Journal of Hydroinformatics

Journal of Hydroinformatics Vol 25 No 3, 1084 doi: 10.2166/hydro.2023.007

Approximation of aeration efficiency at sharp-crested weirs using metaheuristic regression approaches

Akash Jaiswal [a,*], Arun Goel[b] and Parveen Sihag[c]

[a] Water Resources Development and Management, IIT Roorkee, Roorkee, Uttarakhand 247667, India
[b] Civil Engineering Department, NIT Kurukshetra, Kurukshetra, Haryana 136119, India
[c] Civil Engineering Department, Chandigarh University, Chandigarh, Punjab 140413, India
*Corresponding author. E-mail: akaskjaiswal@gmail.com

AJ, 0000-0001-9141-9322

ABSTRACT

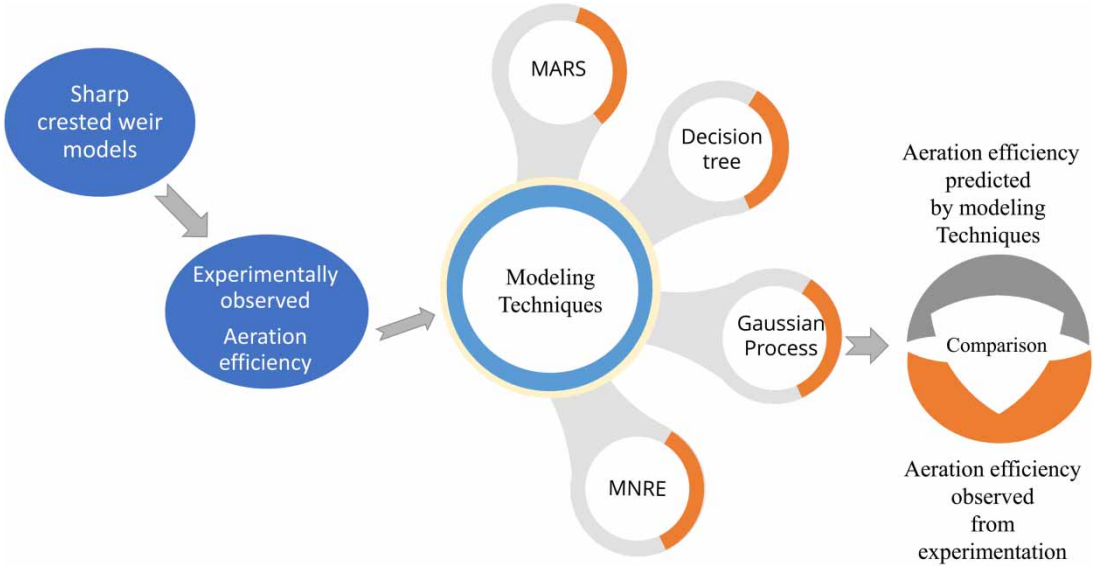
This paper explores the ability of multivariate adaptive regression splines, decision trees, Gaussian processes, and multiple non-linear regression equation approaches to predict the aeration efficiency at various weirs and discusses their results. In total, 126 experimental observations were collected in the laboratory, of which 88 were arbitrarily selected for model training, and the rest were used for model validation. Various graphical presentations and goodness-of-fit parameters were used to assess the performance of the models. Performance evaluation results, Whisker plot, and Taylor's diagram indicated that the GP_rbf-based model was superior to other implemented models in predicting the aeration efficiency of weirs with *CC* (0.9961 and 0.9973), *MAE* (0.0079 and 0.0195), *RMSE* (0.0122 and 0.0251), scattering index (0.0594 and 0.1238), and Nash Sutcliffe model efficiency (0.9923 and 0.9564) values in the training and validating stages, respectively. The predicted values by GP_rbf lie within the $\pm$30% error line in the training and validating stages, with most of it lying at/close to the line of agreement. The random forest model had better predictability than other decision tree models implied. The sensitivity analysis of parameters suggests shape factor and drop height as major influencing factors in predicting the aeration efficiency.

Key words: aeration efficiency, decision tree, Gaussian process, multivariate adaptive regression splines, non-linear regression equation, weirs

HIGHLIGHTS

- Experimental study to evaluate aeration efficiency at various shapes of sharp-crested weir models.
- Application of machine learning techniques to predict aeration efficiency of sharp-crested weirs.
- Introduction of shape factor for different shapes of weirs as an input to ML models.
- Use of graphs and goodness-of-fit parameters to assess the performance of applied ML models.
- Sensitivity analysis.

GRAPHICAL ABSTRACT



INTRODUCTION

Dissolved oxygen (DO) in water is indispensable for all life forms. It is a significant indicator of the water quality needed for human utility and healthy aquatic life (Goel 2013). DO in water should be ≥4 and 5 ppm to sustain warm water and cold water aquatic life, respectively (Baylar *et al.* 2010). Reducing DO concentration below this minimum requirement can significantly stress the natural aquatic cycle.

Aeration replenishes the oxygen deficiency in water by absorbing oxygen from the atmosphere. Aeration efficiency at any weir is the amount of oxygen (ppm) being infused in the water while it flows through the weir. It is expressed in Equation (1) as defined by Gulliver *et al.* (1990),

$$E = \frac{C_d - C_u}{C_s - C_u} \tag{1}$$

The amount of oxygen transferred is a function of temperature (T, °C). Thus, to provide a common base for comparing the performance of different weirs, a normalized aeration efficiency at 20 °C standard temperature is used (Eckenfelder & Ford 1970), as given in the following equation,

$$1 - E_{20} = (1 - E)^{1/f} \tag{2}$$

where C_u and C_d are dissolved oxygen concentration (ppm) upstream and downstream of the weir, respectively, C_s is the saturation concentration of dissolved oxygen of water (ppm), and f is a factor for temperature adjustment calculated as

$$f = 1 + 0.02103(T - 20) + 8.261 \times 10^{-5}(T - 20)^2 \tag{3}$$

Self-aeration, mechanical aeration, aeration through chemicals, and hydraulic jump aeration are among many possible processes to boost DO concentration in water. Sharp-crested weirs are also successfully utilized in assisting oxygen transfer to enhance DO in water. Weir acts as a barrier, thus raising the water level and allowing it to overfall as a nappe, as shown in Figure 1. The surface area of water in contact with the atmosphere increases, thus enhancing DO in the stream by infusing tiny air bubbles (Van der Karoon & Schram 1969a). The increased surface area gives water more time to be in contact with the atmosphere leading to natural aeration without requiring additional operating costs. The amount of oxygen transfer through air entrainment over several kilometers can otherwise be achieved over a single weir in the river (Baylar *et al.* 2010).

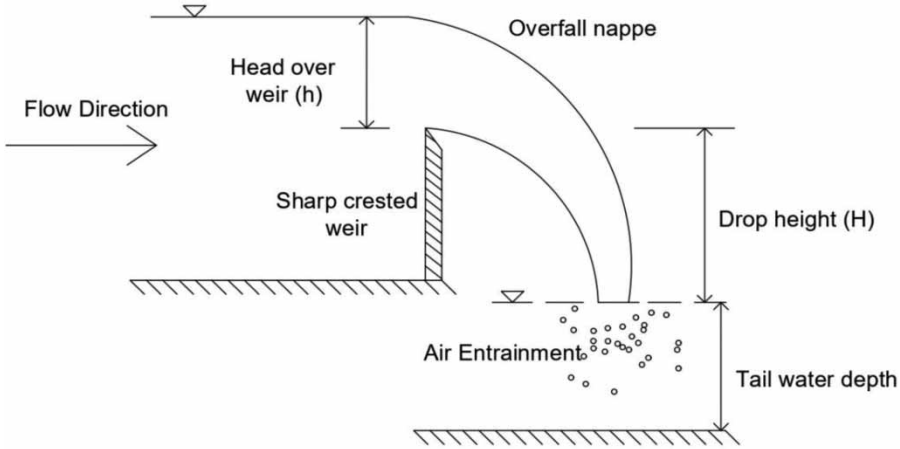


Figure 1 | Sketch showing the process of weir aeration.

Following this, investigation on the effect of the governing factors, such as head over the crest of weirs, drop height, weir configuration, temperature, and discharge, to get the most efficient weir shape for oxygen transfer and empirical relations for aeration efficiency, were proposed by Van der Kroon & Schram (1969a, 1969b); Apted & Novak (1973), Avery & Novak (1978), Nakasone (1987), Gulliver *et al.* (1990, 1998); Gulliver & Rindels (1993), Witt & Gulliver (2012), Baylar *et al.* (2001a, 2001b) and Jaiswal & Goel (2019). Goel (2013), and Jaiswal & Goel (2020) also assessed various machine learning models to predict aeration efficiency at sharp-crested weirs.

The present study investigates the potential of multivariate adaptive regression splines (MARS), Gaussian process (GP), multiple non-linear regression equation (MNRE), and decision tree [random forest (RF), random tree (RT), and M5P] models to predict the aeration efficiency at various weirs. The results of these regression approaches were compared with the multiple non-linear regression equation models. A series of experiments were performed to collect the aeration efficiency at various sharp-crested weirs under different governing parameters. The experiments were conducted on a recirculating rectangular flume in the hydraulics laboratory of CED, NIT, Kurukshetra. The governing parameters include the shape factor of weirs, head over crest of weir, drop height, and discharge. The correlation coefficient, mean absolute error, root mean square error, scattering index (*SI*), and Nash Sutcliffe model efficiency were used as goodness-of-fit parameters to evaluate the performance of the developed models.

MODELING TECHNIQUES

The modeling techniques that were investigated in this paper are briefly discussed in the following.

Multivariate adaptive regression splines

MARS, a non-parametric technique, does not require assumptions about the dependent and independent parameters relation. The space of input parameters in the MARS method is subdivided into several domains, and a linear regression equation is defined for each domain. The border value between the domains is called the knot, and the defined linear regression is called basic function (BF). The basic functions are of forms $(0, x-k)_{\max}$ or $(0, k-x)_{\max}$, wherein x is an independent parameter, and k is a border value. The dependent variable (y) in the MARS method is expressed as a function of x, as given in the following equation

$$y = f(x) = a_o + \sum_{n=1}^{N} a_n \beta_n(x) \tag{4}$$

where a_0 is a constant value, n is the number of BFs, β_n is the basis function, and a_n is coefficient of basic function.

The MARS model is processed in two stages: the first stage is the growing stage, a feed-forward algorithm, wherein a model is developed, and the feature input space is subdivided into various domains. The second stage is the pruning stage, wherein

the basic functions of the first stage that are not significant in enhancing the precision of the model are pruned using generalized cross validation (*GCV*) techniques. The *GCV* is represented by the following equation

$$GCV = \frac{\frac{1}{N}\sum_{i=1}^{N}[y_i - f(x_i)]^2}{[1 - (C(H)/N)]^2} \tag{5}$$

where N is the number of data and $C(H)$ is the complexity penalty that increases by the number of BFs. Complexity penalty $C(H)$, as provided by Parsaie *et al.* (2016), Haghiabi (2017), and Parsaie & Haghiabi (2017), is represented by the following equation

$$C(H) = (h + 1) + dH \tag{6}$$

where d is a penalty number for each basic function, $h + 1$ is a constant term representing a fixed value independent of the number of basic functions, and H is the number of basic functions in the MARS model.

Decision tree

(i) *Random forest*

The RF method is a well-organized collection of tree predictors constructed from input vectors using random vector samples; Breiman (2001), Liaw & Wiener (2002). The RF algorithm, Figure (2), is a well-known general-purpose classification and regression tool based on a hit-or-miss approach, with the variables taken from the best split (Biau & Scornet 2016). This

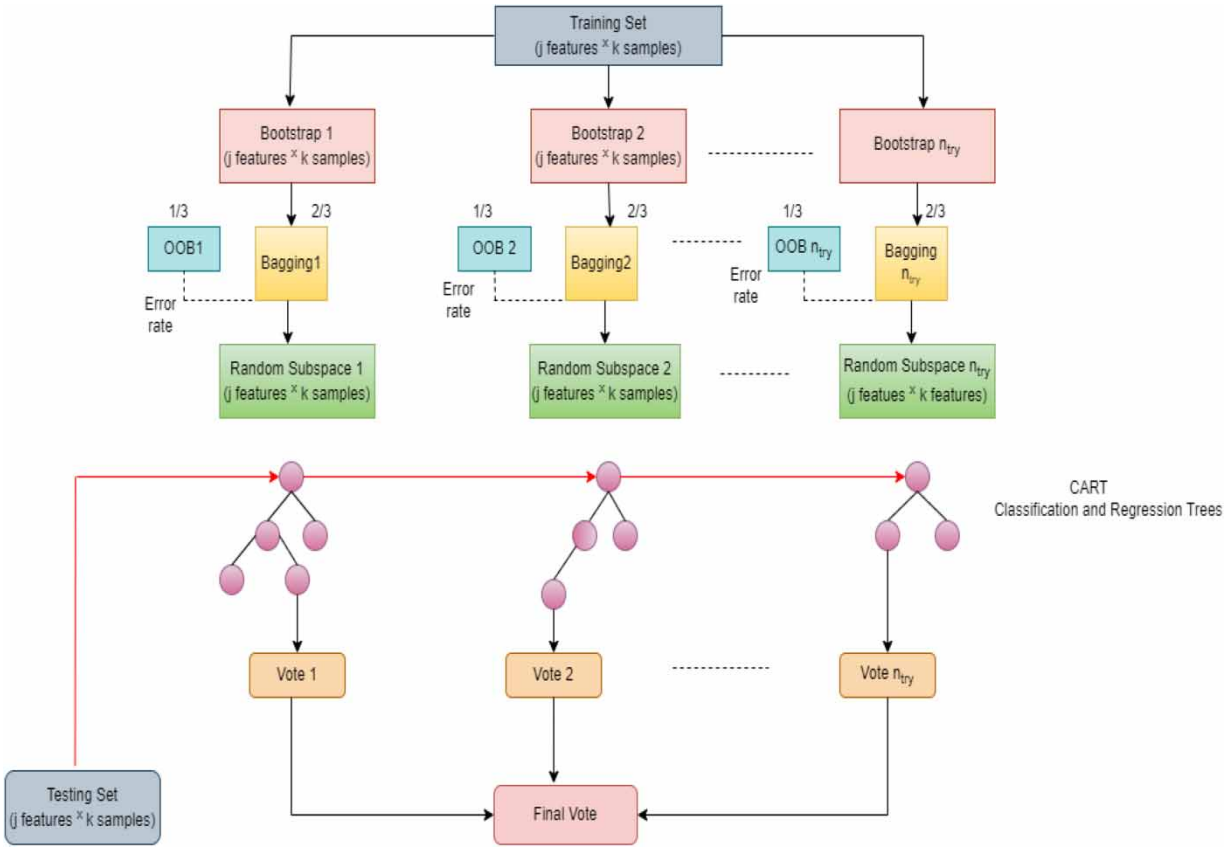


Figure 2 | Random forest algorithm [*source*: Han *et al.* (2018)].

method creates RFs by trapping a group of random trees (Mohanty *et al.* 2019). RF is a combination of bagging and random subspace that integrates weak classification trees and reaches a final decision by a majority vote. The splits for the forest trees are set based on the number of decision trees to be generated (N-tree) and the number of features (j) to be tested to determine the best split. Due to the relative efficiency of the RF classifier and its lack of overfitting, N-tree can be as large as possible (Guan *et al.* 2013). Each tree is grown using 67% of the training data, and the rest, 33%, known as out-of-bag (OOB) data, is used for validation. Hence, RF regression combines k trees, where k represents the number of trees to be produced, which can be any arbitrary value. The classification and regression tree (CART) algorithm generates all the n-trees in the forest without pruning. By combining different factors, RF regression allows the tree to grow to the depth of all the new training data. While creating specific trees, a training set of parameters is drawn from randomly selected data, and a Gini index is used to measure the impurity level in the parameters compared with the output (Breiman *et al.* 1984). RF classifies the variables based on their importance in achieving the optimum RF model.

(ii) *M5P*

Wang & Witten (1996) redesigned the M5 algorithm developed by Quinlan (1992) and proposed a modified algorithm called M5P. Model trees can efficiently handle massive datasets with many characteristics and dimensions and deal with missing data without creating ambiguity. By classifying various nodes into multiple subspaces, the M5P algorithm sets a linear regression at the terminal node and applies it to each sub-location of a multivariate linear regression model. The model tree in the M5P approach is generated in two stages. In the first stage, a decision tree is created using a splitting criterion. They behaved as class values reaching the node as quantifying the error. The M5P tree model algorithm calculates the expected error reduction by evaluating each attribute at that node as branching criteria (Quinlan 1992). The basic tree model is generated from the standard deviations of class values extending to nodes predicted by the separation criteria. The standard deviation approach measures the estimated error at the terminal node and generates linear functions at each node, making the data purer. The relation for the standard deviation ratio (SDR) is given by the following equation

$$SDR = sd(N) - \frac{\sum_{i=1}^{x} |N_i|}{|N|} \times sd(N) \tag{7}$$

where N is the number of samples, N_i is the number of ith samples having potential rise, and sd is the standard deviation.

In the second phase, the generated tree is pruned, as shown in Figure (3). To distinguish which branches should be pruned, the marginalized branches (terminal sections) are excluded in the final stretch, ensuring strong predictability. The new leaves are identified based on the distribution of observations employed in the learning process after they have been pruned and the predictability of the model is then frequently improved by this smoothing approach. A regularization technique is used to solve the anomalies within neighboring linear models in the tree's leaves at the final stage.

(iii) *Random tree*

When pruning is not used, the RT model selects every node based on a specific number of random features. RT barely involves machine learning, mainly exploiting arbitrary knowledge; otherwise, it applies a bagging concept (Hamoud *et al.* 2018). The preceding subsets of each node in the RF should be equally distributed among them. The technique addresses both classification and regression issues. In the classification process, the RT classifier takes the vector of the input property, classes it for every tree in the forest, and then extracts the group mark with the highest votes. A denial result of the classifier's response is the mean of the responses from all the trees in the forest (Cutler *et al.* 2012). Random trees combine two machine learning algorithms: single model trees and RF algorithm. In model trees, the linear layout of every leaf is tailored to its local subdomain, which considerably improves the efficiency of the single stable tree. RF's tree diversity is developed in two ways; at first, the training data are sampled, just like in bagging, by deleting each tree. Also secondly, while creating the tree the best part of that subset is decided by considering a single random subset of all attributes for each node. This contradicts the traditional method of dividing each node according to its potential optimal division. Random model trees are a modified model combining model trees and RFs. RTs use this result as a dividing criterion to simplify the optimization process, encouraging thoughtfully balanced trees with a spherical ridge environment running on all leaves (Barddal & Enembreck 2019).

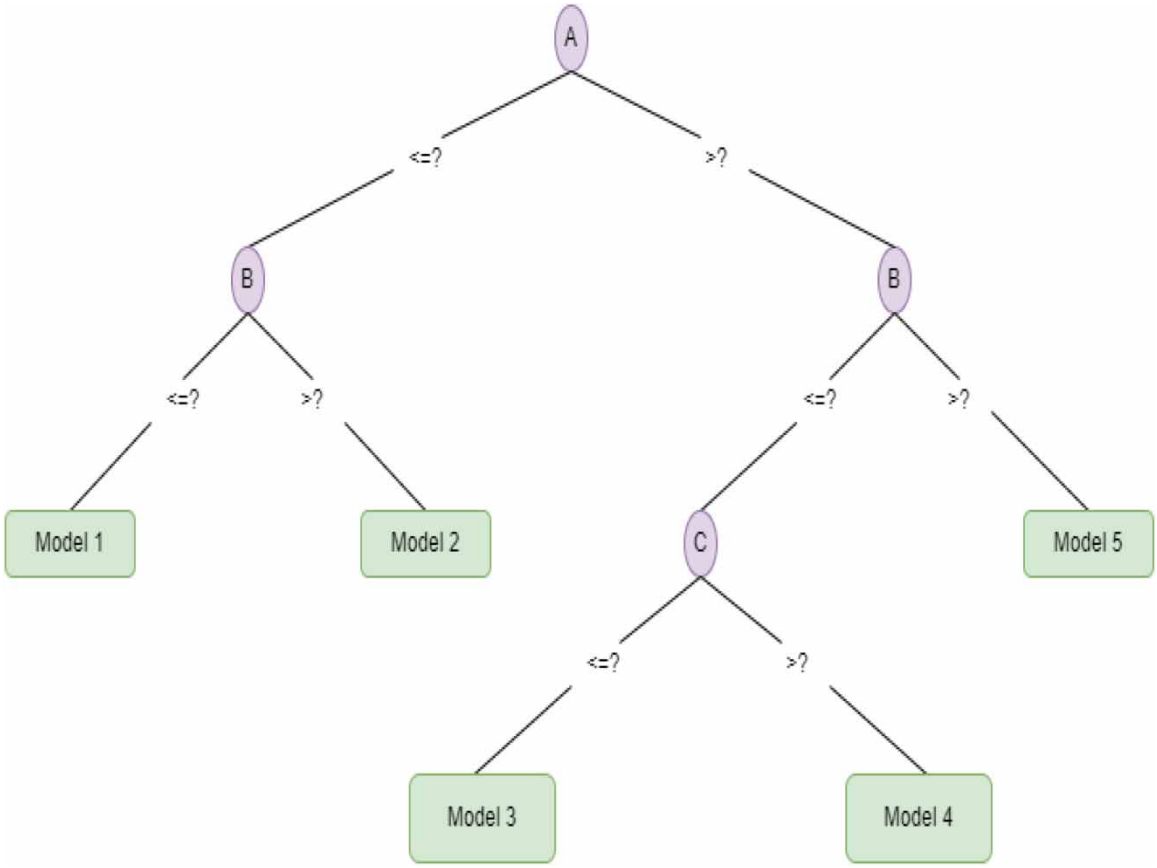


Figure 3 | Schematic representation of M5P pruned tree.

Gaussian process

Gaussian process regression (GPR) is a non-parametric Bayesian approach to regression. It can work well on small datasets and provide uncertainty measurements on the predictions; Quinonero & Rasmussen (2005) assumed that the Gaussian Process Regression could be defined by a linear relation, as shown in the following equation,

$$y = f(x) + \xi \tag{8}$$

where $\xi =$ Gaussian noise $\sim N(0, \sigma^2)$, which is described by a distribution with a zero mean.

The GPR model assumes that adjacent observations transfer information about each other and defines a prior explicitly across function space (Williams & Rasmussen 2006). GPR immediately represents a prior probability over a latent function. The function $f(x)$ is defined using a mean function $m(x)$, and kernel function $k(x, x')$, as given the following equation

$$f(x) = GP[m(x), k(x, x')] \tag{9}$$

The mean vector indicates the function is a central tendency, generally taken as zero (Kuss & Rasmussen 2003). The covariance matrix describes the function's structure and form.

For non-linear decision surfaces, GPR employs the kernel function. The literature presents several kernels, and the perusal of these studies suggests that polynomial kernels, Pearson VII universal kernels, and radial basis kernels perform better (Pal & Mather 2003; (Gill *et al.* 2006). The present study uses the following kernels,

(i) Polynomial (poly) kernel: $[k(x, y) = (x, y)^d]$

(ii) Pearson VII universal kernel (puk): $\left[k(x_i, x_j) = \left(\frac{1}{1 + \left[\left(2\|x_i - x_j\|^2\sqrt{2^{1/\omega} - 1}\right)/\sigma\right]^2}\right)^{\omega}\right]$

(iii) Radial basis kernel (rbf): $[k(x, y) = e^{-\gamma|x-y|^2}]$

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

User-defined parameters (noise, d, γ, σ, and ω) were selected based on a large number of trials that were carried out using different permutations.

Multiple non-linear regression equation

The following relation defines the general form of an MNRE model, as given in Equation (10)

$$y = aX_1^{b1} \times X_2^{b2} \times X_3^{b3} \ldots\ldots X_n^{bn} \tag{10}$$

where y is a dependent variable, i.e., output, $X_1, X_2, X_3 \ldots\ldots X_n$ are predictor variables, i.e., input, a is a constant coefficient, and parameters $b_1, b_2, b_3 \ldots\ldots b_n$ were calculated by minimizing the sum of squares of error in estimation based on the least squares method.

DATA STATISTICS OF GOVERNING PARAMETERS

The present study aimed to assess the applicability of various machine learning algorithms in predicting the aeration efficiency at weirs and identifying the best-performing model. This study used five shapes of sharp-crested weirs, i.e., suppressed, rectangular, trapezoidal, semicircular, and triangular, to evaluate their aeration efficiency. Experiments were performed under various governing parameters and compared to the aeration data when no weir was used.

A total of 126 experimental observations were collected for aeration efficiency at weirs under varying parameters such as head over weir (h), drop height (H), and discharge (Q). The range of parameters used for experimentation is provided in Table 1. It was seen from the experiment that the shape of the weir has a defining role in aeration efficiency. Thus shape factor was defined for weir shapes to inculcate the shapes of weir as an input in modeling, supplied as 1, 2, 3, 4, 5, and 6 for no weir, suppressed, rectangular, trapezoidal, semicircular, and triangular weirs, respectively.

The correlation matrix among governing variables was calculated using Pearson's method and is summarized in Table 2. This table reveals that shape factor has a higher correlation of 0.844 with aeration efficiency (E_{20}) followed by the head of over weir, discharge, and drop height. Thus, the shape factor is the most influential parameter for estimating aeration efficiency.

Table 1 | Range of parameters used for experimentation

Parameter	Notation	Unit	Range
Head over weir	h	cm	1–5.5
Drop height	H	cm	60, 75, and 90
Discharge	Q	l/s	0–6

Table 2 | Correlation matrix among explanatory variables

Variables	Shape factor	h (cm)	H (cm)	Q (l/s)	E_{20}
Shape factor	1				
h (cm)	0	1			
H (cm)	0	0	1		
Q (l/s)	−0.010	0.994	0.013	1	
E_{20}	0.844	0.32349	0.265	−0.313	1

Out of 126 experimental data, 88 were used in the training stage for model generation, and 38 were used for model validation. In the present computational analysis, shape factor, h (cm), H (cm), and Q (l/s) were considered as the input parameters, whereas aeration efficiency was regarded as the output/target parameter. Table 3 summarizes the descriptive data statistics of model components such as minimum, maximum, mean, standard deviation, kurtosis, skewness, and confidence level (95%) for training and validating data sets.

Performance evaluation indices

Five statistical parameters were employed to evaluate the accuracy of machine learning (ML) based models. These statistical parameters are coefficient of correlation (*CC*), mean absolute error (*MAE*), root mean square error (*RMSE*), Nash Sutcliffe model efficiency (*NSE*), and *SI*. These statistical parameters quantify the best fit between the observed and predicted data for all the applied models. The formulas of these indicators are as listed in Equations (11)–(15).

(i)

$$\text{Correlation coefficient } (CC) = \frac{\sum_{i=1}^{N}(R_i - \bar{R})(Q_i - \bar{Q})}{\sqrt{\sum_{i=1}^{N}(R_i - \bar{R})^2 (Q_i - \bar{Q})^2}}, \quad -1 \le CC \le 1 \tag{11}$$

(ii)

$$\text{Mean absolute error } (MAE) = \frac{1}{N}\sum_{i=1}^{N}|R_i - \bar{R}| \tag{12}$$

(iii)

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

(iv)

$$\text{Nash Sutcliffe model efficiency } (NSE) = 1 - \frac{\sum_{i=1}^{N}(R_i - Q_i)^2}{\sum_{i=1}^{N}(Q_i - \bar{Q})^2}, \quad -\infty \le NSE \le 1 \tag{14}$$

Table 3 | Descriptive statistics of governing parameters for training and validating data sets

Statics	Shapes	*h* (cm)	*H* (cm)	*Q* (l/s)	E_{20}	Data set
Minimum	1	1	60	0.486	0.0113	Training
	1	1	60	0.486	0.0112	Validating
Maximum	6	5.33	90	5.981	0.489	Training
	6	5.33	90	5.981	0.432	Validating
Mean	3.511	3.240	75.170	3.048	0.205	Training
	3.474	3.311	74.605	3.135	0.203	Validating
Standard Deviation	1.800	1.504	12.351	1.898	0.139	Training
	1.520	1.515	12.323	1.938	0.122	Validating
Kurtosis	−1.381	−1.402	−1.526	−1.443	−0.994	Training
	−0.968	−1.343	−1.515	−1.401	−0.999	Validating
Skewness	−0.019	−0.116	−0.021	0.087	0.209	Training
	0.053	−0.094	0.050	0.138	0.122	Validating
Confidence Level (95%)	0.381	0.319	2.617	0.402	0.029	Training
	0.499	0.498	4.051	0.637	0.040	Validating

(v)

$$\text{Scattering Index } (SI) = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(R_i - Q_i)^2}}{\bar{Q}} \tag{15}$$

where Q refers to the actual values or values obtained, $\bar{Q}$ refers to the average of observed values, R is the predicted value (model), N is the number of observation.

Specification of software and workstation

The models used for this study were trained and validated to assess their ability to predict aeration efficiency based on four input parameters: shape factor, h, H, and discharge. All the models used in this study were generated using MATLAB 2013a and WEKA 3.9 software on a Dell Vostro 2520 (Intel® CoreTM i3-2348H*Q* CPU, 2.30 GHz (four CPUs), 2048 MB RAM, Windows 8.1 Pro 64-bit) PC.

RESULTS AND DISCUSSION

Implementation and assessment of the MARS model

The dataset from Table 3 was used to develop the MARS model. During MARS model calibration, 15 basic functions were chosen initially, but after pruning non-responsive BFs, an optimal MARS model was generated with six BFs. The general form developed for the MARS model is provided in Equation (16). Table 4 shows the related coefficients involved in Equation (16)

$$E_{20} = 0.11848 + \sum_{M=1}^{8} \beta_m BF_i(x) \tag{16}$$

The pruning measure of 0.0012 in the MARS model was introduced through the GVC constraint. Figures 4(a) and 4(b) compare experimental data with the results of the MARS model for the training and validating stages, respectively. The MARS model can predict the aeration efficiency at weirs with some dependability as the predicted aeration efficiency values are within the $\pm 30\%$ error line. But a few values are outside the $\pm 30\%$ error line, indicating that the model overestimates the aeration efficiency values. Performance evaluation parameters suggest that MARS is suitable for the prediction of aeration efficiency at weirs with *CC* (0.9780 and 0.9519), *MAE* (0.0214 and 0.0284), *RMSE* (0.0289 and 0.0370), *SI* (0.1412 and 0.1823), and *NSE* (0.9565 and 0.9053) values in the training and validating stages, respectively.

Implementation and assessment of the decision tree models

Decision tree models are a cut-and-try process where several trials are attempted to attain the output with optimal user-defined parameters. This study investigated M5P, RF, and RT models for their ability to predict the aeration efficiency at weirs. In M5P, the optimum value user-defined parameter: number of instances was 1. The optimum user-defined parameters used in this study were: the number of features $(m) = 1$ and the number of trees $(k) = 10$. The structure of the M5P based

Table 4 | Basic functions and related coefficients of the MARS model

Sr. No.	Basic function	β_m
1	$BF_1 = \max(0,\ \text{shape factor} - 2)$	+0.038
2	$BF_1 = \max(0,\ 2 - \text{shape factor})$	−0.080
3	$BF_1 = \max(0, h - 2.5)$	−0.016
4	$BF_1 = \max(0,\ 2.5 - h)$	+0.071
5	$BF_1 = BF_1 \times \max(0,\ H - 60)$	+0.001
6	$BF_1 = BF_2 \times \max(0,\ 1.92 - Q)$	−0.089

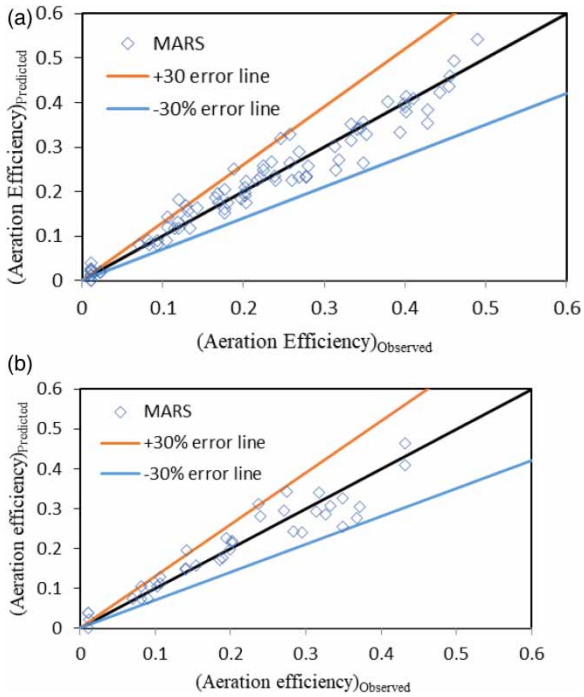


Figure 4 | Plot for observed and predicted aeration efficiency at weirs using the MARS model for (a) training stage and (b) validating stage.

model developed for predicting aeration efficiency (E_{20}) at weirs is shown in Figure 5. Linear equations developed using the M5P based model to predict aeration efficiency (E_{20}) at weirs are listed in Table 5.

The scatter plots for predicted and observed aeration efficiency values at weirs for training and validating stages using the decision tree (M5P, RF, and RT) models are shown in Figure 6. The RF model performed better in predicting aeration

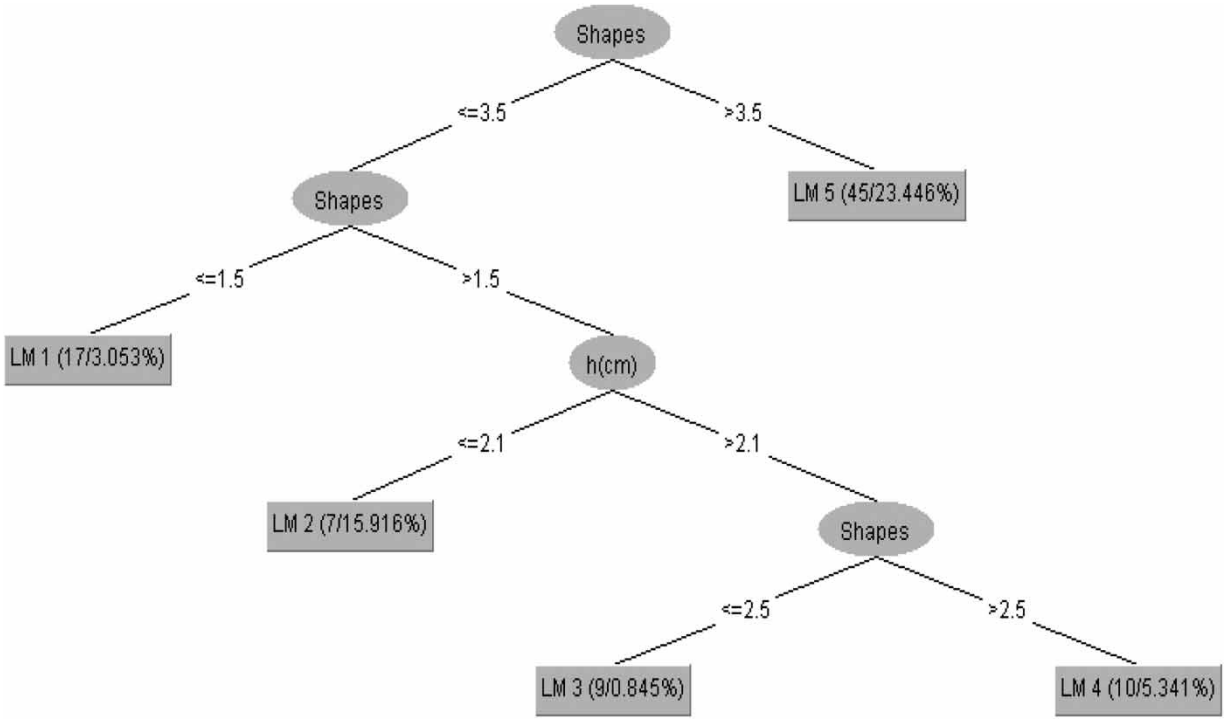


Figure 5 | M5P model developed for predicting aeration efficiency (E_{20}) at weirs.

Table 5 | Linear equations developed using the M5P model

Sr. No.	Linear model number	Linear equation
1	LM:1	$E_{20} = 0.0453 \times \text{shape factor} - 0.0139h + 0.0011H - 0.0582$
2	LM:2	$E_{20} = 0.0658 \times \text{shape factor} - 0.0227h + 0.0016H - 0.0488$
3	LM:3	$E_{20} = 0.0581 \times \text{shape factor} - 0.0230h + 0.0017H - 0.0610$
4	LM:4	$E_{20} = 0.0577 \times \text{shape factor} - 0.0224h + 0.0019H - 0.0725$
5	LM:5	$E_{20} = 0.0634 \times \text{shape factor} - 0.0327h + 0.0042H - 0.2280$

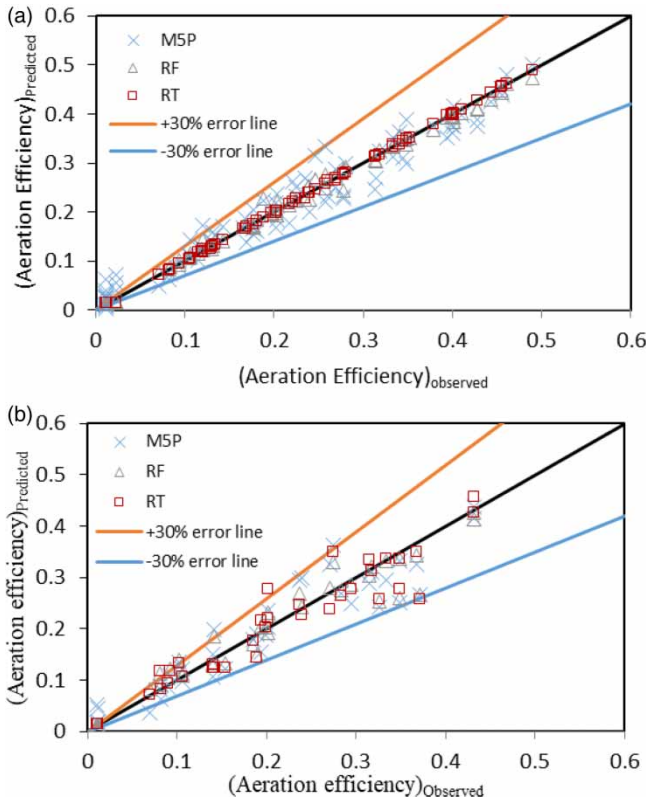


Figure 6 | Plot for observed and predicted aeration efficiency at weirs using decision tree for (a) training data stage and (b) validating data stage.

efficiency at weirs than all the decision tree used in this study. Most of the predicted values of the RF model lie at the line of agreement at the training stage, and in the validating stage, the predicted values are within the $\pm 30\%$ error line.

The values of performance evaluation indices for decision tree models are represented in Table 6. The ability of the RF model to predict the aeration efficiency at weirs is better than the other decision tree model used, with its *CC* (0.9976 and 0.9653), *MAE* (0.0066 and 0.0213), *RMSE* (0.0098 and 0.0322), *SI* (0.0479 and 0.1585), and *NSE* (0.9950 and 0.9285) values in the training and validating stages, respectively.

Implementation and assessment of the Gaussian process model

The generation of the GP model is like the MARS and decision tree models. A number of cuts and tries were performed to attain the optimum values of user-defined parameters. Polynomial kernel, Pearson VII Universal kernel, and radial basis kernel were employed to predict the aeration efficiency at weirs. The scatter plots for predicted and observed aeration efficiency values at weirs for training and validating phases using GP models are shown in Figure 7. The predicted aeration

Table 6 | Performance of the decision tree models in training and validating stages

	Values				
Models	***CC***	***MAE***	***RMSE***	***NSE***	***SI***
Training data set					
M5P	0.9778	0.0240	0.0291	0.9560	0.1420
RF	0.9976	0.0066	0.0098	0.9950	0.0479
RT	0.9998	0.0015	0.024	09997	0.0120
Validating data set					
M5P	0.9383	0.0317	0.0418	0.8793	0.2059
RF	0.9653	0.0213	0.0322	0.9285	0.1585
RT	0.9586	0.0227	0.0345	0.9178	0.1699

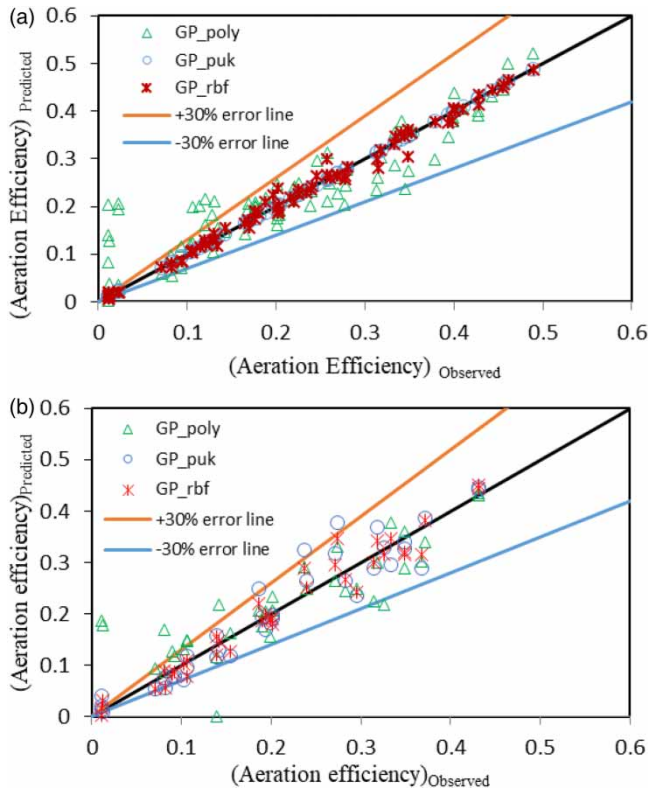


Figure 7 | Plot for observed and predicted aeration efficiency at weirs using GP models for (a) training stage and (b) validating stage.

efficiency values of the GP_rbf model lie at/closer to the line of agreement in training, and they lie within the $\pm 30\%$ error line in validating stage. Thus, ensuring the better predictability of the GP_rbf based model than GP_poly and GP_puk based models in predicting the aeration efficiency at weirs.

The performance evaluation indices values for GP models are presented in Table 7. The GP_rbf model performs better than Gp_poly and GP_puk in predicting the aeration efficiency at weirs with its *CC* (0.9961 and 0.9973), *MAE* (0.0079 and 0.0195), *RMSE* (0.0122 and 0.0251), *SI* (0.0594 and 0.1238), and *NSE* (0.9923 and 0.9564) values in the training and validating stages, respectively.

Table 7 | Performance of the GP models for training and validating stages

	Values				
Models	***CC***	***MAE***	***RMSE***	***NSE***	***SI***
Training data set					
GP-poly	0.9243	0.0375	0.0537	0.8494	0.2626
GP-puk	1.0000	0.0003	0.0003	1.0000	0.0016
GP-rbf	0.9961	0.0079	0.0122	0.9923	0.0594
Validating data set					
GP-poly	0.8604	0.0448	0.0617	0.7368	0.3040
GP-puk	0.9595	0.0259	0.0355	0.9132	0.1746
GP-rbf	0.9793	0.0195	0.0251	0.9564	0.1238

Implementation and assessment of the MNRE model

In this study, a non-linear regression based model was developed based on the least square method using the training data set. The equation for the developed MNRE model is listed as Equation (17).

$$E_{20} = 0.00484 \times \text{shape factor}^{1.054} h^{1.2567} H^{0.4601} Q^{-1.044} \tag{17}$$

The performance of the MNRE-based model is suitable for predicting the aeration efficiency (E_{20}) at weirs with its *CC* (0.9492 and 0.9143), *MAE* (0.0379 and 0.0393), *RMSE* (0.0450 and 0.497), *SI* (0.2201 and 0.2450), and *NSE* (0.8942 and 0.8291) values in the training and validating stages, respectively. The plots for observed aeration efficiency and the predicted values by MNRE for the training and validating phases are given in Figure 8. The predicted values of aeration efficiency by the

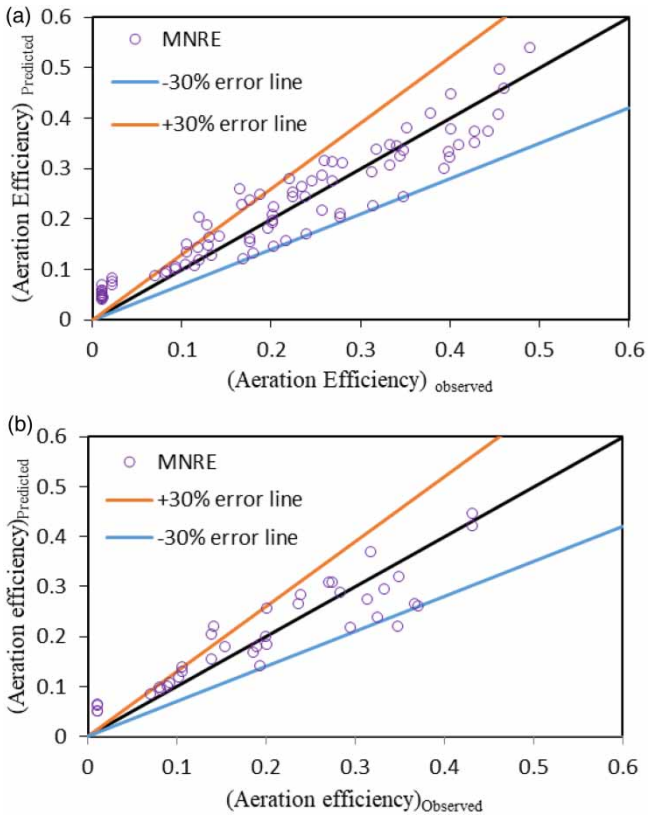


Figure 8 | Plot for observed and predicted aeration efficiency at weirs using the MNRE model for (a) training stage and (b) validating stage.

MNRE model lie outside the ±30% error line for both the training and validating stage. Thus this model overestimates the values of aeration efficiency.

Inter-comparison of the best-performing models

A comparison of the accuracy of the best-performing models (MARS, MNRE, decision trees, and GP) used for this study in predicting the aeration efficiency at weirs is summarized in Table 8. The higher values of *CC* and *NSE* and lower values of *MAE*, *RMSE*, and *SI* indicate that the GP_rbf model outperforms other models in this study.

Figure 9 compares the obtained values of aeration efficiency with the predicted values provided by MARS, RF, GP_rbf, and MNRE models at the training and validating stages. The results indicate that the predicted values of the aeration efficiency by GP_rbf are at/closer to the line of agreement in both the training and validating stages. Thus, it can predict aeration efficiency at weirs with greater accuracy than other tested models. The predicted values by GP_rbf lie within the ±30% error line in both the training and validating stages, with maximum values at/close to the agreement line in the training stage.

Comparison with the literature

In the earlier studies to predict aeration efficiency of triangular weir using the ML model, Goel (2013) used support vector machines (SVMs); Jaiswal & Goel (2020) used GP and M5P models. In both studies, the models used have provided promising results in predicting aeration efficiency at triangular weirs. The present research accesses the performance of several other ML models, including GP and M5P, to predict aeration efficiency at different shapes of weirs. The weir shapes were also taken as input parameters in creating the models by introducing a shape factor. The performance of various ML models applied in the above mentioned literature is presented in Table 9 in terms of *CC* and *RMSE* value for comparison.

The results presented in Tables 6–9 show that the performance of the best-performing models in the literature (M5P; GP_rbf and GP_puk) decreased when the shape factor was introduced as an input parameter in building the model. Wherein GP_rbf outperformed all other models and provided a promising result in predicting the outcomes while taking shape factor as an input parameter with a *CC* and *RMSE* value of 0.9961 and 0.0122, respectively.

STATISTICAL AND GRAPHICAL APPROACHES FOR COMPARING THE MODELS

Analysis of variance through *F*-test

Statistics suggest that ratios of the variances of the samples in each pair must follow the same distribution. The single-factor analysis of variance (ANOVA) test is used to verify whether or not the means of three or more independent groups are equal. The insignificant variation between the two samples is confirmed if the F-value $< F_{critcal}$ and the P-value $> \alpha$ (=0.05). Single-factor ANOVA was performed and is summarized in Table 10. The test results indicate an insignificant variation between observed and predicted values for all the models tested in this paper, having F-values $< F_{critical}$, and P-values > 0.05 in all groups.

Table 8 | Performance evaluation of the MARS, RF, GP_rbf, and MNRE models for training and validating stages

	Values				
Models	***CC***	***MAE***	***RMSE***	***NSE***	***SI***
Training data set					
MARS	0.9780	0.0214	0.0289	0.9565	0.1412
RF	0.9976	0.0066	0.0098	0.9950	0.0479
GP-rbf	0.9961	0.0079	0.0122	0.9923	0.0594
MNRE	0.9492	0.0379	0.0450	0.8942	0.2201
Validating data set					
MARS	0.9519	0.0284	0.0370	0.9053	0.1823
RF	0.9653	0.0213	0.0322	0.9285	0.1585
GP-rbf	0.9793	0.0195	0.0251	0.9564	0.1238
MNRE	0.9143	0.0393	0.0497	0.8291	0.2450

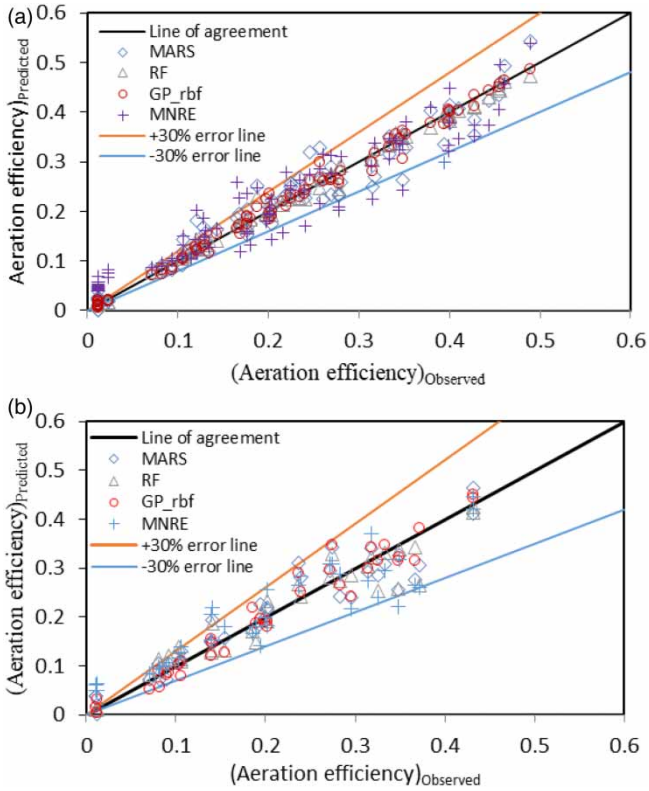


Figure 9 | A plot between observed and predicted values of aeration efficiency at weirs using MARS, RF, GP_rbf, and MNRE for (a) training stage and (b) validating stage.

Table 9 | Predictability of various ML models used in the literature to predict aeration efficiency

Study	Shape of weir	ML model	*CC*	*RMSE*
Goel (2013)	Triangular	SVM (POLY)	0.9828	0.0245
		SVM (RBF)	0.9656	0.0821
		Linear regression	0.9822	0.0249
Jaiswal & Goel (2020)	Triangular	GP_npoly	0.9897	0.0184
		GP_poly	0.9252	0.0496
		GP_puk	0.9998	0.0025
		GP_rbf	0.9998	0.0031
		M5P	0.9998	0.0023

Table 10 | Single-factor ANOVA results among observed and predicted values using all applied models

Source of variation	*F*-value	*P*-value	$F_{critical}$	Insignificant variation between groups
Between observed and MARS	0.000103	0.991939	3.970229	✓
Between observed and M5P	0.022449	0.881306	3.970229	✓
Between observed and RF	0.015778	0.900380	3.970229	✓
Between observed and RT	0.019649	0.888901	3.970229	✓
Between observed and GP_poly	0.055451	0.814486	3.970229	✓
Between observed and GP_puk	2.6E – 09	0.999959	3.970229	✓
Between observed and GP_rbf	0.003854	0.950669	3.970229	✓
Between observed and MNRE	0.00115	0.973035	3.970229	✓

Interquartile range analysis

To evaluate the inconsistency in the estimation of the aeration efficiency at weirs, the 25th, 50th, and 75th percentile values of the observed and predicted aeration efficiency at weirs by the various models were assessed, as tabulated in Table 11. The interquartile range (IQR) is the difference between the 75th and 25th percentiles of the sample. The IQR of values predicted by the GP_rbf based model is in line with the IQR of the observed values. Thus, it confirms that GP_rbf predicts the aeration efficiency at the weirs with greater accuracy than the other models discussed.

Graphical methods

Two graphical methods, the Whisker plot and Taylor's diagram, were used to assess the accuracy of models used in this study in predicting the aeration efficiency at weirs.

(i) *Whisker plot*

A whisker plot displays the minimum, maximum, 25th, 50th, and 75th percentile summary of a set of data. The 25th and 75th percentile values are represented as the lower and upper quartiles of the box, and the 50th percentile values are expressed as the median of the box. A vertical line extending parallel to the boxes indicates variability outside the upper and lower quartiles. The whisker plot, shown in Figure 10 for the validating stage, suggests that the GP_rbf model has the depth of the higher and lower end of boxes almost the same as that of the observed values.

(ii) *Taylor's diagram*

Taylor's diagram popularly used graphical representations to compare models based on their standard deviation from the observed values. It stipulates the degree of correspondence between the observed and predicted aeration efficiency values based on *CC*, *RMSE*, and standard deviation values. In Taylor's diagram, the model near the observed point is the

Table 11 | Quantitative statistics of observed values and predicted values by MARS, RF, GP_rbf, and MNRE models

Statistic	Observed	MARS	RF	GP_rbf	MNRE
Minimum	0.0112	0.0015	0.0120	0.0010	0.0502
Maximum	0.4320	0.4637	0.4260	0.4510	0.4450
1st Quartile (=25%)	0.1038	0.1069	0.1158	0.0930	0.1200
Median (=50%)	0.1968	0.2062	0.1980	0.1905	0.2020
3rd Quartile (=75%)	0.3096	0.2910	0.2788	0.3110	0.2721
IQR	0.2058	0.1841	0.1630	0.2180	0.1521

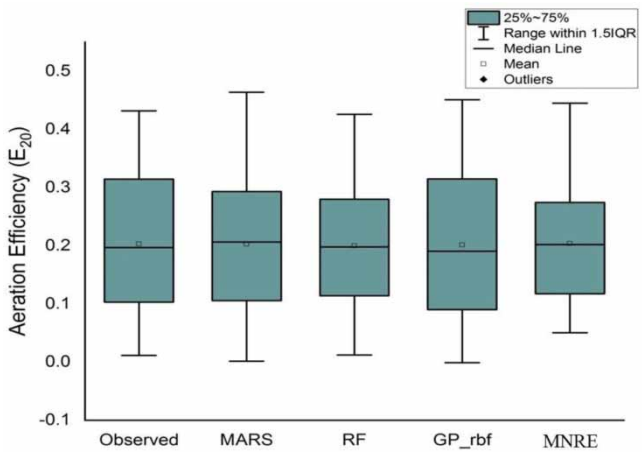


Figure 10 | Whisker plot for the observed and predicted values for the validating stage.

best-performing model. As in Figure 11, the GP_rbf model is closer to the observed point; thus, it has better accuracy in predicting the aeration efficiency than all the models used. The performance of MNRE had the lowest prediction accuracy among MARS-, RF-, and GP_rbf-based models.

Sensitivity analysis

Sensitivity analysis determines how the target variable is affected by changes in input variables. The best-performing model (GP_rbf) was used to observe the influence on its predictability by removing any input parameters. The sensitivity in predicting the aeration efficiency (E_{20}) at weirs values is investigated by examining the response of each input parameter to the output. An input parameter was removed from the training data, and the remaining input combination was supplied to the GP_rbf model. The variations in Performance Evaluation Indices were obtained for each step where any input parameter is removed from the training data set, as shown in Table 12. The shape factor is the most prominent parameter in estimating the aeration efficiency with the highest variation in Performance Evaluation Indices ($CC = 0.3812$, $RMSE = 0.1138$, and $MAE = 0.0895$). The drop height is found to be the second most prominent variable after the shape factor.

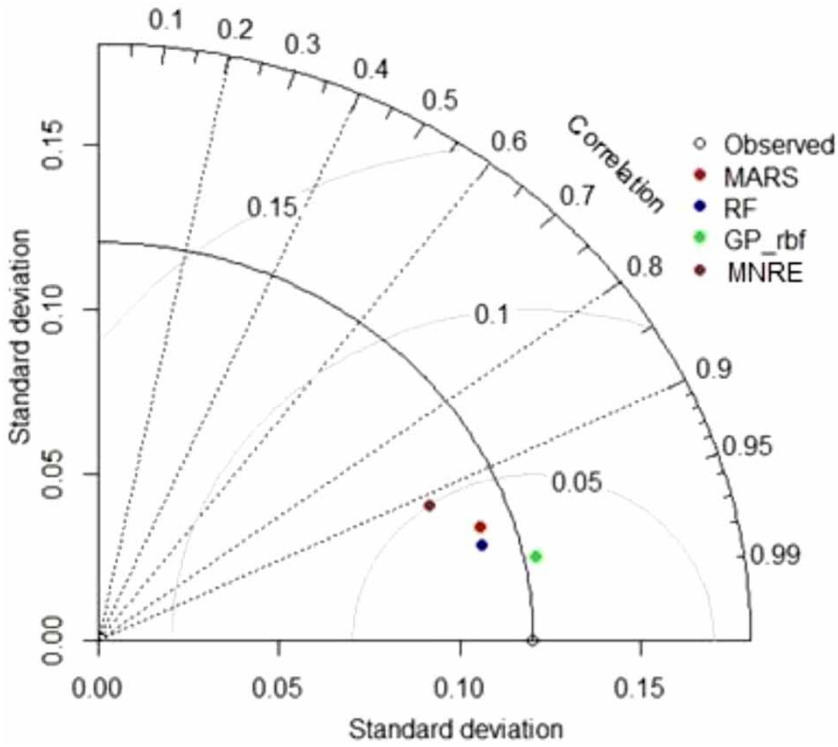


Figure 11 | Taylor's diagram among observed and predicted values for the testing stage.

Table 12 | Sensitivity analysis for parametric variation using the GP_rbf model

Input variable				Target	GP_rbf model		
Shape factor	*h* (cm)	*H* (cm)	*Q* (l/s)	E_{20}	*CC*	*MAE*	*RMSE*
✓	✓	✓	✓	✓	0.9793	0.0195	0.0251
✗	✓	✓	✓	✓	0.3812	0.0895	0.1138
✓	✗	✓	✓	✓	0.9844	0.0165	0.0218
✓	✓	✗	✓	✓	0.9198	0.0341	0.0474
✓	✓	✓	✗	✓	0.9854	0.0165	0.0209

CONCLUSIONS

This investigation aimed to study the ability of various models (MARS, decision tree, and GP) to predict the aeration efficiency at weirs. The same is compared with the results of the MNRE. Various graphical presentations and goodness-of-fit parameters were used to assess the performance of the models used in this study. According to performance evaluation results, the GP_rbf model outperformed the other implemented models in predicting the aeration efficiency at weirs with more promising values of goodness-of-fit parameters. Another significant outcome was that the RF model performed best among other decision tree models used in this study. Furthermore, the performance of the GP_poly model was the worst among all models used. Based on the sensitivity analysis results using the GP_rbf model, the shape factor was the most prominent parameter, followed by drop height in current data.

Through this study, it is found that all the models used to predict the aeration efficiency at weirs are performing well with *CC* values greater than 0.95, except GP_poly, with a *CC* value of 0.86. Out of all the models studied, the GP_rbf model provides the most promising predicted values of aeration efficiency at weirs.

Machine learning models proved their effectiveness in predicting the aeration efficiency at weirs without performing any new experimentation on a similar setup designed to measure aeration efficiency. Using ML models, missing values in experimental data can also be predicted, and incorrect data can be identified. The current study can be extended further with advanced hybrid models (adaptive neuro-fuzzy inference system and genetic algorithm (GA) or particle swarm optimization (PSO), support vector regression (SVR) and GA or PSO, etc.) to assess their applicability in predicting aeration efficiency at weirs. Furthermore, a more extensive range of experimental and field aeration data at the weirs with a larger range of discharges and drop heights should be gathered to refine the accuracy of predictions of the models used for this study. For more precision in prediction of ML models, optimization techniques can be utilized to seek the optimal values of input parameters in order to get highest value of aeration efficiency.

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

Apted, R. W. & Novak, P. 1973 Some studies of oxygen uptake at weirs. In: *Proceedings of the XV Congress, IAHR Paper B23*, Istanbul, pp. 177–186.

Avery, S. T. & Novak, P. 1978 Oxygen transfer at hydraulic structures. *Journal of the Hydraulics Division, ASCE* **104** (11), 1521–1540.

Barddal, J. P. & Enembreck, F. 2019 Learning regularized hoeffding trees from data streams. In: *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*. pp. 574–581.

Baylar, A., Bagatur, T. & Tuna, A. 2001a Aeration performance of triangular notch weirs at recirculating system. *Water Quality Research Journal of Canada* **36** (1), 121–132.

Baylar, A., Bagatur, T. & Tuna, A. 2001b Aeration performance of triangular-notch weirs. *Journal of the Chartered Institution of Water and Environmental Management* **15** (3), 203–206.

Baylar, A., Unsal, M. & Ozkan, F. 2010 Hydraulic structures in water aeration processes. *Water, Air, & Soil Pollution* **210** (1), 87–100.

Biau, G. & Scornet, E. 2016 A RF guided tour. *Test* **25** (2), 197–227.

Breiman, L. 2001 Random forests. *Machine Learning* **45** (1), 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. 1984 Classification and regression trees. In: *Wadsworth Statistics. Probability Series*. (L. Breiman, ed.) Wadsworth, Belmont, California.

Cutler, A., Cutler, D. R. & Stevens, J. R. 2012 Random forests. In: *Ensemble Machine Learning*. (C. Zhang & Y. Ma, eds.) Springer, Boston, MA, pp. 157–175.

Eckenfelder Jr., W. W. & Ford, D. L. 1970 *Water Pollution Control*. Pemberton Press, Austin and New York.

Gill, M. K., Asefa, T., Kemblowski, M. W. & McKee, M. 2006 Soil moisture prediction using support vector machines 1. *JAWRA Journal of the American Water Resources Association* **42** (4), 1033–1046.

Goel, A. 2013 Modeling aeration of sharp crested weirs by using support vector machines. *WASET International Journal of Mechanical, Aerospace, Industrial, Mechatronic and Manufacturing Engineering* **7** (12), 2620–2625.

Guan, H., Li, J., Chapman, M., Deng, F., Ji, Z. & Yang, X. 2013 Integration of orthoimagery and Lidar data for object-based urban thematic mapping using random forests. *International Journal of Remote Sensing* **34** (14), 5166–5186.

Gulliver, J. S. & Rindels, A. J. 1993 Measurement of air-water oxygen transfer at hydraulic structures. *Journal of Hydraulic Engineering* **119** (3), 327–349.

Gulliver, J. S., Thene, J. R. & Rindels, A. J. 1990 Indexing gas transfer in self-aerated flows. *Journal of Environmental Engineering ASCE* **116** (3), 503–523.

Gulliver, J. S., Wilhelms, S. C. & Parkhill, K. L. 1998 Predictive capabilities in oxygen transfer at hydraulic structures. *Journal of Hydraulic Engineering* **124** (7), 664–671.

Haghiabi, A. H. 2017 Prediction of river pipeline scour depth using multivariate adaptive regression splines. *Journal of Pipeline Systems Engineering and Practice* **8** (1), 04016015.

Hamoud, A., Hashim, A. S. & Awadh, W. A. 2018 Predicting student performance in higher education institutions using decision tree analysis. *International Journal of Interactive Multimedia and Artificial Intelligence* **5**, 26–31.

Han, T., Jiang, D., Zhao, Q., Wang, L. & Yin, K. 2013 Comparison of random forest, artificial neural networks and support vector machine for intelligent diagnosis of rotating machinery. *Transactions of the Institute of Measurement and Control* **40** (8), 2681–2693.

Jaiswal, A., Goel, A., 2019 Aeration through weirs – a critical review. In: *Sustainable Engineering. Lecture Notes in Civil Engineering* (Agnihotri, A., Reddy, K. & Bansal, A., eds). Springer, Singapore, p. 30.

Jaiswal, A., Goel, A., 2020 Evaluation of aeration efficiency of triangular weirs by using Gaussian process and M5P approaches. In: *Advanced Engineering Optimization Through Intelligent Techniques. Advances in Intelligent Systems and Computing* (Venkata Rao, R. & Taler, J., eds). Springer Singapore, p. 949.

Liaw, A. & Wiener, M. 2002 Classification and regression by random forest. *R News* **2** (3), 18–22.

Kuss, M. & Rasmussen, C. 2003 Gaussian processes in reinforcement learning. In:. *Advances in neural information processing systems* (S. Thrun, L. Saul & B. Schölkopf, eds.). MIT Press. Cambridge, MA.

Mohanty, S., Roy, N., Singh, S. P. & Sihag, P. 2019 Estimating the strength of stabilized dispersive soil with cement clinker and fly ash. *Geotechnical and Geological Engineering* **37** (4), 2915–2926.

Nakasone, H. 1987 Study of aeration at weirs and cascades. *Journal of Environmental Engineering, ASCE* **113** (1), 64–81.

Pal, M. & Mather, P. M. 2003 An assessment of the effectiveness of decision tree methods for land cover classification. *Remote Sensing of Environment* **86** (4), 554–565.

Parsaie, A. & Haghiabi, A. H. 2017 Improving modelling of discharge coefficient of triangular labyrinth lateral weirs using SVM, GMD*H* and MARS techniques. *Irrigation and Drainage* **66** (4), 636–654.

Parsaie, A., Haghiabi, A. H., Saneie, M. & Torabi, H. 2016 Prediction of energy dissipation on the stepped spillway using the multivariate adaptive regression splines. *ISH Journal of Hydraulic Engineering* **22** (3), 281–292.

Quinlan, J. R. 1992 Learning with continuous classes. In: *5th Australian Joint Conference on Artificial Intelligence*. Vol. 92, pp. 343–348.

Quinonero-Candela, J. & Rasmussen, C. E. 2005 A unifying view of sparse approximate Gaussian process regression. *The Journal of Machine Learning Research* **6**, 1939–1959.

Van der Karoon, G. T. & Schram, A. H. 1969a Weir aeration – part I: single free fall. *H2O* **2** (22), 528–537.

Van der Karoon, G. T. N. & Schram, A. H. 1969b Weir aeration – part II. *H2O* **22**, 538–545.

Wang, Y. & Witten, I. H. 1996 *Induction of Model Trees for Predicting Continuous Classes. Computer Science Working Paper. Report No. 96/23*.

WEKA Software. Waikato Environment for Knowledge Analysis, by the University of Waikato, Hamilton, New Zealand. Available from: https://www.cs.waikato.ac.nz/ml/weka.

Williams, C. K. & Rasmussen, C. E. 2006 *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, Vol. 2, No. 3, p. 4.

Witt, A. M. & Gulliver, J. S. 2012 Predicting oxygen transfer efficiency at low-head gated sill structures. *Journal of Hydraulic Research* **50** (5), 521–531.

First received 11 January 2023; accepted in revised form 19 April 2023. Available online 4 May 2023