International Journal of Concrete Structures and Materials

RESEARCH Open Access

Data-Driven Insights into Controlling the Reactivity of Supplementary Cementitious Materials in Hydrated Cement

Aron Berhanu Degefa[1], Geonyeol Jeon[2], Sooyung Choi[3], JinYeong Bak[3], Seunghee Park[4], Hyungchul Yoon[2] and Solmoi Park[1*]

Abstract

Supplementary cementitious materials (SCMs) play an essential role in sustainable construction due to their potential to reduce carbon emissions, promote circular economy principles, and enhance the properties of concrete. However, the inherent diversity of SCMs makes it challenging to predict their degree of reaction (DOR). This study applies machine learning techniques to predict DOR while exploring key parameters affecting it. Five machine learning models are utilized: linear regression, Gaussian process regression (GPR), decision tree regression, support vector machine and extreme gradient boosting, with GPR providing the most accurate and adaptable prediction. The study delves into the impact of various parameters on DOR, revealing their significance. Silica content emerges as the most critical, followed by particle size distribution, specific gravity, and water-to-cement (W/C) ratio. Optimizing DOR requires extending curing time, reducing particle size distribution, and considering optimal silica content and W/C ratio. This research emphasizes the importance of understanding the relationships between parameters and the DOR of SCMs, providing insights to enhance the efficiency of SCMs in cementitious systems through machine learning and data-driven analysis.

Keywords Supplementary cementitious materials, Degree of reaction, Machine learning, Prediction

1 Introduction

Sustainable construction has emerged as a pressing concern in recent years due to the heightened scrutiny of the environmental consequences of the construction sector. One solution that can ensure sustainability is the use of supplementary cementitious materials (SCMs). SCMs are binding materials used to partially replace Portland cement in the production of concrete and other cement-based products. Their usage plays a vital role in sustainable construction since they can considerably impact resource consumption and carbon emissions (Rahla et al., 2019; Samad & Shah, 2017). In addition, SCMs help to promote a circular economy by ensuring the recycling of waste materials that would otherwise end up in landfills (Diaz-Loya et al., 2019). Furthermore, incorporating SCMs can enhance the properties and performance of cementitious matrices, such as increasing their strength, durability, and workability (Gupta & Chaudhary, 2022; Juenger & Siddique, 2015; Juenger et al., 2019). However, attaining a consistent and predictable degree of reaction (DOR) may be difficult owing to the complex and variable nature of SCMs (Scrivener et al., 2015). Factors such

Journal information: ISSN 1976-0485 / eISSN 2234-1315.

*Correspondence:
Solmoi Park
solmoi.park@pknu.ac.kr
[1] Department of Civil Engineering, Pukyong National University, Busan 48513, Korea
[2] School of Civil Engineering, Chungbuk National University, Cheongju 28644, Korea
[3] Department of Artificial Intelligence, Sungkyunkwan University, Suwon 16419, Korea
[4] School of Civil, Architectural Engineering & Landscape Architecture, Sungkyunkwan University, Suwon 16419, Korea

as water-to-cement (W/C) ratio, chemical composition, physical characteristics, and curing conditions can significantly impact the reactivity of SCMs, leading to variations in the properties and performance of cementitious matrices (Pacewska & Wilińska, 2020; Skibsted & Snellings, 2019). Therefore, comprehending the main properties of SCMs in hydrated cement is imperative in material selection and performance optimization within cementitious matrices.

SCMs are added to concrete to increase their performance and sustainability. Determining the DOR of SCMs is essential in foreseeing the characteristics and performance of the final products. Several key properties of SCMs influence their extent of dissolution in concrete. The properties of SCMs, such as particle size, specific surface area, amount of SCMs being used to replace Portland cement, chemical composition, W/C ratio, curing temperature, and curing time, can significantly impact their performance in cementitious matrices. For instance, the particle size and specific surface area of SCMs can affect the rate and extent of their reaction with cementitious materials. Smaller particle sizes and larger specific surface areas increase the contact area between the SCM and the cementitious materials, leading to a more significant DOR (Hallet et al., 2020; Mirzahosseini & Riding, 2015; Ndahirwa et al., 2022; Sanjuán et al., 2015). The chemical composition of SCMs also affects their compatibility with other materials in the matrix (Sabir et al., 2001; Sanjuán et al., 2015; Tironi et al., 2013). Suraneni et. al. (2019) reported that SCMs with high silica, alumina, and calcium exhibit distinct characteristics regarding their utilization of calcium hydroxide and the amount of heat released. Additionally, the W/C ratio and curing conditions of the system can influence the DOR (Phung et al., 2021). Generally, a higher W/C ratio increases DOR (Escalante et al., 2001; Snellings et al., 2022). However, surpassing a certain level of W/C ratio can result in a more dilute cementitious system with increased particle distance, reducing the DOR (Navarrete et al., 2020). Finally, proper curing conditions, such as temperature and humidity, can improve the DOR by providing the necessary environment for chemical reactions to occur (de Azevedo Basto et al., 2022; Lothenbach et al., 2011; Snellings et al., 2022). The formation and development of hydration products within the cement matrix are influenced by its water content. As hydration progresses, the consumption of water decreases the internal relative humidity of the cement matrix, causing capillary pressure and shrinkage. Consequently, when the relative humidity is low, incomplete hydration may occur, leading to a lower DOR (Skibsted & Snellings, 2019).

Improving the efficiency of cementitious systems hinges on a profound understanding of their properties. Consequently, gaining insights through data-driven analysis becomes crucial, particularly in comprehending the fundamental properties of SCMs that influence the DOR. Leveraging large datasets from diverse sources offers the opportunity to uncover correlations between key SCM properties and their performance within cementitious matrices. Harnessing the power of machine learning (ML) methods further allows for thorough examination and comprehension of extensive data sets, thereby providing deeper insights into the underlying correlations between crucial SCM features and their performance in cementitious matrices. Previous research has already demonstrated the efficient utilization of ML models for parametric investigations, enabling accurate estimations of primary material properties that impact carbonation and compressive strength (Abuodeh et al., 2020; Chen et al., 2022). By adopting this approach, the predictability of SCM influence in hydrated Portland cement can be significantly enhanced by focusing on the major SCM properties affecting the DOR, ultimately resulting in an optimized model.

While several investigations on the DOR of SCMs have been conducted by employing microstructural analysis techniques such as X-ray diffraction (XRD) (Durdziński et al., 2017), scanning electron microscopy (SEM) (Pfingsten et al., 2018), and nuclear magnetic resonance (NMR) (Walkley & Provis, 2019), as well as different testing methods (i.e., selective dissolution (Kocaba et al., 2012) and modified R3 test (Ramanathan et al., 2022)), the existing research still has significant limitations. One of the primary concerns is the substantial variability in DOR observed due to factors such as the source and production process of the SCMs, necessitating independent investigation for reliable conclusions (Ndahirwa et al., 2022). Moreover, the lack of consensus on an appropriate testing method and a standard for evaluating the DOR of SCMs makes it challenging to compare results across studies (Durdziński et al., 2017; Li et al., 2018). Additionally, previous studies predominantly focused on specific characteristics of SCMs, such as their pozzolanic activity (Donatello et al., 2010; Snellings & Scrivener, 2016) or ability to improve concrete durability (Anurag et al., 2021; Ndahirwa et al., 2022), without a comprehensive assessment of their overall reactivity. Thus, further study is essential to highlight the DOR of SCMs in different contexts and develop robust methodologies for evaluating their performance.

This study aimed to identify the essential parameters that affect the DOR of SCMs. Accordingly, five ML methods were employed for predicting the DOR: linear regression, Gaussian process regression (GPR), decision tree (DT) regression, support vector machine (SVM) and extreme gradient boosting (XGBoost). The performance of each model was evaluated using various statistical methodologies. Subsequently, the most accurate, adaptable ML model was selected for a parametric investigation encompassing 22 parameters. The influence of input parameters was studied using the Shapley value. Moreover, the fundamental parameters were set and analyzed with existing theories to determine their potential effect on the DOR. These findings offer valuable insights for optimizing the DOR of SCMs in various applications.

2 Methods

2.1 Data Collection and Description

The experimental data collected to study the DOR in hydrated Portland cement considered various types of binders, encompassing a range of SCMs such as slag, fly ash, metakaolin, limestone, calcium sulfoaluminate cement, silica fumes, magnesia-based cement, glass powder, calcined clay, calcium aluminate cement, and rice husk ash. Several factors affect the DOR, such as the W/C ratio, the oxide composition and proportions of Portland cement and SCMs, curing time and temperature, and physical properties such as particle size distribution, surface area, and specific gravity.

The dataset comprised 247 examples, with 22 input features and DOR as the output. Table 1 summarizes the statistical analysis of these inputs, detailing the units, minimum, maximum, average values, and standard deviations, which offers an essential insight into the range and variability of the features. Moreover, the accompanying histograms aim to illustrate the distribution density of each input, providing a preliminary, straightforward overview of the characteristics of the data. The full dataset is provided as a supplementary material for reference (Additional file 1).

However, the data collected for median particle size diameter (Dv50) were insufficient. To address this limitation, four techniques were employed to compensate for the missing Dv50 values, as described in Table 2. These techniques allowed for a thorough analysis of the impact of missing Dv50 values in machine-learning models.

2.2 Machine Learning Algorithms

Predicting the DOR of SCMs is crucial for optimizing their use in various applications. A range of advanced ML algorithms were utilized to achieve this, including GPR, linear regression, DT, SVM and XGBoost. Before the model development, the collected dataset was randomly divided into training and test groups at a ratio of 80:20 to ensure the robustness and validity of the models. Comprehensive descriptions of each of the ML models employed in this study are provided in the following subsections.

2.2.1 Linear Regression

Linear regression is a statistical analysis tool that is used to describe the relationship between one or more independent variables and a dependent variable. The best-fit line or hyperplane representing the relationship between these variables is determined through linear regression. Equation (1) shows the general formula for linear regression models.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon, \tag{1}$$

where Y is the dependent variable, $\mathrm{x_n}$ values are independent variables, β_n is the regression coefficient, and ε denotes an error (Chou et al., 2014).

2.2.2 Gaussian Process Regression

A Gaussian process refers to a collection of random variables whose joint distribution follows a Gaussian or normal distribution, such that any finite subset of the variables has a joint distribution that is also Gaussian. This is a stochastic process with vector-defined mean and covariance functions expressed as a matrix, as indicated in Eq. (2).

$$f(x) \sim GP\left(\mu(x), k\left(x, x'\right)\right), \tag{2}$$

where f(x) represents the output variable, μ(x) represents the mean function, $\mathrm{k}\left(\mathrm{x},\mathrm{x}'\right)$ represents the covariance function, and GP represents the Gaussian Process (Rasmussen, 2003; Shi & Choi, 2011).

2.2.3 Decision Tree Regression

DT regression is a supervised learning approach that learns basic decision rules based on data characteristics to predict the value of a continuous target variable (Charbuty & Abdulazeez, 2021). This method constructs a tree-structured model with a root node, branches, internal nodes, and leaf nodes. The root node contains the entire dataset and has no incoming branches. The internal nodes reflect the characteristics of the data set, while the branches represent the decision criteria. The leaf nodes reflect the various outcomes of the target variable (Song & Lu, 2015). The method recursively partitions the data into subsets based on their characteristics until the subsets are more homogenous with regard to the target variable. The model then predicts the target variable by averaging the values of the training data in each leaf node of the tree (Pal & Mather, 2001).

2.2.4 Support Vector Machine

SVM is a machine-learning approach that can learn from data and produce predictions based on that data. It can handle regression and classification problems

Table 1 Statistical parameters of the dataset

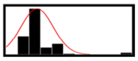
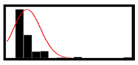
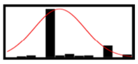
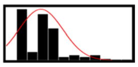
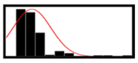
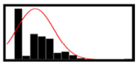
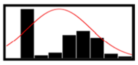
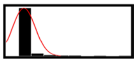
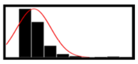
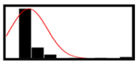
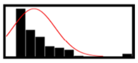
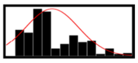
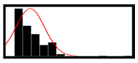
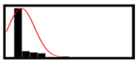
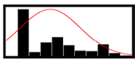
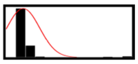
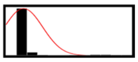
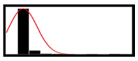
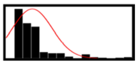
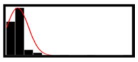
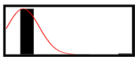
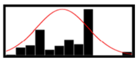
Feature	Units	Histogram	Min.	Max.	Mean	Standard deviation
Water-to-cement ratio	–		0.2	2	0.54	0.28
Curing time	Days		57	1310	140.33	162.73
Curing temperature	°C		5	60	27.28	13.15
SiO_2 content in PC	wt%		0	51.89	11.58	10.37
Al_2O_3 content in PC	wt%		0	22.93	3.3	3.91
Fe_2O_3 content in PC	wt%		0	7.29	1.33	1.17
CaO content in PC	wt%		0	74.82	27.82	22.63
MgO content in PC	wt%		0	40	2.09	4.48
SO_3 content in PC	wt%		0	8.26	1.2	1.35
Na_2O content in PC	wt%		0	1.71	0.16	0.28
K_2O content in PC	wt%		0	2.36	0.37	0.44
SiO_2 content in SCMs	wt%		0	62.53	20.54	14.81
Al_2O_3 content in SCMs	wt%		0	67.68	9.35	8.62
Fe_2O_3 content in SCMs	wt%		0	29.8	1.86	3.36
CaO content in SCMs	wt%		0	45.81	14.31	13.19
MgO content in SCMs	wt%		0	57.3	3.84	8.44
SO_3 content in SCMs	wt%		0	20.82	1.34	3.62
Na_2O content in SCMs	wt%		0	9.97	0.48	1.21
K_2O content in SCMs	wt%		0	2.67	0.43	0.47
Dv50	µm		1.5	130	12.68	11.93
Surface area	m^2/gm		0.3	395	11.29	57.09
Specific gravity	g/cm^3		2.17	3.3	2.68	0.26

Table 1 (continued)

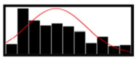
Feature	Units	Histogram	Min.	Max.	Mean	Standard deviation
DOR	%		3	100	44.04	26.42

Table 2 Approaches for representing Dv50

Approach	Description
1	Missing values were predicted based on general material characteristics
2	GPR machine learning was used to predict missing values using other available data
3	Predictions were made solely for samples with available Dv50 data
4	Predictions were made without considering the effect of Dv50 to explore the potential impact of missing data on the models

by determining the optimum function to match the data while minimizing errors. The function is often a linear combination of the input variables, but it may alternatively be a nonlinear transformation based on a kernel function. The kernel function enables the SVM to translate the input into a higher-dimensional space where a linear separator may be found. A linear separator is a hyperplane that splits data into two or more classes with the largest margin attainable. The margin is the distance between the hyperplane and the nearest data points, known as support vectors. The shape and location of the hyperplane are determined by the support vectors (Gholami & Fakhari, 2017; Noble, 2006).

2.2.5 eXtreme Gradient Boosting

XGBoost is a scalable end-to-end tree-boosting system, which is effective for both regression and classification tasks (Chen & Guestrin, 2016). Rooted in the Gradient Boosting (Friedman, 2001) framework, an ensemble learning method, XGBoost harnesses the collective wisdom of multiple weak learners, often represented as simple decision trees, to enhance predictive accuracy. Its iterative approach involves training weak models sequentially, with each subsequent model dedicated to correcting the errors of its forerunners. Beyond its core functionality, XGBoost offers a range of essential features, including integrated regularization for guarding against overfitting, robust handling of missing data, streamlined parallel processing for efficient computation, customizable objective functions to adapt to specific use cases, and the incorporation of tree pruning techniques for fine-tuning model complexity control.

2.3 Evaluation Method

Three independent statistical measures were used to assess the efficiency of the ML models: the root mean square error (RMSE), the mean absolute error (MAE), and the coefficient of determination (R^2). These indicators were used to assess and compare the accuracy and reliability of the performance of the models. Employing these three separate statistical measures provides an advantage in obtaining a fair estimation of accuracy. For instance, RMSE is sensitive to outliners and penalizes large errors, thus facilitating their removal from the dataset (Chai & Draxler, 2014). On the other hand, MAE is more suitable for datasets containing outliers (Willmott & Matsuura, 2005), and R^2 offers more information without being subject to the interpretability limitations of RMSE and MAE (Chicco et al., 2021; Zhang, 2017). The RMSE, MAE, and R^2 equations are expressed below in Eqs. 3, 4, and 5, respectively.

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(X_i - Y_i)^2}, \tag{3}$$

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|X_i - Y_i|, \tag{4}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(X_i - Y_i)^2}{\sum_{i=1}^{m}(Y_m - Y_i)^2}. \tag{5}$$

For Eqs. (3), (4) and (5), X_i is the predicted value, Y_i is the actual value and Y_m is the mean value.

K-fold cross-validation was applied to overcome the problem of overfitting. It involves splitting the dataset into K subsets or "folds" of roughly similar size. The model is then trained and assessed K times, with each fold acting as the validation set once and the remaining folds serving as training folds. This technique contributes to a more robust estimation of the performance of the model by minimizing reliance on a single train-test split (Hastie et al., 2009). Considering the limited size of the dataset, K has been set at 5 to find a balance between evaluating model performance and utilizing the available data.

2.4 Feature Selection

ML predictions can often suffer from the limitation of not being able to recognize the effects of input parameters on the outcome. However, understanding these relationships is crucial as they provide valuable insights into the roles of the input parameters and serve as a foundation for future predictions. In this study, the ML model with the highest prediction performance was utilized, and the order of significance of input features on the desired outcome, which was the DOR, was determined using the Shapley value. The Shapley value is an idea developed from cooperative game theory that allocates a fair distribution of total costs across players of the game (Merrick & Taly, 2020). In the context of ML, the Shapley value can be utilized to quantify the contribution of each feature to a prediction for a given instance. The Shapley value of a parameter is the weighted mean of the marginal contributions of the feature, averaged across all possible feature subsets. The marginal contribution is the difference between the prediction with and without the feature (Cohen et al., 2005).

3 Results and Discussion

The accuracy of ML models for DOR predictions was assessed and summarized in Table 3. It is worth noting that XGBoost produced the most accurate results. However, compared to actual findings, this model produced very inconsistent outputs, which may be attributed to the structure of the dataset. The dataset had a limited number of observations compared to the independent variables. For XGBoost, this imbalance might lead to subpar model performance, possibly stemming from reduced generalization, computational intensity, and overfitting (Barnwal et al., 2022; Ma et al., 2021). As a result, the subsequent analysis utilized the next most accurate model: GPR. GPR demonstrated comparable accuracy to XGBoost while generating interpretable results. Specifically, GPR exhibited outstanding performance with an RMSE of 12.46, an MAE of 8.88, and an R^2 value of 0.79. In contrast, linear regression yielded less favorable results, showcasing an RMSE of 20.24, an MAE of 15.12, and an R^2 value of 0.42.

Table 3 Performance of ML predictions

ML model	Dv50 representation techniques[a]	RMSE	MAE	R^2
Linear regression	1	20.24	15.12	0.42
	2	19.51	14.8	0.46
	3	17.71	7.98	0.49
	4	19.13	14.21	0.48
GPR	1	12.46	8.88	0.79
	2	12.66	8.94	0.77
	3	12.78	5.62	0.74
	4	12.88	9.17	0.77
DT regression	1	17.07	10.91	0.6
	2	16.48	11.09	0.63
	3	16.63	6.74	0.57
	4	16.82	10.92	0.63
SVM	1	20.15	15.04	0.42
	2	18.48	13.76	0.51
	3	19.16	8.74	0.4
	4	12.69	8.99	0.78
XGBoost	1	12.99	8.72	0.73
	2	11.67	8.41	0.79
	3	12.37	8.95	0.72
	4	14.21	10	0.7

[a] The representation techniques for Dv50 are detailed in Table 2

The substantial improvement observed in the prediction accuracy of GPR can likely be attributed to the integration of complete data. Conversely, alternative models showcase superior performance when not considering Dv50 values (DT regression and SVM) or eliminating rows with null Dv50 values (Linear regression).

The experimental and modeled DOR using GPR is compared in Fig. 1. Following the identification of the optimal ML model for capturing the DOR, the significance of the features was further assessed using the Shapley value. Fig. 2 extensively elucidates the relative importance of each feature, providing profound insights into their criticality. Notably, the top five features, listed in order of significance, encompass curing time, SiO_2 content of the SCM, Dv50, specific gravity, and the W/C ratio. Subsequent subsections delve into the detailed descriptions of these features, providing a thorough understanding of their significance and implications.

3.1 Curing Conditions and Their Effect on SCMs Reactivity

Efficient control of the curing process is of paramount importance as it directly impacts the desired material properties and quality. Among the various factors that influence the curing process, curing time and temperature are widely recognized as crucial parameters. Gaining a comprehensive understanding of the relative

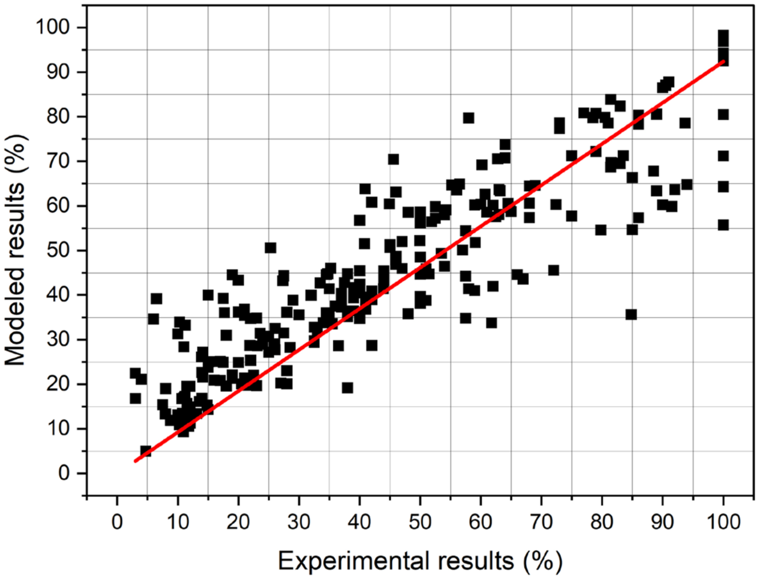


Fig. 1 DOR of modeled versus experimental findings (%). The symbols and lines indicate the experimental results and linear fit of the SCMs modeled results, respectively

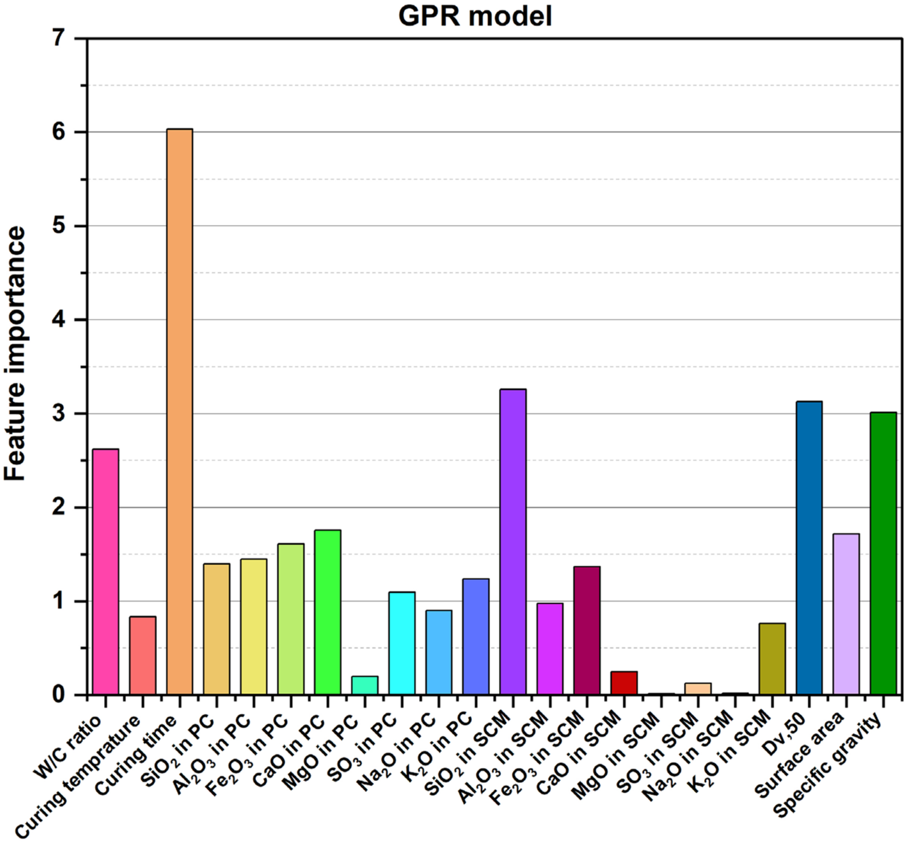


Fig. 2 Quantification of feature importance using Shapley value

significance of these factors can provide valuable insights for optimizing the curing process and enhancing its efficiency. According to the Shapley values, curing time accounts for most of the observed variation in the DOR.

It is important to note that the DOR of SCMs tends to increase as the curing time progresses (Haha et al., 2010; Kocaba et al., 2012). However, it is imperative to differentiate the effect of curing time on DOR from its effect on the rate of DOR. This distinction is essential because the rate of DOR is highly dependent on the type of SCM, exhibiting both increasing and decreasing trends (Skibsted & Snellings, 2019). The curing period of SCMs can be broadly categorized into three main stages to examine the fundamental relationships regarding DOR. Initially, during the early stage, the DOR of SCMs is relatively low as most of the available water is consumed by Portland cement (Skibsted & Snellings, 2019). At the same time, the filler aspects of the SCMs play a significant role (Lothenbach et al., 2011; Schöler et al., 2017). In the intermediate curing stage, the DOR of SCMs gradually increases due to improved water availability and pozzolanic reaction, which contributes significantly to the strength and durability of the concrete (Ahmed, 2019). Finally, the concrete reaches its maximum DOR in the long-term curing stage. However, since curing time is an inherent property of concrete that cannot be altered during the initial formulation of SCM in hydrated Portland cement, greater emphasis should be placed on optimizing other adjustable parameters. In contrast, the influence of curing temperature is relatively less pronounced, indicating that variations in temperature within the considered range do not significantly affect the DOR. While temperature variations can accelerate or retard early-age hydration and affect the stability of specific phases (de Azevedo Basto et al., 2022; Snellings et al., 2022), their impact is minor when compared to other parameters affecting DOR.

3.2 Chemical Composition of Cementitious Matrices and its Effect on SCMs Reactivity

The main oxide compositions of SCMs affecting the DOR in cementitious systems are silica, alumina, and calcium oxide, particularly when exploring viable replacements that can enhance or maintain performance. The presence of silica and alumina generally contributes to the formation of additional hydrates of the form C–A–S–H (Simonsen et al., 2020).

Silica holds high importance (i.e., ranked 2nd), primarily due to silica-based SCMs possessing a high specific surface area and fine particle size. The increased surface area offers more reaction sites, leading to a higher DOR. Silica-based SCMs have higher calcium hydroxide consumption than alumina- or calcium-based SCMs (Suraneni et al., 2019). Additionally, the C–S–H formed from excess silica exhibits a propensity for aluminum uptake, which occurs at the bridging sites within the silicate chains (Lothenbach et al., 2011). Fig. 3a illustrates the relationship between DOR and SiO_2 content of SCM. The DOR generally exhibits an upward trend as the SiO_2 content of SCM increases until it reaches an optimal replacement level. Beyond this point, as depicted in Fig. 3a, additional SiO_2 content becomes redundant, yielding no further changes.

3.3 Physical Properties of Cementitious Matrices and Their Effect on SCMs Reactivity

The physical properties investigated in this research included Dv50, surface area, and specific gravity. The Shapley values indicate that Dv50 and specific gravity possess higher significance. Dv50 measures the particle size distribution of a material, representing the size at which 50% of particle volumes are smaller than the given diameter (Arvaniti & De Belie, 2014). A smaller Dv50 (finer particle size distribution) can improve the DOR of SCMs by providing a larger surface area for interaction between the SCMs and the surrounding cementitious phases, such as calcium hydroxide. Additionally, finer particles facilitate more efficient diffusion and reactant adsorption, leading to improved DOR (Lothenbach et al., 2011; Skibsted & Snellings, 2019). These aspects are also supported by Fig. 3b, which illustrates the decreasing DOR trend with increasing particle size. Similarly, Liu et. al. (2018) demonstrated the increased hydration rate for lower Dv50 values.

In contrast, Fig. 3c shows a direct relationship between DOR and specific gravity. However, establishing a straightforward correlation between the two is challenging since multiple factors influence it. This complexity arises from the diverse physical and chemical transformations within the hydrated cement matrix. Therefore, while the modeled observations give vital insights into the collected dataset, it is crucial to remember that specific gravity can wield positive and negative effects.

3.4 Water-to-Cement Ratio and its Effect on SCMs Reactivity

The W/C ratio can influence the DOR of SCMs through various mechanisms. However, an optimal quantity of these materials must be added to leverage its benefits fully. Attaining an ideal W/C ratio is crucial for effective hydration. A lower ratio may lead to inadequate water supply for cementitious materials, diminishing their reactivity (Snoeck et al., 2014). Conversely, excessive water content can occupy space that should be filled by hydration products, resulting in adverse effects. Workability is another crucial consideration.

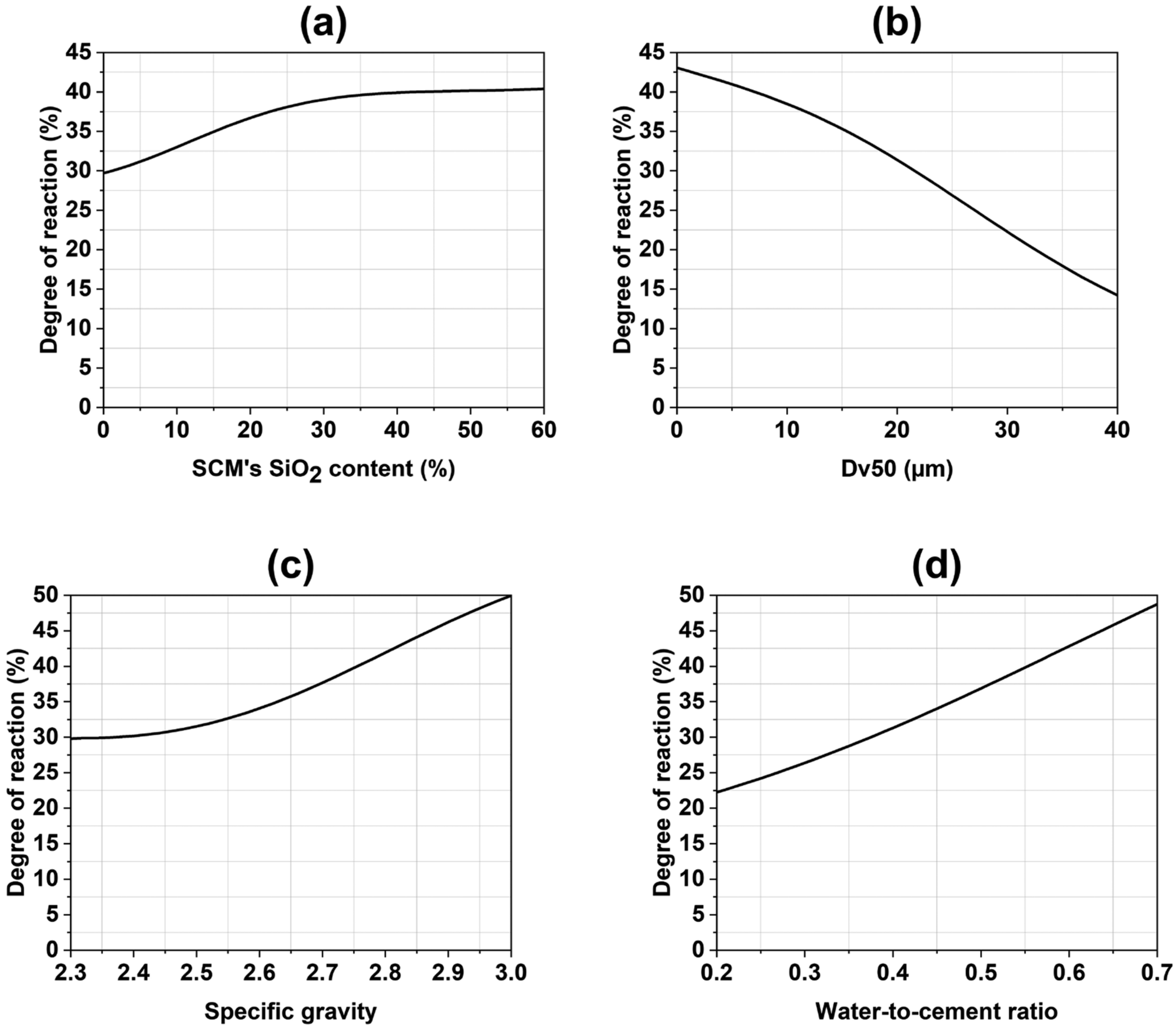


Fig. 3 Influence of key factors on the DOR: **a** SiO_2 content of the SCM, **b** Dv50, **c** specific gravity, and **d** W/C ratio, evaluated based on average oxide composition and physical properties. Conditions: W/C ratio = 0.5, curing time = 180 days and temperature = 25 °C

While surplus moisture can enhance the placement and finishing processes (Reddy & Rao, 2014), it can also contribute to the obstruction of spaces meant for reaction products (Navarrete et al., 2020). Additionally, the W/C ratio impacts the curing process. Lower ratios enhance moisture retention during the initial stages of hydration while negatively influencing the hydration process (Patil & Dubey, 2023).

Moreover, higher W/C ratios are known to increase the porosity of the hydrated cement matrix, yet their optimal utilization remains crucial, as both excessive and inadequate additions can yield adverse effects (Wong et al., 2020). These factors mentioned above collectively wield the potential to significantly influence the DOR of SCMs. Hence, a meticulous selection of the W/C ratio becomes imperative. Fig. 3d shows the effect of the W/C ratio on the DOR of SCMs. For the given W/C ratio ranges, the DOR of SCMs increases as the W/C ratio increases in agreement with previously published papers (Escalante et al., 2001; Snellings et al., 2022).

4 Conclusions

This study focused on investigating the factors that influence the DOR of SCMs in cementitious matrices to optimize their performance. Five different ML models were used: linear regression, GPR, DT regression, SVM, and XGBoost. The model with the best accuracy is further used to analyze the importance of the parameters that affect the DOR. The following conclusions were drawn from the obtained results:

- The GPR model exhibited superior accuracy and interpretability in its predictions compared to other ML models, with RMSE values demonstrating an improvement of more than four units compared to other models, except XGBoost.
- Among the parameters subject to initial adjustments, SCM SiO_2 content was identified as the most critical parameter influencing DOR, followed by Dv50, specific gravity, and W/C ratio.
- Enhancing the DOR of SCMs entails prolonging curing time and reducing Dv50 while simultaneously necessitating optimized values for SiO_2 content of the SCMs and the W/C ratio.
- These findings highlight the importance of a holistic approach to understanding the intricate interplay between various factors affecting the DOR of SCMs. They provide valuable insights into the relationships between the properties of SCMs and their performance in cementitious matrices.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40069-024-00677-w.

Additional file 1. The full dataset is provided as a supplementary material for reference.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2021R1A4A3033128 and 2022R1C1C1007498).

Author contributions

ABD: conceptualization, methodology, validation, formal analysis, investigation, resources, data curation, writing—original draft preparation, writing—review and editing, visualization. GJ: conceptualization, methodology, validation, formal analysis, investigation, resources, data curation, writing—original draft preparation, writing—review and editing, visualization, supervision, project administration, funding acquisition. SC: conceptualization, methodology, validation, formal analysis, investigation, writing—review and editing, visualization, supervision. JB, SP and HY: formal analysis, investigation, writing—review and editing, visualization, supervision. SP: conceptualization, methodology, validation, formal analysis, investigation, resources, data curation, writing—original draft preparation, writing—review and editing, visualization, supervision, project administration, funding acquisition. All authors have read and approved the final manuscript.

Availability of data and materials

Data will be available on request.

Declarations

Consent for publication

All individual participants agreed to be included in the study.

Competing interests

The authors declare they have no financial and competing interests.

Received: 3 January 2024 Accepted: 11 March 2024
Published online: 28 June 2024

References

Abuodeh, O. R., Abdalla, J. A., & Hawileh, R. A. (2020). Assessment of compressive strength of ultra-high performance concrete using deep machine learning techniques. *Applied Soft Computing Journal, 95*, 106552. https://doi.org/10.1016/j.asoc.2020.106552

Ahmed, A. (2019). Chemical reactions in pozzolanic concrete. *Modern Approaches on Material Science, 1*(4), 128–133. https://doi.org/10.32474/mams.2019.01.000120

Anurag, Kumar, R., Goyal, S., & Srivastava, A. (2021). A comprehensive study on the influence of supplementary cementitious materials on physico-mechanical, microstructural and durability properties of low carbon cement composites. *Powder Technology, 394*, 645–668. https://doi.org/10.1016/j.powtec.2021.08.081

Arvaniti, E. C., & De Belie, N. (2014). Method development for the particle size analysis. In *XIII conference on durability of building materials and components* (pp. 679–686).

Barnwal, A., Cho, H., & Hocking, T. (2022). Survival regression with accelerated failure time model in XGBoost. *Journal of Computational and Graphical Statistics, 31*(4), 1292–1302. https://doi.org/10.1080/10618600.2022.2067548

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? Arguments against avoiding RMSE in the literature. *Geoscientific Model Development, 7*(3), 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014

Charbuty, B., & Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends, 2*(01), 20–28. https://doi.org/10.38094/jastt20165

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining, 13–17 August* (pp. 785–794). https://doi.org/10.1145/2939672.2939785

Chen, Z., Lin, J., Sagoe-Crentsil, K., & Duan, W. (2022). Development of hybrid machine learning-based carbonation models with weighting function. *Construction and Building Materials, 321*(November 2021), 126359. https://doi.org/10.1016/j.conbuildmat.2022.126359

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science, 7*, 1–24. https://doi.org/10.7717/PEERJ-CS.623

Chou, J. S., Tsai, C. F., Pham, A. D., & Lu, Y. H. (2014). Machine learning in concrete strength simulations: Multi-nation data analytics. *Construction and Building Materials, 73*, 771–780. https://doi.org/10.1016/j.conbuildmat.2014.09.054

Cohen, S., Ruppin, E., & Dror, G. (2005). Feature selection based on the shapley value. In *IJCAI international joint conference on artificial intelligence* (pp. 665–670).

de Azevedo Basto, P., Estolano de Lima, V., & de Melo Neto, A. A. (2022). Effect of curing temperature in the relative decrease peak intensity of calcium hydroxide pastes for assessing pozzolanicity of supplementary cementitious materials. *Construction and Building Materials, 325*(November 2021), 126767. https://doi.org/10.1016/j.conbuildmat.2022.126767

Diaz-Loya, I., Juenger, M., Seraj, S., & Minkara, R. (2019). Extending supplementary cementitious material resources: Reclaimed and remediated fly ash and natural pozzolans. *Cement and Concrete Composites, 101*, 44–51. https://doi.org/10.1016/j.cemconcomp.2017.06.011

Donatello, S., Tyrer, M., & Cheeseman, C. R. (2010). Comparison of test methods to assess pozzolanic activity. *Cement and Concrete Composites, 32*(2), 121–127. https://doi.org/10.1016/j.cemconcomp.2009.10.008

Durdziński, P. T., Ben Haha, M., Bernal, S. A., De Belie, N., Gruyaert, E., Lothenbach, B., et al. (2017). Outcomes of the RILEM round robin on degree of reaction of slag and fly ash in blended cements. *Materials and*

Structures/Materiaux et Constructions, 50(2), 1–15. https://doi.org/10.1617/s11527-017-1002-1

Escalante, J. I., Gómez, L. Y., Johal, K. K., Mendoza, G., Mancha, H., & Méndez, J. (2001). Reactivity of blast-furnace slag in Portland cement blends hydrated under different conditions. *Cement and Concrete Research, 31*(10), 1403–1409. https://doi.org/10.1016/S0008-8846(01)00587-7

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189–1232.

Gholami, R., & Fakhari, N. (2017). Support vector machine: Principles, parameters, and applications. *Handbook of neural computation* (1st ed.). Elsevier Inc. https://doi.org/10.1016/B978-0-12-811318-9.00027-2

Gupta, S., & Chaudhary, S. (2022). State of the art review on supplementary cementitious materials in India—II: Characteristics of SCMs, effect on concrete and environmental impact. *Journal of Cleaner Production, 357*(April), 131945. https://doi.org/10.1016/j.jclepro.2022.131945

Haha, M. B., De Weerdt, K., & Lothenbach, B. (2010). Quantification of the degree of reaction of fly ash. *Cement and Concrete Research, 40*(11), 1620–1629. https://doi.org/10.1016/j.cemconres.2010.07.004

Hallet, V., De Belie, N., & Pontikes, Y. (2020). The impact of slag fineness on the reactivity of blended cements with high-volume non-ferrous metallurgy slag. *Construction and Building Materials, 257*, 119400. https://doi.org/10.1016/j.conbuildmat.2020.119400

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction. *The elements of statistical learning* (2nd ed., Vol. 27). Springer Science & Business Media.

Juenger, M. C. G., & Siddique, R. (2015). Recent advances in understanding the role of supplementary cementitious materials in concrete. *Cement and Concrete Research, 78*, 71–80. https://doi.org/10.1016/j.cemconres.2015.03.018

Juenger, M. C. G., Snellings, R., & Bernal, S. A. (2019). Supplementary cementitious materials: New sources, characterization, and performance insights. *Cement and Concrete Research, 122*(May 2019), 257–273. https://doi.org/10.1016/j.cemconres.2019.05.008

Kocaba, V., Gallucci, E., & Scrivener, K. L. (2012). Methods for determination of degree of reaction of slag in blended cement pastes. *Cement and Concrete Research, 42*(3), 511–525. https://doi.org/10.1016/j.cemconres.2011.11.010

Li, X., Snellings, R., Antoni, M., Alderete, N. M., Ben Haha, M., Bishnoi, S., et al. (2018). Reactivity tests for supplementary cementitious materials: RILEM TC 267-TRM phase 1. *Materials and Structures/Materiaux et Constructions, 51*(6), 1–14. https://doi.org/10.1617/s11527-018-1269-x

Liu, S., Zhang, T., Guo, Y., Wei, J., & Yu, Q. (2018). Effects of SCMs particles on the compressive strength of micro-structurally designed cement paste: Inherent characteristic effect, particle size refinement effect, and hydration effect. *Powder Technology, 330*, 1–11. https://doi.org/10.1016/j.powtec.2018.01.087

Lothenbach, B., Scrivener, K., & Hooton, R. D. (2011). Supplementary cementitious materials. *Cement and Concrete Research, 41*(12), 1244–1256. https://doi.org/10.1016/j.cemconres.2010.12.001

Ma, M., Zhao, G., He, B., Li, Q., Dong, H., Wang, S., & Wang, Z. (2021). XGBoost-based method for flash flood risk assessment. *Journal of Hydrology, 598*(April), 126382. https://doi.org/10.1016/j.jhydrol.2021.126382

Merrick, L., & Taly, A. (2020). The explanation game: Explaining machine learning models using Shapley values. Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics) (LNCS)*Machine learning and knowledge extraction.* (Vol. 12279). Springer International Publishing. https://doi.org/10.1007/978-3-030-57321-8_2

Mirzahosseini, M., & Riding, K. A. (2015). Influence of different particle sizes on reactivity of finely ground glass as supplementary cementitious material (SCM). *Cement and Concrete Composites, 56*, 95–105. https://doi.org/10.1016/j.cemconcomp.2014.10.004

Navarrete, I., Kurama, Y., Escalona, N., & Lopez, M. (2020). Impact of physical and physicochemical properties of supplementary cementitious materials on structural build-up of cement-based pastes. *Cement and Concrete Research, 130*(February), 105994. https://doi.org/10.1016/j.cemconres.2020.105994

Ndahirwa, D., Zmamou, H., Lenormand, H., & Leblanc, N. (2022). The role of supplementary cementitious materials in hydration, durability and shrinkage of cement-based materials, their environmental and economic benefits: A review. *Cleaner Materials, 5*(June), 100123. https://doi.org/10.1016/j.clema.2022.100123

Noble, W. S. (2006). What is a support vector machine? *Nature Biotechnology, 24*(12), 1565–1567. https://doi.org/10.1038/nbt1206-1565

Pacewska, B., & Wilińska, I. (2020). Usage of supplementary cementitious materials: Advantages and limitations: Part I. C–S–H, C–A–S–H and other products formed in different binding mixtures. *Journal of Thermal Analysis and Calorimetry, 142*(1), 371–393. https://doi.org/10.1007/s10973-020-09907-1

Pal, M., & Mather, P. M. (2001). Decision tree based classification of remotely sensed data. In *Asian conference on remote sensing*.

Patil, M. N., & Dubey, S. K. D. (2023). *Effect of exposure condition, free water–cement ratio on quantities, rheological and mechanical properties of concrete* Lecture notes in civil engineering (Vol. 260). Springer Nature. https://doi.org/10.1007/978-981-19-2145-2_14

Pfingsten, J., Rickert, J., & Lipus, K. (2018). Estimation of the content of ground granulated blast furnace slag and different pozzolanas in hardened concrete. *Construction and Building Materials, 165*, 931–938. https://doi.org/10.1016/j.conbuildmat.2018.01.065

Phung, Q. T., Ferreira, E., Seetharam, S., Nguyen, V. T., Govaerts, J., & Valcke, E. (2021). Understanding hydration heat of mortars containing supplementary cementitious materials with potential to immobilize heavy metal containing waste. *Cement and Concrete Composites, 115*(May 2020), 103859. https://doi.org/10.1016/j.cemconcomp.2020.103859

Rahla, K. M., Mateus, R., & Bragança, L. (2019). Comparative sustainability assessment of binary blended concretes using supplementary cementitious materials (SCMs) and ordinary Portland cement (OPC). *Journal of Cleaner Production, 220*, 445–459. https://doi.org/10.1016/j.jclepro.2019.02.010

Ramanathan, S., Pestana, L. R., & Suraneni, P. (2022). Reaction kinetics of supplementary cementitious materials in reactivity tests. *Cement, 8*(October 2021), 100022. https://doi.org/10.1016/j.cement.2022.100022

Rasmussen, C. E. (2003). Gaussian processes in machine learning. *Advanced lectures on machine learning* (Vol. 3176/2004, pp. 63–71). Springer.

Reddy, V. M., & Rao, D. M. V. S. (2014). Effect of W/C ratio on workability and mechanical properties of high strength self compacting concrete (M70 grade). *IOSR Journal of Mechanical and Civil Engineering, 11*(5), 15–21. https://doi.org/10.9790/1684-11561521

Sabir, B., Wild, S., & Bai, J. (2001). Metakaolin and calcined clays as pozzolans for concrete: A review. *Cement and Concrete Composites, 23*(6), 441–454. https://doi.org/10.1016/S0958-9465(00)00092-5

Samad, S., & Shah, A. (2017). Role of binary cement including supplementary cementitious material (SCM), in production of environmentally sustainable concrete: A critical review. *International Journal of Sustainable Built Environment, 6*(2), 663–674. https://doi.org/10.1016/j.ijsbe.2017.07.003

Sanjuán, M. Á., Argiz, C., Gálvez, J. C., & Moragues, A. (2015). Effect of silica fume fineness on the improvement of Portland cement strength performance. *Construction and Building Materials, 96*, 55–64. https://doi.org/10.1016/j.conbuildmat.2015.07.092

Schöler, A., Lothenbach, B., Winnefeld, F., Haha, M. B., Zajac, M., & Ludwig, H. M. (2017). Early hydration of SCM-blended Portland cements: A pore solution and isothermal calorimetry study. *Cement and Concrete Research, 93*, 71–82. https://doi.org/10.1016/j.cemconres.2016.11.013

Scrivener, K. L., Lothenbach, B., De Belie, N., Gruyaert, E., Skibsted, J., Snellings, R., & Vollpracht, A. (2015). TC 238-SCM: Hydration and microstructure of concrete with SCMs: State of the art on methods to determine degree of reaction of SCMs. *Materials and Structures/Materiaux et Constructions, 48*(4), 835–862. https://doi.org/10.1617/s11527-015-0527-4

Shi, J. Q., & Choi, T. (2011). *Gaussian process regression analysis for functional data* (1st ed.). CRC Press.

Simonsen, A. M. T., Solismaa, S., Hansen, H. K., & Jensen, P. E. (2020). Evaluation of mine tailings' potential as supplementary cementitious materials based on chemical, mineralogical and physical characteristics. *Waste Management, 102*, 710–721. https://doi.org/10.1016/j.wasman.2019.11.037

Skibsted, J., & Snellings, R. (2019). Reactivity of supplementary cementitious materials (SCMs) in cement blends. *Cement and Concrete Research, 124*(June), 105799. https://doi.org/10.1016/j.cemconres.2019.105799

Snellings, R., Machner, A., Bolte, G., Kamyab, H., Durdzinski, P., Teck, P., et al. (2022). Hydration kinetics of ternary slag-limestone cements: Impact of water to binder ratio and curing temperature. *Cement and Concrete*

Research, 151(July 2021), 106647. https://doi.org/10.1016/j.cemconres.2021.106647

Snellings, R., & Scrivener, K. L. (2016). Rapid screening tests for supplementary cementitious materials: Past and future. *Materials and Structures/Materiaux et Constructions, 49*(8), 3265–3279. https://doi.org/10.1617/s11527-015-0718-z

Snoeck, D., Schaubroeck, D., Dubruel, P., & De Belie, N. (2014). Effect of high amounts of superabsorbent polymers and additional water on the workability, microstructure and strength of mortars with a water-to-cement ratio of 0.50. *Construction and Building Materials, 72*, 148–157. https://doi.org/10.1016/j.conbuildmat.2014.09.012

Song, Y. Y., & Lu, Y. (2015). Decision tree methods: Applications for classification and prediction. *Shanghai Archives of Psychiatry, 27*(2), 130–135. https://doi.org/10.11919/j.issn.1002-0829.215044

Suraneni, P., Hajibabaee, A., Ramanathan, S., Wang, Y., & Weiss, J. (2019). New insights from reactivity testing of supplementary cementitious materials. *Cement and Concrete Composites, 103*(May), 331–338. https://doi.org/10.1016/j.cemconcomp.2019.05.017

Tironi, A., Trezza, M. A., Scian, A. N., & Irassar, E. F. (2013). Assessment of pozzolanic activity of different calcined clays. *Cement and Concrete Composites, 37*(1), 319–327. https://doi.org/10.1016/j.cemconcomp.2013.01.002

Walkley, B., & Provis, J. L. (2019). Solid-state nuclear magnetic resonance spectroscopy of cements. *Materials Today Advances, 1*, 100007. https://doi.org/10.1016/j.mtadv.2019.100007

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research, 30*(1), 79–82. https://doi.org/10.3354/cr030079

Wong, H. S., Poole, A. B., Wells, B., Eden, M., Barnes, R., Ferrari, J., et al. (2020). Microscopy techniques for determining water–cement (W/C) ratio in hardened concrete: A round-robin assessment. *Materials and Structures/Materiaux et Constructions, 53*(2), 1–19. https://doi.org/10.1617/s11527-020-1458-2

Zhang, D. (2017). A coefficient of determination for generalized linear models. *American Statistician, 71*(4), 310–316. https://doi.org/10.1080/00031305.2016.1256839

Publisher's Note

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

Aron Berhanu Degefa is a Ph.D. candidate at Department of Civil Engineering, Pukyong National University, South Korea.

Geonyeol Jeon is a Ph.D. candidate at School of Civil Engineering, Chungbuk National University, South Korea.

Sooyung Choi is a Master student at Department of Artificial Intelligence, Sungkyunkwan University, South Korea.

JinYeong Bak is an Assistant professor at Department of Artificial Intelligence, Sungkyunkwan University, South Korea.

Seunghee Park is a Professor at School of Civil, Architectural Engineering& Landscape Architecture, Sungkyunkwan University, South Korea.

Hyungchul Yoon is an Associate Professor at School of Civil Engineering, Chungbuk National University, South Korea.

Solmoi Park is an Associate Professor at Department of Civil Engineering, Pukyong National University, South Korea.