Journal of Engineering and Sustainable Development
Vol. 29, No. 06, November 2025
https://jeasd.uomustansiriyah.edu.iq/index.php/jeasd
ISSN 2520-0917
Research Article
https://doi.org/10.31272/jeasd.3029

Using Binary Tree Algorithms to Predict a Discharge Coefficient of the Type-A Piano Key Weir

Nguyen Minh Ngoc[1*], Bui Hai Phong[2], Le Van Nghi[3], Tran Ngoc Thang[4], Nguyen Thanh Phong[5]

[1.5] Faculty of Urban Environmental and Infrastructural Engineering, Hanoi Architectural University, Hanoi City, Vietnam
[2]Faculty of Information Technology, Hanoi Architectural University, Hanoi City, Vietnam
[3]Key Laboratory of River and Coastal Engineering, Vietnam Academy for Water Resources, Hanoi City, Vietnam
[4]Faculty of Civil Engineering, Hanoi University of Business and Technology, Hanoi City, Vietnam

[*]Email: ngocnm@hau.edu.vn

Article Info

Received	**26/09/2024**
Revised	**25/08/2025**
Accepted	**29/08/2025**

Abstract

Type-A PKW is a type of spillway with inlet and outlet keys connected in a zigzag. The survey shows that previous studies on the flow over PKW mainly focused on establishing empirical equations to calculate C_d. There is no study that meets all research needs. This shows that studies have not taken advantage of data from other studies. Meanwhile, regression Machine Learning (ML) algorithms can solve difficult problems in empirical studies. That is the ability to make maximum use of data sources, and it does not need constraints in predicting C_d values. Therefore, this study was carried out to develop a process to apply ML algorithms to the regression prediction of hydraulic characteristics (including 3 phases). In particular, the application of "Binary tree" algorithms (Decision Tree and Random Forest algorithms) is used to illustrate the study. In addition to determining the advantages of ML algorithms, a comparison was made between applying ML algorithms and using empirical equations to determine C_d values. The study shows that the Random Forest algorithm has better predicting efficiency with statistical indicators very close to the ideal point (R^2 = 0.98, RMSE = 0.046, MAPE = 3.1% and MEA = 0.028).

Keywords: Decision Tree, Discharge coefficient, Machine Learning, Piano Key Weir, Random Forest.

1. Introduction

Piano Key Weir (PKW) is a spillway type with the crest of the dam in a zigzag pattern like the keys of a Piano. It includes outlet keys and water inlet keys. It has been widely applied in Vietnam and internationally, with a discharge capacity that is up to 4-5 times greater than that of conventional spillways [1]. When ground conditions do not allow for extending the length of the dam to increase the flood discharge capacity, the PKW is an effective and economical solution. This type is one of the increasingly popular and applied options [2].

There are many different types of PKW (including four types: A, B, C, and D) [3],[4]. Among them, the type-A PKW has been commonly used in Vietnam and many countries around the world [6][5]-[7]. In Vietnam, several PKWs are operating, such as Van Phong weir (Fig.1), Dak Mi 3 Hydropower Plant (2017), Da Dang weir (2022), etc. Currently, there are several dams under construction, such as the Phu Phong Weir Project (Binh Dinh province), a dam downstream of the Tra Khuc River (Quang Ngai province), and the Bang Lang dam (Lam Dong province) [5]. Around the world, PKWs have been built in India, Australia, France, Switzerland, and so on [6],[7].

Figure 1. Van Phong weir (2015) – Binh Dinh province (by author).

Therefore, studying the type-A PKW has many practical meanings and enriches the ability to choose and apply research on different types of PKW.

The type-A PKW is described by several characteristics (Fig. 2), such as upstream head (H), discharge (Q or q), dam height (P), outlet key width (Wo), inlet key width (W), weir crest length (L), length of a weir unit (Lu), total length of weir (B), down and upstream overhang lengths (Bi and Bo), wall thickness (Ts), etc. [1],[5],[7]. The characteristics of the

spillway are interrelated and affect the efficiency of the flow over the dam.

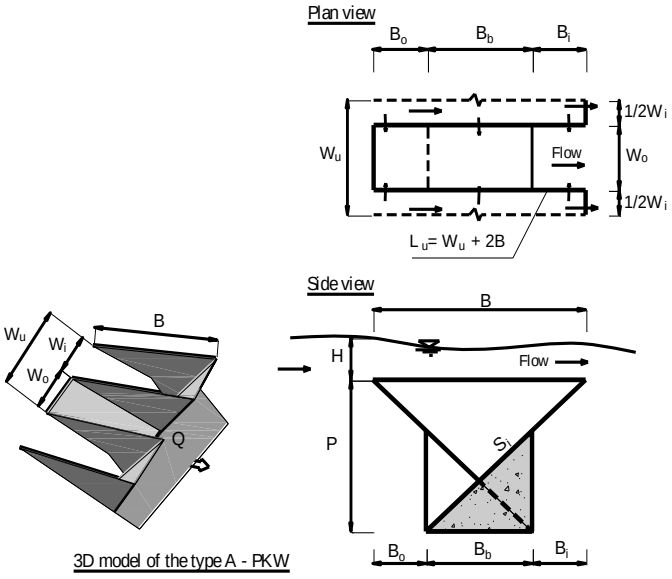

Figure 2. Structure of the type-A PKW.

In fact, studies on flow over PKW focus on establishing empirical equations, where determining the discharge coefficient (C_d) plays a vital role in the calculation [5]. The study by Kabiri-Samani and Javaheri [8] examined geometric characteristics such as L/W, B/P, and Bo/B. The study of Crookston and Tullis [9] considers L/W, with Laugier [10] considering the factors Wo/Wi and Pi/Po. Li et al. [11] only consider the nonlinear rule of H/P. According to Al-Baghdadi and Khassaf [12], they consider H/P and L/W. Al-Shukur and Al-Khafaji [13] base their analysis on H/P, L/W, and Wi/Wo. Guo et al. [14] consider H/P, L/W, Wi/Wo, and B/P; Kumar et al. only consider the linear function of H/P and L/W. Musa and Alghazali [6] only considered the nonlinear function of H/P. Research by Ngoc et al.[5], Yen and Nghi [15], and Nghi and Yen [16] considered the factors H/P, H/Lu, and H/Wu, and classified research according to H/Wo values and other relevant criteria. Each study on the PKW has its own characteristics and diverse influencing factors. However, many databases are not optimally utilized. Studies still have many different influencing factors, leading to inconsistent results.

The assessment of the proposed equations for determining C_d shows that different studies describe different factors affecting C_d. Not only that, applying formulas from one study to another has resulted in various errors, which are sometimes very large [5]. Therefore, integrating different data sources is an urgent requirement. Establishing a calculation method where the conditions for applying the equations are not constrained will play an important role in hydraulic engineering research.

From the analysis of theoretical studies, experimental studies, mathematical model studies (Flow 3D model) [8], [16], and surveys on ML algorithm applications. It shows that the application of Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning (DL) algorithms on PKW is explored in many studies [17]-[19]. As researched by. Khattab et al. [20] applied the Artificial Neural Network (ANN) algorithm in the research and analysis of the flow over the type-C PKW and showed high application efficiency. Bansal et al. [2] used the Gene Expression Programming (GEP) and Extreme Gradient Boosting (XGBoost) algorithms to analyze the energy of the flow over the type C-PKW; the study showed good predictive performance. Dursun et al. [21] estimated the discharge coefficient of a semi-elliptic labyrinth side weir using the Adaptive-Neuro Fuzzy Inference System (ANFIS). The study showed that the performance is quite good and can be used in modeling the discharge coefficient. Zaji et al. [22] applied the combined model of Support Vector Machine regression and the Firefly Optimization algorithm (SVR-FA) to predict the discharge coefficient of the side weir. The results indicate that the SVR-FA model is about 10% more accurate than the SVR model according to the RMSE criterion. Haghbi et al. [23] predicted the discharge coefficient of triangular maze dams using ANN and ANFIS models. Comparing the results of the ANN with the ANFIS model shows that both models have good analytical performance. Gul et al. [24] used the SVM algorithm in selecting parameters for the spillway. Zounemat-Kermani and Mahdavi-Meymand [25] used ANFIS to analyze the discharge coefficient over the spillway, achieving a calculation reliability of 95%. Singh et al. [26] used the GEP algorithm to predict the energy consumption of the type-A PKW, demonstrating good efficiency and significant predictive power for the energy flow over PKW. The above preliminary survey shows that research on ML is still minimal. Especially in the field of construction hydraulics, there is very little research, which limits the ability to apply and expand the research database system.

Practical surveys have shown that the factors affecting the C_d coefficient are also determined in a complex manner; each study will have different groups of influencing factors. Moreover, studies on the application of ML algorithms also show that no algorithm is best for all cases. This study used Decision Tree (DT) and Random Forest (RF) algorithms in regression prediction to evaluate the effectiveness of applying ML algorithms. In addition, comparing this study with studies establishing traditional empirical equations will clarify the meaning of applying ML algorithms [27].

This study is implemented to establish a process for predicting hydraulic characteristics using general ML algorithms. This process is applied to the study of predicting the C_d coefficient of the flow over type-A PKW. It can demonstrate the good applicability of binary tree machine learning algorithms (DT and RF) in predicting C_d coefficient and compare with current traditional experimental studies to clarify the effectiveness of ML algorithms. From there, the experience of applying ML algorithms is used to analyze the flow over a type-A PKW on the physical experimental models. Furthermore, the application of ML helps to quickly predict hydraulic factors of the flow in hydraulic experiments on physical models. It also provides directions for adjusting the physical model right before building the experimental model.

2. Study Method

2.1. Structure of the Study

The study on the application of the "binary tree" algorithm in forecasting the C_d coefficient is approached in many directions, from modern methods (ML algorithms) to traditional methods (empirical equations). The data is collected from various sources and physical models with different dimensions to highlight the study's diversity and richness.

The application of ML in regression predicting studies has been proposed by Sharker [28] with two basic phases (Training and Testing). However, the prediction of hydraulic factors still needs to be clarified. The determination of the method to establish the objective equations and hydraulic characteristics, which need to be collected from physical experimental models, is crucial. So, this study has proposed 3 phases (Fig.3), each stage is clarified and directly oriented to the problem of predicting hydraulic characteristics. It is simulated as follows:

- *Phase 1:* Setting up the ML problem. This phase will use basic equations (such as the momentum equation, energy equation, etc.) to establish the objective functions. The establishment of the objective function is based on the dimensional analysis method and Buckingham's Pi theory. The objective equation helps determine the hydraulic factors that need to be collected in the study, thereby identifying the target variable, the influencing variables, and the quantities that need to be collected from the experimental model. Depending on each different research, the determination of the target equation according to Pi theory is also different.

- *Phase 2:* Establishing the training model. This phase builds on phase 1 to identify the target and data variables, from which the database for the ML model is determined. The next step is to select the ML algorithm to apply and use the training data to train the ML model [28]. At the end of "training", the basic models will be established, and the trained model with good predictive performance will be determined. A model with good predictive performance is one whose statistical indicators approach the "ideal point".

  Statistical indicators will be used as standards in evaluating ML models, including R-squared (R^2), Mean Squared Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). The "ideal point" is the value at which statistical indicators ensure the best agreement between predicted and measured values. At the "ideal point", the R^2 will approach 1, and other statistical indicators will approach "zero" [5], [29].

- *Phase 3:* Using the training model established in phase 1 and the test data (the dataset has a ratio of 20% to 30% compared to the training data [30]) to predict the research results and applying statistical indicators to evaluate the conformity between the predicted values and the measured values of the test dataset.

  At the end of this phase, a suitable ML algorithm for the research problem will be proposed. Then, the training model established in Phase 2 will be used to apply hydraulic characteristic prediction for other studies.

In these 3 phases, Phase 1 is necessary for establishing the objective function, which serves as the scientific basis for creating the data structures of the ML model. Phase 1 will identify the factors to be collected and the data fields to be analyzed.

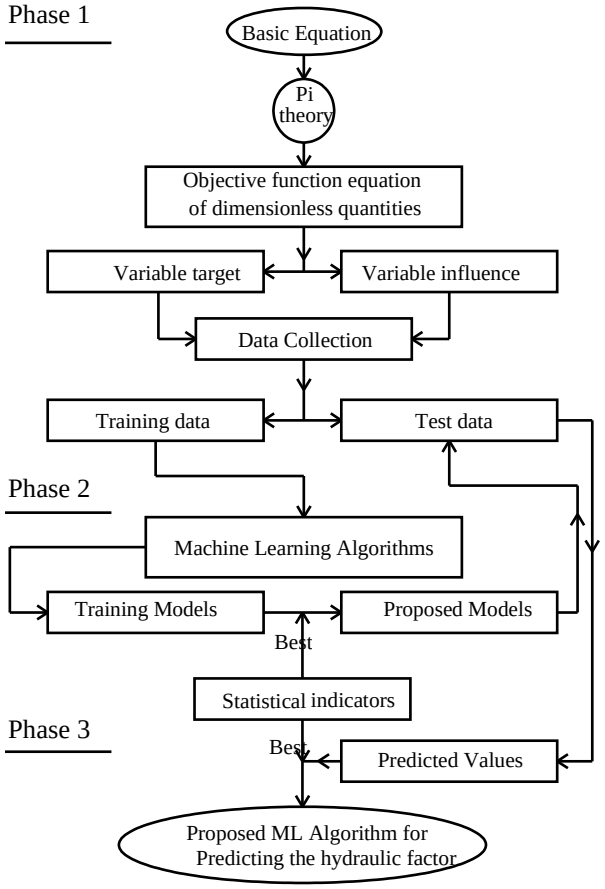

Figure 3. Structure of the process in determining the appropriate ML algorithm in predicting hydraulic factors.

Fig.3 is a general diagram used to simulate the analysis process and determine the best ML algorithm for predicting hydraulic factors. This is the implementation process in this study.

2.2. Establishing the Function Affecting the Study Aims

The structure for analyzing the ML model for the flow over the PKW is shown in Fig. 4. Here, function f is the objective function equation used for analysis in the ML model, and parameters (H/P, H/Wu, và H/Lu) are the input database. The value C_d is the prediction result of the objective function.

Information	Data
Data input	Type of PKW Discharge (m^3/s) Head H (m) Size of PKW (Wu, Lu, P) Conditions of flow state (free or submersion)
Dimensionless quantities	Dimensionless parameter of PKW (H/P, H/Wu, H/Lu)
Result	Predicting the discharge coefficient (C_d)

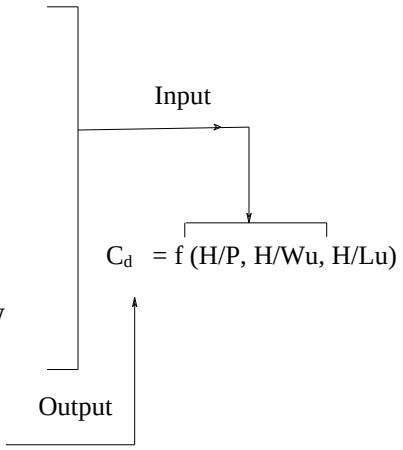

Figure 4. Analyzing the structure of establishing the objective function.

The study with the free flow over the PKW according to the structure of Fig.4, in which all input influencing variables are used and the basic equation for determining the discharge [3], [5], [7] is used as follows:

$$Q = \frac{2}{3} C_d W \sqrt{2gH^3} \quad (1)$$

The quantities considered in establishing the flow equation over the type-A PKW are as follows:

$$F(Q, W_u, W_i, W_o, B, B_i, B_o, L_u, P, \rho, \mu, g, H) = 0 \quad (2)$$

Applying Buckingham's Pi theory will yield the following equation to determine the quantities affecting the flow coefficient [5], [16]:

$$C_d = \Psi\left(\frac{H}{W_u}, \frac{H}{L_u}, \frac{H}{P}\right) \quad (3)$$

Equation (3) is used to assist in determining the target variable and influencing variables in the analysis of the ML algorithm. Furthermore, this is also the equation commonly applied in research methods to establish empirical equations.

The discharge coefficient (C_d) is related to the data fields, and it is related to the values of H, Lu, Wu, and P. This is similar to the studies on Yen and Ngh[15], Guo et al.[14], Rezaei[31], Bhukya et al. [32] and so on.

2.3. Establishing the Function Affecting the Study Aims

The application of the ML model for predicting the discharge coefficient is implemented according to the diagram in Fig.5, including the steps defined as follows

+ Step 1: Collecting data. This includes experimental models, field experiments, and data collection from physical experimental models. Raw data processing and setting up a system of parameters to input into the ML model.

The data collected from experiments on physical models will include: Q, H, P, Bi, Bo, Lu, Wu.

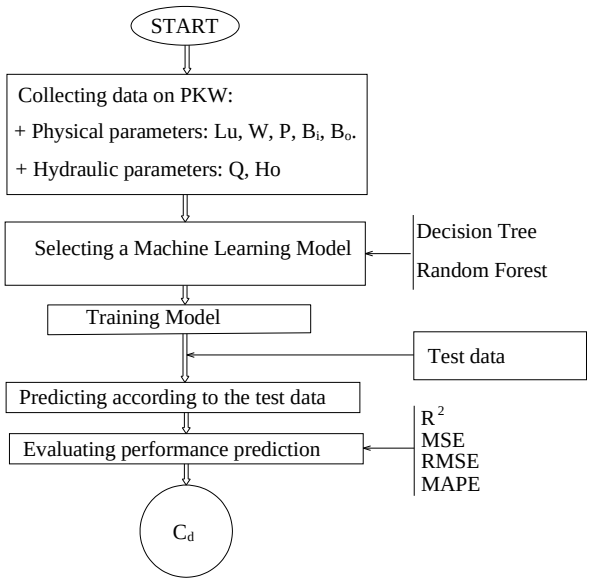

Figure 5. Structure of applying the ML model in predicting C_d

Data processing:

- Removing inappropriate data in physical experiments, this issue is addressed in different experiments. Inappropriate data will be eliminated depending on the experimental implementation process. The law of hydraulic factors change on each physical model will be determined through experiments, and the mutant data due to the influence of the experimental process will be eliminated (this study is based on experiments by different authors that have been widely published, so the data processing stage will not consider it).
- Determining data on the target variable (1): C_d.
- Setting up data on influencing variables, which consist of dimensionless parameters: H/Wu, H/Lu, and H/P.

All data is divided into two types: Target data (parameters to be predicted) and influence data (analyzed variables). For the flow over the PKW, the target data is the discharge coefficient (C_d). The influence variable data are dimensionless quantities, such as H/P, H/Wu, and H/Lu, which are divided into two datasets for two different purposes: Training data (accounting for 80% of the experimental dataset) and test data (accounting for 20% of the experimental dataset [30]).

+ Step 2: Selecting the ML model for analysis. In this study, the DT and RF algorithms are selected and applied to predict the discharge coefficient (C_d) of the flow over the PKW. In this step, the training data will be used to establish training models according to the ML algorithms (the DT and RF training models).

+ Step 3: Evaluating the predictive ability of the ML models and using the training models in Step 2 with the test dataset for research, and in this step, predicting the values of the discharge coefficient (C_d) according to the test data. Then, compare the prediction results with the measured values to evaluate the predictive performance of the models. The predictive performance is evaluated based on statistical indicators [27]. At this step, a suitable ML algorithm will be proposed to predict C_d values for other studies.

3. Machine Learning Algorithms

3.1. Decision Tree Algorithm

Decision Tree (DT) [33],[34] is one of the predictive modeling approaches used in statistics, data mining, and ML. Using DT (as a regression predictive model) to go from observations about a target value of that parameter (represented in the "branches") to conclusions (defined in the "leaves").

The decision of splitting strategy significantly affects the accuracy of the "tree" (Fig.6). The decision criteria are different for classification "tree" and regression "tree". In a regression "tree," the MSE will be used to decide how to split a "Father" node into "child" nodes.

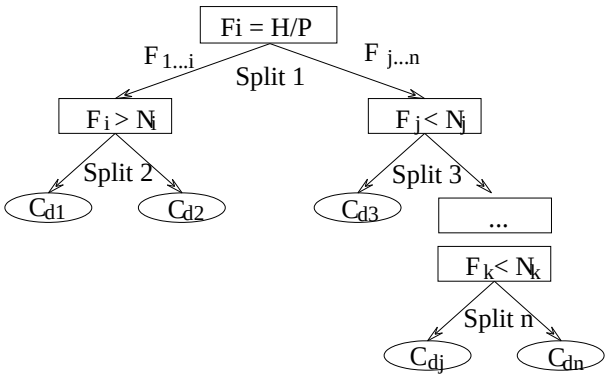

Figure 6. Structure of Binary DT Model

The DT uses a binary algorithm; each "node" is analyzed into a maximum of 2 "leaves". In which the division will separate the data series into different regions, and each region will apply the smallest MSE criterion to evaluate [35]:

$$min_R[\sum(y_i - \widehat{y_R})^2] \quad (4)$$

Where:

y_i is the predicted value in the R region;

y_R is the average value of the R region.

The code of the DT algorithm in MATLAB software is written as follows:

```
Mdl2 = fitrensemble(APKW, BPKW);
tree = fitctree(APKW, BPKW);
result_tree=predict(tree,Ctest);
```

Where:

APKW is the input data of the influencing variables.

BPKW is the input data of the target variable

CTest is the test data

Ketqua_tree is the data file of the predicted values using the DT algorithm.

3.2. Random Forest Algorithm.

Random Forest (RF) [34],[35] is a supervised learning method that can handle problems of classification and prediction of values (regression problems).

The algorithm can be explained mathematically as follows: RF is a collection of many DTs (Fig.7), where each DT is randomly created from resampling (randomly selecting a part of the data to build) and randomly selecting variables from all variables in the data. With such a mechanism, RF provides very high accuracy results, but it cannot grasp the operating mechanism of this algorithm because the model's structure is too complicated.

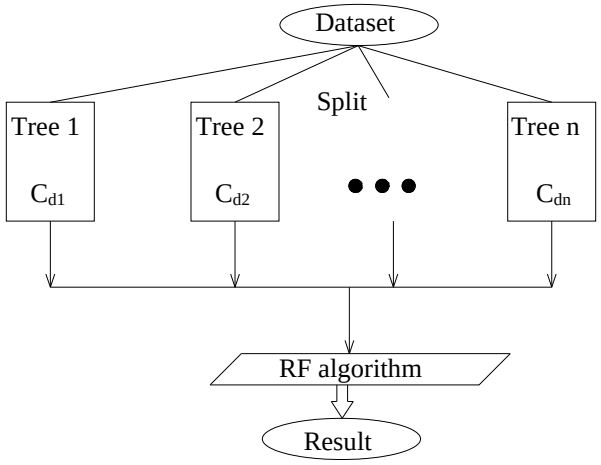

Figure 7. Structure of the Predicting Model using the RF algorithm

The RF algorithm can be considered one of the "Black Box" modeling methods, as it only produces results but cannot explain the model's operating mechanism. This algorithm can be described as follows [27]:

$$g(x) = \sum_{i=1}^{n} \alpha_i f_i(x) \quad (5)$$

where:

fi(x) is the base model of the i-th DT

α_i is the weight of the "Tree" model in the "forest."

$$\sum \alpha_i = 1 \quad (6)$$

This technique is widely used and has good predictive performance. In the RF model, the DT basis models are established independently using a separate data sample.

The code to execute the RF algorithm in MATLAB software is shown as follows:

```
Mdl2PKW = fitrensemble(APKW, BPKW);
YPKWTest = predict (Mdl2PKW, CPKWTest);
```

Where: YPKWTest is the file containing the predicted data according to the RF algorithm.

RF is an improved algorithm compared to the DT algorithm, because RF contains many DTs. In practice, some experiments show that when optimizing the DT algorithm, it can give forecasts with high accuracy. Therefore, it isn't easy to accurately evaluate the best RF or DT algorithm for prediction, as it needs to be verified for each case study.

4. Research Data

4.1. Experimental Model and Input Data

The study collected data from various sources, including research data widely published in domestic and international journals, scientific conferences, doctoral theses, and scientific research topics at all levels. Research data were collected from 6 different authors; the experiments were set up based on eight physical models in the laboratory (Fig.8 shows the type-A PKW and flume built in the laboratory of the Vietnam Academy for Water Resources). Each physical model has different dimensions and ratios between geometric elements. Therefore, the diversity of experimental model structures is vibrant. The geometrical characteristics of PKW (Fig.2) in the studies are shown in Table 1.

Figure 8. Experimenting on the Flow over PKW model at Vietnam Academy for Water Resources (VAWR)

Table 1. Parameters of the experimental model of the type-A PKW.

Authors	Information on physical models			
	L/W	P/W_u	W_i/W_o	B_o/B_i
Kabiri-Samani and Javaheri [8]	6.0 ÷ 8.1	0.63 ÷ 1.79	0.33 ÷ 1.67	0.26 ÷ 1.0
Machiels et al. [3]	5.0	0.33 ÷ 2.0	1.5	1.0
Yen and Nghi [15]	5.0	0.5 ÷ 1.1	1.30	1.0
Hai et al. [36]	4.3 ÷ 8.2	0.3 ÷ 2.4	1.2	1.0
Noui and Ouamane [37]	5.9	0.9	0.96 ÷ 1.53	1.0
Denys [38]	4.4	0.67	1.25	1.0

Based on (3), the experimental data to be collected include the discharge (Q), upstream head over PKW crest (H), and the geometric parameters of the experimental model (Wu, P, Lu, Wo). Each dataset is collected on physical models and is performed independently, ensuring that all research data are independent of each other. During the experiment, the problems of experimental errors and unstable data were addressed. This study used the original published data to ensure objectivity. To avoid the influence of surface tension, the upstream heads at the dam crest must be greater than or equal to 3cm; Ho values less than 3cm are eliminated[39].

The data were collected from experimental plans on physical models. Subsequently, data that did not meet the specific requirements in hydraulic experiments were eliminated, resulting in 360 data sets. The characteristics of the experimental data were established in the form of dimensionless quantities. The experimental data are shown in Table 2. From the data collected in Table 2, establishing the relationship between Cd and the influencing factors (e.g., the relationship between Cd and H/Wu in Fig.9), and evaluating the change in Cd according to hydraulic characteristics.

The study collected real measurement data from physical models. Table 1 shows the size and scale of the physical model, while Table 2 and Fig. 9 present measured data from the physical models. If using a relationship to determine C_d, the correlation coefficient is low (R^2 = 0.69). If using multiple correlations, establishing experimental results is very difficult, and ensuring good computational efficiency is challenging. Therefore, applying the ML Models (the DT and RF algorithms) is an optimal solution, offering many advantages in research.

Table 2. Experimental data of physical models (type-A PKW).

Ref.	q (m^3/s.m)	H (m)	H/P	H/W_o	H/W_u	H/Lu	C_d
Kabiri-Samani and Javaheri [8]	0.05 ÷ 0.20	0.03 ÷ 0.14	0.13 ÷ 0.56	0.16 ÷ 1.87	0.16 ÷ 1.00	0.03 ÷ 0.12	0.66 ÷ 1.98
Machiels et al. [3]	0.04 ÷ 0.41	0.03 ÷ 0.29	0.06 ÷ 2.68	0.26 ÷ 2.45	0.11 ÷ 0.98	0.02 ÷ 0.20	0.53 ÷ 1.88
Yen and Nghi [15][15]	0.03 ÷ 0.32	0.034 ÷ 0.23	0.17 ÷ 2.09	0.31 ÷ 2.08	0.14 ÷ 0.92	0.03 ÷ 0.18	0.58 ÷ 1.53
Hai et al. [36]	0.03 ÷ 0.31	0.032 ÷ 0.232	0.13 ÷ 2.15	0.23 ÷ 3.16	0.11 ÷ 1.36	0.03 ÷ 0.22	0.57 ÷ 1.44
Noui and Ouamane [37]	0.05 ÷ 0.17	0.04 ÷ 0.14	0.15 ÷ 0.95	0.26 ÷ 2.05	0.21 ÷ 0.81	0.04 ÷ 0.14	0.77 ÷ 1.53
Denys [38]	0.1 ÷ 0.75	0.04 ÷ 0.25	0.11 ÷ 0.62	0.18 ÷ 1.03	0.02 ÷ 0.09	0.07 ÷ 0.41	1.37 ÷ 2.54

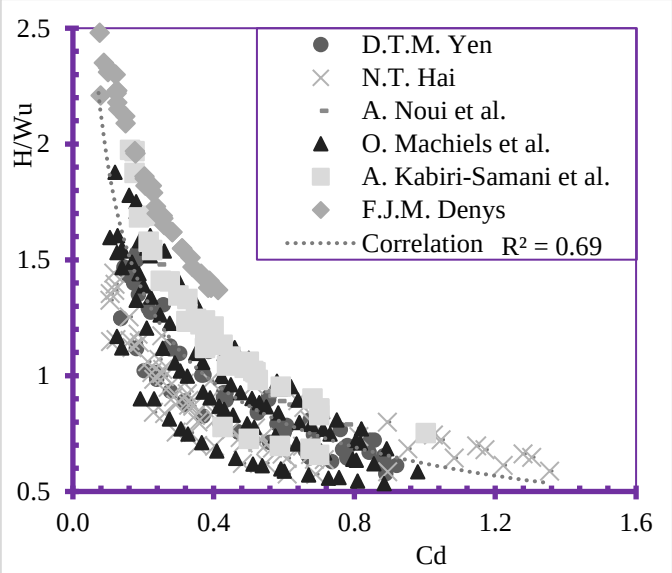

Figure 9. Experimental relationship between C_d and H/W_u

4.2. Establishing baseline data on the study

According to the collected data on the flow over the type-A PKW (Table 2). The data is divided into 2 datasets as follows:

- Training data includes 300 datasets (data characteristics in Table 3 and Fig.10).
- Test data includes 61 datasets (equivalent to 20% of training data, ensuring the stability and efficiency of the research method [30]. The data characteristics are shown in Table 4 and Fig. 11.

The training data and testing data are shown in Tables 3 and 4. The relationship between the target variable (Cd) and some influencing factors is shown in Figs. 10 and 11.

Table 3. Training data of the ML models

Values	q_{tn} (m^3/s.m)	H (m)	H/P	H/W_u	H/L_u	C_d
Min	0.029	0.029	0.06	0.073	0.017	0.534
Max	0.414	0.294	2.68	1.278	0.210	2.540

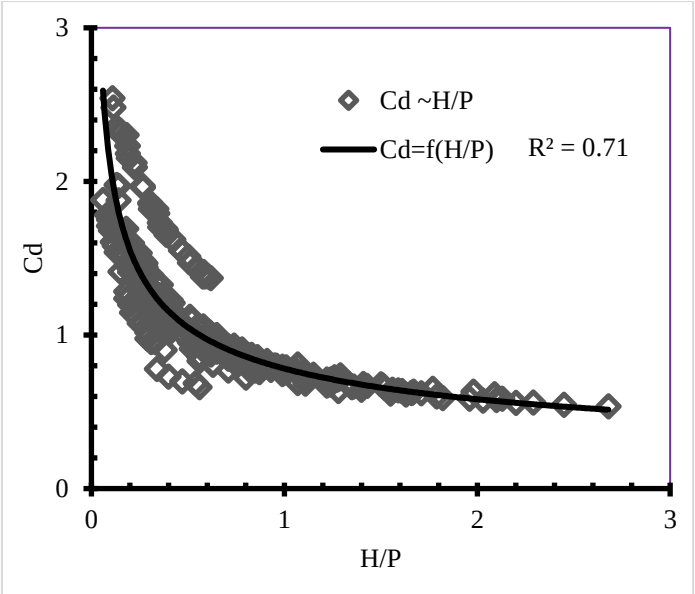

Figure 10. Relationship between C_d and H/P according to training data

Table 4. Test data of the ML model

Values	q_{tn} (m^3/s.m)	H (m)	H/P	H/W_u	H/L_u	C_d
Min	0.034	0.032	0.170	0.106	0.028	0.574
Max	0.367	0.239	2.150	1.355	0.220	2.220

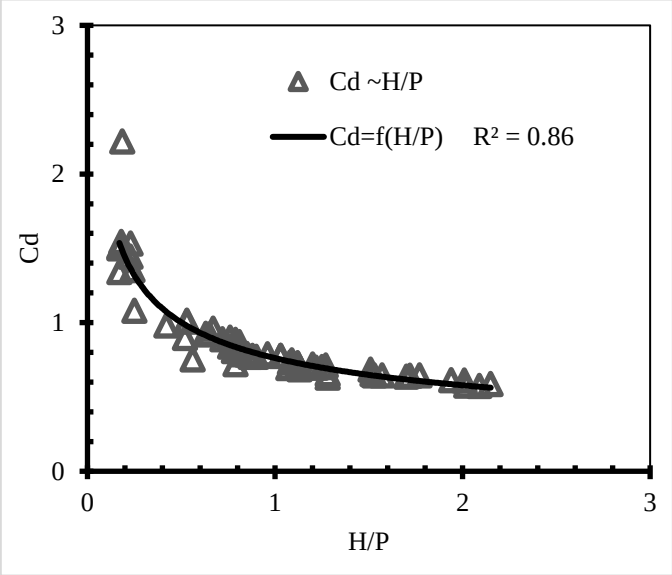

Figure 11. Relationship between C_d and H/P of the test data

Thus, the correlation between C_d and Ho/P in both test and training data is good (R^2> 0.7). However, this correlation coefficient is not very high, so establishing equations based on influencing factors will not guarantee accurate results due to strong dispersion around the average value (Fig.10). Therefore, applying ML algorithms to predict C_d will yield the best results. The training data (Table 3) and test data (Table 4) are used for all the ML models; the predicted results from the test data will be used to evaluate the suitability of the research according to each ML model.

5. Results and discussions

5.1. Setting Up Training and Testing Data for ML Models

Using codes of the DT and RF algorithms in the MATLAB R22b environment, setting up a common data system for both algorithms (Fig.12), including:

- Training data of the influencing variables (APKW): H/P, H/Lu, and H/Wu.
- Training data of the target variable (BPKW): C_d
- Test data of the influencing variables (CPKWTest): H/P, H/Lu, and H/Wu.

The data system (Fig.12) is established and used for standardizing the ML models. The ML algorithms (DT and RF) are used to develop training models. After confirming the training models of the DT and RF, the study used the test data to predict the C_d. Using these expected values to compare with the measured values, thereby evaluating the effectiveness of the predicting models.

Performing the training process by the DT and RF algorithms with the data according to Table 3. The training results are shown in Figs. 13 and 14.

The DT algorithm is built based on a “binary tree”, so the prediction error is quite large (Fig.13). This is due to the automatic division rule at the "nodes". Therefore, to reduce the prediction error, it is necessary to optimize this algorithm.

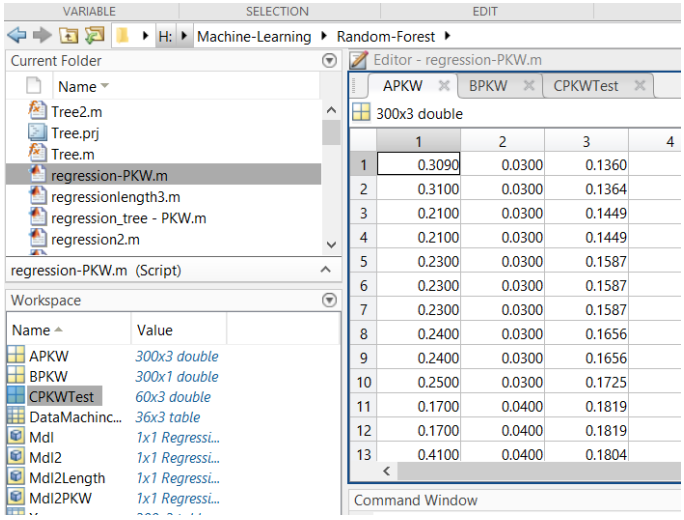

Figure 12. Setting up data for the ML model in MATLAB

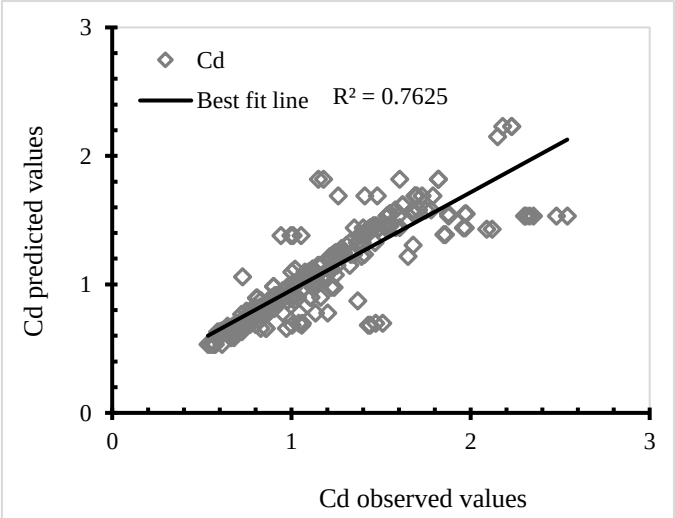

Figure 13. Training results by the DT algorithm

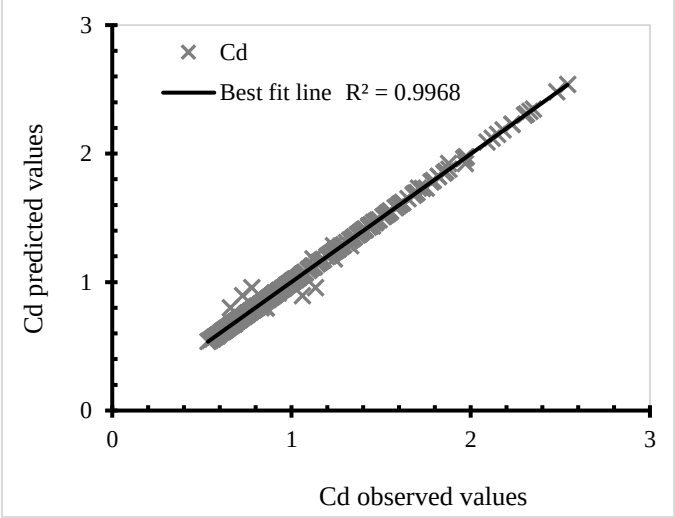

Figure 14. Training results by the RF algorithm

The RF algorithm is built based on establishing many "Binary Trees", so the process of analyzing and determining rules has many advantages. The training results show low prediction error and an R^2 value of 1. After examining the statistical indicators of the test values and the model's forecast results, the evaluation values are shown in Table 5.

Table 5. Statistical indicators of training results according to ML algorithms

Algorithms	MEA	MSE	RMSE	R^2	MAPE (%)
RF	0.006	0.001	0.023	0.997	0.630
DT	0.109	0.047	0.216	0.732	8.319

From Table 5, the statistical indicators for evaluating the training results of the DT and RF algorithms are also very close to the ideal point. However, the RF algorithm is more outstanding in terms of forecasting efficiency. Besides, both algorithms have their own advantages, so the study will use them to analyze the test data and evaluate the algorithms' efficiency in greater depth.

5.2. Study On the Predicting Efficiency of the DT Model

The DT model in the ML, after training, was analyzed with the test data in Table 4, from which the C_d values for the test data were predicted (Fig. 15). The characteristics of the predicted values according to the DT model are shown in Table 6.

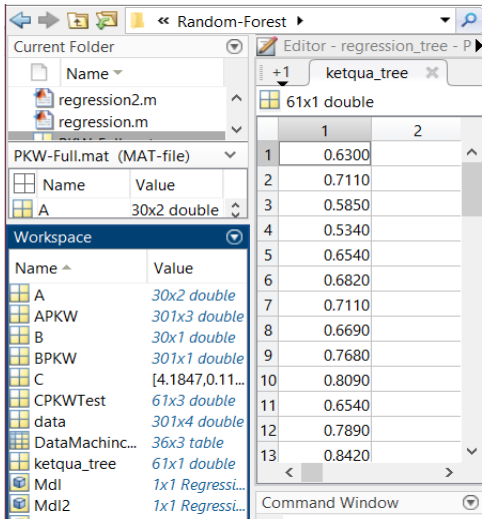

Figure 15. Predicted values of test data according to the DT model

Table 6. Predicted results of C_d according to the DT model

Values	Observed values				Prediction	
	H/P	H/W_u	H/L_u	C_d	C_d	Error (%)
Min	0.170	0.106	0.028	0.574	0.53	0.13
Max	2.150	1.355	0.220	2.220	2.23	14.52

The "Binary Tree" structure of the DT Model in the study on C_d values are depicted in Fig.16. Comparing the predicted values by the DT model with the observed values of C_d from experimenting the physical models, the correlation is shown in Fig.17. As it is displayed in Fig.17, it can be seen that most of the computed data are close to the best fit line and indicate ±10% difference with the observed data. The proportion of values with an error of less than 10% is 83.3% and the error from 10% to 14.5% is 16.7%.

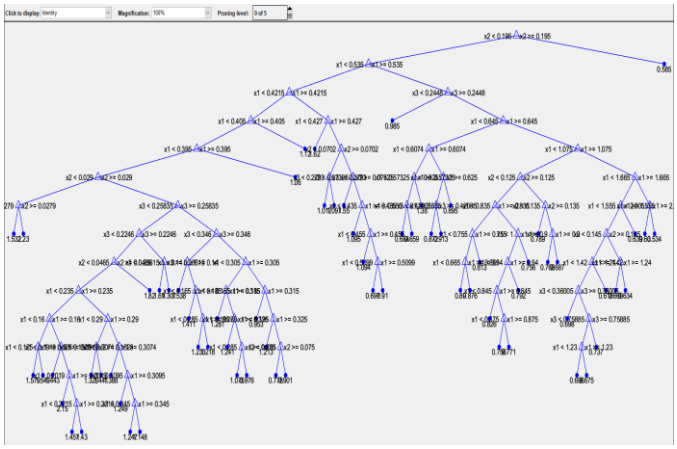

Figure 16. "Tree" diagram of the DT model

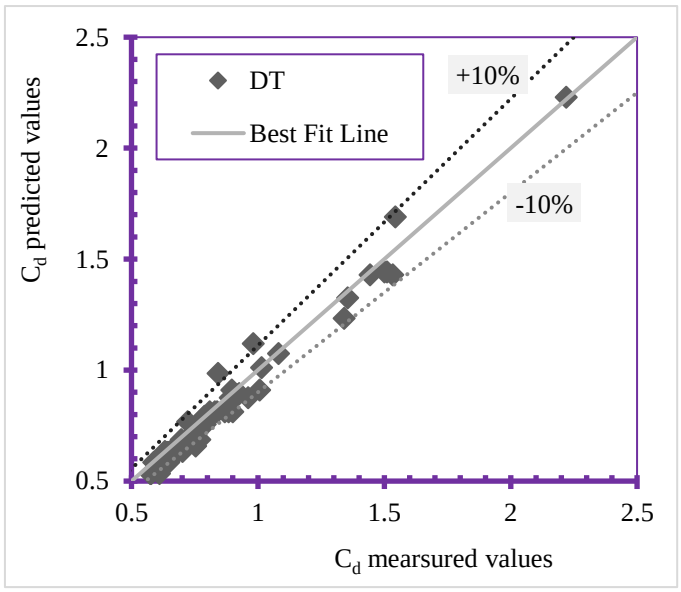

Figure 17. Comparison between calculated data and predicted data of C_d

5.3. Study on the predicting efficiency of the RF model

Predicted C_d values according to the test data in Table 4 are shown in Fig.18; these values are used to compare with the observed values according to the evaluation of statistical indicators. The predicted results for the test data (Table 4) are shown in Table 7.

The correlation between the measured and predicted values of the C_d according to the RF model is shown in Fig.19. As shown

in Fig.19, the most of computed data are close to the agreement line and indicate ±10% difference with the observed data (57/60 data, equivalent to 95% of data), this shows that the predicted and experimental data are in good agreement. The most significant prediction error is 14.1%, but this is only 1 of 3 cases where the error is greater than 10% (accounting for 5% of the predicted data). The comparison shows that the calculation performance of the RF model is better than that of the DT model, as clearly illustrated by the models' algorithms.

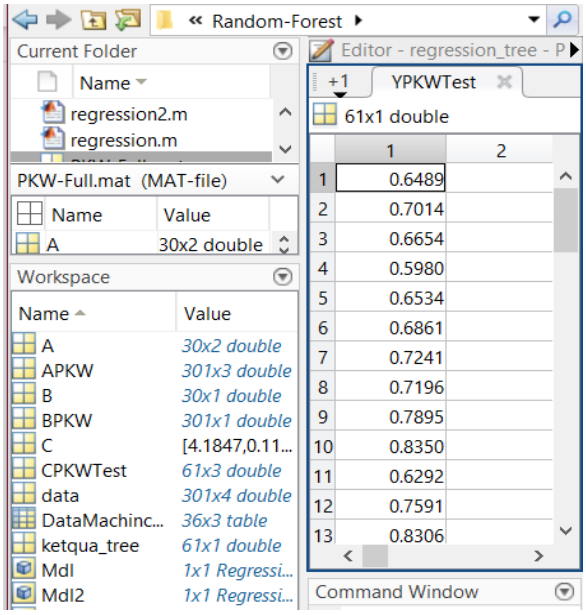

Figure 18. Predicted values of test data by the RF model

Table 7. Predicted results of C_d according to the RF model

Values	Observed values				Predicted values	
	H/P	H/W_u	H/L_u	C_d	C_d	Error (%)
Min	0.170	0.106	0.028	0.574	0.55	0.07
Max	2.150	1.355	0.220	2.220	2.23	14.11

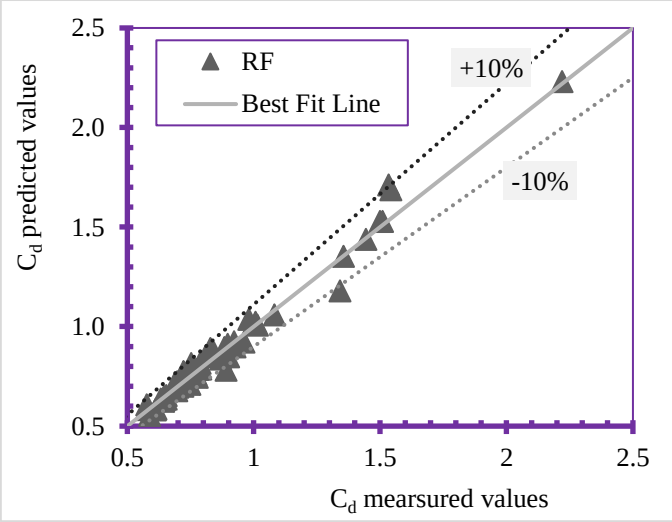

Figure 19. Comparison of C_d between observed data and predicted values by the RF model

There are 3 cases with errors greater than 10% when forecasting using the RF algorithm, but when predicting using the DT algorithm, these cases have errors less than 10% (Table 8). This shows that in the process of automatically splitting data to build "trees" in the RF algorithm, there are rules that are not really optimal. However, in terms of the overall assessment of the entire data series, the statistical indicators were all at the optimal level (close to the ideal point).

Table 8. Predicted values with errors greater than 10%

Observed values				Predicted values			
				RF		DT	
H/P	H/L_u	H/W_u	C_d	C_d	Error (%)	C_d	Error (%)
0.23	0.04	0.207	1.533	1.712	10.4	1.43	7.2
0.17	0.07	0.340	1.341	1.180	13.6	1.233	8.7
0.72	0.06	0.316	0.89	0.780	14.1	0.876	1.6

5.4. Evaluating Prediction Effectiveness Based on Statistical Indicators

Evaluating the predictive ability of the model, analyzing the efficiency, and selecting the optimal model. The study uses the predictive results of the ML models to explore the relationship between the measured and predicted values of the test data series, determining statistical indicators. These indicators evaluate the predictive performance of the ML models. The statistical indicators are shown in Table 9.

Table 9. Statistical indicators of the test data when predicting with the ML models

Algorithms	MEA	MSE	RMSE	R^2	MAPE (%)
DT	0.043	0.003	0.056	0.966	5.0
RF	0.028	0.002	0.046	0.977	3.1

It can be seen from Table 9 that the statistical indicators according to predicted data of the ML models all give good results, the best expected value is the RF model with $R^2 = 0.977$ (strong correlation), MAPE ≈ 3.1% and other statistical indicators very close to the ideal point. Thus, the RF model analizes by the method of averaging many "trees", so the predicting results are closer and more effective than the DT.

This shows that the predictive performance of the ML models is excellent. The performance meets the research well.

5.5. Comparison of Research Results with Empirical Equations

5.5.1. Some empirical equations on the discharge coefficient

The study uses some recently published empirical equations for determining the C_d, which have been evaluated in terms of computational efficiency and practical testing, ensuring the selection of these equations is objective. Some proposed empirical equations are as follows:

+ Study of Ngoc et al.[5]:

- If $H/W_o < 0.5$

$$C_d = 1.856 - 1.729\frac{H}{P} - 0.92\frac{H}{L_u} \quad (4)$$

- If $H/W_o \geq 0.5$

$$C_d = 0.694\frac{P^{0.294}.W_u^{0.148}}{H^{0.442}} \quad (5)$$

+ Study of Guo et al.[14]:

$$C_d = 0.1 + 0.285\left(\frac{L}{W}\right)^{0.45}\left(\frac{B}{P}\right)^{0.1}\left(\frac{W_i}{W_o}\right)^{0.05}\left(\frac{H}{P}\right)^{-0.465} \quad (6)$$

The study will use the above formulas to verify the calculation results and compare them with the predicted values from the ML model.

5.5.2. Calculated results of the discharge coefficient

Using (4), (5), (6), and the test data in Table 4. The calculation of C_d values for each different case and the summary of decomposition results are shown in Table 10.

Table 10. Statistical indicators when tested by empirical equations

Empirical Equations	MEA	MSE	RMSE	R^2	MAPE (%)	ε_{max} (%)
Ngoc et al. [5]	0.044	0.010	0.101	0.89	4.3	47.2
Guo et al. [14]	0.072	0.019	0.138	0.80	7.42	53.3

Comparing the measured values (Table 4) and the calculated values according to the equations (Table 10), it is shown in Fig. 20 and Fig. 21.

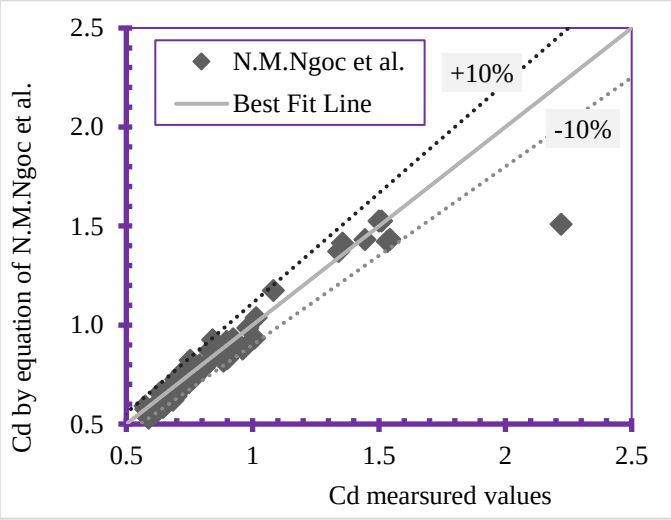

Figure 20. Comparison of measured and calculated data according to the equation of N.M.Ngoc et al.

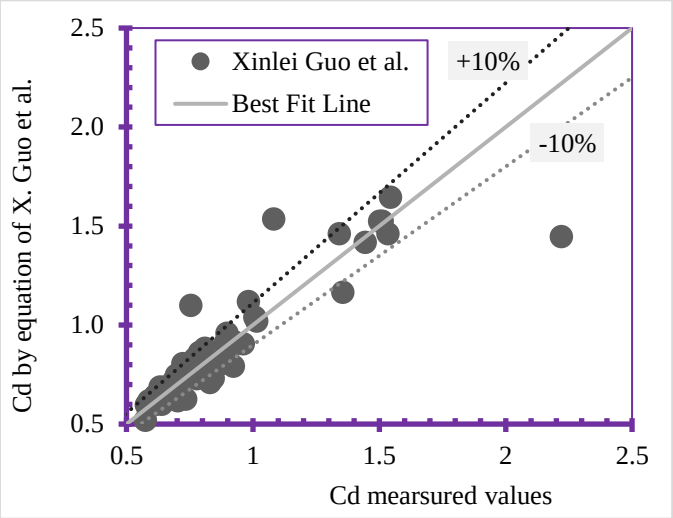

Figure 21. Comparison of measured and calculated data according to the equation of Xinlei Guo et al.

As observed in Table 10, Figs. In 20 and 21, it is evident that the equation of Ngoc et al. demonstrates good calculation efficiency. The statistical indicators are also very close to the "ideal point" (zero), and the correlation coefficient (R^2) is robust ($R^2 = 0.89$). Meanwhile, the equation of Xinlei Gou et al. contains many sudden errors, and the statistical indicators (Table 10) are also not as good as those of the equation of Ngoc et al.[5]. The analysis shows that determining C_d is challenging. Constructing a general equation to determine C_d is challenging, and stabilizing errors is also difficult. Dividing conditions to establish equations for determining C_d is also very complicated to apply. Therefore, using tools with strong discriminant properties and good predictive performance will be more favorable in the study of the C_d of the flow over PKW.

Evaluating the predicted efficiency of the RF model compared to the equation of N.M. Ngoc et al. show that the RF model has better predictive efficiency. Especially when the C_d value is significant, it still gives good forecasting efficiency. Moreover, the RF model does not require conditions in determining the C_d like empirical equations. However, each method has its own advantages and disadvantages. The appearance of values with sudden errors can significantly affect the application of equations to practical work. Therefore, it is necessary to have mutual determination in research and calculations.

When comparing research on determining Cd using the ML algorithm versus traditional empirical equations, the ML algorithm has demonstrated good efficiency. The applied ML algorithm does not have the limiting conditions that are present when applying empirical equations (e.g., Ngoc et al.[5]). This research has integrated many data points from different studies into the training data of the ML algorithm. In the forecast, there are sudden errors of over 10%, mainly due to the use of data from many different sources, leading to significant discrepancies. Therefore, determining Cd values requires

coordination between research using machine learning algorithms and traditional formula methods to reduce possible errors.

6. Conclusions

The ML is an efficient solution with significant potential for application in the study of hydraulic engineering and flow phenomena. The basis for applying ML is the need for complete data; the more, the better. This requires the joint contribution of many scientists to the common database.

This study is the first to apply the "Binary Tree" algorithm in predicting the Cd coefficient of the flow over type-A PKW. Moreover, this study uses diverse data sources and data collected from many physical models of different sizes. These data sources have different data characteristics and variation laws. So, the research results still have many directions for further development. The study has drawn some main conclusions as follows:

- Establishing three basic phases to apply ML algorithms in predicting hydraulic factors, and at the same time, building a research diagram to predict hydraulic factors using the ML models.
- Buckingham's Pi theory ensures scientific appropriateness in setting up data fields in ML (establishing the target variable system and influencing variables).
- The flow datasets over PKW are independent, making them suitable for statistical data characteristics. The evaluation results by statistical indicators are very appropriate. Therefore, applying the ML regression algorithm is highly efficient in predicting the discharge coefficient. In this study, the calculation results from the DT and RF in the ML models, after testing and evaluating with the validation data, showed good predictive values and statistical indicators very close to the ideal point. Meanwhile, the empirical equations had larger and more frequent errors.
- Using ML algorithms to calculate hydraulic characteristics provides fast, accurate, and convenient results. The study has good predictive ability with the geometric conditions of the model: $L/W = 4.0 \div 6.0$ and $Bo/Bi = 1$.
- The research results are applied to predict the flow in experimental models. This helps in choosing the structure of physical models, and the parameters of the model are adjusted quickly. It reduces the time and effort required to build physical models in the laboratory.
- Using the ML algorithms (DT and RF) does not require considering limiting conditions; it only needs appropriate input data.

The study of applying ML to hydraulic engineering in the proposed method is just the basic step, and many other supporting studies are needed to complete the ML model to serve the calculation of flow characteristics over the PKW in practice. In particular, analyzing and comparing the effectiveness and differences between various regression predicting algorithms (such as AI, ML, or DL algorithms), as well as optimizing these ML algorithms, will need to be further researched. In the future, we consider applying the data generation strategies to increase the size of the training and testing datasets.

Conflict of interest

The authors declare that there are no conflicts of interest regarding the publication of this manuscript.

Author Contribution Statement

Nguyen Minh Ngoc conducted the manuscript's structure, proposed the model, and generated the results.

Bui Hai Phong developed the ML model and supervised the findings of this work.

Le Van Nghi collected data, proposed the work, and processed the raw data.

Tran Ngoc Thang and Nguyen Thanh Phong collected data and analyzed the theory using analytical methods and empirical equations.

All authors developed the introduction and contributed to the analysis of the results.

References

[1] M. Iqbal and U. Ghani, "Prediction of the discharge capacity of piano key weirs using artificial neural networks," *Journal of Hydroinformatics*, vol. 26, no. 5, pp. 1167–1188, May 2024. https://doi.org/10.2166/hydro.2024.303

[2] N. Bansal, D. Singh, and M. Kumar, "Computation of energy across the type-C piano key weir using gene expression programming and extreme gradient boosting (XGBoost) algorithm," *Energy Reports*, vol. 9, no. 4, pp. 310–321, Jun. 2023. https://doi.org/10.1016/j.egyr.2023.04.003

[3] O. Machiels, M. Pirotton, P. Archambeau, B. Dewals, and S. Erpicum, "Experimental parametric study and design of Piano Key Weirs," *Journal of Hydraulic Research*, vol. 52, no. 3, pp. 326–335, Mar. 2014. https://doi.org/10.1080/00221686.2013.875070

[4] Q. Rdhaiwi, A. Khoshfetrat, and A. Fathi, "Experimental comparison of flow energy loss in type-B and -C trapezoidal piano key weirs (PKWs)," *J. Eng. Sustain. Dev.*, vol. 28, no. 1, pp. 55–64, Jan. 2024. https://doi.org/10.31272/jeasd.28.1.4

[5] N. M. Ngoc, L. V. Nghi, and D. T. M. Yen, "Experimental Study on the Discharge Coefficient of Type A Piano Key Weir for Water Resource Sustainable Development," *EVERGREEN Joint Journal of Novel Carbon Resource Sciences & Green Asia Strategy*, vol. 10, no. 4, pp. 2299–2307, Dec. 2023. https://doi.org/10.5109/7160907

[6] T. A. Musa and N. O. S. Alghazali, "Enhancing Hydraulic Performance and Discharge Efficiency of a Type-C Piano Key Weir Through Stage Incorporation," *J. Eng. Sustain. Dev.*, vol. 28, no. 5, pp. 630–636, Sep. 2024. https://doi.org/10.31272/jeasd.28.5.8

[7] D. Singh and M. Kumar, "Hydraulic Design and Analysis of Piano Key Weirs: A Review," *Arabian Journal for Science and Engineering*, vol. 47, pp. 5093–5107, Jan. 2022. https://doi.org/10.1007/s13369-021-06370-4

[8] A. Kabiri-Samani and A. Javaheri, "Discharge coefficients for free and submerged flow over Piano Key weirs," *J. Hydraul. Res.*, vol. 50, no. 1, pp. 114–120, Feb. 2012. https://doi.org/10.1080/00221686.2011.647888

[9] B. M. Crookston and B. P. Tullis, "Hydraulic design and analysis of labyrinth weirs. Part 2: Nappe aeration, instability, and vibration," *J. Irrig. Drain. Eng.*, vol. 139, no. 5, pp. 371–377, Oct. 2013. https://doi.org/10.1061/(ASCE)IR.1943-4774.0000553

[10] F. Laugier, "Design and construction of the first Piano Key Weir spillway at Goulours dam," *Int. J. Hydropower Dams*, vol. 14, no. 5, pp. 94–100, Jan. 2007.

[11] S. Li, G. Li, D. Jiang, and J. Ning, "Influence of auxiliary geometric parameters on discharge capacity of piano key weirs," *Flow Meas. Instrum.*, vol. 72, p. 101719, Apr. 2020. https://doi.org/10.1016/j.flowmeasinst.2020.101719

[12] M. B. N. Al-Baghdadi and S. I. Khassaf, "Evaluation of crest length effect on piano key weir discharge coefficient," *Int. J. Energy Environ.*, vol. 9, no. 5, pp. 473–480, Sep. 2018.

[13] H. K. Al-Shukur and G. H. Al-Khafaji, "Experimental study of the hydraulic performance of piano key weir," *Int. J. Energy Environ.*, vol. 9, no. 1, pp. 63–70, Jan. 2018.

[14] X. Guo, Z. Liu, T. Wang, H. Fu, J. Li, Q. Xia, and Y. Guo, "Discharge capacity evaluation and hydraulic design of a piano key weir," *Water Supply*, vol. 19, no. 3, pp. 871–878, May 2019. https://doi.org/10.2166/ws.2018.134

[15] D. T. M. Yen and L. V. Nghi. "Characteristic shape and continuity of flow over the PKW", *Journal of Water Resources Sciences & Technology*, vol. 42/2017, pp. 1-8, Aug. 2017.

[16] L. V. Nghi and D. T. M. Yen, "Establishing an equation for determining the discharge capacity factor for free flow over the PKW," *Journal of Water Resources Sciences & Technology*, vol. 54, pp. 1–8, Jun. 2019.

[17] M. Haghbin and A. Sharafati, "A review of studies on estimating the discharge coefficient of flow control structures based on the soft computing models," *Flow Meas. Instrum.*, vol. 83, p. 102119, Jan. 2022. https://doi.org/10.1016/j.flowmeasinst.2021.102119

[18] D. Shekhar, B. S. Das, K. Devi, J. R. Khuntia, and T. Karmaker, "Discharge estimation in a compound channel with converging and diverging floodplains using ANN–PSO and MARS," *Journal of Hydroinformatics*, vol. 25, no. 6, pp. 2479–2499, Oct. 2023. https://doi.org/10.2166/hydro.2023.145

[19] S. M. Seyedian, A. Haghiabi, and A. Parsaie, "Reliable prediction of the discharge coefficient of triangular labyrinth weir based on soft computing techniques," *Flow Meas. Instrum.*, vol. 92, p. 102403, Aug. 2023. https://doi.org/10.1016/j.flowmeasinst.2023.102403

[20] N. I. Khattab, A. N. Altalib, and M. Y. Mohammed, "Estimating Discharge Coefficient for Piano Key Weir using ANN Technique," *COEC8-2021 Proceedings 24–25 November 2021, Baghdad, Iraq*, pp. 461–1327. https://doi.org/10.1063/5.0132654

[21] O. F. Dursun, N. Kaya, and M. Firat, "Estimating discharge coefficient of semi-elliptical side weir using ANFIS," *Journal of Hydrology*, vol. 426, pp. 55–62, Mar. 2012. https://doi.org/10.1016/j.jhydrol.2012.01.010

[22] H. Zaji, H. Bonakdari, S. R. Khodashenas, and S. Shamshirband, "Firefly optimization algorithm effect on support vector regression prediction improvement of a modified labyrinth side weir's discharge coefficient," *Applied Mathematics and Computation*, vol. 274, pp. 14–19, Feb. 2016. https://doi.org/10.1016/j.amc.2015.10.070

[23] A. H. Haghiabi, A. Parsaie, and S. Ememgholizadeh, "Prediction of discharge coefficient of triangular labyrinth weirs using Adaptive Neuro Fuzzy Inference System," *Alexandria Engineering Journal*, vol. 57, no. 3, pp. 1773–1782, Sep. 2018. https://doi.org/10.1016/j.aej.2017.05.005

[24] E. Gul, M. N. Alpaslan, and M. E. Emiroglu, "Robust optimization of SVM hyper-parameters for spillway type selection," *Ain Shams Engineering Journal*, vol. 12, no. 3, pp. 2413–2423, Sep. 2021. https://doi.org/10.1016/j.asej.2020.10.022

[25] M. Zounemat-Kermani and A. Mahdavi-Meymand, "Hybrid meta-heuristics artificial intelligence models in simulating discharge passing the piano key weirs," *Journal of Hydrology*, vol. 569, pp. 12–21, Feb. 2019. https://doi.org/10.1016/j.jhydrol.2018.11.052

[26] D. Singh and M. Kumar, "Computation of energy dissipation across the type-A piano key weir by using gene expression programming technique," *Water Supply*, vol. 22, no. 8, pp. 6715–6727, Jul. 2022. https://doi.org/10.2166/ws.2022.255

[27] N. M. Ngoc, P. H. Cuong, and B. H. Phong, "Prediction of the conjugate depth of the hydraulic jump in the trapezoidal channel using Random Forest regression," *Journal of Military Science and Technology*, vol. 82, pp. 150–158, Oct. 2022. https://doi.org/10.54939/1859-1043.j.mst.82.2022.150-158

[28] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," *SN Comput. Sci.*, vol. 2, no. 3, p. 160, Mar. 2021. https://doi.org/10.1007/s42979-021-00592-x

[29] K. Hamed and M. K. Elshaarawy, "Soft computing approaches for forecasting discharge over symmetrical piano key weirs," *AI in Civil Engineering*, vol. 4, no. 6, pp. 2730–5392, Mar. 2025. https://doi.org/10.1007/s43503-024-00048-0

[30] M. Shariq, A. Hussain, and Z. Ahmad, "Flow over gabion weir under free and submerged flow conditions," *Flow Meas. Instrum.*, vol. 86, p. 102199, Aug. 2022. https://doi.org/10.1016/j.flowmeasinst.2022.102199

[31] V. Rezaei, S. H. Musavi-Jahromi, A. Khosrowjerdi, and B. Beheshti, "Experimental and Simulation Studies on Water ischarge Coefficients of Rectangular Piano Key Weirs," *International Journal of Technology*, vol. 13, no. 4, pp. 695–705, Oct. 2022. https://doi.org/10.14716/ijtech.v13i4.5377

[32] R. K. Bhukya, M. Pandey, M. Valyrakis, and P. Michalis, "Discharge Estimation over Piano Key Weirs: A Review of Recent Developments," *Water*, vol. 14, no. 19, p. 3029, Sep. 2022. https://doi.org/10.3390/w14193029

[33] F. Aljanabi, M. Dedeoğlu, and C. Şeker, "Environmental Monitoring of Land Use/ Land Cover by Integrating Remote Sensing and Machine Learning Algorithms," *J. Eng. Sustain. Dev.*, vol. 28, no. 4, pp. 455–466, Jul. 2024. https://doi.org/10.31272/jeasd.28.4.4

[34] N. M. Ngoc and B. H. Phong, "Performance Comparison of Prediction of a Hydraulic Jump Depth in a Channel Using Various Machine Learning Models," in *Creative Approaches Towards Development of Computing and Multidisciplinary IT Solutions for Society*, 1st ed., Hoboken, NJ, USA: Scrivener Publishing LLC, pp. 189–205, Sep. 2024. https://doi.org/10.1002/9781394272303.ch11

[35] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001. https://doi.org/10.1023/A:1010933404324

[36] N. T. Hai, "Report on the product of scientific research topic: Study on the application of Piano key weir in the construction of hydraulic works in Vietnam," Southern Institute of Water Resources Research, Ho Chi Minh City, 2014.

[37] I. Noui and A. Ouamane, "Study of optimization of the Piano Key Weir," in *Labyrinth and Piano Key Weirs — PKW 2011*, London, UK: CRC Press, pp. 175–182, 2011.

[38] F. J. M. Denys, "Investigation into Flow-induced Vibrations of Piano Key Weirs," Ph.D. dissertation, Stellenbosch University, South Africa, 2019.

[39] S. Erpicum, B. P. Tullis, M. Lodomez, P. Archambeau, B. J. Dewals, and M. Pirotton, "Scale effects in physical piano key weirs models," *J. Hydraul. Res.*, vol. 54, no. 6, pp. 692–698, Aug. 2016. https://doi.org/10.1080/00221686.2016.1211562