WARSAW UNIVERSITY OF TECHNOLOGY Index 351733
FACULTY OF CIVIL ENGINEERING
COMMITTEE FOR CIVIL AND WATER ENGINEERING
POLISH ACADEMY OF SCIENCES ISSN 1230-2945

DOI: 10.24425/ace.2026.1369

ARCHIVES OF CIVIL ENGINEERING

Vol. LXXII ISSUE 3 2026

pp. 93 –107

Research paper

Application of machine learning algorithm in the prediction of compressive strength of high-rise concrete

Qingzhi Zu[1], Meixia Bai[2]

Abstract: With the continuous progress of civil engineering technology, concrete has become the most widely used material in the civil engineering industry. Its concrete pressure resistance is the core indicator for measuring concrete performance. With the development of big data and artificial intelligence, machine learning algorithms are playing a significant role in concrete strength prediction. at the same time, precise predicting concrete compressive strength has important practical significance. Based on the XGBoost (Extreme Gradient Boosting) algorithm, this study has established a non-linear prediction model of concrete compressive strength. And by collecting and analyzing the setting data of 880 concrete groups, the differences in the predictive model of the four machine learning algorithms of XGBoost, Lightgbm, NGBOOST and Catboost were compared. The research results shown that the performance of models based on the XGBoost algorithm in training and test sets was better than that the other three models. The model based on the XGBoost algorithm has high predictive accuracy and good generalization capabilities. R^2, M_{AE}, and R_{MSE} based on the XGBoost algorithm models were 0.9873, 1.1707, and 1.8657, respectively. R^2, M_{AE} and R_{MSE} based on the XGBoost algorithm models were 0.9141, 3.2671, and 4.8232, respectively. The R^2, M_{AE} and R_{MSEs} obtained based on their data verification were 0.9655, 1.8952, and 2.6778, respectively. The model based on the XGBoost algorithm was excellent for predicting concrete compressive strength. A model based on the XGBoost algorithm was a reliable concrete compressive strength prediction model.

Keywords: concrete compressive strength, XGBoost (Extreme Gradient Boosting) algorithm, predictive model, model verification

[1]Prof., MSc., School of Architectural Engineering, Zhangzhou Institute of Technology, Zhangzhou 363000, China, e-mail: 15392460073@sohu.com, ORCID: 0009-0005-2995-0709

[2]MSc., Zhangzhou Institute of Technology, Zhangzhou 363000, China, e-mail: baimx2008@126.com, ORCID: 0009-0006-4393-979X

1. Introduction

With the continuous progress of civil engineering technology, concrete has become the most widely used material in the civil engineering industry [1]. Concrete compression strength is the core indicator of measurement of its mechanical properties. The compressive strength of concrete is also an important factor that has to be concerned in structural design. The accurate prediction of the compressive strength of concrete is of great significance for improving the efficiency of building materials and ensuring the safety of the building structure [2]. However, many factors affect the pressure resistance of concrete. These factors include the coordination ratio [3] and maintenance conditions between multiple materials [4]. The design method of the traditional concrete-based cooperation ratio relies on experience formulas and a large number of experiments, and it is difficult to meet the development needs of modern engineering technology. The premise of the method based on empirical formulas usually assumes that material composition and environmental factors remain unchanged. It was difficult to adapt to the changing conditions in actual projects. The method of simply relying on experimental data has problems such as high experimental costs and time-consuming [5]. Therefore, it was important to develop a scientific, efficient and precise concrete compressive strength prediction technology [6].22 With the development of big data and artificial intelligence technology, machine-learning technology provides new ideas for prediction of concrete compressive strength. Machine learning can effectively break through the limitations of traditional concrete compressive strength prediction methods by learning existing experimental data, achieving high -precision prediction of concrete compressive strength. Yeh et al. [7] proves the accuracy of the artificial neural network (ANN) in the prediction of concrete compressive strength. Compared with the return analysis model, the prediction results significantly improved the prediction accuracy. Subsequently, a variety of machine-learning models, such as support vector machine (SVM) [8], random forest (RF) [9] and gradients (GBDT) [10], have been adopted in multiple studies, showing excellent performance. Chen et al. [11, 12] have established a set of prediction models based on BP neural network-based powder coal-ash concrete compressive strength, which provides reference and guidance for related analysis. These studies have shown that machine-learning methods have a significant advantage in the process of complex non-linear problems. Although machine-learning methods have shown their superiority, different algorithms have their own advantages and disadvantages in terms of performance and application scope. At present, there was no common machine- learning model that can perform best in all cases [13].

In response to the above problems, this study aims to establish a non-linear mapping relationship between concrete compressive strength index and its influencing factors based on the XGBoost algorithm, thereby accurately predicting the compressive strength of concrete. To this end, this study has collected totally 880 sets of concrete test dataset. In this study, the coordination ratio, content and concrete maintenance period of concrete are the independent variables, and the compressive strength of the concrete is the dependent variable. Based on the XGBoost algorithm model, for multiple indicators, this study compares the predictive performance of different gradient improvement models. Finally, based on its own test data, this study uses the best prediction effect based on the XGBoost algorithm-based model to predict and verify the concrete compressive strength index. XGBoost models provide an efficient nonlinear modeling tool in civil engineering, and through their methodological approach, they drive advancements in concrete strength prediction, enabling applications such as risk warning for high-rise buildings, concrete strength prediction, and decision support, thereby significantly enhancing the safety and sustainability of civil engineering.

1.1. Data analysis

The dataset used in this study is derived from the concrete compressive strength test data provided by Yeh [7] and Rafat et al. [14]. There are 880 sets of data in total. This study takes eight parameters that affect the compressive strength of concrete, namely, cement, slag, fly ash, water, water reducer, coarse aggregate, fine aggregate, and age, the curing age is set at 28 d, The curing condition of the concrete is standard curing. These eight parameters are used as input data. The predicted value of concrete compressive strength was used as output data. The characteristic values of the dataset are shown in Table 1.

Table 1. Descriptive statistics of the dataset

Parameter	Unit	Type	Minimum	Maximum	Mean	Median
cement	kg/m^3	Input	102	540	280.4	272.6
slag	kg/m^3	Input	11	359.4	136.2	135.7
fly ash	kg/m^3	Input	24	261	125	121.9
water	kg/m^3	Input	16.56	247	174	181.7
water reducer	kg/m^3	Input	0.09	32.2	8.8	8.9
coarse aggregate	kg/m^3	Input	621	1145	964.6	965.4
fine aggregate	kg/m^3	Input	478	1079	777.9	780.6
age	day	Input	1	365	44.4	28
compressive strength	MPa	Output	2.33	82.6	36	34.68

The input data and output data are subjected to correlation analysis to obtain the thermal correlation between the variables, as shown in Fig. 1. The degree of correlation between the variables is shown by the color depth of the color block in Fig. 1.

As shown in Fig. 1, the correlation coefficient between cement and concrete compressive strength reached 0.49. This proves that there was a certain degree of positive correlation between the two. The correlation coefficient between water and fine aggregate reached -0.42, which shows that there was a certain degree of negative correlation between the two variables. Overall, the correlation coefficient between the variables was low, indicating that there was no obvious overlap in the characteristics between the variables. This data set can effectively avoid the risk of the prediction model's prediction effect being reduced due to over-reliance on a certain variable.

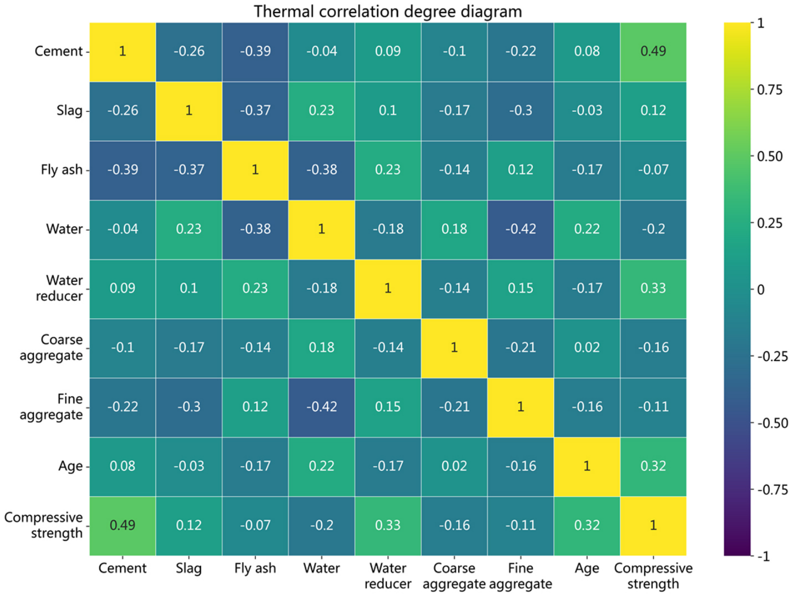

Fig. 1. Thermodynamic correlation between input and output data Thermodynamic correlation diagram

1.2. Data preprocessing

The total of 880 data are randomly divided using the 8:2 principle. In other words, 704 data were training sets for training prediction models. 176 data are test sets for evaluating prediction models. A 5-fold cross-validation was used for the training set, which can effectively reduce the evaluation error and over fitting caused by the randomness of data division.

In order to avoid deviations in model training due to different feature value ranges of feature variables at the same scale, this study uses the Standard Scaler function of the Sklearn Preprocessing software package in the Python language to standardize the feature variables and target quantities so that their mean was 0 and the standard deviation was 1. Accordingly, the calculation formula for data preprocessing is shown in Eq. (1.1).

$$x' = \frac{x - \mu}{\sigma} \tag{1.1}$$

In the formula, x is the value after standardization, x is the current value, μ is the sample mean, σ is the sample standard deviation.

1.3. XGBoost algorithm

The XGBoost (eXtreme Gradient Boosting) algorithm has many parameters that need to be adjusted. Improper parameter settings may lead to over fitting or under fitting. The XGBoost algorithm was sensitive to outliers. If there were many outliers in the data, the prediction performance of the model may decline. Therefore, in the data preprocessing stage, it was necessary to preprocess the data outliers. The quality of the data has a great impact on the prediction performance of the model. Therefore, it was necessary to ensure the accuracy and completeness of the data in the data collection and processing stage. As shown in Fig. 2.

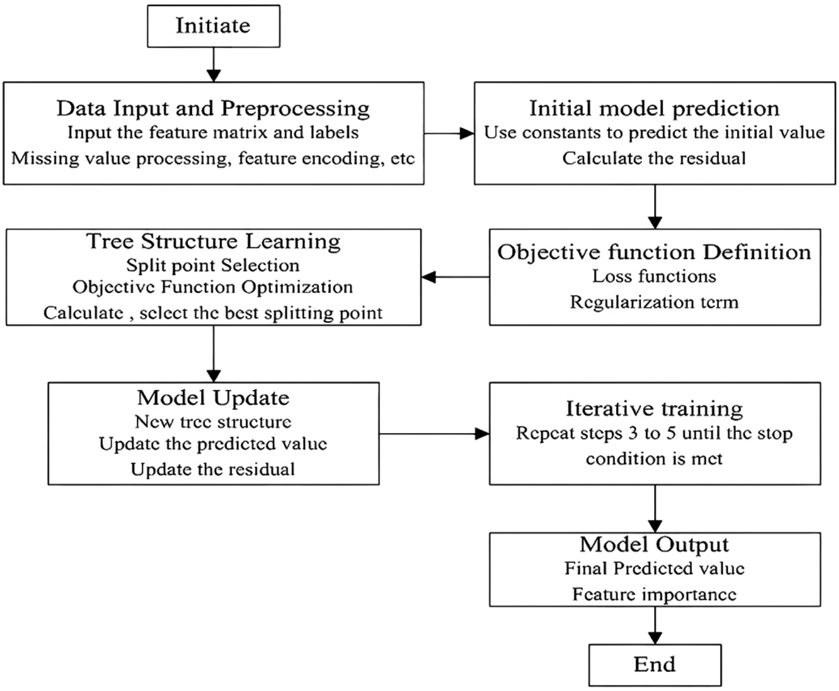

Fig. 2. The algorithm diagram for calculating the XGBoost of pipeline

2. Model establishment

2.1. XGBoost principle

In 2016, The paper [14] proposed the XGBoost (eXtreme Gradient Boosting) algorithm based on the regression tree boosting algorithm. It modifies the previous model by gradually building a new model. In each round of correction, a new model was obtained by fitting the negative gradient of the loss function. The main advantage of the XGBoost algorithm is that regularization is introduced in the objective function to adjust the complexity of the model and avoid overfitting. In addition, the objective function was simplified by the second-order Taylor expansion, and the gradient change trend is further considered to achieve faster and more accurate fitting.

The model established based on the XGBoost (eXtreme Gradient Boosting) algorithm consists of multiple CART decision trees. Each decision tree has its own structure and leaf node weight. The output is the sum of multiple trees. The simplified training process is shown in Fig. 3.

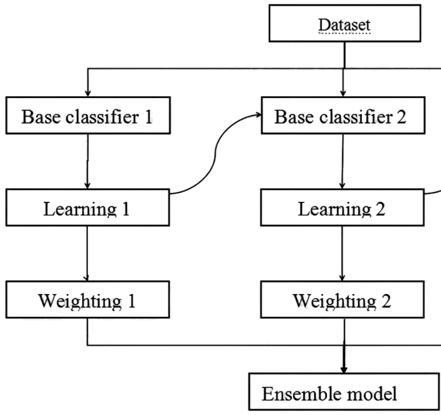

Fig. 3. Training process of the model based on the XGBoost (eXtreme Gradient Boosting) algorithm

Assume that the dataset sample used in training is (x_i, y_i), where $x_i \in R^m, y_i \in R$, then the expression of the XGBoost model was shown in Eq. (2.1).

$$\widehat{y} = F_K(x_i) = F_{K-1}(x_i) + f_K(x_i) \tag{2.1}$$

Where, $\widehat{y}$ is the output set of the XGBoost model, $f_K(x)$ is the Kth decision tree.

The expression of the space F of the CART decision tree is calculated by Eq. (2.2).

$$F = \left\{ f(x) = w_q(x) \right\} \left(q : R^m \rightarrow T, w \in R^T \right) \tag{2.2}$$

Among them, $q : R^m \rightarrow T$ represents the function of the decision tree to map sample data to leaf nodes. T is the number of leaf nodes. And is the score on leaf node i.

To build an XGBoost model, it was necessary to determine the structure and weight of each decision tree. This process requires an objective function to guide the optimization of model training. The objective function of the XGBoost model consists of two parts: the loss function and the regularization term. The expression of the objective function is calculated by Eq. (2.3).

$$Obj = \sum_{i=1}^{n} L(y_i, \widehat{y}_i) + \sum_{K=1}^{K} \Omega(f_k) \tag{2.3}$$

Where, $y_i, \widehat{y}_i$ are the i-th true value and predicted value respectively, L is the loss function, usually mean square error (MSE) or logarithmic loss. Ω is the regularization term used to control the complexity of the model, and its expression is shown in Eq. (2.4).

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|\omega\|^2 \tag{2.4}$$

In the formula, γ, λ is the regularization parameter, which is used to control the complexity of the tree.

The basic idea of the XGBoost algorithm was to rely on the optimization result of the previous step to train the next step. That is, each iteration adds a sub-model on the basis of the previous iteration $f(x_i)$, so as to continuously make the predicted value close to the actual value. After s iterations, the objective function reaches the minimum value. The objective function can be expressed as Eq. (2.5).

$$Obj^{(s)} = \sum_{i=1}^{n} L\left(y_y, \widehat{y}_i^{(s-1)} + f_s(x_i)\right) + \Omega(f_s) \tag{2.5}$$

In order to simplify the objective function, the loss function is expanded by a second-order Taylor to obtain the following Eq. (2.6).

$$Obj^{(s)} = \sum_{i=1}^{n} L\left(y_y, \widehat{y}_i^{(s-1)} + g_i(x_i) + \frac{1}{2} h_i f_s^2(x_i)\right) + \Omega(f_s) \tag{2.6}$$

Where, $g_i = \dfrac{\partial L\left(y_i, \widehat{y}_i^{(s-1)}\right)}{\partial \widehat{y}_i^{(s-1)}}$ is the first-order derivative (gradient); $h_i = \dfrac{\partial^2 L\left(y_i, \widehat{y}_i^{(s-1)}\right)}{\partial \widehat{y}_i^{(s-1)}}$ is the second-order derivative (Hessian).

2.2. *k*-fold cross validation

To avoid the low data utilization caused by randomly dividing the data set by 8:2. This study uses k-fold cross validation ($k = 5$) to use each part of the data for training and testing to make full use of the data, effectively avoid over fitting, and ensure the generalization ability of the model. The main principles of k-fold cross validation are as follows.

In k-fold cross validation ($k = 5$), the original data set was evenly divided into five non-overlapping subsets. In each iteration, one of the subsets was selected as the validation set, and the remaining four subsets are used as training sets. In this way, the model will be trained and validated five times. Each validation uses a different subset as the validation set. Finally, the results of these five validations were averaged to obtain a more reliable and accurate model performance evaluation indicator.

3. Experimental results and analysis

3.1. Evaluation indicators

This study uses quantitative R^2 (coefficient of determination), M_{AE} (mean absolute error) and R_{MSE} (root mean square error) to evaluate the accuracy of the prediction model. The formula was shown in Eq. (3.1), Eq. (3.2) and Eq. (3.3).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{3.1}$$

$$M_{\mathrm{AE}} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3.2}$$

$$R_{\mathrm{MSE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.3}$$

Where, y_i is the i-th measured value, $\hat{y}_i$ is the i-th predicted value, $\overline{y}$ is the mean of the measured values, and n is the number of samples.

The coefficient of determination (R^2) reflects the model's ability to explain the data, and its value ranges from 0 to 1. The closer the R^2 value is to 1, the better the model fits the data. The coefficient of determination indicates how much of the total variation in the actual values can be explained by the model. The mean absolute error (M_{AE}) was the average of the absolute differences between the actual values and the predicted values. M_{AE} provides an intuitive measure of error and can directly reflect the degree of deviation between the predicted values and the actual values. The smaller the M_{AE} value, the better the model's prediction effect. The root mean square error (R_{MSE}) is the square root of the mean square error. The root mean square error (R_{MSE}) had the same dimension as the original data. The smaller the root mean square error (R_{MSE}) value, the better the model's prediction effect. The root mean square error (R_{MSE}) can be regarded as the standard deviation of the average prediction error, so it can provide a more intuitive measure of the error distribution.

3.2. Comparison of concrete performance

To verify the accuracy of the XGBoost algorithm, this study will train and test the training set and test set based on the CatBoost, LightGBM, and NGBoost model reading partitions. The prediction results are compared with those of the model based on the XGBoost algorithm. The four prediction models all use the same data preprocessing method and five-fold cross- validation method to ensure the generalization ability of the model. Fig. 4 shown Comparison of prediction performance between XGBoost model and CatBoost, LightGBM, and NGBoost models.

As can be seen from Fig. 4, the CatBoost model had a general fitting effect on the training set, with R^2 reaching 0.9550, the point distribution was relatively concentrated, and its predicted value and actual value were in good agreement. This shows that the CatBoost model had a strong predictive ability on the training data. However, on the test set, R^2 is slightly lower at 0.8964, indicating that the generalization ability of the model was poor. The LightGBM model has a good fitting effect on the training set. The R^2 value was 0.9673, and its predicted value and actual value were in good agreement. This shows that the LightGBM model could capture the characteristics of the training data well. However, the R^2 value on the test set is only 0.8974, which was quite different from the training set. This shows that the model's generalization ability for new data needs to be improved. The NGBoost model had a relatively poor fitting effect on both the training set and the test set. The R^2 value was significantly lower than other models. This shown that the overall prediction accuracy of the NGBoost model was not as good as other models. The XGBoost model had the best fitting effect on the training set, with an R^2 value of 0.9865, and its predicted value and actual value were in good agreement. Therefore, the effect of fitting the training data was the best. At the same time, the R^2 value on the test set also reached 0.9067, in which shown that the performance of the XGBoost model on both the training set and the test set was better than the prediction effect of other models, As shown in Table 2, The XGBoost model had high prediction accuracy and good generalization ability, The application of the XGBoost model were also discussed by Zhou et al. [16], Zhao et al. [17], and Nguyen Sy [18].

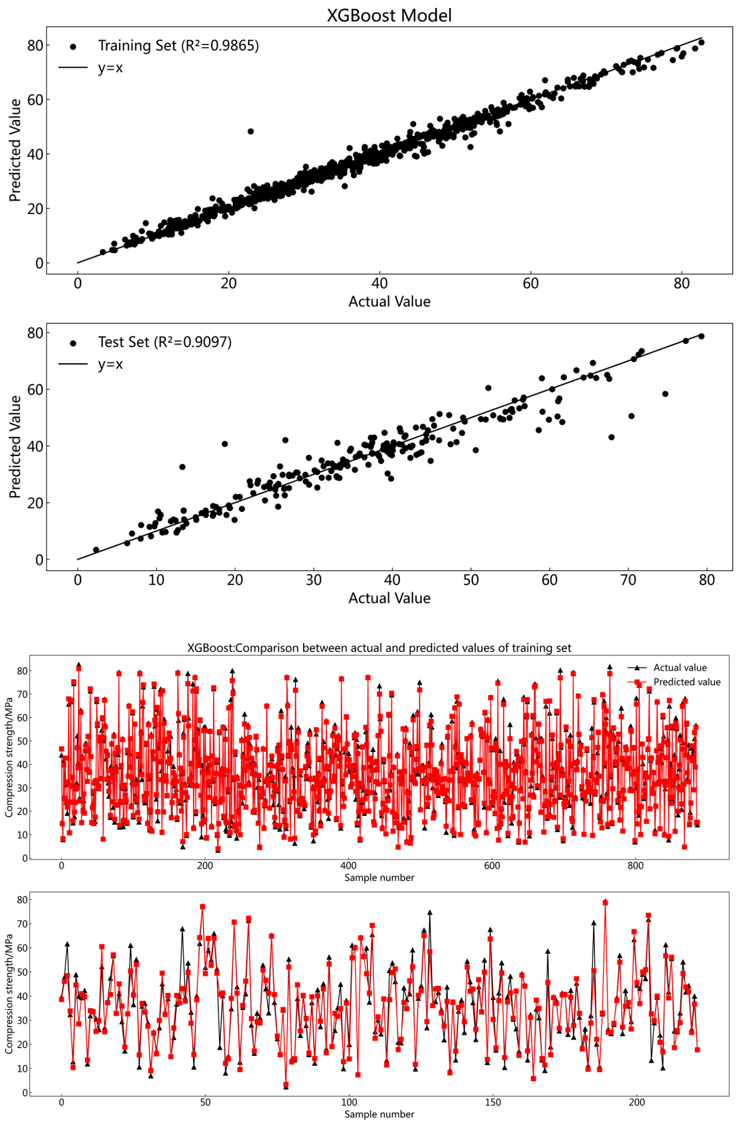

Fig. 4. Comparison of prediction performance between XGBoost model and CatBoost, LightGBM, and NGBoost models

Table 2. Evaluation index values of XGBoost model and CatBoost, LightGBM, and NGBoost model

Category	Training set			Test Set		
Index	$R2$	M_{AE}	R_{MSE}	$R2$	M_{AE}	R_{MSE}
XGBoost	0.9877	1.2126	1.9219	0.9179	3.2825	4.9435
CatBoost	0.9455	3.0567	4.0047	0.8946	3.8036	5.2863
LightGBM	0.9763	2.0686	2.9381	0.8974	3.5586	5.4125
NGBoost	0.9227	3.9403	5.1918	0.8742	4.6123	6.2971

The predicted data and actual data of each model were respectively put into the above Eq. (3.1), Eq. (3.2) and Eq. (3.3) to calculate the evaluation index, and the prediction accuracy of the four models from high to low was as follows.

1. The M_{AE} and R_{MSE} values of the XGBoost model are 1.2215 and 1.9209 on the training set, and 3.2825 and 4.9435 on the test set, respectively, which were lower than the other three models. This shows that the XGBoost model has the highest accuracy and stability in the prediction of concrete compressive strength, and has the strongest generalization ability.
2. The M_{AE} and R_{MSE} of the LightGBM model on the training set were 2.0686 and 2.9381, respectively, and the M_{AE} and R_{MSE} on the test set are 3.5586 and 5.4125, respectively. This shows that the error of the LightGBM model on the test set is relatively large, especially the root mean square error (R_{MSE}) value was high. This shows that the LightGBM model has insufficient prediction performance when dealing with some extreme data points.
3. The M_{AE} and R_{MSE} of the CatBoost model on the training set are 3.0567 and 4.0047, respectively, while the M_{AE} and R_{MSE} on the test set were 3.8036 and 5.2863. This shows that the error of the model on the test set was larger than that on the training set, and the generalization ability of the model needs to be improved;
4. The M_{AE} and R_{MSE} of the NGBoost model on the training set were 3.9403 and 5.1918, respectively, while the M_{AE} and R_{MSE} on the test set were 4.6123 and 6.297, respectively. This shows that the errors of the NGBoost model on both the training set and the test set were large, and the prediction performance and generalization ability were weak.

3.3. Performance comparison after Optuna hyperparameter optimization

To further enhance the performance of the XGBoost model in predicting the compressive strength of concrete, this study utilized Optuna to optimize key hyperparameters of the XGBoost model. Before commencing hyperparameter tuning, it was necessary to outline the range of values for the hyperparameters to be adjusted. Various combinations of different values for multiple hyperparameters will ultimately form a parameter space. The specific hyperparameters optimized and their corresponding ranges were presented in Table 3.

Based on the search ranges outlined in Table 3, the optimal hyperparameters obtained using the Optuna hyperparameter optimization algorithm were presented in Table 4.

Table 3. Optuna hyperparameter tuning range

Number	hyper-parameter	Tuning range
1	max_depth	[3,15]
2	learning_rate	[0.01,0.4]
3	n_estimators	[50,310]
4	subsample	[0.8,1.6]
5	colsample_by tree	[0.7,1.5]
6	min_weight	[2,12]

Table 4. Main parameters of XGBoost model after tuning

Number	hyper-parameter	Optimal value
1	max_depth	9
2	learning_rate	0.14362
3	n_estimators	177
4	subsample	0.79872
5	colsample_by tree	0.91551
6	min_weight	3.5

Incorporate the optimized key parameters into the XGBoost model and compare its performance with the pre-optimization model, As shown in Fig. 5.

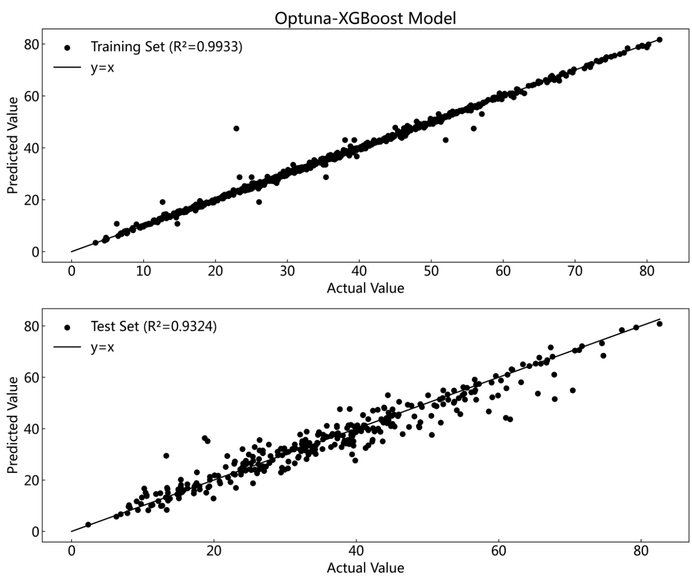

Figure continued on the next page

Figure continued from the previous page

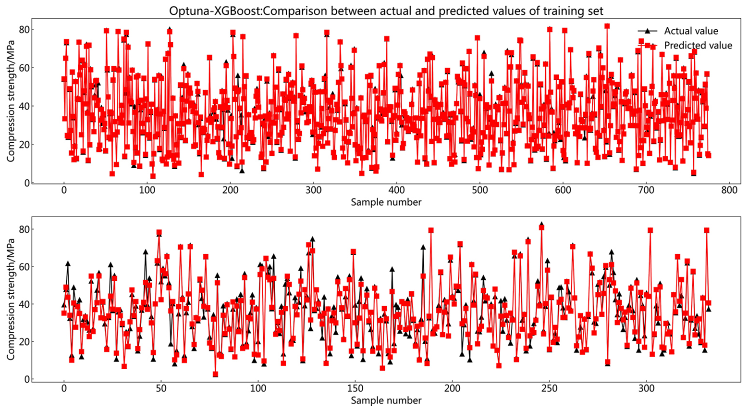

Fig. 5. The test set for the Optuna-optimized XGBoost model

From Fig. 5, the XGBoost model optimized with Optuna demonstrates a better model fit. On the training set, the R^2 of the optimized XGBoost has increased by 0.0071, with the data points being highly concentrated, indicating accurate prediction results. On the test set, the R^2 of the optimized XGBoost has improved by 0.0322, suggesting a significant enhancement in the model's generalization ability. The evaluation metrics were calculated by substituting the training and test data of both models into Eq. (3.2), and Eq. (3.3), the results were presented in Table 5.

Table 5. Metrics for evaluation of optimized XGBoost vs

Categorical Indicator	Training set			Test set		
	R^2	M_{AE}	R_{MSE}	R^2	M_{AE}	R_{MSE}
Optuna-XGBoost	0.9938	0.4527	1.3170	0.9365	2.8998	4.2343
XGBoost	0.9866	1.2226	1.9211	0.9098	3.2876	4.9461

As shown in Table 5, the XGBoost model optimized with the Optuna algorithm had achieved a reduction of approximately 63.03% in M_{AE} and about 31.44% in R_{MSE} on the training set. On the test set, the M_{AE} has decreased by around 11.76%, and the R_{MSE} had decreased by approximately 14.4%. These results indicate that the XGBoost model optimized by Optuna has significantly improved its predictive ability and generalization capability on both the training and test sets.

3.4. Example verification

To further verify the generalization ability of the XGBoost model established above, this study inputs the independent variable's own dataset into the trained XGBoost prediction model to evaluate the prediction performance of the XGBoost model. The Table 6 shows the basic information of the dataset.

Table 6. Descriptive statistics of the own dataset

Parameter	Unit	Type	Minimum	Maximum	Mean
cement	kg/m^3	136.5	541	329.6	311.5
slag	kg/m^3	59.2	169.4	118.6	124.9
fly ash	kg/m^3	141	229	182.8	187
water	kg/m^3	2.1	28.5	10.3	9.92
water reducer	kg/m^3	802	1126	998.5	995.7
coarse aggregate	kg/m^3	78.7	906.4	780	785
fine aggregate	Day	1.2	366	48.3	28.1
age	MPa	6.38	80.2	35.8	34.7

As can be seen from Table 6, this dataset is somewhat different from the dataset provided by Yeh [7], which can effectively verify the generalization ability of the model, as shown in Fig. 6.

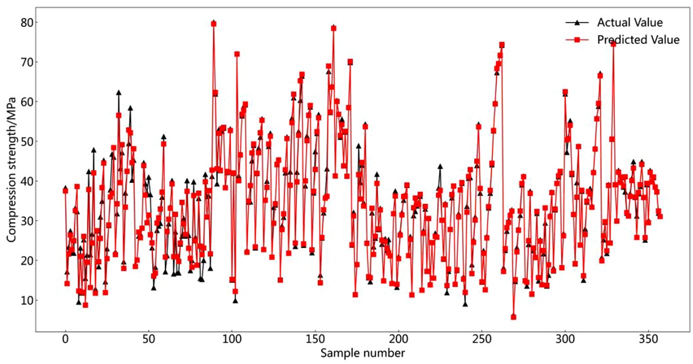

Fig. 6. Comparison of predicted values and actual values on our own dataset

From Fig. 6, the XGBoost model has a high degree of fit between the prediction curve and the actual value curve on its own data set. Specifically, the model performs well in various evaluation indicators. the results were presented in Table 7.

Table 7. Metrics for evaluation of optimized XGBoost vs

Categorical Indicator	Training set			Test set		
	R^2	M_{AE}	R_{MSE}	R^2	M_{AE}	R_{MSE}
XGBoost	0.9098	3.2876	4.9461	0.9826	1.3263	2.2172
CatBoost	0.9455	3.0567	4.0047	0.8946	3.8036	5.2863
LightGBM	0.9763	2.0686	2.9381	0.8974	3.5586	5.4125
NGBoost	0.9227	3.9403	5.1918	0.8742	4.6123	6.2971

These values were significantly lower than other models. The above results show that the concrete compressive strength prediction model based on the XGBoost algorithm constructed in this paper has strong generalization ability and application value.

4. Conclusions

The concrete compressive strength prediction model based on the XGBoost algorithm has been proven to have excellent prediction performance and stability in multiple application scenarios. By integrating multiple decision trees and using the gradient boosting framework for iterative optimization, the model can maintain high efficiency when processing large- scale data sets. The regularization parameters and parallel computing technology provided by the XGBoost algorithm enable the model to perform well in preventing overfitting and improving training speed. The study established a nonlinear prediction model for concrete compressive strength based on the XGBoost algorithm. Through the analysis of 880 sets of concrete test data, the model based on the XGBoost algorithm significantly outperformed the models based on the LightGBM, NGBoost and CatBoost algorithms in terms of prediction accuracy. The verification results based on the own data show that the R^2, M_{AE} and R_{MSE} values of the XGBoost model are 0.9762, 1.3136 and 2.2227 respectively. This further confirms the effectiveness of the XGBoost model prediction and good generalization ability. With the rapid advancement of the Internet of Things (IoT) and intelligent manufacturing technologies, real-time prediction and intelligent optimization have emerged as critical application requirements. By integrating an XGBoost-based concrete compressive strength prediction model with a real-time data acquisition system, real-time monitoring and prediction of concrete compressive strength can be achieved. This will provide robust support for quality control during concrete production and construction processes, thereby further enhancing the reliability and safety of concrete engineering projects. It should be noted that the freeze- resistance (frost resistance) parameter was not considered among the parameters selected in this study, and it will be included as an indicator for further research in the future. Furthermore, explore and construct more features related to concrete compressive strength, such as concrete composition proportions, curing conditions, etc., and screen out the optimal feature subset through feature selection methods.

References

[1] H. Xu, "Analysis of the influence of admixtures on the mechanical properties of concrete", *Technological Innovation and Application*, vol. 21, no. 15, pp. 55–56, 2020.

[2] M.C. Li, M.X. Zhang, and J.R. Zhang, "Research on deformation characteristics and shear strength criteria of normal state-roller compacted concrete combined dam construction materials", *Journal of Hydraulic Engineering*, vol. 50, no. 2, pp. 157–164, 2019.

[3] J. Liu, H.Yu, and B. X. Li, "Research on the influence of mineral admixtures on the durability of PC box girder concret", *Transportation Science and Engineering*, vol. 34, no. 3, pp. 27–30, 2018.

[4] J. Blazy, Ł. Drobiec, and P. Wolka, "Mechanical properties of polymer fibre reinforced concrete in the light of various standards", *Archives of Civil Engineering*, vol. 70, no. 2, pp. 323–342, 2024, doi: 10.24425/ace.2024.149866.

[5] S. Debbarma and R.N.G.D. Ransinchung, "Using artificial neural networks to predict the 28-day compressive strength of roller- compacted concrete pavements containing RAP aggregates", *Road Materials and Pavement Design*, vol. 23, no. 1, pp. 149–167, 2022, doi: 10.1080/14680629.2020.1822202.

[6] M.S. Barkhordari, D.J. Armaghani, A.S. Mohammed, and D.V. Ulrikh, "Data-driven compressive strength prediction of fly ash concrete using ensemble learner algorithms", *Buildings*, vol. 12, no. 2, art. no. 132, 2022, doi: 10.3390/buildings12020132.

[7] K.O. Oriola, A.A. Raheem, and A.B. Ogundele, "Investigation of compressive strength and thermal properties of corn cob ash cement concrete", *Materials Today: Proceedings*, vol. 86, pp. 128–133, 2023, doi: 10.1016/j.matpr.2023.04.520.

[8] Z.H. Lim, F.W. Lee, K.H. Mo, J.H. Lim, M.K. Yew, and K.Z. Kwong, "Compressive strength forecasting of air-entrained rubberized concrete during the hardening process utilizing elastic wave method", *Crystals*, vol. 10, no. 10, art. no. 912, 2020, doi: 10.3390/cryst10100912.

[9] X.N. Cui, Q.C. Wang, M.L. Zhang, J.P. Dai, and C. Xie, "Prediction of compressive strength of high performance concrete based on random forest", *Journal of Lanzhou Jiaotong University*, vol. 40, no. 6, pp. 1–6+14, 2022, doi: 10.3969/j.issn.1001-4373.2021.06.001.

[10] A.A. Alabdullah, M. Iqbal, M. Zahid, K. Khan, M.N. Amin, and F.E. Jalal, "Prediction of rapid chloride penetration resistance of metakaolin based high strength concrete using light GBM and XGBoost models by incorporating SHAP analysis", *Construction and Building Materials*, vol. 345, art. no. 128296, 2022, doi: 10.1016/j.conbuildmat.2022.128296.

[11] H.A. Alkadhim, et al., "Prediction of rapid chloride penetration resistance of metakaolin based concrete using multi-expression programming", *Science of Advanced Materials*, vol. 14, no. 8, pp. 1348–1360, 2022, doi: 10.1166/sam.2022.4341.

[12] P. Kumar, B. Pratap, S. Sharma, and I. Kumar, "Compressive strength prediction of fly ash and blast furnace slag-based geopolymer concrete using convolutional neural network", *Asian Journal of Civil Engineering*, vol. 25, no. 2, pp. 1561–1569, 2024, doi: 10.1007/S42107-023-00861-5.

[13] A. Kumar, S. Sen, S. Sinha, B. Kumar, and C. Nidhi, "Prediction of compressive strength of concrete doped with waste plastic using machine learning-based advanced regularized regression models", *Asian Journal of Civil Engineering*, vol. 26, no. 4, pp. 1723–1741, 2025, doi: 10.1007/S42107-025-01280-4.

[14] R. Siddique, P. Aggarwal, and Y. Aggarwal, "Prediction of compressive strength of self-compacting concrete containing bottom ash using artificial neural networks", *Advances in Engineering Software*, vol. 42, no. 10, pp. 780–786, 2011, doi: 10.1016/j.advengsoft.2011.05.016.

[15] T.Q. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system", in *KDD'16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: ACM, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[16] X.X. Zhou, F. Duan, W.H. Liu, C. Shi, B.Q. Xie, and Z.Q. Yao, "Research on prediction of rheological properties of steel- polypropylene fiber UHPC mortar based on ensemble machine learning", *Construction and Building Materials*, vol. 472, art. no. 140833, 2025, doi: 10.1016/j.conbuildmat.2025.140833.

[17] X. Zhao, P.F. Zhang, D.X. Zhang, Q. Zhao, and Y. Tuerxunmaimaiti, "Prediction of interlaminar shear strength retention of FRP bars in marine concrete environments using XGBoost model", *Journal of Building Engineering*, vol. 105, art. no. 112466, 2025, doi: 10.1016/j.jobe.2025.112466.

[18] T. Nguyen-Sy, "Optimized hybrid XGBoost-CatBoost model for enhanced prediction of concrete strength and reliability analysis using Monte Carlo simulations", *Applied Soft Computing*, vol. 167, Part C, pp. 112490, 2024, doi: 10.1016/j.asoc.2024.112490.

Received: 2025-02-24, Revised: 2025-04-26