Advanced Steel Construction – Vol. 22 No. 2 (2026) 244–254
DOI:10.18057/IJASC.2026.22.2.10

A DATA-CENTRIC STRATEGY TO MITIGATE OVERFITTING OF ML MODELS FOR PREDICTING TORSIONAL CAPACITY FOR CFST COLUMNS

Ming-Xia Dang [1], Meng-Xue Guo [2, *], Ying Li [1], Hua Li [1] and Shi-Lin Yang [3]

[1] *School of Intelligent Construction and Environment, Xi'an Jiaotong University City College, Xi'an, 710018, China*

[2] *School of Civil & Architecture Engineering, Xi'an technological university, Xi'an 710021, China*

[3] *Shaanxi Construction Engineering Group Corporation Limited, Xi'an 710003, China*

** (Corresponding author: E-mail: 18792602146@163.com)*

ABSTRACT

This study investigates the effectiveness of both model-centric and data-centric strategies in addressing the overfitting issue in machine learning (ML) models for predicting the torsional capacity of concrete-filled steel tubular (CFST) columns under combined loading. While prior work has largely focused on optimizing model architectures, our findings reveal that model-centric approaches offer limited improvement when training data is scarce. To address this, we propose a data-centric framework that enhances both the quantity and quality of training data. Specifically, we augment the dataset with synthetic data generated by Conditional Generative Adversarial Networks (CGANs) and finite element analysis (FEA) results. To ensure reliability, we introduce a filtering mechanism that selects high-quality simulated data for model training. Our results reveal that directly incorporating unfiltered synthetic or FEA data into model training can degrade test performance due to the presence of noisy or unreliable samples. In contrast, when high-quality FEA data is carefully filtered and selectively combined with experimental data, the model exhibits a substantial improvement in generalization, reflected by a 5% increase in R^2 with only a marginal 0.45% rise in MAPE. The proposed data selection strategy consistently reduces performance variance across multiple test splits, indicating strong robustness and resistance to overfitting.

ARTICLE HISTORY

Received: 31 March 2025
Revised: 7 July 2025
Accepted: 2 August 2025

KEYWORDS

Data-centric;
Overfitting;
Machine learning;
CTGAN;
CFST columns

1. Introduction

The recent proliferation of machine learning technologies has sparked a paradigm shift in structural engineering research, ushering in a new era of innovative approaches to tackle intricate challenges that were previously intractable using traditional mechanical models[1-4]. Traditionally, engineering practices have been grounded in limit state design criteria, reliant on rigorous experimental validation, intricate finite element modeling, or a synthesis of parameter regression analysis and strength theory derivations to formulate predictive frameworks[5-7]. However, these methodologies are constrained by inherent limitations, notably model dispersion and biases stemming from modeling inaccuracies or the imposition of overly simplistic assumptions. The exponential growth in computational power, coupled with the pressing need for efficient solutions in the engineering sector, has catalyzed significant advancements in machine learning algorithms.

The core of machine learning is inherently data-driven, especially data-centric, with model accuracy and generalization power intimately tied to the diversity and quality of datasets. Suboptimal sample quality can exacerbate model biases by incorporating low-fidelity instances, while limited sample diversity undermines the surrogate models' capacity for generalization[8]. Consequently, the utilization of machine learning in civil engineering has sparked significant concerns within the engineering community, particularly regarding the adequacy and quality of data, along with the risk of overfitting[9]. Models trained on inadequate data are unreliable and prone to overfitting, yielding spurious predictions that undermine their credibility. To address this challenge, the cornerstone lies in constructing comprehensive datasets comprising multi-source, high-quality samples. In civil engineering, where machine learning models operate as black boxes, their reliance on extensive and unbiased physical experimental data for training is paramount. This data must be comprehensively representative, devoid of biases, and encompass a broad range of conditions. However, experimental uncertainties and equipment malfunctions inevitably introduce noisy samples and outliers during testing, limiting the availability of training data in structural engineering and further hindering the acquisition of high-quality data. Consequently, relying solely on data from physical experiments to train machine learning models can be fraught with errors, as limited training data often precipitates overfitting, potentially misleading evolutionary search processes and compromising the models' predictive performance.

The scarcity of experimental samples, insufficient for training robust machine learning models, poses a significant challenge. This limitation underscores the paramount importance of critically evaluating data sources' appropriateness and reliability when designing and deploying machine learning models for civil engineering applications. To address this issue, FEA data are frequently utilized as a complementary resource to experimental data[10]. However, a fundamental gap exists due to the inherent constraints of simulation methods. Despite significant advancements, these methods struggle to fully capture the intricate and dynamic interactions inherent in real-world systems. Factors such as variations in material constitutive laws, mesh discretization techniques, and solution strategies contribute to this discrepancy, rendering it challenging to achieve alignment between simulated and experimental outcomes. To bridge this gap, researchers engage in rigorous validation and calibration of finite element models, striving to develop more precise and generalized machine learning solutions[11]. It is imperative to acknowledge that the inherent discrepancy between non-physical experimental data and real-world data remains an inescapable reality.

In the pursuit of mitigating model biases arising from the limited training data, researchers across disciplines have intensified their focus on tackling data-centric challenges within machine learning[12, 13]. This heightened awareness is particularly pronounced in the realms of medicine and materials science, where the collection, cleansing, and evaluation of data have become paramount. Within structural engineering, scholars have likewise embarked on addressing the dual challenges of data quantity and quality. Notably, Li et al.[14] have employed Gaussian regression to devise a machine learning model that achieves remarkable precision in measuring deformations of reinforced concrete columns, effectively demonstrating its resilience across a broad spectrum of data quality. Additionally, Luo et al.[15] have introduced a pioneering dual-weighted support vector transfer regression methodology, which aims to bolster prediction performance by effectively countering the detrimental effects of limited sample sizes. Moreover, Marani et al.[16] have taken an approach by leveraging physical experimental data to train a Generative Adversarial Network (GAN), enabling the synthesis of novel data that supplements the existing experimental corpus. This technique has facilitated the successful training of machine learning models on a combined dataset, revealing that the utilization of synthetic data significantly reduces the reliance on scarce real-world data during the model training phase. These findings underscore the potential of synthetic data in streamlining and augmenting the machine learning process within structural engineering, thereby enhancing the efficiency and accuracy of predictive models[17].

The advent of existing methodologies has undoubtedly alleviated the burden of data scarcity to a certain degree. Nevertheless, a pivotal yet underexplored challenge looms large: the biases inherent in models fueled by the limited and low-quality data. The difference between physical experimental data, and non-experimental counterparts, including finite element simulations and synthetic data, continues to pose a formidable obstacle. This not only complicates the predictive accuracy of models but also undermines their credibility in real-world applications. Synthetic data, a promising avenue for augmenting data availability, holds the key to bridging the data gap[18]. However, the divide between synthetic and experimental data, albeit narrowing, remains a significant hurdle

that must be overcome. Current endeavors, focused on refining finite element models for heightened realism in simulations and advancing generative models to mirror real-world complexity[19,20]. However, the quest for parity between these data sources and their experimental counterparts persists, emphasizing the need for a more nuanced approach.

This study explores the effect of the model design and data on mitigating overfitting in machine learning models, respectively, i.e., the model-centric and the data-centric training strategy. And, two strategies were used and analyzed in the context of predicting the composite torsional capacity of concrete-filled steel tubular (CFST) columns. Drawing upon two prevalent data augmentation techniques: Conditional Generative Adversarial Network (CTGAN)-based synthetic data[21] and finite element analysis data (FEA). Our approach distinguishes itself from existing research by not focusing on the refinement or development of new models, but rather on devising an efficient strategy for leveraging multiple data sources. Applying this training strategy to the prediction of CFST columns' torsional behavior, we embark on a comprehensive investigation that spans from models to data, with the overarching goal of addressing the issue of overfitting in machine learning models, especially for the predictions of CFST torsional performance. Specifically, our research endeavors encompass:

1) Explore the impact of diverse model-centric strategies on mitigating overfitting in machine learning contexts, shedding light on the effectiveness of various approaches, including model replacement and reduction of non-critical input parameters.

2) Investigate the influence of data-centric strategies including GAN-generated synthetic Data and FE simulation data in training processes, assess their respective contributions to reducing overfitting and enhancing model generalization.

3) Develop and evaluate sample selection and training strategies based on the error minimization, to mitigate overfitting, thereby refine the overall predictive capabilities of the models.

2. Workflow to mitigate overfitting in ML prediction models

Fig. 1 outlines the approach devised in this research to address the challenge of overfitting in data-driven predictive models for assessing the torsional capacity of concrete-filled steel tubular (CFST) columns. This framework encompasses four key stages: data compilation, model evaluation, model-centric optimizations, and data-centric enhancements.

(1) Data Compilation: A comprehensive dataset, initially comprising 243 experimental observations curated from the scientific literature[22-34], serves as the cornerstone for model development. To augment this dataset and enable a deeper exploration of overfitting mitigation strategies, we incorporate 561 FEA data[35-41], and 1670 synthetic samples generated based on CTGAN. This enriched dataset encompasses a wide range of features, including cross-sectional geometries, steel tube dimensions, and loading conditions.

(2) Model-Centric Strategy: A rigorous evaluation framework was established, wherein the initial 243 experimental data points were meticulously partitioned into training (70%) and testing (30%) subsets. Ten state-of-the-art machine learning algorithms were systematically employed for model training and subsequent performance assessment, with the selection of the optimal model being guided by metrics such as the coefficient of determination (R^2). Two model-centric approaches are employed to address overfitting. Firstly, a selection and subsequent substitution of distinct, well-established machine learning algorithms are undertaken, for evaluating their efficacy in extrapolating knowledge from the training samples to unseen data. Secondly, the impact of varying input feature subsets on overfitting was then quantitatively assessed. The elimination of non-critical variables was aligned with the physical insights derived from traditional engineering models.

(3) Data-Centric Strategy: Leveraging the augmented dataset, two distinct strategies (CTGAN- and FEA) are devised to harness the potential of synthetic and FEA data. For these non-experimental data, there was two training strategies was applied, recorded as S1 and S2. For S1, the synthetic and FEA data are directly integrated for model training. In contrast, S2 introduces a sample selection phase. Based on prediction errors, a selective filtering process is applied, the data with the lowest errors at different proportions was utilized for training resulting in datasets. These datasets are then utilized for surrogate model training, aimed at achieving a more robust and generalized model with reduced overfitting tendencies.

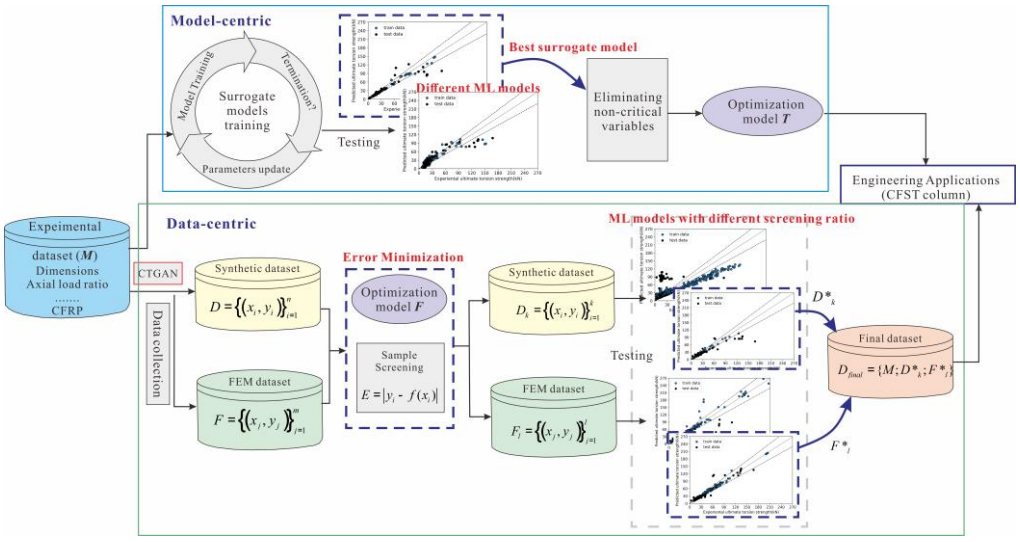


Fig. 1 Workflow to mitigate overfitting in ML prediction model

3. Evaluating model-centric strategies for effective overfitting mitigation

3.1. Update surrogate models

3.1.1. Data collection

The comprehensive dataset employed in this research endeavor comprises 243 rigorously curated experimental data points. The feature parameters intricately woven into this dataset encompass crucial structural design elements alongside the various loading scenarios. These structural design parameters meticulously detail aspects such as cross-sectional shape, steel tube diameter (D), wall thickness (t), steel yield strength (f_y), concrete strength (f_c), longitudinal-bending-torsion ratio (M/T), and axial compression ratio. Table 1 offered a comprehensive statistical summary of the key parameters within the dataset.

3.1.2. Preliminary assessment of traditional ML models

Out of the 243 experimental samples, 167 data points were randomly selected to serve as the training set for the model, while the remaining 76 samples constituted the test set. This strategic partitioning ensures that the test set encapsulates a diverse range of data pertaining to CFST columns under various load combinations, encompassing compression, bending, and torsion. Drawing upon the experimental dataset comprising 243 unique sets of data, we embarked on an initial predictive analysis of the torsional capacity of CFST columns. This endeavor leveraged five well-established machine learning paradigms: Support Vector Regression (SVR), Decision Tree Regression, Random Forest Regression, Linear Regression, and Extreme Gradient Boosting (XGBoost). By establishing predictive models to each of these methodologies, we conducted a comparative analysis to evaluate the corresponding performance. Table 2 provided a comprehensive overview of the prediction results achieved by these models under their hyperparameters configurations.

Upon rigorous validation using 76 genuine experimental data points from the test set, which encompasses both bending-torsion and compression-bending-torsion scenarios, it becomes strikingly apparent that the linear regression model

falls short, demonstrating the intricate nonlinearity inherent in the torsional behavior of reinforced concrete columns under combined loading conditions. This underscores the inadequacy of simplistic linear approaches in capturing complex dynamics. In contrast, the decision tree regression model displays a pronounced overfitting tendency, achieving an impeccable fit with an R^2 of 0.99 on the training dataset, accompanied by remarkably high RMSE, MAE, and MAPE values. However, as Fig. 2 illustrates, this training performance is mirrored by a dramatic deterioration in accuracy and predictive power on the test set, with an MAPE of 43.77%, reflecting a staggering 42.71% drop from its training performance. This phenomenon underscores the model's susceptibility to overfitting noise and nuances within the training data.

Among the myriad of evaluated models, XGBoost emerges as the best performance model, surpassing Adaboost, RF, Linear Regression, and SVM in terms of predictive accuracy. This is due to XGBoost capture the intricacies of the torsional capacity of reinforced concrete columns under pure torsion and compression-torsion loading, a testament to the adequacy of the training samples.

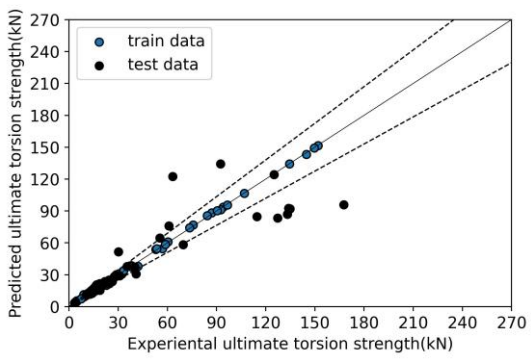
Despite the commendable performance of all five models in predicting structural bearing capacity, as evidenced by their respective performance metrics evaluated on the test set. However, Fig.2 prominently reveals a proportion of samples exhibiting errors exceeding the 15% threshold. This observation underscores the models' struggle to adequately capture specific subsets of data, indicating a limitation in their ability to generalize. The outliers with substantial errors were identified as originating from experimental data pertaining to CFST columns subjected the combined loading of bending and torsion. Even for the model XGBoost, encounters significant challenges when confronted with the 67 test samples encompassing bending-torsion and compression-bending-torsion. This indicated that the models encounter difficulties in learning the intricate and complex interplay between compression, bending and torsion, with a limited availability of bending-torsion experimental data for CFST columns. In other words, the models exhibited pronounced signs of overfitting when confronted with bending-torsion samples, highlighting the need for further refinement and augmentation of the training dataset to better encompass these challenging loading conditions.

Table 1
Statistical information of parameters included 243 experimental data

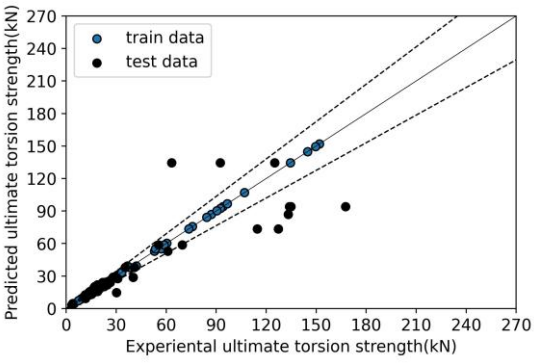
Variable	Count	Mean	Std	Min	25%	50%	75%	Max
Sectional Shape	243	0.71	0.46	0	0	1	1	1
Loading method	243	0.74	0.43	0	1	1	1	1
M/T ratio	243	0.31	0.80	0	0	0	0	4
Axial force rate	243	0.23	0.30	0	0	0	0.3	0.85
Aspect ratio	243	7.23	4.85	1.29	3.00	4	7	20
Wall thickness of steel tube	243	3.40	1.42	1.6	2.1	3.5	4.5	6.5
Yield strength of steel	243	355.3	68.21	242.3	312.8	342.41	397.35	466
Concrete strength	243	32.84	0.10	17.15	24.15	33.10	38.62	54.20
Diameter of steel tube	243	134.4	34.21	90	114	120	160	230
CFRP transverse layers	243	0.57	0.86	0	0	0	0	6
CFRP longitudinal layers	243	0.32	0.68	0	0	0	0	3
Ultimate torsion strength	243	32.86	30.85	3.32	16.8	21.34	28.15	173.5

Table 2
Performance measures for the machine learning models

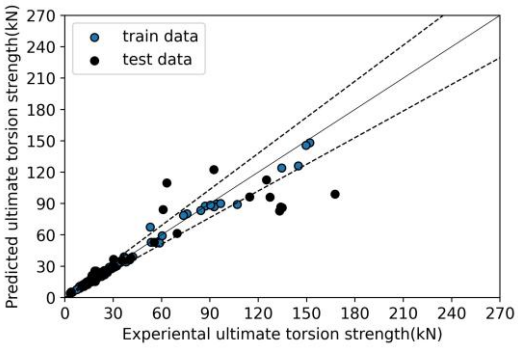
Regression models	Sets	R^2	RMSE(KN)	MAE(KN)	MAPE(%)
Random forest	Test	0.78	17.95	5.25	13.96
	Train	0.99	1.08	0.78	3.97
AdaBoost	Test	0.75	19.85	8.78	13.89
	Train	0.99	0.83	0.40	1.90
XGboost	Test	0.81	16.92	7.60	12.37
	Train	0.98	3.17	1.51	5.05
SVM	Test	0.76	18.95	10.08	26.87
	Train	0.97	4.94	3.93	26.16
LinearRegression	Test	0.81	16.65	10.08	36.07
	Train	0.74	12.80	8.54	42.88



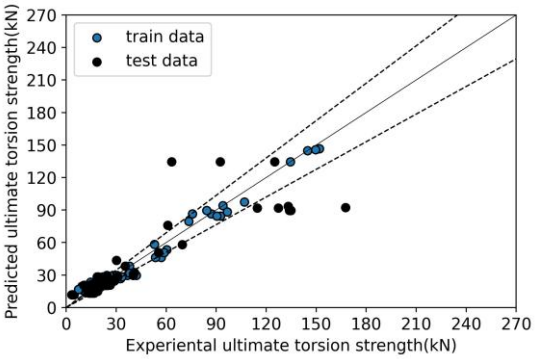
(a) Random forest



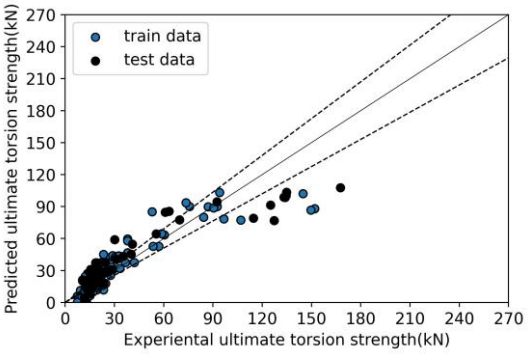
(b) AdaBoost



(c) XGboost



(d) SVM



(e) Linear Regression

Fig. 2 Predicted performance of the machine learning models

3.2. Eliminating non-critical variables

To further explore the potential of mitigating overfitting by reducing model inputs, we implemented a strategy of eliminating non-critical variables. This involved the removal of variables deemed less significant, such as cross-section form and loading method. Table 3 provides a direct comparison of the model's performance under varying numbers of excluded input variables[41]. Based on the 11 characteristic parameters in Table 1, delete the CFRP longitudinal layers, CFRP transverse layers, and loading method in sequence, corresponding to the 10, 9, and 8 characteristic parameters input in Table 3, respectively[41]. Although the removal of non-essential variables exhibited an alleviation of overfitting issues compared to the contrast model with all inputs included, the effect was not pronounced. We concluded that solely optimizing the model, through strategies such as change machine learning models or eliminating non-important variables, proves insufficient to effectively and significantly reduce the model's overfitting problem.

Consequently, we shift our focus from the model itself to the data dimension, intending to adopt a data-centric approach. We aim to achieve fundamental improvements in addressing the model's overfitting phenomenon.

Table 3
Performance measures by reducing model input parameters

Regression models	Sets	R^2	RMSE(KN)	MAE(KN)	MAPE(%)
11	Test	0.81	16.92	7.60	12.37
	Train	0.98	3.17	1.51	5.05
10	Test	0.82	19.85	8.78	11.98
	Train	0.99	2.83	1.42	4.55
9	Test	0.82	16.36	7.60	12.64
	Train	0.99	3.26	1.60	5.10
8	Test	0.83	15.82	7.21	11.91
	Train	0.99	2.88	1.45	4.67

4. Evaluating data-centric strategies for effective overfitting mitigation

4.1. CTGAN-based synthetic data strategy

4.1.1. Establishment of CTGAN

This paper aimed to explore the potential of the non-experimental data in enhancing the availability and quality of data for analysis and modeling, for enhancing prediction accuracy for torsional tests on reinforced concrete columns. The dataset comprises 167 experimental samples meticulously curated from published literature, ensuring the absence of any missing values. It should be noted that, in order to avoid data leakage in the model test set, the generated data based on CTGAN for training the model must only come from the training data. The records are comprehensive, encompassing well-documented information on cross-sectional dimensions, concrete strength, and reinforcement configurations. Utilizing parameters such as concrete strength, steel yield strength, and cross-sectional dimensions as inputs, and the ultimate torsional capacity of reinforced concrete columns as the output, we aim to establish a machine learning model. This approach facilitates the development of predictive capabilities grounded in a robust and comprehensive dataset. To address the potential overfitting issue in machine learning models for predicting the torsional capacity of reinforced concrete columns, arising from data scarcity, we employed the CTGAN methodology to generate 1670 synthetic samples at a ratio of 1:10.

In the pursuit of validating the hypothesis that incorporating additional data, be it synthetic or derived from finite element simulations, into the training process can elevate model performance, it is imperative to emphasize that for all experiments conducted in this study, the test set exclusively comprises authentic experimental data. The selection of model parameters has a significant impact on model performance. In this study, hyperparameter optimization was conducted using the grid search method to identify the optimal parameter combination. Grid search systematically examines all possible combinations of user-defined hyperparameters and selects the configuration that yields the best validation performance. To ensure the reliability of the model and address potential overfitting, external validation was performed using a K-fold cross-validation strategy[42]. In this method, the dataset is divided into K equally sized subsets. In our study, K was set to 10, corresponding to 10-fold cross-validation, as illustrated in Fig. 4.

Table 4 presented the hyperparameter configurations employed for the CTGAN model in this study. CTGAN utilizes fully connected multilayer perceptrons (MLPs) for both the generator and discriminator, consistent with the original architecture proposed by Xu et al [43]. Fig. 3 showed the network structure of CTGAN. The CTGAN model was trained on an Intel i7-13700K CPU (16 cores, 24 threads). Training for 1,000 epochs with a batch size of 16 took approximately 20 minutes, indicating a relatively efficient computation process even without GPU acceleration.

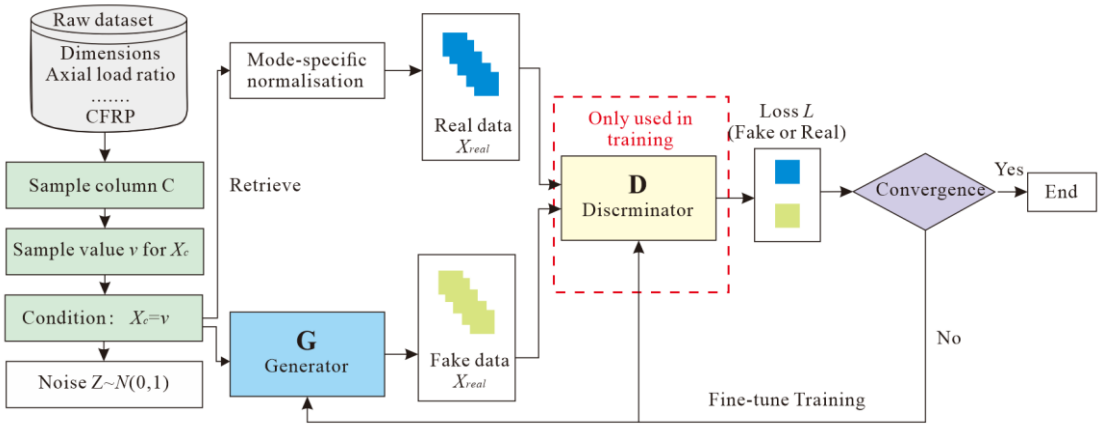


Fig. 3 CTGAN Network Structure

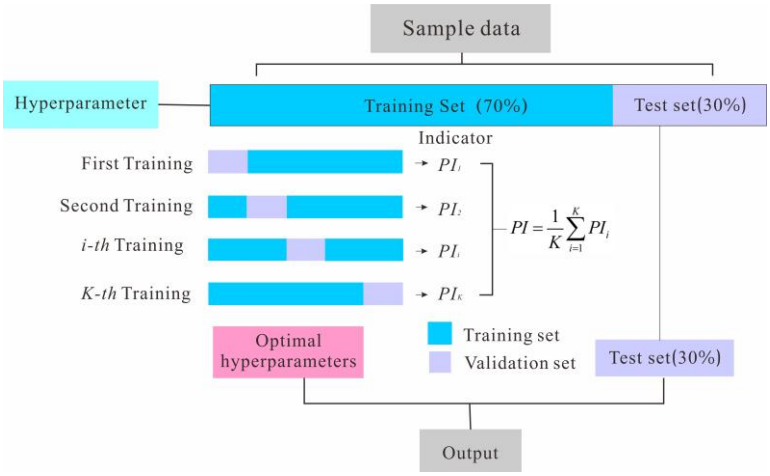
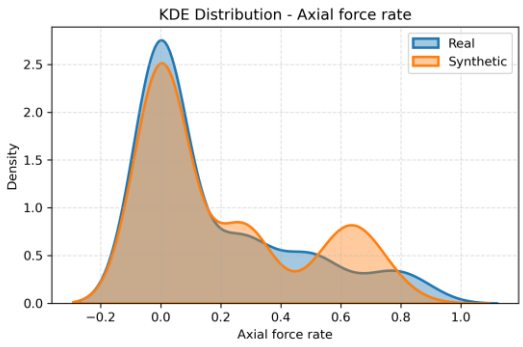


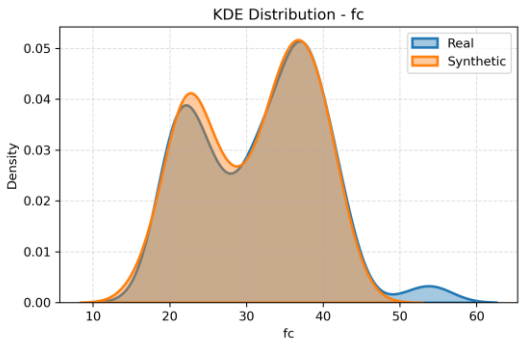
Fig. 4 10-fold cross-validation

Table 4
CTGAN model hyper-parameters settings

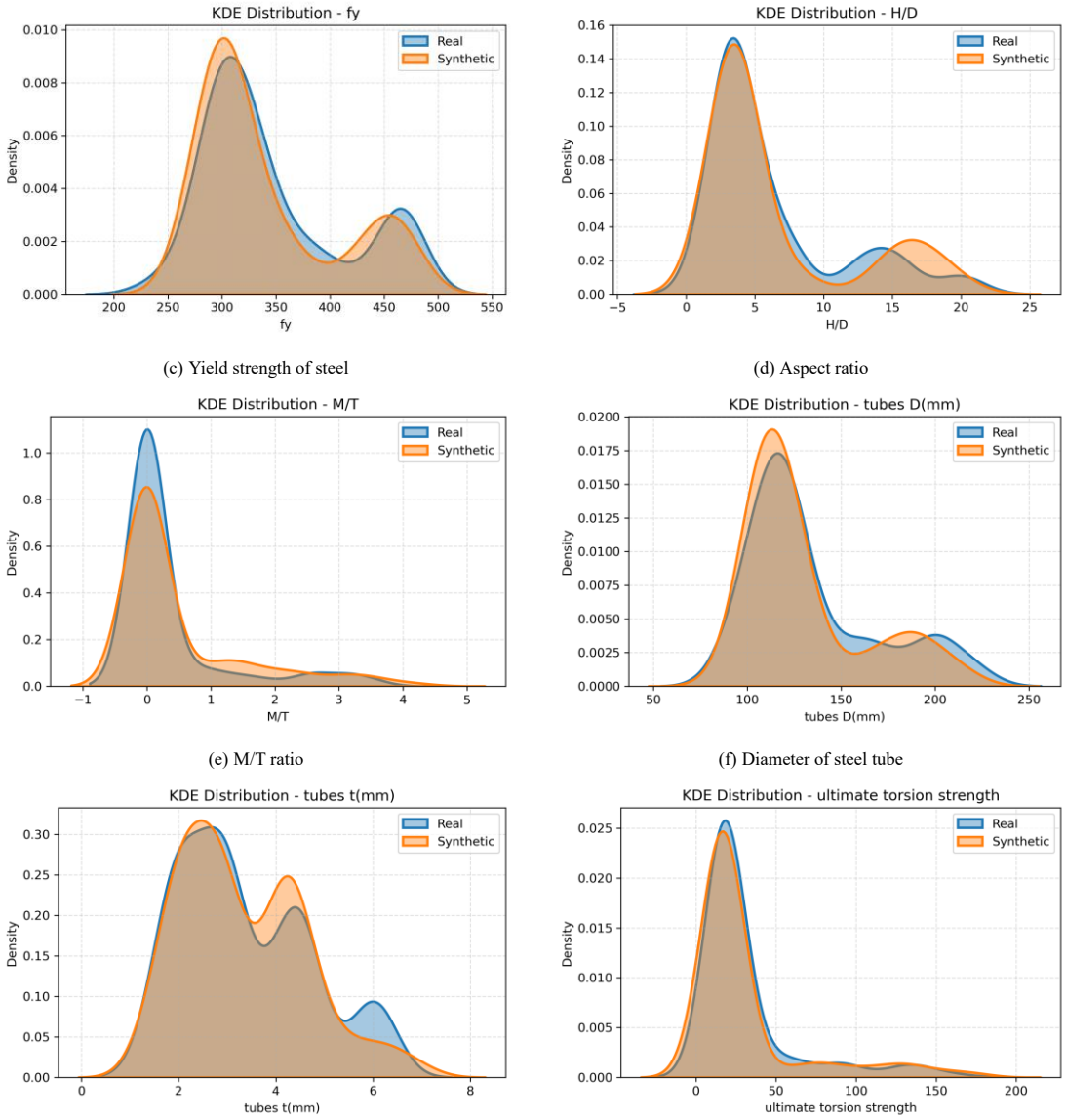
Hyper-parameters	Set Value
RNN cell's units in generator	400
Fully connected layer in generator	100
Layers in discriminator	2
Number of units per layer in discriminator	200
Optimizer	Adam
Learning rate	0.001



(a) Axial force rate



(b) Concrete strength



(c) Yield strength of steel

(d) Aspect ratio

(e) M/T ratio

(f) Diameter of steel tube

(g) Wall thickness of steel tube

(h) Ultimate torsion strength

Fig. 5 Comparison of Synthetic and Real Data Distributions for Key Structural Parameters (KDE Analysis)

To evaluate the quality of the synthetic dataset generated by CTGAN, a kernel density estimation (KDE)-based comparison was carried out between the synthetic and experimental datasets. Several key structural parameters, such as concrete compressive strength, steel yield strength, and aspect ratio, were selected for distributional analysis. As shown in Fig. 5, the marginal distributions of these parameters demonstrate close agreement between the two datasets. These results suggest that the synthetic data can effectively replicate the statistical characteristics of the original experimental samples, thereby validating their representativeness for subsequent structural performance modeling.

4.1.2. Establishment of CTGAN-based synthetic data training strategy

Table 5 illustrated the predictive performance of the ML model trained directly on the entire dataset. However, due to the inherent instability of the generative model, employing the entire sample set (100%) directly for training has resulted in significant bias within the model, ultimately leading to the failure of the training process.

Addressing the issue of model performance collapse caused by directly utilizing vast amounts of synthetic data for training, this study introduces a simple sample screening strategy. The core of the strategy is quantifies the error between predictions generated by a machine learning model, specifically XGBoost (due to its demonstrated superiority in predicting the torsional behavior of CFST columns in this research), and the labels of synthetic data produced by a CTGAN. Subsequently, a subset of samples with the smallest errors is selected for model training.

The error minimization-based sample screening strategy is as follows:

1. Initialization and Model Construction Framework: This study initiates by setting a series of key parameters. Utilizing a comprehensive dataset of CFST columns sets (*M*), XGBoost is chosen to obtain an initial predictive ML model (*F*) through preliminary training.

2. Data Preprocessing and GAN Sample Generation: Following data preprocessing to ensure quality, GAN model parameters are configured, and the number of generated samples is specified. The trained GAN model $G(x)$ then produces a synthetic dataset $D = \{(x_i, y_i)\}_{i=1}^{n}$, where n is the total number of generated samples.

3. Error Evaluation and Sample Ranking: Each synthetic sample is fed into the preliminary predictive model T to obtain a predicted value $f(x_i)$, considered a pseudo-truth in this context. The error between the value y_i and the value $f(x_i)$ is calculated, serving as a critical metric for assessing sample quality. All synthetic samples are then ranked based on their error, ensuring that samples with the smallest errors are prioritized.

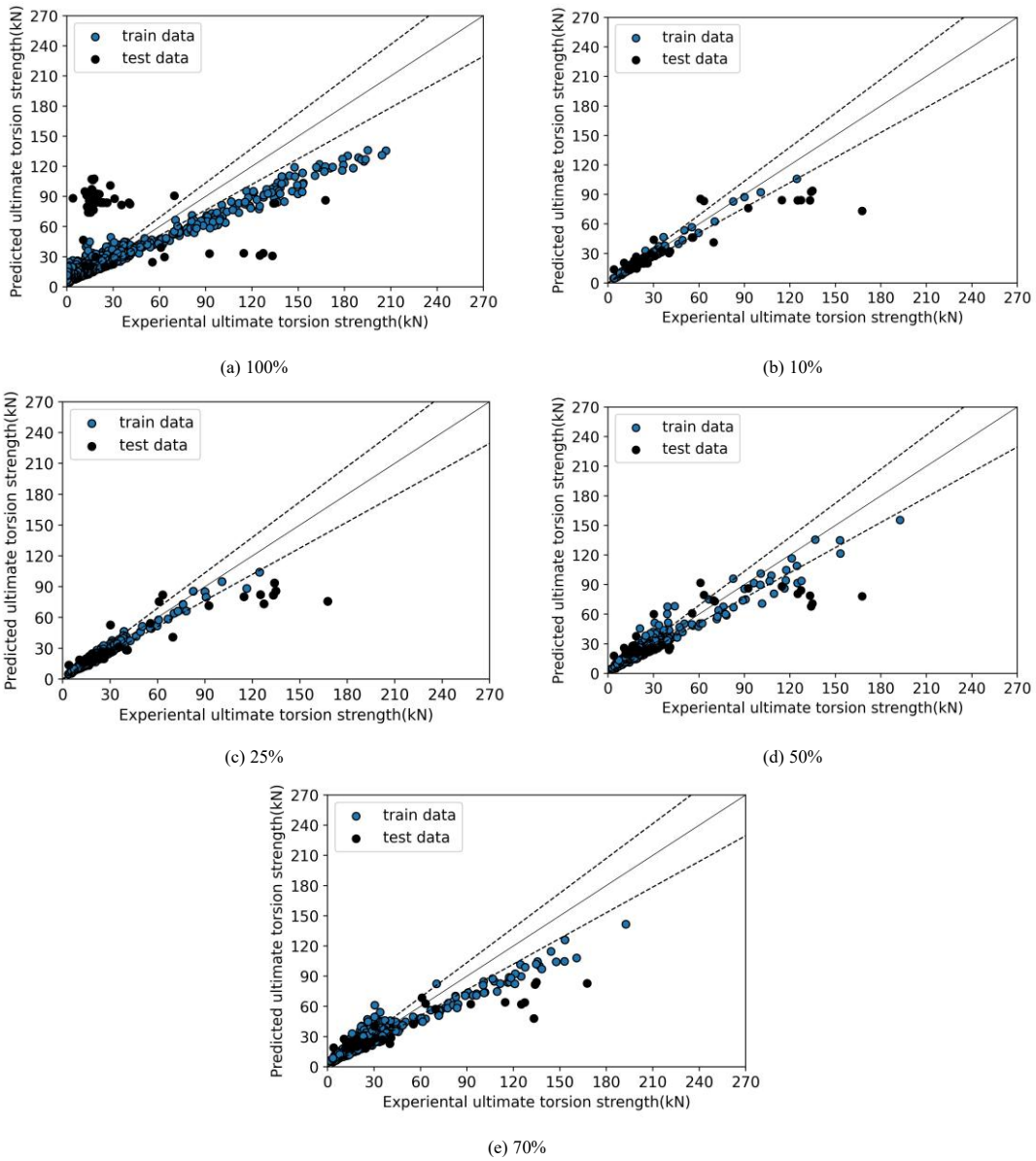
4. Sample Screening Based on Error Minimization: Building upon the ranking, a proportional screening strategy is implemented, selecting the k% samples with the smallest errors to form a new training set $D_k = \{(x_i, y_i)\}_{i=1}^{k}$, where $k \in [0,100]$.

4.1.3. Analysis of CTGAN-based synthetic data training strategy

In pursuit of enhancing model performance, we embarked on an experiment involving a sample selection strategy applied to synthetic data. Specially, the prediction errors of synthetic samples against their true counterparts were calculated,

The top of 10%, 25%, 50%, and 70% data with the lowest errors was categorized and retained resulting in datasets,and subsequent utilized for model

training respectively, shown in Fig. 6. As shown in Table 5, an intriguing trend emerged: as the proportion of retained samples increased, the model exhibited a gradual amplification of bias. Comparative analysis with models trained solely on experimental data revealed a decline in accuracy, marked by a 10.71% reduction in R^2, highlighting the challenges posed by synthetic data. This behavior indicated that while the generative model effectively captures intricate bending-torsion relationships among samples, its inherent instability leads to the generation of samples with significant errors, which subsequently contribute to model bias.



(a) 100%

(b) 10%

(c) 25%

(d) 50%

(e) 70%

Fig. 6 Predicted performance for CTGAN-based synthetic data training strategy

Table 5
Performance measures for CTGAN-based synthetic data training strategy

Regression models	Sets	R^2	RMSE(KN)	MAE(KN)	MAPE(%)
100%	Test	None	63.81	59.59	None
	Train	0.64	9.87	5.64	69.91
10%	Test	0.75	19.29	10.12	24.98
	Train	0.97	2.16	1.13	5.58
25%	Test	0.73	20.01	11.02	28.76
	Train	0.95	2.42	1.52	7.98
50%	Test	0.62	23.87	14.73	41.39
	Train	0.78	6.11	3.61	19.12
70%	Test	0.67	18.68	9.34	43.93
	Train	0.90	4.40	2.67	14.32

4.2. Analysis of FEA-based data training strategy

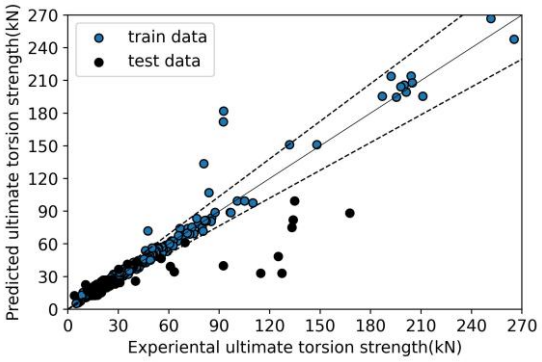
To mitigate these issues, the similar investigation was extended to FEA data. FEA data used in this paper was shown in Table 6. As shown in Fig. 7, when the entire set of FEA data was used for ML model training, an notable improvement in accuracy was achieved. The test set accuracy of 0.55 surpassed that of models trained solely on entire synthetic dataset. Similar FEA-data screening was conducted as mentioned in the error minimization-based sample screening strategy. Upon further scrutinizing the impact of sample selection based on error thresholds, we discovered that retaining simulation samples with errors below 70% achieved the highest model accuracy. When subjected to validation with genuine experimental data, both models trained exclusively on generated data and those utilizing FEA data exhibit a certain level of precision. Nevertheless, a discernible yet consistent gap in performance is observed when compared to models that are trained directly on experimental data. This phenomenon can be attributed to two pivotal factors: (1) the inherent bias in the synthetic data generation model, creating a non-negligible gap between the synthetic and authentic experimental data; (2) the complexity of material constitutive properties and limitations in model solution strategies, collectively impacting the stability of FE simulations, potentially introducing non-physical errors. Additionally, the potential erroneous information embedded within low-quality or non-standard experimental data undoubtedly exacerbates the risk of model overfitting, thereby reducing predictive accuracy.

Table 6
Statistical information of parameters included 561 FEM-based synthetic data

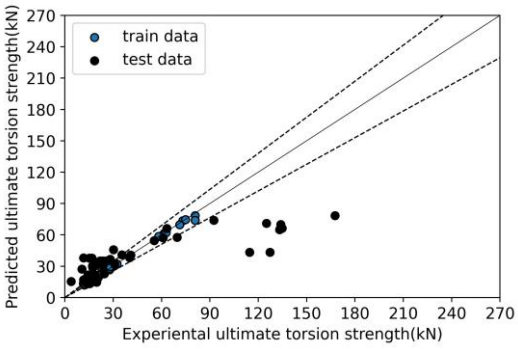
Variable	Count	Mean	Std	Min	25%	50%	75%	Max
Sectional Shape	561	0.96	0.17	0.00	1.00	1.00	1.00	1.00
Loading method	561	0.94	0.15	0.00	1.00	1.00	1.00	1.00
M/T ratio	561	0.70	1.49	0.00	0.00	0.00	1.03	10.25
Axial force rate	561	0.07	0.17	0.00	0.00	0.00	0.00	0.85
Aspect ratio	561	4.01	3.04	0.19	3.00	4.00	4.00	27.00
Wall thickness of steel tube	561	4.31	3.92	1.00	2.40	3.00	5.00	40.00
Yield strength of steel	561	306.1	56.11	234.00	229.00	335.0	345.0	420.0
Concrete strength	561	45.18	14.02	30.00	36.80	36.80	50.00	90.00
Diameter of steel tube	561	190.2	80.11	100.00	120.00	200.0	200.0	400.0
CFRP transverse layers	561	0.87	1.31	0.00	0.00	0.00	2.00	6.00
CFRP longitudinal layers	561	0.18	0.60	0.00	0.00	0.00	0.00	6.00

Table 7
Performance measures for FEM-based data training strategy

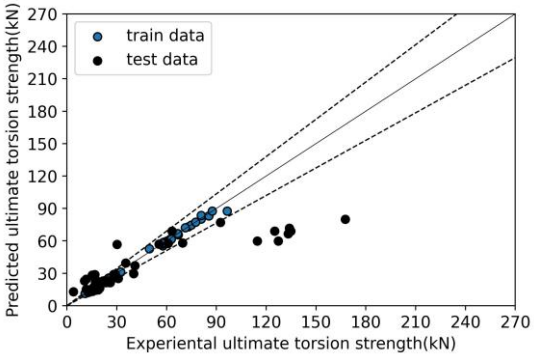
Regression models	Sets	R^2	RMSE(KN)	MAE(KN)	MAPE(%)
100%	Test	0.55	25.93	13.20	29.80
	Train	0.99	21.71	6.03	5.65
10%	Test	0.54	26.19	14.49	42.60
	Train	0.99	1.39	0.72	2.53
25%	Test	0.63	23.53	11.41	26.73
	Train	0.99	1.23	0.60	1.80
50%	Test	0.54	26.31	12.54	26.67
	Train	0.99	2.18	1.21	3.80
70%	Test	0.90	11.89	6.42	23.86
	Train	0.99	2.59	1.46	4.43
80%	Test	0.59	24.86	12.65	28.84
	Train	0.99	3.05	1.62	4.80



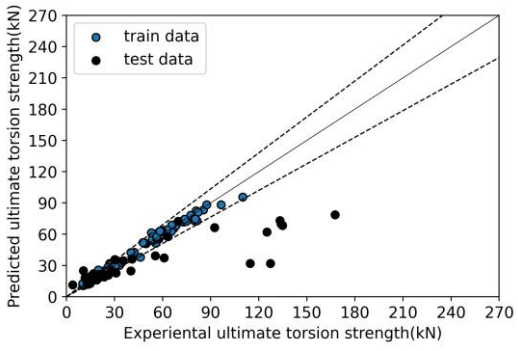
(a) 100%



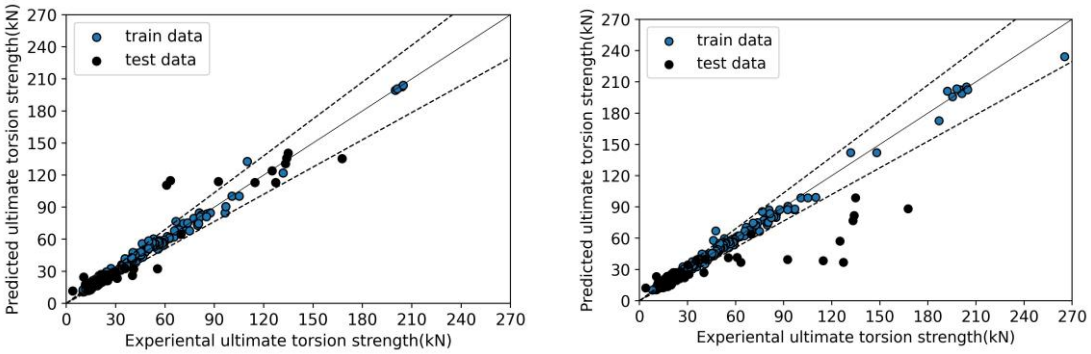
(b) 10%



(c) 25%



(d) 50%

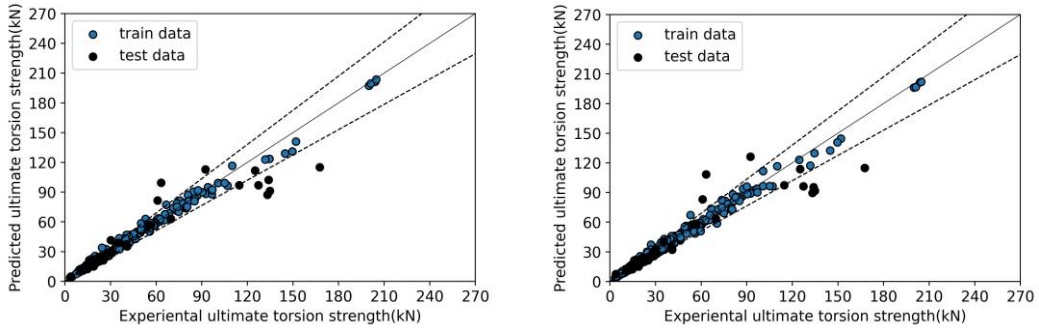


(e) 70% (f) 80%

Fig. 7 Predicted performance for FEM-based data training strategy

4.3. A optimal data-centric strategy to mitigate overfitting in ML models

This section aimed to propose an no-real sample screening strategy that aims to effectively mitigate model overfitting through data-centric preprocessing approaches. Specifically, the XGBoost model trained by the model-centric training strategy due to its exceptional performance was selected as the foundation for our research. Six comparative experimental setups were devised based on the above mentioned analysis results: *Experiment 1*, only the top 10% synthetic data with the lowest errors was utilized for training; *Experiment 2*, the top 70% FEM data with the lowest errors was utilized for training; *Experiment 3*, wherein only experimental data was utilized for training; *Experiment 4*, where a blend of experimental data and the top 10% synthetic data with the lowest errors was utilized for training; *Experiment 5*, where a blend of the top of 70% FEM data with the lowest errors and experimental data was utilized for training, and *Experiment 6*, a blend of experimental data, the top 10% synthetic data with the lowest errors , and the top 70% FEM data with lowest errors was utilized for training.



(a) 70% FEM data and test data (b) 70% FEM data, 10% GAN-data, and test data

Fig. 8 Comparison of predicted performance of different training strategies

Table 8
Performance measures for data-centric training strategies

Experiment	Training data	Sets	R^2	RMSE(kN)	MAE(kN)	MAPE(%)
Experiment 1	10% GAN-data	Test	0.75	19.29	10.12	24.98
		Train	0.97	2.16	1.13	5.58
Experiment 2	70% FEM-data	Test	0.90	11.89	6.42	23.86
		Train	0.99	2.59	1.46	4.43
Experiment 3	Test data	Test	0.83	15.82	7.21	11.91
		Train	0.99	2.88	1.45	4.67
Experiment 4	10% GAN-data+Test data	Test	0.84	18.73	8.01	17.24
		Train	0.99	2.94	1.71	5.15
Experiment 5	**70% FEM-data+Test data**	**Test**	**0.88**	**16.93**	**7.59**	**12.39**
		Train	**0.99**	**2.55**	**1.42**	**4.46**
Experiment 6	70%FEM-data+10%GAN-data+Test data	Test	0.85	14.93	7.07	12.92
		Train	0.99	2.19	1.25	4.63

The performance metrics and validity assessments for *Experiments 1-6* are presented in Table 8. It can be found that, compared with the accuracy to those based exclusively on experimental data of R^2=0.83, the accuracy of the model solely relying on either the top 10% synthetic data with the lowest errors or the 70% FEM data with the lowest errors were 0.75 and 0.90, respectively. However, an enlargement in the RMSE, MAE(kN), and MAPE is observed for the former two experiments when evaluated on the test set.

Furthermore, when the experimental data was introduced to the synthetic or FEM data for model training (as demonstrated in *Experiments 4 and 5*), the accuracy of model with 10% GAN-based data and experimental data increased from 0.75 to 0.84, and there is a marginal decline in the aforementioned error metrics on the test set. The improvement in R^2 signifies an enhancement in the

model's stability and robustness, indicating that the integration of diverse data sources can lead to more resilient predictive models. This can be attributed to the fact that while synthetic data successfully captures the statistical features of real data, its generation relies solely on the currently available physical experiments, which inherently suffer from uneven distributions of certain parameters. Consequently, incorporating synthetic data into training, when directly compared to using solely physical experiments, naturally leads to a slight decrease in test set accuracy, as the synthetic data merely approximates the real data distribution rather than fully capturing its intricacies. CTGAN emerges as an effective approach for generating synthetic data for heterogeneous features and structured tabular datasets; however, it is not without limitations.

Fig. 8 given the prediction performance of the *Experiment 5* and *Experiment 6*. Upon conducting *Experiment 6*, involved integrating all three data sources, i.e., synthetic data, experimental data, and FEA data, as training samples. Specifically, the inclusion of the synthetic data led to a decline in model accuracy compared to *Experiment 5*, and MAPE also exhibited a moderate increased. The minor decline in model performance upon incorporating synthetic data can be attributed to several factors: CTGAN faces challenges in processing high-dimensional features, as the model struggles to learn and generate a large number of unique categories. Additionally, skewed distributions or distributions with a significant proportion of constant values (e.g., a preponderance of zeros in bending-to-torsion and axial compression ratios observed in this study) are difficult for the GAN architecture to capture. For small datasets, synthesis data may be less precise, as CTGAN, like any other deep learning model, thrives on substantial volumes of data. It is important to emphasize that the judicious selection and integration of FEA data with experimental data presents a viable strategy to mitigate the issue of overfitting in predictive models.

5. Conclusion and further work

This study delves into the effectiveness of model-centric and data-centric training strategies in mitigating overfitting issues within ML models, particularly in the context of predicting the torsional capacity of CFST columns. The specific conclusions are articulated as follows:

Firstly, two model-centric strategies were devised: replacing various model and eliminating the input feature parameters. However, under limited sample conditions, these strategies exhibited limited success in mitigating overfitting, highlighting the challenges in achieving significant performance gains solely through model structural optimizations in small datasets.

Subsequently, introduced a data-centric training approach that integrates CTGAN-based synthetic data with FEA-based data, aiming to augment the limited experimental dataset. Notably, training solely on CTGAN-generated data led to model instability and eventual collapse, whereas utilizing FEM data alone sustained model operation but yielded a modest prediction accuracy of 0.55 on the test set.

To further enhance data quality and mitigate model biases stemming from inaccurate training data, a sample selection strategy grounded in the minimum error criterion was devised. This strategy selects synthetic and FEA data that reflect true physical behaviors and exhibit minimal prediction errors, thereby optimizing the datasets. Specifically, a hybrid training approach utilizing the top 70% of FEA samples with the lowest errors, in conjunction with experimental data, notably improved the model's prediction stability and generalization ability, underscoring the efficacy of this strategy in bolstering model robustness.

It is noteworthy that while this study primarily focuses on the efficacy of data-centric training strategies, a detailed exploration of the specific sample selection algorithm was not undertaken. Nevertheless, the proposed method of sample screening based on a minimum error ratio provides a valuable insight and framework for future research endeavoring to efficiently and precisely filter high-quality training samples in big data scenarios, with potential for further development and refinement.

Declaration of competing interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Author statement

Mingxia Dang: Conceptualization, Methodology, Formal Analysis, Writing-Original Draft; Mengxue Guo: Validation, Data Curation, Writing-Original Draft, Supervision, Funding Acquisition; Ying Li: Investigation; Hua Li: Funding Acquisition; Shilin Yang: Resources.

Acknowledgments

This work was supported financially by Natural Science Basic Research Program of Shaanxi Province [Grant No: 2025JC-YBQN-470], Special Scientific Research Program of the Shaanxi Provincial Department of Education [Grant No: 24JX0486], Shaanxi Provincial Department of Science and Technology Talent Program [Grant No: 2024ZC-KJXX-011], Youth Projects of Xi'an Jiaotong University City College [Grant No: 2025Q05] and Research and innovation team of Xi'an Jiaotong University City College: Green Ecological Empowerment and Urban Resilience Innovation Team [Grant No: 037010].

Data availability statement

All data, models, and code generated or used during the study appear in the published article.

Reference

[1] El-Dakhakhni, W. Data Analytics in Structural Engineering. Journal of Structural Engineering, 2021. 147(8): 02021001.
[2] Feng, D.C. Implementing ensemble learning methods to predict the shear strength of RC deep beams with/without web reinforcements. Engineering Structures, 2021. 235: 111979.
[3] Nguyen-Sy T. Predicting the compressive strength of concrete from its compositions and age using the extreme gradient boosting method. Construction and Building Materials, 2020. 260: 119757.
[4] Liu K.H., Xie T.Y., Cai Z.K., et al. Data-driven prediction and optimization of axial compressive strength for FRP-reinforced CFST columns using synthetic data augmentation[J]. Engineering Structures, 2024, 300.
[5] Rahal, K.N. Torsional strength of normal and high strength reinforced concrete beams [J]. Engineering Structures, 2013. 56: 2206-2216.
[6] Deifalla, A. Refining the torsion design of fibered concrete beams reinforced with FRP using multi-variable non-linear regression analysis for experimental results[J]. Engineering Structures, 2021. 226: 111394.
[7] Fiore A , Berardi L , Marano G C. Predicting torsional strength of RC beams by using Evolutionary Polynomial Regression[J].Advances in Engineering Software, 2012, 47(1).
[8] Kim C. Torsional Behavior Evaluation of Reinforced Concrete Beams Using Artificial Neural Network[J].Applied Sciences, 2021, 11.
[9] Zhang T.J., Wang D.L., Lu Y. A data-centric strategy to improve performance of automatic pavement defects detection. Automation in Construction, 2024, 160,105334.
[10] Guo M.X., Huang H., Zhang W., et al. Assessment of RC frame capacity subjected to a loss of corner column[J]. Journal of Structural Engineering, 2022, 148(9):0422122.
[11] Lai D.D., Demartino C., Xiao Y. Interpretable machine-learning models for maximum displacements of RC beams under impact loading predictions[J]. Engineering Structures, 2023,281.
[12] Zakieh A., Hadi G., Amin S., et al. DCServCG: A data-centric sevice code generation using deep learning[J]. Engineering Applications of Artificial Intelligence, 2023, 123,106304.
[13] Sung S.H., Suh J.M., Hwang Y.J., et al. Data-centric arttificial olfactory system based on the eigengraph[J]. Nature Communications, 2024,15:1211.
[14] Li M., Jia G. Multifidelity Gaussian Process Model Integrating Low- and High-Fidelity Data Considering Censoring[J].Journal of Structural Engineering, 2020(3):146.
[15] Luo H. and Paal S.G. Reducing the effect of sample bias for small data sets with double weighted support vector transfer regression. Computer Aided Civil and Infrastructure Engineering, 2021. 36(3): p. 248-263.
[16] Marani A , Nehdi M L .Predicting shear strength of FRP-reinforced concrete beams using novel synthetic data driven deep learning[J].Engineering structures, 2022(Apr.15):257.
[17] Fu B.C., Gao Y.Q., and Wang W. Dual generative adversarial networks for automated component layout design of steel frame-brace structures[J]. Automation in Construction, 2022,146.
[18] Almustafa M.K., Nehdi M.L. Machine learning prediction of structural response for FEP retorfitted RC slabs subjected to blast loading[J]. Engineering Structures, 2021, 244.
[19] Song Z.M., Zhang C., and Lu Y.Y. The methodology for evaluating the fire resistance performance of concrete-filled steel tube columns by integrating conditional tabular generative adversarial networks and random oversampling[J]. Journal of Building Engineering, 2024,97.
[20] Zeng S.H., Wang X., Hua L.Q., et al. Prediction of compressive strength of FRP-confined concrete using machine learning: A novel synthetic data driven framework[J]. Journal of Building Engineering, 2024,94.
[21] Goodfellow, I., Pouget-Abadie, J., Mirza, M., et al. Generative adversarial nets[J],2014. arXiv preprint arXiv:1406.2661.
[22] Beck, J. and O. Kiyomiya. Fundemental pure torsional properties of concrete filled circular steel tubes. Doboku Gakkai Ronbunshu, 2003. 2003(739): 285-296.
[23] Han, L.H., G.-H. Yao, Z. Tao. Performance of concrete-filled thin-walled steel tubes under pure torsion. Thin-Walled Structures, 2007. 45(1): 24-36.
[24] Wang, Y.-H., G.-B. Lu, X.-H. Zhou, Experimental study of the cyclic behavior of concrete-filled double skin steel tube columns subjected to pure torsion. Thin-walled Structures, 2018. 122: 425-438.
[25] Chen, J., W.L. Jin, J. Fu, Experimental investigation of thin-walled centrifugal concrete-filled steel tubes under torsion. Thin-walled structures, 2008. 46(10): 1087-1093.
[26] Nie, X. Ultimate torsional capacity of steel tube confined reinforced concrete columns. Journal of Constructional Steel Research, 2019. 160: 207-222.
[27] Wang, Y.-H. Torsional capacity of concrete-filled steel tube columns circumferentially confined by CFRP. Journal of Constructional Steel Research, 2020. 175: 106320.
[28] Wang Y.H., Guo Y.F., Liu J.P., et al. Experimental study on behavior of concrete filled steel tube columns under torsion and eccentric compression[J]. China Civil Engineering Journal, 2017,50(7):51-61.
[29] Nie J.G., Wang Y.H., Fan J.S. Experimental study on concrete filled steel tubular columns under combined compression, flexure and torsion[J]. Journal of Building Structures, 2012,33(9):1-11.
[30] Wang Y.H., Nie J.G., Fan J.S. Study on the torsion behavior of concrete filled steel tube column with circular section[J]. Engineering Mechanics, 2014,31(3):222-227.

[31] Wang Y.H., Nie J.G., Fan J.S. Cross sectional shear strain disrtibution of rectangular concrete filled steel tubu columns subjected to torsion[J]. Engineering Mechanics, 2014,31(5):101-119.
[32] Wang Y.H., Li S., Zhou X.H., et al. Study on mechanical behavior of concrete filled steel tubular short columns under compound bending-shear-torsion load[J]. Journal of Building Structures, 2017, 38(11):1-12.
[33] Wang Q.L., Ling Z.N., Chen D. Experimental study on torsional behavior of concrete filled CFRP-steel tube with square cross-section [J]. Journal of Building Structures, 2017, 38,S1:478-484.
[34] Jamalpour R. and Hossain K.M.A. Torsion and Combined Torsion-Axial Load Behaviour of Concrete Filled Steel Tube Columns with and without ECC/CFRP Wrap[J]. Journal of Earthquake Engineering, 2024.
[35] Wang Q L, Peng K, Shao Y B. Research on Mechanical Properties of CFRP Confined Concrete-Filled SquareSteel Tubular Under Bending-Torsion Load[J]. Acta Materiae Compositae Sinica,2022, 39(11): 5557–5573.
[36] Wang YH., Wang Y Y., Zhou X H., et al. Coupled ultimate capacity of CFRP confined concrete-filled steel tube columns under compression-bending-torsion load[J]. Structures, 2021,31:558-575.
[37] Wang YH., Nie JG., and Fan JS. Theoretical model and investigation of concrete filled steel tube columns under axial force–torsion combined action[J]. Thin-Walled Structures, 2013,69:1-9.
[38] Nie X., Wang YH., and Li S., et al. Coupled bending-shear-torsion bearing capacity of concrete filled steel tube short columns[J].Thin-Walled Structures,2018,123:305-316.
[39] Yang ZC., Han LH., Zhao HY., et al. Performance of recycled aggregate concrete-filled high-strength steel tubular members under combined compression-bending-torsion[J].Engineering Structures,2025,335:120052.
[40] Zarringol M., Thai H.T. Prediction of the load-shortening curve of CFST columns using ANN-based models[J]. Journal of Buidling Engineering, 2022,51.
[41] Huang H., Xue C.L., Zhang W., et al. Torsion design of CFRP-CFST columns using a data-driven optimization approach[J]. Engineering Structures, 2022, 251:113479.
[42] Feng D.C., Wang W.J., Mangalathu S., et al. Implementing ensemble learning methods to predict the shear strength of RC deep beams with/without web reinforcements[J]. Engineering Structures,2021,235:111979.
[43] Xu L., Maria S., Alfredo C., et al. Modeling Tabular data using Conditional GAN[C]//Advances in Neural Information Processing Systems 32 (NIPS 2019) pre-proceedings, CA: NIPS, 2019.