Diyala Journal of Engineering Sciences

Journal homepage: https://djes.info/index.php/djes

ISSN: 1999-8716 (Print); 2616-6909 (Online)

Study the Efficiency of the XGBoost Algorithm for Squat RC Wall Shear Strength Prediction and Parametric Analysis

Badie. H Sulaiman[1*], Amer M. Ibrahim[1] and Hadeel J. Imran[2]

[1]Department of Civil Engineering, University of Diyala, 32001 Diyala, Iraq
[2]Laser and Optoelectronics Engineering Department, University of Technology, Iraq

ARTICLE INFO

Article history:
Received May 7, 2023
Revised October 8, 2023
Accepted October 29, 2023
Available online March 7, 2024

Keywords:
Machine Learning (ML)
Squat RC walls
Shear strength
XGBoost Model
Empirical Model

ABSTRACT

Squat-reinforced concrete (RC) shear walls with an aspect ratio of less than two are considered effective structural members, where shear is the dominant failure mechanism. Squat shear walls are widely used in nuclear power plants and building construction and feature optimal cost and outstanding performance, due to their lateral strength and high rigidity to resist lateral loads. However, since the accurate evaluation of the shear strength of squat shear walls must meet the design specifications, its calculation may be very complex, challenging, and inaccurate using experimental and theoretical equations due to many influential and overlapping design factors, so it takes more time and higher cost to determine it. This study uses machine learning (ML) methods to build a shear strength prediction efficient model for squat RC walls to address these issues. First, a huge dataset of 1424 RC squat wall test specimens gathered from the literature is utilized for developing an ML model, by employing XGBoost, to predict the shear strength. Results verified that the XGBoost model had the best accuracy and least error while assessing the squat walls' strength at shear. Moreover, an XGBoost optimum algorithm fared better than the empirical models based on mechanics, with a 99.2% accuracy. Finally, to prove that the model can identify the most important variables that significantly affect the shear strength, parameter and sensitivity analyses were performed and the results showed that the wall length is the factor that contributes most to the ultimate shear strength of the squat shear wall as a percentage (7.62%), followed by the yield strength. For the web as a ratio. (6.88%), concrete strength (6.75%), reinforcement ratio information (6.56%), and geometric properties (6.01%), while the axial load represents the smallest contribution, reaching (4.16%).

1. Introduction

Reinforced concrete (RC) walls having an aspect ratio of less than or equal to two are referred to as short or squat walls [1]. These walls, which have strong structural framing and are often employed in conventional buildings and nuclear power plants under safety requirements in areas with high seismic activity and strong winds, significantly boost the lateral load resistance of structures. Squat RC walls are typically available in three various cross-sectional shapes: barbell, flanged, and rectangular. Accurate predictions of squat shear wall capabilities are crucial for earthquake analysis and design procedures. These predictions are challenging to develop because of several elements, including concrete, reinforcing, and flexural shear interactions. Recently, multiple initiatives have been undertaken by researchers to increase experimental equations for shear strength methods in squat walls including the strut and tie approach [2-4], and the softened truss model

* Corresponding author.
E-mail address: badie.hussein@outlook.com
DOI: 10.24237/djes.2024.17110

[5,6], Gulec & Whittaker [7], Adorno Bonilla [8], and Jiaxing Ma [9]. These different models feature simplifying nonlinear behavior and complexity methods for calculating the shear strength of RC walls. Gulick and Whittaker formulas, which are limited to squat-flanged walls with aspect ratios of one or fewer. regarding Adorno-Bonilla and Jiaxing Ma's studies which employed a limited sample of data of 137 and 119 test results, respectively, and the aspect ratios were frequently less than 1.20. As a result of these deficiencies above, their estimations of the shear strength of squat RC walls may be inaccurate biased, and have a broad dispersion in results. Additionally, these research predictions were inaccurate since they only focused on a limited number of crucial essential parameters in their equations for determining shear strength [10-11]. The shear strength of RC walls can be determined empirically, but there are various questions regarding the properties and combinations of the materials, therefore it is problematic to offer an exact empirical model. To obtain the most important parameters that determine and influence the design equations, it is frequently necessary to conduct several experiments, however, these studies are expensive and time-consuming. New investigation avenues have arisen as an alternate remedy in structural and seismic engineering as a result of recent advancements in machine learning (ML) approaches that prompted an increase in study efforts along this line of interest. ML research has lately gained popularity due to the following reasons:

1. Engineering and Construction studies commonly employ closed-form forecasting equations, which exhibit bias and dispersion due to simple modeling and inaccuracies
2. Quick access to specifically preserved prior results from experiments with high-resolution digital simulations
3. ML models have considerably developed, on par alongside human cognitive skills.

Recent studies in the construction engineering sector and its development have widely used these machine learning methods [12]. The shear strength of squat rectangular RC walls was predicted using the artificial neural network and particle swarm optimization algorithm (ANN–PSO), a hybrid model developed by Chen et al. [13] based on a set of 139 test data, they concluded that, when it came to forecasting the strength of shear walls, the recommended model (ANN–PSO) was more accurate than existing models. However, since their study was only limited to the flanged RC wall type, their experiments did not provide a predictive model. Moradi and Hariri Ardebili[14] developed an ANN model to create a library of shear wall datasets to assess the shear strength of typical RC walls. They covered squat and taller walls with rectangular and flanged cross-section shapes within the database. The evaluation and verification dispersal remained rather large despite the fact their results demonstrated the accuracy of the ANN model and no functioning formula or graphical user interface (GUI) tool was made available for the design process. Mangalathu and Jeon have suggested using an artificial neural network (ANN) in their methodology [15] that assessed using a dataset comprised of samples for reinforcement columns with circular sections. Using this method, the shear, flexure, or flexure-shear failure modes of the columns might be anticipated. To forecast the failure mechanism mode of shear walls, Mangalathu and other scholars carried out a similar investigation [16]. Due to their inability to identify the degree of correlation between the predicted failure modes and the input values, the researchers were unable to prove high accuracy. Feng et al. [17] studied performance seismic evaluation designs and employed an ensemble ML technique named adaptive boosting (AdaBoost) to estimate the hinged in a nonlinear analysis of time records the length of the plastic columns is the main variable. The forecasting algorithms model created by Zhang et al. [18] forecasting algorithms was created using a database of 429 RC wall test data and a variety of ML techniques. The outcomes demonstrated that the lateral strength and ultimate drift ratio of RC walls were better predicted by the gradient boosting (GB) and random forest algorithms and that the XGBoost and GB

algorithms performed extremely well in forecasting the failure modes of RC walls. In a study reported by Hemn Ahmed et al. [19], the compressive strength (CS) of geopolymer concrete (GPC) reinforced with nanomaterials was predicted utilizing ML modeling techniques including MEP, FQ, and ANN. Other machine learning techniques were also used to predict the CS of GPC. With one variable used as the output and eleven significant variables employed as the input model parameters, they have been applied to 207 tested CS values. Also, a sensitivity study was conducted to determine whether the input factors affected the CS present in the GPC. According to the restricted numbers of data and inputs, even though the ANN model appeared to be more accurate than other models in calculating the CS of the GPC, additional details about prediction and the influence of design factors needed to be obtained. Previous studies mentioned above showed how machine learning methods might be successful in a variety of circumstances while conquering challenges such as a lack of experimental data and the inability to adapt the model to new circumstances. Considering the knowledge above, the most important objective of this study is to develop an ML model to forecast the shear strength of squat RC walls, allowing the model's predictions to be correlated with the input dataset. To determine the shear strength of squat RC walls, 1424 experimental tests were methodically assembled from previous studies. The data are then typically split using an 80%, and 20% split at random into training and testing sets.

The primary objective of this study is to create an effective Extreme Gradient Boosting (XGBoost) based data model to forecast the shear strength of squat RC walls. The suggested model's outcomes have been compared with those of existing studies and the current design code. It proposed to develop a prediction formula based on the XGBoost model to determine the shear strength of flanged walls while taking into account twenty-five input factors. Moreover, to identify and analyze what parameter is most probable to have an impact on shear strength, sensitivity analysis is carried out, and a variety of conclusions are drawn. Finally, for evaluating the efficiency of models, the statistical metrics criteria are utilized, in addition, the dataset's correlation matrix is also obtained.

2. Empirical Database for Squat RC Walls

The experimental database should be comprehensive for all design parameters and available to create an optimal shear model for RC squat walls. Due to this, data collected from 1424 specimens of RC squat wall tests have been employed in this study [1-3, 20-22]. Twenty-five essential input factors must be taken into account to forecast the shear strength of the walls. The modern database possesses a wide range of squat shear wall features. This in turn improves the prediction accuracy of the trained ML model. Figure 1 shows an exemplary diagram for the squat RC shear wall test for the database. Which has three distinct cross-sectional groupings: Walls might be rectangular, barbell-shaped, or flanged. The four types of input parameters geometric dimensions, reinforcing configurations, material characteristics, and applied loads are depicted in this figure. The specified inputs are variables, are, the concrete compressive strength ($f'c$), vertical reinforcement ratio (ρvbe) and strength ($fyv\ be$), horizontal reinforcement ratio (ρhbe), and strength ($fyhbe$), vertical web reinforcement ratio (ρv), and strength (fyv), horizontal web reinforcement ratio (ρh) and strength (fyh), the ratio of all vertical reinforcement ($\rho vall$), ultimate strengths of the vertical (fuv), and horizontal fuh web reinforcement, the spacing of the vertical and horizontal web reinforcement (Sv and Sh), longitudinal, and horizontal boundary diameter reinforcement ($Dlbe$) and ($\boldsymbol{Dhbe}$), vertical and horizontal web diameter reinforcement (Dwv) and (Dwh), height (hw), length (lw), web thickness (tw), flange height (bf), flange thickness (tf), and, finally, the applied axial load (P). Simply expressed, the output is the actual shear strength (Vn). The details and statistics features of the input variables are shown in Table 1. It uses statistical functions like minimum, maximum, average, standard deviation, and coefficient of variation to show the statistical distribution of each

variable. It is important to keep in mind that the abbreviations for these two concepts are standard deviation (SD) and coefficient of variation (COV). After preprocessing conducted on data cleaning removed duplicates, outliers, and handling with missing data, 1424 test data were selected from 3159 total that were used to construct the histogram's distributions of twenty-five input parameters which are displayed in Figure 2.

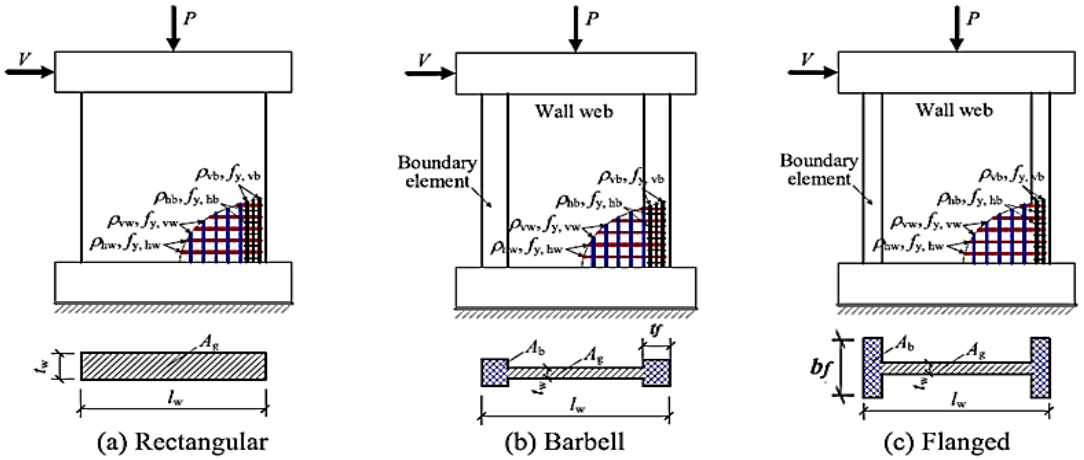
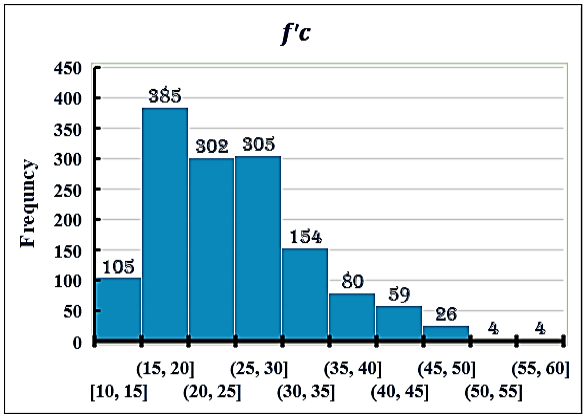


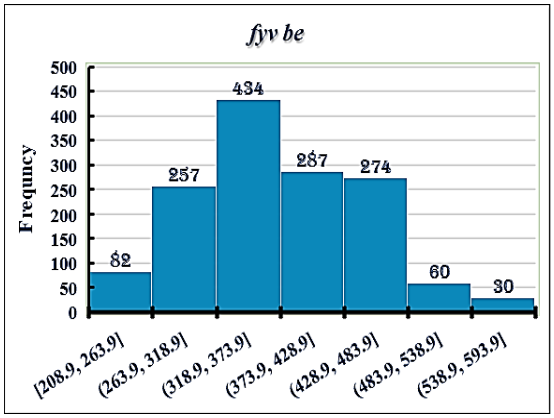
Figure 1. Schematic Diagram of Squat RC Wall Tests [18]

Table 1: Statistics of the empirical input variables

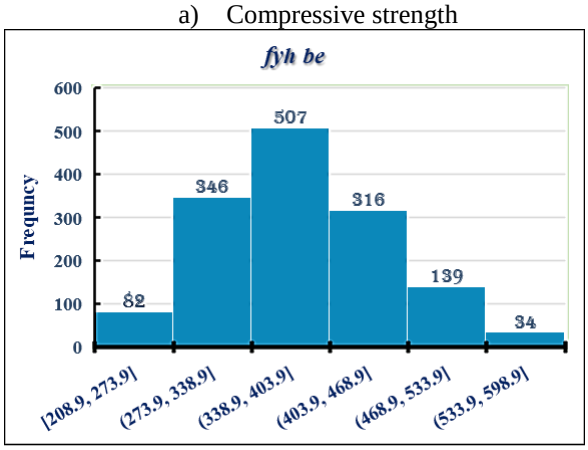
Variable	Unit	Minimum	Maximum	Mean	SD	COV	Type
$f'c$	MPa	10	56	25.07	8.38	0.33	Input 1
fyv be	MPa	208.90	585	375.05	73.83	0.20	Input 2
fyh be	MPa	160.87	529.60	364.22	54.64	0.15	Input 3
fyv	MPa	224	667.00	385.45	87.97	0.23	Input 4
fyh	MPa	222.10	667	386.61	89.10	0.23	Input 5
fuh	MPa	484.61	726.26	634.79	38.00	0.06	Input 6
fuv	MPa	509.09	699.51	635.90	33.77	0.05	Input 7
ρvbe	%	0	8.90	3.09	1.93	0.62	Input 8
ρhbe	%	0	0	0	0	0	Input 9
ρv	%	0	1.63	0.52	0.34	0.66	Input 10
ρh	%	0	1.63	0.54	0.36	0.66	Input 11
$\rho vall$	%	0.30	0.30	0.30	0	0	Input 12
Sv	mm	229	229	229	0	0	Input 13
Sh	mm	203	203	203	0	0	Input 14
Dl be	mm	9.5	9.5	9.5	0	0	Input 15
Dh be	mm	4.95	4.95	4.95	0	0	Input 16
Dwv	mm	6.35	6.35	6.35	0	0	Input 17
Dwh	mm	6.35	6.35	6.35	0	0	Input 18
lw	mm	254	3329.50	1223.55	611.65	0.50	Input 19
hw	mm	150	2760	918.43	535.08	0.58	Input 20
tw	mm	20	203	107.38	29.48	0.27	Input 21
tf	mm	30	260	120.97	58.05	0.48	Input 22
bf	mm	30	610	144.53	98.31	0.68	Input 23
tweb	mm	16	160	69.31	36.93	0.53	Input 24
P	kN	0	830	125.62	197.04	1.57	Input 25
V*n*	kN	0	2668	354.85	373.73	1.05	Output



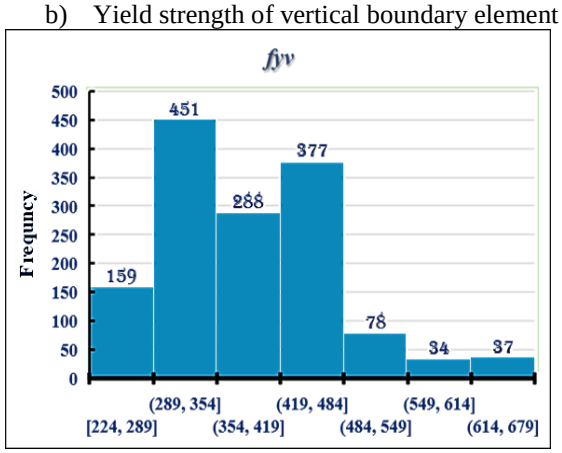
a) Compressive strength



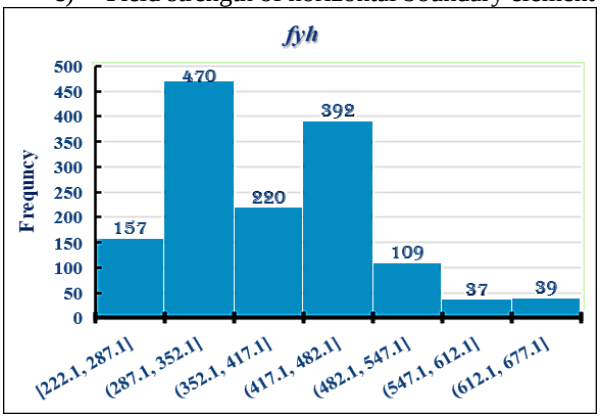
b) Yield strength of vertical boundary element



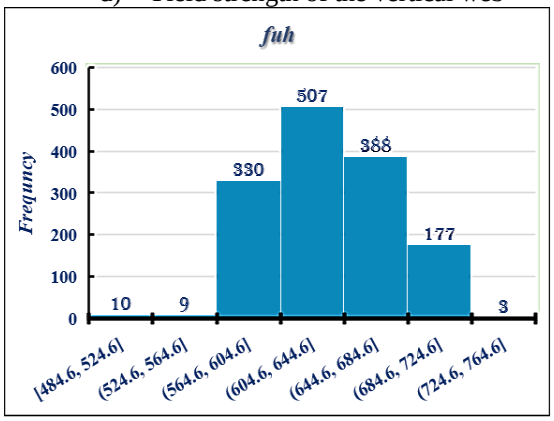
c) Yield strength of horizontal boundary element



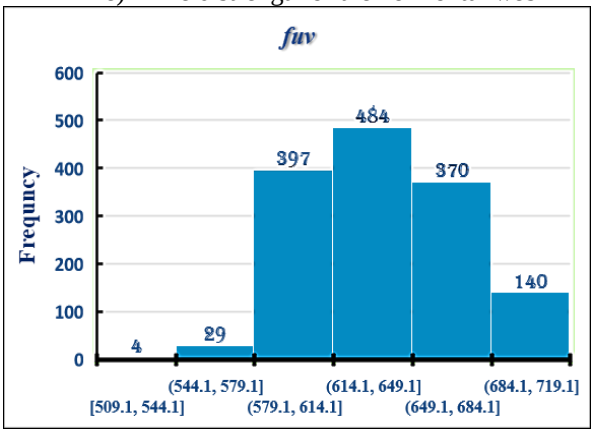
d) Yield strength of the vertical web



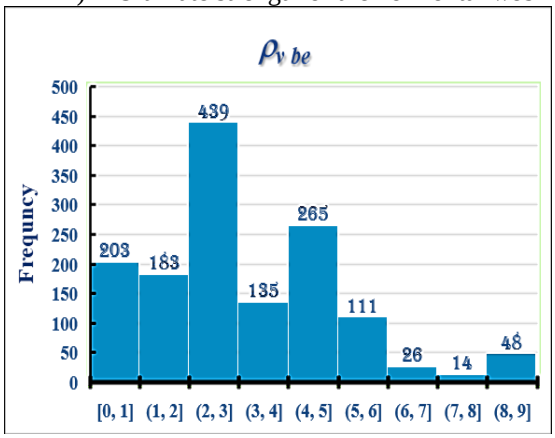
e) Yield strength of the horizontal web



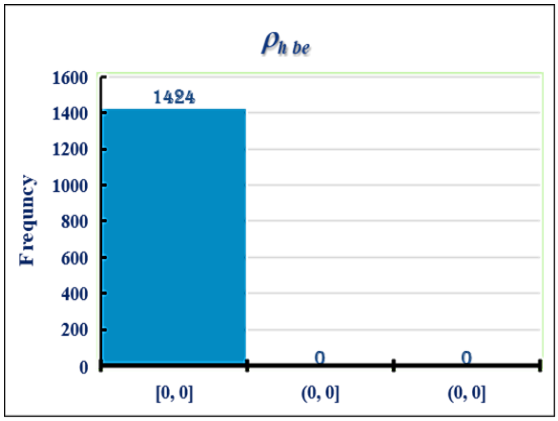
f) Ultimate strength of the horizontal web



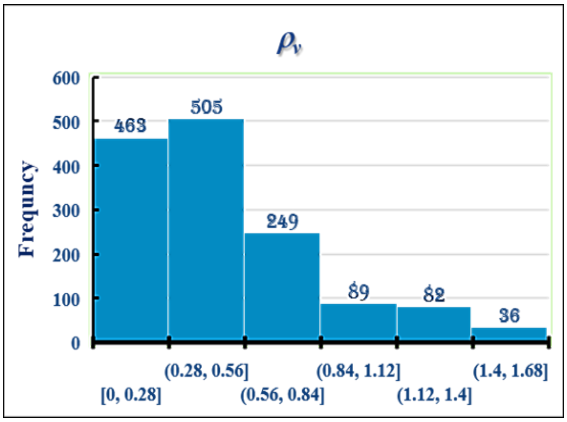
g) Ultimate strength of the vertical web



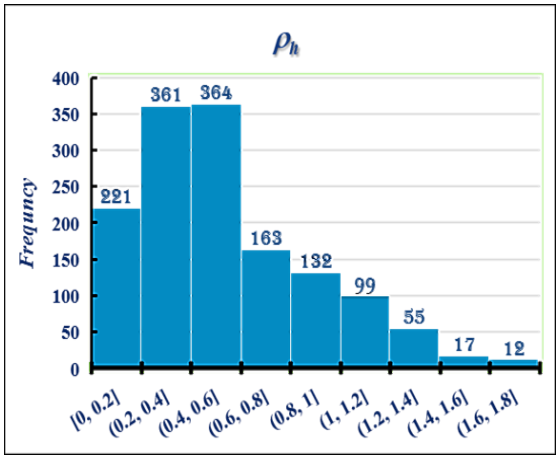
h) Reinforcement ratio of vertical boundary element



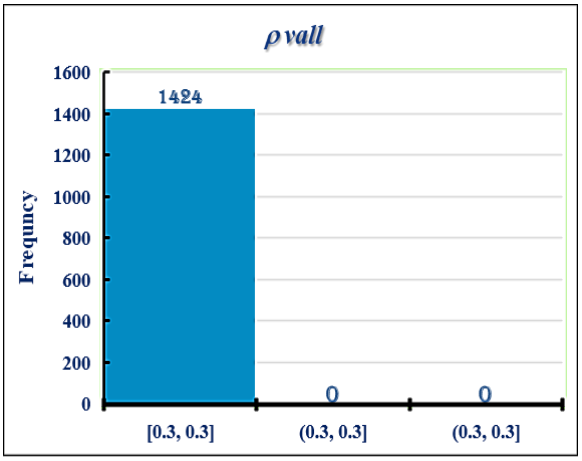
i) Reinforcement ratio of vertical boundary element



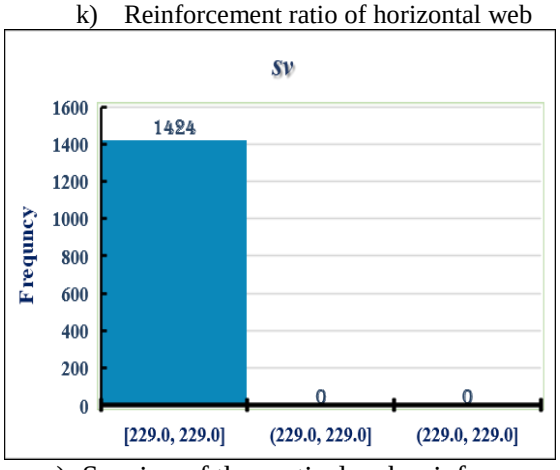
j) Reinforcement ratio of vertical web



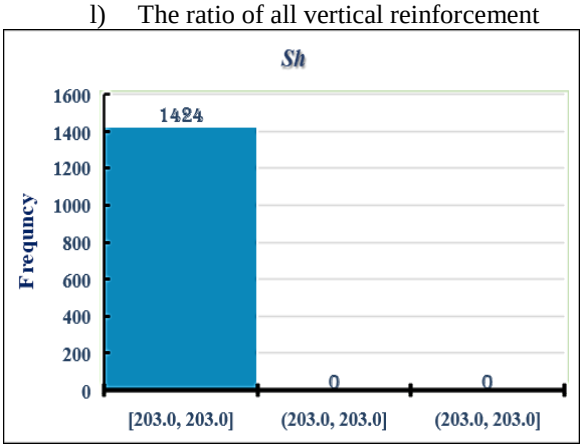
k) Reinforcement ratio of horizontal web



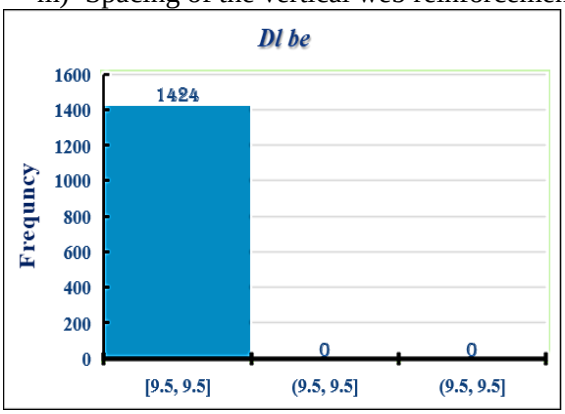
l) The ratio of all vertical reinforcement



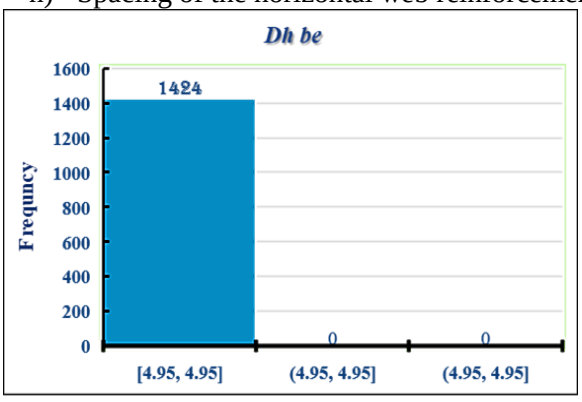
m) Spacing of the vertical web reinforcement



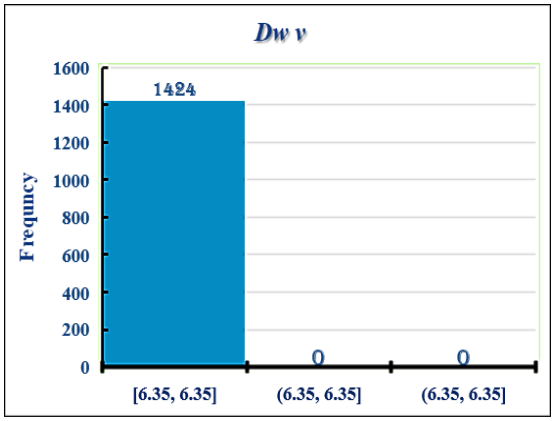
n) Spacing of the horizontal web reinforcement



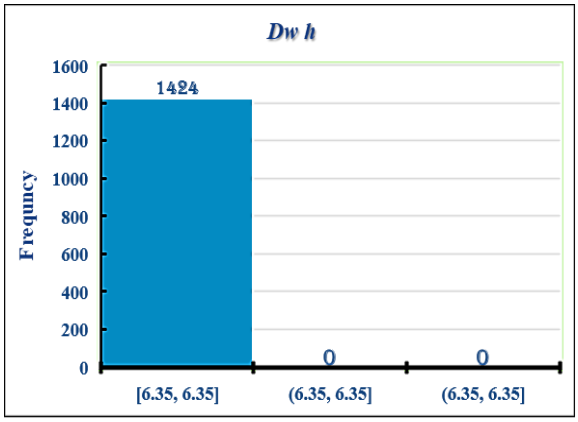
o) Longitudinal boundary diameter reinforcement



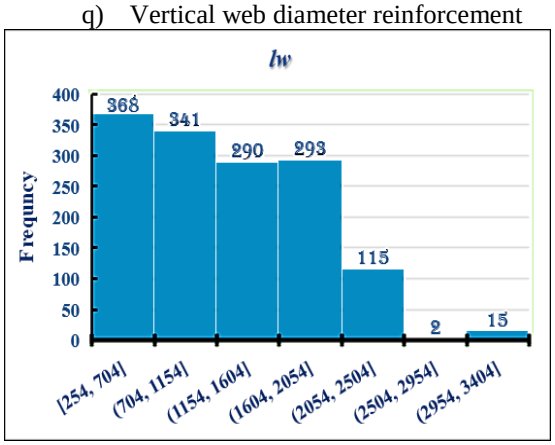
p) Horizontal boundary diameter reinforcement



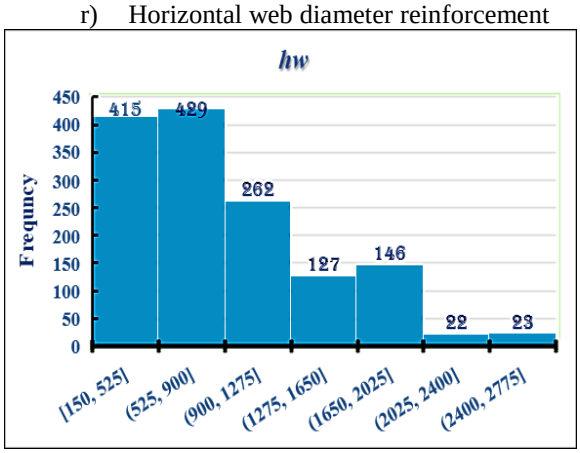
q) Vertical web diameter reinforcement



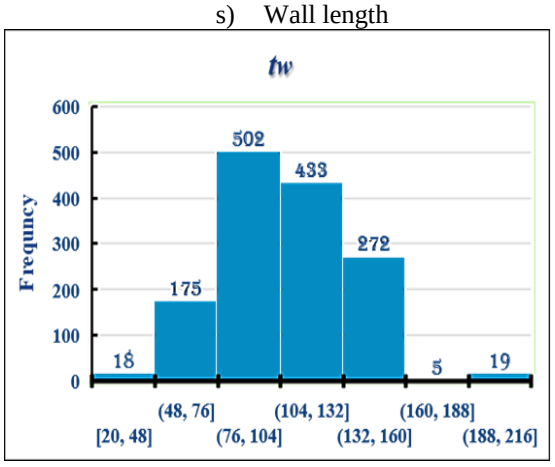
r) Horizontal web diameter reinforcement



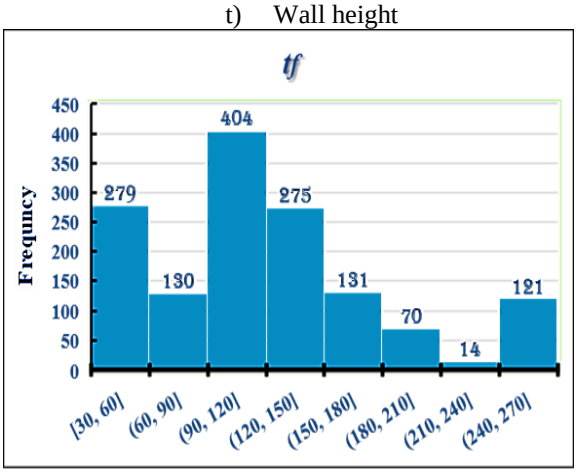
s) Wall length



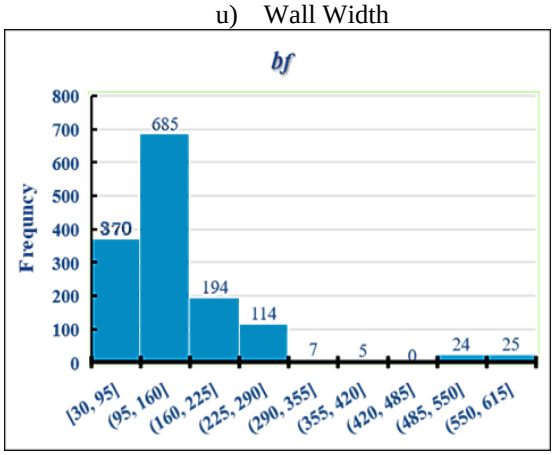
t) Wall height



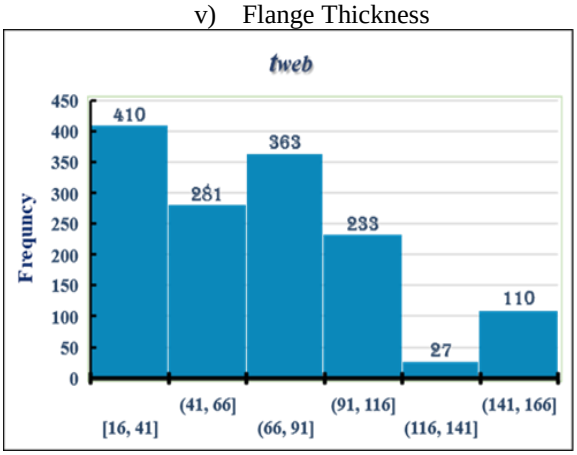
u) Wall Width



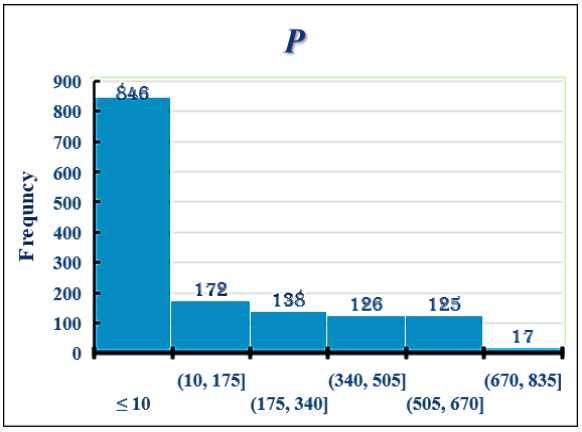
v) Flange Thickness



w) Flange Width



x) Web Thickness



y) Axial load

Figure 2. Histograms of input parameters based on 1424 experimental data

The linear correlation among two variables is commonly determined using the Pearson correlation coefficient, whose value ranges from -1 to +1. While 0 indicates no linear correlation and 1 indicates a perfect linear positive correlation, -1 signifies a perfect linear negative correlation. A coefficient with a value between ±0.50 and ±1 is seen as indicating a significant association. A heatmap of the correlation coefficient between the variables in pairings is shown in Figure.3, It shows that although certain parameters have strong relationships, others have poor correlations. For instance, the correlation coefficient between (*tweb*) and (P) was 0.588, indicating a significant and positive association between the two variables. The correlation between (*fuv*) and (*fuh*) was 0.941, whereas (*fyh be*) and (*ρhbe*) correlated 0.562. Regarding the shear strength (*Vn*), it was discovered that its correlation with (*lw*) alone is substantial; it was 0.704; nevertheless, its correlation with the other variables, including (*hw*), (*fyh be*), (*sv*), (*bf*), and (*f'c*), is poor; they were, respectively, 0.251, -0.112, N/A, 0.415, and 0.254. Figure 3 displays a detailed of the kind of correlation that is present among the input variables, shear strength, and both of them. The statistical analysis of data, histograms, and relationships between variables is known as data exploration, often referred to as exploratory data analysis. It is the method of comprehending and evaluating data via statistical and visual techniques. This technique aids in identifying trends in a dataset. Finding patterns in data distributions, identifying the features of individual variables, and identifying correlations between variables are the three main objectives of data exploration. Histograms and charts are used to visually represent data as part of visualization techniques, making it easier to comprehend the data's numerous relations and structures. This is what took place earlier.

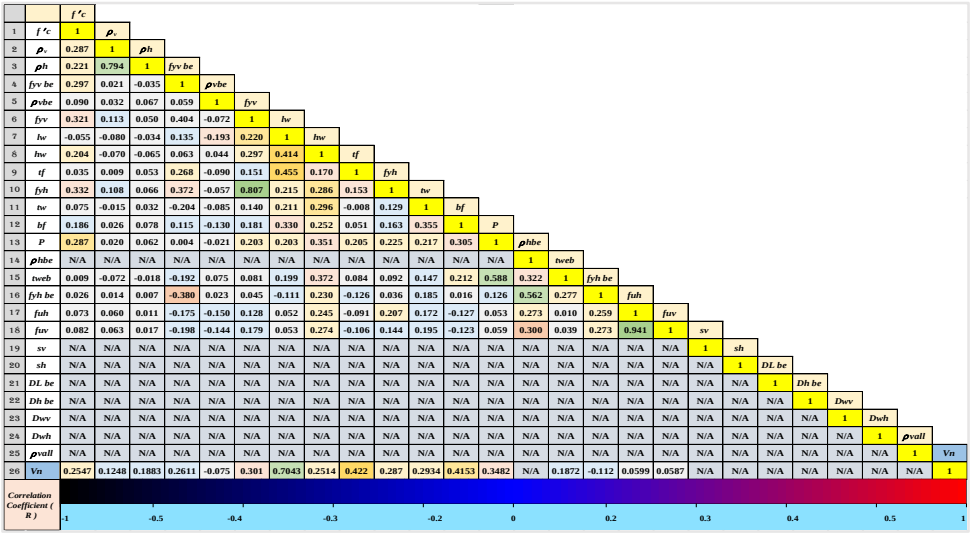


Figure 3. Correlation matrix for input variables and target (output)

3. Existing formulas for RC shear wall strength computation

The present study addresses three widely empirical employed equations for the computation of the shear strength of squat RC walls. These formulae are the results derived from numerous study searches [23], or as specified by design codes [20,24]. Table 2 presents a summary of the formulas used for determining squat RC walls' shear strength. Geometric dimensions, material properties, axially applied loads, and coefficients are four groups of features utilized in Table 2 to calculate the shear strength (*Vn*) of shear walls. The wall height (hw), web length (*lw*), web thickness (*tw*), flange width (*bf*), and flange thickness (*tf*) are within the geometrical variables. Other factors include the total area of the section defined by the length of the mesh, the direction of the shear force (*Acv*), and its thickness. We add to this the vertical reinforcement area in the boundary element, which represents the total shear friction reinforcement area of the walls(*Avf*). A concrete compressive strength (*f'c*), the horizontal and vertical reinforcement yield strength (*fyh* and *fyv*), the walls' reinforcing ratios (*ρh*) and (*ρv*) for both the horizontal and vertical orientations make up the material properties, and axial load (*P*). The two coefficients are utilized in these equations, one represents the percentage of concrete strength to nominal wall shear strength (***αc***), and the other reduces the mechanical properties of lightweight concrete compared to normal-weight concrete with the same compressive strength (**λ**).

4. The model for shear strength of RC walls utilizing XGBoost

The recently created machine learning regression method XGBoost is one of several algorithms in this category that was used to build the model in this instance. A type of ensemble learning technology called XGBoost has been used in multiple studies to forecast and clarify the mechanical behavior of concrete structures [2]. An improved version of XGBoost has been created using the gradient-boosted decision tree (GBDT) ensemble learning technique, which enhances the properties of the loss function and loss optimization process in comparison to the (GBDT). This approach reduces overfitting and controls the complexity of the tree by including a regularization component in the objective function. To prevent overfitting, a strategy called column sampling is employed.

Table 2: Squat RC wall shear strength computation formulas

No.	Models	Formula
1.	Wood[23]	$0.5\sqrt{fc'}Acv \le Vn = \frac{Avf\ fyv}{4} + \rho se\ fyh \le\ 0.83\sqrt{fc'}Acv$
2.	ACI 318-19 Provision[20]	$Vn = (\alpha c\lambda\ \sqrt{fc'} + \rho hfyh)\ Acv$
3.	ASCE 43–05[24]	$Vn = 0.69\sqrt{fc'} - 0.28\sqrt{fc'}\left(\frac{hw}{lw} - 0.5\right) + \frac{P}{4lwtw} + \rho se\ fyh \le\ 1.66\sqrt{fc'}$ where $\rho se = A\rho v + B\rho h$ if $hw/lw < 0.5$, A = 1 and B = 0 if $0.5 < hw/lw < 1.5$, A = $-hw/lw$ +1.5 and B = hw/lw -0.5 if $hw/lw \geqslant 1.5$, A = 0 and B = 1 $Vn = vn*d*tw$ Where $d = 0.6\ lw$

A research effort from the University of Washington, XGBoost was presented in 2016 at the International Conference on Knowledge Discovery and Data Mining (SIGKDD) Conference [25]. Listed below is a brief description of the essential mathematical concepts that XGBoost uses.

where the dataset D {(xi, yi)}, includes input parameters (xi) and output variables (yi). with n samples, m features, and a model that is additive, composed of k fundamental models as the predictive variable, given potential formulations of the following equations:

$$\hat{y}i = \phi(xi) = \sum_{K=1}^{K} \alpha k \, fk(xi) \qquad (1)$$

where ŷi is the prediction value; $\boldsymbol{\phi}$ ($\boldsymbol{xi}$) is the final strong learner; $\boldsymbol{fk}$ ($\boldsymbol{xi}$) is an insufficient learner (the decision tree (DT) technique yields a poor learner); K is the number of weak learners; and αk is the learning rate (Utilizing the learning rate prevented overfitting).

The following form is based on the XGBoost objective function, which comprises a regularization term that represents the model's complexity and a conventional component termed the loss function that is meant to represent the differences between the prediction ŷi and the actual value yi:

$$Obj. = \sum_{i} L\,(yi, \hat{y}i\,) + \sum_{K} \Omega f(k) \qquad (2)$$

The first right-side term, L($\boldsymbol{yi, \hat{y}i}$), denotes an achievable training loss between actual and predicted values. The term regularization refers to the complexity of the model and is the second term on the right side, the second right-side term, Ω($\boldsymbol{k}$), denotes the model's complexity, also known as the regularization term. These two criteria evaluate both the model's complexity and data fit. For the first term, a squared loss function is frequently used. The second term is represented by the tree The number of nodes and the leaf score's L2 standard,

$$L\,(yi, \hat{y}i\,) = (yi - \hat{y}i\,)^2$$

$$\Omega\, f(k) = \gamma T + \frac{1}{2}\lambda \parallel wk \parallel \qquad (3)$$

The fundamental formula then becomes:

$$Obj. = \sum_{i=1}^{m} (yi - \hat{y}i\,)^2 + \sum_{k=1}^{t} \gamma T + \frac{1}{2}\lambda \parallel wk \parallel \qquad (4)$$

whereas T is the number of leaf nodes; wk = weights (or leaf scores); γ and λ are the penalty coefficients. The challenge then becomes identifying the proper learner ft at each step t ≤ K to reduce the loss function. The optimal strategy is determined by solving the second-order Taylor equation, which provides a more precise and straightforward definition of the objective function [2, 26].

5. Validation criteria

In the present study, the Coefficient of Determination (R^2), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and scatter index (SI) were employed as metrics to assess the effectiveness of the prediction models that were given. These concepts are expressed by the following equations.

$$\mathbf{R^2} = \left(\frac{\boldsymbol{n}\sum_{i=1}^{n}\boldsymbol{yi\hat{y}i} - \sum_{1=1}^{n}\boldsymbol{yi}\sum_{i=1}^{n}\boldsymbol{\hat{y}i}}{\sqrt{(\boldsymbol{n}\sum_{i=1}^{n}\boldsymbol{yi^2} - (\sum_{i=1}^{n}\boldsymbol{yi})^2)(\boldsymbol{n}\sum_{i=1}^{n}\boldsymbol{\hat{y}i^2} - (\sum_{i=1}^{n}\boldsymbol{\hat{y}i})^2)}}\right) \qquad (5)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(yi - \hat{y}i)^2} \qquad (6)$$

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{(yi - \hat{y}i}{yi}\right| \qquad (7)$$

$$SI = \frac{RMSE}{\frac{1}{n}\sum_{i=1}^{n} yi} \qquad (8)$$

where n is the number of samples; yi is the actual value of the i^{th} dataset; ŷi is the predicted value of the i^{th} dataset. The variance between the predicted and actual results was measured using the R^2 value. The RMSE number, in the meantime, reflects the average of errors. Additionally, SI measures how dispersed the error is with the dataset's mean, and MAPE is a percentage residual error between the actual and anticipated values. Higher R^2, lower RMSE, and lower MAPE values, overall, demonstrate that the model is more accurate and highly efficient. According to the SI parameter, a model performs poorly when SI > 0.3, fairly well when 0.2 < SI < 0.3, good

performance when 0.1< SI < 0.2, and excellent performance when SI < 0.1[27].

6. Results and discussions

6.1 Model implementation

The basic four phases of the proposed ML model execution are depicted in Figure 4. The initial stage of the data collection involves splitting it into two groups: the training group (80%) and the test group (20%). To prevent the scaling effect, every input has been normalized to fall into the range [0, 1]. During the training phase, a tenfold cross-validation (CV) technique is employed to reduce the bias resulting from the random selection of the training set. The grid-based search method is utilized to identify the optimal hyperparameters. Finally, the model's efficacy on the testing dataset (20%) is assessed using the four measuring tools previously stated. [28]. The KNIME Analytics platform, version 4.7.7, a software program acknowledged as one of the most recent data science and artificial intelligence programs that supports the Python and R languages (computer languages), has been employed in the current study.

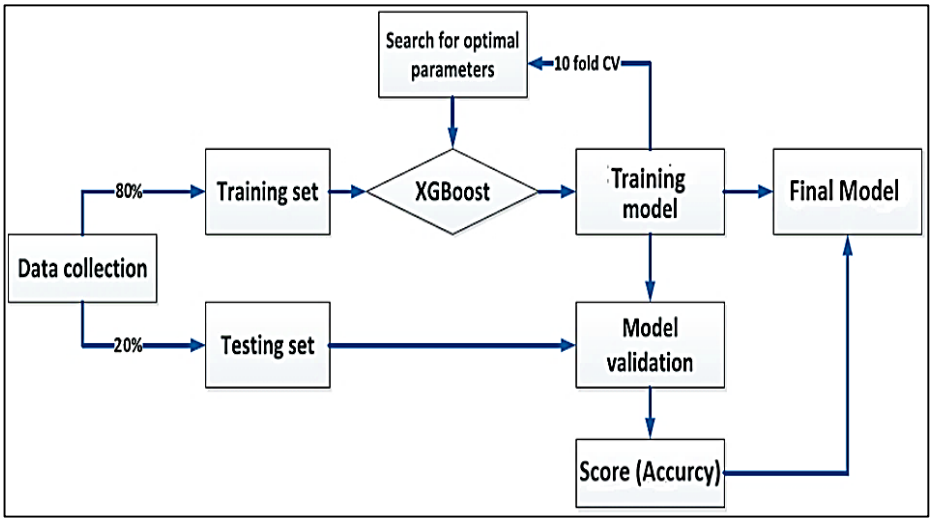


Figure 4. The XGBoost model process flowchart [2]

6.2 The XGBoost model's prediction results

The XGBoost model's predictions for the shear strength of squat shear walls will be explored in the phases that follow. First, multiple tries were made to tune the model's hyperparameters to maximize the shear strength value's prediction accuracy and decrease the error rate by achieving the maximum R^2 and lowest RMSE values. Table 3 illustrates a list of the hyperparameters that were set and the optimum values that were obtained.

Table 3: XGBoost Model Tuning Hyperparameters

Parameter	Lower Bound	Upper Bound	Best Value	Description
Boosting Rounds	1	100	70	The number of trees n estimator
Gamma	0	100	0	Minimum loss to split
Alpha ($\boldsymbol{\gamma}$)	0	10	1	L1 regularization term on weights (Penalty Coefficient)
Lamda ($\boldsymbol{\lambda}$)	0	10	1	L2 regularization term on weights (Penalty Coefficient)
Maximum Depth	0	100	8	Maximum depth of a tree
Minimum Child Weight	0	100	1	The minimum sum of instance weight needed in a child
Maximum Delta Step	0	100	0	Maximum delta step for each leaf output
Subsampling Rate	0.001	1	1	The subsample of rows in training datasets
Column Sampling Rate by Tree	0.5	1	1	The subsample of columns in training datasets

The tenfold CV of mean RMSE with model hyperparameters varies during the training stage, as demonstrated in Figure 5. Since they have a significant effect on the results, only the three most crucial variables (the learning rate, the highest depth of the trees, and the n estimator which is the number of trees) are displayed. The model performs effectively when the parameters (n-estimator = 70, max. depth = 8, and learning rate = 0.3) are applied, as shown in Figure 5 (C). Concerning this, the tenfold CV of mean RMSE is only 32.12 kN. Take note of (70, 32.12), where 70 represents n-estimator and 32.12 represents RMSE.

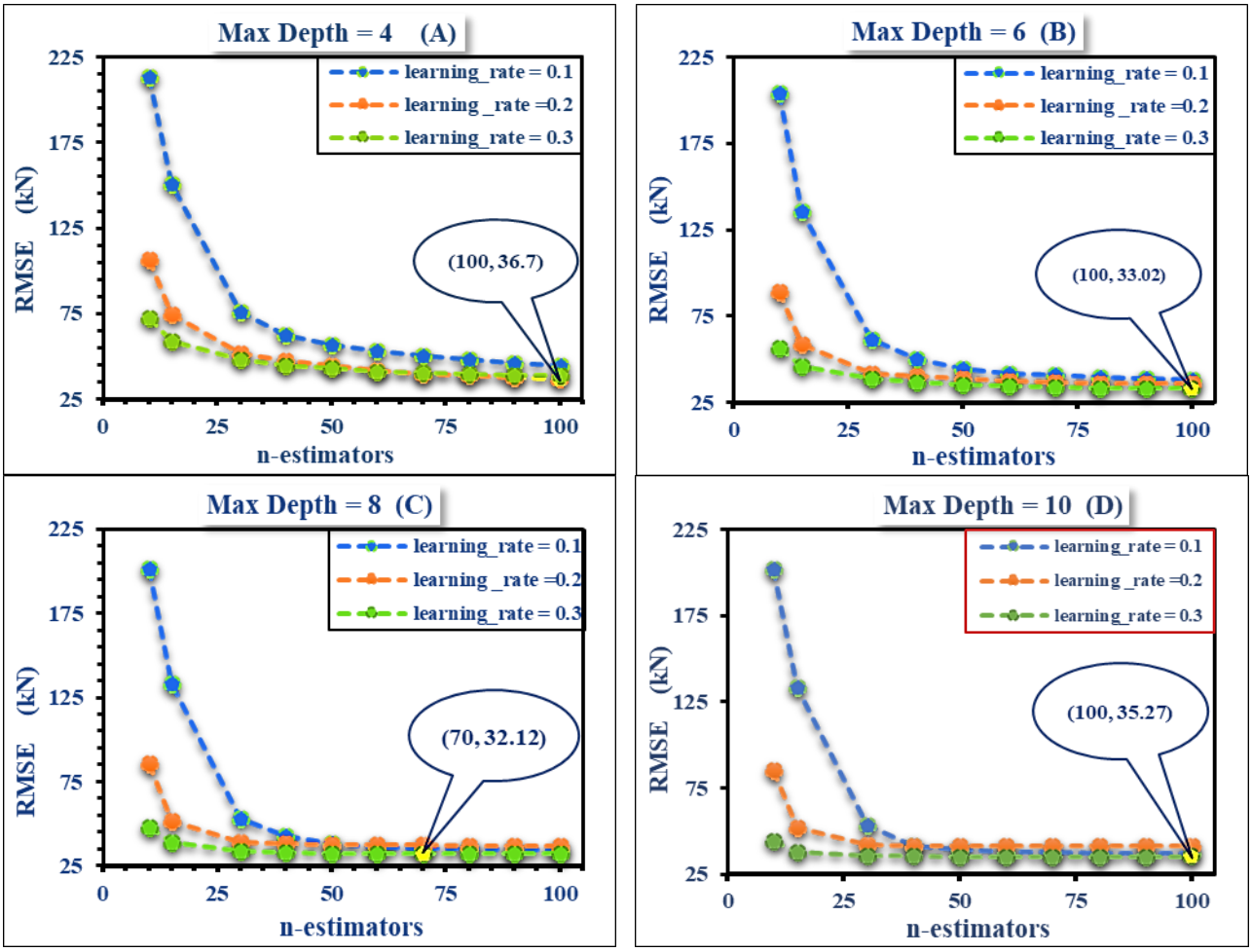


Figure 5. Model development alongside Hyperparameter tuning employing 10-fold CV and grid search

The XGBoost model has been assessed using the testing set after the hyperparameters have been identified. Based on the greatest value of $R^2$and the lowest RMSE, the model offers a high degree of precision in forecasting the shear strength, which is more accurately portrayed by the other three statistical metrics of the model: R^2 = 0.992, MAPE = 6%, and SI = 0.089. Figure 6 exhibits the outcomes of the measurements of the shear strengths in comparison to those that the XGBoost model predicted for the test sets. The dashed line displays the expected values (ideal line y=x), while the bold line shows the scatters' linear regression. The outcome is predicted more precisely the more closely the scattering nearly the ideal line y=x. It has been established that the XGBoost model proposed in this study has a lot less dispersion. Additionally, the data's linear regression line had the lowest MAPE value of 6% and was nearly identical to the ideal line y=x.

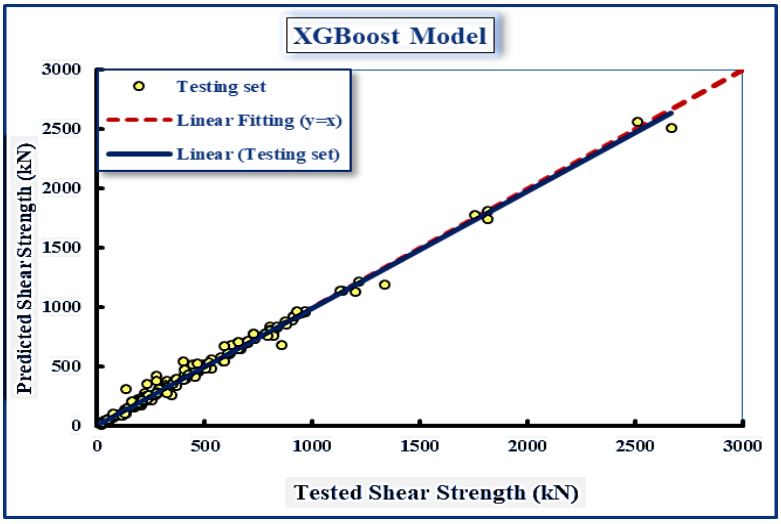


Figure 6. A comparison of tested values and XGBoost model predicted values

The aforementioned results demonstrate the XGBoost model's superior learning and prediction abilities. Figure 7 displays a schematic illustration of every data processing step, including pre-processing, normalization, and segmentation. The algorithm is then given the dataset to create the results.

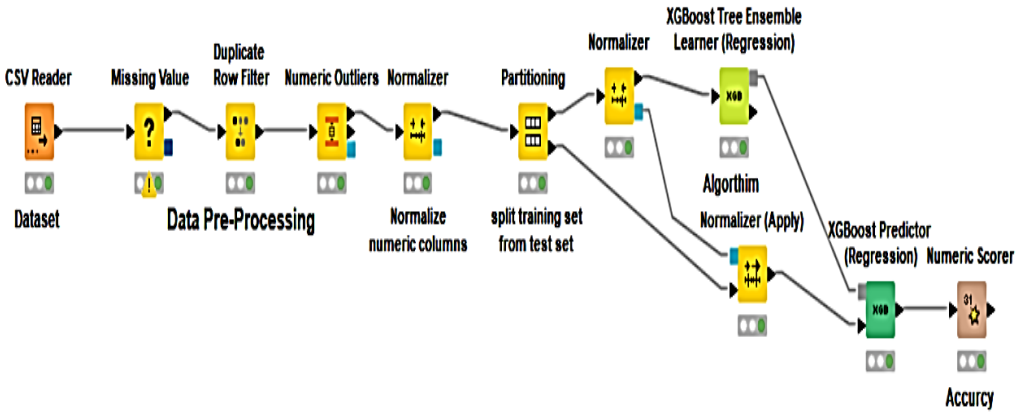


Figure 7. Flowchart Diagram of All Operations Carried out on the data of XGBoost Model [KNIME Program]

6.3 Comparison with empirical mechanic's shear strength models

To emphasize the XGBoost-based prediction model's improved performance, three prominent empirical mechanics models namely, ACI 318-19 (ACI 2019), (ASCE 2005), and Wood (1990) are also employed in comparisons. The dataset was also pre-processed by managing missing values, duplicate values, and outliers before being sent to the model. Processes like partitioning the dataset into training (80%), testing (20%), and normalizing the scale [0, 1] were all carried out afterward these steps. ultimately, receiving the outcome of the forecast. Figure 8 shows the results of the proposed models, while Table 4 provides the precise values of the metrics reflecting the models' performance. Based on the testing data used to assess, it has been verified that The best performance was obtained when using the XGBoost model, while the weakest performance was obtained when applying the ASCE/SEI model.

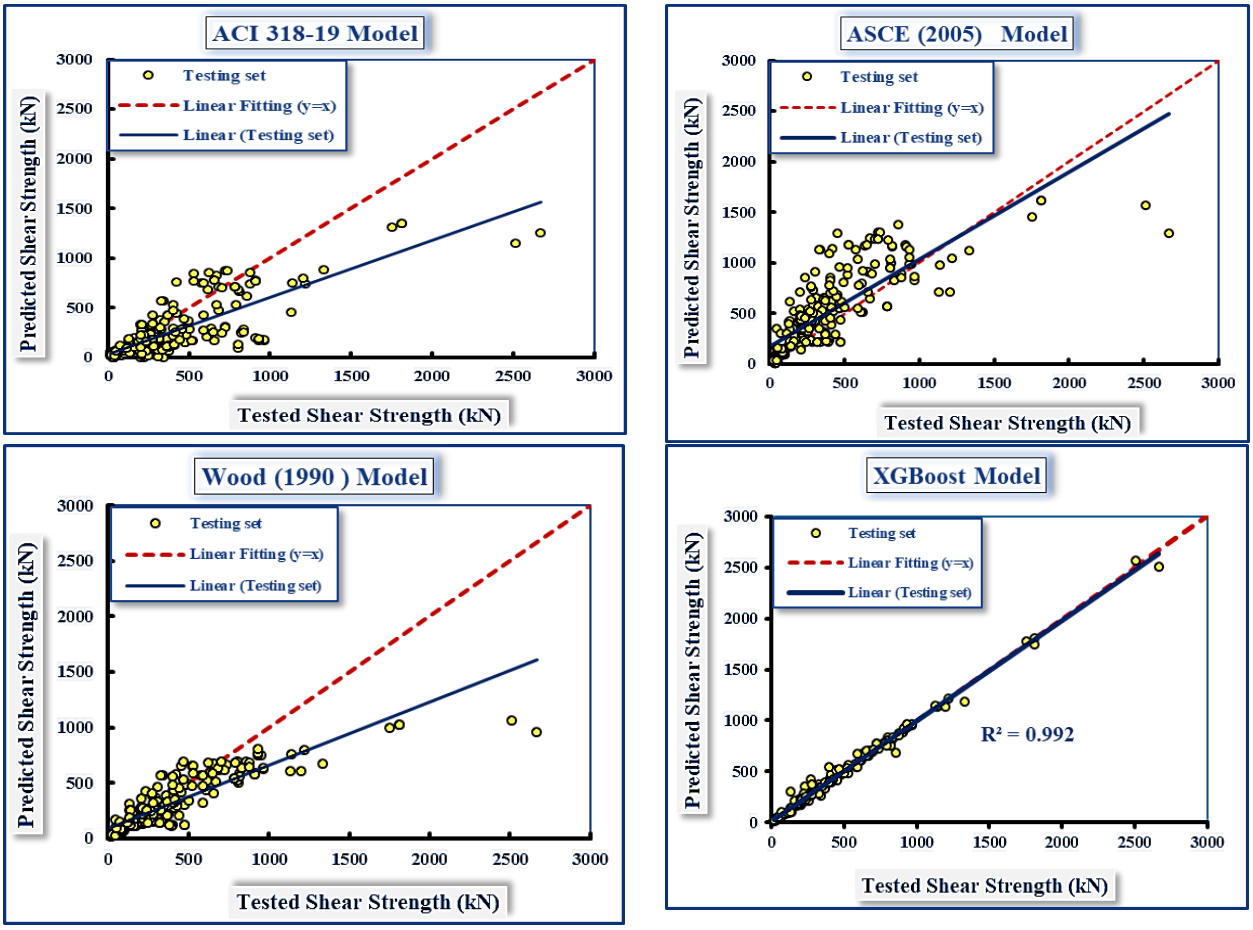


Figure 8. Outcomes on Predictions of Shear Strength of Proposed Models: ACI 318-19, ASCE (2005), Wood (1990), and XGBoost.

Table 4. Comparative findings between the several proposed models

Models	Sets	Measures			
		R^2	RMSE (kN)	MAPE (%)	SI
XGBoost	Testing	0.992	32.12	6	0.089
Wood (1990)	Testing	0.699	202.13	31	0.576
ACI 318-19	Testing	0.539	250.07	44	0.713
ACSE/SEI 43-05	Testing	0.484	264.44	71	0.754

Figure 9 presents the SI evaluation parameter values for the tested versions of the proposed models. The figure clarifies clearly that the SI values for XGBoost, Wood (1990), ACI 318-19, and ASCE (2005) are, respectively, 0.089, 0.576, 0.713, 0754, and 0.365. The XGBoost model showed excellent performance on the test dataset when statistically evaluated by a value of SI=0.089 < 0.1, While the three empirical models performed poorly with an SI value greater than 0.3. Moreover, the results of the scatter interval for residual errors of all developed models are displayed as well in Figure 10.

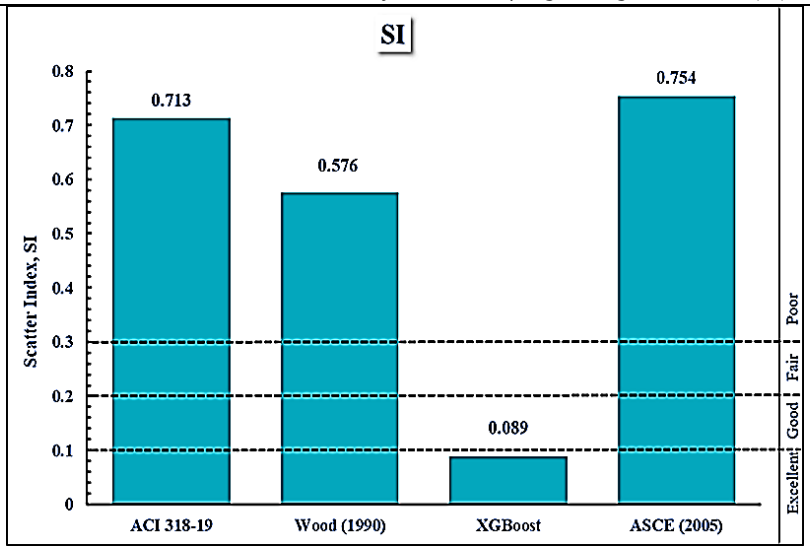


Figure 9. Comparing the SI performance parameters of different developed models

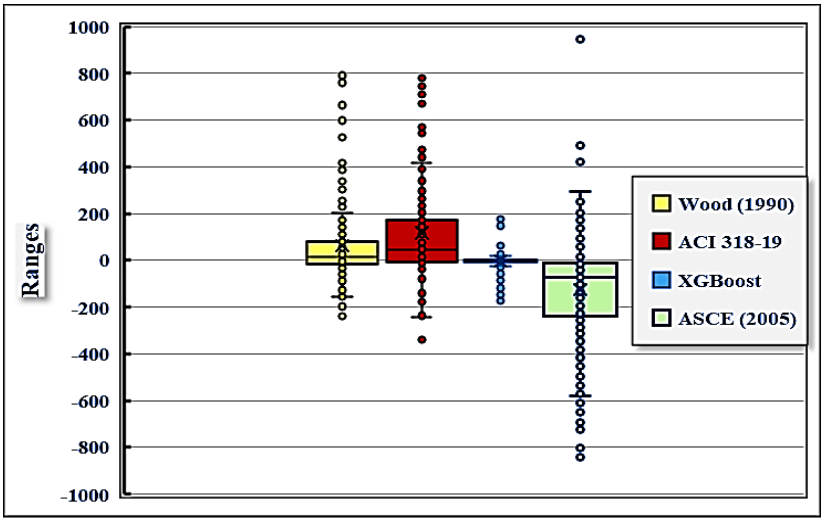


Figure 10. Scatter interval for residual errors of the developed models

In addition, ACI 318-19, ASCE (2005), and Wood (1990), whose MPEs of (31%), (44%), and (71%) respectively, are all outperformed by XGBoost, which obtains the lowest error ratio (MAPE) of 6%. Figure 11 shows a histogram that compares the MAPEs of the developed models. In terms of anticipated accuracy, the XGBoost model performs more effectively than the other three models.

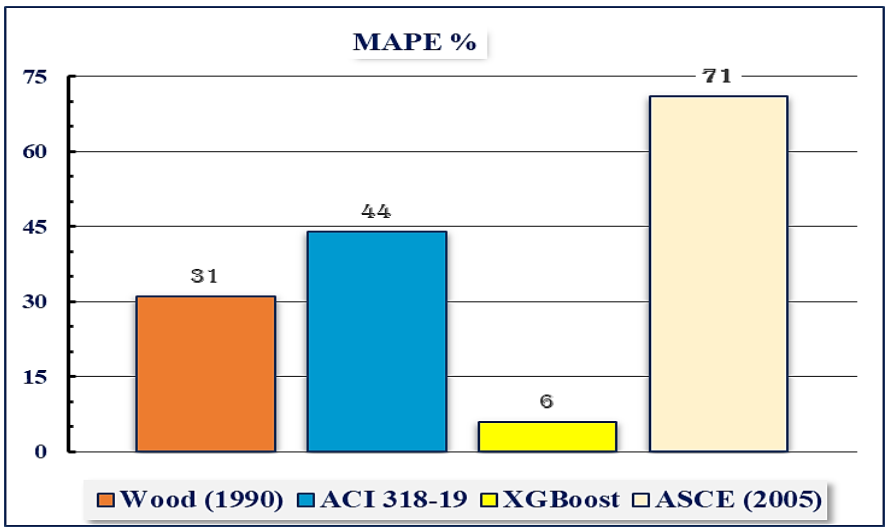


Figure 11. MAPEs of the developed proposed models

Figures. (9, 10, and 11) reveal that the predicted and actual shear strength values for the XGBoost model are closer to one another, evidence of the XGBoost model's superior accuracy and efficiency over the other three empirical models.

The best ML model was selected based on high accuracy and the least scattered in likewise to understand the degree of test data scattering around the average line of predict to test ratio. As a result of the wall aspect ratio (***hw/lw***) ranging from 0.0 to 2.0, Figure 12 shows the anticipated measured ratios of shear strength for the samples in the database along with the model predicted ratio [mean ± standard deviation (St. D.)], and the outcomes of the prediction intervals, which were used to assess these models.

Compared to walls with the highest aspect ratio (between 1.5 and 2), walls with an aspect ratio of less than 1.5 have more data points dispersed around the mean value line. This incident is a result of conclusions made regarding several of the shear failure patterns of low aspect ratio walls. [25]. The coefficient of variation (COV) values for various predicted-to-test ratios can be seen in Table 5 along with minimum, maximum, mean, and standard deviation.

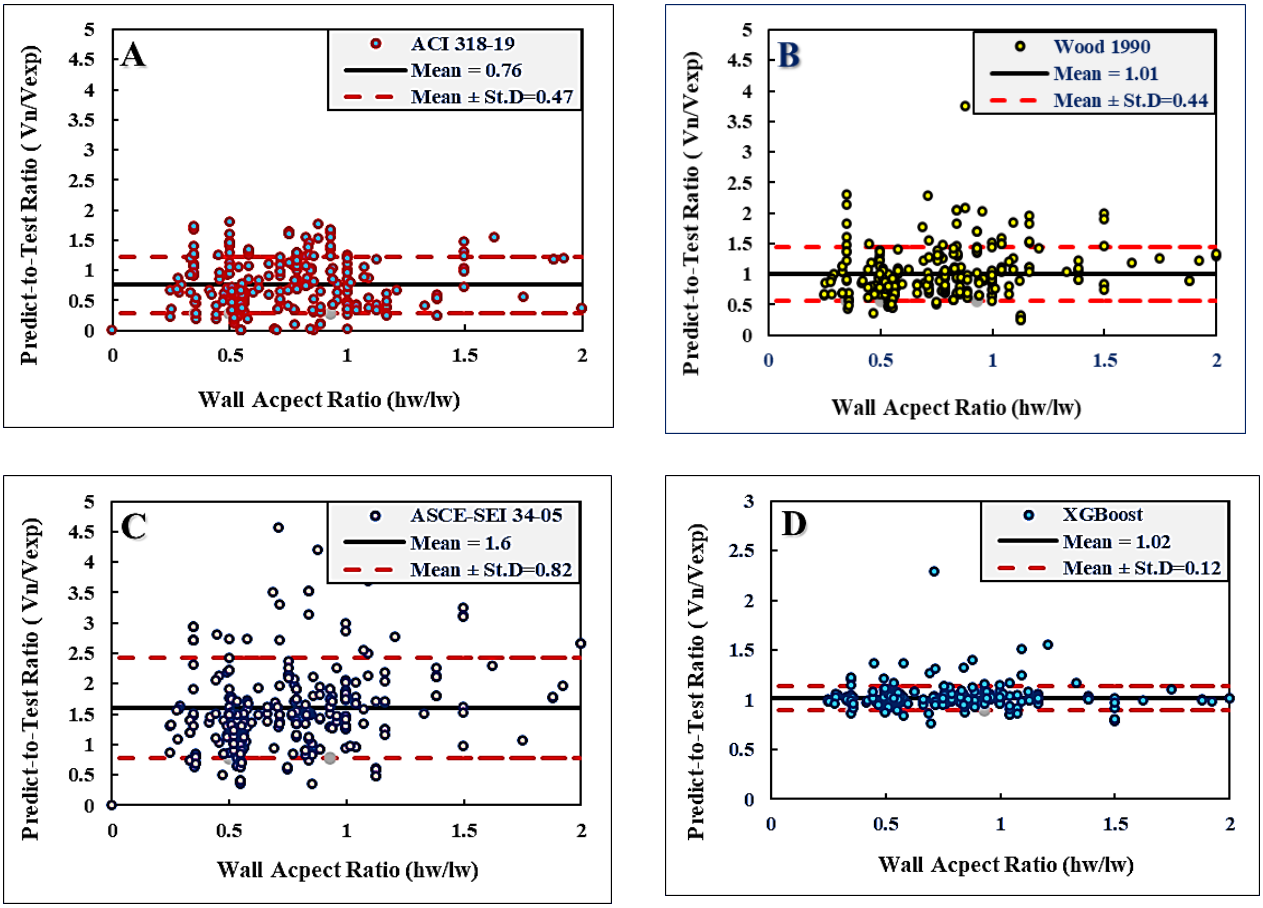


Figure 12. Results predicted by mechanics-based model of shear strength: (A) ACI 318-19; (B) Wood (1990); (C) ASCE/SEI 43-05; (D) XGBoost.

Table 5: Prediction performance of empirical models versus xgboost model

Models	Predict-to-Test Ratio Results				
	Minimum	Maximum	Mean	St.D.	COV %
ACI 318-19	0.01	3.39	0.76	0.47	62.26
ASCE/SEI 43-25	0.34	7.58	1.60	0.82	51.11
Wood (1990)	0.25	3.75	1.01	0.44	43.62
XGBoost	0.66	2.29	1.02	0.12	11.76

The three empirical models' outcomes were highly dispersed, compared to the XGBoost model's results, which had a higher average forecast and less variation in the prediction. The figure and table above show this. Whereas the three mechanistic models' proven COVs are exceptionally high, Wood's (1990) mean estimates of the ratios may be as large as 1.01, which is very near to 1.0. This is true even though the models ACI 318-19 and ASCE/SEI 43-25 achieved distant mean ratios from 1, of 0.76 and 1.6 respectively, which was accomplished with XGBoost equal 1.02, which is extremely close to 1.0 with the lowest COV. Still, the XGBoost model gives the most impressive results in terms of high accuracy and minimal error. Since it included into account all 25 variables related to shear wall design features, which covered all of its failure modes, the XGBoost machine learning model performed better than the other three equations. The XGBoost algorithm has also undergone extensive learning of the wall dataset training. Semi-empirical models, on the other hand, are either numerical formulae constrained to data or mathematical equations with a limited set of variables. As a result, the empirical equations employed to predict wall shear strength in both published research and existing design standards have a major disagreement, which causes their estimated findings to be inaccurate, biased, and scattered [20].

7. Sensitivity analysis and parametric study

In this part, the most important input factors that affect the shear strength of the walls have been identified and investigated by performing a sensitivity analysis. Based upon XGBoost's most precise results predictions. How the model reacts to alteration in the input data reveals the efficiency it performs and, consequently, how well it can correctly reflect reality. Several alternative sets of training data were employed in the sensitivity analysis. When the model had been trained, just one variable from each set was retrieved, and the RMSE was independently calculated for each training dataset. The excluded parameter in the trial with the highest RMSE for the set had the most influence on forecasting shear strength [28]. Table 6 summarizes the outcomes of the sensitivity analysis for the most crucial variables.

Table 6: Analysis of sensitivity employing an XGBoost model

Sr.no	Removed Parameter	R^2	RMSE	Ranking
1	None	0.992	32.12	–
2	fc'	0.97	63.45	3
3	ρv	0.965	61.69	4
4	ρh	0.972	55.6	7
5	*fyv be*	0.985	48.2	11
6	ρv *be*	0.984	57.46	5
7	*fyv*	0.973	64.69	2
8	*lw*	0.967	71.64	1
9	*hw*	0.98	49.3	10

10	*tf*	0.988	50.51	9
11	*fyh*	0.988	45.03	12
12	*tw*	0.979	50.71	8
13	*bf*	0.98	56.52	6
14	*P*	0.989	39.1	14
15	*tweb*	0.984	43.43	13
16	ρhbe	0.992	32.12	N/A
17	$\rho vall$	0.992	32.12	N/A
18	S*v*	0.992	32.12	N/A
19	S*h*	0.992	32.12	N/A
20	*Dl be*	0.992	32.12	N/A
21	*Dh be*	0.992	32.12	N/A
22	*Dwv*	0.992	32.12	N/A
23	*Dwh*	0.992	32.12	N/A

The length of the wall (***lw***), which is present in the eighth row is clearly from the results, the most impactful and sensitive variable for the shear strength prediction of shear walls. This is followed by compressive strength (***f'c***), reinforcement ratio (ρv), and yield strength (***fyv***) of the vertical web. For more, the impact of the geometric features was less significant than the one described above. The remaining factors had a negligible impact, some of the variables such as **s*v***, **s*h***, ***Dl be***, ***Dh be***, ***Dwv***, ***Dwh***, $\rho vall$, *and* ρhbe did not influence the anticipated shear strength of the model, with R^2 and RMSE values of 0.992 and 32.12, respectively, where the same values were used regardless of whether these variables were present before or during the sensitivity analysis. This is because the bulk of the data for these factors that did not affect shear strength were outliers that were removed during the pre-processing of the data before feeding it to the model for training and testing activities. As explained in Table 7 below regarding the feature importance item for XGBoost that was obtained by the KNIME program, the ratios (weights of variables) that contribute to the shear strength of the model XGBoost for the option (importance of features - strength of their contribution), was unknown (?) and did not affect the results before and after it was removed to study their effect on shear strength.

This means that the pre-processing of data in general and outlier data in particular, as well as how to deal with them, have a substantial impact on the results and design parameters. Following sensitivity analysis, the percentage of model parameters contribution was computed, and the results are shown in Figure 13 together with the results of the sensitivity evaluation based on XGBoost models.

Table 7: XGBoost feature importance KNIME

No.	Feature Name	Weight
1	ρhbe	?
2	*sv*	?
3	*sh*	?
4	*Dl be*	?
5	*Dh be*	?
6	*Dwv*	?
7	*Dwh*	?
8	$\rho vall$	?

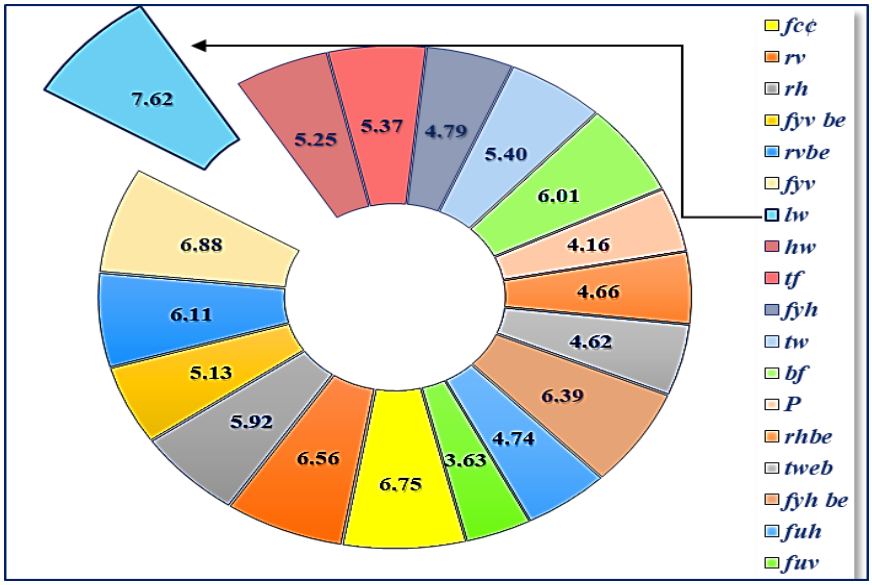


Figure 13. Sensitivity analysis using XGBoost-based model

Figure 13 demonstrates that the results ratio implies that in the above-mentioned data from the optimal model, the crucial influence on the prediction of shear strength is the length of the wall, which is followed by compressive strength, geometrical dimensions, and specifics of the reinforcing attributes while the axial load has the least. However, the height of the wall adversely the strength capacity of the walls. This result corresponds with earlier completed and published in the literature experimental investigations and research initiatives[1, 29-33].

8. Conclusions

In the current study data from 1424 tests, the XGBoost an ML algorithm was utilized to accurately predict the shear strength of squat RC walls. The most effective XGBoost hyperparameters were found using the grid search method, with random choices of 80% and 20% of the data being used for training and testing, respectively. The proposed model's prediction findings were compared with those of semiempirical models based on mechanics. The following conclusions may be taken from this study:

1. The KNIME analytics platform was crucial in handling large amounts of data and delivering quick performance to provide outcomes. Because of its critical role in precision computing operations, its simplicity of handling without difficulty or the need for scripts, its support for the Python and R languages, and its capacity to keep up with emerging techniques for processing and analyzing data, it may be utilized in the disciplines of ML and data science.
2. The shear strength of squat RC walls has been predicted most accurately and with the lowest error within XGBoost. The performance evaluation standards for the testing set were $\mathbf{R^2}$ = 0.992, RMSE =32.12 kN, MAPE = 6%, and SI = 0.089, respectively.
3. Predictions of three semi-empirical models were compared on a mechanical basis with those of XGBoost. It has been demonstrated that the XGBoost model outperformed these three models in terms of results, with a higher mean prediction and a noticeably lower standard deviation. The mean predicted ratio for the test was 1.02, while the COV was just 11.76%.
4. The three most crucial hyperparameters (learning rate, maximum tree depth, and number of trees n-estimator), have the greatest influence on the results and are connected with the XGBoost model's best accuracy and lowest error. The optimal settings for a model are n estimator = 70, max depth = 8, and rate of learning = 0.3,

with a fold of 10 CV average RMSE of just 32.12 kN.

5. Per the results of the sensitivity analysis, the length of the wall is the factor that contributes most to the peak shear strength of the squat shear wall as a ratio (7.62%), followed by the yield strength of the web as a ratio (6.88%), the strength of the concrete (6.75%), the reinforcement ratio information (6.56%), and geometrical properties (6.01%), while the height of the wall (5.25%) reversed effect, and the axial load makes the lowest of a contribution reached as a ratio (4.16%). This is in line with the earlier experimental results.
6. The results of the sensitivity analysis indicated that the input parameters; **s_v**, **s_h**, $D_{l\ be}$, D_{hbe}, D_{wv}, D_{wh}, ρ_{vall}, and ρ_{hbe} Their contribution rate (weights = 0%) did not affect the expected shear strength of the model before and after removing them to study their sensitivity, as the R^2 and RMSE values were 0.992 and 32.12, respectively, and they are the same before and after deleting the variables.
7. Using a machine learning approach to predict the shear strength of squat walls has proven superior to experimental and theoretical models in terms of fast, and accuracy, dealing with all basic variables for designing shear walls in buildings and constructions and using its results in current design work to save time and cost.

References

[1] Gulec, C. K. (2009). Performance-based assessment and design of squat reinforced concrete shear walls. State University of New York at Buffalo.

[2] D.-C. Feng, W.-J. Wang, S. Mangalathu, and E. Taciroglu, "Interpretable XGBoost-SHAP Machine-Learning Model for Shear Strength Prediction of Squat RC Walls," J. Struct. Eng., vol. 147, no. 11, pp. 1–13, 2021, doi: 10.1061/(asce)st.1943-541x.0003115.

[3] P. Chetchotisak, W. Chomchaipol, and J. Teerawong, "Strut-and-tie model for predicting the shear strength of squat shear walls under earthquake loads," Eng. Struct., vol. 256, no. October 2021, p. 114042, 2022, doi: 10.1016/j.engstruct.2022.114042.

[4] W. Kassem, "Shear strength of squat walls: A strut-and-tie model and closed-form design formula," Eng. Struct., vol. 84, pp. 430–438, Feb. 2015, doi: 10.1016/J.ENGSTRUCT.2014.11.027.

[5] L. M. Massone, "Strength prediction of squat structural walls via calibration of a shear–flexure interaction model," Eng. Struct., vol. 32, no. 4, pp. 922–932, Apr. 2010, doi: 10.1016/J.ENGSTRUCT.2009.12.018.

[6] H.-W. Yu and S.-J. Hwang, "Evaluation of Softened Truss Model for Strength Prediction of Reinforced Concrete Squat Walls," J. Eng. Mech., vol. 131, no. 8, pp. 839–846, 2005, doi: 10.1061/(ASCE)0733-9399(2005)131:8(839).

[7] C. K. Gulec and A. S. Whittaker, "Empirical equations for peak shear strength of low aspect ratio reinforced concrete walls," ACI Struct. J., vol. 108, no. 6, p. 777, 2011.

[8] C. M. Adorno-Bonilla, "Shear Strength and Displacement Capacity," 2016.

[9] J. Ma, C.-L. Ning, and B. Li, "Peak Shear Strength of Flanged Reinforced Concrete Squat Walls," J. Struct. Eng., vol. 146, no. 4, pp. 1–11, 2020, doi: 10.1061/(asce)st.1943-541x.0002575.

[10] C. K. Gulec, A. S. Whittaker, B. Stojadinovic, W. W. El-Dakhakhni, B. R. Banting, and S. C. Miller, "Shear Strength of Squat Rectangular Reinforced Concrete Walls," ACI Struct. J., vol. 106, no. 3, pp. 488–497, 2008, doi: 10.1061/(asce)st.1943-541x.0000713.

[11] W. W. El-Dakhakhni, B. R. Banting, and S. C. Miller, "Seismic Performance Parameter Quantification of Shear-Critical Reinforced Concrete Masonry Squat Walls," J. Struct. Eng., vol. 139, no. 6, pp. 957–973, 2013, doi: 10.1061/(asce)st.1943-541x.0000713.

[12] H. Sun, H. V. Burton, and H. Huang, "Machine learning applications for building structural design and performance assessment: State-of-the-art review," J. Build. Eng., vol. 33, p. 101816, Jan. 2021, doi: 10.1016/J.JOBE.2020.101816.

[13] X. L. C. J. P. Fu and J. L. Y. J. F. Gan, "Prediction of shear strength for squat RC walls using a hybrid ANN–PSO model," Eng. Comput., vol. 0, no. 0, p. 0, 2017, doi: 10.1007/s00366-017-0547-5.

[14] M. A. H.-A. Mohammad Javad Moradi, Developing a Library of Shear Walls Database and the Neural Network Based Predictive Meta-Model. 2019. doi: 10.3390/app9122562.

[15] S. Mangalathu and J.-S. Jeon, "Machine Learning–Based Failure Mode Recognition of Circular Reinforced Concrete Bridge Columns: Comparative Study," J. Struct. Eng., vol. 145, no. 10, pp. 1–12, 2019, doi: 10.1061/(asce)st.1943-541x.0002402.

[16] S. Mangalathu, H. Jang, S. Hwang, and J. Jeon, "Data-driven machine-learning-based seismic failure mode identification of reinforced concrete shear walls," Eng. Struct., vol. 208, no. January, p. 110331, 2020, doi: 10.1016/j.engstruct.2020.110331.

[17] D. Feng et al., "Data-Driven Approach to Predict the Plastic Hinge Length of Reinforced Concrete Columns and Its Application," vol. 147, no. 2012, pp. 1–17, 2021, doi: 10.1061/(ASCE)ST.1943-541X.0002852.

[18] H. Zhang, X. Cheng, Y. Li, and X. Du, "Prediction of failure modes, strength, and deformation capacity of RC shear walls through machine learning," J. Build. Eng., vol. 50, no. January, 2022, doi: 10.1016/j.jobe.2022.104145.

[19] H. U. Ahmed et al., "Innovative modeling techniques including MEP, ANN, and FQ to forecast the compressive strength of geopolymer concrete modified with nanoparticles," Neural Comput. Appl., vol. 3, 2023, doi: 10.1007/s00521-023-08378-3.

[20] J. Chou, C. Liu, H. Prayogo, R. R. Khasani, D. Gho, and G. G. Lalitan, "Predicting the nominal shear capacity of a reinforced concrete wall in a building by metaheuristics-optimized machine learning," J. Build. Eng., vol. 61, no. August, p. 105046, 2022, doi: 10.1016/j.jobe.2022.105046.

[21] George E. Brown Jr., "NEES Database: The Shear Wall Database ," DEEDS, 2007. https://datacenterhub.org/dataviewer/view/neesdatabases:db/shearwalldb/

[22] M. U. Santiago Pujol, ACI Subcommittee 445B, Cheng Song, Ying Wang, Aishwarya Puranam, "ACI 445B Shear Wall Database," DEEDS, 2019.

[23] S. L. Wood, "Shear Strength of Low-Rise Reinforced Concrete Walls," Struct. J., vol. 87, no. 1, pp. 99–107, Jan. 1990, doi: 10.14359/2951.

[24] Nuclear Standards Committee, Seismic Design Criteria for Structures, Systems, and Components in Nuclear Facilities. United States of America.: American Society of Civil Engineers, 2005.

[25] A. Farzinpour, E. Mohammadi Dehcheshmeh, V. Broujerdian, S. Nasr Esfahani, and A. H. Gandomi, "Efficient boosting-based algorithms for shear strength prediction of squat RC walls," Case Stud. Constr. Mater., vol. 18, no. February, p. e01928, 20 doi: 10.1016/j.cscm.2023.e01928.

[26] R. H. Faraj, A. A. Mohammed, A. Mohammed, K. M. Omer, and H. Unis, "Systematic multiscale models to predict the compressive strength of self-compacting concrete modified with nano-silica at different curing ages," Eng. Comput., 2021, doi: 10.1007/s00366-021-01385-9.

[27] M. S. Barkhordari and L. M. Massone, "Failure Mode Detection of Reinforced Concrete Shear Walls Using Ensemble Deep Neural Networks," 2022.

[28] H. U. Ahmed, A. S. Mohammed, A. A. Mohammed, and R. H. Faraj, "Systematic multiscale models to predict the compressive strength of fly ash-based geopolymer concrete at various mixture proportions and curing regimes," PLoS One, vol. 16, no. 6 June, pp. 1–26, 2021, doi: 10.1371/journal.pone.0253006.

[29] Jonathan Paulus and Ika Bali, "SHEAR STRENGTH OF REINFORCED CONCRETE WALLS WITH BOUNDARY MEMBER, Semin. Nas. Teknol. dan Sains II 2016, no. August 2016, pp. 1–6, 2016.

[30] J. Chandra, K. Chanthabouala, and S. Teng, "Truss model for shear strength of structural concrete walls," ACI Struct. J., vol. 115, no. 2, pp. 323–335, 2018, doi: 10.14359/51701129.

[31] M. L. Moretti, S. Kono, and T. Obara, "On the shear strength of reinforced concrete walls," ACI Struct. J., vol. 117, no. 4, pp. 293–304, 2020, doi: 10.14359/51724668.

[32] D. Nguyen, V. Tran, D. Ha, V. Nguyen, and T. Lee, "A machine learning-based formulation for predicting the shear capacity of squat flanged RC walls," Structures, vol. 29, no. December 2020, pp. 1734–1747, 2021, doi: 10.1016/j.istruc.2020.12.054.

[33] A. F. Al-Bayati, "Shear Strength of Reinforced Concrete Squat Walls," Civ. Eng. J., vol. 9, no. 2, pp. 273–304, 2023, doi: 10.28991/CEJ-2023-09-02-03.