Transport and Communications Science Journal

DATA AUGMENTATION FOR IMPROVED PREDICTION OF SHEAR STRENGTH CAPACITY OF UHPC-JACKETED RC BEAMS

Le Ha Linh, Ngo Duc Chinh, Le Dac Hien, Hoang Viet Hai*

University of Transport and Communications, No 3 Cau Giay Street, Hanoi, Vietnam

ARTICLE INFO

TYPE: Research Article
Received: 26/01/2026
Revised: 12/03/2026
Accepted: 17/03/2026
Published online: 15/04/2026
https://doi.org/10.47869/tcsj.77.3.6

* *Corresponding author*
Email: hoangviethai@utc.edu.vn

Abstract. Predicting the shear capacity of reinforced concrete (RC) beams strengthened with Ultra-High-Performance Concrete (UHPC) has become a significant subject of international research. However, experimental data remain extremely limited due to the high costs involved. This study proposes a Gaussian-based data augmentation framework to improve the predictive performance of shear resistance for UHPC-strengthened RC beams. Four machine learning models - Random Forest (RF), KNN, LightGBM, and XGBoost—were employed for evaluation. To ensure an unbiased assessment of generalization capability, the testing process was conducted exclusively on the original experimental dataset. The results indicate that models trained solely on limited experimental data exhibit substantial errors. After applying data augmentation, the performance of most models improved significantly, except for the K-Nearest Neighbors model. Among them, the XGBoost model achieved the best performance, with a test R2 of 0.949, MAE of 36.558 kN, and RMSE of 54.737 kN. The findings indicate that the proposed approach effectively addresses data scarcity, providing a dependable solution for assessing and designing RC beams retrofitted with UHPC.

Keywords: Data augmentation; ultra-high-performance concrete; machine learning.

@2026 University of Transport and Communication

Tạp chí Khoa học Giao thông vận tải

ĐÁNH GIÁ KHẢ NĂNG TĂNG CƯỜNG DỮ LIỆU TRONG DỰ BÁO KHẢ NĂNG CHỊU CẮT CỦA DẦM BÊ TÔNG CỐT THÉP ĐƯỢC TĂNG CƯỜNG BỞI BÊ TÔNG UHPC

Lê Hà Linh, Ngô Đức Chinh, Lê Đắc Hiền, Hoàng Việt Hải*

Trường Đại học Giao thông vận tải, Số 3 Cầu Giấy, Hà Nội, Việt Nam

THÔNG TIN BÀI BÁO

CHUYÊN MỤC: Công trình khoa học
Ngày nhận bài: 26/01/2026
Ngày nhận bài sửa: 12/03/2026
Ngày chấp nhận đăng: 17/03/2026
Ngày xuất bản Online: 15/04/2026
https://doi.org/10.47869/tcsj.77.3.6
* *Tác giả liên hệ*
Email: hoangviethai@utc.edu.vn

Tóm tắt. Việc dự báo khả năng chịu cắt của dầm bê tông cốt thép (BTCT) được tăng cường bằng lớp bê tông siêu tính năng (UHPC) đang là hướng nghiên cứu thu hút nhiều quan tâm hiện nay. Tuy nhiên, dữ liệu thực nghiệm còn rất hạn chế do chi phí cao. Nghiên cứu này đề xuất phương án tăng cường dữ liệu dựa trên phân phối Gaussian nhằm nâng cao hiệu suất dự báo sức kháng cắt của dầm BTCT được tăng cường bởi UHPC. Bốn mô hình học máy bao gồm: Random Forest (RF), KNN, LightGBM và XGBoost đã được sử dụng để đánh giá. Quá trình kiểm tra được thực hiện hoàn toàn trên tập dữ liệu thực nghiệm gốc để đảm bảo đánh giá khách quan khả năng khái quát hóa. Kết quả cho thấy các mô hình chỉ huấn luyện trên dữ liệu thực nghiệm bị hạn chế và có sai số khá lớn. Sau khi áp dụng tăng cường dữ liệu, ngoại trừ mô hình KNN, hiệu suất của các mô hình khác được cải thiện rõ rệt. Mô hình XGBoost đạt hiệu quả cao nhất với R2 kiểm tra lên tới 0,949, MAE = 36,558 kN và RMSE = 54,737 kN. Kết quả này giúp giải quyết hiệu quả tình trạng khan hiếm dữ liệu. Cách tiếp cận này cung cấp một phương án tin cậy cho việc đánh giá và thiết kế tăng cường các dầm BTCT bằng bê tông UHPC.

Từ khóa: Tăng cường dữ liệu, Khả năng chịu cắt, UHPC, Mô hình học máy.

@2026 Trường Đại học Giao thông vận tải

1. ĐẶT VẤN ĐỀ

Dầm bê tông cốt thép (BTCT) được sửa chữa, tăng cường bởi bê tông siêu tính năng (Ultra-High-Performance Concrete – UHPC) đang ngày càng thu hút sự quan tâm trong lĩnh vực kết cấu công trình cầu và giúp tăng đáng kể khả năng chịu uốn, cắt, tăng độ bền lâu và khả năng chống lại tác động xâm thực của môi trường. Nhiều nghiên cứu thực nghiệm đã được thực hiện nhằm đánh giá ứng xử uốn, cắt của các dầm BTCT được sửa chữa hoặc tăng cường bằng UHPC trong các trường hợp khác nhau [1], [2], [3]. Các công thức giải tích lý thuyết cũng đã được đề xuất nhằm xem xét ảnh hưởng của UHPC trong trong việc tăng cường khả năng chịu cắt của dầm BTCT. Tuy nhiên, công thức giải tích để tính khả năng chịu cắt cho dầm BTCT được tăng cường UHPC vẫn dựa trên các giả thiết cơ học đơn giản hóa. Những mô hình này thường chưa mô tả đầy đủ được tính phi tuyến, cũng như tương tác phức tạp giữa bê tông cũ và lớp UHPC gia cường, mối quan hệ tương quan giữa các tham số như đặc trưng hình học, đặc trưng vật liệu như bê tông, UHPC và cốt thép.

Mặc dù các tiêu chuẩn thiết kế hiện hành như Eurocode 2 [4] và một số tiêu chuẩn quốc tế cung cấp công thức lý thuyết xác định khả năng chịu cắt của dầm BTCT truyền thống, tuy nhiên như không có công thức tính các cấu kiện được tăng cường bằng UHPC. Các nghiên cứu kinh điển về ứng xử cắt của dầm BTCT và các mô hình phi tuyến tuy đóng góp quan trọng cho hiểu biết cơ bản, nhưng chúng vẫn phụ thuộc nhiều vào tập thông số cố định và khó thích ứng với các tổ hợp vật liệu phức tạp và biến thiên ngẫu nhiên [5], [6].

Để khắc phục hạn chế đó, các nghiên cứu gần đây đã ứng dụng các mô hình tính toán tiên tiến và phương pháp phân tích dựa trên dữ liệu vào phân tích dầm BTCT được tăng cường bởi bê tông UHPC. Các mô hình học máy (Machine learning) cho phép xác định mối quan hệ tốt hơn ứng xử của kết cấu hybrid UHPC và bê tông thường [7]. Đồng thời, các phương pháp học máy và học sâu đã chứng minh tiềm năng vượt trội trong việc nắm bắt mối quan hệ phi tuyến giữa đặc tính vật liệu, thông số hình học và khả năng chịu cắt cực hạn dựa trên cơ sở dữ liệu thí nghiệm [8]. Các mô hình ML dựa trên dữ liệu mang lại một hướng tiếp cận mới đầy hứa hẹn nhằm nâng cao độ chính xác và tính ổn định của dự báo sức kháng cắt của dầm BTCT được tăng cường bằng UHPC.

Tuy nhiên, thách thức lớn đặt ra là số lượng thí nghiệm về dầm BTCT tăng cường UHPC còn hạn chế do chi phí vật liệu cao, yêu cầu gia công mẫu phức tạp và quy trình thử nghiệm tương đối tốn kém. Điều này dẫn đến cơ sở dữ liệu thí nghiệm về vấn đề này nhỏ, gây ảnh hưởng đáng kể đến khả năng dự báo cũng như tổng quát hóa của các mô hình học máy. Nhằm cải thiện vấn đề này, các nghiên cứu gần đây đã quan tâm nhiều hơn đến các kỹ thuật mở rộng dữ liệu (data augmentation) hay còn gọi là tăng cường dữ liệu như một giải pháp hiệu quả để làm giàu các bộ dữ liệu nhỏ mà vẫn bảo toàn được bản chất thống kê của chúng [9]. Trong số đó, các phương pháp tạo sinh dữ liệu dựa trên phân phối Gaussian được xem là một kỹ thuật đơn giản, nhờ khả năng mở rộng dữ liệu mới dựa trên cấu trúc trung bình – phương sai của dữ liệu gốc, đồng thời hạn chế tạo ra các mẫu không thực tế [10]. Phương pháp này đã được sử dụng trong nghiên cứu gần đây nhằm tăng cường dữ liệu để cải thiện thuật toán PSO-ANN để dự báo dự báo dầm BTCT DUL [11]. Ngoài thuật toán Gaussian còn một số nghiên cứu sử dụng TGAN [9], … để tăng cường dữ liệu. Trong nghiên cứu này chúng tôi đi sâu vào phân tích việc sử dụng Gauss để tăng cường dữ liệu.

Nhằm đánh giá tính hiệu quả của phương pháp tạo sinh dữ liệu theo Gaussian trong dự báo khả năng chịu cắt của dầm BTCT được tăng cường bởi bê tông UHPC, nghiên cứu này được

triển khai theo ba bước chính như sau:

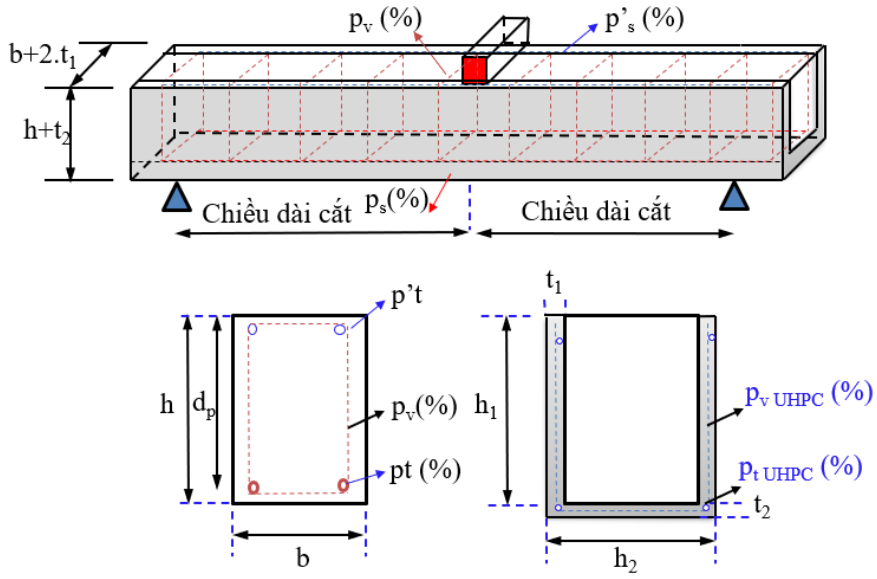
- *Bước 1 – Xây dựng cơ sở dữ liệu thí nghiệm:* Bộ dữ liệu gốc (Dữ liệu A) gồm 69 mẫu được tổng hợp, phản ánh ứng xử chịu cắt của dầm BTCT được tăng cường bởi bê tông siêu tính năng UHPC.
- *Bước 2 – Tạo sinh dữ liệu theo Gaussian:* Phương pháp Gaussian được áp dụng để tạo thêm 276 mẫu dữ liệu tổng hợp, hình thành tập dữ liệu mới (Dữ liệu B) nhằm mở rộng không gian dữ liệu mà vẫn duy trì đặc tính thống kê của bộ dữ liệu gốc.
- *Bước 3 – Xây dựng phương án mô hình hóa:* hai phương án được đề xuất để đánh giá vai trò của việc tăng cường dữ liệu như sau:
  - ✓ Trường hợp 1: Huấn luyện & kiểm tra một số mô hình học máy trên Dữ liệu A.
  - ✓ Trường hợp 2**:** Huấn luyện mô hình trên Dữ liệu B và kiểm tra bằng Dữ liệu A để đánh giá khả năng khái quát hóa của dữ liệu mới được tạo sinh.

Trong mỗi trường hợp, dữ liệu được chia ngẫu nhiên theo tỷ lệ 80% dữ liệu để phục vụ huấn luyện và 20% dữ liệu còn lại sử dụng để kiểm tra. Tối ưu các siêu tham số của các mô hình học máy được thực hiện bằng GridSearch kết hợp 5-fold cross-validation. Cuối cùng, hiệu suất dự báo của bốn thuật toán học máy — Random Forest (RF), k-Nearest Neighbors (k-NN), XGBoost và LightGBM — được đánh giá và so sánh trong mỗi trường hợp.

2. PHƯƠNG PHÁP NGHIÊN CỨU

2.1. Thu thập dữ liệu từ các nghiên cứu thực nghiệm

Một tập dữ liệu gồm 69 mẫu thí nghiệm đã được tổng hợp dựa trên 08 công bố khoa học trước đây [12], [13], [14], [15], [16], [17], [18], [19]. Một quy trình kỹ thuật tiền xử lý và xây dựng đặc trưng (feature engineering) được áp dụng nhằm thiết lập các biến đầu vào, sau cùng 21 biến đầu vào được lựa chọn và 01 biến đầu ra.



Hình 1. Sơ họa các tham số dầm BTCT được tăng cường khả năng chịu cắt bởi bê tông UHPC.

Các mẫu thí nghiệm được thu thập thể hiện sự biến thiên ở nhiều thông số quan trọng, bao gồm kích thước dầm, hàm lượng cốt thép dọc và cốt đai, chiều dày lớp UHPC tăng cường,

cường độ nén của UHPC và cường độ chịu nén của bê tông thường, cường độ chảy của thép cốt, cũng như hàm lượng và tỷ số hình học của sợi thép trong bê tông UHPC. Hình 1 trình bày sơ đồ minh họa cấu tạo điển hình của một dầm BTCT được tăng cường bằng UHPC cùng các thông số hình học – cơ học đặc trưng, trong khi Bảng 1 tóm tắt chi tiết toàn bộ dữ liệu thí nghiệm được tổng hợp sử dụng trong nghiên cứu.

Trong nghiên cứu này, một bộ dữ liệu đã được xây dựng dựa trên các thí nghiệm về dầm BTCT được tăng cường bởi UHPC. Hai mươi hai tham số đầu vào đại diện được xem xét như các biến dự báo khả năng chịu cắt cực hạn (V_u) của dầm BTCT tăng cường UHPC. Các biến số liên tục bao gồm: chiều rộng dầm (b), chiều cao tiết diện (h), chiều dài cao hữu hiệu (d_p), tỷ số giữ chiều dài cắt và chiều cao hữu hiệu (a/d), cường độ nén của bê tông thường (f'c), tỷ lệ cốt thép dọc ở dưới mặt cắt (p_t) tỷ lệ cốt thép dọc ở trên mặt cắt (p'_t)và tỷ lệ cốt đai (p_v) trong vùng bê tông thường, giới hạn chảy cốt thép thường dọc dưới, trên và cốt thép đai (fyt, fyt',fyv), kích thước UHPC tăng cường theo chiều cao và chiều rộng rộng dầm (h_1, t_1, h_2, t_2), diện tích tiết diện UHPC tăng cường (S_{UHPC}), tỷ lệ cốt thép dọc ($p_{t,UHPC}$) và tỷ lệ cốt đai ($p_{v\ UHPC}$) trong lớp UHPC tăng cường, cường độ chảy của cốt thép dọc ($f_{yt\ UHPC}$) và cốt đai ($f_{yv\ UHPC}$) của cốt thép trong lớp UHPC, cường độ nén UHPC ($f'c_{\ UHPC}$) như minh họa ở Hình 3.

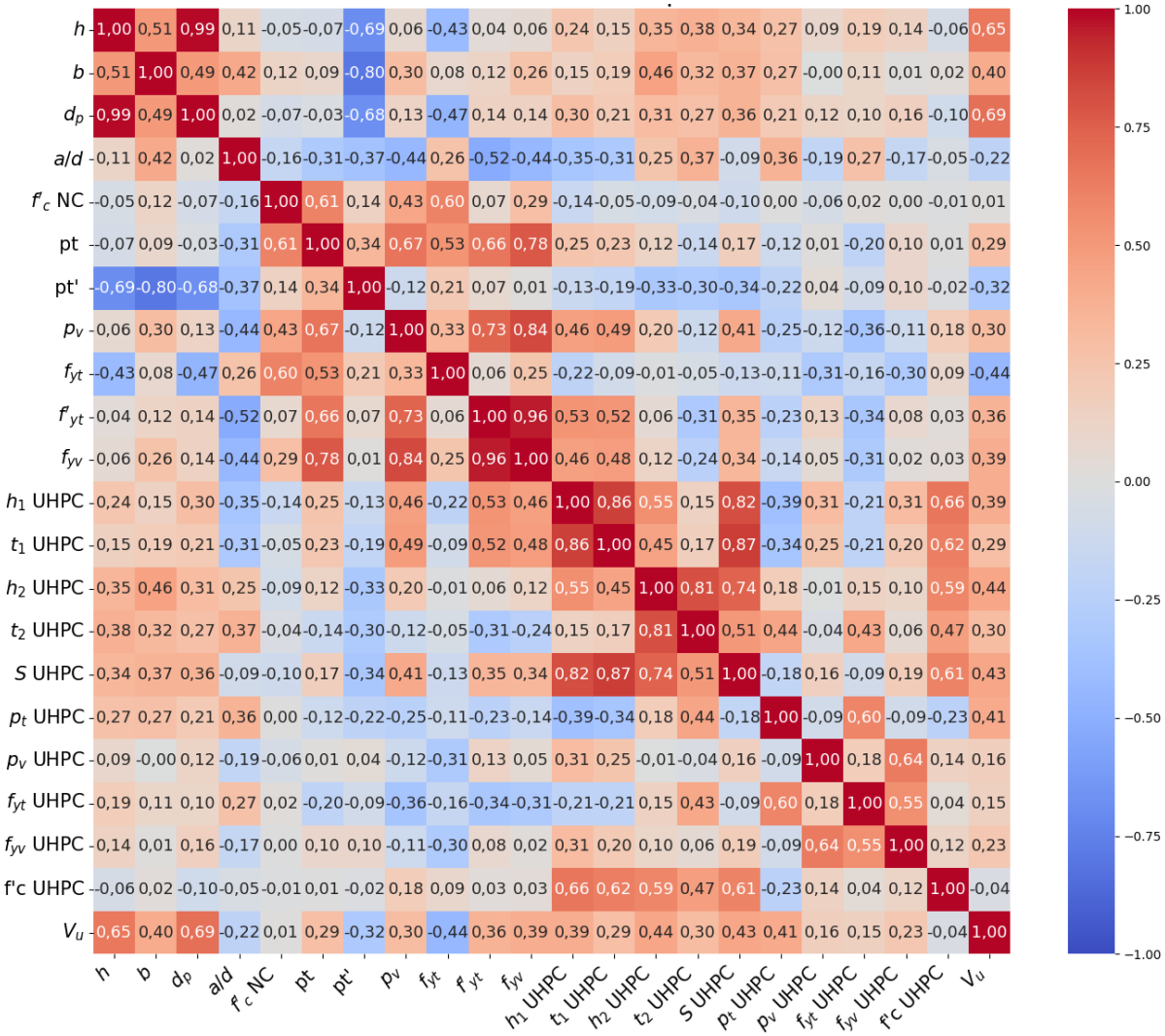
Bảng 1. Thông tin thống kê của các tham số của 69 mẫu thí nghiệm.

No	Feature	Type	Units	Min	Max	Mean	Median	Std
1	h	X1	mm	193	690	337,54	300	109,12
2	b	X2	mm	82	210	155,94	150	32,62
3	d_p	X3	mm	167	540	291,62	275	98,25
4	a/d	X4	-	0,97	3,65	2,44	3	0,87
5	$fc'_{\ NC}$	X5	MPa	20,1	65	39,29	36,3	12,49
6	p_t	X6	%	0,21	4.49	2,28	2,72	1,16
7	pt'	X7	%	0,19	1.15	0,51	0,52	0,25
8	p_v	X8	%	0,00	0.67	0,25	0,20	0,23
9	fyt	X9	MPa	420	610	519,25	537	54,73
10	fyt'	X10	MPa	0	610	383,30	443	236,64
11	fyv'	X11	MPa	0	610	322,72	340	214,88
12	$h_{1\ UHPC}$	X12	mm	0	425	174,77	193	183,28
13	$t_{1\ UHPC}$	X13	mm	0	60	17,25	10	20,06
14	$h_{2\ UHPC}$	X14	mm	0	320	127,77	150	107,91
15	$t_{2\ UHPC}$	X15	mm	0	90	25,72	30	23,25
16	$S_{\ UHPC}$	X16	mm^2	0	67200	17058,12	10000	18264,82
17	$p_{t\ UHPC}$	X17	%	0	9.42	0,86	0	2,19
18	$p_{v\ UHPC}$	X18	%	0	2.39	0,12	0	0,42
19	$f_{yt\ UHPC}$	X19	MPa	0	1550	127,23	0	272,54
20	$f_{yv\ UHPC}$	X20	MPa	0	1550	53,98	0	215,24
21	fc'_{UHPC}	X21	MPa	106,5	158,4	139,02	135,37	13,68
22	V_u	Y	kN	16,9	955	264,94	191,5	229,88

Bộ dữ liệu thí nghiệm ban đầu thể hiện sự biến thiên đáng kể giữa các thông số hình học, cơ học vật liệu và đặc trưng cốt thép của dầm. Hầu hết các biến đầu vào đều có phân bố không đồng đều và có xu hướng lệch nhẹ, phản ánh các lựa chọn thiết kế rời rạc và các ràng buộc trong quá trình thực hiện thí nghiệm hơn là sự biến thiên ngẫu nhiên thuần túy. Các đặc tính vật liệu và tỷ lệ cốt thép chủ yếu tập trung trong những khoảng giá trị tương đối hẹp, trong khi chỉ

một số ít mẫu đại diện cho các cấu hình cường độ cao hoặc mức độ cốt thép lớn (Hình 3).

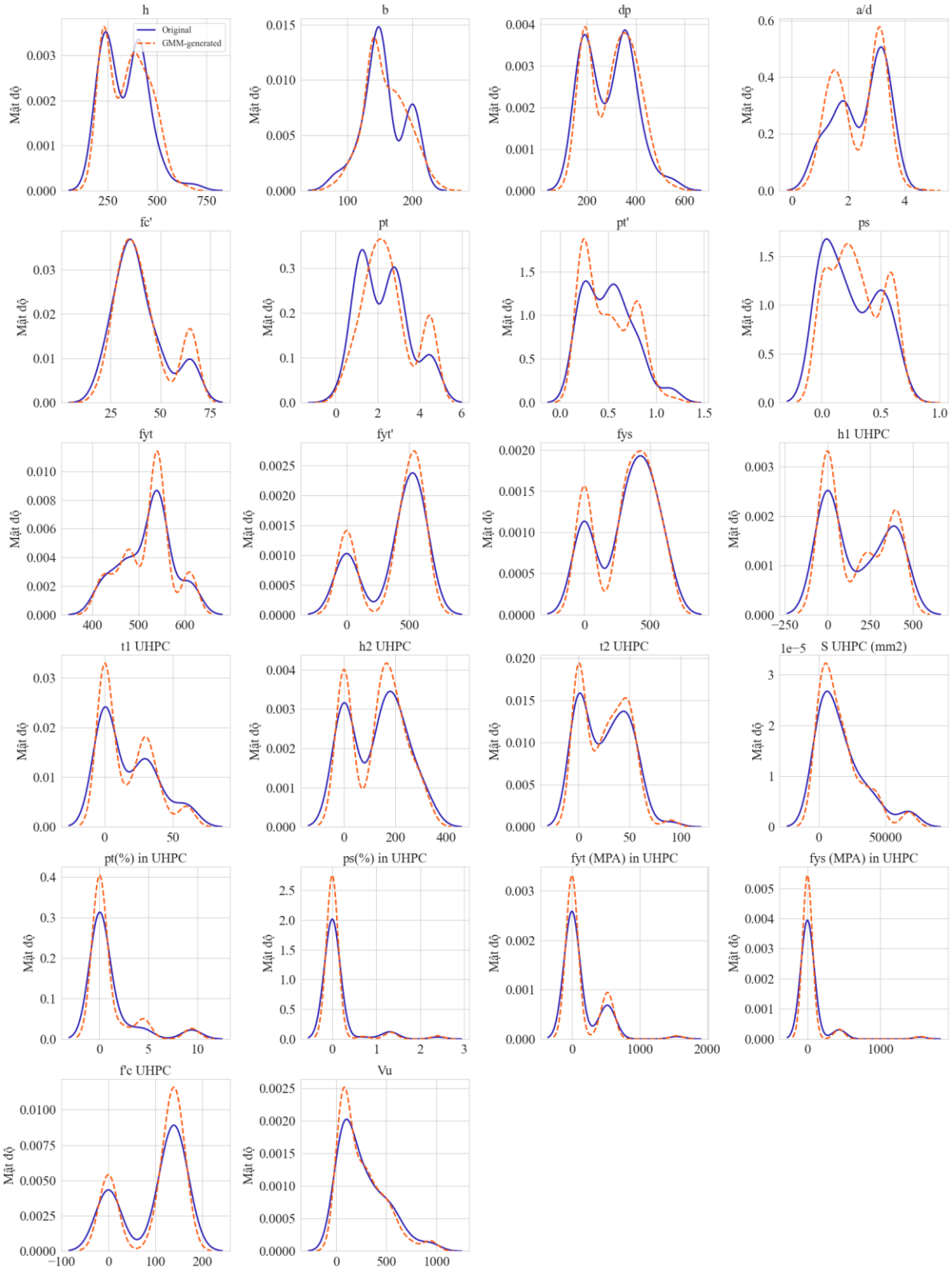
Biến đầu ra – khả năng chịu cắt cực hạn (V_u) – trải rộng trên một phạm vi lớn thay đổi từ 16,9 kN đến 955 kN, cho thấy sự đa dạng của khả năng chịu cắt do ảnh hưởng đồng thời của nhiều yếu tố tương tác. Các đặc điểm phân bố này phản ánh phạm vi dữ liệu thí nghiệm còn hạn chế và sự không đồng nhất vốn có của các mẫu thử, từ đó nhấn mạnh nhu cầu áp dụng các kỹ thuật tăng cường hay mở rộng dữ liệu (data augmentation). Việc mở rộng dữ liệu giúp tăng độ bao phủ của không gian đặc trưng, đồng thời cải thiện độ ổn định và khả năng khái quát của các mô hình học máy được xây dựng trong các bước tiếp theo.



Hình 2. Ma trận tương quan dữ liệu thu thập được.

Hình 2 thể hiện mối quan hệ tương quan giữa các biến, cũng như các biến ảnh hưởng đến khả năng chịu cắt của dầm sau khi được tăng cường bởi UHPC. Các biến có ảnh hưởng lớn nhất đến khả năng chịu cắt của dầm sau khi được tăng cường bởi UHPC lần lượt là chiều cao dầm (0,65); chiều cao hưu hiệu (dp); diện tích S_{UHPC} bổ sung, bề rộng dầm phù hợp với các tham số ảnh hưởng đến khả năng chịu cắt của dầm BTCT.

2.2 Tạo sinh dữ liệu thông qua thuật toán Gaussian Mixture Model (GMM)



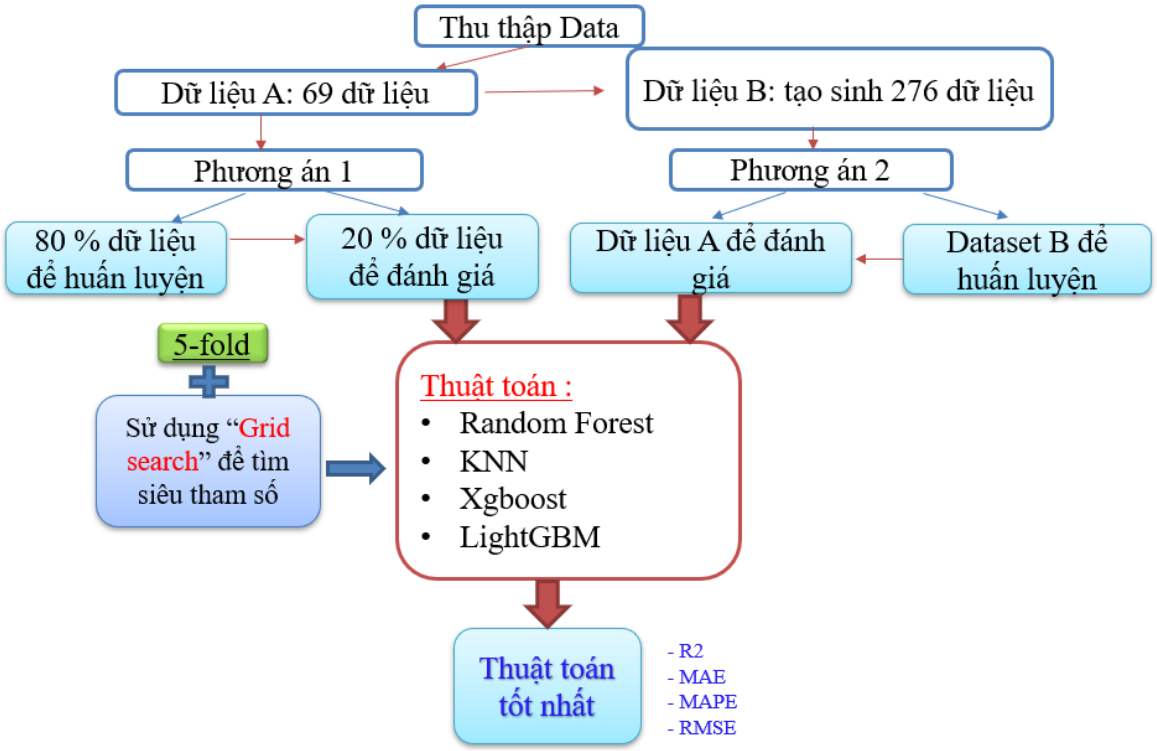
Hình 3. So sánh phân phối mật độ xác suất giữa tập dữ liệu gốc và tập dữ liệu được tạo bởi GMM.

Xuất phát từ thiếu các kết quả thực nghiệm và đặc tính rời rạc của bộ dữ liệu ban đầu, mô hình hỗn hợp Gaussian (Gaussian Mixture Model – GMM) hoặc phương pháp mở rộng dữ liệu dựa trên phân phối Gaussian được áp dụng trong nghiên cứu này nhằm tạo thêm các mẫu dữ liệu mới trong các khoảng giá trị hợp lý về mặt thống kê, qua đó cải thiện độ ổn định và độ tin cậy của các mô hình dự báo [5].

Trong nghiên cứu này, GMM được sử dụng để phát sinh thêm 276 mẫu dữ liệu tổng hợp, gấp 4 lần số lượng dữ liệu thực nghiệm thu thập được. Việc mở rộng này giúp tăng kích thước bộ dữ liệu lên mức đủ lớn để đánh giá đáng tin cậy 69 mẫu thí nghiệm ban đầu, đồng thời nâng cao khả năng sử dụng mô hình học máy khái quát hóa ứng xử chịu cắt của dầm BTCT được tăng cường bằng UHPC. Phân bố dữ liệu trước và sau khi mở rộng dữ liệu được minh họa trong Hình 3. Do việc phân bố dữ liệu gốc như trong Hình 3 có 2-3 đỉnh thay đổi. Nên mô hình GMM trong nghiên cứu này tham số của mô hình Gaussian được lựa chọn như sau: Số lượng thành phần (n_components = 4): Dữ liệu được giả định là sự kết hợp của 4 phân phối chuẩn (Gaussian) thành phần. Việc lựa chọn K=4 cho phép mô hình linh hoạt trong việc bắt kịp các đặc trưng đa đỉnh (multimodal) của dữ liệu thực tế; Loại hiệp phương sai (covariance_type = "full") cho phép mỗi thành phần Gaussian sở hữu một ma trận hiệp phương sai riêng biệt và đầy đủ. Nghiên cứu ảnh hưởng của các loại tham số Gausian đến độ chính xác của việc học sẽ được thực hiện ở các nghiên cứu khác.

Hình 3 cho thấy sự tương đồng chặt chẽ giữa các phân bố mật độ xác suất của dữ liệu gốc và dữ liệu tạo sinh từ GMM, chứng tỏ mô hình GMM có khả năng bảo toàn đặc trưng thống kê của dữ liệu thí nghiệm và phù hợp để sử dụng cho mục đích mở rộng dữ liệu trong các bài toán học máy.

2.3 Trình tự huấn luyện và tính toán



Hình 4. Sơ đồ phương án nghiên cứu.

Dựa trên các bộ dữ liệu đã xây dựng, hai phương án được triển khai nhằm đánh giá một cách có hệ thống hiệu quả của phương pháp tăng cường dữ liệu dựa trên phương pháp Gaussian (Hình 4). Tập Dữ liệu A bao gồm 69 mẫu thí nghiệm gốc, trong khi tập dữ liệu B chứa 276 mẫu dữ liệu được tạo sinh bằng phương pháp Gaussian. Đối với tất cả các mô hình học máy, dữ liệu trong mỗi phương án được chia theo tỷ lệ 80% cho huấn luyện và 20% cho kiểm tra và đánh giá.

• Phương án 1 sử dụng duy nhất dữ liệu A, trong đó các mô hình được huấn luyện và đánh giá dựa trên dữ liệu thí nghiệm gốc.

• Phương án 2 sử dụng Dữ liệu B cho quá trình huấn luyện mô hình, sau đó mô hình đã huấn luyện được kiểm tra, đánh giá Dữ liệu A.

Cấu trúc so sánh này cho phép đánh giá toàn diện ảnh hưởng của phương án tạo sinh dữ liệu Gaussian đến độ chính xác và khả năng tổng quát của mô hình học máy, ảnh hưởng đến độ chính xác của việc dự báo.

Để đảm bảo sự so sánh công bằng và đáng tin cậy, việc tìm các siêu tham số (hyperparameter optimization) được thực hiện cho tất cả các mô hình học máy bằng phương pháp GridSearch kết hợp với kiểm định chéo 5 lần (5-fold cross-validation) trên tập huấn luyện. Quy trình này nhằm tìm ra bộ siêu tham số tối ưu giúp giảm sai lệch mô hình (bias), giảm phương sai (variance), đồng thời cải thiện khả năng khái quát hóa.

Trong nghiên cứu này chúng tôi sử dụng 04 thuật toán học máy phổ biến để đánh giá và so sánh. Các thuật toán được sử dụng là: Random Forest, KNN, XGBoost và LightGBM.

2.3.1. Random Forest Regressor (RF)

Random Forest Regressor (RFR) [20] là một kỹ thuật học máy dạng tổ hợp (ensemble learning) phổ biến, được xây dựng dựa trên việc kết hợp nhiều cây quyết định để cải thiện độ chính xác dự báo. Khác với các mô hình cây đơn lẻ hoặc các biến thể cực đoan khác, RFR tìm kiếm sự cân bằng bằng cách lựa chọn ngưỡng chia tối ưu dựa trên một tập con các đặc trưng được chọn ngẫu nhiên tại mỗi nút. Cơ chế này giúp giảm bớt sự phụ thuộc vào các biến thể mạnh, từ đó giảm phương sai và tăng tính ổn định cho mô hình.

Trong bài toán hồi quy, giá trị dự báo cuối cùng được xác định bằng phương pháp Bagging (Bootstrap Aggregating), tức là lấy trung bình kết quả của toàn bộ các cây thành phần. Bằng cách kết hợp giữa việc lấy mẫu dữ liệu có thay thế và tối ưu hóa cục bộ tại các nút, RFR đạt được khả năng khái quát hóa vượt trội và kiểm soát hiện tượng học thuộc lòng (overfitting) một cách hiệu quả. Nhờ cấu trúc mạnh mẽ và ít nhạy cảm với các tham số mặc định, RFR đã trở thành một công cụ tiêu chuẩn và đáng tin cậy trong việc giải quyết các bài toán kỹ thuật và khoa học dữ liệu có cấu trúc phức tạp.

2.2.2. K-Nearest Neighbors (KNN)

Thuật toán k-Nearest Neighbors (KNN) [21] là một phương pháp học dựa trên ví dụ (instance-based learning), trong đó quá trình dự đoán được thực hiện thông qua mức độ tương đồng giữa các mẫu trong không gian đặc trưng. Khác với các mô hình có giai đoạn huấn luyện rõ ràng, KNN không xây dựng mô hình tường minh mà thay vào đó xác định k mẫu lân cận gần nhất của điểm cần dự đoán dựa trên một thước đo khoảng cách đã chọn (thường dùng khoảng cách Euclid, Minkowski hoặc Manhattan).

Đối với bài toán hồi quy, giá trị dự đoán được tính bằng trung bình đáp ứng của các điểm lân cận. Nhờ bản chất phi tham số, KNN có khả năng mô tả tốt các quan hệ phi tuyến phức tạp trong dữ liệu. Tuy nhiên, hiệu suất của KNN phụ thuộc đáng kể vào việc lựa chọn tham số k, phương pháp chuẩn hóa đặc trưng, và mật độ phân bố dữ liệu; trong đó dữ liệu thưa hoặc có nhiều nhiễu có thể làm suy giảm độ chính xác của mô hình

2.2.3. Extreme Gradient Boosting (XGBoost)

XGBoost [22] là một thuật toán tăng cường (gradient boosting) tiên tiến, xây dựng một tập hợp (ensemble) các cây quyết định theo từng bước tuần tự, trong đó mỗi cây mới được huấn luyện để hiệu chỉnh sai số còn lại của mô hình trước đó. Thuật toán này tích hợp các thành phần regularization trực tiếp vào hàm mục tiêu nhằm kiểm soát độ phức tạp của mô hình và giảm hiện tượng quá khớp.

Nhờ cơ chế xây dựng cây hiệu quả, khả năng xử lý song song và tối ưu hóa quá trình tối thiểu hóa hàm mất mát, XGBoost mang lại độ chính xác dự báo cao và khả năng mở rộng tốt. Những đặc điểm này đặc biệt phù hợp khi xử lý dữ liệu dạng bảng trong các bài toán hồi quy kỹ thuật.

2.2.3. LighGBM

LightGBM (Light Gradient Boosting Machine) [23] là một thuật toán học máy nâng cao dựa trên cấu trúc cây quyết định (GBDT), được tối ưu hóa cho hiệu suất và tốc độ xử lý dữ liệu lớn. Khác với các mô hình truyền thống, LightGBM sử dụng chiến lược phát triển cây theo chiều dọc (Leaf-wise), tập trung tối ưu hóa các nút lá có mức giảm tổn thất cao nhất để tăng độ chính xác. Hai kỹ thuật đột phá là GOSS (lấy mẫu dựa trên độ dốc) và EFB (gom cụm đặc trưng) giúp mô hình giảm đáng kể thời gian huấn luyện và mức tiêu thụ bộ nhớ mà vẫn kiểm soát tốt hiện tượng quá khớp (overfitting). Đây là công cụ mạnh mẽ trong việc xử lý các bài toán phi tuyến phức tạp với độ ổn định cao.

2.2.3. Đánh giá hiệu quả mô hình

Để đánh giá năng lực dự báo của các mô hình học máy được đề xuất, bốn chỉ số thống kê được sử dụng, bao gồm hệ số xác định (R^2), căn bậc hai của sai số bình phương trung bình (RMSE), sai số phần trăm tuyệt đối trung bình (MAPE) và sai số tuyệt đối trung bình (MAE). Các chỉ số này được công nhận rộng rãi trong việc đánh giá hiệu suất hồi quy trong các ứng dụng kỹ thuật kết cấu. Nhìn chung, giá trị R^2 càng lớn và tiến gần đến 1 cho thấy mức độ tương quan mạnh giữa giá trị dự đoán và giá trị thực nghiệm, trong khi các giá trị nhỏ của RMSE, MAPE và MAE phản ánh độ chính xác dự báo cao hơn và sai số ước lượng giảm xuống. Các công thức tính toán của các chỉ số đánh giá này được trình bày dưới đây:

$$R_2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_l)^2} \quad (1)$$

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

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} \quad (3)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (4)$$

3. KẾT QUẢ THÍ NGHIỆM VÀ ĐÁNH GIÁ

3.1. Đánh giá hiệu năng của các mô hình

Các siêu tham số tối ưu thu được cho từng mô hình được tổng hợp trong Bảng 2, trong đó tổng hợp giá trị của các siêu tham số cho hiệu suất tốt nhất trong quá trình Grid Search. Các thiết lập tham số này phản ánh đặc trưng riêng của từng thuật toán và nhấn mạnh vai trò quan trọng của tối ưu siêu tham số trong việc nâng cao độ chính xác dự báo. Sau giai đoạn tối ưu, hiệu suất dự đoán của tất cả các mô hình được đánh giá và so sánh bằng bốn chỉ số thống kê: R^2, MAE, (MAPE) và RMSE. Hiệu suất dự đoán của bốn mô hình học máy theo Trường hợp I (chỉ sử dụng Dữ liệu A) và Trường hợp 2 (huấn luyện trên dữ liệu B và kiểm tra trên Dữ liệu A) lần lượt được tổng hợp trong Bảng 3 và Bảng 4. Sự khác biệt rõ rệt về độ chính xác của mô hình học máy và khả năng khái quát hóa có thể được đánh giá khi so sánh số liệu của hai trường hợp này.

Bảng 2. Siêu tham số tối ưu của các mô hình học máy.

No	**Model**	**Siêu tham số tối ưu của các mô hình học máy**		
		Phương án 1		**Phương án 2**
1	Random Forest	max_depth	5	7
		max_features	sqrt	0.8
		min_samples_leaf	1	1
		min_samples_split	2	2
		n_estimators	500	300
2	KNN	n_neighbors’:	5	5
		p	1	1
		Weights	uniform	uniform
3	XGboost	Learning rate	0.05	0.05
		max_depth	2	3
		n_estimators	600	400
		Subsample	0.7	0.8
		colsample_bytree	0.7	0.9
		reg_alpha	0.1	0.5
		reg_lambda	10	1
4	Light GBM	Max_depth	2	2
		Min_child_sample	10	20
		n_estimators	300	600
		Subsample	0.7	0.7
		Learning rate	0.05	0.05

Bảng 3. Hiệu suất dự đoán trung bình của các mô hình theo Phương án 1.

STT	**Mô hình**	**Tập dữ liệu để huấn luyện**				**Tập dữ liệu để test**			
		R2	**MAE**	**MAPE**	**RMSE**	**R2**	**MAE**	**MAPE**	**RMSE**
1	RF	0.977	23.342	16.107	33.863	0.873	40.504	26.040	56.368
2	KNN	0.879	47.718	24.966	77.858	0.825	40.810	24.381	60.083

3	Light GBM	0.975	24.093	13.199	35.660	0.850	46.906	28.724	67.447
4	XGboost	0.984	17.403	9.382	28.241	0.874	36.558	24.816	54.737

Bảng 4. Hiệu suất dự đoán trung bình của các mô hình theo Phương án 2.

STT	**Mô hình**	**Tập dữ liệu để huấn luyện**				**Tập dữ liệu để test**			
		R2	**MAE**	**MAPE**	**RMSE**	**R2**	**MAE**	**MAPE**	**RMSE**
1	RF	0.991	15.424	12.978	21.514	0.948	40.504	26.040	56.368
2	KNN	0.895	44.918	35.168	71.839	0.791	40.810	24.381	60.083
3	Light GBM	0.978	25.438	17.042	33.135	0.921	46.906	28.724	67.447
4	XGboost	0.997	8.833	8.014	11.400	0.949	36.558	24.816	54.737

Đối với phương án 1, tất cả các mô hình đều thể hiện độ chính xác huấn luyện rất cao, đặc biệt là các thuật toán dạng tập hợp (ensemble) như XGBoost và Random Forest (RF). Hiệu suất huấn luyện: XGBoost đạt hiệu suất ấn tượng nhất trên tập huấn luyện với R^2 = 0,984, đi kèm với sai số thấp nhất trong bảng so sánh (MAE = 17,403 và RMSE = 28,241). Random Forest và LightGBM cũng cho thấy khả năng học mạnh mẽ với giá trị R^2 lần lượt là 0,977 và 0,975. Hiệu suất kiểm tra (Test set): Có sự suy giảm về hiệu suất khi chuyển từ tập huấn luyện sang tập kiểm tra ở tất cả các mô hình, cho thấy sự xuất hiện của hiện tượng quá khớp (overfitting). Tuy nhiên, XGBoost vẫn giữ vững vị thế là mô hình hiệu quả nhất trên tập kiểm tra với R^2 = 0,874 và sai số MAE thấp nhất (36,558). Random Forest bám sát phía sau với R^2 kiểm tra đạt 0,873.Mô hình KNN: Ngược lại, KNN mặc dù có sự chênh lệch giữa hai tập dữ liệu thấp hơn (ít bị overfitting hơn) nhưng lại có hiệu suất tổng thể thấp nhất trong nhóm với R^2 kiểm tra đạt 0,825 và sai số RMSE lên tới 60,083.Nhìn chung, kết quả từ bảng dữ liệu cho thấy mặc dù các mô hình ensemble vẫn duy trì được khả năng khái quát hóa ở mức khá (R^2 > 0,85 đối với RF và XGBoost), nhưng sự chênh lệch sai số giữa hai tập dữ liệu gợi ý rằng việc tinh chỉnh siêu tham số hoặc cải thiện chất lượng dữ liệu đầu vào có thể giúp tối ưu hóa hơn nữa khả năng dự báo.

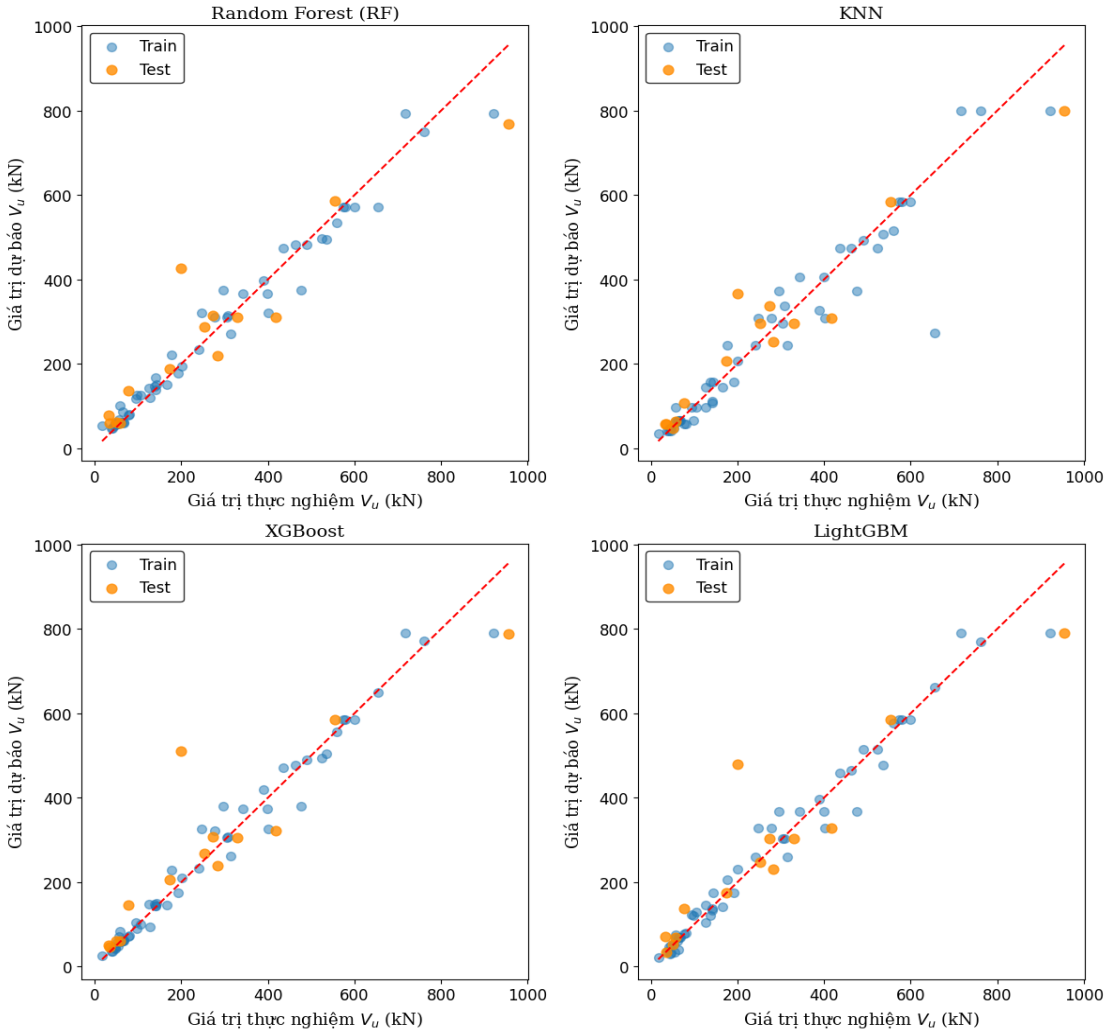
Ngược lại, kết quả từ phương án 2 (Bảng 4) cho thấy sự cải thiện vượt bậc về khả năng khái quát hóa ở tất cả các mô hình khi được huấn luyện bằng bộ dữ liệu tổng hợp và đánh giá trên tập kiểm tra. Trừ thuật toán KNN, các mô hình còn lại đều đạt giá trị R^2 kiểm tra rất cao (đều lớn hơn 0,92), thể hiện mức cải thiện vượt trội so với phương án 1. Trong số các thuật toán, XGBoost cho thấy hiệu suất ấn tượng nhất trên tập kiểm tra với R^2 = 0,949, đi kèm với sai số RMSE thấp nhất (54,737) và MAE thấp nhất (36,558). Random Forest (RF) theo sát phía sau với chỉ số R^2 đạt 0,948, trong khi LighGBM cũng cho thấy khả năng dự báo cực kỳ ổn định với R^2 = 0,921. Đáng chú ý, mặc dù Random Forest và XGBoost đạt mức độ chính xác gần như tuyệt đối trên tập huấn luyện (với R^2 lần lượt là 0,997 và 0,991), chúng vẫn duy trì được hiệu suất cao trên tập kiểm tra mà không gặp hiện tượng quá học thuộc lòng (overfitting) nghiêm trọng như phương án 1. KNN hoạt động không tốt đối với phương án 2.

Như vậy, khi so sánh giữa phương án 1 và phương án 2 cho thấy việc tạo sinh dữ liệu sử dụng GMM đối với cải thiện khả năng khái quát hóa của mô hình. Phương án 1 phản ánh hiện tượng huấn luyện các mẫu dữ liệu hạn chế, dẫn đến hiệu suất kém trên tập kiểm tra. Trong khi đó, phương án 2 chứng minh rằng việc huấn luyện trên một bộ dữ liệu tạo sinh có kích thước đủ lớn và đa dạng giúp cải thiện đáng kể tính ổn định và độ tin cậy của dự đoán. Về mặt ứng dụng thực tiễn, sự so sánh này xác nhận tính khả thi của việc sử dụng các tập tạo sinh như một

nguồn huấn luyện hiệu quả trong các trường hợp dữ liệu thực nghiệm khan hiếm, chẳng hạn như trong các bài toán kỹ thuật kết cấu, việc thí nghiệm tốn kém và khó có thể làm với số lượng lớn.

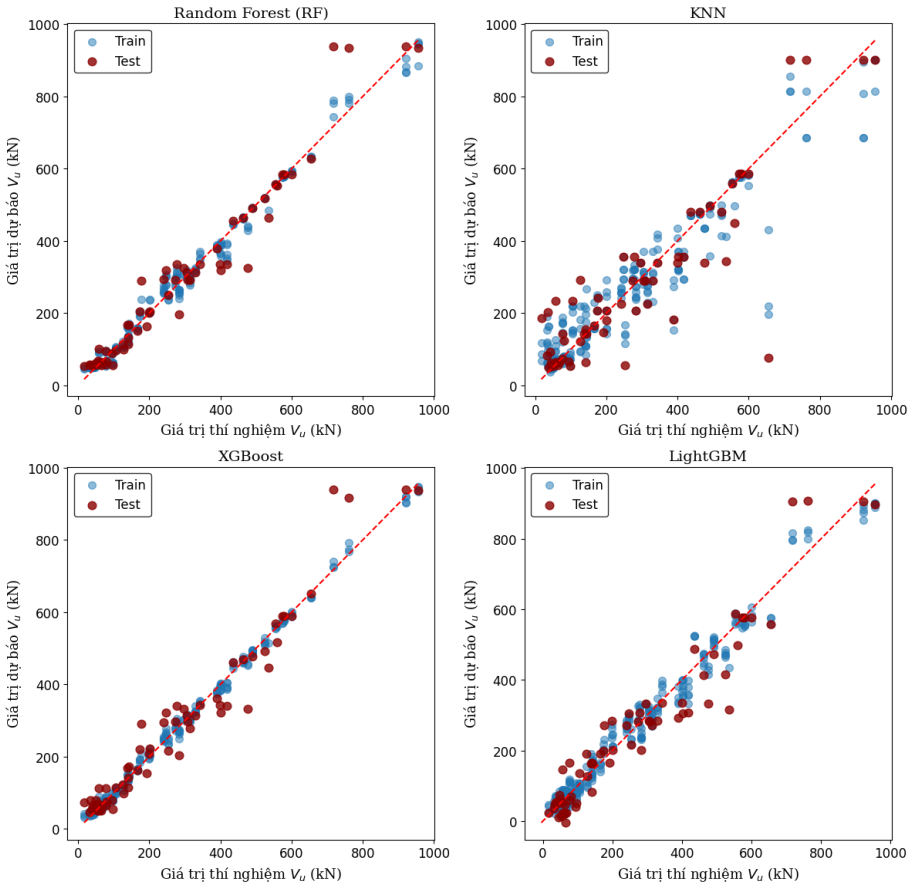
3.2. So sánh giá trị thực tế và dự báo V_u

Hình 5 minh họa mối quan hệ giữa khả năng chịu cắt lớn nhất (V_u) dự đoán và thực nghiệm theo Phương án 1, trong đó tất cả các mô hình đều được huấn luyện và kiểm tra chỉ trên bộ dữ liệu thực nghiệm gốc (Dữ liệu A). Có thể quan sát rõ sự khác biệt rõ rệt giữa hiệu suất huấn luyện và kiểm tra.



Hình 5. Mối quan hệ giữa giá trị thực nghiệm và giá trị dự đoán của bốn mô hình học máy theo phương án 1.

Hình 6 trình bày mối quan hệ giữa giá trị dự đoán và giá trị thực nghiệm theo phương án 2, trong đó các mô hình được huấn luyện trên bộ dữ liệu tổng hợp tạo bởi mô hình GMM (Dữ liệu B) và đánh giá trên dữ liệu thực nghiệm (Dữ liệu A). So với phương án 1, ngoại trừ KNN, các mô hình đều cho thấy sự cải thiện rõ rệt về khả năng khái quát hóa và dự báo.



Hình 6. Mối quan hệ giữa giá trị thực nghiệm và giá trị dự đoán của bốn mô hình học máy theo phương án 2.

Đối chiếu Hình 5 và Hình 6 cho thấy phương án tăng cường dữ liệu giúp cho mô hình XGBoost, LighGBM, Random Forest có khả năng dự báo tốt hơn. Mô hình KNN không phù hợp với tập dữ liệu này.

4. KẾT LUẬN VÀ KIẾN NGHỊ

Nghiên cứu này đã chứng minh hiệu quả của tăng cường dữ liệu dựa trên phân phối Gaussian trong việc nâng cao độ chính xác dự đoán và khả năng khái quát hóa của các mô hình học máy trong bài toán dự đoán khả năng chịu cắt lớn nhất của dầm bê tông cốt thép được tăng cường bằng bê tông UHPC.

So sánh hai phương án huấn luyện cho thấy các mô hình được huấn luyện và kiểm tra chỉ trên bộ dữ liệu gốc (Phương án 1) dễ gặp hiện tượng học thuộc lòng (overfitting), với giá trị R^2 kiểm tra dao động từ 0,825 đến 0,874. Mặc dù đây là mức khá, nhưng sai số dự đoán vẫn còn tương đối cao với RMSE lên tới 67,447. Ngược lại, huấn luyện trên bộ dữ liệu được tăng cường (Phương án 2) và kiểm tra trên dữ liệu thực nghiệm gốc đem lại cải thiện đáng kể. Ngoại trừ thuật toán KNN, Các thuật toán còn lại có chỉ số R^2 tăng lên, giá trị đạt 0,921 – 0,949. Các chỉ số sai số khác giảm xuống rõ rệt, phản ánh khả năng khái quát hóa và độ ổn định dự đoán được nâng cao vượt trội. XGBoost đạt hiệu suất tổng thể tốt nhất sau khi tăng cường dữ liệu với R^2

kiểm tra = 0,949, đồng thời sở hữu sai số thấp nhất nhóm (MAE = 36,558 và RMSE = 54,737). Random Forest (RF) theo sát với R^2 kiểm tra đạt 0,948.

Điểm mới chính của nghiên cứu là việc tăng cường dữ liệu dựa trên thuật toán Gaussian cho dầm BTCT được tăng cường UHPC, kết hợp với quy trình đánh giá, giữ nguyên dữ liệu thực nghiệm gốc để kiểm tra, nhằm đảm bảo đánh giá khả năng khái quát hóa. Từ góc độ kỹ thuật, phương pháp đề xuất trong nghiên cứu này cung cấp một công cụ thực tiễn và hiệu quả để nâng cao khả năng dự đoán trong trường hợp dữ liệu thực nghiệm còn hạn chế, đồng thời cho việc xây dựng các công thức lý thuyết xác định khả năng tăng cường khả năng chịu cắt của dầm BTCT bằng UHPC. Nghiên cứu tiếp theo sử dụng chuẩn hóa dữ liệu, sử dụng các thuật toán tiến tiến để tìm các siêu tham số hoặc sử dụng các thuật toán khác để tăng cường dữ liệu như CTGAN.

LỜI CẢM ƠN

Nghiên cứu này được tài trợ bởi Trường Đại học Giao thông vận tải (ĐH GTVT) trong đề tài mã số T2026-CT-006

TÀI LIỆU THAM KHẢO

[1]. V.H. Hoang, T.A Do, A. T. Tran, X. H Nguyen, Flexural capacity of reinforced concrete slabs retrofitted with ultra-high-performance concrete and fiber-reinforced polymer, Innovative Infrastructure Solutions, 9 (2024). https://doi.org/10.1007/s41062-024-01410-y.
[2]. L. Liu, S. Wan, Flexural bearing capacity of reinforced concrete beams reinforced with carbon fiber reinforced plastics strips and ultra-high performance concrete layers, International Journal of Building Pathology and Adaptation, (2022). https://doi.org/10.1108/IJBPA-04-2022-0056
[3]. Y. Zhang, X. Li, Y. Zhu, X. Shao, Experimental study on flexural behavior of damaged reinforced concrete (RC) beam strengthened by toughness-improved ultra-high-performance concrete (UHPC) layer, Composites Part B: Engineering, 186 (2020). https://doi.org/10.1016/j.compositesb.2020.107834
[4]. Eurocode 2, Design of concrete structures – Part 1: General rules and rules for buildings, Brussels, Belgium, 1992.
[5]. H.R.M. Mohammed, S. Ismail, Proposition of new computer artificial intelligence models for shear strength prediction of reinforced concrete beams, Engineering with Computers, 38 (2022) 3739–3757. https://doi.org/10.1007/s00366-021-01400-z
[6]. A. Kumar, H.C. Arora, N.R. Kapoor et al., Machine learning intelligence to assess the shear capacity of corroded reinforced concrete beams, Scientific Reports, 13 (2023) 2857. https://doi.org/10.1038/s41598-023-30037-9
[7]. W.Z. Taffese, Y. Zhu, Explainable machine learning for predicting flexural capacity of reinforced UHPC beams, Engineering Structures, 343 (2025) 121188. https://doi.org/10.1016/j.engstruct.2025.121188
[8]. V.H. Hoang, M.Q. Tran, V.T. Ngo, Machine learning-based prediction of the axial load capacity of UHPC strengthened reinforced concrete columns: A comparative analysis, PLOS ONE, 21 (2026) e0338120. https://doi.org/10.1371/journal.pone.0338120
[9]. W.Z. Taffese, N. Khodadadi, Y. Zhu, S. Mirjalili, A. Nanni, A generative adversarial network enhanced ensemble learning-based prediction model for moment improvement effect of UHPC strengthened damaged RC beams, Case Studies in Construction Materials, 23 (2025) e05323. https://doi.org/10.1016/j.cscm.2025.e05323
[10]. C. Chokwitthaya, Y. Zhu, S. Mukhopadhyay, A. Jafari, Applying the Gaussian Mixture Model to generate large synthetic data from a small dataset, Conference Proceeding in Construction Research Congress 2020: Computer Applications (2020).

[11]. T.M. Tran, H.L. Le, Enhanced prediction of the flexural capacity of prestressed reinforced concrete beams using an improved PSO-ANN model, Engineering, Technology & Applied Science Research, 16 (2026) 31947–31953. https://doi.org/10.48084/etasr.15746
[12]. M.A. Sakr, A.A. Sleemah, T.M. Khalifa, W.N. Mansour, Shear strengthening of reinforced concrete beams using prefabricated ultra-high performance fiber reinforced concrete plates: Experimental and numerical investigation, Structural Concrete, 20 (2019) 1137–1153. https://doi.org/10.1002/suco.201800137
[13]. A.A. Bahraq, et al., Experimental and numerical investigation of shear behavior of RC beams strengthened by ultra-high-performance concrete, International Journal of Concrete Structures and Materials, 13 (2019) 6. https://doi.org/10.1186/s40069-018-0330-z
[14]. H. Ji, C. Liu, Ultimate shear resistance of ultra-high performance fiber reinforced concrete-normal strength concrete beam, Engineering Structures, 203 (2020) 109825. https://doi.org/10.1016/j.engstruct.2019.109825
[15]. A. Sine, M. Pimentel, S. Nunes, A. Dimande, Shear behaviour of RC-UHPFRC composite beams without transverse reinforcement, Engineering Structures, 257 (2022) 114053. https://doi.org/10.1016/j.engstruct.2022.114053
[16]. L. Ke, et al., Shear performance evaluation of damaged RC beams strengthened with cast-in-place U-shaped UHPFRC shell, Structures, 58 (2023) 105530. https://doi.org/10.1016/j.istruc.2023.105530
[17]. S.-G. Hong, W.-Y. Lim, Strengthening of shear-dominant reinforced concrete beams with ultra-high-performance concrete jacketing, Construction and Building Materials, 365 (2023) 130043. https://doi.org/10.1016/j.conbuildmat.2022.130043
[18]. X. Liu, G.E. Thermou, Shear performance of RC beams strengthened with high-performance fibre-reinforced concrete (HPFRC) under static and fatigue loading, Materials, 17 (2024) 5227. https://doi.org/10.3390/ma17215227.
[19]. A. Abd Elghany, M. Elsayed, A.A. Elsayed, A. Shaheen, Enhancement of the shear capacity of RC deep beams with ultra-high performance fiber-reinforced concrete, Engineering, Technology & Applied Science Research, 15 (2025) 20418–20424. https://doi.org/10.48084/etasr.9792
[20]. L. Breiman, Random forests, Machine Learning, 45 (1) (2001) 5–32.
[21]. Z.J. Zhang, Introduction to machine learning: k-nearest neighbors, Annals of Translational Medicine, 4 (11) (2016) 218. http://dx.doi.org/10.21037/atm.2016.03.37
[22]. J.B. Brownlee, XGBoost with Python: Gradient boosted trees with XGBoost and scikit-learn, Machine Learning Mastery, (2016) 115. https://machinelearningmastery.com/xgboost-with-python/
[23]. G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, et al., LightGBM: a highly efficient gradient boosting decision tree, in: Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS), 2017.