

Original Article

Development of a Data Preprocessing and Feature Engineering Pipeline for Shear Strength of Shear Wall

Sapna Shah¹, Meenal Mategaonkar², Ami Munshi³

^{1,3}Department of Artificial Intelligence, NMIMS' Mukesh Patel School of Technology Management and Engineering, Mumbai, Maharashtra, India.

²Department of Civil Engineering, NMIMS' Mukesh Patel School of Technology Management and Engineering, Mumbai, Maharashtra, India.

¹Corresponding Author : sapna.shah@nmims.edu

Received: 18 February 2026

Revised: 03 September 2026

Accepted: 21 September 2026

Published: 28 October 2026

Abstract - Concrete shear wall plays a crucial role in effective distribution of seismic forces, thereby adding significant value to earthquake resistance behavior for a structural frame. Hence, understanding failure patterns and deriving shear strength of shear wall has been a significant point of attention for researchers to understand and develop an effective system for structures located especially in seismically active zones. Current experimental dataset, which includes geometrical and material parameters that are usually limited in size, is found to be insufficient for prediction. Hence, in this study, experimental data from two sources, one comprising 252 shear wall data and other 182, have been combined and augmented using mathematical methods, including quantile jittering and noise injection. This is followed by data outlier treatment via quantile capping, winzorization, and Interquartile Range (IQR) methods. The resulting dataset now comprises 1736 data points, one output variable, and 13 predictors, which is available for subsequent feature engineering. Web thickness of wall t_w is the most influential parameter in feature engineering conducted using univariate feature selection and RFE (regressive feature elimination with random forest). The contribution of t_w is found to be significant in Shapley Additive Expansion (SHAP) as well. The 13 predictors have been further reduced to 10 orthogonal components using Principal Component Analysis (PCA). This extensive dataset provides a strong basis for machine learning applications and model development that aim to forecast shear strength with greater precision and generalizability. A reliable tool for creating precise predictive models to understand shear wall behavior has been provided by this dataset.

Keywords - Data augmentation, Shear wall, Shear strength, SHAP, Feature engineering.

1. Introduction

In the seismic area, structural elements forecasting the behaviour and failure pattern of the lateral load resisting system, like shear wall, moment frames and braced frames, are having important role. Various studies in this domain have concluded that buildings, including those with wall and frame structural systems, perform poorly due to inadequate confinement, high axial loads, and unfavourable shear compression interactions, despite being designed using modern codes [1].

According to IITK-BMTPC Earthquake Tip 23 (IIT Kanpur, 2005), concrete-framed buildings, such as vertical reinforced concrete walls known as Shear Walls (Figure 1), along with slabs, beams, and columns. These walls begin at the foundation and run the entire height of the structure. According to IITK-BMTPC Earthquake Tips 23 - Ductile Design of Shear Walls Earthquake Tips, 2024, shear wall thickness can vary from 150 mm to 400 mm in high-rise

buildings. Figure 2 shows shear walls are often placed throughout the structure's breadth and length to convey seismic loads to the foundation by acting as broad, vertically oriented beams. In seismically active areas, concrete shear walls are essential lateral load resisting components, particularly when built as squat structures [2].

For structural safety, accurate shear strength prediction is crucial. However, because classic mechanics-based models simplify complicated nonlinear behaviours, they frequently exhibit significant variation and bias.

Predicting shear strength of reinforced concrete shear walls with opening and boundary parts using standard design code requirements is more difficult. To overcome this constraint, researchers have developed an analytical model based on mechanics that reliably predicts the shear capacity of RC walls with openings and boundary elements and shows improved agreement with experimental results [3].



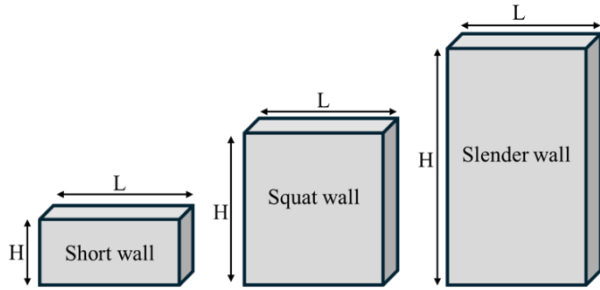


Fig. 1 Types of shear walls

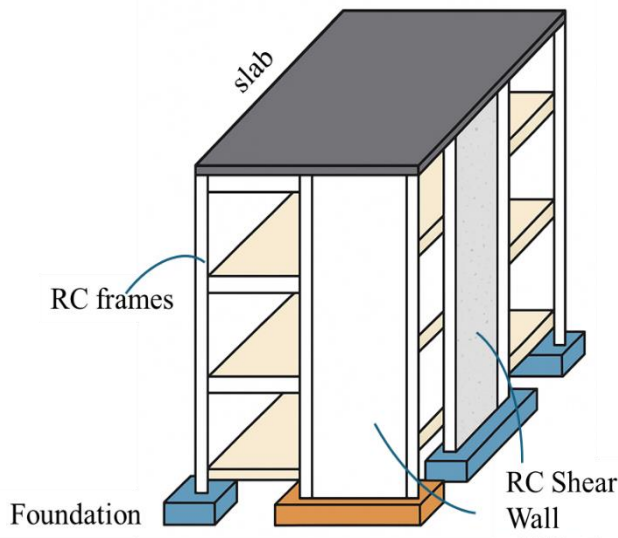


Fig. 2 Lateral load-resisting system

2. Literature Review

Predicting the shear strength of the shear wall remains difficult due to the complex interactions between geometry, material properties, axial load, and reinforcing details [4]. The modified compression field theory, such as strut and tie model and frequently used conventional mechanics-based techniques, such as semi-empirical formulas found in the design code, like ACI 318, 2022, are straightforward approaches that frequently fall short of capturing non-linear material behaviour and multivariate interactions.

In the past, engineers had to rely on their experience and perform calculations in order to create safe and efficient structures. However, new technology related to Artificial Intelligence (AI) and Machine Learning (ML) had gained importance the new era of designing. Various machine learning techniques, such as Deep Neural Network (DNN), Artificial Neural Network (ANN), Decision Tree, ensemble models such as Random Forest and Bagging, Gradient Boosting, etc., are being used to solve structural engineering problems. These methods can easily identify nonlinear correlation and interactions between input parameters. More accurate predictions can be produced using these data-driven

models than traditional empirical formulas [5, 6]. Gondia A. had applied genetic programming to estimate the shear strength of flanged squat walls [7]. As per Naderpour et al. ANN algorithm applied on different experimental datasets for prediction of squat and flanged reinforced concrete walls [8]. Feng et al. used gradient boosting techniques to characterize the nonlinear relationship between wall parameters and shear strength, which yielded more accurate results than code-based predictions [9]. Keshtegar et al. performed adaptive harmony search like optimised neural network to get accuracy based on non-linear behaviour [10]. In 2024, Nguyen et al. used a dataset more than 600 wall specimens and proved that ensemble learning models such as XGBoost, CatBoost and gradient Boosted regression trees comparative outperformed both finite element simulations and code-based methods [4]. Recent development for the machine learning prediction models emphasized on advanced data augmentation for RC shear wall. In 2025, I-Tung Yang and Handy Prayogo has introduces a novel ML framework that integrates a Physics-Aware Conditional Variational Autoencoder (PA-CVAE) for realistic data augmentation, aiming to enhance model robustness and generalization [25]. Recent studies, 2026 by Guoxu Zhao et al. [26] propose a Deep Reinforcement Learning (DRL) method for design of reinforced concrete shear walls based on the Soft Actor-Critic (SAC) algorithm. Their work demonstrated expanding use of AI approaches by successfully combining design code requirements and achieving quick structural design of shear wall. Analytically derived data is used for predicting non linear push over responses of reinforced concrete shear walls by proposed Digital Twin (DT) and Machine Learning (ML) framework [27]. Table 1 summarizes recent studies related to prediction of shear strength of shear wall focusing on researchers methodology, key findings, and limitations to clearly find out the research gap for this study.

3. Motivation and Research Gap

To predict the shear strength of reinforced concrete shear wall, it became necessity to apply recent techniques, especially computational techniques like machine learning and artificial intelligence. Most published studies using Machine Learning (ML) algorithms on shear wall datasets have focused simply on applying ML to an unaddressed problem: how complete they are as a dataset from which to make predictions. Shear wall shear strength prediction has been studied extensively, but the data used for developing these models is limited both in terms of quantity/difficulty because not much data (including shear wall specimens) is very diverse. Most publicly available datasets are too small in sample size and have very little variation in both material properties and test conditions, limiting the generalizability and robustness of trained models. This study created a data pre-processing framework that uses EDA, data augmentation, outlier treatment methods, and feature engineering to improve the quality and usefulness of current shear wall shear strength data. Research gap is there in the consistency, completeness,

and interpretability of datasets as a result of prior research on shear-strength prediction mostly concentrating on model building rather than systematic data curation procedures. This study integrates insights from structural mechanics with state-of-the-art data preparation techniques to provide a more

resilient and refined dataset that future researchers can dependably use for advanced modelling and code-calibration investigations. This concurrent focus on methodological rigor and technical reasoning is advantageous to the fields of structural engineering.

Table 1. Tabulated literature review

Study (Author, Year)	Methodology	Key Findings	Limitations
Kuria, K.K., Kegyes- Brassai, O., 2026	Using SAP2000, SPO2FRAG around 200 shear walls specimens are analysed with ML digital twin based model and RF regression to predict bilinear parameters for nonlinear seismic response.	The model predicted initial stiffness, yield displacement, and post-yield stiffness ratio with very high accuracy. This reduces the time taken for non linear FE analysis to produce pushover curves.	Valid only within the parameter range under study. Accuracy depends on quality of the dataset considered and may not extend beyond the training bounds. It represents flexural behaviour instead of local shear dominated failure modes.
I-Tung Yang, Handy Prayogo., 2025	Classification ML models are generated using data augmentation and synthetic data samples using PACVAE to predict failure modes of RC shear walls and then compared it to ROS, SMOTE, ADASYN, and CTGAN using Bayesian optimization.	Compared to traditional oversampling techniques, the suggested PA-CVAE produced samples that were more physically significant and realistic. Evaluation matrix were all enhanced.	High-quality representative training data is still required for the investigation. A strong experimental database cannot be completely replaced by synthetic augmentation; results are constrained by scarce, imbalanced, or limited RC shear wall samples. A bigger, more consistent dataset is necessary for better generalization.
Tian, S., Ni, X., and Wang, Y., 2025	Used a stacking ensemble of multiple machine learning models to predict seismic shear strength of reinforced concrete shear walls	The stacking model achieved very high prediction accuracy, with reported R2 around 0.98 and a low coefficient of variation compared with traditional shear capacity models.	The model still depends on the quality and coverage of the available dataset; broader and more balanced experimental data are needed for stronger generalization.
Le Nguyen, K., Trinh, H. T., 2024	ML models, e.g. XG Boosts, CatBoosts, GBRT, RF, with Bayesian Optimization used for lateral load resisting system like squat shear walls.	Identified the most suitable ML approach for predicting lateral strength with good practical usefulness.	limited by the specific wall type studied and the squat-wall dataset that is now available; more comprehensive data coverage is still needed. Because models have limited interpretability, the actual explanation of why one model works better than another may not be totally evident for practical engineering usage.
Keshtegar, B., Nehdi, M.L., Kolahchi, R. et al., 2022	Developed a novel hybrid ML model for predicting the shear strength of RC shear walls.	The hybrid model improved prediction accuracy and performed better than conventional empirical methods.	Specific types of wall and related datasets are used, so model is still constrained by the same. So broader validation and diverse training datasets are required.

4. Methodology

Figure 3 illustrates the entire process of study for analyzing experimental data acquired from different sources, and can be used further for prediction models. In addition to encouraging measurement accuracy, an intense data collecting system offers a platform for combining cutting-edge technology, which eventually leads to innovation in both industry and research. In many scientific and industrial applications, where data may be precisely monitored and evaluated to make better decisions, data collecting is an essential procedure.

4.1. Data Acquisition

The dataset used in this study is methodically gathered from two reliable sources presents comprehensive experimentally tested databases that are critical in the data-driven modelling of the shear wall responses. The experimental data contain extensive data sets of shear wall experiments, which were chosen because of their usefulness

and a wide parameter space of wall geometry, material behaviour, and loading cases. The first source used in this study found in Massone and Melo et al. [11]. This dataset contains a large set of 252 experimental data of short wall experiments on Reinforced Concrete (RC) elements under different shear and flexural loads. It was first gathered to assist in the formulation of a generalised analytical model to forecast the shear strength of the shear wall in accordance with panel response. The other dataset is sourced from the published work of Ning and Li (2017), which is accessible via Wiley online library, comprises of 182 experimental records of RC squat walls are extracted from the source, tracking 14 specific structural features, including wall and boundary element geometry, material properties, reinforcement ratios, axial load and experimental shear strength. The Generalized shear-strength database of RC squat walls article published in Earthquake Engineering & Structural Dynamics, which is available in the Wiley online library, is the source of the data used in this work [31].

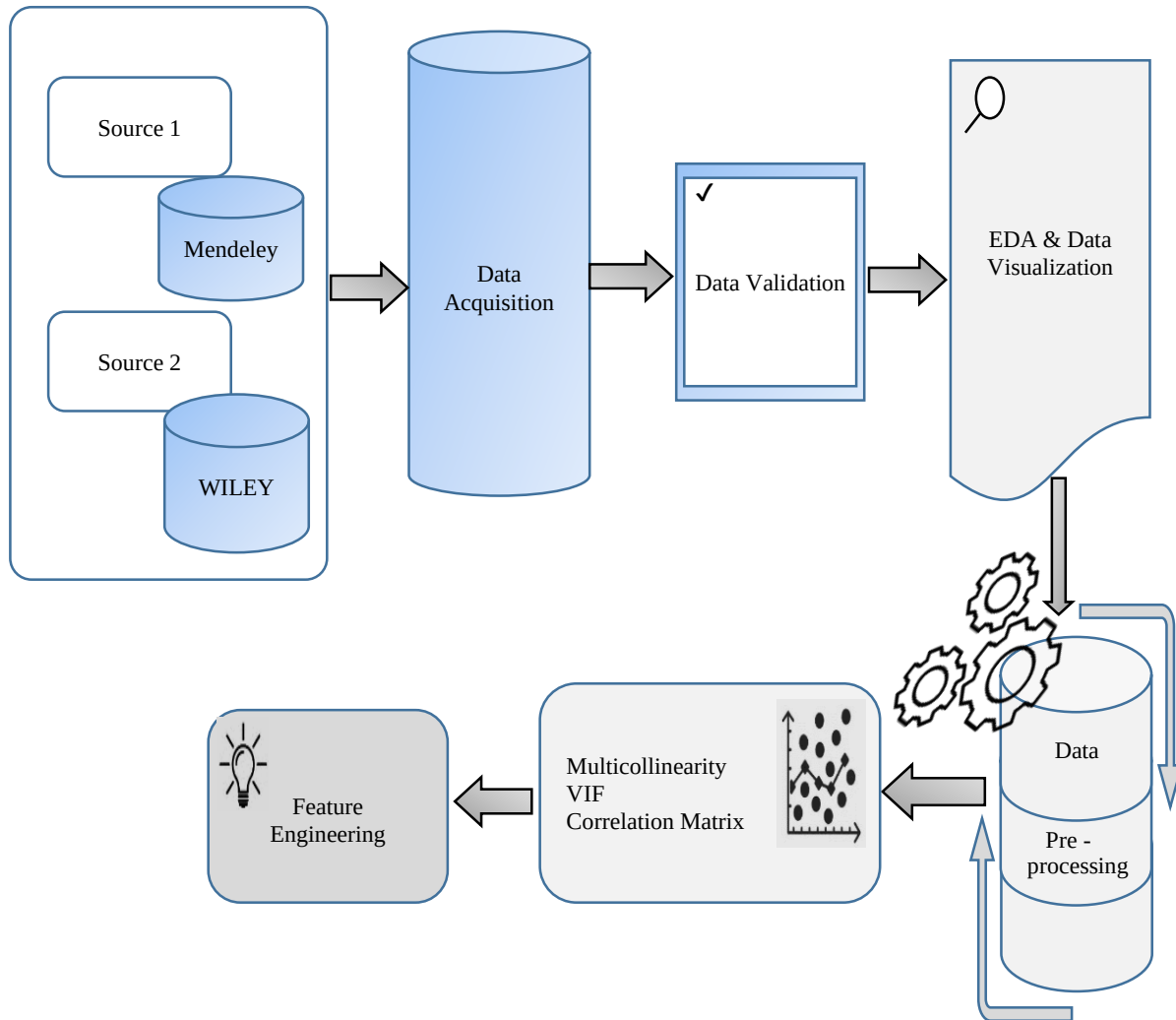


Fig. 3 Methodological process diagram

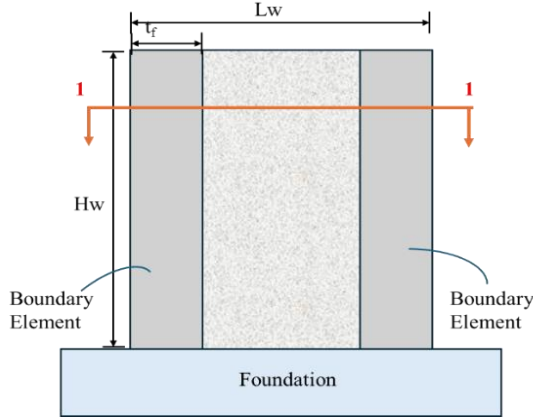


Fig. 4 Cross-section of shear wall with boundary element

The supporting document will also give a comprehensive tabular sum of the experimental test results of Reinforced-Concrete (RC) structural walls, which we have modified to the prediction of the shear strength of shear walls in this particular research. All the parameters of relevance to the specimen and the test (wall height, length, thickness, flange, dimensions, reinforcement ratios, concrete strength and yield strengths, axial loads, shear strength) were digitised based on the tabulated output as given in Table 2.

Table 2. Symbols and description of parameters

Symbols	Description
H_w [mm]	Height of the wall
L_w [mm]	Length of wall
t_w [mm]	Thickness of wall
t_f [mm]	Thickness of boundary element (walls)
L_f [mm]	Length of boundary element (walls)
f'_c [Mpa]	Concrete compressive strength
ρ_t [-]	Distributed transversal reinforcement ratio
ρ_L [-]	Distributed longitudinal reinforcement ratio
ρ_b [-]	Main (boundary) longitudinal reinforcement ratio
f_{yt} [Mpa]	Distributed transversal reinforcement yield stress
f_{yL} [Mpa]	Distributed longitudinal reinforcement yield stress
f_{yb} [Mpa]	Main longitudinal reinforcement yield stress
N [kN]	Axial load
V_{test} [kN]	Experimental shear strength

The finished dataset serves as the basis for reliable, broadly applicable predictive modeling of shear strength in shear walls made of reinforced concrete. The cross-section of a typical shear wall is depicted in Figures 4 and 5, which include every attribute taken into account for the dataset.

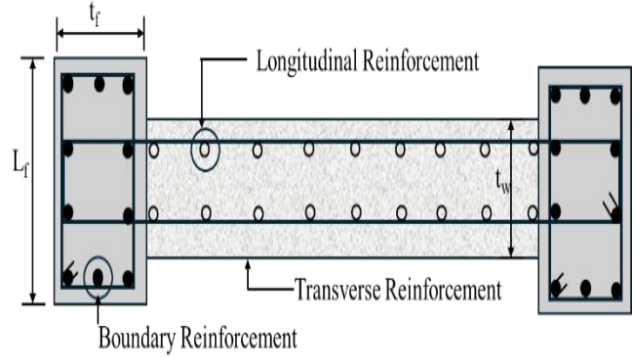


Fig. 5 Section 1-1

4.2. Data Validation

Three code-aligned validation parameters are created and statistically assessed to make sure the input features utilised in the predictive framework adhere to ACI 318 shear design guidelines.

As mentioned in ACI 318 -19, the parameter $\sqrt{f'_c}$ directly reflects the concrete contribution used in ACI shear strength equations ($V_c \propto \sqrt{f'_c}$) as specified in clause 22.5.5.2. The concrete strength range is typical of RC wall applications, according to the dataset's mean value of 4.97 and small dispersion (std = 0.79). Another parameter, web reinforcement $\rho_L f_{yL}$, corresponds to the ACI statement as given by Equation 1 in capturing the transverse reinforcement contribution.

$$V_S = \frac{A_v f_{yv} d}{s} \quad (1)$$

Robust model learning across various confinement conditions is made possible by the observed mean of 1.93 and maximum of 5.28, which confirms sufficient variety in reinforcement levels, from non-reinforced examples to substantially reinforced webs.

$$\frac{v}{b_w l \sqrt{f'_c}} \quad (2)$$

Lastly, the normalised shear stress as given in Equation 2 is created using the stress-format comparison method utilised in ACI 318 -19 clause 18.10.4. as given in Table 3, the low mean value (0.000897) and narrow range seen in dataset imply that all specimens fall within actual shear-demand limitations relative to concrete capacity. All these statistics support the suitability of the modified features for model training and code-based verification, confirming that they accurately

reflect ACI-compliant measurements of concrete strength, web reinforcement efficacy, shear-stress demand.

Table 3. Analytics of features as per ACI compliant

Statistic	$\sqrt{f'_c}$	$\rho_L f_{yL}$	$\frac{V}{b_w l \sqrt{f'_c}}$
Count	434.00	434.00	434.00
Mean	4.965825	1.928743	0.000897
Std	0.793557	1.274994	0.000534
Min	3.507136	0.000000	0.000219
25%	4.312772	0.964432	0.000585
50%	4.919350	1.706513	0.000766
75%	5.474941	2.667930	0.001063
Max	6.848357	5.280000	0.004645

The domain-informed visualisations shown in Figures 6 and 7 aid in determining the impact of important parameters on shear wall behaviour as well as the dataset's agreement with basic mechanics and ACI 318 design assumptions. Figure 6, which shows shear strength vs $\sqrt{f'_c}$ distinguished by different colours coded as thickness (t_w) of the wall, shows a distinct trend of rising shear strength with higher concrete strength, which is in keeping with ACI 318's assumption that shear capacity depends linearly on strength of the concrete. The projected material-driven contribution to shear resistance is validated by the distributed points and their rising direction.

However, Figure 6 shows the platuing behaviour as web reinforcement grows by plotting the normalized shear strength given in Equation 2 against the web reinforcement ratio given as $\rho_L f_{yL}$.

This pattern supports the widely held belief that shear reinforcement loses effectiveness once concrete cracking and aggregate interlock approach their practical limits by indicating diminishing returns in shear capacity above a specific reinforcement threshold.

The collective visualizations support the presence of mechanical consistency within the dataset, confirming that these parameters can be used to develop predictive models of shear strength. An ACI 318-19 compliance analysis confirmed that the data supporting these experimental models adheres to design guidelines established in this code.

Limiting Shear strength is defined as $V_{max} = 0.83\sqrt{f'_c} A_{cv}$ where $A_{cv} = b_w \cdot l$ represents the critical shear

area. This constraint sets an upper bound on the nominal shear capacity of structural walls and acts as a baseline for finding physically unrealistic or non-code-compatible data.

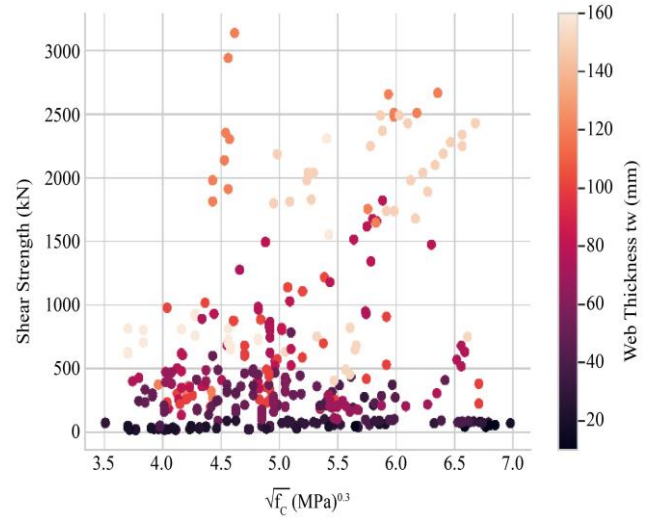


Fig. 6 Domain-informed visualization

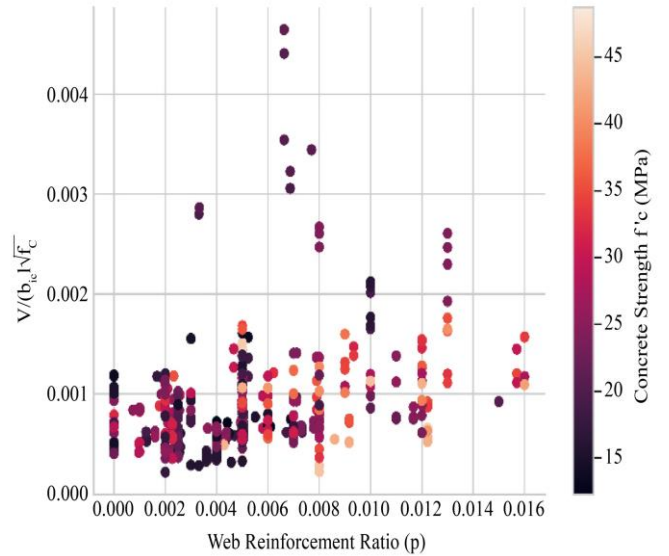


Fig. 7 Web reinforcement variation

The measured shear strengths of all of the samples in the study have been validated against this constraint. All of the samples in the study were validated against this constraint, and the results showed that the measured shear strengths were inside of the ACI-regulated envelope.

The validation process identifies potential anomalies caused by experimental noise, unique loading conditions, and/or atypical wall configurations. In addition, the confidence in the quality of the samples enables the entire dataset to be used as a basis for machine-learning analysis and/or parametric analyses under ACI 318 guidelines.

4.3. Exploratory Data Analysis (EDA)

Exploratory analyses were performed for critical variables, including mean, standard deviation, coefficient of variation, etc., and compared to distributions from reference papers to validate data quality, as well as verify an appropriate degree of dispersion between the datasets and therefore shape the data for ML modeling.

As shown in Table 4, there is considerable variation among the shear wall datasets regarding their geometric and material characteristics.

As given in Table 4, descriptive statistics has performed for the key geometrical characteristics, as well as the axial load and lateral load were included in the shear wall analysis to forecast shear strength.

From the analysis, it has been observed that wall height (H_w) and wall length (L_w) have variations of around 56% and 59 %, respectively. This shows the variety of dimensions present in the dataset for the structural analysis and design concerns. Also, there is a wide range of wall sections, like

thickness of the web (t_w) and flange (t_f), roughly around 64% and 78%, respectively. But flange length (L_f) has more diversity, around 106% in the dataset. This analysis shows how the sample taken for dataset differs, which can affect the structural characterization and design.

According to Table 5, the concrete's compressive strength (f'_c) varies from relatively low to high values. This suggests that the material strengths of various samples vary.

The mean strength of 26.63 shows that it is compatible with a standard grade used for typical structural applications, with variations reflecting changes in concrete quality and curing conditions. It is crucial to incorporate both central tendency and variability when creating shear strength prediction models.

The median values for each variable display distributions that differ somewhat from the mean. These statistics highlight the variability of the data set, which is necessary for creating trustworthy and widely applicable models for predicting shear wall shear strength.

Table 4. Analytics of the geometric features and the force acting on the shear wall

	H_w [mm]	L_w [mm]	t_w [mm]	t_f [mm]	L_f [mm]	N [kN]	V_{test} [kN]
minimum	150.00	420.00	10.00	30.00	30.00	0.00	16.38
maximum	2200.00	3960.00	160.00	360.00	1000.00	2365.00	3138.13
mean	893.75	1331.38	70.50	123.48	211.08	178.15	526.79
median	845.50	1220.00	60.00	102.00	145.00	0.00	281.00
Std. dev.	496.85	782.08	43.13	75.99	224.20	369.63	644.08

Table 5. Reinforcement ratio and material property analytics

	ρ_t [-]	ρ_L [-]	ρ_b [-]	f_{yt} [MPa]	f_{yL} [MPa]	f_{yb} [MPa]	f'_c [MPa]
minimum	0.00	0.00	0.00	0.00	0.00	208.90	12.30
maximum	0.04	0.04	0.11	624.00	624.00	605.00	104.00
mean	0.01	0.01	0.03	360.25	357.75	372.80	26.63
median	0.01	0.01	0.03	351.65	353.00	369.00	24.20
Std. dev.	0.01	0.01	0.02	121.23	119.36	76.12	12.14

The statistics pertaining to reinforcement and its yield stress are compiled in Table 5. The reinforcement ratios (ρ_t , ρ_L , ρ_b) exhibit low mean values that are consistent with traditional minimal reinforcement processes in proportion to wall size, with a coefficient of variation of roughly 25%.

Variability in steel grades is reflected in the yield strengths for transverse and longitudinal reinforcement (f_{yt} , f_{yL} , and f_{yb}), which vary by around 11% around mean values of 360 – 372 MPa. Robust shear wall behaviour modelling is supported by the dataset's varied reinforcing and loading conditions, which are highlighted by this range of properties as well as the wide variance in applied loads and shear capacity.

4.4. Data Visualization

Figure 8 displays the frequency distributions of the primary input features used in the study. The majority of variables exhibit distinct central trends, with data grouping around particular ranges.

A number of input characteristics, including material strengths, wall geometry, and reinforcement ratios, show skewed or unimodal distributions, which indicate typical values that are common in the dataset.

Variability in structural configurations is further highlighted by the observation of extreme or zero values for specific reinforcement scenarios.

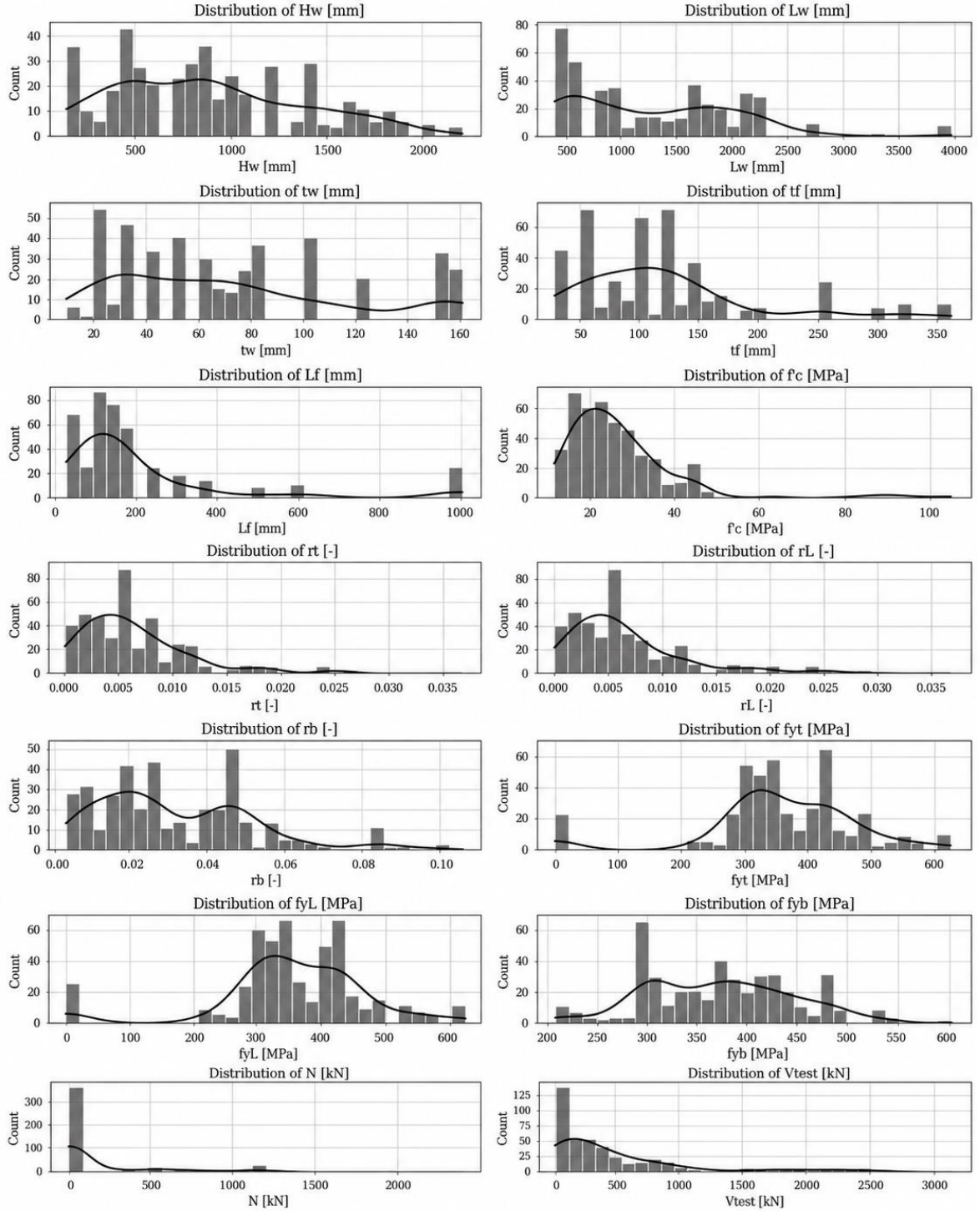


Fig. 8 Distribution graph for predictors

As shown by the box plots, $IQR = Q3 - Q1$, where $Q1$ is the first quartile (25th percentile), and $Q3$ is the third quartile (75th percentile). Table 6 shows the number of outliers found for each parameter, indicating significant variations between features. Wall height (H_w) and web thickness (t_w) showed no outliers, however axial load (N) and flange length (L_f) showed

higher counts with 102 and 60 outliers, respectively. There are also a significant number of outliers in other characteristics, including tested shear strength (V_{test}), tensile strength (f_{yt} , f_{yL}), and total reinforcement ratio (ρ_t , ρ_L) (24, 24, 26, 35, 45, respectively). This quantification guarantees that the abnormal values are identified and dealt with in the analysis.

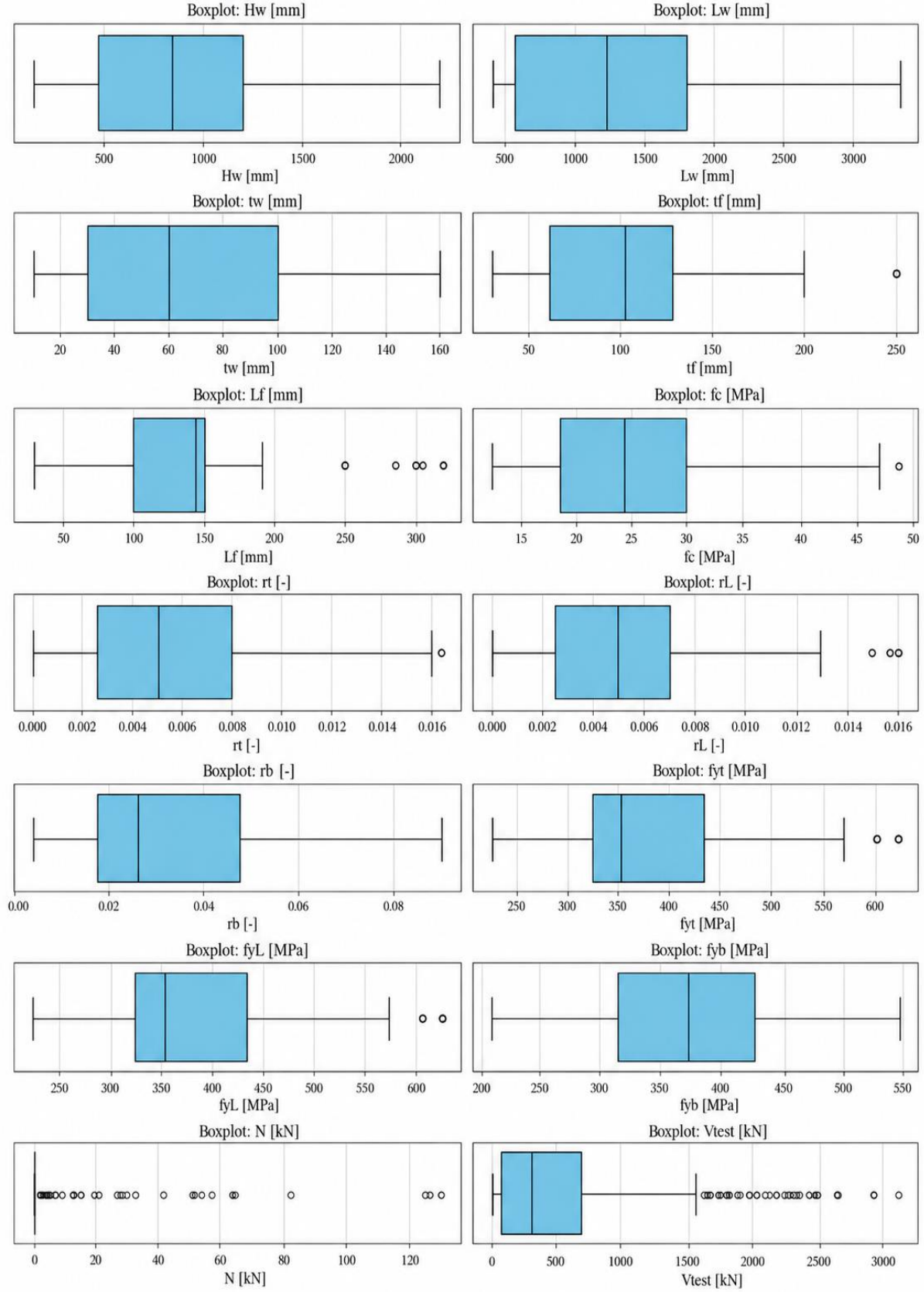


Fig. 9 Outlier detection

Table 6. Feature outliers

H_w mm	L_w mm	t_w mm	t_f mm	L_f mm	f'_c MPa	ρ_t	ρ_L	ρ_b	f_{yt} MPa	f_{yL} MPa	f_{yb} MPa	N kN	V_{test} kN
0	8	0	28	60	9	26	35	4	24	24	2	102	45

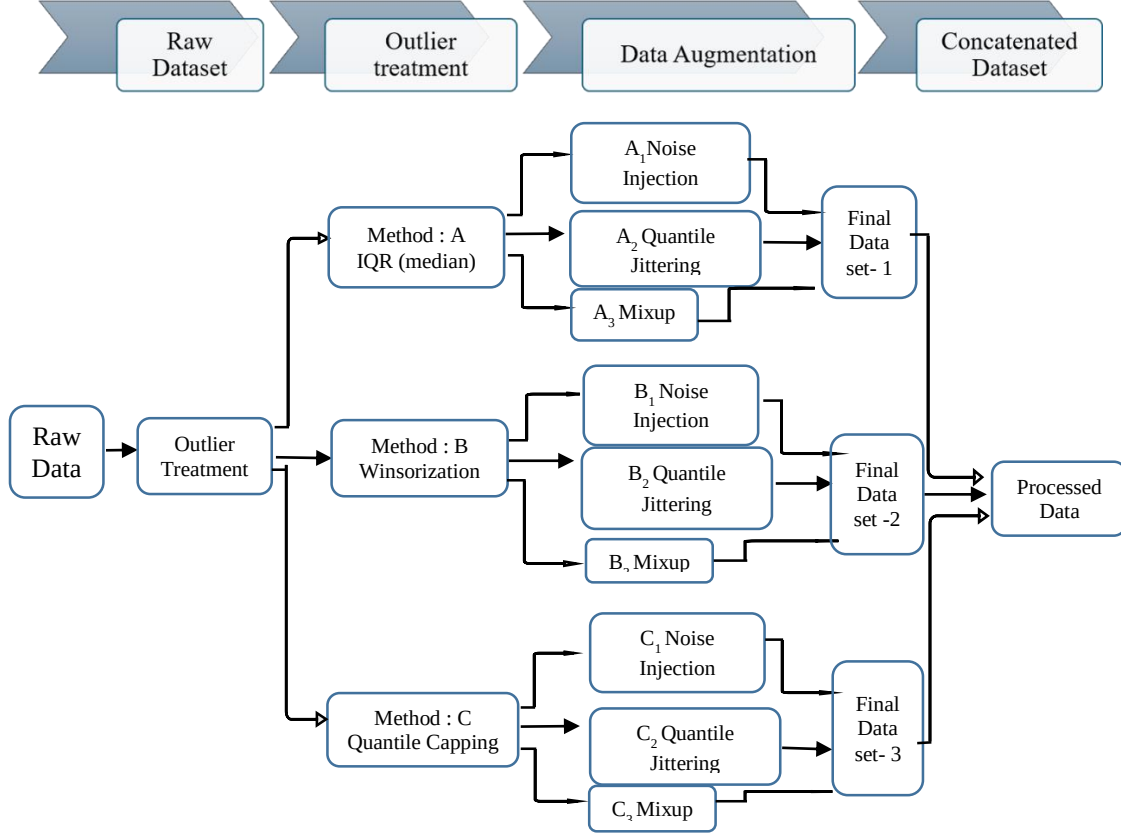


Fig. 10 Data preprocessing and transformation pipeline

4.5. Data Pre-processing

Data preparation is carried out as shown in Figure 10 to guarantee the dataset's integrity and dependability once outliers were found, utilising the aforementioned techniques. To ensure consistency, outlier values are eliminated or imputed using the proper statistical methods. A cleaned and analysis-ready dataset for additional modelling was produced by adding irregular data points, for instance, by replacing them with mean or median values as needed. An essential part of data preprocessing is dealing with outliers. This study employs three outlier handling techniques.

- IQR based outlier treatment using median
- Winsorization
- Quantile Capping

The Interquartile Range $IQR = Q3 - Q1$ is used in the first technique to identify outliers, and any values that fall below $Q1 - 1.5 \times IQR$ or above $Q3 + 1.5 \times IQR$ are deemed extreme. By substituting the median for these outlier values, the distributional structure is maintained while the impact of extreme data is diminished.

Instead of eliminating extreme values, winsorization caps them at predetermined percentile criteria. In order to reduce the impact of outliers while preserving sample size, values below the selected lower percentile (such as the 5th) are substituted with that percentile value, while values above the upper percentile (such as the 95th) are substituted with the corresponding upper percentile. By substituting values outside of predetermined quantile limits with the boundary values themselves, quantile capping corrects outliers. To provide controlled truncation of extreme data points, observations below a selected lower quantile (such as 1st percentile) are set equal to that lower quantile, and observations above the upper quantile (such as the 99th percentile) are made equal to that higher quantile. A 434×14 cleaned dataset was produced following outlier removal and utilising three distinct techniques. Box plots were used to visually examine this imputed dataset in order to verify that there were no extreme values. However, data augmentation became an essential preprocessing step to enable efficient machine learning-based prediction of shear strength in shear walls due to the small sample size.

4.6. Data Augmentation

Data augmentation was used as an essential preprocessing step to artificially enlarge the dataset due to the small amount of observations. Three different outlier treatment methods are initially applied to the original dataset of 434, performed to get the three cleaned data but still, this dataset, 434×14 , is not sufficient for predictions. For each dataset, three augmentation techniques 1. Noise Injection 2. Quantile Jittering 3. Mixup Augmentation are then applied independently. Thus, each technique of outlier generated an augmented dataset containing 1736 specimens (434 original specimens + 3×434 augmented specimens). This procedure resulted in three different datasets, each corresponding to a different outlier handling method, but having the same size, 1736×14 . By increasing the diversity and quantity of data, this method reduces the risk of overfitting that results from small sample sizes and supports accurate machine learning predictions for shear wall shear strength.

4.6.1. Noise Injection

To enhance the data diversity and to replicate the measurement variability, sometimes numerical features are supplemented with noise. The initial value is taken from a normal distribution with a controlled variance and zero mean. Mathematically, this is represented as $x' = x + \epsilon$ where x is the original value, and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ [12]. This method works well for simulating uncertainty in the real world and enhancing model generalisation.

4.6.2. Quantile Jittering

Jittering usually entails random scaling or small shifts based on a uniform distribution, in contrast to Noise Injection, which adds Gaussian noise. To guarantee wider feature coverage, values close to particular quantiles were disturbed. Every data point x in a Q quantile range is transformed as $x' = x + \delta$, where δ is a very small random offset considered from a uniform distribution and not from predefined intervals [13]. It encourages models to learn from various situations, and this technique of data augmentation encourages models to learn from rare situations and reduces the sparsity of data around distribution tails.

4.6.3. Mixup Augmentation

Pairs of data points and their labels were combined linearly to create new samples using Equation (3).

$$\begin{aligned} x' &= \lambda x_i + (1 - \lambda)x_j \\ y' &= \lambda y_i + (1 - \lambda)y_j \end{aligned} \quad (3)$$

Where, x_i, x_j and y_i, y_j are feature and label pairs from two dataset entries, and $\lambda \in [0,1]$ is considered from the beta distribution [14]. This technique strengthens the model's robustness to input changes while reducing overfitting. After all three augmentation techniques are used, the final dataset has 1,736 samples with 14 properties, providing a substantial

and diverse data pool for predictive modeling. This expanded dataset is therefore perfect for creating trustworthy machine learning algorithms to accurately determine the shear strength of shear walls.

4.7. Multicollinearity

Before implementing a machine learning model, multicollinearity must be addressed, and feature selection must be optimized by focused feature engineering. A quantitative and visual assessment of inter-feature interdependence was provided by the analysis of the correlation matrix and related heatmap, which revealed linear relationships among predictors. For every input variable, the Variance Inflation Factor (VIF), which provides a direct indicator of multicollinearity, was also computed. By reducing the negative effects of highly correlated variables, these studies aid in the identification of duplicated features, direct the selection of pertinent predictors, and guarantee robust model performance. The precise numerical values for the correlations between the target and the attributes are displayed in Figure 11.

The correlation matrix shows strong positive intercorrelations among geometric features (up to 0.78) and notable correlations with target (up to 0.69), according to the correlation matrix. Certain characteristics exhibit a negative correlation, which means that as one rises, the other usually falls. For example, the variance indicates that for roughly 15% of their shared pattern in the sample, ρ_b decreases as L_f increases. There is multicollinearity, particularly amongst geometric characteristics. Variables may be carrying redundant information in the model if correlations are greater than 0.7.

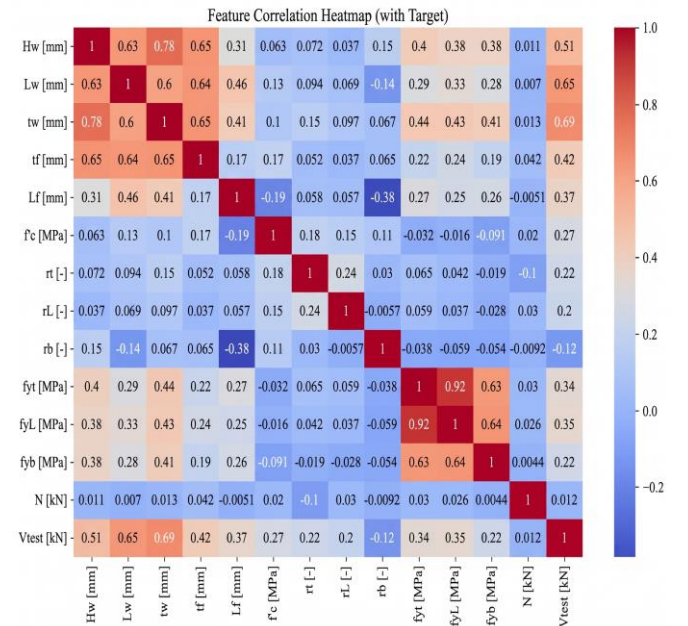


Fig. 11 Feature correlation

To verify that the features are co-linear, the Variance Inflation Factor is also computed. According to the VIF results, f_{yL} [MPa] and f_{yt} [MPa] have the greatest inflation factors, both of which are greater than 7. This shows significant multicollinearity and suggests that these predictors require careful consideration or may require modelling adjustments. Furthermore, t_w [mm] and H_w [mm] exhibit moderate VIF values between 3 and 3.5, indicating that, albeit to a lesser extent, they also contribute to multicollinearity. High multicollinearity can skew model findings and make statistical models harder to interpret, hence predictors with VIFs more than 5 should be reevaluated or integrated.

4.8. Feature Engineering

Selecting features is one of the first stages of developing effective machine-learning models. Consequently, you will need to select features that greatly affect the load capacity, the collapse mechanism, and the displacement characteristics of RC shear walls. Feature and dimension reduction is illustrated in Figure 12. Also, SelectKBest (univariate feature selection), a popular filter-based feature selection method, is used for further refinement. The F-score, which is the ratio of variance within groups to variance between groups, is used in this procedure.

Equation 4 defines F-score for feature selection [15].

$$F(x_j) = \frac{\sum_{k=1}^c n_k (\mu_{j,k} - \mu_j)^2}{\sum_{k=1}^c n_k \sigma_{j,k}^2} \quad (4)$$

Where n_k is the number of samples in class k , $\mu_{j,k}$ is the mean of the j th feature for class k , μ_j is the overall mean of the j th feature, and $\sigma_{j,k}^2$ is the variance of the j th feature within class k [15]. This score compares variation within and between classes to assess each feature's discriminative potential.

In mathematical terms, each feature's F-score measures how effectively it predicts the target variable in regression tasks or distinguishes between the target classes. Top K parameters with maximum F-scores are chosen for the final model once parameters are ranked.

Using this filter-based method, the top five features — t_w [mm], L_w [mm], H_w [mm], t_f [mm], and L_f [mm] — were shown to be the most important predictors in light of the findings, guaranteeing that the model keeps only the most informative inputs for robust analysis.

Recursive Feature Elimination (RFE) with Random Forest Regressor is utilized to evaluate the relative value and ranking of each input feature. The model is recursively fitted in each RFE iteration, features are rated based on their impact (I_j) for variable j was evaluated using Mean Decrease in Impurity (MDI) based on total MSE reduction across T

decision trees [16]. A mathematical depiction of feature significance for Random Forest is given by Equation 5.

$$I_j = \frac{1}{T} \sum_{t=1}^T \Delta MSE_{j,t} \quad (5)$$

Where $\Delta MSE_{j,t}$ represent the weighted reduction in mean squared error achieved by feature j across all split nodes in tree t .

The variables that consistently contributed most to prediction performance were L_w [mm], t_w [mm], $f'c$ [MPa], f_{yt} [MPa], and f_{yL} [MPa], which were ranked best and kept (RFE rank 1, support True, importance >0.03). These contributions are summarized in Table 7, which directs the selection of a more precise and economical model.

4.9. SHAP-based Interpretability and Dimensionality Reduction

In order to create precise machine learning frameworks, the selection of input features is essential. Rigorous feature engineering includes choosing structural and material parameters, such as wall thickness (t_w), wall length (L_w), and concrete compressive strength ($f'c$), to capture both physical relevance and model sensitivity. In structural prediction tasks, feature selection was based on domain expertise to ensure that input variables directly affect target outcomes. SHAP is an additive feature attribution technique that has its roots in game theory; that is, the model's output is a linear sum of the input variables [9]. Feature extraction techniques preserved the original measurement scales in order to preserve the interpretability of SHAP results.

Table 7. Feature engineering rank

Feature	RFE Ranking	Support	Importance
L_w [mm]	1	Yes	0.323031
t_w [mm]	1	Yes	0.434079
$f'c$ [MPa]	1	Yes	0.031079
f_{yt} [MPa]	1	Yes	0.037023
f_{yL} [MPa]	1	Yes	0.06177
L_f [mm]	2	No	0.024955
ρ_L [-]	3	No	0.020779
H_w [mm]	4	No	0.020052
ρ_b [-]	5	No	0.013099
ρ_t [-]	6	No	0.015281
f_{yb} [MPa]	7	No	0.010761
t_f [mm]	8	No	0.007735
N [kN]	9	No	0.000356

As per Equation 6, SHAP values (ϕ_i) decompose a model prediction $f(x)$ in terms of average contributions of each parameter satisfying the efficiency property [17].

$$\sum_{i=1}^M \phi_i = f(x) - \mathbb{E}[f(X)] \quad (6)$$

Where M is number of parameters considered, making sure that the departure from the anticipated model output equals the total of all feature attributions. By using mean absolute SHAP values to provide both local (per-sample) and global (dataset-wide) interpretability, this ensures consistent and locally correct significance assessments. In feature space development for structurally interpretable machine learning models, engineering features like t_w and L_w were prioritized because, as Figure 12 illustrates, they had the biggest SHAP impact. The initial set of 13 correlated features was converted into 10 orthogonal principle components using Principle Component Analysis (PCA), which reduced dimensionality while preserving most of the data variance. The number of PCA components was determined using a cumulative

explained variance threshold of 95%. First 10 principal components were retained, which was reduced from 13 variables to 10 variables while preserving 95% of the variance in the standardized input data. PCA offers a condensed representation of the input data and aids in reducing redundancy among correlated variables, even if it converts the original engineering variables into linear combinations and so decreases their direct physical interpretability. The PCA-transformed components were utilized in the prediction framework, but the original engineering variables were kept for engineering interpretation. By calculating eigenvectors and eigenvalues from the standardized data's covariance matrix and projecting the original features onto directions of highest variance, PCA mathematically finds new axes. This made it possible to extract features efficiently, reduce noise, and enhance the computational efficiency and interpretability of the study's later machine learning models.

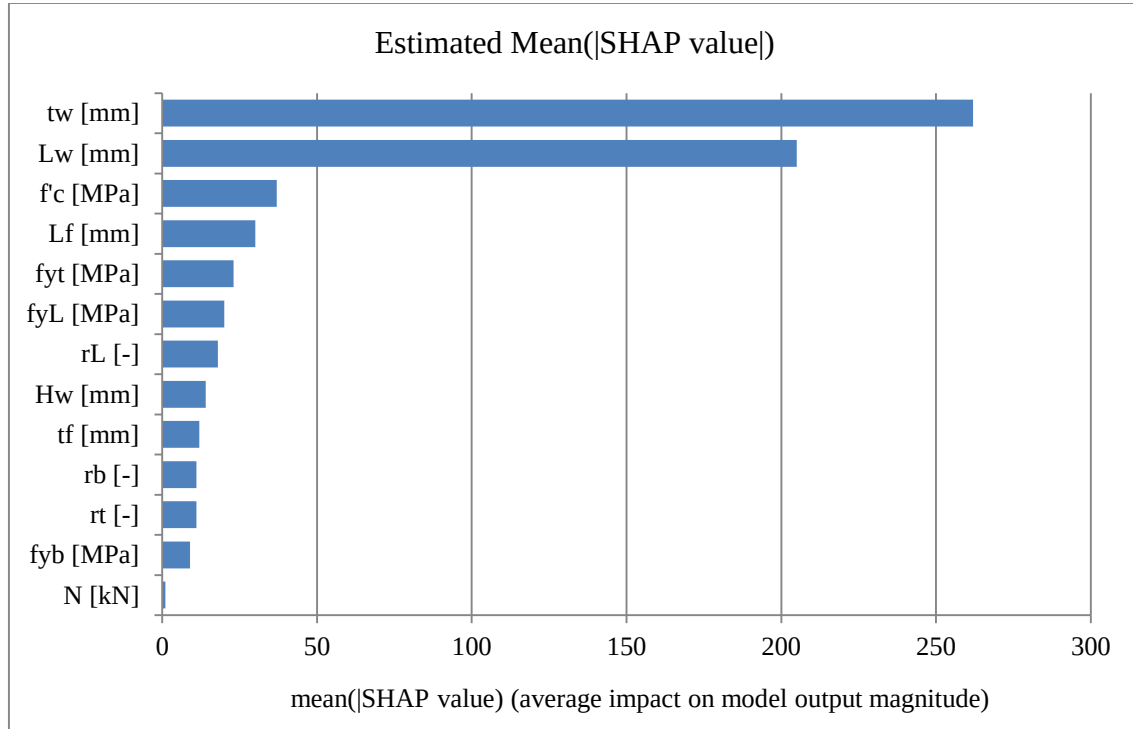


Fig. 12 Mean SHAP value

5. Results and Discussion

Data taken from both sources have been validated with the American standard using various formulas given in ACI 318-19, and it has been observed that the experimental dataset considered is in compliance with the standards. The analysis shows that geometric features in the heatmap typically explain around 40% to 60% of the variation in each other and in the target variable. In comparison, most material properties account for less than 10% of the variation with other features or the target, highlighting their weaker influence in this dataset. There are outliers detected from the exploratory data analysis showed that some outlier handling techniques has to

be done for data preprocessing. From three different methods, it has been observed that method with median replaced all outliers using IQR had removed all the outliers from the dataset. Although Winsorization was applied using the 5th and 95th percentiles as thresholds, the method caps extreme values at these boundary limits rather than removing them. While using the Quantile Capping method with the 1st and 99th percentile boundaries, many outliers were not imputed because only values beyond these extreme thresholds were replaced. Consequently, some observations remained outside the typical range and continued to be identified as outliers during subsequent analysis.

Since answer outliers frequently reflect real variability and informative extremes, the target variable is not taken into account for the treatment. The true link between predictors and the target may be distorted by imputing these values, which could result in biased model training and decreased predictive accuracy. By choosing the appropriate augmentation technique, the unique characteristics of the structural engineering dataset are effectively handled.

The effectiveness of various augmentation techniques in tackling the specific attributes of structural engineering datasets is carefully evaluated. Noise injection mimics measurement errors and material variability, making the augmented data more representative of the real-world conditions encountered in both experimental and field studies. By enhancing coverage of both normal and extreme parameter ranges, quantile jittering plays a crucial role in mimicking rare yet critical structural behaviors. Mixup augmentation enhances model generalization and resilience against overfitting by offering more diverse feature combinations and smooth transitions between samples. When used together, these techniques enhance the reliability and applicability of machine learning models for predicting structural properties.

Multicollinearity of the augmented data set showed that there is strong and positive relation amongst geometrical parameters of the shear wall. The steel yield stress are also highly correlated. The target variable has a moderate positive correlation with dimensions of the web, but in contrast, concrete compressive strength and axial load shows weak correlation with shear strength of the wall.

Feature engineering using three different methods has given top features to be selected for the prediction of the shear strength. Recursive Feature Elimination (RFE) with Random Forest Regressor performed and showed that parameters like L_w [mm], t_w [mm], f'_c [MPa], f_{yt} [MPa], and f_{yL} [MPa] has more importance than 0.03 to be considered for further analysis.

The summary of SHapley Additive Explanations showed two parameters. L_w and t_w have contributed to the most variability for the prediction of shear strength of shear wall. The next influencing factor, as per SHAP analysis, is concrete compressive strength f'_c and L_f followed by the yield stress of the longitudinal and transverse reinforcement. Dimensionality reduction has applied using Principal Component Analysis, which reduced 13 features to 10 orthogonal principal components.

6. State of the Art Comparison

Recent studies mainly focused on the prediction model for estimating shear strength of shear wall. Le Nguyen et al. (2024) compared multiple ML models for lateral strength estimation in squat shear walls, demonstrating that ensemble methods can perform well but remain dependent on the quality

of the available dataset. Tian et al. (2025) used stacking ensemble learning to predict seismic shear strength with high accuracy. Although Yang and Prayogo in (2025) focused on failure mode classification rather than shear strength prediction, the study is relevant to the present study because it demonstrate the application of data augmentation to address the limited RC shear wall dataset. Keshtegar et al. (2022) proposed a hybrid machine learning model for shear strength prediction, their study still relied on small training samples and did not fully address the need for a standardised, reusable dataset. Unlike other efforts that primarily focus on prediction techniques, the current study emphasizes systematic data quality improvement, which is crucial for limited experimental shear wall datasets and for improving downstream machine learning performance.

According to Zaker Esteghamati et al. (2025), generalizability, explainability, and overfitting/underfitting continue to be major obstacles to the practical use of machine learning models in structural engineering, despite their promising performance. This emphasizes the necessity of transparent, consistent data-handling procedures, which is the main contribution of this work.

Strong predictive performance was attained by Gokce et al. (2024) using ML models on a database of 432 non-conforming RC shear walls; however, the study still depends on a fixed experimental database and lacks a standardised, publicly reusable preprocessing pipeline, which restricts wider replication and reuse.

In their review of building ML/DL/RL preprocessing, the S Z Golazad et al. (2024) demonstrated the necessity of cleaning, transforming, scaling, and augmenting raw data prior to model training. The review's primary drawback is the absence of reusable, standardized pretreatment procedures for structural issues unique to a given topic. This demonstrates the necessity of an open data flow for predicting shear-wall strength.

Previous research in structural machine learning has primarily improved the modelling state of the art, but it still relies on ad hoc, tiny, non-standardized experimental datasets with no transparency regarding preprocessing. On the other hand, by offering a standardised feature set, a documented cleaning and augmentation procedure, and a reusable pipeline that may enable further research on shear-wall strength prediction, the current study advances the state of the art in data infrastructure. Additionally, it makes the comparison with earlier work more dependable and significant by strengthening the state of the art in evaluation through more rigorous and repeatable model assessment.

The present study, however, aims to construct an all-encompassing data pipeline that includes steps for preprocessing, cleaning, and augmentation to ensure a more

reliable dataset for future machine learning applications. Since the refined data will be made publicly accessible to foster reproducibility, benchmarking, and further model advancement, this provides a valuable foundation for upcoming researchers focused on predicting the shear strength of shear walls.

7. Conclusion

AI techniques can be useful for the prediction of behavior of shear walls. However, it is challenging to create a thorough and trustworthy dataset. This study focuses on prepping the data systematically, starting with data outlier treatment to deal with outlier elimination to guarantee data quality.

Following conclusions are drawn from the study:

- Shear capabilities vary with a mean of 526 kN throughout a wide range of web heights from 150 to 2200 mm, according to the assembled dataset of 434 shear wall testing findings.
- The majority of variables exhibit a compact distribution with few outliers; for example, the IQR method of outlier detection yielded 28 in t_f , 8 in L_w , 60 in L_f , and 9 in $f'c$, which require careful outlier treatment. With the exception of 8 L_f outliers, IQR method with median has replaced nearly all of the outliers across the three techniques.
- Three data augmentation approaches are used to increase the dataset's resilience and variability, which transformed 434 experimental results with 14 parameters into final augmented data of 1736 x 14. The enriched and unified dataset was created by concatenating the supplemented outputs.
- Multicollinearity is be considered while interpreting model results or selecting features for analysis. Analysis

showed shear capacity has a positive correlation of 0.69 with t_w and on other hand, boundary reinforcement ρ_b has a negative correlation of 0.12 with V_{test} . The correlation matrix pattern shows that geometric dimensions and reinforcement yield strength are the principal drivers of variance in the shear strength of wall. For modeling and design, more emphasis is placed on accurately characterizing these parameters, while the influence of $f'c$ and axial load may be of secondary importance in this specific dataset.

- Feature engineering's univariate feature selection revealed that t_w had the highest F score of 1775.3. Additionally, feature importance utilizing RFE with random forest revealed that t_w had the highest importance. t_w has the most feature impact across all parameters, according to SHAP with random forest ensemble learning feature interaction. PCA has reduced 13 features to 10 orthogonal features for further analysis.

Data preparation amplifies the model dependability in shear strength prediction, and this study highlights how crucial careful feature selection and data preparation is useful. Researchers may apply and compare various AI techniques to assess how well they can predict the shear strength of a concrete shear wall.

For the structural engineering applications, creating more precise predictive frameworks will be aided by this data investigation, which will better understanding of behavior of shear wall for the researchers. Ablation analysis will be performed to systematically evaluate each input parameter's contribution to shear strength prediction and model resilience. The prepared dataset will be utilized in further work to develop models.

References

- [1] Weng Y. Kam, Stefano Pampanin, and Ken Elwood, "Seismic Performance of Reinforced Concrete Buildings in the 22 February Christchurch (Lyttelton) Earthquake," *Bulletin of the New Zealand Society for Earthquake Engineering*, vol. 44, no. 4, pp. 239-278, 2011. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [2] Ahed Habib et al., "Seismic Behavior of Squat Reinforced Concrete Shear Walls: A State-of-the-Art Review," *SDHM Structural Durability and Health Monitoring*, vol. 19, no. 3, pp. 417-439, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [3] Yu-Cheng Hsu et al., "Shear Strength Prediction of Reinforced Concrete Walls with Opening and Boundary Elements," *Journal of Structural Engineering*, vol. 149, no. 7, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [4] Khuong Le Nguyen et al., "Machine Learning Approaches for Lateral Strength Estimation in Squat Shear Walls: A Comparative Study and Practical Implications," *Expert Systems with Applications*, vol. 239, pp. 1-19, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [5] Sujith Mangalathu et al., "Data-Driven Machine-Learning-based Seismic Failure mode Identification of Reinforced Concrete Shear Walls," *Engineering Structures*, vol. 208, 2020. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [6] Siming Tian, Xiangyong Ni, and Yang Wang, "Seismic Shear Strength Prediction of Reinforced Concrete Shear Walls by Stacking Multiple Machine Learning Models," *Applied Sciences*, vol. 15, no. 5, pp. 1-23, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [7] Ahmed Gondia et al., "Mechanics-Guided Genetic Programming Expression for Shear-Strength Prediction of Squat Reinforced Concrete Walls with Boundary Elements," *Journal of Structural Engineering*, vol. 146, no. 11, pp. 2020. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [8] Hosein Naderpour et al., "Shear Strength Prediction of Reinforced Concrete Shear Wall using ANN, GMDH-NN and GEP," *Journal of Soft Computing in Civil Engineering*, vol. 6, no. 1, pp. 66-87, 2022. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]

- [9] De-Cheng Feng et al., “Interpretable XGBoost-SHAP Machine-Learning Model for Shear Strength Prediction of Squat RC Walls,” *Journal of Structural Engineering*, vol. 147, no. 11, 2021. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [10] Behrooz Keshtegar et al., “Novel Hybrid Machine Learning Model for Predicting Shear Strength of Reinforced Concrete Shear Walls,” *Engineering with Computers*, vol. 38, no. 5, pp. 3915-3926, 2022. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [11] Leonardo M. Massone, and Francisco Melo, “General Solution for Shear Strength Estimate of RC Elements based on Panel Response,” *Engineering Structures*, vol. 132, pp. 239-252, 2018. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [12] Gyu Il Kim, and Kyungyong Chung, “Extraction of Features for Time Series Classification Using Noise Injection,” *Sensors*, vol. 24, no. 19, pp. 1-15, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [13] Sumeyra Demir et al., “Data Augmentation for Time Series Regression: Applying Transformations, Autoencoders and Adversarial Networks to Electricity Price Forecasting,” *Applied Energy*, vol. 304, pp. 1-19, 2021. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [14] Hongyi Zhang et al., “mixup: Beyond Empirical Risk Minimization,” *Proceedings of the International Conference on Learning*, pp. 1-13, 2018. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [15] Lin Sun et al., “Feature Selection using Fisher Score and Multilabel Neighborhood Rough Sets for Multilabel Classification,” *Information Sciences*, vol. 578, pp. 887-912, 2021. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [16] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., Springer, pp. 1-745, 2009. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [17] X. L. Chen et al., “Prediction of Shear Strength for Squat RC Walls using a Hybrid ANN–PSO Model,” *Engineering with Computers*, vol. 34, no. 2, pp. 367-383, 2018. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [18] Scott M Lundberg, and Su-In Lee, “A Unified Approach to Interpreting Model Predictions,” *Advances in Neural Information Processing System*, vol. 30, pp. 1-10, 2017. [[Google Scholar](#)] [[Publisher Link](#)]
- [19] *ACI CODE-318-25: Building Code for Structural Concrete—Code Requirements and Commentary (ACI CODE-318-25)*, American Concrete Institute, pp. 1-702, 2025. [[Google Scholar](#)] [[Publisher Link](#)]
- [20] *Eurocode 8: Design of Structures for Earthquake Resistance – Part 1: General Rules, Seismic Actions and Rules for Buildings*, European Committee for Standardization, pp. 1-10, 2005. [[Google Scholar](#)] [[Publisher Link](#)]
- [21] De-Cheng Feng et al., “Machine Learning-Based Compressive Strength Prediction for Concrete: An Adaptive Boosting Approach,” *Construction and Building Materials*, vol. 230, pp. 1-20, 2021. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [22] IIT Kanpur (National Information Centre of Earthquake Engineering), Earthquake Tip 23: Why are buildings with shear walls preferred in seismic regions? Indian Institute of Technology Kanpur, Kanpur, India. Available from NICEE Earthquake Tips series, 2005. [[Publisher Link](#)]
- [23] Ian T. Jolliffe, and Jorge Cadima, “Principal Component Analysis: A Review and Recent Developments,” *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, vol. 374, no. 2065, pp. 1-16, 2016. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [24] Frank J. Vecchio, and Michael P. Collins, “The Modified Compression-Field Theory for Reinforced Concrete Elements Subjected to Shear,” *ACI Journal Proceedings*, vol. 83, no. 2, pp. 219-231, 1986. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [25] I-Tung Yang, and Handy Prayogo, “Interpretable Data-Augmented Machine Learning Framework for Robust Failure Mode Classification of RC Shear Wall,” *Journal of Building Engineering*, vol. 112, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [26] Guoxu Zhao et al., “Automated Design of Shear Wall Structures using Reinforcement Learning with Code Compliance,” *Engineering Structures*, vol. 353, 2026. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [27] Kevin Karanja Kuria, and Orsolya Kegyes-Brassai, “Digital Twin-based Machine Learning Framework for Predicting Nonlinear Seismic Response of Reinforced Concrete Shear Walls using Analytical Data,” *Scientific Reports*, vol. 16, pp. 1-18, 2026. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [28] Mohsen Zaker Esteghamati et al., “Beyond Development: Challenges in Deploying Machine-Learning Models for Structural Engineering Applications,” *Journal of Structural Engineering*, vol. 151, no. 6, pp. 1-28, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [29] Mehmet Tan, Gulsen Taskinkaya, and Zeynep Tuna Deger, “Machine Learning-Based Framework for Failure Forecast and Shear Strength Estimation of Non-Conforming RC Shear Walls,” *Journal of Earthquake Engineering*, vol. 28, no. 12, pp. 3387-3409, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [30] SeyedeZahra Golazad et al., “From Raw to Refined: Data Preprocessing for Construction Machine Learning (ML), Deep Learning (DL), and Reinforcement Learning (RL) Models,” *Automation in Construction*, vol. 168, pp. 1-53, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [31] Chao-Lie Ning, and Bing Li, “Probabilistic Development of Shear Strength Model for Reinforced Concrete squat Walls,” *Earthquake Engineering Structural Dynamic*, vol. 46, no. 6, pp. 877-897, 2017. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]