Journal of Hydroinformatics

Journal of Hydroinformatics Vol 26 No 3, 626 doi: 10.2166/hydro.2024.242

Estimating critical depth and discharge over sloping rough end depth using machine learning

Ahmed Y. Mohammed [a,*] and Parveen Sihag [b]

[a] Dams and Water Resources Engineering Department, College of Engineering, University of Mosul, Mosul, Iraq
[b] Civil Engineering Department, Chandigarh University, Mohali, Punjab, India
*Corresponding author. E-mail: ahmedymaltaee@gmail.com; a.altaee@uomosul.edu.iq

AYM, 0000-0002-2274-9202; PS, 0000-0002-7761-0603

ABSTRACT

This study uses machine learning (ML) to predict the end-depth structure's discharge and critical depth (y_c). Linear regression, M5P, random forest, random tree, reduced error pruning tree, and Gaussian process (GP) are the ML methods used in this investigation. The findings indicate that the radial kernel function-based GP model is most suitable compared to other applied models with the lowest root-mean-square error = 0.0021, 0.007, normalized root-mean-square error = 0.0361, 0.0516 representing mean absolute error = 0.0015, 0.004 and the highest coefficient of correlation = 0.9912, 0.9916, Legates and McCabe's index = 0.8839, 0.9026 Willmott's index = 0.9956, 0.9956, and Nash Sutcliffe model efficiency = 0.9823, 09830 for yc for the end-depth structure (y_c) and discharge (Q) with the testing stage, respectively. Results of the sensitivity study indicate that the friction coefficient is the most significant input variable compared to other parameters for predicting (y_c) and flow running via the thickness model's last stage (Q) using this dataset.

Key words: bed roughness, brink depth, broad-crested weir, critical depth, end-depth, flow discharge, free over-fall, machine learning

HIGHLIGHTS

- The abrupt reduction in the channel bed level is referred to as free overfall.
- Free overfall is used to estimate the discharge flowing via open channels.
- The discharge and critical depth for end depth structures are predicted using machine learning (ML).
- Linear regression, M5P, random forest, random tree, reduced error pruning tree, and Gaussian process (GP) are the ML methods used.
- Radial kernel function-based GP model (GP_RBF) is the most suitable as compared to other applied models.

1. INTRODUCTION

The end-depth structure called free overfall happens when water drops from up to down as a free fall because of a sudden drop or sudden change in the channel bed level caused by a free flow of water. Because of this drop, the brink's pressure distribution is not hydrostatic at the subcritical flow. The flow changed from Gradually Varied Flow (GVF) to Rappelled Varied Flow (RVF) GVF to RVF. The relation between depth over the threshold (brink depth, y_b) and standard depth (y_n), end-depth ratio (EDR), is more critical to predict Q over this structure because it is used as a flow measurement device.

Many investigations dealt with hydraulic characteristics of end-depth structures (Mohammed *et al.* 2007), which have presented a variation of water depth on vertical and skewed free overfall. The results clarified that a tilted outflow free fall is vertically superior, and the depth at the brink for vertical is greater than 11% than skew. The impact of stream channel slope on vertical and inclined free fall was investigated by Mohammed (2009b). The investigation showed that the bed slope affected discharge and the water depth of free overfall. The discharge coefficient for an inclined model is more significant than the vertical 25%. The authors present a theoretical equation to find EDR, water surface profile, and discharge. Mohammed (2009a) investigated a new model of an end-depth structure. He studied triangular shapes with an opposite flow direction with different angles. He compared the results with a standard vertical model. The results showed that the brink depth was more significant than 6% for triangular-shaped compared with vertical. Mohammed (2012) presented a theoretical study to predict EDR and end-depth discharge (EDD) for end-depth models with different end shapes. The

study presents a new empirical equation to calculate EDR and EDD with a percentage error not exceeding 8.5% compared with the experimental data.

There are many studies conducted on channels with different shapes (Dey & Kumar 2002; Irzooki & Hasan 2018; Muhsin & Noori 2021), which investigated the free overfall model in a triangular channel. They focused on the effect of side and bed slopes and bed roughness. The study refers to EDR = 0.695 and 0.755, respectively, and this value increases when the side slope increases. Dey *et al.* (2004), Raikar *et al.* (2004), and Zeidan *et al.* (2021) investigated free overfall in an inverted semi-circular channel. These studies developed new relationships to calculate discharge as well as EDR of 0.81. There are many studies in which free overfall with bed roughness is investigated (Öztürk 2005; Guo *et al.* 2006; Mohammed *et al.* 2011, 2013, 2018; FIRAT 2015). These investigation results referred to EDR values reaching 0.67, the brink-to-y_c ratio increased when the channel slope decreased, and roughness material and distribution increased.

In the last several decades, several investigators have turned to the use of machine learning (ML) algorithms for the analysis of field and laboratory data and have discovered noticeably superior results than those obtained using traditional statistical approaches (Olyaie *et al.* 2019; Yousif *et al.* 2019; Suntaranont *et al.* 2020; Salmasi *et al.* 2021; Thakur *et al.* 2021). Researchers have recently focused on the Gaussian process (GP), random forest (RF), random tree (RT), reduced error pruning (REP) tree, and RF in predicting hydraulic features (Sihag *et al.* 2019, 2020; Salmasi *et al.* 2021). The current work compares the outcomes with models based on linear regression (LR). It introduces M5P, RF, RT, REP tree, and GP as alternative methods for determining y_c for the end-depth structure and Q. Gharehbaghi *et al.* (2023) investigated the influence of a submerged multiple-vane system on the dimensions of the flow separation zone. Several data-driven models are accessible, such as gene expression programming, support vector regression (SVR), Radial Basis Function (RBF), and a robust hybrid SVR with an ant colony optimization algorithm (ACO). Based on statistical metrics, the model grading procedure, scatter plot, and the hybrid SVR (RBF)-ACO model are the most accurate and exact models for predicting the maximum relative length and width. The total grades for these models are 6.75 and 5.8, respectively. The ML method was presented by Latif & Ahmed (2023) to predict the reservoir inflow in the Dokan dam in Iraq and the Warragamba dam in Australia using SVR. The RMSE for the Dokan dam daily inflow is 145.7 and R^2 is 0.85. The findings indicated that the ML had strong performance in Iraq, but its accuracy in Australia was lacking.

2. EXPERIMENTAL METHODOLOGY

The experimental work was done in the Hydraulic Laboratory of Mosul University using a rectangular channel of 10 m long, 450 mm height, and 300 mm wide. The discharge was measured using a rectangular sharp-crested weir by the volumetric method. Five different discharges were used between 5.7 and 20.9 l/s. The end-depth model was made of wood of 300 mm wide, 150 mm height, and 1 m long with an upstream face slope of 10° to ensure the uniform flow over the end-depth model. The bed roughness was made using three different size materials and styles. The first style of roughness uses a 10 mm wide cylindrical wood shape fixed at three rows with a 100 mm distance between each row, one in a straight line and one in a zigzag line, and two rows with a 200 mm distance between each row (Figure 1).

The second and third styles of roughness use 6 mm crushed stone and 2 mm glacier stone, respectively, fixed at three rows at 100 mm distance between each row at a straight line, two rows at 200 mm distance between each row and at the straight line too, as well as total bed roughness fixed at an area of 200 × 300 mm^2 (Figure 2).

There are three channel slopes, 1/100 and 1/200, and leveling cases (0°). The water level was measured using a point gauge with 0.1 mm accuracy for each experiment of the aforementioned instances, estimated standard depth, brink depth, and typical depth above the classic weir. There are 215 runs, 135 for rough, and 80 for smooth models.

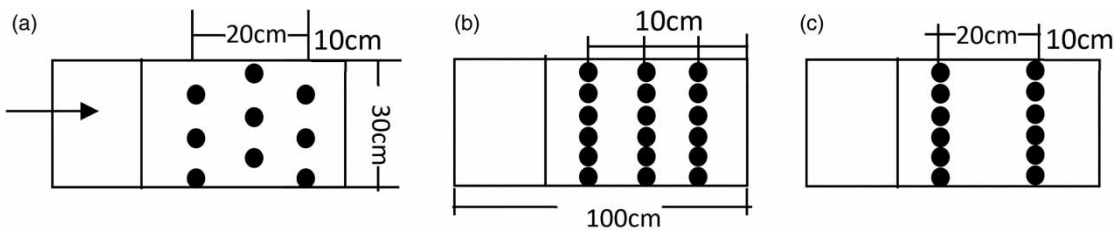

Figure 1 | 10 mm wood roughness style: (a) zigzag, (b) three rows, and (c) two rows.

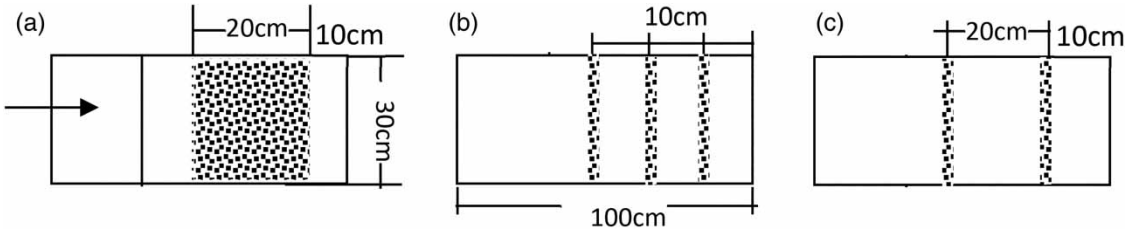

Figure 2 | 2 mm glacier stone and 6 mm crushed stone roughness style: (a) complete, (b) three rows, and (c) two rows.

2.1. Experimental procedures

The standard measurement weir is installed at the channel end until the water head is constant; then, the water level above the weir and measured water volume to time (head-discharge) measurement and used Equation (1) for actual discharge calculated then raise the weir and continue all sizes to ensure not influence on free overfall model, then repeat these procedures for other discharges.

3. THEORETICAL METHODOLOGY

According to experimental works, the actual discharges are calculated using a volumetric method by measuring the volume of water at a specific time and water head above the standard weir, and with the data collected, the following equation predicts the actual Q:

$$Q_{\text{act}} = 0.714 H_w^{1.5} \tag{1}$$

where Q_{act} is the actual discharge in l/s and H_w is the water depth above the standard weir in cm.

The flow over end-depth can be assumed to be similar to the flow over the weir, considering the weir depth ($p = 0$) and head over the weir as normal. The velocity can be calculated at the brink point by applying the Bernoulli equation at the normal depth and brink sections. The Q over end-depth can be measured using the following equation:

$$Q = \int_0^{y_n} C_c b \sqrt{2g(H-z)}\,dz = \frac{2b\sqrt{2g}}{3} C_c [H^{3/2} - (H - y_n)^{3/2}] \tag{2}$$

where y_n and V_n are uniform (normal) depth and velocity, respectively; H is $y_n + {V_n}^2/2g$; g is the gravitational acceleration; z is the vertical distance from a reference level; b is the width of the channel; and C_c is the contraction coefficient.

Equation (2) can be written as follows:

$$\frac{Q}{b g^{1/2} y_n^{3/2}} = \frac{2\sqrt{2g}}{3 g^{1/2} y_n^{3/2}} \frac{y_b}{y_n} [H^{3/2} - (H - y_n)^{3/2}] \tag{3}$$

The discharge per unit length (q) can be calculated using the following equation:

$$q = \frac{Q_{\text{act}}}{b} \tag{4}$$

The critical depth can be measured using the following equation:

$$y_c = \sqrt[3]{q^2/g} \tag{5}$$

4. REVIEW OF REGRESSION AND SOFT COMPUTING TECHNIQUES

4.1. Linear regression

LR is a straightforward form of a mathematical equation. That equation developed a relationship between independent factors and outcome factors. In this study, XLSTAT software is used to formulate LR-based equations. The basic concept behind this is a minor square technique. The general equation of the LR model is shown in Equation (6).

$$D = p_1 a_1 + p_2 a_2 + p_3 a_3 + \quad \ldots\ldots \quad + p_n a_n + z \tag{6}$$

where D is a dependent variable – $a_1, a_2, a_3, \ldots.., a_n$. They are independent variables – $p_1, p_2, p_3, \ldots..$ and p_n They are coefficients, and z is constant in the developed equation.

LR-based techniques for evaluating and forecasting y_c also have been established in the current study. This formula is created using XLSTAT software using the least square approach. The regression-based models lead to the following linear equation:

$$y_c = -0.00724 + 0.2682\, So + 0.5423\, y_b + 0.6437\, y_n - 0.5263\, k \tag{7}$$

LR-based methodologies for estimating the actual flow Q (m^3/s) have also been established in this study. This formula is created using XLSTAT software using the least square approach. The regression-based model's linear formula is as follows:

$$Q = -0.00842 + 0.0687\, S_o + 0.1309\, y_b + 0.2410\, y_n - 0.2024\, k \tag{8}$$

4.2. M5P model

M5P-tree is a regression problem-solving genetic algorithm initially proposed by Quinlan Basser (1992). This tree approach establishes LR attributes on the station node by categorizing or splitting distinct data sections into numerous spaces. It suits a multiple LR model on each sublocation. The M5P-tree technique is based on continuous class issues rather than discontinuous slices, and it can handle functions with many dimensions. It displays the data for each built-in linear model component, making it possible to evaluate the nonlinear relationship between the datasets. Fault evaluation is conducted with the knowledge of the M5P-tree model tree separation criteria per node. The preset value difference of the class entering the node determines the number of errors. Any node solution is obtained using the feature that optimizes the predicted error reduction. Depending on fault estimates per node, details on the M5P-tree model tree splitting criteria are presented. The standard deviation (SD) of the target class at the node is used to compute the M5P error – overfitting results from this division, creating a large tree-like structure. The huge tree is trimmed in the second step, and the chopped subtrees are replaced with LR functions.

4.3. Random forest

The RF theory was established by Breiman (2001). A regression and categorization machine technique builds a collection of tree alternatives at random and predicts the class, which may be the categorization style and separate trees' regressions. To develop a tree, the RF has an assembly of input values at each node (Singh *et al.* 2017). A decision tree is vital in a packing-based RF classifier, a sampling strategy in which the identical specimens are utilized several times and then re-inserted into the database. It is a variation of the bootstrap grouping tree approach known as bagging if only sample bootstrapping is utilized for categorized or regressed and no sampling prediction is made.

An irregular forest-trained model is often used to build models because of its ease of use and excellent performance, even with small datasets. RFs were extensively employed in transportation research. Thakur *et al.* (2021) employed the RF model to forecast the bond strength of Fiber-reinforced plastic (FRP) bars and obtained good results.

4.4. Random tree

The RT method is a regulated teaching technique that creates several separate learners. The collected data are irregularly formed after trying to bag, termed a group teaching algorithm, and every node of an irregular tree is separated. The optimum among the subgroup of predictors is chosen randomly at that node (Aldous 1991). The random trees blend two current ML techniques: binary class trees and RF concepts. Each tree-like bagging is evaluated and regenerated with the training data.

All of the items in the subgroup were irregularly evaluated in the second phase at every node. Following this, the most suitable separation for the subgroup was calculated. For complicated and nonlinear relationships, RT-categorized data and

analytics methods can be utilized (Shi *et al.* 2020). For the investigation, classification and regression approaches were applied. In the clustering algorithm, the classifier receives an input to categorize each tree in a forest group and provides results in a dataset that gets the majority of votes.

4.5. REP tree

The REP tree method is a rapid logical classifier trees methodology that leverages the notion of computer-selected random features to reduce variance inaccuracy (Quinlan 1987; Devasena 2014). The REP tree employs the logistic regression technique and creates several trees via multiple computation processes, from which the most straightforward tree was selected (Devasena 2014). Observing training datasets and reducing the tree's internal structure complexity enables the REP tree to provide a flexible and uncomplicated modeling approach whenever the result is significant (Mohamed *et al.* 2012). During this approach, the pruning algorithm considers the backward overfitting complexity and uses the post-pruning algorithm to encourage the lowest version of the best precision tree logic (Quinlan 1987; Chen *et al.* 2019). It only chooses numbers for mathematical characteristics once (Kalmegh 2015).

4.6. GP regression

A methodology for virtual ML called the vector method (GP regression) allows computer systems to adapt and enhance their capabilities. A method that acts directly above the feature space is GP regression, which is based on the idea that nearby observations must exchange data (Kuß 2006). Other GPs also include the expansion of the core of the Gaussian distribution. The mean and covariance show the Gaussian probability density matrix and the kernel-based correlation vector. GP regression models can predict the unknown input data based on the probability theorem. In addition, they may give expected accuracy, which raises the statistical significance of the predictive model's findings. A GP involves unlimited random variables; hence, procedures are grounded on multivariate Gaussian models. Since the invention of this approach a few years ago, it has been widely used in many study fields, including chemistry, medicine, construction, etc. (Singh *et al.* 2017).

The kernels used in the present study are discussed as follows:

$$\text{RBF kernel} = (e^{-\gamma}|x_i - x_j|^2) \tag{9}$$

$$\text{PUK kernel} = \frac{1}{\left[1 + \left(\frac{\sqrt{||x-y||^2}\sqrt{\left|\left|2\frac{1}{\omega} - 1\right|\right|^2}}{\sigma}\right)\right]^{\omega}} \tag{10}$$

$$\text{Poly kernel} = (1 + (X, Y))^d \tag{11}$$

where σ, ω, γ, and d are kernel-specific parameters.

5. PERFORMANCE EVALUATION PARAMETERS

Calculating the goodness-of-the-fit indices: In the present investigation, numerous necessary statistical measures such as the CC, normalized error (NE), RMSE, normalized root mean square errors (NRMSE), MAE, legates and McCabe's index (LMI), and Willmott's index (WI) were calculated to quantify the fit among the experimental data and the predicted data using applied models. Formulas of these indicators are listed in Equations (12)–(17).

The research also used the analysis of variance (ANOVA) test. When examining changes in means, ANOVA is a set of statistical models and related estimation procedures (such as 'variation' within and across groups). Statistician Ronald Fisher created ANOVA.

$$\text{CC} = \frac{\sum_{i=1}^{N}(R_i - \bar{R})(Q_i - \bar{Q})}{\sqrt{\sum_{i=1}^{N}(R_i - \bar{R})^2 \sum_{i=1}^{N}(Q_i - \bar{Q})^2}} - 1 \le \text{CC} \le 1 \tag{12}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(R_i - Q_i)^2} \tag{13}$$

$$\text{LMI} = 1 - \left[\frac{\sum_{i=1}^{N}|R_i - Q_i|}{\sum_{i=1}^{N}|Q_i - \bar{Q}|}\right] \quad 0 \le \text{LMI} \le 1 \tag{14}$$

$$\text{WI} = 1 - \left[\frac{\sum_{i=1}^{N}(R_i - Q_i)^2}{\sum_{i=1}^{N}(|R_i - \bar{Q}| + |Q_i - \bar{Q}|)^2}\right] \quad 0 \le \text{WI} \le 1 \tag{15}$$

$$\text{NE} = 1 - \left(\frac{\sum_{i=1}^{N}(R_i - Q_i)^2}{\sum_{i=1}^{N}(Q_i - \bar{Q})^2}\right) \quad -\infty \le \text{NS} \le 1 \tag{16}$$

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|R_i - Q_i| \tag{17}$$

where Q is the real data; $\bar{Q}$ is the median of the results; R_i and R are the results expected (model); and N is the investigation quantity.

6. DATASET

This investigation evaluated a significant number (215 experimental observations) of experimental data for modeling the bearing capacity utilizing a tree, GP, and regression-based analysis. Of 215, 150 comments (sample population for training) were used to calibrate the model, and the remaining 65 observations (bundle of data for testing) were utilized to validate the model. The dataset contains channel width (b in m), channel slope (S_o), brink depth (y_b in m), normal depth (y_n in m), and friction coefficient (k) as input parameters, and y_c (in m) and actual Q (in m^3/s) are considered as targets. The correlation matrix of the whole dataset is listed in Table 1. Table 1 indicates that b (m) has no relationship with outputs. Table 2 presents the data statistics of the 120 observations (training dataset) and 56 observations (testing dataset; Figure 3).

7. RESULTS AND DISCUSSION

In this study, the parameters were adapted to predict the y_c (m) and Q flowing through the depth model's terminus (Q (m^3/s), free overfall. The results and discussion of LR, M5P, RF, RT, REP tree, and GP-based models' performance are covered in this portion.

Table 1 | Correlation matrix using a total dataset

	b (m)	S_o	y_b (m)	y_n (m)	K	y_c (m)	Q_{act} (m^3/s)
b (m)	1.0000						
S_o	0.0000	1.0000					
y_b (m)	0.0000	−0.2110	1.0000				
y_n (m)	0.0000	−0.0998	0.9322	1.0000			
K (m)	0.0000	−0.0251	−0.0037	0.2236	1.0000		
y_c (m)	0.0000	−0.0644	0.9586	0.9604	0.0204	1.0000	
Q_{act} (m^3/s)	0.0000	−0.0694	0.9523	0.9599	0.0202	0.9976	1.0000

Table 2 | Dataset characteristics used to predict the bearing capacity of sediment ash beds

Statics	b (m)	S_o	y_b (m)	y_n (m)	k	y_c (m)	Q_{act} (m^3/s)
	Training dataset						
Minimum	0.3000	0.0000	0.0120	0.0460	0.0000	0.0333	0.0057
Maximum	0.3000	0.0100	0.0500	0.1050	0.0100	0.0791	0.0209
Mean	0.3000	0.0037	0.0307	0.0717	0.0050	0.0539	0.0121
SD	0.0000	0.0041	0.0096	0.0162	0.0037	0.0153	0.0051
Kurtosis	−2.0272	−1.3321	−0.8644	−0.7151	−1.4601	−0.9274	−0.7956
Skewness	1.0101	0.5167	0.0672	0.3081	0.1539	0.2332	0.4455
Confidence level (95%)	0.0000	0.0007	0.0016	0.0026	0.0006	0.0025	0.0008
	Testing dataset						
Minimum	0.3000	0.0000	0.0138	0.0470	0.0000	0.0333	0.0057
Maximum	0.3000	0.0100	0.0490	0.1050	0.0100	0.0791	0.0209
Mean	0.3000	0.0042	0.0323	0.0754	0.0051	0.0574	0.0133
SD	0.0000	0.0043	0.0092	0.0166	0.0037	0.0157	0.0053
Kurtosis	−2.0645	−1.5701	−0.8166	−0.9951	−1.4262	−1.1215	−1.1674
Skewness	−1.0238	0.3378	−0.1863	0.0971	0.0549	0.1096	0.2858
Confidence level (95%)	0.0000	0.0011	0.0023	0.0041	0.0009	0.0039	0.0013

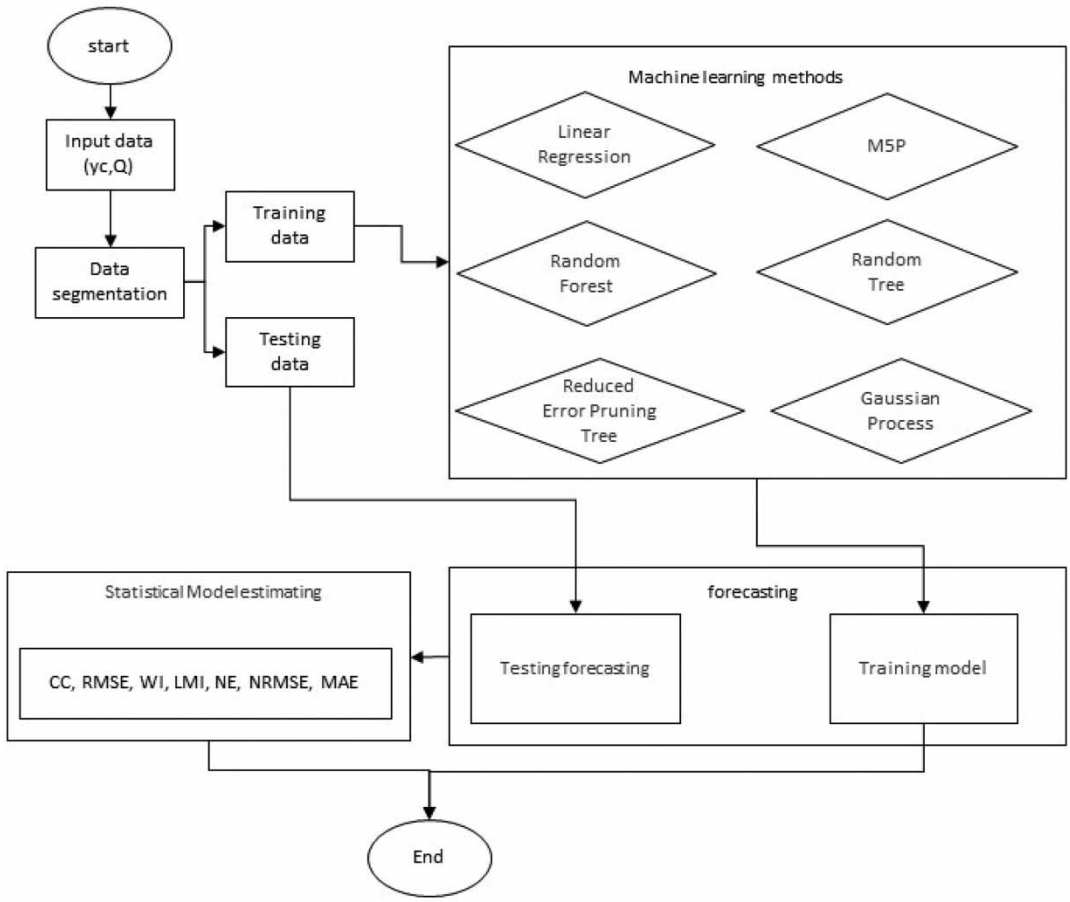

Figure 3 | Flowchart of the machine learning.

7.1. Assessment of regression and soft computing-based model for critical depth y_c

Table 3 shows the achievement evaluation parameters for every one of the models utilized in the training and testing phases. The RT study estimates the y_c better than other models when examining performance assessment indicators during training. This model is more effective than the different applied models. Table 3 recommends that the LR model also performs superior to the GP_Poly model for predicting y_c. Typically, throughout the training stage, the models can be categorized from finest to lowest: RT, RF, Pearson VII kernel function-based GP model, REP tree, M5P, GP_RBF, LR, and GP_Poly.

The GP RBF model outperforms other applicable models in the testing process with the lowest RMSE of 0.0021, NRMSE of 0.0361, MAE of 0.0015 and the highest CC of 0.9912, LMI of 0.8839, WI of 0.9956, NE of 0.9823. This model is more effective than the previous one that has been used. Table 3 shows that the LR model outperforms the M5P, RF, RT, REP tree, and GP Poly models for y_c prediction, with CC values of 0.9837, RMSE values of 0.0028, WI values of 0.9927, LMI values of 0.8264, NE values of 0.9712, NRMSE values of 0.0482, and MAE values of 0.0021 for testing stages. The models can be ordered from good to bad throughout the testing stage: GP_RBF, GP_PUK, LR, RF, M5P, RT, REP tree, and GP_Poly. By using various soft computing algorithms, Appendix A displays the consistency plot of natural versus anticipated y_c values for training and testing phases. These figures indicate that the GP_RBF model is the best-applied model to predict the actual depth. All predicted values using the GP_RBF model lie much closer to the line of perfect agreement ($y = x$), with R^2 values as 0.989 and 0.982 for the training and testing stages, respectively. Results of single-factor ANOVA suggest that there is an insignificant difference among various groups. Table 4 indicates that all predictive models are suitable for predicting y_c using this dataset.

A box plot compares the actual and expected y_c lowest, maximum, first, third, and mean values to analyze how uniformly the projected y_c corresponds to the actual values. Figure 4 shows that in the testing stage, the GP_RBF values are significantly closer to the real data. Overall, assessing Figure 4 shows that, in comparison to fundamental importance in both phases, the breadth of the first and third quadrants of the GP RBF models is almost identical. Figure 4 suggests that GP_RBF is the most appropriate model, and GP_Poly is the model with the lowest predictive accuracy y_c among all applied models.

The Taylor diagram is a robust method to visualize three statistical indicators in a single graph to compare the performance of different models to capture the observation. In this study, the observed and modeled (SD), correlation coefficient (CC), and RMSE in a single point are presented. Figure 5 According to the Taylor diagram, the GP RBF (solid gray circle)-based model

Table 3 | Performance of soft computing and regression-based models for y_c

Models	CC	RMSE	WI	LMI	NE	NRMSE	MAE
	Training dataset						
LR	0.9856	0.0026	0.9927	0.8264	0.9712	0.0482	0.0021
M5P	0.9945	0.0016	0.9971	0.8969	0.9888	0.0301	0.0013
RF	0.9990	0.0007	0.9994	0.9680	0.9977	0.0137	0.0004
RT	1.0000	0.0000	1.0000	1.0000	1.0000	0.0000	0.0000
REP tree	0.9961	0.0013	0.9980	0.9604	0.9922	0.0251	0.0005
GP_PUK	0.9977	0.0011	0.9988	0.9452	0.9952	0.0196	0.0007
GP_Poly	0.8929	0.0073	0.9302	0.5930	0.7690	0.1366	0.0050
GP_RBF	0.9945	0.0016	0.9972	0.9040	0.9889	0.0300	0.0012
	Testing dataset						
LR	0.9837	0.0028	0.9917	0.8247	0.9677	0.0487	0.0023
M5P	0.9720	0.0038	0.9841	0.8159	0.9408	0.0659	0.0024
RF	0.9805	0.0034	0.9864	0.8523	0.9514	0.0597	0.0019
RT	0.9666	0.0041	0.9819	0.9015	0.9291	0.0721	0.0013
REP tree	0.9284	0.0062	0.9555	0.7960	0.8424	0.1075	0.0027
GP_PUK	0.9909	0.0021	0.9954	0.8816	0.9818	0.0366	0.0016
GP_Poly	0.9074	0.0070	0.9454	0.6342	0.7971	0.1220	0.0048
GP_RBF	0.9912	0.0021	0.9956	0.8839	0.9823	0.0361	0.0015

Table 4 | Single-factor ANOVA results among basic and soft computing techniques for predicting (y_c)

Sr. No.	Source of variation	*F*	*P*-value	*F* crit	Insignificant variation
1	Between actual and LR groups	0.00007	0.99352	3.91514	✓
2	Between actual and M5P groups	0.10823	0.74271	3.91514	✓
3	Between actual and RF groups	0.10949	0.74127	3.91514	✓
4	Between actual and RT groups	0.14652	0.70252	3.91514	✓
5	Between actual and REP tree groups	0.67489	0.41288	3.91514	✓
6	Between actual and GP_PUK groups	0.00003	0.99550	3.91514	✓
7	Between actual and GP_Poly groups	0.75192	0.38749	3.91514	✓
8	Between actual and GP_RBF groups	0.00154	0.96877	3.91514	✓
9	Between actual and all applied groups	0.42837	0.90421	1.95446	✓

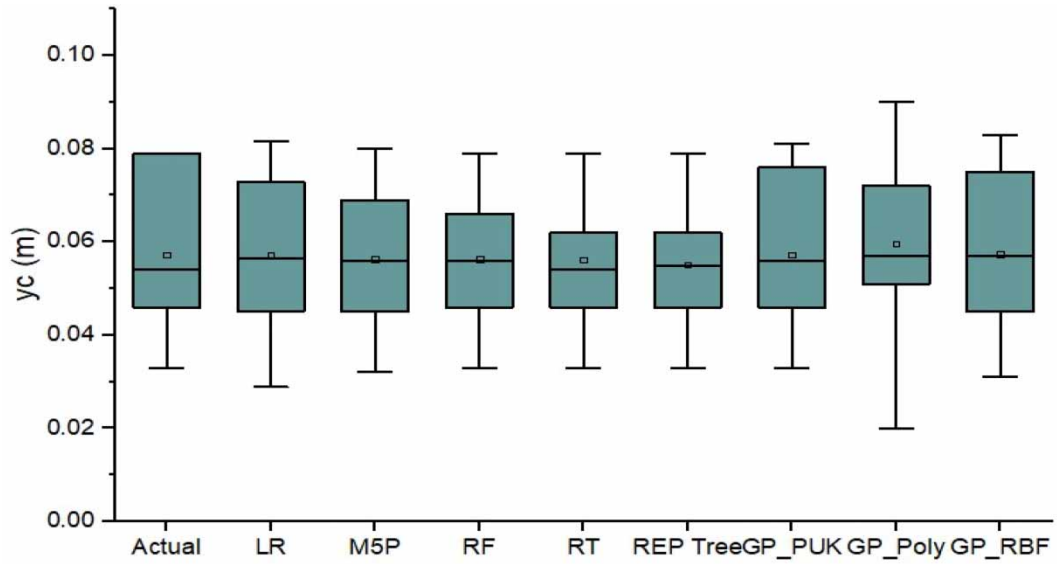

Figure 4 | Box plot for actual and predicted values using soft computing techniques to predict y_c with the testing stage.

is more accurate for predicting y_c than the other applicable models. The overall performance of the GP poly (solid purple ring) model is the worst of all tested models.

7.2. Assessment of regression and soft computing-based model for actual discharge *Q* (m^3/s)

The performance assessment parameters of each of the models used to predict actual Q (m^3/s) in the training and testing stages are listed in Table 5. The measurement of performance evaluation indices in the training stages shows that the RT model predicts the actual Q (m^3/s) superior to other models. The RT model has had the minor RMSE = 0, NRMSE = 0, MAE = 0.0, and the most incredible CC = 1, LMI = 1, WI = 1, and Nash-Sutcliffe efficiency (NSE) = 1 in the training phase, according to the performance indicators. This model is more accurate than the other applied model's model. Table 5 suggests that the LR model is also performing better than the polynomial kernel function-based (GP_Poly)-based model for the prediction of actual Q with CC values as 0.9812, RMSE values as 0.0010, WI values as 0.9902, LMI values as 0.8112, NE values as 0.9620, NRMSE values as 0.0806, and MAE values as 0.0008 for the training stage. In general, throughout the training phase, the models can be categorized from the highest to the lowest in this manner: RT, RF, Pearson VII kernel function-based GP_PUK, REP tree, radial basis kernel function-based (GP_RBF), M5P, LR, and polynomial kernel function-based GP_Poly.

During the testing stage, the GP RBF model outperformed other applicable models in predicting accurate Q, with the lowest RMSE = 0.0007, NRMSE = 0.0516, MAE = 0.0004, and the highest CC = 0.9916, LMI = 0.9026, WI = 0.9956,

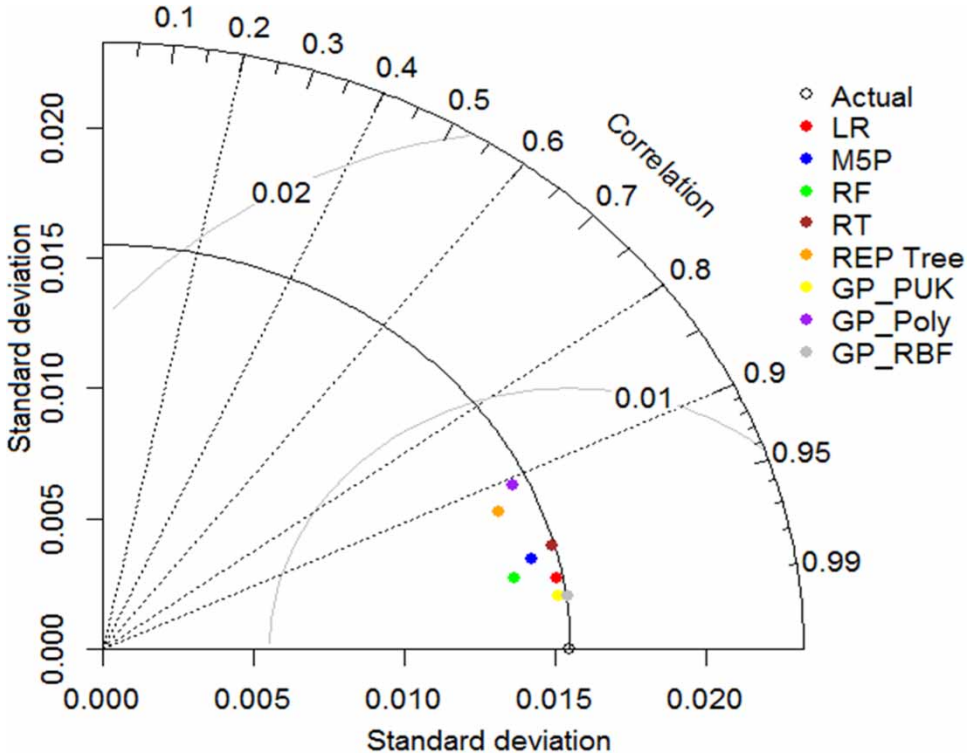

Figure 5 | Taylor diagram for actual and predicted values using soft computing techniques to predict y_c with the testing stage.

Table 5 | Performance evaluation indices for all applied models for actual Q

Models	CC	RMSE	WI	LMI	NE	NRMSE	MAE
	Training dataset						
LR	0.9812	0.0010	0.9902	0.8112	0.9620	0.0806	0.0008
M5P	0.9909	0.0007	0.9951	0.8895	0.9810	0.0569	0.0005
RF	0.9978	0.0003	0.9988	0.9712	0.9953	0.0283	0.0001
RT	1.0000	0.0000	1.0000	1.0000	1.0000	0.0000	0.0000
REP tree	0.9946	0.0005	0.9972	0.9680	0.9891	0.0432	0.0001
GP_PUK	0.9964	0.0004	0.9981	0.9536	0.9925	0.0359	0.0002
GP_Poly	0.9202	0.0021	0.9502	0.6316	0.8312	0.1698	0.0015
GP_RBF	0.9946	0.0005	0.9971	0.9295	0.9886	0.0442	0.0003
	Testing dataset						
LR	0.9796	0.0011	0.9891	0.8189	0.9584	0.0806	0.0008
M5P	0.9672	0.0015	0.9791	0.8120	0.9242	0.1089	0.0009
RF	0.9740	0.0014	0.9808	0.8556	0.9324	0.1028	0.0007
RT	0.9677	0.0013	0.9832	0.9093	0.9357	0.1002	0.0004
REP tree	0.9310	0.0021	0.9552	0.8153	0.8434	0.1564	0.0008
GP_PUK	0.9890	0.0008	0.9940	0.8858	0.9769	0.0601	0.0005
GP_Poly	0.9437	0.0018	0.9684	0.7146	0.8851	0.1340	0.0013
GP_RBF	0.9916	0.0007	0.9956	0.9026	0.9830	0.0516	0.0004

Table 6 | Single-factor ANOVA results among basic and soft computing techniques to predict actual Q

Sr. No.	Source of variation	F	P-value	F crit	Insignificant variation
1	Between actual and LR groups	0.02893	0.86520	3.91514	✓
2	Between actual and M5P groups	0.30136	0.58399	3.91514	✓
3	Between actual and RF groups	0.29215	0.58979	3.91514	✓
4	Between actual and RT groups	0.02249	0.88101	3.91514	✓
5	Between actual and REP tree groups	0.78198	0.37819	3.91514	✓
6	Between actual and GP_PUK groups	0.02792	0.86757	3.91514	✓
7	Between actual and GP_Poly groups	0.18148	0.67082	3.91514	✓
8	Between actual and GP_RBF groups	0.00688	0.93400	3.91514	✓
9	Between actual and all applied groups	0.28975	0.96938	1.95446	✓

NE = 0.9830. This model is much more effective than the others that have been used. Table 5 proposes that the LR model is also outperforming superior M5P, RF, RT, REP tree, and GP_Poly-based models for the prediction of actual Q with CC values as 0.9796, RMSE values as 0.0011, WI values as 0.9891, LMI values as 0.8189, NE values as 0.9584, NRMSE values as 0.0806, and MAE values as 0.0008 for testing stages. The models can be categorized from the best to lowest throughout the testing phase: GP_RBF, GP_PUK, LR, RF, M5P, RT, GP_Poly, and REP tree. Appendix B shows the agreement plot of actual versus predicted values of real Q with various soft computing techniques for training and testing stages. These figures also indicate that the GP_RBF model is the best among all applied models for the prediction of actual Q. All the predicted values using the GP_RBF model lie very close to a line of perfect agreement ($y = x$) with R^2 values as 0.989 and 0.983 for training and testing stages, respectively. Results of single-factor ANOVA suggest that there is an insignificant difference among various groups. Table 6 indicates that all predictive models are suitable for predicting actual Q using this dataset.

The lowest, maximum, first quartile, third quartile, and mean values of the actual and forecasted actual Q were analyzed using a box plot to determine the compatibility of the anticipated y_c with the exact amounts. Figure 6 shows that in the testing stage, the GP_RBF values are significantly closer to the actual data. Overall, assessing Figure 6 suggests that GP_RBF models' first and third quartile widths are almost identical to fundamental importance in both phases. Figure 6 indicates that GP_RBF is the most suitable model for predicting all applied models' actual discharge Q value.

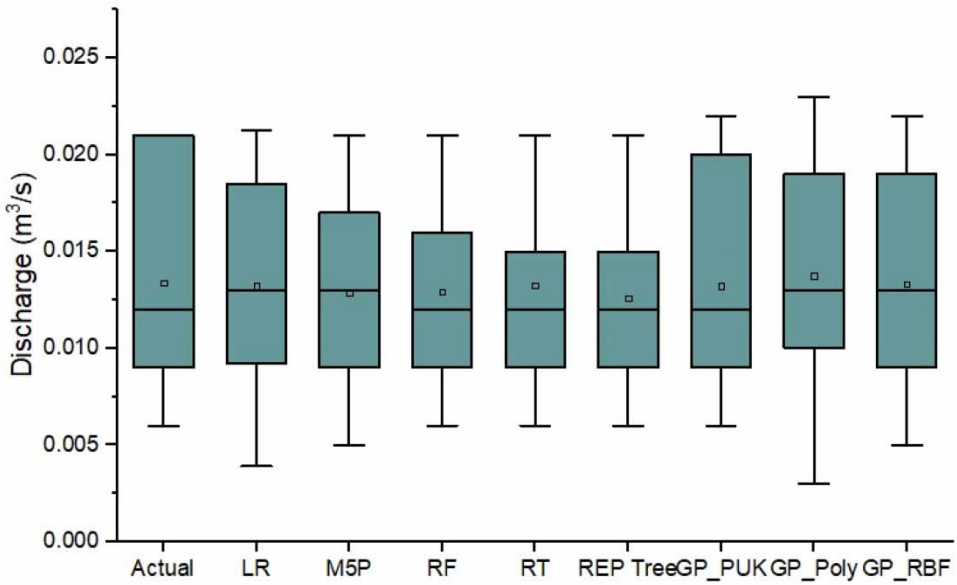

Figure 6 | Box plot for actual and predicted values using soft computing techniques to predict real Q with the testing stage.

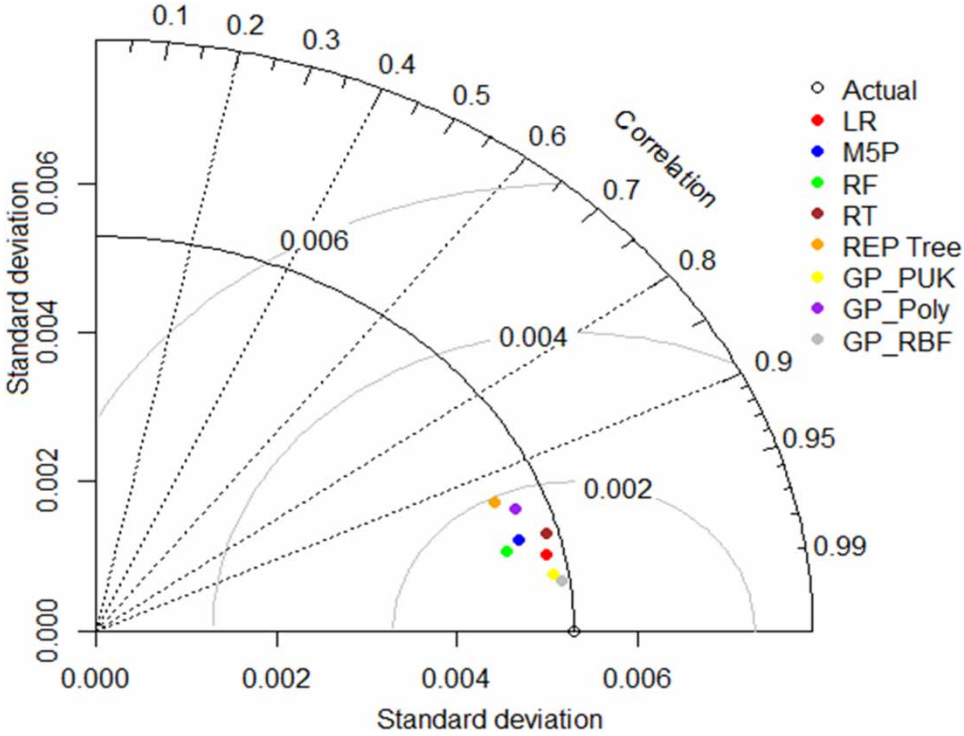

Figure 7 | Taylor diagram for actual and predicted values using soft computing techniques to predict real Q with the testing stage.

A Taylor diagram is selected for fair compassion among various soft computing-based models (Figure 7). The Taylor diagram predicts the actual Q and shows that the GP RBF (solid gray circle) conceptual system has the best efficiency compared to the other applied models. Overall, the REP tree (solid orange circle) performance is the worst among all used models.

7.3. Model optimization for sensitivity analysis

Because the GP RBF-based model outperforms all regression and soft programming models for predicting (y_c) and discharge crossing over the end-depth model (Q), a sensitivity analysis was conducted with the GP RBF model to determine the most sensitive feature among input variables for predicting (y_c) and discharge crossing over the terminal depth model (Q). For (y_c) and discharge crossing above the terminal depth model, input configuration models are generated by removing one input parameter in each case, as shown in Tables 7 and 8 (Q). The root-mean-square error, MAE, and CC are considered while evaluating each model's performance. The friction coefficient is the dataset's most significant input variable for evaluating y_c. Table 7 also indicates that the friction coefficient has a substantial influence in predicting the discharge passing over

Table 7 | The sensitivity analysis results using the GP_RBF-based model for predicting (y_c)

Input combination					Output		GP_RBF-based model			
b (m)	S_o	y_b (m)	Y_n (m)	k	y_c	Eliminated parameter	CC	RMSE	MAE	Rank
✓	✓	✓	✓	✓	✓	Nil	0.9916	0.0021	0.0015	–
✗	✓	✓	✓	✓	✓	b (m)	0.9912	0.0021	0.0017	5
✓	✗	✓	✓	✓	✓	S_o	0.9856	0.0026	0.0022	4
✓	✓	✗	✓	✓	✓	y_b	0.9845	0.0028	0.0022	3
✓	✓	✓	✗	✓	✓	y_n (m)	0.9807	0.0031	0.0023	2
✓	✓	✓	✓	✗	✓	K	0.9770	0.0033	0.0025	1

Table 8 | Sensitivity analysis results using a GP_RBF-based model for predicting discharge passing over the end-depth model (*Q*)

Input combination					Output		GP_RBF-based model			
b (m)	S_o	y_b (m)	y_n (m)	*k*	*Q* (m^3/s)	Eliminated parameter	CC	RMSE	MAE	Rank
✓	✓	✓	✓	✓	✓	Nil	0.9916	0.0007	0.0004	–
✗	✓	✓	✓	✓	✓	*b(m)*	0.9912	0.0007	0.0005	5
✓	✗	✓	✓	✓	✓	*So*	0.9857	0.0009	0.0007	4
✓	✓	✗	✓	✓	✓	*yb(m)*	0.9856	0.0009	0.0007	3
✓	✓	✓	✗	✓	✓	*yn(m)*	0.9790	0.0011	0.0008	2
✓	✓	✓	✓	✗	✓	*K*	0.9743	0.0012	0.0009	1

the end-depth model (*Q*) compared to other input parameters in this study. The friction coefficient and normal depth are the most significant input variables compared to other parameters for predicting (y_c) and discharge crossing above the end-depth model (*Q*).

8. CONCLUSIONS

This study examines how soft computing and regression approaches can forecast (y_c) and discharge over the terminal depth model (*Q*). For estimating (y_c) and flow crossing through the end-depth model, LR, M5P, RF, RT, REP tree, and GP-based models are utilized (*Q*). The behavior of constructed models is assessed using seven distinct goodness-of-fit criteria. According to achievement examination results, it is observed that the radial kernel function-based GP model (GP_RBF) is the most suitable model for the prediction of y_c and *Q* crossing above the end-depth model (*Q*) compared to other applied models with lowest RMSE = 0.0021, 0.007, NRMSE = 0.0361, 0.0516, MAE = 0.0015, 0.004 and the highest CC = 0.9912, 0.9916, LMI = 0.8839, 0.9026, WI = 0.9956, 0.9956 and NE = 0.9823, 09830 for y_c and *Q* crossing above the end-depth model (*Q*), respectively, with testing stage. Another primary outcome of this investigation is that LR-based equations' performance is better than all applied models except GP_PUK and GP_RBF for y_c and *Q*.

Sensitivity analysis results indicate that the friction coefficient is the most significant input variable compared to other parameters for predicting y_c and *Q* crossing above the end-depth model (*Q*) using this dataset.

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Aldous, D. 1991 The continuum random tree. I. *The Annals of Probability* **19** (1). https://doi.org/10.1214/aop/1176990534.

Breiman, L. 2001 Random forests. *Machine Learning* **45** (1), 5–32. https://doi.org/10.1023/A:1010933404324.

Chen, W., Hong, H., Li, S., Shahabi, H., Wang, Y., Wang, X. & Ahmad, B. B. 2019 Flood susceptibility modelling using novel hybrid approach of reduced-error pruning trees with bagging and random subspace ensembles. *Journal of Hydrology* **575**, 864–873. https://doi.org/10.1016/j.jhydrol.2019.05.089.

Devasena L. 2015 Comparative Analysis of Random Forest, REP Tree and J48 Classifiers for Credit Risk Prediction. International Conference on Communication, Computing and Information Technology. ICCCMIT2014, 3 (March 2015), 30–36.

Devasena, L. 2014 Comparative analysis of random forest, REP tree and J48 classifiers for credit risk prediction. *International Journal of Computer Applications*.

Dey, S. & Kumar, B. R. 2002 Hydraulics of free overfall in Δ-shaped channels. *Sadhana – Academy Proceedings in Engineering Sciences* **27** (PART 3), 353–363. https://doi.org/10.1007/BF02703656.

Dey, S., Kumar, D. N. & Singh, D. R. 2004 End-depth in inverted semicircular channels: Experimental and theoretical studies. *Nordic Hydrology* **35** (1), 73–79. https://doi.org/10.2166/nh.2004.0006.

Fırat, C. E. 2004 Effect of roughness on flow measurements in sloping rectangular channels with free overfall [M.S. - Master of Science]. Middle East Technical University.

Firat, C. E. 2015 Effect of roughness on flow measurements in sloping rectangular channels with free overfall. In: *Statewide Agricultural Land Use Baseline 2015*, Vol. 1, Issue February.

Gharehbaghi, A., Ghasemlounia, R., Latif, S. D., Haghiabi, A. H. & Parsaie, A. 2023 Application of data-driven models to predict the dimensions of flow separation zone. *Environmental Science and Pollution Research* **30** (24), 65572–65586. https://doi.org/10.1007/s11356-023-27024-y.

Guo, Y., Zhang, L. & Zhang, J. 2006 Numerical simulation of free overfall in a rough channel. In *European Conference on Computational Fluid Dynamics ECCOMAS CFD*.

Irzooki, R. & Hasan, S. 2018 Characteristics of flow over the free overfall of triangular channel. *MATEC Web of Conferences* **162**, 03006. https://doi.org/10.1051/matecconf/201816203006.

Kalmegh, S. 2015 Analysis of WEKA data mining algorithm REP tree, simple cart and random tree for classification of Indian news. *International Journal of Innovative Science, Engineering & Technology* **2** (2), 438–446.

Kuß, M. 2006 *Gaussian Process Models for Robust Regression, Classification, and Reinforcement Learning*. Ph.D. Thesis. Available from: http://tuprints.ulb.tu-darmstadt.de/674/

Latif, S. D. & Ahmed, A. N. 2023 Ensuring a generalizable machine learning model for forecasting reservoir inflow in Kurdistan region of Iraq and Australia. *Environment, Development and Sustainability* 1–32. https://doi.org/10.1007/s10668-023-03885-8

Mohammed, A. Y. 2009a Hydraulic characteristics of free overfall with triangular end lip. In *33rd IAHR Congress: Water Engineering for A Sustainable Environment*.

Mohammed, A. Y. 2009b Effecting of channel slope on flow characteristics for straight vertical and skew free overfall. *AL-Rafdain Engineering Journal (AREJ)* **17** (1), 80–90. https://doi.org/10.33899/rengj.2009.38694.

Mohammed, A. Y. 2012 Theoretical end depth ratio and end depth discharge relationship for free overfall with different end lip shape. *Jordan Journal of Civil Engineering* **6** (4), 410–417.

Mohammed, A. Y. 2013 Effect of bed roughness distribution and channel slope on rectangular free overfall. *AL-Qadisiya Journal for Engineering Sciences* **6** (2), 115–123.

Mohammed, A. Y. 2018 Artificial neural network (ANN) model for end depth computations. *Journal of Civil & Environmental Engineering* **8** (3). https://doi.org/10.4172/2165-784x.1000316.

Mohammed, A. Y., Khaleel, M. S., & Mohammad, M. Y. 2007 Variation of Water Depth on Normal and Skewed Broad Crested Weirs. *Tikrit Journal of Engineering Sciences* **14** (1), 28–43. https://doi.org/10.25130/tjes.14.1.02.

Mohammed, M. Y., Mohammed, A. Y. & Altalib, A. N. 2011 Gravel roughness and channel slope effects on rectangular free overfall. *Damascus University Journal* **27** (1), 47–54.

Mohamed, W. N. H. W., Salleh, M. N. M. & Omar, A. H. 2012 A comparative study of reduced error pruning method in decision tree algorithms. In: *Proceedings – 2012 IEEE International Conference on Control System, Computing and Engineering, ICCSCE 2012*, pp. 392–397. https://doi.org/10.1109/ICCSCE.2012.6487177

Mohammed, M. Y., Mohammed, A. Y. & Kasem, I. A. 2018 Türk Hidrolik Dergisi/Turkish journal of hydraulic flow measurements in rough free overfall. **2** (1), 8–12. Available from: http://www.dergipark.gov.tr

Muhsin, K. A. & Noori, B. M. A. 2021 Hydraulics of free overfall in smooth triangular channels. *Ain Shams Engineering Journal* **12** (3), 2471–2484. https://doi.org/10.1016/j.asej.2020.11.022.

Olyaie, E., Banejad, H. & Heydari, M. 2019 Estimating discharge coefficient of PK-weir under subcritical conditions based on high-accuracy machine learning approaches. *Iranian Journal of Science and Technology – Transactions of Civil Engineering* **43** (1), 89–101. https://doi.org/10.1007/s40996-018-0150-z.

Öztürk, H. U. 2005 *Discharge Predictions Using Ann in Sloping Rectangular Channels With Free Overfall. A Thesis Submitted to the Graduate School of Natural and Applied Sciences of Middle East Technical University in Partial Fullfillment of the Requirements for the Degree*.

Quinlan, J. R. 1987 Simplifying decision trees. *International Journal of Man-Machine Studies* **27** (3), 221–234. https://doi.org/10.1016/S0020-7373(87)80053-6.

Quinlan, J. R. 1992 Learning with continuous classes. In 5th Australian joint conference on artificial intelligence (Vol. 92, pp. 343–348).

Raikar, R. V., Nagesh Kumar, D. & Dey, S. 2004 End depth computation in inverted semicircular channels using ANNs. *Flow Measurement and Instrumentation* **15** (5–6), 285–293. https://doi.org/10.1016/j.flowmeasinst.2004.06.003.

Salmasi, F., Nouri, M., Sihag, P. & Abraham, J. 2021 Application of SVM, ANN, GRNN, RF, GP and RT models for predicting discharge coefficients of oblique sluice gates using experimental data. *Water Science and Technology: Water Supply* **21** (1), 232–248. https://doi.org/10.2166/ws.2020.226.

Shi, Y., Li, Q., Bu, S., Yang, J. & Zhu, L. 2020 Research on intelligent vehicle path planning based on rapidly-exploring random tree. *Mathematical Problems in Engineering* **2020**. https://doi.org/10.1155/2020/5910503.

Singh, B., Sihag, P. & Singh, K. 2017 Modelling of impact of water quality on infiltration rate of soil by random forest regression. *Modeling Earth Systems and Environment* **3** (3), 999–1004. https://doi.org/10.1007/s40808-017-0347-3.

Sihag, P., Mohsenzadeh Karimi, S. & Angelaki, A. 2019 Random forest, M5P and regression analysis to estimate the field unsaturated hydraulic conductivity. *Appl Water Sci* **9**, 129. https://doi.org/10.1007/s13201-019-1007-8.

Sihag, P., Angelaki, A. & Chaplot, B. 2020 Estimation of the recharging rate of groundwater using random forest technique. *Appl Water Sci* **10**, 182. https://doi.org/10.1007/s13201-020-01267-3.

Suntaranont, B., Aramkul, S., Kaewmoracharoen, M. & Champrasert, P. 2020 Water irrigation decision support system for practicalweir adjustment using artificial intelligence and machine learning techniques. *Sustainability (Switzerland)* **12** (5), 1763. https://doi.org/10.3390/su12051763.

Thakur, M. S., Pandhiani, S. M., Kashyap, V., Upadhya, A. & Sihag, P. 2021 Predicting bond strength of FRP bars in concrete using soft computing techniques. *Arabian Journal for Science and Engineering* **46** (5), 4951–4969. https://doi.org/10.1007/s13369-020-05314-8.

Yousif, A. A., Sulaiman, S. O., Diop, L., Ehteram, M., Shahid, S., Al-Ansari, N. & Yaseen, Z. M. 2019 Open channel sluice gate scouring parameters prediction: Different scenarios of dimensional and non-dimensional input parameters. *Water (Switzerland)* **11** (2). https://doi.org/10.3390/w11020353.

Zeidan, R., Elshemy, M. & Rashwan, I. 2021 Using the brink depth in discharge measurement for inverted semicircular open channels. *Journal of Engineering Research* **5** (1). https://doi.org/10.21608/erjeng.2021.68393.1007.

First received 15 October 2023; accepted in revised form 26 January 2024. Available online 26 February 2024