ISSN 1392-3730/eISSN 1822-3605

VILNIUS TECH
Vilnius Gediminas Technical University

JOURNAL of CIVIL ENGINEERING and MANAGEMENT

2024
Volume 30
Issue 4
Pages 373–390
https://doi.org/10.3846/jcem.2024.21348

APPLICATION OF MACHINE LEARNING MODELS AND GSA METHOD FOR DESIGNING STUD CONNECTORS

Guorui SUN[1]✉, Jiayuan KANG[1], Jun SHI[1,2]

[1] *Key Lab of Structures Dynamic Behavior and Control of the Ministry of Education, Harbin Institute of Technology, Harbin 150090, China*
[2] *School of Civil Engineering, Central South University, Changsha 410075, China*

Article History:
- received 3 November 2023
- accepted 30 January 2024

Abstract. The design of stud connectors is aided by determining the relationship between shear strength and the input variables (number, diameter, height, tensile strength and elastic modulus of the studs, and compressive strength and elastic modulus of the concrete) that influence strength. Since strength is nonlinearly related to the influencing variables, which makes the predictions of the relevant empirical equations unreliable, the use of machine learning (ML) models is preferred. The prediction results of eight machine learning models were evaluated, including linear regression (LR1), ridge regression (RR), lasso regression (LR2), back-propagation artificial neural network (BP ANN), genetic algorithm optimized BP ANN (GA-BP ANN), extreme learning machines (ELM), random forests (RF), and support vector machines (SVM). The results show that the GA-BP ANN model is the most accurate model for prediction with a mean absolute percentage error (MAPE) of 6.17% and an R^2 of 0.9599. Based on the GA-BP ANN model and the global sensitivity analysis (GSA) method, a new parameter importance analysis method was developed to compare the magnitude of the effect of different input variables on strength. It was found that stud diameter had the greatest effect on shear strength.

Keywords: stud connectors, multiple machine-learning model comparisons, global sensitivity analysis, metrics influencing shear strength.

✉Corresponding author. E-mail: *19B933012@stu.hit.edu.cn*

Symbols:

n – Number of studs for a single connector
d – Diameter of stud (mm)
h – The height of the stud (mm)
f_c – Compressive strength of concrete cube (MPa)
f_t – The tensile strength of the stud (MPa)
E_c – Elastic modulus of concrete (MPa)
E_t – Elastic modulus of stud (MPa)
A_s – The cross-sectional area of the stud (mm^2)
f_{ck} – The concrete cylinder compressive strength (MPa)
γ_v – The stud resistance subfactor which value is 0.85
α – The stud height influence factor
φ_{sc} – The resistance factor, equal to 0.85
q_u – Predicted shear strength(N)
q_e – Experimental values of shear strength (N)
μ – Mean value
SD – Standard Deviation
CV – Coefficient of Variation
g_2 – kurtosis
α – regression coefficients
x_i – influencing factor
$\hat{\beta}(k)$ – rigid regression estimation
k – ridge parameter
$\mathbf{X}$ and $\mathbf{Y}$ – the matrix of independent and dependent variables respectively
λ – non-negative positive regular parameter
f – activation function
w_{ij} – the weight of the i-th input and j-th neurons
b_j – hidden layer bias
x_i – the input for the i-th variable
k – the number of single decision tree
x – input variable
$t_i(x)$ – prediction from a single decision tree $f(x, \omega)$ – target prediction
ω – weight vector
b – the threshold
$\Phi(x)$ – the high-dimensional feature space mapping from low-dimensional space x

c – penalty factor
$\|\omega\|^2$ – smoothness or fatness of the function
ε – a non-sensitivity factor
$l(y)$ – loss function
ξ_i and ξ_i^* – slack variable
α_i and α_i^* – Lagrangian multipliers
$K(x, x_i)$ – kernel function
σ – variance of radial basis function
ML – Machine learning
LR1 – Linear regression
RR – Ridge regression
LR2 – Lasso regression
ANN – Artificial neural network
BP – Back Propagation
GA – Genetic algorithm
ELM – Extreme learning machine
RF – Random forests
SVM – Support vector machine
MSE – Mean square error
MAE – Mean absolute error
MAPE – Mean absolute percentage error
RMSE – Root mean square Error
NSE – Nash-sutcliffe efficiency
GSA – Global sensitivity analysis
$f(x_i)$ – Marginal effect
V – Total variance
S_i – Sobol indices of first-order
S_{Ti} – Sobol indices of total-order

1. Introduction

Steel-concrete composite structures are widely used in building structures due to their ability to make full use of the properties of steel and concrete (Ding et al., 2021; Vigneri et al., 2021; Yang et al., 2021). Shear connectors are the key connection parts between steel beams and concrete slabs, and play a huge synergistic role in steel-concrete composite beams (Gu et al., 2019; Kim et al., 2020). Headed studs are the most widely used in bridge engineering due to its convenience of installation, equal shear strength in all directions, satisfactory concrete compaction around the studs, and minimal obstruction to the slab reinforcement (Tm et al., 2019). In order to ensure the safety of the composite structure, it is necessary to investigate the shear strength of the connectors.

The behavior of stud connectors mainly depends on the stud details such as the number of studs, height, diameter and material properties (Wang et al., 2017; Xue et al., 2008). Xue et al. (2012) investigated the different behavior between single and multiple stud connectors by push-out tests. The results show that the ultimate strength of single stud connectors is about 10% greater than that of multi-stud connectors. Wang et al. (2020) analyzed the effect of stud height on the shear performance of the stud connectors and established the shear bearing capacity equation considering the stud height. When the length-diameter ratio of the stud is 4.5~13.2, the shear bearing capacity of the stud increases with the increase of the length-diameter ratio. Hu et al. (2021) concluded that the ratio of stud height to concrete slab thickness has a limited effect on the shear strength of the studs but has a greater effect on the shear stiffness. Currently, small headed studs with a diameter smaller than 22 mm are commonly used in steel-concrete composite bridges for several types of studs whose diameters range from 10 mm to 25 mm are provided in the current specifications. The larger the diameter of the stud, the greater the bearing capacity of the structure (Shim et al., 2004). Similarly, large diameter studs can significantly increase the bearing capacity of the structure, and the shear strength of a 30 mm diameter stud is about 15% higher than that of a 22 mm diameter stud (Wang et al., 2019). The properties of the concrete could also influence the behavior of the stud connectors. In general, most of the concrete slabs of steel-concrete composite beams are made of normal strength concrete, and the compressive strength and elastic modulus of concrete could affect the shear capacity of the studs embedded in concrete (Wu et al., 2021).

Conducting push out tests is one of the main methods for assessing the performance of stud connectors, but it requires a lot of time and money. Finite element simulation is widely used for load carrying capacity calculations, but it requires a high degree of specialization and the output results are highly variable (Farouk et al., 2022, 2023). In order to provide a convenient calculation, a number of formulas have been proposed based on the results of experiments and finite element simulations, including calculations based on test results without considering the effect of damage modes (Luo et al., 2016; Zhang et al., 2020; Zhu et al., 2020). In addition, there are calculation formulas that consider the damage modes of concrete and studs, which have been adopted in different design codes. However, the existing calculation methods are all based on empirical equations obtained by linear regression of experimental data, which are limited by experimental conditions and fail to comprehensively consider the effects of various factors on the bearing capacity. Therefore, there is still a need for a more effective method to reduce the need for push-out tests and to provide a simpler calculation method.

Machine learning (ML) is able to solve complex engineering problems with higher accuracy than the existing methods (Allahyari et al., 2018; Chahnasir et al., 2018; Garzón-Roca et al., 2013; Khalaf et al., 2021; Safa et al., 2016; Tzuc et al., 2021). Slater et al. (2012) combined linear and non-linear regression to calculate the shear strength and found that the former had a smaller error. Hossain et al. (2017) used artificial neural networks (ANN) for shear strength prediction and then verified the reliability of the model using experimental data. Yaseen et al. (2018) found that the combined particle swarm optimization of the support vector machine (SVM) hybrid model in predicting shear strength with high prediction accuracy. Sedghi et al. (2018) used ML model to predict the shear strength of different shear connector and investigated the effect of

different parameters on the ultimate load. The researchers focused on a single ML model and never provided a convincing rationale for the chosen algorithm, so the researchers began comparing different ML models to determine the most suitable one (Setvati & Hicks, 2022; Yosri et al., 2023; Zhang et al., 2023). Table 1 compiles some recently published ML studies and the accuracy of shear strength predictions they have achieved.

It can be seen that the researchers used different ML models to predict the shear strength of the connectors and achieved good results. However, the method of determining the input parameters is not discussed. The predictive model performance of any ML model depends on the input variables used when developing the model. Different input variables lead to different prediction accuracy of the model, so the researchers' determination of the optimal prediction model is also different. Therefore, it is necessary to effectively determine the influencing factors of shear strength on the basis of empirical formulas, and then determine the input variables. Compared with the traditional back-propagation (BP) ANN, Genetic Algorithm Optimized BP ANN (GA-BP ANN) model has higher prediction accuracy and practical effect, because the genetic algorithm can search the global optimal solution, and avoid the problem that BP neural network is easy to fall into the local optimal. Therefore, it is necessary to determine the prediction accuracy of the GA-BP model and compare it with other ML models to determine the optimal prediction model. In addition, various ML algorithms have been used for shear strength prediction models, but the effect of sensitive parameters has not been exploited. The majority of the present approaches are based on experimental data in the analysis of variables, and the influence of variables on shear strength cannot be determined due to the limitations of experimental conditions. Global sensitivity analysis (GSA) provides a quantitative analysis of the effect of different parameters on the structural load carrying capacity (Pianosi et al., 2016; Bernus et al., 2021; Sobol, 1993). Soroush et al. (2020) analyzed the effect of various factors on the performance of concrete slabs by the GSA method and the identified the important variables. Guo et al. (2021) investigated the damage modes of corroded reinforced concrete and the main parameters through global sensitivity analysis. Since GSA method requires a large amount of data, it is necessary to combine ML model with GSA method to propose a parametric analysis method to analyze the influence of various variables on shear strength.

As a solution to the above issues, this paper investigated the shear strength of stud connectors based on experimental, ML model and GSA methods. Determine design factors of shear strength based on previous empirical formula, and establish a database. Then, different ML models are evaluated and their prediction results for shear strength were investigated. Finally, a new parametric analysis method is proposed using GSA combined with ML model to determine the effect of each factor on shear strength.

2. Push-out experiment

2.1. Specimens design

Eight sets of push-out tests were designed to investigate the effect of stud diameter and height on shear strength, as shown in Table 2 and Figure 1. The specimens are numbered N13-H80 to N22-H120, where "N13" and "H80" denote a stud diameter of 13 mm and a stud height of 80 mm, respectively. The dimensions of the H-section steel are 300×300×10×15 mm and the dimensions of the concrete slab are 60×600×150 mm. The yield strength, ultimate strength and elastic modulus of H-section steel plate are 345 MPa, 470 MPa and 210 GPa, respectively. The steel plates are all Q345 type, and the studs are all ML15AL type. The reinforcement bar has a yield strength of 400 MPa, a modulus of elasticity of 200 GPa, a diameter of 8 mm and a type of HRB400.

2.2. Loading setup

The tests were performed by a 5000 kN pressure testing machine. The relative slip was measured by means of a displacement gauge, as shown in Figure 2. The push-out test is carried out by means of multistep loading. In the early stages of loading, the load increment is 20 kN per level. When the load is 0.5~0.8 peak load (P), the load increment is 10 kN per level. When the load is greater than 0.8P, the load increment is 5 kN per level until the specimen is damaged. At the end of the experiment, observe the damaged surface of the specimen, then break the concrete slab and take out the studs to observe their deformation.

2.3. Analysis of test results

The damage modes are shown in Figure 3. The damage mode of all specimens in this test was stud damage. Meanwhile, the studs were deformed only at the root of the shank, while the rest of the studs embedded in the concrete was basically undeformed. Cracks appeared on the surface of the concrete slab as well as at the interface between steel plate and concrete. The larger the diameter of the stud, the greater the width of the concrete crack. The stud height has little effect on the failure mode of specimen.

The load-slip curves are shown in Figure 4. Where, q_e is the experimental value of shear strength. During the initial loading, the slip growth is small. After complete debonding, the load is mainly borne by the studs, which is in the elastic stage, and the slip is approximately linear with the load. Subsequently, the growth rate of slip increases and is not linear, which demonstrates that the specimen enters the elastic-plastic phase. Finally, the stud shank fails and the load on the specimen decreases. The larger the stud diameter, the greater the slip at the ultimate load, but the stud height has less effect on the slip.

The average shear strength of each set of specimens is shown in Figure 5. The larger the stud diameter, the higher the strength of stud connector. The higher the stud height, the higher the strength, but the height has less impact compared to the diameter.

Table 2. Parameters of the specimen

Specimen	n	d (mm)	h (mm)	f_c (MPa)	f_t (MPa)	E_c (GPa)	E_t (GPa)	Note
N13-H80	4	13	80	53	530	37.5	210	n – Number of studs for a single connector; d – Diameter of stud; h – The height of the stud; f_c – Compressive strength of concrete cube; f_t – The tensile strength of the stud; E_c – Elastic modulus of concrete; E_t – Elastic modulus of stud.
N13-H120	4	13	120	53	530	37.5	210	
N16-H80	4	16	80	53	540	37.5	210	
N16-H120	4	16	120	53	540	37.5	210	
N19-H80	4	19	80	53	550	37.5	210	
N19-H120	4	19	120	53	550	37.5	210	
N22-H80	4	22	80	53	560	37.5	210	
N22-H120	4	22	120	53	560	37.5	210	

Table 1. Application of ML model in stud connectors

Reference	ML model	Input variables	Total dataset	R^2	RMSE
Setvati and Hicks (2022)	Linear Regression	Compressive strength of concrete Modulus of elasticity of concrete Tensile strength of stud Diameter of stud Height of stud Diameter of weld collar Height of weld collar	242	0.87	20.18
	Decision Tree			0.87	20.59
	Bagged Ensemble Trees			0.92	16.5
	Super Vector Machine			0.92	16.02
	Gaussian Process Regression			0.92	15.74
	Artificial Neural Network			0.87	20.6
Yosri et al. (2023)	Extreme learning machine	Compressive strength of concrete Stud ultimate strength Stud diameter Stud Height Number of studs Stud spacing	232	0.894	28.41
	Adaptive neuro-fuzzy inference system			0.935	12.76
	Artificial Neural Network			0.906	26.82

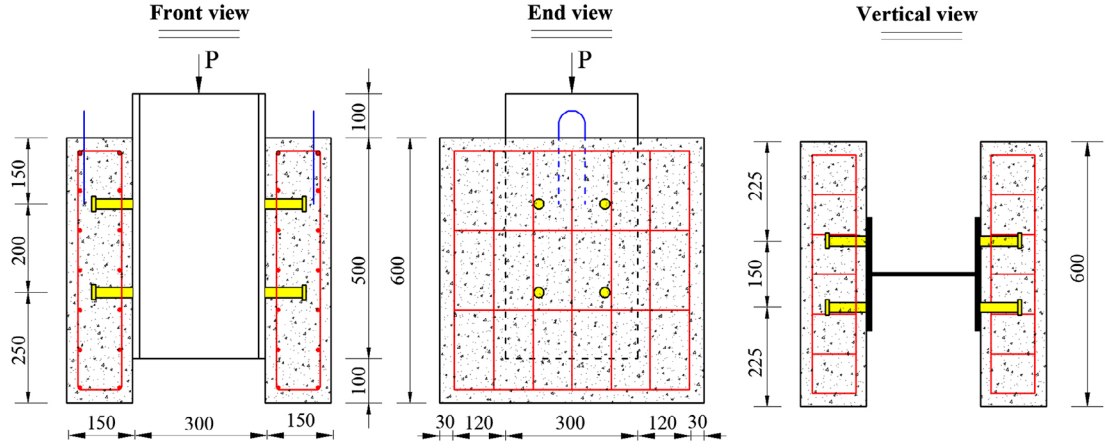

Figure 1. The shape and size of the specimens (mm)

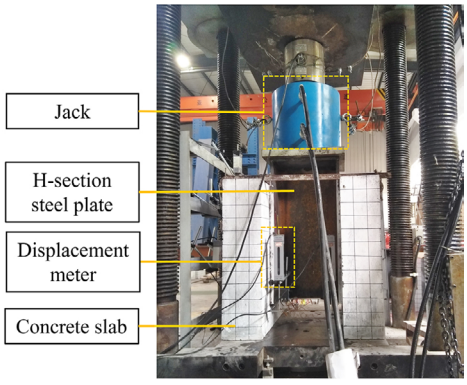

Figure 2. Loading setup

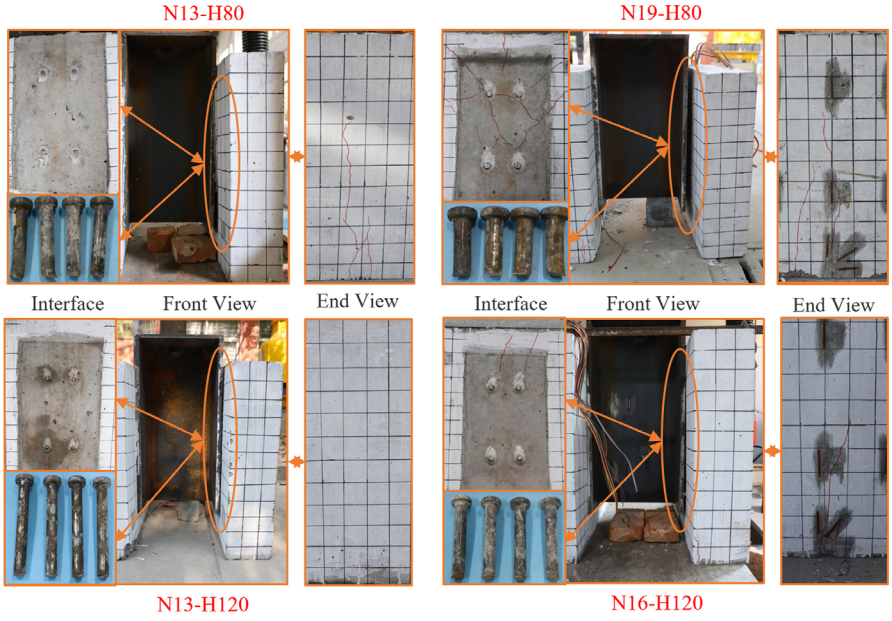

Figure 3. The failure mode of the specimens

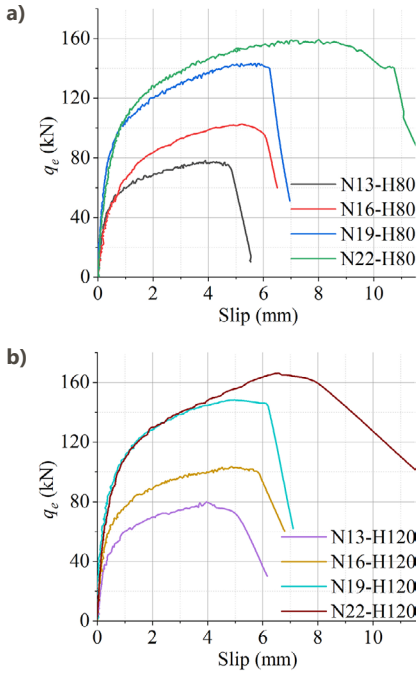

Figure 4. Load-slip curve: a – the height of the stud is 80 mm; b – the height of the stud is 120 mm

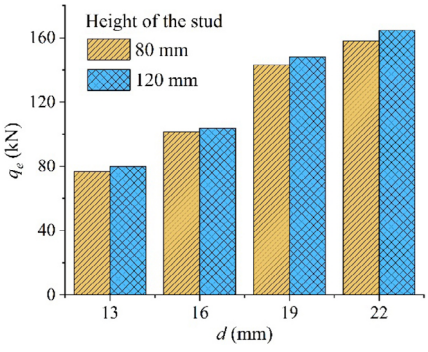

Figure 5. The shear strength of the specimens

3. Database

3.1. Design parameter selection

For stud connectors, a number of national codes propose methods for calculating the shear strength. According to Chinese code GB 50017-2017 (Ministry of Housing and Urban-Rural Development of the People's of China, 2017), the shear strength of single stud is calculated as follows:

$$\begin{cases} q_u = 0.43A_s\sqrt{E_c f_c} \\ q_u = 0.7A_s f_t \end{cases}, \tag{1}$$

where q_u is the calculated value of shear strength (N), A_s is cross-sectional area of the stud (mm^2), f_t is tensile strength of stud (MPa), f_c is cubic compressive strength of concrete (MPa), E_c is elastic modulus of concrete (MPa). Since the damage model of the structure is mainly divided into stud damage and concrete damage, the method for shear strength is mainly divided into two parts, and the minimum value is taken as shear strength of the connectors. According to Chinese code GB 50917-2013 (Ministry of Housing and Urban-Rural Development of the People's of China, 2013), the bearing capacity is calculated as follows:

$$\begin{cases} q_u = 1.19A_s f_t \left(\dfrac{E_c}{E_t}\right)^{0.2}\left(\dfrac{f_c}{f_t}\right)^{0.1} \\ q_u = 0.43\eta A_s\sqrt{E_c f_c} \end{cases}, \tag{2}$$

where E_t is elastic modulus of stud (MPa), η is stud discount factor, related to the spacing between studs and diameter. In Europe, code Eurocode 4 (European Committee for Standardization, 1994) suggests that the shear strength should be calculated as follows:

$$\begin{cases} q_u = \dfrac{0.29\alpha d^2\sqrt{E_c f_{ck}}}{\gamma_v} \\ q_u = \dfrac{0.8A_s f_t}{\gamma_v} \end{cases}, \quad \begin{cases} \alpha = 0.2\left(\dfrac{h}{d}+1\right) \\ \alpha = 1.0 \end{cases}, \tag{3}$$

where α is stud height influence factor, d is the stud diameter (mm), h is the stud height (mm), f_{ck} is concrete cylinder compressive strength (MPa), γ_v is stud resistance subfactor which value is 0.85. The equation takes into account the impact of stud height on the shear strength. In the United States, the code AASHTO LRFD (American Association of State Highway and Transportation Officials, 2017) provides the following equation to calculate the shear strength:

$$\begin{cases} q_u = \varphi_{sc} 0.5 A_s \sqrt{E_c f_{ck}} \\ q_u = \varphi_{sc} A_s f_t \end{cases}, \tag{4}$$

where φ_{sc} is the resistance factor, equal to 0.85. It can be seen that the diameter, height, modulus of elasticity and tensile strength of the bolt affect the shear strength. In addition, the modulus of elasticity and compressive strength of concrete also affect the shear strength. The equations (Xue et al., 2008; Wang et al., 2020) from the references are as follows:

$$q_u = 3\lambda A_s f_t \left(\frac{E_c}{E_t}\right)^{0.4} \left(\frac{f_c}{f_t}\right)^{0.2}, \ \lambda = \begin{cases} 6 - \dfrac{h}{1.05d} & (h/d \leqslant 5) \\ 1 & (5 < h/d < 7); \\ \dfrac{h}{d} - 6 & (h/d \geqslant 7) \end{cases} \tag{5}$$

$$q_u = 17.31 A_s f_t \left(\frac{h}{d}\right)^{0.27} \left(\frac{E_c}{E_t}\right)^{1.75} \left(\frac{f_c}{f_t}\right)^{0.14}. \tag{6}$$

At the same time, the number of studs (n) also influences the strength. Hence, n, d, h, f_c, f_t, E_c and E_t are chosen as the parameters affecting the shear strength. Although different empirical equations have different input variables, all of them consistently have the diameter of the stud in the input variable and the shear strength increases with the diameter of the stud.

3.2. Statistical summary of the data

One hundred test data were selected based on experimental data in this paper and references (Ding et al., 2014, 2017; Shim et al., 2004; Wang & Liu, 2013; Wang et al., 2017, 2020; Xue et al., 2008; Yu et al., 2014). According to the statistical perspective, standard deviation (SD) is usually applied to denote the accuracy and repeatability of the test results. A low SD value indicates that the points are very close to the mean of the set, while a high SD value indicates that the points are distributed over a larger range of values. The calculation equation is as follows:

$$\mathrm{SD} = \sqrt{\frac{1}{100} \sum_{i=1}^{100} \left(X_i - \mu\right)^2}; \tag{7}$$

$$\mathrm{CV} = \frac{\mathrm{SD}}{\mu}; \tag{8}$$

$$g_2 = \frac{\frac{1}{n}\sum_{i=1}^{n} (X_i - \mu)^4}{(\frac{1}{n}\sum_{i=1}^{n} (X_i - \mu)^2)^2} - 3, \tag{9}$$

where, μ is mean value, CV is coefficient of variation, g_2 is kurtosis. As, shown in Table 3, the diameter of the studs is mainly distributed between 16 and 22 mm, which is in accordance with the codes. In addition, studs larger than 25 mm in diameter are available to study the effect of large diameter studs. For similar reasons, studs with a height greater than 200 mm are present. Most of the compressive strengths of the concrete are between 35 MPa and 60 MPa, indicating that the specimens in the database are mainly normal strength concrete. The tensile strength of the studs was concentrated between 350 and 550 MPa. The distribution of the data is more concentrated for the modulus of elasticity of concrete and studs. The coefficient of variation is used to compare the magnitude of dispersion of the data, and the comparison reveals that the number of studs has the greatest degree of dispersion, and the modulus of elasticity of the concrete and the studs has a lesser degree of dispersion. Kurtosis is the characteristic number of peak heights of the distribution curve at the mean, which reflects the sharpness of the peaks. A high kurtosis means that the increase in variance is caused by low frequency of extreme values greater or less than the mean. It can be seen that n, h, f_t and E_c variables show a sharp feature in the probability density distribution because the kurtosis coefficient is greater than zero. On the other hand, since the coefficient of kurtosis is less than zero, the other variables exhibit a flat distribution.

3.3. Calculation accuracy of empirical equations

Based on the database, the experimental values are compared with the results calculated by Eqn (1)~Eqn (6) (see Table 4 and Figure 6). When the value of h/d is greater than 7, the error of Eqn (5) is larger, so it is not discussed

Table 3. Test data statistics

Parameter	n	d (mm)	h (mm)	f_c (MPa)	f_t (MPa)	E_c (GPa)	E_t (GPa)	q_e (kN)
Data	100	100	100	100	100	100	100	100
Max	9.00	30.00	400.00	64.50	675.00	37.50	213.00	330.10
Min	1.00	13.00	50.00	33.10	326.00	29.00	195.00	66.00
μ	2.81	21.16	152.89	46.36	441.76	33.99	206.24	155.27
SD	1.59	4.42	59.96	9.60	60.35	2.10	5.71	54.51
CV	0.57	0.21	0.39	0.21	0.14	0.06	0.03	0.35
g_2	3.09	–0.46	2.07	–0.41	1.22	0.06	–0.45	–0.14

Table 4. The absolute error distribution based on national standard codes (%)

Error	Eqn (1)		Eqn (2)		Eqn (3)		Eqn (4)		Eqn (6)	
	Separate	Cumulative	Separate	Cumulative	Separate	Cumulative	Separate	Cumulative	Separate	Cumulative
0–10	16	16	9	9	5	5	34	34	47	47
11–20	18	34	13	22	12	17	30	64	27	64
20–30	24	58	26	48	16	33	20	84	20	94
30–40	27	85	30	78	35	68	12	96	5	99
40–50	12	97	16	94	23	91	3	99	1	100
50–60	3	100	6	100	9	100	0	99	0	100
Average	26.55		30.09		33.57		17.56		13.16	

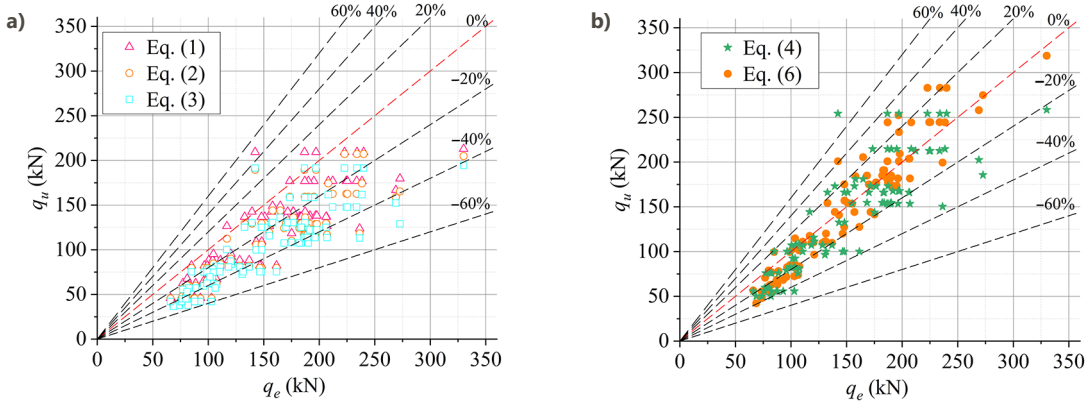

Figure 6. Calculation results: a – Eqns (1)~(3); b – Eqns (4) and (6)

in this paper. The average error of Eqn (1) is 26.5%, and the average error of Eqns (2) and (3) is more than 30%. Eqn (6) has the smallest average error and better error distribution than the other formulas, but the average error is still greater than 10%. In conclusion, the proposed empirical equations for shear strength are very conservative, and most predicted data are less than experimental values. In order to develop a better shear strength prediction model, it is necessary to develop a database in this area and to continuously expand it. To reduce experimental costs, a better model must be developed to calculate the shear strength.

4. Calculation method of bearing capacity

4.1. Description of ML models

Machine Learning (ML) systems are flexible and intelligent computer algorithms that provide data-driven tools for many systems to improve automatic learning and prediction capabilities (Ghorbani et al., 2020). Several ML models are developed in this paper, including linear regression (LR1), ridge regression (RR), lasso regression (LR2), back-propagation artificial neural network (BP ANN), genetic algorithm optimized BP ANN (GA-BP ANN), extreme learning machines (ELM), random forests (RF), and support vector machines (SVM). The software used for this analysis is python and the CPU of the computer is Core i9 and the GPU is GeForce GTX series.

LR1 model is relatively simple to construct and the coefficients of the linear regression model have a clear physical meaning. However, LR1 model requires a linear relationship between the independent variables and the dependent variable and the error term obeys a normal distribution, which may not hold true in practical problems. LR1 regression model is sensitive to outliers, which may lead to model instability and inaccurate prediction results. The RR model is suitable for cases where there is multicollinearity between independent variables or where the number of independent variables is greater than the sample size, and it can prevent overfitting. However, the RR model does not converge to 0 for factors that have a very small effect. The LR2 model is capable of completely eliminating the weights of the least important features, but it is relatively complex to compute, as well as computationally unstable when dealing with highly correlated variables.

The BP ANN model has very strong nonlinear mapping ability and optimization computation ability, strong identification and classification ability for input samples. However, the convergence speed is slow, there are local minima in the objective function, and it is difficult to determine the number of hidden layers and hidden layer nodes. The GA algorithm can optimize the weights and thresholds of BP ANN to overcome the problem that BP neural networks are easy to fall into local minima. Not only can it automatically search for the optimal number of neurons in the hidden layer of the neural network, but

it can also fix the weights and thresholds after GA optimization so that the final results of the network will remain unchanged after many runs.

The ELM model is a feed-forward neuron network whose learning process requires only one forward propagation without iterative updating of weights. It should be noted that the number of neurons in the hidden layer and the choice of activation function may affect the performance of the ELM, and thus need to be adjusted experimentally. RF model has better generalization performance and can effectively reduce the variance of the model. However, it will overfit in some noisy regression problems. SVM is a novel small-sample learning method which, unlike existing statistical methods, achieves efficient inference from training samples to prediction samples. The SVM model utilizes an inner product kernel function instead of a nonlinear mapping into a higher dimensional space, and the nonlinear mapping is the theoretical basis of the method. SVM models are capable of solving high dimensional problems and nonlinear problems, avoiding neural network structure selection and local minima. However, the model requires the selection of appropriate kernel functions and is difficult to implement for large-scale training samples.

4.2. Linear regression

LR1 is used to assess the linear relationship between the independent and dependent parameters (Slater et al., 2012). The linear relationship is interpreted by assigning regression coefficients (α), and the best regression coefficients are selected by gradient descent. The root mean squared error (RMSE) is minimized by choosing the value of regression coefficients to obtain the best-fit relationship. The independent variables in this paper are the seven influencing factors (x_i) and the dependent variable is shear strength:

$$y_{LR1} = \alpha_0 + \sum_{i=1}^{7} x_i \alpha_i. \tag{10}$$

The number of independent variables in multiple linear regression is the main factor affecting the performance of the model. The training time of LR1 model is 0.01s. In this paper, the t-test (Özbayrak et al., 2023) was used to screen the variables that have a significant effect on the regression model. Table 5 lists the model with the highest predictive performance based on the t-test results, which has a P-value of less than 0.05 for all independent variables, indicating that they have a significant effect on the regression equation.

Table 5. Summary of linear regression models for predicting shear strength

Variable	Magnitude of coefficient	t value	P value
Constant	–355.478	–10.344	0
n	–4.372	–2.906	0.00457
d	8.8	16.717	0
h	0.238	6.562	2.96624E–9
f_c	–0.634	–2.349	0.02092
f_t	0.132	3.599	5.14165E–4
E_c	7.844	6.146	1.9612E–8

4.3. Ridge regression

RR is an improved least squares regression method, which is proposed for the case that the least squares regression coefficients cannot solve the singularity of the coefficient matrix of the regular set of equations (Yang & Wen, 2018). It also has the ability to select variables to overcome the effects of multicollinearity based on the ridge trace diagram, eliminate independent variables with small or unstable regression coefficients in ridge regression, and reduce errors. To reduce the mean square error (MSE) of the linear regression model, the ridge estimate of the model is

$$\hat{\beta}(k) = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + k\mathbf{I}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathrm{Y}, \tag{11}$$

where $\hat{\beta}(k)$ is rigid regression estimation, k is a ridge parameter, taking values from 0 to 10, degenerating to linear regression when k = 0, **X** and **Y** is the matrix of independent and dependent variables respectively. The training time of RR model is 0.01 s.

4.4. Lasso regression

LR2 is a regularization method that combines variable selection and parameter estimation simultaneously (Yang & Wen, 2018). The method minimizes the residual sum of squares (RSS) by adding a parametric number as a penalty constraint to the calculation, which enables to produce certain regression coefficients equal to zero. The lasso parameter estimates are defined as follows:

$$\hat{\beta} = \mathrm{argmin}_{\beta}\left[\sum_{i=1}^{t}\left(y_i - \beta_0 - \sum_{j=1}^{7} X_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{7}\left|\beta_j\right|\right], \tag{12}$$

where $\lambda\sum_{j=1}^{7}\left|\beta_j\right|$ is penalty items, λ is a non-negative positive regular parameter. The training time of LR2 model is 0.01 s.

4.5. BP ANN model

ANN model is one of the most widespread ML algorithms for solving nonlinear problems in engineering, although they are known as black boxes for their regression-like computation of hidden layers. ANN model are complex and powerful computational systems composed of a number of simple neurons connecting to each other in some way (Hossain et al., 2017). The signal is adjusted when the information flows through the following ANN as follows:

$$y_i = f\left(\sum_{i=1} w_{ij}x_i + b_j\right), \tag{13}$$

where f is the activation function, w_{ij} is the weight of the i-th input and j-th neurons, b_j is hidden layer bias,

and x_i is the input for the i-th variable. As shown in Figure 7, the back propagation (BP) ANN model consists of an input layer, a hidden layer and an output layer. In this study, the number of neurons in the hidden layer was determined by a trial-and-error procedure (Guan et al., 2019), which compared the accuracy of the ANN with different numbers of neurons in the hidden layer, and the optimal number of neurons with the largest R^2 output was selected for further investigation. The S-shaped tangent function "tansig" was used for the neuron transfer function in the hidden layer and the S-shaped logarithmic function "logsig" was used for the neuron transfer function in the output layer. The model is trained using the function "trainlm". The input layer is the seven variables identified in this paper, the output layer is the shear strength, and the hidden layers is 1. The number of training iterations is 1000, the learning rate is set to 0.1, and the minimum error of the training target is set to 0.00001. The objective function for the training subset is mean squared error (MSE) and the analyze time of BP ANN model is 0.72 s.

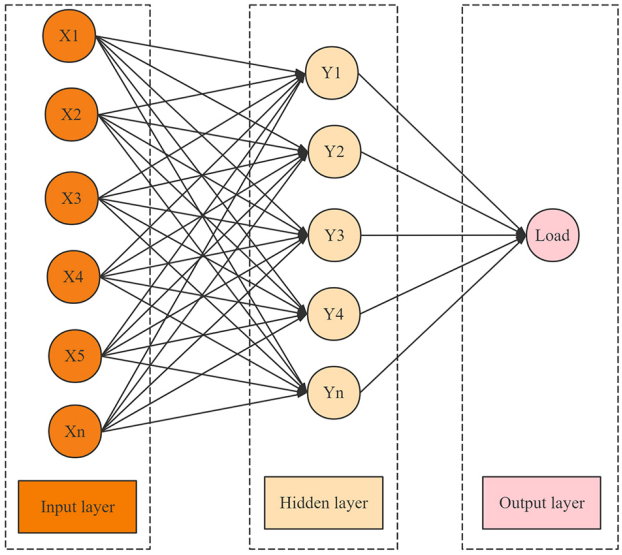

Figure 7. BP ANN model

4.6. GA-BP ANN model

Genetic algorithm is a method that mimics the natural mechanism of inheritance and the theory of biological evolution. The BP ANN model has the problems of slow convergence and the accuracy may not meet the requirements. Genetic algorithm has global search capability, and can avoid the model from into local minimum (Khalaf et al., 2021). The calculation process is shown in Figure 8. The behavior of a genetic algorithm is controlled by a set of hyperparameters such as population size and mutation rate. When the population size is too small, inbreeding occurs, generating pathological genes and preventing the population from evolving to produce the desired ideal population size. When the population size is too large, the results are difficult to converge, leading to wasted resources and reduced robustness. If the mutation probability is too small, the population diversity decreases too quickly, resulting in rapid loss of effective genes that cannot be easily repaired. When the mutation probability is too large, the probability of higher-order patterns being destroyed increases. Similar to the mutation probability, when the crossover probability is too large, it will frequently destroy existing favorable patterns, increase stochasticity, and miss optimal individuals. A crossover probability that is too small cannot effectively update the population. If the number of evolutionary generations is too small, the algorithm does not converge easily and the population is not yet mature. If the number of evolutionary generations is too large, the algorithm is already proficient or the population has converged prematurely, there is no point in continuing to evolve, it will only increase the time expenditure and waste of resources. Hence, the population size is set to 100, the number of evolutionary iterations is 50, the crossover probability is 0.9, the variation probability is 0.1. The analyze time of GA-BP ANN model is 1.65 s.

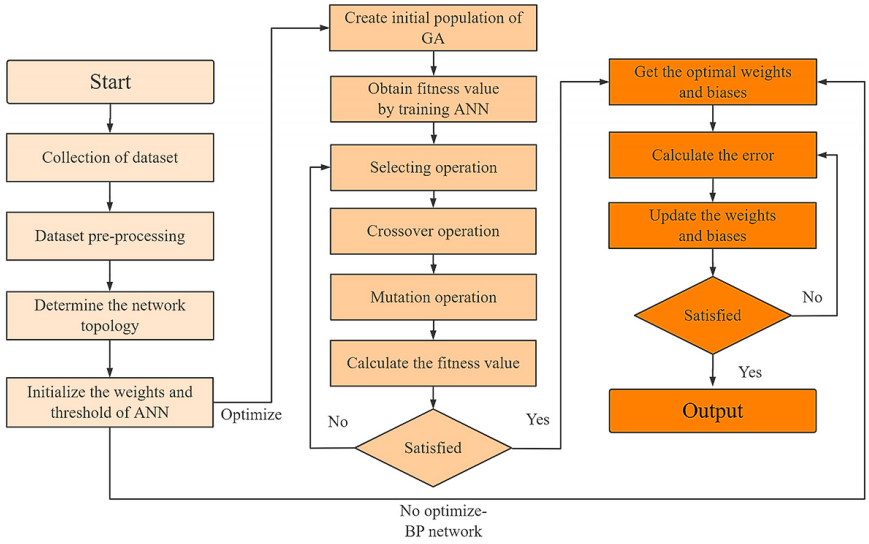

Figure 8. Calculation process of GA-BP ANN model

4.7. Extreme learning machine

ELM model is no longer a gradient-based algorithm in the training stage, but utilizes random input layer weights and biases (Shen et al., 2020). During data training, the randomly generated values do not need to be changed, and only the number of neurons in the hidden layer needs to be adjusted to arrive at the optimal solution. The calculation process is as follows: (1) Data preprocessing, using normalization; (2) Train the ELM model to find the connection weights of the hidden layer and the output layer; (3) Prediction of data using the obtained output layer weights. The analyze time of ELM model is 2.33 s.

4.8. Random forest

RF are the algorithms that combine multiple decision trees together (Breiman, 2001). It is a supervised prediction algorithm capable of making regression predictions based on data from input and output variables. This approach reduces the possibility of over-decision compared to algorithms based on a single decision tree. It also reduces the variance and bias of the predictions without compromising the accuracy of the decisions by collectively evaluating the predictions of all decision trees (Barjouei et al., 2021). During the calculation of RF model, *k* sample sets are extracted from the original sample set, and then the corresponding decision trees are formed by training the *k* sample sets respectively. Finally, the output of the k decision trees is combined with the strategy to get the final model output. The input variables are categorized by each decision tree. The RF model classification tree is set to 800 trees with 5 leaves. The objective function is MSE, and the example analyze time is 9.45 s. The prediction function for the RF algorithm is expressed as:

$$\hat{f}_{RF}^{K}(x)=\frac{1}{K}\sum_{k=1}^{K}T_i(x), \tag{14}$$

where K is the number of single decision tree, X is input variable and $T_i(x)$ is prediction from a single decision tree.

4.9. Support vector machine

SVM model achieves classification and prediction by modeling the mapping between the input feature vectors and the vectors of the output (Yaseen et al., 2018). That is, given an input sample, the relevant output result is available under the mapping relation. The SVM model requires datasets to define its input variables and corresponding output variables, and the predictions of the model are achieved by fitting the regression function accurately. The learning function used by the model to approximate the target value is as follows:

$$f(x,\omega)=\omega\phi(x)+b\,, \tag{15}$$

where $f(x, \omega)$ is target prediction, ω is weight vector, b is the threshold and $\Phi(x)$ is the high-dimensional feature space mapping from low-dimensional space x. The penalty risk function is calculated as follows:

$$\begin{cases} R(C)=C\dfrac{1}{l}\sum_{i=1}^{n}L(y)+\dfrac{1}{2}\|\omega\|^2 \\ L(y)=\begin{cases}\left|f(x_i,\omega)-y_i\right|-\varepsilon & \left|f(x_i,\omega)-y_i\right|>\varepsilon \\ 0 & \left|f(x_i,\omega)-y_i\right|\le\varepsilon\end{cases}\end{cases}, \tag{16}$$

where C is penalty factor, a larger value indicates a higher degree of concern for the total error throughout the optimization process, which in this paper takes the value of 100, $\|\omega\|^2$ is smoothness or fatness of the function, ε is a non-sensitivity factor which ignores the error within a certain distance from the true value, and in this paper takes the value of 0.1, $L(y)$ is loss function. Introducing the slack variable (ξ_i and ξ_i^*), Eqn (16) is changed to:

$$\min\frac{1}{2}\|\omega\|^2+C\frac{1}{l}\sum_{i=1}^{l}\left(\xi_i-\xi_i^*\right) \quad ST\begin{cases} y_i-f\left(x_i,\omega\right)\le\varepsilon+\xi_i \\ f\left(x_i,\omega\right)-y_i\le\varepsilon+\xi_i^* \\ \xi_i,\xi_i^*>0\end{cases}. \tag{17}$$

After constructing the Lagrangian function through Lagrange multipliers, the functional expression of the SVM model is:

$$f\left(x\right)=\sum_{i=1}^{N}\left(\alpha_i-\alpha_i^*\right)K\left(x,x_i\right)+b, \tag{18}$$

where α_i and α_i^* is Lagrangian multipliers and $K(x, x_i)$ is kernel function. The kernel function affects the prediction performance of the SVM model, and in this paper the radial basis function is used. It is a localized kernel function which maps a sample to a higher dimensional space, and it is one of the most widely used, with relatively good performance for both large and small samples. The formula for radial basis function is expressed as:

$$K\left(x,x_i\right)=\exp\left(-\frac{\left\|x_i-x_j{}^2\right\|}{2\sigma^2}\right), \tag{19}$$

where σ is variance of radial basis function. The analyze time of SVM model is 1.43 s.

5. Analysis of prediction results

5.1. Evaluation standards

To determine the gap between the predicted values of the model and the true values of the sample, mean absolute error (MAE), mean absolute percentage error (MAPE), root mean square error (RMSE), R-squared (R^2) and nash-sutcliffe efficiency (NSE) is introduced:

$$\mathrm{MAE}=\frac{1}{t}\sum_{i=1}^{t}\left|q_{e(i)}-q_{u(i)}\right|; \tag{20}$$

$$\mathrm{MAPE}=\frac{1}{t}\sum_{i=1}^{t}\left|\frac{q_{e\,(i)}-q_{u\,(i)}}{q_{e\,(i)}}\right|\times100\%; \tag{21}$$

$$\mathrm{RMSE}=\sqrt{\frac{1}{t}\sum_{i=1}^{t}\left|q_{e(i)}-q_{u(i)}\right|^2}; \tag{22}$$

$$R^2=\frac{\left[\sum_{i=1}^{t}\left(q_{e(i)}-q'_e\right)\left(q_{u(i)}-q'_u\right)\right]^2}{\sum_{i=1}^{t}(q_{e(i)}-q'_e)^2\sum_{i=1}^{t}\left(q_{u(i)}-q'_u\right)^2}; \tag{23}$$

$$\mathrm{NSE}=1-\frac{\sum_{i=1}^{t}(q_{e(i)}-q_{u(i)})^2}{\sum_{i=1}^{t}(q_{e(i)}-q'_e)^2}, \tag{24}$$

where t is quantity of samples, $q_{u(i)}$ is the i-th predicted value, $q_{e(i)}$ is the i-th experimental value, q'_u is average of all the predicted values and q'_e is average of all the experimental values. Smaller prediction errors do not necessarily guarantee better results. Therefore, the MAPE, RMSE, R^2 and NSE need to be considered to guarantee the uniqueness of solution. The lower values of MAE, MAPE and RMSE values and the values of R^2 and NSE closer to 1.00 indicate that the model predicts better results. Among them, the evaluation criteria (Bayram & Çıtakoğlu, 2023) of MAPE, R^2 and NSE are shown in Table 6.

5.2. Comparison of prediction methods

The division of training set was kept at 80% for all ML models and 20% for the testing set. The results of different ML models are shown in Table 7. Based on the judgement criteria, it can be seen that the value of MAPE for all ML models is less than 10, indicating that all ML models have "high prediction". For R^2 and NSE, the GA-BP model, BP model, RF model, ELM model and SVM model perform "very good". Among them, the GA-BP ANN model has the best prediction accuracy, the training set and test set have the lowest value of MAE and RMSE. The value of R^2 (0.9629 for the training set and 0.9548 for the test set) and NSE (0.9594 for the training set and 0.9523 for the test set) are the highest, and the value of MAPE (6.01 for the training set and 7.75 for the test set) is only higher than that of SVM model. LR1 model, RR model and LR2 model have the lowest R^2 and NSE values, and the R^2 of training sets are all below 0.9. Meanwhile, the MAE values, MAPE values and RMSE values of LR1, RR and LR2 models are higher than the other models.

The distribution of test results and predicted results of different models is shown in Figure 9. The maximum error of machine learning models does not exceed 50%, among which the maximum error of BP ANN model, GA-BP model, ELM model, RF model and SVM model is less than 40%. The trend of the measured data versus the predicted data was roughly on a 45° line in each model, and each of the machine learning prediction models achieved an acceptable level of accuracy. In the case of the GABP-ANN and SVM models, the gap between predicted and experimental values was smaller than for the other ML models. As shown in Figure 10, the MAE of all machine learning models is less than 10%, and the prediction error of most data (over 90%) is less than 20%. For the prediction performance of GA-BP ANN model, the percentage of prediction error less than 20% is more than 95%, and the percentage of prediction error less than 10% is more than 78%.

Taylor diagrams are given in Figure 11 for evaluating and comparing the test results of the models. Taylor diagrams compare the predictive performance of different models based on standard deviation, center root mean square error (CRMSE) and correlation with test values (Citakoglu & Demir, 2023; Coşkun & Citakoglu, 2023).

Table 6. Range of values for MAPE, R^2 and NSE performance standards

Range of MAPE	Performance	Range of R^2	Performance	Range of NSE	Performance
MAPE ≤ 10	High prediction	$R^2 \le 0.6$	Unsatisfactory	NSE ≤ 0.4	Unsatisfactory
10 < MAPE ≤ 20	Good prediction	$0.6 < R^2 \le 0.75$	Regular	0.4 < NSE ≤ 0.5	Acceptable
20 < MAPE ≤ 40	Reasonable prediction	$0.75 < R^2 \le 0.9$	Good	0.5 < NSE ≤ 0.6	Satisfactory
40 < MAPE	Inaccurate prediction	$0.9 < R^2 \le 1.0$	Very good	0.6 < NSE ≤ 0.7	Good
–	–	–	–	0.7 < NSE ≤ 1.0	Very good

Table 7. Predicted results of ML model

Model	Training data					Test data				
	MAE	MAPE	RMSE	R^2	NSE	MAE	MAPE	RMSE	R^2	NSE
LR1	12.12	9.04	15.93	0.9241	0.8922	13.44	9.39	19.30	0.8866	0.8647
RR	12.09	8.90	15.56	0.9188	0.8933	12.98	9.18	18.90	0.8819	0.8729
LR2	12.29	9.20	15.45	0.9195	0.8912	12.97	9.18	18.87	0.8821	0.8799
BP	12.02	8.33	14.19	0.9331	0.9312	12.19	9.20	16.13	0.9230	0.9123
GA-BP	8.02	6.01	11.25	0.9629	0.9594	10.30	7.75	12.91	0.9548	0.9523
ELM	10.88	7.80	15.06	0.9347	0.9212	13.23	9.35	15.19	0.9204	0.9123
RF	11.85	8.68	15.64	0.9250	0.9112	12.47	8.21	16.96	0.9132	0.9024
SVM	8.44	5.70	13.74	0.9436	0.9387	11.50	7.77	13.82	0.9414	0.9311

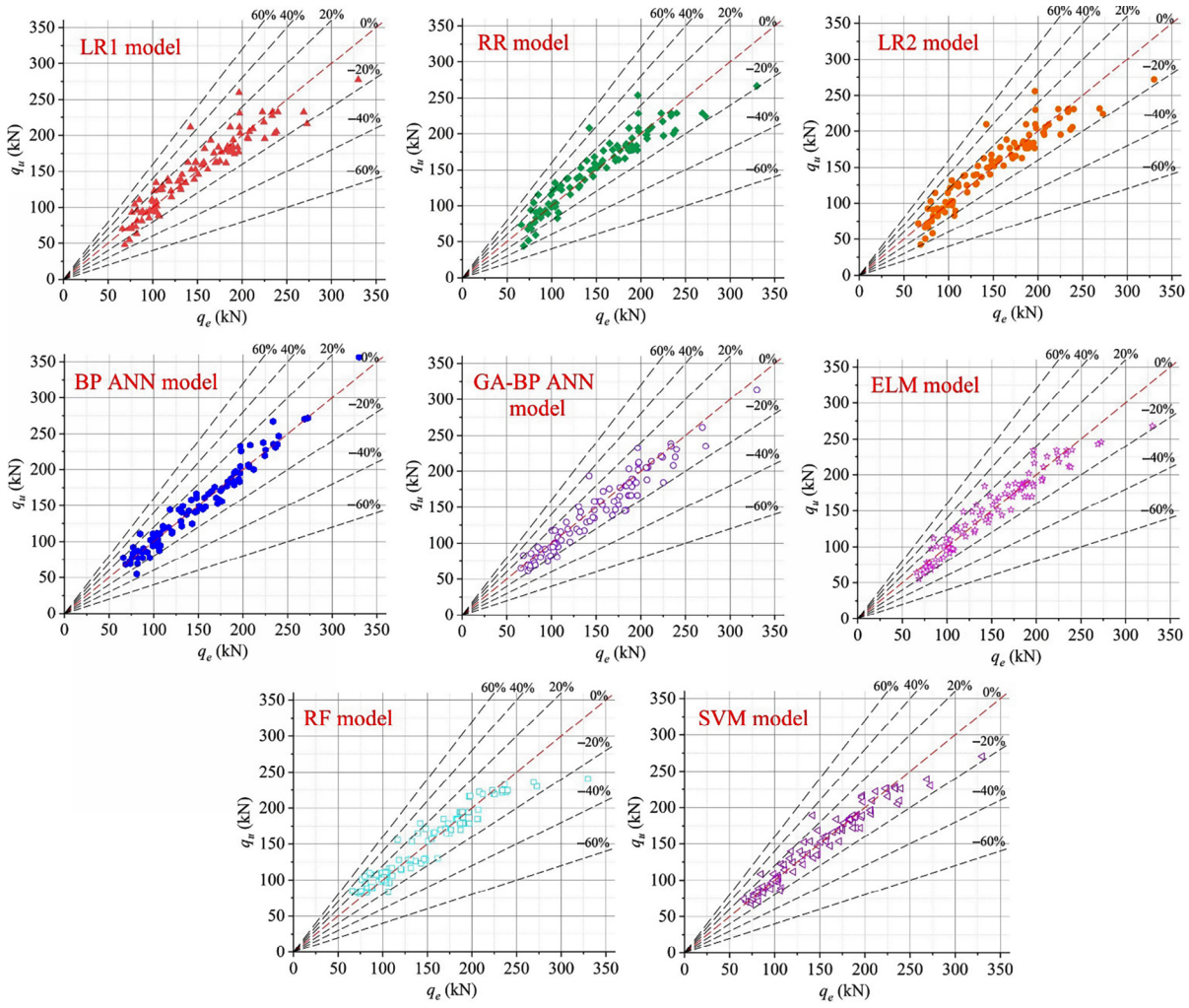

Figure 9. Distribution of experimental results and predicted results

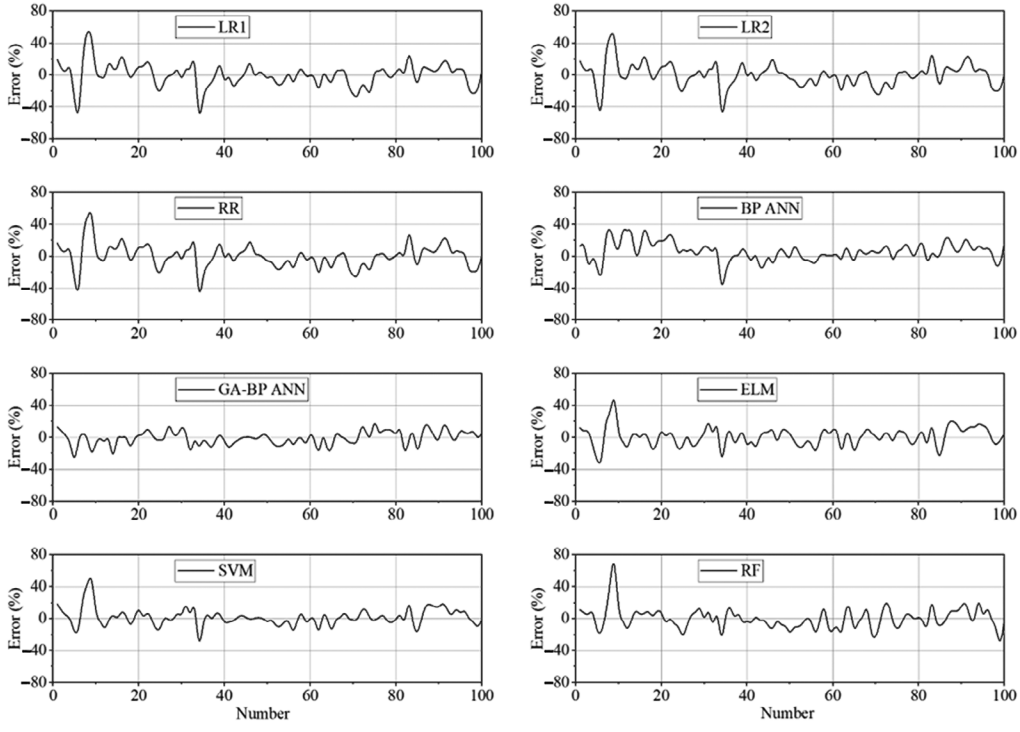

Figure 10. Prediction errors of different ML models

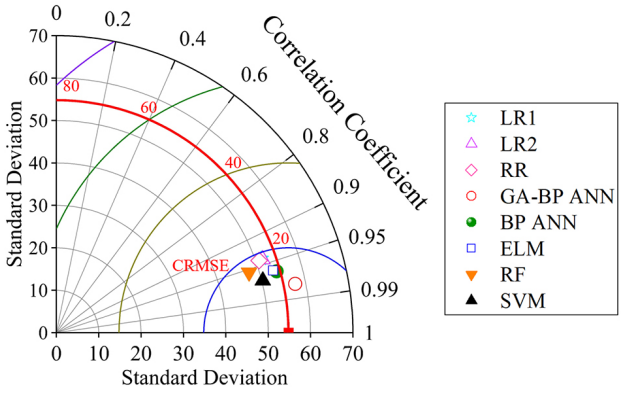

Figure 11. Taylor diagram for ML models

The proximity to the observations indicates better predictive performance of the model. It can be seen that the correlation coefficients and CRSME values of the GA-BP ANN model are very close to the observation point. Therefore, the prediction performance of GA-BP ANN model is better than other models. According to the criteria of CRMSE, the BP model, ELM model and SVM model have similar distances from the observation point, indicating that the predictive performance of these models is similar. The LR1 model, RR model and LR2 model have similar distances from the observation point and these models are furthest away from the observation point, indicating that the predictive performance of these models is poorer than the other ML models.

As shown in Figure 12, 10-fold cross-validation was performed on all models. The error rates for the individual cross-validation training ranged between 4.1% and 12.3%. The average error for the training set ranged from 6.05% to 9.82%, and the average error for the test set ranged from 5.96% to 9.62%. Therefore, cross-validation shows that all ML models have good prediction results. The GA-BP ANN model has the best performance, followed by the SVM model. The LR1, RR and LR2 models display the least accuracy.

The Kruskal-Wallis test is used to analyze the predictions of all ML models to determine if there is a significant difference between the predicted and experimental values. This test has been previously used to compare the distribution of ML model predictions to measurements (Zouzou & Citakoglu, 2023). The Kruskal-Wallis test is a nonparametric alternative to the traditional one-way ANOVA that tests whether the data in all groups are from the same population by comparing the medians of the data sets. In the Kruskal-Wallis test, the values in each group are ranked and the rank is used to calculate the test statistic rather than the actual value. The significance level of this test is 0.05 and the hypothesis of H_0 is that the predicted and test values are from the same group. As shown in Table 8, all ML models passed the Kruskal-Wallis test, indicating that the predicted and experimental values of the ML models are not significantly different.

The prediction results of the GA-BP ANN model were compared with the calculation results of the empirical equations, as shown in Table 9 and Figure 13. The results of the dataset for the GA-BP ANN model show less dispersion compared to the empirical formulations. Specifically, Eqn (6) shows an underestimation at lower shear strengths (<130 kN) and an overestimation at higher shear strengths (>150 kN). There is no overall trend of overestimation or underestimation for the prediction results of the GA-BP ANN model. The prediction results of the machine learning models are more accurate compared to the empirical formulas. The comparison with other machine models and formulas demonstrates the superiority of the developed GA-BP ANN model.

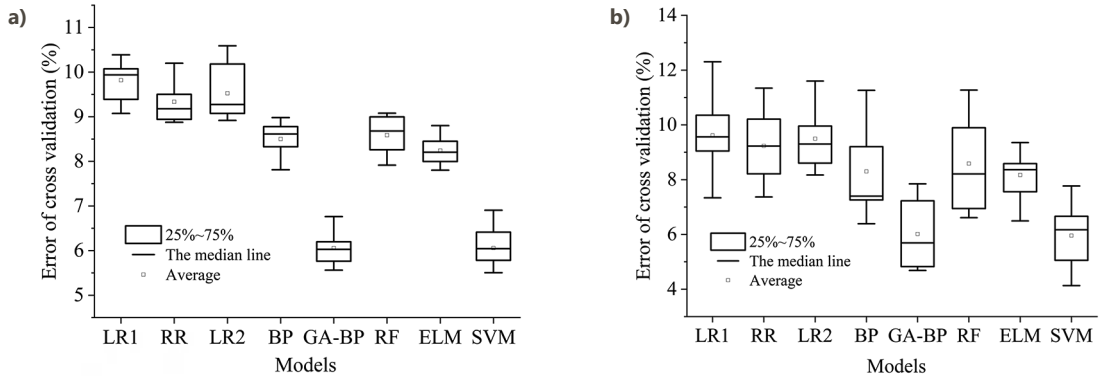

Figure 12. Error of cross validation: a – training data; b – test data

Table 8. Kruskal-Wallis test results

Model	LR1	RR	LR2	BP	GA-BP	ELM	RF	SVM
P value	0.8912	0.8604	0.8989	0.4006	0.9805	0.9435	0.845	0.8989
H_0	Accept	Accept	Accept	Accept	Accept	Accept	Accept	Accept

Table 9. Predicted results

Model	MAE	MAPE	RMSE	R^2	NSE	Equation	MAE	MAPE	RMSE	R^2	NSE
BP	12.06	8.50	15.76	0.9283	0.9164	Eqn (1)	38.48	26.55	44.94	0.7581	0.3203
GA-BP	8.47	6.17	11.60	0.9599	0.9546	Eqn (2)	43.98	30.09	50.30	0.7725	0.1484
ELM	11.35	8.11	15.09	0.9239	0.9234	Eqn (3)	49.44	33.57	56.16	0.7297	0.1123
RF	11.98	8.59	16.90	0.9111	0.9039	Eqn (4)	25.45	17.56	32.39	0.7455	0.6469
SVM	9.05	5.96	13.76	0.9404	0.9363	Eqn (6)	17.63	13.16	22.40	0.9077	0.8311

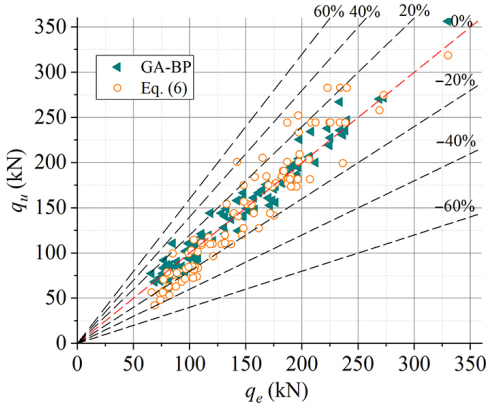

Figure 13. Comparison of GA-BP ANN model and Eqn (6)

6. Parameter analysis

6.1. Global sensitivity analysis based on Machine learning

The purpose of global sensitivity analysis (GSA) is to investigate the corresponding relationship between the model output response and the input parameters, and to provide guidance for choosing a more reasonable and effective solution to reduce the uncertainty of the model output response. The frequently used global sensitivity analysis index is the variance-based Sobol global sensitivity index, and it has been widely used in practical engineering problems. Sobol's method aims to identify important parameters and to assess the extent to which they affect the response of interest. Since the global sensitivity analysis method requires a large amount of data, this paper combines the GA-BP ANN model with GSA method to propose a new parametric analysis method, as shown in Figure 14.

The model is $Y = f(x)$, $x = [x_1, x_2, \cdots, x_k]$. x is the dataset, and the range of values for each parameter is determined in Table 2. $f(x)$ is as follows:

$$f(x) = f_0 + \sum_{i=1}^{k} f_i(x_i) + \sum_{1<i<j\le k} f_{i,j}(x_i, x_j) + \cdots + f_{1,2,\cdots,k}(x_1, x_2, \cdots, x_k), \quad (25)$$

where $\sum_{i=1}^{k} f_i(x_i)$ is the sum of the main effect functions, and $\sum_{1<i<j\le k} f_{i,j}(x_i, x_j) + \cdots + f_{1,2,\cdots,k}(x_1, x_2, \cdots, x_k)$ is the sum of all interactions. In addition, the equation must satisfy the features listed below:

$$\int_0^1 f_{i_1,i_2,\cdots,i_n}(x_{i1}, \cdots, x_{in})\mathrm{d}x_{ij} = 0, \ 1 \le j \le n. \quad (26)$$

The equation can be decomposed as follows:

$$f_0 = \int_0^1 f(x)\mathrm{d}x; \quad (27)$$

$$f_i(x_i) = \int_0^1 \cdots \int_0^1 f(x)\mathrm{d}x_{-i} - f_0; \quad (28)$$

$$f_{ij}(x_i, x_j) = \int_0^1 \cdots \int_0^1 f(x)\mathrm{d}x_{-(i,j)} - f_0 - f_i(x_i) - f_j(x_j), \quad (29)$$

where $f(x_i)$ is marginal effect. The total variance (V) is:

$$V = Var[f(x)] = \int_0^1 f^2(x)\mathrm{d}x - f_0^2 = E[f^2(x)] - E[h(x)]^2, \quad (30)$$

where $E[]$ is expected value, $Var[]$ is variance. The partial $(V_{i_1}, \cdots, i_s)$ is:

$$V_{i_1}, \cdots, i_s = \int_0^1 \cdots \int_0^1 h^2_{i_1,\ldots,i_n}(x_{i_1}, \ldots, x_{i_s})\mathrm{d}x_{i_1} \cdots \mathrm{d}x_{i_n}, \quad 1 \le i_s \le k; \quad (31)$$

$$V = \sum_{i=1}^{n} V_i + \sum_{1\le i<j\le n} V_{ij} + \ldots + V_{1,2,\ldots,n}. \quad (32)$$

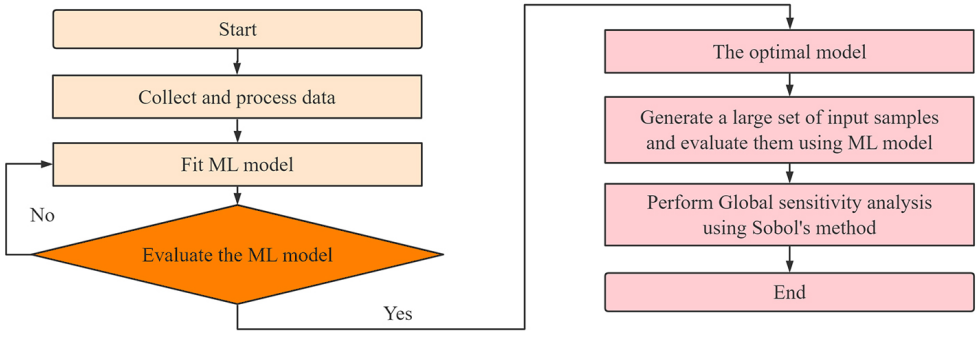

Figure 14. Calculation process of parametric analysis method

The estimation of the Sobol indices including first-order (S_i) and total-order (S_{Ti}) is:

$$S_i = \frac{V_{i_1,\cdots,i_s}}{V} = \frac{Var(f(x)) - E_x\left[Var_{x_{\sim i}}\left(f(x) \mid x_i\right)\right]}{Var(f(x))}; \quad (33)$$

$$S_{Ti} = 1 - \frac{V_{\sim i}}{V} = \frac{Var_{x_{\sim j}}\left[E_{x_i}\left(f(x) \mid x_{\sim i}\right)\right]}{Var(f(x))}, \quad (34)$$

where $V_{\sim i}$ represents the variance of all input parameters except i-th, $x_{\sim i}$ is all but the i-th input factor. S_i is the main effect. S_{Ti} is all contributions of the input variable to the output variance.

6.2. Analysis of significance

Because of the superior predictive performance of GA-BP ANN model, the model was used to generate the 4000 data needed for the global sensitivity analysis. S_i and S_{Ti} of the factors were normalized to study the contribution of each variable, as shown in Figure 15. The stud diameter has the greatest influence on bearing capacity. The tensile strength, height of studs and the strength of concrete have similar effects on shear strength. The strength of the concrete has a higher impact on the shear performance of the structure than the strength of the studs. It should be noted that the number of studs has a relatively small effect on the shear strength since the shear strength of a single stud is calculated in this paper.

The bearing capacity under each parameter is predicted by GA-BP ANN model, as shown in Figure 16. As the stud diameter increases, the shear strength improves. The higher the concrete strength, the more pronounced the effect of stud diameter on strength. At constant stud diameter, the higher the stud, the higher the shear strength. As the stud diameter increases, the effect of stud height on shear strength decreases. When the stud diameter is small, the tensile strength of the stud affects the shear strength. Due to group shear effect, the number of studs has different effects on shear strength. The number of studs affects the shear strength differently for different stud diameters. Comparing the stud diameter with other variables, it can be seen that the effect of stud diameter on shear strength is greater than other variables.

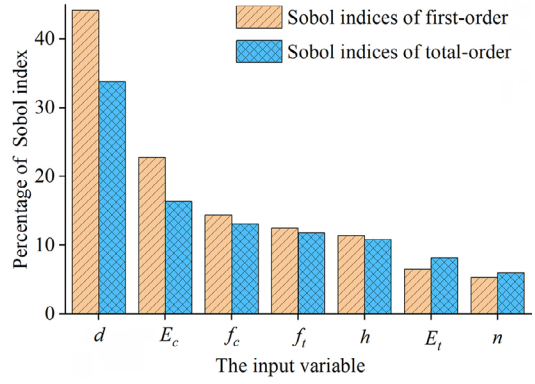

Figure 15. Analysis results of GSA method

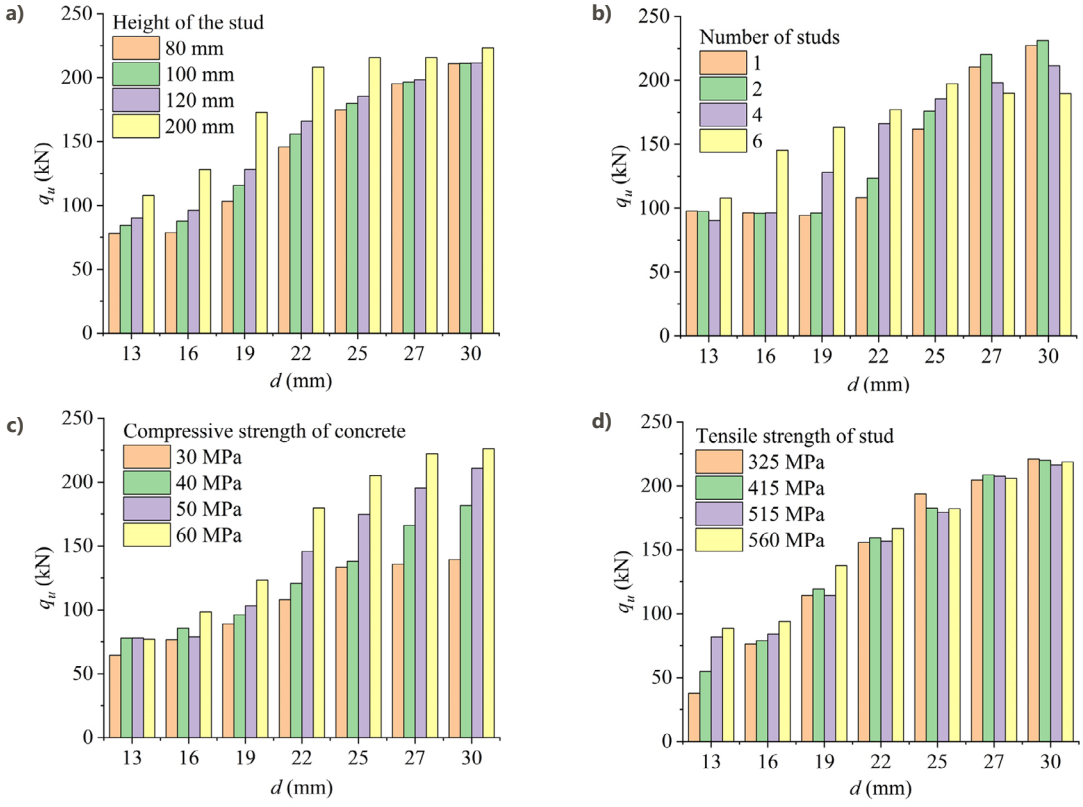

Figure 16. Prediction results with different input parameters: a – height of the stud; b – number of studs; c – compressive strength of concrete; d – tensile strength of the stud

7. Conclusions

In this paper, the prediction method of shear strength of stud connectors is established by machine learning (ML) model and global sensitivity analysis (GSA) method. The influencing factors of shear strength were determined based on empirical formulas and the prediction accuracies of different ML models were compared. Finally, the ML model was combined with the GSA method to investigate the influence of each parameter on shear strength. The main conclusions are as follows:

(1) Based on empirical equations, it is known that the number, height, diameter, tensile strength and elastic modulus of the studs affect the shear strength. In addition, the compressive strength and modulus of elasticity of concrete also affect the shear strength. Although different empirical equations have different input variables, all of them have stud diameters in their input variables and the shear strength increases with increasing stud diameter. The traditional empirical equations have mean errors greater than 10% and most of the calculated values are smaller than the experimental values.

(2) The prediction performance of different ML models including linear regression (LR1), ridge regression (RR), lasso regression (LR2), back-propagation artificial neural network (BP ANN), genetic algorithm optimized BP ANN (GA-BP ANN), extreme learning machines (ELM), random forests (RF), and support vector machines (SVM) model were evaluated. According to the judgement criteria, it can be seen that the MAPE values of all ML models are less than 10, indicating that all ML models have high prediction performance. In terms of R^2 and NSE, the GA-BP model, BP model, RF model, ELM model and SVM model all have values greater than 0.9, which is a very good performance.

(3) The GA-BP ANN model has the best prediction performance, with NSE and R^2 values closest to 1, and MAE and RMSE values lowest among ML models. The SVM model also has excellent prediction performance, with MAPE values lowest among all models, NSE and R^2 values only lower than the GA-BP ANN model, and MAE and RMSE values only higher than the GA-BP ANN model. The traditional empirical equations have higher MAE, RMSE and MAPE values than the ML model, and lower NSE and R2 values than the ML model, so the prediction performance is much lower than the ML model.

(4) Based on the GA-BP ANN model and the global sensitivity analysis (GSA) method, a new parameter importance analysis method was developed to quantify the extent of each variable's effect on shear strength. The results show that stud diameter has the greatest influence on shear strength. The strength of the concrete has a higher impact on the shear performance of the structure than the strength of the studs. It should be noted that the number of studs has a relatively small effect on the shear strength since the shear strength of a single stud is calculated in this paper.

(5) The limitation of this study is that only the compressive strength and modulus of elasticity of normal strength concrete were used to predict the strength of the connectors. With the continuous development of technology, materials such as ultra-high performance concrete are being used more and more widely in construction projects, so future research will need to expand the database to investigate the effect of the type of concrete on the strength. In addition, this paper did not consider combining input variables to establish new input variables and did not determine the effect of the combined variables on strength.

Acknowledgements

The authors would like to thank the members of the CSU 1004 office for their selfless help and useful suggestions.

Author contributions

Guorui Sun conceived the study and were responsible for the design and development of the data analysis. Jiayuan Kang were responsible for data collection and analysis. Jun Shi were responsible for data interpretation. Guorui Sun wrote the first draft of the article.

Disclosure statement

The authors declare no conflict of interest.

References

Allahyari, H., Nikbin, I., Rahimi, S., & Heidarpour, A. (2018). A new approach to determine strength of Perfobond rib shear connector in steel-concrete composite structures by employing neural network. *Engineering Structures*, *157*, 235–249. https://doi.org/10.1016/j.engstruct.2017.12.007

American Association of State Highway and Transportation Officials. (2017). *AASHTO LRFD bridge design specifi*cations (AASHTO LRFDUS-2017) (8th ed.). Washington, DC, USA.

Barjouei, H. S., Ghorbani, H., Mohamadian, N., Wood, D. A., Davoodi, S., Moghadasi, J., & Saberi, H. (2021). Prediction performance advantages of deep machine learning algorithms for two-phase flow rates through wellhead chokes. *Journal of Petroleum Exploration and Production*, *11*, 1233–1261. https://doi.org/10.1007/s13202-021-01087-4

Bayram, S., & Çıtakoğlu, H. (2023). Modeling monthly reference evapotranspiration process in Turkey: Application of machine learning methods. *Environmental Monitoring and Assessment*, *195*(1), Article 67. https://doi.org/10.1007/s10661-022-10662-z

Bernus, A., Ottlé, C., & Raoult, N. (2021). Variance based sensitivity analysis of FLake lake model for global land surface modeling. *Journal of Geophysical Research - Atmospheres*, *126*, Article e2019JD031928. https://doi.org/10.1029/2019JD031928

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chahnasir, E. S., Zandi, Y., Shariati, M., Dehghani, E., Toghroli, A., Mohamed, E. T., Shariati, A., Safa, M., Wakil, K., & Khorami, M. (2018). Application of support vector machine with firefly algorithm for investigation of the factors affecting the shear strength of angle shear connectors. *Smart Structures and Systems*, *22*(4), 413–424.

Citakoglu, H., & Demir, V. (2023). Developing numerical equality to regional intensity-duration-frequency curves using evolutionary algorithms and multi-gene genetic programming. *Acta Geophysica*, *71*(1), 469–488. https://doi.org/10.1007/s11600-022-00883-8

Coşkun, Ö., & Citakoglu, H. (2023). Prediction of the standardized precipitation index based on the long short-term memory and empirical mode decomposition-extreme learning machine models: The Case of Sakarya, Türkiye. *Physics and Chemistry of the Earth*, *131*, Article 103418. https://doi.org/10.1016/j.pce.2023.103418

Ding, F., Ni, M., Gong, Y., Yu, Z., Zhou, Z., & Zhou, L. (2014). Experimental study on slip behavior and calculation of shear bearing capacity for shear stud connectors. *Journal of Building Structures*, *35*(9), 98–106 (in Chinese).

Ding, F., Yin, G., Wang, H., Wang, L., & Guo, Q. (2017). Static behavior of stud connectors in bi-direction push-off tests. *Thin-Walled Structures*, *120*, 307–318. https://doi.org/10.1016/j.tws.2017.09.011

Ding, J., Li, Y., Xing, W., Ren, P., & Yuan, C. (2021). Mechanical properties and engineering application of single-span steel-concrete double-sided composite beams. *Journal of Building Engineering*, *1*, Article 102644. https://doi.org/10.1016/j.jobe.2021.102644

European Committee for Standardization. (1994). *Eurocode 4: Design of composite steel and concrete structures* (EN 1994-1-1). Brussels, Belgium.

Farouk, A. I. B., Zhu, J. S., & Gu, Y. H. (2022). Finite element analysis of the shear performance of box-groove interface of ultra-high-performance concrete (UHPC)-normal strength concrete (NSC) composite girder. *Innovative Infrastructure Solutions*, *7*, Article 212. https://doi.org/10.1007/s41062-022-00815-x

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

Garzón-Roca, J., Marco, C. O., & Adam, J. M. (2013). Compressive strength of masonry made of clay bricks and cement mortar: Estimation based on Neural Networks and Fuzzy Logic. *Engineering Structures*, *48*, 21–27. https://doi.org/10.1016/j.engstruct.2012.09.029

Ghorbani, H., Wood, D. A., Choubineh, A., Tatar, A., Abarghoyi, P. G., Madani, M., & Mohamadian, N. (2020). Prediction of oil flow rate through an orifice flow meter: Artificial intelligence alternatives compared. *Petroleum*, *6*, 404–419. https://doi.org/10.1016/j.petlm.2018.09.003

Gu, J., Liu, D., Deng, W., & Zhang, J. (2019). Experimental study on the shear resistance of a comb-type perfobond rib shear connector. *Journal of Constructional Steel Research*, *158*, 279–289. https://doi.org/10.1016/j.jcsr.2019.03.032

Guan, C., Duan, Y. Z., Zhai, J. Q., & Han, D. (2019). Hydraulic dynamics in split fuel injection on a common rail system and their artificial neural network prediction. *Fuel*, *255*, Article 115792. https://doi.org/10.1016/j.fuel.2019.115792

Guo, H. Y., Dong, Y., Bastidas-Arteaga, E., & Gu, X. L. (2021). Probabilistic failure analysis, performance assessment, and sensitivity analysis of corroded reinforced concrete structures. *Engineering Failure Analysis*, *124*, Article 105328. https://doi.org/10.1016/j.engfailanal.2021.105328

Hossain, K., Gladson, L., & Anwar, M. (2017). Modeling shear strength of medium-to ultra-high-strength steel fiber-reinforced concrete beams using artificial neural network. *Neural Computing & Applications*, *28*(1), 1119–1130. https://doi.org/1007/s00521-016-2417-2

Hu, Y. Q., Qiu, M. H., Chen, L. L., Zhong, R., & Wang, J. (2021). Experimental and analytical study of the shear strength and stiffness of studs embedded in high strength concrete. *Engineering Structures*, *236*, Article 111792. https://doi.org/10.1016/j.engstruct.2020.111792

Khalaf, J., Majeed, A., Aldlemy, M., Ali, Z., & Yaseen, Z. (2021). Hybridized deep learning model for perfobond rib shear strength connector prediction. *Complexity*, *2021*, Article 6611885. https://doi.org/10.1155/2021/6611885

Kim, K., Han, O., Heo, W., & Kim, S. (2020). Behavior of Y-type perfobond rib shear connection under different cyclic loading conditions. *Structures*, *26*, 562–571. https://doi.org/10.1016/j.istruc.2020.04.053

Luo, Y., Hoki, K., Hayashi, K., & Nakashima, M. (2016). Behavior and strength of headed stud-SFRCC shear connection. I: Experimental study. *Journal of Structural Engineering*, *142*(2), Article 4015112. https://doi.org/10.1061/(ASCE)ST.1943-541X.0001363

Ministry of Housing and Urban-Rural Development of the People's of China. (2013). *Code for design of steel and concrete composite bridges* (GB 50917-2013). Beijing, China (in Chinese).

Ministry of Housing and Urban-Rural Development of the People's of China. (2017). *Standard for design of steel structures* (GB 50017-2017). Beijing, China (in Chinese).

Özbayrak, A., Ali, M. K., & Çıtakoğlu, H. (2023). Buckling load estimation using multiple linear regression analysis and multigene genetic programming method in cantilever beams with transverse stiffeners. *Arabian Journal for Science and Engineering*, *48*(4), 5347–5370. https://doi.org/10.1007/s13369-022-07445-6

Pianosi, F., Beven, K., Freer, J., Hall, J. W., Rougier, J., Stephenson, D. B., & Wagener, T. (2016). Sensitivity analysis of environmental models: A systematic review with practical workflow. *Environmental Modelling & Software*, *79*, 214–232. https://doi.org/10.1016/j.envsoft.2016.02.008

Safa, M., Shariati, M., Ibrahim, Z., Toghroli, A., & Petkovic, D. (2016). Potential of adaptive neuro fuzzy inference system for evaluating the factors affecting steel-concrete composite beam's shear strength. *Steel and Composite Structures*, *21*(3), 679–688. https://doi.org/10.12989/scs.2016.21.3.679

Sedghi, Y., Zandi, Y., Shariati, M., Ahmadi, E., Azar, V. M., Toghroli, A., Safa, M., Mohamad, E. T., Khorami, M., & Wakil, K. (2018). Application of ANFIS technique on performance of C and L shaped angle shear connectors. *Smart Structures and Systems*, *22*(3), 335–340. https://doi.org/10.12989/sss.2018.22.3.335

Setvati, M. R., & Hicks, S. J. (2022). Machine learning models for predicting resistance of headed studs embedded in concrete. *Engineering Structures*, *254*, Article 113803. https://doi.org/10.1016/j.engstruct.2021.113803

Shim, C. S., Lee, P. G., & Yoon, T. Y. (2004). Static behavior of large stud shear connectors. *Engineering Structures*, *26*(12), 1853–1860. https://doi.org/10.1016/j.engstruct.2004.07.011

Shen, Y. W., Yap, K. S., & Li, X. (2020). A new probabilistic output constrained optimization extreme learning machine. *IEEE Access*, *8*, 28934–28946. https://doi.org/10.1109/ACCESS.2020.2971012

Slater, E., Moni, M., & Alam, M. (2012). Predicting the shear strength of steel fiber reinforced concrete beams. *Construction and Building Materials*, *26*(1), 423–436. https://doi.org/10.1016/j.conbuildmat.2011.06.042

Soroush, Z., Brian, T., & Abdollah, S. (2020). Significant variables affecting the performance of concrete panels impacted by wind-borne projectiles: A global sensitivity analysis. *International Journal of Impact Engineering*, *144*, Article 03650. https://doi.org/10.1016/j.ijimpeng.2020.103650

Sobol, I. M. (1993). Sensitivity estimates for nonlinear mathematical models. *Mathematical and Computer Modelling*, *1*(4), 407–414.

Tm, A., Jd, C., & Bra, B. (2019). Shear resistance of headed shear studs welded on welded plates in composite floors. *Engineering Structures*, *197*, Article 109412. https://doi.org/10.1016/j.engstruct.2019.109412

Tzuc, O. M., Gamboa, O. R., Rosel, R. A., Poot, M. C., Edelman, H., Torres, M. J., & Bassam, A. (2021). Modeling of hygrothermal behavior for green facade's concrete wall exposed to nordic climate using artificial intelligence and global sensitivity analysis. *Journal of Building Engineering*, *33*, Article 101625. https://doi.org/10.1016/j.jobe.2020.101625

Vigneri, V., Odenbreit, C., & Romero-Gu, Z. A. (2021). Numerical study on design rules for minimum degree of shear connection in propped steel-concrete composite beams. *Engineering Structures*, *241*(4), Article 112466. https://doi.org/10.1016/j.engstruct.2021.112466

Wang, Q., & Liu, Y. (2013). Experimental study of shear capacity of stud connector. *Journal of Tongji University* (*Natural Science*), *41*(5), 659–663 (in Chinese).

Wang, J., Guo, J., Jia, L., Chen, S., & Dong, Y. (2017). Push-out tests of demountable headed stud shear connectors in steel-UHPC composite structures. *Composite Structures*, *170*, 69–79. https://doi.org/10.1016/j.compstruct.2017.03.004

Wang, J., Qi, J., Tong, T., Xu, Q., & Xiu, H. (2019). Static behavior of large stud shear connectors in steel-UHPC composite structures. *Engineering Structures*, *178*, 534–542. https://doi.org/10.1016/j.engstruct.2018.07.058

Wang, J., Zhang, A., & Wang, W. (2020). Effects of stud height on shear behavior of stud connectors. *Journal of Zhejiang University* (*Engineering Science*), *54*(11), 2076–2084 (in Chinese).

Wu, F., Tang, W., Xue, C., Sun, G., & Zhang, H. (2021). Experimental investigation on the static performance of stud connectors in steel-HSFRC composite beams. *Materials*, *14*(11), Article 2744. https://doi.org/10.3390/ma14112744

Xue, W., Ding, M., Wang, H., & Luo, Z. (2008). Static behavior and theoretical model of stud shear connectors. *Journal of Bridge Engineering*, *13*(6), 623–634. https://doi.org/10.1061/(ASCE)1084-0702(2008)13:6(623)

Xue, D., Liu, Y., Zhen, Y., & He, J. (2012). Static behavior of multi-stud shear connectors for steel-concrete composite bridge. *Journal of Constructional Steel Research*, *74*(8), 1–7. https://doi.org/10.1016/j.jcsr.2011.09.017

Yang, X., & Wen, W. (2018). Ridge and Lasso regression models for cross-version defect prediction. *IEEE Transactions on Reliability*, *67*(3), 885–896. https://doi.org/10.1109/TR.2018.2847353

Yang, Y., Liang, W., Yang, Q., & Cheng, Y. (2021). Flexural behavior of web embedded steel-concrete composite beam. *Engineering Structures*, *240*, Article 112345. https://doi.org/10.1016/j.engstruct.2021.112345

Yaseen, Z., Tran, M., Kim, S., Bakhshpoori, T., & Deo, R. (2018). Shear strength prediction of steel fiber reinforced concrete beam using hybrid intelligence models: A new approach. *Engineering Structures*, *177*, 244–255. https://doi.org/10.1016/j.engstruct.2018.09.074

Yosri, A. M., Farouk, A. I. B., Haruna, S. I., Deifalla, A. F., & Shaaban, W. M. (2023) Sensitivity and robustness analysis of adaptive neuro-fuzzy inference system (ANFIS) for shear strength prediction of stud connectors in concrete. *Case Studies in Construction Materials*, *18*, Article e02096. https://doi.org/10.1016/j.cscm.2023.e02096

Yu, Z., Shi, W., & Kuang, Y. (2014). Experimental study on mechanical properties of corroded stud. *Journal of Central South University* (*Science and Technology*), *45*(1), 249–255 (in Chinese).

Zhang, Y., Liu, A., Chen, B., Zhang, J., Pi, Y., & Bradford, M. (2020). Experimental and numerical study of shear connection in composite beams of steel and steel-fibre reinforced concrete. *Engineering Structures*, *215*, Article 110707. https://doi.org/10.1016/j.engstruct.2020.110707

Zhang, F., Wang, C., Zou, X., Yang, W., Chen, D., Wang, Q., & Wang, L. (2023). Prediction of the shear resistance of headed studs embedded in precast Steel-Concrete structures based on an interpretable machine learning method. *Buildings*, *13*(2), Article 496. https://doi.org/10.3390/buildings13020496

Zhu, L., Wang, J. J., Li, X., Tang, L., & Yu, B. Y. (2020). Experimental and numerical study of curved SFRC and ECC composite beams with various connectors. *Thin-Walled Structures*, *155*, Article 106938. https://doi.org/10.1016/j.tws.2020.106938

Zouzou, Y., & Citakoglu, H. (2023). General and regional cross-station assessment of machine learning models for estimating reference evapotranspiration. *Acta Geophysica*, *71*(2), 927–947. https://doi.org/10.1007/s11600-022-00939-9