

Determination of the optimal hyperparameters of binary tree algorithms for predicting hydraulic jump characteristics in a smooth trapezoidal channel

Nguyen Kien Quyet¹, Nguyen Quang Hung², Nguyen Minh Ngoc^{3*}

¹ University of Transport Technology, Hanoi, Vietnam

² Faculty of Civil Engineering, Thuyloi University, Hanoi, Vietnam

³ Faculty of Urban Environmental and Infrastructural Engineering, Hanoi Architectural University, Hanoi, Vietnam

* Corresponding author's e-mail: ngocnm@hau.edu.vn

ABSTRACT

The algorithms based on binary trees can be used for the applications involving regression prediction and classification. Decision trees, boosted trees, and bagged trees are among the algorithms examined in this study. The study analyzed the algorithmic structure and developed a procedure for analyzing and predicting a hydraulic parameter using these algorithms. The Buckingham Pi theorem was applied to define the objective function, thereby identifying the input data fields required for the machine learning models and determining the experimental data to be collected. The proposed machine learning models were subsequently applied to predict the hydraulic jump characteristics in a trapezoidal channel. The results indicate that machine learning-based modeling is highly suitable for predicting hydraulic parameters. Evaluation using statistical performance indicators showed strong to very strong predictive performance, with the coefficient of determination R^2 ranging from 0.95 to 1.00. Other statistical indicators, including RMSE, MSE, and MAE, were all very close to their ideal value of zero. The study further showed that the proposed Boosted Tree 7 model (with “minimum leaf size” equal to 1 and “number of learners” equal to 800) achieved the best predictive performance. For the training dataset, the model produced an R^2 value approximately equal to 1, while the other statistical indicators were nearly zero. For the testing dataset, the model also achieved a very strong coefficient of determination, with R^2 approximately equal to 0.98. The maximum prediction error for hydraulic jump length was 5.2%. These findings demonstrate that machine learning algorithms are highly suitable and effective for predicting hydraulic characteristics in hydraulic engineering applications.

Keywords: machine learning, binary-tree, boosted tree, bagged tree, Pi theory, hydraulic jump, length jump, sequent depth.

INTRODUCTION

Machine learning (ML) is a subset of artificial intelligence (AI). ML is a branch of computer science that was created in the 1980s. Without the need for particular programming procedures for the research object, it can self-learn based on the input database. ML is “trained” using vast amounts of data and algorithms, which enable it to understand how to accomplish tasks, as opposed to manually generating software commands with

a specific set of instructions to complete a certain task [1]. ML has a very close relationship with the field of statistics [2]. ML uses statistical models to “memorize” the distribution of data, simulate generalization and inference. From there, provide forecast results based on optimal statistical criteria (R^2 , MSE, RMSE, etc.) [3].

ML is considered to be well applied in the fields of analytical research for fluids, providing good support in analyzing experimental data of physical models and facilitating data sharing

[4]. ML algorithms are very rich with system structures from simple to complex, including algorithms such as DT, RF, SVM (Support Vector Machine), KNN (k-nearest neighbors), etc. [5]. DT and RF applications have been applied in many fields, such as analyzing hydrological and hydraulic factors of river basins, thereby predicting water levels and flow of the Garonne River (France) [6]. Analyzing water prices, forecasting the possibility of water price growth (predicted efficiency reaches (R-squared) $R^2 > 0.8$), thereby helping to adjust water prices appropriately and quickly [7]. In hydraulic calculations, the analysis for dirty drainage networks also gives good performance ($R^2 = 0.994$) [8], or flood analysis also helps reduce calculation time and costs [9]. The analysis for the water supply network has shown an effective reduction in water loss compared to the conventional plan by up to 40% [10].

For the study of the hydraulic jump on a rectangular channel with a roughness bed, the decision tree classifier (J48) algorithms and a multi-layer neural network (NN) in ML have shown high performance in classifying jumping water [11], or applying Gene Expression Programming (GEP), ANN and SVM algorithms also bring high efficiency in predicting water hydraulic characteristics (y_2/y_1 and L_j/y_1) [12], or good performance in predicting the jump length using expression programming (GEP), adaptive-neuro-fuzzy inference system with grid partition (ANFIS-GP), and neural networks (ANNs) algorithms [13]. In addition, when predicting the jump length using the GEP algorithm in the slop channel [14] or applying ANN in smooth bed [15], the results of evaluation of the suitability are also very high.

The application of ML algorithms to predicting hydraulic jump length has been widely investigated in recent years. Previous studies have mainly focused on predicting hydraulic jump length in rectangular channels with smooth, rough, sloping, or gradually expanding beds, among other configurations [16–18]. Existing studies have primarily relied on comparing different algorithms to identify the most suitable model for each specific case. However, the determination of optimal hyperparameters for each algorithm has not yet been sufficiently investigated. For example, in a study predicting hydraulic jump length over a rough channel bed using seven ML algorithms, the results after parameter optimization showed that XGBoost and CatBoost achieved the best performance [19].

Another study applied six ML algorithms to predict hydraulic jump length in a stilling basin and found that ANN and CatBoost provided better predictive performance. The study also indicated that binary tree-based algorithms do not automatically optimize their hyperparameters [20].

Based on a review of previous studies on the application of machine learning to hydraulic engineering problems, most existing research has focused primarily on using ML algorithms to predict various hydraulic characteristics, while comparatively little attention has been given to optimizing the predictive accuracy of each algorithm. Even when the same algorithm is employed, different hyperparameter settings can produce substantially different levels of predictive performance. Therefore, evaluating and tuning ML algorithms under different parameter configurations is essential for identifying the models with higher accuracy and better generalization capability. In practice, in the field of hydraulics and structural hydraulics, there are very few studies analyzing the optimization of binary tree algorithms, which limit their applicability and the potential for expanding research capabilities. Binary tree algorithms have a wide range of applications and do not require overly stringent input data, ensuring accurate results for regression prediction studies.

Therefore, this study focuses on the evaluation and analysis of the Binary tree algorithms. The research includes theory, data characteristics in ML, determining data fields in ML from Buckingham's Pi theory, and analytical examples of geometric characteristics (the length of the jump) of the jump in the smooth horizontal trapezoidal channel.

METHODOLOGY

Decision tree algorithm

Decision tree (DT) [21] is a type of prediction model, it is used in statistics, data mining and ML. Using a decision tree (as a prediction model) to go from observations of a target value (represented in the branches) to the predicted outcome (represented in the leaves). DT is the most powerful algorithm of the supervised algorithm type. The main entities of the decision tree are “nodes” and “leaves”, this is where the data is divided and left, where the prediction results appear. The

decision to divide the strategy greatly affects the accuracy of the “tree” (Figure 1).

Considering the binary tree algorithms, the algorithm first selects a value and divides the data into two subsets. For each subset, a separate mean square error (MSE) value is calculated.

Expanding algorithms from binary trees

Expanding the application of binary tree algorithms, which is the process of combining and reconstructing binary trees. It is described as the Random Forest algorithm. It can be seen that the Random forest (RF) [21] is a supervised learning method, so it can handle problems of classification and prediction of values. RF is a set of many “trees”, in which each “tree” is randomly created from resampling (randomly selecting a part of the data to build the tree model) and randomly selecting variables from all variables in the data. With such a mechanism, RF yields very high accuracy, but it will be impossible to grasp the operating mechanism of this algorithm (due to the too complex structure of the model). This algorithm can

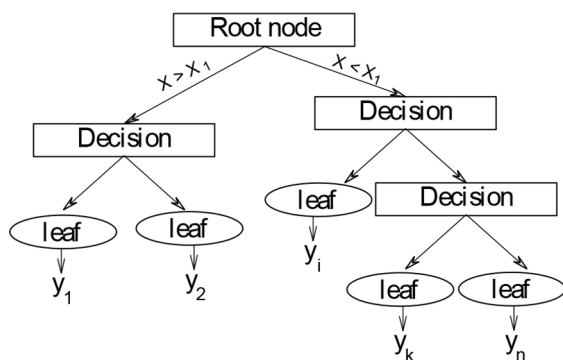
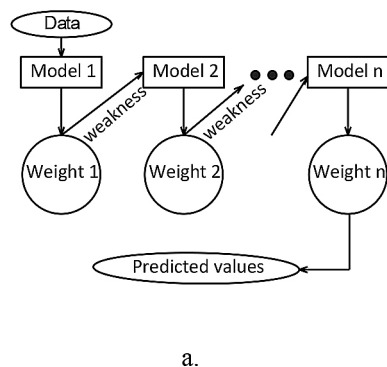
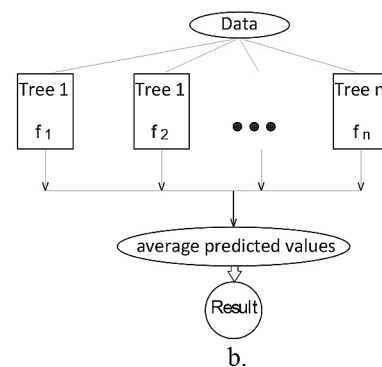


Figure 1. Algorithmic structure of binary decision tree [21]



a.



b.

Figure 2. Structure of models in the Random Forest algorithms: a) structure of the Boosted Tree model[22], b) structure of the Bagged Tree model [23,24]

be considered as one of the “black box” modeling methods and can only draw results, but cannot explain the operating mechanism of the model.

Random Forest is employed as a bagged tree-based ML model for predicting the hydraulic characteristics of hydraulic jump (the sequent depth and jump length). The model performance is optimized by tuning key hyperparameters, particularly the number of “learners” and the number of “leaves”, which control the ensemble size and the complexity of individual decision trees. For comparison, boosted tree models may also be considered, as they differ from Random Forest by constructing trees sequentially and adjusting the contribution of each “learner” based on prediction errors from previous iterations[15,21].

The Boosted Tree model determines the weights of “trees” by iteratively adjusting these weights according to each analysis step (Figure 2a). The Bagged Tree model (Figure 2b) determines the weight for each “tree” set up in the model, and selecting a sufficient learning class will ensure the analysis results [22].

These algorithms were proposed by Friedman in 1996 and improved with the contributions of scientists [25,26]. Determining the minimum number of “leaves” and the number of “learners” plays an important role in building regression forecasts. For the Bagged Tree model, determining the number of “learners” is limited for effective prediction; the research shows that the effectiveness of the number of classes is about 100 [27].

Dimensional analysis and study method

Dimensional analysis is a means used to solve a physical problem, which it does by homogenizing parameters to provide an indication of the

factors influencing the phenomenon under study [24,28]. Let $q_1, q_2, q_3 \dots q_n$ be n -dimensional variables, it is determined in a given problem, and it is related by a basic theoretical equation, then they are expressed in the form of related functional equations as follows:

$$F(q_1, q_2, \dots, q_n) = 0 \quad (1)$$

Choosing k main variables and Equation 1 will be decomposed into $(n - k)$ dimensionless and independent variables, called “Pi group” (Π). Applying Buckingham’s Pi theory, Equation 1 is transformed into the following relationship:

$$\Pi_1 = \Psi(\Pi_2, \Pi_3 \dots \Pi_{n-k}) \quad (2)$$

Equation 2 is used to analyze and evaluate the factors affecting the quantity to be studied (Π_1). At the same time, point out the data characteristics that need to be collected in experimental research on physical models. In ML problems, Equation 2 is the basis for determining the target variables (Π_1) and influencing variables in the regression problem ($\Pi_2, \Pi_3, \dots, \Pi_{n-k}$). The analysis of ML models is based on Equation 2, showing a complete view of the influencing factors in the problem of prediction using ML models [15, 24, 29].

Determining the predictive value of the ML model, the value and influencing factors must first be determined. Therefore, determining the objective function as content plays an important role in obtaining accurate forecasts. Ensuring accurate predictions of the ML models, this research proposed a method to apply Buckingham’s Pi theory to establish the objective function, as well as determine the influencing factors and data collection from the physical experiment model.

Figure 3 simulates the method of applying ML models in research, especially suitable for predicting with regression models in studying hydraulic construction problems. In this study, the research structure in Figure 3 was applied to illustrate the suitability of ML algorithms in predicting hydraulic problems, and ML models were set up in MATLAB software.

The predictive performance of the optimized ML models was evaluated using statistical indicators that quantify the agreement between measured and predicted values. In MATLAB, ML models such as Decision Tree, Boosted Tree, and Bagged Tree models provide built-in statistical metrics for assessing prediction accuracy. Similar evaluation approaches have also been widely reported

in previous studies [29]. The commonly used performance indicators include the root mean square error (RMSE), mean square error (MSE), coefficient of determination (R^2), mean absolute error (MAE), and percentage error (%) [2,29].

ILLUSTRATING THE MACHINE LEARNING PROBLEM VIA THE STUDY ON PREDICTING THE HYDRAULIC CHARACTERISTICS OF THE JUMP

Theoretical equation of geometric characteristics of the jump

The hydraulic jump is a phenomenon when the flow depth changes suddenly from low ($y_1 < y_c$) to high depths ($y_2 > y_c$). In other words, the flow state changes from a supercritical flow to a subcritical flow. Geometric characteristics of the jump include the jump length (L_j) and the sequent depth of the jump (y_1, y_2). Studies on the jump (Figure 4), the equations of momentum and mass are applied and meet the changing characteristics of the hydraulic factors.

The study analyzed the structure of the jump in the smooth trapezoidal channel (stilling basin, bed width - b , side slope - m), the momentum equation for the zone of the hydraulic jump [24, 29]:

$$P_1 - P_2 - F_{ms} = \rho Q \alpha_{02} V_2 - \rho Q \alpha_{01} V_1 \quad (3)$$

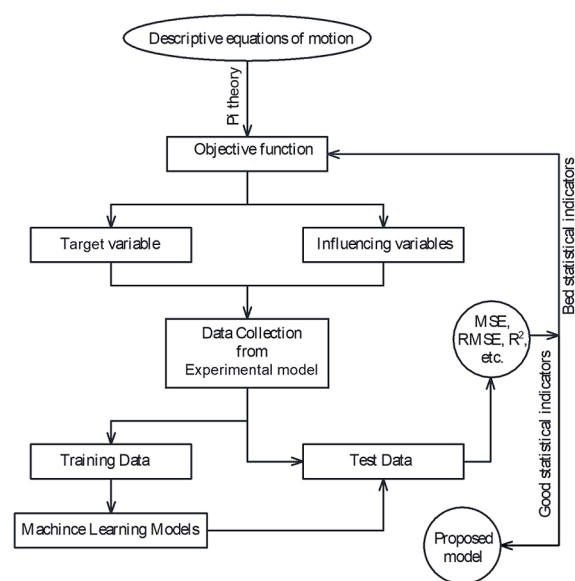


Figure 3. Simulation structure on applying ML models in this study

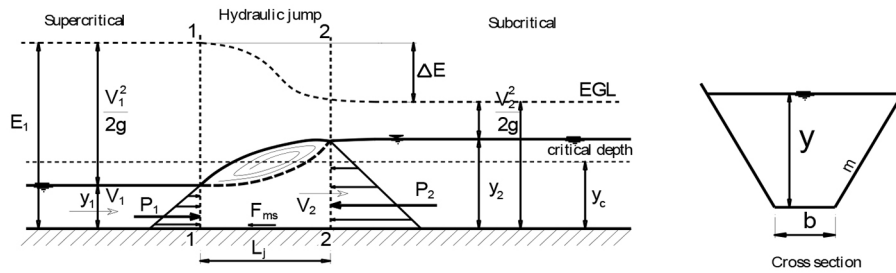


Figure 4. The Jump structure on a horizontal trapezoidal channel

where: P_1, P_2 – pressure force at sections 1 and 2, respectively (N); F_{ms} – friction loss (N); Q – discharge (m^3/s or l/s); σ_{01}, σ_{02} – kinetic coefficient at sections 1 and 2, respectively; V_1, V_2 – mean velocity of flow at sections 1 and 2, respectively (m/s).

Equation 3 is encountered in many different studies. Depending on the research objective (the jump length or the sequent depths), the friction force component (F_{ms}) is determined accordingly. The study considered two analytical cases for Equation 3, the analyses are shown in Table 1.

Based on the objective function, the target data system (y_2/y_1 or L_j/y_1) and data of influencing factors in training the ML model were determined. Equations 9 and 10 help determine the training data system, thereby serving as a basis for collecting input data. This is similar to the study of determining empirical or semi-empirical equations. However, in equations, the distribution law does not change, but in ML, data is divided into groups, and determining the optimal rules for forming output data will be based on the optimization of statistical indicators (MSE, RMSE, R^2 , etc.) [29].

Experimental model and research data

The physical experiments were conducted using symmetric trapezoidal channels with bottom widths of 33.5 cm and 55 cm at the Key Laboratory of River and Coastal Engineering, Vietnam Academy for Water Resources. The experimental models were designed to represent the geometry of practical hydraulic structures, with a symmetric trapezoidal cross-section adopted for the stilling basin downstream of a spillway. The data obtained from channels with different bottom widths were included to minimize the influence of model-scale effects on the experimental results. In addition, datasets reported by other researchers were incorporated to broaden the range and diversity of the database used in the analysis. Since the design parameters of the stilling basin were determined based on the characteristics of a stable hydraulic jump, the present study was limited to Froude numbers, Fr , ranging from 4.5 to 9.0.

The study used 3 physical models (with bed width b is 20 cm, 33.5 cm, and 55 cm) with the same conditions (the horizontal trapezoidal

Table 1. Objective function analysis for geometric characteristics of the jump according to Buckingham theorem

a. Study on the sequent depths	b. Study on the jump length
+ Friction force Equation 3 only considers hydrostatic pressure (P_1, P_2), average velocity (V_1, V_2), discharge (Q), momentum coefficient (α_0), density (ρ), and viscosity (μ).	
The friction force from turbulent flow has little effect on the depth characteristics, so the friction force component (F_{ms}) can be ignored, and this factor may be hidden in the experimental coefficients of Froude number in the study.	Using a frictional shear stress theory (τ_0) to analyze the friction force components according to the length (L_j) and average wetted circumference ($\bar{P}$) [24], there are: $F_{ms} = \tau_0 L_j \bar{P} \quad (4)$
+ The equation to determine the influencing factors is as follows [15]:	
$f(y_2, y_1, V_1, V_2, m, b, \alpha_0, \rho, \mu, g) = 0 \quad (5)$	$f(L_j, y_1, y_2, V_1, V_2, m, \rho, \mu, g, b, \alpha_0) = 0 \quad (6)$
+ Applying dimensional analysis theory to determine the equation of influencing factors [22],[23]:	
$\beta(y_2/y_1, b/y_1, m, Fr_1) = 0 \quad (7)$	$\beta(L_j/y_1, y_2/y_1, m, b/y_1, Fr_1) = 0 \quad (8)$
+ According to Buckingham's Pi theory to determine the objective function in studies	
$y_2/y_1 = \Psi(M_1, Fr_1) \quad (9)$	$L_j/y_1 = \Phi(y_2/y_1, Fr_1, M_1) \quad (10)$

channel; a side slope of the channel is m , it is 1, and smooth bed and the channel made of organic glass) to collect data in Table 2 [24,29,30].

About 10% of research data (Table 2, 3) were separated to create test data; random sampling and ensuring test data were spread evenly within the research data range (the study used the Index function in Excel to separate data objectively). Thus, with a total of 109 datasets, 12 datasets (11% of the total data) were separated for testing. The test data is shown in Table 4.

Figure 5 represents the research data collected from 3 experimental models (Tables 2, 3). Each relationship between hydraulic factors tends to converge. Although the coefficient of determination ($R^2 \gg 0.94$) is large, on the chart the values still have a large dispersion around the average line. Similarly, the test values (Table 4 and Figure 6) also have a strong correlation and have a relatively clear change rule. The experimental data show a relatively uniform distribution of the hydraulic characteristics over the investigated Froude-number range of $Fr = 4.5\text{--}9.0$. In particular, both the training and testing datasets adequately cover the range of hydraulic conditions considered in this study, ensuring that their distributions are consistent with the intended scope of the predictive analysis.

The experimental data were obtained from independent tests. Each of the 121 data records corresponds to a separate experimental run conducted under a distinct set of input conditions. This experimental design provides an appropriate degree of independence among the observations

and ensures that the dataset is representative of the investigated hydraulic conditions. Consequently, the dataset provides a reliable basis for developing, training, and evaluating predictive models using different machine-learning algorithms. This shows that regression studies on prediction are highly feasible. ML models in MATLAB will use these data to set up training models and run test data to evaluate the suitability of the ML model in studying hydraulic construction.

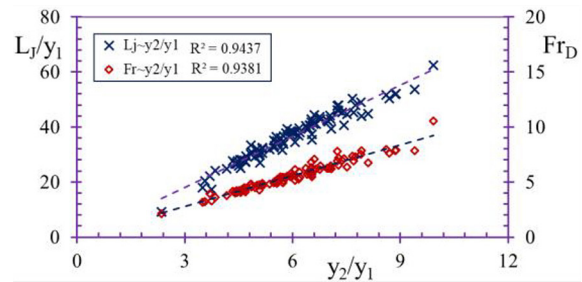


Figure 5. Relationships among hydraulic parameters in the experimental dataset

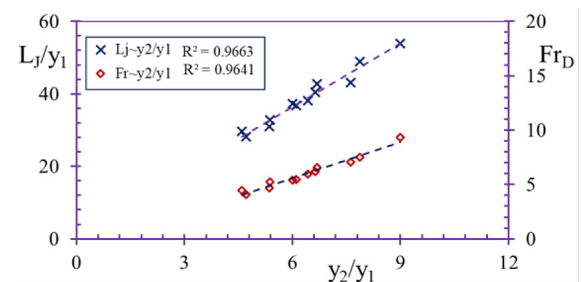


Figure 6. Relationships among hydraulic parameters in the testing dataset

Table 2. Experimental data from physical models

Ref.	Bed width (cm)	Length of plum (m)	Q (l/s)	y_1 (m)	y_2 (m)	L_j (m)
[24]	55	4	$60 \div 201$	$0.033 \div 0.081$	$0.182 \div 0.488$	$1.0 \div 2.87$
	33.5		$65 \div 167$	$0.041 \div 0.092$	$0.258 \div 0.490$	$1.5 \div 3.18$
[30]	20	6	$20.1 \div 98$	$0.0405 \div 0.0812$	$0.142 \div 0.441$	$0.55 \div 3.0$

Table 3. Research data

Values	b (cm)	y_2/y_1	$Fr = V_1/(g \cdot y_1)^{0.5}$	$M_1 = m y_1/b$	L_j/y_1
Min	20	2.347	2.125	0.060	9.091
Max	55	9.926	10.570	0.406	62.50

Table 4. Testing data for ML models

Values	b (cm)	Fr_D	M_1	y_2/y_1	L_j/y_1
Min	20	4.587	4.062	0.087	28.125
Max	55	7.879	7.776	0.406	48.966

RESULTS AND DISCUSSION

Machine learning models applied in the study

A decision tree algorithm is represented by nodes and leaves. In this study, the decision tree model was initially simulated using three default configurations: fine tree, medium tree, and coarse tree. However, the model can be further refined by pruning the leaves and adjusting the decision nodes. Therefore, several decision tree models with different pruning settings were analyzed.

In addition, the Random Forest model also provides default ensemble-based configurations, such as boosted and bagged tree approaches. For the Random Forest model, the splitting process and tree reconstruction can be tuned to generate multiple decision trees and combine them into an ensemble forest. In this study, several splitting criteria and combinations of decision trees within the Random Forest framework were analyzed, as presented in Table 5.

Table 5 presents the configuration of the tree-based ML models used in this study. For the Tree model, the model structure was defined by specifying the number of leaves and decision nodes. For the Boosted Tree and Bagged Tree

models, the number of leaves and learners was determined. All model configurations were tested and analyzed to identify the optimal structure for each model type. Based on these results, suitable ML models were proposed for predicting the hydraulic characteristics of hydraulic jumps in symmetric trapezoidal channels.

Analyzing the problem of predicting the sequent depth

Analysis using the tree model

For the binary decision tree model, the model configurations were defined by adjusting the number of leaves and decision nodes. Six tree models were investigated in this study. After the training process, the basic dataset was used to analyze the predictive performance of each model and to compare the measured values with the predicted values. Based on this comparison, statistical indicators were determined to evaluate the prediction accuracy and overall performance of the models. The analysis results are presented in Table 6.

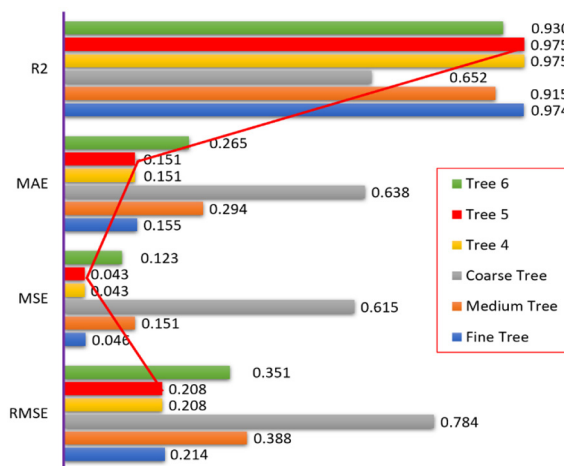
Based on Table 6 and Figure 7, the statistical indicators show that the Tree 5 model achieved the best predictive performance, with a very strong coefficient of determination R^2 ($R^2 > 0.9$),

Table 5. Selection of model parameters used in the study

Models		Nodes	Minimum leaf size	Number of learners
Default decision tree model	Fine tree	10	4	
	Medium tree	10	12	
	Coarse tree	10	36	
Decision tree model with tuned nodes and leaves	Tree 4	9	3	
	Tree 5	10	1	
	Tree 6	30	10	
Boosted Tree	Boosted Tree 2		8	30
	Boosted Tree 3		1	30
	Boosted Tree 4		5	30
	Boosted Tree 5		1	100
	Boosted Tree 6		1	500
	Boosted Tree 7		1	800
Bagged Tree	Bagged Tree 2		8	30
	Bagged Tree 3		5	30
	Bagged Tree 4		1	30
	Bagged Tree 5		1	100
	Bagged Tree 6		1	150
	Bagged Tree 7		1	200
	Bagged Tree 8		1	250
	Bagged Tree 9		1	300

Table 6. Results of the binary decision tree models

Algorithms	RMSE	MSE	R ²	MAE
Fine tree	0.21392	0.04576	0.97412	0.1552
Medium tree	0.38838	0.15084	0.91469	0.29449
Coarse tree	0.78435	0.61521	0.65206	0.63829
Tree 4	0.20832	0.043396	0.97547	0.15101
Tree 5	0.2083	0.04339	0.97548	0.15053
Tree 6	0.35134	0.12344	0.93019	0.26462


Figure 7. Statistical performance indicators of the tree models

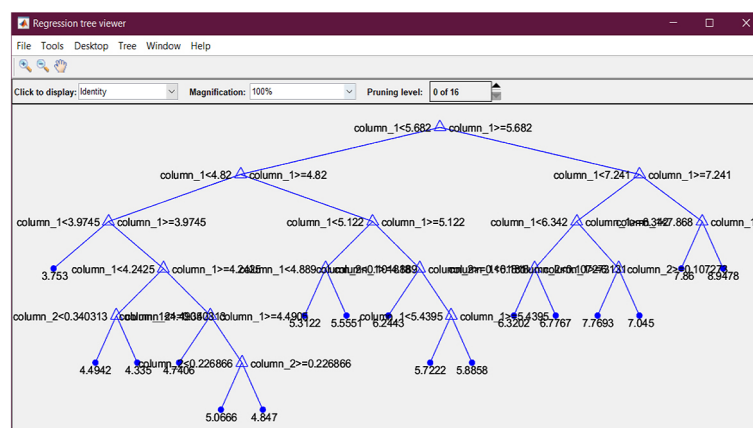
while the other error indicators were close to their ideal values, namely the zero point. The tree 5 model, with 10 decision nodes and a minimum leaf size of 1 per node, was therefore identified as the most effective model for prediction using the decision tree algorithm.

The tree 5 model was then applied to evaluate the testing dataset. The analysis results for the testing data presented in Table 3 are shown in Figures 7 and 8 as well as Table 6.

As shown in Figures 7 and 8, the tree 5 model selected the Froude number variable (Column_1) as the main predictor for splitting the decision nodes. The model structure was defined with a minimum leaf size of 1 and a maximum of 10 decision nodes. This configuration was selected based on the RMSE optimization results, which indicated the best predictive performance for the decision tree model.

The prediction of the sequent depth of hydraulic jumps using the tree model indicates that the proposed ML approach is suitable for predicting the geometric characteristics of hydraulic jumps. The tree model classified and predicted the data into separate classes (Figure 9), with each classified data class represented by one row of predictions from Model 5. For the tree 5 model, the statistical indicators approached their ideal values, with ($R^2 = 0.97$), which is very close to 1, while the other error indicators were close to 0, including RMSE = 0.208, MSE = 0.043, and MAE = 0.15. The maximum error range was approximately 0.82, as shown in Figure 5.

In addition, the input data order was reversed compared with the initial analysis (In MATLAB software, the information includes: Column_1 represented the value of M_1 and Column_2


Figure 8. Structure of the tree 5 decision tree model for sequent depth prediction

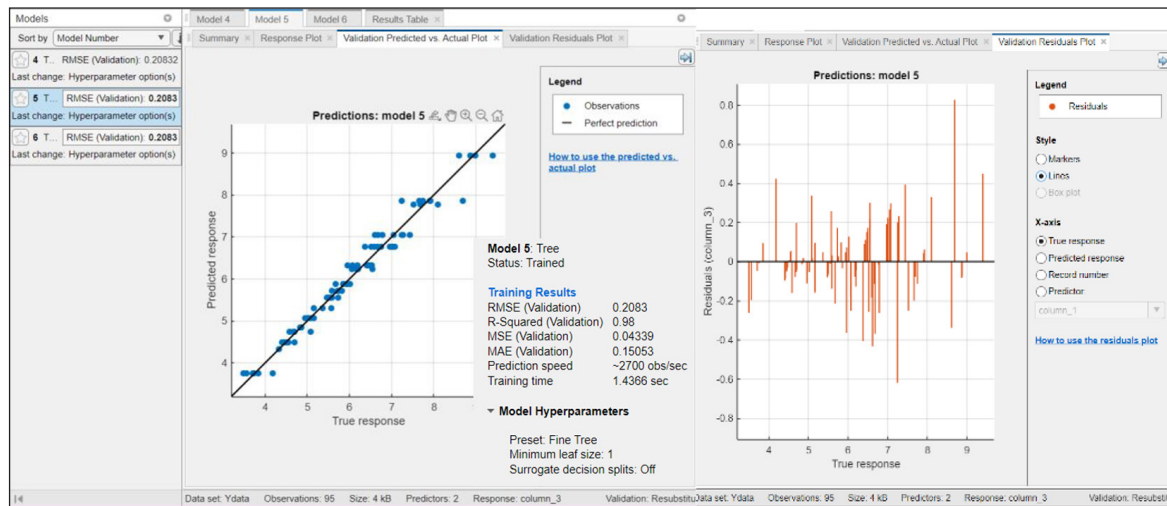


Figure 9. Training results of the tree 5 model

represented the value of Fr). The analysis was then performed again using the tree 5 model. However, the results showed no significant difference. Verification through the decision tree model indicated that the tree model automatically selected the appropriate data column as the basis for splitting the decision nodes. The Froude number data column was used for node splitting, regardless of the order of the input variables used in the training process (Figure 9).

This study reused the testing dataset from [29] to analyze the tree 5 model. The results showed that the maximum prediction error was 17%, which is considerably smaller than that obtained using the “decision tree” model with the LS_Boost algorithm proposed by [31], where the maximum error reached 32.4%. This improvement can be attributed to the fact that the Tree model developed in the present study was pruned and optimized by adjusting the minimum number of leaves and the number of nodes to obtain the optimal model structure. In contrast, the study by [29] did not adjust the leaf and node parameters, whereas the specific node and leaf structure were not clearly

identified during the model analysis using the source code. In addition, the training dataset used for the tree 5 model in the present study was expanded, with approximately 15% more data than that used in [29].

Analysis using the Boosted Tree model

The Boosted Tree model was investigated using different configurations of the minimum leaf size and the number of learners, as presented in Table 5. The training process was performed using the same dataset as that used for the Binary Tree model, based on the LS_Boost algorithm [31]. The training results were evaluated using statistical performance indicators analyzed in MATLAB.

Based on the results presented in Table 7 and Figure 10, the Boosted Tree 7 model (Figure 11) achieved the best predictive performance according to the statistical indicators. The analysis showed that, for the Boosted Tree models, when the number of learners was 500 or higher, the statistical indicators approached their ideal values. Therefore, the Boosted Tree 7 model, with 800 learners, was selected for model validation in this study.

Table 7. Performance analysis of the Boosted Tree ML model

Algorithms	RMSE	MSE	R ²	MAE
Boosted Tree 2	0.31667	0.10028	0.94328	0.27017
Boosted Tree 3	0.27647	0.076436	0.95677	0.25416
Boosted Tree 4	0.28913	0.083597	0.95272	0.25508
Boosted Tree 5	0.02724	0.000742	0.99958	0.021641
Boosted Tree 6	0.000154	2.37E-08	1	0.00013
Boosted Tree 7	3.84E-06	1.47E-11	1	3.21E-06

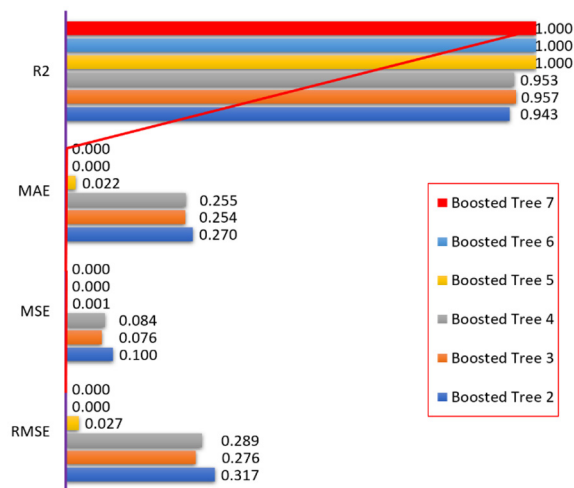


Figure 10. Statistical performance indicators of the Boosted Tree models

Analysis using the Bagged Tree model

Based on Table 5, the sequent depth of the hydraulic jump was analyzed using the Bagged Tree algorithm with different adjustments of the “leaf” and “learner” parameters. The analytical

results are presented in Table 8. As shown in Table 8 and Figure 12, the Bagged Tree model provided the best prediction performance when the number of learners was set to 250. However, the Bagged Tree 5 model, with 100 learners, also produced satisfactory prediction results, with statistical indicators that were nearly equivalent to those of the Bagged Tree 8 model using 250 learners. This result is consistent with recommendations on selecting an appropriate number of learners in several studies [27].

As shown in Tables 6, 7, and 8, the Tree 5 model, after hyperparameter adjustment, provides more accurate predictions than the baseline Bagged Tree and Boosted Tree models. This improvement is reflected in the statistical agreement between the predicted and measured values, as indicated by the R^2 , MSE, and RMSE metrics. However, when the optimized versions of all three algorithms are considered collectively, the Bagged Tree algorithm demonstrates superior predictive performance. This is particularly evident from the prediction results of the Bagged Tree 8 model, which achieves

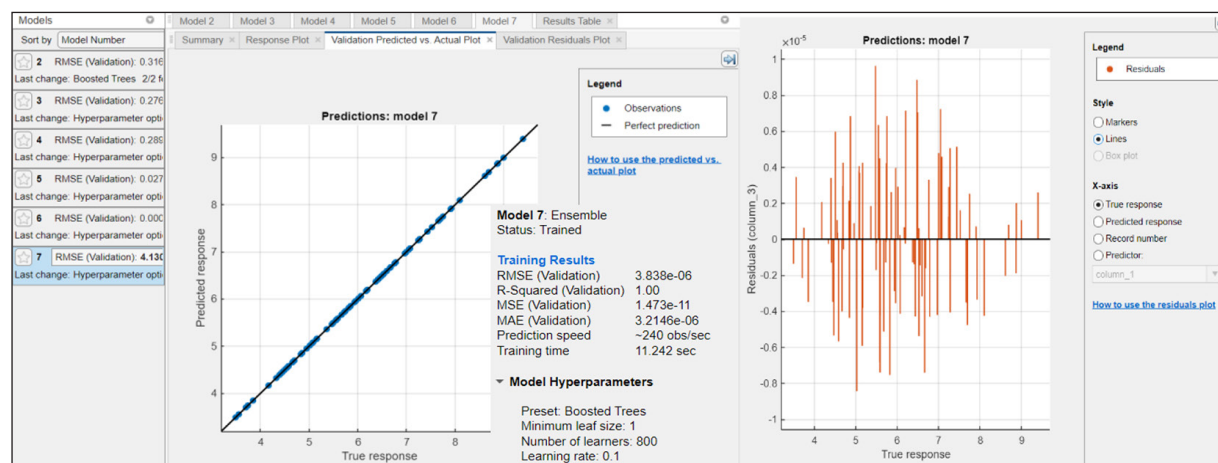


Figure 11. Model training results for the Boosted Tree model

Table 8. Analysis of the Bagged Tree ML model

Algorithms	RMSE	MSE	R ²	MAE
Bagged Tree 2	0.3714	0.1379	0.9220	0.2414
Bagged Tree 3	0.3183	0.1013	0.9427	0.2218
Bagged Tree 4	0.2414	0.0583	0.9671	0.1553
Bagged Tree 5	0.2213	0.0490	0.9723	0.1594
Bagged Tree 6	0.2509	0.0629	0.9644	0.1736
Bagged Tree 7	0.2360	0.0557	0.9685	0.1620
Bagged Tree 8	0.2193	0.0481	0.9728	0.1520
Bagged Tree 9	0.2345	0.0550	0.9689	0.1632

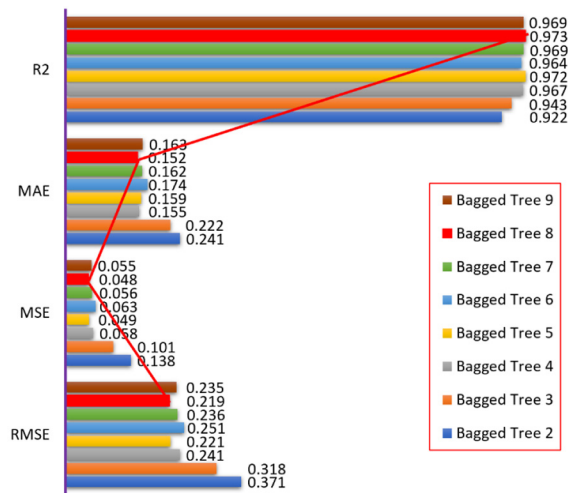


Figure 12. Comparison of statistical performance indicators among the Bagged Tree models

the best overall agreement with the experimental data and the most favorable evaluation metrics among the investigated models. Figure 13 indicates that the model performs well when the sequent depth ratio of the hydraulic jump (y_2/y_1) ranges from 4 to 8. When ($y_2/y_1 > 8$), the predicted values tend to be underestimated, whereas when ($y_2/y_1 < 4$), the predicted values tend to be overestimated. Compared with the empirical formulation proposed in the previous study [32], the Bagged Tree 8 model demonstrates higher predictive accuracy. Specifically, the empirical equation reported in [32] achieved a coefficient of determination of $R^2 = 0.94$ and an $RMSE = 0.327$, whereas Bagged Tree 8 yielded more favorable performance metrics, indicating closer agreement between the predicted and experimental values. This improvement is particularly

meaningful because empirical equations are generally derived by assuming a predefined functional relationship and may therefore have limited ability to represent the nonlinear behavior and complex interactions among the hydraulic parameters governing hydraulic-jump characteristics. This analysis shows that the model with 250 learners, namely the Bagged Tree 8 model, provides better predictive performance. Therefore, this model is proposed for further testing and evaluation of its effectiveness in predicting the hydraulic characteristics of hydraulic jumps in symmetric trapezoidal channels.

Analyzing the problem of predicting the jump length

The proposed ML models, namely Tree 5, Boosted Tree 7, and Bagged Tree 8, were applied to the above case to establish ML training models for predicting hydraulic jump length. The input variables consisted of Fr , y_2/y_1 , and M_1 , while the output variable was L_j/y_1 , as presented in Tables 2 and 3. After training the three proposed ML models, the statistical performance indicators used to evaluate the agreement between the predicted and experimental data are presented in Table 9.

Based on the results presented in Table 9, all proposed models exhibited strong predictive performance, with the coefficient R^2 of determination ranging from 0.97 to 1. In particular, the Boosted Tree 7 model exhibits outstanding predictive accuracy, as reflected by its statistical performance indicators, particularly the high R^2 value and the low MSE and RMSE values, with all performance metrics approaching their ideal values. The model achieved a very high coefficient

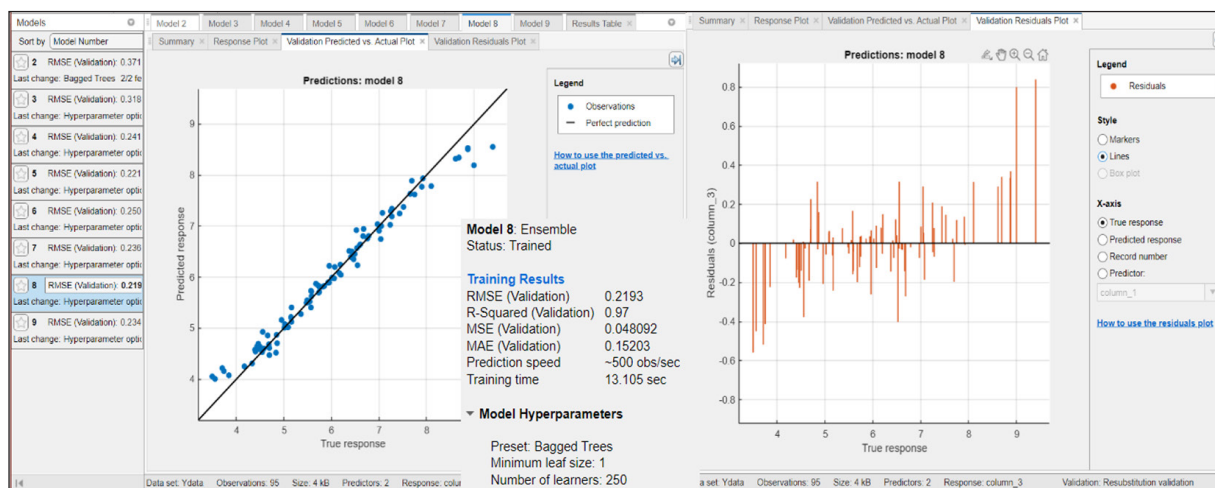


Figure 13. Model training results for the Bagged Tree 8 model

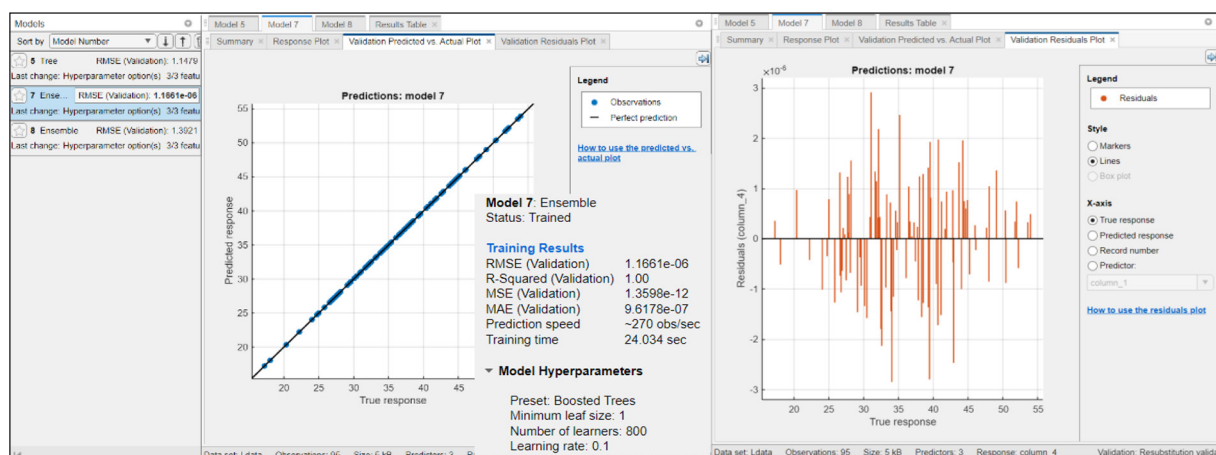


Figure 14. Analysis of the hydraulic jump length using the Boosted Tree 7

of determination ($R^2 \gg 1$), while other statistical indicators were nearly zero.

The training results obtained using the Boosted Tree 7 model were analyzed as follows: As shown in Figure 14, the Boosted Tree 7 model exhibits superior predictive performance compared with the other models, as evidenced by stronger agreement between predicted and measured values and more favorable statistical performance indicators. The predicted results were in very good agreement with the experimental data, with all analyzed points located close to the optimal line. The residual errors were less than $3 \cdot 10^{-5}$. Compared with the empirical formulation reported in [24], the Boosted Tree 7 model demonstrates higher predictive accuracy. The empirical equation yielded a coefficient of determination of $R^2 = 0.89$ and an RMSE = 0.113, whereas the Boosted Tree 7 model achieved more favorable statistical performance. This improvement indicates that the optimized Boosted Tree model is more capable of capturing the nonlinear relationships and complex interactions among the hydraulic parameters governing the investigated phenomenon. Unlike an empirical equation, which is generally constrained by a predefined functional form, the Boosted Tree approach iteratively combines multiple weak learners to reduce prediction errors and progressively improve model performance. Therefore, the superior performance of Boosted Tree 7 suggests that hyperparameter optimization enhances the model's ability to represent the underlying hydraulic behavior and provides a more accurate data-driven alternative for prediction within the investigated range.

Therefore, all ML models achieved good predictive performance; however, the Boosted Tree

7 model provided the best prediction accuracy for both hydraulic jump length and sequent depth.

VALIDATION OF THE PROPOSED BINARY TREE-BASED ALGORITHMS

Based on the proposed trained models, namely Tree 5, Boosted Tree 7, and Bagged Tree 8, the determination of the minimum leaf size and the limits of the number of nodes or learners plays an important role in the analysis results. When the minimum leaf size was set to 2 or higher, the predictive performance decreased significantly, as shown in Tables 6, 7, and 8. In contrast, the models achieved the best performance when the minimum leaf size was set to 1.

In this study, a common dataset was used to analyze two hydraulic jump characteristics, including sequent depth and hydraulic jump length. The results indicate that the Boosted Tree 7 model provided the optimal performance in analyzing hydraulic parameters, with a minimum leaf size of 1 and 800 learners. The proposed ML models were then validated using an independent dataset, as presented in Table 4. The validation performance of the trained models was evaluated using statistical indicators.

The validation results of the prediction performance of the ML models are presented in Tables 10 and 11.

As shown in Table 10, the Boosted Tree 7 model achieved the best predictive performance, with a prediction error of 3.6% for the testing dataset. The other models produced larger errors, with the maximum error reaching 8.1% for the

Table 9. Analysis of the hydraulic jump length using the proposed ML models

Algorithms	Statistical indicators			
	RMSE	MSE	R ²	MAE
Tree 5	1.1479	1.3177	0.9807	0.85596
Boosted Tree 7	1.17E-06	1.36E-12	1	9.62E-07
Bagged Tree 8	1.3921	1.9379	0.97161	0.95977

Table 10. Prediction of the hydraulic jump sequent depth using different models

Measured values		Predicted values					
b (m)	y_2/y_1	Tree 5		Boosted Tree 7		Bagged Tree 8	
		y_2/y_1	Error (%)	y_2/y_1	Error (%)	y_2/y_1	Error (%)
0.20	6.518	7.045	6.15	6.562	0.7	6.568	0.7
0.20	5.377	5.722	6.03	5.502	2.3	5.025	7.0
0.20	4.587	4.741	3.24	4.550	0.8	4.394	4.4
0.335	7.621	7.045	8.17	7.353	3.6	7.311	4.2
0.335	7.879	7.860	0.25	7.876	0.0	7.702	2.3
0.335	7.388	7.045	4.87	7.266	1.7	7.315	1.0
0.335	6.625	6.777	2.24	6.862	3.4	6.826	3.0
0.335	6.426	6.320	1.67	6.491	1.0	6.486	0.9
0.55	6.113	5.886	3.85	6.021	1.5	5.953	2.7
0.55	6.000	5.722	4.85	5.936	1.1	5.932	1.1
0.55	4.708	4.494	4.76	4.669	0.9	4.959	5.1
0.55	5.357	5.067	5.74	5.185	3.3	5.117	4.7
Max	0.406	7.86	8.17	7.88	3.64	7.70	7.00
Min	0.087	4.49	0.25	4.55	0.05	4.39	0.67

Table 11. Prediction of the hydraulic jump length using the testing dataset

Measured values		Predicted values					
b (m)	L_j/y_1	Tree 5		Boosted Tree 7		Bagged Tree 8	
		L_j/y_1	Error (%)	L_j/y_1	Error (%)	L_j/y_1	Error (%)
0.20	42.904	42.764	0.33	43.020	0.3	41.162	7.9
0.20	32.787	32.482	0.94	32.314	1.5	30.694	6.8
0.20	29.593	27.977	5.78	28.828	2.7	27.057	9.4
0.335	43.103	44.926	4.06	45.495	5.3	45.875	6.0
0.335	48.966	47.821	2.39	48.950	0.0	48.375	1.2
0.335	45.306	44.926	0.85	45.177	0.3	45.164	0.3
0.335	40.417	41.419	2.42	41.073	1.6	41.457	2.5
0.335	38.148	39.977	4.57	39.746	4.0	39.771	4.1
0.55	36.761	38.194	3.75	37.147	1.0	38.057	3.4
0.55	37.273	38.194	2.41	36.985	0.8	37.304	0.1
0.55	28.125	29.961	6.13	27.789	1.2	27.919	0.7
0.55	31.071	31.278	0.66	30.558	1.7	31.304	0.7
Max	48.966	47.82	6.13	48.95	5.26	48.38	9.4
Min	28.125	27.98	0.33	27.79	0.03	27.06	0.1

Tree 5 model. Due to differences in model structures, the maximum errors did not occur at the same analysis point. The three proposed models were then applied to the testing dataset presented in Table 3. The prediction results for the testing series are shown in Table 11. Based on Table 11, the error analysis shows that the Boosted Tree 7 model produced the best results, with the maximum difference between the predicted and measured values of hydraulic jump length being 5.2%.

The statistical indicators obtained from the comparison between the predicted data and the experimental data are presented in Table 12.

As shown in Table 12, the Boosted Tree 7 model produced the best statistical performance in predicting the hydraulic characteristics of the hydraulic jump.

As shown in Figures 15, 16, and 17, the Boosted Tree 7 model provided the best prediction performance for both the sequent depth and the hydraulic jump length, with maximum errors of 3.6% and 5.3%, respectively. The statistical indicators were also close to their ideal values. This can be attributed to the Boosted Tree algorithm, which iteratively combines the data to construct new trees and updates the weights of the

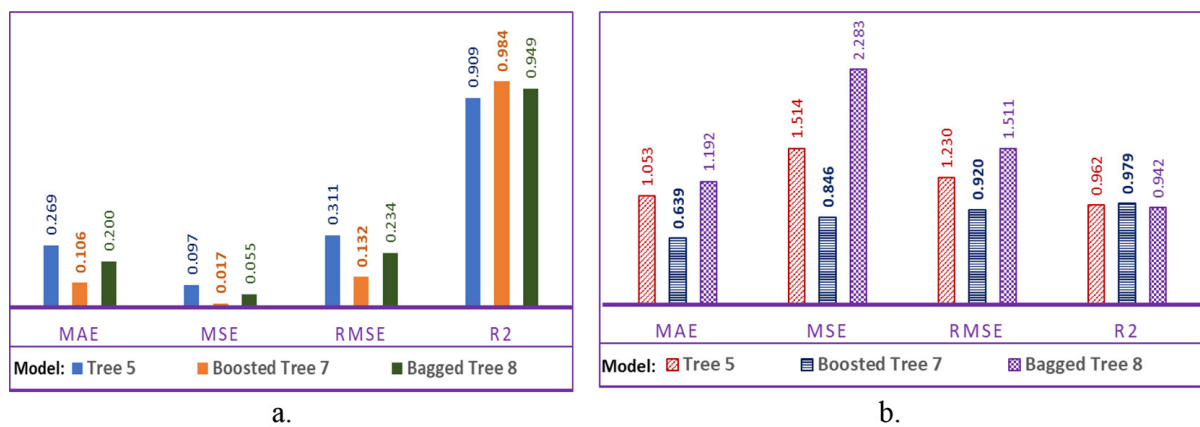


Figure 15. Statistical indicators obtained from the analysis using the testing dataset, a. Analysis of the sequent Depth, b. Analysis of the jump length

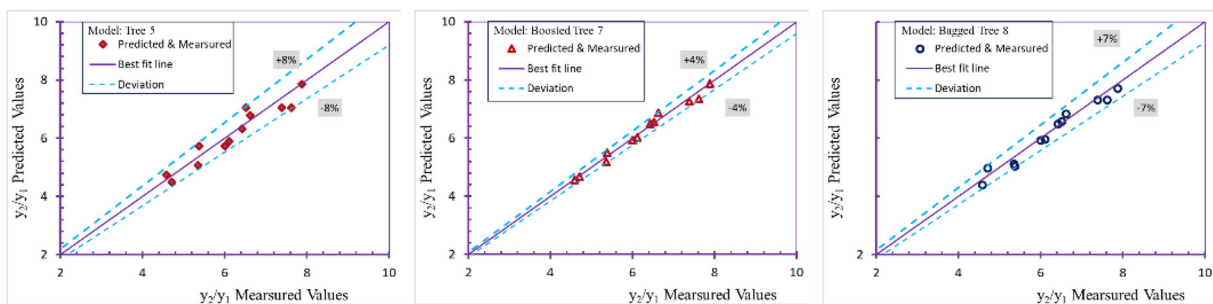


Figure 16. Comparison of measured and predicted sequent depth values in the testing dataset using ML models

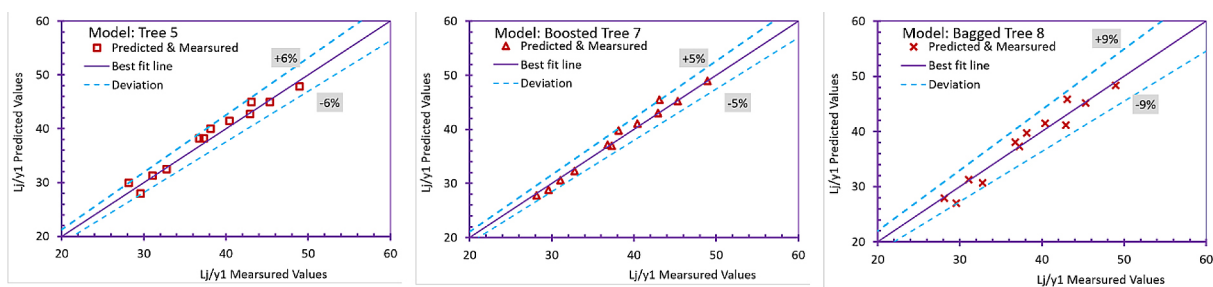


Figure 17. Comparison of the measured and predicted jump length values using ML models for the testing dataset

Table 12. Statistical indicators for model validation using the testing dataset

Models	MAE	MSE	RMSE	R ²
Sequent depth				
Tree 5	0.269	0.097	0.311	0.909
Boosted Tree 7	0.106	0.017	0.132	0.984
Bagged Tree 8	0.200	0.055	0.234	0.949
Hydraulic jump length				
Tree 5	1.053	1.514	1.230	0.962
Boosted Tree 7	0.639	0.846	0.920	0.979
Bagged Tree 8	1.192	2.283	1.511	0.942

trees, thereby allowing the model to be progressively optimized according to the statistical performance indicators. For the Boosted Tree model, when the number of learners was 500 or higher, the prediction results showed no significant difference. Therefore, this study recommends using a minimum leaf size of 1 and 800 learners to achieve good prediction performance.

The application of ML algorithms in predicting hydraulic parameters proved to be effective, with the results showing good convergence across the analyses. However, the determination of the objective function remains a complex issue, requiring a scientific approach that is consistent with the underlying physical laws. Previous studies have indicated that Buckingham's Pi theorem is a highly effective method for supporting the development of ML models. This theorem helps identify and analyze all influential variables involved in predicting the target variable.

CONCLUSIONS

This study demonstrated that ML models, including Tree, Boosted Tree, and Bagged Tree models, are suitable algorithms for predicting hydraulic parameters in hydraulic engineering. However, no single model can be considered optimal for all types of prediction problems; rather, each problem tends to have a model that is more appropriate than the others. In this study, the hydraulic characteristics of hydraulic jumps were predicted, and the results showed good computational performance, with relatively small prediction errors. The maximum prediction errors ranged from 3.6% to 9.4%. The statistical indicators were also close to their ideal values, namely the zero-error benchmark, while the coefficient of determination ranged from strong ($R^2 > 0.9$) to very strong ($R^2 \approx 1$). The results of this study

can be used to predict the downstream sequent depth and hydraulic-jump length for Froude numbers ranging from $Fr = 4.5$ to 9.0 , corresponding to stable hydraulic-jump conditions. These two parameters are essential for the hydraulic design and sizing of trapezoidal stilling basins located downstream of spillways, as they directly influence the required basin depth and length for effective energy dissipation. The main conclusions of this study are as follows:

1. This study contributes to the development of an automated approach for determining hydraulic-jump characteristics in trapezoidal stilling basins, supporting downstream flow management and the design of energy-dissipation structures.
2. The ML models demonstrated higher predictive accuracy than conventional approaches based on empirical equations, highlighting the potential of ML to capture the complex nonlinear relationships governing hydraulic characteristics..
3. The Buckingham Pi theorem provided a systematic basis for deriving dimensionless target and predictor variables for the hydraulic-jump models.
4. The hyperparameter-tuning results indicate that, within the scope of this study, a relatively simple Tree model can achieve higher predictive accuracy than more complex models such as Bagged Tree or Boosted Tree when appropriate hyperparameters are selected. This finding suggests that predictive performance depends not only on model complexity but also strongly on the proper selection and optimization of hyperparameters. Therefore, hyperparameter tuning plays a critical role in selecting suitable machine-learning models for predicting hydraulic characteristics of engineering structures.
5. The determination of optimal parameters for ML models is relatively complex. Each model type has different characteristics in terms of

parameter optimization; however, a common basis identified in this study is that the minimum leaf size should be set to 1. For the Boosted Tree model, a larger number of learners generally improves prediction performance. In contrast, the Bagged Tree model showed good performance when the number of learners ranged from 100 to 250.

6. The ML models implemented in MATLAB can automatically select the appropriate input variable for splitting the prediction regions at the decision nodes. This process is not dependent on the order of the input data columns during model training.
7. For a ML problem to achieve good predictive performance, it is necessary not only to identify all relevant influencing factors but also to use a sufficiently rich dataset. This study represents an initial attempt to analyze, propose, and test ML models for predicting hydraulic jump characteristics using a dataset of limited size. Therefore, for practical applications, the predicted results should be further compared with empirical formulas in order to select the most appropriate values.
8. Regression-based ML models have a wide range of applications. When implemented in computational environments such as MATLAB, these models can provide high efficiency, a convenient research workflow, and rapid testing of different calculation scenarios, thereby supporting the selection of the most suitable model. These findings indicate that ML algorithms are highly suitable for prediction studies in hydraulic engineering and have strong potential for practical application.

In practice, hydraulic characteristics vary from one study to another, and the suitability of prediction algorithms is therefore problem-dependent. Thus, no single algorithm can be conclusively regarded as universally applicable to the prediction of hydraulic characteristics. Instead, determining the optimal hyperparameters of each algorithm is necessary to enhance the predictive performance for specific hydraulic problems. This study provided an overview of binary tree-based algorithms and determined their optimal hyperparameters for the considered prediction task. Future research should extend the analysis to other ML algorithms to identify the most suitable modeling approach for predicting the hydraulic characteristics of hydraulic jumps.

REFERENCES

1. Barbierato E, Gatti A. The challenges of machine learning: A critical review. *Electronics* 2024;13:416. <https://doi.org/10.3390/electronics13020416>
2. Bzdok D, Altman N, Krzywinski M. Statistics versus machine learning. *Nat Methods* 2018;15:233–4. <https://doi.org/10.1038/nmeth.4642>
3. Grebovic M, Filipovic L, Katnic I, Vukotic M, Popovic T. Machine learning models for statistical analysis. *Int Arab J Inf Technol* 2023;20. <https://doi.org/10.34028/iajit/20/3A/8>
4. Brunton SL, Noack BR, Koumoutsakos P. Machine learning for fluid mechanics. *Annu Rev Fluid Mech* 2020;52:477–508. <https://doi.org/10.1146/annurev-fluid-010719-060214>
5. Sarker IH. Machine learning: Algorithms, real-world applications and research directions. *SN Comput Sci* 2021;2:160. <https://doi.org/10.1007/s42979-021-00592-x>
6. Lapeyre CJ, Cazard N, Roy PT, Ricci S, Zaoui F. Reconstruction of Hydraulic Data by Machine Learning. In: Gourbesville P, Caignaert G, editors. *Adv. Hydroinformatics*, Singapore: Springer; 2020, p. 701–15. https://doi.org/10.1007/978-981-15-5436-0_54
7. Xu Z, Lian J, Bin L, Hua K, Xu K, Chan HY. Water price prediction for increasing market efficiency using random forest regression: A case study in the Western United States. *Water* 2019;11:228. <https://doi.org/10.3390/w11020228>
8. Granata F, Gargano R, Tricarico C, Di Nunno F. Machine learning models in complex wastewater hydraulics problems. 5th IAHR Eur. Congr., Trento, Italy: 2018, p. 745–6. https://doi.org/10.3850/978-981-11-2731-1_183-cd
9. Hosseiny H, Nazari F, Smith V, Nataraj C. A framework for modeling flood depth using a hybrid of hydraulics and machine learning. *Sci Rep* 2020;10:8222. <https://doi.org/10.1038/s41598-020-65232-5>
10. Gupta G. Monitoring water distribution network using machine learning. Stockholm, Sweden: KTH Royal Institute Of Technology; 2017.
11. Mahtabi G, Chaplot B, Azamathulla HM, Pal M. Classification of hydraulic jump in rough beds. *Water* 2020;12:2249. <https://doi.org/10.3390/w12082249>
12. Karbasi M, Azamathulla HMd. GEP to predict characteristics of a hydraulic jump over a rough bed. *KSCE J Civ Eng* 2016;20:3006–11. <https://doi.org/10.1007/s12205-016-0821-x>
13. Khosravinia P, Sanikhani H, Abdi C. Predicting hydraulic jump length on rough beds using data-driven models. *J Rehabil Civ Eng* 2018;6:139–53. <https://doi.org/10.22075/jrce.2017.11047.1180>
14. Pasandideh I, Rajabi A, Yosefvand F, Shabanlou

- S. Modeling length of hydraulic jump on sloping rough bed using gene expression programming. *J AI Data Min* 2020;8:535–44. <https://doi.org/10.22044/jadm.2020.7444.1886>
15. Naseri M, Othman F. Determination of the length of hydraulic jumps using artificial neural networks. *Adv Eng Softw* 2012;48:27–31. <https://doi.org/10.1016/j.advengsoft.2012.01.003>
16. Ho HV. Determination of the surface roller length of hydraulic jumps in horizontal rectangular channels using the machine learning method. *Stoch Environ Res Risk Assess* 2024;38:2539–62. <https://doi.org/10.1007/s00477-024-02697-0>
17. Ziari M, Karami H, Ostadi A, Ghazvinian H. Simulation and prediction of hydraulic jump characteristics over expanding rough beds using FLOW-3D and soft computing techniques. *J Hydroinformatics* 2025;27:88–106. <https://doi.org/10.2166/hydro.2025.270>
18. Ty, T.C. Prediction of jump length, bed shear stress, and energy loss using machine learning models for a steady, free hydraulic jump on rough beds. *Journal of Materials and Construction* 2026; 16(1): 225-231. <https://doi.org/10.54772/jomc.01.2026.1168>
19. Elshaarawy MK, Hamed AK. Machine learning and interactive GUI for estimating roller length of hydraulic jumps. *Neural Comput Appl* 2025;37:4247–76. <https://doi.org/10.1007/s00521-024-10846-3>
20. Rehman A, Rehman H ur, Khan NM. Machine learning based prediction of hydraulic jump lengths in stilling basins for barrages and low-head hydraulic structures. *Mehran Univ Res J Eng Technol* 2026;45:47–61. <https://doi.org/10.22581/0320>
21. Breiman L, Friedman J, Olshen RA, Stone CJ. *Classification and Regression Trees*. New York: Chapman and Hall/CRC; 2017. <https://doi.org/10.1201/9781315139470>
22. Bagga PJ, Patel KM, Makhesana MA, Şirin Ş, Khanna N, Krolczyk GM, Pala AD, Chauhan KC. Machine vision-based gradient-boosted tree and support vector regression for tool life prediction in turning. *Int J Adv Manuf Technol* 2023;126:471–85. <https://doi.org/10.1007/s00170-023-11137-2>
23. Dimitriadis SI, Liparas D, Initiative for the ADN. How random is the random forest? Random forest algorithm on the service of structural imaging biomarkers for Alzheimer’s disease: from Alzheimer’s disease neuroimaging initiative (ADNI) database. *Neural Regen Res* 2018;13:962. <https://doi.org/10.4103/1673-5374.233433>
24. Ngoc NM, Cuong PH, Son TT, Nam NV, Phong NT. Experimental study of the hydraulic jump length in a smooth trapezoidal channel. *Sci Rev Eng Environ Sci SREES* 2022;31:63–76. <https://doi.org/10.22630/srees.2334>
25. Leo B. Bagging predictors 1996;24:123–40.
26. Breiman L. Heuristics of instability and stabilization in model selection. *Ann Stat* 1996;24:2350–83.
27. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer; 2017.
28. G.Jacoby W. *Data Theory and Dimensional Analysis*. SAGE Publications, Inc.; 1991. <https://doi.org/10.4135/9781412983860>
29. Ngoc NM, Cuong PH, Phong BHP. Prediction of the conjugate depth of the hydraulic jump in the trapezoidal channel using Random Forest regression. *J Mil Sci Technol* 2022;150–8. <https://doi.org/10.54939/1859-1043.j.mst.82.2022.150-158>
30. Wanoschek R, Hager WH. Hydraulic jump in trapezoidal channel. *J Hydraul Res* 1989;27:429–46. <https://doi.org/10.1080/00221688909499175>
31. Friedman JH. Greedy function approximation: A gradient boosting machine. *Ann Stat* 2001;29:1189–232. <https://doi.org/10.1214/aos/1013203451>
32. Ngoc, N.M. and Thang, T.N. A study of an empirical sequent depths equation of the hydraulic jump in a horizontal trapezoidal channel. *Engineering, Technology & Applied Science Research*. 2023; 13, 6, 12307–12313. <https://doi.org/10.48084/etasr.6431>