



APPLICATION OF MACHINE LEARNING MODELS TO PREDICT THE SURFACE ROLLER LENGTH OF HYDRAULIC JUMPS IN RECTANGULAR CHANNELS

Hung Viet Ho*

Faculty of Water Resources Engineering, Thuyloi University, 175 Tay Son Street, Hanoi, Viet Nam

ARTICLE INFO

TYPE: Research Article

Received: 17/01/2026

Revised: 07/10/2026

Accepted: 08/10/2026

Published online: 15/10/2026

<https://doi.org/10.47869/tcsj.77.8.1>

* *Corresponding author*

Email: hohung.thuyluc@tlu.edu.vn

Abstract. The surface roller length of a hydraulic jump is an important geometric characteristic that directly influences the design of energy dissipation structures. Accurate determination of this parameter is of great practical significance in water resources engineering. Existing empirical formulas yield inconsistent results. Therefore, this study proposes and evaluates the predictive performance of four decision-tree ensemble models: Categorical Boosting (CB), Extra Trees (ET), Gradient Boosting (GB), and Extreme Gradient Boosting (XGB). Unlike previous studies, the hyperparameters were optimized using Bayesian optimization combined with repeated 10-fold cross-validation, the contribution of each input variable was quantified using SHAP values, and the stability of the results was verified through 30 random data splits. The results show that the machine learning models produce generally lower predictive errors than empirical formulas. On the representative split, CB ranked first in all four metrics (NSE = 0.984; MAE = 1.695; RMSE = 2.207; MAPE = 7.13%). Over 30 splits, CB retained the lowest mean MAE, RMSE and MAPE. Hence, CB is the most suitable predictive method, complementing traditional empirical formulas in cases where multiple hydraulic factors need to be considered simultaneously.

Keywords: hydraulic jump, roller length, machine learning model, Categorical Boosting, SHAP.



ỨNG DỤNG CÁC MÔ HÌNH HỌC MÁY DỰ BÁO CHIỀU DÀI KHU XOÁY NƯỚC NHẢY TRONG KÊNH CHỮ NHẬT

Hồ Việt Hùng*

Khoa Kỹ thuật tài nguyên nước, Trường Đại học Thủy lợi, 175 Tây Sơn, Hà Nội, Việt Nam

THÔNG TIN BÀI BÁO

CHUYÊN MỤC: Công trình khoa học

Ngày nhận bài: 17/01/2026

Ngày nhận bài sửa: 07/10/2026

Ngày chấp nhận đăng: 08/10/2026

Ngày xuất bản Online: 15/10/2026

<https://doi.org/10.47869/tcsj.77.8.1>

* Tác giả liên hệ

Email: hohung.thuyluc@tlu.edu.vn

Tóm tắt. Chiều dài khu xoáy nước nhảy là một đặc trưng hình học quan trọng, ảnh hưởng trực tiếp đến thiết kế công trình tiêu năng. Việc xác định chính xác đại lượng này có ý nghĩa thực tiễn lớn trong lĩnh vực thủy lợi. Các công thức thực nghiệm hiện có cho kết quả không nhất quán. Vì vậy, nghiên cứu này đề xuất và đánh giá hiệu suất dự báo của bốn mô hình tổ hợp cây quyết định: Tăng cường phân loại (Categorical Boosting - CB), Cây bổ sung (Extra Trees - ET), Tăng cường độ dốc (Gradient Boosting - GB) và Tăng cường độ dốc cực đại (Extreme Gradient Boosting - XGB). Khác với các nghiên cứu trước, siêu tham số được tối ưu bằng thuật toán Bayes kết hợp kiểm định chéo lặp 10 nhóm, mức độ quan trọng của các biến đầu vào được lượng hóa bằng giá trị SHAP, và độ ổn định của kết quả được kiểm chứng qua 30 lần chia ngẫu nhiên dữ liệu. Kết quả cho thấy các mô hình học máy cho sai số dự báo nhìn chung thấp hơn các công thức thực nghiệm. Trên lần chia đại diện, CB dẫn đầu ở cả bốn chỉ số ($NSE = 0,984$; $MAE = 1,695$; $RMSE = 2,207$; $MAPE = 7,13\%$). Qua 30 lần chia, CB giữ MAE, RMSE và MAPE trung bình thấp nhất. Do đó, CB là phương pháp dự báo phù hợp nhất, bổ sung cho các công thức thực nghiệm truyền thống trong các trường hợp cần xét đồng thời nhiều yếu tố thủy lực.

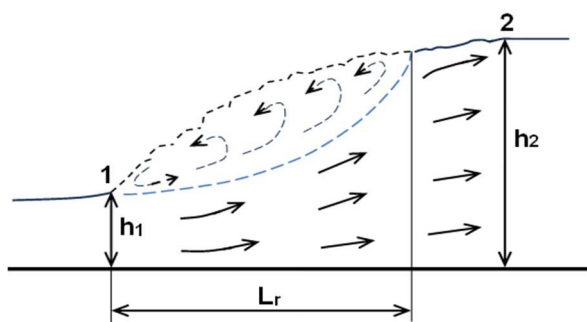
Từ khóa: nước nhảy, chiều dài khu xoáy, mô hình học máy, Categorical Boosting, SHAP.

@ 2026 Trường Đại học Giao thông vận tải

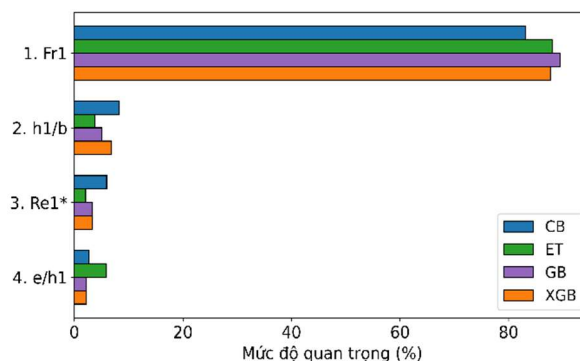
1. ĐẶT VẤN ĐỀ

Trong các công trình tháo nước như đập tràn, cống lộ thiên hay cống ngầm, dòng chảy xiết thường chuyển sang trạng thái chảy êm ở hạ lưu thông qua hiện tượng nước nhảy: theo chiều dòng chảy, vận tốc giảm mạnh, độ sâu tăng nhanh [1]. Khi tính toán thủy lực các công trình thủy lợi, giao thông, cần xác định độ sâu trước và sau nước nhảy, chiều dài khu xoáy và chiều dài toàn bộ nước nhảy [2]. Trong số đó, chiều dài khu xoáy L_r (Hình 1) là đại lượng trực quan nhất: có thể quan sát thực tế, đo đạc trong máng thí nghiệm và ảnh hưởng trực tiếp chiều dài bề tiêu năng cũng như quy mô công trình, vì vậy cần được tính toán với độ chính xác cao.

Trước đây, để ước tính L_r trong kênh chữ nhật nằm ngang, ta chủ yếu dựa vào các công thức kinh nghiệm với biến chính là số Froude và độ sâu trước nước nhảy [3, 4], hoặc kết hợp thêm hai độ sâu liên hiệp [5]. Các công thức tiêu biểu thuộc về Safranez, Pietrkowski, Schroder, Sarma và Newnam, và Hager [4]; ngoài ra Peterka [6] và Bretz [7] cũng công bố kết quả thí nghiệm trong máng kính. Do được xây dựng từ những điều kiện thí nghiệm khác nhau, mỗi công thức chỉ đúng trong một phạm vi nhất định, kết quả giữa các công thức không thống nhất và sai số nhìn chung còn lớn (xem Bảng 8). Thực tế đó đặt ra nhu cầu về một công cụ tính L_r có độ chính xác cao hơn và phản ánh được nhiều yếu tố thủy lực hơn.



Hình 1. Độ sâu nước nhảy h_1 , h_2 , chiều dài L_r .



Hình 2. Mức độ quan trọng của 4 biến đầu vào.

Ở giai đoạn hiện nay, các thuật toán học máy (ML) thuộc nhóm học có giám sát đang trở thành công cụ chủ lực cho các bài toán hồi quy trong thủy văn - thủy lực. Mạng nơ-ron nhân tạo (ANN) đã được dùng để ước lượng chiều dài nước nhảy từ năm 2012 [8]; đến năm 2021, nhóm mô hình ANFIS (adaptive neuro-fuzzy inference system), ANFIS-PSO (particle swarm optimization) và LASSO (least absolute shrinkage and selection operator) được áp dụng cho tỷ số hai độ sâu liên hiệp (h_1 , h_2) trong kênh chữ nhật [9]. Ở trong nước và trên thế giới, nhiều thuật toán khác như Rừng cây ngẫu nhiên (RF), Tăng cường độ dốc (GB), Cây bổ sung (ET), Máy Vector hỗ trợ (SVM), Tăng cường thích ứng (AdaBoost) và Tăng cường phân loại (CB) cũng lần lượt được thử nghiệm cho chiều dài nước nhảy [10, 11], chiều dài khu xoáy [12] và tỷ số hai độ sâu liên hiệp [13], với độ chính xác nhìn chung vượt các công thức thực nghiệm. Gần đây, hướng tiếp cận dựa trên dữ liệu tiếp tục được mở rộng cho các đặc trưng khác của nước nhảy, như hệ số dao động áp suất trong nước nhảy [14] hay tổn thất năng lượng dòng chảy qua đoạn thất hẹp [15], cho thấy đây vẫn là hướng nghiên cứu đang phát triển nhanh.

Trên cơ sở đó, nghiên cứu này xây dựng và kiểm định bốn mô hình ML gồm CB, ET, GB, và Tăng cường độ dốc cực đại (XGB) để ước tính chiều dài khu xoáy của nước nhảy tự do trong kênh hình chữ nhật nằm ngang, sử dụng Python 3.7 cùng các thư viện Numpy, Pandas và Matplotlib. Hiệu quả dự báo của bốn mô hình được đối chiếu lẫn nhau và so sánh với hai công thức thực nghiệm nhằm chọn ra mô hình phù hợp nhất.

Nghiên cứu này kế thừa nguồn dữ liệu thí nghiệm gốc của Hager và cộng sự [4] và Bretz [7], vốn đã được sử dụng trong hai công bố trước đây: nghiên cứu quốc tế đánh giá bảy mô hình ML [12] và nghiên cứu về tám cấu trúc mạng ANN [16]. So với hai công trình đó, bài báo này có các điểm phát triển sau, đồng thời là khoảng trống nghiên cứu được tập trung giải quyết: (i) hai thuật toán CB và XGB lần đầu được xây dựng như các mô hình dự báo L_r hoàn chỉnh, trong [16], hai thuật toán này mới chỉ được dùng để chấm điểm tầm quan trọng của biến đầu vào, chưa được kiểm định như mô hình dự báo; (ii) các siêu tham số được tối ưu bằng tối ưu hóa Bayes kết hợp kiểm định chéo lặp 10 nhóm, thay cho phương pháp dò lưới trong [12]; (iii) bốn mô hình tổ hợp cây quyết định được so sánh trên cùng một quy trình dữ liệu và đánh giá, cho phép rút ra kết luận công bằng về thuật toán tối ưu trong họ này; (iv) tầm quan trọng của các biến đầu vào được lượng hóa bằng giá trị SHAP thay cho điểm số dựa trên mức giảm tạp chất trong [16]; (v) khác biệt giữa các mô hình được kiểm chứng bằng kiểm định Wilcoxon dấu-hạng có hiệu chỉnh Holm trên 30 lần chia. Kết quả của nghiên cứu này vì vậy bổ sung và phát triển các công bố đã có, không thay thế hay lặp lại chúng.

2. PHƯƠNG PHÁP NGHIÊN CỨU VÀ CÁC DỮ LIỆU

2.1. Phương pháp phân tích thứ nguyên

Xét đoạn dòng chảy 1-2 bao trọn nước nhảy trên kênh chữ nhật nằm ngang (Hình 1) và viết phương trình động lượng cùng phương trình liên tục cho đoạn này [1, 2], có thể chỉ ra rằng chiều dài khu xoáy L_r chịu chỉ phối của tám đại lượng: độ sâu h_1 và vận tốc trung bình V_1 tại mặt cắt trước nước nhảy, chiều rộng b và độ nhám e của lòng dẫn, khối lượng riêng ρ và hệ số nhớt μ của nước, cùng gia tốc trọng trường g . Quan hệ tổng quát này được thể hiện như sau.

$$L_r = f(h_1, V_1, b, \rho, \mu, e, g) \quad (1)$$

Vì tám biến trên chứa đủ ba thứ nguyên cơ bản, Định lý π -Buckingham [17] cho phép rút gọn chúng về năm nhóm không thứ nguyên (hàm π) có quan hệ với nhau theo phương trình (2).

$$\pi_1 = \Phi(\pi_2, \pi_3, \pi_4, \pi_5) \quad (2)$$

Sau khi giải một hệ 5 phương trình cho kết quả là 5 hàm π dưới đây:

$$\pi_1 = L_r/h_1; \pi_2 = Fr_1; \pi_3 = Re_1^*; \pi_4 = e/h_1; \pi_5 = h_1/b \quad (3)$$

Trong đó: Fr_1 là số Froude trước nước nhảy tính theo (4); Re_1^* là số Reynold hiệu chỉnh tại mặt cắt trước nước nhảy tính theo (5) với ν là hệ số nhớt động học của nước; π_1 được gọi là chiều dài khu xoáy tương đối; π_4 là độ nhám tương đối; π_5 là độ sâu tương đối trước nước nhảy. Về ý nghĩa thủy lực, số Fr_1 đặc trưng cho tương quan giữa lực quán tính và trọng lực, Re_1^* phản ánh tương quan giữa lực quán tính và lực cản nhớt, π_4 biểu thị ảnh hưởng của lực cản nhám, còn π_5 phản ánh ảnh hưởng của đặc điểm hình học mặt cắt kênh đến sự phát triển của nước nhảy.

$$Fr_1 = \frac{V_1}{\sqrt{gh_1}} \quad (4)$$

$$Re_1^* = \frac{V_1 h_1}{\nu} \quad (5)$$

Tóm lại, L_r/h_1 là hàm số của nhiều yếu tố thủy lực khác nhau, được biểu thị bằng phương trình (6). Phương trình này cho thấy bốn đầu vào và một đầu ra của mô hình ML.

$$\frac{L_r}{h_1} = \Phi \left(Fr_1, Re_1^*, \frac{e}{h_1}, \frac{h_1}{b} \right) \quad (6)$$

2.2. Các công thức thực nghiệm so sánh với mô hình ML

Theo tổng kết của Hager và cộng sự [4], L_r chịu ảnh hưởng của kích thước và hình dạng lòng dẫn, điều kiện thượng-hạ lưu và đặc điểm thành bên; tuy nhiên cho đến nay đại lượng này vẫn chỉ được ước tính bằng công thức thực nghiệm vì chưa có lời giải lý thuyết trọn vẹn. Bảng 1 thống kê các công thức tính L_r chỉ từ số Fr_1 , không cần độ nhám lòng dẫn, tính nhớt của chất lỏng và chiều rộng kênh - đây là nhóm công thức được nghiên cứu này dùng để so sánh.

Bảng 1. Các công thức thực nghiệm tính chiều dài khu xoáy nước nhảy.

TT	Tác giả	Số Froude Fr_1	Công thức tính
1	Safranez (1929)	1,72 - 19,1	$\frac{L_r}{h_1} = 6Fr_1$
2	Pietrkowski (1932)	5,5 - 19,8	$\frac{L_r}{h_1} = 5,9Fr_1$
3	Schroder (1963)	1,83 - 9,93	$\frac{L_r}{h_1} = 40 \times erf \left\{ \frac{1}{16} (1 + 8Fr_1^2)^{1/2} - 3 \right\}$
4	Sarma - Newnham (1973)	1,21 - 3,79	$\frac{L_r}{h_1} = 6,73(Fr_1 - 1)$
5	Hager (1990)	2,5 - 8	$\frac{L_r}{h_1} = 8(Fr_1 - 1,5)$
		$Fr_1 > 8$	$\frac{L_r}{h_1} = -12 + 160 \tanh(Fr_1 / 20) : \frac{h_1}{b} < 0,1$
		$Fr_1 > 8$	$\frac{L_r}{h_1} = -12 + 100 \tanh(Fr_1 / 12,5) : 0,1 < \frac{h_1}{b} < 0,7$

Trong Bảng 1, hai công thức Safranez và Hager có miền số Froude áp dụng rộng nhất, còn công thức Sarma - Newnham bị bó hẹp nhất. Dù cấu trúc gần giống nhau, công thức Safranez thường cho trị số lớn hơn Pietrkowski; riêng công thức Hager bổ sung ảnh hưởng của bề rộng đáy kênh khi số Fr_1 vượt trên 8, nhờ đó linh hoạt hơn trong một số tình huống đặc thù.

Ưu thế của nhóm công thức này là gọn nhẹ, dễ áp dụng vào tính toán thực hành; nhược điểm là mỗi công thức gắn với một bộ điều kiện thí nghiệm riêng nên kết quả phân tán, đồng thời việc không thể hiện rõ ảnh hưởng của lực ma sát khiến trị số L_r tính ra có xu hướng lớn hơn thực tế, cần chú ý khi sử dụng.

2.3. Phương pháp mô hình toán ứng dụng ML

Bốn thuật toán khảo sát đều thuộc họ tổ hợp cây quyết định. GB, CB và XGB theo cơ chế tăng cường - các cây được huấn luyện nối tiếp, cây sau học phần dư của chuỗi cây trước và dự báo cuối cùng là tổng có trọng số của các cây [18]; ET theo cơ chế tổ hợp song song, các cây sinh độc lập rồi lấy trung bình. Dưới đây chỉ nêu đặc điểm phân biệt của từng thuật toán.

2.3.1. Mô hình Tăng cường độ dốc (GB)

GB là dạng cơ bản của nhóm tăng cường: mỗi cây khớp phần dư theo hướng ngược gradient của hàm mất mát, tốc độ học điều tiết mức đóng góp của từng cây [18]. Thuật toán không có số hạng chính quy hóa tường minh nên việc kiểm soát quá khớp phụ thuộc vào độ sâu cây, số mẫu tối thiểu tại nút và tỷ lệ lấy mẫu con.

2.3.2. Mô hình Tăng cường phân loại (CB)

CB khác GB ở kỹ thuật tăng cường có thứ tự (ordered boosting): gradient của mỗi mẫu được ước lượng từ mô hình chỉ học trên các mẫu đứng trước nó theo một hoán vị ngẫu nhiên, nhờ đó khử được thiên lệch vốn càng lớn khi bộ dữ liệu càng nhỏ [19].

2.3.3. Mô hình Cây bổ sung (ET)

ET sinh đồng thời nhiều cây không cắt tỉa, ngưỡng phân tách tại mỗi nút được chọn hoàn toàn ngẫu nhiên và mọi cây đều học trên toàn bộ mẫu gốc chứ không trên mẫu tái lập [20]. Phương sai vì vậy được khử bằng trung bình hóa nhiều cây chứ không bằng cắt tỉa từng cây.

2.3.4. Mô hình Tăng cường độ dốc cực đại (XGB)

XGB bổ sung vào hàm mục tiêu các số hạng chính quy hóa trên trọng số lá và số hạng phạt theo số lá, đồng thời lấy mẫu con theo cả hàng và cột ở mỗi vòng lặp [21]. Các cơ chế này hướng tới kiểm soát quá khớp trên dữ liệu lớn, nhiều chiều.

2.4. Phương pháp đánh giá sai số

Độ chính xác của các mô hình được định lượng qua năm chỉ số thống kê quen thuộc trong mô phỏng thủy lực: Sai số tuyệt đối trung bình (MAE), Sai số căn bậc hai trung bình (RMSE), Sai số phần trăm tuyệt đối trung bình (MAPE), Sai số tương đối (RE) và hệ số Nash-Sutcliffe (NSE). Bốn chỉ số sai số càng nhỏ (gần đến 0) và NSE càng tiến về 1 thì kết quả dự báo càng bám sát số liệu thí nghiệm. Công thức tính của từng chỉ số được nêu trong Bảng 2, giúp so sánh hiệu quả của các mô hình với nhau và với công thức thực nghiệm.

Bảng 2. Các chỉ số đánh giá và công thức tính toán.

Chỉ số	Công thức tính	Chỉ số	Công thức tính
NSE	$NSE = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$	RMSE	$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}}$
MAE	$MAE = \frac{\sum_{i=1}^n y_i - \hat{y}_i }{n}$	MAPE	$MAPE = \left(\frac{1}{n} \sum_{i=1}^n \frac{ y_i - \hat{y}_i }{y_i} \right) \times 100\%$
RE	$RE = \frac{y_i - \hat{y}_i}{y_i} \times 100\%$		

Trong các công thức trên, y_i và $\hat{y}_i$ là giá trị thực đo và giá trị ước tính; $\bar{y}$ là trị số trung bình của các giá trị thực đo; n là số lần quan sát.

2.5. Số liệu thí nghiệm

Nguồn dữ liệu của nghiên cứu này là năm chuỗi thí nghiệm nước nhảy do Hager và cộng sự [4, 22] cùng Bretz [7] công bố, thực hiện trên các máng chữ nhật nằm ngang có hai bề rộng 98 mm và 500 mm (Bảng 3); đáy máng và một thành bên bằng nhựa PVC, thành còn lại bằng

kính để tiện quan sát, nên độ nhám tuyệt đối của lòng dẫn chỉ khoảng 0,005 mm [22]. Nhiệt độ nước trong phòng thí nghiệm luôn duy trì ổn định ở 16-18° C, ứng với hệ số nhớt động học là $1,1 \times 10^{-6} \text{ m}^2/\text{s}$. Sau khi loại bỏ sai số và nhiễu, 156 bộ số liệu thí nghiệm được giữ lại, toàn bộ thuộc nước chảy tự do (không ngập), bao quát dải số Froude $Fr_1 = 2,26-15,96$ - phủ các trạng thái nước chảy thường gặp ở hạ lưu tràn, cống, công trình tiêu năng, phù hợp với điều kiện của các công trình thủy lợi Việt Nam.

Toàn bộ 156 mẫu được gộp chung rồi phân chia ngẫu nhiên bằng hàm *train_test_split* thành 109 mẫu huấn luyện (khoảng 70%) và 47 mẫu kiểm định (hơn 30%), với hạt giống ngẫu nhiên (*random_state=1*); 47 mẫu này được dùng thống nhất cho cả bốn mô hình để bảo đảm so sánh khách quan, không phụ thuộc ý chí chủ quan của người sử dụng. Việc tách tập kiểm định được thực hiện trước mọi thao tác tối ưu siêu tham số, và kiểm định chéo chỉ chạy trên 109 mẫu huấn luyện, nhờ đó loại trừ hiện tượng rò rỉ dữ liệu (data leakage). Ngoài ra, toàn bộ quy trình còn được lặp độc lập 30 lần với các hạt giống khác nhau để đánh giá độ ổn định của mô hình. Cần lưu ý hai điểm: thứ nhất, cách gộp chung dữ liệu của Hager và Bretz rồi chia ngẫu nhiên trong nghiên cứu này khác với quy trình của [12], nơi 12 mẫu Bretz được giữ riêng làm tập thử nghiệm độc lập - khác biệt này ảnh hưởng trực tiếp đến việc so sánh các chỉ số giữa hai nghiên cứu (xem mục 4). Thứ hai, phạm vi áp dụng của mô hình ML bị giới hạn trong miền dữ liệu huấn luyện: $Fr_1 = 2,26-15,96$; $h_1/b = 0,01-0,65$; $e/h_1 = 0,00004-0,001$; $Re_1^* = (0,05-4,55) \times 10^5$, ứng với lòng dẫn có độ nhám nhỏ; việc ngoại suy ra ngoài miền này (độ nhám lớn, điều kiện biên hạ lưu phức tạp) cần được kiểm chứng bổ sung trước khi áp dụng mô hình vào tính toán thiết kế. Thông số chi tiết của từng chuỗi thí nghiệm được trình bày trong Bảng 3.

Bảng 3. Các thông số cơ bản của các thí nghiệm.

Chiều rộng b (mm)	h_1/b [-]	e/h_1 [-]	$Re_1^* \times 10^{-5}$ [-]	Fr_1 [-]
500	0,012-0,078	0,0008-0,0001	0,09-0,95	3,33-15,31
98	0,1-0,65	0,0005-0,0001	0,09-2,00	2,45-10,87
500	0,055-0,109	0,0002-0,0001	0,52-3,12	2,26-11,12
500	0,01-0,072	0,001-0,0001	0,05-1,04	2,38-15,96
500	0,068-0,237	0,00015-0,00004	1,09-4,55	3,91-6,22

3. CÁC SIÊU THAM SỐ MÔ HÌNH VÀ ĐÁNH GIÁ ẢNH HƯỞNG CỦA CÁC BIẾN ĐẦU VÀO

3.1. Lựa chọn các siêu tham số của mô hình ML

Siêu tham số quyết định phần lớn chất lượng dự báo của mô hình ML. Khác với [12] sử dụng tìm kiếm lưới, nghiên cứu này chọn phương pháp tối ưu hóa Bayes (Bayesian optimization - BO). Quá trình tối ưu thực hiện 50 lần đánh giá mô hình (*random_state = 1*) với hàm mục tiêu là hệ số NSE, kết hợp kiểm định chéo lặp RepeatedKfold 10 nhóm $\times$ 3 lần và chỉ chạy trên 109 mẫu huấn luyện, nhằm bảo đảm tính ổn định và hạn chế quá khớp. Với các mô hình, giá trị mục tiêu tốt nhất xuất hiện ở lần đánh giá thứ 19 (CB), 26 (ET), 38 (GB), 31 (XGB) và không được cải thiện thêm trong các lần còn lại, cho thấy 50 lần đánh giá là đủ để quá trình hội tụ. Bảng 4 tổng hợp miền giá trị khảo sát và trị số tối ưu được chọn cho từng mô hình.

Các trị số tối ưu không theo một khuôn mẫu chung mà phản ánh đúng cơ chế khử phương sai của từng thuật toán. CB chọn cây nông nhất đi kèm chính quy hóa mạnh ($l2_leaf_reg=9,057$). ET chọn cây sâu, không cắt tỉa ($min_samples_leaf=1$; $min_samples_split=2$) - đúng như thiết kế của thuật toán, nơi phương sai được khử bằng trung bình hóa nhiều cây ngẫu nhiên hoàn toàn chứ không bằng cắt tỉa từng cây [20]. GB và XGB chọn độ sâu trung gian nhưng bù lại bằng ràng buộc số mẫu tối thiểu tại nút phân tách ($min_samples_split=15$, GB) và bằng chính quy hóa cùng lấy mẫu con ($gamma=0,342$; $reg_lambda=10$; $subsample=0,60$, XGB). Cần lưu ý rằng max_depth chỉ là cận trên của độ sâu chứ không phải độ sâu thực tế: với 109 mẫu huấn luyện, số nút lá của một cây không thể vượt quá 109, trong khi một cây phát triển đủ 8 mức về lý thuyết có tới $2^8 = 256$ nút lá. Do đó phần lớn các nhánh buộc phải dừng sớm hơn cận trên, và nguy cơ quá khớp được kiểm soát chủ yếu bởi các tham số chính quy hóa nêu trên chứ không bởi riêng cận trên độ sâu.

Bảng 4. Các siêu tham số chính của bốn mô hình học máy.

TT	Mô hình	Tham số chính	Khoảng giá trị	Giá trị tối ưu
1	CB	iterations	100 - 1000	587
		learning_rate	0,01 - 0,30 (log)	0,298
		depth	2 - 8	5
		$l2_leaf_reg$	1,0 - 10,0	9,057
		bagging_temperature	0,0 - 1,0	0,823
2	ET	n_estimators	100 - 1000	100
		max_depth	2 - 12	12
		min_samples_split	2 - 20	2
		min_samples_leaf	1 - 10	1
		bootstrap	cố định	False
3	GB	n_estimators	100 - 1000	765
		learning_rate	0,01 - 0,30 (log)	0,300
		max_depth	2 - 8	7
		min_samples_split	2 - 20	15
		min_samples_leaf	1 - 10	2
		subsample	0,6 - 1,0	1,000
4	XGB	n_estimators	100 - 1000	180
		learning_rate	0,01 - 0,30 (log)	0,119
		max_depth	2 - 8	8
		subsample	0,6 - 1,0	0,600
		gamma	0,0 - 1,0	0,342
		reg_lambda	0,1 - 10,0 (log)	10,000

3.2. Đánh giá ảnh hưởng của các biến đầu vào bằng giá trị SHAP

Điểm số dựa trên mức giảm tạp chất (impurity-based feature importance) - cách chấm điểm phổ biến trước đây - thiên vị các biến nhận nhiều giá trị phân biệt và không so sánh trực tiếp được giữa các thuật toán có cấu trúc cây khác nhau. Nghiên cứu này vì vậy sử dụng giá trị SHAP (SHapley Additive exPlanations) [23] - cách quy đổi đóng góp của từng biến dựa trên giá trị Shapley trong lý thuyết trò chơi hợp tác. Nhờ tính cộng tính và tính nhất quán của giá trị Shapley, đóng góp của một biến được đo bằng chính đơn vị của biến mục tiêu và xếp hạng thu

được so sánh được giữa bốn thuật toán. Giá trị SHAP được tính bằng bộ ước lượng TreeExplainer cho cả bốn mô hình. Để kết quả diễn giải phản ánh toàn bộ miền dữ liệu chứ không riêng một tập con, mỗi mô hình được huấn luyện lại trên toàn bộ 156 mẫu bằng chính bộ siêu tham số đã tối ưu ở Bảng 4. Thao tác này không tham gia vào việc lựa chọn siêu tham số cũng như đánh giá sai số ở mục 4, nên không tạo ra rò rỉ dữ liệu đối với các kết quả kiểm định. Với mỗi mô hình, trị tuyệt đối của giá trị SHAP được lấy trung bình trên toàn bộ mẫu rồi chuẩn hóa để tổng của bốn biến bằng 100%.

Bảng 5 và Hình 2 trình bày kết quả tính toán. Cả bốn mô hình đều cho rằng số Fr_1 chi phối áp đảo chiều dài khu xoáy, chiếm từ 83,10% (CB) đến 89,50% (GB), trung bình 87,10%. Kết quả này phù hợp với bản chất của nước nhảy trong dòng chảy hở, nơi tương quan giữa lực quán tính và trọng lực quyết định quy mô khu xoáy. Ba biến còn lại đóng vai trò thứ yếu nhưng không thể bỏ qua, tổng cộng khoảng 13%: độ sâu tương đối h_1/b đứng thứ hai (trung bình 5,98%), tiếp đến là số Reynolds hiệu chỉnh Re_1^* (3,67%) và độ nhám tương đối e/h_1 (3,26%). Ngoài trừ ET, ba mô hình còn lại đều có chung một đánh giá về thứ tự xếp hạng các biến đầu vào. Thứ tự này nhất quán với điều kiện của bộ số liệu: do máng thí nghiệm có độ nhám tuyệt đối chỉ khoảng 0,005 mm, e/h_1 biến thiên trong dải hẹp nhất trong bốn biến (Bảng 3) nên mang ít thông tin phân biệt nhất; ngược lại h_1/b thay đổi gần 65 lần giữa hai bề rộng máng, nên ảnh hưởng của hình học mặt cắt được các mô hình nhận diện rõ hơn ảnh hưởng của độ nhám. Đáng chú ý, mức phân tán của xếp hạng đầu vào giữa bốn mô hình khá nhỏ. Với biến e/h_1 , tỷ lệ giữa mô hình đánh giá cao nhất (ET: 5,93%) và thấp nhất (GB và XGB: 2,18%) chỉ khoảng 2,7 lần, so với hơn 40 lần theo cách chấm điểm dựa trên mức giảm tạp chất.

Bảng 5. Giá trị SHAP đánh giá mức độ quan trọng của các biến đầu vào.

TT	Đầu vào	CB (%)	ET (%)	GB (%)	XGB (%)	Trung bình (%)
1	Fr_1	83,10	88,11	89,50	87,68	87,10
2	h_1/b	8,20	3,85	5,07	6,78	5,98
3	Re_1^*	5,96	2,11	3,25	3,36	3,67
4	e/h_1	2,74	5,93	2,18	2,18	3,26

4. KẾT QUẢ VÀ THẢO LUẬN

4.1. Kết quả huấn luyện mô hình

Sử dụng các siêu tham số đã được lựa chọn, bốn mô hình ML được huấn luyện và kiểm định trên lần chia đại diện; kết quả trình bày ở Bảng 6. Các chỉ số ở cột "Huấn luyện (KĐC - kiểm định chéo)" không phải sai số khớp lại trên chính dữ liệu mô hình đã học, mà là trung bình điểm số của 30 phần kiểm định thuộc quy trình kiểm định chéo lặp RepeatedKFold 10 nhóm $\times$ 3 lần trên 109 mẫu huấn luyện. Ở giai đoạn này, MAE dao động từ 2,244 đến 2,565 và RMSE từ 3,518 đến 4,580, phản ánh mức độ sai lệch ít và độ phân tán thấp so với số liệu thí nghiệm.

Xét thứ hạng, ET đạt NSE cao nhất (0,974) và RMSE thấp nhất (3,518), thể hiện mức độ phù hợp cao giữa số liệu thí nghiệm và dự báo. CB có MAE và MAPE thấp nhất (2,244 và 6,65%) với NSE 0,971 ngang GB; XGB xếp cuối ở cả bốn chỉ số. Ba mô hình ET, CB và GB vì vậy có hiệu suất tương đương nhau và cùng vượt XGB.

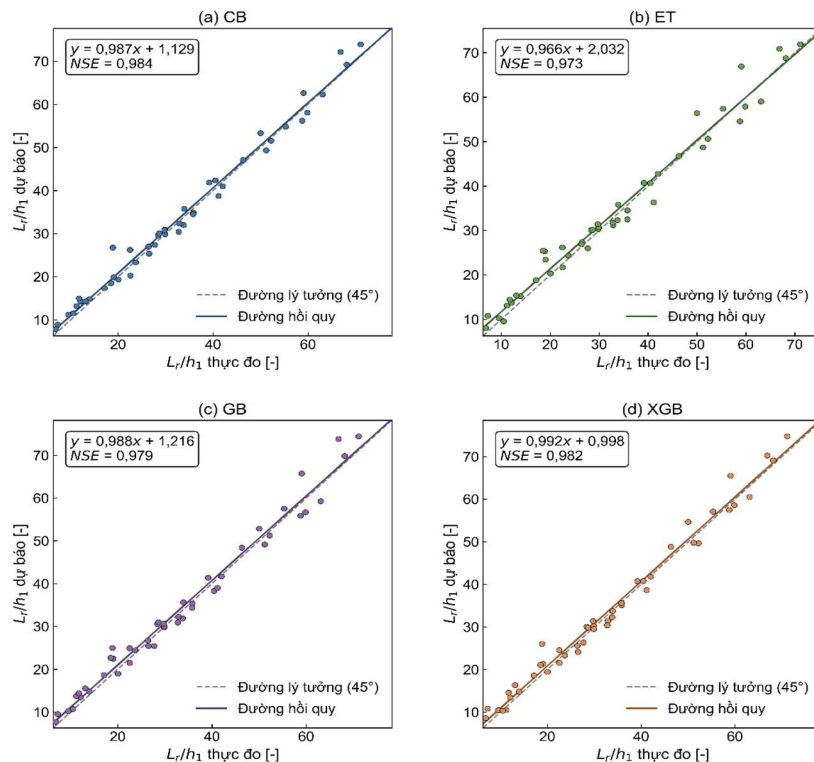
Bảng 6. Đánh giá kết quả huấn luyện và kiểm định bốn mô hình học máy.

TT	Mô hình	MAE [-]		RMSE [-]		NSE [-]		MAPE	
		Huấn luyện (KĐC)	Kiểm định	Huấn luyện (KĐC)	Kiểm định	Huấn luyện (KĐC)	Kiểm định	Huấn luyện (KĐC)	Kiểm định
1	CB	2,244	1,695	3,735	2,207	0,971	0,984	6,65%	7,13%
2	ET	2,393	2,232	3,518	2,899	0,974	0,973	6,66%	9,29%
3	GB	2,466	2,064	3,680	2,578	0,971	0,979	6,93%	8,03%
4	XGB	2,565	1,832	4,580	2,350	0,956	0,982	7,51%	7,88%

4.2. Kết quả kiểm định mô hình

Sau giai đoạn huấn luyện, các mô hình ML được kiểm định để đánh giá độ chính xác dự báo. Bảng 6 trình bày hiệu suất của các mô hình theo bốn chỉ số thống kê. Kết quả cho thấy cả bốn mô hình đều đạt NSE từ 0,973 đến 0,984, thể hiện mức độ phù hợp rất cao giữa giá trị dự báo và số liệu thí nghiệm. Về sai số tương đối, mô hình CB đạt giá trị MAPE thấp nhất (7,13%), tiếp đến là XGB (7,88%), GB (8,03%) và ET (9,29%). CB đồng thời có MAE (1,695) và RMSE (2,207) thấp nhất, tức dẫn đầu ở cả bốn chỉ số trên lần chia này.

Hình 3 minh họa mối quan hệ giữa số liệu thí nghiệm và giá trị dự báo của bốn mô hình: phần lớn các điểm phân bố sát đường lý tưởng (đường 45°), hệ số góc của đường hồi quy nằm trong khoảng 0,966 - 0,992 và hệ số chặn từ 0,998 đến 2,032, cho thấy tương quan tuyến tính mạnh và độ lệch hệ thống nhỏ. ET có hệ số góc nhỏ nhất (0,966) đi kèm hệ số chặn lớn nhất (2,032), tức có xu hướng dự báo thiên lớn ở vùng L_r/h_1 nhỏ - đây cũng là nguyên nhân khiến MAPE của mô hình này cao nhất.



Hình 3 (a, b, c, d). So sánh số liệu thí nghiệm với dự báo trong giai đoạn kiểm định mô hình.

Đối chiếu hai cột của Bảng 6 cho thấy MAE và RMSE trên tập kiểm định đều thấp hơn trên tập huấn luyện ở cả bốn mô hình, còn MAPE thì ngược lại - cao hơn ở tập kiểm định. Hiện tượng này không điển hình và cần được lý giải thận trọng. Trước hết, không có rò rỉ dữ liệu: tập kiểm định được tách riêng trước khi tinh chỉnh siêu tham số và kiểm định chéo chỉ chạy trên 109 mẫu huấn luyện (mục 2.5). Nguyên nhân chính nằm ở quy mô nhỏ của tập kiểm định (47 mẫu): tập này có cùng phân bố với dữ liệu huấn luyện nhưng chứa ít điểm biên khó dự báo hơn, nên các chỉ số tính theo giá trị tuyệt đối (MAE, RMSE) được cải thiện; ngược lại, tỷ lệ mẫu có L_T/h_1 nhỏ trong tập kiểm định lại cao hơn, mà sai số tương đối bị khuếch đại chính ở vùng giá trị nhỏ, nên MAPE tăng lên. Việc hai nhóm chỉ số biến thiên ngược chiều nhau cho thấy đây không phải là một lần chia thuận lợi một chiều. Ngoài ra, để chứng minh kết quả không phụ thuộc vào một lần chia dữ liệu, quá trình chia-huấn luyện-kiểm định được lặp độc lập 30 lần với các hạt giống ngẫu nhiên khác nhau ($random_state = 0 - 29$); Bảng 7 trình bày giá trị trung bình $\pm$ độ lệch chuẩn (σ) của các chỉ số trên tập kiểm định.

Hơn thế nữa, các chỉ số của lần chia đại diện được đối chiếu với phân bố của 30 lần chia ở Bảng 7. Hệ số NSE của cả bốn mô hình đều nằm trong phạm vi một độ lệch chuẩn quanh giá trị trung bình; MAE và RMSE tốt hơn trung bình từ $0,4\sigma$ đến $1,4\sigma$, trong khi MAPE lại kém hơn trung bình, riêng ET kém tới $+2,21\sigma$ - đây là lý do ET có MAPE cao bất thường (9,29%) ở lần chia này. Như vậy, kết quả ở Bảng 6 là của một mẫu đại diện phân bố chứ không phải một lần chia may mắn được lựa chọn.

Bảng 7. Trung bình $\pm$ độ lệch chuẩn của các chỉ số khi kiểm định qua 30 lần chia dữ liệu ngẫu nhiên.

TT	Mô hình	MAE [-]	RMSE [-]	NSE [-]	MAPE
1	CB	$2,206 \pm 0,362$	$3,515 \pm 1,342$	$0,968 \pm 0,018$	$6,81\% \pm 0,78\%$
2	ET	$2,379 \pm 0,378$	$3,538 \pm 1,141$	$0,968 \pm 0,015$	$7,43\% \pm 0,84\%$
3	GB	$2,428 \pm 0,429$	$3,863 \pm 1,304$	$0,961 \pm 0,019$	$7,35\% \pm 0,88\%$
4	XGB	$2,478 \pm 0,481$	$4,056 \pm 1,607$	$0,957 \pm 0,025$	$7,63\% \pm 0,99\%$

Bảng 7 cho thấy CB có MAE (2,206), RMSE (3,515) và MAPE (6,81%) trung bình thấp nhất, còn NSE trung bình của CB và ET bằng nhau (0,968). Xét thứ hạng trung bình gộp cả bốn chỉ số, trật tự là CB (1,70), ET (2,52), GB (2,72) và XGB (3,06); CB đứng đầu về MAE và về MAPE ở 18/30 lần chia, và không lần nào xếp cuối ở bất kỳ chỉ số nào.

Để tránh kết luận chỉ dựa trên trị số trung bình, khác biệt giữa các cặp mô hình được kiểm chứng bằng kiểm định Wilcoxon dấu-hạng trên 30 cặp giá trị [24], có hiệu chỉnh Holm [25] cho sáu phép so sánh ở mỗi chỉ số. Kết quả cho thấy CB vượt trội ET, GB và XGB một cách có ý nghĩa thống kê về MAE ($p \leq 4,4 \times 10^{-4}$) và MAPE ($p \leq 1,8 \times 10^{-4}$); về RMSE và NSE, CB tốt hơn GB ($p = 0,011$ và $0,010$) và XGB ($p = 2,6 \times 10^{-4}$ và $1,6 \times 10^{-4}$), trong khi không khác biệt có ý nghĩa so với ET ($p = 0,685$ và $0,761$). Ba cặp ET-GB, ET-XGB và GB-XGB cũng không khác biệt có ý nghĩa ở cả bốn chỉ số. Như vậy, CB là mô hình tốt nhất, đặc biệt về sai số tuyệt đối và tương đối; ở RMSE và NSE, CB và ET có hiệu suất tương đương về mặt thống kê. XGB có hiệu suất thấp nhất ở cả giai đoạn huấn luyện (kiểm định chéo) và trung bình 30 lần chia, đồng thời có độ phân tán lớn nhất, phù hợp với nhận định rằng cơ chế chính quy hóa [21] của XGB phát huy hiệu quả hơn trên dữ liệu lớn và nhiều chiều.

4.3. So sánh mô hình ML với công thức thực nghiệm và thảo luận

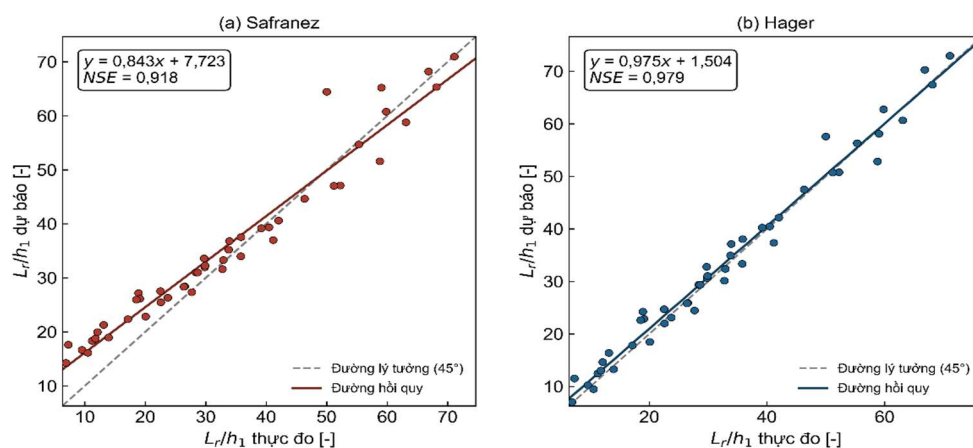
Nghiên cứu này so sánh khả năng dự báo chiều dài khu xoáy nước nhảy của các mô hình ML với hai công thức thực nghiệm. Hình 4 và Bảng 8 thể hiện rằng, công thức Hager có độ

chính xác cao hơn công thức Safranez ở tất cả các chỉ số đánh giá, đặc biệt là sai số MAPE thấp hơn đáng kể, chứng tỏ khả năng dự báo của công thức Hager tốt hơn.

Bảng 8. Kết quả kiểm định các công thức thực nghiệm.

TT	Công thức	MAE [-]	RMSE [-]	NSE [-]	MAPE
1	Safranez	4,043	5,076	0,918	22,86%
2	Hager	1,965	2,556	0,979	8,26%

Để làm rõ hơn đặc trưng sai số của các mô hình toán, Hình 5 trình bày biểu đồ tần suất sai số tương đối (RE) trong giai đoạn kiểm định của bốn mô hình ML. Cả bốn mô hình đều có phân bố lệch nhẹ về phía âm, tức có xu hướng dự báo thiên lớn, nhưng mức độ khác nhau rõ rệt. CB và XGB tập trung chặt nhất quanh giá trị không, với cột tần suất cao nhất chứa lần lượt 16 và 18 trong số 47 mẫu; GB có phân bố rộng và bẹt hơn; ET phân tán nhất, phù hợp với MAPE cao nhất trong bốn mô hình ở lần chia này. Đuôi âm của cả bốn mô hình kéo dài tới khoảng -45% đến -50% , ứng với một vài mẫu có L_r/h_1 nhỏ, nơi sai số tương đối bị khuếch đại. Để tiện đối chiếu, biểu đồ tần suất sai số tương đối của hai công thức thực nghiệm Safranez và Hager được trình bày trong Hình 6. Phân bố sai số của công thức Safranez lệch rõ về phía âm với đuôi kéo dài tới khoảng -145% (thiên lớn hệ thống), phù hợp với nhận định công thức này thường cho chiều dài khu xoáy lớn hơn giá trị thực đo; trong khi đó công thức Hager có phân bố khá cân đối quanh giá trị không nhưng vẫn phân tán rộng hơn đáng kể so với mô hình CB.

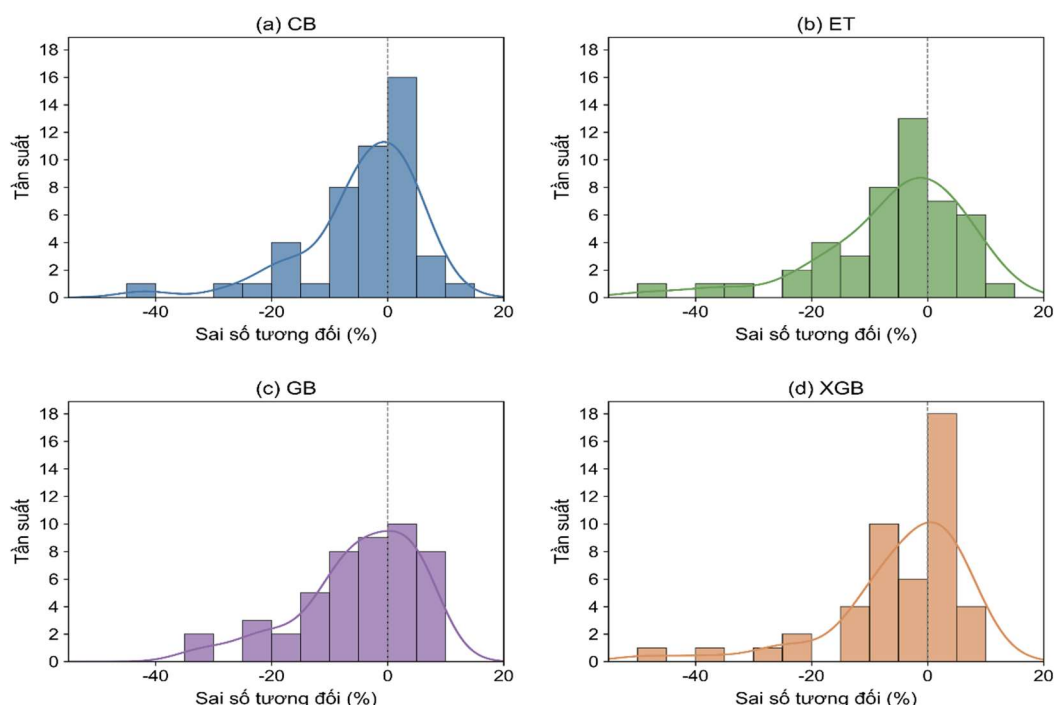


Hình 4 (a, b). Biểu đồ so sánh số liệu thí nghiệm với kết quả tính toán kiểm định công thức.

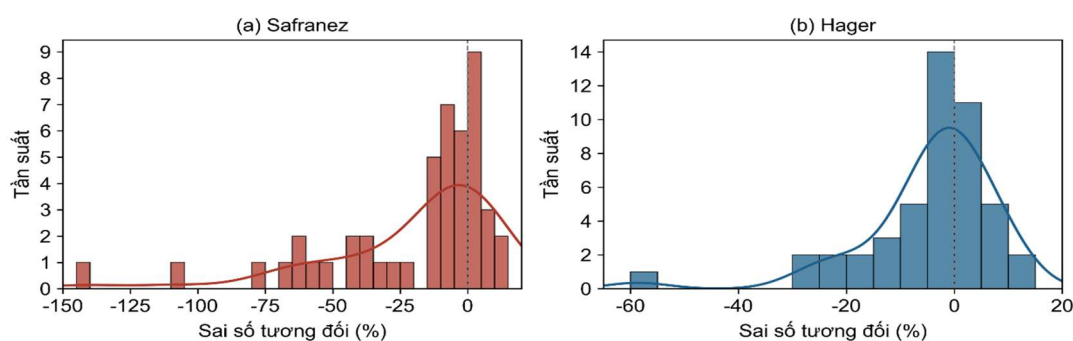
Bảng 6 và Bảng 8 cho thấy, mô hình CB vượt Hager ở cả bốn chỉ số, XGB cũng vượt ở cả bốn chỉ số, GB tương đương (NSE bằng nhau, MAPE thấp hơn nhưng MAE và RMSE cao hơn chút ít), riêng ET kém hơn Hager ở cả bốn chỉ số trên lần chia này. Kết quả bất lợi của ET mang tính cục bộ ở lần chia này, còn xét trung bình 30 lần thì MAPE của ET là 7,43%, vẫn thấp hơn giá trị 8,26% của Hager. Công thức Safranez kém xa cả hai nhóm (NSE = 0,918; MAPE = 22,86%), phù hợp với nhận xét ở mục 2.2 rằng công thức này cho trị số L_r thiên lớn.

Như vậy, lợi thế của phương pháp ML so với công thức thực nghiệm là có thật nhưng cần được thể hiện đúng mức. Mô hình tốt nhất (CB) giảm khoảng 14% ở cả ba chỉ số MAE, RMSE và MAPE so với công thức tốt nhất (Hager); nhưng không phải mọi mô hình ML đều vượt mọi công thức thực nghiệm trên mọi lần chia dữ liệu. Giá trị thực tế của cách tiếp cận ML nằm ở chỗ nó cho phép đưa đồng thời độ nhám lòng dẫn, độ nhót chất lỏng và bề rộng lòng dẫn vào quan hệ dự báo - những yếu tố mà nhóm công thức ở Bảng 1 không thể hiện - chứ không nằm ở một mức cải thiện sai số áp đảo.

So sánh với công bố [12] sử dụng cùng nguồn dữ liệu gốc, hệ số NSE của mô hình ET trong nghiên cứu này ($0,973$ ở lần chia đại diện và $0,968 \pm 0,015$ trung bình qua 30 lần chia) cao hơn giá trị $0,922$ mà [12] báo cáo ở giai đoạn thử nghiệm. Cần nhấn mạnh rằng hai con số này không đo trong cùng một điều kiện: ở [12], các mô hình được huấn luyện và hiệu chỉnh trên 144 mẫu của Hager rồi thử nghiệm trên 12 mẫu độc lập của Bretz - một phép thử ngoài phân bố khá khế hơn; còn nghiên cứu này gộp cả hai nguồn thành 156 mẫu rồi chia ngẫu nhiên, nên 47 mẫu kiểm định có cùng phân bố với dữ liệu huấn luyện. Giá trị bổ sung của bài báo so với [12] nằm ở chỗ: hai mô hình CB và XGB lần đầu được đánh giá như các mô hình dự báo L_r hoàn chỉnh, và CB được xác định là thuật toán tốt nhất trong họ cây tăng cường trên cùng một quy trình so sánh. So với cấu trúc ANN tốt nhất trên cùng bộ dữ liệu [16] ($NSE = 0,992$; $MAPE = 4,32\%$), mô hình CB có sai số kiểm định cao hơn. Đóng góp của bài báo này so với [16] vì vậy không nằm ở việc hạ thấp sai số, mà ở ba điểm khác: tầm quan trọng của từng biến đầu vào được lượng hóa bằng giá trị SHAP thay cho điểm số dựa trên mức giảm tạp chât; độ ổn định của kết quả được kiểm chứng qua 30 lần chia dữ liệu; và thứ hạng giữa các mô hình được kiểm định bằng thống kê thay vì chỉ so sánh trị số trung bình.



Hình 5 (a, b, c, d). Biểu đồ tần suất sai số tương đối trong giai đoạn kiểm định các mô hình ML.



Hình 6 (a, b). Biểu đồ tần suất sai số tương đối của hai công thức thực nghiệm.

Nhược điểm cố hữu của phương pháp ML là kết quả có thể dao động giữa các lần chạy do yếu tố ngẫu nhiên trong khởi tạo và chia mẫu, vì vậy mỗi mô hình cần được chạy lặp nhiều lần trước khi rút ra kết luận. Qua 30 lần chia, CB có độ lệch chuẩn nhỏ nhất ở MAE ($\pm 0,362$) và MAPE ($\pm 0,78\%$), còn ET nhỏ nhất ở RMSE ($\pm 1,141$) và NSE ($\pm 0,015$); hai mô hình này vì vậy là hai mô hình tái lập tốt nhất giữa các lần chạy. Ngược lại, XGB có độ lệch chuẩn lớn nhất ở cả bốn chỉ số, tức kém ổn định nhất.

5. KẾT LUẬN

Nghiên cứu này đánh giá hiệu suất của bốn mô hình ML trong việc dự báo chiều dài khu xoáy nước nhảy, với trọng tâm là độ chính xác. Mục tiêu chính của nghiên cứu là xây dựng, phân tích và so sánh hiệu quả của bốn mô hình ML gồm CB, ET, GB và XGB, nhằm xác định mô hình tối ưu. Các mô hình này cũng được so sánh với hai công thức thực nghiệm để đánh giá độ chính xác. Kết quả dự báo được phân tích cả định tính và định lượng nhằm làm rõ năng lực dự báo của từng mô hình. Những phát hiện chính của nghiên cứu bao gồm:

- Tối ưu hóa các siêu tham số đóng vai trò then chốt trong việc nâng cao hiệu suất mô hình. Phương pháp tối ưu Bayes kết hợp kiểm định chéo lặp cho phép dò tìm siêu tham số với chi phí tính toán thấp: chỉ 50 lần đánh giá mô hình đã đủ để quá trình hội tụ.
- Phân tích bằng giá trị SHAP cho thấy số Froude Fr_1 là yếu tố ảnh hưởng lớn nhất, tiếp theo là tỷ số h_1/b , số Reynolds hiệu chỉnh Re_1^* và độ nhám tương đối e/h_1 .
- Mô hình CB đạt hiệu suất cao nhất. Ở lần chia đại diện, CB dẫn đầu cả bốn chỉ số và giảm khoảng 14% sai số so với công thức Hager. Qua 30 lần chia, kiểm định Wilcoxon có hiệu chỉnh Holm xác nhận CB tốt hơn ET, GB và XGB về MAE và MAPE; tốt hơn GB và XGB về RMSE và NSE. Riêng với ET, hai mô hình tương đương ở hai chỉ số RMSE và NSE.
- Đối với bộ dữ liệu nhỏ, một lần chia dữ liệu chưa đủ để đánh giá độ tin cậy. Thứ hạng mô hình thay đổi giữa các lần chia, với độ lệch chuẩn của NSE lên tới $\pm 0,025$. Vì vậy, việc lựa chọn mô hình tốt nhất cần dựa trên nhiều lần chia dữ liệu kết hợp với kiểm định thống kê, thay vì một kết quả đơn lẻ.

Tóm lại, mô hình CB đã được chứng minh là giải pháp hiệu quả trong dự báo chiều dài khu xoáy nước nhảy. Mô hình này không nhằm thay thế các công thức thực nghiệm - vốn vẫn giữ giá trị nhờ tính đơn giản và phạm vi áp dụng đã được kiểm chứng - mà bổ sung một công cụ dự báo chính xác hơn cho các trường hợp cần xét đồng thời nhiều yếu tố thủy lực.

Bên cạnh các kết quả đạt được, nghiên cứu vẫn còn một số hạn chế: bộ dữ liệu kế thừa từ các thí nghiệm trong máng kính có độ nhám nhỏ, chưa bao quát lòng dẫn tự nhiên; các mô hình được so sánh đều thuộc họ cây tăng cường, chưa đối chiếu trực tiếp với các họ thuật toán khác (như ANN, SVM) trên cùng một quy trình đánh giá. Trong các nghiên cứu tiếp theo, nhóm tác giả sẽ mở rộng bộ số liệu thí nghiệm và so sánh liên họ thuật toán.

TÀI LIỆU THAM KHẢO

- [1]. W.H. Hager, Energy Dissipators and Hydraulic Jump, Kluwer Academic Publishers, Water Science and Technology Library, 8 (1992). <https://link.springer.com/book/10.1007/978-94-015-8048-9>
- [2]. W.H. Hager, Hydraulic jump, Energy dissipators, (1995) 43-60. <https://doi.org/10.1201/9780203757512-4>
- [3]. K.V.N. Sarma, D.A. Newnham, Surface profiles of hydraulic jump for Froude numbers less than four, Water Power, 25 (1973) 139-142.
- [4]. W.H. Hager, R. Bremen, N. Kawagoshi, Length of roller - Ressaut hydraulique classique: Longueur du rouleau, J Hydraul Res, 28 (1990) 591-608. <https://doi.org/10.1080/00221689009499048>

- [5]. H.E. Schulz, A.L.A. Simões, J.D. Nóbrega, Roller lengths, sequent depths, surface profiles for pre-design of dissipation basins, IWHS 2015, (2015) 81. <https://doi.org/10.13140/RG.2.1.3027.0889>
- [6]. A.J. Peterka, Hydraulic Design of Stilling Basins and Energy Dissipators, Monograph E, A Water Resources Technical Publication, USBR, 25 (1984) 240.
- [7]. N.V. Bretz, Hydraulic Jump Forced by Sill, Ecole Polytechnique Fédérale de Lausanne, 1988.
- [8]. M. Naseri, F. Othman, Determination of the length of hydraulic jumps using artificial neural networks, Adv Eng Softw, 48 (2012) 27-31. <https://doi.org/10.1016/j.advengsoft.2012.01.003>
- [9]. S. Baharvand, A. Jozaghi, R. Fatahi-Alkouhi, S. Karimzadeh, R. Nasiri, B. Lashkar-Ara, Comparative Study on the Machine Learning and Regression-Based Approaches to Predict the Hydraulic Jump Sequent Depth Ratio, Iran J Sci Technol - Trans Civ Eng, 45 (2021) 2719-2732. <https://doi.org/10.1007/s40996-020-00526-2>
- [10]. Lê Văn Nghi, Nguyễn Minh Ngọc, Ứng dụng thuật toán SVM dự báo chiều dài nước nhảy trên kênh hình thang cân, Tạp chí Khoa học và Công nghệ Thủy lợi, (2023) 43-49.
- [11]. Hồ Việt Hùng, Ứng dụng phương pháp học máy tính toán chiều dài nước nhảy trong kênh lăng trụ mặt cắt hình chữ nhật, Tạp chí Khoa học và Công nghệ Thủy lợi, 84 (2024) 76-86.
- [12]. H.V. Ho, Determination of the surface roller length of hydraulic jumps in horizontal rectangular channels using the machine learning method, Stoch Environ Res Risk Assess, 38 (2024) 2539-2562. <https://doi.org/10.1007/s00477-024-02697-0>
- [13]. Hồ Việt Hùng, Ứng dụng các thuật toán học máy xác định độ sâu sau nước nhảy trong kênh chữ nhật có xét đến ảnh hưởng của lực ma sát, Tạp chí Khoa học Giao thông vận tải, 76 (2025) 172-186. <https://doi.org/10.47869/tcsj.76.2.5>
- [14]. A. Azma, A.G.L. Borthwick, Y. Liu, D. Zhang, E. Afaridegan, Assessment of Data-Driven Models for Modeling the Coefficient of Pressure Fluctuations in a Hydraulic Jump, Int J Comput Intell Syst, 19 (2026) 126. <https://doi.org/10.1007/s44196-026-01259-2>
- [15]. H. Abbaszadeh, R. Daneshfaraz, V. Sume, J. Abraham, Experimental investigation and application of soft computing models for predicting flow energy loss in arc-shaped constrictions, AQUA - Water Infrastructure, Ecosystems and Society, 73 (2024) 637-661. <https://doi.org/10.2166/aqua.2024.010>
- [16]. Hồ Việt Hùng, Áp dụng các mô hình mạng nơ-ron nhân tạo (ANN) tính toán chiều dài khu xoáy của nước nhảy trong kênh chữ nhật nằm ngang, Tạp chí Khoa học Giao thông vận tải, 77 (2026) 254-268. <https://doi.org/10.47869/tcsj.77.3.3>
- [17]. E. Buckingham, On physically similar systems; illustrations of the use of dimensional equations, Phys Rev, 4 (1914) 345-376. <https://doi.org/10.1103/PhysRev.4.345>
- [18]. J.H. Friedman, Greedy function approximation: A gradient boosting machine, Ann Statist, 29 (2001) 1189-1232. <https://doi.org/10.1214/aos/1013203451>
- [19]. L. Prokhorenkova, G. Gusev, A. Vorobev, A.V. Dorogush, A. Gulin, CatBoost: Unbiased boosting with categorical features, Adv Neural Inf Process Syst, 31 (2018) 6638-6648.
- [20]. P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees, Mach Learn, 63 (2006) 3-42. <https://doi.org/10.1007/s10994-006-6226-1>
- [21]. T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, 2016, ACM, 785-794. <https://doi.org/10.1145/2939672.2939785>
- [22]. W.H. Hager, R. Bremen, Sequent depths: Le ressaut hydraulique classique: étude des hauteurs conjuguées, J Hydraul Res, 27 (1989) 565-585. <https://doi.org/10.1080/00221688909499111>
- [23]. S.M. Lundberg, S.I. Lee, A unified approach to interpreting model predictions, Adv Neural Inf Process Syst, 30 (2017) 4765-4774. <https://doi.org/10.48550/arXiv.1705.07874>
- [24]. F. Wilcoxon, Individual comparisons by ranking methods, Biometrics Bulletin, 1 (1945) 80-83. <https://doi.org/10.2307/3001968>
- [25]. S. Holm, A simple sequentially rejective multiple test procedure, Scand J Stat, 6 (1979) 65-70.