Structural Durability & Health Monitoring

Tech Science Press

https://doi.org/10.32604/sdhm.2026.080742

ARTICLE

Study on Prediction of Grouting Material Curing Age Based on SAFT and Hyperparameter-Optimized XGBoost

Pengcheng Xia[1], **Zhihong Pan**[1,*], **Ruoyu Chen**[1] and **Linyuan Wang**[2]

[1]School of Civil Engineering and Architecture, Jiangsu University of Science and Technology, Zhenjiang, China
[2]Department of Agricultural and Biological Engineering, Bagley College of Engineering, Starkville, MS, USA

*Corresponding Author: Zhihong Pan. Email: zhhpan@just.edu.cn

Received: 13 February 2026; Accepted: 30 April 2026; Published: 24 August 2026

ABSTRACT: The grouting sleeves in prefabricated structures critically depend on the strength development of the grout; however, existing non-destructive testing methods struggle to capture its time-dependent evolution. This study proposes a hybrid prediction framework that combines the Synthetic Aperture Focusing Technique (SAFT) with a hyperparameter-optimized XGBoost model. Ultrasonic signals were collected at five curing stages (0, 1, 3, 7, and 28 days), from which SAFT-derived features and the area ratios of six color regions were extracted as input variables, with the curing age serving as the model output. Following a correlation analysis with compressive strength, three optimization algorithms—Random Search (RS), Bayesian Optimization (BO), and Adaptive Particle Swarm Optimization (APSO)—were employed to tune the hyperparameters of XGBoost. Among these, the BO-XGBoost model achieved the lowest prediction error within the data range of this study (R^2 = 0.9881, MAE = 0.2037), outperforming the other comparative models. Further validation using leave-one-specimen-out cross-validation (LOSO-CV) indicated that, under the current laboratory conditions, the model maintained good prediction consistency across the three available sleeve specimens. Feature importance and partial dependence plots were used to further verify the physical relevance of the SAFT-derived features and to elucidate their positive or negative effects on strength. This method provides a preliminary interpretable framework for the non-destructive evaluation of grout strength development in prefabricated structures.

KEYWORDS: Ultrasonic testing; synthetic aperture focusing technique; grouting material; Bayesian optimization; machine learning

1 Introduction

Prefabricated buildings are increasingly becoming mainstream in the modern construction industry due to their advantages in construction efficiency and quality control [1–3]. Particularly regarding reinforcement connections, grouted sleeve connections are widely used for joining precast components to each other or to cast-in-place sections [4–8]. The reliability of these grouted sleeves heavily depends on the quality of the grouting material, especially its strength [9,10]. However, issues such as arbitrary increases in water content or the use of non-compliant materials often lead to substandard strength in practical engineering, consequently compromising structural safety [11,12]. Therefore, developing accurate and efficient methods for detecting the strength of grouting materials is crucial for ensuring the reliability of connections in prefabricated buildings and promoting the healthy development of this technological system.

Nondestructive testing (NDT) encompasses technical methods for evaluating the quality and performance of materials without causing damage to the tested object. Common NDT methods include acoustic emission (AE) [13], X-ray testing [14], and the ultrasonic method [15], among others. Data obtained through NDT can provide insights into structural service life, strength, and material quality [16,17].

Currently, NDT techniques have been extensively introduced into research on grouted sleeves [18]. The ultrasonic method, characterized by its rapid detection speed, operational simplicity, and reliable results, has become a favored key tool in the field of structural NDT for engineering structures. Cao et al. [19] proposed a nondestructive testing method based on variations in the wavelet packet energy ratio, combining finite element analysis and experiments to detect the compactness of grouted sleeves in precast shear walls. This method effectively addresses the challenges of difficult compactness detection and grouting quality control. Li and Liu [20] proposed using ultrasonic guided waves, studying the influence of defects on guided wave signals through finite element simulation and experiments. They constructed Damage Index (DI) and Weighted Damage Index (WDI), established correlations between these indices and mechanical performance via monotonic tensile tests on sleeves, and subsequently completed structural health assessment using the Analytic Hierarchy Process. Liu et al. [21] proposed a Grouting Material Vibration Device (GMVD) to enhance the quality of Half-Grouted Sleeve Connections (HGSCs). Ultrasonic testing verification showed that the initial wave propagation time in specimens was shortened after vibration, indicating improved grouting compactness and quality. Notably, while these methods can detect defects within the sleeve's grouting material, such as the location and size of defects or the compactness of the grouted connection, they less frequently address the assessment of the grouting material's strength.

In the field of concrete materials, numerous researchers have utilized ultrasonic testing technology combined with machine learning methods to estimate concrete strength and quality, considering various influencing factors, offering significant advantages in terms of time and cost [22,23]. Günaydın et al. [24] employed Logistic Regression (LR), Multilayer Perceptron (MLP), Support Vector Machine (SVM), and k-Nearest Neighbors (k-NN) models to analyze the strength of 63 concrete samples with different curing ages (7, 28, 90 days) and humidity levels, using NDT-derived data such as Schmidt hammer hardness and ultrasonic pulse velocity. Gan et al. [25] acquired time-domain and frequency spectrum curves using an ultrasonic tester, built an ultrasonic database containing 623 data entries combined with compressive strength tests, and designed a hybrid model integrating convolutional layers, an attention mechanism, and a residual network. The optimal model was obtained by optimizing hyperparameters through SHAP analysis, providing a reliable solution for assessing the strength of in-service concrete. Albuthbahak and Hiswa [26] developed seven supervised machine learning regression models based on 436 concrete datasets, incorporating Ultrasonic Pulse Velocity (UPV) and mix proportion parameters like water-cement ratio to predict concrete compressive strength. Asteris et al. [27] employed six soft computing models to estimate the strength of concrete samples based on data obtained from NDT (ultrasonic pulse velocity and rebound hammer methods).

In summary, drawing on the extensive achievements in concrete strength prediction, machine learning-based regression methods offer a new approach for predicting the strength of grouting materials within sleeves. Although the hardening process, stress state, and multi-factor coupling effects (such as mix proportion and curing environment) of sleeve grouting materials differentiate them from ordinary concrete, the propagation characteristics of ultrasonic waves in materials are closely related to parameters like density and elastic modulus, which in turn are intrinsically linked to strength. Therefore, combining ultrasonic testing with machine learning algorithms holds promise for achieving accurate and efficient prediction of grouting material strength.

To achieve this goal, this study aims to construct a reliable machine learning model that uses SAFT-extracted features as inputs and the curing age as the output to predict the curing stage of grouting material inside sleeves. Given the significant evolution of grout strength with curing age, the predicted curing age can further serve as an indirect indicator of the strength development state of the grout, thereby reducing reliance on time-consuming and labor-intensive laboratory tests. Crucially, existing studies [24–27] have largely focused on features and models themselves, while generally overlooking the hyperparameter optimization step, which has a decisive impact on model performance. These studies often rely on default settings, failing to fully realize the predictive potential of the models. To address this research gap, the core innovations and contributions of this paper are as follows:

1. Establishment of a systematic comparative framework for hyperparameter optimization: Based on the same SAFT feature system and XGBoost model, three optimization strategies—Random Search, Bayesian Optimization, and Adaptive Particle Swarm Optimization—are implemented in parallel and systematically for the first time, providing a preliminary reference for algorithm comparison and model design under larger sample conditions in future work.

2. Quantitative validation of the performance advantages of optimization algorithms: Through rigorous experimental comparison, Bayesian Optimization achieved relatively superior predictive performance under the data conditions of this study, indicating that hyperparameter optimization significantly impacts this type of task. Furthermore, the predictive stability of the model under limited independent component conditions was preliminarily examined through supplementary leave-one-specimen-out cross-validation (LOSO-CV).

3. Clarification of the critical role of hyperparameter optimization in the "ultrasonic feature–strength prediction" pipeline: This provides a reusable analytical framework for model design and algorithm selection under larger sample conditions in future research.

Through the work outlined above, this study not only strives to improve prediction accuracy and provide an efficient tool for engineering decision-making but also aims to deepen the methodological practice of machine learning in the field of construction materials, promoting its evolution from mere "usability" to "optimal application."

2 Research Methodology

To achieve the research objectives, the first step involved collecting ultrasonic testing data before grouting and at curing ages of 1, 3, 7, and 28 days. High-resolution imaging of the grouting material inside the sleeves was performed, and the SAFT feature value along with its internal distribution (represented by the area percentages of six color regions from purple to red) were extracted to create the dataset. Concurrently, standard specimens of the same age were subjected to compressive strength tests. The compressive strength results were then compared and analyzed against the SAFT feature values to validate the correlation between the SAFT feature value and the development of compressive strength in the grouting material during curing. The dataset was partitioned into training, validation, and test sets. During the training phase, three optimization algorithms were employed to identify the optimal hyperparameters for the XGBoost model, resulting in the best hyperparameter combination. Subsequently, the trained model was used for time-dependent prediction analysis on the test set, followed by an analysis of the input features.

2.1 Synthetic Aperture Focusing Technique (SAFT)

The Synthetic Aperture Focusing Technique (SAFT) is a signal processing method used in ultrasonic imaging, primarily aimed at enhancing image resolution by focusing reflected ultrasonic signals [28]. This technique involves moving a single probe or an array probe to different positions for transmitting/receiving signals. The echo data measured from multiple acquisition positions are weighted and reconstructed to synthesize a higher-resolution image, yielding clearer and more detailed zonal information [29].

Based on the propagation distance and ultrasonic velocity, the resolution of SAFT images is improved using methods that account for signal time delays and beamforming. The principle of SAFT is illustrated in Fig. 1. The total propagation distance of the ultrasonic wave is expressed as:

$$d_{aa}\left(x_O, z_o\right) = 2\sqrt{\left(x_a - x_o\right)^2 + \left(z_a - z_o\right)^2} \tag{1}$$

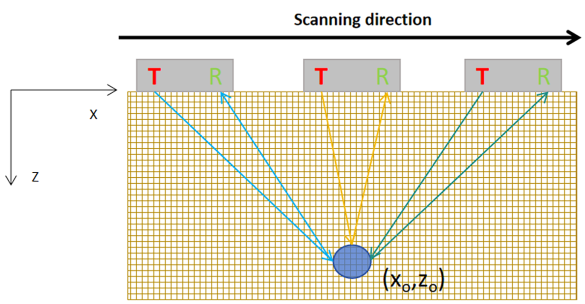

Figure 1: Schematic diagram of synthetic aperture focusing technology (SAFT) principle.

The corresponding time-of-flight $t_{aa}\left(x_o, z_o\right)$ is given by:

$$t_{aa}\left(x_o, z_o\right) = \frac{d_{aa}\left(x_o, z_o\right)}{c} \tag{2}$$

where c is the propagation velocity of the ultrasonic wave in the material, which can be calculated based on the difference in arrival times between the first and second wave reflections.

The pixel intensity $I\left(x_o, z_o\right)$ for each focal point is obtained by summing the signals from each transducer position corresponding to the time-of-flight to the focal point. The pixel intensity at each focal point is related to the time-of-flight and can be expressed as:

$$I\left(x_o, z_o\right) = \left|\sum_{i=1}^{n} h\left[s\left(t_a\left(x_o, z_o\right) + t_w\right)\right]\right| \tag{3}$$

where $s\left(t_a\left(x_o, z_o\right)\right)$ represents the signal received by the transducer, and n denotes the number of scans performed by the transducer along the surface.

2.2 XGBoost Regression Model

XGBoost is a gradient-boosted tree-based regression model known for its high computational efficiency and accuracy. In the XGBoost model, optimization is performed through an iterative process of weighted tree addition [30]. In each iteration, the model updates its parameters based on the residuals—the differences between predicted and actual values [31]. Assuming K is the total number of trees, the final prediction is obtained by a weighted summation of the outputs from all individual trees.

$$\hat{y}_i = \sum_{k=1}^{K} f_k\left(x_i\right) \tag{4}$$

In this equation, $f_k\left(x_i\right)$ is the output of the k-th tree, x_i is the i-th sample, and K is the number of trees. Each tree $f_k\left(x\right)$ can be represented as a decision tree, given by:

$$f_k\left(x\right) = w_{q(x)}\left(x\right)\left(q: R^m \rightarrow T, w \in R^T\right) \tag{5}$$

where q(x) represents the tree structure that maps the input x to a leaf node, T denotes the number of leaves in the tree, and w is the weight associated with each leaf node. The objective function of XGBoost is defined as:

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \left[\mathcal{L}(y_i, \hat{y}_i) + \Omega(f_i) \right] \tag{6}$$

here, N is the number of samples, $\mathcal{L}(y_i, \hat{y}_i)$ is the loss function that measures the model's fit to the data, and $\Omega(f_i)$ is the regularization term that penalizes model complexity.

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{7}$$

In this expression, T is the number of leaf nodes in the tree, w_j is the weight of the j-th leaf node, and γ and λ are regularization coefficients that control the complexity of the tree. Given the demonstrated performance of the XGBoost regression model in various regression prediction tasks, it was selected as the base model for this study.

2.3 Hyperparameter Tuning Optimization Algorithms

In regression prediction using the XGBoost model, hyperparameter tuning plays a critical role as it directly influences the model's performance and generalization capability. An appropriate hyperparameter configuration can enhance the model's ability to fit the current data, more accurately uncover the intrinsic relationships between features and the target, effectively prevent overfitting and underfitting, thereby improving the model's stability and predictive accuracy on unseen data. Furthermore, proper hyperparameter selection helps reduce model training time, conserve computational resources, and improve the efficiency of model deployment and practical application.

Commonly used hyperparameter tuning algorithms include Grid Search and various optimization algorithms. Although the former is straightforward to use, it requires substantial computational resources and becomes time-consuming in high-dimensional parameter spaces. In contrast, other optimization algorithms offer significant advantages in terms of efficiency, adaptability, and global search capability, making them more favored in practical applications. Therefore, this study selects the Random Search algorithm, Bayesian Optimization, and the Adaptive Particle Swarm Optimization algorithm to tune the key hyperparameters of the XGBoost model. These optimization algorithms each have distinct characteristics: the Random Search algorithm efficiently approximates the optimal solution with fewer trials by randomly sampling and evaluating combinations within the parameter space, avoiding an exhaustive search; Bayesian Optimization guides the selection of subsequent hyperparameters by building a surrogate model based on historical evaluation results, thereby finding the best parameter combination with fewer trials; the Adaptive Particle Swarm Optimization algorithm simulates the social behavior of bird flocks, dynamically adjusting learning factors and inertia weights during the search process to effectively balance global exploration and local exploitation, enhancing convergence speed and optimization precision.

2.3.1 Random Search Algorithm

The Random Search algorithm is an efficient and easily implementable method for hyperparameter optimization. Its core concept involves defining a hyperparameter space and randomly sampling candidate configurations from this space according to specified probability distributions. The model performance of each configuration is evaluated, and ultimately the hyperparameter combination with the best performance is selected [32].

First, a search space Θ containing all hyperparameters to be optimized is defined. It is the Cartesian product of the domains Λ_i of each hyperparameter:

$$\Theta = \Lambda_1 \times \Lambda_2 \times \ldots \times \Lambda_k \tag{8}$$

here, each Λ_i can be a continuous interval (e.g., learning rate) or a discrete set (e.g., tree depth).

Subsequently, random sampling is performed. The algorithm independently and randomly performs N samplings from the defined space Θ. The hyperparameter combination obtained from the j-th sampling is:

$$\theta^{(j)} \sim P(\Theta) \tag{9}$$

here, $P(\Theta)$ is the probability distribution defined over the hyperparameter space, typically a uniform distribution or log-uniform distribution.

Finally, for each sampled point $\theta(j)$ its loss L (e.g., Mean Squared Error) is evaluated on the validation set D_{val}. The hyperparameter combination that minimizes the loss is selected as the output:

$$\theta^* = \arg\min_{\theta(j)} L\left(f_{\theta(j)}, D_{\mathrm{val}}\right) \tag{10}$$

The sampling process in Random Search can be based on uniform distributions, log-uniform distributions, or other prior distributions, depending on the nature of the hyperparameters.

2.3.2 Bayesian Optimization Algorithm

Bayesian Optimization is an efficient global optimization algorithm and stands as one of the most advanced and promising techniques in the fields of probabilistic machine learning and artificial intelligence [33]. It is a sequential optimization strategy based on Bayes' theorem, particularly suited for black-box optimization problems where evaluating the objective function is computationally expensive. Its core idea involves constructing a probabilistic surrogate model of the objective function and utilizing an acquisition function to balance exploration and exploitation, thereby efficiently searching for the global optimum.

Unlike traditional methods such as Grid Search, Bayesian hyperparameter optimization leverages information from previous evaluations to guide the selection of subsequent hyperparameters [34]. Through the surrogate function, Bayesian Optimization effectively models the relationship between hyperparameters and the model's objective function, iteratively selects potentially superior hyperparameter combinations, and ultimately identifies the set yielding the best performance. Consequently, it offers significant advantages in terms of search efficiency and accuracy [35].

In Bayesian Optimization, Expected Improvement (EI) is a commonly used acquisition function. The core idea of the EI acquisition function is to select, at each step, the point offering the largest expected improvement over the current best value as the next point to evaluate. This expected improvement is calculated based on the predictive mean and standard deviation of the current model, considering the expected amount by which the objective function value at a new point might exceed the current best value. The formula for the EI acquisition function is as follows:

$$\mathrm{EI}(\mathrm{x}) = \mathrm{E}\left[\max\left(0, \mathrm{f}(\mathrm{x}) - \mathrm{f}_{\mathrm{best}}\right)\right] \tag{11}$$

where: $\mathrm{f}_{\mathrm{best}}$ is the best known value of the objective function. $\mathrm{f}(\mathrm{x})$ is the predicted value of the objective function at point x.

When using a Gaussian Process to model the objective function, $f(x)$ is assumed to follow a Gaussian distribution. Given the predictive mean $\mu(x)$ and standard deviation $\sigma(x)$ at point x, the EI can be expressed as:

$$\mathrm{EI}(x) = \sigma(x)\phi\left(\frac{\mu(x) - f_{\mathrm{best}}}{\sigma(x)}\right) + \left(\mu(x) - f_{\mathrm{best}}\right)\Phi\left(\frac{\mu(x) - f_{\mathrm{best}}}{\sigma(x)}\right) \tag{12}$$

where, $\mu(x)$ is the predictive mean of the objective function at point x; $\sigma(x)$ is the standard deviation of the predictive value of the objective function at point x, representing the uncertainty in the prediction; f_{best} is the best observed value of the objective function so far; $\phi(\cdot)$ is the probability density function (PDF) of the standard normal distribution; $\Phi(x)$ is the cumulative distribution function (CDF) of the standard normal distribution.

2.3.3 Adaptive Particle Swarm Optimization Algorithm

The Particle Swarm Optimization (PSO) algorithm is inspired by the collective behavior of bird flocks or fish schools, finding the optimal solution by balancing individual experience with collective group knowledge. Adaptive Particle Swarm Optimization (APSO) introduces adaptive mechanisms based on the standard PSO algorithm, significantly enhancing its convergence performance and robustness [36].

The APSO algorithm maintains a swarm of particles, where each particle represents a potential solution, i.e., a hyperparameter configuration. In a D-dimensional search space, the position of the i-th particle is denoted as $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iD})$, and its velocity is denoted as $\mathbf{v}_i = (v_{i1}, v_{i2}, \ldots, v_{iD})$. The algorithm updates the state of each particle using the following formulas:

1. Velocity Update Formula:

$$v_{id}^{t+1} = w^t \cdot v_{id}^t + c_1 \cdot r_1 \cdot \left(p_{id}^t - x_{id}^t\right) + c_2 \cdot r_2 \cdot \left(g_d^t - x_{id}^t\right) \tag{13}$$

2. Position Update Formula:

$$x_{id}^{t+1} = x_{id}^t + v_{id}^{t+1} \tag{14}$$

here, p_{id}^t is the personal historical best position of particle i, g_d^t is the global best position found by the entire swarm, r_1 and r_2 are random numbers within the range [0, 1], and c_1 and c_2 are learning factors.

The core innovation of the APSO algorithm lies in its adaptive parameter adjustment strategy:

Adaptive Inertia Weight: The inertia weight w^t is dynamically adjusted based on the algorithm's convergence state:

$$w^t = w_{\max} - \left(w_{\max} - w_{\min}\right) \cdot \frac{t}{T_{\max}} + \alpha \cdot \sigma^t \tag{15}$$

where σ^t represents a population diversity metric.

Adaptive Learning Factors: The cognitive and social learning factors are adjusted based on the search progress:

$$c_1^t = c_{1,\mathrm{initial}} + \left(c_{1,\mathrm{final}} - c_{1,\mathrm{initial}}\right) \cdot \frac{t}{T_{\max}} \tag{16}$$

$$c_2^t = c_{2,\mathrm{initial}} + \left(c_{2,\mathrm{final}} - c_{2,\mathrm{initial}}\right) \cdot \frac{t}{T_{\max}} \tag{17}$$

This adaptive mechanism allows the algorithm to emphasize global exploration in the early stages and focus on local fine-tuning in the later stages, effectively balancing the conflict between exploration and exploitation. The modeling and training process for the three hybrid models proposed in this study is illustrated in Fig. 2.

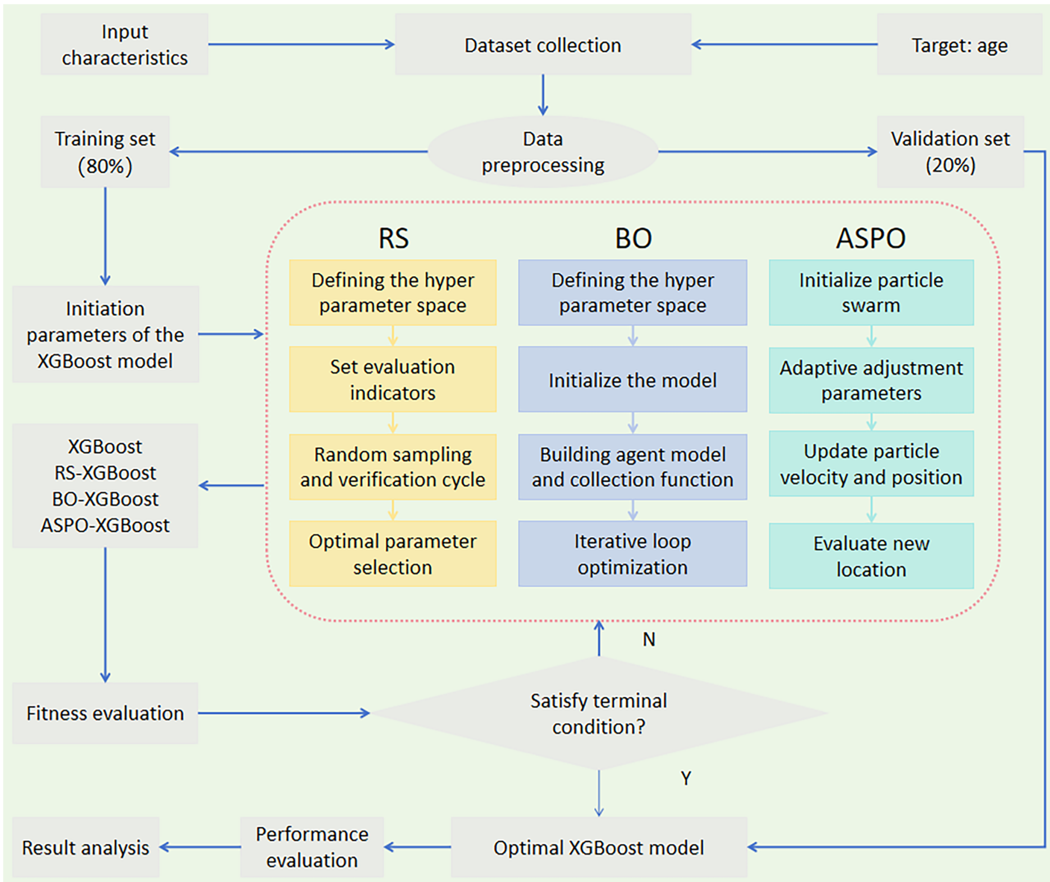

Figure 2: Framework diagram of the method in this paper.

3 Laboratory Experimental Verification

3.1 Dataset Collection

To collect ultrasonic data and corresponding strength data samples of grouting material at different curing ages, experiments were conducted using C85 grouting material with a water-to-material ratio of 12% to prepare standard specimens. Precast concrete columns measuring 500 mm × 500 mm × 700 mm (Fig. 3a) and precast concrete walls measuring 200 mm × 900 mm × 700 mm (Fig. 4a) were selected as the research subjects. The specimen design strictly followed the Technical Specification for Precast Concrete Structures (JGJ 1-2014). The columns were equipped with half-grouting sleeves with a diameter of 45 mm, length of 310 mm, and wall thickness of 10.2 mm, while the walls utilized full-grouting sleeves with a diameter of 42 mm, length of 310 mm, and wall thickness of 8.5 mm. All sleeves were arranged symmetrically in accordance with engineering standards (Figs. 3b and 4b). Typical defects such as grout leakage, voids, and reinforcement encapsulation were preset during specimen fabrication. An A1040 MIRA 3D ultrasonic acquisition instrument (Fig. 5) was used, which features a 4 × 8 dry-point contact shear wave transducer matrix. It employs electronic beam focusing and a spring-mounted design to adapt to surface roughness, ensuring stable coupling.

(a) (b)

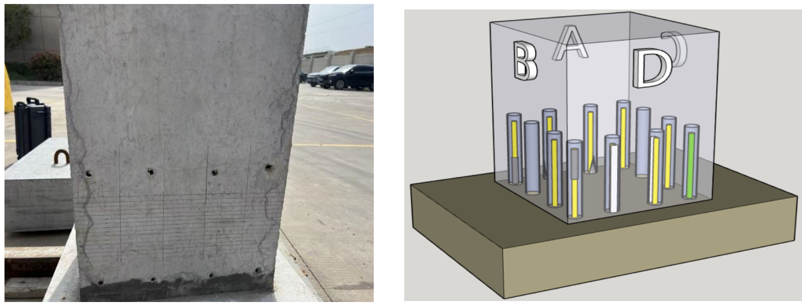

Figure 3: (**a**) Prefabricated concrete column (**b**) Internal spatial distribution diagram of prefabricated concrete column.

(a) (b)

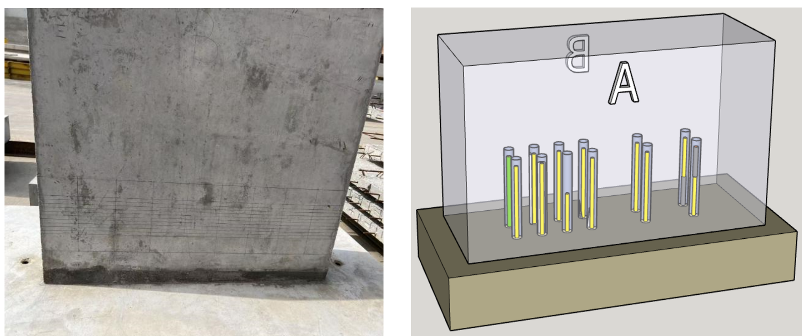

Figure 4: (**a**) Prefabricated concrete wall (**b**) Internal spatial distribution diagram of prefabricated concrete wall.

Figure 5: A1040 MIRA 3D acquisition data diagram.

Based on the requirements of other experimental studies, specimens were fabricated with preset typical defects, including grout leakage, voids, and reinforcement encapsulation. Consequently, two normally grouted sleeves were selected from the columns (Column Sleeve A and Column Sleeve B), and one normally grouted sleeve was selected from the wall (Wall Sleeve C). Ultrasonic data were systematically collected before grouting and at curing ages of 1, 3, 7, and 28 days. For each curing age, 16 sets of data were collected per sleeve in the columns, and 11 sets per sleeve in the wall. At each measurement point, five independent repeated ultrasonic acquisitions were performed under the same acquisition parameters. The SAFT feature values and color-region area ratios used for model training and testing were obtained from these five repeated acquisitions. Therefore, the 215 observations reported in this study refer to distinct measurement positions and curing-age conditions, rather than the total number of raw repeated acquisitions. This scheme comprehensively documented the evolution of ultrasonic characteristics during the strength development of the grouting material, providing data support for constructing a curing age prediction model based on ultrasonic features and a model for characterizing the strength development state of the grouting material.

Although a total of 215 ultrasonic observations were obtained in this study, these data originated from repeated measurements at different spatial locations and at different curing ages on only three physical sleeve specimens. Therefore, these observations are not completely independent in a statistical sense: they share the same material batch, grouting construction process, curing environment, and laboratory conditions. The setup that the spacing between scanning positions exceeds three times the ultrasonic beam width was primarily intended to reduce acoustic interference between adjacent measurement points and improve local measurement resolution, but it does not serve as a strict basis for statistical independence.

Consequently, this study should be regarded as a proof-of-concept investigation under controlled laboratory conditions. The purpose of repeated scanning is to describe the spatiotemporal evolution of ultrasonic features within each sleeve, rather than to require a large number of independent samples. To evaluate the cross-specimen performance of the model to the greatest extent possible, this paper further supplements a leave-one-specimen-out cross-validation (LOSO-CV) analysis.

The standard specimen size used in this study is 40 mm × 40 mm × 160 mm, with testing ages set at 1, 3, 7, and 28 days. The specific performance parameters of the grouting material are listed in Table 1. For each curing age, two sets of standard test blocks were prepared, with each set containing three parallel specimens. After reaching the designated curing age, the specimens were taken out of the standard curing room and allowed to sit in a natural environment for 6 h to ensure sufficient surface drying. During the testing process, a dividing line was first marked along the central axis of the long side of each specimen, dividing it into two parts, A and B. Subsequently, a three-point bending test was conducted to fracture the specimen along the marked line, obtaining separate A and B portions. Compressive strength tests were then performed on both portions, and the compressive strength values were recorded separately. The specific testing procedure is illustrated in Fig. 6.

Table 1: Performance parameters of grouting materials.

Standard Water-Binder Ratio	**12%**
Design strength	85 MPa
3–24 h Vertical expansion rate	0.02%–0.40%
Bleeding rate	0%
Chloride ion content	≤0.03%

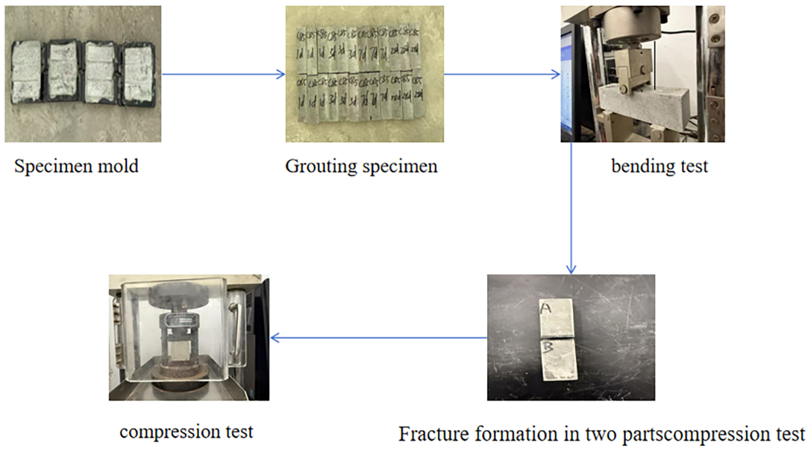

Figure 6: Experimental process.

3.2 Analysis of Experimental Data

In this study, the Synthetic Aperture Focusing Technique (SAFT) is employed to quantitatively analyze the grouting quality of precast concrete column sleeves. This is achieved by comparing B-scan images before grouting and at different curing ages (1, 3, 7, and 28 days) after grouting, as illustrated in Fig. 7a.

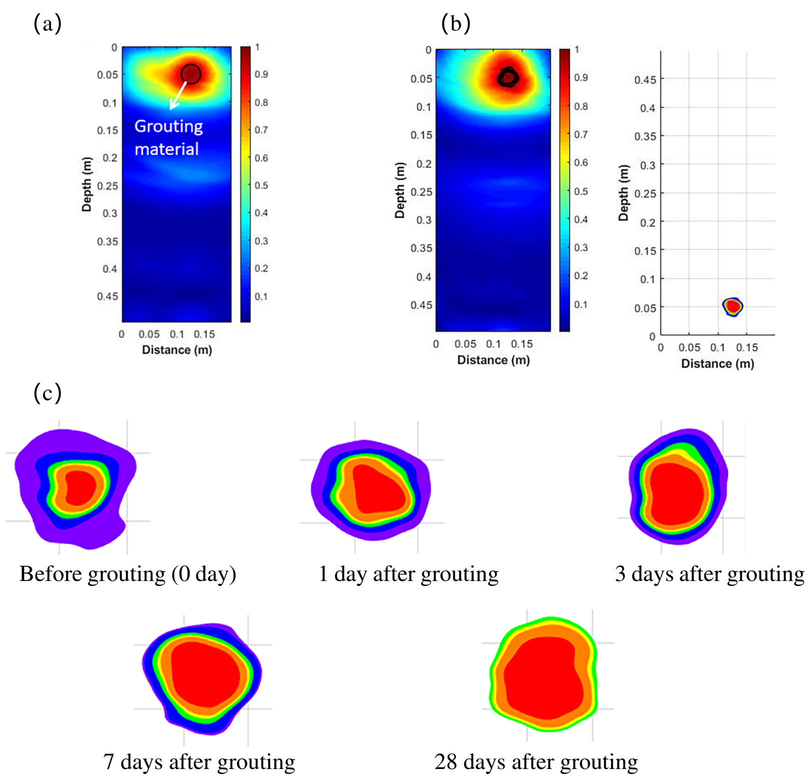

Figure 7: (**a**) Samples of precast concrete column B-scan (**b**) Samples of B scanning layered diagram (**c**) Samples of B scanning layered enlarged view.

This paper proposes a method for extracting SAFT feature values based on the equivalent area. The specific procedure is as follows. First, each original SAFT B-scan image is denoted as $I_n(x, y)$, and the pixel intensities are linearly mapped to the interval [0, 1] using min–max normalization:

$$I_n\left(x, y\right) = \frac{I\left(x, y\right) - I_{\min}}{I_{\max} - I_{\min}} \tag{18}$$

where $I_{\min}$ and $I_{\max}$ represent the minimum and maximum pixel values in the image, respectively, and $I_n(x, y)$ denotes the normalized pixel intensity. Normalization reduces the impact of absolute amplitude differences between images on subsequent feature extraction.

Based on the normalized image, the position of the grouting sleeve within the B-scan image is further determined. According to previous research [37], after noise-suppressed SAFT imaging, the grouting sleeve region appears as a distinct high-reflection area in the image. By incorporating the known structural information that the sleeve is located approximately 5 cm from the concrete inspection surface, the region with the highest local reflection intensity within the corresponding depth range is identified. The geometric center of this strongest reflection region is then taken as the center coordinates of the grouting sleeve (x_c, y_c). This localization method leverages the significant reflection characteristics generated by the acoustic impedance differences between the sleeve, its internal grouting material, and the surrounding concrete, thereby achieving stable identification of the sleeve position.

After determining the sleeve center, an equivalent circular area is constructed in the image based on the actual structural dimensions of the sleeve. Assuming the equivalent radius of the sleeve in the image is r, an equivalent circle region C is defined with (x_c, y_c) as the center and r as the radius, serving as the target region for subsequent feature extraction. In this paper, the average normalized pixel value within region C is calculated as the SAFT feature value for that sample:

$$F_{SAFT} = \frac{1}{N_C} \sum\nolimits_{(x,y)\in C} I_n\left(x, y\right) \tag{19}$$

where N_C is the total number of pixels within the equivalent circle region C. This feature value characterizes the overall ultrasonic response level of the grouting sleeve region.

For the extraction of color region area ratio features, this paper employs a single set of globally fixed thresholds for all specimens and all curing ages, rather than adaptively calculating thresholds for individual images. The threshold set is defined as:

$$T = \{t_0, t_1, t_2, t_3, t_4, t_5, t_6\} \tag{20}$$

where $t_0 = 0$ and $t_6 = 1$. The five internal thresholds t_1 to t_5 were calibrated only using the average SAFT feature values of the two column sleeves, namely Column Sleeve A and Column Sleeve B, before grouting (0 d) and after grouting at 1, 3, 7, and 28 d. The Wall Sleeve C data were not used in the threshold calibration process. After calibration, these thresholds were fixed and then applied unchanged to all samples, including Wall Sleeve C, for extracting the color-region area ratio features, respectively:

$t_1 = 0.9301$, $t_2 = 0.9441$, $t_3 = 0.9512$, $t_4 = 0.9555$, $t_5 = 0.9652$

Accordingly, the pixels within the equivalent circle region are divided into six intensity intervals:

$$[t_0, t_1), [t_1, t_2), [t_2, t_3), [t_3, t_4), [t_4, t_5), [t_5, t_6]$$

These correspond to the six color regions described in the text, namely:

Purple: [0, 0.9301), Blue: [0.9301, 0.9441), Green: [0.9441, 0.9512), Yellow: [0.9512, 0.9555), Orange: [0.9555, 0.9652), Red: [0.9652, 1]

The area ratio of each color region is defined as the proportion of the area corresponding to that intensity interval to the total area of the equivalent circle region.

As shown in Fig. 7b,c, quantifying the area ratios of the various color regions reveals the evolution of the grouting material strength over time. The reference values are derived from the average SAFT feature values of samples before grouting (0 d) and after grouting at 1, 3, 7, and 28 d, with the evolution trend illustrated in Fig. 8a. The five internal thresholds mentioned above remain fixed once determined and are uniformly applied to classify the color regions in all images, without individual adjustments for different specimens or curing ages. Therefore, the extracted color region area ratios in this study share a consistent methodological basis for comparison.

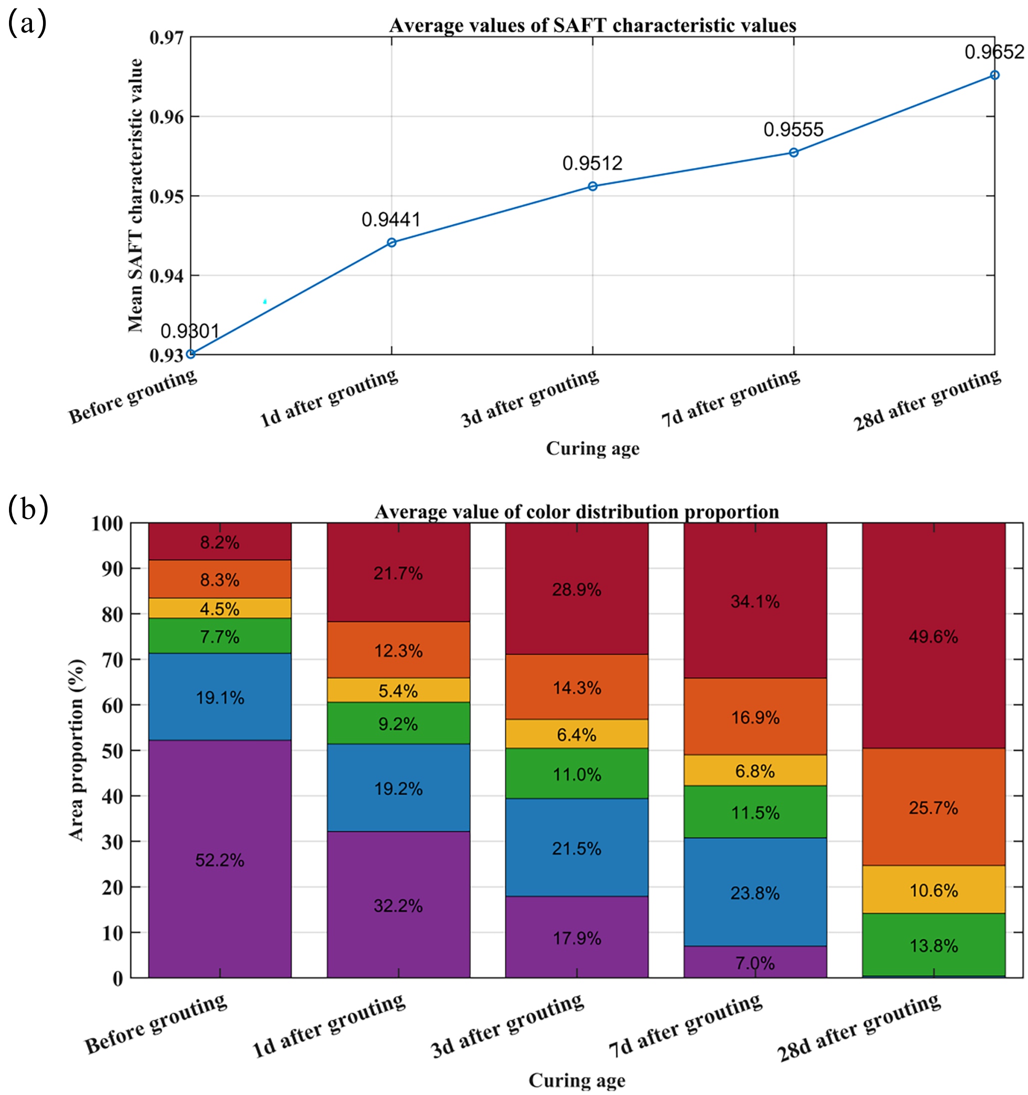

Figure 8: (**a**) Average values of SAFT characteristic values (**b**) Average value of color distribution proportion.

The statistical characteristics and distribution of these samples are presented in Fig. 8a,b. Data analysis indicates that the average SAFT feature value gradually increases from 0.93010 before grouting (0 d) to 0.96520 at 28 d after grouting, showing an overall upward trend with extended curing age. This result suggests that as the hydration reaction of the cementitious grouting material progresses, the internal pore structure of the material gradually improves, and the medium uniformity and compactness continuously increase, thereby enhancing the ultrasonic response intensity in the sleeve region. In other words, the increase in the SAFT feature value reflects, to some extent, the development process of the grouting material from an early dispersed, heterogeneous state to a denser, more uniform state in the later stages.

All data used in this study were acquired using the A1040 MIRA 3D device under fixed acquisition parameters, including consistent probe frequency, gain, and sampling rate, thereby ensuring consistency in the conditions for SAFT image acquisition and the basis for feature extraction. On this foundation, the six color region thresholds employed in this paper are all established based on uniformly normalized SAFT images and remain unchanged across all specimens and curing age samples, thus providing a unified basis for comparison. It should be noted that this threshold system is established under the specific testing equipment and acquisition parameters used in this study. Should the testing equipment, acquisition parameters, or imaging conditions change significantly, it is advisable to reapply data normalization and recalibrate the threshold intervals accordingly to ensure the comparability of the extracted features.

As shown in Figs. 7c and 8b, the variations in the area ratios of different feature value intervals clearly illustrate the staged evolution of the internal structure of the grouting material. In the early stage (1–3 days), the proportion of low feature value regions (purple, blue) sharply decreased from 71.34% to 17.89%, indicating the activation of hydration reactions, the formation of the initial structure, and the rapid reduction of unreacted areas. In the middle stage (3–7 days), the proportions of medium-to-high feature value regions (green, yellow) tended to stabilize or increased slightly, while the high feature value regions (orange, red) began to expand, reflecting the processes of accelerated hydration and pore refinement. In the later stage (7–28 days), the proportion of high feature value regions significantly increased from 51.98% to 75.29%, while low-value regions nearly disappeared, indicating that the material had become highly dense and mature. This distributional shift from a “broad peak with low values” to a “narrow peak with high values” non-destructively characterizes the continuous improvement in the internal structural uniformity and mechanical properties of the grouting material.

Both the SAFT feature value and the compressive strength exhibited significant increases with curing age, as shown in Fig. 9, and the evolution patterns of the two were highly consistent: the fastest growth occurred in the early stage, slowed in the middle stage, and tended to stabilize in the later stage. The SAFT feature value not only reflects the overall trend of strength development through its increase in mean value but also accurately depicts the internal densification process of the material through the quantitative changes characterized by the shrinkage of low-value regions and the expansion of high-value regions in the spatial distribution. Therefore, the SAFT feature value and its distribution characteristics constitute a sensitive and reliable non-destructive monitoring framework capable of dynamically tracking the internal structural evolution of the grouting material and providing an effective technical means for assessing its strength development state.

It is important to note that the primary role of min–max normalization is to unify the numerical range of pixel intensities and improve the consistency of feature extraction within the same acquisition system; it does not physically eliminate the influences of factors such as sleeve diameter, wall thickness, surrounding concrete volume, sleeve–concrete interface geometry, and lateral confinement on shear wave propagation and SAFT imaging results. Therefore, this study does not presuppose the inherent transferability of the normalized SAFT features across different sleeve configurations, but rather provides a preliminary examination of their empirical applicability under the existing limited specimen conditions.

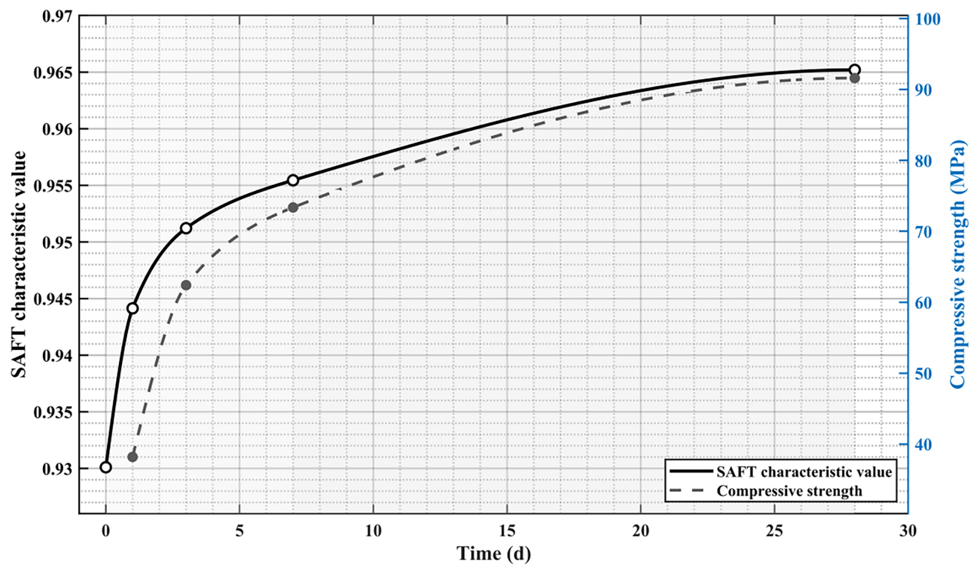

Figure 9: Shows the variation of SAFT characteristic values and compressive strength of grouting materials with age.

3.3 Model Hyperparameter Tuning

Among the three optimization algorithms adopted in this study, three hyperparameters that significantly influence the XGBoost model—namely, maximum depth (Max_depth), learning rate (Learning rate), and number of estimators (Number of estimators)—were selected for tuning. Their search ranges and configurations are presented in Table 2. The model was implemented in Python, with the operating environment consisting of a CPU: Intel (R) Core (TM) i7-14700K CPU @ 5.6 GHz and a GPU: NVIDIA GeForce RTX 4070. In this section, the model training dataset was derived from the ultrasonic inspection data of two half-grouting sleeves (Column Sleeve A and Column Sleeve B) in the precast concrete columns. To evaluate model performance, a stratified sampling strategy was employed to divide the training dataset into a training set (80%) and a validation set (20%). The evaluation of cross-specimen predictive capability was supplemented by the leave-one-specimen-out cross-validation (LOSO-CV) analysis conducted in Section 3.7, using sleeves as the grouping units. Within the gradient boosting decision tree framework, the model parameters were iteratively optimized during training. The optimal hyperparameter combinations for the different models are listed in Table 2, with all parameters adopting the default values provided by the Python libraries.

3.4 Training Results under Different Optimized Models

To comprehensively evaluate the performance of each model after hyperparameter tuning, this study compares the performance of the baseline XGBoost model with models optimized by three optimization algorithms (RS, BO, and APSO) on both the training set and the validation set. Mean Squared Error (MSE) and the coefficient of determination (R^2) were selected as the core evaluation metrics. The former measures the deviation between predicted and actual values, while the latter indicates the model's ability to explain the target variable. Detailed results are presented in Table 3. In terms of overall performance, the models tuned by optimization algorithms, especially BO-XGBoost, achieved superior error metrics and fitting results on the validation set. Specifically, BO-XGBoost achieved the lowest MAE (0.3117) and the highest R^2 (0.9583) on the validation set, indicating good predictive consistency under the current training data partition. APSO-XGBoost performed second best on the validation set (MAE = 0.3623, R^2 = 0.9560), and its generalization error was also better than that of the baseline model. In contrast, the baseline XGBoost and RS-XGBoost both had MAE values exceeding 0.42 on the validation set, with relatively lower R^2 values, suggesting certain limitations in their generalization capability.

Table 2: Search range and optimal combination of hyper-parameters.

Model	Parameter	Search Range	Optimal Value	Cost Time
XGBoost	/	/	Default value	5.621 s
RS-XGBoost	Max_depth	[1, 20]	5	26.315 s
	Learning rate	[0.01, 1]	0.01	
	Number of estimators	[10, 1000]	953	
BO-XGBoost	Max_depth	[1, 20]	8	35.753 s
	Learning rate	[0.01, 1]	0.18	
	Number of estimators	[10, 1000]	886	
ASPO-XGBoost	(Max_depth)	[1, 20]	12	314.291 s
	Learning rate	[0.01, 1]	0.1	
	Number of estimators	[10, 1000]	203	

Table 3: Training evaluation indexes of different models.

Model	Dataset	MAE	R^2
XGBoost	Training set	0.1968	0.9973
	Validation set	0.4256	0.9476
RS-XGBoost	Training set	0.4013	0.9497
	Validation set	0.4296	0.9452
BO-XGBoost	Training set	0.2982	0.9657
	Validation set	0.3117	0.9583
ASPO-XGBoost	Training set	0.2876	0.9791
	Validation set	0.3623	0.9560

A deeper analysis of the performance disparity between the training set and validation set reveals that the baseline XGBoost model achieved an extremely high R^2 (0.9973) and a very low MAE (0.1968) on the training set, but its validation set performance showed a significant gap (MAE: 0.4256, R^2: 0.9476), indicating pronounced overfitting. This suggests that the model memorized the training data to some extent rather than learning generalizable patterns. The improvement in generalization capability for RS-XGBoost was limited, with its validation set MAE (0.4296) even slightly higher than that of the baseline model, and its training set metrics were relatively lower (MAE: 0.4013, R^2: 0.9497). In contrast, for BO-XGBoost and APSO-XGBoost, the MAE and R^2 metrics on the training and validation sets were much closer. Notably, BO-XGBoost exhibited the most balanced performance between the two sets, demonstrating the significant advantage of Bayesian optimization in finding the optimal hyperparameter combination that balances model complexity and predictive capability.

From an algorithm selection perspective, BO-XGBoost demonstrated superior overall efficiency in this study. For the XGBoost tuning task described in this paper, only three key hyperparameters were involved, and the search ranges were relatively limited. Therefore, this task can be considered a low-dimensional black-box optimization problem with a constrained search space. The results in Tables 2 and 3 indicate that Bayesian optimization achieved the best test set metrics at a lower computational cost. In contrast, although APSO possesses strong global search capabilities, it incurred significantly higher computational cost for this task (314.291 vs. 35.753 s), and its final test performance was still slightly inferior to that of BO-XGBoost. Consequently, the results of this study do not support considering APSO as the preferred optimization strategy for this task; instead, they indicate that Bayesian optimization offers greater practicality for such low-dimensional XGBoost hyperparameter optimization problems. It should be noted that the potential advantages of APSO are more likely to manifest in optimization scenarios characterized by higher dimensionality, more numerous peaks, or more irregular evaluation functions, though this was not validated on the dataset used in this study.

In summary, under the data partitioning conditions of this study, hyperparameter optimization contributed to improving the empirical predictive performance of the XGBoost model. Among the variants, BO-XGBoost achieved relatively better error metrics on the validation set and was therefore selected as the preferred model configuration for subsequent analyses.

3.5 Comparison of Prediction Results from Different Optimized Models

To objectively and comprehensively evaluate the final performance of the XGBoost models optimized by different algorithms, this section conducts predictive analysis on an independent test set that was not involved in training or hyperparameter tuning. The dataset consists of ultrasonic inspection data from a full-grouting sleeve (Wall Sleeve C) in a precast concrete wall. To further compare the predictive performance of the four models, four error evaluation metrics were introduced to quantify the performance of the optimized XGBoost models, including Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). Detailed results are shown in Fig. 10 and Table 4.

Among all compared models, the BO-XGBoost (Bayesian Optimization) model demonstrated outstanding and comprehensive predictive performance. It achieved the lowest values across all error metrics (MAE: 0.2037, MAPE: 5.56%, MSE: 0.2130, RMSE: 0.3339) while maintaining the highest coefficient of determination (R^2: 0.9881). This result indicates that BO-XGBoost achieved the lowest prediction error for curing age on the current independent sleeve, suggesting that the model can effectively capture the correspondence between SAFT features and the curing stage. Considering the consistency between curing age and compressive strength development, this indicates that the model has potential for characterizing the strength development state of the grouting material. However, since this study has not specifically characterized repeated SAFT measurements under the same sleeve, same curing age, and same scanning position, the MAPE of 5.56% should still be understood as an empirical error under the current data conditions, rather than directly interpreted as the true prediction accuracy. Considering further that the test involved only one independent sleeve, this result is more suitable as preliminary evidence of methodological feasibility rather than sufficient proof of engineering generalizability.

The APSO-XGBoost model also performed excellently, achieving an R^2 value (0.9812) second only to BO-XGBoost and significantly higher than the baseline model and RS-XGBoost. In terms of MAPE (6.93%) and MAE (0.2510), it also exhibited low empirical error on the current test set. Although its MSE and RMSE were slightly higher than those of BO-XGBoost, indicating slightly inferior handling of individual outliers, overall, it still demonstrates the effectiveness of this optimization algorithm under the conditions of this study.

In contrast, the baseline XGBoost model and the RS-XGBoost model showed relatively inferior predictive performance. Although the baseline model achieved a reasonable R^2 (0.9783), its MAE (0.2935) and MAPE (8.76%) were significantly higher than those of BO-XGBoost and APSO-XGBoost, revealing the limitations of its default parameters in terms of generalization capability. RS-XGBoost performed the worst among all optimized models, with its highest MAPE (9.56%) and RMSE (0.5654) indicating insufficient generalization capability of the parameter combination obtained through random search and poor prediction stability, failing to effectively improve the model's performance on unseen data.

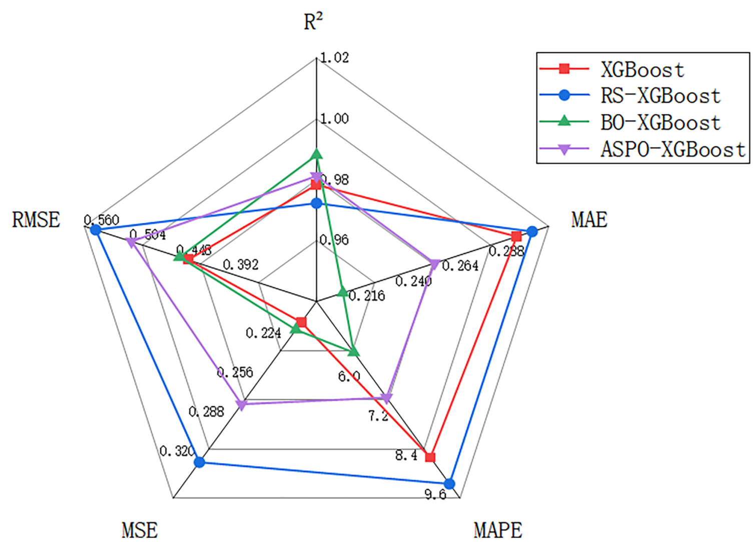

Figure 10: Radar chart of prediction indicators.

Table 4: Predictive evaluation indicators of different models.

Model	R^2	MAE	MAPE	MSE	RMSE
XGBoost	0.9783	0.2935	8.76%	0.2069	0.4549
RS-XGBoost	0.9723	0.3015	9.56%	0.3207	0.5663
BO-XGBoost	0.9881	0.2037	5.56%	0.2130	0.4615
ASPO-XGBoost	0.9812	0.2510	6.93%	0.2736	0.5231

From the perspective of engineering standards, the performance requirements for grouting materials for steel bar connections are typically expressed in terms of minimum strength indicators rather than prescribed error limits for predictive models. According to JGJ 1-2014, the compressive strength requirements for normal-temperature sleeve grouting materials are no less than 35 MPa at 1 day, no less than 60 MPa at 3 days, and no less than 85 MPa at 28 days. In this study, the MAPE of the optimal model on the test set was 5.56%; however, this value was calculated for curing-age prediction rather than direct strength prediction. Because the relationship between curing age and compressive strength is nonlinear, the age-prediction error cannot be directly converted into an absolute strength error by multiplying it by the 28-day strength requirement. Therefore, the present results should be interpreted as indicating the potential of the proposed method for characterizing strength development trends under controlled laboratory conditions, rather than as providing a direct MPa-level strength prediction accuracy. Further studies should establish an experimentally calibrated age–strength relationship and conduct appropriate error propagation before quantifying the corresponding strength prediction error.

Based on the comprehensive analysis above, under the current dataset and testing conditions, BO-XGBoost achieved relatively optimal empirical prediction metrics; APSO-XGBoost also demonstrated good competitiveness. This result indicates that hyperparameter optimization strategies hold application value for the task in this study.

3.6 Statistical Testing of Model Prediction Errors

To avoid evaluating model performance solely based on point estimate metrics such as R^2, MAE, and MAPE, this study further conducts statistical tests using the per-sample absolute prediction errors of each model on the same test set. Specifically, using 55 test samples, the per-sample absolute prediction error is calculated for each model, and paired error differences are constructed between BO-XGBoost and each comparison model, i.e., (per-sample absolute error of BO-XGBoost)−(per-sample absolute error of the comparison model). On this basis, a two-sided paired t-test is employed to compare the error differences between models, and the results are presented in Table 5. In the paired t-test, the paired t-statistic reflects the direction and relative magnitude of the difference in per-sample absolute errors, whereas the p-value indicates the probability that such a difference arises from random fluctuation. Since the statistical test is based on 55 test samples, the degrees of freedom are 54. When both the mean error difference and the paired t-statistic are negative, it indicates that the per-sample absolute prediction error of BO-XGBoost is smaller than that of the corresponding comparison model.

Table 5: Results of paired t-tests for absolute errors per sample across different models.

Comparison Model	Mean Difference (BO-Contrast Model)	Paired t-Value	p Value	Result Note
BO-XGBoost vs. XGBoost	−0.0821	−5.2	3.14×10^{-6}	BO-XGBoost has smaller errors
BO-XGBoost vs. RS-XGBoost	−0.1074	−6.43	3.46×10^{-8}	BO-XGBoost has smaller errors
BO-XGBoost vs. ASPO-XGBoost	−0.056	−8.89	3.71×10^{-12}	BO-XGBoost has smaller errors

As shown in Table 5, the mean per-sample absolute error differences between BO-XGBoost and the baseline XGBoost, RS-XGBoost, and APSO-XGBoost are −0.0821, −0.1074, and −0.0560, respectively, with corresponding paired t-statistics of −5.20, −6.43, and −8.89, and two-sided p-values of 3.14×10^{-6}, 3.46×10^{-8}, and 3.71×10^{-12}. These results demonstrate that, under the current test set, BO-XGBoost achieves significantly smaller per-sample absolute prediction errors than the other three models, indicating superior error control capability.

The above findings show that the advantage of BO-XGBoost is reflected not only in comprehensive evaluation metrics such as R^2, MAE, and MAPE, but is also supported by statistical tests at the per-sample error level. In particular, compared with APSO-XGBoost, although the differences in comprehensive metrics are relatively small, the per-sample absolute error comparison still reveals better predictive performance for BO-XGBoost. Therefore, the conclusion that BO-XGBoost delivers the best predictive performance is based not only on point estimate metrics and graphical analysis, but is also corroborated by the statistical comparison of errors. Nevertheless, given the limited test sample size and the relatively homogeneous data source in this study, these conclusions apply primarily to the current data conditions; their robustness in broader engineering scenarios remains to be further validated.

3.7 Cross-Specimen Cross-Validation

To avoid drawing inferences solely based on the results obtained from a single "column sleeve training—wall sleeve testing" partition, this study supplements the original analysis with leave-one-specimen-out cross-validation (LOSO-CV). Given that all ultrasonic data originate from three independent sleeve specimens, namely Column Sleeve A, Column Sleeve B, and Wall Sleeve C, the cross-validation was conducted using the sleeve specimens as grouping units: in each round, all scan data from one of the sleeves were sequentially selected as the test set, while the data from the remaining two sleeves served as the training set. This grouping strategy prevents repeated observations from the same sleeve from simultaneously appearing in both the training and test sets, thereby better reflecting the model's predictive performance when confronted with unseen entire validation components compared to the original point-wise partitioning approach. In the LOSO-CV analysis, the BO-XGBoost hyperparameters were not re-optimized in each round; instead, the same hyperparameter configuration obtained from Bayesian optimization was used for all three LOSO rounds.

It should be noted that although this strategy is more stringent than the original single hold-out test, the total number of physically independent specimens in this study remains only three. Therefore, the results of this cross-validation should be understood as a preliminary cross-specimen validation under limited component conditions, rather than being equated with sufficient statistical proof of the model's robust generalization capability under large-sample conditions. In other words, while the multiple scanning positions and curing ages in this study provide rich spatiotemporal observational information for the model, these observations still share the same or similar material systems, experimental operations, and laboratory environments, and thus cannot be simply regarded as completely independent statistical samples.

Table 6 presents the prediction results of the BO-XGBoost model in the three rounds of leave-one-specimen-out cross-validation. It can be observed that when Wall Sleeve C was used as the test set and Column Sleeves A and B as the training set, the model achieved the highest prediction accuracy, with an R^2 of 0.9881, MAE of 0.2037, and MAPE of 5.56%. When Column Sleeve A served as the test set, the R^2 was 0.9873, MAE was 0.2113, and MAPE was 5.96%. When Column Sleeve B served as the test set, the R^2 was 0.9868, MAE was 0.2161, and MAPE was 6.13%. All three rounds maintained high R^2 values and low error levels, indicating that under the current experimental conditions, the model exhibits good stability and consistency in predicting strength across different independent sleeve components.

Table 6: Results of leave-one-specimen-out cross-validation for the BO-XGBoost model.

Test Set Component	Training Set Component	R^2	MAE	MAPE	MSE	RMSE
Wall Sleeve C	Column Sleeve A + Column Sleeve B	0.9881	0.2037	5.56%	0.2130	0.4615
Column Sleeve A	Column Sleeve B + Wall Sleeve C	0.9873	0.2113	5.96%	0.2266	0.4760
Column Sleeve B	Column Sleeve A + Wall Sleeve C	0.9868	0.2161	6.13%	0.2456	0.4956

Further comparison of the three rounds of validation results reveals that the evaluation metrics fluctuated only slightly across different test components, suggesting that BO-XGBoost is not effective only for a specific sleeve but can capture the relationship between SAFT features and grouting material strength well across the range of the three types of sleeves in this study. This result, to some extent, enhances the reliability of the original hold-out test conclusions and also indicates that the hyperparameter combination obtained via Bayesian optimization possesses good cross-component adaptability under the current experimental conditions.

3.8 Analysis of Feature Importance and Partial Dependence

In the process of optimizing machine learning models to characterize the strength development state of grouting materials as a function of curing age, although the models demonstrate strong predictive performance, explaining and elucidating the model prediction process is equally crucial. Traditional machine learning models are often regarded as "black boxes" lacking interpretability. Therefore, introducing model interpretation methods not only helps guide model optimization and decision-making strategies but also enhances the trust of model users in the trained model. To this end, this paper employs two methods—feature importance analysis and partial dependence analysis—to achieve a systematic explanation of the model's prediction mechanism.

Feature importance analysis is one of the most commonly used methods for model interpretation. It quantifies the contribution of each feature to the model's prediction results, thereby ranking the key influencing factors. Specifically, the greater the influence of a given feature on the prediction outcome, the higher its corresponding importance score. Fig. 11 presents the relative feature importance results of the BO-XGBoost model for the task of predicting curing age. The analysis shows that the SAFT feature value has the highest relative importance, with a normalized importance of 100%. This is followed by the Red feature, with an importance approximately 35.5% of that of the SAFT feature value. The Purple feature ranks third, with an importance of approximately 22.6%. The Blue feature ranks fourth, with an importance of approximately 12.2%. The Orange feature ranks fifth, with an importance of approximately 9.9%. The Green feature ranks sixth, with an importance of approximately 2.6%. The Yellow feature ranks seventh, with an importance of approximately 1.5%. The results indicate that the SAFT feature value and the area ratios of the various color regions contribute to varying degrees to the identification of curing age, with the SAFT feature value contributing the most, suggesting that the overall enhancement of the ultrasonic response within the material is a key basis for distinguishing curing stages.

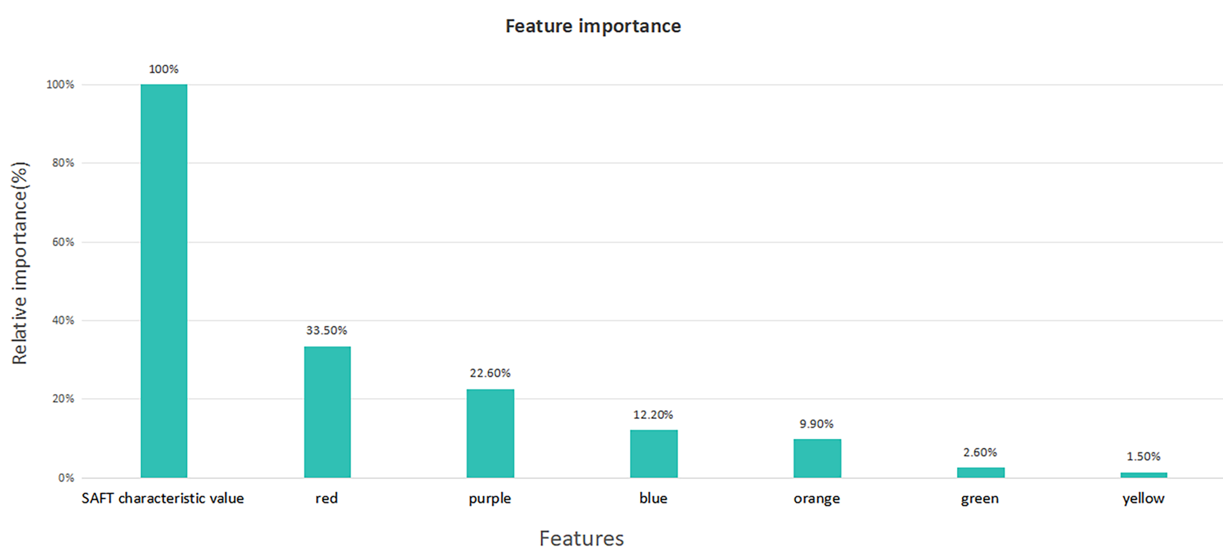

Figure 11: Feature importance.

However, feature importance analysis only identifies the degree of importance of features and cannot reveal how they specifically influence the prediction outcome. To further explore the intrinsic relationship between features and the output, this paper further employs partial dependence analysis. This method reflects the trend of the model output's response to changes in a given feature, but the results should be understood as the average response learned by the model under the current data conditions, rather than a strict causal relationship between the variable and the target. Changes in the partial dependence values intuitively reflect variations in the model output. Fig. 12 shows partial dependence plots for several input features, where the shaded areas represent confidence intervals. The results reveal that the SAFT feature value and the Orange feature generally exhibit a distinctly positive association, while the Purple and Blue features show an overall negative association, which is largely consistent with the general evolution trend of gradual shrinkage of low-strength regions and expansion of high-strength regions. In contrast, the Green and Yellow features did not exhibit simple monotonic relationships: the Green feature showed a slight decrease before increasing, while the Yellow feature reached a local peak at a value around 4–5, then decreased, with some recovery at higher intervals. This indicates that the influence of the area ratio of the medium-strength regions on the prediction output is characterized by staging and non-linearity. This phenomenon suggests that simply interpreting the Green and Yellow features as "larger values correspond to stronger characteristics during the mid-hydration stage and thus higher strength" would not fully capture the model's response. A more plausible interpretation is that these two types of intermediate-strength regions may correspond to transitional states in the process of shifting from low-value to high-value regions, and their area ratios do not necessarily change monotonically across different curing stages. Concurrently, the possibility that the model captures local data patterns under limited sample conditions cannot be ruled out. Therefore, this paper defines the physical meaning of the Green and Yellow features as empirical characterizations related to mid-stage hydration evolution, rather than as well-established monotonic mechanisms. This result suggests that subsequent research still needs to combine larger sample sizes, repeated experiments, and coupled acoustic-hydration analysis to further validate the true physical meaning and robustness of these intermediate color region features.

3.9 Relationship with Traditional Strength Assessment Methods and Discussion on Applicability

In engineering practice, the early strength development of grouting materials or cement-based materials is often evaluated using methods such as the maturity method, the rebound hammer method, and localized destructive/semi-destructive tests (e.g., pull-out tests). Each of these methods has its own scope of application and can also serve as an important reference benchmark for the non-destructive evaluation framework proposed in this paper.

Among them, the maturity method, which is based on the empirical relationship between the temperature–time history and strength development, offers the advantages of simple implementation and suitability for continuous monitoring during the early age. For the issue of grouting material age development addressed in this study, the maturity method is conceptually relevant, as the strength gain of grouting materials is also governed by the hydration process, which is closely related to the temperature history. However, the maturity method typically requires pre-establishing a maturity–strength calibration relationship under the same material system and curing conditions. Its prediction results are sensitive to factors such as material mix proportion, ambient temperature, component dimensions, and curing regime. For grouted sleeve connections, which are constrained by steel sleeves, characterized by complex local geometries, and whose internal state is not easily observed directly, relying solely on temperature history often fails to capture the effects of local defects, interface conditions, and spatial heterogeneity on the ultrasonic response and actual service state.

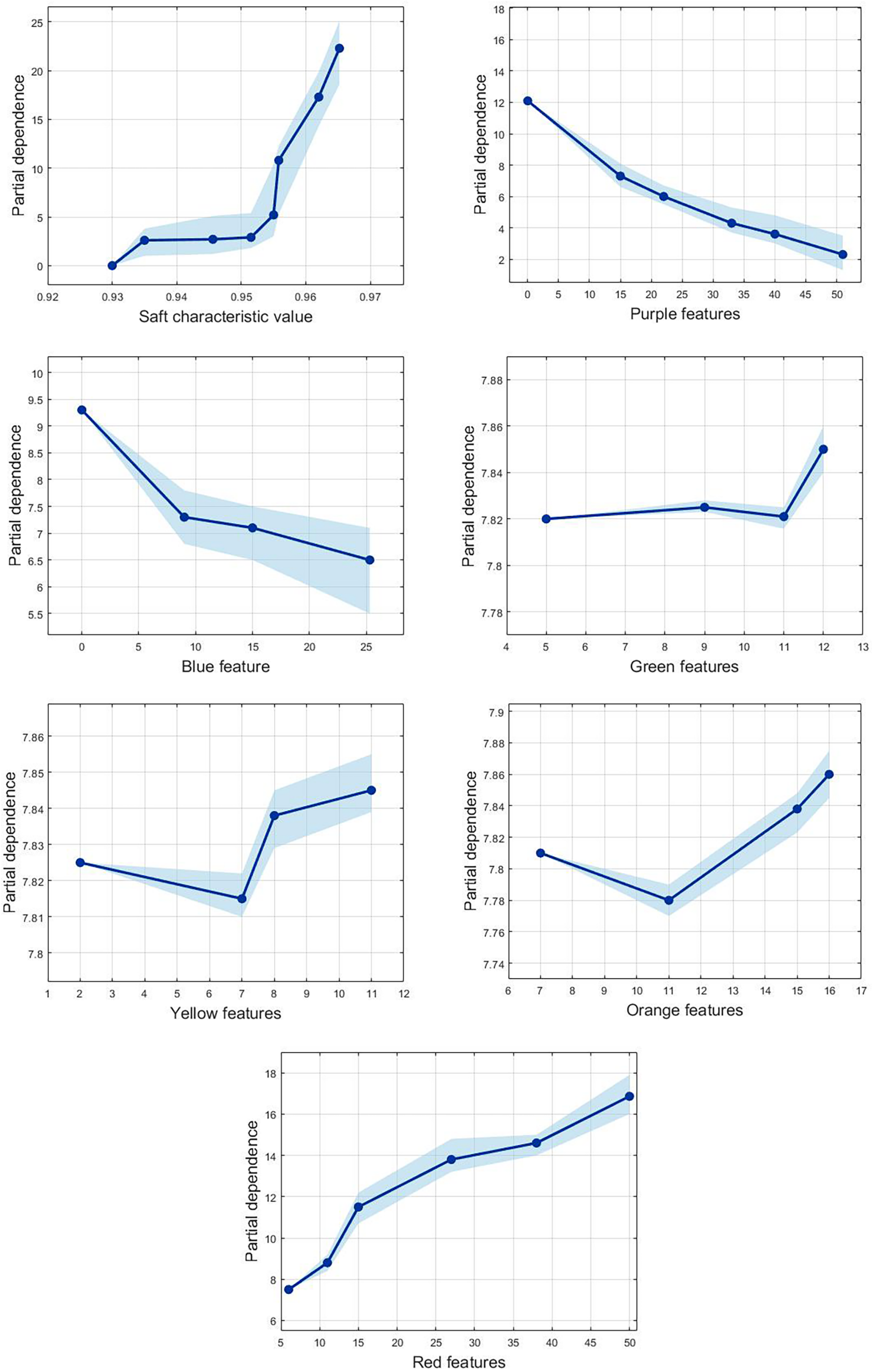

Figure 12: Partial dependence plots for seven input variables.

In contrast, the method proposed in this paper is characterized by its direct use of SAFT image features to represent the spatial acoustic response evolution within the grouted sleeve region. Therefore, in addition to information related to curing age, it also has the potential to reflect local densification, defect distribution, and heterogeneity changes. Nevertheless, it should be noted that this study has not yet simultaneously established comparative data from methods such as the maturity method, rebound hammer method, or pull-out tests. Consequently, it cannot yet quantitatively demonstrate the accuracy advantage or engineering superiority of the proposed method over traditional approaches. Based on this, this paper positions the proposed method as a preliminary non-destructive characterization framework for assessing the strength development state of grouting materials in sleeves. The complementary relationship between this method and traditional approaches, as well as their comparative performance, still requires further validation through synchronized experiments in subsequent research. It should be emphasized that this study did not conduct a direct experimental comparison with conventional strength assessment methods, such as the maturity method, rebound hammer method, or pull-out test; therefore, no quantitative conclusion can be drawn regarding the relative accuracy or superiority of the proposed SAFT–machine learning framework compared with these methods.

It should also be noted that the identification of the sleeve center based on the region with the highest local reflection intensity was implemented with reference to the known sleeve depth and the visual characteristics of the SAFT image. In the present study, this procedure was not fully automated by a predefined intensity threshold, fixed search-window size, or tie-breaking rule. Therefore, the localization of the sleeve region may involve a certain degree of operator dependence. Future work should develop a more standardized and automated localization procedure, such as threshold-based segmentation, template matching, or intensity-weighted centroid extraction, to further improve the reproducibility of SAFT feature extraction.

4 Conclusion

This study systematically investigates the age-related prediction method for grouting materials based on ultrasonic testing and machine learning approaches. The main conclusions are as follows:

1. Within the limited sample dataset constructed in this study, hyperparameter optimization improved the error performance of the XGBoost model on the hold-out data, with BO-XGBoost achieving relatively optimal prediction results.

2. Feature importance and partial dependence analyses indicate that the SAFT feature value and the distribution of color regions can effectively characterize the internal structural evolution of grouting materials during the curing process and contribute significantly to the prediction of curing age.

3. By leveraging the consistent relationship between curing age and compressive strength development, this study demonstrates the feasibility of using "SAFT features + machine learning" for the non-destructive characterization of the strength development state of grouting materials. However, since the data originate from repeated spatiotemporal observations of only three sleeve specimens, the current results remain a preliminary validation. Furthermore, future work should incorporate acoustic numerical simulations or sensitivity analyses to systematically evaluate the impact of variations in sleeve geometry and boundary conditions on the comparability of SAFT features. It should also be noted that the age–strength calibration relationship established in this study is specific to the C85 grouting material system and the laboratory curing regime adopted herein; recalibration is required before applying the proposed method under field conditions.

4. The leave-one-specimen-out cross-validation results provide a preliminary assessment of cross-specimen consistency among the three available sleeve specimens. Although the R^2 values in the three LOSO rounds differed only slightly, this result should not be interpreted as evidence of robust transferability to broader engineering scenarios. In particular, the present LOSO design did not test transferability across

different material batches, operators, grouting procedures, curing regimes, or field environmental conditions. Future studies should include more independent specimens and more diverse experimental conditions to further evaluate the engineering generalizability of the proposed method.

Acknowledgement: This research was supported by the Key Technology Research and Development Project for Self-Repairing and Intelligent Monitoring Prefabricated Products for Intelligent Transportation (Grant No. GY2023028). The authors would like to express their sincere gratitude to the research team for their contributions to this project. No other non-author contributors are identified.

Funding Statement: This work was supported by the Key Technology Research and Development Project for Self-Repairing and Intelligent Monitoring Prefabricated Products for Intelligent Transportation (Grant No. GY2023028), awarded to Pengcheng Xia, and Ruoyu Chen.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Pengcheng Xia, Ruoyu Chen; data collection: Pengcheng Xia, Zhihong Pan, Ruoyu Chen; analysis and interpretation of results: Pengcheng Xia, Zhihong Pan, Ruoyu Chen, Linyuan Wang; draft manuscript preparation: Pengcheng Xia; critical revision: Zhihong Pan, Ruoyu Chen. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The authors declare that the data supporting the findings of this study are available within the paper. Should any raw data files be needed in another format they are available from the corresponding author upon reasonable request.

Ethics Approval: This study did not involve human participants, human tissue, or animal subjects. The research focused solely on ultrasonic testing of grouting materials and precast concrete components. Therefore, ethical approval was not required.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Luo X, Ma F, Deng C, Zheng X, Liao C. Research on the design method of 3D parts library of prefabricated concrete composite wall-slab system based on BIM. Buildings. 2025;15(16):2897. doi:10.3390/buildings15162897.
2. Bi W. Research on the way of prefabricated building information sharing based on computer software BIM. J Phys Conf Ser. 2021;1744(2):022078. doi:10.1088/1742-6596/1744/2/022078.
3. Song X, Li Z, Liu Y. Research on strategies for enhancing system security resilience of prefabricated building engineering. Sci Rep. 2025;15(1):32309. doi:10.1038/s41598-025-18161-0.
4. Li Q, Yi M, Zhang B. Mechanical properties of grouted sleeve splice with horizontal grouting defect after exposure to high temperature. Structures. 2025;79:109443. doi:10.1016/j.istruc.2025.109443.
5. Nie X, Long L, Fu Q, Xiao Q, Xu Q, Chen G, et al. Experimental study on the effect of supplementary grouting on the mechanical properties of grout-deficient sleeve connections. Structures. 2025;77:108986. doi:10.1016/j.istruc.2025.108986.
6. Lu JX, Wu H, Cheng YH, Chen GQ. Blast resistance of grouting sleeve connected precast concrete columns under close-in explosions. Int J Impact Eng. 2024;187(6):104908. doi:10.1016/j.ijimpeng.2024.104908.
7. Yang C, Chen H, Wang X, Wang Q, Zhang B, Ou J. Pull-out performance of grouting sleeve connection with short studs shear keys in modular steel building. J Build Eng. 2024;84:108553. doi:10.1016/j.jobe.2024.108553.
8. Gu S, Yang J, Shen S, Li X. Investigation on a novel reinforcement method of grouting sleeve connection considering the absence of reserved reinforcing bars in the transition layer. Materials. 2024;17(23):5961. doi:10.3390/ma17235961.
9. Liu X, Qu H, Liu B, Zhou Y, Li J, Deng W, et al. Research on rheological behavior and strength characteristics of cement-based grouting materials. Buildings. 2025;15(11):1796. doi:10.3390/buildings15111796.

10. Luo X, Lu Q, Zhang B. Experimental study on modification of grouting material for joints of prefabricated buildings. Int J Concr Struct Mater. 2024;18(1):12. doi:10.1186/s40069-023-00649-6.
11. Gao F, Wang C, Tang Y, Zhang P. Sleeve grouting process risk forecasting based on system dynamics for industrialized building systems. ASCE-ASME J Risk Uncertainty Eng Syst Part A Civ Eng. 2025;11(4):04025064. doi:10.1061/ajrua6.rueng-1384.
12. Zhi Q, Yuan Z, Zheng Y, Jia L, Guo Z. Experimental research on seismic performance of precast concrete shear walls with a novel grouted sleeve used in the connection. J Earthq Eng. 2024;28(5):1379–403. doi:10.1080/13632469.2023.2233030.
13. Xie K, Huang K, Liang D, Huang L, Wang Q. Axial compressive damage and failure early warning method of concrete-filled steel tube with defects based on acoustic emission. Structures. 2026;86:111411. doi:10.1016/j.istruc.2026.111411.
14. Li Z, Sugita A, Abe M, Hasegawa S. Non-destructive characterization of grouted tendon ducts in bridge structures via high-energy X-ray imaging. J Civ Struct Health Monit. 2025;15(8):3187–203. doi:10.1007/s13349-025-00973-3.
15. Cao D, Pan Z, Zhang Z, Zeng B. Defect detection of grouting sleeve connection with energy ratio change in frequency domain. Appl Acoust. 2024;221(1):110037. doi:10.1016/j.apacoust.2024.110037.
16. Mohammed AA, Rafiq SK, Hamid NA. The assessment of concrete subjected to preloading using non destructive testing methods. Case Stud Constr Mater. 2021;15(9):e00705. doi:10.1016/j.cscm.2021.e00705.
17. Saleh E, Tarawneh A, Dwairi H, AlHamaydeh M. Guide to non-destructive concrete strength assessment: homogeneity tests and sampling plans. J Build Eng. 2022;49(6):104047. doi:10.1016/j.jobe.2022.104047.
18. Liu J, Li D, Cui X. Research status and future directions of defect detection in grouted splice sleeves: a review. Constr Build Mater. 2023;402(22):133010. doi:10.1016/j.conbuildmat.2023.133010.
19. Cao D, Pan Z, Zhang Z, Zeng B. Study on non-destructive testing method of grouting sleeve compactness with wavelet packet energy ratio change. Constr Build Mater. 2023;389(9):131767. doi:10.1016/j.conbuildmat.2023.131767.
20. Li D, Liu H. Detection of sleeve grouting connection defects in fabricated structural joints based on ultrasonic guided waves. Smart Mater Struct. 2019;28(8):085033. doi:10.1088/1361-665X/ab29b0.
21. Liu Y, Xu N, Zhao S, Feng A, Yang H, Kang Y, et al. Proposal and experimental verification of enhancement method for half-grouted sleeve connection quality. Constr Build Mater. 2024;416(4):135161. doi:10.1016/j.conbuildmat.2024.135161.
22. Somdutta, Rai B. Hybrid machine learning models for predicting compressive strength of self-compacting concrete: an integration of ANFIS and Metaheuristic algorithm. Nondestruct Test Eval. 2026;41(4):1789–821. doi:10.1080/10589759.2025.2483375.
23. Shishegaran A, Varaee H, Rabczuk T, Shishegaran G. High correlated variables creator machine: prediction of the compressive strength of concrete. Comput Struct. 2021;247:106479. doi:10.1016/j.compstruc.2021.106479.
24. Günaydın O, Akbaş E, Özbeyaz A, Güçlüer K. Machine learning based evaluation of concrete strength from saturated to dry by non-destructive methods. J Build Eng. 2023;76:107174. doi:10.1016/j.jobe.2023.107174.
25. Gan X, Wang W, Jiang C, Ye L, Chen F, Zhou T, et al. Ultrasonic detection and deep learning for high-precision concrete strength prediction. J Build Eng. 2025;104(19):112372. doi:10.1016/j.jobe.2025.112372.
26. Albuthbahak MO, Hiswa RMAA. Prediction of concrete compressive strength using supervised machine learning models through ultrasonic pulse velocity and mix parameters. Rom J Mater. 2019;49(2):232–43. doi:10.1016/j.autcon.2005.07.003.
27. Asteris PG, Skentou AD, Bardhan A, Samui P, Lourenço PB. Soft computing techniques for the prediction of concrete compressive strength using Non-Destructive tests. Constr Build Mater. 2021;303(1):124450. doi:10.1016/j.conbuildmat.2021.124450.
28. Dinh K, Tran K, Brown J, Ferraro CC, Parida S. Ray-based SAFT for imaging concrete structures from ultrasonic data. In: SPIE Smart Structures + Nondestructive Evaluation. Proceeding of the SPIE 12487, Nondestructive Characterization and Monitoring of Advanced Materials, Aerospace, Civil Infrastructure, and Transportation XVII. Long Beach, CA, USA; 2023. p. 50–63. doi:10.1117/12.2659061.
29. Yun D, Kim J, Jhang KY. Imaging of contact acoustic nonlinearity using synthetic aperture technique. Ultrasonics. 2013;53(7):1349–54. doi:10.1016/j.ultras.2013.04.002.

30. Wan S, Li S, Chen Z, Tang Y. An ultrasonic-AI hybrid approach for predicting void defects in concrete-filled steel tubes via enhanced XGBoost with Bayesian optimization. Case Stud Constr Mater. 2025;22(6):e04359. doi:10.1016/j.cscm.2025.e04359.
31. Jodeiri Shokri B, Mirzaghorbanali A, McDougall K, Karunasena W, Nourizadeh H, Entezam S, et al. Data-driven optimised XGBoost for predicting the performance of axial load bearing capacity of fully cementitious grouted rock bolting systems. Appl Sci. 2024;14(21):9925. doi:10.3390/app14219925.
32. Jiang J, Ma T, Chen X, Wu S, Chen G. Application of different machine learning algorithms in post-earthquake failure probability assessment of underground structures. Soil Dyn Earthq Eng. 2026;200:109823. doi:10.1016/j.soildyn.2025.109823.
33. Abdi K, Kebaili N, Djouhri M. Bayesian networks modelling for predicting compressive strength and slump flow of self-compacting concrete. Asian J Civ Eng. 2024;25(3):2567–78. doi:10.1007/s42107-023-00928-3.
34. Xie D, Wang L, Li S, Khodaparast HH, Friswell MI, Lu ZR. Efficient two-stage modal identification for structures with closely spaced modes by Bayesian FFT and joint approximate diagonalization. J Sound Vib. 2025;618(4):119365. doi:10.1016/j.jsv.2025.119365.
35. Ma Y, Li J, Wang D, Mi J, Dai K. Probabilistic prediction of hysteretic curves of corroded reinforced concrete columns based on Bayesian theory. Soil Dyn Earthq Eng. 2025;193(4):109308. doi:10.1016/j.soildyn.2025.109308.
36. Jia H, Liu Y, Jia S, Liu Q. An improved adaptive particle swarm optimization algorithm with interactions between particles for path planning of underwater vehicles. Trans Inst Meas Control. 2026;48(2):395–408. doi:10.1177/01423312241287257.
37. Xia P, Pan Z, Chen R. Research on imaging of precast concrete structure grouting sleeve based on noise suppression SAFT filtering algorithm. Res Nondestruct Eval. 2026;37(1):17–42. doi:10.1080/09349847.2026.2614555.