Tạp chí điện tử
Khoa học và Công nghệ Giao thông
Trang website: https://jstt.vn/index.php/vn

Compressive Strength Prediction and Parametric Sensitivity Analysis of Basalt Fiber-Reinforced Concrete: An Explainable Machine Learning Approach

Nguyen Thuy Anh*, Do Quoc Chien, Cao Nam Anh, Nguyen Manh Ha
*University of Transport Technology, Hanoi, Vietnam

Article info
Type of article:
Original research paper

DOI:
https://doi.org/10.58845/jstt.utt.2026.vn.6.1.48-63

***Corresponding author:**
Email address:
anhnt@utt.edu.vn

Received: 19/01/2026
Received in Revised Form: 08/03/2026
Accepted: 10/03/2026

Abstract: Compressive strength is the most critical mechanical property determining the load-bearing capacity of basalt fiber-reinforced concrete (BFRC). Determining this parameter through traditional experimental methods is time-consuming, costly, and labor-intensive. This paper applies three machine learning algorithms, including Random Forest, HistGradientBoosting, and CatBoost, to predict the compressive strength of BFRC using a dataset of 267 samples. Hyperparameter optimization is performed via grid search, and the results demonstrate that all three models achieve high reliability. Among them, Random Forest exhibits the best stability and generalization capability with the lowest RMSE (2.05 MPa). Additionally, SHAP analysis is integrated to elucidate the influence of input variables. The results indicate that cement content, silica fume, and fine aggregate are the most influential variables on compressive strength, with primarily positive effects from cement and silica fume, while water content and basalt fiber typically exhibit negative impacts if they exceed optimal thresholds. This research provides a reliable decision-support tool for rapid material property prediction, assisting in the practical optimization of mix designs.

Keywords: Basalt fiber-reinforced concrete, compressive strength, machine learning

Tạp chí điện tử
Khoa học và Công nghệ Giao thông
Trang website: https://jstt.vn/index.php/vn

Dự báo cường độ nén và phân tích ảnh hưởng thành phần cấp phối bê tông sợi basalt: Tiếp cận bằng học máy giải thích

Nguyễn Thùy Anh*, Đỗ Quốc Chiến, Cao Nam Anh, Nguyễn Mạnh Hà
*Trường Đại học Công nghệ Giao thông vận tải, Hà Nội, Việt Nam

Thông tin bài viết
Dạng bài viết:
Bài báo nghiên cứu

DOI:
https://doi.org/10.58845/jstt.utt.2026.vn.6.1.48-63

***Tác giả liên hệ:**
Địa chỉ Email:
anhnt@utt.edu.vn

Ngày nộp bài: 19/01/2026
Ngày nộp bài sửa: 08/03/2026
Ngày chấp nhận: 10/03/2026

Tóm tắt: Cường độ nén là chỉ tiêu cơ lý quan trọng nhất quyết định khả năng chịu lực của bê tông sợi basalt (BFRC). Việc xác định chỉ tiêu này bằng thí nghiệm nén mẫu truyền thống tốn kém nhiều thời gian, chi phí và nhân lực. Bài báo ứng dụng ba thuật toán học máy gồm Random Forest, HistGradientBoosting và CatBoost để dự báo cường độ nén BFRC trên tập dữ liệu 267 mẫu. Thông qua kỹ thuật tìm kiếm lưới để tối ưu siêu tham số, kết quả cho thấy cả ba mô hình đều đạt độ tin cậy cao. Trong đó, Random Forest thể hiện tính ổn định và khả năng tổng quát hóa tốt nhất với sai số RMSE thấp nhất (2.05 MPa). Bên cạnh đó, phân tích SHAP được tích hợp để làm rõ mức độ ảnh hưởng của các biến đầu vào. Kết quả chỉ ra rằng hàm lượng xi măng, muội silic và cốt liệu mịn là các biến có ảnh hưởng lớn nhất đến cường độ nén, với hướng tác động chủ yếu tích cực từ xi măng và muội silic, trong khi hàm lượng nước và sợi basalt thường mang tính chất tiêu cực nếu vượt ngưỡng tối ưu. Nghiên cứu cung cấp một công cụ hỗ trợ tin cậy cho việc dự báo nhanh tính chất vật liệu, hỗ trợ tối ưu hóa cấp phối trong thực tiễn.
Từ khóa: Bê tông sợi basalt, cường độ nén, học máy.

1. Giới thiệu

Bê tông là vật liệu được sử dụng rộng rãi trong ngành xây dựng do tính bền vững, khả năng linh hoạt và chi phí sản xuất hợp lý [1]. Tuy nhiên, bê tông thông thường tồn tại hạn chế về cường độ kéo thấp và tính giòn, dẫn đến nguy cơ nứt vỡ dưới tải trọng động hoặc kéo [2]. Để khắc phục những hạn chế này, việc bổ sung sợi vào hỗn hợp bê tông đã được đề xuất như một giải pháp cải thiện tính chất cơ học [3,4]. Trong số các loại sợi, sợi basalt được sản xuất từ đá bazan núi lửa, nổi bật nhờ tính chất thân thiện với môi trường, khả năng chống ăn mòn cao và chi phí thấp hơn so với sợi thủy tinh hoặc sợi carbon [5]. Do đó, bê tông sợi basalt (Basalt fiber reinforced concrete - BFRC) đã nhận được sự quan tâm ngày càng tăng, với các nghiên cứu cho thấy nó cải thiện cường độ kéo, khả năng chống nứt và độ bền tổng thể so với bê tông thông thường [6,7]. Cường độ nén là một trong những chỉ số cơ học quan trọng nhất của BFRC, quyết định khả năng chịu tải trọng chính trong các công trình xây dựng.

Các nghiên cứu thực nghiệm trước đây đã chỉ ra rằng việc bổ sung sợi basalt với liều lượng và chiều dài phù hợp có thể tăng cường độ nén của bê tông, đồng thời cải thiện hành vi ứng suất-biến dạng [7–9] . Tuy nhiên, phương pháp xác định cường độ nén thông qua thí nghiệm truyền thống, chẳng hạn như thử nén sau 28 ngày bảo dưỡng, đòi hỏi thời gian dài, tiêu tốn chi phí nguyên vật liệu

và nhân lực [10]. Bên cạnh đó, mối quan hệ giữa các thành phần cấp phối (hàm lượng xi măng, nước, phụ gia khoáng, sợi Basalt...) và cường độ nén đầu ra thường mang tính phi tuyến tính phức tạp [11]. Các công thức thực nghiệm hoặc bán thực nghiệm truyền thống thường chỉ đúng trong phạm vi hẹp của dữ liệu khảo sát và khó đạt độ chính xác cao khi áp dụng cho các cấp phối mới. Những hạn chế này dẫn đến nhu cầu phát triển các phương pháp dự đoán thay thế, nhằm giảm thiểu số lượng thí nghiệm và tối ưu hóa thiết kế hỗn hợp BFRC.

Những năm gần đây, sự phát triển của trí tuệ nhân tạo và kỹ thuật học máy (Machine Learning - ML) đã mở ra hướng tiếp cận mới trong việc dự báo tính chất vật liệu xây dựng. Nhiều nghiên cứu đã chỉ ra rằng các thuật toán học máy có khả năng mô hình hóa các mối quan hệ phi tuyến phức tạp mà không cần giả định trước về dạng hàm số học [12–14]. Các mô hình như Mạng nơ-ron nhân tạo (ANN) hay Máy vector hỗ trợ (SVM) đã được áp dụng để dự báo cường độ bê tông. Gần đây, các thuật toán học máy dạng tổ hợp (Ensemble Learning) như Random Forest hay Boosting (Gradient Boosting, CatBoost) đã chứng minh tính ưu việt trong bài toán hồi quy dữ liệu dạng bảng [15,16]. Tuy nhiên, các nghiên cứu áp dụng cụ thể cho bê tông cốt sợi Basalt tại Việt Nam còn hạn chế, đặc biệt là việc giải thích cơ chế hoạt động của mô hình "hộp đen".

Xuất phát từ thực tế trên, nghiên cứu này tập trung ứng dụng và so sánh hiệu quả của ba thuật toán học máy gồm Random Forest, HistGradientBoosting và CatBoost để xây dựng mô hình dự báo cường độ nén của bê tông cốt sợi Basalt. Dựa trên bộ dữ liệu thực nghiệm thu thập được, nghiên cứu nhằm mục tiêu tìm ra mô hình dự báo tối ưu, hỗ trợ các kỹ sư và nhà nghiên cứu xác định nhanh cường độ bê tông, giảm thiểu khối lượng thí nghiệm thử dần và tối ưu hóa công tác thiết kế cấp phối.

2. Cơ sở dữ liệu và phương pháp nghiên cứu

2.1. Mô tả bộ dữ liệu

Để xây dựng và kiểm chứng các mô hình học máy, nghiên cứu sử dụng bộ dữ liệu thực nghiệm bao gồm 267 mẫu BFRC được thu thập trong nghiên cứu của Wang và cộng sự [17]. Dữ liệu bao gồm các thành phần cấp phối cơ bản và các thông số liên quan đến sợi Basalt. Mỗi mẫu dữ liệu được đặc trưng bởi 10 biến đầu vào bao gồm ba nhóm chính: (1) Vật liệu kết dính (Xi măng và các phụ gia khoáng), (2) Cốt liệu và nước, (3) Đặc trưng sợi Basalt, cùng với 1 biến đầu ra là cường độ nén ở tuổi 28 ngày. Sợi Basalt được sử dụng trong bộ dữ liệu là dạng sợi cắt ngắn được phân tán ngẫu nhiên trong nền bê tông. Đây là loại vật liệu gia cường phổ biến, giúp cải thiện đáng kể tính dẻo và khả năng chống nứt cho hỗn hợp bê tông. Về cơ chế ảnh hưởng đến cường độ nén, sợi basalt không trực tiếp tăng khả năng chịu tải nén như trong trường hợp chịu kéo hoặc uốn, mà chủ yếu kiểm soát sự hình thành vi nứt và lan truyền nứt trong ma trận bê tông, cải thiện độ dẻo và độ bền tổng thể [8,9]. Tuy nhiên, khi hàm lượng sợi tăng cao (vượt 0.2–0.3%), hiện tượng vón cục hoặc phân tán không đồng đều có thể xảy ra, dẫn đến giảm mật độ đóng gói hạt và tạo ra các khuyết tật rỗng trong cấu trúc, từ đó làm suy giảm cường độ nén thay vì chỉ "vượt ngưỡng tối ưu". Thống kê mô tả về giải dữ liệu, giá trị trung bình và độ lệch chuẩn của các biến được trình bày chi tiết tại Bảng 1. Để trực quan hóa sự phân bố dữ liệu, Hình 1 thể hiện biểu đồ hộp của tất cả các thông số đầu vào và đầu ra. Giá trị cường độ nén trong tập dữ liệu nghiên cứu trải rộng từ 23.3 MPa đến 69.9 MPa; xác định phạm vi áp dụng hữu hiệu của các mô hình đề xuất. Các hộp dữ liệu trên biểu đồ minh họa độ phân tán rộng của các biến thành phần, phản ánh tính đa dạng của các cấp phối bê tông được khảo sát. Trong biểu đồ hộp (Hình 1), các điểm đỏ đại diện cho các giá trị ngoại lai, phản ánh sự biến thiên đặc biệt trong cấp phối thực nghiệm, chẳng hạn như hàm lượng muội silic hoặc nước cao bất thường có thể xuất hiện do các điều kiện thử nghiệm đặc thù. Những giá trị ngoại lai này này không bị loại bỏ để giữ tính đa dạng của dữ liệu, nhưng được kiểm soát trong quá trình huấn luyện mô hình để tránh ảnh hưởng tiêu cực đến độ chính

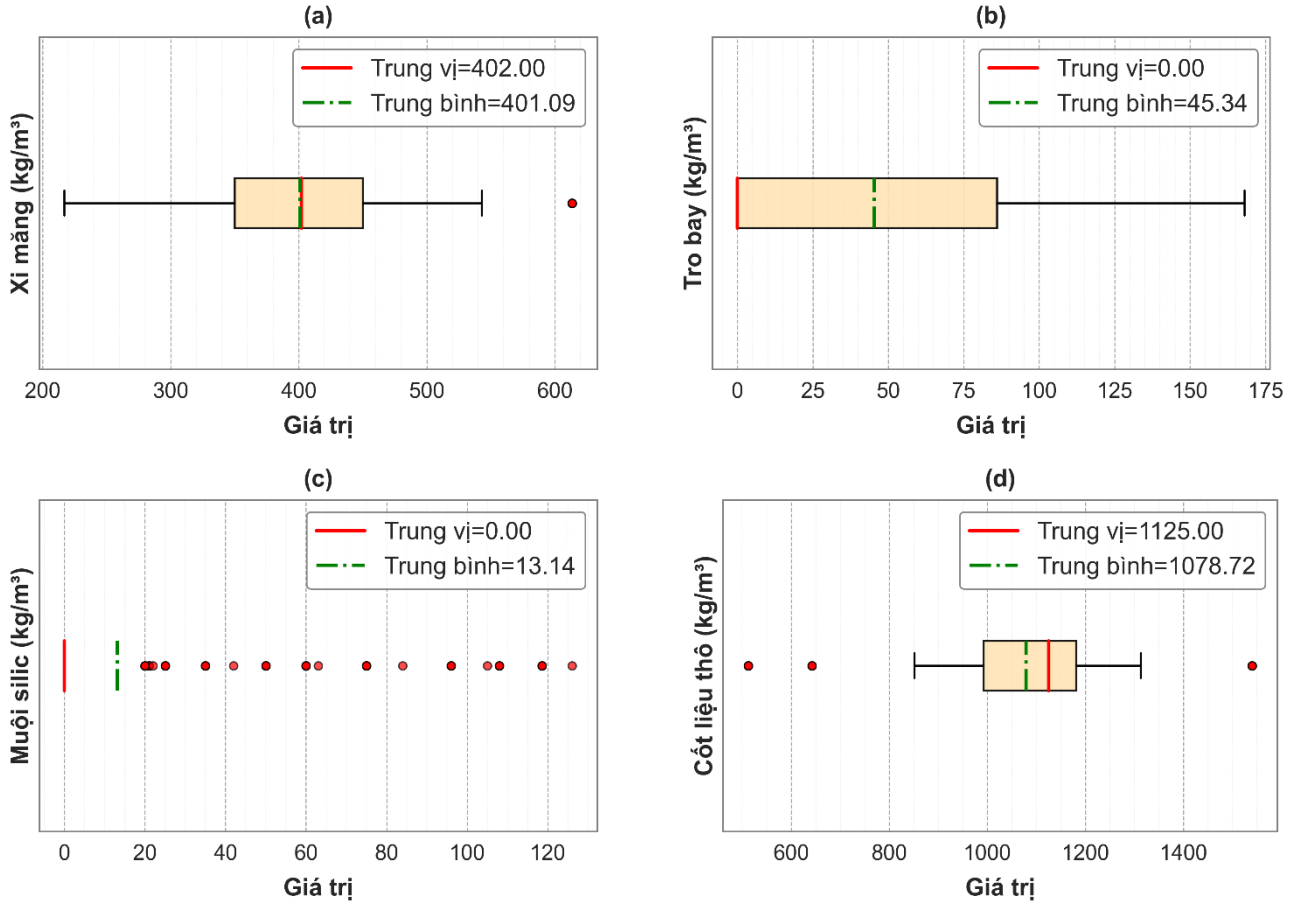
xác dự báo, phù hợp với các nghiên cứu về dữ liệu vật liệu xây dựng nơi biến thiên thực tế là phổ biến [17,18].

Tập dữ liệu được phân chia ngẫu nhiên thành tập huấn luyện (70%) và tập kiểm tra (30%). Quá trình tối ưu hóa siêu tham số sử dụng kỹ thuật kiểm chứng chéo 5 lần trên tập huấn luyện để đảm bảo độ tin cậy. Tập kiểm tra được giữ nguyên vẹn và chỉ được sử dụng một lần duy nhất sau khi quá trình huấn luyện hoàn tất, nhằm cung cấp một đánh giá khách quan và trung thực nhất về khả năng tổng quát hóa của mô hình cuối cùng.

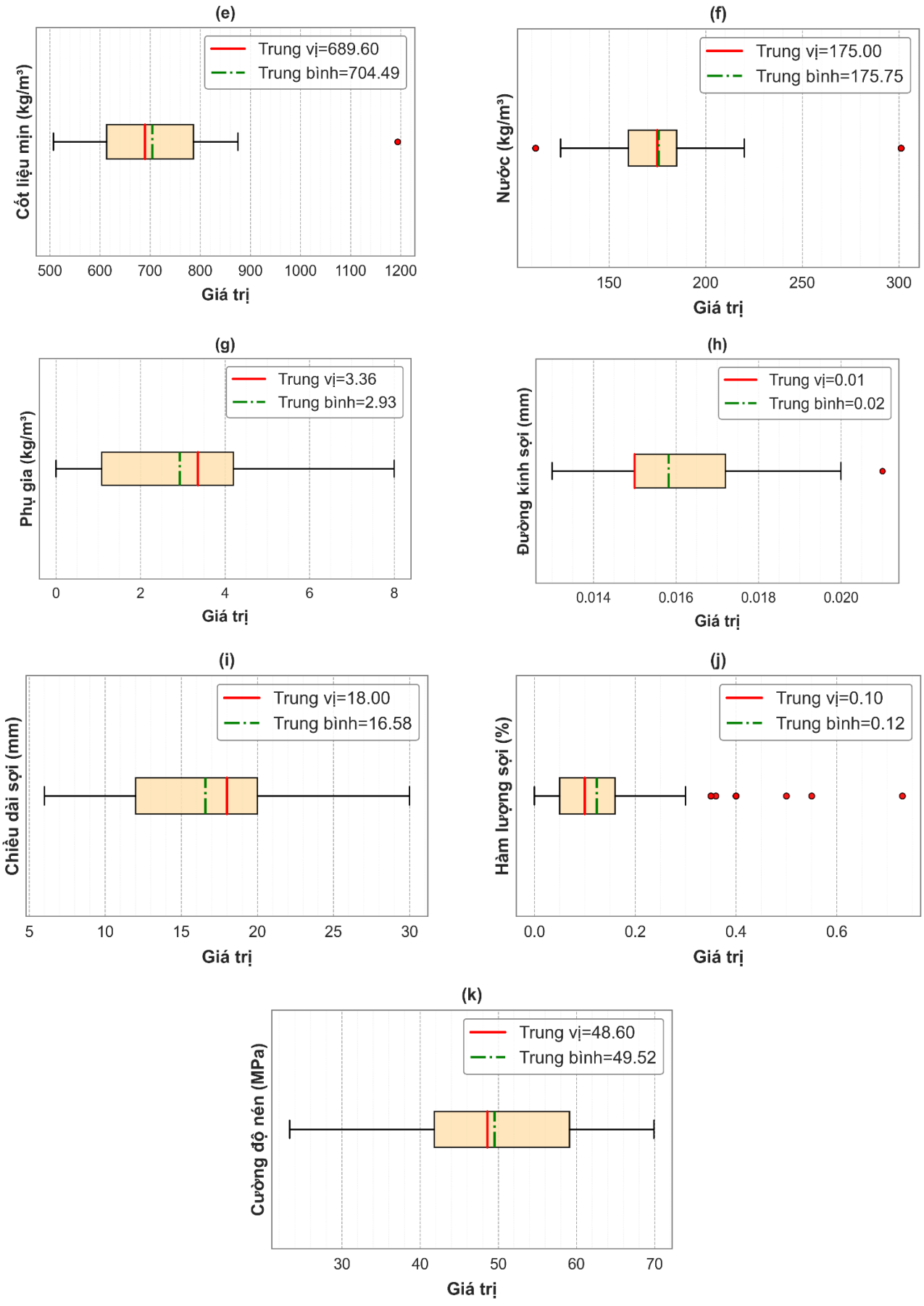
Bảng 1. Các thông số đầu vào và đầu ra của bộ dữ liệu

	Đơn vị	GTNN	Trung vị	Trung bình	GTLN	Std	Sk
Xi măng (X_1)	kg/m^3	217.0	402.0	401.1	613.3	75.9	-0.07
Tro bay (X_2)	kg/m^3	0	0	45.3	168.0	56.5	0.79
Muội silic (X_3)	kg/m^3	0	0	13.1	126.0	28.869	2.37
Cốt liệu thô (X_4)	kg/m^3	512.0	1125.0	1078.7	1540.0	176.9	-0.62
Cốt liệu mịn (X_5)	kg/m^3	507.0	689.6	704.5	1193.7	108.88	1.21
Nước (X_6)	kg/m^3	112.0	175.0	175.8	301.0	30.9	1.79
Phụ gia (X_7)	kg/m^3	0	3.36	2.9	8.0	1.94	-0.34
Đường kính sợi (X_8)	mm	0.013	0.02	0.016	0.021	0.002	1.11
Chiều dài sợi (X_9)	mm	6	18.0	16.58	30.0	5.93	0.47
Hàm lượng sợi (X_{10})	%	0	0.1	0.12	0.73	0.12	1.97
Cường độ nén (Y)	MPa	23.3	48.6	49.5	69.9	11.4	-0.04

Sk=Độ lệch; Std=Độ lệch chuẩn.



Hình 1. Biểu đồ hộp thể hiện phân bố các thông số của bộ dữ liệu nghiên cứu.


(e)
Trung vị=689.60
Trung bình=704.49
Cốt liệu mịn (kg/m³)
500 600 700 800 900 1000 1100 1200
Giá trị
(f)
Trung vị=175.00
Trung bình=175.75
Nước (kg/m³)
150 200 250 300
Giá trị
(g)
Trung vị=3.36
Trung bình=2.93
Phụ gia (kg/m³)
0 2 4 6 8
Giá trị
(h)
Trung vị=0.01
Trung bình=0.02
Đường kính sợi (mm)
0.014 0.016 0.018 0.020
Giá trị
(i)
Trung vị=18.00
Trung bình=16.58
Chiều dài sợi (mm)
5 10 15 20 25 30
Giá trị
(j)
Trung vị=0.10
Trung bình=0.12
Hàm lượng sợi (%)
0.0 0.2 0.4 0.6
Giá trị
(k)
Trung vị=48.60
Trung bình=49.52
Cường độ nén (MPa)
30 40 50 60 70
Giá trị

Hình 1. (tiếp)

2.2. Các thuật toán học máy

Nghiên cứu lựa chọn ba thuật toán học máy thuộc nhóm học tập giám sát, bao gồm Random Forest và hai biến thể mạnh mẽ của Gradient Boosting là HistGradientBoosting và CatBoost.

Random Forest: Là một thuật toán học máy tổ hợp dựa trên kỹ thuật Bagging. Random Forest xây dựng nhiều cây quyết định trong quá trình huấn luyện và đưa ra kết quả dự báo bằng cách lấy trung bình cộng kết quả của các cây thành phần. Ưu điểm của Random Forest là khả năng giảm thiểu hiện tượng quá khớp (ovefitting) so với cây quyết định đơn lẻ và hoạt động ổn định trên các bộ dữ liệu có nhiễu.

HistGradientBoosting: Là một biến thể tối ưu của Gradient Boosting, được thiết kế để tăng tốc độ huấn luyện trên các tập dữ liệu lớn. HistGradientBoosting hoạt động bằng cách nhóm các giá trị liên tục của biến đầu vào vào các biểu đồ tần suất rời rạc, giúp giảm đáng kể độ phức tạp tính toán khi tìm điểm phân chia tối ưu cho cây quyết định. Thuật toán này đặc biệt hiệu quả trong việc xử lý dữ liệu dạng bảng và có cơ chế tích hợp để xử lý các giá trị khuyết thiếu (nếu có).

CatBoost: Là một thư viện Gradient Boosting mã nguồn mở được phát triển bởi Yandex. Điểm khác biệt của CatBoost nằm ở kỹ thuật xử lý biến phân loại và sử dụng cơ chế "Ordered Boosting" để giải quyết vấn đề rò rỉ dữ liệu và dịch chuyển dự báo. CatBoost thường cho độ chính xác cao và khả năng tổng quát hóa tốt ngay cả với các bộ dữ liệu có kích thước vừa và nhỏ mà không cần tinh chỉnh tham số quá phức tạp.

2.3. Tối ưu hóa siêu tham số bằng tìm kiếm lưới

Để khai thác tối đa tiềm năng của các thuật toán, nghiên cứu này sử dụng kỹ thuật tìm kiếm lưới (Grid search) kết hợp với kiểm chứng chéo. Quy trình này thực hiện tìm kiếm toàn diện trên một không gian tham số được xác định trước. Tại mỗi lưới tham số, mô hình được đánh giá hiệu suất và bộ tham số cho kết quả sai số dự báo thấp nhất (dựa trên chỉ số RMSE hoặc R^2 trung bình) sẽ được chọn làm mô hình cuối cùng để huấn luyện lại trên toàn bộ tập dữ liệu huấn luyện trước khi kiểm tra trên tập kiểm tra độc lập. Các siêu tham số chính và không gian tìm kiếm chi tiết cho từng thuật toán được được trình bày cụ thể trong Bảng 2.

Bảng 2. Không gian tìm kiếm siêu tham số cho các mô hình

Thuật toán	Siêu tham số	Giá trị tìm kiếm	Ý nghĩa
Random Forest	n_estimators	[100, 300, 500]	Số lượng cây.
	max_depth	[None, 10, 20]	Độ sâu tối đa.
	min_samples_leaf	[1, 5, 10]	Số mẫu tối thiểu tại lá.
	max_features	["sqrt", 0.5, 1.0]	Số đặc trưng tối đa tại nút.
HistGradientBoosting	max_iter	[100, 300, 500]	Số vòng lặp tối đa.
	learning_rate	[0.01, 0.05, 0.1]	Tốc độ học.
	max_depth	[None, 5, 10]	Độ sâu tối đa.
	l2_regularization	[0, 1, 10]	Hệ số điều chuẩn
CatBoost	iterations	[100, 300, 500]	Số vòng lặp.
	learning_rate	[0.01, 0.05, 0.1]	Tốc độ học.
	depth	[4, 6, 8]	Độ sâu của cây.
	l2_leaf_reg	[1, 3, 9]	Hệ số điều chuẩn.

Việc thiết lập không gian tìm kiếm siêu tham số tại Bảng 2 được thực hiện dựa trên kinh nghiệm thực nghiệm và tham chiếu từ các nghiên cứu về dự báo tính chất cơ lý bê tông bằng mô hình tổ hợp [19,20], nơi các khoảng giá trị tương tự đã chứng minh hiệu quả trong việc cân bằng giữa độ phức tạp mô hình và khả năng tổng quát hóa. Với quy mô dữ liệu n = 267, việc giới hạn không gian tìm kiếm ở mức hợp lý là cần thiết để kiểm soát chi phí tính toán, đồng thời tránh việc mở rộng tham số quá mức gây ra hiện tượng nhiễu trong quá trình tối ưu hóa.

2.4. Đánh giá hiệu suất mô hình

Hiệu suất dự báo của các mô hình được đánh giá thông qua ba chỉ số thống kê phổ biến: hệ số xác định (R^2), sai số bình phương trung bình (RMSE) và sai số tuyệt đối trung bình (MAE). Các chỉ số này được tính toán như sau:

$$R^2=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \quad (1)$$

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \quad (2)$$

$$MAE=\frac{1}{n}|y_i-\hat{y}_i| \quad (3)$$

Trong đó:

y_i: Giá trị cường độ nén thực nghiệm của mẫu thứ i.

$\hat{y}_i$: Giá trị cường độ nén dự báo bởi mô hình.

$\bar{y}$: Giá trị trung bình của cường độ nén thực nghiệm.

n: Số lượng mẫu dữ liệu.

Mô hình tốt nhất là mô hình có giá trị R^2 tiệm cận 1, đồng thời RMSE và MAE có giá trị nhỏ nhất.

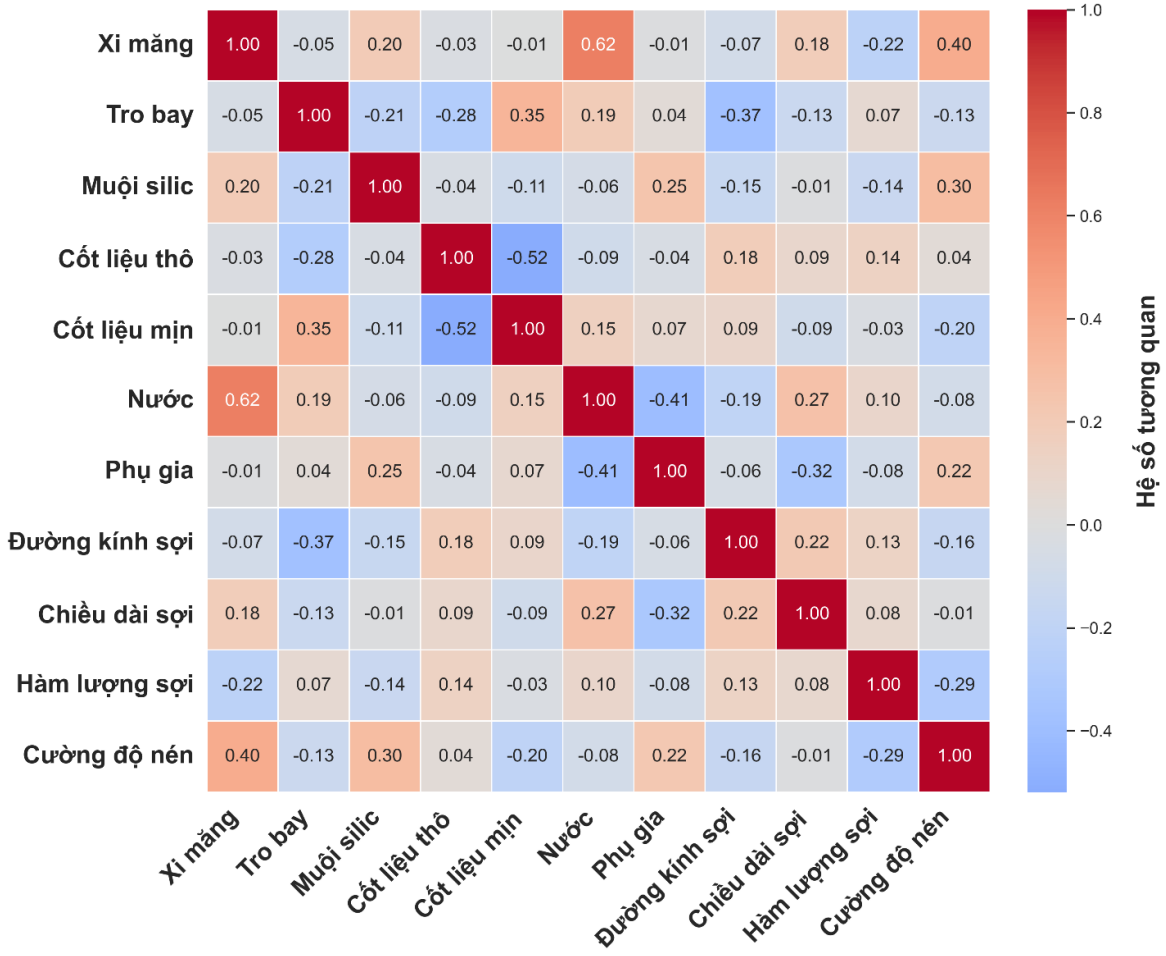
3. Kết quả và thảo luận

3.1. Phân tích tương quan

Trước khi tiến hành xây dựng các mô hình dự báo, phân tích tương quan Pearson được thực hiện nhằm đánh giá mức độ phụ thuộc tuyến tính giữa các biến đầu vào và cường độ nén mục tiêu, đồng thời kiểm soát hiện tượng đa cộng tuyến. Hình 2 trình bày ma trận nhiệt thể hiện các hệ số tương quan giữa các biến. Kết quả phân tích ma trận tương quan giữa các biến đầu vào và cường độ nén của BFRC cho thấy những đặc điểm lý tính đáng chú ý của loại vật liệu này. Trong số các thành phần hỗn hợp, hàm lượng xi măng thể hiện mối tương quan thuận mạnh nhất với cường độ nén (hệ số 0.40), phản ánh đúng bản chất của quá trình thủy hóa tạo nên cường độ bê tông. Sự đóng góp tích cực của các biến muội silic (0.30) và phụ gia (0.22) cũng minh chứng cho hiệu quả của việc cải thiện cấu trúc rỗng và tăng cường độ đặc chắc thông qua các phản ứng pozzolanic và giảm tỷ lệ nước/chất kết dính.

Ngược lại, hàm lượng nước và tro bay thể hiện mối tương quan nghịch với hệ số lần lượt là -0.08 và -0.13, cho thấy sự gia tăng các thành phần này trong điều kiện không kiểm soát có thể dẫn đến sự suy giảm cường độ nén của mẫu. Đối với các đặc tính của sợi basalt, hàm lượng sợi và đường kính sợi đều có tương quan nghịch với cường độ nén (lần lượt là -0.29 và -0.16). Điều này có thể được giải thích do sự gia tăng hàm lượng sợi vượt quá mức tối ưu thường gây ra hiện tượng vón cục hoặc tạo ra các vùng rỗng trong cấu trúc bê tông nếu không được phân tán tốt, làm gián đoạn tính liên tục của bê tông và dẫn đến giảm khả năng chịu nén. Bên cạnh đó, chiều dài sợi gần như không có sự tương quan tuyến tính rõ rệt với cường độ nén (hệ số -0.01), cho thấy ảnh hưởng của chiều dài sợi đến cường độ nén phức tạp hơn và có thể mang tính phi tuyến cao. Hệ số tương quan tuyến tính thấp của biến chiều dài sợi nhấn mạnh sự giới hạn của các phương pháp hồi quy tuyến tính truyền thống, vốn giả định mối quan hệ tuyến tính giữa các biến và không thể nắm bắt hiệu quả các tương tác phi tuyến phức tạp trong bê tông sợi basalt, chẳng hạn như ảnh hưởng ngưỡng của chiều dài sợi đến sự phân tán và kiểm soát nứt [7,9]. Do đó, việc ứng dụng các thuật toán học máy phi tuyến như Random Forest và Gradient Boosting là cần thiết, vì chúng có khả năng mô hình hóa các mối quan hệ không tuyến tính và tương

tác giữa các biến mà không cần giả định dạng hàm số trước, dẫn đến độ chính xác dự báo cao hơn so với hồi quy cổ điển, như đã chứng minh trong các nghiên cứu tương tự về vật liệu composite



Hình 2. Ma trận tương quan giữa các biến

Về mối quan hệ giữa các biến độc lập, hệ số tương quan giữa cốt liệu mịn và cốt liệu thô đạt giá trị -0.52. Đây là mức tương quan nghịch đáng kể nhất giữa các biến đầu vào, phản ánh quy luật thiết kế cấp phối thực tế khi sự gia tăng của thành phần cốt liệu này thường đi kèm với việc giảm thành phần cốt liệu kia để đảm bảo tính ổn định về thể tích của hỗn hợp. Nhìn chung, các hệ số tương quan giữa các biến độc lập đều nằm trong ngưỡng an toàn (dưới 0.70), cho thấy không có hiện tượng đa cộng tuyến nghiêm trọng, tạo cơ sở vững chắc cho việc huấn luyện các mô hình học máy đạt độ chính xác và ổn định cao.

3.2. Kết quả tối ưu hóa siêu tham số

Để đảm bảo các mô hình học máy đạt hiệu suất cao nhất và tránh hiện tượng quá khớp, nghiên cứu đã áp dụng kỹ thuật tìm kiếm lưới kết hợp với kiểm chứng chéo 5 lần trên tập dữ liệu huấn luyện. Quá trình này cho phép xác định tổ hợp các siêu tham số tối ưu nhất cho từng thuật toán Random Forest, HistGradientBoosting và CatBoost. Tiêu chí đánh giá để lựa chọn bộ tham số tốt nhất là tối đa hóa hệ số xác định trung bình ($\overline{R^2}$) trên tập kiểm chứng chéo. Kết quả tìm kiếm trên không gian tham số (Bảng 2) ghi nhận bộ siêu tham số tối ưu cho từng thuật toán, và kết quả cho thấy việc tinh chỉnh tham số mang lại hiệu quả rõ rệt, với cả ba mô hình đều đạt mức độ tin cậy cao ($\overline{R^2}$ > 0.8). Chi tiết về bộ siêu tham số tối ưu nhất cùng giá trị $\overline{R^2}$ trên tập xác thực của từng mô hình được tổng hợp trong Bảng 3.

Bảng 3. Giá trị các siêu tham số tối ưu

Thuật toán	Siêu tham số	Giá trị tối ưu	$\overline{R^2}$
Random Forest	n_estimators	500	0.880
	max_depth	None	
	min_samples_leaf	1	
	max_features	"sqrt"	
HistGradientBoosting	max_iter	500	0.887
	learning_rate	0.1	
	max_depth	None	
	l2_regularization	1	
CatBoost	iterations	500	0.847
	learning_rate	0.1	
	depth	4	
	l2_leaf_reg	3	

3.3. So sánh hiệu suất các mô hình

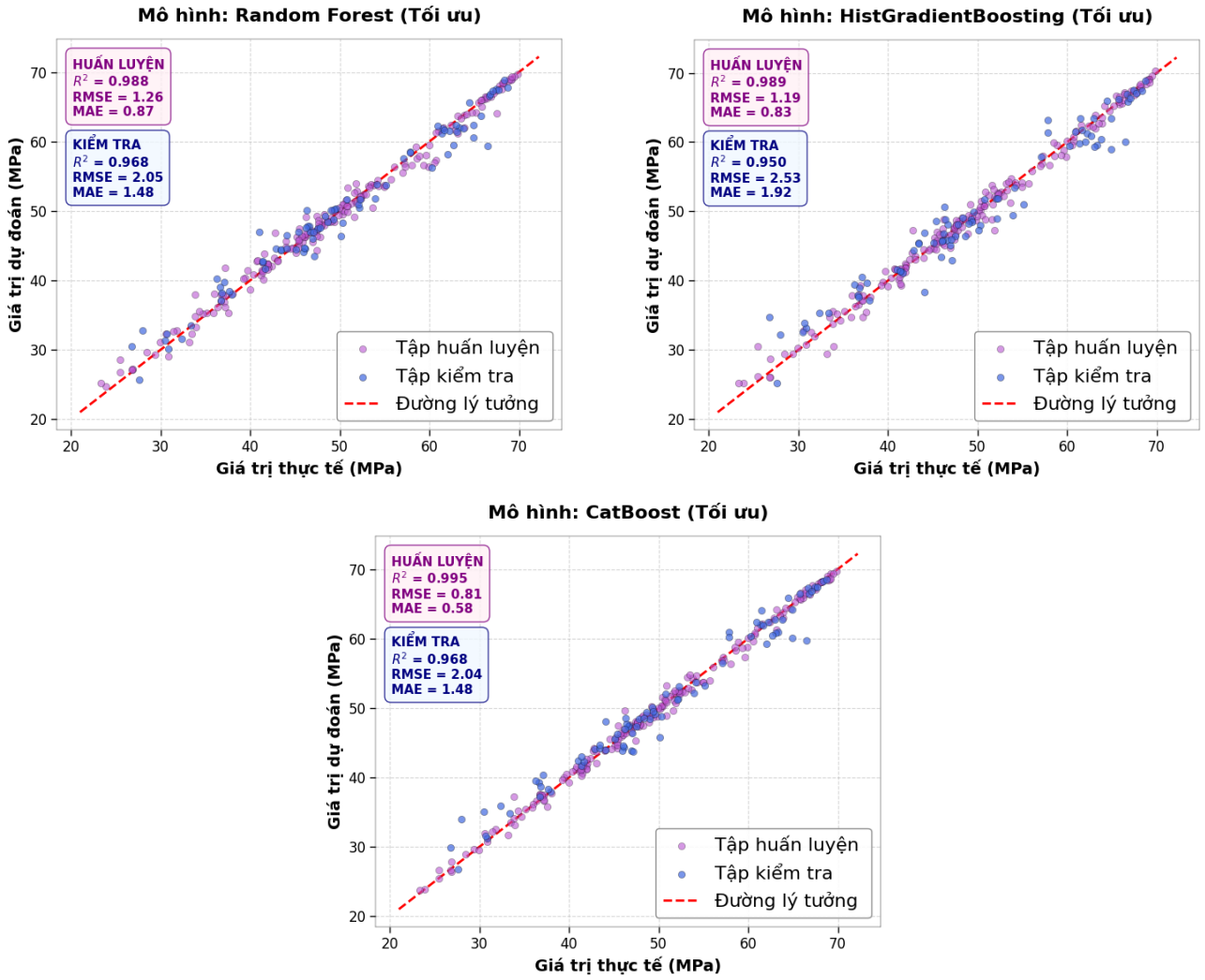
Sau khi xác định được bộ siêu tham số tối ưu, các mô hình Random Forest, HistGradientBoosting và CatBoost được tiến hành huấn luyện lại trên toàn bộ tập huấn luyện (70%) và đánh giá độc lập trên tập kiểm tra (30%). Kết quả định lượng về độ chính xác dự báo thông qua ba chỉ số R^2, RMSE và MAE được tổng hợp chi tiết trong Bảng 4. Trên tập huấn luyện, các mô hình đều đạt độ hội tụ cao ($R^2 \approx 0.99$). Tuy nhiên, đánh giá trên tập kiểm tra độc lập cho thấy sự khác biệt về khả năng tổng quát hóa. Mô hình HistGradientBoosting có kết quả rất tốt trên tập xác thực (như đã đề cập ở mục 3.2), nhưng trên tập kiểm tra độc lập, hiệu năng của nó thấp hơn một chút so với hai mô hình còn lại, với sai số RMSE tăng lên mức 2.45 MPa. Mặc dù CatBoost có R^2 huấn luyện cao nhất (0.992), nhưng mô hình Random Forest lại cho thấy sai số thấp hơn và ổn định hơn trên tập kiểm tra độc lập (RMSE = 2.05 MPa). Trong kỹ thuật học máy cho vật liệu, khả năng tổng quát hóa trên dữ liệu mới được ưu tiên cao hơn độ khớp trên tập huấn luyện nhằm đảm bảo tính tin cậy khi ứng dụng thực tế. Sự chênh lệch này phản ánh đặc trưng về sự đánh đổi giữa bias và variance, khẳng định Random Forest là lựa chọn tối ưu để kiểm soát hiện tượng quá khớp trong nghiên cứu này.

Bảng 4. Tổng hợp kết quả đánh giá hiệu suất các mô hình

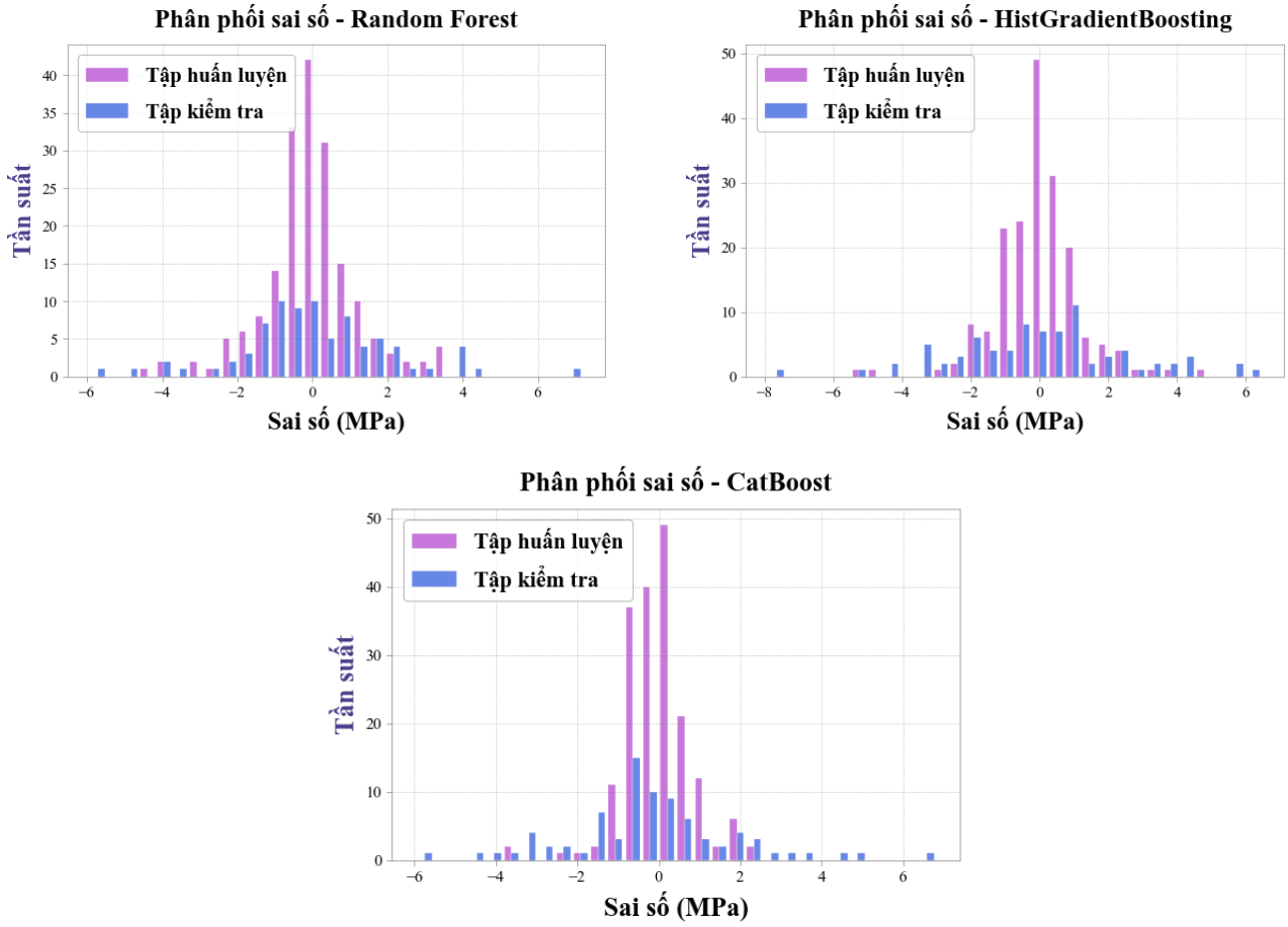
Mô hình	Tập dữ liệu	R^2	RMSE (MPa)	MAE (MPa)
Random Forest	Huấn luyện	0.988	1.26	0.87
	Kiểm tra	0.968	2.05	1.48
HistGradientBoosting	Huấn luyện	0.989	1.17	0.81
	Kiểm tra	0.954	2.45	1.82
CatBoost	Huấn luyện	0.992	1.01	0.77
	Kiểm tra	0.965	2.12	1.59

Trên tập huấn luyện, các mô hình đều đạt độ hội tụ cao ($R^2 \approx 0.99$). Tuy nhiên, đánh giá trên tập kiểm tra độc lập cho thấy sự khác biệt về khả năng tổng quát hóa. Mô hình HistGradientBoosting có kết quả rất tốt trên tập xác thực (như đã đề cập ở mục 3.2), nhưng trên tập kiểm tra độc lập, hiệu năng của nó thấp hơn một chút so với hai mô hình còn lại, với sai số RMSE tăng lên mức 2.45 MPa. Mặc dù CatBoost có R^2 huấn luyện cao nhất (0.992), nhưng mô hình Random Forest lại cho thấy sai số thấp hơn và ổn định hơn trên tập kiểm tra độc lập (RMSE = 2.05 MPa). Trong kỹ thuật học máy cho vật liệu, khả năng tổng quát hóa trên dữ liệu mới được ưu tiên cao hơn độ khớp trên tập huấn luyện nhằm đảm bảo tính tin cậy khi ứng dụng thực tế. Sự chênh lệch này phản ánh đặc trưng về sự đánh đổi giữa bias và variance, khẳng định Random Forest là lựa chọn tối ưu để kiểm soát hiện tượng quá khớp trong nghiên cứu này. Để minh họa rõ hơn về độ chính xác của các mô hình, Hình 3 thể hiện biểu đồ hồi quy tuyến tính so sánh giữa giá trị thực nghiệm (trục hoành) và giá trị dự báo (trục tung) cho cả tập huấn luyện và tập kiểm tra.

Trên biểu đồ, đường chéo nét đứt (y=x) đại diện cho trạng thái dự báo lý tưởng, nơi giá trị dự báo trùng khớp hoàn toàn với thực tế. Biểu đồ hồi quy (Hình 3) cho thấy các điểm dữ liệu của Random Forest trên tập kiểm tra bám sát đường chéo lý tưởng y=x tương tự như tập huấn luyện, trong khi HistGradientBoosting có xu hướng phân tán rộng hơn ở vùng cường độ cao



Hình 3. Biểu đồ hồi quy so sánh giá trị thực nghiệm và dự báo của các mô hình trên tập huấn luyện và kiểm tra



Hình 4. Biểu đồ phân bố sai số dự báo của các mô hình

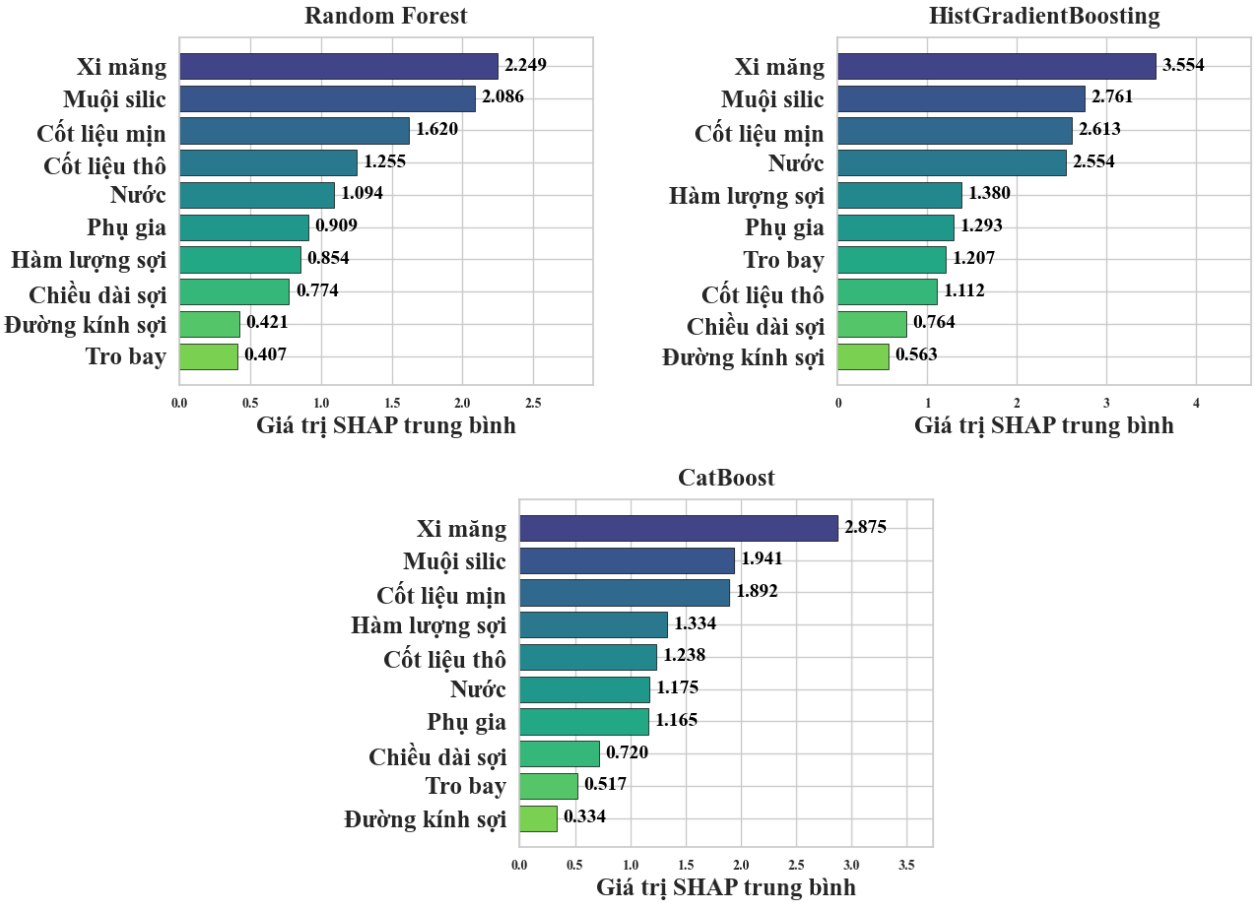
Phân bố dữ liệu tại Hình 3 cho thấy mô hình đạt độ chính xác cao nhất trong dải cường độ 40–60 MPa. Đây là dải cường độ phổ biến nhất của bê tông kết cấu trong hạ tầng giao thông. Tương tự, biểu đồ phân bố sai số (Hình 4) xác nhận sai số dự báo của Random Forest tập trung chủ yếu quanh giá trị 0. Có thể thấy, mặc dù mô hình CatBoost có khả năng học tập dữ liệu huấn luyện tốt nhất, nhưng Random Forest được xác định là mô hình tối ưu nhất cho bộ dữ liệu này nhờ sự cân bằng giữa độ chính xác cao và khả năng tổng quát hóa ổn định trên dữ liệu mới.

3.4. Phân tích mức độ ảnh hưởng của các biến

Để đánh giá mức độ ảnh hưởng của các biến đầu vào đối với cường độ nén dự báo, nghiên cứu áp dụng kỹ thuật SHAP (SHapley Additive exPlanations), một phương pháp giải thích mô hình học máy dựa trên lý thuyết trò chơi. SHAP tính toán giá trị đóng góp biên của từng biến đối với mỗi dự báo, sau đó tổng hợp để xác định tầm quan trọng tổng thể và hướng ảnh hưởng (tích cực hoặc tiêu cực) của biến đó. Phương pháp này cung cấp cái nhìn toàn diện về cách các thành phần cấp phối ảnh hưởng đến đầu ra mô hình, đồng thời giúp xác thực tính nhất quán giữa các thuật toán khác nhau. Hình 5 thể hiện giá trị SHAP tuyệt đối trung bình nhằm xếp hạng tầm quan trọng của các biến cho cả ba mô Random Forest, HistGradientBoosting và CatBoost. Hình 6 thể hiện biểu đồ phân bố giá trị SHAP cho từng biến, minh họa sự phân tán và hướng ảnh hưởng của chúng đối với cường độ nén, trong đó mỗi điểm đại diện cho một mẫu dữ liệu, màu sắc thể hiện giá trị của biến (đỏ là cao, xanh là thấp) và vị trí trên trục hoành biểu thị mức độ đóng góp vào kết quả dự báo (dương là làm tăng cường độ, âm là làm giảm cường độ). Kết quả từ Hình 5 cho thấy sự nhất quán giữa ba mô hình trong việc xác định các biến quan trọng nhất. Cụ thể, ba biến quan trọng nhất được cả ba mô hình thống nhất nhận diện là Xi măng, Muội silic và Cốt

liệu mịn. Kết quả này có thể được giải thích dựa trên các tính chất cơ bản của vật liệu: Xi măng là thành phần chính tạo nên chất kết dính và cường độ; Muội Silic với kích thước hạt nhỏ giúp lấp đầy các lỗ rỗng và tham gia phản ứng làm đặc chắc cấu trúc; Cốt liệu mịn giúp điền đầy khoảng trống giữa các hạt cốt liệu thô, làm tăng độ đặc của hỗn hợp.

Các kết quả này phù hợp với nhận định trong nghiên cứu [18]. Sự thay đổi nhẹ trong thứ tự quan trọng của các biến giữa Random Forest và hai mô hình Boosting (HistGradientBoosting và CatBoost) có thể được giải thích bởi bản chất thuật toán: Random Forest dựa trên bagging với phân chia ngẫu nhiên, dẫn đến trọng số đồng đều hơn cho các biến, trong khi các mô hình Boosting tập trung vào việc điều chỉnh lỗi dần dần, làm nổi bật các tương tác phức tạp hơn giữa các biến như sợi basalt và cốt liệu [15,16]. Dù vậy, sự nhất quán tổng thể (với xi măng, muội silic và cốt liệu mịn luôn dẫn đầu) khẳng định tính ổn định của phân tích SHAP, phản ánh rằng sự khác biệt chủ yếu xuất phát từ cơ chế học tập chứ không phải nhiễu dữ liệu

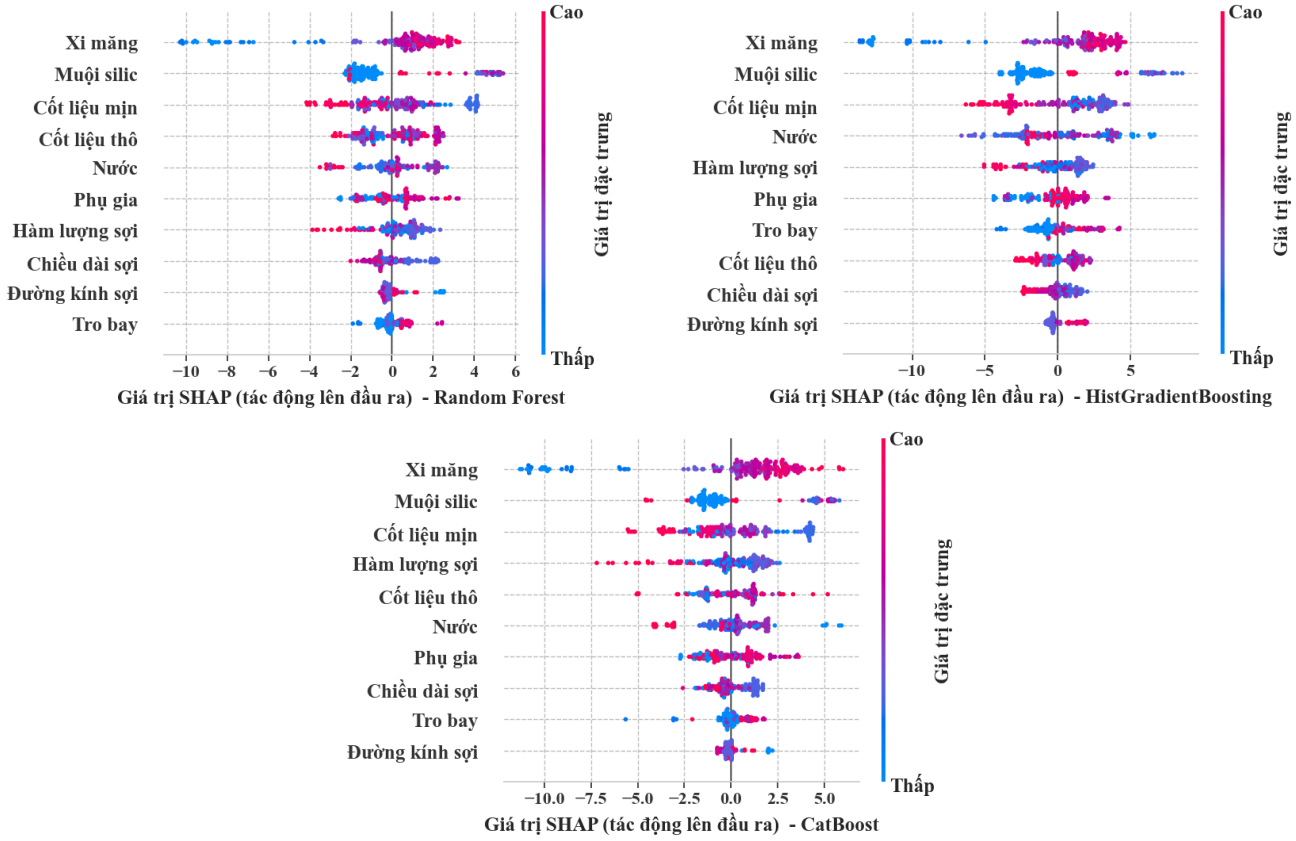


Hình 5. Giá trị SHAP tuyệt đối trung bình của các biến đầu vào

Phân tích giá trị SHAP (Hình 6) cho thấy xu hướng tác động vật lý rõ rệt và nhất quán giữa các mô hình. Cụ thể, các giá trị cao (màu đỏ) của biến xi măng và muội silic tập trung về phía dương trục hoành, phản ánh quy luật thủy hóa xi măng và phản ứng pozzolanic – tạo ra các hợp chất C-S-H bổ sung, tăng độ đặc chắc và giảm độ rỗng mao dẫn trong ma trận bê tông. Sự tương đồng này giữa kết quả học máy và bản chất vật lý của vật liệu khẳng định tính hợp lý của mô hình, vì SHAP không chỉ xác nhận vai trò lấp đầy lỗ rỗng mà còn làm nổi bật tương tác tích cực với xi măng, giúp cải thiện cường độ nén tổng thể trong BFRC, phù hợp với các nghiên cứu thực nghiệm về phụ gia khoáng hoạt tính [2,10]. Biến hàm lượng sợi basalt cho thấy tác động hai chiều: khi hàm lượng sợi thấp (màu xanh), SHAP dương cho thấy sợi hỗ trợ kiểm soát vi nứt và tăng độ dẻo, nhưng khi hàm lượng

cao (màu đỏ), SHAP chuyển sang âm, phản ánh sự hình thành cụm sợi làm gián đoạn tính liên tục của ma trận bê tông, tăng độ rỗng và giảm cường độ nén [7,21,22].

Phân tích này không chỉ xác nhận cơ chế vật lý – nơi phân tán không đồng đều dẫn đến khuyết tật cấu trúc – mà còn chứng minh khả năng của học máy trong việc phát hiện ngưỡng tối ưu (khoảng 0.10–0.20%), góp phần mở rộng ứng dụng thực tiễn để tránh lãng phí vật liệu và tối ưu hóa cấp phối BFRC. Trong khi đó, biến nước thể hiện phân bố giá trị cao ở phía âm, phù hợp với quy luật gia tăng độ rỗng khi tỷ lệ nước/chất kết dính cao [23].



Hình 6. Biểu đồ SHAP mô tả chiều hướng tác động của các biến đầu vào đến cường độ nén

Phân tích SHAP không chỉ xác nhận các quy luật cơ bản đã biết mà còn làm rõ các ngưỡng ảnh hưởng và tương tác giữa các biến, nâng cao giá trị giải thích cơ chế dựa trên dữ liệu. Ví dụ, tác động tích cực của xi măng và muội silic có thể bị suy giảm khi kết hợp với hàm lượng nước lớn (tương tác nghịch với độ rỗng mao dẫn) [23]; đồng thời, hàm lượng sợi basalt tương tác với cốt liệu mịn trong việc kiểm soát độ đặc chắc của hỗn hợp [18,22]. Các mô hình Boosting (CatBoost và HistGradientBoosting) nổi bật hơn Random Forest ở khả năng phân tách chi tiết các tương tác phức tạp này, phù hợp với xu hướng nghiên cứu hiện nay về học máy giải thích trong lĩnh vực vật liệu xây dựng [24,25]. Việc kết hợp ba thuật toán cùng phương pháp SHAP giúp thống nhất và làm rõ quy trình dự báo, cung cấp cơ sở khoa học vững chắc để tối ưu hóa cấp phối BFRC – chẳng hạn như duy trì hàm lượng xi măng và muội silic cao kết hợp với sợi basalt ở mức tối ưu nhằm cân bằng giữa cường độ nén, độ bền và hiệu quả kinh tế.

4. Kết luận và kiến nghị

Nghiên cứu này đã chứng minh tính khả thi của việc ứng dụng các mô hình học máy bao gồm Random Forest, HistGradientBoosting (HGB) và CatBoost để dự báo cường độ nén của bê tông sợi basalt (BFRC) dựa trên bộ dữ liệu 267 mẫu thí nghiệm. Kết quả cho thấy cả ba mô hình, sau khi được tối ưu hóa bằng kỹ thuật tìm kiếm lưới, đều

đạt được độ chính xác cao và khả năng dự báo tin cậy. Trong đó, mô hình Random Forest được xác định là công cụ tối ưu nhất cho bộ dữ liệu này nhờ sự cân bằng giữa độ chính xác cao và khả năng tổng quát hóa ổn định, giảm thiểu tối đa hiện tượng quá khớp với sai số bình phương trung bình thấp nhất (RMSE = 2.05 MPa). Bên cạnh đó, việc tích hợp kỹ thuật SHAP đã giúp làm rõ "hộp đen" của các thuật toán học máy, xác định được nhóm biến có ảnh hưởng lớn đến cường độ nén bao gồm hàm lượng xi măng, muội silic và cốt liệu mịn. Sự tương đồng trong kết quả giải thích của cả ba mô hình đã khẳng định tính khoa học của bộ dữ liệu và sự phù hợp của các phương pháp tiếp cận được đề xuất. Từ những kết quả đạt được này, các hướng nghiên cứu tiếp theo được đề xuất nên tập trung mở rộng quy mô bộ dữ liệu với các loại phụ gia và điều kiện bảo dưỡng đa dạng hơn, đồng thời kết hợp mô hình dự báo với các thuật toán tối ưu hóa đa mục tiêu nhằm tìm ra thiết kế cấp phối BFRC không chỉ đạt yêu cầu về cường độ mà còn tối ưu về chi phí và hiệu quả môi trường.

Tài liệu tham khảo

[1] C. Meyer. (2009). The greening of the concrete industry. *Cement and Concrete Composites*, 31(8), 601–605. https://doi.org/10.1016/j.cemconcomp.2008.12.010

[2] F. Köksal, A. Beycioğlu, M. Dobiszewska. (2022). Optimization based on toughness and splitting tensile strength of steel-fiber-reinforced concrete incorporating silica fume using response surface method. *Materials*, 15(18), 6218. https://doi.org/10.3390/ma15186218

[3] S. Grzesiak, M. Pahn, M. Schultz-Cornelius, S. Harenberg, C. Hahn. (2021). Influence of fiber addition on the properties of high-performance concrete. *Materials*, 14(13), 3736. https://doi.org/10.3390/ma14133736?urlappend=%3Futm_source%3Dresearchgate.net%26utm_medium%3Darticle

[4] F. Korkut, M. Karalar. (2023). Investigational and numerical examination on bending response of reinforced rubberized concrete beams including plastic waste. *Materials*, 16(16), 5538. https://doi.org/10.3390/ma16165538

[5] B.N. Al-Kharabsheh, M.M. Arbili, A. Majdi, S.M. Alogla, A. Hakamy, J. Ahmad, A.F. Deifalla. (2022). Basalt fibers reinforced concrete: strength and failure modes. *Materials*, 15(20), 7350. https://doi.org/10.3390/ma15207350

[6] Y. Li, Z. Gu, B. Zhao, J. Zhang, X. Zou. (2022). Experimental study on mechanical properties of basalt fiber concrete after cryogenic freeze-thaw cycles. *Polymers*, 15, 196. https://doi.org/10.3390/polym15010196

[7] X. Wang, J. He, A.S. Mosallam, C. Li, H. Xin. (2019). The Effects of Fiber Length and Volume on Material Properties and Crack Resistance of Basalt Fiber Reinforced Concrete (BFRC). *Advances in Materials Science and Engineering,* 1–17, 7520549. https://doi.org/10.1155/2019/7520549

[8] D. Wang, Y. Ju, H. Shen, L. Xu. (2019). Mechanical properties of high performance concrete reinforced with basalt fiber and polypropylene fiber. *Construction and Building Materials,* 197, 464–473. https://doi.org/10.1016/j.conbuildmat.2018.11.181

[9] W. Chen, Z.C. Zhu, J. Wang, J. Chen, Y. Mo. (2019). Numerical analysis of mechanical properties of chopped basalt fiber reinforced concrete. *Key Engineering Materials*, 815, 175–181. https://doi.org/10.4028/www.scientific.net/KEM.815.175

[10] A. Najmoddin, H. Etemadfard, M. Ghalehnovi. (2024). Multi-output machine learning for predicting the mechanical properties of BFRC. *Case Studies in Construction Materials,* 20, e02818. https://doi.org/10.1016/j.cscm.2023.e02818

[11] J. Zheng, T. Yao, J. Yue, M. Wang, S. Xia. (2023). Compressive strength prediction of BFRC based on a novel hybrid machine learning model. *Buildings* 13(8), 1934.

https://doi.org/10.3390/buildings13081934?urlappend=%3Futm_source%3Dresearchgate.net%26utm_medium%3Darticle

[12] H.-B. Ly. (2026). An interpretable machine learning-based measurement system for predicting the compressive strength of concrete at elevated temperatures. *Measurement,* 264, 120273. https://doi.org/10.1016/j.measurement.2025.120273

[13] M. Hasanipanah, R.A. Abdullah, M. Iqbal, H.-B. Ly. (2023). Predicting rubberized concrete compressive strength using machine learning: A feature importance and partial dependence analysis. *Journal of Science and Transport Technology,* 3(1), 26–43. https://doi.org/10.58845/jstt.utt.2023.en.3.1.26-43

[14] H.-B. Ly, T.-A. Nguyen, V.Q. Tran. (2021). Development of deep neural network model to predict the compressive strength of rubber concrete. *Construction and Building Materials*, 301, 124081. https://doi.org/10.1016/j.conbuildmat.2021.124081

[15] C. Bentéjac, A. Csörgő, G. Martínez-Muñoz. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review,* 54, 1937–1967. https://doi.org/10.1007/s10462-020-09896-5

[16] M. Fernández-Delgado, M.S. Sirsat, E. Cernadas, S. Alawadi, S. Barro, M. Febrero-Bande. (2019). An extensive experimental survey of regression methods. *Neural Networks*, 111, 11–34. https://doi.org/10.1016/j.neunet.2018.12.010

[17] M. Wang. (2022) Mechanical properties dataset of BFRC for strength prediction with machine learning. Mendeley Data 1.

[18] M.E. Haque, M. Arifuzzaman, K. Khan, A.K.M. Azad, A.E. Alluqmani, A. Kashem. (2025). Machine learning models for mechanical properties prediction of basalt fiber-reinforced concrete incorporating graphical user interface. *Scientific Reports,* 15, 37029. https://doi.org/10.1038/s41598-025-20304-2

[19] J. Kim, D. Lee. (2025). Comparative study on hyperparameter tuning for predicting concrete compressive strength. *Buildings,* 15, 2173. https://doi.org/10.3390/buildings15132173?urlappend=%3Futm_source%3Dresearchgate.net%26utm_medium%3Darticle

[20] Z. Zhang, T. Zeng, Y. Zeng, P. Zhu. (2025). Explainable Prediction of UHPC Tensile Strength Using Machine Learning with Engineered Features and Multi-Algorithm Comparative Evaluation. *Buildings*, 15(17), 3217. DOI: 10.3390/buildings15173217

[21] B.N. Al-Kharabsheh, M.M. Arbili, A. Majdi, S.M. Alogla, A. Hakamy, J. Ahmad, A.F. Deifalla. (2023). Basalt fiber reinforced concrete: A compressive review on durability aspects. *Materials*, 16(1), 429. https://doi.org/10.3390/ma16010429?urlappend=%3Futm_source%3Dresearchgate.net%26utm_medium%3Darticle

[22] K. Khan, W. Ahmad, M.N. Amin, A. Ahmad, S. Nazar, A.A. Alabdullah. (2022). Compressive strength estimation of steel-fiber-reinforced concrete and raw material interactions using advanced algorithms. *Polymers,* 14, 3065. https://doi.org/10.3390/polym14153065?urlappend=%3Futm_source%3Dresearchgate.net%26utm_medium%3Darticle

[23] K.C. Onyelowe, A.M. Ebid, S. Hanandeh, V. Kamchoom, P. Awoyera, S. Avudaiappan. (2025). Modeling the compressive strength behavior of concrete reinforced with basalt fiber. *Scientific Reports,* 15, 11493. https://doi.org/10.1038/s41598-025-96343-6

[24] I.U. Ekanayake, D.P.P. Meddage, U. Rathnayake. (2022). A novel approach to explain the black-box nature of machine learning in compressive strength predictions of concrete using Shapley additive explanations (SHAP). *Case Studies in Construction*

Materials, 16, e01059. https://doi.org/10.1016/j.cscm.2022.e01059

[25] X. Wang, Z. Ke, W. Liu, P. Zhang, S. Cui, N. Zhao, W. He. (2025). Compressive Strength Prediction of Basalt Fiber Reinforced Concrete Based on Interpretive Machine Learning Using SHAP Analysis. *Iranian Journal of Science and Technology, Transaction of Civil Engineering*, 49, 2461–2480. https://doi.org/10.1007/s40996-024-01594-4