Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Combinatorial Press

Application of Least Squares Support Vector Regression Analysis to Green Concrete with Oil Palm Husk Aggregate

Jinshuai Lu[1], Shuhao Zhang[1], Jin Ma[1], Wenying You[1,✉]

[1] *Weifang Engineering Vocational College, Qingzhou, Shandong, 262500, China*

ABSTRACT
Least Squares Support Vector Regression (LSSVR) machine has the advantages of small sample, nonlinearity and high dimensionality, which can solve the problem of predicting the compressive strength of green concrete with oil palm shell aggregate. In this paper, the error sum of squares instead of the error sum is used as the objective function, IFFA is used to find the optimization of the kernel function parameters and penalty factors of LSSVR, and the PWLCM-based chaotic search is used to initialize the population, and ultimately the improved auricular fox algorithm is realized for the optimization of the least squares support vector regression algorithm, which makes it have strong fitting and generalization abilities, and significantly reduces the burden of computation, thus improving the Computational efficiency. Application of the designed combined algorithm for compressive strength prediction of concrete reveals that the R^2, MAPE and RMSE values obtained by this paper's model on the training dataset are 98.71%, 5.92% and 1.0823 MPa, respectively. The correlation coefficients predicted by the model are much closer to 1 as compared to that of the baseline model, which suggests that this paper's model possesses a superior generalization capability, making it more effective in dealing with complex and invisible data. The adopted method is practical and innovative, and is of guiding significance for practical engineering.

Keywords: least squares, support vector, auricular fox optimization, compressive strength

1. Introduction

With global industrialization and urbanization, the consumption of non-renewable energy and the emission of greenhouse gases (GHGs) have led to an increase in global temperatures and caused many environmental problems. Up to now, the average CO2 concentration in the earth's atmosphere has increased from 285 ppm to 417 ppm [1-3]. In order to keep the global temperature change within 1.5°C, China has pledged to peak its carbon emissions by 2030 and eventually achieve carbon neutrality by 2060. As one of the major CO2 emitting industries, the construction industry pays more

✉ Corresponding author.
E-mail address: ywy0536@163.com (W. You).
Received 05 March 2024; Revised 15 May 2024; Accepted 20 December 2024; Published Online 16 April 2025.
DOI: 10.61091/jcmcc127b-512

attention to green building, sustainable materials, carbon reducing materials and carbon reduction throughout the building life cycle [4-6].

Concrete is an important material for infrastructure construction, the most consumed material after water, and the most consumed man-made material in the world [7-8]. And CO2 emissions from the concrete industry have a huge impact on the environment. It has been calculated that most of the CO2 emissions in the concrete industry come from the production of cement and transportation of materials, and each ton of cement produced generates an equal amount of CO2 emissions [9-10]. The production of other raw materials in concrete, such as aggregates, admixtures and construction processes, also contributes to the high carbon emissions of the industry. In addition, the concrete industry is one of the major consumers of natural resources, and the increased production of concrete puts great pressure on the reserves of these natural resources [11-12]. With the promotion of energy saving and emission reduction policies in the construction industry, the engineering community is exploring sustainable development measures for concrete to reduce CO2 emissions in this industry, including the use of alternative materials such as waste materials, optimization of the concrete production process as well as the adoption of alternative energy sources, promotion of high-performance concrete technologies, and the development of green construction, among other methods [13-14].

In order to ensure the requirements of green concrete ratio [15-16], relevant designers should not only focus on the selection of concrete raw materials, but also analyze the impact of different ratios on concrete materials, improve the performance of concrete materials, improve the utilization rate of concrete materials, reduce resource waste, and lay the foundation for promoting the development of China's construction engineering industry [17-20].

In this paper, the improved auricular fox algorithm is used to optimize the least squares support vector regression algorithm and applied to the prediction of compressive strength of concrete. Firstly, the algorithm is modeled on the principle of least squares support vector regression machine, and then the LSSVR model is used to establish the component prediction model of the sequence, and the improved Auricular Fox algorithm is proposed to optimize the key hyper-parameters of the LSSVR model, and the predictions are added together to obtain the final prediction results. The designed algorithm is applied to an aggregate concrete example to analyze the model prediction performance and compare it with the best hyperparameters entered by different machine learning models such as ANN, SVM and SVR.

2. Method

From the point of view of intelligent statistical learning theory, the least support vector regression model (LSSVR) has great potential to reduce the simulation task and improve the fitting accuracy. RBF is chosen as the kernel function of LSSVR. When RBF is used to train LSSVR, the penalty factor and kernel function parameters are the key hyperparameters that determine the performance of the model. Therefore, in this paper, the LSSVR kernel parameters are optimized using the improved auricular fox method for the prediction of concrete performance.

2.1. Least Squares Support Vector Regression Machine

2.1.1. Statistical learning theory. Statistical learning theory is a combination of computer science, pattern recognition and applied statistics, and is a theory for studying machine learning in the case of small data sample sets, which includes:

Principle of empirical risk minimization: using computable empirical risk approximation to replace the incomputable expected risk, and finding the optimal estimation of expected risk by studying empirical risk minimization.

VC dimension theory: VC dimension represents the learning ability of a decision function candidate set, based on the minimization of empirical risk, VC dimension can be used as the basis for selecting the decision function candidate set.

Principle of structural risk minimization: in reducing the empirical risk, the candidate functions increase and the confidence interval increases.

In order to half-balance the different elements, these ten thousand facets should be taken into account to achieve the purpose of structural risk minimization.

The algorithm is realized as follows:

Suppose a training dataset containing i sample point:

$$S = \{(X_1, y_1), (X_2, y_2), \ldots (X_i, y_i)\}, i = 1, 2, \cdots N \tag{1}$$

where each sample point contains n observation of the input, i.e., $X_i = (x_i^1, x_i^2, \cdots x_i^n), n = N$; y_i is the output. The test dataset and the training dataset are generated independently and identically according to the joint probability distribution $P(X, y)$. By using the sample points of the training dataset to train the learning, the decision function f is obtained. Input into the test dataset X_i, the predicted value of y_i is obtained $\dot{y}_i = f(X_i)$. Introducing the loss function to evaluate the $\dot{y}_i$ compared to the size of the error of y_i, the squared loss function is more broadly applied, which is calculated by the formula:

$$L(y, \dot{y}) = (y - \dot{y})^2 \tag{2}$$

A larger value of $L(y, \dot{y})$ indicates a larger prediction error. The expected value of $L(y, \dot{y})$ is the expected risk:

$$R_{\exp}(f) = \int L(y, \dot{y}) dP(X, y) \tag{3}$$

In practice the joint probability distribution $P(X, y)$ is usually unknown, so it is not feasible to calculate the expected value of $L(y, \dot{y})$ directly. According to the law of large numbers, the expected risk can be replaced by the empirical risk using n sample point data:

$$R_{emp}(f) = \frac{1}{n} \sum_{i=1}^{n} L(y_i, \dot{y}_i) \tag{4}$$

When n takes a larger value, i.e., when the sample size is larger, R_{emp} converges to $R_{\exp}(f)$, and traditional learning methods, such as great likelihood estimation and least squares, use the principle of empirical risk minimization to find the optimal solution. However, unlike traditional mathematical problems, the sample points in the actual problem are often limited, especially when the sample size is small, the difference between R_{emp} and $R_{exp}(f)$ usually does not meet the accuracy requirements. The structural risk minimization (SRM) principle is used in this problem to improve the accuracy. The structural risk is defined as:

$$R_{sm}(f) = \frac{1}{n} \sum_{i=1}^{n} L(y_i, \dot{y}_i) + \lambda J(f) \tag{5}$$

where $\lambda J(f)$ is the regularization term and $\lambda \geq 0$ serves to balance the relationship between the regularization term and the empirical risk. The larger the regularization value, the higher the model complexity.

2.1.2. Support vector regression machine. Support vector regression machine (SVR) is an important branch of SVM in the field of regression. Unlike classification which searches for the interval-maximum hyperplane, the strategy of SVR is to find a hyperplane that is closest to the totality of all sample points [21]. SVR requires the introduction of an insensitive loss parameter ε, which is defined as a loss of 0

if the deviation value of the sample points from the hyperplane is less than ε. As in equation (1) sample dataset, the insensitive loss function is expressed as:

$$L(f(x_i), y_i) = |f(x_i) - y_i|_\kappa = \begin{cases} 0, & |f(x_i) - y_i| < \varepsilon \\ |f(x_i) - y_i| - \varepsilon, & |f(x_i) - y_i| < \varepsilon \end{cases} \tag{6}$$

where $f(x)$ is the closest hyperplane to the totality of all sample points to be found. Like SVM, SVR requires the introduction of a slack variable $\xi \geq 0$ to improve the fitting accuracy. In this way the solution problem of SVR is called:

$$\begin{gathered} \min_{\omega,b,\xi_i,\tilde{\xi}_i} \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}(\xi_i + \tilde{\xi}_i) \\ s.t.\ f(x_i) - y_i \leq \varepsilon + \xi_i \\ y_i - f(x_i) \leq \varepsilon + \xi_i \\ i = 1, 2, \cdots n \end{gathered} \tag{7}$$

By introducing the insensitive loss parameter ε and the relaxation variable ξ, the sample points are categorized into three classes.

1) Sample points whose distance from the hyperplane falls within the allowed range of the insensitive loss parameter ε and no loss is calculated. (2) The distance between the sample point and the hyperplane does not fall within the allowable range of the insensitive loss parameter ε, but the error from the range is less than ξ, and no loss is calculated. (3) Other sample points calculate the loss.

Same as SVM, the introduction of Lagrange multipliers can be used to solve the dyadic problem of Eq. (7)

$$\begin{gathered} \max_{a,a} \sum_{i,j=1}^{n}(\tilde{a}_i - a_i r)(\tilde{a}_j - a_j r)x_i^T x_j + \sum_{i=1}^{n}[y_i((\tilde{a}_i - a_i) - \varepsilon(\tilde{a}_i + a_i))] \\ s.t.\sum_{i=1}^{n}(\tilde{a}_i - a_i) = 0, 0 \leq a_i, \tilde{a}_i \leq C, i = 1, 2, \cdots n \end{gathered} \tag{8}$$

where $a = (a_1, \cdots a_n)^T$ and $\tilde{a} = (\tilde{a}_1, \cdots \tilde{a}_n)^T$ are the Lagrange multipliers.

Satisfying the Lagrangian function at the saddle point, the product of the constraints and the Lagrangian multipliers under the KKT (Karush-Kuhn-Tucher) condition is 0. Thus it can be derived that a_i and corresponds to the product of $\tilde{a}_i$ is 0. From this, the support vector can be derived, and further ω^* and b^* can be obtained, which results in the prediction function as:

$$f(x) = \sum_{i=1}^{n}(a_i^* - \tilde{a}_i^*)(x_i^T x) + b^* \tag{9}$$

After adapting the kernel function $K(x_i, x)$ to the specific problem is introduced, the prediction function is finally called:

$$f(x) = \sum_{i=1}^{n}(a_i^* - \tilde{a}_i^*)K(x_i, x) + b^* \tag{10}$$

2.1.3. Least Squares Support Vector Regressors. SVR is essentially a quadratic programming problem of Eq. (8), and when the sample size increases, the training complexity will increase dramatically, leading to an increase in computational time consumption. Least Squares Support Vector Regressor (LSSVR) changes the constraints into equational constraints, transforms the quadratic programming problem into solving a system of linear equations, and replaces the slack variable ξ with the error squared to solve the analytical solution of the parameters [22]. LSSVR reduces the

computational effort of the modeling, while improving the fitting accuracy of the regression problem with small samples.

For a given training dataset, the regression prediction function of LSSVR is:

$$f(x) = \omega^T x + b \tag{11}$$

The objective function of LSSVR after mapping the sample points to the higher dimensional space using the mapping function is:

$$\begin{gathered} \min_{\omega,b,\xi} \frac{1}{2} \| \omega \|^2 + \frac{1}{2} \gamma \sum_{i=1}^{n} \xi_i^2 \\ s.t.\, y_i = \omega^T \phi(x_i) + b + \xi_i \\ i = 1, 2, \cdots n \end{gathered} \tag{12}$$

where ξ_i is the error factor; γ is the penalty parameter. A large increase in γ will increase the penalty for high error sample points and also lead to overfitting phenomenon at the same time.

Lagrange multipliers are introduced to solve Eq. (12):

$$L(\omega, b, \xi, a) = \frac{1}{2} \| \omega \|^2 + \frac{1}{2} \gamma \sum_{i=1}^{n} \xi_i^2 - \sum_{i=1}^{n} a_i [\omega^{\mathrm{T}} \phi(x_i) + b + \xi_i - y_i] \tag{13}$$

where a is the Lagrange multiplier. The partial derivatives of the solution are solved separately for the Lagrangian function and the solution is obtained by making the partial derivatives zero:

$$\begin{cases} \omega = \sum_{i=1}^{n} a_i \phi(x_i), \\ \sum_{i=1}^{n} a_i = 0, \\ a_i = \gamma \xi_i, i = 1, 2, \cdots n, \\ \omega^T \phi(x_i) + b + \xi_i - y_i = 0, i = 1, 2, \cdots n. \end{cases} \tag{14}$$

Elimination elements ξ_i and ω are obtained:

$$\begin{bmatrix} 0 & 1_n^{\mathrm{T}} \\ 1_n & \Omega + \frac{1}{\gamma} I_n \end{bmatrix} \begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix} \tag{15}$$

where 1_n is the n-dimensional unit vector; I_n is the unit matrix of $n \times n$; y is the output vector; and Ω is the kernel matrix, i.e., $\Omega_{ij} = K(x_i, x_j)$.

The final regression prediction function is obtained:

$$f(x) = \sum_{i=1}^{n} a_i K(x_i, x) + b \tag{16}$$

The relationship between the junction temperature and the combined thermal parameters is linear in general and locally influenced by the environment. Taken together, the radial basis kernel function is selected for on-line detection of progressive junction temperature using LSSVR. In the LSSVR with radial basis kernel as the kernel function, the parameters to be selected mainly include the penalty parameter γ and the kernel function width parameter σ in the kernel function.

2.2. *Optimization of the Improved Auricular Fox Algorithm for Least Squares Support Vector Regression*

2.2.1. Auricular Fox Optimization Algorithm. Auricular Fox Optimization Algorithm (FFA) is a new group intelligence optimization algorithm [23]. The specific steps of the auricular fox optimization algorithm are as follows:

The positions of the auricular fox population members in FFA can be represented by the following matrix:

$$X=\begin{bmatrix}X_1\\ \vdots\\ X_i\\ \vdots\\ X_N\end{bmatrix}=\begin{bmatrix}x_{1,1} & \cdots & x_{1,j} & \cdots & x_{1,m}\\ \vdots & \ddots & \vdots & \iddots & \vdots\\ x_{i,1} & \cdots & x_{i,j} & \cdots & x_{i,m}\\ \vdots & \iddots & \vdots & \ddots & \vdots\\ x_{N,1} & \cdots & x_{N,j} & \cdots & x_{N,m}\end{bmatrix}_{N\times m} \tag{17}$$

where X is the population matrix of eared foxes; X_i is the i th fox in the population of range N.

The fitness values for all eared foxes can be represented by the following vector:

$$F=\begin{bmatrix}F_1\\ \vdots\\ F_i\\ \vdots\\ F_N\end{bmatrix}_{N\times 1}=\begin{bmatrix}F(X_1)\\ \vdots\\ F(X_i)\\ \vdots\\ F(X_N)\end{bmatrix}_{N\times 1} \tag{18}$$

where F is the vector of auricular fox adaptation values; F_i is the adaptation value of the i rd auricular fox.

In order to simulate the behavior of the eared fox in the excavation process, a neighborhood with radius R around the actual position is set up.The key point in FFA is that the position of the prey is randomly generated in the search space, and this way increases the exploration ability of FFA in the search space. The above process can be represented by equations (19) to (21).

$$x_{i,j}^{P1}=x_{i,j}+(2\cdot r-1)\cdot R_{i,j} \tag{19}$$

$$R_{i,j}=\alpha\cdot\left(1-\frac{t}{T}\right)\cdot x_{i,j} \tag{20}$$

$$X_i=\begin{cases}X_i^{P1}, F_i^{P1}<F_i\\ X_i, \quad else\end{cases} \tag{21}$$

where X_i^{P1} is the new state of the i nd eared fox proposed according to the first stage; $x_{i,j}$ is the position of the i th eared fox in the j th dimension; $R_{i,j}$ is the radius of the neighborhood of $x_{i,j}$; T is the total number of iterations, and α is a constant set to 0.2.

The second stage of the population update of the FFA can be expressed in equations (22) to (24).

$$X_{i,j}^{rand}: x_{i,j}^{rand}=x_{k,j}, k\in\{1,2,\cdots,N\}, i=1,2,\cdots,N \tag{22}$$

$$X_{i,j}^{P2}=\begin{cases}x_{i,j}+r\cdot(x_{i,j}^{rand}-I\cdot x_{i,j}), F_i^{rand}<F_i\\ x_{i,j}+r\cdot(x_{i,j}-x_{i,j}^{rand}), \quad else\end{cases} \tag{23}$$

$$X_i = \begin{cases} X_i^{P2}, F_i^{P2} < F_i \\ X_i, \quad else \end{cases} \tag{24}$$

where $X_{i,j}^{ramd}$ is the target location from which the i nd eared fox intends to escape, $x_{i,j}^{rand}$ is its jth dimension; $x_{i,j}^{P2}$ is the new location of the i th eared fox according to the proposed state in the second stage; and I is a number randomly equal to 1 or 2, and the parameter I affects the exploratory ability of the FFA.

2.2.2. Improvement of Auricular Fox Optimization Algorithm:

1) Chaos search based on PWLCM. PWLCM is a commonly used form of mapping in chaos theory, and in one-dimensional scenarios, PWLCM generates more uniformly distributed solution sets. Compared with rand random numbers, PWLCM can generate more uniformly distributed random numbers, which can accelerate the convergence of the algorithm. Therefore, PWLCM is used in this chapter to generate the initial population.The expression of PWLCM is shown in equation (25).

$$k(t^*+1) = \begin{cases} \dfrac{k(t^*)}{P} & ,0 \le k(t^*) < P \\ \dfrac{k(t^*)-P}{0.5-P} & , \quad P \le k(t^*) < 0.5 \\ \dfrac{1-P-k(t^*)}{0.5-P} & , \quad 0.5 \le k(t^*) < 1-P \\ \dfrac{1-k(t^*)}{P} & , \quad 1-P \le k(t^*) < 1 \end{cases} \tag{25}$$

where $k(t)$ is a random number between 0 and 1; $P = 0.25$ is taken to make the distribution of values generated by PWLCM more uniform.

2) Nonlinear adaptive incremental inertia weight factor. There were originally some problems in the FFA algorithm, and one of the main problems is that it is difficult to effectively balance the global and local searches throughout the algorithm. To solve this problem, nonlinear adaptive incremental inertia weight factors are introduced.The population of the FFA algorithm performs global search at the beginning and then gradually shifts to local search. In this process, the introduction of nonlinear adaptive incremental inertia weight factors ω allows the population to better adapt to different evolutionary periods. This is achieved by calculating ω based on the current context. Specifically, the value of ω is adjusted according to the current state of the algorithm in order to make the trade-off between global and local search more rational. The nonlinear adaptive incremental inertia weight factor ω is calculated as shown in equation (26):

$$\omega = \begin{cases} (\alpha + \beta \times b) \times \sin\left(\pi \times \dfrac{t}{T}\right)^3 & m\left(\dfrac{t}{6}\right) \ge 3 \\ 1 & m\left(\dfrac{t}{6}\right) < 3 \end{cases} \tag{26}$$

where α and β are the selection factors for the initial optimal eared fox and secondary eared fox; b is a random number from 0 to 1; and $m\left(\dfrac{t}{6}\right)$ is the remaining number of iterations divided by 6.

The nonlinear adaptive incremental inertia weighting factor ω is then introduced into the global search phase and local escape phase auricular fox position calculation formulas; this process is shown in Eqs. (27) and (28).

$$x_{i,j}^{P1} = \omega \cdot x_{i,j} + (2 \cdot r - 1) \cdot R_{i,j} \tag{27}$$

$$X_{i,j}^{P2} = \begin{cases} x_{i,j} + r \cdot (x_{i,j}^{rand} - I \cdot \omega \cdot x_{i,j}), F_i^{rand} < F_i \\ x_{i,j} + r \cdot (\omega \cdot x_{i,j} - x_{i,j}^{rand}), \quad else \end{cases} \tag{28}$$

The above formulation updates the position of the eared foxes by optimally taking into account the evolutionary differences of the eared fox populations during the evolutionary process, and adaptively assigns inertia weight factors of different sizes to them. When the inertia weight factor ω is increased, the global optimization ability of the algorithm is significantly enhanced. However, the local search capability is reduced and the accuracy of the solution may be lower. When the inertia weight factor ω decreases, the global optimization ability decreases while the local optimization ability increases and the accuracy of the solution is higher. The inertia weight factor helps the eared foxes to search at different convergence rates until they are close to the optimal value for the next iteration.

Under the premise that the overall search behavior remains unchanged, the inertia weight factor improves the search ability of the eared fox population at different stages. In the early stage of the population, having a strong global search ability, the local search ability was improved appropriately, which increased the search accuracy and convergence speed of the FFA. In the late stage of the population, with strong local search ability, the global search ability is appropriately improved, which effectively avoids the FFA from falling into local optimal solutions. These features satisfy the needs of FFA for global exploration ability and local exploitation ability in different evolutionary periods.

2.2.3. IFFA optimization of LSSVR models. When RBF is used to train LSSVR, the penalty factor ξ and kernel function parameter σ are the key model hyperparameters, which have an important impact on the prediction performance. If the kernel function parameters and penalty factors are not properly selected, the generalization ability and prediction performance of LSSVR will be seriously weakened, so it is necessary to optimize the kernel function parameters and penalty factors of LSSVR. In this chapter, IFFA is used to optimize the kernel function parameters and penalty factors of LSSVR, and the specific steps of the process are as follows:

1) Initialize the relevant parameters of the auricular fox algorithm, including the number of populations N, the maximum number of iterations $T_{\max}$, and the incremental inertia weight factor ω for nonlinear adaptive;

2) Initialize the population using PWLCM-based chaotic search:

3) Randomly generate prey locations;

4) The first stage of the auricular fox optimization algorithm, in which the auricular fox digs under the sand to find prey and then moves towards this area, calculates the new state of the auricular fox according to Eq. (27), and updates the auricular fox position according to Eq. (21);

5) The second phase of the auricular fox optimization algorithm, the strategy of the auricular fox to avoid the predator attack, evaluating its objective function according to Eq. (22), oxen into the target position of the auricular gallery fox to escape, calculating the new state of the auricular fox according to Eq. (28) and updating the position of the auricular fox according to Eq. (24):

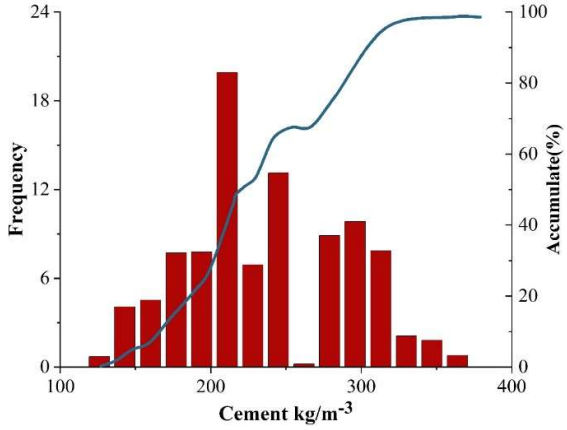
6) Determine whether the maximum number of iterations is reached, if it is satisfied, output the optimal solution, if not, return to step (3);

7) Input the optimal solution into the LSSVR model.

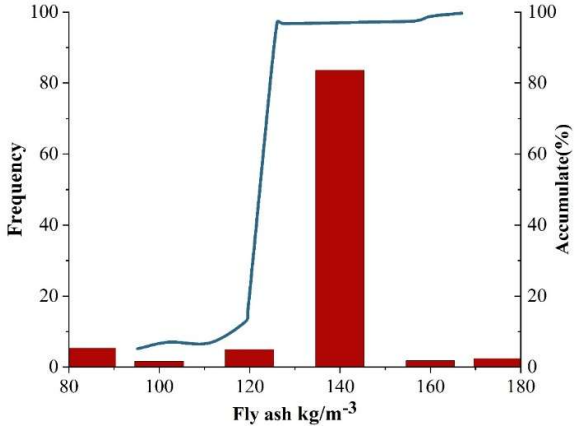
3. Results and Discussion

3.1. Calculated experimental data and model parameter optimization

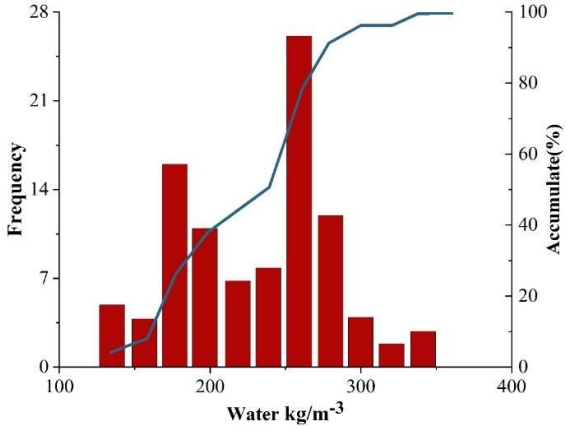
3.1.1. Statistical characterization of the data set. A total of 98 concrete cylindrical specimens were selected for this paper's calculations, each with a height of 200 mm and a diameter of 100 mm; the specimens were made in the same tray, demolded 24 h after the completion of the casting, and tested for compressive strength under standard curing conditions for specimens of different ages. For each concrete specimen, six components (cement, fly ash, water, water reducer, coarse aggregate and fine aggregate) and age were used as input variables in the dataset, and compressive strength was used as an output variable. The fineness of the fly ash in the material was <34%, the bulk stability was 0.05%, and the dry shrinkage was 0.06%; the percentage of components in the fly ash was 52.1% SiO_2, 4.26% Al_2O_3, 2.32% CaO, 26.9% Fe_2O_3, 0.11% Na_2O, 1.51% MgO, 1.48% K_2O, and 1% loss on burnout. The distribution of input and output variables is shown in Fig. 1, and (a) to (h) are the components. The statistical characteristics of the input and output variables are shown in Table 1, where the ratio of the training set to the test set is 80%:20%. From Table 1, it can be seen that the compressive strength varies greatly with different combinations of input variables, and the lowest to highest compressive strength interval in the test set is 10MPa~42MPa.



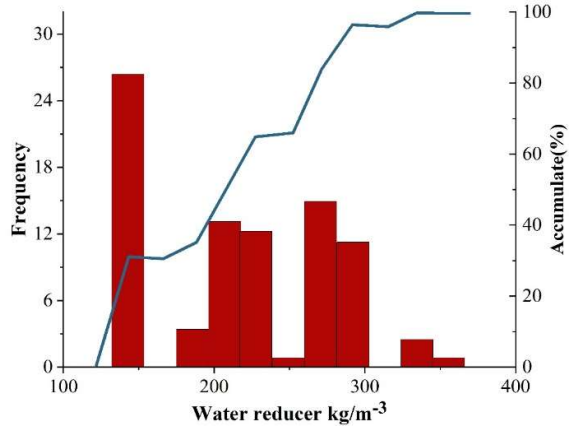
(a) Cement (kg/m^{-3})



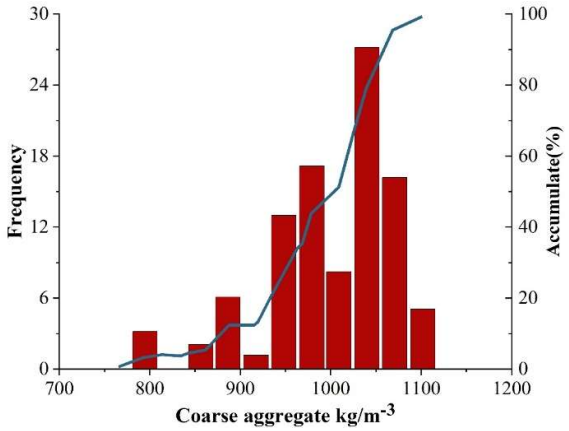
(b) Fly ash (kg/m^{-3})



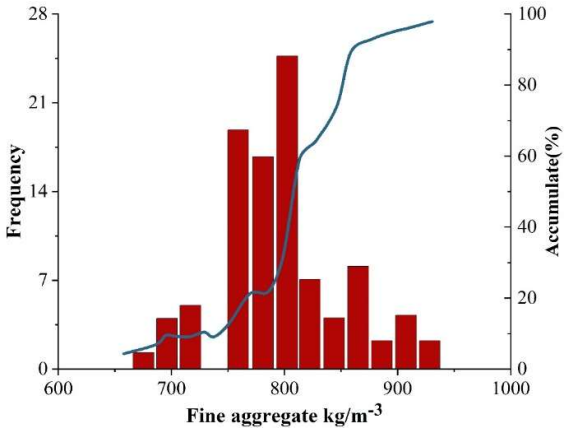
(c) Water (kg/m^{-3})



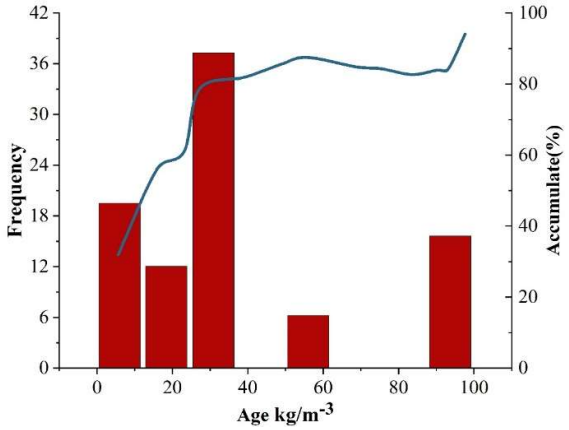
(d) Water reducer (kg/m^{-3})



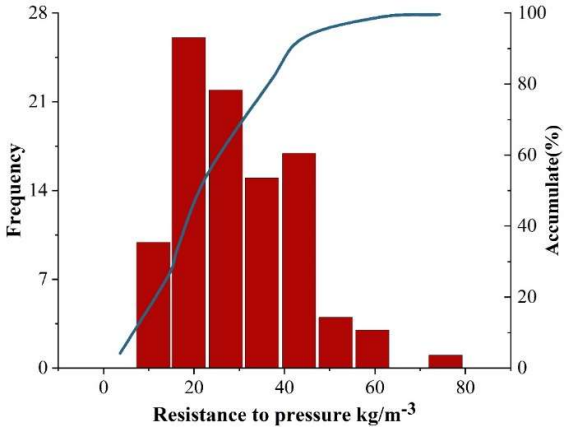
(e) Coarse aggregate (kg/m^{-3})



(f) Fine aggregate (kg/m^{-3})



(g) Age (kg/m^{-3})



(h) Resistance to pressure (kg/m^{-3})

Fig. 1. Input and output variable distribution

Table 1. Statistics of input and output variables

Data	Characteristic statistics	Cement	Fly ash	Water	Water reducer	Coarse aggregate	Fine aggregate	Age	Resistance to pressure
Training set	Max	377	165	214	19	1114	908	92	75
	Min	141	95	143	0	805	702	4	11
	SD	55	8	17	5	68	43	28	14
	mean	242	126	176	7	1013	801	35	32
Test set	Max	347	169	221	14	1117	834	91	42
	Min	148	97	159	0	803	691	4	10
	SD	57	12	16	6	78	44	26	8
	mean	236	138	179	7	984	775	33	26

The Pearson correlation coefficients between the input and output variables are shown in Figure 2. In the current dataset, the linear correlation between any of the input and output variables is weak, and none of them exceeds 0.8, which indicates that there is a complex nonlinear relationship between compressive strength and these input parameters. Therefore, the relationship between compressive strength and these parameters is difficult to be reflected by a clear explicit equation, which is the significance of the model in this paper.

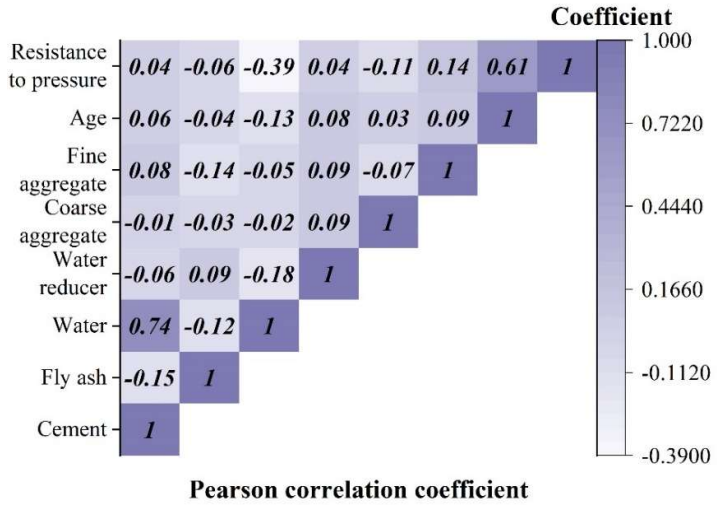


Fig. 2. Pearson correlation coefficient

3.1.2. Evaluation indicators. In order to describe and compare the performance of the models, five evaluation metrics are introduced, namely, linear correlation coefficient (R), mean square error (MSE), root mean square error (RMSE), mean absolute percentage error (MASE), and mean absolute precision error (MAPE).

3.2. Comparative analysis of forecast results

3.2.1. Model Prediction Performance and Validation Analysis. Table 2 shows the statistical results of the performance of the model in this paper and the baseline model. The results of the performance metrics are shown in Figures 3 and 4, where Figures 3 and 4 show the statistical results of the performance of the training and test datasets, respectively.

In the test set, this paper's model has the best performance with the highest R^2 value of 97.28% and the lowest RMSE and MAPE of 1.604 MPa and 9.08%, respectively. On the training dataset the model of this paper obtained R^2, MAPE and RMSE values of 98.71%, 5.92% and 1.0823 MPa respectively.This

indicates that the model of this paper has a superior generalization ability which makes it more effective in dealing with complex invisible data.

Table 2. Performance evaluation of all models

Data set	Model	R^2	RMSE	MAE	MAPE	A20-index
Training set	ANN	0.8563	1.9895	1.3645	0.2025	0.7922
	SVM	0.874	1.7935	1.3015	0.1358	0.8255
	SVR	0.9341	1.3865	0.9718	0.0574	0.8797
	This model	0.9871	1.0823	0.6298	0.0592	0.9674
Test set	ANN	0.8115	2.2432	1.9058	0.2475	0.7413
	SVM	0.8566	2.1092	1.5373	0.174	0.8147
	SVR	0.9309	1.9885	1.3704	0.1187	0.8557
	This model	0.9728	1.604	1.2016	0.0908	0.8991

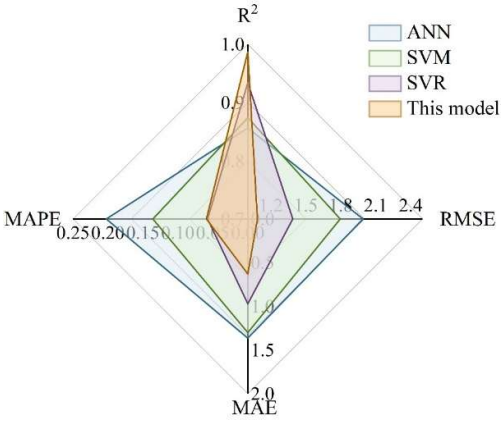


Fig. 3. Training set performance indicators

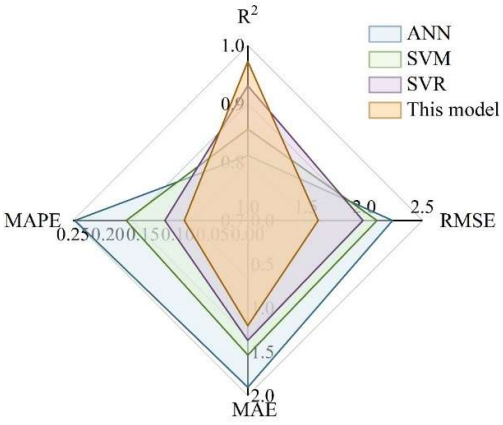
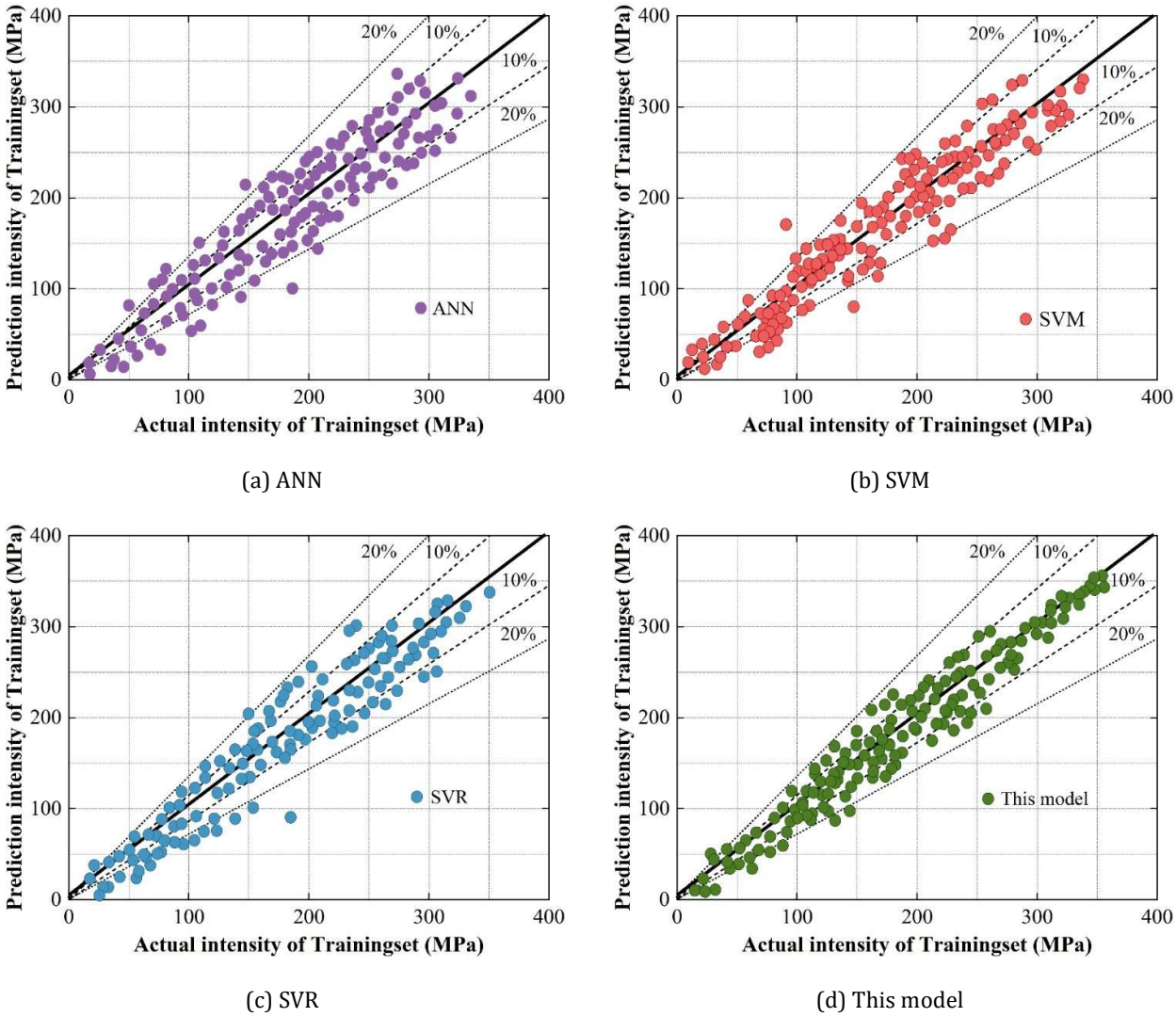


Fig. 4. Test set performance indicators

Figure 5 shows the regression plots for all models in the training phase. The horizontal axis of each plot represents the observed values in the training sample and the vertical axis represents the predicted values of the models. The black line in each plot is the line where the observed and predicted values agree 100%, and the other radial lines are the lines that differ from the black line by 10% and 20%. Therefore, if all the points are placed on the black line of the equation y = x, this means that the model is able to predict the actual values without any error.

As shown in the figure, the model in this paper not only has the highest R^2 value but also has the most similar equation to y=x. After this model, there are SVR, SVM and ANN models, respectively.

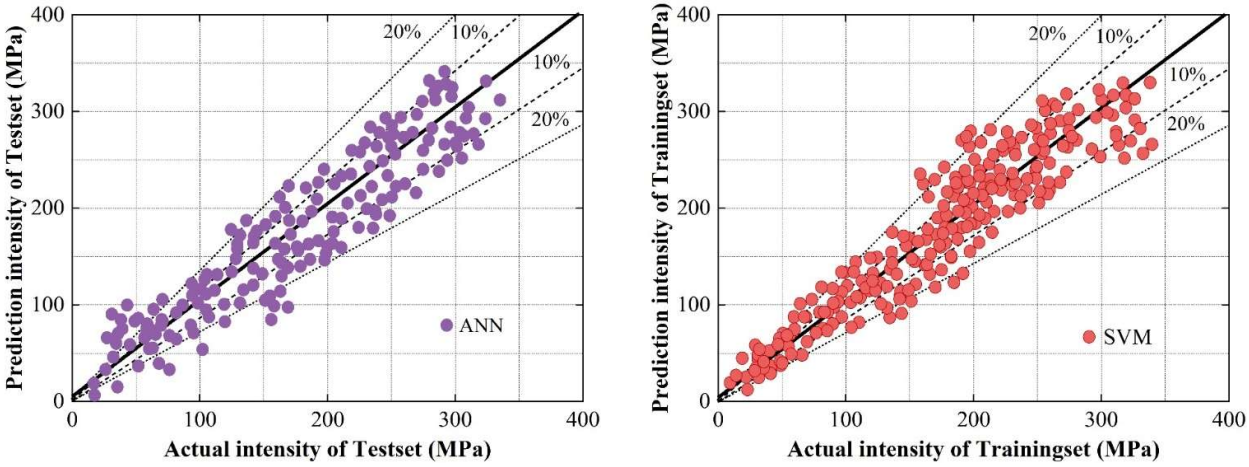


(a) ANN (b) SVM

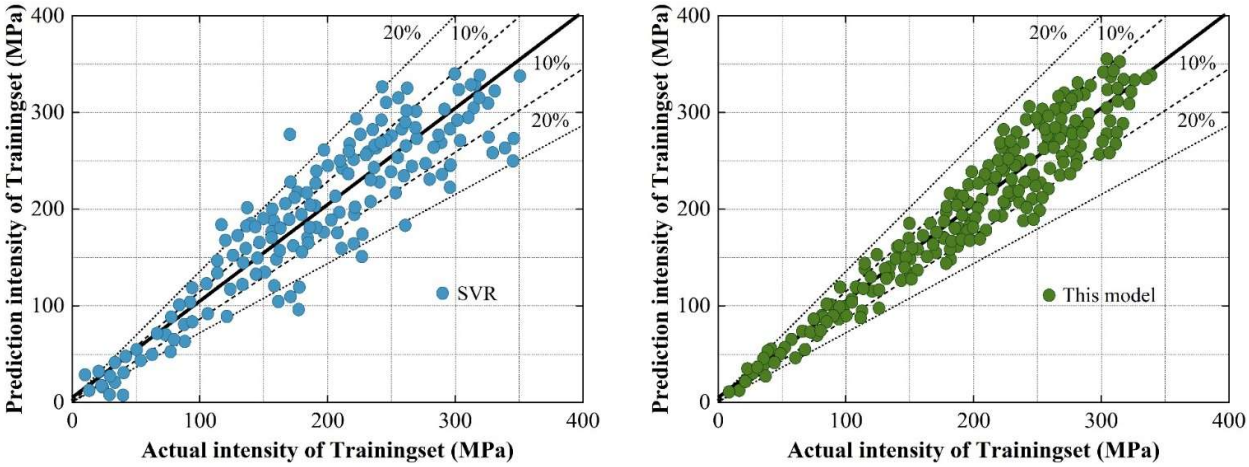
(c) SVR (d) This model

Fig. 5. Regression diagrams of all models during training phase

Figure 6 shows the regression plot of the model in the testing phase. The model in this paper can achieve better performance as it has the highest R^2 value and it has collected the data points in the graph more compactly along the black line (100% consistent line).The R^2 values are ranked after the model in this paper are SVR, SVM and ANN models respectively.



(a) ANN (b) SVM



(c) SVR (d) This model

Fig. 6. Regression diagrams of all models during training phase

3.2.2. Model Hyperparameter Analysis. In order to ensure that the hyperparameter adjustment process is fair and unbiased, this paper adopts a unified treatment, i.e., the adjustment is carried out within the same hyperparameter space, and the results are shown in Table 3. Table 4 shows the results of the optimal hyperparameter analysis for different machine learning models such as ANN, SVM and SVR. This process is mainly carried out through an iterative trial-and-error approach, where the hyperparameters of each model are finely tuned, changing only one parameter at a time while keeping the other parameters unchanged, in order to compute the prediction accuracy of the model under the current parameter configuration. By this method, the parameter settings are gradually optimized to achieve the highest prediction accuracy for each model. Ultimately, it is shown that the model in this paper has the best performance among the models, with a prediction accuracy of 0.8926 for the SVR model, followed by the SVM model and the ANN model in that order. A detailed record of the optimal hyperparameter configuration for each model can be found in Table 4.

Table 3. The hyperparameter spaces of all models

	ANN	SVM	SVR	This model
Number of trees	[1,600]	[1,400]	[1,400]	[1,300]
Learning rate	[1,10]	[1,12]	[1,12]	[1,12]
Max features	[0.0,10]	[0.0,10]	[0.0,10]	[0.0,10]
Min samples splid	0.001-0.10]	[0.001-0.10]	[0.001-0.10]	[0.001-0.10]
Min samples leaf	0.001-0.10]	[0.001-0.10]	[0.001-0.10]	[0.001-0.10]
Max depth	[1,10]	[1,6]	[1,6]	[1,10]
Data used	10Fold	CV	10Fold	CV
Performance index	R^2	R^2	R^2	R^2

Table 4. The optimal hyperparameters of all models

	ANN	SVM	SVR	This model
Number of trees	[1,600]	240	220	180
Learning rate	0.08	0.11	0.12	0.13
Max features	3	2	2	5
Min samples splid	4.3	4.5	4.9	5.6
Min samples leaf	1.65	1.71	1.69	1.92
Max depth	3	4	8	9
Performance index	0.7619	0.7972	0.8105	0.8926

3.2.3. Model interpretation. In this chapter, the model prediction results are interpreted using the SHAP method.SHAP is a method for interpreting the output of machine learning models. It is based on the Shapley value in cooperative game theory, which is able to assign a contribution value to each feature to explain its impact on the model prediction results.The core idea of the SHAP method is to calculate the marginal contribution of each feature in different combinations of features, and to obtain the Shapley value of each feature by weighting the average of all possible combinations.

According to the established model, the contribution analysis of prediction results based on SHAP values can be divided into two levels. Figure 7 shows the interpretation of the global level of the model in this paper. The results show that cement content has the most significant effect on the mechanical properties of concrete and slag ash has the least effect on the mechanical properties.

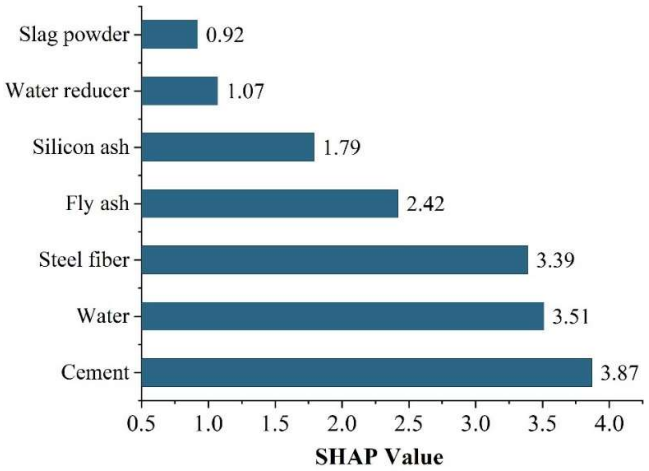


Fig. 7. Feature importance

4. Conclusion

In this paper, the least squares support vector regression algorithm is optimized using the improved auricular fox algorithm, and the designed combinatorial algorithm is applied to the selected concrete compressive strength examples for analysis. The research conclusions obtained are as follows:

1) The model of this paper has the best performance in the test set comparison, with the highest R^2 value of 97.28%, and both RMSE and MAPE are the lowest among the models, with 1.604 MPa and 9.08%, respectively.

2) The convergence of this paper's model in the regression plot is excellent, not only with the highest R^2 value, but also with the most similar equation to y=x, which can be regarded as the optimal performance for the prediction of the degree of concrete compressive strength.

3) Cement content has the most significant effect on the mechanical properties of concrete, reaching 3.87%, and slag ash has the least effect on the mechanical properties, only 0.92%.

References

[1] Hertwich, E. G., Ali, S., Ciacci, L., Fishman, T., Heeren, N., Masanet, E., ... & Wolfram, P. (2019). Material efficiency strategies to reducing greenhouse gas emissions associated with buildings, vehicles, and electronics—a review. Environmental Research Letters, 14(4), 043004.

[2] Tan, X., Lai, H., Gu, B., Zeng, Y., & Li, H. (2018). Carbon emission and abatement potential outlook in China's building sector through 2050. Energy Policy, 118, 429-439.

[3] Ahmed Ali, K., Ahmad, M. I., & Yusup, Y. (2020). Issues, impacts, and mitigations of carbon dioxide emissions in the building sector. Sustainability, 12(18), 7427.

[4] Teng, Y., Li, K., Pan, W., & Ng, T. (2018). Reducing building life cycle carbon emissions through prefabrication: Evidence from and gaps in empirical studies. Building and Environment, 132, 125-136.

[5] Wang, Q., Guo, W., Xu, X., Deng, R., Ding, X., & Chen, T. (2023). Analysis of carbon emission reduction paths for the production of prefabricated building components based on evolutionary game theory. Buildings, 13(6), 1557.

[6] Zhou, N., Khanna, N., Feng, W., Ke, J., & Levine, M. (2018). Scenarios of energy efficiency and CO2 emissions reduction potential in the buildings sector in China to year 2050. Nature Energy, 3(11), 978-984.

[7] Liew, K. M., Sojobi, A. O., & Zhang, L. W. (2017). Green concrete: Prospects and challenges. Construction and building materials, 156, 1063-1095.

[8] TANASH, A. O. A. A. R., & Muthusamy, K. (2022). Concrete industry, environment issue, and green concrete: a review. Construction, 2(1), 01-09.

[9] Sivakrishna, A., Adesina, A., Awoyera, P. O., & Kumar, K. R. (2020). Green concrete: A review of recent developments. Materials Today: Proceedings, 27, 54-58.

[10] Vishwakarma, V., & Ramachandran, D. (2018). Green Concrete mix using solid waste and nanoparticles as alternatives–A review. Construction and Building Materials, 162, 96-103.

[11] Hashmi, A. F., Khan, M. S., Bilal, M., Shariq, M., & Baqi, A. (2022). Green concrete: an eco-friendly alternative to the OPC concrete. Construction, 2(2), 93-103.

[12] Golewski, G. L. (2018). Green concrete composite incorporating fly ash with high strength and fracture toughness. Journal of cleaner production, 172, 218-226.

[13] Thomas, B. S. (2018). Green concrete partially comprised of rice husk ash as a supplementary cementitious material–A comprehensive review. Renewable and Sustainable Energy Reviews, 82, 3913-3923.

[14] Al-Hamrani, A., Kucukvar, M., Alnahhal, W., Mahdi, E., & Onat, N. C. (2021). Green concrete for a circular economy: A review on sustainability, durability, and structural properties. Materials, 14(2), 351.

[15] Makul, N., Fediuk, R., Amran, M., Zeyad, A. M., Murali, G., Vatin, N., ... & Vasilev, Y. (2021). Use of recycled concrete aggregates in production of green cement-based concrete composites: A review. Crystals, 11(3), 232.

[16] Zhu, K., Zhang, Y., Meng, X., & Ma, Z. (2022). Structural Assembly Analysis of Concrete Buildings with Intelligent Finite Element Analysis. Journal of Sensors, 2022(1), 6701021.

[17] Tzortzi, J. N., Hasbini, R. A., Ballottari, M., & Bellamoli, F. (2024). The Living Concrete Experiment: Cultivation of Photosynthetically Active Microalgal on Concrete Finish Blocks. Sustainability, 16(5), 2147.

[18] Bonde, A., Agrawal, P. S., Chaware, M., Dode, S., Gupta, S., & Ahmad, F. Evolution of green concrete in the building industry: A sustainable approach for nanomaterials: development and application. In Nanofluids Technology for Thermal Sciences and Engineering (pp. 292-328). CRC Press.

[19] Wang, N., Wang, Q., Xu, S., & Lei, L. (2021). Green fabrication of mechanically stable superhydrophobic concrete with anti-corrosion property. Journal of Cleaner Production, 312, 127836.

[20] Salama, W. (2017). Design of concrete buildings for disassembly: An explorative review. International Journal of Sustainable Built Environment, 6(2), 617-635.

[21] Seyed Hamidreza Mirsalari,Afrooz Haghbin,Mehdi Khatir & Farbod Razzazi. (2024). Least Squares Support Vector Regression-Based Channel Estimation for OFDM Systems in the Presence of Impulsive Noise. Wireless Personal Communications(2),1-16.

[22] Xiaoming Han,Xin Zhao,Yecheng Wu,Zhengwei Qu & Guofeng Li.(2024).A least squares–support vector machine for learning solution to multi-physical transient-state field coupled problems.Engineering Applications of Artificial Intelligence(PA),109321-109321.

[23] Huining Pei,Jingru Cao,Man Ding,Ziyu Wang & Yunfeng Chen.(2025).A assessment method for ergonomic risk based on fennec fox optimization algorithm and generalized regression neural network.Displays102905-102905.