

Article

Predicting Carbonation Depth of Recycled Aggregate Concrete Using Optuna-Optimized Explainable Machine Learning

Yuxin Chen ¹, Xiaoyuan Li ^{1,*}, Enming Li ^{2,*} and Jian Zhou ¹ 

¹ School of Resources and Safety Engineering, Central South University, Changsha 410083, China; 245501013@csu.edu.cn (Y.C.); j.zhou@csu.edu.cn (J.Z.)

² ETSI Minas y Energía, Universidad Politécnica de Madrid, Ríos Rosas 21, 28003 Madrid, Spain

* Correspondence: 225502046@csu.edu.cn (X.L.); enming.li@alumnos.upm.es (E.L.)

Abstract

Accurately predicting the carbonation depth of recycled aggregate (RA) concrete is essential for durability assessment. Based on a dataset of 682 experimental samples, this study employed seven machine learning algorithms to develop prediction models for the carbonation depth of RA concrete. The Optuna framework was utilized to conduct 500 trials of hyperparameter optimization for these models, with the objective of minimizing the 5-fold cross-validated mean squared error. Results indicate that model performance improved significantly after optimization. Among them, the XGBoost model achieved the best performance, with a coefficient of determination (R^2) of 0.9789, root mean squared error (RMSE) of 1.0811, mean absolute error (MAE) of 0.6972, mean absolute percentage error (MAPE) of 8.7932%, variance accounted for (VAF) of 97.8966%, and mean bias error (MBE) of 0.0641 on the test set. Explainability analysis using SHapley Additive exPlanations (SHAP) further revealed that exposure time is the most significant factor influencing the carbonation depth prediction. Additionally, considering that the database incorporates both natural and accelerated carbonation conditions, the samples were partitioned based on CO_2 concentration and conducts a stratified performance evaluation. The results demonstrate that the model maintains high predictive accuracy under natural carbonation as well as across different accelerated carbonation intervals, indicating that, within the scope covered by the current dataset, the proposed approach provides a highly accurate and interpretable tool for predicting the carbonation depth of recycled aggregate concrete.

Keywords: recycled aggregate concrete; carbonation depth; machine learning; explainability; hyperparameter optimization



Academic Editor: Flavio Stochino

Received: 4 December 2025

Revised: 9 January 2026

Accepted: 10 January 2026

Published: 14 January 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Concrete is the most widely used construction material globally. While its production and use generate substantial economic and social benefits, they are also accompanied by a series of environmental issues, such as the depletion of natural aggregates, accumulation of construction and demolition waste, and greenhouse gas emissions [1–4]. Processing construction demolition waste into recycled aggregates and using them in recycled aggregate (RA) concrete has become one of the important approaches to realizing resource recycling and advancing the carbon-neutrality goals of the construction industry [5,6]. However, the old mortar adhered to the RA surface results in a more complex pore structure and significantly higher water absorption [7–9]. This not only affects the mechanical properties of the concrete but also accelerates its carbonation process, thereby posing a serious threat to

the structural durability. Concrete carbonation is a chemical process wherein atmospheric CO_2 diffuses through the pore structure of concrete and reacts with cement hydration products (primarily calcium hydroxide), forming calcium carbonate and water [10]. This reaction continuously consumes alkaline species in the pore solution, leading to a gradual decrease in pH (i.e., neutralization). When the carbonation front reaches the surface of the reinforcing steel and the local pH falls below approximately 9, the passive film on the steel is destroyed, and corrosion is initiated under the combined action of oxygen and moisture [11–13]. This potential risk greatly restricts the large-scale application of RA concrete in structures intended for long-term service. Consequently, accurately predicting the carbonation depth of RA concrete has become a critical issue for assessing its long-term performance and service life, as well as for ensuring its safe and reliable use in engineering practice.

Traditionally, the prediction of carbonation depth has primarily relied on theoretical models based on Fick's first law of diffusion or empirical formulas derived from limited experimental data [14–19]. Although these methods possess a certain physical basis, they typically require long-term and costly experimental support and struggle to effectively capture the complex coupling effects among multiple factors, such as the water-to-binder ratio, aggregate characteristics, ambient CO_2 concentration, and exposure time [20,21]. This is particularly true for RA concrete, whose carbonation process is significantly influenced by parameters including the water absorption of RA, replacement ratio, paste properties, and admixture dosage [22,23]. These factors markedly alter the material's micro-pore structure and carbonation reaction path, further undermining the applicability and predictive accuracy of traditional models [24]. With the continuous accumulation of available carbonation test data, constructing data-driven models for high-accuracy prediction of carbonation depth has become a vital research direction for enhancing the quantitative durability assessment of RA concrete [25–28].

In recent years, machine learning has emerged as a pivotal tool for data-driven research on concrete performance. Methodological frameworks centered on ensemble learning, automated optimization, and interpretability analysis have demonstrated considerable versatility and effectiveness in addressing complex challenges such as predicting the mechanical properties of preplaced aggregate concrete [29]. Within the specific domain of predicting the carbonation depth of RA concrete, related studies have also demonstrated considerable potential [30–35]. For instance, Biswas et al. [20] used a dataset of 300 samples and combined support vector regression (SVR) with four metaheuristic algorithms, including chicken swarm optimization (CSO), to develop hybrid models for predicting the carbonation depth of fly-ash concrete. Among these models, the CSO–SVR model performed best. Núñez and Nehdi [25] employed a gradient boosting regression tree (GBRT) model to predict the carbonation depth of RA concrete incorporating mineral admixtures and demonstrated that this model achieved higher accuracy than several analytical formulas based on Fick's law. Liu et al. [26] proposed a hybrid artificial neural network model optimized by the whale optimization algorithm (WOANN) to predict the carbonation depth of RA concrete, and its accuracy was significantly superior to that of traditional code-based models and standalone machine learning models. Moghaddas et al. [36] adopted artificial bee colony-based expression programming (ABCEP), a novel automated regression technique, to derive an explicit predictive formula for the carbonation depth of RA concrete, which showed superior performance to conventional models on a relatively large dataset. Xi et al. [37] developed a hybrid machine learning model for predicting the carbonation depth of RA concrete by integrating extreme gradient boosting (XGBoost) with the multi-verse optimizer (MVO). The results indicated that the model achieved high accuracy, with an R^2 of 0.9398 on the test set.

Although existing studies have confirmed the potential of machine learning in this area, a systematic research framework has yet to be established. Specifically, current work still leaves room for further exploration in three aspects: (i) the breadth of model selection and the systematic nature of comparative evaluation, (ii) the automation and consistency of hyperparameter optimization, and (iii) the physically meaningful interpretability of model decision mechanisms. Compared with advanced frameworks that have been successfully applied to predicting other concrete properties, research on RA concrete carbonation—a critical durability issue governed by complex multi-factor coupling—still lacks an integrated framework that combines systematic model screening, automated hyperparameter tuning, and machine learning interpretability.

To address this gap, this study sought to establish an interpretable machine learning framework driven by Optuna for hyperparameter optimization and systematically applied it to predict the carbonation depth of RA concrete. Specifically, seven representative types of models—k-nearest neighbors regression (KNN), SVR, random forest (RF), extremely randomized trees (ET), XGBoost, light gradient boosting machine (LGBM), and categorical boosting (CatBoost)—were selected for development. The Optuna framework was employed to perform 500 trials of automated hyperparameter optimization for these models, using the mean squared error (MSE) from 5-fold cross-validation as the unified objective function. Subsequently, the optimal model was selected through a multi-metric performance comparison, and its prediction mechanisms and feature contributions were interpreted in depth at both global and local levels using the SHapley Additive exPlanations (SHAP) method. Through this systematic work, the present study aimed not only to deliver a high-accuracy predictive tool, but also to offer a reliable, transparent, and transferable data-driven methodology for the durability design and assessment of RA concrete.

2. Materials

The data used to develop the carbonation depth prediction model for RA concrete in this study were sourced from references [36,37]. To ensure data integrity for modeling, the original dataset underwent rigorous cleaning and preprocessing, which involved removing samples with any missing or unreported values in the input features or the target variable (carbonation depth). This process resulted in a final set of 682 complete and valid experimental samples that form the foundation for this study. These data were sourced from 21 published studies and cover 11 countries, including China, Portugal, and the United Kingdom (Figure 1), demonstrating good representativeness and diversity in terms of geographical distribution and experimental conditions. Notably, more than 80% of the samples incorporate RA at varying proportions, indicating that the model was trained and validated predominantly on RA-concrete cases that are directly relevant to the research objective. The input features of the dataset include RA water absorption (RAWA, %), water-to-binder ratio (WBR), fine aggregate content (FAC, kg/m³), gravel content (GC, kg/m³), RA content (RAC, kg/m³), superplasticizer (SP, kg/m³), carbon concentration (CC, %), and exposure time (T, d). The output variable is the carbonation depth (CD, mm). Regarding feature selection, it should be noted that although the original studies may have reported additional variables (e.g., relative humidity and cement mineralogical composition), substantial discrepancies exist across studies in terms of reporting formats and data completeness for these parameters. To enable the development of a CD prediction model for RA concrete on a relatively large sample set, this study ultimately adopted the above eight engineering parameters, which are more commonly reported in the literature and are defined in a clearer and more consistent manner. Moreover, this feature set emphasizes practical applicability, relying primarily on information that can be determined at the mix-design stage or is relatively easy to obtain in engineering practice. The statistical descriptions of all variables

are provided in Table 1. Notably, T ranges from 7 to 3650 days, covering both short-term accelerated carbonation and long-term natural carbonation processes. Detailed data are provided in the Supplementary Materials.

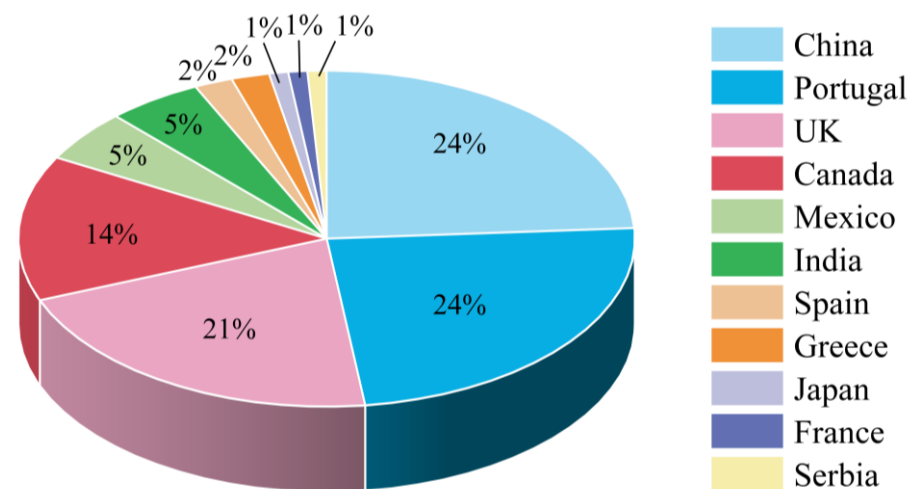


Figure 1. Percentage of data contributed by different countries.

Table 1. Descriptive statistics of the variables used for model development.

Feature	Mean *	Min *	25% *	50% *	75% *	Max *	STD *
RAWA	5.79	0.20	4.70	5.30	6.27	16.58	2.39
WBR	0.51	0.25	0.45	0.50	0.55	1.00	0.11
FAC	645.59	357.66	550.00	625.00	787.00	998.00	170.90
GC	448.52	0.00	0.00	454.93	846.45	1311.00	435.71
RAC	586.69	0.00	198.16	635.00	953.00	1280.00	406.73
SP	0.89	0.00	0.00	0.00	0.73	7.31	1.82
CC	5.31	0.05	3.00	3.50	5.00	50.00	6.48
T	176.98	7.00	28.00	56.00	91.00	3650.00	538.77
CD	10.19	0.10	4.80	8.34	13.30	50.05	7.89

Note. * Mean, Min, 25%, 50%, 75%, Max, and STD denote the arithmetic mean, minimum, 25th percentile (Q1), median (50th percentile, Q2), 75th percentile (Q3), maximum, and standard deviation of each variable, respectively.

To facilitate model development and validation, the entire dataset was randomly split into a training set and a test set at a ratio of 4:1. Figure 2 presents the results of the correlation analysis among all variables. The lower-left triangular panels show the scatter distributions of the training and test sets together with their linear fits; the diagonal panels depict the probability density distributions of each variable; and the upper-right triangular panels report the Pearson correlation coefficients between pairs of features. The analysis indicated that, apart from a relatively strong negative correlation between GC and RAC, the correlations among the other input features were generally weak. The linear correlations between the input features and CD were also generally low, suggesting that carbonation behavior was strongly affected by the coupled influence of multiple factors and exhibited a complex nonlinear relationship. It should be emphasized that GC and RAC represent the contents of natural coarse aggregate and recycled coarse aggregate, respectively. Both variables have clear and independent physical meanings and may exert differentiated effects on carbonation resistance through mechanisms such as pore structure and the interfacial transition zone. Therefore, although GC and RAC exhibit a strong statistical correlation, they were both retained as model inputs in this study to avoid a priori information loss during preprocessing and to preserve essential information for

subsequent mechanistic interpretation and feature-contribution analysis. In addition, the feature distributions of the training and test sets are highly consistent, indicating that the random split was reasonable and conducive to a reliable evaluation of the model's subsequent generalization performance.

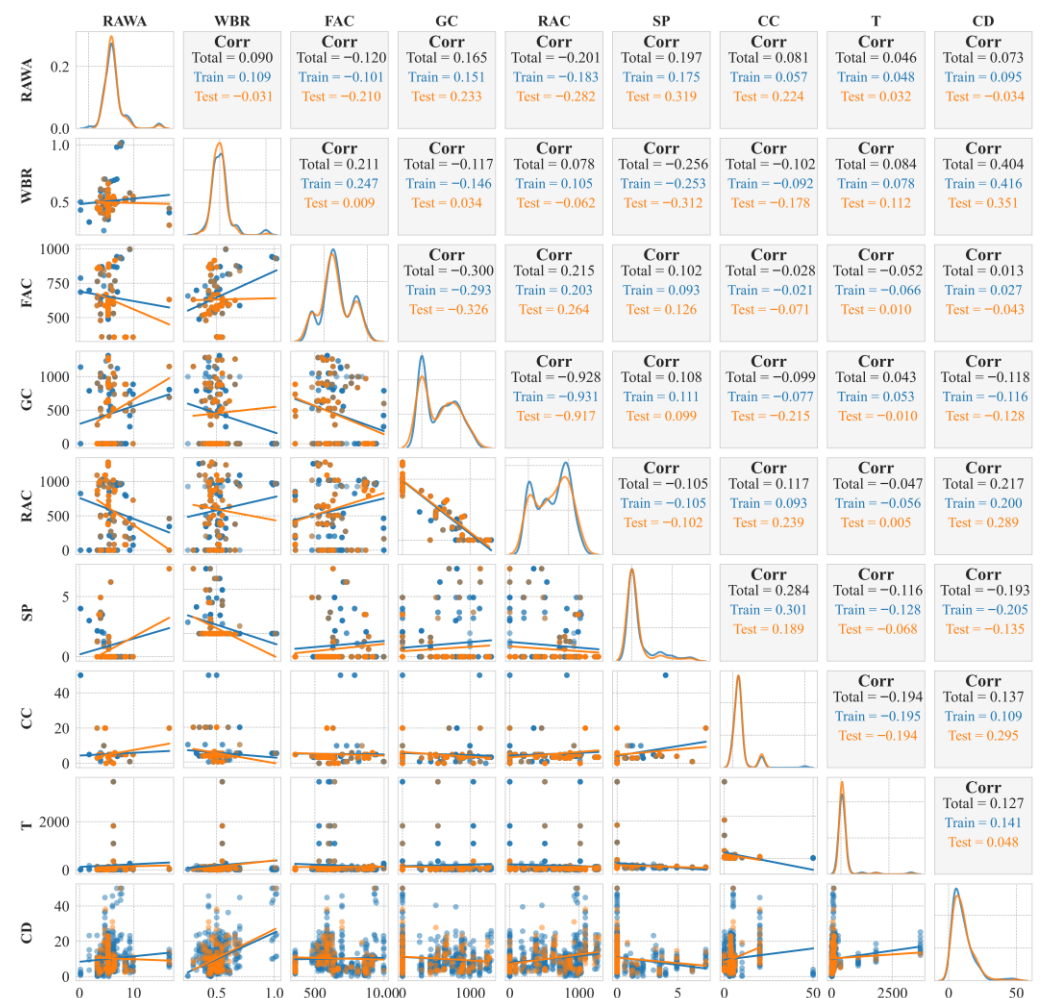


Figure 2. Feature correlation analysis.

3. Methods

3.1. Machine Learning Models

3.1.1. Instance-Based Lazy Learning

K-nearest neighbors (KNN) [38] regression is an instance-based, non-parametric method belonging to the family of lazy learning algorithms. Its guiding principle is that “similar inputs yield similar outputs”. During prediction, the algorithm first calculates the distance between the test sample and all training samples. It then selects the K nearest neighbors, and finally obtains the prediction through a weighted average of the target values of these neighbors.

3.1.2. Kernel-Based Models

Support vector regression (SVR) [39] is the application of support vector machines to regression problems. The core idea of SVR is to find an optimal hyperplane such that the majority of the data points lie within a margin band (called the ϵ -insensitive band) centered around this hyperplane. It focuses only on those outliers that fall outside the

margin band and the few critical samples that define the margin boundaries—namely, the support vectors. The fundamental objective function of SVR is formulated as follows:

$$\begin{aligned} \min_{w, b, \xi_i, \xi_i^*} \quad & \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n (\xi_i + \xi_i^*) \\ \text{subject to} \quad & \begin{cases} y_i - (w^T \Phi(x_i) + b) \leq \varepsilon + \xi_i, \\ (w^T \Phi(x_i) + b) - y_i \leq \varepsilon + \xi_i^*, \\ \xi_i, \xi_i^* \geq 0, \quad i = 1, \dots, n. \end{cases} \end{aligned} \quad (1)$$

where C is the penalty parameter that balances the flatness of the ε -insensitive tube and the allowed deviations (quantified by the slack variables ξ_i and ξ_i^*); w is the weight vector in the mapped feature space; and $\Phi(x_i)$ denotes the feature mapping from the original input space to a higher-dimensional space, which is implemented via the kernel function.

3.1.3. Tree-Based Ensemble Models: Bagging Methods

Random Forest (RF) regression [40] is an ensemble learning method based on bootstrap aggregation and the random subspace method. Its main procedure is as follows: first, multiple bootstrap samples are repeatedly drawn from the original training set to generate several subsampled datasets; subsequently, a separate CART regression tree is trained independently on each subsample, without pruning (or with only minimal pruning). During node splitting in each tree, the algorithm considers only a randomly selected subset of features and chooses the split point that maximizes the reduction in variance. The final prediction is obtained by averaging the outputs of all decision trees, thereby reducing the high variance associated with a single tree.

Extremely Randomized Trees (ET) [41] is a tree ensemble model that employs a “strong randomization” strategy. Unlike RF, at each node it first randomly selects a subset of features, then generates random candidate split thresholds for each feature, and finally chooses the split that maximizes variance reduction. This method typically does not rely on bootstrap sampling; instead, it uses randomly generated thresholds to substantially reduce the correlation among trees, thereby achieving very high training efficiency. The trade-off is a slightly higher bias for each individual tree; however, ensemble aggregation often reduces the overall model variance, leading to generalization performance on noisy or high-dimensional data that is comparable to, or even better than, that of RF.

3.1.4. Tree-Based Ensembles: Boosting Methods

Extreme gradient boosting (XGBoost) [42] is a high-performance gradient boosting decision tree (GBDT) ensemble algorithm. Its core principle is to construct multiple decision trees sequentially through additive training, where each tree learns to correct the errors of the preceding trees, and the final prediction is obtained by taking a weighted sum of the outputs of all trees. The objective function is given by:

$$\begin{aligned} \text{Obj}(\theta) &= \sum_i L(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \\ \text{where } \Omega(f) &= \gamma T + \frac{1}{2} \lambda \|\omega\|^2 \end{aligned} \quad (2)$$

where $\sum_i L(\hat{y}_i, y_i)$ is the loss function that measures the discrepancy between the predicted values $\hat{y}_i$ and the true values y_i ; $\Omega(f)$ is the regularization term; T is the number of leaf nodes in the tree; ω represents the weights of the leaf nodes; and γ and λ are regularization coefficients that control the tree complexity to prevent overfitting.

In contrast to traditional gradient boosting that uses only first-order derivatives, XGBoost performs a second-order Taylor expansion of the loss function during optimization,

utilizing both the first-order gradients (g_i) and second-order Hessian values (h_i) to guide model convergence more rapidly and accurately.

Light gradient boosting machine (LGBM) is an efficient machine learning algorithm built on GBDT [43]. While retaining first- and second-order gradient-based optimization, it enhances training speed and memory efficiency through four key mechanisms: (i) A histogram-based algorithm that discretizes continuous features into bins, substantially reducing the computational cost of split-point finding. (ii) Gradient-based one-side sampling (GOSS), which preserves samples with large gradients while applying weighted random sampling to small-gradient samples, thereby balancing training efficiency and unbiased estimation; (iii) Exclusive feature bundling (EFB), which effectively mitigates the curse of dimensionality in high-dimensional sparse settings; (iv) A leaf-wise growth strategy with depth constraints, in which the leaf with the largest gain is split at each step, achieving higher accuracy for the same number of splits, but requiring appropriate hyperparameter tuning to control model complexity and prevent overfitting.

Categorical boosting (CatBoost) [44] is a GBDT algorithm specifically optimized for handling categorical features. It employs the ordered target statistics technique to encode categorical variables by introducing random permutations and prior terms to avoid target leakage. Concurrently, it utilizes an ordered boosting mechanism, which computes unbiased gradient estimates based on multiple permutations of the training data, thereby significantly mitigating prediction shift. CatBoost uses symmetric (oblivious) trees as base learners, which not only accelerate prediction but also provide a regularization effect. The model natively supports missing values, sparse data, and categorical features without requiring extensive one-hot encoding. Although CatBoost performs particularly well on tabular data rich in categorical features, it also maintains excellent predictive performance on purely numerical datasets, owing to its ordered boosting mechanism and symmetric tree structure.

3.2. Hyperparameter Optimization Method: Optuna

Optuna is an advanced hyperparameter optimization framework specifically designed for machine learning [45]. It is built around three core concepts: Study, Trial, and Objective. A Study serves as the container for the optimization process, recording the complete trial history as well as the best results; a Trial corresponds to a single run of training and validation under a given hyperparameter configuration θ ; and the Objective function is defined by the user, with the validation performance metric it returns treated as the quantity to be minimized or maximized by the optimizer.

3.2.1. Definition of the Search Space

Hyperparameter optimization first requires an explicit definition of the search space, namely the admissible range and type of each parameter to be tuned. For each machine learning model, this study defined a corresponding search space, the details of which are provided in Table 2. All parameter ranges were implemented in code using Optuna's `trial.suggest_*` functions (e.g., `suggest_int`, `suggest_float`). To reduce numerical discrepancies in reporting and to facilitate reproducibility, all continuous float hyperparameters were stored and presented with values rounded to four decimal places.

Table 2. Definition of the hyperparameter search space.

Model	Hyperparameter	Search Space	Value Type
KNN	n_neighbors	[1, 50]	integer
	weights	uniform or distance	categorical
	p	1 or 2	integer
SVR	C	[0.001, 100]	float
	epsilon	[0.001, 10]	float
	gamma	[0.001, 10]	float
RF and ET	n_estimators	[100, 1000]	integer
	max_depth	[1, 50]	integer
	min_samples_split	[2, 20]	integer
	min_samples_leaf	[1, 20]	integer
XGBoost	n_estimators	[100, 1000]	integer
	max_depth	[1, 50]	integer
	learning_rate	[0.01, 0.5]	float
	subsample	[0.6, 1]	float
	colsample_bytree	[0.6, 1]	float
	reg_alpha	[0, 10]	float
	reg_lambda	[0, 10]	float
	min_child_weight	[1, 10]	integer
LGBM	gamma	[0, 5]	float
	n_estimators	[100, 1000]	integer
	max_depth	[1, 50]	integer
	learning_rate	[0.01, 0.5]	float
	subsample	[0.6, 1]	float
	colsample_bytree	[0.6, 1]	float
	reg_alpha	[0, 10]	float
	reg_lambda	[0, 10]	float
CatBoost	iterations	[10, 400]	integer
	depth	[1, 16]	integer
	learning_rate	[0.01, 0.5]	float
	l2_leaf_reg	[0.1, 10]	float
	border_count	[32, 255]	integer

3.2.2. Sampling Strategy

Optuna by default employs the Tree-structured Parzen Estimator (TPE) for sequential sampling. The core idea of TPE is to model the conditional probability $p(x|y)$ of the hyperparameters x given the objective value y , rather than modeling the objective function directly. Specifically, the algorithm ranks all historical observations by their objective values and uses a quantile threshold γ to split them into a “promising” group, $\ell(x)$, and a “non-promising” group, $g(x)$. It then performs kernel density estimation on the hyperparameter distributions of these two groups separately. When minimizing the objective function, TPE selects the hyperparameter combination that maximizes the expected improvement criterion, which is equivalent to choosing the point that maximizes the likelihood ratio $\ell(x)/g(x)$, as the candidate for the next trial:

$$\theta^* \approx \operatorname{argmax}_{\theta} \frac{\ell(\theta)}{g(\theta)} \quad (3)$$

In practice, Optuna first conducts a set of initial “warm-start trials” using either random sampling or Latin hypercube sampling. It then iteratively updates the density estimates for $\ell(x)$ and $g(x)$ using the accumulated observation data, focusing subsequent sampling on the parameter regions that are most promising for performance gains.

3.2.3. Pruning and Stopping Criterion

To control optimization efficiency, a fixed trial budget was adopted as the primary stopping criterion. Specifically, trials = 500 was set for each model, providing a practical balance between sufficiently exploring the search space and maintaining reasonable computational cost. Because the objective value was defined as the MSE averaged over 5-fold cross-validation, each trial was required to complete all folds to obtain a stable and reliable performance estimate. Therefore, to avoid potential bias introduced by prematurely terminating trials based on intermediate results and to ensure consistent evaluation across different hyperparameter configurations, no dynamic pruning strategy was applied. This setting ensured that all candidate configurations were assessed in a complete and fair manner.

3.3. Interpretability Analysis Method

SHapley Additive exPlanations (SHAP) [46] is theoretically grounded in the Shapley value from cooperative game theory and decomposes the prediction of an individual sample into the sum of the marginal contributions of each feature. In this way, it quantifies the influence of features on the predictions at both the local (single-sample) and global (overall) levels, thereby alleviating the “black-box” nature of the models.

Let the model output be $f(x)$, the complete feature set be F , and the total number of features be $M = |F|$. Select a background dataset B , with the baseline value defined as:

$$\phi_0 = \mathbb{E}_{X \sim B}[f(X)] \quad (4)$$

For any sample x , SHAP assigns a local attribution $\phi_j(x)$ to each feature $j \in F$ such that:

$$f(x) = \phi_0 + \sum_{j=1}^M \phi_j(x) \quad (5)$$

holds true (in regression tasks, each ϕ_j shares the same unit of measurement as the target variable). For the j -th feature of sample x , its SHAP value $\phi_j(x)$ is obtained by computing the expected marginal contribution of that feature across all possible feature subsets, expressed mathematically as:

$$\phi_j(x) = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(M - |S| - 1)!}{M!} (f_{S \cup \{j\}}(x) - f_S(x)) \quad (6)$$

where $f_S(x)$ denotes the expected model output when only the feature subset S is utilized. This computational framework ensures equitable allocation of feature contributions and rigorously satisfies three theoretical properties: local accuracy, missingness, and consistency.

3.4. Model Evaluation Metrics

To comprehensively assess the overall performance of the predictive models, this study employed six statistical metrics for evaluation from different perspectives, including goodness-of-fit, accuracy metrics, and bias analysis. These metrics include the coefficient of determination (R^2), variance accounted for (VAF), root mean squared error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and mean bias error

(MBE) [47–51]. Let the dataset be $\{(x_i, y_i)\}_{i=1}^n$, the model predictions be $\hat{y}_i$, and the sample mean be $\bar{y}$. The metrics are defined as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (7)$$

$$\text{VAF} = \left(1 - \frac{\text{Var}(y - \hat{y})}{\text{Var}(y)}\right) \times 100\% \quad (8)$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (9)$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (10)$$

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100\% \quad (11)$$

$$\text{MBE} = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i) \quad (12)$$

where R^2 ranges from $(-\infty, 1]$, with values closer to 1 indicating a better model fit. VAF ranges from $(-\infty, 100\%]$, and values closer to 100% are preferred. RMSE, MAE, and MAPE all range from $[0, +\infty)$, where smaller values indicate higher predictive accuracy. MBE reflects the direction of systematic bias: positive values indicate overall overestimation, negative values indicate overall underestimation, and ideally it should be close to 0.

3.5. Model Development Workflow

The construction workflow of the proposed prediction models is illustrated in Figure 3, and can be summarized as follows: First, the original dataset was randomly split into a training set and a test set at a ratio of 4:1. Next, systematic hyperparameter optimization was conducted on the training set for seven machine learning models using the Optuna framework. In this stage, the hyperparameter search space for each model was explicitly defined, and the TPE sampler was employed to guide the search. The optimization objective was uniformly set to minimize the MSE averaged over 5-fold cross-validation, and a fixed budget of 500 trials was specified as the stopping criterion for each model to ensure sufficient exploration of the search space. Subsequently, each optimized model was retrained on the full training set using the best hyperparameters identified by Optuna, and its predictive performance was evaluated on both the training set and the independent test set. By comprehensively comparing prediction accuracy and generalization performance on the test set, the model with the best overall performance was selected. Finally, SHAP was applied to interpret the predictive mechanism and feature contributions of the selected model at both the global and local levels.

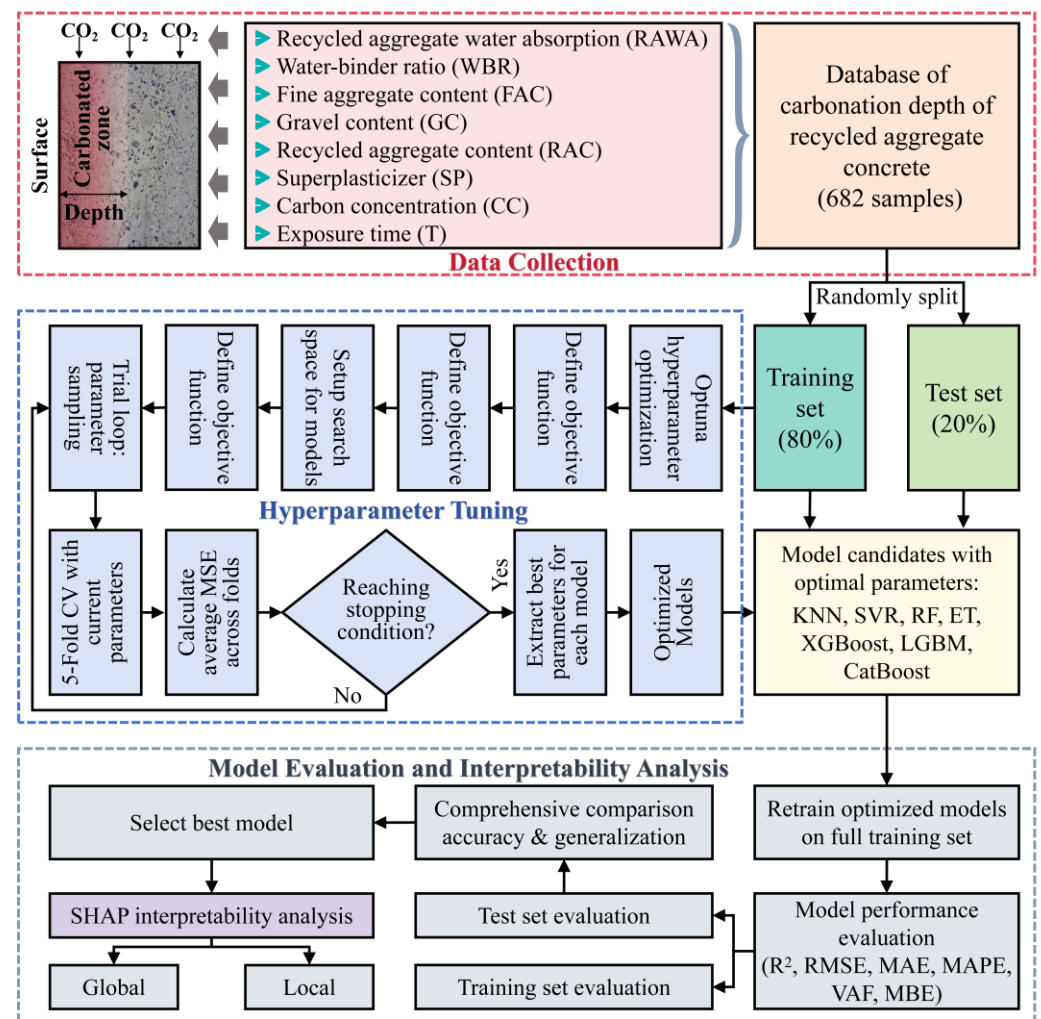


Figure 3. Workflow for developing the CD prediction model.

4. Results

4.1. Hyperparameter Optimization Results

Figure 4 illustrates the convergence process of the objective function values over 500 optimization trials. It was observed that all models achieved rapid decreases in the objective value during the initial phase (first 50 trials), accomplishing most of the performance improvements. After approximately 100 trials, the models entered a phase of diminishing marginal returns, where the curves gradually flattened. In terms of the optimization results, CatBoost ultimately converged to the lowest objective value, followed by XGBoost. The optimal objective values and the corresponding hyperparameter configurations for each model are summarized in Table 3.

After obtaining the optimal hyperparameters, they were applied to the corresponding models, which were then retrained on the full training set. The predictive performance of these models was subsequently evaluated on both the training and test sets. Figure 5 shows a comparison of the evaluation metrics on the test set before and after model optimization. The results indicated that hyperparameter optimization with Optuna substantially improved the predictive performance of the KNN, SVR, XGBoost, LGBM, and CatBoost models. Among them, SVR showed the most notable improvement, with R^2 increased by 177.38%, while RMSE, MAE, and MAPE decreased by approximately 69.67%, 68.95%, and 65.10%, respectively. KNN also exhibited substantial gains. Although the gradient boosting models (XGBoost, LGBM, and CatBoost) already demonstrated high accuracy under

their default parameter settings, their error metrics decreased further after optimization; for instance, XGBoost achieved reductions in RMSE, MAE, and MAPE of about 23.02%, 25.57%, and 24.81%, respectively. In contrast, RF and ET showed only minor changes in metrics before and after optimization. RF even exhibited a slight decline in some metrics, suggesting that for the dataset used in this study, the default hyperparameters of these bagging-based models were already near optimal. Their performance was less sensitive to hyperparameter tuning, and further optimization may introduce additional variance with limited gains.

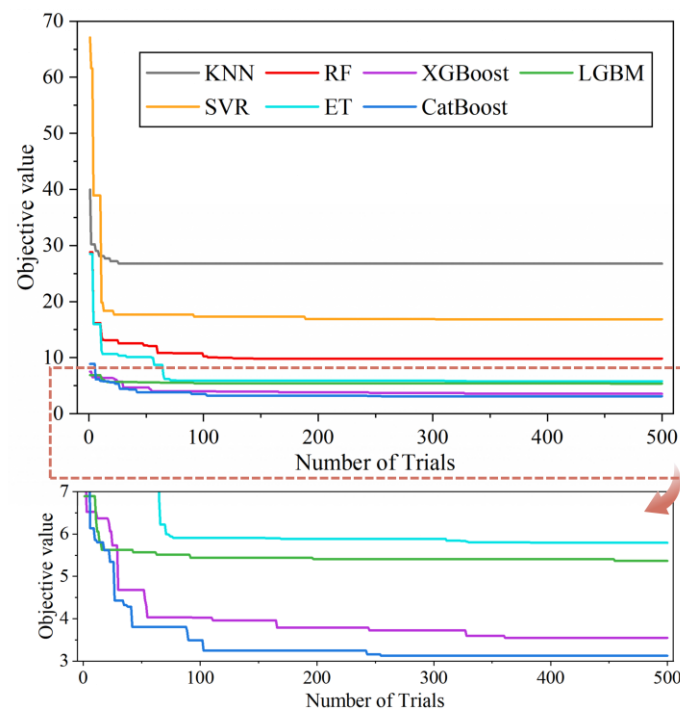


Figure 4. Iterative convergence of the objective function value over 500 optimization trials.

Table 3. Results of model hyperparameter optimization.

Model	Best Objective Value	Optimal Trial Number	Optimal Hyperparameters
KNN	26.7809	26	n_neighbors: 6; weights: distance; p: 2
SVR	16.8423	303	C: 99.7026; epsilon: 0.0137; gamma: 9.9384
RF	9.8207	144	n_estimators: 138; max_depth: 48; min_samples_split: 2; min_samples_leaf: 1
ET	5.7907	383	n_estimators: 116; max_depth: 28; min_samples_split: 2; min_samples_leaf: 1
XGBoost	3.5484	361	n_estimators: 1000; max_depth: 26; learning_rate: 0.1335; subsample: 0.6746; colsample_bytree: 0.6502; reg_alpha: 0.1466; reg_lambda: 2.0844; min_child_weight: 6; gamma: 0.0056
LGBM	5.3658	455	n_estimators: 962; max_depth: 31; learning_rate: 0.1961; subsample: 0.8597; colsample_bytree: 0.7895; reg_alpha: 0.2351; reg_lambda: 2.3606
CatBoost	3.1254	255	Iterations: 399; depth: 3; learning_rate: 0.453; l2_leaf_reg: 2.2635; border_count: 47

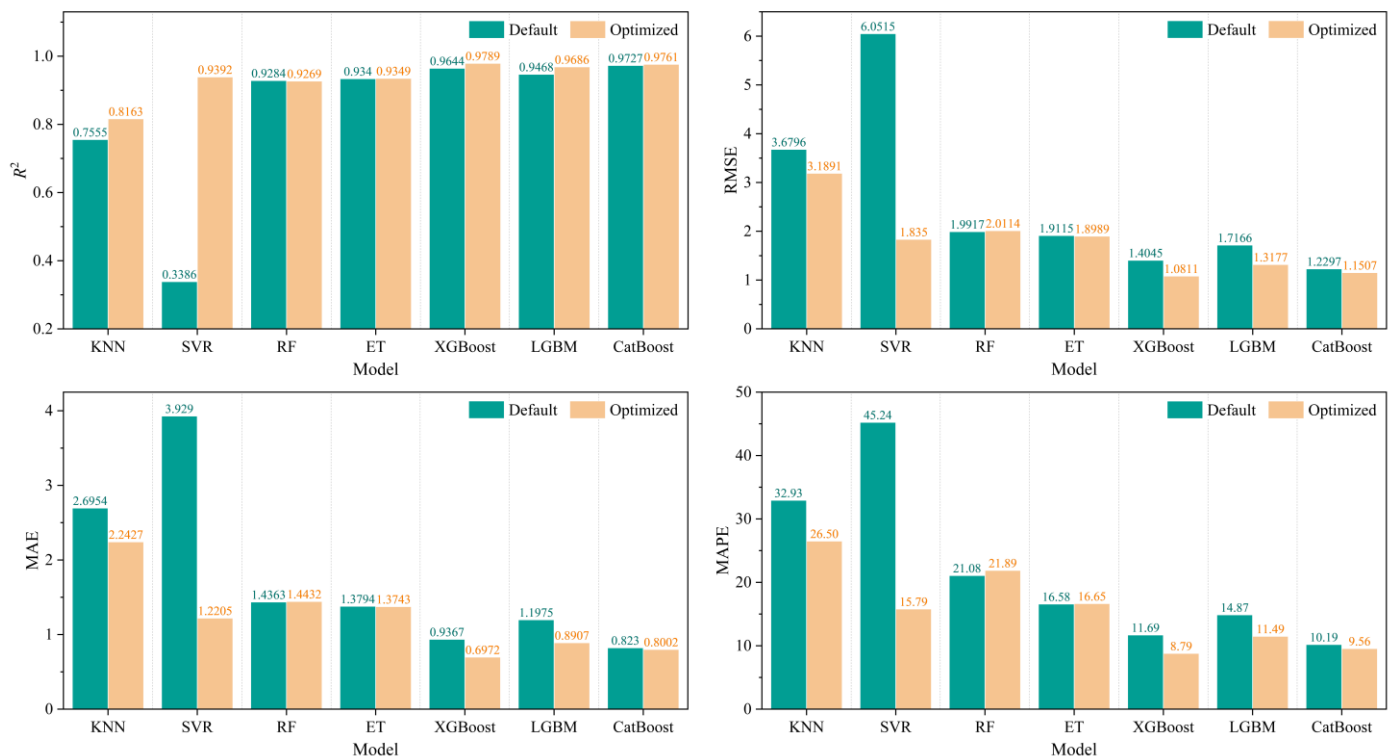


Figure 5. Comparison of model performance before and after optimization.

4.2. Comparison of Model Predictive Performance

After Optuna-based hyperparameter optimization was completed for all models, this section systematically evaluated and compared the overall predictive performance of seven machine learning algorithms for predicting the CD of RA concrete. Multiple evaluation metrics—including R^2 , RMSE, MAE, MAPE, VAF, and MBE—were used to conduct a comprehensive quantitative analysis from the perspectives of prediction accuracy, error distribution, and bias characteristics. The model prediction results are shown in Figure 6. With the exception of KNN, all models demonstrated a high goodness-of-fit on both the training and test sets. On the test set, most models achieved R^2 values generally above 0.93, VAF values exceeding 92%, and MBE values close to zero. The scatter points are largely distributed along the $y = x$ line, indicating no significant systematic bias in the predictions. Although both KNN and ET achieved an R^2 of 0.9953 on the training set, KNN's R^2 dropped to 0.8163 on the test set, accompanied by marked increases in RMSE and MAPE and a long-tailed residual distribution. This pattern reflected overfitting to the training samples and insufficient generalization capability. SVR, RF, and ET showed relatively balanced performance across the training and test sets, though their error levels were slightly higher than those of the gradient boosting models. In contrast, the gradient boosting models—represented by XGBoost, LGBM, and CatBoost—delivered the best overall performance. Their training and test results were closely aligned, with concentrated residual distributions, indicating that these models can not only effectively capture the complex nonlinear relationships governing CD but also possess strong generalization stability.

To quantify the uncertainty of model predictions, Table 4 summarizes the residual statistics and uncertainty metrics for each model on the test set. Regarding systematic bias, the MBE of all models was close to zero, and their respective 95% confidence intervals (95% CIs) all covered zero, indicating that no statistically significant overall trend of overestimation or underestimation was observed at the 95% confidence level. Among them, CatBoost achieved the closest MBE to zero (-0.0084), demonstrating optimal bias

control. LGBM (0.0330) and XGBoost (0.0641) also exhibited only a very slight positive bias, implying an overall marginal tendency toward overestimation that remained statistically insignificant. In contrast, KNN showed a relatively larger absolute MBE (-0.4344), and its median residual was also negative (-0.4461), suggesting a greater tendency toward underestimation on the test set. Although its 95% CI still spanned zero, both the point estimate and the central location of its error distribution indicated that the magnitude and stability of its bias were inferior compared to the gradient boosting models. Further considering error dispersion (uncertainty), XGBoost exhibited the smallest residual standard deviation ($SD = 1.0831$) and interquartile range ($IQR = 0.8404$), and its 95% confidence interval for MBE was also the narrowest ($[-0.1173, 0.2455]$). This indicated that the model not only maintained controllable bias but also achieved the lowest error variability and the best prediction stability. CatBoost and LGBM likewise showed relatively low SD and IQR values, demonstrating good robustness. In contrast, KNN yielded the largest SD and IQR and a wider MBE interval, suggesting a more dispersed error distribution, higher uncertainty, and relatively poorer generalization stability. Overall, considering both bias (MBE/Median) and uncertainty (SD/IQR/CI), the gradient-boosting models outperformed the other methods; among them, XGBoost performed best in terms of stability and uncertainty control, while CatBoost stood out in bias control.

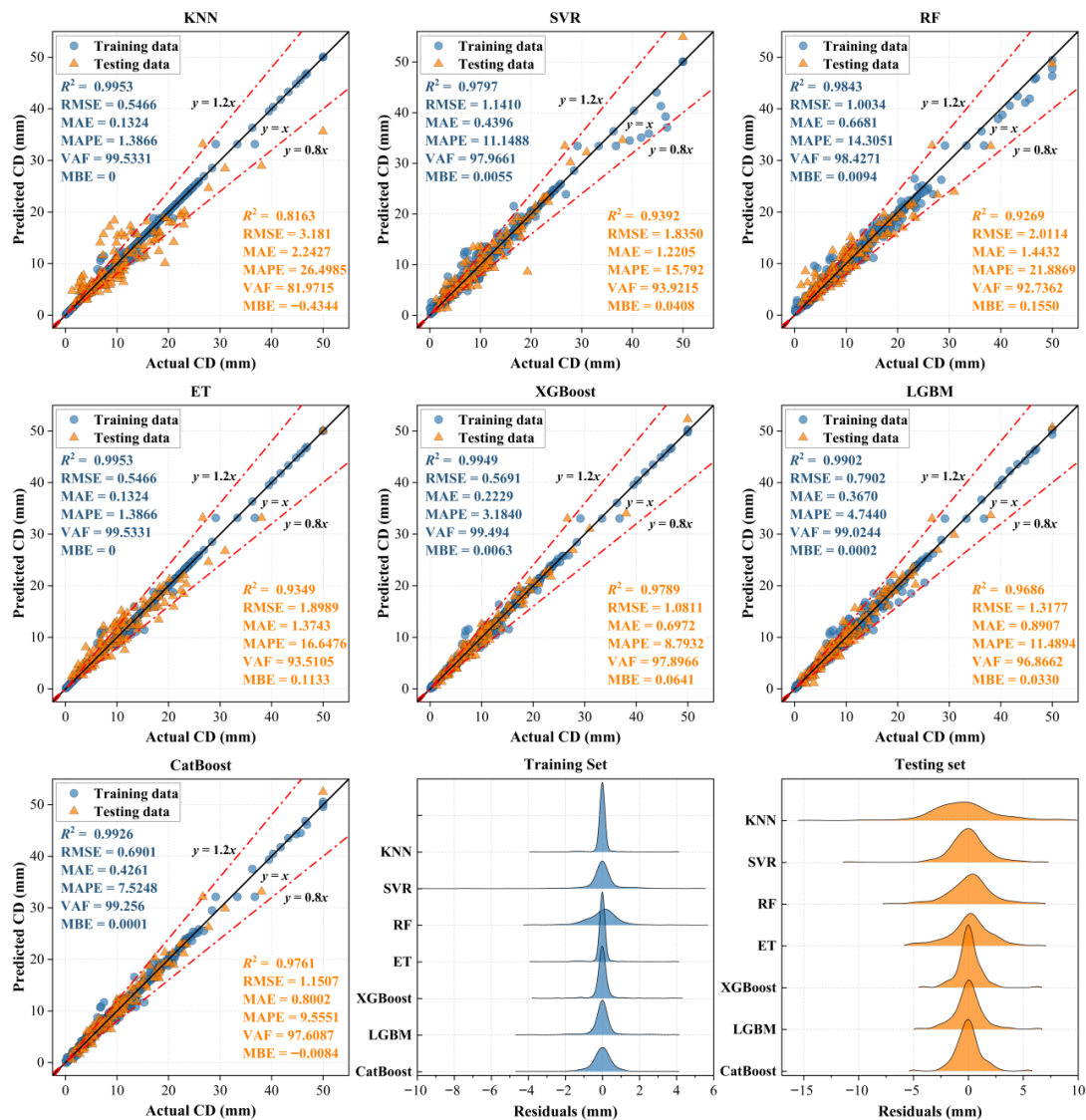


Figure 6. Model prediction results on training and test sets.

Table 4. Test-set residual statistics and uncertainty measures for different models.

Model	MBE	SD *	95% CI *	Median *	IQR *
KNN	−0.4344	3.1710	[−0.9654, 0.0966]	−0.4461	3.1483
SVR	0.0408	1.8413	[−0.2676, 0.3491]	0.0504	1.6022
RF	0.1550	2.0128	[−0.1820, 0.4921]	0.2690	1.9866
ET	0.1133	1.9025	[−0.2053, 0.4319]	0.1774	1.6920
XGBoost	0.0641	1.0831	[−0.1173, 0.2455]	0.0309	0.8404
LGBM	0.0330	1.3221	[−0.1883, 0.2544]	0.0886	1.1036
CatBoost	−0.0084	1.1549	[−0.2018, 0.1850]	−0.0726	1.0788

Note. * SD denotes the standard deviation of the test-set prediction residuals, characterizing the dispersion of errors; 95% CI denotes the two-sided 95% confidence interval of the test-set MBE, where an interval covering zero indicates that the model exhibits no statistically significant systematic bias; Median is the median of the test-set residuals, representing a robust typical signed error that is less sensitive to outliers; and IQR is the interquartile range of the test-set residuals ($Q3 - Q1$), quantifying the spread of the middle 50% of errors.

The performance of a model on the test set reflects its generalization capability to unseen samples. To enable an intuitive multi-dimensional comparison, all evaluation metrics are normalized and visualized in a multi-metric heatmap, such that, on a unified scale, larger values consistently indicate better performance. As shown in Figure 7, XGBoost achieved a normalized score of 1 on five metrics: R^2 , RMSE, MAE, MAPE, and VAF, with only a slightly lower score of 0.869 on the bias metric (MBE), which indicated its overall superiority in both accuracy and stability. CatBoost also achieved scores close to 1 across all metrics and performed best in terms of bias control. LGBM was slightly inferior to the former two, but still maintained scores generally above 0.85 on all metrics, placing its overall performance in the top tier. Taken together, these results support XGBoost as the preferred model for CD prediction in this study, whereas CatBoost and LGBM served as important comparative and complementary models.

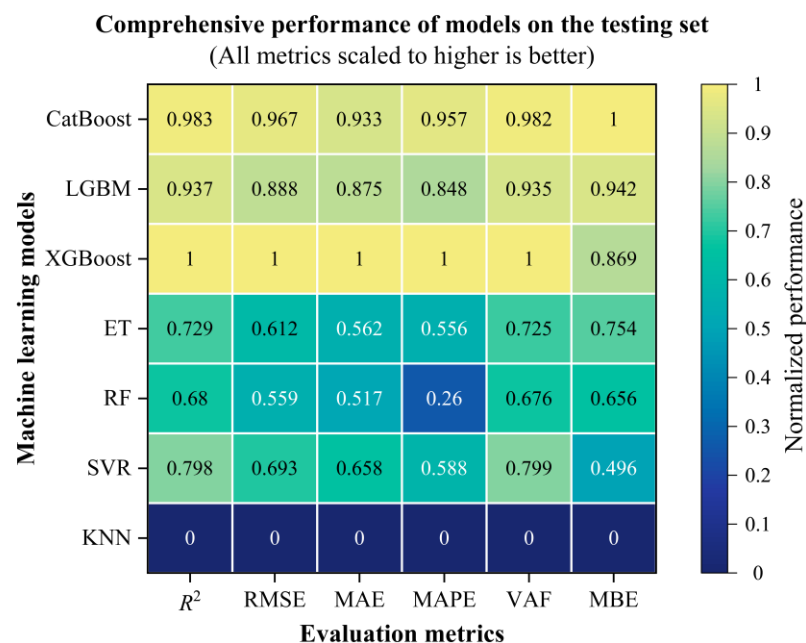
**Figure 7.** Multi-metric heatmap of model performance on the test set.

Table 5 presents a direct performance comparison between the proposed Optuna-XGBoost model and the CD prediction models developed by Xi et al. [37] using the same database. Compared with the best-performing model reported in that study (XGB-MVO), the proposed Optuna-XGBoost exhibited consistently superior performance: R^2 increased from 0.9398 to 0.9789 (an improvement of approximately 4.13%), while RMSE, MAE, and

MAPE decreased from 1.7565 mm, 1.0688 mm, and 18.63% to 1.0811 mm, 0.6972 mm, and 8.7932%, respectively (corresponding reductions of 38.28%, 33.37%, and 59.77%). These results demonstrated that, on an identical data basis, the Optuna-based hyperparameter optimization framework was more effective in unlocking the potential of XGBoost, thereby achieving markedly higher predictive accuracy and stability.

Table 5. Performance comparison with CD prediction models reported in previous studies.

Model	R^2	RMSE	MAE	MAPE (%)
XGB-GA model [37]	0.9401	1.7516	1.0464	21.86
XGB-MVO model [37]	0.9398	1.7565	1.0688	18.63
XGB-SSA model [37]	0.9373	1.7916	1.0277	21.52
This paper (Optuna-XGBoost)	0.9789	1.0811	0.6972	8.7932

4.3. Model Performance Under Different CO₂ Concentration Regimes

The dataset compiled in this study covers a wide range of CO₂ concentrations (0.05–50%), corresponding to both natural and accelerated carbonation test conditions. Natural carbonation refers to the process that occurs when concrete is exposed to atmospheric CO₂ over extended periods under real service conditions. In this study, the outdoor exposure test groups in the dataset are characterized by a CO₂ concentration of 0.05%, representing the natural carbonation condition. Accelerated carbonation, on the other hand, refers to laboratory tests in which the CO₂ concentration is elevated (e.g., 1%, 3%, 5% or higher) to shorten the test duration. Previous studies have indicated that as CO₂ concentration increases, the mechanisms and kinetics of accelerated carbonation may increasingly deviate from those of natural carbonation, particularly when the CO₂ concentration exceeds about 3%—a condition under which accelerated test results may become misleading [52–54]. This discrepancy could potentially affect the ability of a model trained on mixed-regime data to accurately learn the underlying patterns of natural carbonation. To examine whether the inclusion of high-CO₂ accelerated carbonation data in the dataset would impair the model's predictive performance for natural carbonation scenarios, a stratified performance evaluation was conducted for the Optuna-XGBoost model.

Following the approach outlined in relevant literature [52–54], the dataset was divided into three intervals based on CO₂ concentration (CC): (i) natural carbonation (CC = 0.05%); (ii) low-CO₂ accelerated carbonation (0.05% < CC ≤ 3%, whose mechanism is relatively closer to that of natural carbonation); and (iii) high-CO₂ accelerated carbonation (CC > 3%, where more pronounced mechanistic deviations may occur). The prediction results of the Optuna-XGBoost model across these three concentration intervals are summarized in Figure 8 and Table 6. The results showed that the model achieved high predictive accuracy on the test sets of all three CO₂ regimes (R^2 = 0.9640–0.9780). Notably, under natural carbonation conditions, the model still performed excellently on the test set (R^2 = 0.9780, MAPE = 7.9975%), indicating that the inclusion of high-CO₂ accelerated carbonation samples in the training data did not substantially impair the model's predictive capability for the key engineering scenario of natural carbonation. Furthermore, although the absolute errors (RMSE/MAE) in the high-CO₂ regime were significantly larger than those in the other regimes—which could be attributed to the wider range of carbonation depths and the greater data scatter induced by extreme CO₂ conditions—the MAPE remained at a level of 8–10%, demonstrating that the model retains relatively stable predictive performance even under high-CO₂ conditions. In summary, the Optuna-XGBoost model exhibits good reliability and practical potential for predicting the CD of RA concrete across different CO₂ exposure environments.

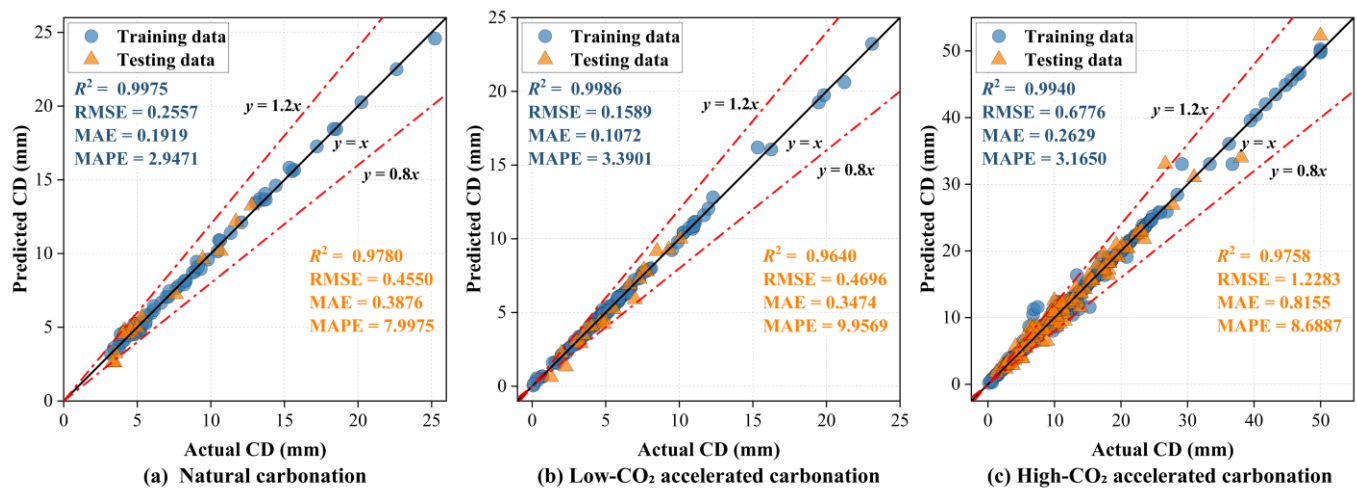


Figure 8. Predicted versus actual CD for the Optuna-XGBoost model under different CO₂ concentration regimes.

Table 6. Performance of the Optuna-XGBoost model across CO₂ concentration regimes.

CO ₂ Concentration Regime	Dataset	R ²	RMSE	MAE	MAPE (%)	VAF	MBE
Natural carbonation	Training	0.9975	0.2557	0.1919	2.9471	99.7471	0.0110
	Testing	0.9780	0.4550	0.3876	7.9975	97.8031	0.0144
Low-CO ₂ accelerated carbonation	Training	0.9986	0.1589	0.1072	3.3901	99.8588	−0.0003
	Testing	0.9640	0.4696	0.3474	9.9569	96.9329	−0.1800
High-CO ₂ accelerated carbonation	Training	0.9940	0.6776	0.2629	3.1650	99.4019	0.0074
	Testing	0.9758	1.2283	0.8155	8.6887	97.5988	0.1203

4.4. SHAP Interpretability Analysis Results

To systematically elucidate the prediction logic and feature-contribution mechanisms of the Optuna-XGBoost model, SHAP was employed for interpretability analysis. SHAP provides two complementary perspectives—global and local. At the global level, it quantifies the overall contribution magnitude and influence direction of each input feature on the model predictions; at the local level, it decomposes an individual prediction into feature-wise contributions, illustrating how multiple features jointly reinforce or offset one another under specific sample conditions and potentially indicating interaction effects. This dual-perspective analysis improved the transparency and interpretability of the model conclusions.

4.4.1. Global Interpretation

Figure 9 shows the SHAP global summary of the Optuna-XGBoost model (beeswarm plot and mean |SHAP value| bar chart), which collectively describes the overall contribution magnitude and effect direction of each input variable on CD. The gray bars denote global feature importance, quantified as the mean absolute SHAP value across all samples. In the beeswarm plot, each feature is represented by a column of points showing the distribution of SHAP values over the dataset: positive SHAP values indicate an increasing contribution to the predicted CD, whereas negative values indicate a decreasing contribution. The color of each dot reflects the relative magnitude of the feature's value.

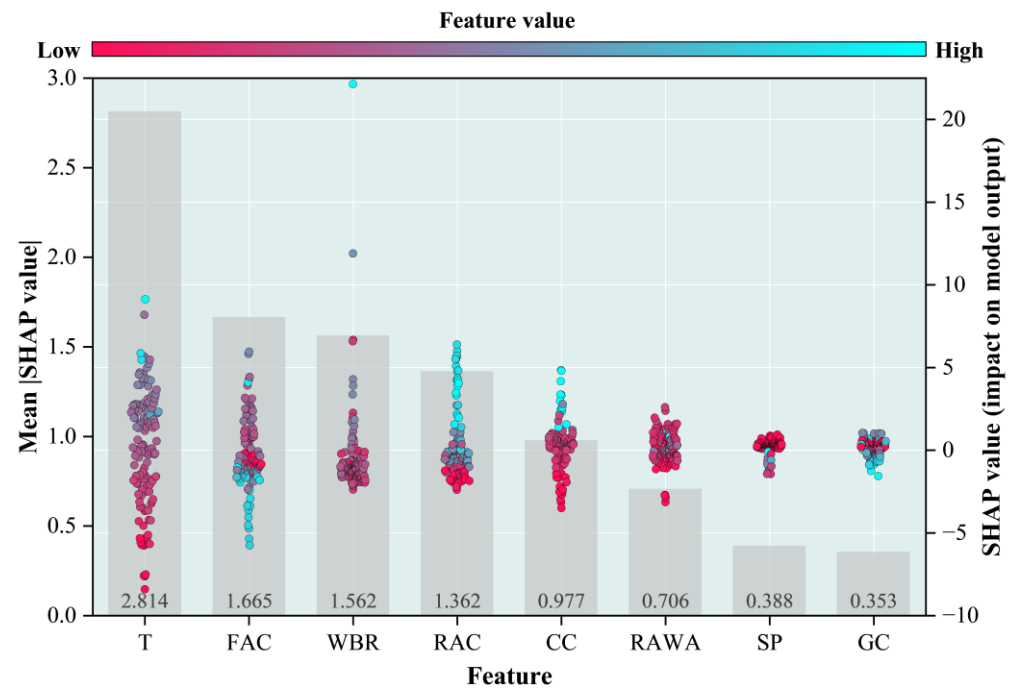


Figure 9. Global SHAP explanation.

Based on the ranking of mean $|SHAP|$, exposure time (T) exhibited the highest global importance (mean $|SHAP\ value| = 2.814$) and served as the dominant factor influencing the prediction of CD. This was followed by FAC (1.665), WBR (1.562), and RAC (1.362). CC (0.977) and RAWA (0.706) showed a moderate level of influence, whereas the marginal contributions of SP (0.388) and GC (0.353) were relatively small. By further considering the correspondence between point color (feature value) and the sign of SHAP values in the beeswarm plot, an overall effect trend was observed: higher values of T, WBR, RAC, and CC were more likely to be associated with positive SHAP values, indicating that, under the joint influence of other variables, these factors tended to increase the predicted CD. In contrast, higher FAC values tended to yield negative SHAP values for most samples, suggesting that increasing FAC was generally associated with a lower predicted CD. It was also noteworthy that some features exhibited substantial dispersion of SHAP values even at similar feature values, implying that the learned response relationships were partly nonlinear and might be jointly modulated by combinations of other variables.

Based on this, to further elucidate the influence pattern of CC on carbonation depth prediction and its coupling effect with T, the SHAP dependence plots for CC and T are plotted in this study (Figure 10), aiming to intuitively reveal their marginal contributions to the predictions and the interaction effect between them. As shown in Figure 10a, the SHAP values for CC generally exhibit an upward trend with increasing CC. Under natural CO_2 conditions ($CC = 0.05\%$), $SHAP(CC)$ values are mostly close to zero or negative, indicating a relatively weak marginal contribution of CC on the predicted carbonation depth. As CC increases, $SHAP(CC)$ gradually shifts from negative to positive and strengthens notably, showing a more pronounced positive contribution especially when CC exceeds approximately 3%. This suggested that the model had learned the response characteristic of a more significant carbonation acceleration effect under high CO_2 conditions. Meanwhile, $SHAP(CC)$ still shows a certain degree of dispersion around similar CC levels, and the dispersion becomes more pronounced in the high- CO_2 range, implying that the effect of CC was not a simple linear and monotonic superposition but was jointly modulated by combinations of other input variables, resulting in a more complex nonlinear response. These findings demonstrated that the model not only identified the main effects of CC

and T individually but also captured the positive interaction mechanism between them in driving the carbonation depth. Combined with the regime-wise performance evaluation in Section 4.3, the aforementioned dependence relationships provided a reasonable explanation for the model's behavior under different carbonation conditions. The stronger dispersion of SHAP(CC) in the high- CO_2 regime was consistent with the wider range of carbonation depths and greater experimental variability within this regime, thereby offering a plausible interpretation for its relatively higher absolute errors (RMSE/MAE). Conversely, under natural/near-natural conditions, the model still exhibited stable marginal responses and high predictive accuracy, demonstrating that its predictive capability for natural carbonation scenarios had not been significantly compromised despite being trained on mixed-regime data.

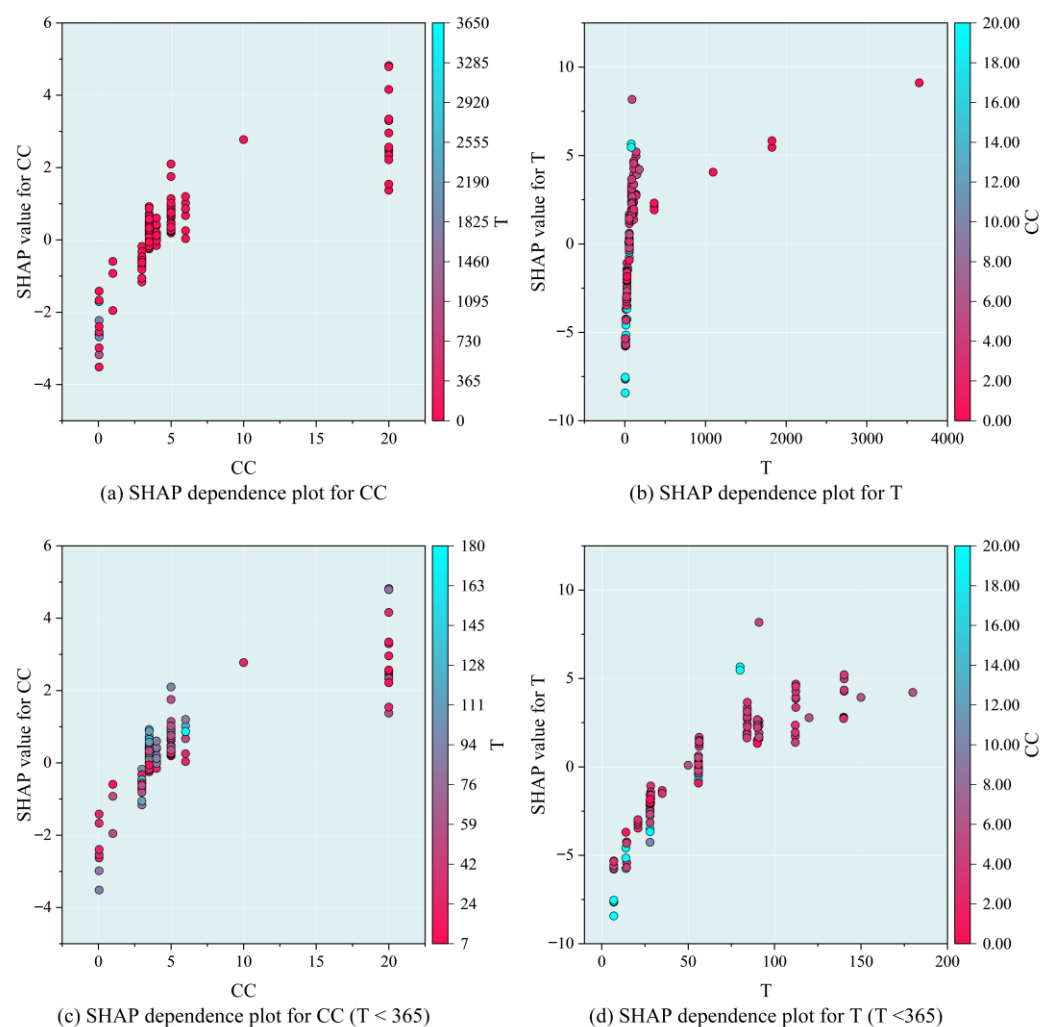


Figure 10. SHAP dependence plot for CC and T.

4.4.2. Local Interpretation

To further reveal the model's decision path at the sample level and the joint mechanisms of multiple features, two representative samples from the test set were selected for local interpretation analysis. Their SHAP waterfall plots are presented in Figure 11.

Each waterfall plot starts from the baseline prediction $E[f(X)]$ (10.112 mm in the figure), which is the expected model output over the entire dataset. The SHAP contribution of each feature, conditioned on the specific values of that sample, is then added sequentially to yield the final prediction $f(x)$ for that sample (e.g., 9.481 mm for sample ① and 12.775 mm for sample ②). In the plot, a red bar indicates that the variable increases the predicted

CD relative to the baseline (positive contribution), whereas a blue bar indicates a decrease (negative contribution); the length of the bar reflects the magnitude of the contribution. The sample-specific value of each feature is also annotated on the left. For instance, in sample ①, the exposure time $T = 120$ d corresponds to a positive contribution of +2.780, implying that—given the fixed values of the other features (RAWA, WBR, FAC, GC, RAC, SP, and CC)—a longer exposure time significantly drives the predicted CD upward.

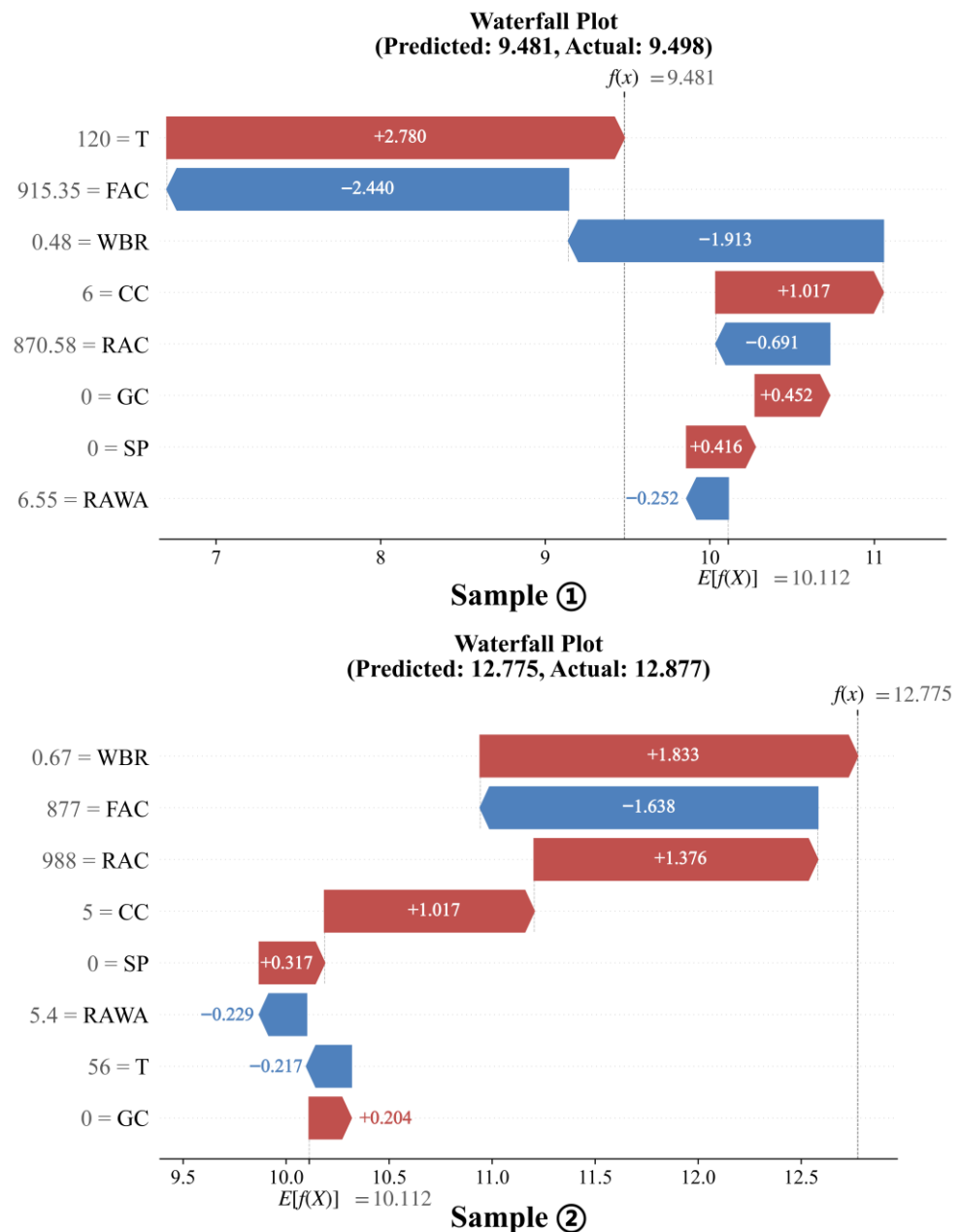


Figure 11. Local SHAP explanation.

The waterfall plots provide an intuitive visualization of how multiple variables reinforce or offset one another under the conditions of a given sample. For Sample ①, although variables such as T (d) and CC (%) exerted strong positive contributions to CD (e.g., $T = 120$ d corresponds to +2.780 and $CC = 6\%$ corresponds to +1.017), these effects were counteracted by the pronounced negative contributions of $WBR = 0.48$ (−1.913) and $FAC = 915.35$ kg/m³ (−2.440), resulting in a final prediction of $f(x) = 9.481$ mm, which is below the baseline. In Sample ②, by contrast, the positive contributions of

WBR = 0.67 (+1.833), RAC = 988 kg/m³ (+1.376), and CC = 5% (+1.017) added in the same direction; although factors such as FAC = 877 kg/m³ (−1.638) exerted a certain suppressing effect on CD, the overall net effect remained positive, yielding a higher prediction of $f(x)$ = 12.775 mm. Notably, the direction and magnitude of a given variable's contribution can vary across samples (e.g., WBR contributed negatively in Sample ① but positively in Sample ②, and T showed a strong positive contribution in Sample ① but a slightly negative contribution in Sample ②). This suggested that the effect of a given variable on CD might vary with different combinations of other mixture parameters (RAWA, FAC, GC, RAC, and SP) and environmental factors (CC and T), thereby indirectly indicating the presence of nonlinear relationships or feature interactions in the model.

Overall, the global SHAP importance ranking and the local decision pathways jointly suggested that the model's response patterns to key variables were broadly consistent with established empirical understanding, which supported the interpretability and credibility of the model results alongside its high predictive accuracy.

5. Limitations and Prospects

Based on 682 experimental samples, systematic hyperparameter optimization was conducted for multiple machine learning models using the Optuna framework, and the best-performing model was interpreted using SHAP. This provided a high-accuracy and interpretable tool for predicting the CD of RA concrete. Nevertheless, several limitations should be acknowledged to clarify the scope of applicability and to outline directions for future research.

First, the input variables in this work mainly included engineering-accessible parameters such as mix proportion, material dosage, and exposure time, which could effectively capture the macroscopic variation trend of CD. However, beyond these factors, the carbonation process of concrete is also significantly affected by relative humidity and temperature, cementitious composition, pore structure, and other factors. Due to incomplete reporting or inconsistent definitions of such information in the multi-source literature-compiled dataset, forcibly incorporating these variables was not feasible as it would substantially reduce the usable sample size and undermine cross-study comparability. Therefore, the interpretability results of this study should be understood as conditional attributions within the current feature framework, rather than a complete elucidation of the carbonation mechanism. Future work may construct a more comprehensive and standardized database and, on that basis, incorporate more representative features to enhance the model's mechanistic representational capability and interpretive depth.

Secondly, the dataset compiled in this study was obtained from different literature sources. Differences in experimental protocols, environmental control methods, specimen preparation, and measurement standards inevitably introduce additional uncertainty. To address the wide span of CO₂ concentrations, the model's learning behavior across different concentration intervals was examined through CO₂-regime-wise performance evaluation and CC–T SHAP dependence analysis. It should be emphasized, however, that the model performance was grounded in the range covered by the training dataset, and extrapolative predictions for extreme conditions beyond the training distribution should therefore be used with caution.

Third, there remains room for improvement in evaluation robustness and external validation. The model comparison in this study relied on a single random train–test split. Although 5-fold cross-validation was used during hyperparameter optimization to reduce overfitting risk, a single split might not have fully represented performance estimation variability. Furthermore, a fully independent external dataset from a unified experimental protocol is lacking for rigorous validation. Future work should prioritize incorporating

such independent data to externally validate the model, further assessing its transferability and reliability in practical engineering scenarios.

In addition, SHAP was used in this study to interpret the model from both global and local perspectives, effectively enhancing the transparency of the predictions. However, it should be noted that post hoc interpretability methods may struggle to fully disentangle main effects from interaction effects when features are correlated or when complex non-linear interactions exist [55]. Therefore, the interpretability results reported in this work should be regarded as a statistical contribution decomposition specific to the feature set and data distribution, rather than strict causal inference. Future work may combine multiple interpretability approaches or conduct more in-depth sensitivity analyses to further elucidate the influence of complex interactions and correlations on the mechanisms underlying CD prediction.

In summary, future work should prioritize expanding and standardizing datasets, refining the feature system, performing rigorous robustness evaluation and external validation, and advancing interaction-sensitive mechanistic interpretation alongside uncertainty modeling. These efforts will enhance the models' predictive reliability and their value as a reference for engineering applications.

6. Conclusions

This study developed a prediction model for the carbonation depth of recycled aggregate concrete by employing the Optuna framework for hyperparameter optimization across seven machine learning algorithms and interpreting the optimal model using the SHAP method, based on a dataset of 682 experimental samples. The main findings are summarized as follows:

- (1) The Optuna-based hyperparameter optimization significantly enhanced the predictive performance of most models, particularly for SVR, KNN, and the gradient boosting models (XGBoost, LGBM, and CatBoost). However, for bagging-based models such as Random Forest and Extremely Randomized Trees, their default parameter settings were already near-optimal, and further optimization yielded limited performance gains. This indicates that systematic hyperparameter tuning is crucial for certain model types, yet its application should be tailored according to the specific characteristics of each algorithm.
- (2) Among the seven models, XGBoost performed best on the test set ($R^2 = 0.9789$, RMSE = 1.0811, MAE = 0.6972, MAPE = 8.7932%, VAF = 97.8966%, and MBE = 0.0641), which demonstrated excellent fitting accuracy and generalization within the range of mix proportions and exposure conditions covered by the dataset in this study. Furthermore, after partitioning the samples by CO₂ concentration for stratified evaluation, the model maintained high predictive accuracy across all test subsets, including under natural carbonation conditions, which indicated that training on mixed-condition data did not significantly compromise its predictive capability for the key engineering scenario of natural carbonation.
- (3) The SHAP-based interpretability analysis (global and local) indicated that exposure time was the most critical factor influencing carbonation depth, followed by fine aggregate content, water-to-binder ratio, and recycled aggregate content. The local waterfall plots further revealed the superposition and offsetting of multiple feature contributions at the individual-sample level, suggesting that the model could capture the complex nonlinear relationships between carbonation depth and the combined effects of mix-design variables and environmental exposure conditions, thereby improving the transparency and engineering acceptability of the prediction results.

- (4) The proposed “Optuna optimization + multi-model comparison + SHAP interpretation” framework enables effective prediction of the carbonation depth of recycled aggregate concrete within the scope of the compiled data, while enhancing the interpretability and credibility of the prediction results. This framework provides a reproducible methodology for data-driven modeling of similar durability issues in construction materials, offering certain engineering reference value. However, its applicability to broader material systems and service environments still requires further validation.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/buildings16020349/s1>.

Author Contributions: Conceptualization, X.L. and E.L.; methodology, Y.C.; software, Y.C.; validation, Y.C., X.L. and E.L.; formal analysis, Y.C.; writing—original draft preparation, Y.C.; writing—review and editing, J.Z., X.L. and E.L.; visualization, Y.C.; supervision, J.Z.; funding acquisition, J.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Deep Earth Probe and Mineral Resources Exploration—National Science and Technology Major Project (2025ZD1010703).

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Rao, A.; Jha, K.N.; Misra, S. Use of aggregates from recycled construction and demolition waste in concrete. *Resour. Conserv. Recycl.* **2007**, *50*, 71–81. [\[CrossRef\]](#)
2. Andrew, R.M. Global CO₂ emissions from cement production, 1928–2018. *Earth Syst. Sci. Data* **2019**, *11*, 1675–1710. [\[CrossRef\]](#)
3. Sáez, P.V.; Osmani, M. A diagnosis of construction and demolition waste generation and recovery practice in the European Union. *J. Clean. Prod.* **2019**, *241*, 118400. [\[CrossRef\]](#)
4. Tran, C.N.N.; Illankoon, I.M.C.S.; Tam, V.W.Y. Decoding Concrete’s Environmental Impact: A Path Toward Sustainable Construction. *Buildings* **2025**, *15*, 442. [\[CrossRef\]](#)
5. Naderpour, H.; Rafiean, A.H.; Fakharian, P. Compressive strength prediction of environmentally friendly concrete using artificial neural networks. *J. Build. Eng.* **2018**, *16*, 213–219. [\[CrossRef\]](#)
6. Tam, V.W.; Butera, A.; Le, K.N.; Li, W. Utilising CO₂ technologies for recycled aggregate concrete: A critical review. *Constr. Build. Mater.* **2020**, *250*, 118903. [\[CrossRef\]](#)
7. Kou, S.C.; Poon, C.S. Long-term mechanical and durability properties of recycled aggregate concrete prepared with the incorporation of fly ash. *Cem. Concr. Compos.* **2013**, *37*, 12–19. [\[CrossRef\]](#)
8. Zhang, J.; Shi, C.; Li, Y.; Pan, X.; Poon, C.S.; Xie, Z. Influence of carbonated recycled concrete aggregate on properties of cement mortar. *Constr. Build. Mater.* **2015**, *98*, 1–7. [\[CrossRef\]](#)
9. Kurda, R.; de Brito, J.; Silvestre, J.D. Water absorption and electrical resistivity of concrete with recycled concrete aggregates and fly ash. *Cem. Concr. Compos.* **2019**, *95*, 169–182. [\[CrossRef\]](#)
10. Malami, S.I.; Anwar, F.H.; Abdulrahman, S.; Haruna, S.I.; Ali, S.I.A.; Abba, S.I. 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. [\[CrossRef\]](#)
11. Zhang, N.; Xi, B.; Li, J.; Liu, L.; Song, G. Utilization of CO₂ into recycled construction materials: A systematic literature review. *J. Mater. Cycles Waste Manag.* **2022**, *24*, 2108–2125. [\[CrossRef\]](#)
12. Carević, V.; Ignjatović, I.; Dragaš, J. Model for practical carbonation depth prediction for high volume fly ash concrete and recycled aggregate concrete. *Constr. Build. Mater.* **2019**, *213*, 194–208. [\[CrossRef\]](#)
13. Mi, T.; Li, Y.; Liu, W.; Dong, Z.; Gong, Q.; Min, C.; Chu, S.H. The effect of carbonation on chloride redistribution and corrosion of steel reinforcement. *Constr. Build. Mater.* **2023**, *363*, 129641. [\[CrossRef\]](#)
14. Jiang, L.; Lin, B.; Cai, Y. A model for predicting carbonation of high-volume fly ash concrete. *Cem. Concr. Res.* **2000**, *30*, 699–702. [\[CrossRef\]](#)
15. Eklou, S.O. Model for practical prediction of natural carbonation in reinforced concrete: Part 1—Formulation. *Cem. Concr. Compos.* **2018**, *86*, 40–56. [\[CrossRef\]](#)

16. Bouzoubaâ, N.; Bilodeau, A.; Tamtsia, B.; Foo, S. Carbonation of fly ash concrete: Laboratory and field data. *Can. J. Civ. Eng.* **2010**, *37*, 1535–1549. [[CrossRef](#)]
17. Zhang, K.; Xiao, J. Prediction model of carbonation depth for recycled aggregate concrete. *Cem. Concr. Compos.* **2018**, *88*, 86–99. [[CrossRef](#)]
18. Zou, Z.; Yang, G. A model of carbonation depth of recycled coarse aggregate concrete under axial compressive stress. *Eur. J. Environ. Civ. Eng.* **2022**, *26*, 5196–5203. [[CrossRef](#)]
19. Saetta, A.V.; Vitaliani, R.V. Experimental investigation and numerical modeling of carbonation process in reinforced concrete structures: Part I: Theoretical formulation. *Cem. Concr. Res.* **2004**, *34*, 571–579. [[CrossRef](#)]
20. Biswas, R.; Li, E.; Zhang, N.; Kumar, S.; Rai, B.; Zhou, J. Development of hybrid models using metaheuristic optimization techniques to predict the carbonation depth of fly ash concrete. *Constr. Build. Mater.* **2022**, *346*, 128483. [[CrossRef](#)]
21. Golafshani, E.M.; Behnood, A.; Kim, T.; Ngo, T.; Kashani, A. Metaheuristic optimization based-ensemble learners for the carbonation assessment of recycled aggregate concrete. *Appl. Soft Comput.* **2024**, *159*, 111661. [[CrossRef](#)]
22. Xiao, J.Z.; Lei, B.; Zhang, C.Z. On carbonation behavior of recycled aggregate concrete. *Sci. China Technol. Sci.* **2012**, *55*, 2609–2616. [[CrossRef](#)]
23. Silva, R.V.; Neves, R.; de Brito, J.; Dhir, R.K. Carbonation behaviour of recycled aggregate concrete. *Cem. Concr. Compos.* **2015**, *62*, 22–32. [[CrossRef](#)]
24. Guo, H.; Shi, C.; Guan, X.; Zhu, J.; Ding, Y.; Ling, T.C.; Zhang, H.; Wang, Y. Durability of recycled aggregate concrete—A review. *Cem. Concr. Compos.* **2018**, *89*, 251–259. [[CrossRef](#)]
25. Núñez, I.; Nehdi, M.L. Machine learning prediction of carbonation depth in recycled aggregate concrete incorporating SCMs. *Constr. Build. Mater.* **2021**, *287*, 123027. [[CrossRef](#)]
26. Liu, K.; Alam, M.S.; Zhu, J.; Zheng, J.; Chi, L. Prediction of carbonation depth for recycled aggregate concrete using ANN hybridized with swarm intelligence algorithms. *Constr. Build. Mater.* **2021**, *301*, 124382. [[CrossRef](#)]
27. Concha, N.C. A robust carbonation depth model in recycled aggregate concrete (RAC) using neural network. *Expert Syst. Appl.* **2024**, *237*, 121650. [[CrossRef](#)]
28. Wang, D.; Tan, Q.; Wang, Y.; Liu, G.; Lu, Z.; Zhu, C.; Sun, B. Carbonation depth prediction and parameter influential analysis of recycled concrete buildings. *J. CO₂ Util.* **2024**, *85*, 102877. [[CrossRef](#)]
29. Saleh, M.A.; Kazemi, F.; Abdelgader, H.S.; Isleem, H.F. Optimization-Based Multitarget Stacked Machine-Learning Model for Estimating Mechanical Properties of Conventional and Fiber-Reinforced Preplaced Aggregate Concrete. *Arch. Civ. Mech. Eng.* **2025**, *25*, 185. [[CrossRef](#)]
30. Taffese, W.Z.; Sistonen, E.; Puttonen, J. CaPrM: Carbonation prediction model for reinforced concrete using machine learning methods. *Constr. Build. Mater.* **2015**, *100*, 70–82. [[CrossRef](#)]
31. Lee, H.; Lee, H.S.; Suraneni, P. Evaluation of carbonation progress using AIJ model, FEM analysis, and machine learning algorithms. *Constr. Build. Mater.* **2020**, *259*, 119703. [[CrossRef](#)]
32. Ehsani, M.; Ostovari, M.; Mansouri, S.; Naseri, H.; Jahanbakhsh, H.; Nejad, F.M. Machine learning for predicting concrete carbonation depth: A comparative analysis and a novel feature selection. *Constr. Build. Mater.* **2024**, *417*, 135331. [[CrossRef](#)]
33. Wang, X.; Yang, Q.; Peng, X.; Qin, F. A review of concrete carbonation depth evaluation models. *Coatings* **2024**, *14*, 386. [[CrossRef](#)]
34. Chen, X.; Liu, X.; Cheng, S.; Bian, X.; Bai, X.; Zheng, X.; Xu, X.; Xu, Z. Machine learning-based modelling and analysis of carbonation depth of recycled aggregate concrete. *Case Stud. Constr. Mater.* **2025**, *22*, e04162. [[CrossRef](#)]
35. Alizamir, M.; Gholampour, A.; Kim, S.; Heddami, S.; Kim, J. Designing a robust extreme gradient boosting model with SHAP-based interpretation for predicting carbonation depth in recycled aggregate concrete. *Artif. Intell. Rev.* **2026**, *59*, 4. [[CrossRef](#)]
36. Moghaddas, S.A.; Nekoei, M.; Golafshani, E.M.; Nehdi, M.; Arashpour, M. Modeling carbonation depth of recycled aggregate concrete using novel automatic regression technique. *J. Clean. Prod.* **2022**, *371*, 133522. [[CrossRef](#)]
37. Xi, B.; Zhang, N.; Li, E.; Li, J.; Zhou, J.; Segarra, P. A comprehensive comparison of different regression techniques and nature-inspired optimization algorithms to predict carbonation depth of recycled aggregate concrete. *Front. Struct. Civ. Eng.* **2024**, *18*, 30–50. [[CrossRef](#)]
38. Ertuğrul, Ö.F.; Tağluk, M.E. A novel version of k nearest neighbor: Dependent nearest neighbor. *Appl. Soft Comput.* **2017**, *55*, 480–490. [[CrossRef](#)]
39. Smola, A.J.; Schölkopf, B. A tutorial on support vector regression. *Stat. Comput.* **2004**, *14*, 199–222. [[CrossRef](#)]
40. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
41. Geurts, P.; Ernst, D.; Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **2006**, *63*, 3–42. [[CrossRef](#)]
42. 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 (KDD 2016), San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. [[CrossRef](#)]

43. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A highly efficient gradient boosting decision tree. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 3149–3157.
44. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: Unbiased boosting with categorical features. In Proceedings of the NeurIPS @ Montréal: The Thirty-Second Annual Conference on Neural Information Processing Systems (NIPS), Montreal, QC, Canada, 3–8 December 2018; Volume 31, pp. 6638–6648.
45. Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD 2019), Anchorage, AK, USA, 4–8 August 2019; pp. 2623–2631. [\[CrossRef\]](#)
46. Ekanayake, I.U.; Meddage, D.P.P.; Rathnayake, U. A novel approach to explain the black-box nature of machine learning in compressive strength predictions of concrete using Shapley additive explanations (SHAP). *Case Stud. Constr. Mater.* **2022**, *16*, e01059. [\[CrossRef\]](#)
47. Nguyen, H.; Vu, T.; Vo, T.P.; Thai, H.T. Efficient machine learning models for prediction of concrete strengths. *Constr. Build. Mater.* **2021**, *266*, 120950. [\[CrossRef\]](#)
48. Zhou, J.; Chen, Y.; Li, C.; Qiu, Y.; Huang, S.; Tao, M. Machine learning models to predict the tunnel wall convergence. *Transp. Geotech.* **2023**, *41*, 101022. [\[CrossRef\]](#)
49. Li, E.; Zhang, N.; Xi, B.; Yu, Z.; Fissaha, Y.; Taiwo, B.O.; Segarra, P.; Feng, H.; Zhou, J. Analysis and modelling of gas relative permeability in reservoir by hybrid KELM methods. *Earth Sci. Inform.* **2024**, *17*, 3163–3190. [\[CrossRef\]](#)
50. Chen, Y.; Zhang, Y.; Li, C.; Zhou, J. Application of XGBoost model optimized by multi-algorithm ensemble in predicting FRP–concrete interfacial bond strength. *Materials* **2025**, *18*, 2868. [\[CrossRef\]](#)
51. Chen, Y.; Kadkhodaei, M.H.; Zhou, J. Development of the Optuna-NGBoost-SHAP model for estimating ground settlement during tunnel excavation. *Undergr. Space* **2025**, *24*, 60–78. [\[CrossRef\]](#)
52. Von Greve-Dierfeld, S.; Lothenbach, B.; Vollpracht, A.; Wu, B.; Huet, B.; Andrade, C.; Medina, C.; Thiel, C.; Gruyaert, E.; Vanoutrive, H.; et al. Understanding the carbonation of concrete with supplementary cementitious materials: A critical review by RILEM TC 281-CCC. *Mater. Struct.* **2020**, *53*, 136. [\[CrossRef\]](#)
53. Vollpracht, A.; Gluth, G.J.G.; Rogiers, B.; Uwanuakwa, I.D.; Phung, Q.T.; Villagran Zaccardi, Y.; Thiel, C.; Vanoutrive, H.; Etcheverry, J.M.; Gruyaert, E.; et al. Report of RILEM TC 281-CCC: Insights into Factors Affecting the Carbonation Rate of Concrete with SCMs Revealed from Data Mining and Machine Learning Approaches. *Mater. Struct.* **2024**, *57*, 206. [\[CrossRef\]](#)
54. Uwanuakwa, I.D.; Akpınar, P. Enhancing the Reliability and Accuracy of Machine Learning Models for Predicting Carbonation Progress in Fly Ash-Concrete: A Multifaceted Approach. *Struct. Concr.* **2024**, *25*, 3020–3034. [\[CrossRef\]](#)
55. Molnar, C.; König, G.; Herbringer, J.; Freiesleben, T.; Dandl, S.; Scholbeck, C.A.; Casalicchio, G.; Grosse-Wentrup, M.; Bischl, B. General Pitfalls of Model-Agnostic Interpretation Methods for Machine Learning Models. In *xxAI—Beyond Explainable AI*; Holzinger, A., Goebel, R., Fong, R., Moon, T., Müller, K.-R., Samek, W., Eds.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2020; Volume 13200, pp. 39–68. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.