buildings

MDPI

*Article*

# Compressive Strength Prediction of BFRC Based on a Novel Hybrid Machine Learning Model

**Jiayan Zheng [1,*], Tianchen Yao [1], Jianhong Yue [2], Minghui Wang [1] and Shuangchen Xia [2]**

[1] School of Civil Engineering, Chongqing Jiaotong University, Chongqing 400074, China; ytc2497220206@163.com (T.Y.); 17839217509@126.com (M.W.)
[2] Sichuan Chengqiongya Expressway Co., Ltd., Chengdu 610000, China; yjh20230010728@163.com (J.Y.); Churze@outlook.com (S.X.)
* Correspondence: jiayanzheng@cqjtu.edu.cn

**Abstract:** Basalt fiber-reinforced concrete (BFRC) represents a form of high-performance concrete. In structural design, a 28-day resting period is required to achieve compressive strength. This study extended an extreme gradient boosting tree (XGBoost) hybrid model by incorporating genetic algorithm (GA) optimization, named GA-XGBoost, for the projection of compressive strength (CS) on BFRC. GA optimization may reduce many debugging efforts and provide optimal parameter combinations for machine learning (ML) algorithms. The XGBoost is a powerful integrated learning algorithm with efficient, accurate, and scalable features. First, we created and provided a common dataset using test data on BFRC strength from the literature. We segmented and scaled this dataset to enhance the robustness of the ML model. Second, to better predict and evaluate the CS of BFRC, we simultaneously used five other regression models: XGBoost, random forest (RF), gradient-boosted decision tree (GBDT) regressor, AdaBoost, and support vector regression (SVR). The analysis results of test sets indicated that the correlation coefficient and mean absolute error were 0.9483 and 2.0564, respectively, when using the GA-XGBoost model. The GA-XGBoost model demonstrated superior performance, while the AdaBoost model exhibited the poorest performance. In addition, we verified the accuracy and feasibility of the GA-XGBoost model through SHAP analysis. The findings indicated that the water–binder ratio (W/B), fine aggregate (FA), and water–cement ratio (W/C) in BFRC were the variables that had the greatest effect on CS, while silica fume (SF) had the least effect on CS. The results demonstrated that GA-XGBoost exhibits exceptional accuracy in predicting the CS of BFRC, which offers a valuable reference for the engineering domain.

**Keywords:** BFRC; compressive strength; genetic algorithm; machine learning



**Citation:** Zheng, J.; Yao, T.; Yue, J.; Wang, M.; Xia, S. Compressive Strength Prediction of BFRC Based on a Novel Hybrid Machine Learning Model. *Buildings* **2023**, *13*, 1934. https://doi.org/10.3390/buildings13081934

Academic Editor: Rajai Zuheir Al-Rousan

Received: 2 July 2023
Revised: 22 July 2023
Accepted: 27 July 2023
Published: 29 July 2023



## 1. Introduction

Concrete is the most widely used material in the building-structure industry, but the demand for its performance is increasing, thus the trend of research and application of advanced and eco-materials is getting more and more attention and development. Many methods have been proposed to improve the performance of concrete materials, and the incorporation of fibers into concrete is a promising solution. Among them, basalt fiber (BF) is an inorganic and environment-friendly green material that is often used as a concrete additive to improve its tensile properties [1–4]. Basalt-fiber-reinforced concrete (BFRC) has the advantage of superior mechanical properties, durability, crack resistance, and frost resistance and the raw material is readily available, which makes its application promising. It is essential to accurately predict the basic mechanical performance of BFRC to ensure the safety of engineering structures.

Wang D et al. [5] investigated the impacts of BF, polypropylene fibers, and blended fibers on the compressive strength (CS), flexural strength, splitting tensile strength, and stress–strain curves of HPC. The results indicated that fiber admixture could moderately

enhance the CS of concrete. Wang X et al. [6] investigated the effects of fiber length and admixture amount on the mechanical properties and crack resistance of BFRC. The results demonstrated that adding the appropriate amount of fiber can increased the CS of BFRC. Chen et al. [7] conducted a comprehensive study on the influence of BF content on basic mechanical properties of concrete. The investigation involved performing compression and cracking tensile experiments. The results demonstrated that BF can enhance the CS of concrete by a moderate amount. Most studies have evaluated the mechanical behavior of BFRC by using experimental methods [8–13]. However, BFRC is a mixture of a variety of materials in precise proportions such as basalt fibers, cement, water and aggregate. If the test results are not sufficiently reliable due to mixture design failures, the entire concrete-mixture design procedure has to be restarted. However, experimental studies require a large amount of instrumentation and equipment, which is costly.

In order to surpass the limitations associated with experimental methods, several researchers have turned to machine learning (ML) methods as a means to predict concrete strength [14–16]. Hong et al. [17] verified that the RF algorithm can predict the compressive strength of BFRC well. Severcan et al. [18] employed GEP to effectively capture the splitting tensile strength of concrete. Their study demonstrated that GEP exhibited superior performance in predicting the splitting tensile strength compared to other approaches. Nguyen et al. [19] used RF, decision trees (DTs), and XGBoost for the prediction of CS on fly-ash-based polymer concrete and concluded that the XGBoost model outperformed the other two models. In a study by Gupta et al. [20], the optimal proportions of concrete mixes were determined using the Gaussian process, M5P model, random forest (RF), and random tree (RT) techniques and different applied models were evaluated. It was found that Gaussian process regression using RBF kernel produced better results than other models. Asteris et al. [21] used AdaBoost, support vector machine (SVM), RF, DT, and K nearest neighbors to evaluate the behavior of cement-based mortar CS. Kang et al. [22] extended and compared various ML models to forecast flexural strength of steel-fiber-reinforced concrete. Finally, it was found that the CS prediction performance was generally better than the flexural-strength prediction performance regardless of the machine-learning algorithm used. A forest deep neural network (FDNN) algorithm proposed by Altayeb et al. [23] outperformed previous algorithms in predicting the mechanical properties of cementitious composites. Armaghani et al. [24] used models such as artificial neural networks (ANNs), and ANFIS to forecast the CS of cementitious mortar materials with or without biased kaolinite. Ahmed et al. [25] investigated the CS of concrete mixed with fly ash using various techniques such as linear regression, genetic algorithms, and particle-swarm optimization. Nazar et al. [26] used random forest regression, DTs and GEP-model multiple machine-learning approaches for the prediction and valuation of CS in nano-modified concrete and showed that the RFT model performed better in terms of accuracy and precision of the results. Esmaeili and Benemaran [27] used particle-swarm optimization (PSO) and black-widow optimization algorithms (BWOAs) to optimally determine the variables to generate two XGB structures that predicted the modulus of elasticity (MR) of the modified substrate under wet and dry cycling conditions. The results showed that the combination of variables in the M3 model were the most appropriate and that the BWO algorithm was competent in determining the optimal values of the XGB parameters. Benemaran et al. [28] used four different optimization methods (particle-swarm optimization, social-spider optimization, sinusoidal cosine algorithm, and multi-universe optimization) based on the extreme gradient-boosting model to predict the modulus of recovery of flexible pavement foundations. The results showed that all optimization models worked well, but the limit gradient-boosting model based on particle-swarm optimization exhibited the best prediction accuracy. Li et al. [29] developed four hybrid models based on the aquila-optimizer algorithm using an adaptive neuro-fuzzy inference system, support vector regression, random forests, and limiting gradient boosting to accurately predict the unconfined compressive strength of marine clay and recycled tile blends. The results of the study showed

that the model of extreme gradient boosting with the aquila-optimizer algorithm had the best performance with a small scatter index and better generalization.

In summary, existing studies have mainly focused on single-scale models without considering the effects of extensive laboratory-work data or factors in the concrete CS. Single algorithms often lead to locally optimal solutions that are difficult to generalize which can ultimately limit their practical application. Many factors influent the CS of concrete, including the mix ratio, water–cement ratio, and shape of the mixture. Therefore, designing a hybrid model integrating multiple algorithms can consider these factors more comprehensively and enhance prediction accuracy and generalizability of concrete CS. Genetic algorithms (GAs) have good global search capability and robustness and can avoid getting trapped in locally optimal solutions. XGBoost is a powerful integrated learning algorithm with efficient, accurate, and scalable features. However, its grid search increases the time significantly as the parameters increase when performing parameter tuning. Therefore, a hybrid model was designed to extend the XGBoost model by using genetic-algorithm parameter optimization (GA-XGBoost) for better and faster prediction of CS on BFRC as well as optimization of concrete mix ratio designs. The main research content was as follows:

1. A large dataset on the basic mechanical properties of BFRC was constructed using experimental data on BFRC strength from published literature and made available to the public;
2. GA-XGBoost was developed and applied to predict CS on BFRC, and the model was validated by SHAP analysis;
3. Six independent regression models—XGBoost, gradient-boosted decision tree (GBDT) regressor, AdaBoost, RF, SVR, and GA-XGBoost—were adopted to predict the concrete CS, and the accuracy of these models' predictions was compared.

## 2. Data Preprocessing

Before constructing a model, the dataset must be assembled and preprocessed. Many researchers [14–17,30–34] have attempted machine-learning predictions, but no studies have been conducted to build shared BFRC model datasets. New datasets needed to be obtained from strength tests of BFRC to develop models for predicting CS.

Since most design criteria for concrete are concerned with strength, this study decided to use and collect test data from several BFRC strengths and form a common dataset. This dataset contained a large amount of data regarding the design parameters of BFRC concrete mixes, as well as several strength parameters. The database included water–cement ratio (W/C), coarse aggregate (CA), silica fume (SF), water–binder ratio (W/B), high-efficiency water reducing agent (S), fine aggregate (FA), and fly ash (F) as concrete-mix design parameters. The ratio of length to diameter of fibers (FL/FD) and fiber content (FC) were used as fiber-property parameters. Finally, the strength parameter recorded in this dataset was compressive strength (CS). The dataset contained 12 parameters, 11 of which are features, and 1 a target. Table 1 presents descriptive statistics for the dataset utilized in the development of the model.

To enhance the visualization of the graphs, we used Python to create Figures 1 and 2. Figure 1 is a graphical representation of the pairing matrix, and demonstrates how the magnitude of the correlation between the variables changed as different factors changed. Among them, the increase in W/C, W/B, and FA led to a decrease in the concrete CS, while the improvement in C, S, and SF enhanced the CS, and the effect of other factors on CS was not significant. Similar conclusions can be drawn from the correlation coefficient matrix (Figure 2), where W/C, W/B, and FA were negatively correlated with CS with correlation factors of $-0.52$, $-0.52$, and $-0.2$, while C, SF, and S were positively correlated with CS to a comparable extent with correlation factors of 0.32, 0.32, and 0.28. In particular, it was observed from Figure 2 that FC was negatively correlated with CS ($-0.15$).

**Table 1.** Dataset descriptive statistics.

| Feature | Units | Mean | Std | Min | Max | Count |
|---|---|---|---|---|---|---|
| W/C | - | 0.450 | 0.0722 | 0.280 | 0.717 | 346 |
| W/B | - | 0.400 | 0.0753 | 0.241 | 0.573 | 346 |
| C | $kg/m^3$ | 395.3 | 69.25 | 217 | 613.3 | 346 |
| F | $kg/m^3$ | 40.92 | 53.98 | 0 | 168 | 346 |
| SF | $kg/m^3$ | 13.14 | 28.28 | 0 | 126 | 346 |
| CA | $kg/m^3$ | 1093 | 170.6 | 512 | 1540 | 346 |
| FA | $kg/m^3$ | 697.9 | 110.7 | 507 | 1194 | 346 |
| W | $kg/m^3$ | 175.2 | 29.15 | 112 | 301 | 346 |
| S | $kg/m^3$ | 3.088 | 2.292 | 0 | 8.360 | 346 |
| FL/FD | $10^3$ | 1.037 | 0.377 | 0.345 | 2 | 346 |
| FC | % | 0.141 | 0.131 | 0 | 0.730 | 346 |
| CS | MPa | 50.15 | 11.80 | 15.52 | 96.25 | 346 |

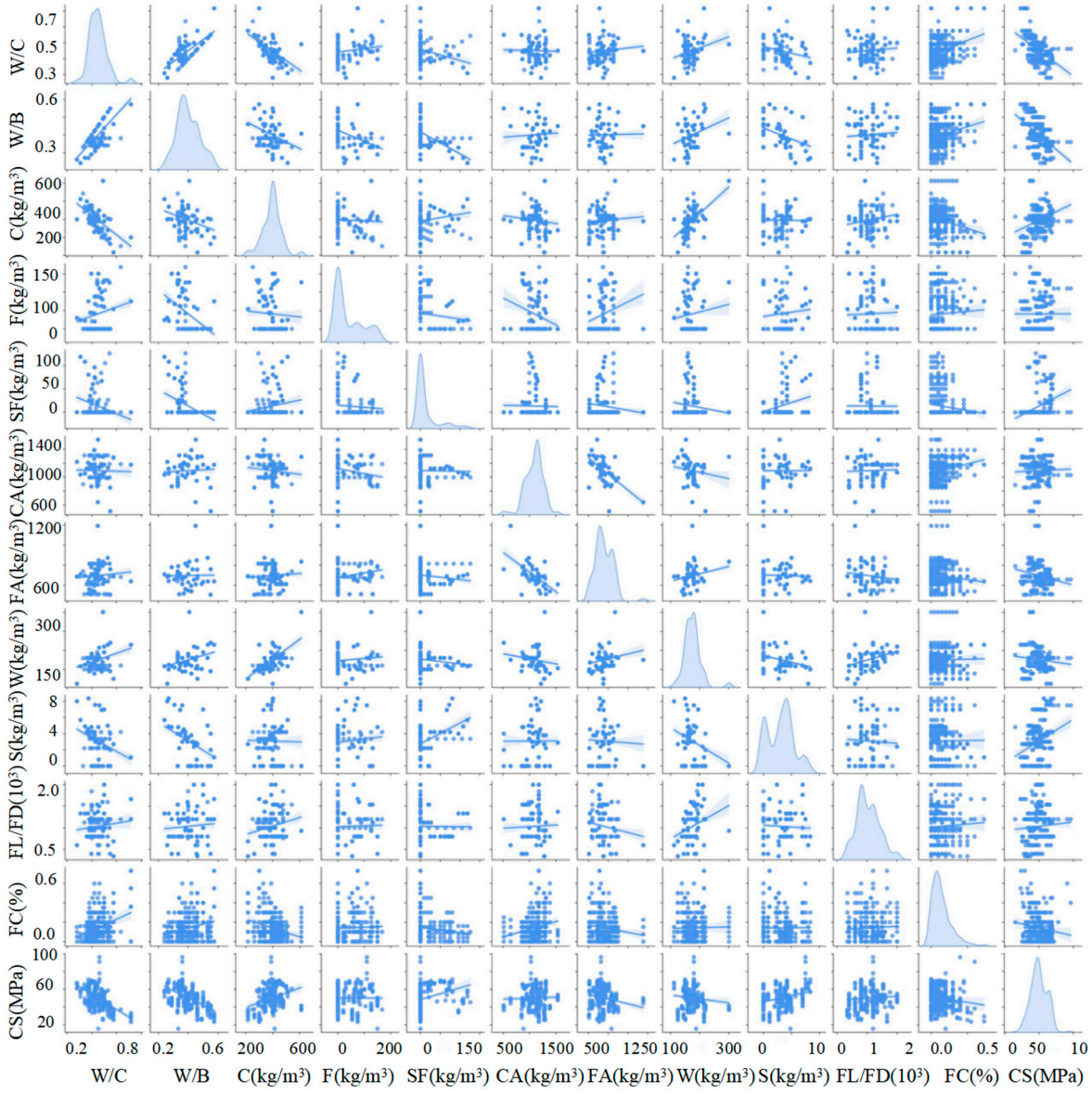


**Figure 1.** Pair diagrams showing variations of magnitudes of variables with each other.

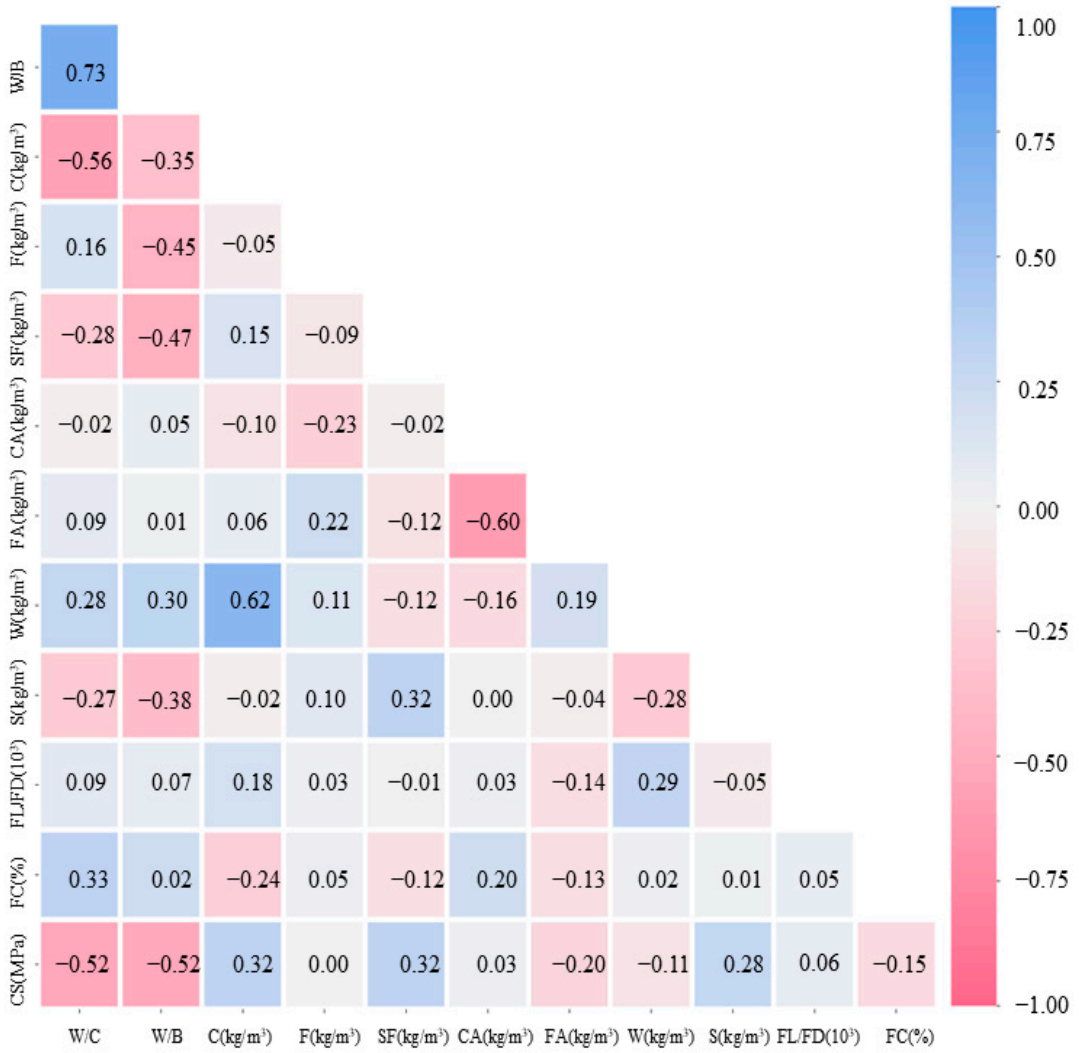


**Figure 2.** Correlation coefficient matrix.

According to Oey et al. [35], scaling the training and test sets before segmentation can potentially result in data leakage. In this study, following the separation of the training and test sets, the dataset was scaled using the robust-scaler function from the sklearn library. By utilizing median and quartile scaling, the robustness of machine-learning algorithm was enhanced to minimize the impact of outliers on the model. Then, we could significantly enhance the accuracy of prediction results.

## 3. Methodology

### *3.1. Machine-Learning Algorithm*

To make it easier to use these machine-learning algorithms, we adopted python as the IDE and used the scikit-learn library from it. The purpose of our study was to validate the GA-XGBoost performance in predicting the CS of BFRC and to compare it with XGBoost, GBDT, AdaBoost, RF, and SVR. To ensure the reliability of the GA-XGBoost model, feature-importance analysis and SHAP analysis were performed, and the findings were analyzed in comparison with previous studies.

#### 3.1.1. Brief Description of RF

RF is described as a consolidated learning model which decreases the variance of an individual decision tree by building the integration of multiple decision trees [36]. This algorithm leverages the definition of bagging, which involves aggregating randomly

selected similar datasets from the training set into a forest [37]. In order to randomly select enough features among the nodes, the RF algorithm uses bagging to optimize them. The decision tree partitions the randomly selected sample features into left and right subtrees. This method reduces the instability of the decision tree and improves its generalization ability, but leads to over-fitting in some noisy classification and regression problems. Figure 3 presents the schematic diagram depicting the bagging algorithm.

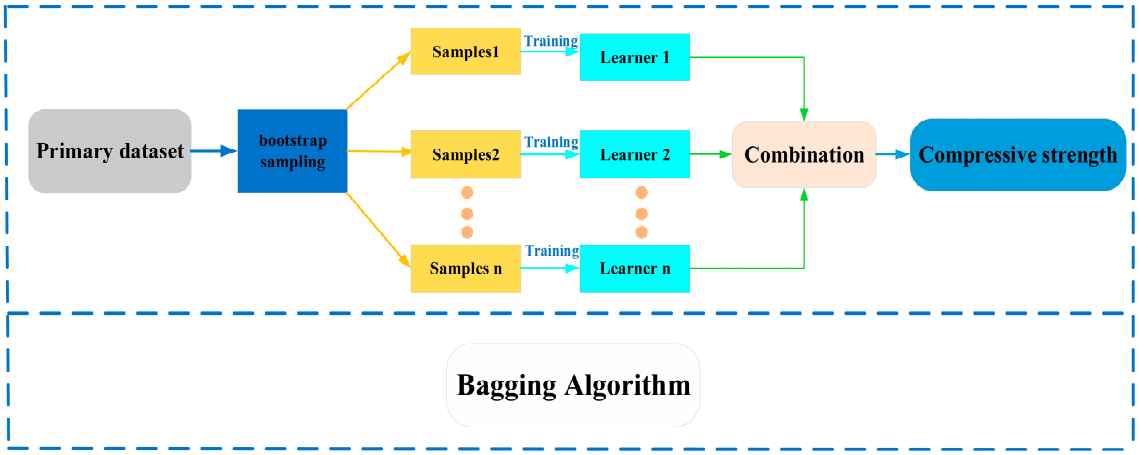


**Figure 3.** Illustration of bagging method.

### 3.1.2. Brief Description of AdaBoost

The boosting method is a consolidated learning model that constructs a strong learner based on jointing multiple weak learners. In the boosting method, the sample weights of follow-up models are increased based on the learning from the previous model. This iterative process results in the creation of strong learners [38]. AdaBoost is a boosting algorithm that enables training to focus on hard-to-predict samples by dynamically weighting the training samples, ultimately achieving the optimal weak learner [39]. In each iteration, each weak learner dynamically modifies the weights of each sample based on the previously obtained prediction accuracy and trains the new dataset. This algorithm makes the combination of weak learners more likely to produce accurate prediction results. Figure 4 shows the schematic diagram of the boosting algorithm.

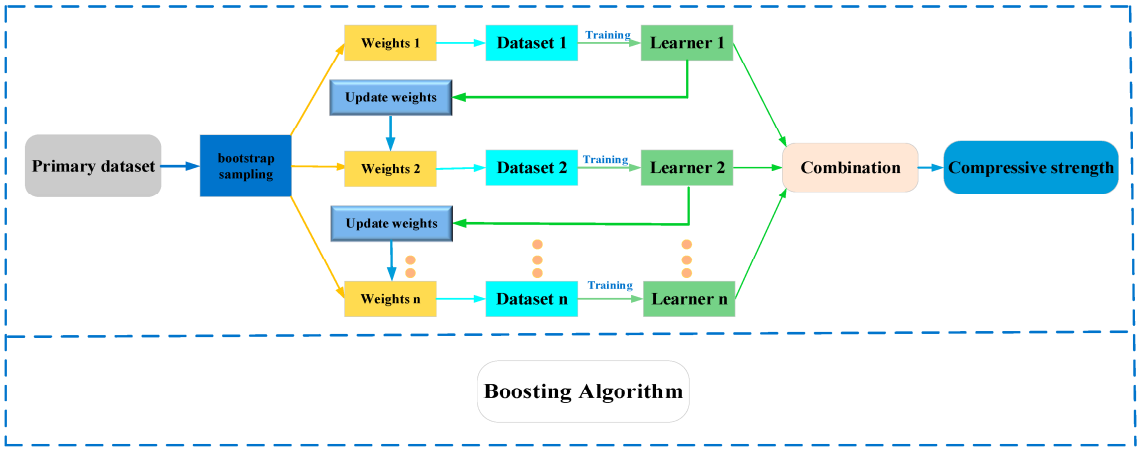


**Figure 4.** Illustration of boosting method.

### 3.1.3. Brief Description of GBDT

The definition of the GBDT is a decision-tree-based synthesis algorithm designed to fit the training data by continuously reducing the residuals. Following each round of training

for the weak learner, the GBDT adjusts itself by moving in the direction of a decreasing gradient of the loss function [40].

Similar to AdaBoost, the GBDT uses the gradient rather than the weights of error points to locate weak learner deficiencies. The GBDT can use a wider range of objective functions than AdaBoost.

#### 3.1.4. Brief Description of SVR

SVR is an algorithm for solving linear and nonlinear regression problems with a strong generalization capability [41,42]. It predicts export values by combining multiple kernel functions to construct reliable regression models. Another advantage of this algorithm is that the optimal kernel function can be applied to enhance the forecast result accuracy. In addition to regression problems, SVR can be used for pattern recognition and data classification.

#### 3.1.5. Brief Description of XGBoost

XGBoost is an upgraded version of the boosting algorithm based on the GBDT algorithm, and it allows for more accurate and faster fitting by directly using first- and second-order gradients to extend the loss function [43]. This algorithm has been widely used in several fields and offers greater problem-solving capabilities and fewer usage constraints. In contrast to the GBDT, XGBoost exhibits shorter learning times and superior predictive power, particularly when confronted with large-scale datasets. However, the disadvantages of XGBoost are that it requires a lot of memory and computational resources during the training process, and may have overfitting problems for certain datasets. Therefore, careful hyperparameter selection and model optimization are required when using XGBoost.

#### 3.1.6. Features, Advantages, and Disadvantages of the above Five Models

As mentioned above, RF, AdaBoost, GBDT Regressor and XGBoost are all integrated learning algorithms based on decision trees, while SVR is a support vector machine based on a regression algorithm. In general, the DT is the foundation and combinations of DTs creates a RF that has higher accuracy. AdaBoost adjusts the sample weights at each iteration to strengthen those samples that are misclassified. GBDT is designed to fit the training data by continuously reducing the residuals. SVR is a powerful regression algorithm for dealing with nonlinear regression problems. Finally, in order to address the overfitting problem, XGBoost extended and possessed the highest accuracy of the four decision tree methods, while SVR has a high sensitivity to parameters. The main features, advantages, and disadvantages of the above five models are presented in Table 2.

#### 3.1.7. Combination of Genetic Algorithm and XGBoost

Genetic algorithms (GAs), which originated from computer simulations of biological systems, are stochastic, efficient, and parallel global search and optimization methods. They are able to automatically generate and accumulate knowledge about the search space during the search process and adaptively control the search process to find the optimal solution.

Although XGBoost excels in all aspects, it does have drawbacks, one of which is that many of the parameters are customizable and results can vary from configuration to configuration. When using grid search to adjust parameters, the number of parameters added is correlated with the search time spent. Due to its inherent parallelism and ability to perform distributed computations, GAs quickly traverse all solving methods in the solution space without falling into the trap of rapidly decreasing locally optimal solutions. Consequently, GAs allow faster optimization of many machine-learning parameters and can achieve higher efficiency, outperforming grid searches.

**Table 2.** Summary of the above models [44–48].

| Model | Main Features | Advantages | Disadvantages |
| --- | --- | --- | --- |
| RF | RF is an integrated model consisting of multiple decision trees, which can reduce the variance of individual decision trees. | 1. Improves model accuracy by reducing overfitting;<br>2. Handles both classification and regression problems;<br>3. Robust to noise and missing values in the data sets. | 1. The model training time is longer;<br>2. Larger storage space is required to store multiple decision trees;<br>3. The presence of outliers in the dataset can influent the accuracy of the model. |
| AdaBoost | AdaBoost is an integrated model consisting of several weak classifiers for binary and multivariate classification problems. | 1. Manages overfitting to solve dichotomous and multiclassification problems;<br>2. Handles high-dimensional, unbalanced datasets. | 1. Sensitive to outliers, which may have a significant impact on the model;<br>2. Longer training time;<br>3. Tendency to overfit when the data set is too noisy. |
| GBDT | GBDT Regressor is an integrated decision-tree-based learning algorithm that improves the accuracy and generalization performance through gradient boosting method. | 1. Handles effectively nonlinear relationships and complex feature interactions;<br>2. Missing values and outliers are handled automatically and with good robustness;<br>3. Generates feature importance rankings to facilitate subsequent feature engineering. | 1. Sensitive to noise and outliers and prone to overfitting;<br>2. Relatively slow training speed;<br>3. Prone to prediction bias for unbalanced datasets. |
| SVR | SVR is a support vector machine (SVM)-based regression algorithm that can be used to handle nonlinear regression problems. | 1. Handles high-dimensional data sets and non-linear regression problems;<br>2. There is no requirement for the uniqueness of the solution of the model;<br>3. Setting the kernel function to be well adapted to different data types. | 1. Time- and memory-consuming when dealing with large data sets;<br>2. Sensitive to noise and easy to over-fit;<br>3. Need to solve convex optimization problems with high solution complexity. |
| XGBoost | XGBoost is a decision-tree-based gradient boosting algorithm that also uses regularization to enhance the accuracy and generalization performance. | 1. Automatic feature selection and excellent performance in handling large scale, high-dimensional data;<br>2. Handles missing values, outliers, and unbalanced datasets, and is less prone to overfitting;<br>3. Supports parallel computing and distributed computing with fast processing time. | 1. Requires long training time and large amount of computational resources to handle large-scale datasets;<br>2. Parameter tuning may lead to over-fitting or under-fitting;<br>3. Vulnerable to outliers. |

To address the problems of long preparation time and many parameters in XGBoost models, we adopted GAs to improve the efficiency of parameter search process in XGBoost. The individual population of GAs is usually defined as the XGBoost parameters and the fitness function is determined as the average score achieved during XGBoost training. Also, selection, crossover, and variation are used to enhance the operational efficiency of the model. Figure 5 shows a schematic representation of GA-XGBoost.

As a result, GAs should be used to optimize the XGBoost model, so that we find the optimal solution faster, reduce the tuning time, and ease the burden of parameter tuning.

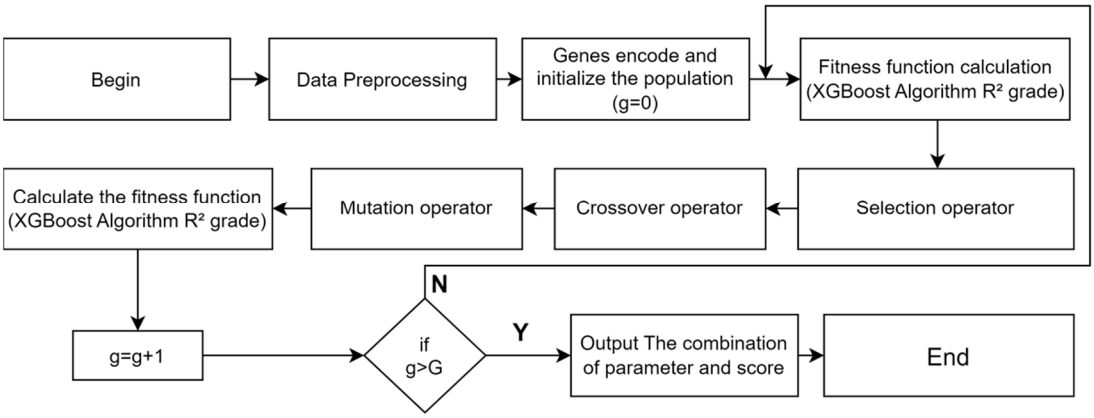


**Figure 5.** GA−XGBoost algorithm flow chart.

### *3.2. Model Performance Evaluation*

For machine-learning algorithms, we commonly use mean square error (*MSE*), coefficient of determination ($R^2$), root mean square error (*RMSE*), and mean absolute error (*MAE*) to describe the accuracy and stability of a model.

In these regression models, the key to estimation performance is the variance component of the response features, which can have values from negative infinity to 1. Thus, if the algorithm's model satisfies the data perfectly, its value will be 1 and also explain the data variability. The formulae for the four evaluation indicators are shown below.

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

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \tag{2}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{3}$$

$$MAE = \frac{\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|}{n} \tag{4}$$

where $\hat{y}_i$ and $y_i$ are the predicted value and true value, $\bar{y}$ is the sample mean, and $n$ is the count of data samples.

### *3.3. Methodology Flowchart*

#### 3.3.1. Data Collection

The flowchart of the current work included the four main steps above, as shown in Figure 6.

The database consisted of 346 experimental results collected from 28 published articles about the CS of BFRC. This dataset was randomly divided into a training dataset (accounting for 70%) and a test dataset (the remaining 30%).

#### 3.3.2. Model Training

We used the training dataset to train the ML models mentioned above. These algorithms are described in Section 3.1. The training process was repeated until the above-mentioned models were successfully trained.

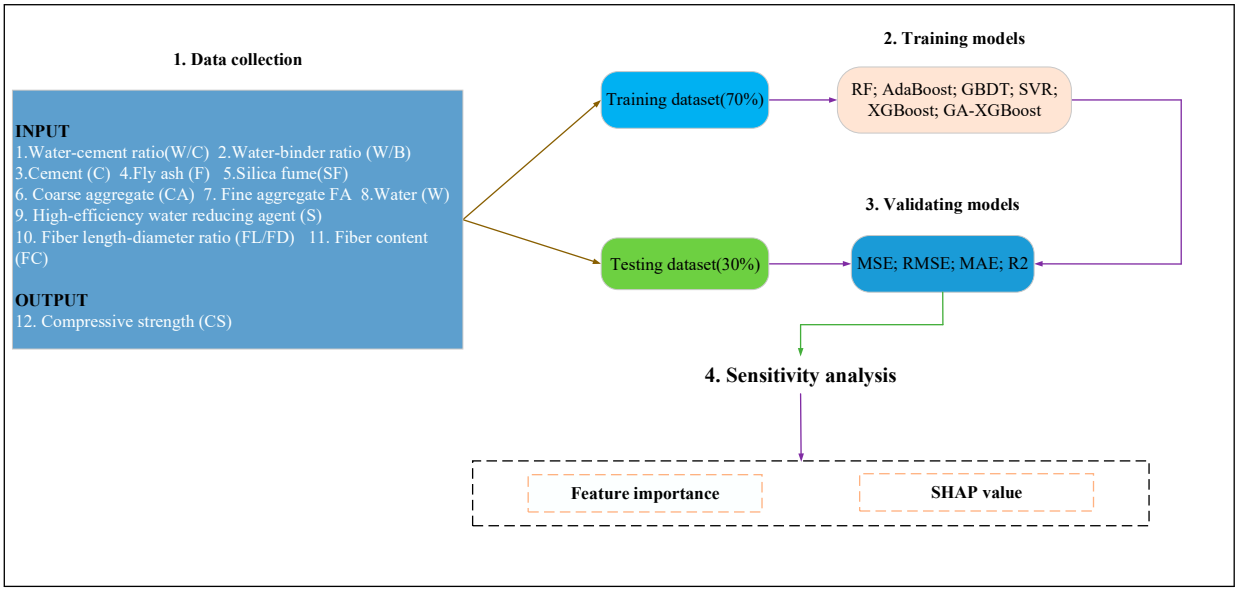


**Figure 6.** Methodology flowchart.

#### 3.3.3. Model Verification

After step 2, we used the testing set to validate the above models. That is, we used the GA-XGBoost model proposed in this paper to compare with five other existing models. Additionally, we used the repeated holdout method to ensure the model prediction performance. Model prediction ability was evaluated using four criteria detailed in Section 3.1.

#### 3.3.4. Sensitivity Analysis

Finally, we evaluated the influence of input variables with feature importance and SHAP values to assess the influence of the 11 input parameters.

First, we obtained the original dataset and selected the most relevant features using the feature selection function with reference to Figures 1 and 2. Second, in order to evaluate the performance of the model and ensure the consistency and stability of the model, we randomly divided the dataset into a training set and a test set, and adjusted the random-state parameter of the model to 1. In order to improve the speed and accuracy of the model training, we scaled the training set and the test set by using the median and interquartile spacing. Then, the corresponding datasets were trained and tested, and tuned by traditional grid-search methods to obtain five single machine-learning models. The default settings for the other models' parameters were used. Finally, we utilize GAs for parameter tuning of the XGBoost models to obtain an integrated model, and the flow is shown in Figure 5.

## 4. Results and Discussion

### *4.1. Evaluation of Six Models*

Table 3 shows the results of statistical validation of the model performance on the dataset under four different metrics. Here, the computed results of four performance criteria ($R^2$, *MSE*, *RMSE*, and *MAE*) were used to assess the accuracy of the model, a score-based system was designed to assign a score (from 1 to 6) based on the criterion values, and the total score described the ranking of the model. A smaller total score represented a better model performance. Compared to the XGBoost model, the GA-XGBoost model achieved smaller metric errors on the training set (*MSE* = 2.2596, *RMSE* = 1.5032, *MAE* = 1.0116, $R^2$ = 0.9834). The results indicated that GAs can be used to maximize performance and enhance the learning ability of the XGBoost model. Also, GA-XGBoost showed the highest score on the test set with a metric error of (*MSE* = 7.6962, *RMSE* = 2.7742, *MAE* = 2.0564, $R^2$ = 0.9483). It is reasonable to use GAs to improve parameter tuning of XGBoost to maximize its performance. The performance of other models using grid-search optimization

showed that grid-search optimization is also adequate to predict the CS of concrete with considerable model accuracy.

**Table 3.** Statistical evaluation metrics of models.

| Type of Set | Metrics | XGBoost | GBDT | AdaBoost | RF | SVR | GA-XGBoost |
|---|---|---|---|---|---|---|---|
| Train | $R^2$ | 0.9307 | 0.987 | 0.7822 | 0.96 | 0.9527 | 0.9834 |
| | Rank | 5 | 1 | 6 | 3 | 4 | 2 |
| | *MSE* | 9.4145 | 1.7631 | 29.6077 | 5.4331 | 6.4333 | 2.2596 |
| | Rank | 5 | 1 | 6 | 3 | 4 | 2 |
| | *RMSE* | 3.0683 | 1.3278 | 5.4413 | 2.3309 | 2.5364 | 1.5032 |
| | Rank | 5 | 1 | 6 | 3 | 4 | 2 |
| | *MAE* | 1.9637 | 0.8235 | 4.4515 | 1.6056 | 0.758 | 1.0116 |
| | Rank | 5 | 2 | 6 | 4 | 1 | 3 |
| Test | $R^2$ | 0.9133 | 0.914 | 0.826 | 0.9322 | 0.9123 | 0.9483 |
| | Rank | 4 | 3 | 6 | 2 | 5 | 1 |
| | *MSE* | 12.6259 | 12.5174 | 25.321 | 9.8671 | 12.7635 | 7.6962 |
| | Rank | 4 | 3 | 6 | 2 | 5 | 1 |
| | *RMSE* | 3.5533 | 3.538 | 5.032 | 3.1412 | 3.5726 | 2.7742 |
| | Rank | 4 | 3 | 6 | 2 | 5 | 1 |
| | *MAE* | 2.5726 | 2.473 | 4.0276 | 2.2175 | 2.5728 | 2.0564 |
| | Rank | 4 | 3 | 6 | 2 | 5 | 1 |
| Total rank score | | 36 | 17 | 48 | 21 | 33 | 13 |

According to the experimental results shown in Figure 7, all models except the AdaBoost model performed well in predicting the CS on BFRC. The predicted values of AdaBoost model showed greater deviations relative to their experimental values due to adaptively increasing the weight of the prediction error. The GA-XGBoost had the best performance and stability, while the RF, GBDT, XGBoost, and SVR performed slightly differently. However, after grid-search optimization, the AdaBoost model still performed poorly because of its lack of sensitivity to noise and outliers. Therefore, we can conclude that GA-XGBoost is the most suitable ML algorithm for predicting the CS on BFRC.

### *4.2. SHAP Analysis*

SHAP analysis is an algorithm used to interpret the prediction results of ML models by providing the magnitude of each feature's impact on the prediction results. The core of SHAP analysis is based on the Shapley value principle, which assigns each feature's contribution to all possible subsets of features and calculates the expected value of the feature contribution to obtain the magnitude of each feature's impact on the prediction results of the model. Shapley values can be used not only to assess the importance of each input variable, but also to calculate the impact of individual input variables on the final result, which is presented in the form of a Shapley plot. Feature-importance analysis is a method used to determine which features or variables have the most influence on the prediction results of a model to help optimize and explain the model's performance and results.

SHAP analysis can help us interpret the model prediction results and determine which features have the most influence on the prediction results. Also, SHAP can help us validate the reliability of the model to determine if the model uses reasonable features in making predictions on the input data. Finally, SHAP analysis can also help us enhance the model performance by analyzing the impact of each feature on the model's prediction results and determining which features need more attention and optimization to enhance the prediction performance. For BFRC fit design, it is very important to know the impact of model input variables. In addition, the proposed algorithm should be described in detail to understand how to analyze and calculate the predicted concrete CS. For the

global interpretation, we performed SHAP and feature-importance analysis on the best-performing GA-XGBoost model.

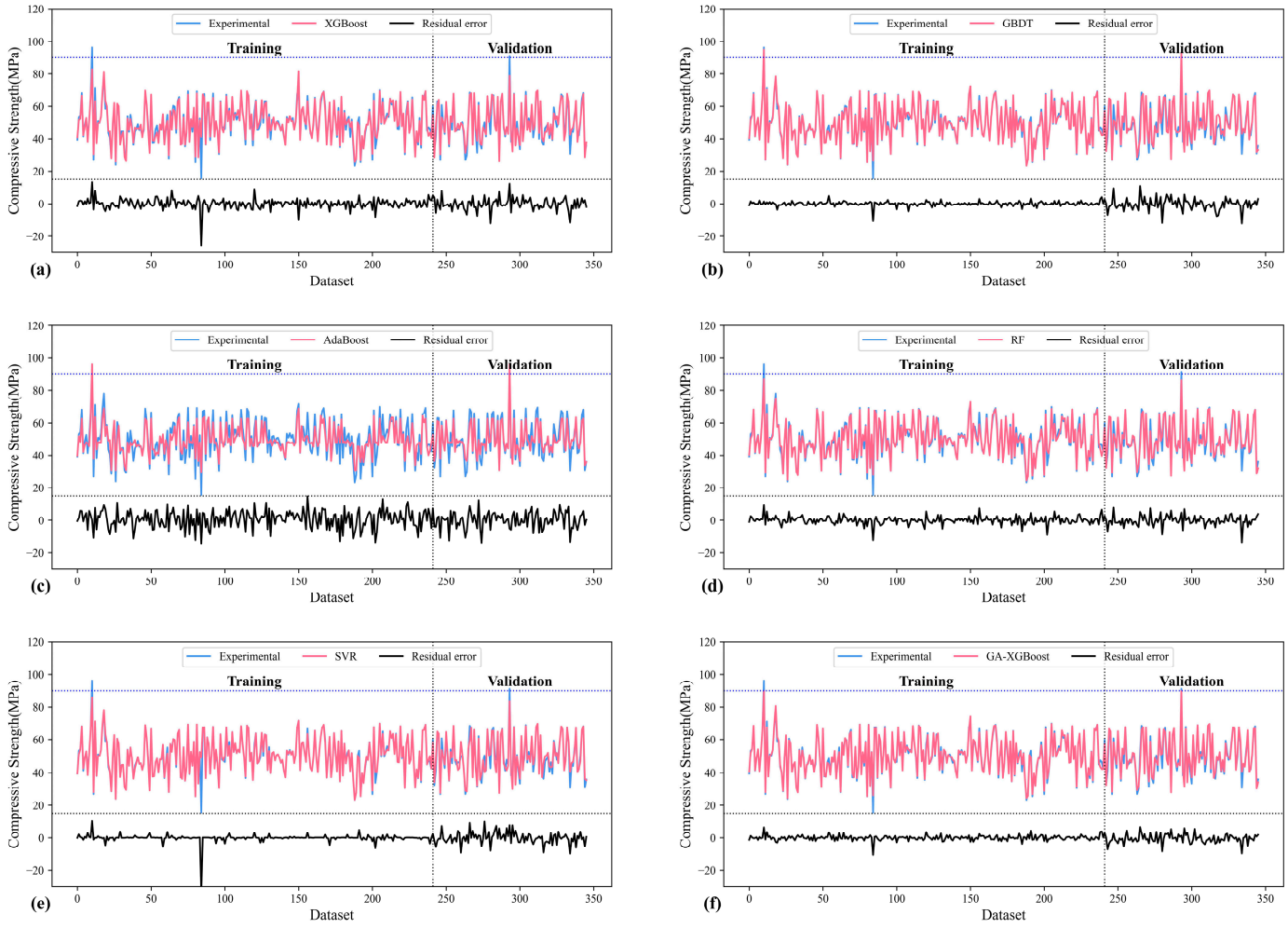


**Figure 7.** Comparison of prediction CS with six algorithms. (The subfigures (**a**–**f**) represent respectively the comparison between the experimental results and the predictions of the XGBoost, GBDT, AdaBoost, RF, SVR, GA-XGBoost and their residual errors).

To test the effect of features on the prediction of the target variables, Figure 8 indicates the significance of input variables. The results were obtained by averaging the Shapley values throughout the data collection process. We found that W/B dominated the CS of BFRC, which is consistent with the experimental results. The variables FA and W/C were in the second and third place and, in that order, of considerable importance for CS. It is also clear from this figure that CA content was the least important variable, while SF and FL/FD were slightly more important than CA.

Aggregate plots summarize the relative importance of input factors and their relationship with the independent variables and therefore play an important role in SHAP analysis. Figure 9 indicates global interpretation of GA-XGBoost using the SHAP interpretation, showing the distribution of SHAP values for all model features and the main trends of the variables. The figure shows SHAP values of 11 input features, where positive values indicate an aggressive gain to CS and negative values indicate the opposite. Red dots represent those samples with high eigenvalues and blue dots are the opposite.

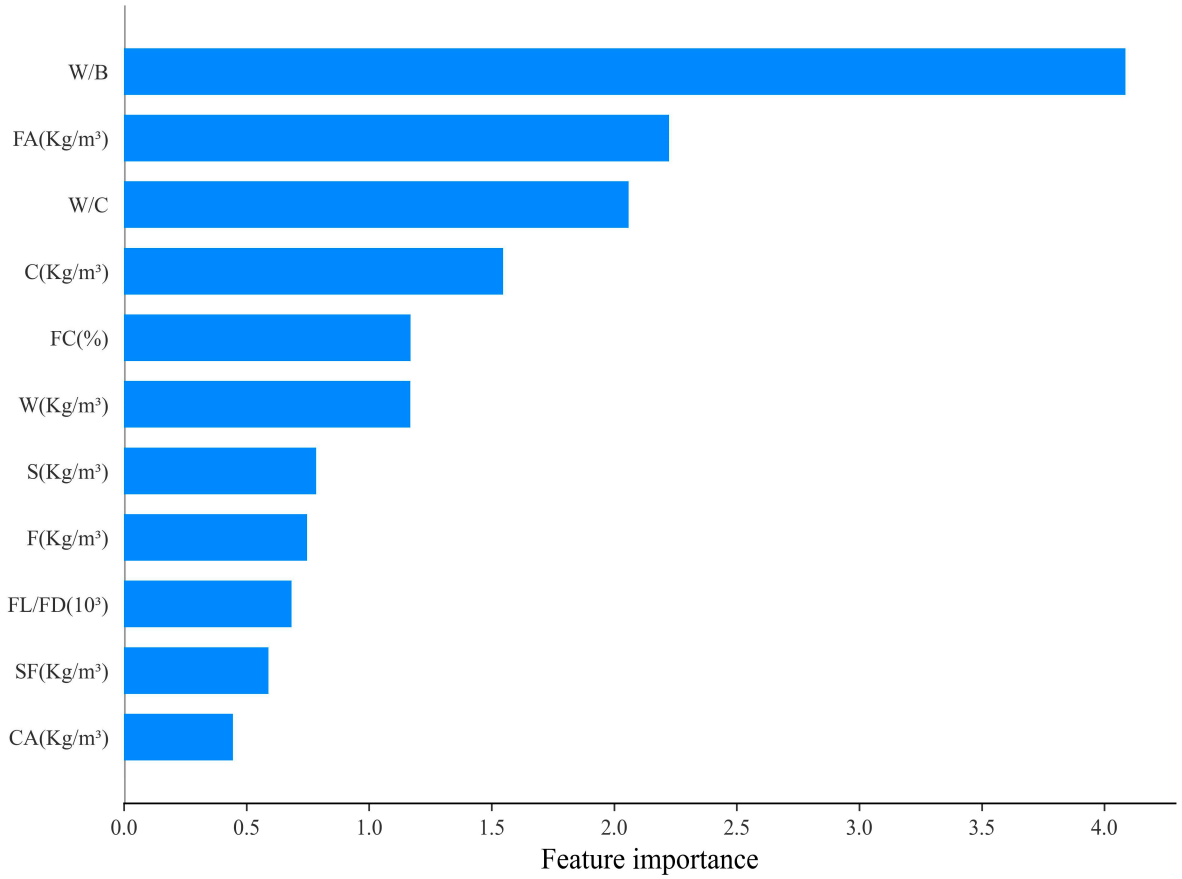


**Figure 8.** Feature importance of CS.

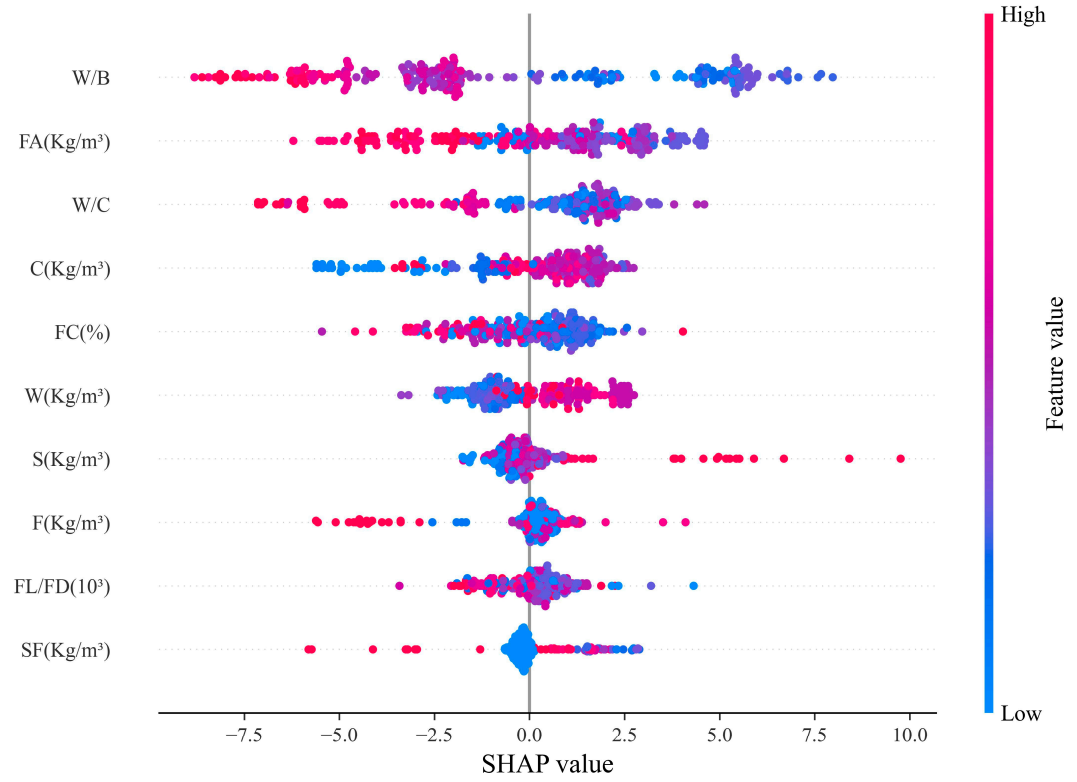


**Figure 9.** SHAP summary plot.

Figure 9 shows that the W/B variable has a low sample eigenvalue and a high SHAP value. Therefore, the W/B variable with smaller sample eigenvalues has a larger positive gain in improving the CS of BFRC. For the FA and W/C variables, SHAP values with lower sample eigenvalues are positive, which indicates that reducing FA and W/C can improve the CS of BFRC to some extent. Additionally, we can derive from Figure 9 that the SF and FL/FD variables have smaller sample eigenvalues and SHAP values close to 0, which indicates that both of them have little influence on improving the CS of BFRC.

## 5. Conclusions and Limitations

In order to avoid local optimal solutions and reduce the time for model parameter debugging, an extended XGBoost model (GA-XGBoost) with adoption of GA optimization parameters was designed in this study to facilitate the accuracy and stability of predicting the compressive strength of BFRC. For comparative analysis, we also used XGBoost, GBDT regressor, AdaBoost, RF, and SVR—five other regression models. The following conclusions were reached:

(1) Compared to other regression models, the GA-XGBoost model shows the best accuracy and stability in predicting CS of BFRC. For the test dataset, the $R^2$, *MSE*, *RMSE*, and *MAE* of GA-XGBoost were 0.9483, 7.6962 MPa, 2.7742 MPa, and 2.0564 MPa, and the errors were within the acceptable range.
(2) By using GAs to tune the parameters in the ML algorithm, a lot of debugging work can be avoided and the best combination of parameters can be obtained. For engineering applications involving ML algorithms, this can greatly assist in developing practical solutions.
(3) According to SHAP analysis, W/B of BFRC is the most important variable that dominates CS, followed by FA and W/C. The variable FC has some influence on CS, while other variables, such as CA and SF, have less influence on CS. This can provide some reference for the design of BFRC fits.

These results show that the model has high prediction accuracy and stability, and has several application values:

(1) It can guide the calculation of BFRC compressive strength required for engineering;
(2) It effectively reduces the difficulty of obtaining BFRC compressive strength, reduces the experimental workload, saves time and cost, and is more economical and environmentally friendly;
(3) We developed a genetic algorithm for parameter optimization to determine the key parameters of the prediction model, which can provide an effective reference for the optimization of other machine models.

While this study used a large data set containing 11 input variables to build the model, this also increased the complexity of the model. Having more input variables leads to more complex models, which may be detrimental to the generalization ability of the model. At the same time, the presence of these non-independent selected inputs in our input features may have led to redundant information in the data, affecting the performance and generalization ability of the model. We can consider optimizing these non-independent variables by means of feature engineering and quantitative simplification. Therefore, there may be limitations to machine-learning algorithms for practical engineering applications. Future research needs to explore whether input-variable reduction can enhance the accuracy and generalization ability of ML models.

**Author Contributions:** J.Z., conceptualization, resources, funding acquisition; T.Y., writing, validation, original draft, and formal analysis; J.Y., data curation, and writing—review and editing; M.W., supervision, investigation, and reviewing; S.X., writing—review and editing, and validation. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by the National Natural Science Foundation of China (Grant No. 12102073), the Education Commission Project of Chongqing (Grant No. KJQN202000711), and the Transportation Science and Technology Project of Sichuan Province (Grant No. 2020-ZL-B1). We also acknowledge the State Key Laboratory of Mountain Bridge and Tunnel Engineering for providing experimental equipment.

**Data Availability Statement:** Publicly available datasets were analyzed in this study. This data can be found here: Wang, Minghui (2022), "Mechanical Properties Dataset of BFRC for strength prediction with machine learning", Mendeley Data, V1, DOI: 10.17632/b5s8ywwgwr.1.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Nomenclature

| | |
|---|---|
| BFRC | Basalt-fiber-reinforced concrete |
| XGBoost | Extreme gradient boosting tree |
| ML | Machine learning |
| GBDT | Gradient-boosted decision tree |
| AdaBoost | Adaptive gradient boosting |
| BF | Basalt fiber |
| *MSE* | Mean square error |
| W/C | Water–cement ratio |
| SF | Silica fume |
| S | High-efficiency water reducing agent |
| FL/FD | Ratio of length to diameter of fibers |
| F | Fly ash |
| CS | Compressive strength |
| GA | Genetic algorithm |
| RF | Random forest |
| SVR | Support vector regression |
| *MAE* | Mean absolute error |
| $R^2$ | Coefficient of determination |
| *RMSE* | Root mean square error |
| CA | Coarse aggregate |
| W/B | Water–binder ratio |
| FA | Fine aggregate |
| FC | Fiber content |

## References

1. Jalasutram, S.; Sahoo, D.R.; Matsagar, V. Experimental investigation of the mechanical properties of basalt fiber-reinforced concrete. *Struct. Concr.* **2017**, *18*, 292–302. [CrossRef]
2. Arslan, M.E. Effects of basalt and glass chopped fibers addition on fracture energy and mechanical properties of ordinary concrete: CMOD measurement. *Constr. Build. Mater.* **2016**, *114*, 383–391. [CrossRef]
3. Zhao, Y.-R.; Wang, L.; Lei, Z.-K.; Han, X.-F.; Shi, J.-N. Study on bending damage and failure of basalt fiber reinforced concrete under freeze-thaw cycles. *Constr. Build. Mater.* **2018**, *163*, 460–470. [CrossRef]
4. Li, M.; Gong, F.; Wu, Z. Study on mechanical properties of alkali-resistant basalt fiber reinforced concrete. *Constr. Build. Mater.* **2020**, *245*, 118424. [CrossRef]
5. Wang, D.; Ju, Y.; Shen, H.; Xu, L. Mechanical properties of high performance concrete reinforced with basalt fiber and polypropylene fiber. *Constr. Build. Mater.* **2019**, *197*, 464–473. [CrossRef]
6. Wang, X.; He, J.; Mosallam, A.S.; Li, C.; Xin, H. The Effects of Fiber Length and Volume on Material Properties and Crack Resistance of Basalt Fiber Reinforced Concrete (BFRC). *Adv. Mater. Sci. Eng.* **2019**, *2019*, 7520549. [CrossRef]
7. Chen, W.; Zhu, Z.C.; Wang, J.; Chen, J.; Mo, Y. Numerical Analysis of Mechanical Properties of Chopped Basalt Fiber Reinforced Concrete. *Key Eng. Mater.* **2019**, *815*, 175–181. [CrossRef]
8. Jiang, C.; Fan, K.; Wu, F.; Chen, D. Experimental study on the mechanical properties and microstructure of chopped basalt fibre reinforced concrete. *Mater. Des.* **2014**, *58*, 187–193. [CrossRef]
9. Kizilkanat, A.B.; Kabay, N.; Akyüncü, V.; Chowdhury, S.; Akça, A.H. Mechanical properties and fracture behavior of basalt and glass fiber reinforced concrete: An experimental study. *Constr. Build. Mater.* **2015**, *100*, 218–224. [CrossRef]
10. Pehlivanli, Z.O.; Uzun, I.; Demir, I. Mechanical and microstructural features of autoclaved aerated concrete reinforced with autoclaved polypropylene, carbon, basalt and glass fiber. *Constr. Build. Mater.* **2015**, *96*, 428–433. [CrossRef]

11. Katkhuda, H.; Shatarat, N. Improving the mechanical properties of recycled concrete aggregate using chopped basalt fibers and acid treatment. *Constr. Build. Mater.* **2017**, *140*, 328–335. [CrossRef]
12. Ahmad, M.R.; Chen, B. Effect of silica fume and basalt fiber on the mechanical properties and microstructure of magnesium phosphate cement (MPC) mortar. *Constr. Build. Mater.* **2018**, *190*, 466–478. [CrossRef]
13. Sun, X.; Gao, Z.; Cao, P.; Zhou, C. Mechanical properties tests and multiscale numerical simulations for basalt fiber reinforced concrete. *Constr. Build. Mater.* **2019**, *202*, 58–72. [CrossRef]
14. Naser, M.Z. Machine Learning Assessment of FRP-Strengthened and Reinforced Concrete Members. *ACI Struct. J.* **2020**, *117*, 237–251. [CrossRef]
15. Aravind, N.; Nagajothi, S.; Elavenil, S. Machine learning model for predicting the crack detection and pattern recognition of geopolymer concrete beams. *Constr. Build. Mater.* **2021**, *297*, 123785. [CrossRef]
16. Basaran, B.; Kalkan, I.; Bergil, E.; Erdal, E. Estimation of the FRP-concrete bond strength with code formulations and machine learning algorithms. *Compos. Struct.* **2021**, *268*, 113972. [CrossRef]
17. Li, H.; Lin, J.; Lei, X.; Wei, T. Compressive strength prediction of basalt fiber reinforced concrete via random forest algorithm. *Mater. Today Commun.* **2022**, *30*, 103117. [CrossRef]
18. Severcan, M.H. Prediction of splitting tensile strength from the compressive strength of concrete using GEP. *Neural Comput. Appl.* **2011**, *21*, 1937–1945. [CrossRef]
19. Nguyen, M.H.; Mai, H.-V.T.; Trinh, S.H.; Ly, H.-B. A comparative assessment of tree-based predictive models to estimate geopolymer concrete compressive strength. *Neural Comput. Appl.* **2022**, *35*, 6569–6588. [CrossRef]
20. Gupta, S.; Sihag, P. Prediction of the compressive strength of concrete using various predictive modeling techniques. *Neural Comput. Appl.* **2022**, *34*, 6535–6545. [CrossRef]
21. Asteris, P.G.; Koopialipoor, M.; Armaghani, D.J.; Kotsonis, E.A.; Lourenço, P.B. Prediction of cement-based mortars compressive strength using machine learning techniques. *Neural Comput. Appl.* **2021**, *33*, 13089–13121. [CrossRef]
22. Kang, M.-C.; Yoo, D.-Y.; Gupta, R. Machine learning-based prediction for compressive and flexural strengths of steel fiber-reinforced concrete. *Constr. Build. Mater.* **2020**, *266*, 121117. [CrossRef]
23. Altayeb, M.; Wang, X.; Musa, T.H. An ensemble method for predicting the mechanical properties of strain hardening cementitious composites. *Constr. Build. Mater.* **2021**, *286*, 122807. [CrossRef]
24. Armaghani, D.J.; Asteris, P.G. A comparative study of ANN and ANFIS models for the prediction of cement-based mortar materials compressive strength. *Neural Comput. Appl.* **2021**, *33*, 4501–4532. [CrossRef]
25. Ahmed, H.U.; Mostafa, R.R.; Mohammed, A.; Sihag, P.; Qadir, A. Support vector regression (SVR) and grey wolf optimization (GWO) to predict the compressive strength of GGBFS-based geopolymer concrete. *Neural Comput. Appl.* **2023**, *35*, 2909–2926. [CrossRef]
26. Nazar, S.; Yang, J.; Ahmad, W.; Javed, M.F.; Alabduljabbar, H.; Deifalla, A.F. Development of the New Prediction Models for the Compressive Strength of Nanomodified Concrete Using Novel Machine Learning Techniques. *Buildings* **2022**, *12*, 2160. [CrossRef]
27. Esmaeili-Falak, M.; Benemaran, R.S. Ensemble deep learning-based models to predict the resilient modulus of modified base materials subjected to wet-dry cycles. *Geomech. Eng.* **2023**, *32*, 583–600.
28. Benemaran, R.S.; Esmaeili-Falak, M.; Javadi, A. Predicting resilient modulus of flexible pavement foundation using extreme gradient boosting based optimised models. *Int. J. Pavement Eng.* **2022**, *24*, 1–20. [CrossRef]
29. Li, D.; Zhang, X.; Kang, Q.; Tavakkol, E. Estimation of unconfined compressive strength of marine clay modified with recycled tiles using hybridized extreme gradient boosting method. *Constr. Build. Mater.* **2023**, *393*, 131992. [CrossRef]
30. Malami, S.I.; Anwar, F.H.; Abdulrahman, S.; Haruna, S.; Ali, S.I.A.; Abba, S. Implementation of hybrid neuro-fuzzy and self-turning predictive model for the prediction of concrete carbonation depth: A soft computing technique. *Results Eng.* **2021**, *10*, 100228. [CrossRef]
31. Iqbal, M.; Zhang, D.; Jalal, F.E.; Javed, M.F. Computational AI prediction models for residual tensile strength of GFRP bars aged in the alkaline concrete environment. *Ocean Eng.* **2021**, *232*, 109134. [CrossRef]
32. Liu, Q.-F.; Iqbal, M.F.; Yang, J.; Lu, X.-Y.; Zhang, P.; Rauf, M. Prediction of chloride diffusivity in concrete using artificial neural network: Modelling and performance evaluation. *Constr. Build. Mater.* **2021**, *268*, 121082. [CrossRef]
33. Salami, B.A.; Olayiwola, T.; Oyehan, T.A.; Raji, I.A. Data-driven model for ternary-blend concrete compressive strength prediction using machine learning approach. *Constr. Build. Mater.* **2021**, *301*, 124152. [CrossRef]
34. Zhang, Y.; Aslani, F. Compressive strength prediction models of lightweight aggregate concretes using ultrasonic pulse velocity. *Constr. Build. Mater.* **2021**, *292*, 123419. [CrossRef]
35. Oey, T.; Jones, S.; Bullard, J.W.; Sant, G. Machine learning can predict setting behavior and strength evolution of hydrating cement systems. *J. Am. Ceram. Soc.* **2019**, *103*, 480–490. [CrossRef]
36. Fawagreh, K.; Gaber, M.M.; Elyan, E. Random forests: From early developments to recent advancements. *Syst. Sci. Control Eng. Open Access J.* **2014**, *2*, 602–609. [CrossRef]
37. Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123–140. [CrossRef]
38. Freund, Y.; Schapire, R.E. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *J. Comput. Syst. Sci.* **1997**, *55*, 119–139. [CrossRef]
39. Hastie, T.; Rosset, S.; Zhu, J.; Zou, H. Multi-class adaboost. *Stat. Its Interface* **2009**, *2*, 349–360. [CrossRef]
40. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]

41. Yuvaraj, P.; Murthy, A.R.; Iyer, N.R.; Sekar, S.; Samui, P. Support vector regression based models to predict fracture characteristics of high strength and ultra high strength concrete beams. *Eng. Fract. Mech.* **2013**, *98*, 29–43. [CrossRef]
42. Sun, J.; Zhang, J.; Gu, Y.; Huang, Y.; Sun, Y.; Ma, G. Prediction of permeability and unconfined compressive strength of pervious concrete using evolved support vector regression. *Constr. Build. Mater.* **2019**, *207*, 440–449. [CrossRef]
43. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
44. Zhang, W.; Zhang, R.; Wu, C.; Goh, A.T.C.; Lacasse, S.; Liu, Z.; Liu, H. State-of-the-art review of soft computing applications in underground excavations. *Geosci. Front.* **2020**, *11*, 1095–1106. [CrossRef]
45. Charbuty, B.; Abdulazeez, A. Classification Based on Decision Tree Algorithm for Machine Learning. *J. Appl. Sci. Technol. Trends* **2021**, *2*, 20–28. [CrossRef]
46. Jafarzadeh, H.; Mahdianpari, M.; Gill, E.; Mohammadimanesh, F.; Homayouni, S. Bagging and Boosting Ensemble Classifiers for Classification of Multispectral, Hyperspectral and PolSAR Data: A Comparative Evaluation. *Remote Sens.* **2021**, *13*, 4405. [CrossRef]
47. Cao, J.; Kwong, S.; Wang, R. A noise-detection based AdaBoost algorithm for mislabeled data. *Pattern Recognit.* **2012**, *45*, 4451–4465. [CrossRef]
48. Sun, Y.; Ding, S.; Zhang, Z.; Jia, W. An improved grid search algorithm to optimize SVR for prediction. *Soft Comput.* **2021**, *25*, 5633–5644. [CrossRef]