Transport and Communications Science Journal, Vol. 77, Issue 4 (05/2026), 371-384

Transport and Communications Science Journal

GAUSSIAN-BASED DATA AUGMENTATION FOR IMPROVED PREDICTION OF AXIAL CAPACITY OF UHPC-JACKETED RECTANGULAR RC COLUMNS

Hoang Viet Hai, Le Dac Hien, Tran Thi Bich Thao, Bui Thanh Tung

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

ARTICLE INFO

TYPE: Research Article
Received: 20/12/2025
Revised: 09/05/2026
Accepted: 12/05/2026
Published online: 15/05/2026
https://doi.org/10.47869/tcsj.77.4.3

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

Abstract. Accurate prediction of the axial load-carrying capacity of reinforced concrete (RC) columns strengthened with ultra-high-performance concrete (UHPC) jackets is essential for reliable structural design. However, experimental data for UHPC-jacketed RC columns are scarce due to high costs and complex testing procedures, limiting the generalization of data-driven models. This study proposes a Gaussian-based data augmentation framework to enhance the predictive performance of machine learning models for estimating the axial capacity of UHPC-jacketed rectangular RC columns. An experimental database compiled from the literature is statistically analyzed, and Gaussian distribution-based techniques are employed to generate synthetic samples while preserving the statistical characteristics of the original data. Several machine learning models are developed and evaluated, with testing performed exclusively on the original experimental dataset to ensure unbiased generalization assessment. The results show that the best-performing model, CatBoost, exhibits poor generalization when trained solely on experimental data, achieving a test R^2 of 0.526 with MAE = 328.1 kN, MAPE = 33.5%, and RMSE = 540.9 kN. After Gaussian-based data augmentation, CatBoost performance improves substantially, reaching a test R^2 of 0.943, with MAE = 145.51 kN, MAPE = 10.5%, and RMSE = 222.1 kN. These results confirm that Gaussian-based data augmentation significantly enhances prediction accuracy, robustness, and generalization. The proposed framework offers a practical solution for mitigating data scarcity and supports reliable design and assessment of UHPC-strengthened RC columns.

Keywords: Gaussian-based data augmentation; ultra-high-performance concrete; machine learning.

1. INTRODUCTION

Reinforced concrete (RC) columns strengthened with ultra-high-performance concrete (UHPC) have attracted increasing attention in bridge and building structures due to their significantly enhanced load-carrying capacity, ductility, durability, and resistance to environmental degradation [1, 2]. Numerous experimental, analytical, and numerical studies have been conducted to evaluate the axial and combined axial–flexural capacity of UHPC-strengthened RC columns under various loading and boundary conditions [3]. For instance, several researchers have proposed sectional analysis approaches that account for the confinement effect and superior mechanical properties of UHPC to predict the ultimate strength and deformation capacity of strengthened columns. Experimental investigations have also demonstrated that UHPC jacketing or overlay techniques can effectively delay concrete crushing, improve reinforcement utilization, and enhance post-peak behavior [4].

Nevertheless, most conventional analytical models for strengthened RC columns are based on simplified mechanical assumptions and deterministic material properties, which may not adequately capture the highly nonlinear behavior, confinement effects, and damage evolution mechanisms associated with UHPC–RC composite systems [5]. The interaction between existing concrete, reinforcing steel, and the UHPC layer introduces additional complexities that are often neglected in traditional formulations [6].

Current design provisions, such as those provided in Eurocode 2 and other international standards [7], offer general guidelines for RC column design; however, they provide limited guidance for UHPC-strengthened members and often fail to account for material heterogeneity, interface behavior, and uncertainties related to construction quality and long-term performance. Classical analytical studies on RC column behavior and nonlinear modeling frameworks have contributed substantially to understanding axial and flexural responses, yet these approaches typically rely on fixed input parameters and lack adaptability when dealing with complex material interactions and probabilistic variations.

To address these limitations, recent research has increasingly incorporated advanced computational and data-driven techniques into the analysis of UHPC-strengthened RC columns. Hybrid modeling frameworks combining classical mechanics with numerical methods have been proposed to better represent the composite action between UHPC and conventional concrete [8]. More recently, machine learning and deep learning approaches have shown strong potential in capturing nonlinear relationships between material properties, geometric parameters, and ultimate load capacity based on experimental databases [9]. Such data-driven models offer a promising alternative for improving the accuracy and robustness of strength predictions for UHPC-strengthened RC columns under complex loading conditions.

However, it should be emphasized that experimental investigations on UHPC-strengthened RC columns are often constrained by high material costs, complex specimen preparation, and demanding testing procedures. As a result, the available experimental databases are frequently limited in size, which may restrict the generalization capability and reliability of advanced data-driven models. To overcome this challenge, recent studies have increasingly explored data augmentation techniques as an effective strategy to enrich small datasets while preserving their underlying statistical characteristics. Among various augmentation approaches, Gaussian-based data generation [10, 11] has emerged as a simple yet robust solution, as it enables the synthetic expansion of experimental data by leveraging the mean and variance structure of the

original samples, thereby enhancing data diversity without introducing unrealistic patterns. This approach is particularly suitable for structural engineering applications, where experimental variables typically exhibit approximately continuous and statistically distributed behavior.

To systematically investigate the effectiveness of Gaussian-based data augmentation in structural prediction, the present study is conducted through the following sequential steps:

Step 1- Construction of the experimental database: An experimental dataset (Dataset A) consisting of 60 samples is compiled to represent the real structural behaviour and serve as the reference data source.

Step 2- Gaussian-based data generation: A Gaussian-based approach is applied to Dataset A to generate an additional 240 synthetic samples, forming Dataset B. This step aims to expand the data space while maintaining the essential statistical characteristics of the original experimental observations.

Step 3- Definition of modelling strategies: Three modelling strategies are established to evaluate the role of data augmentation:

- **Strategy 1:** Machine learning models are trained and tested using Dataset A only to establish baseline performance.
- **Strategy 2:** Models are trained on Dataset B and evaluated on Dataset A to examine the generalization capability of synthetic-data-driven learning.

For each strategy, the available data are randomly divided into 80% for training and 20% for testing, and hyperparameters are optimized using GridSearch with 5-fold cross-validation. Finally, the predictive performance of four machine learning algorithms— Extra Trees Regressor (ETR), k-Nearest Neighbors, XGBoost, and CatBoost—is systematically evaluated and compared.

2. METHODS

2.1. Data collected from experimental test.

To estimate the axial load-carrying capacity of reinforced concrete (RC) columns reinforced with ultra-high-performance concrete (UHPC) jackets, this study will develop predictive models. In structural engineering, precise estimation of the capacity of such retrofitting components is crucial, as it supports safer and more economical design alternatives. To gather experimental data on UHPC-jacketed RC columns, a thorough literature review was conducted. Based on this review, data from reputable published sources were systematically combined to create a new, extensive database containing 60 test results [1], [12], [13], [2], [14], [15], [4], [16], [17]. This dataset offers a cohesive framework for examining how UHPC-retrofitted RC columns behave. To guarantee accurate capacity prediction, machine learning techniques were used to process and analyze the collected data. To refine raw data and convert them into appropriate input parameters for model training, an organized feature engineering process was utilized.

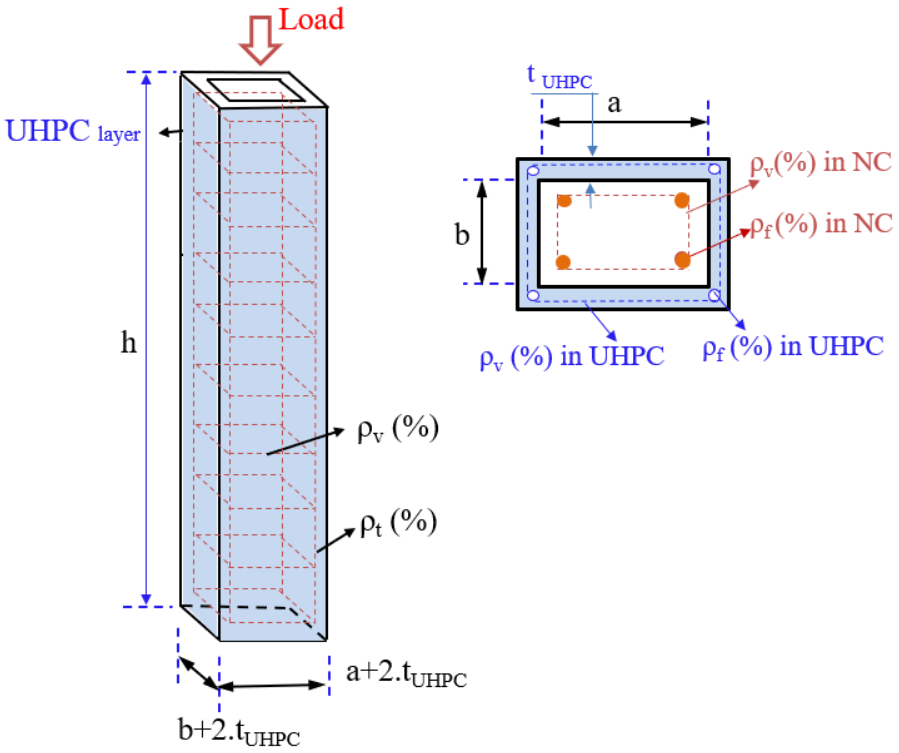

Figure 1. Schematic of RC columns strengthened with UHPC.

The collected specimens demonstrate variability in several key parameters, including column dimensions, longitudinal and transverse reinforcement ratios, jacket thickness, compressive strengths of both UHPC and the original concrete, yield strength of reinforcing steel, and fiber volume fraction and aspect ratio within the UHPC matrix. Figure 1 presents a schematic of a conventional UHPC-reinforced RC column, along with its defining parameters, whilst Table 1 provides a comprehensive description of the experimental dataset.

This study established a comprehensive dataset derived from experimental investigations of reinforced concrete columns strengthened with ultra-high-performance concrete jackets. Seventeen typical input features, including one category and sixteen numerical variables, were evaluated as predictors of the ultimate axial load capacity (P_u) of UHPC-strengthened RC columns. The numerical features include the following: column width (b); column length (a) for rectangular sections; column height (h); sectional area of normal concrete (S_{NC}); Compressive strength of normal concrete ($f'c$); Longitudinal reinforcement ratio (ρ_t) and transverse reinforcement ratio (ρ_v) in normal concrete; sectional area of Ultra-High Performance Concrete (UHPC) (S_{UHPC}); Longitudinal reinforcement ratio (ρ_t in UHPC) and transverse reinforcement ratio (ρ_v in UHPC) in the UHPC jacket; Yield strength of longitudinal (f_{yt}) and transverse (f_{yv}) reinforcement; Compressive strength of UHPC ($f'_{c\ UHPC}$); Thickness of the UHPC jacket (t_{UHPC}); Fiber dosage (volume fraction of steel fibers, % *fiber*) in the UHPC matrix. These features are illustrated in Figure 1.

The original experimental dataset exhibits considerable variability across geometric, material, and reinforcement-related parameters. Most input variables show non-uniform and mildly skewed distributions, reflecting discrete design choices and practical constraints in experimental programs rather than purely random variation. Material properties and reinforcement ratios are mainly concentrated within limited ranges, while a smaller number

of specimens represent higher-strength or heavily reinforced configurations.

Table 1. Statistical information for parameters in the database.

No	Feature	Type	Units	Min	Max	Mean	Median	Std	Skewness	Kurtosis
1	*a*	X1	mm	100	300	152.17	150	49.03	0.70	-0.13
2	*b*	X2	mm	100	300	158.50	150	49.31	0.42	-0.48
4	*h*	X3	mm	300	1000	536.67	500	225.84	0.57	-0.65
5	S_{NC}	X4	mm^2	10000	90000	26215.00	22500	16407.68	1.37	2.67
6	*fc'*	X5	MPa	22.2	48.2	31.48	27.01	9.51	0.91	-0.69
7	ρ_t *in NC*	X6	%	0	3.141	1.63	1.858	0.95	0.01	-0.74
8	ρ_v *in NC*	X7	%	0	3.66	0.94	0.536	0.93	1.20	0.61
9	S_{UHPC}	X8	mm^2	0	31600	10576.53	12500	8969.47	0.29	-0.68
10	ρ_t *in UHPC*	X9	%	0	2.513	0.49	0	0.77	1.32	0.39
11	ρ_v *in UHPC*	X10	%	0	3.66	0.36	0	0.79	3.01	9.38
12	f_{yt}	X11	MPa	360	644	478.31	502	89.08	0.36	-0.70
13	f_{yv}	X12	MPa	240	1173	568.17	469	352.50	0.81	-0.89
14	fc'_{UHPC}	X13	MPa	81.6	189.97	112.78	112	23.33	1.73	4.43
15	*% fiber*	X14	%	0	2.3	0.90	0.5	0.93	0.27	-1.80
16	t_{UHPC}	X15	mm	0	40	17.38	20	13.62	-0.16	-1.33
17	P_u	Y	kN	294.2	3866	1706.82	1398.77	939.51	0.67	-0.51

The output variable, ultimate load (P_u), spans a wide range, indicating diverse structural capacities influenced by multiple interacting factors. These distributional characteristics highlight the limited coverage and inherent heterogeneity of the experimental data, which motivates the application of data augmentation techniques to enhance the robustness and generalization capability of subsequent machine learning models.

2.2. Data Generation Using Gaussian Mixture Model (GMM)

Considering the experimental constraints and discrete nature of the original dataset, Gaussian Mixture Model (GMM) or Gaussian-based augmentation is introduced to generate additional samples within statistically plausible ranges, thereby improving the robustness of subsequent predictive modeling [10].

In this study, GMM was employed to produce an additional 240 synthetic samples. This augmentation aims to expand the dataset to a sufficient size that enables reliable evaluation of the original 60 experimental observations while simultaneously enhancing the model's ability to generalize the underlying structural behavior. The distributions of the dataset before and after augmentation are illustrated in Figure 2.

Figure 2 shows a close match between the probability density distributions of the original and GMM-generated data, demonstrating that the Gaussian Mixture Model preserves the statistical properties of the experimental dataset and is suitable for data augmentation in machine learning applications

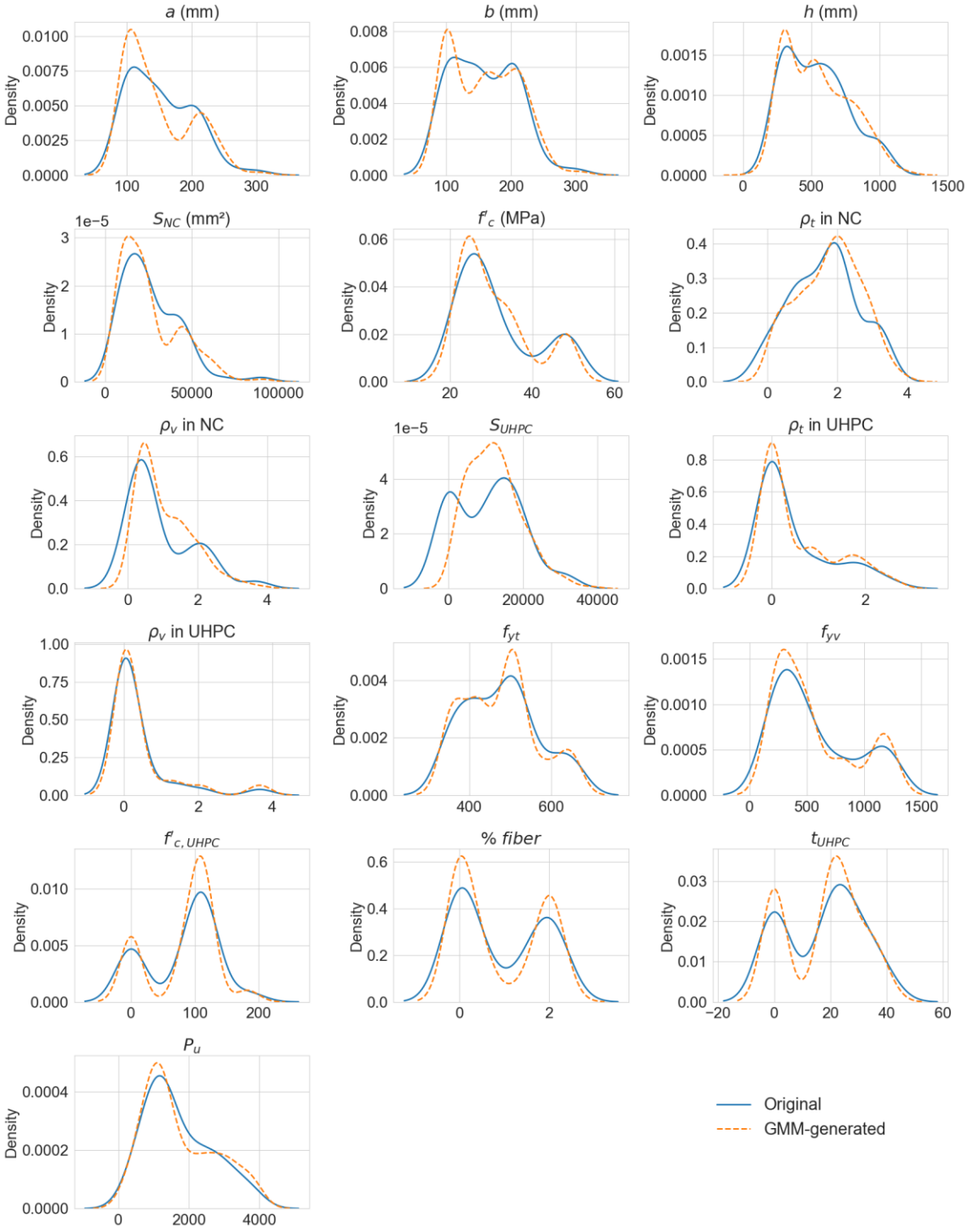

Figure 2. Comparison of probability density distributions between original and GMM-generated datasets.

2.3. The proposed method

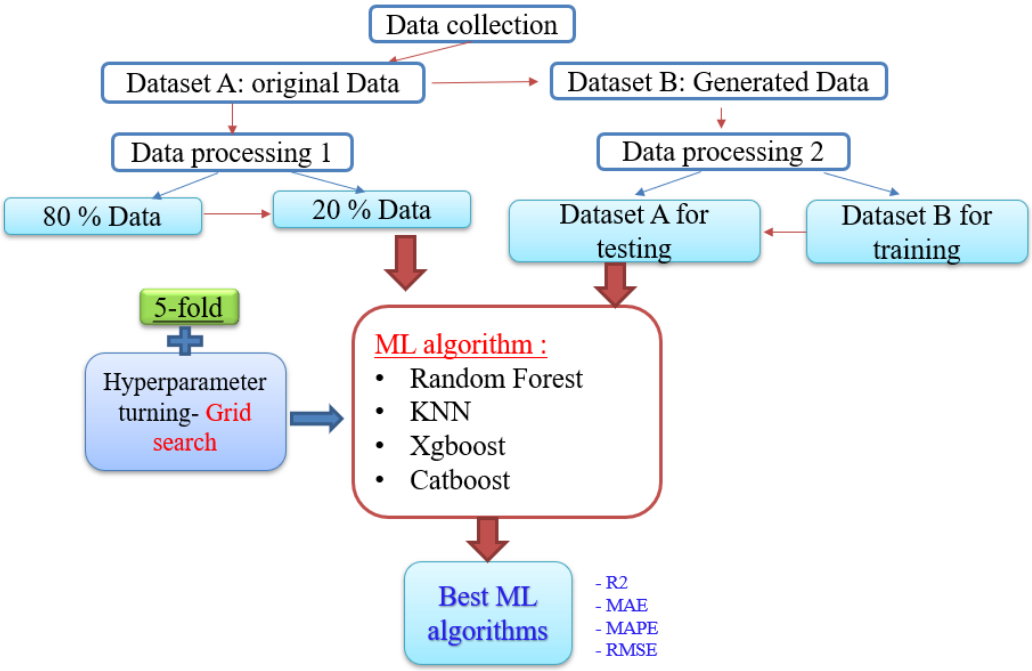

Figure 3. Methodology of research.

Based on the constructed datasets, two distinct modeling strategies are implemented to systematically evaluate the effectiveness of Gaussian-based data augmentation (Figure 3). Dataset A consists of the original 60 experimental samples, while Dataset B comprises 240 synthetically generated samples obtained using the Gaussian approach. For all machine learning models, the available data in each strategy are randomly divided into 80% for training and 20% for testing.

- Strategy 1 employ Dataset A only, where the models are trained and tested using the original experimental data to establish a baseline predictive performance.
- Strategy 2 utilizes Dataset B for model training, while the trained models are subsequently evaluated on Dataset A to examine the generalization capability of models trained solely on synthetic data.

This structured comparison framework allows a comprehensive assessment of the influence of Gaussian-based data generation on model accuracy and generalization when applied to real experimental data. To ensure a fair and reliable comparison, hyperparameter optimization is performed for all machine learning models using GridSearch combined with 5-fold cross-validation (5 K-fold) on the training data. This procedure aims to identify the optimal set of hyperparameters that minimizes model bias and variance while improving generalization performance. This study employs four machine learning algorithms—Extra Trees Regressor (ETR), k-Nearest Neighbors (KNN), XGBoost, and CatBoost—covering a broad range of tree-based, ensemble, instance-based, and gradient boosting approaches. The optimized models are subsequently evaluated on the corresponding test datasets according to each strategy.

2.3.1. Extra Trees Regressor (ETR)

Extra Trees Regressor (ETR) is an advanced ensemble learning technique that enhances randomization beyond the standard Random Forest approach. Unlike traditional methods that

search for optimal splitting thresholds, ETR selects thresholds entirely at random for each feature at every node. This mechanism significantly lowers model variance and minimizes the impact of data noise.

For regression, the final prediction is the average of all individual tree outputs. By prioritizing stochasticity over local optimization, ETR offers superior computational efficiency and a smoother decision boundary. Its strong generalization capability and effective control over overfitting make it a robust tool for modeling complex engineering and scientific data.

2.3.2. K-Nearest Neighbors (KNN)

The k-Nearest Neighbors (KNN) algorithm is an instance-based learning method that makes predictions based on the similarity between samples in the feature space. Instead of constructing an explicit model during training, KNN identifies the k closest training instances to a query point using a predefined distance metric. In regression tasks, the predicted value is computed as the average response of the selected neighbors. Owing to its non-parametric nature, KNN is capable of capturing complex nonlinear relationships; however, its performance is sensitive to the choice of k, feature scaling, and data sparsity.

2.3.3. Extreme Gradient Boosting (XGBoost)

XGBoost is an advanced gradient boosting framework that builds an ensemble of decision trees in a sequential manner, where each new tree is trained to correct the residual errors of the preceding ensemble. The algorithm incorporates regularization terms into the objective function to control model complexity and prevent overfitting. Through efficient tree construction, parallel computation, and optimized loss minimization, XGBoost achieves high predictive accuracy and scalability. These characteristics make it particularly effective for structured tabular data in regression-oriented engineering applications.

2.3.4. CatBoost

CatBoost is a gradient boosting algorithm specifically designed to improve training stability and reduce prediction bias in tree-based ensembles. It employs an ordered boosting strategy that prevents information leakage during training and enhances generalization performance, especially for small or heterogeneous datasets. For regression problems, predictions are obtained by aggregating the outputs of multiple symmetric decision trees. Due to its robustness, minimal need for extensive preprocessing, and strong performance on tabular data, CatBoost has gained increasing attention in engineering prediction tasks.

CatBoost is advantageous for forecasting the axial resistance of UHPC-reinforced RC columns because of numerous complexes, nonlinear interaction parameters, including core area, UHPC jacket area, material strength, and reinforcement ratio. CatBoost effectively captures nonlinear interactions, minimizes prediction errors, and maintains excellent generalization. This elucidates why CatBoost frequently surpasses other boosting models in numerous applied investigations.

2.3.5. Performance Evaluation Metrics

To evaluate the predictive capability of the proposed machine learning models, four statistical metrics are adopted, including the coefficient of determination (R^2), root mean square error (RMSE), mean absolute percentage error (MAPE), and mean absolute error (MAE). These indicators are widely recognized for assessing regression performance in structural engineering applications. In general, a larger R^2 value, approaching unity, signifies a stronger agreement between predicted and observed values, whereas smaller values of RMSE, MAPE, and MAE

correspond to higher prediction accuracy and reduced estimation errors. The formulations of these evaluation metrics are provided below:

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

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

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

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

3. RESULTS AND DISCUSSION

3.1. Evaluation of Model

Table 2 summarizes the ideal hyperparameters for each model, detailing the tuned configurations that achieved the best result during the Grid Search procedure. These parameter configurations illustrate the distinct attributes of each algorithm and underscore the importance of hyperparameter tuning for improving predictive accuracy. After the optimization phase, the predictive performance of all models is assessed and compared using four statistical metrics: the coefficient of determination (R^2), mean absolute error (MAE), mean absolute percentage error (MAPE), and root mean square error (RMSE). The predictive performances of the four machine learning models under Strategy I (Dataset A only) and Strategy II (training on Dataset B and testing on Dataset A) are summarized in Tables 3 and 4, respectively. A clear distinction in model behavior and generalization capability can be observed between the two strategies.

Table 2. Optimal hyperparameters.

No	Model	Hypeparameter	Search space	Optimal hyperparameter	
				Strategy 1	Strategy 2
1	ETR	max_depth	[3,5,7,8]	8	8
		max_features	[sqrt, 0.8]	0.8	0.8
		min_samples_leaf	[1,2,5]	1	1
		min_samples_split	[2,5,10]	2	2
		n_estimators	[200, 300, 500]	200	500
2	KNN	n_neighbors':	[3,5, 7, 9, 11]	5	3
		p	[1,2]	2	1
		Weights	[uniform]	uniform	uniform
3	XGboost	Learning rate	[0.01, 0.03, 0.05]	0.03	0.01
		max_depth	[2, 3, 4]	3	3
		n_estimators	[300, 500, 800]	800	800
		Subsample	[0.6, 0.7, 0.8]	0.8	0.8
		gamma	[0, 0.1, 0.3]	0	0
4	Catboost	Depth	[1, 2, 3]	3	3
		Iteration	[300, 500, 800]	500	800
		l2_leaf_reg	[2, 3, 4, 5]	2	4
		Learning rate	[0.01, 0.05, 0.1]	0.1	0.05

Table 3. Average predictive performance of the model obtained in Strategy I.

No	Model	Data Train				Data Test			
		R^2	MAE	MAPE	RMSE	R2	MAE	MAPE	RMSE
1	ETR	0.965	113.511	7.422	172.386	0.467	397.886	40.751	573.413
2	KNN	0.846	267.146	15.890	366.06	0.645	331.578	33.720	467.871
3	XGboost	0.979	56.836	2.887	135.27	0.437	356.737	40.414	589.457
4	Catboost	0.979	51.289	2.432	134.37	0.526	328.089	33.502	540.860

Table 4. Average predictive performance of the model obtained in Strategy II.

No	Model	Data Train				Data Test			
		R^2	MAE	MAPE	RMSE	R2	MAE	MAPE	RMSE
1	ETR	0.991	72.44	6.9	97.44	0.942	145.42	10.7	223.51
2	KNN	0.887	223.49	20.9	338.02	0.574	417.29	27.1	608.17
3	XGboost	0.991	73.98	7.0	95.04	0.942	146.49	10.5	223.57
4	Catboost	0.995	52.90	5.1	67.89	0.943	145.51	10.9	222.42

Under Strategy I, all models exhibit very high training accuracy, particularly the ensemble-based algorithms (XGBoost and CatBoost), with training R^2 values reaching 0.979 and very low MAE and RMSE values. Extra Trees Regressor (ETR) also exhibits strong training performance with $R^2 = 0.965$. However, a pronounced deterioration in predictive performance is observed on the test set for most models. Despite their excellent training accuracy, XGBoost and CatBoost suffer a significant drop in test performance, with R^2 values decreasing to 0.437 and 0.526, respectively, indicating limited generalization capability. ETR shows similar behavior, yielding a relatively low test $R^2 = 0.467$ and large prediction errors. In contrast, KNN demonstrates comparatively better robustness, achieving the highest test $R^2 = 0.645$, although its MAE, MAPE, and RMSE remain relatively high. These findings indicate that training solely on the original experimental dataset leads to evident overfitting, which can be attributed to the small size and limited variability of Dataset A.

In contrast, Strategy II shows substantial improvements in generalization performance across most models when trained on the synthetic Dataset B and evaluated on Dataset A. The ensemble-based models exhibit remarkable predictive capability, with test R2R^2R2 values exceeding 0.94 for ETR, XGBoost, and CatBoost. Specifically, ETR achieves a test $R^2 = 0.942$ with MAE = 145.42 and RMSE = 223.51, while XGBoost delivers a similar performance with a test $R^2 = 0.942$, MAE = 146.49, and RMSE = 223.57. CatBoost provides the best predictive performance among all models, achieving a test $R^2 = 0.943$ with MAE = 145.51 and RMSE = 222.42. These results indicate a significant improvement in predictive capability compared with Strategy I. Although KNN still shows comparatively weaker performance, its results remain acceptable with a test R^2=0.574, MAE = 417.29, and RMSE = 608.17.

The comparison between Strategy I and Strategy II clearly highlights the critical role of synthetic data augmentation in enhancing model generalization. Strategy I primarily reflect memorization of limited experimental data, leading to poor test performance. Strategy II confirms that training on a sufficiently large and diverse synthetic dataset can significantly improve robustness and predictive reliability, even when models are evaluated on unseen real experimental samples. From a practical perspective, this comparison validates the feasibility of using synthetically generated datasets as an effective training source for machine learning models in scenarios where experimental data are scarce, such as in structural engineering

applications.

3.2. Comparison of actual and predicted values P_u

Figure 4 illustrates the relationship between the predicted and actual ultimate load values obtained using Strategy 1, in which all models are trained and tested solely on the original experimental dataset (Dataset A). A pronounced discrepancy between training and testing performance can be clearly observed.

For all models, the training data points are closely clustered around the ideal diagonal line, indicating very high fitting accuracy. This behavior is particularly evident for XGBoost and CatBoost, which achieve near-perfect alignment on the training set. However, the test data points exhibit a much wider dispersion, especially for ETR and XGBoost, reflecting a substantial loss of predictive accuracy on unseen samples.

Random Forest shows the most severe overfitting behavior, with test points widely scattered and deviating significantly from the diagonal line, consistent with its extremely low test R^2. KNN demonstrates relatively better robustness, with test points moderately aligned along the reference line, although noticeable deviations remain at higher load values. Despite their excellent training performance, both XGBoost and CatBoost fail to generalize effectively, as evidenced by the visible spread of test points and reduced test accuracy.

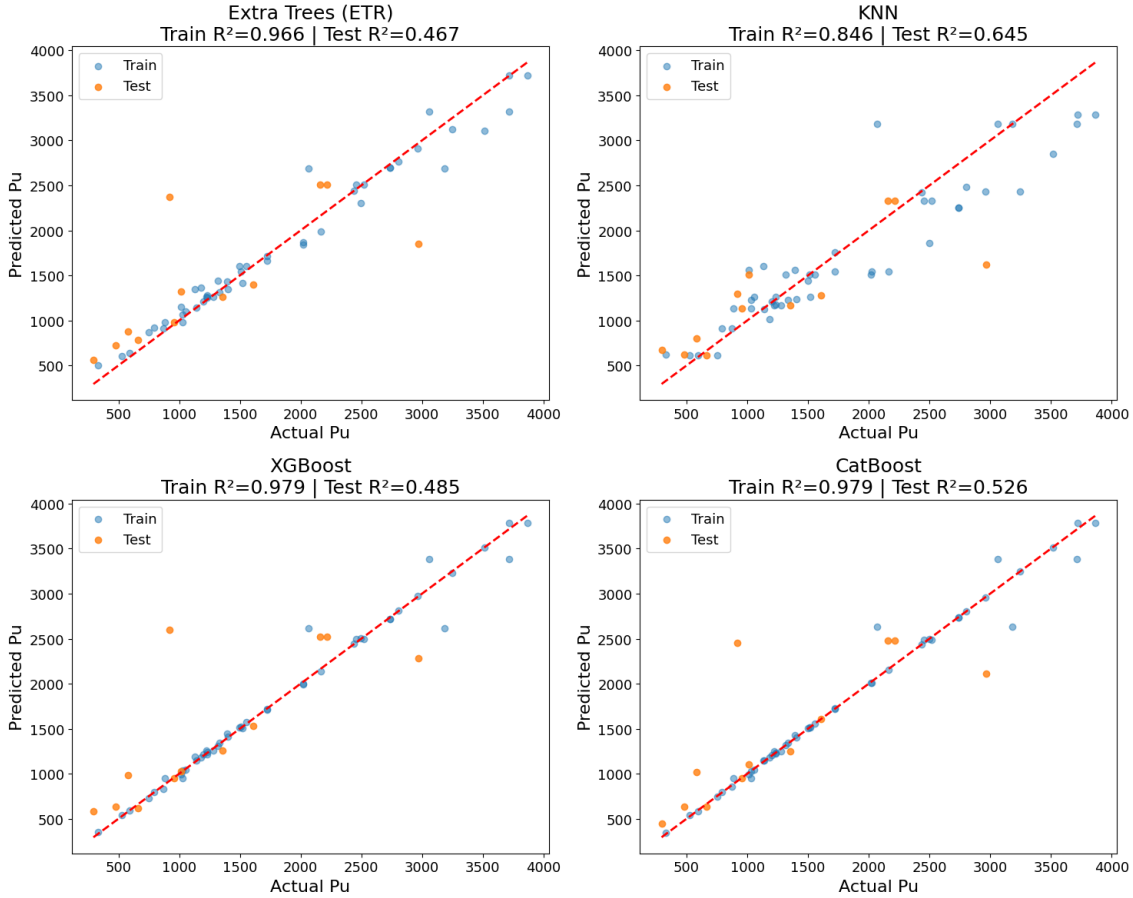

Figure 4. Relationship between actual and predicted values for the four machine learning models under Strategy 1.

This figure highlights that training exclusively on the limited experimental dataset leads to strong memorization rather than true learning, resulting in poor generalization capability across most models.

Figure 5 presents the predicted versus actual values obtained under Strategy 2, where the models are trained on the GMM-generated synthetic dataset (Dataset B) and evaluated on the real experimental data (Dataset A). Compared to Strategy I, a remarkable improvement in generalization performance is observed for all models.

The test data points show a much tighter alignment along the diagonal reference line, indicating significantly enhanced predictive consistency. Random Forest exhibits a notable reduction in prediction dispersion, with test points closely following the ideal trend. KNN achieves excellent agreement between predicted and actual values across the entire load range, demonstrating strong stability and robustness.

XGBoost and CatBoost display the most favorable behavior, with both training and testing points forming a dense cluster around the diagonal line. The reduced spread of test samples suggests that these models are able to capture the underlying physical relationships more effectively when trained on a sufficiently large and diverse synthetic dataset.

This figure clearly demonstrates that synthetic data augmentation using Gaussian Mixture Model (GMM) substantially mitigates overfitting, enabling machine learning models to generalize well to real experimental data despite the scarcity of original measurements.

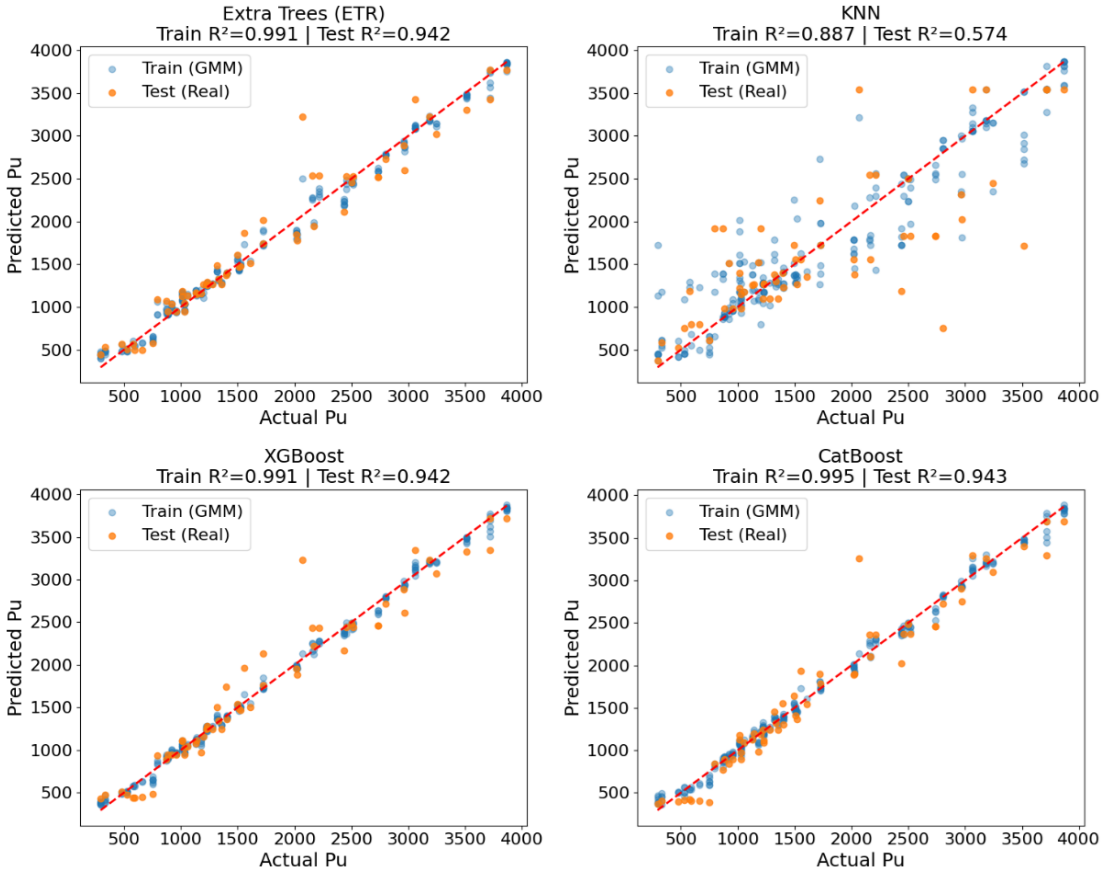

Figure 5. Relationship between actual and predicted values for the four machine learning models under Strategy 2.

4. CONCLUSION

This study demonstrates the effectiveness of Gaussian-based data augmentation in enhancing the predictive accuracy and generalization capability of machine learning models for estimating the axial load-carrying capacity of UHPC-jacketed rectangular RC columns. Owing to the limited availability of experimental data, a Gaussian distribution-based augmentation strategy was employed to expand the original dataset from 60 experimental samples to 240 synthetic–experimental samples, enabling more robust and reliable model training.

A comparison of the two strategies reveals that models trained only on the original dataset exhibit clear overfitting. Although the training accuracy is high (R^2 = 0.846 ÷ 0.979), the test performance drops significantly with R^2= 0.437÷ 0.645 and large prediction errors. In contrast, the Gaussian-augmented training strategy substantially improves model generalization. The ensemble models achieve test R^2 values of approximately 0.94 with significantly reduced errors (MAE = 145÷149 kN, RMSE = 222÷224 kN).

Among the models, CatBoost delivers the best performance with R^2 = 0.943, MAE = 145.51 kN, MAPE = 10.5%, and RMSE = 222.42 kN. These results demonstrate that Gaussian-based data augmentation combined with advanced ensemble learning provides an effective framework for improving prediction reliability when experimental data are limited.

The key novelty of this study lies in the systematic validation of a Gaussian-based data augmentation framework for UHPC-strengthened RC columns, coupled with a strict evaluation protocol that preserves the original experimental data for testing to ensure unbiased generalization assessment. From an engineering perspective, the proposed approach provides a practical and efficient tool for improving prediction reliability under data-scarce conditions, offering valuable support for preliminary design, structural safety evaluation, and decision-making in UHPC strengthening applications.

ACKNOWLEDGMENTS

This research is funded by the University of Transport and Communications (UTC) under grant number T2025-CT-004TD

REFERENCES

[1]. D. Z. Helles, Strengthening of Square Reinforced Concrete Columns with Fibrous Ultra High Performance Self-Compacting Concrete Jacketing, PhD Thesis, The Islamic University Gaza, Palestine, 2014.

[2]. S. A. Dadvar, D. Mostofinejad, H. Bahmani, Strengthening of RC columns by ultra-high performance fiber reinforced concrete (UHPFRC) jacketing, Construction and Building Materials, 235 (2020) 117485. https://doi.org/10.1016/j.conbuildmat.2019.117485

[3]. L.D. Tolentino, Effect of ultra-high performance concrete repair layer thickness on the behavior of concrete columns, 25 (2024) 1801-1818. https://doi.org/10.1002/suco.202300165

[4]. H. W. Tian, Experimental and numerical investigation on square concrete-filled UHPC tubular columns under axial compression. Structures, 70 (2024) 107655. https://doi.org/10.1016/j.istruc.2024.107655

[5]. M. Bolbolvand, S.M. Tavakkoli, F.J. Alaee, Prediction of compressive and flexural strengths of ultra-high-performance concrete (UHPC) using machine learning for various fiber types, Construction and Building Materials, 493 (2025) 143135. https://doi.org/10.1016/j.conbuildmat.2025.143135

[6]. 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

[7]. Eurocode 2: Design of concrete structures - Part 1: General rules and rules for buildings. Brussels, Belgium, 1992.

[8]. 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

[9]. T.G. Wakjira, A. Abushanab, and M.S. Alam, Hybrid machine learning model and predictive equations for compressive stress-strain constitutive modelling of confined ultra-high-performance concrete (UHPC) with normal-strength steel and high-strength steel spirals. Engineering Structures, 304 (2024) 117633. https://doi.org/10.1016/j.engstruct.2024.117633

[10].C. Bishop, Pattern Recognition and Machine Learning, Springer, 2006, p.p 140-155.

[11].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.

[12].R.M. ENAMI, Reforço de pilares curtos de concreto armado por encamisamento com concreto de ultra-alto desempenho, in Escola de Engenharia de São Carlos, PhD Thesis, University of São Paulo, Brazil, 2017.

[13].J. Chen, Z. Wang, A. Xu, J. Zhou, Compressive Behavior of Corroded RC Columns Strengthened With Ultra-High Performance Jacket, Frontiers in Materials, 9 (2022). https://doi.org/10.3389/fmats.2022.859620

[14].R.M.I.R. Susilorini and Y. Kusumawardaningsih, Advanced Study of Columns Confined by Ultra-High-Performance Concrete and Ultra-High-Performance Fiber-Reinforced Concrete Confinements, Fibers, (2023) 11. https://doi.org/10.3390/fib11050044

[15].M.A. Alamoodi, M. Zahid, B.H. Adu Bakar, B.A. Tayeh, A.M. Zeyad, Behavior of damaged reinforced concrete columns retrofitted with ultra-high performance fiber reinforced concrete jackets under uniaxial loading, Journal of Building Engineering, 108 (2025) 112837. https://doi.org/10.1016/j.jobe.2025.112837

[16].A.I.B. Farouk, W. Rong, J. Zhu, Compressive behavior of ultra-high-performance-normal strength concrete (UHPC-NSC) column with the longitudinal grooved contact surface, Journal of Building Engineering, 68 (2023) 106074. https://doi.org/10.1016/j.jobe.2023.106074.

[17].H. Shehab, A. Eisa, A. M. Wahba, P. Sabol, D. Katunsky, Strengthening of Reinforced Concrete Columns Using Ultra-High Performance Fiber-Reinforced Concrete Jacket. 13 (2023) 2036. https://doi.org/10.3390/buildings13082036