scientific reports

OPEN

Toward the accurate estimation of elliptical side orifice discharge coefficient applying two rigorous kernel-based data-intelligence paradigms

Masoud Karbasi[1], Mehdi Jamei[2], Iman Ahmadianfar[3] & Amin Asadi[4,5]✉

In the present study, two kernel-based data-intelligence paradigms, namely, Gaussian Process Regression (GPR) and Kernel Extreme Learning Machine (KELM) along with Generalized Regression Neural Network (GRNN) and Response Surface Methodology (RSM), as the validated schemes, employed to precisely estimate the elliptical side orifice discharge coefficient in rectangular channels. A total of 588 laboratory data in various geometric and hydraulic conditions were used to develop the models. The discharge coefficient was considered as a function of five dimensionless hydraulically and geometrical variables. The results showed that the machine learning models used in this study had shown good performance compared to the regression-based relationships. Comparison between machine learning models showed that GPR (RMSE = 0.0081, R = 0.958, MAPE = 1.3242) and KELM (RMSE = 0.0082, R = 0.9564, MAPE = 1.3499) models provide higher accuracy. Base on the RSM model, a new practical equation was developed to predict the discharge coefficient. Also, the sensitivity analysis of the input parameters showed that the main channel width to orifice height ratio (B/b) has the most significant effect on determining the discharge coefficient. The leveraged approach was applied to identify outlier data and applicability domain.

Water diversion structures comprised of side weirs, side intakes, side orifices, and sluice gates are the most crucial and widespread devices in sewerage, irrigation, and agricultural systems which controllably divert and transfer the flowing water from the main channel to a tributary channel[1]. The flow regime in water diversion devices is categorized into the spatially varied flow with decreasing flow discharge[2,3]. In the past four decades, numerous investigations have been conducted on hydraulic characteristics side structures, including side wires[1,4–8], and Side Sluice Gate[9–12].

As one of the most important diversion structures, the side orifices are usually placed on the side of the open channel, at a specific height from the bed, to distribute the flow discharge to outside the channel in the aim of irrigation and drainage systems and wastewater treatment plants[13]. The importance of measuring the lateral flow through the orifices in hydraulic systems has led to considerable research efforts devoted to laboratory investigations on various shapes of side orifice in open channels[14]. Ramamurthy et al.[15], as a pioneer researcher, studied the flow mechanism through a rectangular lateral orifice in a rectangular open channel. Besides, they theoretically and experimentally analyzed the characteristics of flow through the lateral weir-orifice unit[16]. Gill[17] and Ojha and Subbaiah[18] presented the theoretical analysis of flow characteristics through various side slots in free-surface and pressure flow conditions.

Moreover, Swamee[19] studied discharge coefficient estimation for flow through weir-orifice units under free-flow conditions. Recently, Hussain et al.[20] experimentally and theoretically accomplished the sharp-crested circular side orifices in open rectangular channels under free-flow conditions. Besides, they extended their investigation on the lateral circular orifice under both free and submerged flow conditions. They derived various

[1]Water Engineering Department, Faculty of Agriculture, University of Zanjan, Zanjan, Iran. [2]Faculty of Engineering, Shohadaye Hoveizeh Campus of Technology, Shahid Chamran University of Ahvaz, Dasht-e Azadegan, Susangerd, Iran. [3]Department of Civil Engineering, Behbahan Khatam Alanbia University of Technology, Behbahan, Iran. [4]Institute of Research and Development, Duy Tan University, Da Nang 550000, Vietnam. [5]Faculty of Natural Sciences, Duy Tan University, Da Nang 550000, Vietnam. ✉email: aminasadi64@gmail.com

discharge coefficient relationships with acceptable accuracy, which usually were depended on upstream hydraulic characteristics[21]. Also, Hussain et al.[22] reported an analytical and laboratory study on the hydraulic characteristics of flow-through side square orifices in rectangular open channels. They provided a new discharge coefficient relationship based on the approach flow Froude number and ratio of orifice and channel width. Hussain et al.[23] conducted extensive laboratory and theoretical research on the performance of sharp-crested rectangular side orifice under the free-flow condition compared to square and circular side orifice. They found that the circular orifice is more efficient to divert flow than square side orifice by the same opening area. Besides, they developed their research in the aim of modifying the concept used by Ramamurthy[15] in the derivation of discharge coefficient relationship for flow through lateral side rectangular orifice.

More recently, Vatankhah and Mirnia[13] conducted an experimental and analytical study to predict the discharge coefficient of sharp-crested side triangular orifices based on 750 laboratory experiments under free-flow conditions. They proposed different discharge coefficient relationships for two scenarios: the approach Froude number and the absence of one. Furthermore, Vatankhah and Rafeifar[3] assessed the operation of the elliptical sharp-crested side orifices for diverting flow from the horizontal open channel. This research examined both small and large elliptical side orifices for deriving the regression-based discharge coefficient relationships in two scenarios depending on Froude's approach and without it. In the current research, it should be mentioned that 588 laboratory experiment tests for data-driven based assessment of elliptical side orifice discharge coefficient.

Artificial intelligence and machine learning models in different engineering problems such as hydraulic[24–26], geotechnical[27–31], and mechanical[32] engineerings have become very popular. In two recent decades, laboratory equipment and human errors, on the one hand, the complexity and nonlinear behavior of spatially varied flow through these facilities and the insufficient accuracy of classical regression-based methods, on the other hand, has caused several researchers to turn their attention to the data-driven and machine learning techniques[33–36]. Numerous attempts focus on the application of conventional artificial intelligence (AI), such as an artificial neural network (ANN) and adaptive neuro-fuzzy inference system (ANFIS), to measure the coefficient of discharge through the side lateral weirs and side orifices. For instance, the mentioned approaches implemented to predict the discharge coefficient for rectangular sharp-crested weirs[37–40]; for measuring the discharge capacity of rectangular sharp-crested side weirs in sewer systems[41]; to assess the discharge coefficient of triangular and trapezoidal labyrinth side weirs; for estimating the discharge coefficient for a semi-elliptical labyrinth side weirs[42]; to accurate determination of the discharge coefficient for a triangular side weir under subcritical flow conduction[26,43], and predict the discharge of rectangular and circular side orifices in a rectangular channel[44]. The bedside, Gene expression programming (GEP) paradigm has been employed to determine the discharge coefficient of rectangular side weirs in various flow regimes along the rectangular and trapezoidal channels[45,46]. In other cases, a support vector machine (SVM) model was employed to model the discharge coefficient of a side weir in a rectangular[47] and trapezoidal channel[48]; multivariate adaptive regression splines (MARS) and the group method of data handling (GMDH) have been utilized to predict the discharge coefficient of Weir-Gat[49].

Furthermore, More recently, the self-adaptive extreme learning machine (SAELM) as a novel ML approach has been employed to model the side weirs discharge on converging channels[50] and circular and rectangular side orifices along the open channel[51–53]. Jamei et al. applied three data-driven approaches, multiple linear regression with interaction (MLRI), locally weighted learning regression (LWLR), and multiple linear regression (MLR), to estimate the discharge coefficient of a triangular side orifice. Their results demonstrated the high capability of LWLR and MLRI models to estimate discharge coefficient[54].

Literature review inferred that different ML-based methods had been used to model the discharge coefficient of hydraulic diversion devices. However, some structures have the complexity of hydraulic characteristics due to their particular shape. The accurate estimation of the discharge coefficient in them requires applying robust and efficient data-driven approaches.

In this research, the estimation of discharge coefficient in elliptical side orifice under free flow conduction, for the first time, has been assessed using two kernel-based data-intelligence paradigms, namely, Gaussian process regression (GPR) and kernel extreme learning method (KELM). Here, the response surface methodology (RSM), generalized regression neural network (GRNN), and empirical methods were adopted to validate the provided schemes. To the best of our knowledge, the proposed data-driven approaches have not yet been used for discharge measuring the hydraulics structures. Furthermore, the applicability domains of the provided paradigms were examined using the leverage approach, and a sensitivity analysis was performed to identify the most influential variables. Models were evaluated based on statistical indices, and the results were presented as tables and figures. This paper describes the experimental data and machine learning models in the second part (materials and methods). Results and discussion are presented in the third part, and at last, the conclusions are offered.

Need for research.

Determining the lateral flow in the side orifices is essential for water management, water pricing, and environmental engineering objectives. In the present study, to increase the accuracy of flow determination, using machine learning methods, GPR, KELM, GRNN, and RSM, models and relationships were developed to determine the flow coefficient of the elliptical side orifice.

Material and methods

Dimensional analysis and data preparation.

Factors affecting the elliptical side orifices are (1) Dimension of the elliptic orifice (a is horizontal semi-axis and b is vertical semi-axis) (2) orifice crest height (the distance between the channel bed and the orifice) (W) (3) velocity in the main channel ($V1$) which is calculated as $V1 = Q/A_1$ (4) orifice upstream ($y1$) and downstream ($y2$) depths (5) Main channel width (B) (6) Gravity acceleration (g) (7) Water surface tension (σ) (8) Water dynamic viscosity (μ) and (9) Water density (ρ)[3]. Fig-

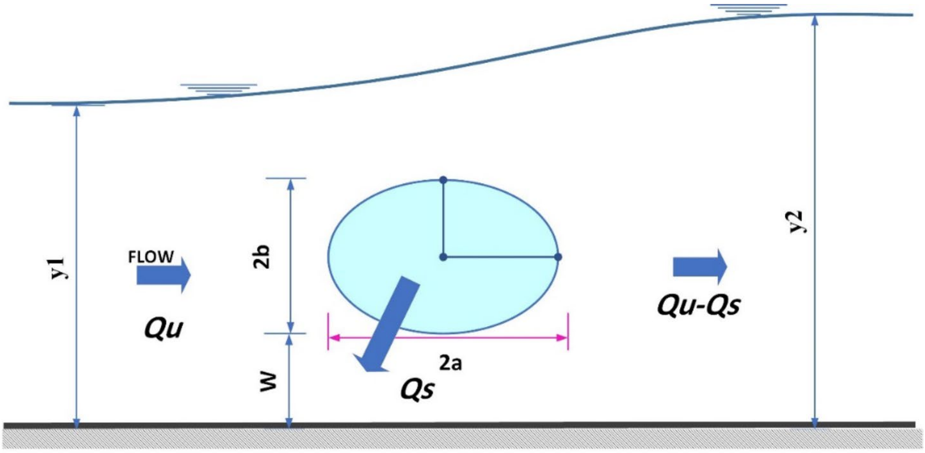

Figure 1. Schematic view of elliptic side orifice and its geometrical parameters.

Data	Statistic	W/b	B/a	B/b	$b/y1$	$Fr1$	C_d
Train data	Mean	2.665	1.463	9.047	0.172	0.484	0.518
	Std	1.212	0.208	2.611	0.039	0.121	0.026
	Min	1.25	1.25	6.25	0.093	0.219	0.405
	Max	5	1.667	12.5	0.252	0.777	0.569
Test data	Mean	2.837	1.446	8.971	0.168	0.472	0.517
	Std	1.246	0.209	2.576	0.039	0.108	0.027
	Min	1.25	1.25	6.25	0.093	0.273	0.426
	Max	5	1.667	12.5	0.256	0.731	0.562

Table 1. Statistical specifications of train and test datasets.

ure 1 shows an elliptical side orifice and the geometric parameters used in it. According to the variables affecting the discharge coefficient of elliptical side orifice, a relation can be written as follows:

$$Cd = f1\left(a, b, W, B, y2, V1, g, \rho, \sigma, \mu\right). \tag{1}$$

Using Buckingham's π theory, effective dimensionless parameters can be obtained as follows:

$$Cd = f2\left(\frac{W}{b}, \frac{B}{a}, \frac{B}{b}, \frac{y1}{b}, Fr1 = \frac{V1}{\sqrt{g \cdot y1}}, Re = \frac{\rho \cdot V1 \cdot a}{\mu}, We = \frac{\sigma}{\rho, g \cdot a^2}\right). \tag{2}$$

Given that in open channels, the most critical effective force is gravity, the effect of viscosity and surface tension can be ignored[5,55], so Reynolds and Weber numbers can be removed from the above equation.

$$Cd = f3\left(\frac{W}{b}, \frac{B}{a}, \frac{B}{b}, \frac{y1}{b}, Fr1 = \frac{V1}{\sqrt{g \cdot y1}}\right) \tag{3}$$

In the present study, Vatankhah and Rafeifar[3] laboratory data, which includes 588 series of data, were used. They studied the effect of different geometrical and hydraulic parameters on the elliptical side orifice discharge coefficient. A horizontal rectangular channel (12 m length, 0.25 m width, and 0.5 m height) was used to perform experiments. Two types of rectangular and triangular weirs were used to measure the flow through the orifice (Qs) orifice and the upstream flow of the orifice (Qu). Two different lengths of orifice (a = 15, 20 cm), three heights (b = 2,3,4 cm) and 2 crest heights (w = 5, 10 cm) were used. In total, 12 different geometric shapes were created. (Qu) values ranged from 13.8 to 39.6 l/s, Qs ranged from 3.66 to 21.41 l/s, and the Froude number in the main channel ranged from 0.22 to 0.77. Finally, the discharge coefficient can be calculated as $C_d = Q/\pi ab\sqrt{2gh_c}$ where $h_c = y_1 - W - b$. 588 laboratory data were randomly divided into two parts: training (75%) and test (25%). Table 1 shows the statistical specifications of training and test datasets.

Figure 2 shows the relationship between the output variable (Cd) and independent variables for the dataset used in this study. The numbers in the figure represent the linear relationship between variables using the Pearson correlation coefficient. The value of this coefficient varies from −1 to 1. Positive values indicate a direct connection between variables, and negative values indicate an inverse relationship between variables. According to Fig. 2,

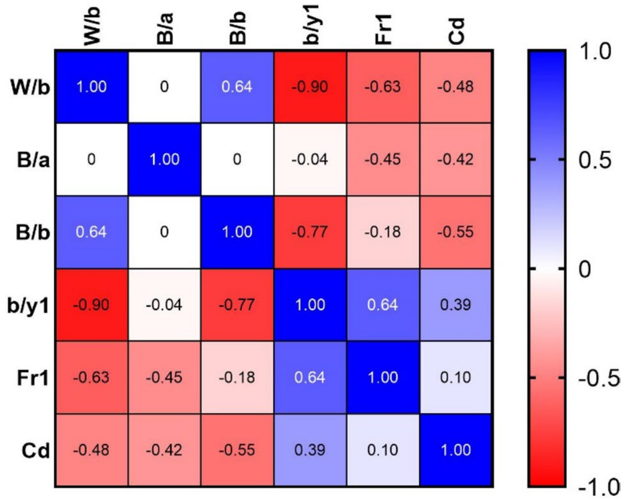

Figure 2. Correlation matrix of input(independent) variables and output (dependent) variable.

the two variables $b/y1$ ($r_p = 0.39$) and $Fr1$ ($r_p = 0.10$) directly affects the discharge coefficient, which means that as they increase, the discharge coefficient increases. The three variables W/b ($r_p = -0.48$), B/a ($r_p = -0.42$) and B/b ($r_p = -0.55$) have the inverse effect on the discharge coefficient, and as they increase, the discharge coefficient decreases. According to Fig. 2, the variables B/b and W/b have the highest absolute correlation with the elliptical side orifice discharge coefficient.

Machine learning technics.

Gaussian process regression (GPR). The Gaussian process regression (GPR) model falls into the category of supervised machine learning methods[56]. GPR is a kernel-based non-parametric based on Bayes, with high computational efficiency and accuracy which its operation is easy for users[57]. This approach can solve classification and regression problems. This method has a high capability in modeling complex nonlinear issues[58]. A Gaussian process is expressed by the mean function $m(x)$ and the covariance function $k(x_i, x_j)$ as follows[59,60]:

$$f(x) \sim GP\left(m(x), k\left(x_i, x_j\right)\right). \tag{4}$$

In the regression problem, y is defined as observations and ε as noise. This noise has an average of zero and σ_n^2 variance. As a result, the Gaussian process regression model can be expressed as follows:

$$y = f(x) + \varepsilon, \varepsilon \sim N\left(0, \sigma_n^2\right). \tag{5}$$

In the above equation, x is the input data matrix, y is the output data vector, and f is the values of the GPR function. The joint distribution is defined by the kernel function as follows:

$$\begin{bmatrix} y \\ f_* \end{bmatrix} \sim N\left(0, \begin{bmatrix} K(X,X) + \sigma_n^2 I_n & K(x_*, X)^T \\ K(x_*, X) & K(x_*, x_*) \end{bmatrix}\right) \tag{6}$$

where $K(x_*, X)$ is equal to

$$K(x_*, X) = [k(x_*, x_1), k(x_*, x_2), \cdots, k(x_*, x_n)] \tag{7}$$

X is the training input matrix $X = [x_1, x_2, \ldots, x_n]^T$, y is the training output vector $y = \left[y_1, y_2, \ldots, y_n\right]^T$, x_* is the test input vector and f_* is the output for the test input data vector.

Finally, the predictor distribution is expressed by the following equation:

$$f_*|X, y, x_* \sim N\left(\left(\overline{f_*}\right), cov\left(f_*\right)\right) \tag{8}$$

where $\overline{f_*}$ and $cov\left(f_*\right)$ are defined as follows:

$$\overline{f_*} = K(x_*, X)\left[K(X,X) + \sigma_n^2 I_n\right]^{-1} y \tag{9}$$

$$cov\left(f_*\right) = K(x_*, x_*) - K(x_*, X)\left[K(X,X) + \sigma_n^2 I_n\right]^{-1} K(x_*, X)^T. \tag{10}$$

The covariance function is used to measure the effect of data points on each other[57]. This function shows the number of coordinated changes between the two variables. The proper selection of kernel function (covariance) is one of the essential factors affecting the performance of the GPR model. Numerous kernel functions

Kernel function	Kernel equation
Rational quadratic	$k\left(x_i, x_j\middle|\theta\right) = \sigma_f^2\left(1 + \frac{r^2}{2\alpha\sigma_l^2}\right)^{-\alpha}$
Exponential Kernel	$k\left(x_i, x_j\middle|\theta\right) = \sigma_f^2 exp\left[-\frac{r}{\sigma_l}\right]$
Matern 3/2	$k\left(x_i, x_j\middle|\theta\right) = \sigma_f^2\left(1 + \frac{\sqrt{3}r}{\sigma_l}\right)exp\left[-\frac{\sqrt{3}r}{\sigma_l}\right]$
Squared exponential	$k\left(x_i, x_j\middle|\theta\right) = \sigma_f^2 exp\left[-\frac{1}{2}\frac{\left(x_i - x_j\right)^T\left(x_i - x_j\right)}{\sigma_l^2}\right]$
Matern 5/2	$k\left(x_i, x_j\middle|\theta\right) = \sigma_f^2\left(1 + \frac{\sqrt{5}r}{\sigma_l} + \frac{5r^2}{3\sigma_l^2}\right)exp\left[-\frac{\sqrt{5}r}{\sigma_l}\right]$
ARD squared exponential	$k\left(x_i, x_j\middle|\theta\right) = \sigma_f^2 exp\left[-\frac{1}{2}\sum_{m=1}^{d}\frac{\left(x_{im} - x_{jm}\right)^2}{\sigma_m^2}\right]$
ARD exponential	$k\left(x_i, x_j\middle|\theta\right) = \sigma_f^2 exp[-\varphi]$
ARD matern 3/2	$k\left(x_i, x_j\middle|\theta\right) = \sigma_f^2\left(1 + \sqrt{3}\varphi\right)exp\left[-\sqrt{3}\varphi\right]$
ARD matern 5/2	$k\left(x_i, x_j\middle|\theta\right) = \sigma_f^2(1 + \sqrt{5}\varphi + 5\varphi^2)exp[-\sqrt{5}\varphi]$
ARD rational quadratic	$k\left(x_i, x_j\middle|\theta\right) = \sigma_f^2\left(1 + \frac{1}{2\alpha}\sum_{m=1}^{d}\frac{\left(x_{im} - x_{jm}\right)^2}{\sigma_m^2}\right)^{-\alpha}$

Table 2. List of Kernel functions used for GPR model. Where σ_f is the signal standard deviation (Std), σ_l is the characteristic length scale, r is the Euclidean distance between x_i and x_j which is defined by $r = \sqrt{\left(x_i - x_j\right)^T\left(x_i - x_j\right)}$, α is a positive-valued scale-mixture parameter, $\varphi = \sqrt{\sum_{m=1}^{d}\frac{\left(x_{im} - x_{jm}\right)^2}{\sigma_m^2}}$ and σ_m is a separate length scale for each predictor m, $m = 1, 2, \ldots, d$. The values of $\theta = \left\{l, \sigma_f^2, \sigma_l^2\right\}$ (hyper-parameters) is calculated by maximizing the marginal likelihood[61,62].

are defined for use in the GPR model[59]. In the present study, ten types of kernels were examined and evaluated. Table 2 shows the kernel equations used in the present study.

Kernel extreme learning machine (KELM). ELM is a developed version of the single-layer feed-forward network (SLFN) with a random nature presented by Huang et al.[63]. The ELM consists of three layers: input, hidden, and output layers. The main advantages of this method are (1) easy implementation, (2) fast training speed, and (3) powerful generalization capability (Huang et al. 2012).

The mathematical formulation of the ELM for a dataset including M samples, by considering the number of hidden layer nodes equal to H, can be expressed as,

$$y_l = \sum_{j=1}^{H}\rho_j g(b_j x_l + c_j), \quad l = 1, 2, \ldots, M \tag{11}$$

where ρ_j denotes the output weight vector, which connects the jth hidden layer node and output layer node. $g(x)$ represents the ELM activation function (AF), b_j is the weight of the input dataset, and c_j is the bias value for the jth hidden layer node. Equation (8) can be defined as,

$$Y = G\rho \tag{12}$$

where Y denotes the model output, G is the matrix of hidden layer output, which is expressed as,

$$G = \begin{bmatrix} g(b_1x_1 + c_1) & \ldots & g(b_Hx_1 + c_H) \\ \vdots & \ldots & \vdots \\ g(b_1x_M + c_1) & \ldots & g(b_Hx_M + c_H) \end{bmatrix}. \tag{13}$$

The ELM uses a fitness function to determine the optimum value for the ρ, which is given as,

$$F = \sum_{l=1}^{M}\left(\sum_{j=1}^{H}\rho_j g(b_j x_l + c_j\right) - T_l)^2 \tag{14}$$

where T is the target vector.

Based on the generalized inverse theory, the solution of Eq. (12) is defined as,

$$\rho = G^{\dagger}Y \tag{15}$$

where $G^{\dagger}$ refers to the Moore–Penrose inverse matrix (MPIM) of G. Regarding the orthogonal projection technique and theory of ridge regression[64], the regularization factor (RF) was applied in the process of optimization so that the ρ can be achieved as,

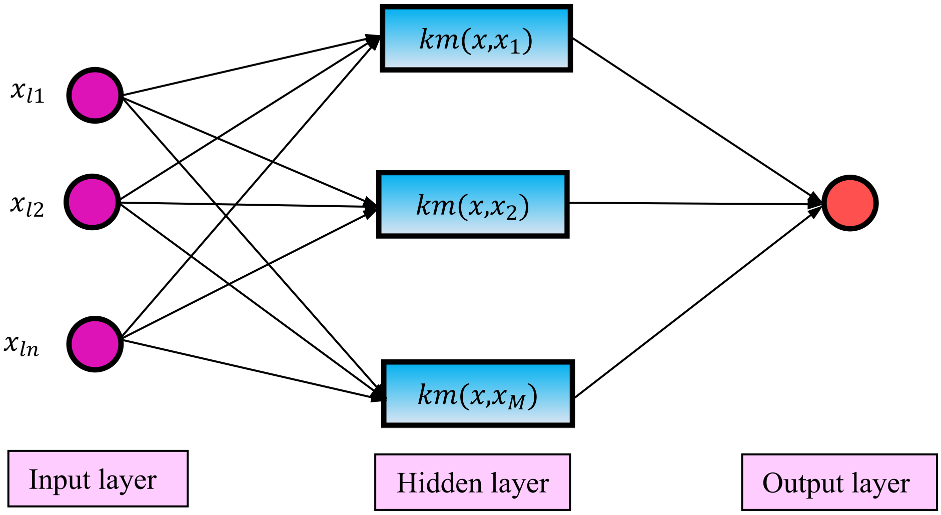

Figure. 3. KELM structure.

$$\rho = \left(G^T G + \frac{I}{RF}\right)^{-1} G^T Y \tag{16}$$

where I denotes the identity matrix. Accordingly, the ELM output function is defined as,

$$z(x) = g(x)\rho = g(x)\left(G^T G + \frac{I}{RF}\right)^{-1} G^T Y. \tag{17}$$

Notwithstanding the suitable efficiency of the ELM, but because this method is random, it may be trapped in the local optima. Therefore, the kernel extreme learning machine (KELM) was presented by Huang et al.[65]. The main structure of the KELM is displayed in Fig. 3. In this method, a kernel matrix (KM) ($KM(x, x_j)$) is employed instead of the AF ($g(x)$). The KM can be formulated based on Eq. (18).

$$\phi = G^T G : \phi_{j,l} = g(x_j) g(x_l) = KM(x_j, x_l) \tag{18}$$

The output function of the proposed KELM is expressed as,

$$f(x) = \begin{bmatrix} KM(x, x_1) \\ \vdots \\ KM(x, x_H) \end{bmatrix} \left(\phi + \frac{I}{RF}\right)^{-1} Y. \tag{19}$$

In this work, the radial basis function (RBF) is utilized as a KM, which can be obtained as,

$$KM(x_j, x_l) = \exp\left(-\frac{||x_j - x_l||^2}{\mu}\right) \tag{20}$$

where μ is a constant number.

Generalization regression neural network (GRNN). Generalized regression neural network (GRNN) is a kind of radial basis function network (RBFN) that is based on kernel regression[66]. Unlike the conventional neural networks (CNN), the GRNN does not need a repetitive training process like the back-propagation technique. The GRNN does not stick to local solutions[67–69]. This method comprises four layers: input, pattern, summation, and output layers.

The input layer receives the input dataset (x). In this layer, the number of neuros is equal to the dimension of the input dataset. In the pattern layer, neurons using a nonlinear function transform the input dataset (x) to p_k (i.e., the output of the pattern layer) based on the following equation:

$$p_k = \exp\left[\frac{(x - x_k)^T (x - x_k)}{2\rho^2}\right] \tag{21}$$

where x_k denotes the training sample of the kth neuron in the input layer. ρ is the spread factor.

The third layer (i.e., summation layer) consists of two types of neurons: (1) one simple neuron and (2) m weighted neurons, which are specified by S_o and S_t. These kinds of neurons are defined as,

$$S_o = \sum_{k=1}^{M} p_k, \quad S_t = \sum_{k=1}^{M} y_k p_k \tag{22}$$

where y_k is the target dataset.

The output layer (i.e., output layer) divides the summation layer results to achieve the output predicted result, which is expressed as,

$$Y = \frac{S_t}{S_o}. \tag{23}$$

Surface response methodology (RSM). In the present study, RSM was used to investigate the effect of independent variables (geometric and hydraulic conditions) on the output (response) variable (side orifice discharge coefficient) and also to provide an optimal regression relationship for the elliptical side orifice discharge coefficient prediction. The RSM method is a statistical tool for modeling and analyzing the behavior of the process (input) variables on the response (output) variable[70]. Using RSM, most information can be obtained with a minimum of experimental data. The 2nd order RSM model includes linear, quadratic, and the interaction of input variables sentences. The RSM model for the above case can be expressed as follows[71,72]:

$$y = \sum_{i=1}^{k} \alpha_i X_i + \sum_{i=1}^{k} \alpha_{ii} X_i^2 + \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \alpha_{ij} X_i X_j + \varepsilon \tag{24}$$

where X is the input data matrix, y is the output data estimation vector, ε is a random error vector and α_i, α_{ii} *and* α_{ij} are regression coefficients which the following equation can calculate:

$$\alpha = \left(X^T X\right)^{-1} X^T Y \tag{25}$$

A flowchart of the machine learning models for the discharge coefficients of the elliptical side weir can be depicted in Fig. 4. In all models, the input is normalized using the following formula:

$$x_{nor} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{26}$$

where x is the value of variable and x_{min} and x_{max} are the minimum and maximum value of the variable, respectively.

Accuracy criteria of approaches.

Five statistical indices evaluated the models: root mean square error (RMSE), mean average percentage error (MAPE), correlation coefficient (R), normalized root mean square error (NRMSE), and mean bias error (MBE). The relationships of each of the mentioned parameters are presented below:

$$R = \frac{\sum_{i=1}^{N}(Cdo_i - \overline{Cdo})\left(Cdp_i - \overline{Cdp}\right)}{\sqrt{\sum_{i=1}^{N}(Cdo_i - \overline{Cdo})^2 \sum_{i=1}^{N}(Cdp_i - \overline{Cdp})^2}} \tag{27}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(Cdo_i - Cdp_i)^2}{N}} \tag{28}$$

$$MBE = \frac{1}{N}\sum_{i=1}^{N}\left(Cdo_i - Cdp_i\right) \tag{29}$$

$$NRMSE = \frac{RMSE}{\overline{Cdo}} \tag{30}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{Cdo_i - Cdp_i}{Cdo_i}\right| \times 100. \tag{31}$$

At the above equations, Cdo_i and Cdp_i respectively are observed and predicted values of discharge coefficient of elliptical side orifice, $\overline{Cdo}$ is the mean value of observations, $\overline{Cdp}$ is the mean value of predictions, and N is the number of data.

Outlier detection with leverage approach.

Through developing a mathematical model, it is necessary to detect outlier data obtained from the model. Several methods have been proposed to identify and detect outlier data. Among these, the leverage approach is one of the most well-established and widely used approaches.

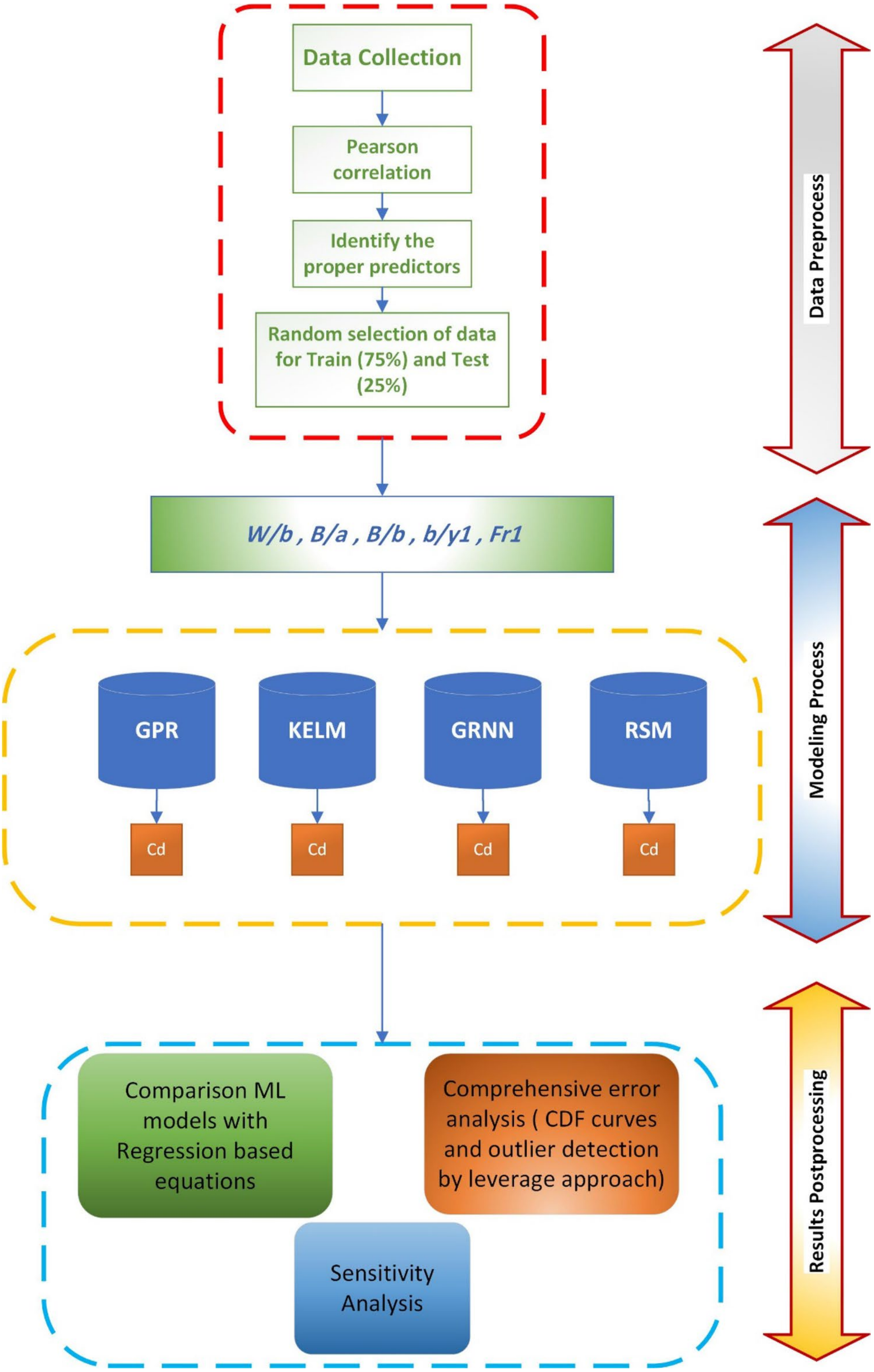
Data Collection
Pearson correlation
Identify the proper predictors
Random selection of data for Train (75%) and Test (25%)
W/b , B/a , B/b , b/y1 , Fr1
GPR
KELM
GRNN
RSM
Cd
Cd
Cd
Cd
Comparison ML models with Regression based equations
Comprehensive error analysis (CDF curves and outlier detection by leverage approach)
Sensitivity Analysis
Data Preprocess
Modeling Process
Results Postprocessing

Figure 4. The flowchart of predicting discharge coefficient of elliptical side orifice by different machine learning (ML) models.

In this method, the difference between the actual data and the data obtained from the model is defined as the residual. To calculate the leverage index (*hat*) the following matrix must be calculated:

$$H = X\left(X^{T}X\right)X^{-1}. \tag{32}$$

Kernel	RMSE	R	MAPE	MBE	NRMSE
Exponential	0.00875	0.95098	1.40634	−0.00062	1.69007
Squaredexponential	0.00845	0.95423	1.36432	−0.00054	1.63369
Matern3/2	0.00857	0.95299	1.38097	−0.00061	1.65591
Matern5/2	0.00853	0.95346	1.37472	−0.00059	1.64759
Rational quadratic	0.00845	0.95423	1.36432	−0.00054	1.63370
ARDexponential	0.00840	0.95474	1.38167	−0.00031	1.62279
ARDsquaredexponential	0.00810	0.95797	1.32428	−0.00029	1.56469
ARDmatern3/2	0.00821	0.95674	1.35327	−0.00028	1.58723
ARDmatern5/2	0.00817	0.95723	1.34383	−0.00028	1.57823
ARDrationalquadratic	0.00818	0.95715	1.34447	−0.00028	1.57975

Table 3. Effect of GPR kernel type on model accuracy.

In the above equation, X is an $n \times k$ matrix, where n is the number of samples and k is the number of model variables. The diagonal elements of the H matrix are known as the hat or leverage index. The standard residual percent (R) is plotted against the *hat* to indicate the domain of application of the model and the outlier data. The warning value of the leverage H_* is calculated from the following equation:

$$H_* = 3(k+1)/n. \tag{33}$$

The plot of R versus *hat* is known as the Williams diagram. If most of the data are in the range of $-3 < R < 3$ and $0 < H < H_*$ It indicates the appropriate application of the model in the mentioned range and, therefore, shows the developed model's statistical validity[73,74].

Results and discussion

This section discusses and evaluates the results obtained from GPR, KELM, GRNN, and RSM models and regression-based models. There will also be a comprehensive comparison between the mentioned AI models and regression-based models. Error analysis was performed using CDF curves, relative error, and leverage approach. Finally, sensitivity analysis will be performed to determine the parameters affecting the elliptical side orifice discharge coefficient. All models are performed in the MATLAB 2020a software on a personal computer (Intel Core i7 2.6 GHz processor and 16 GB RAM).

Gaussian process regression (GPR) model.

The GPR model was created using the dimensionless variables mentioned in the previous section as input and the discharge coefficient (Cd) parameter as output. The most important factor in the performance of the GPR model is the type of kernel and its parameters. In the present study, ten kernels and LBFGS-based quasi-Newton methods were used to optimize kernel parameters. Table 3 shows the results obtained from different kernels with their optimal parameter. The results obtained by different kernels were compared using R and RMSE statistical parameters for the test data series. According to Table 3, the *ARDsquaredexponential* kernel with R = 0.9579, RMSE = 0.0081 and MAPE = 1.3243% had the best performance in estimating the orifice discharge coefficient. The *ARDMatern* 5/2 kernel with R = 0.9571, RMSE = 0.0087 and MAPE = 1.5782% was the second model with high accuracy. The weakest performance was provided by *exponential* kernel with R = 0.9509, RMSE = 0.0087 and MAPE error percentage = 1.4063%. The results of the optimal GPR model for the training and test data series are presented in Fig. 5.

Kernel extreme learning machine (KELM).

In the KELM model, the RBF kernel was considered as the model kernel[75,76]. The RBF kernel has one parameter as σ, and the KELM model has one parameter as an adjustment parameter (C). The grid serach method was used to obtain σ and C. The values of σ and C were changed from 0.01 to 3 and 1 to 1000, respectively. Finally, the optimal values of $\sigma = 0.1$ and C = 600 were obtained for the test data series. According to the optimal parameters of the kernel and KELM model, the best model was obtained with R = 0.9564, RMSE = 0.0082, and MAPE = 1.3499%. The results of the optimal KELM model are presented in Fig. 6 for the test and training datasets.

Generalized regression neural network (GRNN).

The only parameter in the GRNN model is the Spread parameter[77]. To obtain the optimal spread value, its values were changed between 0.01 and 10 with 0.01 intervals. The results showed that the optimal value of this parameter is 0.05. R = 0.929, RMSE = 0.0106 and MAPE = 1.6971% were obtained for the optimal GRNN model. The results of the optimal GRNN model are presented in Fig. 7 for the test and training datasets.

Response surface methodology (RSM).

The effect of independent variables on the side elliptical orifice discharge coefficient was evaluated using the RSM model. One of the advantages of the RSM is presenting a regression relationship between input and output variables. The RSM model is based on the number of inde-

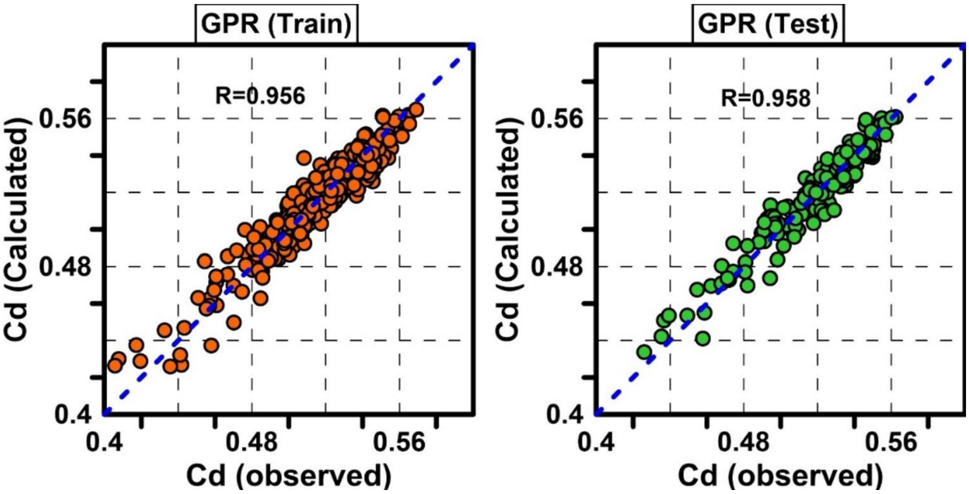

Figure 5. Scatter plots of observed Cd against predicted Cd by GPR model for train and test data.

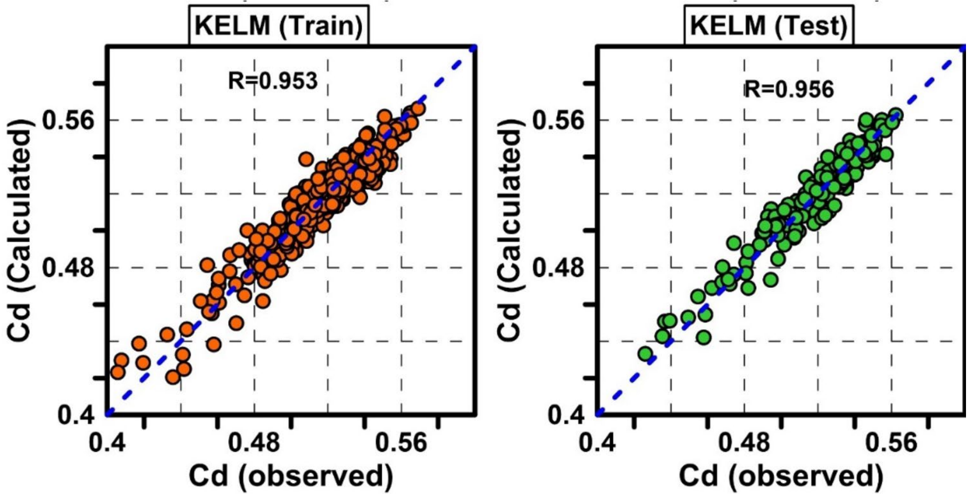

Figure 6. Scatter plots of observed Cd against predicted Cd by KELM model for train and test data.

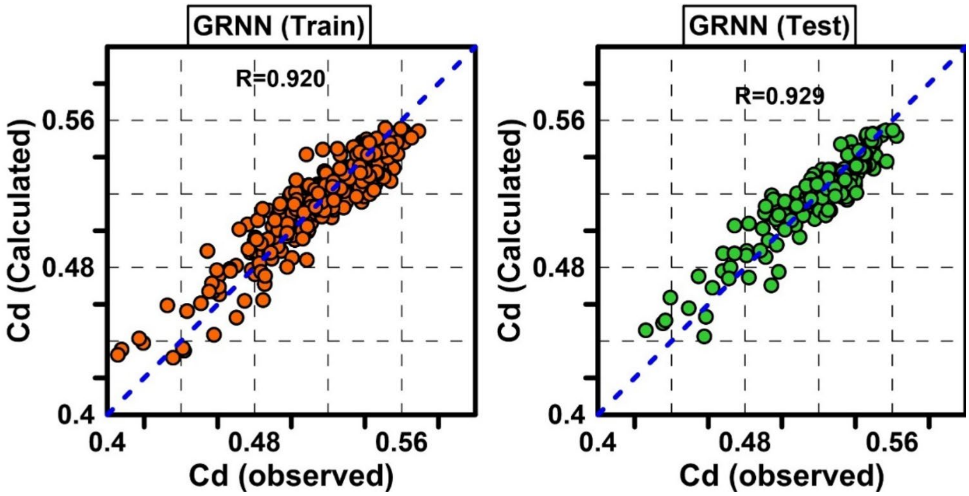

Figure 7. Scatter plots of observed Cd against predicted Cd by GRNN model for train and test data.

Variables	Coefficient	Sum of squares	F-value	*p* value
W/b	0.18167	0.06822	695.09123	0.00000
B/a	0.13982	0.09677	986.01148	0.00000
B/b	0.00772	0.00170	17.35857	0.00004
$b/y1$	3.93829	0.00163	16.65797	0.00005
$Fr1$	−0.26858	0.03962	403.65941	0.00000
$W/b \times B/a$	−0.02340	0.00205	20.87685	0.00001
$W/b \times B/b$	−0.00743	0.00371	37.77965	0.00000
$W/b \times b/y1$	−0.65597	0.00503	51.25667	0.00000
$B/a \times B/b$	−0.00870	0.00078	7.95540	0.00502
$B/a \times b/y1$	−1.35512	0.00152	15.45248	0.00010
$B/a \times Fr1$	0.26254	0.00127	12.95195	0.00036
$B/b \times b/y1$	0.11749	0.00199	20.24286	0.00001
$B/b \times Fr1$	−0.04114	0.00197	20.08285	0.00001
$b/y1 \times Fr1$	−3.38425	0.00159	16.21657	0.00007
$(B/b)^2$	0.00118	0.00128	13.02753	0.00034
$(Fr1)^2$	0.65565	0.00319	32.46088	0.00000

Table 4. ANOVA results for determining the effective variable interactions in the RSM model.

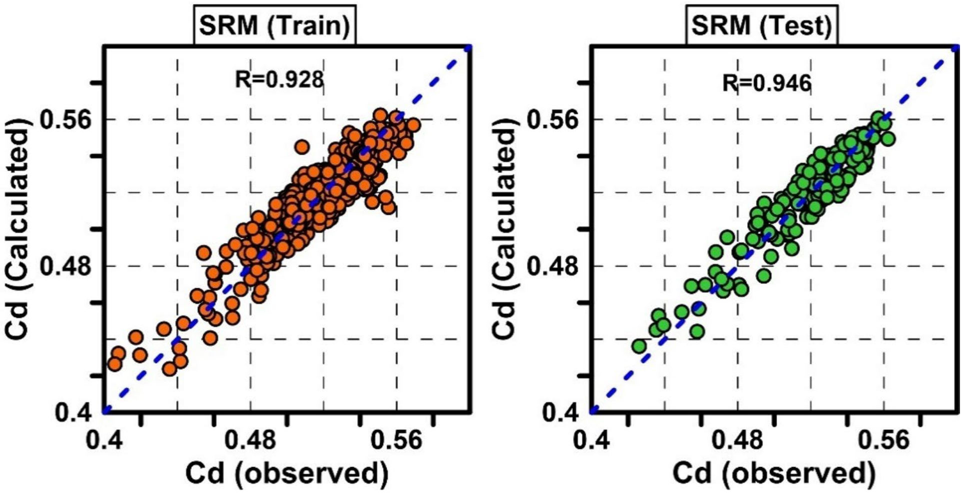

Figure 8. Scatter plots of observed Cd against predicted Cd by MLRI model for train and test data.

pendent variables, their squares, and their relationship (in pairs). The equation obtained from this model is as follows:

$$\begin{aligned} Cd = {} & 0.0717 + 0.1817W/b + 0.1398B/a - 0.0077B/b + 3.9383b/y1 \\ & - 0.2686Fr1 - 0.0234W/b \cdot B/a - 0.0074W/b \cdot B/b - 0.6560W/b \cdot b/y1 \\ & - 0.0087B/a \cdot B/b - 1.3551B/a \cdot b/y1 - 0.0411B/b \cdot Fr1 \\ & - 3.3842b/y1 \cdot Fr1 + 0.0012(B/b)^2 + 0.6557(Fr1)^2. \end{aligned} \tag{34}$$

Table 4 shows the ANOVA analysis of variance for the equation and its coefficients. According to this table, all coefficients are significant (p value < 0.05). In the RSM model, R = 0.9456, RMSE = 0.0092 and MAPE = 1.4921% were obtained for the test dataset. Figure 8 shows the performance of the RSM model for training and test data.

Regression-based equations. Vatankhah and Rafeifar[3] presented five regression-based models to calculate the elliptical side orifice discharge coefficient.

$$Equation\ 1: C_d = 0.64 - 0.1\left(\frac{W}{b}\right)^{0.572}\left(\frac{B}{a}\right)^{1.27}\left(\frac{B}{b}\right)^{0.04}\left(\frac{b}{y1}\right)^{0.13}Fr1^{0.85} \tag{35}$$

Equations	RMSE	R	MAPE	MBE	NRMSE
Equation 1	0.0106	0.9277	1.6846	0.0003	2.0464
Equation 2	0.0107	0.9254	1.6993	0.0007	2.0731
Equation 3	0.0234	0.5713	3.6714	−0.0014	4.5212
Equation 4	0.0273	0.5149	4.3799	0.0113	5.2789
Equation 5	0.0202	0.7008	3.1016	−0.0011	3.9039

Table 5. Results of regression-based equations.

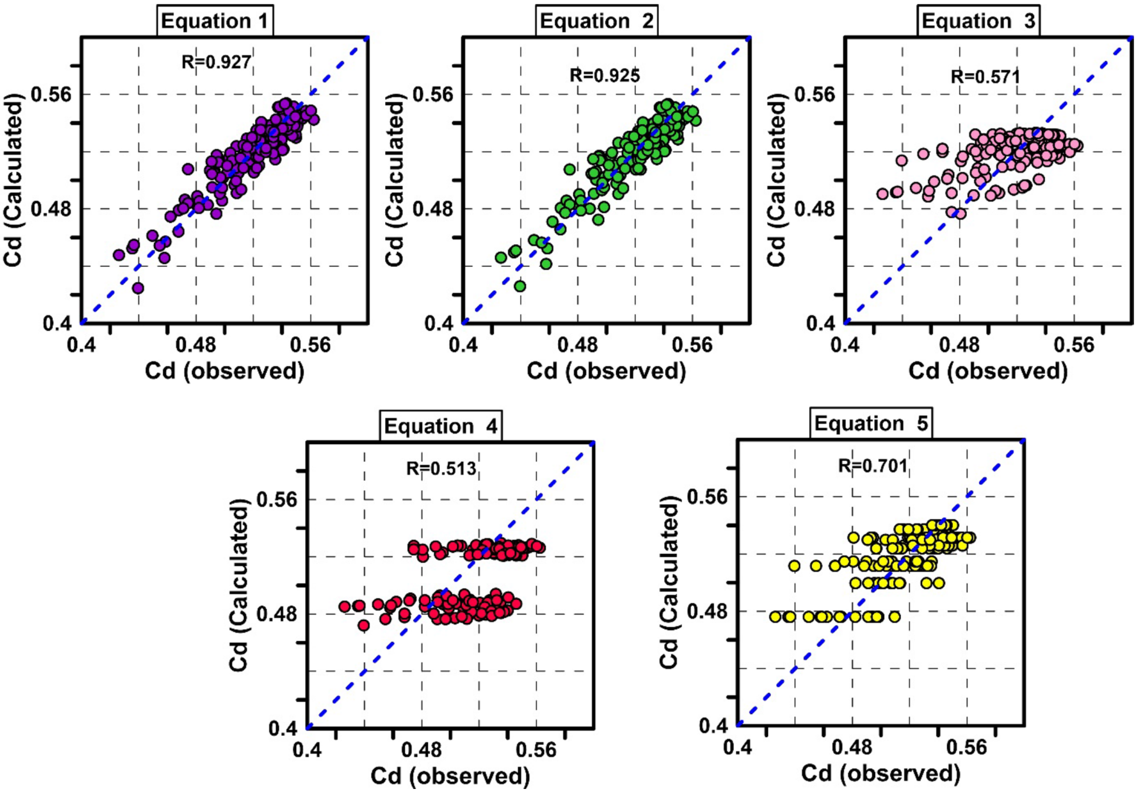

Figure 9. Scatter plots of regression-based equations.

$$Equation\ 2: C_d = 0.635 - 0.085\left(\frac{W}{b}\right)^{0.57}\left(\frac{B}{a}\right)^{1.33} Fr1^{0.91} \quad (36)$$

$$Equation\ 3: C_d = 0.536 - 0.005\left(\frac{W}{b}\right)^{2.11} Fr1^{1.35} \quad (37)$$

$$Equation\ 4: C_d = 0.578 - 0.039\left(\frac{B}{a}\right)^{1.47} Fr1^{0.0.24} \quad (38)$$

$$Equation\ 5: C_d = 0.549 - 0.004\left(\frac{W}{b}\right)^{0.969}\left(\frac{B}{a}\right)^{2.631} \quad (39)$$

Table 5 shows the results obtained from these five regression-based models. According to Table 5, Eq. (1) in which all effective parameters are involved with R = 0.9277, RMSE = 0.0106 and MAPE = 1.6846% had the best performance. Equation 2 takes into account the parameters $Fr1$, B/a and w/b as input with R = 0.9254, RMSE = 0.0107 and MAPE = 1.6993% is in the second rank. In equations, 3 to 5, which consider the parameters $Fr1 - w/b, Fr1 - B/a$ and $w/b - B/a$ as input, respectively, the accuracy of the equations is not acceptable, and the value of R is $R \leq 0.7$. The MAPE error in Eqs. (3) to (5) models is more than 3%. Figure 9 shows the

Statistical criteria	Models GPR	GRNN	K-ELM	RSM	Equation 1
Training stage					
R	0.9556	0.9202	0.9530	0.9279	0.8962
RMSE	0.0077	0.0104	0.0079	0.0097	0.0116
MAPE	1.1781	1.5769	1.2128	1.4790	1.7496
MBE	0	0.0001	0	0	0.0005
NRMSE	1.4844	2.0043	1.5230	1.8742	2.2231
AVE rank	1	4	2	3	5
Testing stage					
R	0.9580	0.9291	0.9564	0.9456	0.9277
RMSE	0.0081	0.0106	0.0082	0.0092	0.0106
MAPE	1.3243	1.6971	1.3499	1.4921	1.6846
MBE	−0.0003	−0.0004	−0.0002	−0.0001	0.0003
NRMSE	1.5647	2.0532	1.5929	1.7738	2.0464
AVE rank	1	4	2	3	5

Table 6. Performance of AI models and best Vatankhah and Rafeifar[3] regression-based equation in predicting discharge coefficient.

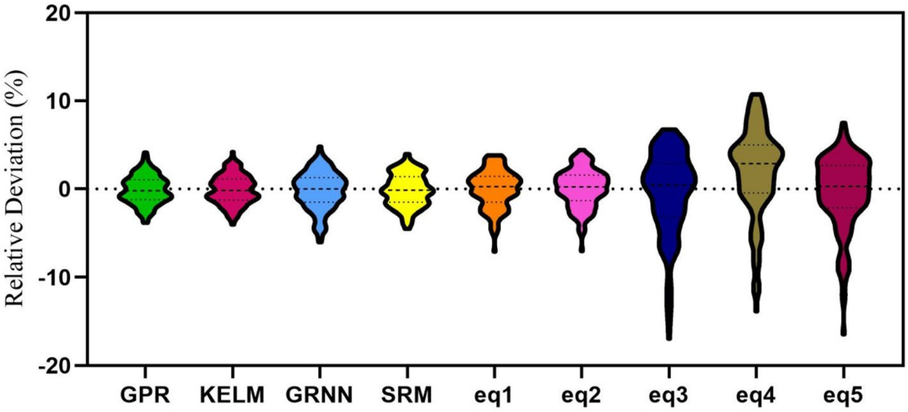

Figure 10. The error distribution of four developed AI models and regression-based equations.

scatter plots of experimental data and regression-based equation. As shown in the figure, Eqs. (3) to (5) did not provide good results.

Comparison between models.

GPR, KELM, GRNN, and RSM models were developed to predict the side elliptic orifice discharge coefficient in the previous section, and their optimal parameters were obtained. This section will compare the machine learning models developed in the previous section and the top regression model. Table 6 shows the statistical parameters of the best results obtained from the machine learning models and the best regression model for the training and test datasets. According to Table 6, all machine learning models performed better than the regression-based model. Comparison between machine learning models also shows that the GPR model, with R = 0.9556, RMSE = 0.0077 for training data, and R = 0.9580 and RMSE = 0.0081 for test data, had the highest accuracy in estimating the orifice discharge coefficient. The KELM model is in the second rank with a slight difference (R = 0.953 and RMSE = 0.0080 for training data and R = 0.9564 and RMSE = 0.0082 for test data). The GRNN model had the lowest accuracy among machine learning models (R = 0.9202 and RMSE = 0.0104 for training data and R = 0.9291 and RMSE = 0.0106 for test data). The RSM model also had a good accuracy in estimating the elliptical side orifice (R = 0.9279 and RMSE = 0.0097 for training data and R = 0.9456 and RMSE = 0.0092 for test data) by presenting a regression relationship.

Figure 10 shows the error distribution in a violin graph for machine learning models and five regression equations studied in the present study. According to the figure, the lowest error range is related to the GPR model [− 3.78% to + 4.146%]. After the GPR model, the KELM model is in second place with an error range [− 3.981% to + 4.222%]. The GRNN model with the error range [− 5.99% to + 4.833%] has the highest error range among machine learning models. According to Fig. 10, regression-based models have more error ranges. The best regression model (Eq. 1) has an error range of [− 7.057% to + 3.835%]. Equations (3) to (5) have the largest error range.

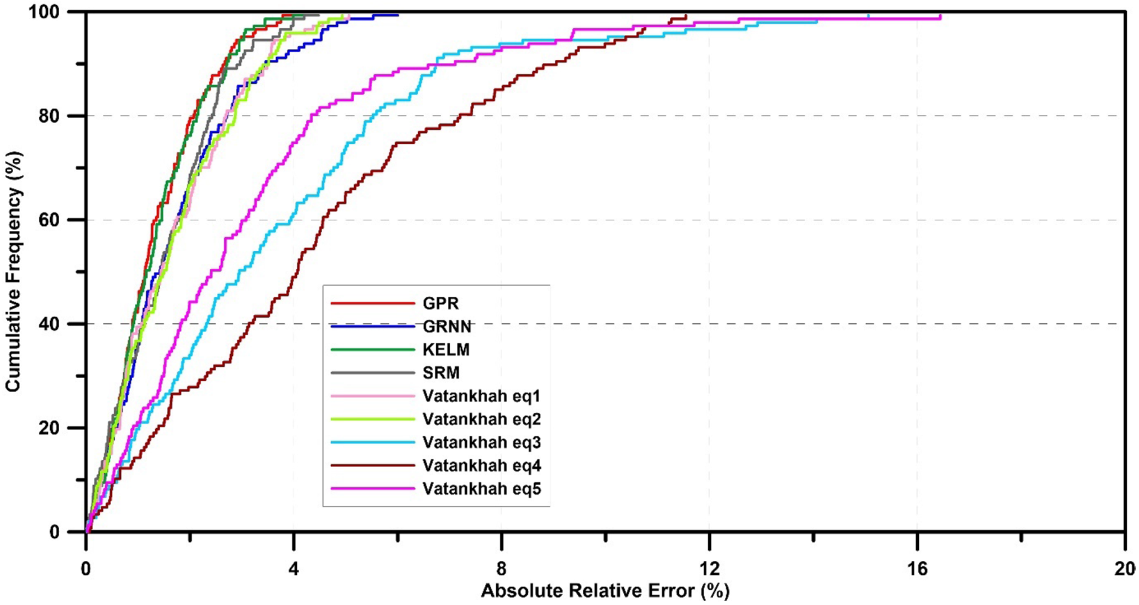

Figure 11. The cumulative frequency (%) of absolute relative error (%) for AI models and regression-based equations.

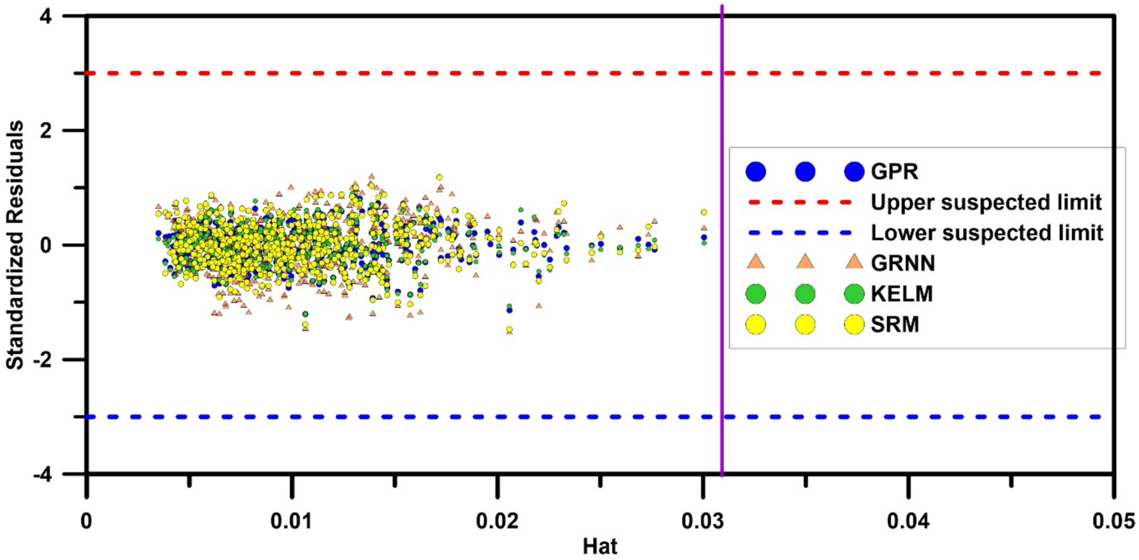

Figure 12. Williams plot to identify the application domain of machine learning models.

Figure 11 shows the cumulative frequency versus absolute error percentage. According to Fig. 11, the GPR model provides an error of less than 1.7% for 70% of the data. This number is 1.74% for the KELM model and 2.16 and 2.03% for GRNN and RSM models. As a result, the GPR model is more accurate and reliable in estimating the elliptical side orifice discharge coefficient. In regression models, Eqs. (1) and (2) for 70% of the data represent an absolute error percentage of less than 2.2%. In Eqs. (3) to (5), the values of this number are equal to 4.76%, 5.7%, and 3.65%, respectively. The mentioned results of the analysis of the cumulative frequency curve against the absolute percentage of error show the superiority of machine learning models over regression-based models.

Finally, to ensure the statistical validity of the developed models, the values of H matrix, leverage index (hat), standard residual percentage R and warning value of leverage H_* was calculated according to the leverage approach, and the Williams diagram was plotted for all machine learning and regression-based models. Figure 12 shows the Williams diagram for the GPR, GRNN, KELM, and RSM machine learning models. According to Fig. 12 in all models, the data obtained from the models are in the range of $-3 < R < 3$ and $0 < H < H_*$ And are therefore statistically valid. Figure 13 shows the Williams diagram for regression-based models. As can be seen from Fig. 13, Eqs. (1), (2), and (5) are statistically valid and are in the range of $-3 < R < 3$ and $0 < H < H_*$ But the Eqs. (3) and (4) are not in the range of confidence, and therefore their application is not recommended in estimating the discharge coefficient of elliptical side orifice.

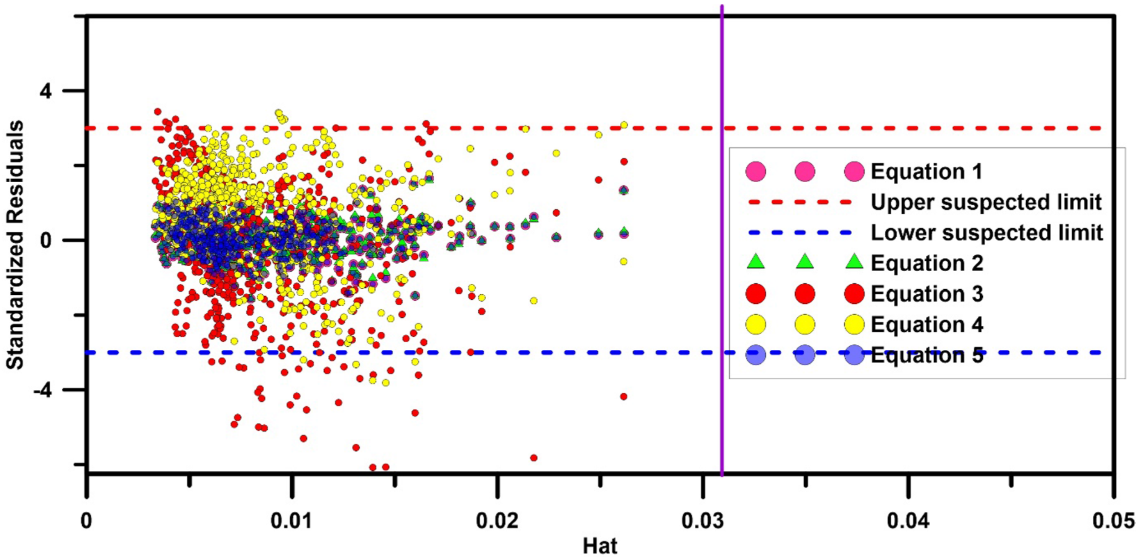

Figure 13. Williams plot for identifying the application domain of regression-based models.

Metrics	All-W/b	All-B/a	All-B/b	All-$b/y1$	All-$Fr1$	All
R	0.9052	0.7932	0.9576	0.9432	0.8968	0.9580
RMSE	0.0120	0.0174	0.0081	0.0094	0.0125	0.0081
MAPE	1.7022	2.4944	1.3336	1.4860	1.9147	1.3243
MBE	− 0.0003	− 0.0012	− 0.0003	− 0.0007	− 0.0007	− 0.0003
NRMSE	2.3179	3.3595	1.5723	1.8159	2.4168	1.5647
AVE rank	3	1	5	4	2	–

Table 7. The statistical measures for sensitivity analysis situations.

Sensitivity analysis. A sensitivity analysis was performed on the data using the GPR model (superior model) to determine the variables affecting the elliptical side orifice discharge coefficient. One of the reliable methods in sensitivity analysis is omitting each data variable and determining statistical parameters in the absence of this variable in model[78]. Table 7 shows the sensitivity analysis results of the variables affecting the elliptical side orifice discharge coefficient. According to Table 7, omitting the parameter B/a (channel width to orifice length) had the greatest effect on reducing the model accuracy (R = 0.7932). Therefore B/a is the most effective parameter in determining the elliptical side orifice discharge coefficient. The Froude number ($Fr1$) with R = 0.8968 is the second parameter affecting the discharge coefficient. The parameters w/b with R = 0.9052, $b/y1$ with R = 0.9432 and B/b with R = 0.9576 are in the rank of 3 to 5 parameters affecting the discharge coefficient.

Conclusion

In the present study, four machine learning methods KELM, GPR, GRNN, and RSM, were used to estimate the elliptical side orifice discharge coefficient. The results were compared with the proposed regression-based equations. The data used to develop the models included 588 series of laboratory data. Five dimensionless parameters: orifice crest height to orifice height ratio (W/b), main channel width to orifice length ratio (B/a), main channel width to orifice height ratio (B/b), upstream orifice depth (y1) to orifice height ratio (y1/b) and upstream orifice Froude number (Fr1) as the model input and the discharge coefficient of side elliptical orifice (Cd) were considered as model output. The results obtained from the statistical parameters of the test dataset showed that all four machine learning models had performed well in estimating the elliptical side orifice discharge coefficient, and the R-value varies between 0.9580 for the GPR model (the strongest model) to 0.9291 for the GRNN model (the weakest model). Comparing machine learning models and regression-based models showed the superiority of artificial intelligence models in estimating the orifice discharge coefficient. The highest accuracy belongs to GPR (RMSE = 0.0081, R = 0.958, MAPE = 1.3242) and KELM (RMSE = 0.0082, R = 0.9564, MAPE = 1.3499) models. The RSM model had good accuracy and provided a functional regression equation for calculating the discharge coefficient. Error analysis using cumulative error distribution curves and relative error distribution function also shows the superiority of the GPR model over other methods used in the present study. Using the RSM model, this study developed a new practical regression equation to predict the elliptic side orifice's discharge coefficient. The leveraged approach was applied to detect outliers and the model applicability domain. Results showed that all proposed machine learning models are statistically valid. Also, the sensitivity analysis result of model input parameters showed that the (B/a) parameter has the most impact on model performance and the (B/b) parameter has the least impact on model performance. The present study results can be used to refine the delivered flow measurement for optimal management of water consumption by the elliptical side orifice structure.

Limitations and future scope. The results of this research are valid for the range of data used, and it is most used in a variety of elliptical sharp-crested side orifices. Therefore, to calculate the discharge coefficient related to different types of circular sections, more effort is needed to collect data sets related to them. The future scope can be investigated by providing an individual model capable of estimating the discharge coefficient of both circular and elliptical orifice by combining corresponded experimental data sets. Also, developing an ensemble model for integrating the advantage of each developed standalone model could be effective in enhancing the accuracy of discharge coefficient computation.

Data availability

The used dataset and codes in this research are available upon reasonable request from the corresponding author.

Received: 7 May 2021; Accepted: 14 September 2021
Published online: 05 October 2021

References

1. Shariq, A., Hussain, A. & Ansari, M. A. Lateral flow through the sharp crested side rectangular weirs in open channels. *Flow Meas. Instrum.* **59**, 8–17 (2018).
2. Hussain, A. & Haroon, A. Numerical analysis for free flow through side rectangular orifice in an open channel. *ISH J. Hydraul. Eng.* **00**, 1–8 (2019).
3. Vatankhah, A. R. & Rafeifar, F. Analytical and experimental study of flow through elliptical side orifices. *Flow Meas. Instrum.* **72**, 101712 (2020).
4. Borghei, S. M., Jalili, M. R. & Ghodsian, M. Discharge coefficient for sharp-crested side weir in subcritical flow. *J. Hydraul. Eng.* **125**, 1051–1056 (1999).
5. Shariq, A., Hussain, A. & Ansari, M. A. Discharge coefficient for side compound weirs in open channel. One Day National Conference on "Recent Innovations in Science, Technology and Engineering" At: National Institute of Technology, Srinagar, Jammu and Kashmir, India, 177–186 (2017).
6. Coşar, A. & Agaccioglu, H. Discharge coefficient of a triangular side-weir located on a curved channel. *J. Irrig. Drain. Eng.* **130**, 410–423 (2004).
7. Emiroglu, M. E., Agaccioglu, H. & Kaya, N. Discharging capacity of rectangular side weirs in straight open channels. *Flow Meas. Instrum.* **22**, 319–330 (2011).
8. Aghayari, F., Honar, T. & Keshavarzi, A. A study of spatial variation of discharge coefficient in broad-crested inclined side weirs. *Irrig. Drain. J. Int. Comm. Irrig. Drain.* **58**, 246–254 (2009).
9. Esmailzadeh, M., Heidarpour, M. & Eslamian, S. S. Flow characteristics of a sharp-crested side sluice gate. *J. Irrig. Drain. Eng.* **141**,1–6 (2015).
10. Ghodsian, M. Flow through side sluice gate. *J. Irrig. Drain. Eng.* **129**, 458–463 (2003).
11. Swamee, P. K. Sluice-gate discharge equations. *J. Irrig. Drain. Eng.* **118**, 56–60 (1992).
12. Bijankhan, M., Kouchakzadeh, S. & Belaud, G. Application of the submerged experimental velocity profiles for the sluice gate's stage-discharge relationship. *Flow Meas. Instrum.* **54**, 97–108 (2017).
13. Vatankhah, A. R. & Mirnia, S. H. Predicting discharge coefficient of triangular side orifice under free flow conditions. *J. Irrig. Drain. Eng.* **144**, 1–14 (2018).
14. Eghbalzadeh, A., Javan, M., Hayati, M. & Amini, A. Discharge prediction of circular and rectangular side orifices using artificial neural networks. *KSCE J. Civ. Eng.* **20**, 990–996 (2015).
15. Ramamurthy, A. S., Tim, U. S., & Sarraf, S. Rectangular Lateral Orifices in Open Channels. *J. Environ. Eng.* **2**, 292–300 (1986).
16. Ramamurthy, A. S., Tim, U. S. & Rao, M. V. J. Weir-orifice units for uniform flow distribution. *J. Environ. Eng. (U. S.)* **113**, 155–166 (1987).
17. Gill, M. A. Flow through side slots. *J. Environ. Eng. (U. S.)* **113**, 1047–1057 (1987).
18. Ojha, C. S. P. & Subbaiah, D. Analysis of flow through lateral slot. *J. Irrig. Drain. Eng.* **123**, 402–405 (1997).
19. Swamee, P. K., Pathak, S. K. & Ali, M. S. Weir orifice units for uniform flow distribution. *J. Irrig. Drain. Eng.* **119**, 1026–1035 (1993).
20. Hussain, A., Ahmad, Z. & Asawa, G. L. Discharge characteristics of sharp-crested circular side orifices in open channels. *Flow Meas. Instrum.* **21**, 418–424 (2010).
21. Hussain, A., Ahmad, Z. & Ojha, C. S. P. Flow through lateral circular orifice under free and submerged flow conditions. *Flow Meas. Instrum.* **52**, 57–66 (2016).
22. Taylor, P. Flow through side square orifices in open channels. *ISH J. Hydraul. Eng.* https://doi.org/10.1080/09715010.2011.10515056 (2011).
23. Hussain, A., Ahmad, Z. & Ojha, C. S. P. Analysis of flow through lateral rectangular orifices in open channels. *Flow Meas. Instrum.* **36**, 32–35 (2014).
24. Akbari, M., Salmasi, F., Arvanaghi, H., Karbasi, M. & Farsadizadeh, D. Application of Gaussian process regression model to predict discharge coefficient of Gated Piano Key Weir. *Water Resour. Manag.* **33**, 3929–3947 (2019).
25. Karbasi, M. & Azamathulla, H. M. Prediction of scour caused by 2D horizontal jets using soft computing techniques. *Ain Shams Eng. J.* **8**, 559–570 (2017).
26. Zaji, A. H. & Bonakdari, H. Performance evaluation of two different neural network and particle swarm optimization methods for prediction of discharge capacity of modified triangular side weirs. *Flow Meas. Instrum.* **40**, 149–156 (2014).
27. Zhou, J., Huang, S., Wang, M. & Qiu, Y. Performance evaluation of hybrid GA–SVM and GWO–SVM models to predict earthquake-induced liquefaction potential of soil: a multi-dataset investigation. *Eng. Comput.* **37**, 1–19 (2021).
28. Zhou, J. *et al.* Optimization of support vector machine through the use of metaheuristic algorithms in forecasting TBM advance rate. *Eng. Appl. Artif. Intell.* **97**, 104015 (2021).
29. Zhou, J. *et al.* Predicting TBM penetration rate in hard rock condition: A comparative study among six XGB-based metaheuristic techniques. *Geosci. Front.* **12**, 101091 (2021).
30. Wang, S. *et al.* Rockburst prediction in hard rock mines developing bagging and boosting tree-based ensemble techniques. *J. Cent. South Univ.* **28**, 527–542 (2021).
31. Momeni, E., Yarivand, A., Dowlatshahi, M. B. & Armaghani, D. J. An efficient optimal neural network based on gravitational search algorithm in predicting the deformation of geogrid-reinforced soil structures. *Transp. Geotech.* **26**, 100446 (2021).

32. Geetha, N. K. & Bridjesh, P. Overview of machine learning and its adaptability in mechanical engineering. *Mater. Today Proc.* (2020).
33. Ahmadianfar, I., Jamei, M. & Chu, X. Prediction of local scour around circular piles under waves using a novel artificial intelligence approach. *Mar. Georesour. Geotechnol.* **39**, 1–12 (2019).
34. Jamei, M. & Ahmadianfar, I. Prediction of scour depth at piers with debris accumulation effects using linear genetic programming. *Mar. Georesour. Geotechnol.* **38**, 468–479 (2020).
35. Lodomez, M. *et al.* Nappe oscillations on free-overfall structures, data from laboratory experiments. *Sci. Data* **7**, 1–10 (2020).
36. Yaseen, Z. M., Ali, M., Sharafati, A., Al-Ansari, N. & Shahid, S. Forecasting standardized precipitation index using data intelligence models: Regional investigation of Bangladesh. *Sci. Rep.* **11**, 1–25 (2021).
37. Ebtehaj, I., Bonakdari, H., Zaji, A. H., Azimi, H. & Khoshbin, F. GMDH-type neural network approach for modeling the discharge coefficient of rectangular sharp-crested side weirs. *Eng. Sci. Technol. Int. J.* **18**, 746–757 (2015).
38. Khoshbin, F. *et al.* Adaptive neuro-fuzzy inference system multi-objective optimization using the genetic algorithm/singular value decomposition method for modelling the discharge coefficient in rectangular sharp-crested side weirs. *Eng. Optim.* **48**, 933–948 (2016).
39. Asadi, M. J., Shabanlou, S., Najarchi, M. & Najafizadeh, M. M. A hybrid intelligent model and computational fluid dynamics to simulate discharge coefficient of circular side orifices. *Iran. J. Sci. Technol. Trans. Civ. Eng.* **45**, 1–26 (2020).
40. Parsaie, A., Haghiabi, A. H., Saneie, M. & Torabi, H. Predication of discharge coefficient of cylindrical weir-gate using adaptive neuro fuzzy inference systems (ANFIS). *Front. Struct. Civ. Eng.* **11**, 111–122 (2017).
41. Khorchani, M. & Blanpain, O. Development of a discharge equation for side weirs using artificial neural networks. *J. Hydroinform.* **7**, 31–39 (2005).
42. Dursun, O. F., Kaya, N. & Firat, M. Estimating discharge coefficient of semi-elliptical side weir using ANFIS. *J. Hydrol.* **426–427**, 55–62 (2012).
43. Parsaie, A. & Haghiabi, A. H. Prediction of discharge coefficient of side weir using adaptive neuro-fuzzy inference system. *Sustain. Water Resour. Manag.* **2**, 257–264 (2016).
44. Eghbalzadeh, A., Javan, M., Hayati, M. & Amini, A. Discharge prediction of circular and rectangular side orifices using artificial neural networks. *KSCE J. Civ. Eng.* **20**, 990–996 (2016).
45. Bonakdari, H., Gharabaghi, B., Ebtehaj, I. & Sharifi, A. A New Approach to estimate the discharge coefficient in sharp-crested rectangular side orifices using gene expression programming. in *Science and Information Conference* 77–96 (Springer, 2020).
46. Azimi, H., Bonakdari, H. & Ebtehaj, I. A highly efficient gene expression programming model for predicting the discharge coefficient in a side weir along a trapezoidal canal. *Irrig. Drain.* **66**, 655–666 (2017).
47. Azamathulla, H. M., Haghiabi, A. H. & Parsaie, A. Prediction of side weir discharge coefficient by support vector machine technique. *Water Sci. Technol. Water Supply* **16**, 1002–1016 (2016).
48. Azimi, H., Bonakdari, H. & Ebtehaj, I. Design of radial basis function-based support vector regression in predicting the discharge coefficient of a side weir in a trapezoidal channel. *Appl. Water Sci.* **9**, 78 (2019).
49. Parsaie, A., Azamathulla, H. M. & Haghiabi, A. H. Prediction of discharge coefficient of cylindrical weir—Gate using GMDH-PSO. *ISH J. Hydraul. Eng.* **24**, 116–123 (2017).
50. Zarei, S., Yosefvand, F. & Shabanlou, S. Discharge coefficient of side weirs on converging channels using extreme learning machine modeling method. *Measurement* **152**, 107321 (2020).
51. Mahmodian, A. R., Rajabi, A., Izadbakhsh, M. A. & Shabanlou, S. Evaluation of side orifices shape factor using the novel approach self-adaptive extreme learning machine. *Model. Earth Syst. Environ.* **5**, 925–935 (2019).
52. Gharib, R., Heydari, M., Kardar, S. & Shabanlou, S. Simulation of discharge coefficient of side weirs placed on convergent canals using modern self-adaptive extreme learning machine. *Appl. Water Sci.* **10**, 50 (2020).
53. Reza, A., Ahmad, M., Mohammad, R., Izadbakhsh, A. & Shabanlou, S. Evaluation of side orifices shape factor using the novel approach self- adaptive extreme learning machine. *Model. Earth Syst. Environ.***5**, 925–935 (2019).
54. Jamei, M., Ahmadianfar, I., Chu, X. & Yaseen, Z. M. Estimation of triangular side orifice discharge coefficient under a free flow condition using data-driven models. *Flow Meas. Instrum.* **77**, 1–19 (2020).
55. Ramamurthy, A. S., Tim, U. S. & Sarraf, S. Rectangular lateral orifices in open channels. *J. Environ. Eng.* **112**, 292–300 (1986).
56. Liu, Y., Guo, J., Wang, Q. & Huang, D. Prediction of filamentous sludge bulking using a state-based Gaussian processes regression model. *Sci. Rep.* **6**, 31303 (2016).
57. Rasmussen, C. E. Gaussian processes in machine learning. in *Summer School on Machine Learning* 63–71 (Springer, 2003).
58. Fu, J.-C., Huang, H.-Y., Jang, J.-H. & Huang, P.-H. River stage forecasting using multiple additive regression trees. *Water Resour. Manag.* **33**, 4491–4507 (2019).
59. Williams, C. K. I. & Rasmussen, C. E. *Gaussian Processes for Machine Learning* Vol. 2 (MIT Press, 2006).
60. Noack, M. M. *et al.* Autonomous materials discovery driven by Gaussian process regression with inhomogeneous measurement noise and anisotropic kernels. *Sci. Rep.* **10**, 1–16 (2020).
61. Schulz, E., Speekenbrink, M. & Krause, A. A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions. *J. Math. Psychol.* **85**, 1–16 (2018).
62. Momeni, E., Dowlatshahi, M. B., Omidinasab, F., Maizir, H. & Armaghani, D. J. Gaussian process regression technique to estimate the pile bearing capacity. *Arab. J. Sci. Eng.* **45**, 8255–8267 (2020).
63. Huang, G.-B., Zhu, Q.-Y. & Siew, C.-K. Extreme learning machine: Theory and applications. *Neurocomputing* **70**, 489–501 (2006).
64. Hoerl, A. E. & Kennard, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67 (1970).
65. Huang, G.-B., Zhou, H., Ding, X. & Zhang, R. Extreme learning machine for regression and multiclass classification. *IEEE Trans. Syst. Man Cybern. Part B* **42**, 513–529 (2011).
66. Specht, D. F. A general regression neural network. *IEEE Trans. Neural Netw.* **2**, 568–576 (1991).
67. Kisi, O. The potential of different ANN techniques in evapotranspiration modelling. *Hydrol. Process. Int. J.* **22**, 2449–2460 (2008).
68. Kim, S., Shiri, J. & Kisi, O. Pan evaporation modeling using neural computing approach for different climatic zones. *Water Resour. Manag.* **26**, 3231–3249 (2012).
69. Jamei, M., Ahmadianfar, I., Olumegbon, I. A., Karbasi, M. & Asadi, A. On the assessment of specific heat capacity of nanofluids for solar energy applications: Application of Gaussian process regression (GPR) approach. *J. Energy Storage* **33**, 102067 (2020).
70. Bucher, C. G. & Bourgund, U. A fast and efficient response surface approach for structural reliability problems. *Struct. Saf.* **7**, 57–66 (1990).
71. Hamid, H. A., Jenidi, Y., Thielemans, W., Somerfield, C. & Gomes, R. L. Predicting the capability of carboxylated cellulose nanowhiskers for the remediation of copper from water using response surface methodology (RSM) and artificial neural network (ANN) models. *Ind. Crops Prod.* **93**, 108–120 (2016).
72. Mehmood, T., Ramzan, M., Howari, F., Kadry, S. & Chu, Y.-M. Application of response surface methodology on the nanofluid flow over a rotating disk with autocatalytic chemical reaction and entropy generation optimization. *Sci. Rep.* **11**, 1–18 (2021).
73. Rousseeuw, P. J. & Leroy, A. M. *Robust Regression and Outlier Detection* Vol. 589 (Wiley, 2005).
74. Shateri, M., Ghorbani, S., Hemmati-Sarapardeh, A. & Mohammadi, A. H. Application of Wilcoxon generalized radial basis function network for prediction of natural gas compressibility factor. *J. Taiwan Inst. Chem. Eng.* **50**, 131–141 (2015).
75. Su, X., An, J., Zhang, Y., Zhu, P. & Zhu, B. Prediction of ozone hourly concentrations by support vector machine and kernel extreme learning machine using wavelet transformation and partial least squares methods. *Atmos. Pollut. Res.* **11**, 51–60 (2020).

76. Fu, Q. *et al.* Prediction of the diet energy digestion using kernel extreme learning machine: A case study with Holstein dry cows. *Comput. Electron. Agric.* **169**, 105231 (2020).
77. He, F. *et al.* Construction and evaluation of two computational models for predicting the incidence of influenza in Nagasaki Prefecture, Japan. *Sci. Rep.* **7**, 7192 (2017).
78. Deng, Z. *et al.* Data-driven state of charge estimation for lithium-ion battery packs based on Gaussian process regression. *Energy* **205**, 118000 (2020).

Author contributions

All authors contributed to discussions and preparation of the manuscript, M.K. and M.J. were responsible for performing the simulations and analyzing the data, and I.A. proposed the research methodology and drafted the manuscript, and A.A. edited the manuscript, answered the reviewers' comments, and led the Project.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to A.A.

Reprints and permissions information is available at www.nature.com/reprints.

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