*Data-Centric Engineering* (2026), 7: e34
doi:10.1017/dce.2026.10075

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

# Data-driven modeling of compressive strength in sustainable self-compacting concrete incorporating recycled aggregates using ensemble learning techniques

Arslan Qayyum Khan[1], Muhammad Dawood Rasheed[2], Muhammad Huzaifa Naveed[2] and Amorn Pimanmas[3]

[1]Department of Civil and Environmental Engineering, Florida International University, USA
[2]Department of Civil Engineering, The University of Lahore, Pakistan
[3]Department of Civil Engineering, Kasetsart University, Thailand
**Corresponding author:** Amorn Pimanmas; Email: amorn.pi@ku.th

**Received:** 02 February 2026 UTC; **Revised:** 02 July 2026 UTC; **Accepted:** 14 July 2026 UTC

**Keywords:** compressive strength; machine learning; recycled aggregates; self-compacting concrete; XGBoost

## Abstract

This study develops a robust framework for estimating the compressive strength of self-compacting concrete (SCC) incorporating recycled aggregates using supervised machine learning (ML) techniques. A comprehensive experimental database comprising 582 concrete mix designs was used, encompassing diverse input variables including binder content, water, coarse and fine aggregates, recycled aggregate proportion, superplasticizer dosage, and curing time. Seven ML algorithms—XGBoost, CatBoost, AdaBoost, Extra Trees, Bagging Regressor, K-Nearest Neighbors, and Radius Neighbors—were systematically trained using a stratified 70/15/15 data split and optimized via grid search with five-fold cross-validation. Model performance was evaluated using coefficient of determination ($R^2$), root mean squared error, and MAE across training, validation, and testing datasets. Among all models, XGBoost demonstrated the highest accuracy, achieving an average $R^2$ of 0.9799, RMSE of 2.87 MPa, and mean absolute error of 1.97 MPa. The Permutation Feature Importance analysis revealed that binder content, water, and coarse aggregate were the most influential predictors of strength. This study confirms that ensemble ML models, particularly XGBoost, can reliably predict the compressive strength of SCC with recycled aggregates, while offering transparent insights into material behavior. The results provide a valuable tool for sustainable mix design optimization and practical implementation in eco-efficient concrete construction.

## Impact statement

This study demonstrates how large, heterogeneous experimental datasets can be systematically leveraged to improve predictive modeling in sustainable construction materials. By compiling and curating a high-quality database of 582 self-compacting concrete mixes incorporating recycled aggregates, the study shows that ensemble learning—particularly XGBoost—can achieve both high predictive accuracy and interpretability. Beyond performance, the study emphasizes data-centric practices such as robust preprocessing, cross-validation, and feature importance analysis to enhance model generalizability and engineering trust. The proposed framework supports data-driven mix design optimization, enabling more reliable adoption of

This research article was awarded Open Data badge for transparent practices. See the Data Availability Statement for details.

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/dce.2026.10075 Published online by Cambridge University Press

recycled materials in structural concrete while reducing experimental trial-and-error. More broadly, the study illustrates how data-centric machine learning can bridge experimental materials science and practical engineering decision-making.

## 1. Introduction

The construction sector is a cornerstone of global development, but is also responsible for an alarming share of natural resource depletion and environmental degradation (Liu et al., 2021; Ridho et al., 2021). Cement production alone contributes nearly 8% of global $CO_2$ emissions, while aggregate extraction and concrete manufacturing strain ecosystems and deplete non-renewable resources (Khan et al., 2023a). With the surge in infrastructure development, especially in rapidly urbanizing regions, these environmental concerns are intensifying. In this context, there has been a significant push toward sustainable construction practices, with a particular focus on resource conservation, emission reduction, and the implementation of a circular economy (Khan et al., 2026a). One promising approach is the reuse of construction and demolition waste (CDW), particularly recycled aggregates, in new concrete applications (Gontia et al., 2019).

Recycled aggregates—comprising crushed concrete, bricks, and other demolition byproducts—have been increasingly investigated as partial or full replacements for natural coarse and fine aggregates (Xiao et al., 2012; Saha and Rajasekaran, 2016; Bui et al., 2017). This approach significantly reduces the demand for pure materials and mitigates the environmental burden of waste disposal. Moreover, the use of recycled concrete aggregate supports sustainable construction by lowering the carbon footprint and promoting closed-loop material cycles (Poon et al., 2004a; McNeil and Kang, 2013). However, recycled aggregates are often characterized by inferior mechanical and physical properties compared to natural aggregates. These include higher water absorption, increased porosity, variable strength, and the presence of adhered old mortar, all of which pose challenges in maintaining consistent performance in structural concrete, especially in high-performance or flow-critical applications such as self-compacting concrete (SCC) (Etxeberria et al., 2007; Rahal, 2007).

SCC is a highly flowable, non-segregating concrete that can fill formwork under its own weight without the need for vibration (Khan et al., 2025a). Developed in Japan in the late 1980s, SCC is particularly suited for precast applications, densely reinforced sections, and architectural elements requiring excellent surface finishes (Ahmad, 2008). Its fresh state properties, high filling ability, passing ability, and segregation resistance are achieved through careful mix proportioning and the use of superplasticizers and viscosity-modifying agents (Shi et al., 2015). However, SCC's unique rheological requirements make it sensitive to variations in material properties, especially when recycled aggregate is introduced. The unpredictable moisture content and rough surface texture of recycled aggregates impact workability, leading to inconsistent placement and compaction. More critically, the compressive strength of SCC containing recycled aggregates can vary widely due to multiple interacting parameters, including binder content, water-to-binder ratio, aggregate gradation, admixture dosage, and curing conditions (Bui et al., 2017).

Predicting the compressive strength of SCC with recycled aggregates using traditional empirical methods or linear regression models has proven unreliable, especially when extended across diverse datasets or mix compositions (Chaabene et al., 2020; Nunez et al., 2021). These classical models often fail to capture the nonlinearities and interaction effects present in concrete systems with recycled constituents. In recent years, machine learning (ML) has emerged as a powerful alternative for material property prediction due to its ability to learn complex patterns from data without being constrained by rigid assumptions (Khan et al., 2025b, 2026b). Unlike traditional statistical methods, ML algorithms can model high-dimensional relationships and are adaptable to large, heterogeneous datasets typical of real-world experimental research in concrete materials (Nafees et al., 2021).

Numerous studies have demonstrated the potential of ML in predicting the mechanical and durability properties of concrete. Techniques such as Artificial Neural Networks (ANNs), Support Vector Regression (SVR), and Ensemble Learning methods (e.g., Random Forests and Gradient Boosting Machines)

https://doi.org/10.1017/dce.2026.10075 Published online by Cambridge University Press

have been employed to predict compressive strength, modulus of elasticity, split tensile strength, and durability indices like carbonation depth and chloride penetration (Asteris et al., 2021; Huang et al., 2023). For example, Khan et al. (2023b) and Nasir et al. (2020) developed ANN-based models for predicting compressive strength in conventional and modified concrete, demonstrating improved performance over regression methods. More recently, the use of ensemble learning, particularly XGBoost, CatBoost, and Extra Trees, has shown superior accuracy and generalization capability due to their ability to handle feature interactions and reduce overfitting through built-in regularization techniques (de-Prado-Gil et al., 2022; Rajakarunakaran et al., 2022).

However, several limitations persist in existing literature. Many studies focus on conventional concrete or use datasets limited in size, diversity, or feature space (de-Prado-Gil et al., 2022; Jagadesh et al., 2023). Others evaluate only a single ML algorithm without comparison to alternative approaches, which limits generalizability and practical guidance for model selection. Importantly, very few studies have applied ML to predict the compressive strength of SCC containing recycled aggregates, despite the increasing relevance of such systems in sustainable infrastructure. Furthermore, interpretability remains a concern. Most ML models, particularly black-box types such as neural networks, offer little insight into which input variables most significantly affect the predictions, making it difficult for engineers to trust or apply these models in practical design workflows (Liu et al., 2021).

To bridge these gaps, this study makes three distinct contributions relative to prior ML studies on SCC. First, it employs a substantially larger and more heterogeneous database—582 unique mixes compiled from more than 35 independent experimental studies—than the single-source or limited datasets used in earlier work. Second, rather than evaluating a single algorithm, it systematically benchmarks seven models spanning three algorithmic families (bagging, boosting, and instance-based) under an identical training and validation protocol, providing direct guidance for model selection. Third, it integrates Permutation Feature Importance (PFI) to convert the best-performing model into an interpretable, engineering-relevant tool rather than a black-box predictor. A large-scale database comprising 582 unique SCC mix designs with recycled aggregates was utilized, representing a broad spectrum of input variables, including binder content, water content, fine and coarse aggregates, recycled aggregate proportion, superplasticizer dosage, and curing time. The study evaluates and compares the performance of seven ML algorithms, including bagging-based models (Bagging Regressor and Extra Trees Regressor), boosting-based models (XGBoost, CatBoost, AdaBoost), and instance-based models (K-Nearest Neighbors (KNN) and Radius Neighbors (RN)). Each model was rigorously trained and validated using stratified data splits and optimized through a grid search and cross-validation process to ensure robust generalization. Model performance was assessed using standard metrics—coefficient of determination ($R^2$), root mean squared error (RMSE), and mean absolute error (MAE)—across training, validation, and testing datasets. Beyond predictive accuracy, the study employed PFI to quantify the contribution of each input variable toward the predicted compressive strength, offering transparency and aiding mix design decisions. By systematically comparing multiple ML algorithms and interpreting their outputs through both statistical and engineering lenses, this study contributes a robust and interpretable predictive model for the compressive strength of SCC with recycled aggregates.

## 2. Ml models and evaluation metrics

### *2.1. ML models*

In artificial intelligence, ML is a fundamental method allowing computers to learn from data without explicit programming (Khan et al., 2026c). Jovel and Greiner (2021) and Saraswat and Raj (2021) classify ML techniques as supervised, unsupervised, and reinforcement learning. While unsupervised learning identifies trends in unlabeled data (Jovel and Greiner, 2021), supervised learning trains models on labeled data to provide predictions. In several sectors, including biological research and drug discovery, ML algorithms can analyze vast amounts of data and produce faster and more accurate results (Khan et al., 2025c). Classification and regression challenges provide additional divisions of supervised learning (Bakshi and Bakshi, 2018). The effectiveness of ML models depends on factors such as data quality,

https://doi.org/10.1017/dce.2026.10075 Published online by Cambridge University Press

algorithm selection, and proper training to avoid issues like overfitting (Khan et al., 2024). ML provides robust tools for leveraging extensive data and addressing complex issues across various fields as it continues to evolve.

### *2.1.1. Bagging-based models*

Bagging embodies an ensemble learning approach that enhances the effectiveness of unstable estimators or classifiers, especially in the context of high-dimensional data (Bühlmann and Yu, 2002). The method operates by generating several bootstrap samples from the training data, utilizing the learning algorithm on each sample, and subsequently averaging the predictions that result from this process (Buja and Stuetzle, 2002). Bagging serves to diminish variance and mean squared error by softening difficult decisions; however, it simultaneously introduces an increase in bias (Bühlmann and Yu, 2002; Buja and Stuetzle, 2002). A more cost-effective approach known as subagging, which relies on subsampling, demonstrates comparable accuracy to bagging (Bühlmann and Yu, 2002). Bagging has found extensive application in the field of remote sensing, serving important roles in areas such as crop yield prediction, estimation of forest structure parameters, mapping of natural hazards, and the spatial downscaling of climate parameters and land surface temperature (Zhang et al., 2022). It is noteworthy that both bagging with replacement and without replacement can yield comparable outcomes when suitable resample sizes are selected (Buja and Stuetzle, 2002).

Extra Trees represent a collective ML approach specifically designed to address supervised classification and regression challenges. This approach expands upon the concept of random forests by incorporating additional randomization into the construction of trees (Geurts et al., 2006). This algorithm effectively randomizes the selection of attributes and cuts points while splitting tree nodes, which may lead to entirely randomized trees that operate independently of the output values (Geurts et al., 2006). Like other tree-based approaches, Extra Trees can handle various types of predictor variables, address missing data, and inherently account for interaction effects among predictors (Elith et al., 2008). This approach demonstrates computational efficiency and resilience to overfitting (Liaw and Wiener, 2002). Ensemble methods such as Extra Trees and other techniques like random forests and boosted regression trees typically demonstrate superior predictive performance compared to conventional single-tree models (Elith et al., 2008; Loh, 2011). Applying these methods is especially beneficial for representing intricate and nonlinear relationships within various ecological and scientific fields.

### *2.1.2. Boosting-based models*

XGBoost (eXtreme Gradient Boosting) embodies a robust ML algorithm known for its exceptional prediction accuracy, interpretability, and versatility in classification tasks (Ali et al., 2023). This method is an ensemble approach that constructs decision trees incrementally, focusing on the optimization of a chosen differentiable loss function (Fafalios et al., 2020). The advantages of XGBoost are evident in its ability to scale, its portability, and its capacity for distributed processing, which enables it to manage billions of instances with remarkable efficiency (Chen and Guestrin, 2016). The algorithm utilizes innovative methods, including a sparsity-aware approach designed for handling sparse data and a weighted quantile sketch for approximate tree learning (Chen and Guestrin, 2016). XGBoost has gained significant popularity in addressing regression and classification challenges, especially within the realms of scientific computing, data science, and ML applications (Ali et al., 2023). The increasing popularity of this approach in ML competitions has played a significant role in its widespread adoption in recent years (Fafalios et al., 2020). XGBoost integrates both *L1* (Lasso) and *L2* (Ridge) regularization terms into its objective function, which helps mitigate overfitting and improve the model's generalization, making it especially suitable for high-dimensional and noisy datasets, such as those encountered in concrete material research.

The objective function in XGBoost is composed of two key elements: a differentiable convex loss function $l(yi, yi)$, which quantifies the discrepancy between the predicted and actual values, and a regularization term $\Omega(f)$, which penalizes the model's complexity. The regularization component serves

https://doi.org/10.1017/dce.2026.10075 Published online by Cambridge University Press

to mitigate the formation of excessively complex trees by imposing constraints on both the number of leaves and the magnitude of leaf weights, as illustrated in Eq. (1):

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \tag{1}$$

where $T$ refers to the total count of leaves, while $\omega$ signifies the vector comprising the weights assigned to each leaf, the symbol $\lambda$ indicates the coefficient associated with *L2* regularization, and $\gamma$ denotes the cost incurred in terms of complexity for each additional leaf introduced. In order to enhance the efficiency of this function, XGBoost employs a second-order Taylor expansion of the loss function. This approach transforms the problem into a format that is amenable to quadratic optimization, as illustrated in Eq. (2):

$$\mathcal{L}_k \cong \sum_{i=1}^{n}\left[g_i f_k(x_i) + \frac{1}{2}h_i {f_k}^2(x_i)\right] + \Omega(f_k,) \tag{2}$$

In this expression, $g_\mathrm{i}$ and $h_\mathrm{i}$ represent the first- and second-order gradients of the loss function concerning the predicted output for the $i$-th instance.

This formulation enables an accurate assessment of tree structure and leaf weights. Additional simplification is attained by consolidating gradients across the instances $\mathrm{i} \in I_j$ that are categorized within leaf $j$. The function is presented in a more concise manner, as illustrated in Eq. (3).

$$\mathcal{L}_k \cong \sum_{j=1}^{T}\left[\left(\sum i \in Ijg_i\right)w_j + \frac{1}{2}\left(\sum i \in I_j h_i + \lambda\right){w_j}^2\right] + \gamma(T) \tag{3}$$

This regularized formulation enhances computational efficiency while also improving generalization by reducing overfitting.

CatBoost is an ensemble ML model that utilizes gradient-boosted decision trees and demonstrates a remarkable ability to manage categorical features (Dorogush et al., 2018). The findings indicate that it has exhibited remarkable performance across a range of applications, characterized by notable accuracy (Joshi et al., 2021). CatBoost demonstrates notable efficacy in handling large-scale data challenges that encompass diverse and categorical datasets (Hancock and Khoshgoftaar, 2020). CatBoost, functioning as a boosting algorithm, integrates predictions from various base models to attain exceptional performance across a range of machine-learning applications (Mienye and Sun, 2022). The model demonstrates notable strengths, particularly in its Graphics Processing Unit (GPU) implementation for learning and Central Processing Unit (CPU) implementation for scoring, both of which offer considerable speed advantages over other gradient-boosting libraries (Dorogush et al., 2018). Nonetheless, the effectiveness of CatBoost is notably influenced by the tuning of hyperparameters, highlighting the critical role of optimization (Hancock and Khoshgoftaar, 2020).

AdaBoost (Adaptive Boosting) embodies a significant advancement in ML, as it systematically improves the performance of weak learners by concentrating on instances that have been misclassified (Freund and Schapire, 1999). The process involves allocating weights to training instances, which are subsequently adjusted based on the performance of each weak learner. This method culminates in integrating these learners into a robust classifier (Schapire, 1999). The effectiveness of AdaBoost can be attributed to its capacity to mitigate overfitting and its connection to support vector machines (Freund and Schapire, 1999). AdaBoost was specifically designed to filter out instances that exhibit high relative estimation errors (Solomatine and Shrestha, 2004). This model has enhanced performance relative to other boosting methods, bagging, and ANNs across various applications, including hydrological modeling (Solomatine and Shrestha, 2004). The adaptability and efficacy of AdaBoost have contributed to its widespread adoption within ML (Tu et al., 2017).

### *2.1.3. Instance-based models*

KNN represents a straightforward yet powerful ML algorithm that is employed for both classification and regression tasks (Taunk et al., 2019). The system functions by retaining all accessible data points and

https://doi.org/10.1017/dce.2026.10075 Published online by Cambridge University Press

categorizing new instances according to their resemblance to the pre-existing dataset (Gangula and Venkateswarlu, 2024). The strength of KNN is found in its straightforwardness and efficiency, especially when applied to large datasets characterized by low dimensions (Kramer and Kramer, 2013). The algorithm's performance is shaped by various factors, including the selection of the $k$ value, the method used for distance calculation, and the suitability of the predictors (Zhang, 2016). Although KNN finds extensive use in numerous fields, it is essential to acknowledge its limitations (Gangula and Venkateswarlu, 2024). In response to these limitations, scholars have created enhanced iterations of the KNN algorithm (Taunk et al., 2019). The algorithm demonstrates a remarkable adaptability, applicable to various tasks such as multi-label classification, regression, and semi-supervised learning challenges (Kramer and Kramer, 2013). Although KNN is straightforward in its approach, it holds significant value within the dynamic field of ML and artificial intelligence (Gangula and Venkateswarlu, 2024).

The RN algorithm is a supervised, instance-based ML approach that combines simplicity with effectiveness in analyzing extensive datasets, while ensuring rapid execution and minimal use of runtime space (Rossikova et al., 2016). This approach exhibits parallels with the KNN algorithm. This method is extensively used in pattern recognition and ML for classification and regression (Kramer and Kramer, 2013). Both methods rely on classification based on proximity, where available data points are stored, and new cases are classified according to their similarity to the existing data (Gangula and Venkateswarlu, 2024). KNN and RN demonstrate notable effectiveness when applied to large datasets with low dimensionality (Kramer and Kramer, 2013). Careful consideration of various factors is essential for these algorithms, including the number of neighbors, methods for calculating distance, and selecting suitable predictors, all of which contribute to optimizing model performance (Zhang, 2016). Although KNN and RN find utility in various fields, their performance may be affected by the particular nature of the problem and the attributes of the dataset involved.

### *2.2. Evaluation metrics for the performance of ML models*

Performance metrics play a crucial role in evaluating ML models. Although metrics such as RMSE and MAE are frequently employed, they possess certain limitations regarding their interpretability (Chicco et al., 2021). $R^2$ is suggested as a more insightful and accurate measure for assessing regression analysis (Chicco et al., 2021). The frameworks in question enable a comprehensive evaluation of models, assist in the development of tailored regression models, and provide support for the processes of model selection, parameter tuning, and optimization (Van Thieu, 2024). Various metrics enable researchers to assess, compare, and refine ML algorithms and models in a meaningful way. In Eqs. (4–6), where n represents the total number of datasets, the predicted values are denoted by $\widehat{y}_i$ and the actual values $y_i$, with $\overline{y}$ represents the mean of the actual values.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}\Sigma(y_i - \overline{y}_i)^2} \tag{4}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2} \tag{5}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y}_i| \tag{6}$$

The RMSE and MAE have the same units (MPa) as the compressive strength. In contrast, $R^2$ is represented as a unitless value. The lower values of RMSE and MAE, alongside the higher values of $R^2$, suggest an outstanding accuracy in the predictive outcomes of ML models.

https://doi.org/10.1017/dce.2026.10075 Published online by Cambridge University Press

## 3. Research methodology

The primary objective of this study is to develop a robust data-driven model for predicting the compressive strength of recycled aggregate-based SCC. As illustrated in Figure 1, the methodology begins with the establishment of a comprehensive experimental database, which involves selecting relevant input parameters and collecting data from previous experimental studies. These parameters include binder content, water, coarse aggregate, fine aggregate, recycled aggregate, superplasticizer, and curing time. Following the formulation of the dataset, a series of supervised ML models are trained and evaluated, including ensemble models from the boosting and bagging families as well as instance-based algorithms. The performance of these models is rigorously assessed using statistical metrics, including $R^2$, RMSE, and MAE. After model evaluation, a comparative analysis is conducted within each algorithmic group to identify the model with the best generalization performance. Subsequently, PFI is employed to quantify the relative contribution of each input variable toward strength prediction. The final stage involves selecting the most accurate and stable prediction model, which provides a reliable tool for optimizing mix designs and promoting the sustainable use of recycled materials in concrete construction.

### *3.1. Dataset compilation and description*

The foundation of this study lies in the development of a comprehensive and diverse experimental dataset for modeling the compressive strength of SCC incorporating recycled aggregates. A total of 582 unique SCC mix designs were compiled from over 35 peer-reviewed publications spanning the last two decades.

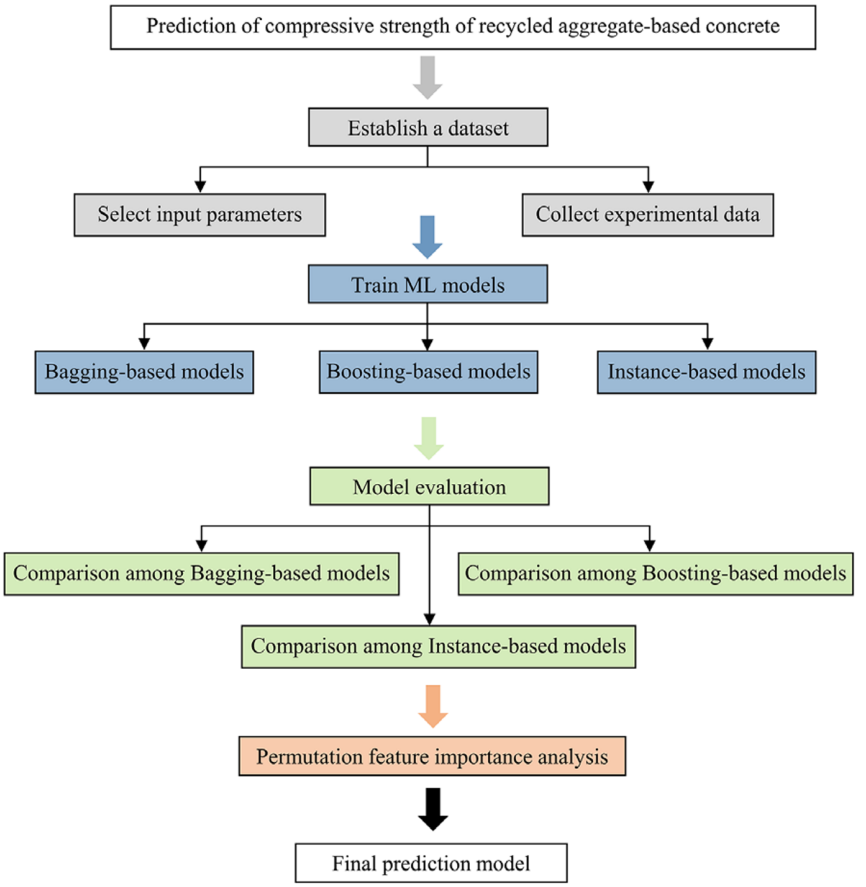



***Figure 1.*** *Methodological framework.*

https://doi.org/10.1017/dce.2026.10075 Published online by Cambridge University Press

The number of records contributed by each source ranges from 1 to 80 (median ≈ 10); the five largest sources account for approximately 43% of the records and the ten largest for approximately 64%, reflecting the concentration of SCC-with-recycled-aggregate testing within a limited number of research groups. The record-level source attribution for all 582 mixes is provided in the publicly available dataset to ensure full transparency of the dataset composition (Lam et al., 1998; Poon et al., 2000, 2001, 2002, 2004a, 2004b, 2007; sun Poon and Azhar, 2003; Poon and Ho, 2004; Kou et al., 2007, 2011; Kou and Poon, 2008; Güneyisi et al., 2010; Schoppe, 2011; Uysal and Sumer, 2011; Ali and Al-Tersawy, 2012; Revathi et al., 2013; Gesoglu et al., 2015; Kou and Poon, 2015; Ranjbar and Mousavi, 2015; Yang et al., 2015; Sadrmomtazi et al., 2016; Aslani et al., 2018; Duan et al., 2018; Fiol et al., 2018; Hilal and Kareem, 2018; Mohamed, 2018; Wan et al., 2018; Chunchu and Putta, 2019; Mohammed et al., 2019; Sadeghi-Nik et al., 2019; Sasanipour and Aslani, 2019; Bidabadi et al., 2020; Duan et al., 2020; Poongodi et al., 2021; Faraj et al., 2022; Mohammed et al., 2022). These sources were selected to ensure diversity in material types, mix proportions, testing methods, and geographic conditions, thereby capturing a broad range of real-world scenarios. We note, however, that although more than 35 studies are represented, a portion of the records originates from a smaller number of research groups, reflecting the concentration of SCC-with-recycled-aggregate testing in the literature. A breakdown of the number of records per source is provided to make this composition transparent. Each mix entry in the dataset includes corresponding input variables and target output values derived from laboratory testing of hardened concrete. The resulting database of 582 records is larger than, and at least comparable in breadth to, the datasets used in most prior ML studies on SCC with recycled aggregates, which typically draw on a few hundred records from a smaller number of sources. More importantly, its wide coverage of binder, water, aggregate, admixture, and curing ranges ensures that the parameter space relevant to practical SCC mix design is well represented, supporting the development of generalizable models.

The input parameters extracted from the literature include binder content ($kg/m^3$), fine aggregate ($kg/m^3$), coarse aggregate ($kg/m^3$), recycled aggregate ($kg/m^3$), water content ($kg/m^3$), superplasticizer dosage ($kg/m^3$), and curing time (days). These seven features were selected based on their established influence on concrete workability and mechanical properties, particularly in SCC mixtures, where flow characteristics and segregation resistance are highly sensitive to material ratios. The compressive strength (MPa), measured at various curing ages (ranging from 7 to 400 days), serves as the dependent variable in the modeling process. Curing age is incorporated explicitly as one of the seven input features (curing time, in days); consequently, the model accounts directly for the influence of age on strength development, and the same mix design tested at different ages is represented as distinct records with their respective measured strengths.

The collected data was carefully curated to eliminate inconsistencies and ensure dimensional uniformity. Units were standardized across all sources, and mix designs with missing or incomplete strength measurements were excluded. When available, mixtures tested under standard curing conditions (20 ± 2 °C, 95% relative humidity) were prioritized to maintain consistency. Special attention was given to maintaining the full heterogeneity of the input variables in order to improve the generalizability of the ML models across different concrete compositions and performance ranges. Because the database aggregates results from more than 35 laboratories, variability in raw-material sources, specimen geometry, and testing protocols inevitably introduces a degree of inter-study noise. Unit standardization, exclusion of incomplete or inconsistent records, prioritization of standard curing conditions, and training-set-only normalization were applied to reduce systematic inconsistencies, while genuine physical variability was deliberately preserved so that the models learn relationships that hold across diverse practical conditions rather than the idiosyncrasies of a single source.

Table 1 presents a sample of the compiled dataset, illustrating the range and variety of SCC-based mix designs included in this study. The table reflects significant variation in key input parameters: binder content ranged from 300 to 702 $kg/m^3$, water content from 112 to 271 $kg/m^3$, superplasticizer dosage from 0 to 35.1 $kg/m^3$, and curing time from 7 to 400 days. This variation is crucial for training robust models that can handle the complex interactions inherent in recycled concrete systems.

https://doi.org/10.1017/dce.2026.10075 Published online by Cambridge University Press

https://doi.org/10.1017/dce.2026.10075 Published online by Cambridge University Press

***Table 1.*** *Sample entries from the dataset*

| Binder content ($kg/m^3$) | Fine aggregate ($kg/m^3$) | Coarse aggregate ($kg/m^3$) | Recycled aggregate ($kg/m^3$) | Water ($kg/m^3$) | Superplasticizer ($kg/m^3$) | Curing time (Days) | Compressive strength (MPa) | References |
|---|---|---|---|---|---|---|---|---|
| 500 | 725 | 1087 | 0 | 150 | 0.5 | 28 | 77.2 | Poon et al. (2002) |
| 702 | 641 | 949 | 0 | 135 | 35.1 | 28 | 77.44 | Poon et al. (2000) |
| 390 | 615 | 1143 | 0 | 195 | 0 | 28 | 31.44 | sun Poon and Azhar (2003) |
| 500 | 758 | 927 | 0 | 150 | 2.5 | 28 | 68.72 | Poon et al. (2001) |
| 415 | 870 | 750 | 0 | 160 | 9.25 | 28 | 51.6 | Poon and Ho (2004) |
| 400 | 621 | 1157 | 0 | 160 | 6.4 | 28 | 29.84 | Lam et al. (1998) |
| 410 | 628 | 1048 | 0 | 225 | 0 | 28 | 37.68 | Poon et al. (2007) |
| 410 | 642 | 0 | 1017 | 225 | 0 | 28 | 30.48 | Kou et al. (2007) |
| 355 | 690 | 0 | 1068 | 195 | 0 | 28 | 28.48 | Kou and Poon (2008) |
| 350 | 706 | 0 | 1041 | 180 | 0 | 28 | 33.6 | Duan et al. (2018) |
| 429 | 737 | 0 | 909 | 193 | 0.94 | 28 | 27.28 | Schoppe (2011) |
| 525 | 668 | 0 | 923 | 184 | 3.67 | 28 | 58.16 | Kou and Poon (2015) |
| 379 | 590 | 0 | 1171 | 190 | 0 | 28 | 26.08 | Poon et al. (2004b) |
| 353 | 649 | 531 | 531 | 207 | 0 | 28 | 30.48 | Poon et al. (2004a) |
| 390 | 640 | 527 | 539 | 195 | 0 | 28 | 34.8 | Kou et al. (2011) |
| 495 | 861 | 800 | 20 | 174.73 | 7 | 7 | 46.14 | Mohammed et al. (2019) |
| 430 | 829 | 781 | 2 | 194 | 3 | 28 | 19.1 | Aslani et al. (2018) |
| 550 | 784 | 745 | 14 | 176 | 7 | 28 | 76.4 | Faraj et al. (2022) |
| 450 | 751 | 759 | 3.14 | 202.5 | 3.25 | 28 | 28.92 | Wan et al. (2018) |
| 444 | 517 | 387 | 4.76 | 195.36 | 2.5 | 90 | 18.7 | Ranjbar and Mousavi (2015) |
| 520 | 798 | 818 | 15.4 | 187.2 | 3.6 | 28 | 50 | Hilal and Kareem (2018) |
| 497 | 776.18 | 758.45 | 34.48 | 178.92 | 5.65 | 7 | 29.76 | Chunchu and Putta (2019) |
| 450 | 777.8 | 770 | 72.2 | 193.5 | 7.4 | 60 | 22.58 | Sadrmomtazi et al. (2016) |
| 523 | 633 | 300 | 26.2 | 188.28 | 5.3 | 28 | 24.72 | Yang et al. (2015) |
| 550 | 840.2 | 840.2 | 0 | 137.5 | 21.7 | 28 | 60.19 | Mohammed et al. (2022) |
| 480 | 896.1 | 800 | 0 | 173 | 7 | 28 | 65.3 | Mohamed (2018) |

*Continued*

https://doi.org/10.1017/dce.2026.10075 Published online by Cambridge University Press



***Table 1.*** *Continued*

| Binder content ($kg/m^3$) | Fine aggregate ($kg/m^3$) | Coarse aggregate ($kg/m^3$) | Recycled aggregate ($kg/m^3$) | Water ($kg/m^3$) | Superplasticizer ($kg/m^3$) | Curing time (Days) | Compressive strength (MPa) | References |
|---|---|---|---|---|---|---|---|---|
| 550 | 717 | 922 | 0 | 176 | 6 | 90 | 88.4 | Güneyisi et al. (2010) |
| 550 | 866 | 775 | 0 | 181.5 | 8.8 | 400 | 101.8 | Uysal and Sumer (2011) |
| 440 | 801 | 890 | 75 | 160.0 | 6.80 | 28 | 59.4 | Ali and Al-Tersawy (2012) |
| 410 | 482.5 | 820 | 482.5 | 185.0 | 2.50 | 28 | 34.18 | Bidabadi et al. (2020) |
| 420 | 810 | 702 | 468 | 180.0 | 0.00 | 28 | 49.48 | Poongodi et al. (2021) |
| 512.8 | 372 | 284.8 | 985.2 | 183.0 | 7.00 | 28 | 40.1 | Sadeghi-Nik et al. (2019) |
| 420 | 1204.4 | 401.4 | 119.5 | 168.0 | 4.45 | 28 | 35.7 | Sasanipour and Aslani (2019) |
| 469 | 761 | 656 | 206 | 191.0 | 4.20 | 28 | 35.4 | Revathi et al. (2013) |
| 590 | 650 | 1150 | 0 | 112.0 | 1.80 | 28 | 58.3 | Fiol et al. (2018) |
| 570 | 765 | 0 | 749.2 | 171.0 | 4.95 | 56 | 68.67 | Gesoglu et al. (2015) |
| 615 | 660 | 585 | 195 | 246.0 | 1.53 | 28 | 40.73 | Duan et al. (2020) |

***Table 2.*** *Descriptive statistics of the input and output variables used in model training*

| Notation | Input parameter | Mean | Standard deviation | Range/categories | | |
|---|---|---|---|---|---|---|
| | | | | Min | Max | Range |
| $X_1$ | Binder content (kg/m$^3$) | 475.2477 | 68.41153 | 300 | 702 | 402 |
| $X_2$ | Fine aggregate (kg/m$^3$) | 743.7917 | 158.3686 | 0 | 1204.4 | 1204.4 |
| $X_3$ | Coarse aggregate (kg/m$^3$) | 698.597 | 311.5101 | 0 | 1237 | 1237 |
| $X_4$ | Recycled aggregate (kg/m$^3$) | 185.8861 | 347.5521 | 0 | 1642 | 1642 |
| $X_5$ | Water (kg/m$^3$) | 182.1219 | 21.16723 | 112 | 271 | 159 |
| $X_6$ | Superplasticizer (kg/m$^3$) | 5.379407 | 4.7245 | 0 | 35.1 | 35.1 |
| $X_7$ | Curing time (Days) | 52.33432 | 70.25353 | 7 | 400 | 393 |
| $X_8$ | Compressive strength (MPa) | 48.83982 | 20.31599 | 9.89 | 105.7 | 95.81 |

To further understand the statistical characteristics of the dataset, Table 2 summarizes the mean, standard deviation, minimum, and maximum values of each input and output parameter. For instance, the average compressive strength across all mixes was 48.84 MPa, with a standard deviation of 20.32 MPa, highlighting the broad strength spectrum, from early-age low-strength SCC to high-performance concrete exceeding 100 MPa. Similarly, the recycled aggregate content ranged from 0 to 1642 kg/m$^3$, indicating that the dataset covers mixtures with partial to complete replacement of recycled aggregate. The records with zero recycled aggregate content correspond to control (reference) mixtures produced entirely with natural aggregate; these were intentionally retained so that the models learn behavior across the full range of replacement levels, from conventional SCC to fully recycled-aggregate mixes, and do not indicate a dataset inconsistency.

The wide dispersion and variability of the dataset not only support the development of generalizable ML models, but also provide a realistic representation of challenges faced in the practical design of SCC using recycled materials. The dataset's heterogeneity ensures that the trained models are applicable across different mix proportions, materials, and curing conditions, making the predictive framework developed in this study valuable for both academic research and industrial implementation in sustainable concrete design.

### *3.2. Data splitting and preprocessing*

To ensure unbiased model training and accurate performance evaluation, the compiled dataset comprising 582 SCC mix designs was systematically partitioned into three subsets: training (70%), validation (15%), and testing (15%). This stratified split ensures that each subgroup maintains the statistical diversity and distribution of the original dataset, including a broad range of compressive strengths and input variable combinations. The training set was used to develop and fit the models, the validation set was employed for hyperparameter tuning and model selection, and the testing set was reserved for final performance evaluation on unseen data.

Prior to model development, the dataset underwent essential preprocessing steps to improve model robustness and computational efficiency. All input variables were standardized using z-score normalization, which transforms each feature to have a mean of zero and a standard deviation of one. This step is critical for distance-based algorithms, such as KNN and RN, and also enhances the convergence and stability of gradient-based models, including XGBoost and CatBoost. The standardization was performed using statistics computed from the training set and applied consistently to the validation and testing sets to prevent data leakage.

During the preprocessing phase, the dataset was also examined for missing values, outliers, and redundancies. Entries with missing values or inconsistent units were excluded to ensure data integrity and accuracy. The database was additionally screened for exact duplicate records (entries identical across all input variables and the measured compressive strength); only a negligible number were identified

https://doi.org/10.1017/dce.2026.10075 Published online by Cambridge University Press

(approximately 0.7% of records), which does not materially affect the reported results. Records sharing identical input values but differing in measured strength were retained, as they correspond to the same mix design tested at different curing ages or to replicate specimens rather than to true duplicates. While no significant outliers were removed, extreme values were retained to preserve the natural variability in concrete mix design parameters, thereby enhancing the model's ability to generalize across a wide range of practical mix conditions.

The dataset encompasses a diverse range of mix characteristics, including binder content ranging from 300 to 702 kg/m$^3$, water content from 112 to 271 kg/m$^3$, recycled aggregate content from 0 to 1642 kg/m$^3$, and compressive strength values from 9.89 to 105.7 MPa (Figure 2). Such variability ensures that the ML models are trained on a sufficiently broad parameter space, allowing them to learn both linear and nonlinear relationships among input features.

To assess model generalizability, all performance metrics—$R^2$, RMSE, and MAE—were calculated separately for the training, validation, and testing subsets. This approach allowed the evaluation of each model's fitting behavior, overfitting tendencies, and real-world predictive capability. Furthermore, five-fold cross-validation was performed on the training set during hyperparameter optimization to reduce variance in performance estimates further.

Through careful data preprocessing and dataset splitting, this study aims to train and evaluate the ML models under realistic conditions. It should be noted that, because the database aggregates records from multiple source studies, observations originating from the same study may be correlated; the reported metrics

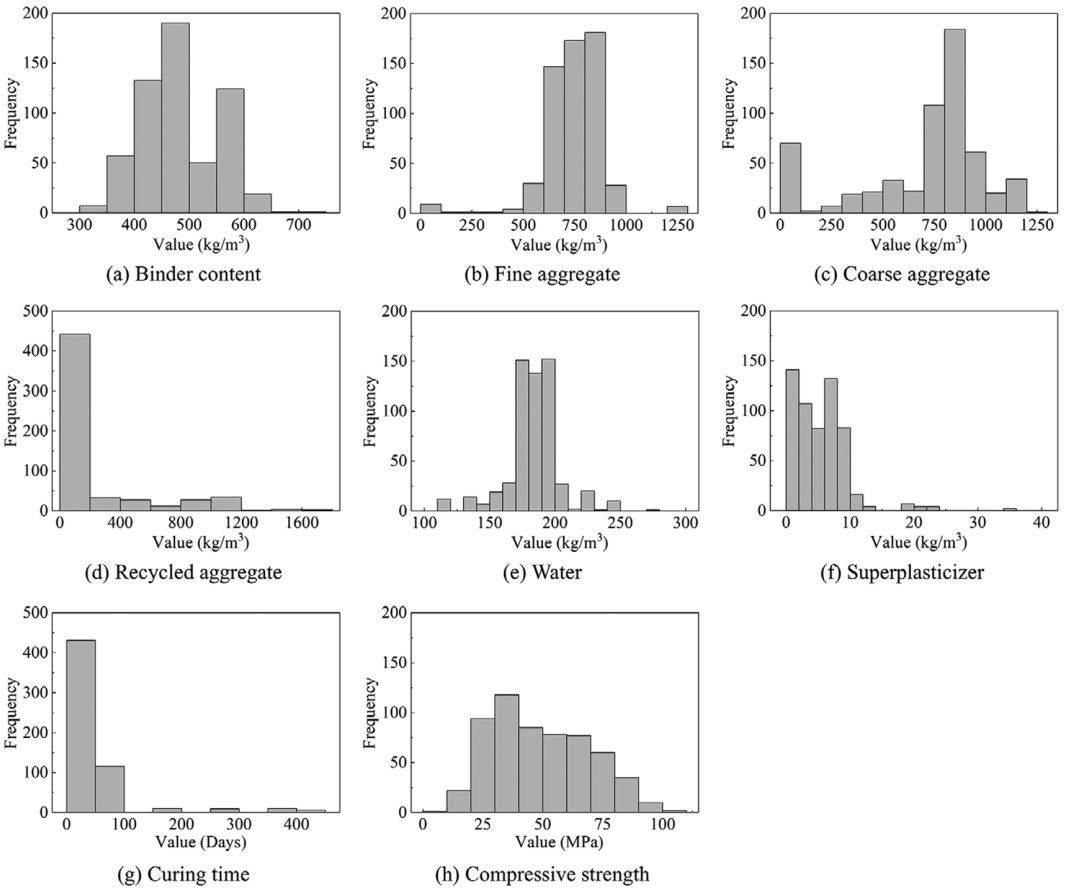



***Figure 2.*** *Distribution histograms of all input variables and output.*

https://doi.org/10.1017/dce.2026.10075 Published online by Cambridge University Press

therefore reflect performance under a random-split protocol and may be optimistic relative to a source-independent (grouped) evaluation. As the full dataset is publicly available, performance under alternative splitting schemes can be independently assessed, which we identify as a valuable direction for future work.

### *3.3. Model development*

The development of ML models in this study was structured around three algorithmic categories: bagging-based, boosting-based, and instance-based models. Each model was trained using the standardized dataset and optimized through grid search and five-fold cross-validation (Table 3). The primary objective during model development was to minimize prediction errors while maintaining generalizability across the wide variability of SCC mix designs with recycled aggregates.

Model training was conducted on the 70% training subset, and hyperparameters were tuned using the 15% validation set. Following optimization, the models were assessed on the independent 15% test set to evaluate their real-world predictive capacity. All models were developed using Python's scikit-learn and XGBoost libraries, and their predictions were compared against experimental compressive strength values using $R^2$, RMSE, and MAE.

#### *3.3.1. Bagging-based Models*

Bagging demonstrated optimal performance using 100 estimators, incorporating 75% of the features at each split ("max_features = 0.75") and employing the complete dataset for each base learner

***Table 3.*** *Optimized hyperparameters for all machine learning models*

| Model | Hyperparameter | Value |
|---|---|---|
| Bagging | max_features | 0.75 |
| | max_samples | 1 |
| | *n*_estimators | 100 |
| Extra trees | max_depth | 30 |
| | max_features | sqrt |
| | min_samples_leaf | 1 |
| | min_samples_split | 2 |
| | *n*_estimators | 150 |
| CatBoost | depth | 6 |
| | iterations | 300 |
| | learning_rate | 0.1 |
| AdaBoost | estimator__max_depth | 7 |
| | learning_rate | 0.1 |
| | *n*_estimators | 150 |
| XGBoost | colsample_bytree | 0.8 |
| | learning_rate | 0.05 |
| | max_depth | 7 |
| | *n*_estimators | 150 |
| | subsample | 1 |
| KNN | *n*_neighbors | 9 |
| | weights | distance |
| | metric ($p$) | manhattan ($p = 1$) |
| RN | radius | 5.0 |
| | weights | distance |
| | metric ($p$) | manhattan ($p = 1$) |

https://doi.org/10.1017/dce.2026.10075 Published online by Cambridge University Press

("max_samples = 1.0"). This approach effectively enhanced ensemble diversity and the stability of the model. Lastly, the grid search for the Extra Trees identified an optimal maximum depth of 30, accompanied by 150 estimators. The model utilized the square root of the total number of features for each split ("max_features = sqrt"), establishing a minimum sample requirement of 2 for splitting nodes and 1 for leaf nodes. The implemented settings proved to be effective in utilizing the model's randomization capabilities to minimize variance.

#### *3.3.2. Boosting-based models*

The grid search process for the AdaBoost model identified an optimal learning rate of 0.1, 150 estimators, and a maximum depth of 7 for the base estimator. This approach enabled the model to identify nonlinear patterns while successfully managing the risk of overfitting. Similarly, CatBoost exhibited notable efficiency when configured with 300 boosting iterations, a depth of 6, and a learning rate of 0.1. The parameters facilitated precise learning on tabular data while minimizing the need for extensive preprocessing. The optimal configuration for the XGBoost model was identified as a learning rate of 0.05, a maximum depth of 7, and 150 estimators. Furthermore, the model demonstrated optimal performance when utilizing a column subsampling rate of 0.8 ("colsample_bytree") alongside a row subsampling rate of 1.0 ("subsample"). These hyperparameters played a significant role in improving generalization while simultaneously minimizing the risk of overfitting.

#### *3.3.3. Instance-based models*

In addition to the ensemble models, two instance-based learning algorithms were developed and refined to enhance the prediction of the compressive strength of concrete made from recycled aggregates. The KNN model demonstrated optimal performance with the utilization of neighbors (n_neighbors = 9), the application of the Manhattan distance metric ($p$ = 1), and the implementation of a distance-based weighting scheme (weights = "distance"). This configuration enabled the algorithm to place greater emphasis on closer neighbors, thereby enhancing the local interpolation of compressive strength values. Similarly, the RN model demonstrated optimal performance when utilizing a fixed radius of 5.0, while also incorporating the Manhattan distance ($p$ = 1) alongside distance-based weighting. The chosen hyperparameters enabled both models to effectively adjust to the spatial distribution of data points, particularly in areas with varying densities, thereby improving their predictive accuracy.

## 4. Results and discussion

This study systematically evaluated the predictive capabilities of seven ML models: Extra Trees, Bagging, XGBoost, CatBoost, AdaBoost, KNN, and RN, in estimating the compressive strength of SCC that incorporates recycled aggregates. Previous studies have indicated that $R^2$ values falling between 0.60 and 0.75 are typically considered satisfactory. Values ranging from 0.75 to 0.95 suggest good predictive ability, while those exceeding 0.95 demonstrate excellent prediction performance. On the other hand, values that fall below 0.60 are regarded as indicative of unsatisfactory model performance (Marani and Nehdi, 2020). The established thresholds served as benchmarks for evaluating the accuracy and reliability of the models developed in this study. Alongside the performance evaluation, a sensitivity analysis utilizing the PFI was conducted to assess the relative impact of each input variable on the predicted compressive strength. The results of the model performance metrics and the sensitivity analysis are thoroughly presented and discussed in the following sections. Figure 3 provides an overall comparison of all seven models across $R^2$, RMSE, and MAE, and is introduced here to give the reader a consolidated overview before the category-wise analysis that follows.

### *4.1. Predictive performance of bagging-based models*

This study investigated two widely used ensemble learning models under the bagging paradigm: the Bagging Regressor and the Extra Trees Regressor. Both models are designed to reduce the variance of

https://doi.org/10.1017/dce.2026.10075 Published online by Cambridge University Press

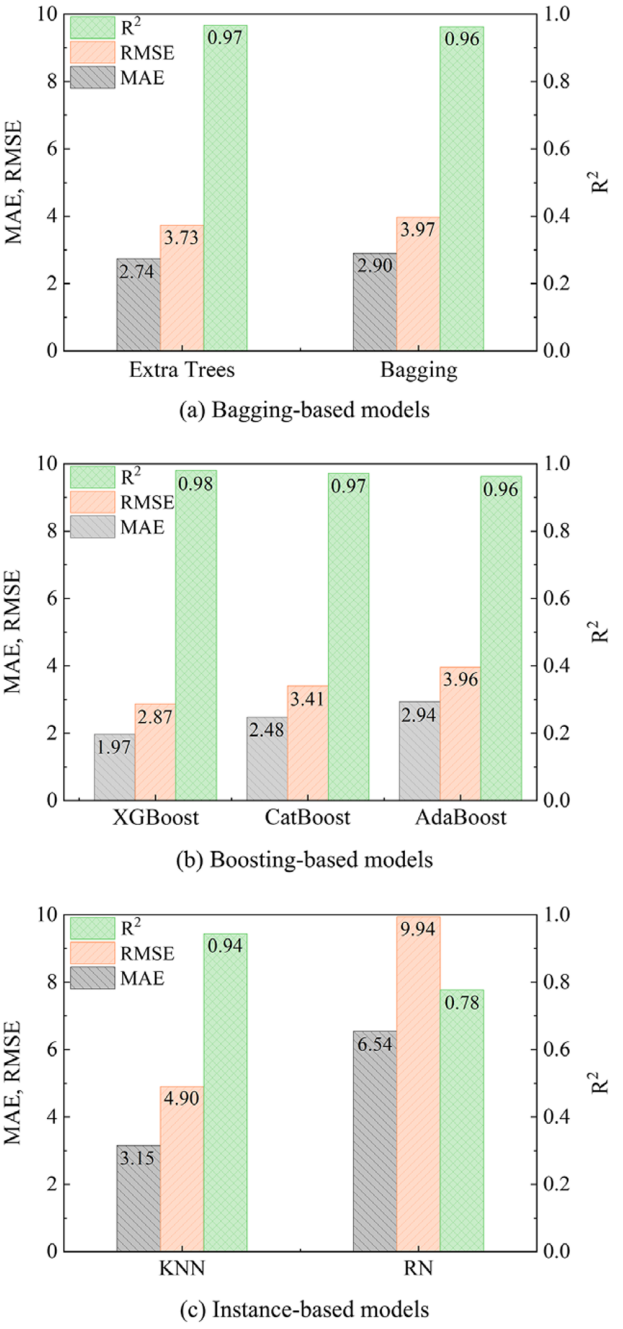



***Figure 3.*** *Comparison of model performance across three categories using* $R^2$*, RMSE, and MAE.*

predictions by combining multiple decision tree learners; however, they differ in their approach to handling randomness during training. The Bagging Regressor aggregates predictions from various decision trees trained on bootstrapped samples, whereas the Extra Trees algorithm introduces additional stochasticity by also randomizing the splitting thresholds during node formation. This distinction plays a crucial role in predictive performance.

https://doi.org/10.1017/dce.2026.10075 Published online by Cambridge University Press

***Table 4.*** *Model performance comparison using* $R^2$*, RMSE, and MAE*

| | | Bagging-based models | | Boosting-based models | | | Instance-based models | |
|---|---|---|---|---|---|---|---|---|
| Datasets | Metrics | Extra Trees | Bagging | XGBoost | CatBoost | AdaBoost | KNN | RN |
| Training | $R^2$ | 0.9648 | 0.9599 | 0.9756 | 0.9682 | 0.959 | 0.9782 | 0.9783 |
| | RMSE | 3.6788 | 3.9235 | 3.0593 | 3.4974 | 3.9665 | 2.8937 | 2.8873 |
| | MAE | 2.4158 | 2.6399 | 1.5967 | 2.2761 | 2.703 | 0.9284 | 0.9284 |
| Validation | $R^2$ | 0.974 | 0.9729 | 0.9887 | 0.9826 | 0.9748 | 0.949 | 0.6922 |
| | RMSE | 3.6015 | 3.6764 | 2.3718 | 2.9461 | 3.5438 | 5.0431 | 12.3939 |
| | MAE | 2.7087 | 2.6678 | 1.8896 | 2.2343 | 2.6992 | 3.7867 | 9.5057 |
| Testing | $R^2$ | 0.9615 | 0.9533 | 0.9754 | 0.9649 | 0.9529 | 0.9032 | 0.6601 |
| | RMSE | 3.9014 | 4.2983 | 3.1183 | 3.7273 | 4.3178 | 6.1906 | 11.5980 |
| | MAE | 3.1016 | 3.3879 | 2.4148 | 2.923 | 3.4083 | 4.7431 | 9.1992 |
| Average | $R^2$ | 0.9668 | 0.962 | 0.9799 | 0.9719 | 0.9623 | 0.9435 | 0.7768 |
| | RMSE | 3.7294 | 3.9743 | 2.8699 | 3.4061 | 3.9554 | 4.9034 | 9.9408 |
| | MAE | 2.742 | 2.8985 | 1.967 | 2.4778 | 2.9368 | 3.1527 | 6.5444 |

Based on the results in Table 4, the Extra Trees consistently exhibited superior performance across the training, validation, and testing datasets. On average, it achieved an $R^2$ of 0.9668, an RMSE of 3.7294 MPa, and an MAE of 2.742 MPa. In comparison, Bagging yielded slightly lower performance, with $R^2$ = 0.962, RMSE = 3.9743 MPa, and MAE = 2.8985 MPa. These results demonstrate the improved generalization of Extra Trees in modeling the compressive strength of SCC containing recycled aggregates.

The performance trends are visually confirmed in Figure 4a,b, which present predicted versus actual compressive strength values for Extra Trees and Bagging Regressors, respectively. The Extra Trees plot shows a tighter clustering of points along the ideal diagonal line, especially for mid- to high-strength concretes (40–80 MPa). Furthermore, Figure 4a displays a more uniform distribution of residuals, indicating the absence of significant underestimation or overestimation zones. The Bagging model, while generally effective, exhibits slightly more dispersion in the low and high strength ranges, suggesting a minor predictive bias under extreme mixing conditions.

From an engineering perspective, these results highlight the benefits of enhanced randomization in high-dimensional concrete datasets. The Extra Trees model effectively captures the intricate nonlinear behavior and interaction effects inherent in SCC mixtures containing recycled aggregates. Notably, SCC mixes with higher curing times or elevated binder content—which tend to produce more dispersed strength outcomes—are more accurately modeled by Extra Trees, owing to its ability to avoid overfitting while maintaining precision in regression splits.

### *4.2. Predictive performance of boosting-based models*

The boosting-based models examined in this study include XGBoost, CatBoost, and AdaBoost, each representing state-of-the-art ensemble learning techniques that can capture complex, nonlinear relationships in engineering data. These models improve prediction accuracy by sequentially training a series of weak learners—typically decision trees—where each subsequent model attempts to correct the errors of its predecessor. As shown in Table 4, among the boosting models, XGBoost demonstrated the highest overall accuracy across all three data subsets. It achieved an average $R^2$ of 0.9799, with an RMSE of 2.8699 MPa and an MAE of 1.967 MPa, outperforming both CatBoost and AdaBoost. XGBoost's predictive superiority is attributed to its use of second-order gradient optimization, regularization techniques (L1 and L2), and internal mechanisms to handle sparse data. These features enable it to generalize effectively across diverse mix design scenarios.

https://doi.org/10.1017/dce.2026.10075 Published online by Cambridge University Press

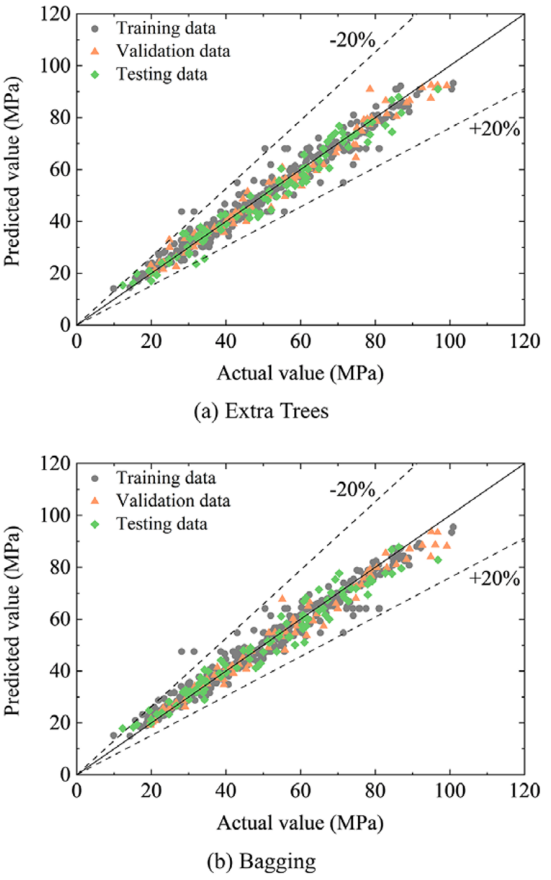



***Figure 4.*** *Predicted versus actual compressive strength plots for bagging-based models.*

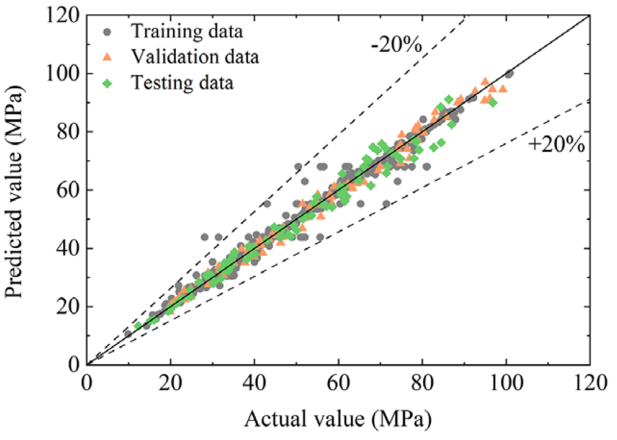
The CatBoost model also demonstrated strong predictive capability, with an average $R^2$ of 0.9719, RMSE of 3.4061 MPa, and MAE of 2.4778 MPa. Although slightly less accurate than XGBoost, CatBoost's competitive performance is noteworthy, especially given its built-in handling of categorical data and resistance to overfitting through ordered boosting. However, in this case, since all input variables are continuous and normalized, the edge CatBoost usually has with categorical inputs may not be fully leveraged.

AdaBoost, while demonstrating acceptable accuracy ($R^2$ = 0.9623, RMSE = 3.9554 MPa, and MAE = 2.9368 MPa), was clearly outperformed by XGBoost and CatBoost. AdaBoost's strategy of focusing on hard-to-predict samples can lead to overfitting in noisy datasets or when the weak learners (decision stumps or shallow trees) fail to capture high-order interactions. This is evident in Figure 5c, where AdaBoost predictions exhibit greater variance and larger residuals at both low and high compressive strength ranges.

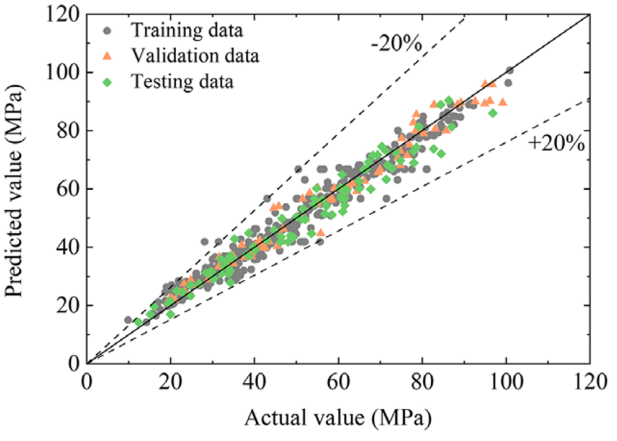
Visual comparisons are presented in Figure 5a–c. Figure 5a (XGBoost) illustrates an almost perfect alignment between the predicted and actual compressive strengths, with the majority of points tightly clustered around the ideal 1:1 line. The plot also reveals consistent performance across strength levels (10–100 MPa), emphasizing the model's ability to generalize across varied mix designs. Figure 5b (CatBoost) displays a similarly tight pattern, though with slightly more scatter, especially in the mid-strength range (30–60 MPa). In contrast, Figure 5c (AdaBoost) reveals larger prediction errors, especially in mixes with higher recycled aggregate content and extended curing times.

From an engineering standpoint, the exceptional performance of XGBoost confirms its suitability for modeling the complex relationships between concrete constituents and resulting compressive strength.

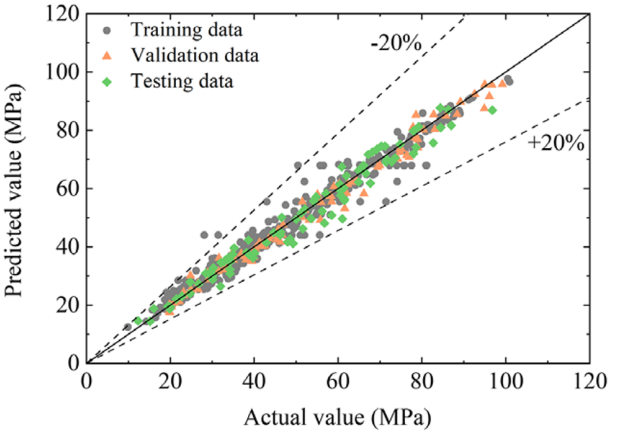
https://doi.org/10.1017/dce.2026.10075 Published online by Cambridge University Press




(a) XGBoost




(b) CatBoost




(c) AdaBoost

***Figure 5.*** *Predicted versus actual compressive strength plots for boosting-based models.*

https://doi.org/10.1017/dce.2026.10075 Published online by Cambridge University Press

The regularization terms in XGBoost prevent overfitting, making it especially effective when dealing with datasets collected from various sources with different experimental setups and material characteristics. This robustness is vital in practice, where mix design data often exhibits heterogeneity due to variations in recycled aggregate quality, moisture absorption, and curing environments.

### *4.3. Predictive performance of instance-based models*

The supervised instance-based models evaluated in this study include the KNN algorithm and its variant, the RN model. Unlike ensemble models, supervised instance-based approaches are non-parametric and rely directly on the training dataset during prediction. According to Table 4, the KNN model demonstrated moderately strong predictive performance, achieving an average $R^2$ of 0.9435, RMSE of 4.9034 MPa, and MAE of 3.1527 MPa. The model performed well in the training phase ($R^2$ = 0.9782 and MAE = 0.9284 MPa), indicating high fit to known data. However, performance deteriorated during validation and testing, particularly in terms of error magnitudes, suggesting signs of overfitting and limited generalization. KNN's reliance on local neighbors makes it prone to error when faced with sparsely populated or highly nonlinear input regions—a common issue when modeling SCC mixes with extreme proportions of recycled aggregate or unusually long curing durations.

The RN model, in contrast, yielded substantially lower predictive power. Its average $R^2$ was only 0.7768, with an RMSE of 9.9408 MPa and an MAE of 6.5444 MPa, indicating that the model struggled to produce accurate predictions across all datasets. This poor performance is likely due to inconsistent neighborhood density within the high-dimensional feature space. Many test points either had too few neighbors within the fixed radius or encountered overly heterogeneous local patterns, leading to unstable predictions. In some cases, the RN model effectively defaulted to averaging weakly related points, amplifying noise and bias.

Figure 6a,b offer visual validation of these observations. In Figure 6a, the KNN-predicted strengths align reasonably with actual values in the mid-strength range (30–70 MPa), but outliers become more pronounced at both ends of the spectrum. The RN model (Figure 6b) exhibits a more dispersed pattern, with many predictions deviating significantly from the ideal 1:1 line, particularly in mixes with high recycled aggregate content, where material heterogeneity leads to significant local variance.

From an engineering perspective, while KNN may be adequate for exploratory analysis or cases where rapid prediction from well-clustered data is required, it does not offer the level of reliability necessary for robust prediction of compressive strength across diverse SCC mix designs. The weakness of RN further reinforces this conclusion, showing that instance-based models are ill-suited for tasks where data heterogeneity, interaction effects, and nonlinearity dominate—conditions typical in recycled aggregate concrete design.

### *4.4. Predictive performance of optimal ML models*

To identify the most accurate and reliable predictive model for predicting the compressive strength of SCC with recycled aggregates, a comparative evaluation of the top-performing algorithms was conducted. This included the best bagging-based model (Extra Trees), the leading boosting-based model (XGBoost), and the strongest instance-based model (K-Nearest Neighbors), as summarized in Table 5.

Among these, XGBoost emerged as the overall best performer, consistently delivering superior results across the training, validation, and testing phases. Specifically, XGBoost achieved:

- $R^2$ = 0.9756 (training), 0.9887 (validation), and 0.9754 (testing)
- RMSE = 3.0593 MPa, 2.3718 MPa, and 3.1183 MPa, respectively
- MAE = 1.5967 MPa, 1.8896 MPa, and 2.4148 MPa, respectively

This level of performance reflects XGBoost's ability to model complex, nonlinear relationships and subtle interaction effects between input variables, which is critical in SCC systems with recycled aggregates, where compressive strength is influenced by interdependent factors such as binder type, recycled aggregate content, curing conditions, and superplasticizer dosage.

https://doi.org/10.1017/dce.2026.10075 Published online by Cambridge University Press

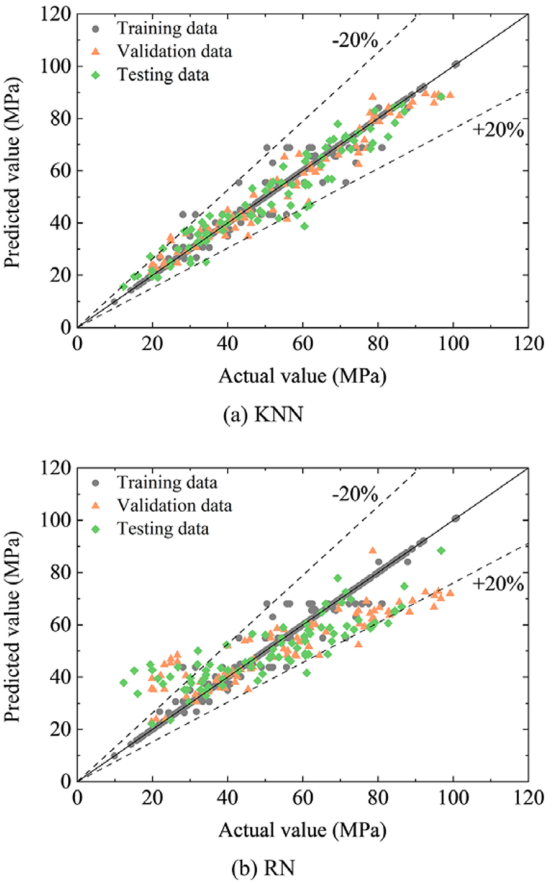



***Figure 6.*** *Predicted versus actual compressive strength plots for instance-based models.*

***Table 5.*** *Summary of top-performing models by category and their prediction accuracy*

| | | Best models | | |
|---|---|---|---|---|
| Datasets | Metrics | Extra trees | XGBoost | KNN |
| Training | $R^2$ | 0.9648 | 0.9756 | 0.9782 |
| | RMSE | 3.6788 | 3.0593 | 2.8937 |
| | MAE | 2.4158 | 1.5967 | 0.9284 |
| Validation | $R^2$ | 0.974 | 0.9887 | 0.949 |
| | RMSE | 3.6015 | 2.3718 | 5.0431 |
| | MAE | 2.7087 | 1.8896 | 3.7867 |
| Testing | $R^2$ | 0.9615 | 0.9754 | 0.9032 |
| | RMSE | 3.9014 | 3.1183 | 6.1906 |
| | MAE | 3.1016 | 2.4148 | 4.7431 |
| Average | $R^2$ | 0.9668 | 0.9799 | 0.9435 |
| | RMSE | 3.7294 | 2.8699 | 4.9034 |
| | MAE | 2.742 | 1.967 | 3.1527 |

https://doi.org/10.1017/dce.2026.10075 Published online by Cambridge University Press

The Extra Trees, although slightly less accurate, maintained high performance with an average $R^2$ of 0.9668 and low error metrics (RMSE = 3.7294 MPa and MAE = 2.742 MPa). It proved more stable than KNN and less sensitive to data scaling. Its advantage lies in its randomized tree splits and efficient ensemble averaging, which provide robustness against noisy data. On the other hand, the KNN model, while demonstrating strong training accuracy ($R^2$ = 0.9782), suffered from notable performance degradation in the validation and testing subsets ($R^2$ = 0.949 and 0.9032, respectively), with corresponding increases in RMSE and MAE. This overfitting tendency reflects KNN's sensitivity to the density and distribution of training samples. Its reliance on local proximity without a generalized learning framework restricts its utility for high-dimensional, heterogeneous SCC datasets.

Figure 3 illustrates the overall comparison across all models, while Figures 4–6 provide predicted versus actual plots highlighting the precision of each model's outputs. Notably, Figure 5a (XGBoost) shows nearly perfect alignment along the 1:1 line, while Figures 4a (Extra Trees) and 6a (KNN) exhibit more spread in residuals, particularly in regions corresponding to low- or high-strength mixes. Sample-wise prediction errors (Figure 7) support this conclusion, as XGBoost consistently maintains low error magnitudes, whereas KNN and, especially, RN exhibit high variability.

In practical engineering terms, the superiority of XGBoost translates to more reliable performance under varying mix designs, including:

- High variability in recycled aggregate quality
- Diverse curing durations (ranging from 7 to 400 days)
- Mixes with and without superplasticizer use
- Wide binder content range (300–702 kg/m$^3$)

The regularization capabilities of XGBoost (through L1 and L2 penalties), its ability to handle collinearity and noise, and its gradient-based optimization make it a fit-for-purpose tool in real-world applications where data inconsistencies and multidimensionality are typical. Moreover, its minimal tuning overhead and high interpretability, thanks to built-in feature importance metrics, make it highly suitable for deployment in design-stage decision support systems for sustainable SCC production. In practice, the model can be applied in two ways. First, at the mix-design stage, it serves as a rapid screening tool, allowing engineers to estimate the compressive strength of candidate proportions and narrow the design space before committing to laboratory trials. Second, because it relies on only seven routinely available mix-proportion inputs, the trained model can be integrated directly into a spreadsheet or a lightweight software tool to provide instant strength estimates within an existing mix-design workflow.

In summary, XGBoost is the recommended model for predicting the strength of SCC with recycled aggregates, offering a balance of predictive power, robustness, and practical utility across diverse material and environmental conditions. It should be emphasized that the model is intended for use within the documented range of the training data (e.g., binder 300–702 kg/m$^3$, recycled aggregate 0–1642 kg/m$^3$, and curing 7–400 days). As the ensemble models cannot reliably extrapolate beyond these bounds, predictions for inputs outside the training domain should be regarded as indicative only and verified experimentally.

## 5. Input parameters and their effects

### *5.1. Feature importance analysis*

To further understand the underlying relationships between input parameters and the predicted compressive strength of SCC containing recycled aggregates, a PFI analysis was conducted using the XGBoost model, which was previously identified as the most accurate among the ML models evaluated. The PFI technique involves systematically shuffling each input variable and quantifying the resulting increase in prediction error, thus indicating how heavily the model relies on each feature for accurate predictions. PFI was chosen because it is model-agnostic, directly grounded in predictive performance, computationally efficient, and readily interpretable by practitioners. It should be noted, however, that PFI can be affected

https://doi.org/10.1017/dce.2026.10075 Published online by Cambridge University Press

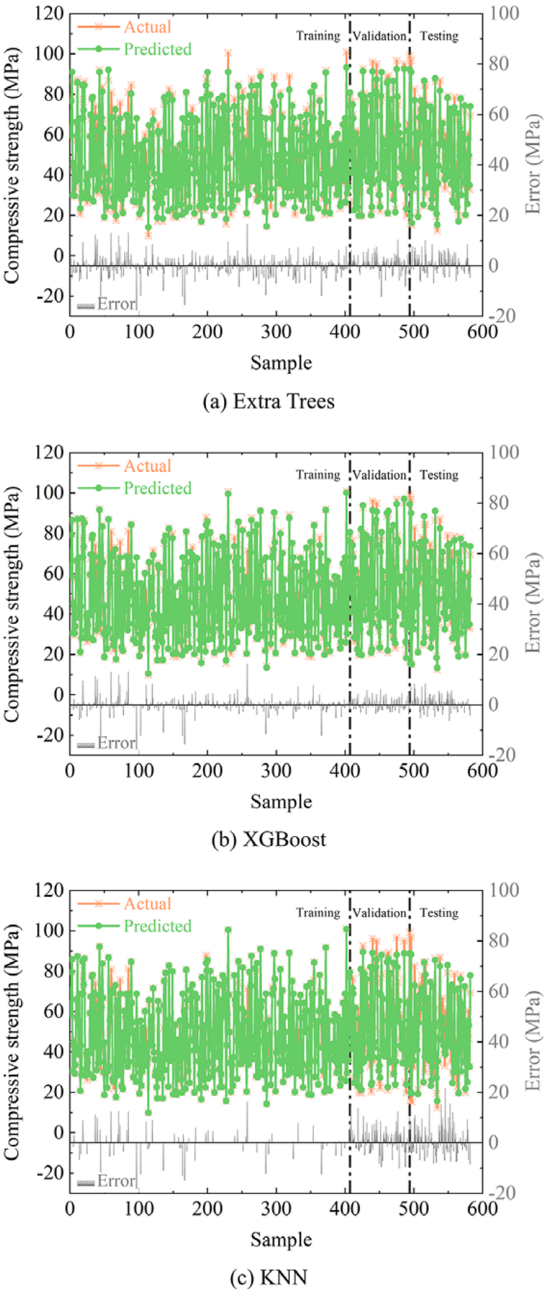



***Figure 7.*** *Sample-wise prediction errors for top models in each category.*

by correlation among input variables, since permuting one feature may create physically implausible combinations and redistribute importance among correlated predictors; results are therefore interpreted at the level of broad parameter influence rather than exact rank order. Alternative attribution methods such as SHAP offer additive, locally consistent explanations but at higher computational cost and under their own dependence assumptions, and represent a worthwhile direction for future work.

According to the results summarized in Table 6 and visualized in Figure 8, binder content was by far the most influential variable, producing a permutation-induced MAE increase of approximately 11.05 MPa

https://doi.org/10.1017/dce.2026.10075 Published online by Cambridge University Press

***Table 6.*** *PFI results showing the effect of individual input variables on compressive strength prediction*

| Input parameter | PFI | $MAE_{Perm}$ |
|---|---|---|
| Binder content (kg/m$^3$) | 0.633420435 | 11.05359394 |
| Water (kg/m$^3$) | 0.256107346 | 7.492656324 |
| Coarse aggregate (kg/m$^3$) | 0.244573898 | 7.431748137 |
| Curing time (Days) | 0.135342857 | 5.004351431 |
| Recycled aggregate (kg/m$^3$) | 0.092099077 | 4.692300509 |
| Fine aggregate (kg/m$^3$) | 0.068643885 | 4.258144094 |
| Superplasticizer (kg/m$^3$) | 0.027678152 | 3.083275328 |

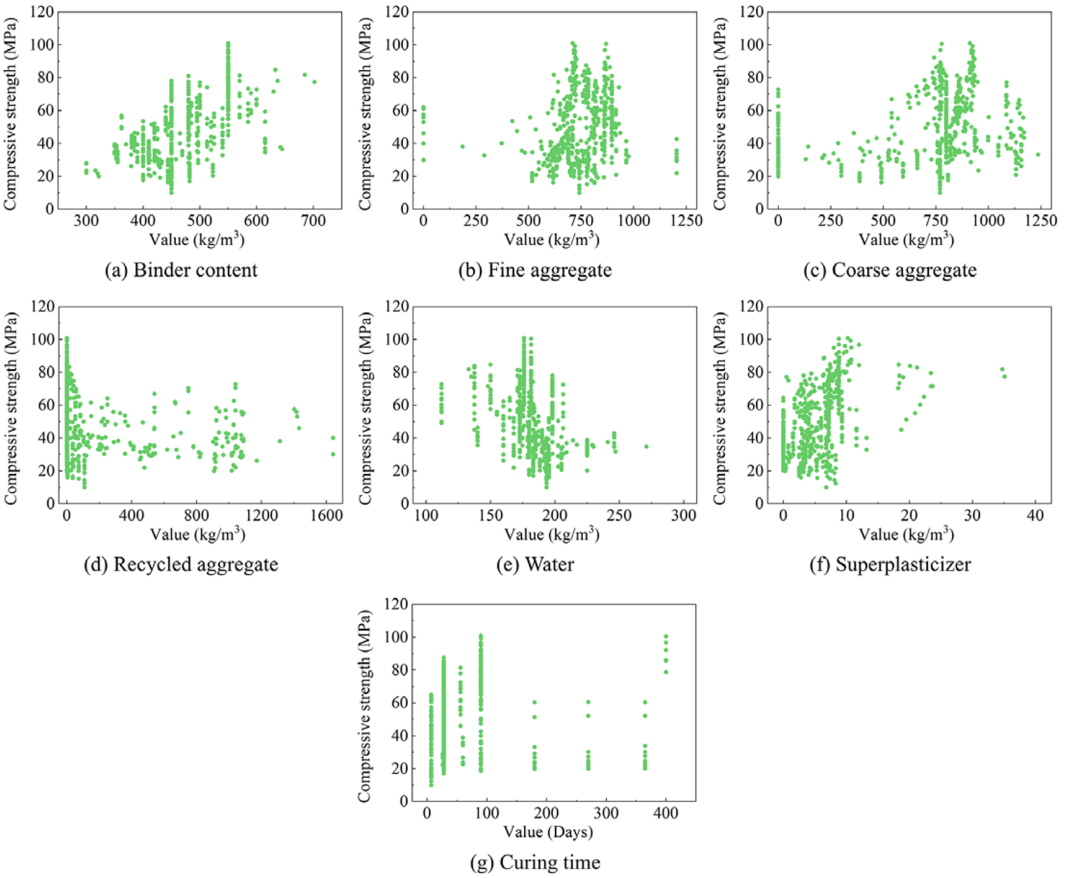



***Figure 8.*** *PFI results showing the relative impact of each input variable on compressive strength prediction.*

(normalized PFI = 0.633)—more than 1.4 times the effect of the next most important feature. Water content (≈ 7.49 MPa; 0.256) and coarse aggregate (≈ 7.43 MPa; 0.245) followed at comparable levels, while curing time (≈ 5.00 MPa; 0.135) and recycled aggregate content (≈ 4.69 MPa; 0.092) had moderate influence. Fine aggregate (≈ 4.26 MPa; 0.069) and superplasticizer dosage (≈ 3.08 MPa; 0.028) were the least influential, each contributing well below half the effect of binder content. This is consistent with conventional concrete mix design knowledge, as the binder governs paste volume and directly affects hydration reactions, strength

https://doi.org/10.1017/dce.2026.10075 Published online by Cambridge University Press

development, and microstructural densification. In SCC mixes using recycled aggregates, which often have irregular surfaces and lower intrinsic strength, the role of binder becomes even more critical in compensating for potential weaknesses in the interfacial transition zone (ITZ).

The second most important variable was water content. Its strong influence on the model reflects the well-established importance of the water-to-binder ratio in determining porosity, workability, and ultimately strength. Especially in SCC, where high flowability is required, water content must be carefully balanced to avoid excessive bleeding and segregation, both of which can degrade mechanical properties. The model's high sensitivity to this parameter indicates its role not just in fresh concrete behavior but also in the development of compressive strength.

Coarse aggregate content followed closely in importance. Its impact is rooted in its function as the primary load-bearing component in concrete. In the case of SCC with recycled aggregates, the variability in recycled coarse aggregate properties, such as residual mortar and water absorption, complicates predictions. Nevertheless, the model was able to account for these complexities effectively. Curing time was also recognized as a significant contributor to prediction accuracy. This is particularly notable for SCC, which incorporates recycled materials, where strength gain can be delayed or exhibit non-standard behavior due to the influence of RA on hydration kinetics and internal curing dynamics.

The content of recycled aggregate itself was found to have a moderate but meaningful influence. While high levels of recycled aggregates are generally associated with reduced strength due to weaker aggregate-matrix bonds and elevated porosity, the model's ability to compensate for these effects through other parameters such as binder and water content suggests a nuanced interaction that traditional empirical models may fail to capture. It should also be noted that binder content, water content, and recycled aggregate content are physically interrelated rather than independent: the high water absorption of recycled aggregate modifies the effective water available for hydration, coupling its effect with both the water content and the binder demand needed to preserve the effective water-to-binder ratio. Although PFI reports the marginal influence of each variable and does not isolate such interactions explicitly, the tree-ensemble models capture them implicitly through hierarchical splitting, consistent with the models' ability to offset high recycled aggregate content through binder and water adjustments. Fine aggregate content and superplasticizer dosage were among the least important variables. Fine aggregates, while important for rheology and particle packing, play a relatively minor role in determining the strength of hardened concrete. Superplasticizer, although vital for flowability in SCC, typically has an indirect effect on strength unless overdosed, which can lead to segregation or unexpected reductions in water content.

From an engineering perspective, these findings provide valuable insights into which mix parameters should be prioritized during SCC design involving recycled aggregates. The model not only affirms established concrete science but also reveals the complex interplay between variables, enabling data-driven optimization of sustainable concrete mixes. The apparent dominance of binder, water, and coarse aggregate content in governing strength confirms the central role of paste quality and aggregate interactions, particularly in recycled systems. The application of PFI thus provides a transparent and interpretable approach to understanding and refining ML predictions in concrete engineering.

### *5.2. Engineering interpretation and literature comparison*

The results of this study offer several valuable engineering insights into the behavior of SCC incorporating recycled aggregates, particularly in relation to how ML models capture and interpret the underlying physical mechanisms that influence compressive strength. The superior performance of ensemble models, especially XGBoost, demonstrates that data-driven approaches can successfully model the complex, nonlinear interactions among binder content, aggregate proportions, water dosage, curing time, and admixtures. From an engineering standpoint, this is particularly significant for recycled aggregate concrete, where variability in material properties and deviations from traditional mix designs present challenges to conventional empirical or semi-empirical models.

The prominence of binder content in the feature importance analysis aligns with the long-established role of paste quality and volume in determining concrete strength, especially in SCC, where the mix must

https://doi.org/10.1017/dce.2026.10075 Published online by Cambridge University Press

flow without segregation. This finding is consistent with those of Dinakar et al. (2013) and Kou and Poon (2008), who observed that increasing the binder fraction can mitigate the adverse effects of using recycled aggregates by enhancing matrix density and improving the quality of the ITZ. Likewise, the strong influence of water content and curing time observed in this study supports the work of Saha and Rajasekaran (2016), who reported that recycled aggregates' mixes are particularly sensitive to variations in moisture conditions and curing periods due to the porous nature and high water absorption capacity of recycled aggregates.

The model's nuanced treatment of recycled aggregate content, where it is neither the most dominant nor negligible, reflects the balancing act between sustainability goals and structural performance. Prior studies, such as those by Jagadesh et al. (2021) and Bui et al. (2017), reported that while increased recycled aggregate content typically reduces strength, its effects can be effectively managed through proportioning and supplementary cementitious materials. The ML models in this study appeared to capture this compensatory mechanism, particularly in scenarios where high binder content and proper curing neutralized the potential strength loss due to recycled aggregates.

When comparing model performance across the literature, this study's XGBoost model achieved an average $R^2$ of 0.9799 and MAE as low as 1.96 MPa, outperforming the ANN models reported by Gesoglu et al. (2015) and Ridho et al. (2021), which typically yielded $R^2$ values in the range of 0.90–0.95 with higher error margins. Moreover, unlike many previous works that focused on a narrow dataset or a single ML algorithm, this research incorporates 582 diverse mix designs sourced from over 35 independent studies. This provides robustness and generalizability, making the findings applicable across a broad spectrum of real-world SCC applications.

In addition, this study supports the growing consensus in the literature, as highlighted by de Prado-Gil et al. (2022) and Ekanayake et al. (2022)—that ensemble learning techniques, particularly gradient boosting methods, offer superior accuracy and interpretability for concrete property prediction. The integration of PFI further enhances transparency, allowing engineers to understand and trust the model's decision-making process—a critical step toward the widespread adoption of ML tools in structural material design.

## 6. Conclusion

This study presented a comprehensive approach for predicting the compressive strength of SCC incorporating recycled aggregates using multiple supervised ML models. A large and diverse experimental database of 582 SCC mix designs was employed, capturing a wide range of binder compositions, aggregate proportions, curing times, and admixture dosages. Seven ML models, belonging to bagging, boosting, and instance-based families, were trained and evaluated using standard performance metrics including $R^2$, RMSE, and MAE.

Among all models, XGBoost demonstrated the highest predictive accuracy and consistency across training, validation, and testing datasets, achieving an average $R^2$ of 0.9799 and MAE of just 1.97 MPa. The model not only outperformed other boosting techniques, such as CatBoost and AdaBoost, but also surpassed the performance of bagging-based models, like Extra Trees, and instance-based models, like KNN. The XGBoost model's strength lies in its ability to capture nonlinear relationships and interactions among variables while maintaining robustness and generalizability through embedded regularization techniques.

The feature importance analysis, performed using PFI, revealed that binder content, water content, and coarse aggregate quantity were the most influential parameters affecting compressive strength predictions. This finding is consistent with established engineering knowledge, reaffirming that strength in SCC mixes is highly dependent on paste quality and aggregate packing. Curing time and recycled aggregate content also played meaningful roles, highlighting the delayed hydration characteristics and material variability associated with the use of recycled aggregates.

By integrating advanced ML with domain-specific understanding of concrete technology, this study illustrates the practical value of data-driven tools in guiding sustainable construction practices. The XGBoost

https://doi.org/10.1017/dce.2026.10075 Published online by Cambridge University Press

model, in particular, offers a reliable framework for predicting SCC strength across a wide range of mix compositions, thereby facilitating more efficient, economical, and environmentally responsible mix design processes.

Several limitations should be acknowledged. The framework predicts compressive strength only and does not capture durability parameters, and it does not explicitly represent supplementary cementitious materials or mineral admixtures. Because the dataset is compiled from more than 35 independent studies, it carries inherent inter-study heterogeneity in materials and testing protocols; predictions are valid only within the documented range of the training data; and, as the models were evaluated using a random data partition, performance under a source-independent (grouped) validation scheme may differ. A portion of the records originates from a limited number of research groups, which may bias the models toward the mix-design conventions and testing practices of those groups and could affect generalization to mixtures produced under different conditions; source-aware validation is identified as a means of evaluating this in future work. Future work may explore the incorporation of additional input features such as mineral admixtures, curing temperature, and durability indices, the adoption of source-aware validation, or the extension of the model to forecast other mechanical and durability properties of SCC containing recycled aggregates, including multi-output prediction.

**Data availability statement.** The dataset supporting the findings of this study, including all input parameters and output variables required to reproduce the reported results, is publicly available in the Zenodo repository at: https://doi.org/10.5281/zenodo.18459938 (Khan et al., 2026d).

**Author contribution.** Conceptualization: A.Q.K.; Data curation: A.Q.K., M.D.R.; Formal analysis: A.Q.K, M.H.N.; Investigation: A.Q.K., M.H.N.; Methodology: A.Q.K., M.D.R.; Project administration: M.H.N., A.P.; Resources: A.P.; Supervision: A.P.; Visualization: A.Q.K., M.D.R.; Writing – original draft: A.Q.K., M.D.R.; Writing – review & editing: M.H.N., A.P.

**Funding statement.** This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

**Competing interests.** The authors have no relevant financial or non-financial interests to disclose.

## References

**Ahmad S** (2008) A study of self-compacting concrete made with marginal aggregates. *Arabian Journal for Science ...* [Online]. Available at https://www.academia.edu/30259383/A_study_of_self_compacting_concrete_made_with_marginal_aggregates (accessed 24 May 2025).

**Ali ZA**, **Abduljabbar ZH**, **Tahir HA**, **Bibo Sallow A and Almufti SM** (2023) eXtreme gradient boosting algorithm with machine learning: A review. *Academic Journal of Nawroz University 12*(2), 320–334. https://doi.org/10.25007/AJNU.V12N2A1612.

**Ali EE and Al-Tersawy SH** (2012) Recycled glass as a partial replacement for fine aggregate in self compacting concrete. *Construction and Building Materials 35*, 785–791. https://doi.org/10.1016/j.conbuildmat.2012.04.117.

**Aslani F**, **Ma G**, **Wan DLY and Muselin G** (2018) Development of high-performance self-compacting concrete using waste recycled concrete aggregates and rubber granules. *Journal of Cleaner Production 182*, 553–566. https://doi.org/10.1016/J.JCLEPRO.2018.02.074.

**Asteris PG**, **Skentou AD**, **Bardhan A**, **Samui P and Pilakoutas K** (2021) Predicting concrete compressive strength using hybrid ensembling of surrogate machine learning models. *Cement and Concrete Research*. https://doi.org/10.1016/j.cemconres.2021.106449.

**Bakshi K and Bakshi K** (2018) Considerations for artificial intelligence and machine learning: Approaches and use cases. IEEE Aerospace Conference, 1–9. https://doi.org/10.1109/AERO.2018.8396488.

**Bidabadi MS**, **Akbari M and Panahi O** (2020) Optimum mix design of recycled concrete based on the fresh and hardened properties of concrete. *Journal of Building Engineering 32*, 101483. https://doi.org/10.1016/j.jobe.2020.101483.

**Bühlmann P and Yu B** (2002) Analyzing bagging. *The Annals of Statistics 30*(4), 927–961. https://doi.org/10.1214/aos/1031689014.

**Bui NK**, **Satomi T and Takahashi H** (2017) Improvement of mechanical properties of recycled aggregate concrete basing on a new combination method between recycled aggregate and natural aggregate. *Construction and Building Materials 148*, 376–385. https://doi.org/10.1016/J.CONBUILDMAT.2017.05.084.

**Buja A and Stuetzle W** (2002) Observations on Bagging. [Online]. Available at http://stat.wharton.upenn.edu/~buja/PAPERS/sinica-bagging-buja-stuetzle.pdf (accessed 24 May 2025).

**Chaabene WB**, **Flah M and Nehdi ML** (2020) Machine learning prediction of mechanical properties of concrete: Critical review. *Construction and Building Materials 260*, 119889. https://doi.org/10.1016/J.CONBUILDMAT.2020.119889.

**Chen T and Guestrin C** (2016) XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), San Francisco, CA, USA, 13–17 August 2016*, pp. 785–794. New York, NY, USA: Association for Computing Machinery (ACM). https://doi.org/10.1145/2939672.2939785.

**Chicco D**, **Warrens MJ and 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/SUPP-1.

**Chunchu BRK and Putta J** (2019) Rheological and strength behavior of binary blended SCC replacing partial fine aggregate with plastic E-waste as high impact polystyrene. *Buildings 9*(2), 50. https://doi.org/10.3390/BUILDINGS9020050.

**de-Prado-Gil J**, **Palencia C**, **Silva-Monteiro N and Martínez-García R** (2022) To predict the compressive strength of self compacting concrete with recycled aggregates utilizing ensemble machine learning models. *Case Studies in Construction Materials 16*, e01046. https://doi.org/10.1016/J.CSCM.2022.E01046.

**Dinakar P**, **Sethy KP and Sahoo UC** (2013) Design of self-compacting concrete with ground granulated blast furnace slag. *Materials and Design 43*, 161–169. https://doi.org/10.1016/J.MATDES.2012.06.049.

**Dorogush AV**, **Ershov V and Gulin A** (2018) CatBoost: Gradient Boosting with Categorical Features Support. arXiv preprint arXiv:1810.11363. https://doi.org/10.48550/arXiv.1810.11363.

**Duan S**, **Liao H**, **Ma Z**, **Cheng F**, **Fang L**, **Gao H and Yang H** (2018) The relevance of ultrafine fly ash properties and mechanical properties in its fly ash-cement gelation blocks via static pressure forming. *Construction and Building Materials 186*, 1064–1071. https://doi.org/10.1016/J.CONBUILDMAT.2018.08.035.

**Duan Z**, **Singh A**, **Xiao J and Hou S** (2020) Combined use of recycled powder and recycled coarse aggregate derived from construction and demolition waste in self-compacting concrete. *Construction and Building Materials 254*, 119323. https://doi.org/10.1016/j.conbuildmat.2020.119323.

**Ekanayake IU**, **Meddage DPP and Rathnayake U** (2022) A novel approach to explain the black-box nature of machine learning in compressive strength predictions of concrete using Shapley additive explanations (SHAP). *Case Studies in Construction Materials 16*, e01059. https://doi.org/10.1016/J.CSCM.2022.E01059.

**Elith J**, **Leathwick JR and Hastie T** (2008) A working guide to boosted regression trees. *Journal of Animal Ecology 77*(4), 802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x.

**Etxeberria M**, **Vázquez E**, **Marí A and Barra M** (2007) Influence of amount of recycled coarse aggregates and production process on properties of recycled aggregate concrete. *Cement and Concrete Research 37*(5), 735–742. https://doi.org/10.1016/J.CEMCONRES.2007.02.002.

**Fafalios S**, **Charonyktakis P and Tsamardinos I** (2020) *Gradient Boosting Trees*. Gnosis Data Analysis PC, Heraklion, Greece.

**Faraj RH**, **Mohammed AA**, **Mohammed A**, **Omer KM and Ahmed HU** (2022) Systematic multiscale models to predict the compressive strength of self-compacting concretes modified with nanosilica at different curing ages. *Engineering with Computers 38*(S3), 2365–2388. https://doi.org/10.1007/S00366-021-01385-9.

**Fiol F**, **Thomas C**, **Muñoz C**, **Ortega-López V and Manso JM** (2018) The influence of recycled aggregates from precast elements on the mechanical properties of structural self-compacting concrete. *Construction and Building Materials 182*, 309–323. https://doi.org/10.1016/j.conbuildmat.2018.06.132.

**Freund Y and Schapire R** (1999) A Short Introduction to Boosting.

**Gangula R and Venkateswarlu B** (2024) Exploring the power and practical applications of K-nearest neighbours (KNN) in machine learning. *Journal of Computer Allied Intelligence 2*(1), 8–15. https://doi.org/10.69996/JCAI.2024002.

**Gesoglu M**, **Güneyisi E**, **Öz HÖ**, **Taha I and Yasemin MT** (2015) Failure characteristics of self-compacting concretes made with recycled aggregates. *Construction and Building Materials 98*, 334–344. https://doi.org/10.1016/j.conbuildmat.2015.08.036.

**Geurts P**, **Ernst D and Wehenkel L** (2006) Extremely randomized trees. *Machine Learning 63*(1), 3–42. https://doi.org/10.1007/s10994-006-6226-1.

**Gontia P**, **Thuvander L**, **Ebrahimi B**, **Vinas V**, **Rosado L and Wallbaum H** (2019) Spatial analysis of urban material stock with clustering algorithms: A Northern European case study. *Journal of Industrial Ecology 23*(6), 1328–1343. https://doi.org/10.1111/JIEC.12939.

**Güneyisi E**, **Gesolu M and Özbay E** (2010) Strength and drying shrinkage properties of self-compacting concretes incorporating multi-system blended mineral admixtures. *Construction and Building Materials 24*(10), 1878–1887. https://doi.org/10.1016/J.CONBUILDMAT.2010.04.015.

**Hancock JT and Khoshgoftaar TM** (2020) CatBoost for big data: An interdisciplinary review. *Journal of Big Data 7*(1), 1–45. https://doi.org/10.1186/S40537-020-00369-8/FIGURES/9.

**Hilal MTNNN and Kareem Q** (2018) Influence of polyethylene waste on some fresh& mechanical properties of self - compacting concrete. *Journal of Engineering and Applied Sciences 13*(14), 10901–10911. https://doi.org/10.3923/JEASCI.2018.10901.10911.

**Huang P**, **Dai K and Yu X** (2023) Machine learning approach for investigating compressive strength of self-compacting concrete containing supplementary cementitious materials and recycled aggregate. *Journal of Building Engineering 79*, 107904. https://doi.org/10.1016/J.JOBE.2023.107904.

**Jagadesh P**, **de Prado-Gil J**, **Silva-Monteiro N and Martínez-García R** (2023) Assessing the compressive strength of self-compacting concrete with recycled aggregates from mix ratio using machine learning approach. *Journal of Materials Research and Technology 24*, 1483–1498. https://doi.org/10.1016/j.jmrt.2023.03.037.

https://doi.org/10.1017/dce.2026.10075 Published online by Cambridge University Press

**Jagadesh P**, **Juan-Valdés A**, **Guerra-Romero MI**, **Pozo JMM-D**, **García-González J and Martínez-García R** (2021) Effect of design parameters on compressive and split tensile strength of self-compacting concrete with recycled aggregate: An overview. *Applied Sciences 11*(13), 6028. https://doi.org/10.3390/APP11136028.

**Joshi A**, **Saggar P**, **Jain R**, **Sharma M**, **Gupta D and Khanna A** (2021) CatBoost—An ensemble machine learning model for prediction and classification of student academic performance. *Advances in Data Science and Adaptive Analysis 13*. https://doi.org/10.1142/S2424922X21410023.

**Jovel J and Greiner R** (2021) An introduction to machine learning approaches for biomedical research. *Frontiers in Medicine 8*. https://doi.org/10.3389/fmed.2021.771607.

**Khan AQ**, **Naveed MH**, **Rasheed MD and Miao P** (2023a) Prediction of compressive strength of fly ash-based geopolymer concrete using supervised machine learning methods. *Arabian Journal for Science and Engineering*. https://doi.org/10.1007/s13369-023-08283-w.

**Khan AQ**, **Awan HA**, **Rasul M**, **Siddiqi ZA and Pimanmas A** (2023b) Optimized artificial neural network model for accurate prediction of compressive strength of normal and high strength concrete. *Cleaner Materials 10*, 100211. https://doi.org/10.1016/J.CLEMA.2023.100211.

**Khan AQ**, **Naveed MH**, **Rasheed MD and Pimanmas A** (2024) Prediction of stress–strain behavior of PET FRP-confined concrete using machine learning models. *Arabian Journal for Science and Engineering*. https://doi.org/10.1007/s13369-024-09086-3.

**Khan AQ**, **Muhammad SG**, **Raza A and Pimanmas A** (2025a) Advanced machine learning techniques for predicting mechanical properties of eco-friendly self-compacting concrete. *Journal of Road Engineering 5*(2), 213–229.

**Khan AQ**, **Muhammad SG**, **Raza A**, **Chaimahawan P and Pimanmas A** (2025b) Advanced machine learning techniques for predicting compressive strength of ultra-high performance concrete. *Frontiers of Structural and Civil Engineering 19*(4), 503–523. https://doi.org/10.1007/s11709-025-1169-4.

**Khan AQ**, **Naveed MH**, **Rasheed MD and Pimanmas A** (2025c) Enhancing predictive accuracy in shear strength of RC deep beams: A comprehensive analysis using ensemble machine learning models. *Arabian Journal for Science and Engineering 51*, 12783–12801. https://doi.org/10.1007/s13369-025-10763-0.

**Khan AQ**, **Rasheed MD**, **Naveed MH and Pimanmas A** (2026a) Accurate compressive strength prediction of fly ash geopolymers using advanced ensemble models and Morris analysis. *Computer Modeling in Engineering & Sciences 147*(2), 1–10. https://doi.org/10.32604/cmes.2026.083654.

**Khan AQ**, **Muhammad SG**, **Raza A and Pimanmas A** (2026b) Machine learning models for predicting carbonation depth in fly ash concrete: Performance and interpretability insights. *Journal of Road Engineering 6*(1), 74–90.

**Khan AQ**, **Rasul M and Pimanmas A** (2026c) Domain-specific prediction of punching shear capacity in slab-column connections using explainable XGBoost models and SHAP analysis. *Data-Centric Engineering 7*, e6. https://doi.org/10.1017/dce.2026.10041.

**Khan AQ**, **Rasheed MD**, **Naveed MH and Pimanmas A** (2026d) Dataset for: Data-Driven Modeling of Compressive Strength in Sustainable Self-Compacting Concrete Incorporating Recycled Aggregates Using Ensemble Learning Techniques, Zenodo, V1. https://doi.org/10.5281/zenodo.18459938.

**Kou SC and Poon CS** (2008) Mechanical properties of 5-year-old concrete prepared with recycled aggregates obtained from three different sources. *Magazine of Concrete Research 60*(1), 57–64. https://doi.org/10.1680/macr.2007.00052.

**Kou SC and Poon CS** (2015) Effect of the quality of parent concrete on the properties of high performance recycled aggregate concrete. *Construction and Building Materials 77*, 501–508. https://doi.org/10.1016/J.CONBUILDMAT.2014.12.035.

**Kou SC**, **Poon CS and Agrela F** (2011) Comparisons of natural and recycled aggregate concretes prepared with the addition of different mineral admixtures. *Cement and Concrete Composites 33*(8), 788–795. https://doi.org/10.1016/J.CEMCONCOMP.2011.05.009.

**Kou SC**, **Poon CS and Chan D** (2007) Influence of fly ash as cement replacement on the properties of recycled aggregate concrete. *Journal of Materials in Civil Engineering 19*(9), 709–717. https://doi.org/10.1061/(ASCE)0899-1561(2007)19:9(709).

**Kramer O** (2013) K-nearest Neighbors. In: *Dimensionality Reduction with Unsupervised Nearest Neighbors. Intelligent Systems Reference Library*, Vol. *51*. Berlin, Heidelberg: Springer, pp. 13–23. https://doi.org/10.1007/978-3-642-38652-7_2

**Lam L**, **Wong YL and Poon CS** (1998) Effect of fly ash and silica fume on compressive and fracture behaviors of concrete. *Cement and Concrete Research 28*(2), 271–283. https://doi.org/10.1016/S0008-8846(97)00269-X.

**Liaw A and Wiener M** (2002) Classification and Regression by randomForest. *R News 2*(3), 18–22.

**Liu K**, **Alam MS**, **Zhu J**, **Zheng J and Chi L** (2021) Prediction of carbonation depth for recycled aggregate concrete using ANN hybridized with swarm intelligence algorithms. *Construction and Building Materials 301*, 124382. https://doi.org/10.1016/J.CONBUILDMAT.2021.124382.

**Loh WY** (2011) Classification and regression trees. *WIREs Data Mining and Knowledge Discovery 1*(1), 14–23. https://doi.org/10.1002/WIDM.8.

**Marani A and Nehdi ML** (2020) Machine learning prediction of compressive strength for phase change materials integrated cementitious composites. *Construction and Building Materials 265*, 120286. https://doi.org/10.1016/J.CONBUILDMAT.2020.120286.

**McNeil K and Kang THK** (2013) Recycled concrete aggregates: A review. *International Journal of Concrete Structures and Materials 7*(1), 61–69. https://doi.org/10.1007/s40069-013-0032-5.

**Mienye ID and Sun Y** (2022) A survey of ensemble learning: Concepts, algorithms, applications, and prospects. *IEEE Access 10*, 99129–99149. https://doi.org/10.1109/ACCESS.2022.3207287.

https://doi.org/10.1017/dce.2026.10075 Published online by Cambridge University Press

**Mohamed O** (2018) Durability and compressive strength of high cement replacement ratio self-consolidating concrete. *Buildings 8*(11), 153. https://doi.org/10.3390/BUILDINGS8110153.

**Mohammed MK**, **Al-Hadithi AI and Mohammed MH** (2019) Production and optimization of eco-efficient self compacting concrete SCC with limestone and PET. *Construction and Building Materials 197*, 734–746. https://doi.org/10.1016/J.CONBUILDMAT.2018.11.189.

**Mohammed AM**, **Asaad DS and Al-Hadithi AI** (2022) Experimental and statistical evaluation of rheological properties of self-compacting concrete containing fly ash and ground granulated blast furnace slag. *Journal of King Saud University - Engineering Sciences 34*(6), 388–397. https://doi.org/10.1016/J.JKSUES.2020.12.005.

**Nafees A**, **Amin MN**, **Khan K**, **Nazir K**, **Ali M**, **Javed MF**, **Aslam F**, **Musarat MA and Vatin NI** (2021) Modeling of mechanical properties of silica fume-based green concrete using machine learning techniques. *Polymers 14*(1). https://doi.org/10.3390/polym14010030.

**Nasir M**, **Gazder U**, **Maslehuddin M**, **Al-Amoudi OSB and Syed IA** (2020) Prediction of properties of concrete cured under hot weather using multivariate regression and ANN models. *Arabian Journal for Science and Engineering 45*(5), 4111–4123. https://doi.org/10.1007/s13369-020-04403-y.

**Nunez I**, **Marani A**, **Flah M and Nehdi ML** (2021) Estimating compressive strength of modern concrete mixtures using computational intelligence: A systematic review. *Construction and Building Materials 310*, 125279. https://doi.org/10.1016/J.CONBUILDMAT.2021.125279.

**Poon CS**, **Azhar S**, **Anson M and Wong YL** (2001) Comparison of the strength and durability performance of normal- and high-strength pozzolanic concretes at elevated temperatures. *Cement and Concrete Research 31*(9), 1291–1300. https://doi.org/10.1016/S0008-8846(01)00580-4.

**Poon CS and Ho DWS** (2004) A feasibility study on the utilization of r-FA in SCC. *Cement and Concrete Research 34*(12), 2337–2339. https://doi.org/10.1016/J.CEMCONRES.2004.02.013.

**Poon CS**, **Kou SC and Lam L** (2002) Pore size distribution of high performance metakaolin concrete. *Journal Wuhan University of Technology, Materials Science Edition 17*(1), 42–46. https://doi.org/10.1007/BF02852633.

**Poon CS**, **Kou SC and Lam L** (2007) Influence of recycled aggregate on slump and bleeding of fresh concrete. *Materials and Structures/Materiaux et Constructions 40*(9), 981–988. https://doi.org/10.1617/S11527-006-9192-Y/METRICS.

**Poon CS**, **Lam L and Wong YL** (2000) A study on high strength concrete prepared with large volumes of low calcium fly ash. *Cement and Concrete Research 30*(3), 447–455. https://doi.org/10.1016/S0008-8846(99)00271-9.

**Poon CS**, **Shui ZH**, **Lam L**, **Fok H and Kou SC** (2004a) Influence of moisture states of natural and recycled aggregates on the slump and compressive strength of concrete. *Cement and Concrete Research 34*(1), 31–36. https://doi.org/10.1016/S0008-8846(03)00186-8.

**Poon CS**, **Shui ZH and Lam L** (2004b) Effect of microstructure of ITZ on compressive strength of concrete prepared with recycled aggregates. *Construction and Building Materials 18*(6), 461–468. https://doi.org/10.1016/J.CONBUILDMAT.2004.03.005.

**Poongodi K**, **Murthi P**, **Awoyera PO**, **Gobinath R and Olalusi OB** (2021) Durability properties of self-compacting concrete made with recycled aggregate for pavement application. *Silicon*, *13*(8), 2727–2735. https://doi.org/10.1007/s12633-020-00635-7

**Rahal K** (2007) Mechanical properties of concrete with recycled coarse aggregate. *Building and Environment 42*(1), 407–415. https://doi.org/10.1016/J.BUILDENV.2005.07.033.

**Rajakarunakaran SA**, **Lourdu AR**, **Muthusamy S**, **Panchal H**, **Jawad Alrubaie A**, **Musa Jaber M**, **Ali MH**, **Tlili I**, **Maseleno A**, **Majdi A and Ali SHM** (2022) Prediction of strength and analysis in self-compacting concrete using machine learning based regression techniques. *Advances in Engineering Software 173*, 103267. https://doi.org/10.1016/j.advengsoft.2022.103267.

**Ranjbar MM and Mousavi SY** (2015) Strength and durability assessment of self-compacted lightweight concrete containing expanded polystyrene. *Materials and Structures/Materiaux et Constructions 48*(4), 1001–1011. https://doi.org/10.1617/S11527-013-0210-6/METRICS.

**Revathi P**, **Selvi RS and Velin SS** (2013) Investigations on fresh and hardened properties of recycled aggregate self compacting concrete. *Journal of The Institution of Engineers 94*(3), 179–185. https://doi.org/10.1007/s40030-014-0051-5.

**Ridho BKAMA**, **Ngamkhanong C**, **Wu Y and Kaewunruen S** (2021) Recycled aggregates concrete compressive strength prediction using artificial neural networks (ANNs). *Infrastructures 6*(2), 17. https://doi.org/10.3390/INFRASTRUCTURES6020017.

**Rossikova Y**, **Li JJ and Morreale P** (2016) Intelligent Data Mining for Translator Correctness Prediction. In: Qiu M (ed.) *Proceedings of the 2nd IEEE International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS), New York, NY, USA, 9–10 April 2016*. Piscataway, NJ, USA: IEEE, 394–399. https://doi.org/10.1109/BigDataSecurity-HPSC-IDS.2016.19

**Sadeghi-Nik A**, **Berenjian J**, **Alimohammadi S**, **Lotfi-Omran O**, **Sadeghi-Nik A and Karimaei M** (2019) The effect of recycled concrete aggregates and metakaolin on the mechanical properties of self-compacting concrete containing nanoparticles. *Iranian Journal of Science and Technology - Transactions of Civil Engineering 43*(S1), 503–515. https://doi.org/10.1007/s40996-018-0182-4.

**Sadrmomtazi A**, **Dolati-Milehsara S**, **Lotfi-Omran O and Sadeghi-Nik A** (2016) The combined effects of waste polyethylene terephthalate (PET) particles and pozzolanic materials on the properties of self-compacting concrete. *Journal of Cleaner Production 112*, 2363–2373. https://doi.org/10.1016/J.JCLEPRO.2015.09.107.

https://doi.org/10.1017/dce.2026.10075 Published online by Cambridge University Press

**Saha S and Rajasekaran C** (2016) Mechanical properties of recycled aggregate concrete produced with Portland pozzolana cement. *Advances in Concrete Construction 4*(1), 27–35. https://doi.org/10.12989/ACC.2016.4.1.027.

**Saraswat P and Raj S** (2021) A brief review on machine learning and its various techniques. *International Journal of Innovative Research in Computer Science & Technology*. https://doi.org/10.55524/ijircst.2021.9.6.25.

**Sasanipour H and Aslani F** (2019) Effect of specimen shape, silica fume, and curing age on durability properties of self-compacting concrete incorporating coarse recycled concrete aggregates. *Construction and Building Materials 228*, 117054. https://doi.org/10.1016/j.conbuildmat.2019.117054.

**Schapire RE** (1999) A Brief Introduction to Boosting. In: *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI–99), Stockholm, Sweden, 31 July–6 August 1999*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., Vol. *2*, 1401–1406.

**Schoppe BM** (2011) *Shrinkage & Modulus of Elasticity in Concrete with Recycled Aggregates*, *Master's Theses*, https://doi.org/10.15368/theses.2011.62.

**Shi C**, **Wu Z**, **Lv K and Wu L** (2015) A review on mixture design methods for self-compacting concrete. *Construction and Building Materials 84*, 387–398. https://doi.org/10.1016/J.CONBUILDMAT.2015.03.079.

**Solomatine DP and Shrestha DL** (2004) AdaBoost.RT: a boosting algorithm for regression problems, *IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541) 2*, 1163–1168. https://doi.org/10.1109/IJCNN.2004.1380102.

**sun Poon C and Azhar S** (2003) Deterioration and recovery of metakaolin blended concrete subjected to high temperature. *Fire Technology 39*(1), 35–45. https://doi.org/10.1023/A:1021775110075.

**Taunk K**, **De S**, **Verma S and Swetapadma A** (2019) A brief review of nearest neighbor algorithm for learning and classification. *International Conference on Intelligent Computing and Control Systems, ICCS 2019*, 1255–1260, https://doi.org/10.1109/ICCS45141.2019.9065747.

**Tu C**, **Liu H and Xu B** (2017) AdaBoost typical algorithm and its application research. *MATEC Web of Conferences 139*. https://doi.org/10.1051/MATECCONF/201713900222.

**Uysal M and Sumer M** (2011) Performance of self-compacting concrete containing different mineral admixtures. *Construction and Building Materials 25*(11), 4112–4120. https://doi.org/10.1016/J.CONBUILDMAT.2011.04.032.

**Van Thieu N** (2024) PerMetrics: A framework of performance metrics for machine learning models. *The Journal of Open Source Software 9*(95), 6143. https://doi.org/10.21105/JOSS.06143.

**Wan DSLY**, **Aslani F and Ma G** (2018) Lightweight self-compacting concrete incorporating perlite, scoria, and polystyrene aggregates. *Journal of Materials in Civil Engineering 30*(8), 04018178. https://doi.org/10.1061/(ASCE)MT.1943-5533.0002350.

**Xiao J**, **Li W and Poon C** (2012) Recent studies on mechanical properties of recycled aggregate concrete in China-a review. *SCIENCE CHINA Technological Sciences 55*(6), 1463–1480. https://doi.org/10.1007/s11431-012-4786-9.

**Yang S**, **Yue X**, **Liu X and Tong Y** (2015) Properties of self-compacting lightweight concrete containing recycled plastic particles. *Construction and Building Materials 84*, 444–453. https://doi.org/10.1016/J.CONBUILDMAT.2015.03.038.

**Zhang Z** (2016) Introduction to machine learning: k-nearest neighbors. *Annals of Translational Medicine 4*(11), 218. https://doi.org/10.21037/ATM.2016.03.37.

**Zhang Y**, **Liu J and Shen W** (2022) A review of ensemble learning algorithms used in remote sensing applications. *Applied Sciences 12*. https://doi.org/10.3390/app12178654.

**Cite this article:** Khan AQ, Rasheed MD, Naveed MH and Pimanmas A (2026). Data-driven modeling of compressive strength in sustainable self-compacting concrete incorporating recycled aggregates using ensemble learning techniques. *Data-Centric Engineering*, 7, e34. doi:10.1017/dce.2026.10075

https://doi.org/10.1017/dce.2026.10075 Published online by Cambridge University Press