

<https://doi.org/10.1038/s44296-025-00053-z>

Machine learning for efficient CO₂ sequestration in cementitious materials: a data-driven method



SUN Yanjie¹, ZHANG Chen^{1,2}, WEI Yuan-Hao¹, JIN Haoliang³, SHEN Peiliang⁴, POON Chi Sun⁴, YAN He⁵ & WEI Xiao-Yong^{1,6}✉

Extensive experimental work has proved that CO₂ sequestration by cementitious materials offers a promising venue for addressing the rising carbon emissions problem. However, relying merely on experiments on specific materials or some simple empirical methods makes it difficult to provide a comprehensive understanding. To address these challenges, this paper applies three advanced machine-learning techniques (Decision Tree, Random Forest, and eXtreme Gradient Boosting (XGBoost)), with existing datasets coupling with data collected from the literature. The results show that the XGBoost model significantly outperforms traditional linear regression approaches. In addition, aiding in the SHapley Additive exPlanations (SHAP), apart from the widely recognized factors, cement type was also investigated and shown its crucial role in affecting carbonation depth. CEM II/B-LL and CEM II/B-M are two types having high carbonation potential. The results enable the identification of key factors influencing CO₂ sequestration through cement and provide insights into optimizing experimental design.

Greenhouse gases, particularly carbon dioxide (CO₂), are major and well-agreed contributors to climate change, leading to a global push for strategies aimed at reducing carbon emissions. This has resulted in significant efforts focused on carbon capture, utilization, and storage (CCUS), which seek to mitigate CO₂ levels in the atmosphere and reduce the impacts of climate change¹. Among the various CCUS strategies, one of the most promising approaches to close the carbon loop involves the sequestration of CO₂ through hydration products in cementitious materials, which was previously considered the major source of CO₂ emission^{2–4}. The main advantage of cementitious material carbonation is its favorable thermodynamics⁵. However, one of the main challenges is the slow kinetics of the carbonation reaction, which limits the efficiency and overall sequestration capacity of these materials.

The CO₂ absorption capacity of cementitious materials is collectively influenced by various factors, including the carbonation environment⁶, relative humidity (RH)⁷, water-to-binder(w/b) ratio⁸, carbonation type⁹, etc. However, based on the complexity of the cement system, conventional experimental approaches have typically focused on examining the impact of individual factors. Although some traditional studies tried to analyze the combination effects of these factors, these analyses were primarily empirical

and struggled to generalize to new data. A summary of the influencing factors studied in the literature is provided in Table 1. Limited studies have focused on combinations of the effect of multiple factors on carbonation due to the constraints of analytical methods and the challenges of examining larger factor combinations. For instance, by investigating the influence of factor CO₂ concentration, the values of factors relative humidity and temperature have to be controlled to ensure the reliability and reproducibility of experimental results¹⁰. Linear methods, such as fitting empirical laws, have been utilized for analysis⁷.

However, assessing the combined effects of multiple factors across varying degrees remains challenging. For instance, Liu et al.¹¹ conducted numerous tests to study the coupling effect of relative humidity and CO₂ concentration, attempting to model the relationship. Despite their efforts, the study was constrained by the limited data available for fitting the effects of three factors (temperature, relative humidity, and CO₂ concentration). Also, the conclusions drawn were restricted to the specific conditions of the experimental batch, limiting their broader applicability. This limitation impedes the identification of the optimal combination when considering multiple variables, as the best setting for individual factors may not necessarily lead to the best outcome when they are combined. For example,

¹Department of Computing, The Hong Kong Polytechnic University, Hong Kong SAR, China. ²School of Hotel and Tourism Management, The Hong Kong Polytechnic University, Hong Kong SAR, China. ³Department of Materials Science and Engineering, University of Sheffield, Sheffield, UK. ⁴Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University, Hong Kong SAR, China. ⁵Department of Chemistry, Hong Kong University of Science and Technology, Hong Kong SAR, China. ⁶College of Computer Science, Sichuan University, Chengdu, China. ✉e-mail: x1wei@polyu.edu.hk

Table 1 | Influencing factors of carbonation rate reported in the literature

Ref.	Data scale	Influence factors	Basic conclusions
26	16 groups	RH, Cement Type	RH has a positive effect at low values and reaches a peak in the range of 60% to 70%. The type of cement significantly affects carbonation; specifically, the use of silica fume decreases the carbonation depth.
6	9 groups	CO ₂	Elevated CO ₂ concentrations result in increased consumption of calcium silicate hydrate (C-S-H) and clinker.
31	4 groups	Carbonation type	Accelerated testing alters the ranking of cement types based on their influence.
32	13 groups	CO ₂ , RH	A high concentration of CO ₂ and moderate RH (around 50%) are essential factors for enhancing the degree of carbonation.
33	14 groups	Temperature	The rate of carbonation was significantly influenced by temperature, with the optimal carbonation occurring at approximately 100 °C.
34	16 groups	Water/cement ratio	Water/cement ratio is the most important influence factor.
35	9 groups	Water/cement ratio, time	Carbonation depth increases with water/cement ratio and the age of concrete.

Table 2 | Performance of each model

Model	Training set <i>MSE</i>	Training set <i>RMSE</i>	Test set <i>MSE</i>	Test set <i>RMSE</i>
Decision Tree	1.494×10^{-33}	3.865×10^{-17}	3.589	1.890
Random Forest	0.173	0.416	1.947	1.395
XGBoost	0.005	0.073	1.774	1.332

Leemann et al. indicated that if the RH is controlled at 57%, the increase in CO₂ concentration shows a limited carbonation effect. However, if the RH was CO₂ increased to 70% or above, there would be a positive relationship between concentration and carbonation effect¹⁰. We addressed these issues by reorganizing existing data into a dataset that meets the requirements of advanced machine-learning methods. By applying machine learning to analyze these complex variables, such as the influence of environmental factors (e.g., CO₂ concentration, relative humidity, temperature) and cement type, on carbonation depth, our research offers new insights and predictive capabilities that can contribute to more efficient and sustainable cement design practices.

In terms of methodology, these challenges come from the simplicity of conventional methods. These methods are based on linear models, and thus, it is hard to capture the multivariate effects. The inability to construct computational models arises from the reliance on the regression of empirical functions with the experimental data. This motivates us to conduct a study to address these issues by leveraging advanced machine-learning technology. Machine-learning approaches have been proven across diverse fields for their ability to effectively model intricate multivariate and inter-factor relationships, and also in the field of cement research, such as compressive strength^{12–14}, alkali-silica reaction expansion¹⁵, geopolymerization process¹⁶. As for the cement carbonation research, machine learning is also beginning to receive increasing attention^{17–19}. However, most of the research only focused on their own experimental results, and the data size is limited. Thus, in this study, we incorporated a large-scale dataset to establish a more generalized prediction model. We have customized three machine-learning models based on decision tree (DT)²⁰, random forest (RF)²¹, and eXtreme Gradient Boosting (XGBoost) models²² by refining with empirical functions, to predict the CO₂ sequestration capacity of cementitious materials. To further explain the model's internal mechanism, SHapley Additive exPlanations (SHAP). SHAP is a widely recognized method for interpreting model outputs by quantifying the contribution of each feature²³, and it has been extensively applied across various domains, including cementitious material carbonation, where it has been used to evaluate the influence of material composition²⁴.

The proposed method demonstrates its capability not only in identifying the best combination but also in facilitating cost management. With computational models that can generalize to new factor combinations, we overcome the limitations of traditional empirical analysis. To exemplify this, two applications in cement CO₂ sequestration and multi-variable optimization have been integrated. This study aims to enhance our understanding

of the effects of various factors, thereby informing the design of carbon capture strategies in cementitious materials.

Results

Model development and evaluation

As shown in Table 2, for the training set, the DT model has the lowest *MSE* and *RMSE* on the training set, indicating it fits the training data well. RF and XGBoost have higher *MSE* and *RMSE* values on the training set, indicating a less precise fit compared to the DT. Decision tree may create overly complicated trees that capture noise. RF and XGBoost show much lower *MSE* and *RMSE* values on the test set, with XGBoost having the lowest values among all models (*MSE* and *RMSE* were decreased by 51.57% and 29.52% compared with the performances of DT model). The Taylor's diagram was also plotted, as shown in Fig. 1, and the results are consistent with our findings. It shows that XGBoost performs the best on the test set, followed by RF, with DT showing the lowest performance. This suggests that these models generalize better to new data and are more robust in predicting CO₂ sequestration capacity. The performance of each model can be further illustrated by Fig. 2. The green points indicate the model performance on the training set; The DT model demonstrates superior performance compared to RF and XGBoost. But when it comes to the performance on the test set, a clear dispersion can be found on the DT model. While the performances of RF and XGBoost are better at generalization. Overall, XGBoost is the most effective model for predicting CO₂ sequestration capacity in cementitious materials, balancing both training accuracy and test set performance. One key advantage of XGBoost over the other two models is its ability to correct errors during its intermediate steps, which prevents the propagation of errors and the snowball effect that might occur in the other models. This feature allows XGBoost to fit the existing data more effectively.

When compared to previous machine-learning models applied to cement carbonation, the XGBoost model has a good performance in both accuracy and complex interaction processing. The high effectiveness of tree-based methods like XGBoost enables it to address problems for some traditional regression models. Thus, the selected XGBoost model is also compared with the traditional empirical regression method using a subset of the dataset. The randomly selected factors were RH (65%), w/b ratio (0.37), CO₂ concentration (0.045%), temperature (20 °C), carbonation type (NAC), cement type (CEM II/B-V), cement strength class (42.5), cement strength development (R), cement content (75%), and addition type (fly ash). To evaluate the carbonation depth variation depending on carbonation time, the comparison is shown in Fig. 3. The results show that the XGBoost model

significantly outperforms the traditional empirical regression, with the *MSE* and *RMSE* reduced by 99.79% and 95.43%, respectively. This highlights the practical advantage of our model, as the traditional empirical regression makes it hard to capture the complex non-linear relationships in the data, while XGBoost effectively models them. So, in the following sections, the XGBoost model was selected to perform model interpretation and carbonation depth prediction.

SHAP interpretation and feature analysis

The SHAP analysis of the XGBoost model reveals a comprehensive order of feature importance, as illustrated in Fig. 4a. The results align with the established understanding that carbonation time is the most significant factor influencing carbonation depth (contributing 21.8% to the model's predictions); longer duration results in deeper carbonation depth²⁵. In Addition to carbonation time, *RHRH*²⁶, which accounts for 16.6%, the *w/b* ratio⁸ at 15.3%, and *CO*₂ concentration⁶ at 11.9% are identified as the three most critical factors affecting concrete carbonation, aside from carbonation time. Figure 4b, a SHAP waterfall summary, provides deeper insights into the influence of each feature. This visualization shows the both magnitude

and direction of influence factors. For relative humidity, both extremely low and excessively high levels negatively impact carbonation. When the RH is too high, the transport of reactants is impeded, thereby slowing down the carbonation process⁷. For the *w/b* ratio, it positively correlates with carbonation, a higher *w/b* ratio promotes carbonation. This is because an increased *w/b* ratio leads to higher porosity and a greater degree of cement hydration, which in turn enhances the carbonation reaction⁸. Furthermore, the figure emphasizes the importance of *CO*₂ concentration in accelerating the carbonation reaction. A higher *CO*₂ concentration environment will contribute to the reaction of calcium hydroxide. An important finding from this research is the significant role of cement type, which is also a key factor influencing carbonation. This will be explored in detail in the following sections.

Quantitative analysis

In the quantitative analysis using SHAP values, interactions between features were observed. Given that the calculation of SHAP values for individual features is independent, it was deemed appropriate to aggregate the SHAP values of certain features. Some factors can be easily adjusted in practical applications, and analyzing these factors enables the model to assist in experimental design, thereby reducing the need for extensive experimentation.

Prior to conducting the pair effect analysis, the importance of the top 10 feature pairs, excluding carbonation time, was identified (Fig. 5). Carbonation

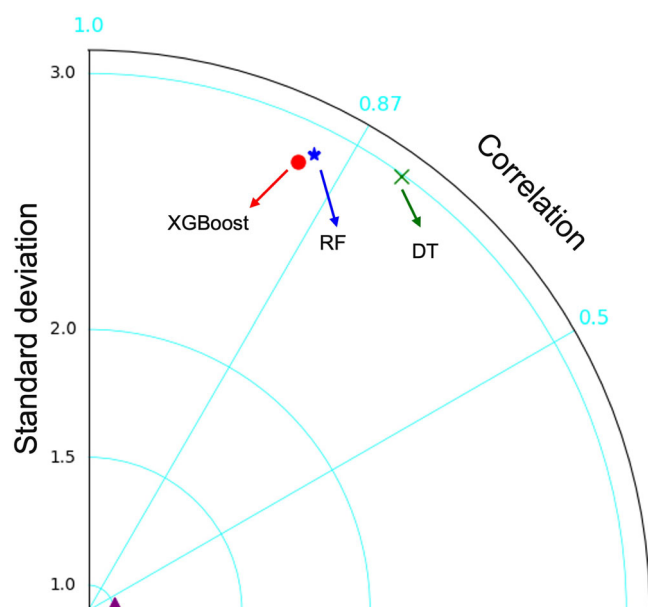


Fig. 1 | Model comparison using Taylor's diagram. The purple triangle represents the observed values, the red circle represents XGBoost results, the blue asterisk represents RF results, and the green cross represents DT results.

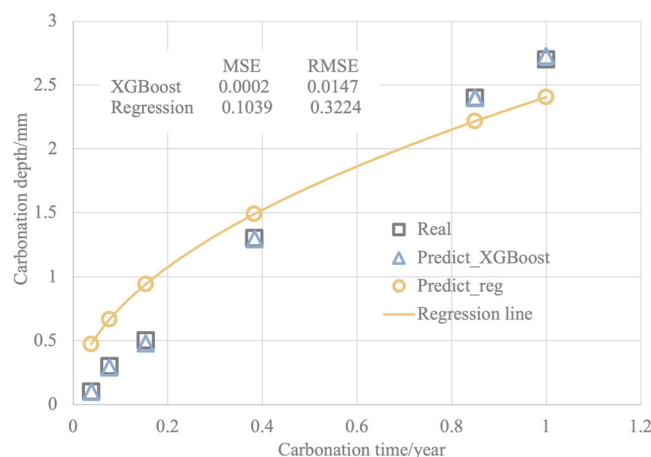


Fig. 3 | The comparison between traditional empirical regression and XGBoost model with selected data. The black rectangle represents the real data, the blue triangle represents the results of XGBoost prediction, the yellow circle represents the regression results, and the yellow line represents the regression line.

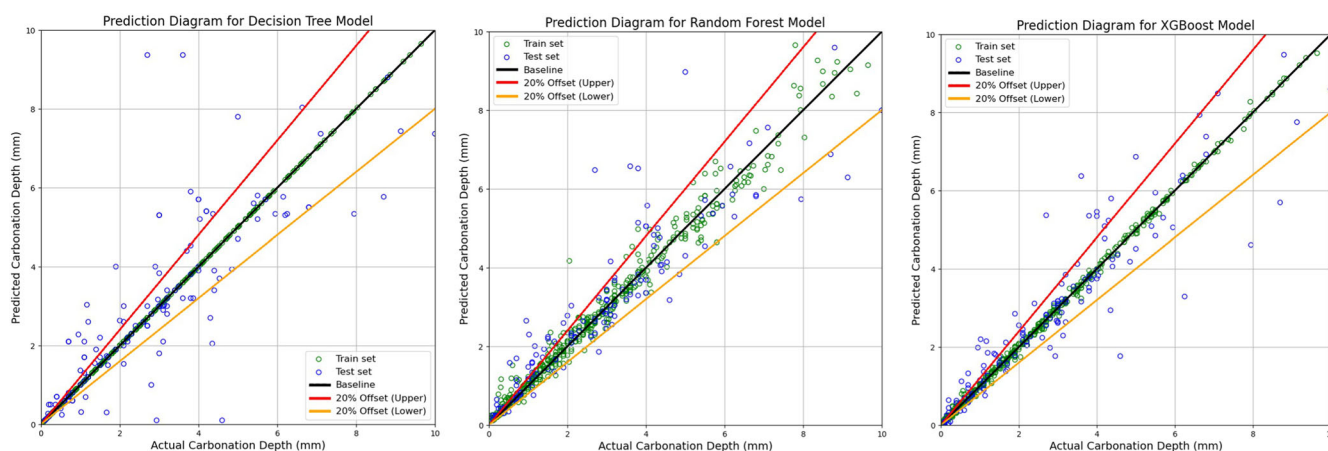


Fig. 2 | The prediction diagram for train and test sets of each model (DT, RF, and XGBoost). The black line represents the baseline, the red line indicates the 20% upper offset, the yellow line indicates the 20% lower offset, the green circle marks the training set, and the blue circle marks the test set.

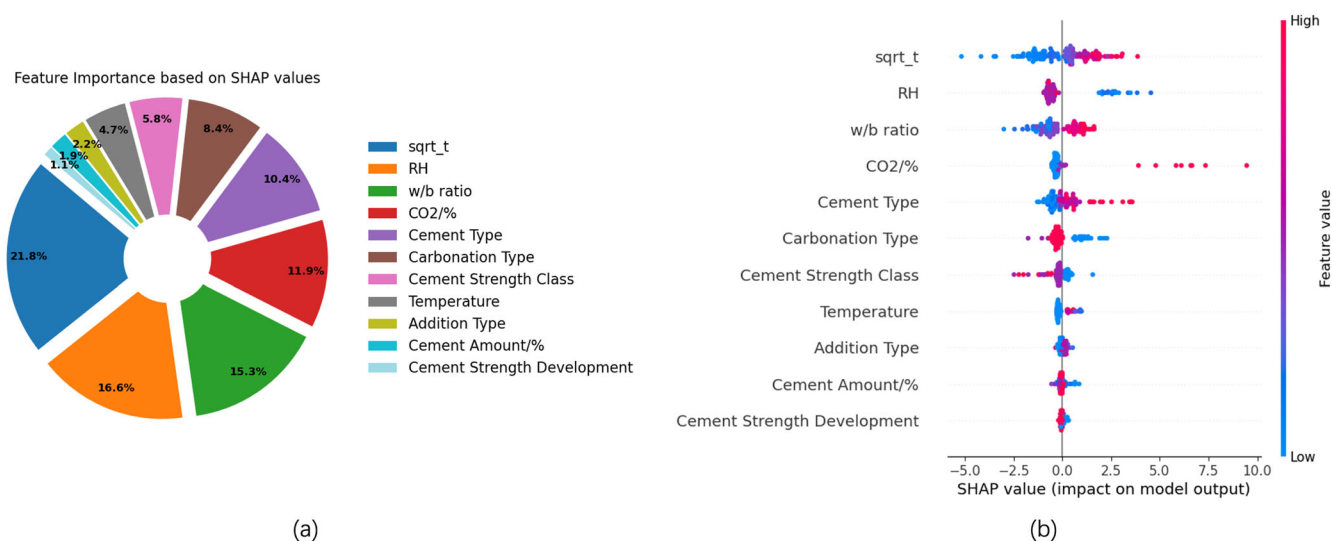
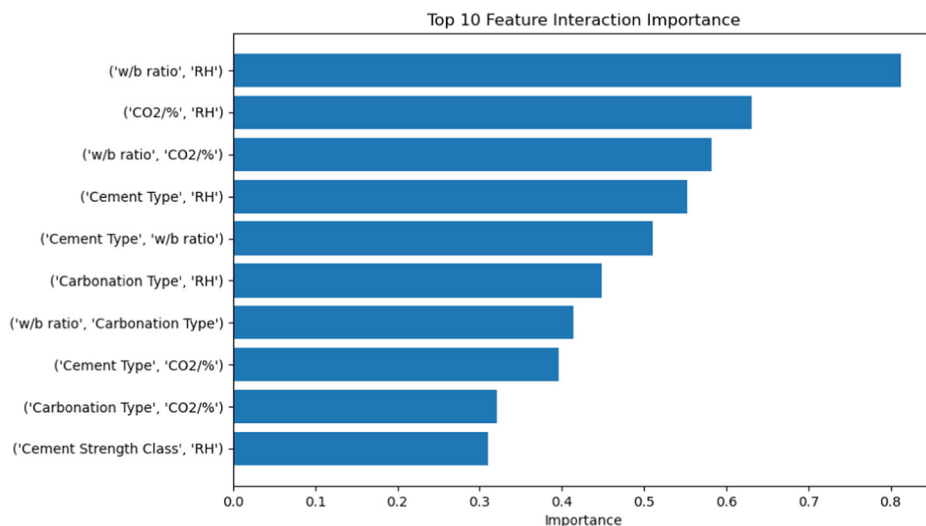


Fig. 4 | SHAP interpretation of the XGBoost model. a Feature importance. **b** SHAP waterfall.

Fig. 5 | The importance of the combination effects.



time was excluded as it is widely understood that longer carbonation time leads to greater carbonation depth. The analysis indicates that the pair consisting of the w/b ratio and RH is the most influential. These two features are commonly encountered and easily adjustable factors in carbonation control.

As illustrated in Fig. 6a, the combined effects of the w/b ratio and RH reveal that the highest SHAP value is achieved with a w/b ratio of 0.58 and an RH of 80%. Figure 6b depicts the relationship between CO₂ concentration and RH, indicating that a higher CO₂ concentration combined with an RH between 70% and 75% is preferable. Figure 6c presents the relationship between the w/b ratio and CO₂ concentration, demonstrating that both factors positively influence carbonation, with a w/b ratio of 0.78 and a CO₂ concentration of 4.2% achieving the highest carbonation depth. As mentioned earlier, cement type is another crucial factor. Its relationship with RH and the w/b ratio is analyzed in Fig. 6d, e. The trends for RH and the w/b ratio are similar, with cement types 6 and 7, corresponding to CEM II/B-LL and CEM II/B-M, showing superior carbonation performance.

Case study

After establishing and evaluating the machine-learning model, it was selected to assist in the experimental design process. To assess the model's effectiveness in this context, a case study was conducted to evaluate its suitability for guiding experimental design. A new dataset from an academic

study¹⁰ was utilized for this purpose. The dataset includes varied features such as Cement Type (CEM I, CEM III/A, CEM II/B-LL), Cement Strength Class (52.5, 42.5, 32.5), Cement Strength Development (N, R), and w/b ratio (0.65, 0.40). Importantly, this data had not been previously trained by the model. After inputting the data, the results, as shown in Fig. 7, revealed some discrepancies between experimental and predicted outcomes. For example, for CEM III/A 42.5 N 0.65, the deviation between the experiment and the prediction is 29.4%, representing a poor performance of the model. But for CEM III/A 42.5 N 0.40, the deviation is only 0.8%. This is reasonable, given that the dataset was unseen by the model, and there may be other unaccounted features influencing the results. However, the results are still meaningful, with certain clear trends consistent between experimental and predicted values. For instance, the earlier analysis indicated that CEM II/B-LL has a more significant effect on carbonation, which is also observed in the case study, where the *K* value for this cement type is much higher than for the other two. This demonstrates that the model is sensitive to cement type, making it a useful tool for cement selection. Also, the model is more accurate to higher w/b ratios (the mean deviation for w/b 0.65 groups is 3.0%), where the carbonation trends for different cements become more distinct. However, the model shows less sensitivity to Cement Strength Development, resulting in lower accuracy for these predictions, as also suggested by the earlier SHAP analysis.

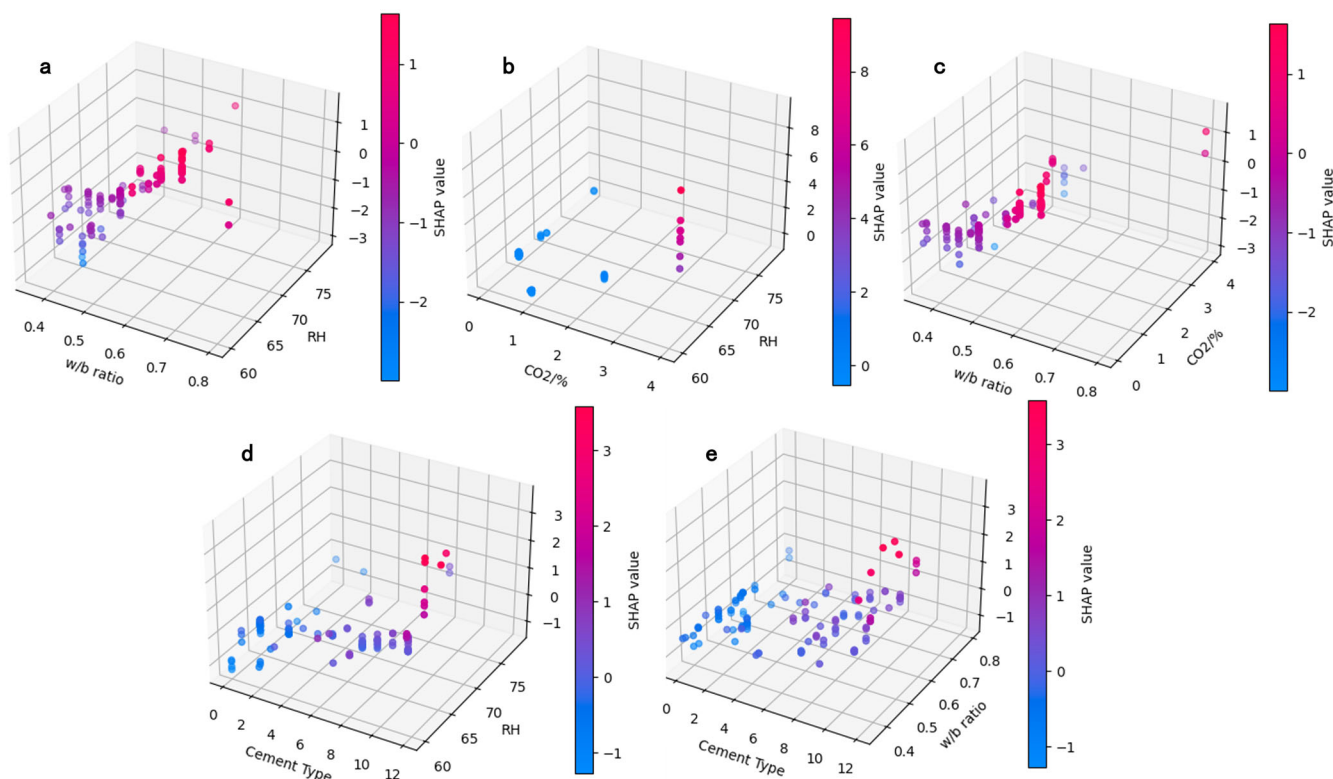
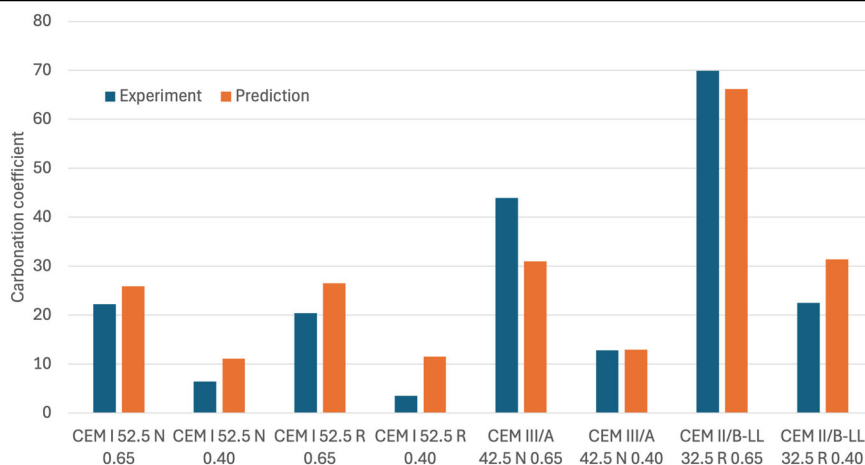


Fig. 6 | Combined effect analysis of the top 5 pairs. a w/b ratio-RH, **b** CO₂-RH, **c** w/b ratio-CO₂, **d** cement type-RH, and **e** cement type-w/b ratio.

Fig. 7 | Case study results. The blue bar chart represents the experimental results, and the orange bar chart represents the prediction results.



Another case study is to use the prediction model to generate more data and project the results onto a three-dimensional space to visualize the combination effects of different factors. Take the influence factors RH and w/b ratio as an example, the results are shown in Fig. 8. The findings indicate that the trend does not strictly follow the general assumption that an increase in the w/b ratio leads to a decrease in the carbonation coefficient. The carbonation coefficient increases with the w/b ratio at the RH around 65%–70%. But when RH reached over 75%, the effect of the w/b ratio can be omitted. With such an analysis, the combined effects of other factors can be obtained. And by controlling for certain variables, a more accurate trend can be observed. Based on the above case studies, we can see that machine learning can assist in cement carbonation design. With limited data, these methods can provide a preliminary estimate. While for more detailed data, they can offer a more visually insightful representation.

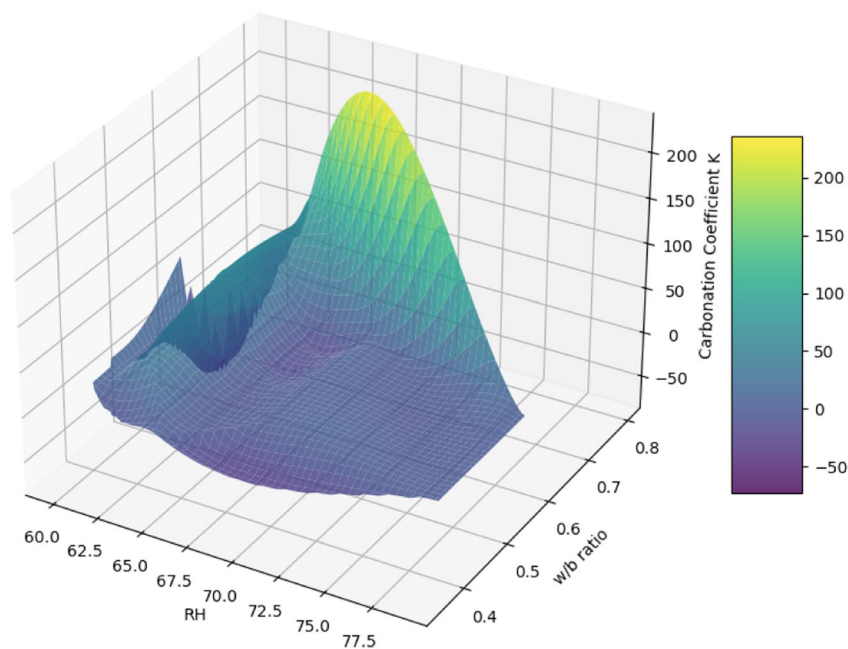
Discussion

In this study, experimental results from the literature were analyzed using three different machine-learning models, and the performance of each model was evaluated. Further interpretation was performed using SHAP analysis. The XGBoost model was found to outperform both the decision tree and the random forest models. The main cause was attributed to its nature as an ensemble model, which combined the predictions of multiple individual decision trees. Notably, XGBoost employed a boosting technique where trees were built sequentially, with each subsequent tree attempting to correct the errors of the previous trees, leading to a model that can capture complex patterns more effectively.

The model identified carbonation time, relative humidity, w/b ratio, and CO₂ concentration as the most influential factors, collectively accounting for over 50% of the feature importance. The new finding was that

Fig. 8 | Case study RH VS w/b ratio.

Smoothed Surface Plot for RH, w/b ratio, and Carbonation Coefficient K



cement type was also found to significantly affect carbonation, with CEM II/B-LL and CEM II/B-M exhibiting faster carbonation rates.

The application of machine-learning enabled the analysis of co-effects, facilitating the identification of optimized experimental conditions. The case study suggested that the model is generally reliable and sensitive to cement type.

However, this work also had some limitations; the model was not that accurate when applied to unseen datasets. So, the model can be further improved through hyperparameter optimization or exploring a new model. Also, the dataset used was relatively small compared to the scale required for practical machine-learning applications. To address these, we will enhance our model and expand the dataset to improve generalizability.

Methods

Data collection and processing

The data was collected from existing databases [CarboDB²⁷](#). It is an open-access repository specializing in concrete carbonation data. Compared to the limited scope of experimental results shown in Table 1, this large-scale dataset allows for a broader and more generalized analysis. This dataset was chosen because it is currently the largest available, including a wide range of experimental results from previous studies. It contained extensive multi-factor combination relationships. A total of 1619 data entries were gathered. The dataset includes detailed information on various factors influencing concrete carbonation, such as cement type, RH, temperature, CO₂ concentration, w/b ratio, carbonation time, and so on. Prior to the analysis, the data underwent a series of processing procedures to ensure quality and consistency. This involved:

- **Features selection:** There were various features of the raw data, but we only focused on a limited type of features: 'Time/d', 'Cement Type', 'Cement Strength Class (the compression strength at a specific age)', 'Cement Strength Development (the rate at which the compressive strength of cement develops over time)', 'Cement Amount/%', 'Addition Type', 'w/b ratio', 'Carbonation Type', 'CO₂/%', 'RH', 'Temperature', 'Depth average [mm]'. These selected features represent a mix of material properties and environmental factors, ensuring that the analysis captures both the characteristics of cement and the external conditions influencing carbonation.
- **Handling missing data:** for 'cement amount/%', the missing data was set as 100, assuming that if the cement amount is not mentioned, no other materials have been added to replace cement. for 'addition type',

the missing data was set as 'n'. for rh and temperature, the missing data were set as the common values 65 and 20. since these are commonly used in practice.

- **Delete data:** Removal of data that missing the carbonation depth value.
- **Encoding Categorical Variables:** Categorical variables, such as cement type, were encoded, and the mappings were stored. Categorical variables were encoded into numeric values. It assigned an integer code to each category. After encoding was complete, these numeric values were used for machine-learning algorithms. The mappings stored during the encoding process can be referenced later.

The processed dataset thus represents a substantial and reliable foundation for the subsequent machine-learning analysis.

Machine-learning modeling and interpretation

In essence, the various methods used to estimate carbonation depth aim to construct a function f that when provided with a set of factors and their corresponding values as inputs (e.g., carbonation time, RH), generates an estimate of the carbonation depth d . Representing the factor values as $v_1, v_2, \dots, v_m$, the function can be expressed as:

$$d = f(v_1, v_2, \dots, v_m). \quad (1)$$

The carbonation process is influenced by a range of factors, as mentioned earlier, with carbonation time standing out as particularly significant.²⁸ Traditionally, in accordance with Fick's first law of diffusion, the carbonation depth is commonly assumed to adhere to the Eq. (2)²⁹.

$$d = K\sqrt{t} \quad (2)$$

where d represents the carbonation depth in millimeters, while K signifies the carbonation coefficient in millimeters per square root of the year. The parameter K is anticipated to exhibit a nuanced relationship with a series of factors. To derive a specific K , a regression is performed on the observed d at given $\sqrt{t}$. For each set of factor values, an individual regression is necessary to determine an independent K , resulting in a lookup table as shown in Table 3, where a carbonation depth can be obtained by looking up the experimental results, and using them to derive a specific K .

It is crucial to emphasize that in conventional methodologies relying on Eq. (2), the actual values of the influencing factors are not explicitly utilized in the regression procedure. The function f acquired through traditional techniques essentially serves as a lookup function derived from empirical findings in the table, as opposed to a computational model. The drawback of this non-computational function is the absence of a modeled multivariate distribution of the factors. Consequently, predicting the carbonation depth d for factor values not in the table requires additional experiments to gather data for regression. This approach is not only impractical but also costly.

The reason why traditional methods cannot establish a computational model lies primarily in the fact that the regression techniques require prior knowledge of the expression of the regression function, which is typically very challenging, especially when dealing with a large number of factors. Therefore, approaches similar to Eq. (2) adopt a simple linear assumption to construct the expression, while avoiding the issue of involving factor values in the calculations. This was an inevitable choice in the early days when computational power was limited. Machine-learning methods are ideal for addressing these challenges.

Let's define the input as a vector $X = [v_1, v_2, \dots, v_n]^T$ and the output as $Y = d$. In machine learning, a complex function can be learned (i.e., $Y = f(X)$) without knowing its exact expression, given that input-output pairs (i.e., (X, Y)) are available. The learning process begins by initializing a random model and using it to compute an estimation $f'(X)$. Subsequently, this result is compared to the expected output Y . A loss is then computed to gauge how far the estimation deviates from the expected value. In this study, Mean Square Error (MSE) and Root Mean Square Error (RMSE) will be utilized to calculate the loss as follows:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (3)$$

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

where y_i is the actual value, $\hat{y}_i$ is the predicted value, and n is the number of observations.

Table 3 | Traditional lookup table

v_1 (Carbonation time/d)	v_2 (RH/%)	...	v_n (Cement Type)	d (Carbonation depth/mm)
56	65	...	CEM I	2.0
...	...	...	...	...
280	70	...	CEM III/B	7.3

The initial model is adjusted iteratively to minimize this loss. This iterative process continues until the minimum loss is achieved.

The procedure is generally similar to the regression process, with a major difference in the fact that the explicit form of f is not necessarily required. Taking the Decision Tree (DT) methods as an example²⁰, these methods assume the function is a decision tree (Fig. 9) in which a final optimal estimation is reached after evaluating the input values in an orderly manner. The DT algorithm assesses loss and refines the tree's structure to optimize both the ranges of factor values and the sequence in which factors are evaluated, aiming to make decisions that minimize losses. A notable advantage is that an explicit function f (i.e., structure for the decision tree) is not necessarily required beforehand; the DT algorithm can autonomously determine it. Moreover, the utilization of decision trees provides an avenue to capture non-linear relationships among variables.

Another prevalent option known as Random Forest (RF) is constructed based on decision trees²¹. The concept involves training numerous decision trees concurrently, consolidating and contrasting their decisions to arrive at a more holistic decision. The algorithm commences by dividing the data into multiple subsets, with each subset used to train a decision tree (Fig. 10). Ultimately, the collective insights of the trees are leveraged to make the final decision, akin to tapping into the collective wisdom of a crowd.

XGBoost is another widely used technique²². In contrast to Random Forest, which simultaneously learns decision trees, XGBoost sequentially trains a series of trees (Fig. 10). The aim is for each new tree to rectify errors made by its predecessors. This is achieved by assigning a weight to each dataset, regulating its influence on the learning process. These weights are adjusted at each iteration: increased if the prediction deviates significantly from the expected value (indicating a need for more focus in the next iteration) or decreased otherwise. The progressive integration of decision trees in a sequential manner leads to a continuous enhancement in performance. XGBoost has a series of outstanding features that enable its high-performance behavior. It can handle complex relationships through the use of decision trees. XGBoost addresses the overfitting issue through regularization (L1 and L2), which penalizes overly complex trees. As for the predictive performance, XGBoost's sequential tree-building process, regularization, and iterative error correction improve its performance. For all models, default hyperparameters were applied. However, we did attempt to optimize the hyperparameters of the XGBoost model, and the results showed an improvement in performance. Since hyperparameter optimization was not the focus of this study, we plan to explore it further in future research.

Traditional ways of interpreting the effects of a single-factor typically rely on linear regression model²⁶. While it offers an understanding of individual effects, it has the limitation of capturing the interactions between multiple factors. Also, the traditional single-factor regression struggles to give an insight into how different factors contribute to a particular outcome. To solve such a limitation of the traditional interpretation method, we adopted SHAP analysis³⁰ for a more comprehensive interpretation of our

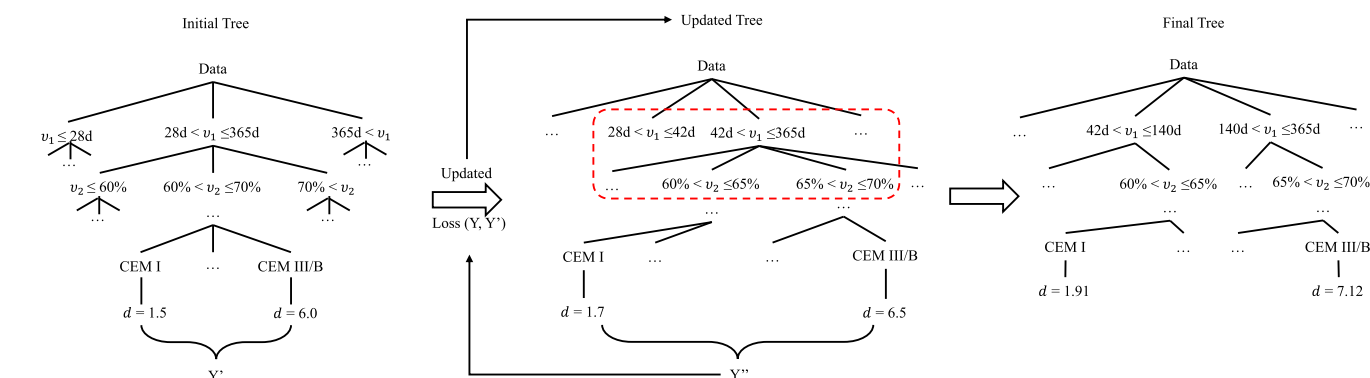


Fig. 9 | Mechanism of DT model, there is an initial tree and the data are input and go through each branch, the prediction carbonation depth Y' will initially be obtained. By comparing the loss between the real combination depth Y and Y' , the

decision tree will be further updated, e.g., by increasing the branches as indicated in the red frame. Finally, a final tree is supposed to get an acceptable prediction value.

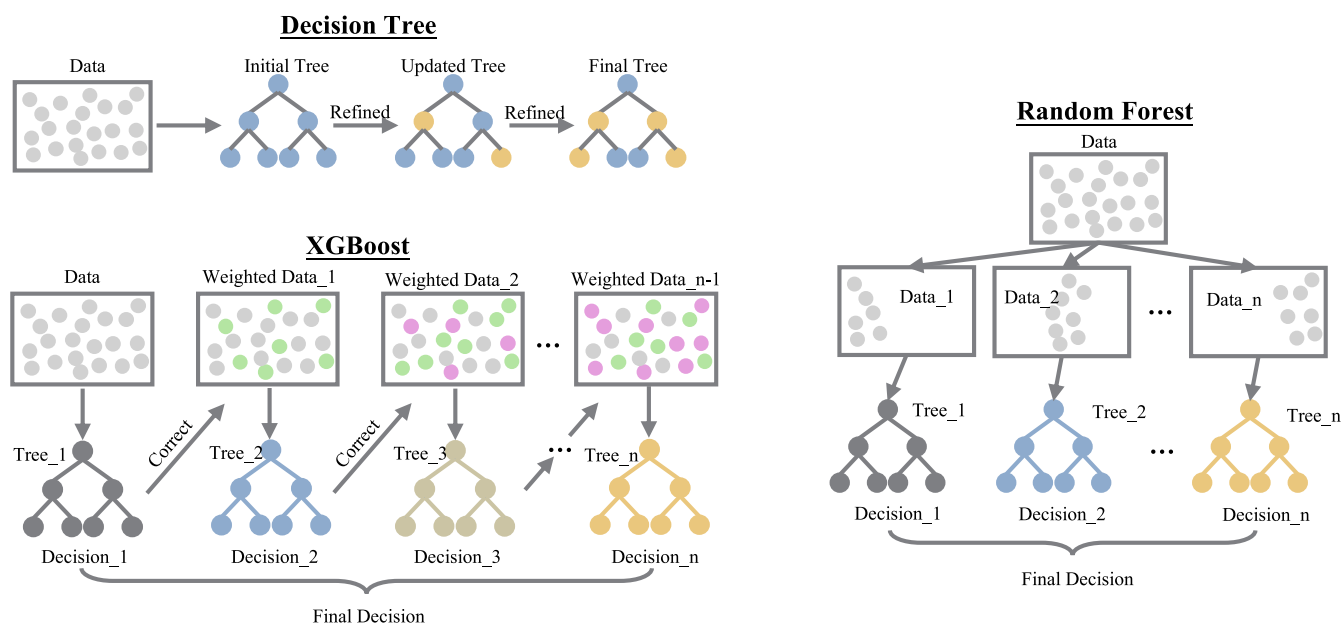


Fig. 10 | Illustration of three models.

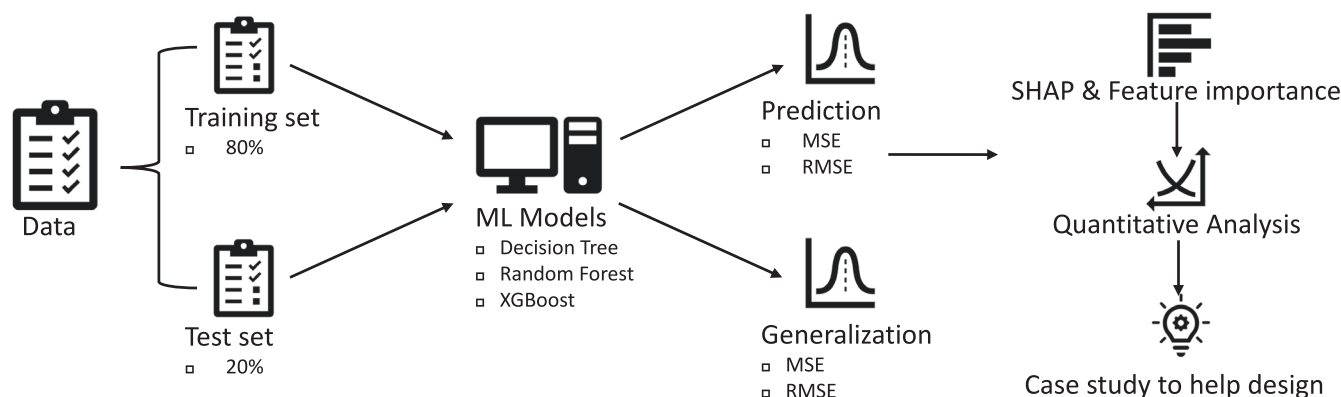


Fig. 11 | Framework of conceptual model.

machine-learning models. SHAP is an approach that assigns each factor a value to identify its contribution to the final prediction. This interpretation method allows us to analyze the effect of multiple factors in a visible way. In combination with machine-learning models in this study, a comprehensive framework for addressing the limitations of traditional empirical methods is pointed out. It enables us to include all relevant factors in the analysis.

Overall, the workflow for studying the carbonation behavior by machine learning is shown in Fig. 11. The process begins with a data split, 80% of the data is set as training set, the rest as test set. The data is then used to train three machine-learning models, as mentioned before. The loss is then calculated to evaluate the accuracy of prediction and the generalization. Besides the prediction performance, SHAP is employed to interpret the feature importance. After that, the interpretation will be adopted to help a quantitative analysis. Finally, case studies are given to indicate how this work helps material design.

Data availability

The dataset used in this study can be accessed from CarboDB at <https://carbodb.bgu.tum.de/#/>.

Code availability

Code will be provided upon request.

Received: 11 November 2024; Accepted: 10 March 2025;
Published online: 04 April 2025

References

- Chen, S., Liu, J., Zhang, Q., Teng, F. & McLellan, B. C. A critical review on deployment planning and risk analysis of carbon capture, utilization, and storage (CCUS) toward carbon neutrality. *Renew. Sustain. Energy Rev.* **167**, 112537 (2022).
- Hills, T., Leeson, D., Florin, N. & Fennell, P. Carbon capture in the cement industry: technologies, progress, and retrofitting. *Environ. Sci. Technol.* **50**, 368–377 (2016).
- Walker, I., Bell, R. & Rippey, K. Mineralization of alkaline waste for ccus. *npj Mater. Sustain.* **2**, 28 (2024).
- Jin, F., Zhao, M., Xu, M. & Mo, L. Maximising the benefits of calcium carbonate in sustainable cements: opportunities and challenges associated with alkaline waste carbonation. *npj Mater. Sustain.* **2**, 1 (2024).
- Ramirez-Corredores, M. M. Sustainable production of CO₂-derived materials. *npj Mater. Sustain.* **2**, 35 (2024).
- Liu, Z. et al. Carbonation of blast furnace slag concrete at different CO₂ concentrations: Carbonation rate, phase assemblage, microstructure and thermodynamic modelling. *Cem. Concr. Res.* **169**, 107161 (2023).

7. Xu, Z. et al. Effects of temperature, humidity and CO_2 concentration on carbonation of cement-based materials: a review. *Constr. Build. Mater.* **346**, 128399 (2022).
8. Mehdizadeh, H., Jia, X., Mo, K. H. & Ling, T.-C. Effect of water-to-cement ratio induced hydration on the accelerated carbonation of cement pastes. *Environ. Pollut.* **280**, 116914 (2021).
9. Sahoo, P., Rao, N., Jain, S. K. & Gupta, S. Carbon sequestration in earth-based alkali-activated mortar: phase changes and performance after natural exposure. *npj Mater. Sustain.* **2**, 34 (2024).
10. Leemann, A. & Moro, F. Carbonation of concrete: the role of CO_2 concentration, relative humidity and CO_2 buffer capacity. *Mater. Struct.* **50**, 1–14 (2017).
11. Liu, P., Yu, Z. & Chen, Y. Carbonation depth model and carbonated acceleration rate of concrete under different environment. *Cem. Concr. Compos.* **114**, 103736 (2020).
12. Parhi, S. K. & Patro, S. K. Prediction of compressive strength of geopolymer concrete using a hybrid ensemble of grey wolf optimized machine learning estimators. *J. Build. Eng.* **71**, 106521 (2023).
13. Parhi, S. K., Nanda, A. & Panigrahi, S. K. Multi-objective optimization and prediction of strength along with durability in acid-resistant self-compacting alkali-activated concrete. *Constr. Build. Mater.* **456**, 139235 (2024).
14. Parhi, S. K. & Patro, S. K. Compressive strength prediction of pet fiber-reinforced concrete using dolphin echolocation optimized decision tree-based machine learning algorithms. *Asian J. Civ. Eng.* **25**, 977–996 (2024).
15. Parhi, S. K. & Panigrahi, S. K. Alkali-silica reaction expansion prediction in concrete using hybrid metaheuristic optimized machine learning algorithms. *Asian J. Civ. Eng.* **25**, 1091–1113 (2024).
16. Parhi, S. K. & Patro, S. K. Parametric analysis and prediction of geopolymerization process. *Mater. Today Commun.* **41**, 111047 (2024).
17. Peng, Y. & Unluer, C. Interpretable machine learning-based analysis of hydration and carbonation of carbonated reactive magnesia cement mixes. *J. Clean. Prod.* **434**, 140054 (2024).
18. Tran, V. Q., Mai, H.-V. T., To, Q. T. & Nguyen, M. H. Machine learning approach in investigating carbonation depth of concrete containing fly ash. *Struct. Concr.* **24**, 2145–2169 (2023).
19. Ehsani, M. et al. Machine learning for predicting concrete carbonation depth: A comparative analysis and a novel feature selection. *Constr. Build. Mater.* **417**, 135331 (2024).
20. Wu, X. et al. Top 10 algorithms in data mining. *Knowl. Inf. Syst.* **14**, 1–37 (2008).
21. Ho, T. K. Random decision forests. In: *Proceedings of 3rd International Conference on Document Analysis and Recognition*. Vol. 1, 278–282 (IEEE, 1995).
22. 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*, Association for Computing Machinery, 785–794 (2016).
23. Mi, J.-X., Li, A.-D. & Zhou, L.-F. Review study of interpretation methods for future interpretable machine learning. *IEEE Access* **8**, 191969–191985 (2020).
24. He, B. et al. Interpretation and prediction of the CO_2 sequestration of steel slag by machine learning. *Environ. Sci. Technol.* **57**, 17940–17949 (2023).
25. Xi, F. et al. Substantial global carbon uptake by cement carbonation. *Nat. Geosci.* **9**, 880–883 (2016).
26. Elsalamawy, M., Mohamed, A. R. & Kamal, E. M. The role of relative humidity and cement type on carbonation resistance of concrete. *Alex. Eng. J.* **58**, 1257–1264 (2019).
27. Thiel, C., Haynack, A., Geyer, S., Braun, A. & Gehlen, C. CarboDB—Open Access Database for Concrete Carbonation. In: *Proceedings of the 3rd RILEM Spring Convention and Conference*, Springer International Publishing, 1, 79–90 (2020).
28. Qiu, Q. A state-of-the-art review on the carbonation process in cementitious materials: Fundamentals and characterization techniques. *Constr. Build. Mater.* **247**, 118503 (2020).
29. Yoon, I.-S., Çopuroğlu, O. & Park, K.-B. Effect of global climatic change on carbonation progress of concrete. *Atmos. Environ.* **41**, 7274–7285 (2007).
30. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **30**, 4768–4777 (2017).
31. Sanjuán, M., Andrade, C. & Cheyrezy, M. Concrete carbonation tests in natural and accelerated conditions. *Adv. Cem. Res.* **15**, 171–180 (2003).
32. Xuan, D., Zhan, B. & Poon, C. S. A maturity approach to estimate compressive strength development of CO_2 -cured concrete blocks. *Cem. Concr. Compos.* **85**, 153–160 (2018).
33. Wang, D., Noguchi, T. & Nozaki, T. Increasing efficiency of carbon dioxide sequestration through high temperature carbonation of cement-based materials. *J. Clean. Prod.* **238**, 117980 (2019).
34. Ho, D. & Lewis, R. Carbonation of concrete and its prediction. *Cem. Concr. Res.* **17**, 489–504 (1987).
35. Lo, Y. & Lee, H. Curing effects on carbonation of concrete using a phenolphthalein indicator and Fourier-transform infrared spectroscopy. *Build. Environ.* **37**, 507–514 (2002).

Acknowledgements

The work described in this paper was supported by The Hong Kong Polytechnic University. We would like to thank Hong Mengze for providing professional advice for improving this paper.

Author contributions

S.Y. pointed out the idea, Z.C. and W.X. provided suggestions about machine-learning models, and S.P. and P.C.S. provided suggestions about cementitious material carbonation. W.Y. and J.H. helped collect the data, and together with S.Y., they worked on the data processing. S.Y., Z.C., and W.X. wrote the main manuscript text and the model code. Y.H. helped revise the manuscript. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Xiao-Yong WEI.

Reprints and permissions information is available at <http://www.nature.com/reprints>

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

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

© The Author(s) 2025