Using data mining methods to improve discharge coefficient prediction in Piano Key and Labyrinth weirs

Mahdi Majedi-Asl a,*, Mehdi Fuladipanah b, Venkat Arun c and Ravi Prakash Tripathi d

a Department of Civil Engineering, Faculty of Engineering, University of Maragheh, Maragheh, Iran
b Department of Civil Engineering, Ramhormoz Branch, Islamic Azad University, Ramhormoz, Iran
c Department of Civil Engineering, Haramaya Institute of Technology, Haramaya University, Haramaya, Ethiopia
d Department of Civil Engineering, Rajkiya Engineering College, Sonbhadra, Varnasi, India

*Corresponding author. E-mail: mehdi.majedi@gmail.com

ABSTRACT

As a remarkable parameter, the discharge coefficient (Cd) plays an important role in determining weirs’ passing capacity. In this research work, the support vector machine (SVM) and the gene expression programming (GEP) algorithms were assessed to predict Cd of piano key weir (PKW), rectangular labyrinth weir (RLW), and trapezoidal labyrinth weir (TLW) with gathered experimental data set. Using dimensional analysis, various combinations of hydraulic and geometric non-dimensional parameters were extracted to perform simulation. The superior model for the SVM and the GEP predictor for PKW, RLW, and TLW included Ho = P, N, W = P(ð), Ho = P, W = P, N ð and H o = P, Lc = W, A = W, Fr(ð) respectively. The results showed that both algorithms are potential in predicting discharge coefficient, but the coefficient of determination (RMSE, R2, Cd(DDR)max) illustrated the superiority of the GEP performance over the SVM. The results of the sensitivity analysis determined the highest effective parameters for PKW, RLW, and TLW in predicting discharge coefficients are Ho = P, Ho = P, and Fr respectively.

Key words: data-driven solver, error distribution, nonlinear weir, sensitivity analysis

HIGHLIGHTS

• Three different types of weirs have been studied in this paper.
• Two SVM and GEP algorithms have been implemented to predict the discharge coefficient of three weirs.
• Eighteen combinations of dimensionless parameters have been tested to achieve optimum prediction of the discharge coefficient.
• An equation for a superior model has been extracted to simulate discharge coefficient.

1. INTRODUCTION

Weirs are a type of hydraulic structure implemented for water surface control, flow measurement, and passing excess water volume. The shape and the length of the weir crest have a remarkable role in determining overflow capacity. Having more crest length, nonlinear weirs named labyrinth type are a form of weir that can convey more flow for an equal hydraulic head than linear types. Their crest is often triangular, rectangular, trapezoidal, and arched in shape. The functional differences of labyrinth weirs are illustrated with discharge coefficient, Cd, dependent on the hydraulic and geometric parameters. Research in recent decades indicated that because of the complexity and diversity of effective parameters on the Cd, numerical simulation has been developed to predict discharge coefficient. The data-driven (soft computing) approach is another technique that has also been used to predict the Cd using experimental data during train and test phases. These models are capable of extracting complex and hidden relationships among dependent and independent variables (Samadi et al. 2014, 2015, 2020).

This is an Open Access article distributed under the terms of the Creative Commons Attribution Licence (CC BY 4.0), which permits copying, adaptation and redistribution, provided the original work is properly cited (http://creativecommons.org/licenses/by/4.0/).
was reported by Kisi (2013) and Emiroglu & Kisi (2013). Ebtehaj et al. (2015a, 2015b, 2015c) utilized data-mining methods to determine the $C_d$ of side weirs located on the sidewalls of rectangular channels. Their results proved the significant acceptable accuracy of these models. Azamathulla et al. (2016) used the SVM to predict the $C_d$ for side weir. The determination coefficients for the train and the test phases were obtained as 0.96 and 0.93 respectively, implying more accuracy of the simulation. Parsaei & Haghiabi (2017) performed a comparative study using the GMDH, the multivariate adaptive regression splines (MARS), and the SVM to predict a combined weir-gate discharge coefficient. Their results confirmed not only that all models have the capability of prediction but also the performance of the SVM algorithm is superior to other methods. Norouzi et al. (2019) performed a comparative study to predict discharge coefficient using the SVM and the ANN of trapezoidal labyrinth weir. The results of their study showed not only the good performance of the two models but also the remarkable superiority of the SVM. Azimi et al. (2019) employed the SVR to predict the $C_d$ of a side weir in a trapezoidal channel. They claimed that the SVR model simulates the $C_d$ with acceptable accuracy. Mohammed & Sharifi (2020) proposed using the GEP to predict discharge coefficient for oblique side weir. Zhou et al. (2015), Zaji et al. (2016), Zaji & Bonakdari (2017), Roushangar et al. (2017), Nadiri et al. (2018), Majedi Asl & Fuladipanah (2019), Sadeghfam et al. (2019), Kumar et al. (2019) and Roushangar et al. (2021) proposed to evaluate discharge coefficient using data-driven methods.

Given the significance of the data-driven methods, this paper probes the exert of the GEP and the SVM algorithms to acquire the discharge coefficient of three different weirs, i.e. piano key, rectangular labyrinth, and trapezoidal labyrinth. Further, a new regression equation is extracted to predict the discharge coefficient. Finally, using the sensitivity analysis, the order of the effective parameters on the $C_d$ is determined.

2. METHODS AND MATERIALS

2.1. Experimental models description

Two SVM and GEP models have been used to predict the $C_d$ for piano key weir (PKW), rectangular labyrinth weir (RLW), and trapezoidal labyrinth weir (TLW). Experimental data were gathered from Rostami et al. (2018) and Seamons (2014). Figures 1 and 2 show the geometric characteristics of the weirs.
2.2. Discharge coefficient function

Rostami et al. (2018) performed their experiments in a rectangular flume of length 10 m, width 0.3 m, and height 0.6 m at the Khuzestan Water and Power Authority Laboratory, Iran. The range of experimental geometric characteristics of the weir has been presented in Table 1. In this table, W is the width of each cycle, P is the weir height, L is the effective length of the weir, Ts is the weir thickness, Wi and Wo are the width of inlet and outlet keys of the weir, respectively, and N is the number of keys in the weir.

They extracted and expressed discharge coefficient $C_d$ as the following equation for the PKW:

$$C_d = F\left(\frac{H_o}{P}, N, \frac{W}{P}\right)$$  \hspace{1cm} (1)

where $H_o$ is the total upstream head of the weir. Seamons (2014) conducted the experiments in a rectangular flume 14.6 m in length, 1.2 m in width, and 0.9 m in depth at the Water Research Laboratory, Utah State University, Logan, UT, USA. The total number of data sets was 313 by conducting experiments on 13 labyrinth weirs. Table 2 has summarized the variation range of the applied parameters. In this table, $H_T$ displays the total head of flow above the weir crest. Geometric parameters have been introduced in Figure 2.

Seamons (2014) presented $C_d$ as a function of geometry characteristics and hydraulic conditions for the RLW and TLW as shown by Equations (2) and (3), respectively:

$$C_d = F\left(\frac{H_o}{P}, N, \frac{w}{P}\right)$$  \hspace{1cm} (2)

$$C_d = F\left(\frac{H_o}{P}, \frac{L_c}{W}, \alpha, \frac{w}{P}, \frac{A}{w}, F_r\right)$$  \hspace{1cm} (3)

2.3. Overview of Support Vector Machine (SVM)

Developed by Cortes & Vapink (1995), the SVM is an optimization algorithm for binary classification that separates two classes using a border. In this optimization-based algorithm, samples are determined to form the bound of the two classes. These samples are named as support vectors. Similar to other regression models, the dependent variable $y$ is predicted

Table 1 | The range of geometric characteristics of the piano key weir (Rostami et al. 2018)

| Model        | W   | W/P | L (cm) | Ts (mm) | N  | Wi/Wo | Number of experiments |
|--------------|-----|-----|--------|---------|----|-------|----------------------|
| Piano key    | 15  | 2, 2.5, 3 | 148.5  | 5       | 2  | 1.25  | 36                   |
| Piano key    | 30  | 2, 2.5, 3 | 148.5  | 5       | 1  | 1.25  | 36                   |

Table 2 | The range of hydraulic and geometric variables of weirs (Seamons 2014)

| Variables  | Range          | Values   |
|------------|----------------|----------|
| Q (m$^3$/s) | 0.02–0.64      | –        |
| $H_T$      | 0.051–0.835    | –        |
| $C_d$      | 0.309–0.684    | –        |
| W (m)      | –              | 1.15, 1.23|
| N          | –              | 2        |
| $\alpha$ (degree) | –    | 12, 15   |
| $t_w$ (m)  | –              | 0.04     |
| P (m)      | –              | 0.3, 0.38, 0.46 |
using several independent variables \( x_i \). Finally, the following regression equation is formed:

\[
y = \mathbf{W}^T \Phi(x) + b + \text{noise} \tag{4}
\]

where the amount of the noise is determined using an amount of allowable error \( \varepsilon \). In Equation (4), \( \mathbf{W} \) is the coefficients vector, \( b \) is a constant value, and \( \Phi \) is the nuclear function. This algorithm aims to find the \( f(x) \) so that optimization of the error function is performed using the model training with a data series. The optimizing process is carried out by minimizing Equation (5) under mentioned conditions in Equation (6):

\[
\frac{1}{2} \mathbf{W}^T \cdot \mathbf{W} + C \sum_{i=1}^{N} \xi_i + C \sum_{i=1}^{N} \xi_i^d \tag{5}
\]

\[
\begin{align*}
\mathbf{W}^T \cdot \Phi(x_i) + b - y_i & \leq \varepsilon + \xi_i \\
\mathbf{W}^T \cdot \Phi(x_i) - b & \leq \varepsilon + \xi_i^d \\
\xi_i, \xi_i^d & \geq 0
\end{align*} \tag{6}
\]

where \( C \) indicates the penalty while occurring error, \( N \) is the number of the samples, \( \xi_i \) and \( \xi_i^d \) are deficiency coefficients. The algorithm processes the considered function by minimizing three terms in Equation (5). The nuclear function, \( \Phi(x) \) called the Kernel function, in Equation (5) is defined as follows:

\[
\Phi(x_i)^T = \mathbf{K}(x_i, x) \tag{7}
\]

Some of the most important nuclear functions are presented in Table 3. In this table \( \gamma, C \) and \( d \) are kernel parameters. The SVM generalization performance (estimation accuracy) depends on a good set of meta-parameters parameters \( \gamma, C \) and \( d \) are the kernel parameters. The choices \( \gamma, r \), and \( d \) control the prediction (regression) model complexity. Of all nuclear functions, the radial-based function (RBF) is the most used one and is proposed by notable researchers. In RBF, \( \sigma \) is the function parameter. The characteristic parameters of the SVM, i.e., \( \varepsilon \) and \( C \), are optimized, and the value of the setting parameter, \( \gamma \), is determined with the trade-off between the minimum fitting error and the estimated function. The user, via trial and error, determines the SVM parameters. In the RBF function, \( \varepsilon \) is the radius of the tube within which the regression function must lie.

2.4. Overview of gene expression programming

Invented by Ferreira (2001, 2006), GEP is an artificial procedure to solve genotype systems. It is used for solving complex real-world problems (Azamathullah 2012; Samadi et al. 2021). Computer programs with different shapes and lengths encoded in linear chromosomes with fixed sizes are progressed with GEP. Encoding of chromosome information is the final step of the GEP algorithm leading to the tree expression, called the translation process. The GEP algorithm involves four main steps: (1) initialize the population by creating the chromosomes (individuals); (2) identify a suitable fitness function to evaluate the best individual; (3) conduct genetic operations to modify the individuals to achieve the optimal solution in the next generation; (4) check the stop conditions. The flowchart of the GEP algorithm is illustrated in Figure 3.

**Table 3 | Kernel functions**

| Kernel's function name            | Function                                      |
|-----------------------------------|-----------------------------------------------|
| Linear                           | \( \mathbf{K}(x_i, x_j) = x_i^T x_j \)        |
| Poly-nominal function            | \( \mathbf{K}(x_i, x_j) = |y x_i^T x_j + C|^{1/d} \) |
| Radial basis function            | \( \mathbf{K}(x_i, x_j) = \exp(-|y x_i - x_j|^2) \) |
| Sigmoid function                 | \( \mathbf{K}(x_i, x_j) = \tanh(y x_i^T x_j + C) \) |
2.5. Evaluation criteria

The efficiency of the classic regression and data-driven models have been assessed using the following indices:

\[
RMSE = \sqrt{\frac{\sum_{i=1}^{N} (x_{o,i} - x_{p,i})^2}{N}} \tag{8}
\]

\[
R^2 = \frac{\sum_{i=1}^{N} (x_{o,i} - \bar{x}_{o})(x_{p,i} - \bar{x}_{p})^2}{\sum_{i=1}^{N} (x_{o,i} - \bar{x}_{o})^2 \sum_{i=1}^{N} (x_{p,i} - \bar{x}_{p})^2} \tag{9}
\]

Here, \(x_{o,i}\), \(x_{p,i}\), and \(N\) are the observed discharge coefficient, the predicted discharge coefficient, and test number, respectively.

However, on the other side, mentioned statistical indices only show mean error in model performance without error distribution. Therefore, it is important to test the model using some other performance evaluation criterion to check its robustness. Another criterion has been applied in this paper, based on discrepancy ratio (DR) defined by White et al. (1973), to check the model robustness as following:

\[
DR = \frac{\text{Predicted value}}{\text{Observed value}} \tag{10}
\]

DR is commonly used as an error measure in the literature and is used widely by many researchers such as Seo & Cheong (1998), Deng et al. (2002), KashefiPour & Falconer (2002), Tayfur & Singh (2005), and others. However, it is not utilizable for negative and zero values. To remedy this problem, Noori et al. (2010) have presented the developed discrepancy ratio (DDR) statistic:

\[
DDR = \frac{\text{Predicted value}}{\text{Observed value}} - 1 \tag{11}
\]

For better judgment and visualization, the Gaussian function of DDR values could be calculated and illustrated in a standard normal distribution format. For this reason, firstly, the DDR values must be standardized and then using the Gaussian
function, the normalized value of DDR ($C_d^{(DDR)}$) is calculated. Generally, more tendencies in the error distribution graph to the centerline and larger value of the maximum $C_d^{(DDR)}$ are equal to more accuracy.

### 3. RESULTS AND DISCUSSION

A different combination of dimensionless parameters has been implemented to assess various models to predict discharge coefficient using SVM and GEP algorithms for the PKW, RLW, and TLW. These combinations have been presented in Table 4.

#### 3.1. SVM solver

The total number of data used for modeling was 72. Of all measured data, the share of train and test phases were 75% and 25%, respectively. The results of the SVM algorithm for the PKW showed model 1 has the best performance among the five models. The setting parameters $C$, $\gamma$, and $\varepsilon$ were obtained as 40, 4, and 0.1 for the optimum model with $R^2 = 0.9785$ and RMSE = 0.02418 for the train stage and $R^2 = 0.9789$ and RMSE = 0.027 for the test phase. These results were obtained with the RBF kernel function. Figure 4 illustrates a scatter plot of measured and predicted values of $C_d$ and conformity of data points for $C_d$ during train and test phases for model 1. The distribution of predicted and observed data around the fit

---

**Table 4** | Various combinations of dimensionless parameters to predict discharge coefficient of PKW, RLW, and TLW

| Weir type | Model number | Dimensionless II | Weir type | Model number | Dimensionless II | Weir type | Model number | Dimensionless II |
|-----------|--------------|------------------|-----------|--------------|------------------|-----------|--------------|------------------|
| PKW       | Model 1      | $H_o \frac{P}{F}$, $N$, $W$ | RLW       | Model 1      | $H_o \frac{P}{F}$, $N$, $W$ | TLW       | Model 1      | $H_o \frac{P}{F}$, $L_c$, $\alpha$, $W$, $A$, $W$, $F_r$ |
|           | Model 2      | $H_o \frac{P}{F}$, $W$ |           | Model 2      | $H_o \frac{P}{F}$, $N$ |           | Model 2      | $H_o \frac{P}{F}$, $L_c$, $W$, $A$, $W$, $F_r$ |
|           | Model 3      | $H_o \frac{P}{F}$, $N$ |           | Model 3      | $N$, $W$ |           | Model 3      | $H_o \frac{P}{F}$, $L_c$, $A$, $W$, $F_r$ |
|           | Model 4      | $N$, $W$ |           | Model 4      | $H_o \frac{P}{F}$, $W$ |           | Model 4      | $H_o \frac{P}{F}$, $\alpha$, $A$, $W$, $F_r$ |
|           | Model 5      | $H_o \frac{P}{F}$ |           | Model 5      | $H_o \frac{P}{F}$ |           | Model 5      | $H_o \frac{P}{F}$, $\alpha$, $A$, $W$ |
|           | Model 6      | $H_o \frac{P}{F}$, $\alpha$, $W$ |           | Model 6      | $H_o \frac{P}{F}$ |           | Model 6      | $H_o \frac{P}{F}$, $\alpha$, $W$ |
|           | Model 7      | $H_o \frac{P}{F}$, $F_r$ |           | Model 7      | $H_o \frac{P}{F}$ |           | Model 7      | $H_o \frac{P}{F}$ |
|           | Model 8      | $H_o \frac{P}{F}$ |           | Model 8      | $H_o \frac{P}{F}$ |           | Model 8      | $H_o \frac{P}{F}$ |

---

**Figure 4** | SVM solver’s performance for model 1 for the PKW.
line and their conformity during the train and the test phases determine the high accuracy of the SVM model. More statistical indices have been presented in Table 5, demonstrating the fitness among measured and predicted values of the discharge coefficient. As it is seen, all indices have very close values through the test and train stages.

The distribution of the standard distribution of $C_d(DDR)$ values for the SVM model through train and test phases has been shown in Figure 5. The maximum values of $C_d(DDR)$ for the train and the test phases were obtained at 8.153 and 8.698, respectively. These values show that the model has better performance during the test stage than the train.

A sensitivity analysis was performed for model 1 in Table 6 to evaluate the impact of input variables on the estimated discharge coefficient. In this analysis, each time a parameter was omitted from the model inputs, then the model was implemented, and accuracy was evaluated as the most significant parameter. The deleted parameter with the most effect on declined model precision and boosted model error was rated as the most remarkable parameter. A look at the analysis of the results reveals that dropping the parameter $H_o/P$ has increased the RMSE to 0.1569 and 0.1905 and has decreased $R^2$ to 0.1151 and 0.0127 through train and test phases, respectively. The calculation demonstrated that the prominent and sensitive parameter for the SVM predictor is $H_o/P$. The second and the third most important and effective parameters are $w/P$ and $N$, respectively.

The next simulation of $C_d$ belongs to RLW. The number of data for the RLW is 33. The share of the train and the test phases are 73% and 27%, respectively. The first model of Table 4 has had the best performance to predict the discharge coefficient. The setting parameters $C$, $\gamma$, and $\epsilon$ were obtained as 60, 1.4, and 0.1, respectively. The values of $(R^2, \text{RMSE})$ for the train and the test phases were calculated as (0.9745, 0.0208) and (0.9734, 0.0138), respectively. A scatter plot of observed versus predicted $C_d$ and a plot of data point against discharge coefficient values for the train and the test stages have been presented in

**Table 5 | Summary of the SVM performance assessment for PKW**

| Statistical indices | Train phase | | Test phase | |
|---------------------|-------------|----------------|-------------|----------------|
|                     | Measured data | Predicted data | Measured data | Predicted data |
| Number of data      | 54          | 18             | 18           | 18             |
| Correlation coefficient | 0.9892     | 0.9894         |              |                |
| Mean                | 0.62        | 0.63           | 0.58         | 0.58           |
| Maximum             | 0.930       | 1              | 0.89         | 0.86           |
| Minimum             | 0.34        | 0.37           | 0.38         | 0.38           |
| STDEV               | 0.1680      | 0.1627         | 0.1829       | 0.1733         |

**Figure 5 | Distribution of $C_d(DDR)$ against $Z_{DDR}$ for the SVM model for PKW.**

Downloaded from http://iwaponline.com/ws/article-pdf/22/2/1964/100966/ws022021964.pdf by guest
Figure 6. As it is clear, the data are less scattered around the fit line and have significant conformity during the test and train phases, indicating remarkable acceptance performance for the SVM model. Table 7 displays the summary of the statistical indices for the performance of the SVM algorithm for a superior combination of parameters. Very little difference among indices proves the good simulation of the SVM model.

Figure 7 illustrates the $C_d(DDR)$ vs. $Z_{DDR}$ for the superior model through train and test stages. The maximum $C_d(DDR)$ for the train and the test stages are 8.256 and 14.255, respectively. These values verify the better performance of the SVM during the test phase than the train, indicating correct operation. A sensitivity analysis has been performed for model 1 (Table 8). Removing $H_pP$ has changed the model function dramatically. The sharp difference is evident between the outputs of the model

**Table 6** | Sensitivity analysis of the SVM model for PKW

| Top model | Omitted variable | Training phase | Test phase |
|-----------|-----------------|----------------|------------|
|           |                 | $R^2$ | RMSE | $R^2$ | RMSE |
| Model 1   | None            | 0.9785 | 0.0248 | 0.9789 | 0.0270 |
|           | $w$             | 0.9112 | 0.0502 | 0.8699 | 0.0657 |
|           | $F$             | 0.9511 | 0.0369 | 0.9602 | 0.0366 |
|           | $N$             | 0.1151 | 0.1569 | 0.0127 | 0.1905 |

**Table 7** | Summary of the SVM performance assessment for RLW

| Statistical indices | Train phase | Test phase |
|---------------------|-------------|------------|
|                     | Measured data | Predicted data | Measured data | Predicted data |
| Number of data      | 24          | 9           |
| Correlation coefficient | 0.9871         | 0.9917     |
| Mean                | 0.43         | 0.43        | 0.37          | 0.38          |
| Maximum             | 0.64         | 0.62        | 0.57          | 0.56          |
| Minimum             | 0.27         | 0.26        | 0.26          | 0.26          |
| STDEV               | 0.1255       | 0.1298      | 0.0955        | 0.0955        |
preference by the omission of HoP with decreasing R² from 0.9745 to 0.0636 and from 0.9834 to 0.3682 for the train and test stages. Increasing RMSE from 0.0208 to 0.1209 and from 0.0138 to 0.1019 through the train and the test phases illustrate the highest effect of parameter Ho

The discharge coefficient of the TLW is the third modeling process with the SVM. The share of the train and the test phases from the total observed data (140 data) were 80% and 20%, respectively. Among the eight mentioned combinations in Table 4, combination 3 had the highest adaptation with measured discharge coefficients. This model includes Ho/P, Lc/W, A/W and Fr parameters. Setting parameters C, \( \gamma \), and \( \varepsilon \) of the SVM algorithm was obtained 100, 1, and 0.1, respectively. The values of the (R², RMSE) for the train and test phases were (0.9896, 0.009) and (0.9886, 0.0144). The measured versus the predicted values of \( C_d \) have been illustrated in Figure 8. High values of well performance indices are well known from this figure. All measured and predicted data have scattered around line 1:1 and conformity of datasets during the train and the test stages indicate the high values of correlation and fewer values of inaccuracy. The summary of statistical characteristics of the predicted and measured values for \( C_d \) has been presented in Table 9, an improved SVM performance for the TLW. Figure 9 illustrates the distribution of standardized \( C_d(DDR) \) vs. Z_DDR for the TLW. The maximum values of \( C_d(DDR) \) for the train and the test stages are 25.724 and 33.524, respectively. The better simulation for the test phase proves the correct performance of modeling. The sensitivity analysis (Table 10) shows Fr is the most effective parameter because of the most decreasing R² and most increasing RMSE by omitting this parameter from the modeling process. The other effective parameters in descending order are Ho/P, Lc/w and A/w, respectively.

3.2. GEP solver

In this section, the result of the GEP’s simulation has been presented for the three weirs. Three mentioned models of PKW in Table 4 were examined, and the first one has the best output based on the presented chromosome’s properties in Table 11. The share of the train and the test phases of all data are 75 and 25%, respectively. The values of RMSE as fitness function error of the GEP for the train and the test phases are 0.0297 and 0.0425, respectively. The tree expression of the GEP algorithm is presented in Figure 10, including mathematical functions and operators.
The correspondence equation of the tree expression is:

\[
(d_{10}^{10} - d_1 c_0 d_1)^{-1} \left( \frac{c_0 d_1}{d_0} - d_0 \right) + d_0^2 + e \left[ \frac{1}{e^{d_2} + (c_1 - c_0 - d_2)d_0} \right] + \frac{d_0 + c_0}{(c_1 + d_1)(e^{-d_2} + (c_1 - c_0 - d_2)d_0)}
\]

(12)

Figure 8 | SVM solver’s performance for model 3 of TLW.

Table 9 | Summary of the SVM performance assessment for TLW

| Statistical indices | Train phase | Test phase |
|---------------------|-------------|------------|
|                     | Measured data | Predicted data | Measured data | Predicted data |
| Number of data      | 112         | 28         |
| Correlation coefficient | 0.9947     | 0.9942  |
| Mean                | 0.55        | 0.55       |
| Maximum             | 0.70        | 0.68       |
| Minimum             | 0.34        | 0.34       |
| STDEV               | 0.0795      | 0.0792     |

Figure 9 | Standardized normal distribution graph of DDR values for SVM models of TLW.

The correspondence equation of the tree expression is:
Table 10 | Sensitivity analysis of SVM for the TLW

| Top model | Omitted variable | Training phase | Test phase |
|-----------|-----------------|----------------|-----------|
|           |                 | $R^2$ | RMSE | $R^2$ | RMSE |
| Model 3   | None            | 0.9896 | 0.009 | 0.9886 | 0.0114 |
|           | Fr              | 0.8978 | 0.027 | 0.899 | 0.0309 |
|           | $A_w$           | 0.9721 | 0.0133 | 0.9717 | 0.0176 |
|           | $L_w$           | 0.953 | 0.0183 | 0.9602 | 0.0197 |
|           | $H_o/P$         | 0.9123 | 0.023 | 0.9132 | 0.0265 |

Table 11 | Parameter values used to predict $C_d$ in PKW

| Parameters                  | Values |
|-----------------------------|--------|
| Head size                   | 10     |
| Number of chromosomes       | 30     |
| Number of genes             | 3      |
| Mutation rate               | 0.044  |
| Inversion rate              | 0.1    |
| One-point recombination rate| 0.3    |
| Two-point recombination rate| 0.3    |
| Gene recombination rate     | 0.1    |
| Gene transposition rate     | 0.1    |
| IS transposition rate       | 0.1    |
| RIS transposition rate      | 0.1    |
| Fitness function error type | RMSE   |
| Linking function            | $+$    |

where $d_o$, $d_1$ and $d_2$ are $w/P$, $N$ and $H_o/P$, respectively. The constant values of $(c_o, c_1)$ for gene 1–gene 3 are $(5.580322, -5.624054)$, $(8.224731,1.193298)$ and $(-0.984681, 6.815644)$. A comparison among statistical indices has been presented in Table 12. Given the analysis results, there is good fitness between observed and predicted values of discharge coefficient. The test phase is more accurate than the train.

Variation of observed and predicted values of $C_d$ through the train and test phases has been displayed in Figure 11. It can be deduced that the GEP model performance is acceptable because of the adaptation of datasets during the train and the test phases. Figure 12 illustrates the standardized normal distribution of the DDR for the PKW using the GEP algorithm. The maximum values of $C_d(DDR)$ for the train and the test phases are 7.531 and 9.305, respectively. Although both bell diagrams have almost the same focus around the vertical axis, the maximum value of the test stage proves the better performance than the test. The sensitivity analysis (Table 13) for the GEP illustrates the highest impact of $H_o/P$ because of the crucial changes in accuracy indices so that in the test phase RMSE has increased from 0.0425 to 0.2411 and $R^2$ has dropped from 0.9767 to 0.1403.

The discharge capacity of RLW was the second one modeled using the GEP. Chromosome’s parameters have been presented in Table 14. The values of RMSE for the train and test phases were 0.01179 and 0.026, respectively. A tree expression of the GEP predictor has been illustrated in Figure 13.

The correspondence formula is described by Equation (10):

$$2d_1 + c_1 - \left( \frac{1}{d_0 + d_1 * c_0} \right)^{12} + \frac{((c_1 + d_0) - c_0^{0.5}) + (d_0 - c_1)}{(d_0 + d_2^2) + (d_2 + c_0) + (d_1 + d_2)} \tag{13}$$
Figure 10 | Expression tree for the PKW.

Table 12 | Summary of the GEP performance assessment for the PKW

| Statistical indices   | Train phase | Test phase |
|-----------------------|-------------|------------|
|                       | Measured data | Predicted data | Measured data | Predicted data |
| Number of data        | 54          | 17         |
| Correlation coefficient | 0.9859        | 0.9883     |
| Mean                  | 0.64        | 0.64       | 0.53        | 0.56         |
| Maximum               | 0.93        | 0.90       | 0.71        | 0.75         |
| Minimum               | 0.36        | 0.29       | 0.35        | 0.30         |
| STDEV                 | 0.1756      | 0.1676     | 0.1267      | 0.1494       |
where $d_0$, $d_1$ and $d_2$ are $w/P$, $N$ and $H_0/P$, respectively. The constant values $(c_0, c_1)$ for gene 1–gene 3 are $(2.832062, 3.009155)$, $(-6.081268, -2.36441)$ and $(-6.081268, -3.693909)$. For more detailed analysis, some statistical criteria of the GEP performance assessment have been presented in Table 15. A comparison between observed and predicted values of $C_d$ has been presented in Figure 14. There is good agreement between measured and predicted values of the discharge coefficient. Improving evaluation criteria during the test phase proves the correct modeling process of the GEP simulator. The
The normal distribution of the DDR for the RLW has been presented in Figure 15. The highest values of $C_d(DDR)$ for the train and the test stages are 12.23 and 13.23, respectively. These figures determine the better operation through the test phase. According to Table 16, the most effective parameter for the GEP modeling is $H_o/P$. Figure 16 shows the tree expression of TLW.

The simulation of the discharge coefficient for TLW has been performed based on the chromosome characteristics mentioned in Table 17. The values of RMSE for the train and the test phases were 0.00849 and 0.0823, respectively. Figure 10 presents the tree expression of the GEP modeling for TLW.

The correspondence formula is shown by Equation (11):

$$
C_d = \left[\frac{c_1^2}{d_2 + c_0} + \left[\frac{d_0}{d_3} - (d_5 + d_0)\right] - d_1\right]^{-1} + \left[\frac{2d_0}{d_4(d_5 + c_1)}\right] * \frac{Lnd_2}{c_1} + \left[\frac{d_5}{d_4 + d_5 + c_0 + d_5}\right]^{-1} (14)
$$
where \( d_0, d_1, d_2, d_3, d_4 \) are \( F_r, A/W, L_c/w \) and \( H_0/P \), respectively. The constant values of \((c_0, c_1)\) for gene 1–gene 4 are \((-8.993226, -0.424591), (6.883759, 0.423791), (6.913514, 6.413147), (-5.898438, 6.419434).\) A comparison between measured and predicted values has been shown in Table 18. In this table, statistical indices show good agreement between

| Statistic                  | Train phase | Test phase |
|----------------------------|-------------|------------|
|                            | Measured    | Predicted  | Measured    | Predicted  |
| Number of data             | 24          | 9          |
| Correlation coefficient    | 0.9935      | 0.9923     |
| Mean                       | 0.41        | 0.42       |
| Maximum                    | 0.59        | 0.64       |
| Minimum                    | 0.26        | 0.26       |
| STDEV                      | 0.1065      | 0.1396     |

Figure 14 | GEP solver’s performance for model 1 of RLW.

Figure 15 | Standardized normal distribution graph of DDR values for GEP models of RLW.
measured and predicted values, especially the test phase. To evaluate the model performance, the distribution of error values based on the standardized $C_{d(DDR)}$ has been shown in Figure 17. The maximum values of $C_{d(DDR)}$ for the train and test phases are 25.75 and 39.5, respectively, indicating the superior performance of the test stage for the GEP. Given the sensitivity analysis results in Table 19, the hydraulic condition, $Fr$, has the most effect on the GEP performance. The other important parameters in descending order are $H_o/P$, $A/w$ and $L_c/w$.

**Table 16 | Sensitivity analysis of the GEP for the RLW**

| Top model | Omitted variable | Training phase | Test phase |
|-----------|-----------------|----------------|------------|
| Model 1   | None            | $R^2$ | RMSE | $R^2$ | RMSE |
|           | $w$             | 0.9870 | 0.01179 | 0.9846 | 0.0260 |
|           | $P$             | 0.9389 | 0.03965 | 0.9581 | 0.02987 |
|           | N               | 0.9630 | 0.03805 | 0.9706 | 0.0131 |
|           | $H_o$           | 0.1362 | 0.2597  | 0.1321 | 0.2623 |

**Figure 16 | Expression tree for the TLW.**
Table 17 | Parameter values used to predict $C_d$ in TLW

| Parameters                        | Values |
|----------------------------------|--------|
| Head size                        | 9      |
| Chromosomes numbers              | 33     |
| Number of genes                  | 4      |
| Mutation rate                    | 0.04   |
| Inversion rate                   | 0.1    |
| One-point recombination rate     | 0.3    |
| Two-point recombination rate     | 0.3    |
| Gene recombination rate          | 0.1    |
| Gene transposition rate          | 0.1    |
| IS transposition rate            | 0.1    |
| RIS transposition rate           | 0.1    |
| Fitness function error type      | RMSE   |
| Linking function                 | +      |

Table 18 | Summary of the GEP performance assessment for the TLW

| Statistical indices | Train phase | Test phase |
|---------------------|-------------|------------|
|                     | Measured data | Predicted data | Measured data | Predicted data |
| Number of data      | 74          | 21         |
| Correlation coefficient | 0.9956     | 0.9983     |
| Mean                | 0.54        | 0.54       | 0.55        | 0.63          |
| Maximum             | 0.70        | 0.69       | 0.62        | 0.72          |
| Minimum             | 0.34        | 0.34       | 0.40        | 0.46          |
| STDEV               | 0.0901      | 0.0888     | 0.0767      | 0.0906        |

Figure 17 | Standardized normal distribution graph of DDR values for GEP models of TLW.
4. CONCLUSION

This research work investigates the capability of the SVM and the GEP algorithms in predicting the discharge coefficient of PKW, RLW, and TLW using gathered experimental data sets. Furthermore, a regression equation was extracted for each weir to simulate and predict the $C_d$ based on effective dimensionless parameters. Finally, a sensitivity test was performed to determine the order of the effective parameters on the discharge coefficient. Two algorithms have the potential to simulate discharge coefficients with acceptable accuracy. The results can be summed up as follows:

- Two SVM and GEP algorithms perform well in predicting the discharge coefficient of PKW based on statistical performance evaluation indices for PKW. But a comparison between maximum values of $C_d(DDR)$ indicating priority of GEP over SVM.
- Obtained analysis results prove the capability of SVM and GEP in simulating the discharge coefficient of RLW. The superiority of the GEP over the SVM is due to a large number of maximum values of $C_d(DDR)$.
- The results of statistical indices and the maximum values of $C_d(DDR)$ indicate the remarkable superiority of the GEP over the SVM for TLW.
- The sensitivity analysis suggests that the SVM and the GEP are capable of predicting the discharge coefficient of nonlinear weir well using hydraulic and geometric parameters. The most influential parameter for PKW and RLW was $H_0/P$. The most significant parameter for the TLW was $Fr$.

DATA AVAILABILITY STATEMENT

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

REFERENCES

ASCE 2000 Task committee on application of artificial neural networks in hydrology. I: preliminary concepts. *Journal of Hydrologic Engineering ASCE* **5**, 115–123.

Azamathulla, H. Md. 2012 Gene expression programming for prediction of scour depth downstream of sills. *Journal of Hydrology* **460–461**, 156–159.

Azamathulla, H. M., Haghiabi, A. H. & Parsaie, A. 2016 Prediction of side weir discharge coefficient by support vector machine technique. *Water Science and Technology: Water Supply* **16**(4), 1002–1016.

Azimi, H., Bonakdari, H. & Ebtehaj, I. 2019 Design of radial basis function-based support vector regression in predicting the discharge coefficient of a side weir in a trapezoidal channel. *Applied Water Science* **9**(4), 78.

Cortes, C. & Vapink, V. 1995 Support-vector networks. *Machine Learning* **20**, 273–297.

Deng, Z. Q., Bengtsson, L., Singh, V. P. & Adrian, D. D. 2002 Longitudinal dispersion coefficient in single-channel streams. *Journal of Hydraulic Engineering* **128**, 901–916.

Ebtehaj, I., Bonakdari, H., Khoshbin, F. & Azimi, H. 2015a Pareto genetic design of group method of data handling type neural network for prediction discharge coefficient in rectangular side orifices. *Flow Measurement and Instrumentation* **41**, 67–74.

Ebtehaj, I., Bonakdari, H., Zaji, A. H., Azimi, H. & Khoshbin, F. 2015b GMDH-type neural network approach for modeling the discharge coefficient of rectangular sharp-crested side weirs. *Engineering Science and Technology, an International Journal* **18**, 746–757.

Ebtehaj, I., Bonakdari, H., Zaji, A. H., Azimi, H. & Sharifi, A. 2015c Gene expression programming to predict the discharge coefficient in rectangular side weirs. *Applied Soft Computing* **35**, 618–628.
Emiroglu, M. E. & Kisi, O. 2013 Prediction of discharge coincident for trapezoidal labyrinth side weir using a neuro-fuzzy approach. *Water Resources Management* **27**, 1473–1488.

Ferreira, C. 2001 Gene expression programming: a new adaptive algorithm for solving problems. *Complex Systems* **13** (2), 87–129.

Ferreira, C. 2006 Gene Expression Programming: Mathematical Modeling by an Artificial Intelligence. 2nd ed. Springer, Berlin–Heidelberg, Germany.

Juma, I. A., Hussein, H. & AL-Sarraj, M. 2014 Analysis of hydraulic characteristics for hollow semi-circular weirs using artificial neural networks. *Flow Measurement and Instrumentation* **38**, 49–53.

Kashefpour, M. S. & Falconer, R. A. 2002 Longitudinal dispersion coefficients in natural channels. *Water Research* **36** (6), 1596–1608.

Kisi, O. 2013 ANFIS to estimate discharge capacity of rectangular side weir. *Water Management* **166** (WM9), 479–487.

Kumar, S., Ahmad, Z. & Mansoor, T. 2011 A new approach to improve the discharging capacity of sharp-crested triangular plan form weirs. *Journal of Flow Measurement and Instrumentation, Elsevier* **22**, 175–180.

Kumar, B., Kadia, S. & Ahmad, Z. 2019 Evaluation of discharge equations of the piano key weirs. *Flow Measurement and Instrumentation*. https://doi.org/10.1016/j.flowmeasinst.2019.10157

Majedi Asl, M. & Fuladipanah, M. 2019 Application of the evolutionary methods in determining the discharge coefficient of triangular labyrinth weirs. *JWSS* **22** (4), 279–290.

Mohammed, A. Y. & Sharifi, A. 2020 Gene Expression Programming (GEP) to predict coefficient of discharge for oblique side weir. *Applied Water Science* **10** (145). https://doi.org/10.1007/s13201-020-01211-5

Nadiri, A. A., Sedghi, Z., Khatabi, R. & Sadeghfam, S. 2018 Mapping specific vulnerability of multiple confined and unconfined aquifers by using artificial intelligence to learn from multiple DRASTIC frameworks. *Journal of Environmental Management* **227**, 415–428.

Noori, R., Karbassi, A. R. & Sabahi, M. S. 2010 Evaluation of PCA and Gamma test techniques on ANN operation for weekly solid waste predicting. *Journal of Environmental Management* **91**, 767–771.

Norouzi, R., Daneshfaraz, R. & Ghaderi, A. 2019 Investigation of discharge coefficient of trapezoidal labyrinth weirs using artificial neural networks and support vector machines. *Applied Water Science* **9** (148). https://doi.org/10.1007/s13201-019-1026-5

Parsaei, A. & Haghiabi, A. H. 2017 Improving modeling discharge coefficient of triangular labyrinth lateral weir using SVM, GMDH and MARS techniques. *Irrigation and Drainage* **66** (4), 656–654.

Rostami, H., Heidarragad, M., Purmohammadi, M. H., Kamanbedast, A. A. & Bordbar, A. 2018 Laboratory study of discharge coefficients of one and two-cycle piano key weirs and comparison with rectangular labyrinth weir. *Irrigation and Drainage Structures Engineering Research* **19** (71), 51–66.

Roushangar, K., Alami, M. T., Majedi Asl, M. & Shiriz, J. 2017 Determining discharge coefficient of the labyrinth and arced labyrinth weirs using support vector machine. *Hydrology Research* **49** (3), 924–938.

Roushangar, K., Majedi-Asl, M. & Shahnazi, S. 2021 Hydraulic performance of PK weirs based on experimental study and kernel-based modeling. *Water Resources Management*. https://doi.org/10.1007/s11269-021-02905-4.

Sadeghfam, S., Daneshfaraz, R., Khatabi, R. & Minaei, O. 2019 Experimental studies on scour of supercritical flow jets in upstream of screens and modeling scouring dimensions using artificial intelligence to combine multiple models (AIMM). *Journal of Hydroinformatics* **21** (5), 893–907.

Samadi, M., Jabbari, E. & Azamathulla, H. M. 2014 Assessment of MS’ model tree and classification and regression trees for prediction of scour depth below free overfall spillways. *Neural Computing and Applications* **24** (2), 357–366.

Samadi, M., Jabbari, E., Azamathulla, H. M. & Mojjallal, M. 2015 Estimation of scour depth below free overfall spillways using multivariate adaptive regression splines and artificial neural networks. *Engineering Applications of Computational Fluid Mechanics* **9** (1), 291–300.

Samadi, M., Sarkardeh, H. & Jabbari, E. 2020 Explicit data-driven models for prediction of pressure fluctuations occur during turbulent flows on sloping channels. *Stochastic Environmental Research and Risk Assessment* **34** (5), 691–707.

Samadi, M., Sarkardeh, H. & Jabbari, E. 2021 Prediction of the dynamic pressure distribution in hydraulic structures using soft computing methods. *Soft Computing* **25** (5), 3873–3888.

Seamons, T. R. 2014 *Labyrinth Weir: A Look Into Geometric Variation and its Effect on Efficiency and Design Method Predictions*. M.S. Thesis, Utah State University, Logan, UT, USA.

Seo, I. W. & Cheong, T. S. 1998 Predicting longitudinal dispersion coefficient in natural streams. *Journal of Hydraulics Engineering* **124**, 25–32.

Tayfur, G. & Singh, V. P. 2005 Predicting longitudinal dispersion coefficient in natural streams by artificial neural network. *Journal of Hydraulic Engineering* **131**, 991–1000.

White, W. R., Milli, H. & Crabbe, A. D. 1973 Sediment transport: An appraisal method, Vol. 2: Performance of theoretical methods when applied to flume and field data. Hydr. Res. Station Rep., No. 1T119, Wallingford, UK.

Zaji, A. H., Bonakdari, H. & Shamshirband, S. 2016 Support vector regression for modified oblique side weirs discharge coefficient prediction. *Flow Measurement and Instrumentation* **51**, 1–7.

Zaji, A. H. & Bonakdari, H. 2017 Optimum support vector regression for discharge coefficient of modified side weirs prediction. *INAE Lett* **2** (1), 25–53.

Zhou, Q., Zhou, H., Zhou, Q., Yang, F., Luo, L. & Li, T. 2015 Structural damage detection based on posterior probability support vector machine and Dempster-Shafer evidence theory. *Applied Soft Computing* **36**, 368–374.

First received 19 July 2021; accepted in revised form 30 August 2021. Available online 13 September 2021