gradient and data-sparse catchments. J Hydrol 524:489–506
Jayanthi SK, Sasikala S (2013) Reptree classifier for identifying link spam in web search engines. IJSC 3:498–505
Jiménez F, Sánchez G, Palma J et al (2021) Multivariate feature ranking of gene expression data. arXiv Prepr arXiv211102357
Kalantari Z, Ferreira CSS, Koutsouris AJ et al (2019) Assessing flood probability for transportation infrastructure based on catchment characteristics, sediment connectivity and remotely sensed soil moisture. Sci Total Environ 661:393–406
Keesstra S, Nunes JP, Saco P et al (2018) The way forward: can connectivity be useful to design better measuring and modelling schemes for water and sediment dynamics? Sci Total Environ 644:1557–1572
Khosravi K, Cooper JR, Daggupati P et al (2020) Bedload transport rate prediction: application of novel hybrid data mining techniques. J Hydrol 585
Khozani ZS, Khosravi K, Pham BT et al (2019) Determination of compound channel apparent shear stress: application of novel data mining models. J Hydroinformatics 21:798–811
Koci J, Sidle RC, Jarihani B, Cashman MJ (2020) Linking hydrological connectivity to gully erosion in savanna rangelands tributary to the great barrier reef using structure-from-motion photogrammetry. L Degrad Dev 31:20–36
LaValle S (1998) Rapidly-exploring random trees: a new tool for path planning. Res Rep 9811
López-Vicente M, González-Romero J, Lucas-Borja ME (2020) Forest fire effects on sediment connectivity in headwater sub-catchments: evaluation of indices performance. Sci Total Environ 732
Marchamalo M, Hooke JM, Sandercock PJ (2016) Flow and sediment connectivity in semi-arid landscapes in SE Spain: patterns and controls. L Degrad Dev 27:1032–1044
Martini L, Baggio T, Torresani L et al (2022) R_IC: a novel and versatile implementation of the index of connectivity in R. Environ Model Softw 155
Masselink RJH, Keesstra SD, Temme AJAM et al (2016) Modelling discharge and sediment yield at catchment scale using connectivity components. L Degrad Dev 27:933–945
Mohamed WNHW, Salleh MNM, Omar AH (2012) A comparative study of reduced error pruning method in decision tree algorithms. In: 2012 IEEE International conference on control system, computing and engineering. IEEE, pp 392–397
Moriasi DN, Arnold JG, Van Liew MW, Bingner RL, Harmel RD, Veith TL (2007) Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Trans. ASABE 50:885–900
Mutanga O, Adam E, Cho MA (2012) High density biomass estimation for wetland vegetation using WorldView-2 imagery and random forest regression algorithm. Int J Appl Earth Obs Geoinf 18:399–406
Najafi S, Sadeghi SH, Heckmann T (2021) Analysis of sediment accessibility and availability concepts based on sediment connectivity throughout a watershed. L Degrad Dev 32:3023–3044
Parsons AJ, Bracken L, Poeppl RE et al (2015) Introduction to special issue on connectivity in water and sediment dynamics. Earth Surf Process Landforms 40:1275–1277
Phillips JD, Schwanghart W, Heckmann T (2015) Graph theory in the geosciences. Earth Sci Rev 143:147–160
Quinlan JR (1992) Learning with continuous classes. In: 5th Australian joint conference on artificial intelligence. World Scientific pp 343–348
Quinlan JR (1987) Simplifying decision trees. Int J Man Mach Stud 27:221–234
Renard KG (1997) Predicting soil erosion by water: a guide to conservation planning with the Revised Universal Soil Loss Equation (RUSLE). United States Government Printing
Senagi K, Jouandeau N (2022) Parallel construction of Random Forest on GPU. J Supercomput 78:10480–10500
Sichingabula HM (1998) Factors controlling variations in suspended sediment concentration for single-valued sediment rating curves, Fraser River, British Columbia, Canada. Hydrol Process 12:1869–1894
Tehrany MS, Pradhan B, Jebur MN (2013) Spatial prediction of flood susceptible areas using rule based decision tree (DT) and a novel ensemble bivariate and multivariate statistical models in GIS. J Hydrol 504:69–79

Tiranti D, Crema S, Cavalli M, Deangeli C (2018) An integrated study to evaluate debris flow hazard in alpine environment. Front Earth Sci 6:60

Tehrany MS, Pradhan B, Jebur MN (2014) Flood susceptibility mapping using a novel ensemble weights-of-evidence and support vector machine models in GIS. J Hydrol 512:332–343

Turnbull L, Hütt M-T, Ioannides AA et al (2018) Connectivity and complex systems: learning from a multi-disciplinary perspective. Appl Netw Sci 3:1–49

Wang Y, Witten IH (1997) April. Inducing model trees for continuous classes. In: Proceedings of the Ninth European Conference on Machine Learning 128–137

Wischmeier WH, Smith DD (1978) Predicting rainfall erosion losses: a guide to conservation planning. Department of Agriculture, Science and Education Administration

Wohl E, Brierley G, Cadol D et al (2019) Connectivity as an emergent property of geomorphic systems. Earth Surf Process Landforms 44:4–26

Yousefi S, Pourghasemi HR, Avand M et al (2021) Assessment of land degradation using machine-learning techniques: a case of declining rangelands. L Degrad Dev 32:1452–1466

Youssef AM, Pradhan B, Pourghasemi HR, Abdullahi S (2015) Landslide susceptibility assessment at Wadi Jawrah Basin, Jizan region, Saudi Arabia using two bivariate models in GIS. Geosci J 19:449–469

Zhan C, Gan A, Hadi M (2011) Prediction of lane clearance time of freeway incidents using the M5P tree algorithm. IEEE Trans Intell Transp Syst 12:1549–1557

Zhao C, Liu C, Li W, Tang Y, Yang F, Xu Y, ... Hu C (2023) Simulation of urban flood process based on a hybrid LSTM-SWMM model. Water Resour Manag 37(13):5171–5187

Ziegler AD, Benner SG, Tantasirin C et al (2014) Turbidity-based sediment monitoring in northern Thailand: hysteresis, variability, and uncertainty. J Hydrol 519:2020–2039

Zounemat-Kermani M, Kis¸i ¨O, Adamowski J, Ramezani-Charmahineh A (2016) Evaluation of data driven models for river suspended sediment concentration modeling. J Hydrol 535:457–472