



# OPEN Prediction of HPC compressive strength based on machine learning

Libing Jin<sup>1✉</sup>, Jie Duan<sup>1,3</sup>, Yichen Jin<sup>2</sup>, Pengfei Xue<sup>1</sup> & Pin Zhou<sup>1</sup>

There is a complex high-dimensional nonlinear mapping relationship between the compressive strength of High-Performance Concrete (HPC) and its components, which has great influence on the accurate prediction of compressive strength. In this paper, an efficient robust software calculation strategy combining BP Neural Network (BPNN), Support Vector Machine (SVM) and Genetic Algorithm (GA) is proposed for the prediction of compressive strength of HPC. 8 features were extracted from the previous literature, and a compressive strength database containing 454 sets of data was constructed. The model was trained and tested, and the performance of 4 Machine Learning (ML) models, namely BPNN, SVM, GA-BPNN and GA-SVM, was compared. The results show that the coupled model is superior to the single model. Moreover, because GA-SVM has better generalization ability and theoretical basis, its convergence speed and prediction accuracy are better than GA-BPNN. Then Grey Relational Analysis (GRA) and Shapley analysis were used to verify the interpretability of the GA-SVM model, which showed that the water-binder ratio had the most significant influence on the compressive strength. Finally, the combination of multiple input variables to evaluate the compressive strength supplemented this research, and again verified the significant influence of water-binder ratio, providing reference value for subsequent research.

**Keywords** Genetic algorithm, Machine learning model, High-performance concrete, Compressive strength, Parameter analysis

HPC is a widely utilized cement-based composite material<sup>1</sup>, consisting of water, cement, coarse aggregate, and additional mineral admixtures and chemical admixtures. It is produced through mixing, vibrating, forming, curing and solidification in specific proportions based on engineering experience or test formulas. Its widespread use is attributed to its excellent plasticity, fire resistance, corrosion resistance and utilization of local materials. The concept of HPC was first proposed by the National Institute of Technology and the American Concrete Institute in 1950. In contrast to ordinary concrete, HPC offers superior durability, workability and volume stability<sup>2</sup>.

At the same time, the mapping relationship between each component of HPC and its compressive strength becomes more complicated and multi-dimensional. Traditional modeling methods are based on scientists' descriptions of objective things and often need to oversimplify practical problems. This will cause the model to fail to capture enough mapping relationships and the error will be too large to achieve the desired effect. In addition, traditional modeling methods are faced with many difficulties in data acquisition, processing and final verification. Therefore, in the face of more dimensional mapping relationship between components of HPC and compressive strength, even on the basis of existing empirical regression formulas<sup>3-5</sup> and a large number of test sample data, it is difficult to derive a universal equation that can reach the accuracy of compressive strength measured in laboratory environment<sup>6-9</sup>.

With the continuous development of artificial intelligence, the powerful data identification and nonlinear fitting capabilities of ML make it have certain advantages in solving complex engineering problems<sup>10</sup>. Among them, a lot of research has been done on the prediction of the basic performance of concrete. With the development of concrete, ML has become a new tool for its performance research, and the results are fruitful.

Adding various fibers, nanoparticles or other substances to concrete can improve its mechanical properties<sup>11</sup> to form a new type of concrete, but there are few studies on its performance prediction. Chithra et al.<sup>12</sup> studied the use of artificial neural networks to accurately predict the compressive strength of concrete containing nano-silica particles and copper slag. Ashrafi et al.<sup>13</sup> used a heuristic regression method to predict the strength and

<sup>1</sup>School of Civil Engineering, Henan University of Technology, Zhengzhou 450000, China. <sup>2</sup>School of Physical Science and Engineering, Beijing Jiaotong University, Beijing 100044, China. <sup>3</sup>China MCC5 Group Corp. Ltd, Chengdu 610000, China. ✉email: jinlb@haut.edu.cn

ultrasonic pulse velocity of fiber-reinforced concrete. 3D printed concrete has attracted wide attention in recent years. Its short construction time, low labor cost and high design freedom make it suitable for various fields<sup>14</sup>. However, its anisotropy makes its compressive strength unreliable in predicting. Then, the beetle antenna search technology proposed by<sup>15</sup> can automatically search the optimal hyperparameters, thus realizing the accurate prediction of the compressive strength of 3D printed concrete under steam curing conditions. Long et al.<sup>16</sup> established a long- and short-term memory network model to further improve the accuracy of predicting the dynamic compressive strength of concrete-like materials at high strain rates. The above research provides certain ideas for the performance prediction of other types of concrete and has practical guiding significance. More researches on ML are shown in Table 1 below.

In short, ML as a sign of the age of intelligence, it opens up countless possibilities in civil engineering. It not only improves production efficiency, but also creates unprecedented changes in many aspects such as structural quality control and structural safety monitoring. Especially in the era of "double carbon", machine learning technology plays a huge role in the development and performance improvement of building materials. Therefore, to enable civil engineering, it will be able to help the early achievement of the "double carbon" goal, and help realize the diversified and intelligent development of civil engineering.

In this paper, the single model of BPNN and SVM is combined with GA optimization algorithm to make up for the lack of parallel computing capability of SVM, easy to fall into local optimal solution and easy to disappear gradient during BPNN training. At the same time, more input variables are introduced to further improve the prediction accuracy and reliability of the prediction model. In addition, the interpretability of the model was verified by GRA and Shapley analysis, which ensured the reliability of the model interpretation. It provides guidance for subsequent researchers to evaluate the compressive strength of HPC.

## Description and analysis of database

For machine learning techniques, database compilation is a fundamental and critical step. In order to make the research results universal, the composition of HPC is studied in this paper. Experimental data samples were collected from 9 literatures, and outlier data sets with water-binder ratio less than 0.1 or greater than 0.5 were eliminated, so as to establish a 28-day HPC cube compressive strength database. Table 2 lists detailed statistics of the data used in this article.

Previous work	Model	Input variables	Forecast target	R <sup>2</sup>	Data points
Chithra et al. <sup>12</sup>	ANN1 ANN2 ANN3	C; NS; A; SP C; NS; Fa; CS; A; SP C; NS; Fa; CS; A; SP	Compressive strength	0.9944 0.9987 0.9975	54 180 264
Aiyer et al. <sup>17</sup>	LSSVM RVM	C; A; W/B; SP; Fa; Ca	Compressive strength	0.8836 0.9082	56
Motamedi et al. <sup>18</sup>	ANFIS SVM-RBF kernel SVM- linear kernel	CoS; C; Fa; A	Unconfined compressive strength	0.9996 0.9980 0.8622	810
Pham et al. <sup>19</sup>	ANN SVM FA-LSVM	C; Fa; SCa; MCa; W; SP; A	Compressive strength	0.76 0.79 0.87	239
Omran et al. <sup>20</sup>	Multiple layer perceptron (ANN)	C; A; FA; Fa; pea gravel, Haydite LWA, and Micro Air	Compressive strength	0.97	144
Ashrafian et al. <sup>13</sup>	MLP-ANN LSSVM	$V_f$ ; S/A; LP; NS; W/B; SP/B; A	Compressive strength	0.8892 0.9120	175
Zhang et al. <sup>21</sup>	BAS-RF	W/B; PP fiber; Steel fiber; Scoria; Fa; Ca; SP	uniaxial compressive strength	0.9477	131
Chen et al. <sup>7</sup>	LSSVM-NSGA-II	W/B; C; FA; Fa; Ca; SP	Frost resistance Permeability resistance	0.94084 0.9443	100
Yuan et al. <sup>22</sup>	GA-BPANN ANFIS	C; BFS; FA; W; SP; Ca; Fa	Compressive strength	0.813 0.950	180
Chandwani <sup>23</sup>	GA-ANN	W/B; C; FA; Fa; Ca; Ad	Slump	0.9586	560
Yan et al. <sup>24</sup>	GA-ANN	$S$ ; $P$ ; $d_b$ ; $c/d_b$ ; $l_d/d_b$ ; $\rho$ ; $\sqrt{f_c}$	Bond strength	0.950	687
Fan et al. <sup>25</sup>	ANN GA-ANN	C; SF; LP; Fa; W/C	Compressive strength	0.8306 0.9704	80
Wang et al. <sup>26</sup>	ANN GA-ANN PSO-ANN	$c_s/h_s$ ; $c_s/h_s$ ; $b/h_s$ ; $l_b/h_s$ ; $f_t$ $\rho_s$ ; $\rho_{sv}$	Bond strength	0.7225 0.8281 0.8649	191

**Table 1.** A synopsis of models from earlier research. Fine aggregate (Fa); coarse aggregate (Ca); blast furnace slag (BFS); admixtures (Ad); bar surface (S); bar position (P); bar diameter ( $d_b$ ); concrete cover to bar diameter ratio ( $c/d_b$ ); embedment length to bar diameter ratio ( $l_d/d_b$ ); transverse reinforcement ratio ( $\rho$ ); limestone powder (LP); water cement ratio (W/C); concrete cover ( $c_s$ ) and side cover ( $c_s$ ); relative steel section width ( $b/h_s$ ); relative bonded length ( $l_b/h_s$ ); concrete tensile strength ( $f_t$ ); relative size of steel section ( $\rho_s$ ); stirrup ratio ( $\rho_{sv}$ ); nano silica content (NS); copper slag content (CS), age of specimen (A); Cockle shell (CoS); small coarse aggregate (SCa); medium coarse aggregate (MCa); Volumetric Percentage of Fiber ( $V_f$ ); Superplasticizer to Binder Ratio (SP/B).

Data source	Parameters	Unit	Min	Q <sub>25%</sub>	Mean	Median	Q <sub>75%</sub>	Max	Std	Skw
Asteris et al. <sup>27</sup>	X <sub>1</sub> : W	kg/m <sup>3</sup>	135	160	168.50	170	180	214	15.61	0.42
Baykasoglu et al. <sup>28</sup>	X <sub>2</sub> : PC	kg/m <sup>3</sup>	220	365.78	441.54	426.67	510	708	96.78	0.36
Chindaprasirt et al. <sup>29</sup>	X <sub>3</sub> : W/B	(ratio)%	0.18	0.3	0.33	0.33	0.4	0.5	0.08	-0.001
Elahi et al. <sup>30</sup>	X <sub>4</sub> : S/A	(ratio)%	0.28	0.38	0.44	0.44	0.49	0.64	0.06	0.50
Jiao et al. <sup>31</sup>	X <sub>5</sub> : SF	kg/m <sup>3</sup>	0	0	23.34	0	30.00	208	44.07	2.04
Lim et al. <sup>32</sup>	X <sub>6</sub> : FA	kg/m <sup>3</sup>	0	0	57.43	45	91.32	275	65.46	1.24
Pala et al. <sup>33</sup>	X <sub>7</sub> : AE	kg/m <sup>3</sup>	0	0	0.03	0.04	0.06	0.078	0.03	0.28
Prasad et al. <sup>34</sup>	X <sub>8</sub> : SP	kg/m <sup>3</sup>	0	3.68	8.00	6.55	8.89	36.5	6.42	1.72
Yen et al. <sup>35</sup>	Y: CS	MPa	38	50	67.63	63	79.14	123	21.19	0.80

**Table 2.** HPC compressive strength test parameters . Skw = Skewness; Std = Standard deviation.

The distribution of input variables greatly affects the degree of generalization of the constructed model<sup>36</sup>. The cumulative frequency range histogram in Fig. 1 illustrates how input factors affect the compressive strength of HPC. From the graph, it can be seen that the frequency of the vector is appropriately high, and the range of input variables is wide. Cement content ranges from 220 to 708 (kg/m<sup>3</sup>) while most values are between 300 and 650 (kg/m<sup>3</sup>). Compared with cement content, the water content range is relatively narrow, and most are between 160 and 185 (kg/m<sup>3</sup>). Regarding the content of silica fume and fly ash, there are many cases where the content is zero, and the dispersed content of the two is between 0 and 200 (kg/m<sup>3</sup>) and 0–275 (kg/m<sup>3</sup>), respectively, and the change is relatively large. The water-binder ratio is another important vector influencing the compressive strength of HPC<sup>37,38</sup>, and its value is mainly concentrated around 0.3. In this research, the compressive strength of HPC ranges from 38 to 123 MPa and is concentrate between 40 and 80 MPa.

Figure 2. shows the multiple-correlation matrix. Different colors represent different correlation values. The horizontal and vertical cross factors have a positive association when the value is positive, and a negative correlation when it is negative. Among the 8 input variables selected, the correlation between superplasticizer and silica fume content was the strongest (R = 0.84), which is consistent with previous research. It is commonly known that the strength and mechanical characteristics of HPC can be considerably impacted by the dosage of silica fume and superplasticizer<sup>37,38</sup>. Regarding the relation between input and output variables, superplasticizer and compressive strength have the largest positive association, with silica fume and cement following closely behind. In general, there is little difference in the positive and negative relevance found between the input and output variables. Consequently, to guarantee the model's accuracy, all eight input variables were employed.

## Machine learning approaches

### BPNN

ANN is a calculation model based on simulating the connection and excitation suppression of neurons in the human brain. Typically, its structure is composed of three layers of neurons: the input layer, hidden layer, and output layer. Neurons in the input layer are responsible for receiving information from the surrounding environment, and neurons in hidden layer and output layer are in charge of carrying out linear and nonlinear approximations of the system under investigation. A single microprocessor will add up the weighted values received from the input layer neurons during the linear phase of ANN. The function of activation is used to the sum during the nonlinear phase, and the outcome is sent as the microprocessor's output. In an ANN, weights are used to connect each layer's neurons to one another sequentially. Figure 3 illustrates the ANN's structural layout. The total amount of neurons in the input layer and output layer, respectively, is represented by the amount of features and tags in data set. It is possible to ascertain the amount of neurons in the hidden layer using an empirical formula method or a trial-and-error approach.

### Support vector machine regression (SVM)

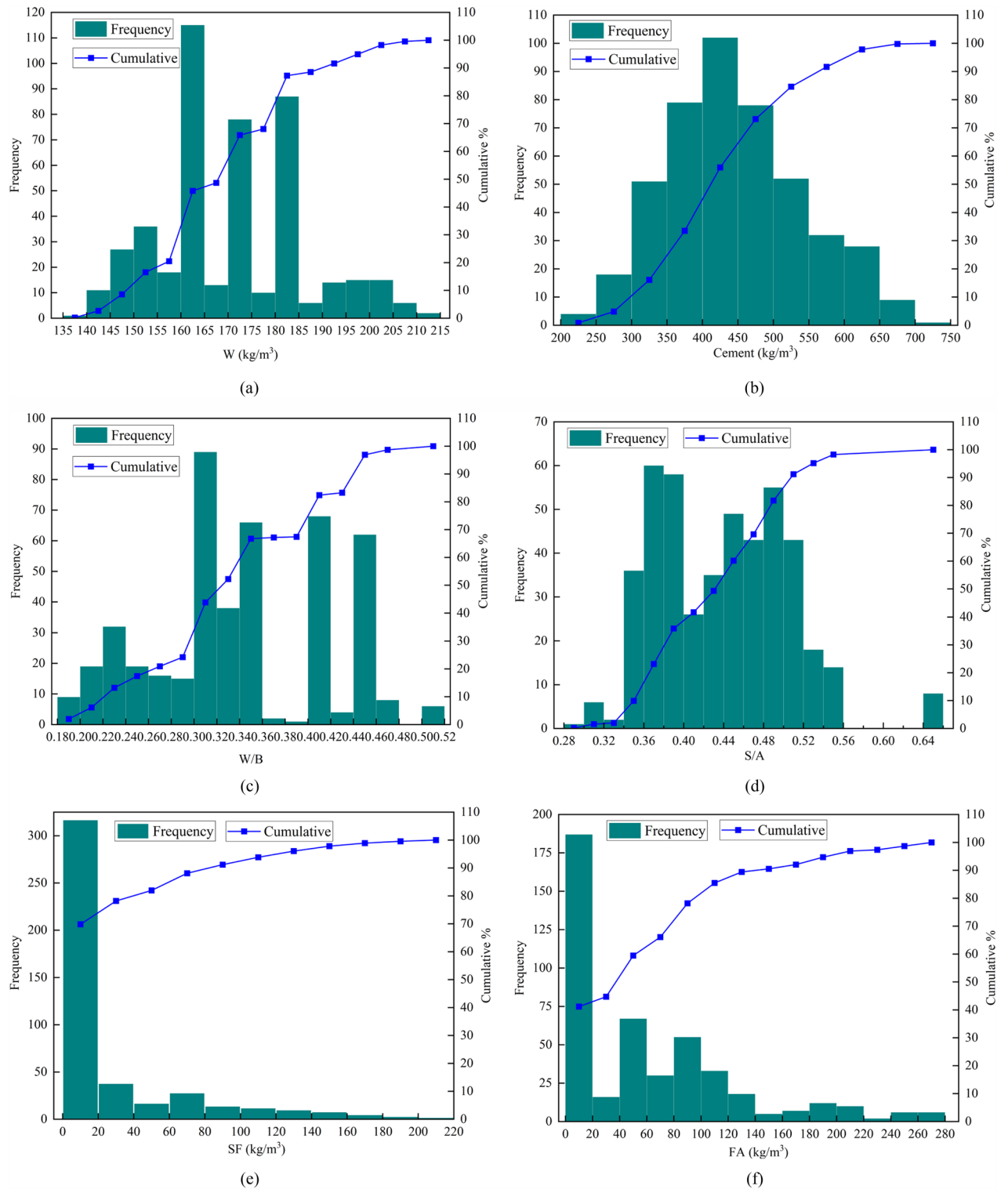
Cortes and Vapnik<sup>39</sup> present SVM as a solution for categorization issues. In order to solve regression and prediction issues, SVM is a significant branch that SVM extended. These kinds of regression and prediction models usually compute model losses using the difference value between the model's expected and actual output values. The loss is only zero when the predicted value of the model matches the true value. As shown in Fig. 4, the low dimensional feature space of the sample can be transformed into a high-dimensional feature space using Gaussian radial basis functions to enhance the performance of the regression model and better match the collected data. The output of the SVM is expressed as a linear function, as shown below:

$$f(x) = \langle \omega \varphi(x) \rangle + b \quad (1)$$

where  $\langle \cdot \rangle$  represents point function;  $\omega$  is the minimal value obtained from the following equation.

$\varphi(x)$  is a mapping of input features to higher-dimensional feature spaces;  $b$  is one of the function's parameters vectors.

$$\min imise \left[ \frac{1}{2} \|\omega\|^2 + C \sum_{j=1}^n (\xi_j + \xi_j^*) \right] \quad (2)$$



**Figure 1.** Statistical distributions of the input/output variables: (a)–(i).

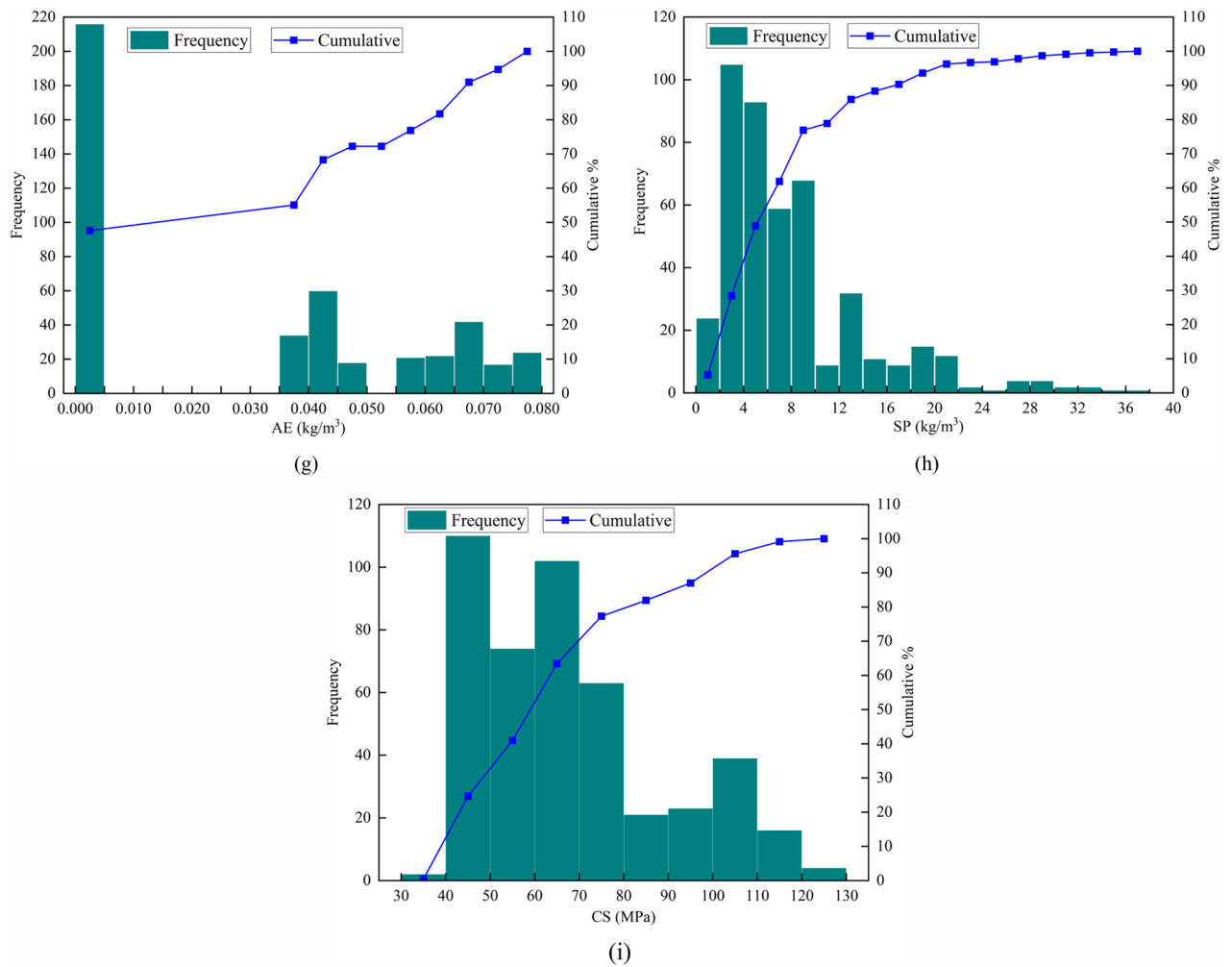


Figure 1. (continued)

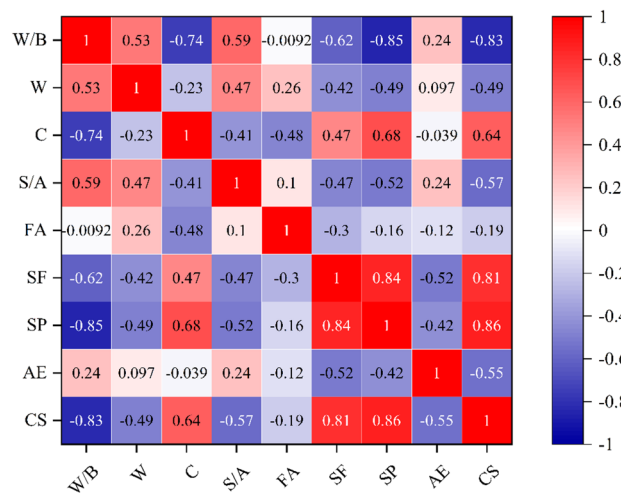
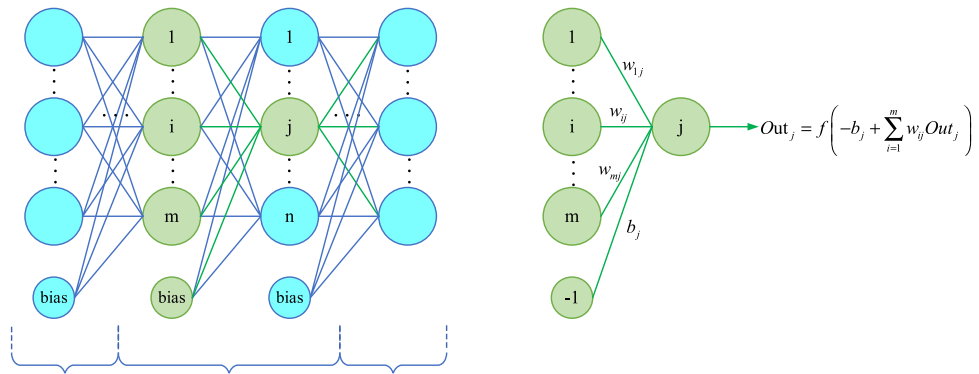
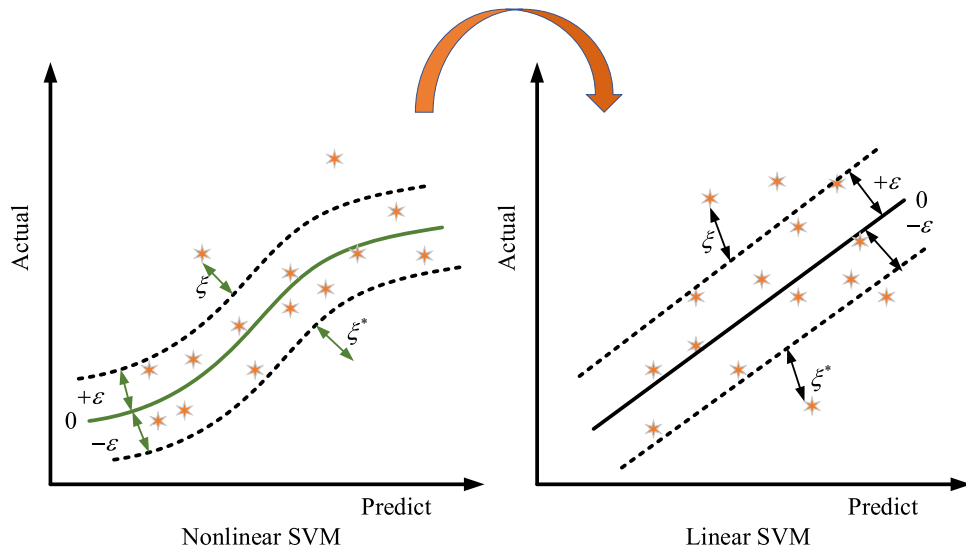


Figure 2. Multiple correlations of input variables.



**Figure 3.** Architecture of ANN.



**Figure 4.** Support vector machine algorithm structure.

$$\text{subject to } \{y_i - \omega\varphi(x_i) - b \leq \varepsilon + \xi_j \omega\varphi(x_i) + b - y_i \leq \varepsilon + \xi_j^* \xi_j, \xi_j, \xi_j^* \geq 0$$

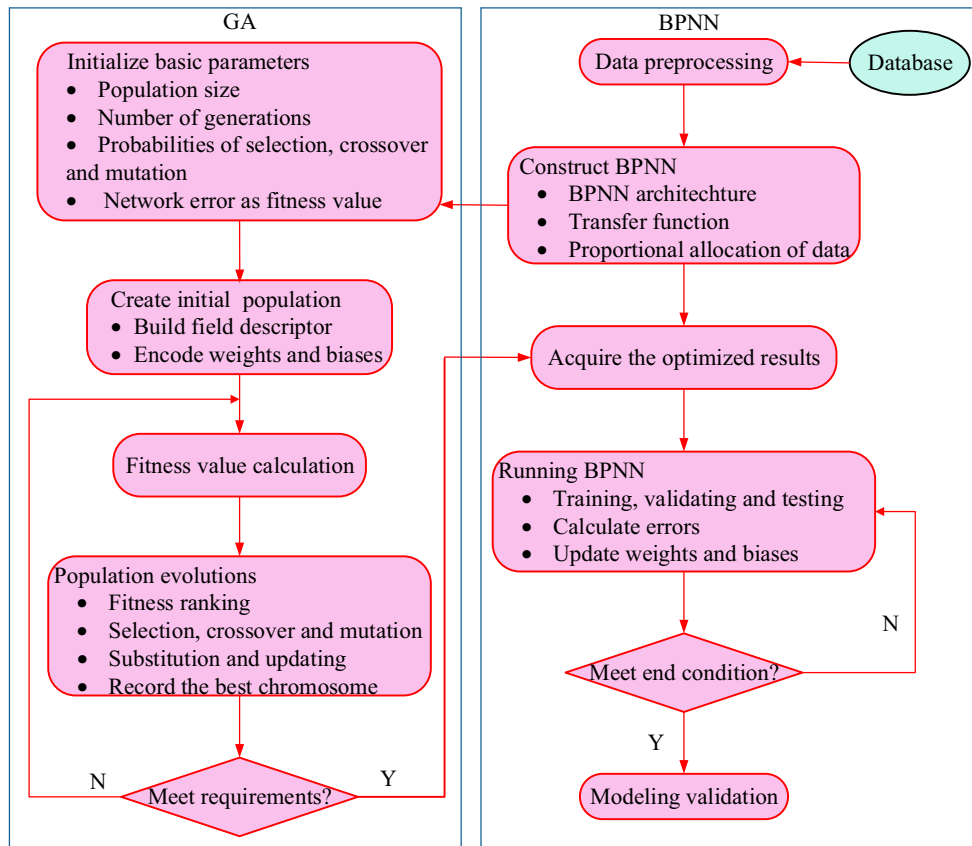
where  $C$  for the regularization parameter;  $n$  represents the quantity of samples;  $\xi_j, \xi_j^*$  are slack variables;  $\varepsilon$  is insensitive loss function;  $y_i$  is experimental value.

**Principle of genetic algorithm (GA)**

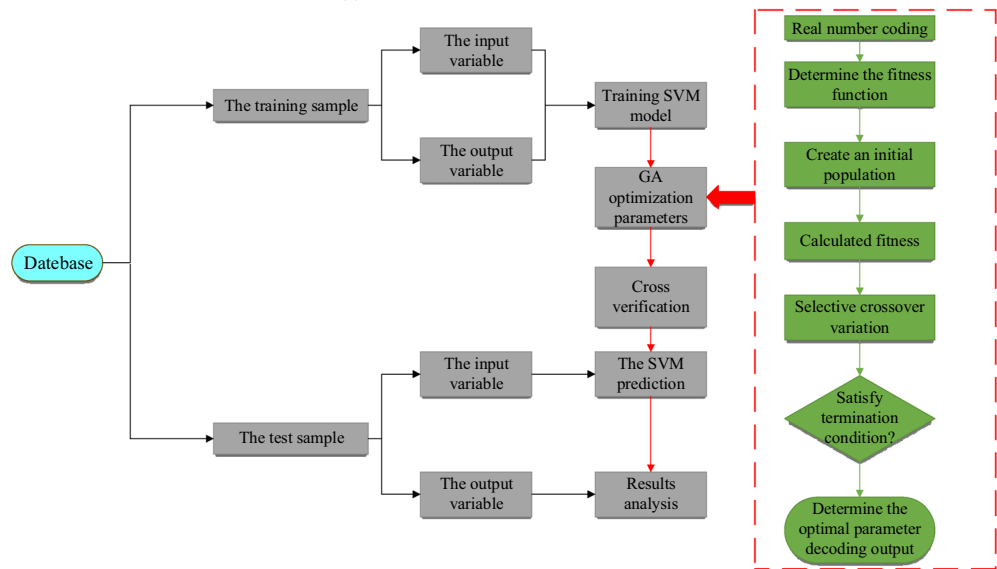
The natural biological evolution system’s computer numerical simulation technology is where GA originated. It is a technique for competition based on random global search and optimization that was created by modeling how organisms naturally evolve<sup>40</sup>. It is utilized for hyperparameter modification in various machine learning algorithms and is based on Mendelian genetics and Darwin’s theory of evolution. In an iterative search process, it can automatically search, gather information about the search space, and adaptively regulate the search to find the best solution. Consequently, it can be considered a worldwide, efficient, parallel search heuristic method. Because the conventional BPNN has a tendency to enter a local minimum, the SVM hyperparameter adjustment is crucial. The baseline weights and biases of the BPNN are adjusted in this work using the random global search and optimization capabilities of the GA. Additionally, the global search and optimization of the SVM’s hyperparameters is employed to improve the model’s precision. Figure 5 illustrates how BPNN and SVM are optimized using GA.

**Assessment of model function**

Three widely used statistical indicators—Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Squared Correlation Coefficient ( $R^2$ )—were utilized to evaluate the built prediction model in order to determine its performance<sup>28,41,42</sup>. While both MAE and RMSE provide insight on the size of the participation error<sup>43</sup>, the RMSE data are mostly used to choose the optimal prediction model. With lower MAE and RMSE values, the established prediction model performs better.  $R^2$  is a statistical metric that quantifies the extent to which the



(a)



(b)

**Figure 5.** GA optimization flowchart of BPNN and SVM: (a) and (b).

independent variable can account for variations in the dependent variable. Its values fall between 0 and 1. A regression line is said to be closer to each sample test point if the  $R^2$  value is endlessly close to 1. This suggests that the larger the ratio of the regression's sum of squares to the population's sum of squares. Regression fitting works best when changes in independent variable  $x$  can completely account for changes in the value of the dependent variable,  $y$ <sup>42</sup>. The formulas of the three indicators are as follows:

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i| \quad (3)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2} \quad (4)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (5)$$

where  $\hat{y}_i$ ,  $y_i$  are predicted value and tested value respectively;  $\bar{y}$  is the mean of all tested values; the test set's overall amount of samples is denoted by  $n$ .

### Research framework process

The frame structure of machine learning model for studying compressive strength of HPC in this paper is shown in Fig. 6. There are generally four steps:

Step 1: Database construction.

There are 454 samples overall in the built database, including PC, W, SP, SF, FA, AE, W/B, S/A, and HPC compressive strengths. The database was subjected to multi-correlation matrix analysis to ascertain the level of relationship among the input variables. Then, utilizing 70% of database for training and 30% for testing, the machine learning model was put to use.

Step 2: Training of the machine learning model.

The training set was applied in this step to train two independent machine learning models of the suggested BPNN and SVM. Furthermore, GA was employed to optimize the model's weights, thresholds, and hyperparameters, producing two coupled machine learning models: GA-BPNN and GA-SVM.

Step 3: Defining the optimal machine learning model.

The test dataset is applied to validate the machine learning model. Next, using statistical markers (MAE, RMSE,  $R^2$ ), the optimal model is chosen.

Step 4: Parameter analysis.

GRA and shapley analysis were used to test the characteristics and significance of input variables, and the influence of the number and value of input variables on the model was investigated. Finally, the actual experimental results are analyzed and compared with the predicted results generated by the six combinations.

## The results of the research and discussion

### Performance assessment of models

The quantity of neurons in hidden layer of BPNN influences its property, whereas the  $c$  and  $g$  parameters have an impact on SVM performance. BPNN may approximate any nonlinear model of topological structure with three layers<sup>44,45</sup>. As a result, this paper only used one hidden layer. If there are not enough neurons in the network, it cannot accurately depict the connection between input and output. On the other hand, an excess of neurons may result in an extended running period and the over-fitting phenomenon. Generally, the empirical formula (6) is used to determine the quantity of neurons. The results of comparing and selecting the neuron count of hidden layers on network performance (MSE) are shown in the Fig. 7. The parameters  $c$  and  $g$  were optimized through cross validation and genetic algorithm, and the results are shown in Fig. 8.

Table 3 shows the course of trial and error that was utilized to determine the optimal hyperparameters for the machine learning. While certain parameters were left as default, each ideal parameter utilized in the model was fine-tuned using a combination of expertise and trial-and-error to get its best value. Next, each model's forecast precision was computed.

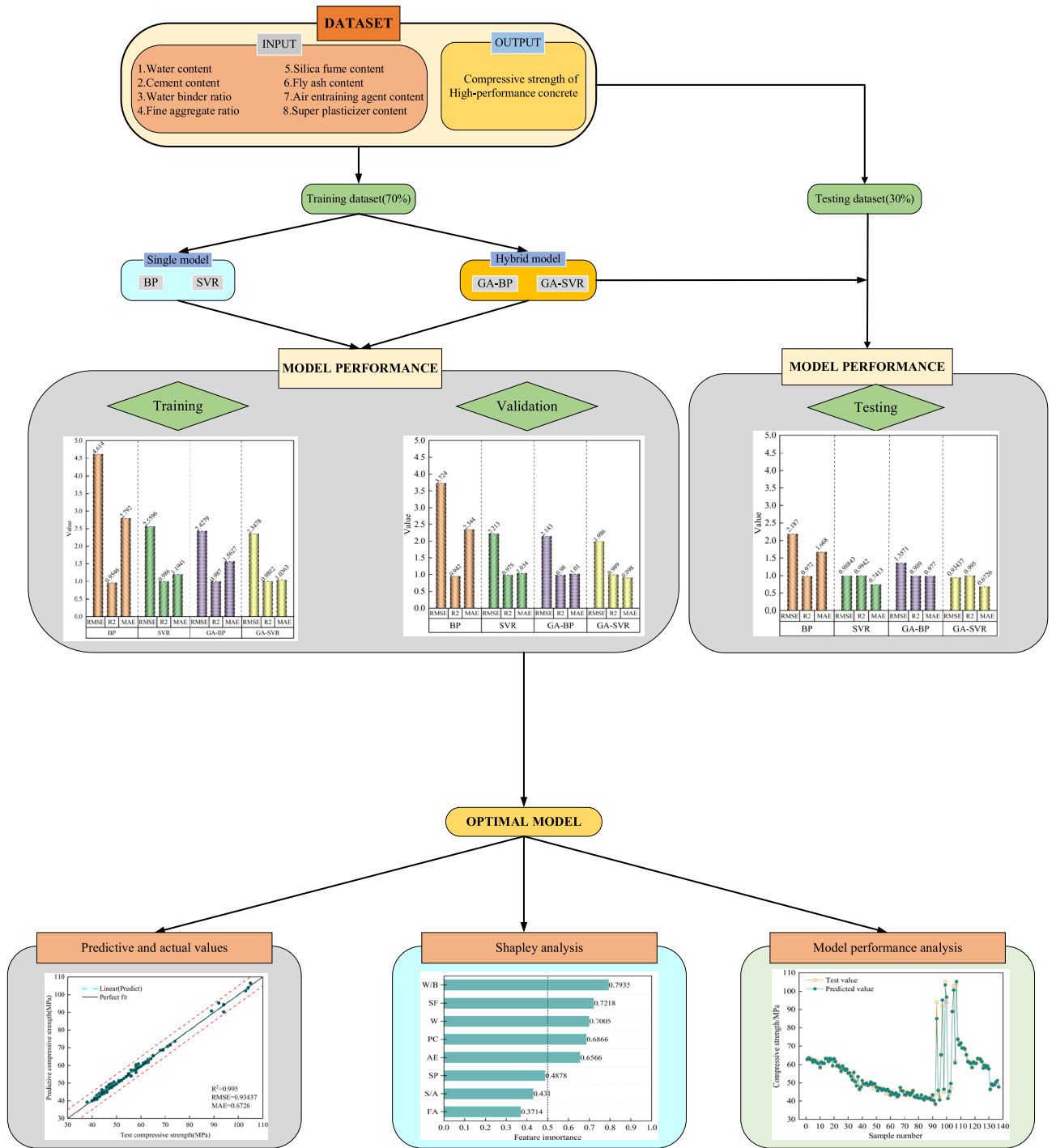
$$L = \sqrt{n + m} + a \quad (6)$$

The variables  $L$ ,  $n$ , and  $m$  represent the quantity of neurons in hidden layer, output layer, and input layer, respectively, of a BPNN. The constant,  $a$ , can be any of the following values. Figure 7 illustrates how the quantity of neurons in hidden layer affects the network's MSE. In this research, the ANN was run multiple times to get the train and test sets' MSE values. Their average was then computed to determine the ideal number of hidden layer nodes.

The comparison of the four developed machine learning models is displayed in Fig. 9. The training set's  $R^2$  value is consistently less than the test set's, even for all employed algorithms, albeit the difference is negligible. The model that employs BPNN has the lowest accuracy since it has the lowest  $R^2$  value and bigger RMSE and MAE values than the SVM model. Overall, all algorithm of this study exhibit strong performance, with accuracy in predicting surpassing that of a single model. But the GA-SVM combination fared better than the others, with the lowest MAE and RMSE values as well as the greatest  $R^2$  value.

The advantages of the evolutionary algorithm in stochastic global search and optimization are evident in the hybrid model's superior performance over the single model, as demonstrated by the comparative findings in Fig. 9<sup>40</sup>. GA, for instance, is difficult to trap in a local minimum. Thus, GA can enhance the efficacy of a single predictive model by reducing prediction error; in this instance, a combination of models was applied to estimate

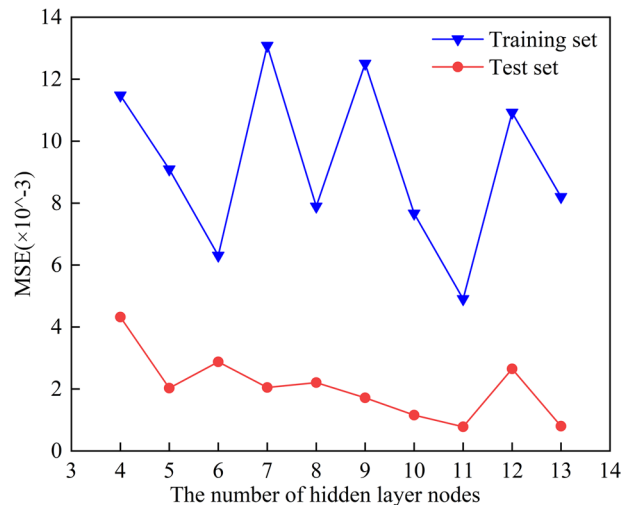




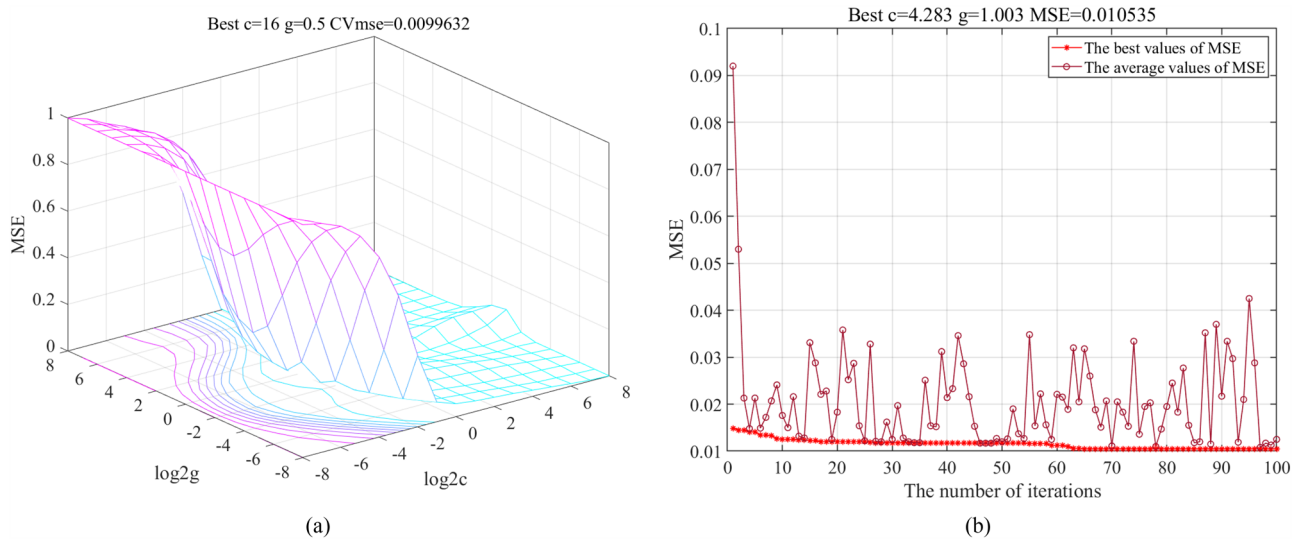
**Figure 6.** HPC compressive strength machine learning framework.

the HPC property. For the train and test sets of the database, Fig. 10 presents a comparison and comparison of the actual and predictive values of compressive strength.

Figure 10 displays the goodness of fit of the tested and anticipated compressive strength of HPC. The figure's perfect fitting curve is represented by  $y = x$ , and the 5 MPa model prediction accuracy error boundary is indicated by the crossing line  $y = x \pm 5$ . The training and test sets' compressive strengths, as predicted by the two coupling models, are demonstrated in terms of their accuracy and goodness of fit using the fitting curve and 5 MPa error boundary. The findings demonstrated that nearly all of the data points were situated close to the curve  $y = x$ . The test set's values of  $R^2$  were 0.982 and 0.995, respectively, while the two model training sets'  $R^2$  values were  $R^2 = 0.987$  and  $R^2 = 0.989$ . This shows that the test set's goodness of fit is higher than the training sets', and the model's predicted and tested values have a better goodness of fit. The error border shows that all of the test set's data points are inside the boundary, but only a small portion of the training set's data points for the two models



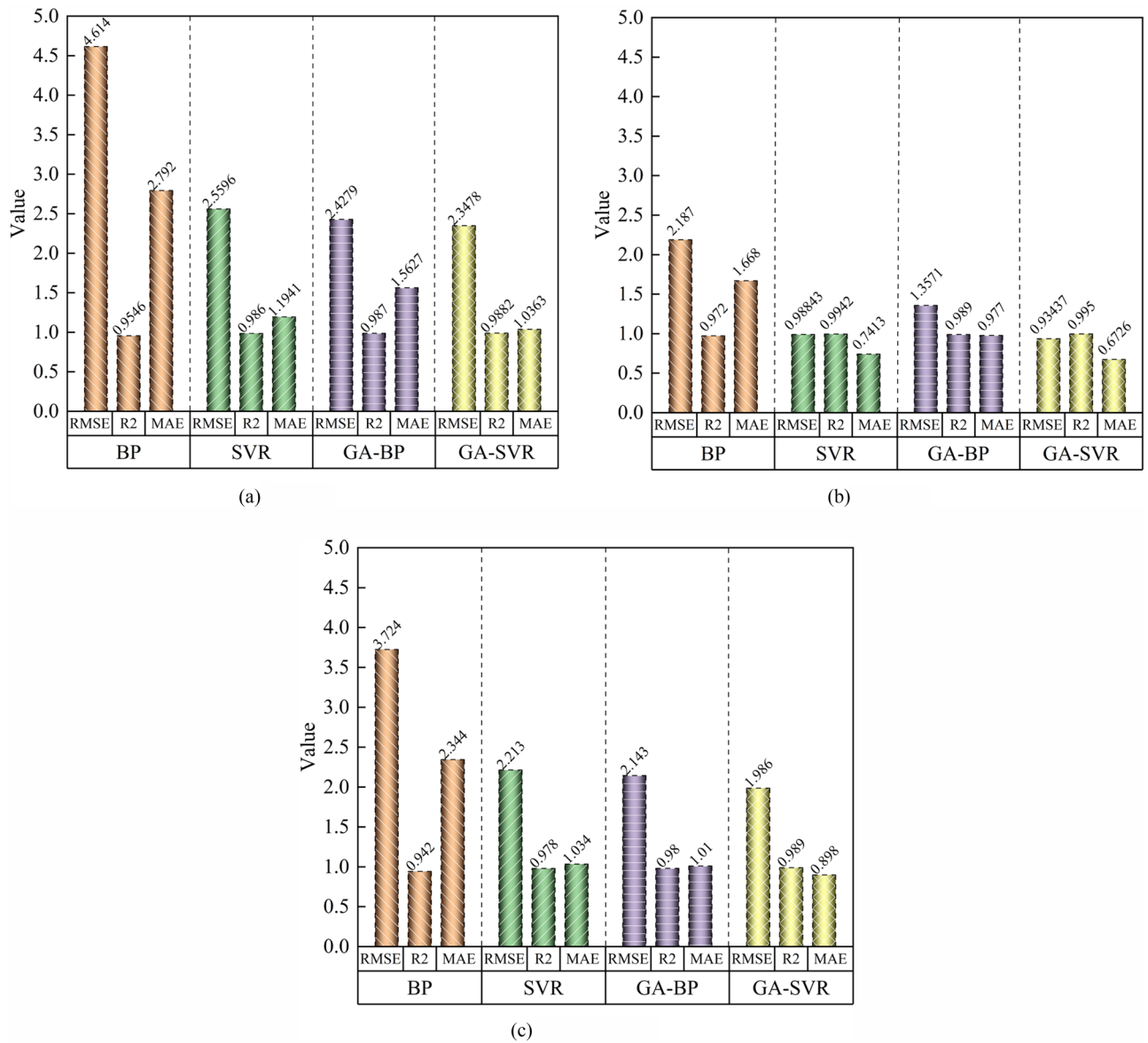
**Figure 7.** Best hidden layer neurons nodes.



**Figure 8.** c, g parameter optimization: (a) and (b).

GA-BP model		GA-SVM model					
BP parameters		GA parameters		SVM cgForRegress		GA SVMcgForRegress	
Epochs	1000	Sizepop	30	$c_{min}$	-8	Sizepop	20
Learning rate	0.01	Maxgen	50	$c_{max}$	8	Maxgen	200
Goal	1e-5	pCrossover	0.8	$g_{min}$	-8	pCrossover	0.4
Mc	0.01	pMutation	0.2	$g_{max}$	8	pMutation	0.01
Min grad	1e-6	-	-	v	5	v	5
Max fail	6	-	-	$c_{step}$	1	cbound	[1,100]
				$g_{step}$	1	gbound	[1,1000]
				$mse_{step}$	0.06	ggap	0.9

**Table 3.** Optimal hyperparameters for machine learning models.



**Figure 9.** Performance of machine learning: (a)–(c).

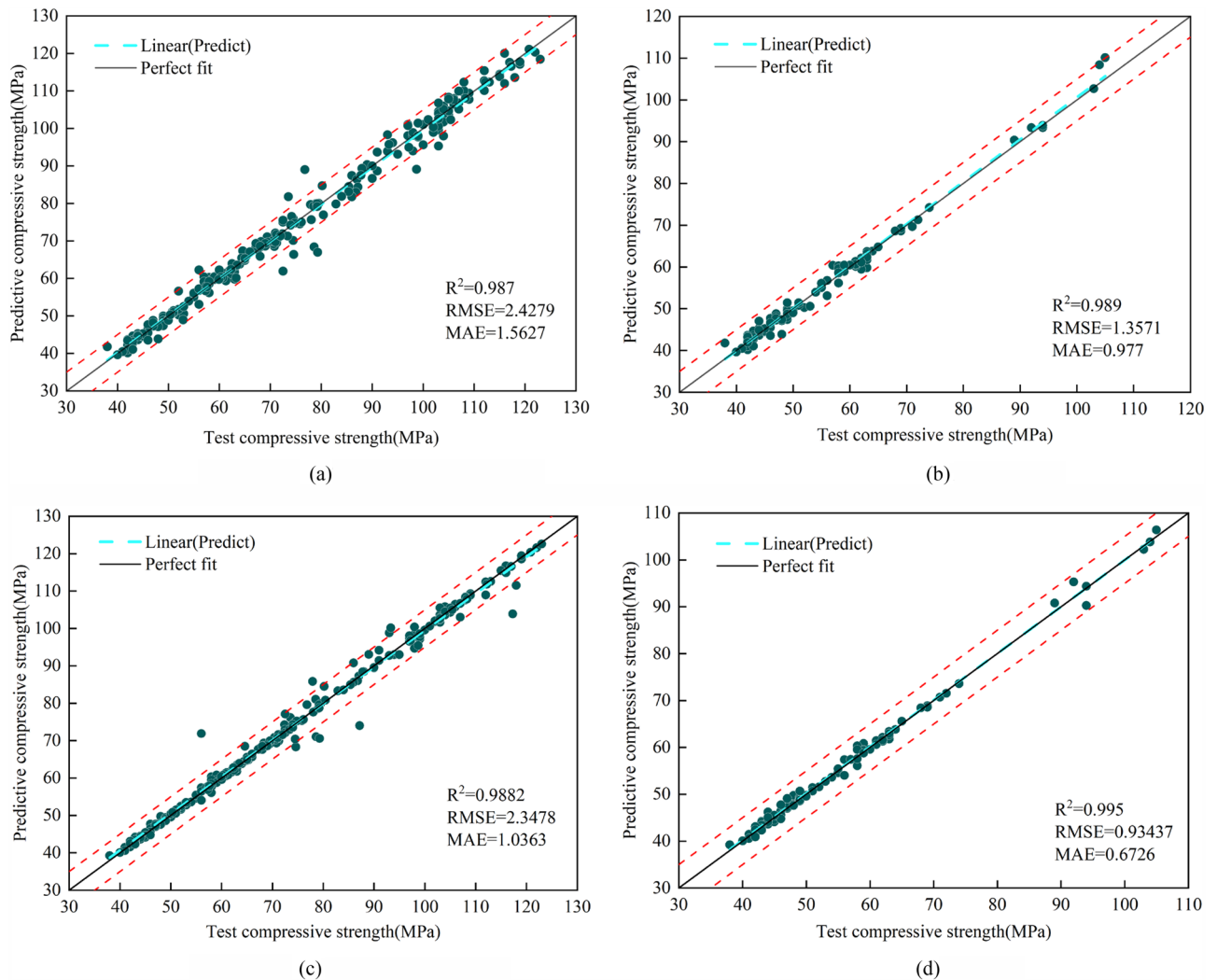
are outside of it. GA-SVM’s RMSE and MAE values are, respectively, 0.93437 MPa and 0.6726 MPa. According to the results, GA-SVM has the highest prediction precision in the hybrid model. This is primarily because genetic algorithms have the ability to search globally and optimize them, while SVM algorithms require fewer parameters and minimize duplicate learning.

### Characteristic importance analysis

Although the GA-SVM mixed model in this study can accurately predict the compressive strength of HPC, it cannot test the influence of different input factors on the compressive strength of HPC. In order to study the importance of each input vector to the compressive strength of HPC in GA-SVM coupling model. In this paper, GRA and shapely analysis methods were used to combine the features to study the influence of the input features on the model performance<sup>42</sup>.

#### Theory of GRA

By measuring the correlation between a parent sequence (reference sequence) and a sub reference (comparison sequence) using the sequence curve that the current data set forms, GRA is able to perform quantitative analysis. The research object’s data set serves as the reference sequence, and the pertinent variables that have an impact on the research object serve as the GRA comparison sequence. It is possible to ascertain the degree of correlation by comparing the two sequence curves. This method uses the degree of relevance—which is a quantification of the geometric form similarity—to reflect the degree of similarity between the geometric shapes of sequence curves. The largest impact on the system’s reference sequence correlation can be identified using correlation calculations<sup>46–48</sup>. The primary idea is to look into the degree of connection between comparison sequences and



**Figure 10.** Comparison of predictive and tested values of hybrid model: (a)–(d).

reference sequence pairs using a simple model. Therefore, the best possible sorting of the comparison sequence is the aim of influence. It is assumed that the comparison and reference sequences are RS and CS, respectively.

$$X_0 = \{x_0(1), x_0(2) \dots, x_0(n)\} \text{ (RS)}$$

$$X_i = \{x_i(1), x_i(2) \dots, x_i(n)\} \text{ (CS)}$$

$i$  is the amount of CS, which is 8 in this paper.

The following are the stages involved in operating GRA:

1. Dimensionless method: Since the range and units of initial data collection vary, dimensionless processing is required. However, due to the existence of 0 information in the data of this study, the dimensionality cannot be treated by means of averaging. In this paper, the dimensionless data is processed in a standardized way.

$$x_i(z)' = \frac{x_i(z) - \mu}{\sigma} \tag{7}$$

where  $z = 1, 2, \dots, n-1, n$ .  $\mu$  is the mean value of the data sample and  $\sigma$  is the standard deviation of the data sample.

2. To solve the gray relation coefficient:

$$\zeta_i(z) = \frac{\min_i \min_z |x_0(z) - x_i(z)| + \rho \cdot \max_i \max_z |x_0(z) - x_i(z)|}{|x_0(z) - x_i(z)| + \rho \cdot \max_i \max_z |x_0(z) - x_i(z)|} \tag{8}$$

where  $\rho \in (0,1)$  represents the resolution coefficient. The resolution will increase as the resolution coefficient decreases. The resolution coefficient is set as 0.2 in this paper.

3. To solve the degree of gray relation:

$$\gamma_i = \frac{1}{n} \sum_{z=1}^n \zeta_i(z) \quad (9)$$

where  $z=1, 2, \dots, n-1, n$ .

4. Sorting by gray relationship degree: The measure of sequence proximity and the definition of the connection level order between the parent and subsequences are based on the level of gray relation solved.

#### *GRA analysis enhances model interpretability*

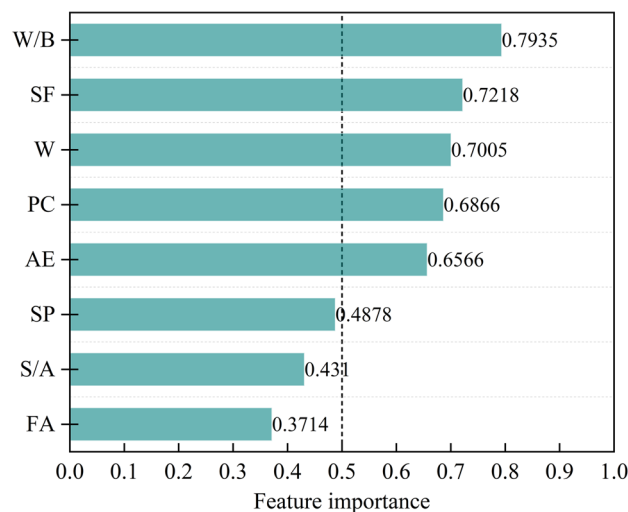
The standard compressive strength of HPC cube at 28 days of age is taken as reference sequence, and its 8 components are used as factors to establish a comparison sequence, and then the GRA model of HPC compressive strength is established by MATLAB. The model research results are shown in Fig. 11.

A correlation analysis of the GA-SVM prediction model's input variables and output values is presented in Fig. 11. A correlation degree more than 0.7 is typically considered significant, 0.5 to 0.7 is considered rather significant, and the remaining correlation degrees are considered insignificant. The research results show that the main factor affecting the compressive strength of HPC is W/B<sup>49-51</sup>. W/B too small will make the concrete easy to crack, not conducive to site construction operations. Conversely, excessive W/B will reduce the compressive strength. Therefore, appropriate W/B is crucial. Secondly, although SF, W, PC and AE have less influence on compressive strength than W/B, the importance value of the features still belongs to a fairly significant range. However, SP, S/A and FA belong to the non-significant range because their feature importance values are lower than 0.5. Although these characteristics have little influence on the strength, it is feasible to appropriately improve the compressive strength of HPC. For example, the addition of SP at an appropriate water-cement ratio can change the void structure inside the concrete and the final form of the hydration product. The mechanical properties, workability and durability of HPC are affected by S/A. When the concrete mixture falls within the tolerable range, it can achieve increased fluidity, maintain good cohesion and water retention, and thus improve strength. According to GRA theory, the closer the correlation value is to 1, the greater the impact of this feature on the compressive strength of HPC<sup>49</sup>. It can be concluded that all the characteristics have an effect on the compressive strength of HPC, among which W/B, SF and W are the most influential. Therefore, in order to improve the calculation speed of the model and reduce the workload of researchers, the less important elements can be removed when predicting the compressive strength of HPC.

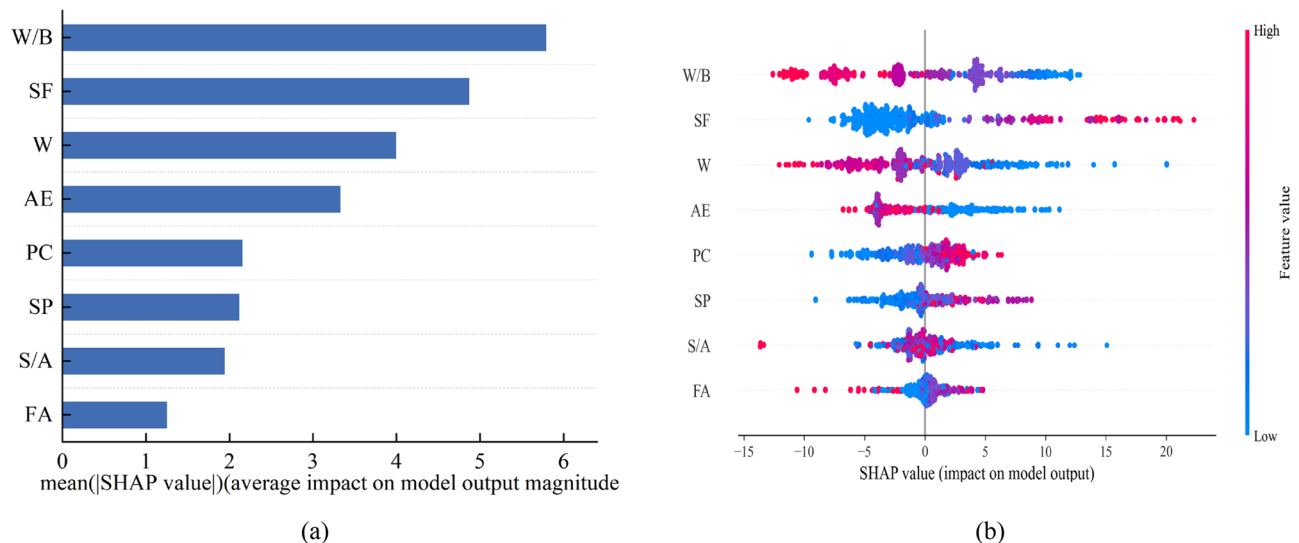
#### *Shapley analysis enhances model interpretability*

Shapley analysis is an analytical method that enhances the interpretability of machine learning models. It uses the feature importance indicator to describe the importance of the input variable of the database to affect the output value. Shapley analysis was carried out in this paper, and the results of feature importance are shown in the figure. Among the 8 characteristics, W/B has the most significant influence on compressive strength, and its value can reach 5.794. In addition, the importance value of SF feature was 4.872, which also contributed more to compressive strength, followed by AE, PC, SP, S/A and FA. Compared with the analysis results of GRA, the two analysis results are roughly the same.

It is crucial to understand the specific impact of each feature in the database on the machine learning output target. Therefore, the summary of SHAP values in Fig. 12 shows the reader the impact of each feature on the output target. Each point in the diagram represents the SHAP value of the variable. The color of the dots from pink to blue indicates that the influence of each feature is from strong to weak. The x horizontal axis represents



**Figure 11.** Gray relational rank of input variables.



**Figure 12.** Shapley feature importance analysis diagram:(a)–(b).

the SHAP value of each feature, and the  $y$  vertical axis displays each feature in descending order. In addition, the data point located in the negative region has a negative correlation effect on the output, and vice versa has a positive correlation effect. Therefore, a lower W/B will increase the compressive strength of HPC, and a lower SF content will decrease the compressive strength of HPC.

#### *Effect of the amount of input variables*

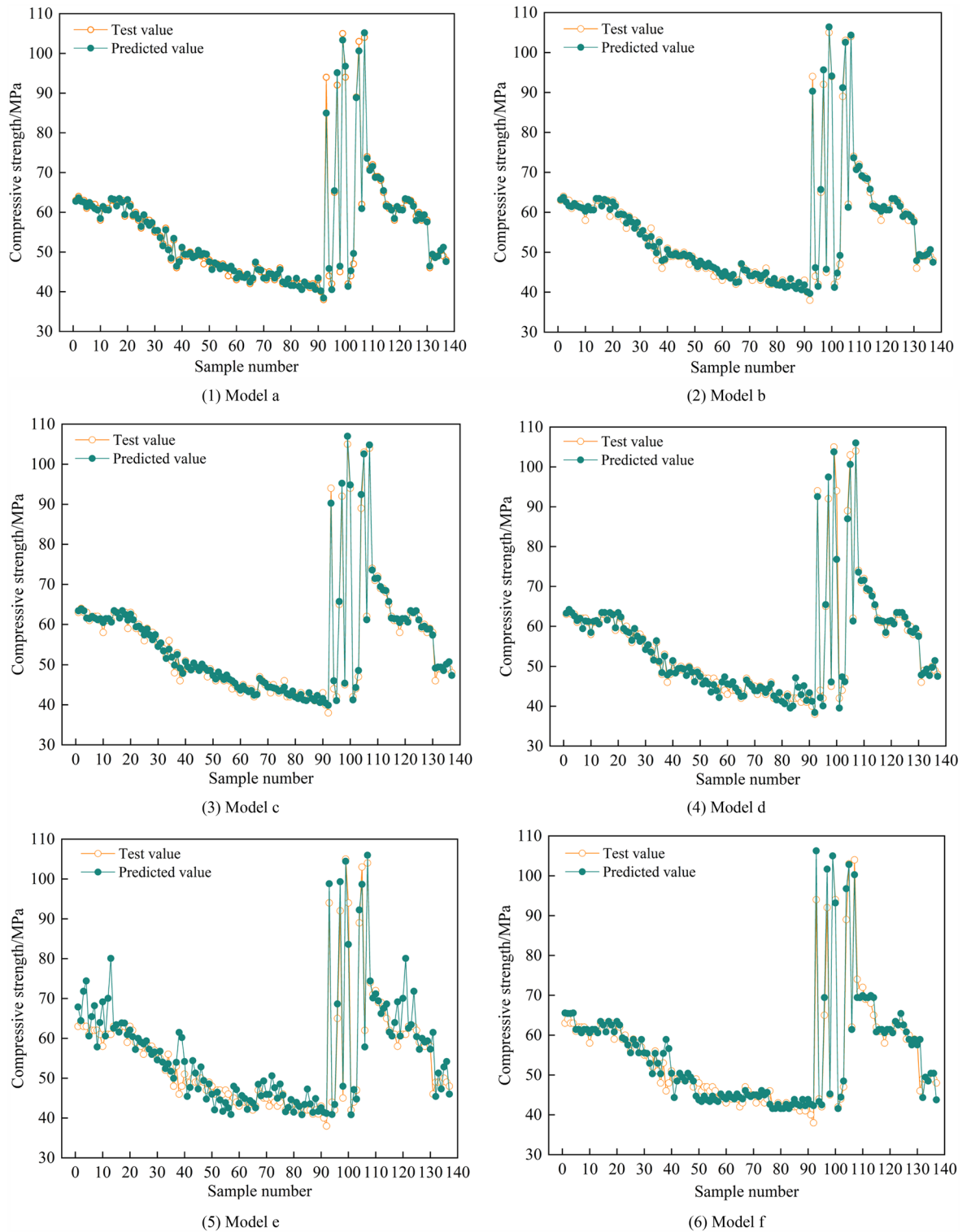
In this part, the performance of the GA-SVM model is studied by combining the results of the feature importance analysis of the database with the best prediction model GA-SVM. The compressive strength predicted by the numerical model is affected by the number of features and their significance<sup>42</sup>. Therefore, on the basis of the feature importance analysis results, this study gradually ignored some features, and finally only retained three input variables, AE, PC and SP, and established six numerical prediction models. The goal is to ignore certain variables and study how the performance of the model changes. Table 4 shows the performance evaluation indexes of 6 kinds of combination models, and the specific test values and predicted values are shown in Fig. 13.

Figure 13 shows the predictions of six models created according to the importance of compressive strength for each input feature. Combined with Table 4, it can be seen that with the reduction of the number of input features, the evaluation indicators of the model show an overall decline trend, with the maximum decline rate of 11.97%. This is because model a has all the input features and model e has the fewest input features. Model f also has the least number of features, but its performance is much better than model e. This is because model f has higher feature importance AE, and model e has lower feature importance. The influence of AE feature importance on the model can be compared from models d and e. Therefore, the importance and number of input features together determine the predictive performance of the model.

In engineering practice, it is crucial to quickly find important information from large amounts of data. The main goal of GRA and shapley's analysis is to accomplish this critical step by using mathematical models to assess the importance of database features. Therefore, GRA and shapley analysis are used to reduce the dimensionality of the database by reducing the number of input features. This allows the model to run faster while maintaining accuracy and identifying important data.

Model	Performance index		
	R <sup>2</sup>	RMSE	MAE
a:W/B、 SF、 W、 AE、 PC、 SP、 S/A、 FA	0.995	0.93437	0.6726
b: SF、 W、 AE、 PC、 SP、 S/A、 FA	0.99463	0.95853	0.91878
c: W/B、 SF、 W、 AE、 PC、 SP、 S/A	0.99368	1.0397	1.0809
d: AE、 PC、 SP、 S/A	0.97765	1.9551	3.8225
e: PC、 SP、 S/A	0.8758	4.6091	21.2434
f: AE、 PC、 SP	0.95254	2.8493	8.1185

**Table 4.** Combination performance of different input variables.



**Figure 13.** Results of model prediction comparison: (1)–(6).

**Conclusion**

In this paper, machine learning technology is used to study the nonlinear relationship between the HPC compressive strength and its influencing factors under large data samples, and the prediction model of its compressive strength is developed. The accuracy and performance of the model are evaluated based on a variety of statistical indexes. Finally, GRA analysis and Shapley analysis are used to verify and understand the relative importance and influence of each input feature on the output target.

1. The four machine learning models constructed in this study all have good prediction accuracy for the compressive strength of HPC. Moreover, the prediction accuracy and performance of GA-SVM model are higher than other models in both training and testing stages. The  $R^2$  value of the training phase is 0.9882, and the  $R^2$  value of the test phase is 0.995. Obviously, the accuracy and performance of the model testing phase are better than that of the training phase. This shows that the developed GA-SVM model has a strong data fitting ability and can fully capture the nonlinear relationship between features and output targets.
2. GRA and Shapley analysis together show that the importance of the characteristics of the HPC compressive strength database built in this study for the characteristics of compressive strength is sorted as follows: W/B, SF, W, AE, PC, SP, S/A, FA. This is of great significance for understanding the structure and characteristics of concrete.
3. The performance of the six models constructed by the combination of feature importance ranking shows that the performance of the model shows a decreasing trend with the decrease of the number of input features. Model (e) has the largest decline, with a decrease of up to 11.97% compared to model A. For one thing, model (e) has the fewest input features. On the other hand, AE has a high degree of contribution to model e. This confirms that the feature importance ranking analyzed in this paper is correct, and can provide help for the subsequent research on the optimization of the mix ratio of high performance concrete.
4. In the follow-up work, GA-SVM machine learning model can be used to study the durability of HPC. Then combined with NSGA-II multi-objective optimization algorithm, the mix ratio optimization design of high performance concrete is carried out. On the basis of the improvement of durability and compressive strength, the manufacturing cost of HPC concrete is reduced, which makes a certain contribution to the goal of "double carbon".

### Data availability

Since the data in this article will be used in subsequent studies, the data are not publicly available. However, it may be obtained from the corresponding author (jinlb@haut.edu.cn) at his reasonable request.

Received: 7 March 2024; Accepted: 16 July 2024

Published online: 22 July 2024

### References

1. He, H. *et al.* Research progress in mechanisms, influence factors and improvement routes of chloride binding for cement composites. *J. Build. Eng.* **86**, 108978 (2024).
2. Chang, T. P., Chuang, F. C. & Lin, H. C. A mix proportioning methodology for high-performance concrete. *J. Chin. Inst. Eng.* **19**(6), 645–655 (1996).
3. Bhanja, S. & Sengupta, B. Investigations on the compressive strength of silica fume concrete using statistical methods. *Cem. Concr. Res.* **32**(9), 1391–1394 (2002).
4. Bharatkumar, B. H. *et al.* Mix proportioning of high-performance concrete. *Cement Concr. Compos.* **23**(1), 71–80 (2001).
5. Zain, F. M. & M, M Abd S.,. Multiple regression model for compressive strength prediction of high-performance concrete. *J. Appl. Sci.* **9**(1), 155–160 (2009).
6. Bischoff, P. H. & Perry, S. H. Compressive behaviour of concrete at high strain rates. *Mater. Struct.* **24**(6), 425–450 (1991).
7. Chen, H. *et al.* An RF and LSSVM-NSGA-II method for the multi-objective optimization of high-performance concrete durability. *Cement Concr. Compos.* **129**, 104446 (2022).
8. Lessard, M., Challal, O. & Aticin, P. C. Testing high-strength concrete compressive strength. *Mater. J.* **90**(4), 303–307 (1993).
9. Shi, H., Xu, B. & Zhou, X. Influence of mineral admixtures on compressive strength, gas permeability and carbonation of high-performance concrete. *Constr. Build. Mater.* **23**(5), 1980–1985 (2009).
10. Salehi, H. & Burgueño, R. Emerging artificial intelligence methods in structural engineering. *Eng. Struct.* **171**, 170–189 (2018).
11. Huang, H. *et al.* Property assessment of high-performance concrete containing three types of fibers. *Int. J. Concr. Struct. Mater.* **15**, 1–17 (2021).
12. Chithra, S. *et al.* A comparative study on the compressive strength prediction models for high performance concrete containing nano silica and copper slag using regression analysis and artificial neural networks. *Constr. Build. Mater.* **114**, 528–535 (2016).
13. Ashrafián, A. *et al.* Prediction of compressive strength and ultrasonic pulse velocity of fiber reinforced concrete incorporating nano silica using heuristic regression methods. *Constr. Build. Mater.* **190**, 479–494 (2018).
14. Singh, A. *et al.* Utilization of antimony tailings in fiber-reinforced 3D printed concrete: A sustainable approach for construction materials. *Constr. Build. Mater.* **408**, 133689 (2023).
15. Yao, X. *et al.* AI-based performance prediction for 3D-printed concrete considering anisotropy and steam curing condition. *Constr. Build. Mater.* **375**, 130898 (2023).
16. Long, X. *et al.* Machine learning method to predict dynamic compressive response of concrete-like material at high strain rates. *Def. Technol.* **23**, 100–111 (2023).
17. Aiyer, B. G. *et al.* Prediction of compressive strength of self-compacting concrete using least square support vector machine and relevance vector machine. *KSCSE J. Civ. Eng.* **18**(6), 1753–1758 (2014).
18. Motamedi, S. *et al.* RETRACTED: Estimating unconfined compressive strength of cockle shell–cement–sand mixtures using soft computing methodologies. *Eng. Struct.* **98**, 49 (2015).
19. Pham, A. D., Hoang, N. D. & Nguyen, Q. T. Predicting compressive strength of high-performance concrete using metaheuristic-optimized least squares support vector regression. *J. Comput. Civ. Eng.* **30**(3), 06015002 (2016).
20. Omran, B. A., Chen, Q. & Jin, R. Comparison of data mining techniques for predicting compressive strength of environmentally friendly concrete. *J. Comput. Civ. Eng.* **30**(6), 04016029 (2016).
21. Zhang, J. *et al.* Modelling uniaxial compressive strength of lightweight self-compacting concrete using random forest regression. *Constr. Build. Mater.* **210**, 713–719 (2019).
22. Yuan, Z., Wang, L. N. & Ji, X. Prediction of concrete compressive strength: Research on hybrid models genetic based algorithms and ANFIS. *Adv. Eng. Softw.* **67**, 156–163 (2014).
23. Chandwani, V., Agrawal, V. & Nagar, R. Modeling slump of ready-mix concrete using genetic algorithms assisted training of artificial neural networks. *Expert Syst. Appl.* **42**(2), 885–893 (2015).
24. Yan, F. *et al.* Evaluation and prediction of bond strength of GFRP-bar reinforced concrete using artificial neural network optimized with genetic algorithm. *Compos. Struct.* **161**, 441–452 (2017).



25. Fan, D. *et al.* Precise design and characteristics prediction of ultra-high-performance concrete (UHPC) based on artificial intelligence techniques. *Cement Concr. Compos.* **122**, 104171 (2021).
26. Wang, X., Liu, Y. & Xin, H. Bond strength prediction of concrete-encased steel structures using hybrid machine learning method[C]//Structures. *Elsevier* **32**, 2279–2292 (2021).
27. Asteris, P. G. *et al.* Prediction of self-compacting concrete strength using artificial neural networks. *Eur. J. Environ. Civ. Eng.* **20**(sup1), s102–s122 (2016).
28. Baykasoğlu, A., Öztaş, A. & Özbay, E. Prediction and multi-objective optimization of high-strength concrete parameters via soft computing approaches. *Expert Syst. Appl.* **36**(3), 6145–6155 (2009).
29. Chindaprasirt, P. *et al.* Influence of fly ash fineness on the chloride penetration of concrete. *Constr. Build. Mater.* **21**(2), 356–361 (2007).
30. Elahi, A. *et al.* Mechanical and durability properties of high-performance concretes containing supplementary cementitious materials. *Constr. Build. Mater.* **24**(3), 292–299 (2010).
31. Chujié, J., Wenhua, Z. & Juan, H. Slump and strength model of high-strength concrete. *J. Southeast Univ.: Nat. Sci. Edition* **40**(S2), 144–149 (2010).
32. Lim, C. H., Yoon, Y. S. & Kim, J. H. Genetic algorithm in mix proportioning of high-performance concrete. *Cement Concr. Res.* **34**(3), 409–420 (2004).
33. Pala, M. *et al.* Appraisal of long-term effects of fly ash and silica fume on compressive strength of concrete by neural networks. *Constr. Build. Mater.* **21**(2), 384–394 (2007).
34. Prasad, B. K. R., Eskandari, H. & Reddy, B. V. V. Prediction of compressive strength of SCC and HPC with high volume fly ash using ANN. *Constr. Build. Mater.* **23**(1), 117–128 (2009).
35. Yen, T. *et al.* Influence of class F fly ash on the abrasion–erosion resistance of high-strength concrete. *Constr. Build. Mater.* **21**(2), 458–463 (2007).
36. Iqbal, M. F. *et al.* Sustainable utilization of foundry waste: Forecasting mechanical properties of foundry sand-based concrete using multi-expression programming. *Sci. Total Environ.* **780**, 146524 (2021).
37. Shi, C. *et al.* A review on ultra-high-performance concrete: Part I. Raw materials and mixture design. *Constr. Build. Mater.* **101**, 741–751 (2015).
38. Wang, D. *et al.* A review on ultra-high-performance concrete: Part II. Hydration, microstructure and properties. *Constr. Build. Mater.* **96**, 368–377 (2015).
39. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**(3), 273–297 (1995).
40. Suratgar, A. A., Tavakoli, M. B. & Hoseinabadi, A. Modified Levenberg-Marquardt method for neural networks training. *World Acad. Sci. Eng. Technol.* **6**(1), 46–48 (2005).
41. Alshihri, M. M., Azmy, A. M. & El-Bisy, M. S. Neural networks for predicting compression strength of structural light weight concrete. *Constr. Build. Mater.* **23**(6), 2214–2219 (2009).
42. Feng, D. C. *et al.* Machine learning-based compressive strength prediction for concrete: An adaptive boosting approach. *Constr. Build. Mater.* **230**, 117000 (2020).
43. Chai, T. & Draxler, R. R. Root mean square error (RMSE) or mean absolute error (MAE)—Arguments against avoiding RMSE in the literature. *Geosci. Model Dev.* **7**(3), 1247–1250 (2014).
44. Bengio, Y. & Le Cun, Y. Scaling learning algorithms towards AI. *Large-Scale Kernel Mach.* **34**(5), 1–41 (2007).
45. Hornik, K., Stinchcombe, M. & White, H. Multilayer feedforward networks are universal approximators. *Neural Netw.* **2**(5), 359–366 (1989).
46. Prusty, J. K. & Pradhan, B. Multi-response optimization using Taguchi-Grey relational analysis for composition of fly ash-ground granulated blast furnace slag based geopolymer concrete. *Constr. Build. Mater.* **241**, 118049 (2020).
47. Zhang, P., Liu, C. & Li, Q. Application of gray relational analysis for chloride permeability and freeze-thaw resistance of high-performance concrete containing nanoparticles. *J. Mater. Civ. Eng.* **23**(12), 1760–1763 (2011).
48. Zhu, L., Zhao, C. & Dai, J. Prediction of compressive strength of recycled aggregate concrete based on gray correlation analysis. *Constr. Build. Mater.* **273**, 121750 (2021).
49. Abuodeh, O. R., Abdalla, J. A. & Hawileh, R. A. Assessment of compressive strength of ultra-high-performance concrete using deep machine learning techniques. *Appl. Soft Comput.* **95**, 106552 (2020).
50. Kaloop, M. R. *et al.* Compressive strength prediction of high-performance concrete using gradient tree boosting machine. *Constr. Build. Mater.* **264**, 120198 (2020).
51. Shi, Y. *et al.* Design and preparation of ultra-high-performance concrete with low environmental impact. *J. Cleaner Prod.* **214**, 633–643 (2019).

## Acknowledgements

This research was supported by the National Natural Science Foundation of China (52178171), the Special Focus on the Development and Promotion of Henan Province (232103810080, 232102320014, 232102110266), the fund of Henan Key Laboratory of Grain and Oil Storage Facility & Safety (2023KF08), and the Innovative Funds Plan of Henan University of Technology (2022ZKCJ06). The authors express gratitude for their support.

## Author contributions

L.J.: Conceptualization, Supervision, Funding acquisition. J.D.: Methodology, Software, Writing—original draft, Validation, Investigation. Y.J.: writing—review and editing. P.X.: Funding acquisition, Resources. P.Z.: data and image processing. The author's name is correctly listed as "P.Z.". All authors have read and agreed to the published version of the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to L.J.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024