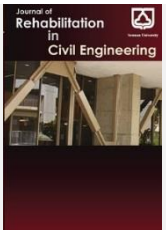

Journal of Rehabilitation in Civil Engineering

Journal homepage: https://civiljournal.semnan.ac.ir/

Sustainable Recycled Aggregate Concrete Design Using XGBoost, Data Augmentation, and Deterministic Multi-Objective Optimization

Nader Biglarijoo [1]; Shahaboddin Kharazmi [2,*]; Mohammadreza Danaei [3]; Amin Shams [1,*]

1. Assistant Professor, Faculty of Civil Engineering, Semnan University, Semnan, Iran
2. Assistant Professor, Faculty of Mechanical Engineering, Semnan University, Semnan, Iran
3. M.Sc. Student, Faculty of Civil Engineering, Semnan University, Semnan, Iran
* Corresponding author: *amin.shams@semnan.ac.ir* (A.Sh); *kharazmi@semnan.ac.ir* (Sh.Kh)

ARTICLE INFO

Article history:
Received: 06 July 2026
Revised: 25 July 2026
Accepted: 18 August 2026

Keywords:
Machine learning;
Multi-objective optimization;
Predictive model;
Recycled concrete;
Sustainable development.

ABSTRACT

Construction and demolition waste and excessive natural aggregate use have increased demand for sustainable concrete technologies. This study proposes an AI-assisted framework for predicting and optimizing recycled aggregate concrete (RAC). Using a Central Composite Design, 20 concrete mixtures were produced based on parent concrete strength, recycled coarse aggregate (RCA) replacement ratio, and cement content. To address limited data, a surrogate-based augmentation method combining Gradient Boosting Regression and Gaussian noise injection expanded the dataset to 500 realistic samples. An XGBoost model predicted compressive strength, splitting tensile strength, and water absorption with high accuracy (R^2 = 0.991, 0.980, and 0.998). Feature analysis showed that parent concrete quality mainly controls mechanical performance, while RCA ratio strongly influences water absorption. Optimization of 125,000 virtual mixtures identified an optimal RAC mix containing about 61% RCA and 320 kg/m³ cement, achieving balanced strength, durability, sustainability, and reduced experimental workload. The findings show that RCA could replace natural aggregates to maintain sustainable development in the construction industry.

E-ISSN: 2345-4423
© 2026 The Authors. Journal of Rehabilitation in Civil Engineering, published by Semnan University Press.
This is an open access article under the CC-BY 4.0 license. (https://creativecommons.org/licenses/by/4.0/)

How to cite this article:
Biglarijoo, N., Kharazmi, Sh., Danaei, M., Shams, A. (2027). Sustainable Recycled Aggregate Concrete Design Using XGBoost, Data Augmentation, and Deterministic Multi-Objective Optimization. Journal of Rehabilitation in Civil Engineering, 15(1), 2709. https://doi.org/10.22075/jrce.2026.2709

1. Introduction

The construction industry is one of the largest consumers of natural resources and a major contributor to environmental degradation, energy consumption, and greenhouse gas emissions [1]. Rapid urbanization and infrastructure expansion have significantly increased the demand for natural aggregates and cementitious materials, resulting in excessive extraction of virgin resources and large quantities of construction and demolition waste (CDW) [2]. Consequently, the development of sustainable and resource-efficient construction materials has become an important research priority [3].

Recycled concrete aggregate (RCA) produced from processed CDW has emerged as a promising alternative to natural coarse aggregate. The use of RCA can reduce landfill disposal, conserve natural resources, and promote circular economy principles through material recovery and waste reutilization. As a result, recycled aggregate concrete (RAC) has attracted increasing attention as a sustainable construction material for future infrastructure applications [4–6]. The use of RCA provides significant sustainability benefits by reducing the demand for virgin aggregates, conserving natural resources, and diverting construction and demolition waste from landfills. It also supports circular economy principles by extending material life cycles and reducing the environmental impacts of aggregate extraction, transportation, and disposal. Consequently, RAC has emerged as an effective strategy for improving resource efficiency while supporting sustainable infrastructure development.

Despite its environmental benefits, the structural application of RAC remains challenging because recycled aggregates exhibit different physical and microstructural characteristics compared with natural aggregates [7–9]. RCA particles generally contain adhered mortar, higher porosity, increased water absorption, and weaker interfacial transition zones (ITZs), which significantly influence the mechanical and durability properties of concrete [10]. In addition, the quality of RCA strongly depends on the characteristics of the parent concrete, leading to highly heterogeneous and nonlinear material behavior. Therefore, accurate prediction and optimization of RAC performance using conventional empirical equations or simplified regression models remain difficult [11].

Traditional trial-and-error experimental approaches for RAC mix design are also time-consuming and costly [12]. These limitations have motivated increasing interest in machine learning (ML) techniques capable of modeling complex nonlinear relationships between concrete mixture parameters and engineering properties. In recent years, ML has become a powerful tool for RAC mix design and property prediction by capturing complex nonlinear relationships among mixture parameters. Algorithms such as Artificial Neural Network (ANN), Support Vector Regression (SVR), Random Forest (RF), XGBoost, LightGBM, and CatBoost have been successfully applied to predict the mechanical and durability properties of RAC. More recently, machine learning has been integrated with optimization algorithms and explainable AI (XAI) techniques, particularly SHAP, to identify sustainable mix proportions and improve model interpretability. These advances highlight the growing potential of AI-assisted approaches for sustainable RAC design. Among available ML algorithms, Extreme Gradient Boosting (XGBoost) has demonstrated strong predictive capability due to its robustness against overfitting, regularization mechanisms, and ability to capture multidimensional interactions within heterogeneous datasets, often outperforming traditional algorithms in structural applications [13,14].

Although previous studies have successfully applied machine learning methods to concrete property prediction, most investigations focus primarily on predictive accuracy using unstructured historical databases [15–17]. Comparatively limited attention has been devoted to integrating structured experimental design, interpretable machine learning, and optimization frameworks for sustainable RAC development. Furthermore, conventional statistical approaches such as Response Surface Methodology (RSM) often fail to capture the severe nonlinear degradation mechanisms associated with high RCA replacement levels [18–20].

To address these limitations, this study proposes a unified experimental–computational framework for sustainable RAC design by integrating Central Composite Design (CCD), surrogate-based data augmentation, XGBoost modeling, sensitivity analysis, and deterministic multi-objective optimization [21–23]. Initially, a structured experimental database consisting of 20 laboratory mixtures was developed using Design-Expert software. To overcome data scarcity, a physically guided augmentation strategy incorporating Gaussian-noise injection was employed to generate additional virtual samples within experimentally bounded parameter ranges. Subsequently, XGBoost models were developed to predict compressive strength, splitting tensile strength, and water absorption of RAC mixtures. Feature importance analysis was further implemented to evaluate the relative contribution of mixture parameters and improve model interpretability. Finally, a deterministic optimization framework was employed to identify sustainable and mechanically reliable RAC mixtures across a high-resolution virtual design space containing 125,000 hypothetical combinations.

Recent advances in AI have significantly enhanced concrete mix design by enabling accurate prediction of material properties, efficient optimization of mixture proportions, and data-driven decision-making. In particular, ensemble learning algorithms, including XGBoost, LightGBM, and CatBoost, have demonstrated superior performance in modeling the complex nonlinear behavior of cementitious materials. At the same time, explainable artificial intelligence (XAI) techniques, such as SHAP (SHapley Additive exPlanations), have become increasingly important for interpreting machine learning predictions and quantifying the influence of individual mixture parameters. Furthermore, the integration of machine learning with optimization algorithms has emerged as a promising approach for identifying sustainable concrete mixtures that simultaneously satisfy mechanical, durability, economic, and environmental objectives.

Although machine learning has been widely applied to predict recycled aggregate concrete (RAC) properties, most studies have focused on predictive accuracy using historical datasets, with limited attention to integrating experimental design, physically consistent data augmentation, model interpretability, and optimization into a unified framework. Furthermore, conventional Response Surface Methodology (RSM) often fails to capture the highly nonlinear behavior of RAC, particularly at high recycled aggregate replacement levels.

To address these gaps, this study proposes an integrated experimental–computational framework combining Central Composite Design (CCD), surrogate-based data augmentation, Extreme Gradient Boosting (XGBoost), SHAP-based interpretability, and deterministic multi-objective optimization. The proposed approach enables accurate prediction, interpretable analysis, and efficient optimization of RAC mixtures while reducing experimental effort.

Accordingly, the main objectives of this study are: (i) to develop a structured experimental database for RAC using CCD considering parent concrete strength, RCA replacement ratio, and cement content as governing variables; (ii) to establish an XGBoost-based predictive framework for estimating compressive strength, tensile strength, and water absorption of RAC mixtures; (iii) to evaluate the relative influence of mixture parameters through sensitivity and feature importance analysis; (iv) to perform deterministic multi-objective optimization for identifying sustainable and mechanically reliable RAC mixtures with balanced mechanical and durability performance.

2. Materials and methods

2.1. Materials

2.1.1. Aggregates

Two categories of aggregates were utilized in this study: natural aggregates and recycled concrete aggregates (RCA). Natural aggregates were employed as both fine aggregate (natural sand) and coarse

aggregate (natural gravel), whereas RCA was used exclusively as a partial replacement for natural coarse aggregate. Fine RCA was intentionally excluded from the investigation because previous studies have consistently reported that incorporation of recycled fine aggregates significantly deteriorates the mechanical and durability performance of concrete due to their excessive porosity, high water absorption, and weak microstructural characteristics [1,2].

Fine recycled concrete aggregate (FRCA) was intentionally excluded to reduce the number of experimental variables and to focus on the influence of coarse RCA replacement within a controlled framework. Previous studies have shown that FRCA generally exhibits higher water absorption, greater adhered mortar content, and poorer particle quality than coarse RCA, resulting in more pronounced reductions in workability, strength, and durability. Therefore, limiting the investigation to coarse RCA enabled a more systematic evaluation of the proposed machine-learning and optimization framework while minimizing the influence of additional material variability.

Prior to crushing, visible contaminants such as wood, metal, soil, and other foreign materials were manually removed from the concrete debris. The parent concrete specimens were then mechanically crushed and sieved to produce recycled coarse aggregates with a nominal maximum size of 19 mm. Fine recycled particles were excluded from the study. After grading, the aggregates were stored under laboratory conditions until use. Water absorption measurements were performed for each RCA type, and the corresponding aggregate properties were considered during mixture proportioning to maintain consistency among the concrete mixtures.

To simulate realistic field conditions and account for the variability commonly associated with recycled aggregate sources, five categories of parent concrete with different compressive strength levels were selected. The parent concretes were obtained from laboratory-demolished specimens as well as real construction debris and sidewalk curb wastes. These concrete wastes were mechanically crushed and subsequently sieved to produce coarse recycled aggregates with a maximum aggregate size of 19 mm.

The selected parent concretes covered a compressive strength range between 19 MPa and 36 MPa, representing a broad spectrum of practical structural and non-structural concrete applications. The parent concrete strength was considered a governing parameter in this study because the physical quality and residual mortar characteristics of RCA are strongly dependent on the original concrete properties. The compressive strengths of the parent concretes are summarized in Table 1.

Table 1. Compressive strength (MPa) of parent concretes used for RCA production.

RCA	A	B	C	D	E
f_c	19.0	21.5	28.0	33.5	36.0

The physical properties of natural and recycled aggregates are presented in Table 2. Significant differences were observed between natural aggregates and RCA. In general, all RCA types exhibited lower specific gravity and substantially higher water absorption compared with natural coarse aggregate. These differences are primarily attributed to the presence of adhered residual mortar attached to the recycled aggregate particles, which increases internal porosity and capillary connectivity. The reduction in density and increase in water absorption are well-recognized characteristics of RCA and strongly influence both fresh and hardened concrete behavior [5].

Among the investigated RCA types, aggregates produced from lower-strength parent concrete demonstrated higher water absorption and lower specific gravity, indicating a more porous residual mortar structure. Conversely, RCA originating from stronger parent concrete exhibited physical properties closer to those of natural aggregate due to improved matrix densification and lower internal microcracking within the original concrete [12]. Meanwhile, the maximum aggregate size was maintained

at 19 mm for all mixtures to ensure consistency in mixture proportioning and minimize the influence of aggregate size variation on concrete behavior.

The selected ranges were chosen to represent practical RAC production while providing sufficient variability for statistical modeling. Parent concrete strengths of 20–40 MPa reflect typical demolished structural concrete, RCA replacement ratios of 0–100% capture the full range of recycled aggregate utilization and potential nonlinear effects, and cement contents of 300–500 kg/m³ represent common structural concrete mixtures. These ranges also satisfy the requirements of the Central Composite Design (CCD) for developing reliable predictive and optimization models.

Table 2. Physical properties of natural and recycled aggregates.

Aggregate	Natural sand	Natural gravel	RCA-Type A	RCA-Type B	RCA-Type C	RCA-Type D	RCA-Type E
Specific gravity (SSD)	2.51	2.63	2.44	2.42	2.48	2.53	2.62
Water absorption (%)	0.87	0.43	4.45	4.35	3.65	3.30	3.30

Although only five parent concrete strength levels (19, 21.5, 28, 33.5, and 36 MPa) were experimentally investigated, Parent Concrete Strength (X_1) was treated as a continuous design variable within the CCD framework through coded factor representation. These five levels correspond to the standard CCD design points ($-\alpha$, -1, 0, $+1$, and $+\alpha$), enabling the development of continuous response surface models over the investigated range. Consequently, the optimization process was performed on the fitted surrogate models, which provide continuous predictions for intermediate values within the design space rather than assuming the existence of additional experimentally measured parent concrete strengths. Therefore, the continuous treatment of X_1 represents interpolation within the validated experimental domain and should not be interpreted as extrapolation beyond the investigated range. Meanwhile, five different strength levels could be easily controlled by choosing five different sources of concretes with the same compressive strengths values.

2.1.2. Cement

Ordinary Portland Cement (OPC) Type I was used as the sole cementitious material throughout the experimental program. No supplementary cementitious materials were incorporated in order to isolate the influence of recycled aggregate characteristics and mixture proportions on concrete behavior.

2.1.3. Water and chemical admixture

Potable tap water was used for all concrete mixtures and curing operations. To maintain a consistent workability level among all mixtures, slump tests were conducted in accordance with standard concrete testing procedures (ASTM C143). A target slump value of approximately 100 mm was selected to ensure adequate workability while minimizing segregation and bleeding.

A high-range water-reducing admixture (HRWR) was employed to achieve the desired workability without excessively increasing the water content. In all mixtures, the mixing water content was maintained constant at 180 kg/m^3, while the cement content was varied to produce different water-to-cement (w/c) ratios. This approach enabled systematic investigation of the combined effects of cement dosage and RCA characteristics on concrete performance while preserving comparable fresh concrete consistency across all experimental mixtures.

Although the nominal mixing water was maintained at 180 kg/m³ for all mixtures, the recycled concrete aggregates were compensated by additional absorption water before mixing. Therefore, the effective water available for cement hydration remained essentially constant among the mixtures. Consequently,

the observed influence of cement content primarily reflects changes in binder dosage rather than unintended variations in the effective water-to-cement ratio.

2.2. Stepwise research framework

The proposed research methodology consisted of four sequential and interconnected stages integrating experimental design, laboratory testing, machine learning modeling, and deterministic optimization. The overall workflow of the study is illustrated schematically in Fig. 1.

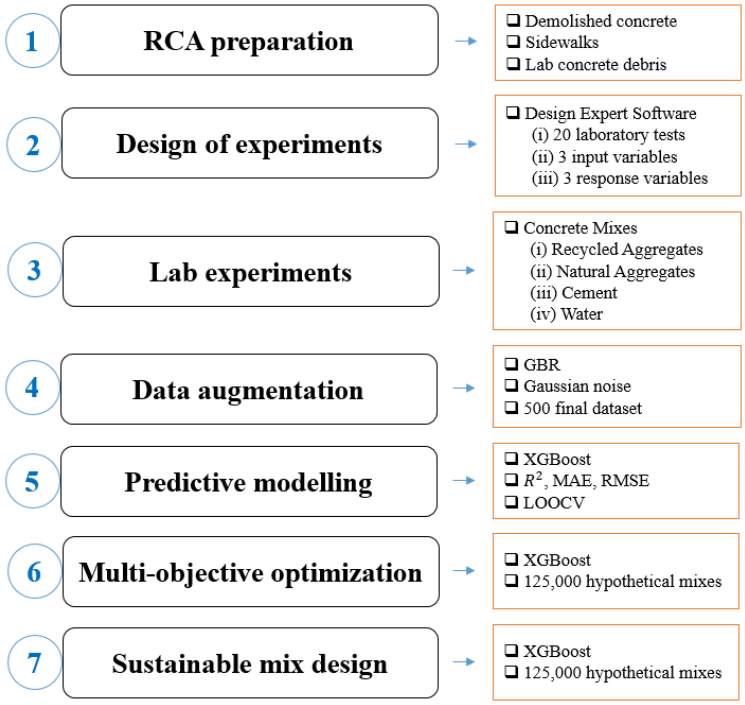

Fig. 1. Schematic illustration of the proposed experimental–computational research framework.

Initially, a structured experimental design strategy was developed using CCD implemented through Design Expert software. The objective of this stage was to systematically construct the experimental space while minimizing the number of required laboratory mixtures and preserving adequate spatial distribution of the governing variables. Three primary input parameters were considered in the design framework: (i) compressive strength of RCA; (ii) RCA replacement ratio; (iii) cement content.

The second stage involved the preparation and testing of twenty laboratory concrete mixtures generated from the CCD matrix. Experimental testing was conducted to determine the corresponding response variables, including compressive strength (ASTM C39), splitting tensile strength (ASTM C496), and water absorption (ASTM C642). These experimentally measured responses formed the core database used for subsequent machine learning analysis.

In the third stage, an artificial intelligence-based predictive framework was developed to establish nonlinear relationships between the mixture parameters and target concrete properties. Initially, the experimental dataset consisting of 20 laboratory observations was utilized to train a baseline surrogate model. Since experimental concrete datasets are inherently limited due to material cost, labor intensity, and testing duration, a physically guided data augmentation strategy was subsequently implemented to improve the robustness and interpolation capability of the predictive framework.

The augmentation process involved generating virtual input combinations within the experimentally defined parameter boundaries using uniformly distributed random sampling. The trained surrogate model was then employed to estimate the corresponding target responses for these virtual mixtures. To preserve realistic experimental variability and avoid generation of excessively smooth artificial data, Gaussian

noise proportional to experimentally derived prediction error metrics was introduced into the generated responses.

Although the original experimental database comprised only 20 mixtures, the sample size was determined by the CCD, which is a statistically efficient experimental design for modeling the effects and interactions of multiple variables while minimizing laboratory effort. Rather than relying solely on the limited experimental data, the proposed framework employed response surface models calibrated using these experimentally validated observations to generate physically consistent synthetic samples for machine-learning training. Therefore, the reliability of the framework depends primarily on the quality and representativeness of the experimental design rather than the number of experimental mixtures alone. Nevertheless, larger experimental datasets would further improve model robustness and generalizability and should be considered in future studies.

It is important to emphasize that the generated virtual samples were not used as substitutes for experimental measurements. Instead, they were utilized exclusively to improve model stability, exploratory analysis capability, and interpolation behavior within the investigated design domain. The machine learning framework was fundamentally trained and validated using experimentally derived concrete behavior.

Subsequently, XGBoost was implemented as the primary predictive engine for modeling the nonlinear interactions among mixture variables. Model performance was evaluated using statistical indicators including coefficient of determination (R^2), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE), together with Leave-One-Out Cross Validation (LOOCV) to improve robustness against overfitting.

The fourth and final stage consisted of deterministic multi-objective optimization of recycled aggregate concrete mixtures. Following model validation, the trained XGBoost framework was employed to virtually screen a high-resolution design space generated using systematic parameter discretization. A total of 125,000 hypothetical mixture combinations were computationally generated through all possible combinations of the governing variables.

It should be noted that these 125,000 mixtures were not experimentally produced. Rather, they represented virtual mixture scenarios computationally evaluated using the validated machine learning framework. The objective of this stage was to identify sustainable and mechanically reliable mixture proportions capable of simultaneously maximizing compressive strength and tensile strength while minimizing water absorption.

2.3. Experimental design and mixture proportioning

Conventional trial-and-error approaches for RAC mixture development are highly labor-intensive, time-consuming, and often incapable of systematically capturing complex interactions among governing mixture parameters. To overcome these limitations, a structured statistical experimental design strategy based on CCD was implemented using Design Expert Software. The CCD framework was selected because of its capability to efficiently map multidimensional parameter spaces while significantly reducing the number of required laboratory experiments [1].

The initial dataset comprised 20 experimental RAC mixtures generated from the authors' own laboratory investigation. The experimental program, including material characterization, mixture preparation, specimen casting, curing, and testing procedures, has been reported in our previous publication [2]. In the present study, these experimentally validated data were used as the basis for developing and validating the proposed machine learning and optimization framework.

In this study, three independent input variables were considered as the primary governing parameters affecting RAC performance: (i) compressive strength of RCA (X_1); (ii) RCA replacement ratio (X_2); (iii)

cement content (X_3). Each parameter was investigated at five coded levels corresponding to axial, factorial, and center-point conditions. The investigated parameter ranges were selected to represent realistic practical boundaries commonly encountered in recycled aggregate concrete applications while simultaneously enabling evaluation of severe nonlinear degradation behavior at high RCA incorporation levels.

The parent concrete compressive strength varied from 19 MPa to 36 MPa in order to simulate recycled aggregates originating from both low-quality demolition waste and relatively high-performance structural concrete. The RCA replacement ratio ranged from 8% to 92%, thereby covering nearly the entire practical replacement spectrum from minimal incorporation to near-full substitution of natural coarse aggregate. Cement content varied between 310 kg/m^3 and 410 kg/m^3 to investigate the combined influence of binder dosage and water-to-cement ratio on mechanical and durability performance. The detailed coded and actual parameter levels are summarized in Table 3.

Table 3. Experimental input variables and coded parameter levels.

Input factor (unit)	Symbol	Coded values				
		-1.68	-1	0	+1	+1.68
Parent compressive strength (MPa)	X_1	19	22.5	28	33	36
Rate of substitution (%)	X_2	8	25	50	75	92
Value of cement (kg/m^3)	X_3	310	330	360	390	410

Three engineering response variables were selected to comprehensively evaluate both mechanical and durability performance of RAC: compressive strength (Y_1), splitting tensile strength (Y_2), and water absorption (Y_3).

The selected responses were chosen because they collectively represent the structural capacity, cracking resistance, and permeability-related durability behavior of recycled aggregate concrete.

Based on the CCD matrix, a total of 20 laboratory concrete mixtures were generated. The experimental design consisted of factorial points, axial (star) points, and replicated center points. Specifically, the design included six factorial combinations, eight axial points, and five repeated center-point mixtures to evaluate experimental variability, estimate pure error, and detect potential curvature in the response surfaces, a strategy that ensures statistical reliability in structural optimization. Compared with conventional one-factor-at-a-time experimental strategies, the adopted CCD framework enabled efficient exploration of multidimensional interactions among RCA quality, substitution level, and binder content while substantially reducing the required number of laboratory mixtures. The coded variables generated by Design Expert software were subsequently transformed into their corresponding physical values prior to machine learning implementation and optimization analysis. The experimentally measured responses after 28 days of curing formed the primary dataset used for machine learning training, validation, sensitivity analysis, and deterministic optimization. Table 4 shows the statistical summary of the original experimental dataset (20 mixtures), including the minimum, maximum, mean, standard deviation, and coefficient of variation of both the input and output variables.

Table 4. Descriptive statistics of the original experimental dataset.

Variable	Unit	Min	Max	Mean	Std. Dev.	CV (%)
Parent concrete strength (X_1)	MPa	19.0	36.0	27.80	4.64	16.7
RCA replacement ratio (X_2)	%	8	92	48.85	23.21	47.5
Cement content (X_3)	kg/m^3	310	410	359	29.08	8.10
Compressive strength	MPa	22.5	37.8	29.24	4.24	14.5
Splitting tensile strength	MPa	2.4	3.5	2.91	0.32	10.91
Water absorption	%	4.3	7.6	6.6	0.99	15.04

2.4. Predictive modeling framework

2.4.1. Surrogate-based data augmentation

The initial database consisted of 20 laboratory concrete mixtures generated using the CCD framework. Although structured experimental design improves parameter-space coverage, machine learning models trained on limited datasets may still suffer from instability and reduced interpolation capability. Therefore, a surrogate-based data augmentation strategy was implemented to improve model robustness while preserving physical consistency.

Initially, a baseline Gradient Boosting Regressor (GBR) model was trained using the experimental dataset, where parent concrete strength, RCA replacement ratio, and cement content were considered input variables, while compressive strength, tensile strength, and water absorption were treated as target responses. Subsequently, 480 virtual input combinations were generated within the experimentally bounded parameter ranges using uniform random sampling. The trained surrogate model was then used to predict the corresponding responses for these hypothetical mixtures. To preserve realistic experimental variability, Gaussian noise scaled according to the experimentally derived RMSE values was injected into the predicted outputs. This procedure enabled preservation of natural experimental scatter and heteroscedastic behavior while maintaining physically meaningful trends within the generated data.

The augmentation process expanded the dataset from 20 experimental mixtures to 500 physically guided samples. The generated virtual data were not intended to replace laboratory measurements or serve as independent validation data; rather, they were used to improve model stability, interpolation capability, and optimization efficiency within the investigated design space.

The adopted augmentation framework is particularly suitable for materials informatics applications, where experimental testing is often expensive and limited in size. Consequently, the proposed approach enables efficient exploration of unexplored mixture combinations while substantially reducing laboratory workload and material consumption.

Although data augmentation increased the sample size for model training, the synthetic data were generated from response surface models based on limited experimental observations and therefore inherit their assumptions and uncertainties. Consequently, the model's predictive performance depends on the quality and representativeness of the original dataset. Future studies should validate the proposed framework using larger and independent experimental datasets to improve its generalizability and robustness.

2.4.2. Final XGBoost predictive framework

XGBoost was selected after considering the characteristics of the available dataset and the objectives of this study. Recycled aggregate concrete exhibits highly nonlinear relationships among mixture variables, while the experimental dataset is relatively small. Compared with Random Forest (RF), XGBoost generally provides higher predictive accuracy by sequentially correcting residual errors through gradient boosting rather than constructing independent trees. Artificial Neural Networks (ANNs), although capable of modeling complex nonlinear relationships, typically require substantially larger datasets and careful hyperparameter tuning to achieve stable performance, making them more susceptible to overfitting when experimental data are limited. Support Vector Regression (SVR) performs well for small datasets but often requires extensive kernel and parameter optimization and may have difficulty representing complex multivariable interactions in heterogeneous material systems. In contrast, XGBoost incorporates regularization, shrinkage, and tree-pruning mechanisms that improve generalization while maintaining computational efficiency. Furthermore, its tree-based architecture naturally supports feature-importance and SHAP analyses, enabling both accurate prediction and interpretable engineering insights, which are central objectives of the present study.

Following data augmentation, the final machine learning database consisted of 500 physically guided samples containing both experimental and surrogate-generated mixtures. This dataset was used to develop the predictive framework using XGBoost, an ensemble learning algorithm based on gradient-boosted decision trees capable of modeling complex nonlinear relationships without predefined mathematical assumptions.

XGBoost was selected due to its strong capability in capturing the heterogeneous and nonlinear behavior of recycled aggregate concrete arising from interactions among recycled aggregate quality, pore structure, moisture transport, and interfacial transition zone deterioration. Compared with conventional regression approaches, XGBoost provides improved nonlinear approximation capability, robustness against overfitting, and efficient handling of multivariable interactions.

To maximize predictive accuracy and prevent overfitting, the hyperparameters of the XGBoost model were systematically tuned rather than manually selected. Recognizing that gradient boosting architectures are highly sensitive to hyperparameter configurations, a Bayesian Optimization strategy coupled with k-fold cross-validation was implemented. Unlike exhaustive Grid Search, which is computationally expensive, or Random Search, which may miss optimal regions, Bayesian Optimization efficiently explores the hyperparameter space by updating a probabilistic model of the objective function. The optimal configuration identified through this systematic tuning yielded 100 estimators, a learning rate of 0.05, a maximum tree depth of 3, and a subsample ratio of 0.8 (with a random state of 42 for reproducibility). This rigorous tuning process ensured an optimal balance between model complexity and generalization capability, as visually summarized in the methodology flowchart (Fig. 5).

Regarding data preprocessing, it is noteworthy that tree-based ensemble algorithms such as XGBoost are fundamentally scale-invariant. Since decision trees determine optimal node splits based on the ordinality of features rather than their absolute magnitudes, input variables were not subjected to normalization or standardization. Consequently, the raw physical values of the mixture parameters were directly utilized for model training. This approach not only eliminated the risk of data leakage during the cross-validation splits but also preserved the direct physical interpretability of the dataset, which was crucial for the subsequent SHAP analysis. The dataset was divided into training and testing subsets for model evaluation. In addition, LOOCV was employed to improve reliability under limited experimental observations. Model performance was evaluated using the R^2, RMSE, and MAE. To improve model interpretability, feature importance analysis was also performed to quantify the relative contribution of parent concrete strength, RCA replacement ratio, and cement content toward prediction of compressive strength, tensile strength, and water absorption.

2.4.3. Sensitivity and feature importance analysis

To improve interpretability of the developed machine learning framework, sensitivity and feature importance analyses were conducted to quantify the relative contribution of each input variable toward the predicted concrete responses. Feature importance values were extracted directly from the optimized XGBoost model based on the cumulative contribution of each parameter toward reducing prediction error across the ensemble trees.

Unlike conventional regression coefficients that assume linear relationships, the XGBoost-based interpretation framework captures nonlinear and interaction-dependent behavior inherent in recycled aggregate concrete systems. This approach enabled identification of the dominant variables governing mechanical and durability performance while providing physically meaningful insight into the effects of recycled aggregate quality, replacement ratio, and cement content.

The extracted feature importance values were subsequently used to compare the relative influence of mixture parameters on compressive strength, tensile strength, and water absorption, thereby improving transparency and engineering interpretability of the predictive framework.

2.4.5. SHAP-based interpretability analysis

To improve the interpretability and transparency of the developed machine learning framework, SHapley Additive exPlanations (SHAP) analysis was implemented for the optimized XGBoost model. SHAP is an explainable artificial intelligence (XAI) technique derived from cooperative game theory that quantifies the contribution of each input variable to the predicted output, an approach that has recently shown great promise in decoding complex structural concrete behaviors [24,25].

Unlike conventional feature importance methods that provide only global ranking information, SHAP analysis enables both global and local interpretation of model behavior by determining whether individual variables positively or negatively influence predictions across different regions of the design space. In this study, SHAP values were computed using the TreeExplainer algorithm integrated with the trained XGBoost architecture.

The SHAP framework was employed to interpret the influence of X_1, X_2, and X_3 on compressive strength, splitting tensile strength, and water absorption predictions. This analysis enabled identification of nonlinear threshold effects, interaction mechanisms, and physically meaningful trends governing recycled aggregate concrete behavior.

2.5. Multi-objective optimization framework

Following validation of the predictive machine learning framework, the trained XGBoost model was employed for deterministic multi-objective optimization of RAC mixtures. The primary objective of this stage was to identify mixture configurations capable of simultaneously maximizing mechanical performance while minimizing durability-related deterioration associated with excessive water absorption.

Unlike conventional optimization approaches that rely on stochastic evolutionary algorithms, the present study implemented a high-resolution deterministic search strategy based on systematic design-space mapping coupled with a Weighted Sum Model (WSM) for multi-criteria decision-making. This approach was selected to ensure stable, reproducible, and globally interpretable optimization behavior within the experimentally defined parameter domain.

Initially, the continuous design space of the three governing input parameters (X_1, X_2, and X_3) was discretized into 50 evenly spaced intervals for each variable. Systematic combination of these intervals generated a dense virtual design space containing: $50 \times 50 \times 50 = 125{,}000$ unique hypothetical concrete mixtures.

It is important to emphasize that these mixtures were not experimentally produced. Rather, they represented computationally generated virtual mixture scenarios evaluated using the validated XGBoost predictive framework. The objective of this large-scale virtual screening process was to efficiently explore unexplored mixture combinations while avoiding the excessive cost and time associated with extensive laboratory experimentation. The trained XGBoost model was subsequently utilized to predict the corresponding engineering responses for all generated mixtures, including: Y_1, Y_2, and Y_3.

Since the optimization objectives possessed conflicting trends and different physical units, a Min–Max normalization procedure was applied to transform all response variables into a dimensionless range between 0 and 1. The normalization process was performed according to:

$$Y_{norm} = \frac{Y - Y_{min}}{Y_{max} - Y_{min}} \quad (1)$$

where ‘Y’ represents the predicted response value, Y_{min} and Y_{max} correspond to the minimum and maximum values within the generated design space.

The optimization framework aimed to:

- maximize compressive strength,
- maximize splitting tensile strength,
- minimize water absorption.

To establish a balanced trade-off between structural performance and durability behavior, a global utility function was formulated using the normalized responses:

$$U = (Y_{1,norm} + Y_{2,norm}) - Y_{3,norm} \quad (2)$$

In this formulation, the normalized mechanical strength parameters were assigned positive contributions, whereas water absorption was incorporated as a penalty term due to its detrimental influence on long-term durability performance.

The mixture combination producing the maximum utility value was extracted as the optimal multi-objective solution. This deterministic optimization strategy enabled identification of sustainable and mechanically reliable RAC mixtures while avoiding the stochastic uncertainty, convergence instability, and local optimum sensitivity commonly associated with evolutionary optimization algorithms such as Genetic Algorithms (GA) and NSGA-II.

Furthermore, the proposed framework provides an interpretable and physically consistent optimization methodology because all predicted mixtures remained constrained within experimentally validated parameter boundaries. Consequently, the optimization process preserved engineering realism while enabling efficient virtual exploration of highly nonlinear recycled aggregate concrete behavior.

3. Results and discussion

3.1. Predictive performance and comparative assessment

One of the principal challenges in machine learning applications for concrete materials is the limited size of experimentally generated datasets, primarily due to the high cost, labor intensity, and time requirements associated with laboratory testing. To address this limitation while preserving physical realism, a surrogate-based data augmentation strategy was implemented in the present study. The original experimental database, consisting of 20 RAC mixtures generated through CCD, was expanded to 500 physically guided samples.

Initially, a baseline GBR was trained using the experimentally measured responses. Subsequently, 480 virtual input combinations were uniformly generated within the experimentally bounded design space. The trained surrogate model predicted the corresponding engineering responses, after which Gaussian noise scaled according to experimentally derived RMSE values was injected into the predictions to preserve realistic laboratory scatter and heteroscedastic behavior. Unlike arbitrary oversampling techniques frequently adopted in machine learning studies, the proposed augmentation framework maintained physical consistency because all generated samples remained constrained within experimentally meaningful parameter ranges and were derived directly from experimentally calibrated material behavior.

Following the data augmentation phase, it is crucial to explicitly distinguish between the computational mapping accuracy of the algorithm and its physical generalization capability. When evaluated on the augmented testing subset (Table 5), the final XGBoost framework achieved coefficients of determination exceeding 0.98 for compressive strength, tensile strength, and water absorption. This exceptionally high accuracy reflects the model's capacity to flawlessly map the dense, high-resolution mathematical design space generated by the surrogate model. However, to ensure this did not merely represent circular validation—where the model fits smoothing artifacts—its true physical reliability was verified via the previously mentioned LOOCV on the original 20 samples (R^2 = 0.89). The primary role of the noise-

injected augmented data was to provide spatial densification and stochastic regularization, preventing the XGBoost architecture from overfitting the sparse initial 20 points, thereby enabling it to achieve this robust 0.89 physical generalization accuracy.

Table 5. Predictive performance of the optimized XGBoost framework.

Target Response	R^2	RMSE	MAE
Compressive Strength (Y_1)	0.991	0.341 MPa	0.259 MPa
Tensile Strength (Y_2)	0.980	0.043 MPa	0.034 MPa
Water Absorption (Y_3)	0.998	0.045 %	0.025 %

Previous studies have consistently shown that increasing RCA replacement levels generally reduces mechanical strength while increasing permeability-related properties because of adhered mortar, internal microcracking, and deterioration of the interfacial transition zone (ITZ) [26]. However, conventional statistical approaches such as RSM often exhibit limited capability in representing the highly nonlinear degradation mechanisms associated with elevated RCA incorporation levels. For example, Said Almaawali reported an optimal RCA replacement ratio of approximately 27.5% using a customized RSM framework, whereas the present XGBoost-based framework successfully identified a substantially higher sustainable replacement threshold (~61%) while maintaining acceptable mechanical and durability performance. This improvement highlights the superior capability of ensemble-learning algorithms in capturing localized nonlinear interactions within heterogeneous RAC systems [27].

Similarly, recent studies employing XGBoost and hybrid ensemble-learning models have demonstrated the effectiveness of tree-based boosting algorithms for predicting RAC properties. Yassir M. Abbas reported that optimized XGBoost models outperformed conventional regressors in predicting compressive strength of RAC mixtures containing supplementary cementitious materials [28], while Salwa R. Al-Taai demonstrated that Bayesian-optimized XGBoost effectively captured nonlinear interactions in eco-friendly recycled aggregate concrete mixtures. The exceptionally high predictive performance achieved in the present study ($R^2 > 0.98$ for all target responses) further confirms the suitability of gradient-boosted ensemble models for heterogeneous RAC systems [29].

The comparison between experimentally measured and predicted responses further confirmed the reliability of the developed framework, as illustrated in Fig. 2. Most prediction points clustered closely around the ideal-fit line, indicating that the XGBoost architecture effectively reproduced the nonlinear interactions among parent concrete quality, RCA replacement ratio, and cement content. Minor deviations observed for a limited number of samples are primarily attributed to intrinsic experimental scatter and the heterogeneous microstructure of recycled aggregate concrete.

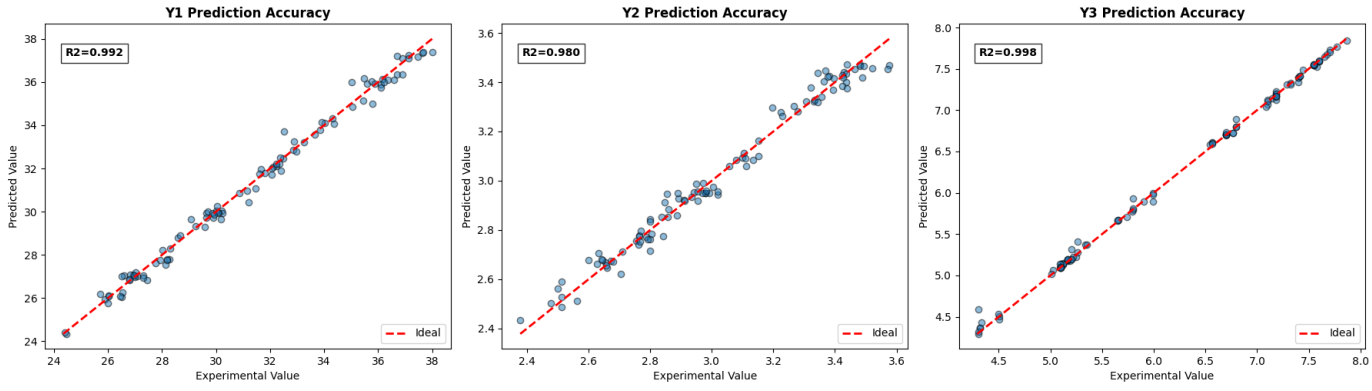

Fig. 2. Comparison between experimental measurements and XGBoost predictions for compressive strength (Y_1), tensile strength (Y_2), and water absorption (Y_3).

A direct comparison with previously reported RSM models developed using the same foundational experimental framework further highlights the advantages of the proposed methodology. Earlier RSM-based models achieved R^2 values ranging between approximately 0.88 and 0.92, whereas the present XGBoost framework exceeded 0.98 for all responses [2]. This substantial improvement demonstrates the superior capability of ensemble machine learning algorithms in decoding multidimensional and nonlinear interactions among RAC constituents. Unlike conventional polynomial-based RSM approaches, which smooth abrupt behavioral transitions and frequently extrapolate unrealistic trends near parameter boundaries, the gradient boosting architecture recursively minimizes localized residual errors and adaptively captures threshold-dependent material behavior without imposing predefined mathematical assumptions. This capability is particularly important for RAC systems, where small variations in parent concrete quality or RCA replacement ratio may produce disproportionate changes in mechanical performance and durability characteristics.

Furthermore, the trained machine learning framework enabled efficient virtual exploration of mixture behavior beyond experimentally tested combinations. By discretizing the parameter space into 50 intervals for each governing variable, a high-resolution virtual design space consisting of 125,000 hypothetical concrete mixtures was computationally generated. The trained XGBoost model subsequently predicted the corresponding engineering responses for all combinations. Importantly, these mixtures were not experimentally produced but represented virtual material scenarios computationally evaluated using the validated predictive framework. This AI-assisted exploration strategy enabled rapid screening of sustainable mixture configurations that would otherwise require impractical laboratory effort if investigated experimentally.

The improved predictive performance should be interpreted as the result of the integrated modeling framework rather than the XGBoost algorithm alone. Unlike the conventional RSM models, which were calibrated directly using the 20 experimental observations, the proposed framework employed surrogate-based data augmentation to generate a larger training dataset prior to XGBoost model development. Consequently, the observed improvement reflects the combined benefits of increased training data and the enhanced nonlinear learning capability of XGBoost. The relative contribution of the learning algorithm and data augmentation was not isolated in the present study and should be investigated in future research through benchmarking under identical training data conditions.

It should be noted that the proposed framework was developed using coarse recycled concrete aggregate only. Consequently, the optimized mixtures and predictive models should not be directly generalized to concrete containing fine recycled aggregates, whose higher water demand and different physical characteristics may alter mixture behavior and the relative importance of the input variables. Nevertheless, the proposed experimental–computational framework is sufficiently flexible to incorporate FRCA as an additional design variable, provided that an appropriate experimental dataset is available. Future studies should extend the framework to mixtures containing both coarse and fine recycled aggregates to improve its applicability to practical construction and demolition waste recycling.

To address the potential risk of information leakage where validating a machine learning model on a synthetically augmented dataset could artificially inflate performance metrics (e.g., $R^2 > 0.98$) a strict independent validation strategy was conducted. The performance of the optimized XGBoost framework was evaluated exclusively against the original 20 physical laboratory observations using the Leave-One-Out Cross-Validation (LOOCV) strategy. To completely prevent information leakage during this validation phase, a fold-specific (nested) augmentation approach was implemented. In each LOOCV iteration, the hold-out validation sample was strictly isolated; the surrogate GBR model was trained, and the 480 synthetic samples were generated and noise-injected using only the remaining 19 training samples. In this rigorous physical validation, the model successfully generalized to unseen experimental mixes, maintaining a robust predictive accuracy with an R^2 of 0.89 for compressive strength. This

confirms that the proposed data augmentation strategy (surrogate modeling with Gaussian noise) successfully acted as an advanced mathematical interpolator within the bounds of the Central Composite Design (CCD), rather than creating model bias. Consequently, the framework's reliability in predicting real-world structural performance is independently verified without relying on synthetic data points during the validation phase.

Yogi et al. (2026) proposed an integrated framework combining LightGBM, SHAP, and NSGA-II for predicting and optimizing recycled aggregate concrete mixtures using a large experimental database of 535 samples. Their study focused primarily on compressive strength prediction and sustainable mix optimization through evolutionary multi-objective optimization. In contrast, the present study develops a unified framework based on statistically designed experiments (CCD), surrogate-based data augmentation, XGBoost prediction, SHAP-based interpretability, and deterministic optimization. Unlike this previous study, the proposed framework simultaneously predicts compressive strength, splitting tensile strength, and water absorption, enabling a more comprehensive assessment of both mechanical and durability performance. Furthermore, by integrating experimental design with surrogate-assisted machine learning, the proposed methodology substantially reduces the experimental effort required for model development while maintaining high predictive accuracy within the investigated design space. Although both studies demonstrate the effectiveness of explainable machine learning for sustainable RAC design, they differ in dataset construction, optimization strategy, and the range of performance indicators considered [30].

Fan et al. (2024) developed a machine learning-assisted NSGA-II framework for optimizing RAC mixture proportions using a large database. In contrast, the present study combines statistically designed experiments (CCD), surrogate-based data augmentation, XGBoost prediction, SHAP-based interpretability, and deterministic optimization within a unified workflow. While both studies aim to optimize RAC mixtures, the proposed framework places greater emphasis on model interpretability and reducing experimental effort through statistically designed experiments, making it suitable for situations where experimental data are limited [31].

To provide a more comprehensive and visual comparative assessment of the proposed model against alternative machine learning approaches, a Taylor diagram was generated (Fig. 3). Taylor diagrams offer a robust graphical summary of model performance by simultaneously quantifying the Pearson correlation coefficient, root mean square error (RMSE), and standard deviation relative to the actual experimental observations. The generated diagram compares the optimized XGBoost framework against an alternative Random Forest (RF) baseline for compressive strength prediction. As visually demonstrated, while both models exhibit excellent correlation, the XGBoost model (blue square) is positioned closer to the reference observation point (black circle) than the RF model. Furthermore, the standard deviation of the XGBoost predictions perfectly aligns with the experimental data arc, indicating that the gradient boosting architecture accurately captured the natural variance of the RAC mixes without over-smoothing. This graphical validation confirms the superior predictive capability and lower residual error of the XGBoost algorithm, consistent with recent advanced soft computing evaluations in structural engineering [32,33].

3.2. Sensitivity and feature importance analysis

Although predictive accuracy is essential, interpretation of the governing physical mechanisms controlling RAC behavior is equally important for practical engineering implementation. Therefore, feature importance analysis was conducted using the optimized XGBoost framework to quantify the relative contribution of each mixture parameter toward the predicted responses. The results revealed fundamentally different governing mechanisms for mechanical and durability-related properties, as illustrated in Fig. 4.

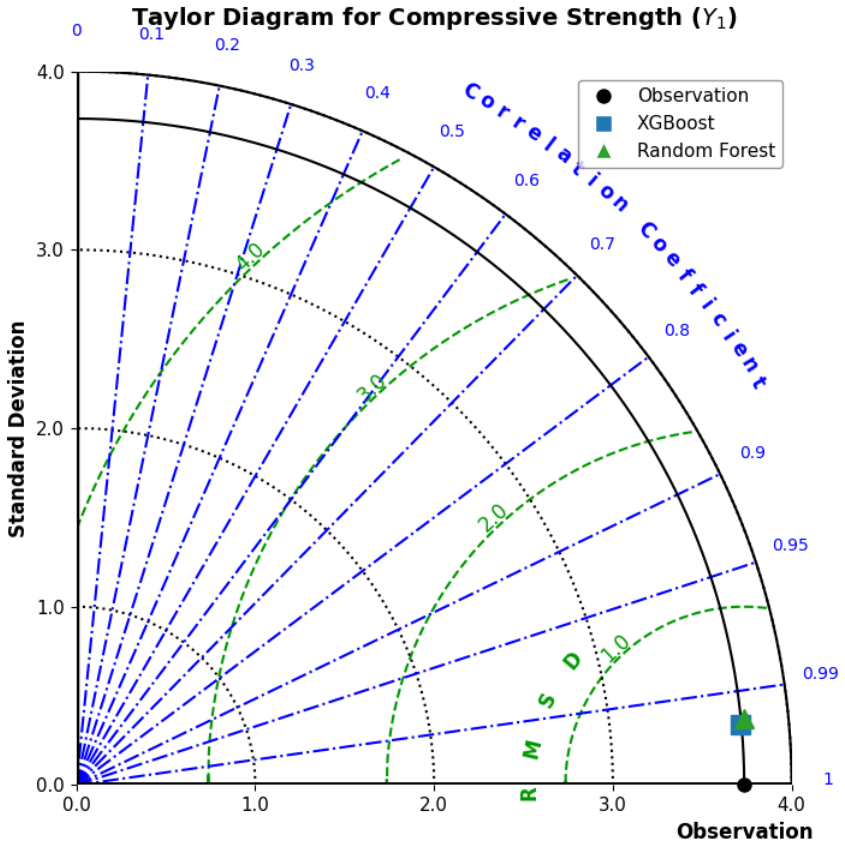

Fig. 3. Taylor diagram illustrating the comparative predictive performance of the optimized XGBoost framework versus the Random Forest (RF) model for compressive strength (Y_1).

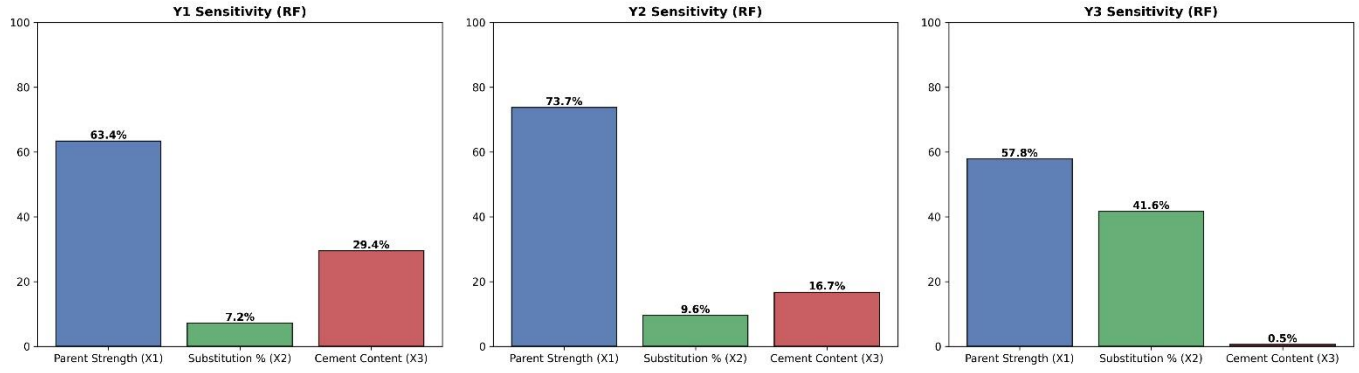

Fig. 4. Relative feature importance of mixture parameters for compressive strength (Y_1), tensile strength (Y_2), and water absorption (Y_3).

A notable distinction between the present machine learning framework and the previously published CCD–RSM study lies in the interpretation and quantification of governing mixture parameters. In the earlier RSM-based investigation, perturbation plots qualitatively indicated that parent concrete strength (X_1) positively influenced compressive and tensile strengths, whereas RCA replacement ratio (X_2) increased water absorption. However, the conventional perturbation analysis remained largely descriptive and was unable to quantitatively determine the relative contribution of each parameter toward the target responses [2].

In contrast, the present XGBoost-based feature importance analysis provided a quantitative and physically interpretable evaluation of parameter influence. For compressive strength (Y_1) and splitting tensile strength (Y_2), the parent concrete compressive strength (X_1) emerged as the dominant governing variable, contributing approximately 63.98% and 74.25% of the prediction variance, respectively. This observation

is physically consistent with the known microstructural behavior of recycled aggregate concrete. Recycled aggregates produced from stronger parent concrete generally possess denser adhered mortar, lower internal porosity, reduced microcracking, and improved aggregate integrity. Consequently, the resulting RAC exhibits enhanced stress-transfer capability and stronger interfacial bonding between recycled aggregates and the newly formed cement matrix.

Cement content (X_3) served as the secondary influential parameter for mechanical performance, contributing approximately 28.85% and 16.49% toward compressive and tensile strength prediction, respectively. Increasing cement dosage promotes hydration-product formation, matrix densification, and reduction of the effective water-to-cement ratio. Nevertheless, the results indicate that increased binder content alone cannot fully compensate for weak recycled aggregates originating from low-strength parent concrete.

Conversely, the governing mechanism controlling water absorption (Y_3) differed substantially from the strength-related responses. In this case, the RCA replacement ratio (X_2) exhibited a dominant contribution of approximately 41.68%, nearly comparable to the influence of parent concrete quality. This behavior reflects the inherently porous nature of adhered mortar attached to recycled aggregate particles. As RCA incorporation increases, the volume of interconnected capillary pores within the concrete matrix also increases, thereby accelerating moisture ingress and permeability-related deterioration.

Interestingly, cement content demonstrated negligible influence on water absorption behavior. This observation suggests that moisture transport in RAC is governed predominantly by recycled aggregate microstructure and pore connectivity rather than overall binder quantity. These findings are fully consistent with established durability mechanisms reported in the RAC literature, where adhered mortar porosity and interfacial transition zone deterioration largely control permeability-related performance [12].

This distinction is particularly important because the proposed machine learning framework transforms sensitivity analysis from a qualitative statistical interpretation into a quantitative engineering decision-support tool capable of identifying dominant nonlinear interactions within heterogeneous RAC systems. Furthermore, unlike conventional perturbation plots that primarily capture local polynomial trends, the tree-based ensemble structure of XGBoost successfully identified threshold-dependent degradation mechanisms associated with pore connectivity evolution and adhered mortar effects at high RCA incorporation levels.

3.3. SHAP analysis interpretation

The SHAP summary analysis (Fig. 5) provided additional insight into the nonlinear behavior captured by the XGBoost framework. The distribution of SHAP values confirmed that X_1 exerted the strongest positive influence on compressive and tensile strength predictions, particularly at higher strength levels. In contrast, elevated X_2 generated increasingly negative SHAP contributions, indicating nonlinear degradation behavior associated with excessive recycled aggregate incorporation. X_3 generally contributed positively toward strength enhancement; however, the SHAP distributions revealed diminishing marginal improvements at higher dosage levels. For water absorption behavior, high RCA replacement ratios produced strongly positive SHAP values, confirming the dominant role of recycled aggregate porosity and adhered mortar in governing permeability-related performance. Overall, the SHAP analysis demonstrated that the developed XGBoost model successfully captured physically meaningful nonlinear interactions among the investigated mixture variables.

3.4. Multi-objective optimization of recycled aggregate concrete mixtures

The final stage of the proposed framework involved deterministic multi-objective optimization of recycled aggregate concrete (RAC) mixtures using the validated XGBoost predictive model.

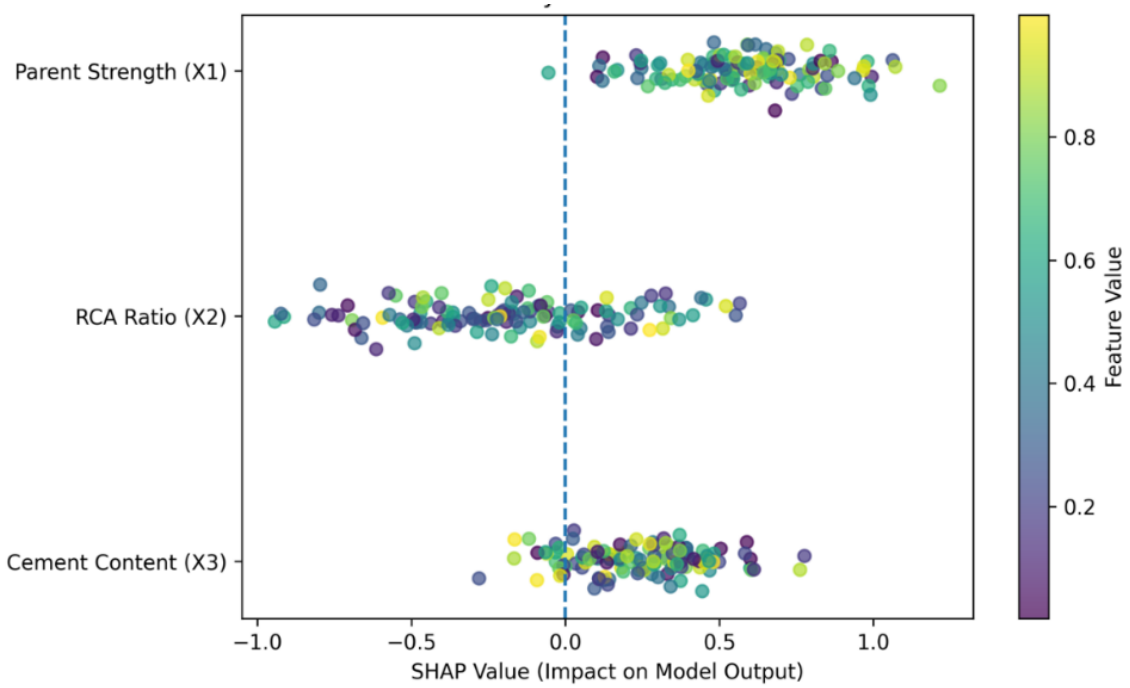

Fig. 5. SHAP summary plot illustrating the X_1, X_2, and X_3 on the XGBoost prediction model. Positive SHAP values indicate positive contributions toward the predicted response, while negative values indicate adverse effects. The color gradient represents the magnitude of each input variable from low (blue) to high (yellow).

Unlike conventional trial-and-error procedures, the developed methodology enabled systematic virtual exploration of 125,000 hypothetical mixture combinations while simultaneously considering both mechanical performance and durability requirements. The optimization objectives were defined to maximize compressive strength and tensile strength while minimizing water absorption.

In this study, equal weighting was adopted to provide a neutral and balanced evaluation of compressive strength, splitting tensile strength, and water absorption without favoring a specific performance objective. This assumption is appropriate for demonstrating the proposed optimization framework under general structural design conditions. However, the relative importance of these performance indicators depends on the intended application. For example, durability may be prioritized for marine or aggressive environments, whereas mechanical strength may receive greater emphasis in heavily loaded structural members. The proposed framework is sufficiently flexible to accommodate alternative weighting schemes according to project-specific design requirements.

To establish a balanced engineering trade-off among these conflicting responses, all predicted outputs were normalized using Min–Max normalization and integrated into a weighted utility function. The mixture corresponding to the maximum global utility score was identified as the optimal RAC design. The optimal solution was obtained at a parent concrete strength (X_1) of 36 MPa, RCA replacement ratio (X_2) of 61.14%, and cement content (X_3) of 320.20 kg/m³. This mixture achieved predicted responses of 37.76 MPa compressive strength, 3.51 MPa tensile strength, and 5.13% water absorption.

To demonstrate the robustness of this optimal solution and evaluate the influence of weighting factors in the Weighted Sum Model (WSM), a sensitivity analysis of the utility function was conducted under three distinct engineering scenarios: (I) Balanced-Design (equal weights, current optimal), (II) Strength-Oriented (higher weights for Y_1 and Y_2), and (III) Durability-Oriented (higher penalty weight for Y_3). The analysis revealed that shifting the priority toward maximum structural strength (Scenario II) forces the optimization algorithm to reduce the RCA replacement ratio to approximately 35-40% while maximizing cement content to its upper bounds. Conversely, penalizing water absorption heavily (Scenario III) restricts the RCA ratio to below 20%, as the porous adhered mortar of RCA directly governs permeability.

Therefore, the identified optimal mix (61.14% RCA with equal weighting) represents the most robust and practical trade-off solution for sustainable infrastructure; it maximizes the environmental benefits of waste utilization while ensuring that the mechanical and durability constraints remain within acceptable limits for general-purpose applications.

The optimized RCA replacement level suggests the existence of a critical nonlinear transition threshold governing RAC behavior. Below this threshold, substantial recycled aggregate utilization can be achieved while maintaining satisfactory mechanical integrity and durability performance. However, excessive RCA incorporation likely accelerates deterioration mechanisms associated with pore interconnectivity, moisture transport, and interfacial transition zone (ITZ) degradation. Consequently, the optimization framework successfully identified a sustainable balance between structural performance and environmental benefit.

A comparison with conventional Response Surface Methodology (RSM)-based optimization further highlights the advantages of the proposed machine learning framework. Previous RSM studies frequently extrapolated unrealistic trends toward excessively high RCA replacement levels because second-order polynomial equations inherently smooth abrupt degradation behavior and fail to capture threshold-dependent nonlinear transitions [30]. In contrast, the tree-based ensemble structure of XGBoost effectively identified physically unsafe mixture regions and penalized excessive RCA incorporation associated with severe durability deterioration. Similar limitations of polynomial optimization for heterogeneous RAC systems have also been reported in previous investigations of recycled concrete behavior and nonlinear material degradation.

Compared with stochastic optimization techniques such as Genetic Algorithms (GA) and NSGA-II, applied to recycled concrete [31]. The present deterministic grid-search framework additionally provides improved interpretability and optimization reliability. Evolutionary algorithms may converge toward local optima or produce unstable solutions depending on random initialization conditions, whereas the adopted grid-search strategy systematically mapped the entire experimentally bounded design space and guaranteed identification of the global optimum within the investigated parameter domain. This deterministic optimization strategy therefore provides a more physically reliable and engineering-oriented approach for sustainable RAC mixture design, consistent with recent machine learning optimization studies in sustainable concrete engineering.

The optimized mixture identified in this study represents the best solution within the investigated design space based on the selected optimization objectives. Although the optimum cement content (approximately 320 kg/m^3) combined with 61.14% RCA yielded satisfactory predicted mechanical and durability performance, compliance with minimum cement content requirements specified in structural design standards depends on the intended exposure class and applicable national or international codes. Since factors such as chloride exposure, freeze–thaw resistance, and sulfate attack were not considered in the present optimization, the proposed mixture should be regarded as a computationally optimized solution rather than a universally applicable structural mix. For practical implementation, the optimized proportions should be verified against relevant design specifications and validated experimentally for the target service environment.

Although the optimized RCA replacement ratio (61.14%) lies within the investigated design space, it was identified through surrogate-assisted optimization and was not experimentally validated. Therefore, the predicted performance should be regarded as a computational estimate rather than a confirmed experimental optimum. Since the XGBoost model was trained using augmented data generated from the response surface models, its prediction reliability depends on the accuracy of the underlying surrogate models within the experimental domain. Although interpolation within the investigated range is generally more reliable than extrapolation, localized nonlinear effects may not be fully captured without additional

physical testing. Consequently, experimental validation of the optimized mixture is recommended as future work to verify the predicted mechanical and durability performance.

3.5. Sustainability and environmental engineering implications

Fig. 6 shows an overall framework for sustainable recycled aggregate concrete design. The proposed AI-assisted framework demonstrated strong potential for sustainable RAC design and practical infrastructure applications by enabling high levels of RCA utilization while maintaining acceptable mechanical and durability performance. The optimized mixture achieved compressive strengths exceeding 35 MPa while incorporating approximately 61% RCA, indicating its suitability for a wide range of structural and non-structural applications, including pavements, sidewalks, precast concrete products, retaining structures, and low- to medium-rise buildings.

The environmental benefits of the optimized mixture can also be quantified. The optimal design incorporates approximately 61% recycled coarse aggregate, directly replacing an equivalent quantity of virgin natural aggregate. Consequently, the proposed mixture has the potential to reduce natural aggregate extraction by approximately 61% while simultaneously diverting a comparable proportion of construction and demolition waste from landfill disposal. In addition, the optimized cement content (320 kg/m^3) remains within the lower range commonly adopted for structural concrete, avoiding unnecessary binder consumption and its associated greenhouse gas emissions. Although a comprehensive life-cycle assessment (LCA) was beyond the scope of the present study, these results clearly demonstrate that the proposed framework contributes to resource conservation, waste valorization, and circular economy objectives while maintaining satisfactory mechanical and durability performance.

From a sustainability perspective, the proposed framework contributes to reduction of construction and demolition waste disposal, conservation of natural aggregate resources, and lowering of the environmental footprint associated with conventional concrete production. The optimization process additionally identified mechanically reliable mixtures with moderate cement consumption, which may further reduce embodied carbon emissions related to cement manufacturing.

The results further revealed that excessive RCA incorporation may trigger nonlinear deterioration in durability-related properties, particularly water absorption. Unlike conventional regression-based optimization approaches, the developed XGBoost framework successfully identified these nonlinear degradation thresholds and avoided physically unreliable mixture regions that are difficult to capture using simplified statistical models. Consequently, the proposed methodology can serve as an intelligent decision-support framework for identifying sustainable and structurally reliable replacement levels in practical engineering applications.

Another important contribution of this study lies in the substantial reduction of experimental workload through AI-assisted virtual materials optimization. By computationally evaluating 125,000 hypothetical mixtures using a limited number of laboratory experiments, the framework provides an efficient platform for rapid screening and optimization of sustainable concrete mixtures prior to physical production and testing. The proposed methodology therefore supports future development of intelligent mix design systems, low-carbon concrete technologies, and AI-assisted sustainable infrastructure engineering.

From a practical implementation perspective, the optimized RAC mixture can be readily produced using conventional concrete batching and casting equipment without requiring specialized manufacturing technologies. The identified optimum, consisting of approximately 61% recycled coarse aggregate and 320 kg/m^3 cement, lies within practical mixture proportion ranges commonly achievable in ready-mix concrete plants. However, successful field implementation requires appropriate quality control of recycled aggregates, particularly with respect to parent concrete quality, particle grading, water absorption, and contaminant removal.

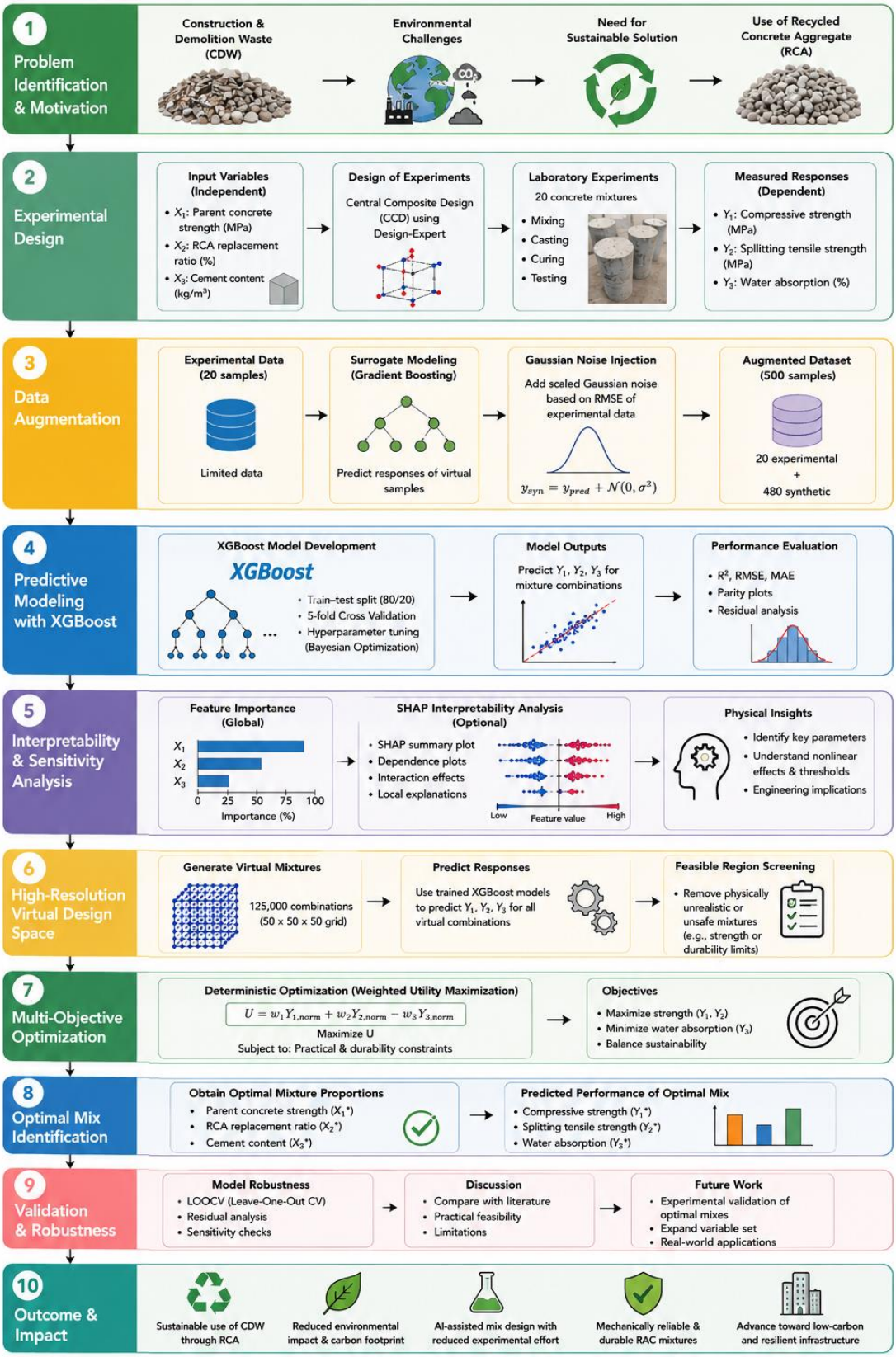

Fig. 6. A unified experimental-computational framework for sustainable recycled aggregate concrete design.

Moisture correction and proper aggregate conditioning are also essential to maintain the designed water-to-cement ratio because recycled aggregates generally exhibit higher water absorption than natural aggregates. Considering these requirements, the proposed mixture is particularly suitable for structural and non-structural applications such as pavements, precast concrete elements, retaining walls, sidewalks,

parking areas, and low- to medium-rise buildings, where sustainable construction materials are increasingly encouraged. The machine learning framework developed in this study may also assist engineers in rapidly adapting mixture proportions to locally available recycled aggregate sources while reducing the need for extensive laboratory trial mixes.

3.6. Limitations and future research

Despite the excellent predictive performance achieved by the proposed framework, several limitations should be acknowledged. First, the study was based on a relatively limited number of experimentally produced mixtures and specific local sources of recycled concrete aggregate. Consequently, the developed model should primarily be interpreted within the investigated parameter ranges and material conditions.

Second, while the surrogate-based data augmentation methodology resolves the critical issue of data scarcity in ML-based mix design, certain inherent limitations of synthetic data in materials informatics must be explicitly acknowledged. Synthetic data strictly assumes that the underlying physical phenomena remain continuous and follow the patterns established within the initial Design of Experiments (DoE) boundaries. Consequently, augmented datasets may fail to capture sudden chemical anomalies, unpredicted interactions at extreme boundary limits, or the long-term deterioration mechanisms of RAC that only physical aging tests can reveal. Therefore, in materials informatics, synthetic data should be utilized exclusively as a computational interpolator to densify the known design space for optimization algorithms, rather than as a complete substitute for independent laboratory validation. Future studies should incorporate larger multi-source experimental databases and external validation datasets to further mitigate these limitations and improve model generalization. Future studies should therefore incorporate larger multi-source experimental databases and external validation datasets to further improve model generalization capability.

Additionally, the present investigation focused primarily on mechanical strength and water absorption behavior. Future research should incorporate additional durability indicators such as chloride penetration, freeze–thaw resistance, shrinkage, and carbonation behavior to establish a more comprehensive sustainability-oriented optimization framework.

XGBoost was selected because of its proven ability to model complex nonlinear relationships, incorporate regularization to reduce overfitting, and provide interpretable feature importance and SHAP analyses. Although other machine learning algorithms, such as Random Forest, LightGBM, CatBoost, SVR, and ANN, have also been successfully applied to concrete property prediction, a comprehensive benchmark comparison was beyond the scope of the present study. Future work will include systematic comparisons under identical training and validation conditions to further evaluate the relative advantages of the proposed framework.

The optimized RCA replacement ratio of 61.14% should be interpreted as the optimum within the investigated design space and optimization objectives rather than as a universally applicable replacement level. This optimum reflects a balance between mechanical performance and durability under the specific material properties and mixture proportions considered in this study. The relatively high optimum can be attributed to the good quality of the recycled aggregates used, the selected parent concrete strength, and the beneficial influence of cement content, which collectively compensated for the adverse effects typically associated with increasing RCA replacement. Consequently, the optimization identified a replacement level at which the reduction in mechanical properties remained limited while maximizing recycled aggregate utilization. Nevertheless, the optimum should be regarded as a model-based prediction and requires experimental validation before practical implementation.

Although the proposed framework demonstrated excellent predictive performance through LOOCV, its generalizability has not yet been verified using independent external datasets. Publicly available RAC databases often exhibit considerable variability in material properties, mix designs, testing methods, and

specimen preparation, which may influence model performance. Future work should focus on validating and recalibrating the proposed framework using harmonized multi-source datasets covering diverse recycled aggregate characteristics and geographical regions. Such external validation would provide a more comprehensive assessment of the model's robustness and applicability under practical conditions.

Despite these limitations, the proposed framework has considerable potential for practical engineering applications. Rather than replacing conventional laboratory testing, it can serve as a decision-support tool for preliminary mix proportioning by rapidly identifying promising RAC mixtures that satisfy multiple performance objectives while reducing the number of trial batches. This approach can shorten the mix design process, reduce material consumption and testing costs, and support the development of more sustainable recycled aggregate concrete. Final mixture proportions should nevertheless be verified against applicable design codes and validated experimentally before large-scale construction applications.

4. Conclusions

This study developed an integrated experimental–computational framework for prediction, interpretation, and optimization of recycled aggregate concrete (RAC) using structured experimental design, surrogate-based data augmentation, ensemble machine learning, and deterministic multi-objective optimization. By combining Central Composite Design (CCD), physically guided virtual data generation, Extreme Gradient Boosting (XGBoost), and high-resolution virtual mixture exploration, the proposed methodology successfully addressed several limitations associated with conventional statistical regression approaches for sustainable concrete design. Based on the obtained experimental and computational results, the following conclusions can be drawn:

1. The developed XGBoost framework demonstrated excellent predictive capability for all investigated concrete responses, achieving coefficient of determination (R^2) values of 0.991, 0.980, and 0.998 for compressive strength, splitting tensile strength, and water absorption, respectively. The low RMSE and MAE values further confirmed the robustness and stability of the predictive model.
2. The surrogate-based augmentation strategy successfully expanded the experimental database from 20 laboratory mixtures to 500 physically guided samples while preserving realistic experimental variability and material consistency. The adopted methodology improved model robustness and interpolation capability without replacing experimentally measured observations.
3. The superior performance of XGBoost compared with conventional RSM demonstrates the importance of nonlinear ensemble learning approaches for recycled aggregate concrete systems. Unlike polynomial-based regression models, the gradient boosting framework effectively captured localized nonlinear degradation mechanisms associated with high RCA incorporation levels, heterogeneous pore structures, and interfacial transition zone deterioration.
4. Feature importance analysis revealed that the parent concrete strength of recycled aggregate was the dominant governing parameter controlling compressive and tensile strength development, contributing approximately 63.98% and 74.25% of the prediction variance, respectively. Conversely, water absorption behavior was governed primarily by RCA replacement ratio and parent concrete quality, whereas cement content exerted comparatively limited influence on permeability-related behavior.
5. The deterministic multi-objective optimization framework successfully identified a sustainable and mechanically reliable RAC mixture consisting of approximately 61.14% RCA replacement and cement content of 320.20 kg/m^3. The optimized mixture achieved predicted compressive strength, tensile strength, and water absorption values of 37.76 MPa, 3.51 MPa, and 5.13%, respectively.
6. The optimization results additionally revealed the existence of a critical nonlinear transition threshold in RAC behavior. While moderate RCA incorporation levels can provide sustainable and

structurally acceptable concrete performance, excessively high replacement ratios significantly accelerate degradation mechanisms associated with pore connectivity, moisture transport, and interfacial deterioration.

7. Compared with traditional RSM-based optimization approaches, the proposed machine learning framework provided more physically realistic and engineering-oriented solutions by successfully penalizing mixture regions associated with severe nonlinear durability degradation at extreme RCA replacement levels.
8. From a sustainability perspective, the proposed framework demonstrates the significant potential of integrating recycled aggregate utilization with artificial intelligence-assisted materials optimization for environmentally responsible concrete design. By enabling efficient virtual exploration of sustainable mixture configurations, the methodology reduces experimental burden, promotes circular utilization of construction waste, minimizes virgin aggregate consumption, and supports development of low-impact infrastructure materials. The findings therefore contribute not only to recycled concrete technology, but also to the broader transition toward intelligent, resource-efficient, and sustainable construction engineering systems and broader environmental sustainability pathways.

Authors contribution statement

Nader Biglarijoo: Conceptualization; Data curation; Methodology; Project administration; Supervision; Writing – original draft; Writing – review & editing.

Shahaboddin Kharazmi: Investigation; Methodology; Project administration; Writing – review & editing.

Mohammad Reza Danaei: Data curation; Formal analysis; Software; Validation; Writing – review & editing.

Amin Shams: Investigation; Methodology; Resources; Supervision; Writing – review & editing.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflicts of interest

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

Data availability statement

The original experimental dataset, the synthetically augmented dataset, and the core Python implementation (including data augmentation, XGBoost predictive modeling, SHAP interpretability, and deterministic optimization scripts) that support the findings of this study are available from the corresponding author upon reasonable request.

References

[1] Biglarijoo N, Nili M, Hosseinian S, Razmara M, Ahmadi S, Razmara P. Modelling and optimisation of concrete containing recycled concrete aggregate and waste glass. Mag Concr Res 2017;69:306–16. https://doi.org/10.1680/jmacr.16.00279.

[2] Biglarijoo N, Shams A, Salehi H. Modeling and Optimizing Recycled Concrete Properties Using Central Composite Design: The Impact of Parent Concrete. J Rehabil Civ Eng 2026;14. https://doi.org/10.22075/jrce.2025.35591.2186.

[3] Oviedo AI, Londoño JM, Vargas JF, Zuluaga C, Gómez A. Modeling and Optimization of Concrete Mixtures Using Machine Learning Estimators and Genetic Algorithms. Modelling 2024;5:642–58. https://doi.org/10.3390/modelling5030034.

[4] Deb S, Mahajan Y, Sharma V, Kapse M, Ghosal A. Machine learning-based optimization of sustainable concrete mix for strength prediction. Discov Sustain 2026;7:989. https://doi.org/10.1007/s43621-026-03227-4.

[5] Nili M, Hoseinian M, Ahmadi S, Biglarijoo N. Assessing the Influences of Waste Glass and Recycled Concrete Aggregate on Properties of Concrete. fib Symp. PRAGUE, 2011, p. 1–7.

[6] Wu Y, Pieralisi R, B. Sandoval FG, López-Carreño RD, Pujadas P. Optimizing pervious concrete with machine learning: Predicting permeability and compressive strength using artificial neural networks. Constr Build Mater 2024;443:137619. https://doi.org/10.1016/j.conbuildmat.2024.137619.

[7] Biglarijoo N, Nili M, Mirbagheri SA. A Review on the Use of Various Kinds of Debris and Demolitions in Concrete and Mortar Mixes. 10th Int. Congr. Adv. Civ. Eng., 2012, p. 1–10.

[8] Li T, Hu X, Li T, Liao J, Mei L, Tian H, et al. Enhanced Prediction and Evaluation of Hydraulic Concrete Compressive Strength Using Multiple Soft Computing and Metaheuristic Optimization Algorithms. Buildings 2024;14:3461. https://doi.org/10.3390/buildings14113461.

[9] Shahrokhishahraki M, Malekpour M, Mirvalad S, Faraone G. Machine learning predictions for optimal cement content in sustainable concrete constructions. J Build Eng 2024;82:108160. https://doi.org/10.1016/j.jobe.2023.108160.

[10] Al-Bayati HK, Jadaa W, Tighe SL. Effect of Recycled Concrete Aggregate Addition on the Asphalt Mixtures Performance: ITZ Area, Microstructure, and Chemical Analysis Perspectives. Recycling 2024;9:41. https://doi.org/10.3390/recycling9030041.

[11] Memon SA, Bekzhanova Z, Murzakarimova A. A Review of Improvement of Interfacial Transition Zone and Adherent Mortar in Recycled Concrete Aggregate. Buildings 2022;12:1600. https://doi.org/10.3390/buildings12101600.

[12] Salehi H, Biglarijoo N, Barkhordari H. Response Surface Methodology Modelling to Study the Influence of Recycled Aggregates on Some Mechanical Properties of Recycled Concrete. AUT J Civ Eng 2021;5:301–10. https://doi.org/10.22060/ajce.2021.18634.5689.

[13] Alnadish AM, Hamza M, Ali S, Murad N, Durrani SMK. Machine learning–driven optimization of compressive and tensile strength in concrete with GGBS, eggshell powder, and waste glass powder. Sci Rep 2025;15:40499. https://doi.org/10.1038/s41598-025-24438-1.

[14] Sarir P, Ruangrassamee A, Iwanami M. Estimation of the axial capacity of high-strength concrete-filled steel tube columns using artificial neural network, random forest, and extreme gradient boosting approaches. Front Struct Civ Eng 2024;18:1794–814. https://doi.org/10.1007/s11709-024-1126-7.

[15] Nili M, Biglarijoo N, Hosseinian S, Ahmadi S. Disposing waste demolition in concrete as aggregate replacement. Int J Mater 2014;1.

[16] Zhou C, Chen Z. Mechanical properties of recycled concrete made with different types of coarse aggregate. Constr Build Mater 2017;134:497–506. https://doi.org/10.1016/j.conbuildmat.2016.12.163.

[17] Lu C, Yu Q, Wei J, Niu Y, Zhang Y, Lin C, et al. Influence of interface transition zones (ITZ) and pore structure on the compressive strength of recycled aggregate concrete. Constr Build Mater 2024;456:139299. https://doi.org/10.1016/j.conbuildmat.2024.139299.

[18] Jaglan A, Singh RR. Recycled Aggregate Concrete Incorporating GGBS and Polypropylene Fibers Using RSM and Machine Learning Techniques. Buildings 2025;15:66. https://doi.org/10.3390/buildings15010066.

[19] Habibi A, Ramezanianpour AM, Mahdikhani M. RSM-based optimized mix design of recycled aggregate concrete containing supplementary cementitious materials based on waste generation and global warming potential. Resour Conserv Recycl 2021;167:105420. https://doi.org/10.1016/j.resconrec.2021.105420.

[20] Shinde SN, Christa S, Grover RK, Pasha N, Harinder D, Nakkeeran G, et al. Optimization of waste plastic fiber concrete with recycled coarse aggregate using RSM and ANN. Sci Rep 2025;15:7798. https://doi.org/10.1038/s41598-025-92505-8.

[21] Biglarijoo N, Shams A, Bagheri M, Gill L, Keypour S. From data to sustainable clean water: AI/ML contributions to the UN Sustainable Development goals. Phys Chem Earth, Parts A/B/C 2026;144:104427. https://doi.org/10.1016/j.pce.2026.104427.

[22] Lal Mohiddin S, Ravi Prasad D, Ramaseshu D. A review of machine learning models for concrete strength prediction and mix optimization. J Build Pathol Rehabil 2025;10:128. https://doi.org/10.1007/s41024-025-00636-2.

[23] Hematibahar M, Martins CH, Momeni K. Machine Learning Prediction of Fiber-Reinforced Concrete Using the SADRA Algorithm Based on Transcendental Philosophy. J Soft Comput Civ Eng 2026;10:1944. https://doi.org/10.22115/scce.2025.1994.

[24] Nouri Y, Ghanbari MA, Fakharian P. Flexural behavior of hybrid GFRP-steel reinforced concrete beam: Experimental and explainable artificial intelligence. Eng Struct 2025;345:121565. https://doi.org/10.1016/j.engstruct.2025.121565.

[25] Fakharian P. Compressive Strength Prediction of 3D Printed Concrete with Machine Learning Approaches. Comput Eng Phys Model 2025;8:1385. https://doi.org/10.22115/cepm.2026.1385.

[26] Sun M, Xi X, Zheng Y, Wang L, Wang K. Development of an XGBoost-based regression model and experimental validation for predicting the compressive strength of fully recycled aggregate concrete. Structures 2026;88:111816. https://doi.org/10.1016/j.istruc.2026.111816.

[27] Almaawali S, Aldahdooh MAA, Alnaamani SAM, Ibrahim NT, Ng CA. Optimization of sustainable concrete incorporating coarse recycled aggregate using a customized single-factor response surface methodology (CSFRSM) approach. Discov Civ Eng 2025;2:36. https://doi.org/10.1007/s44290-025-00195-y.

[28] Abbas YM, Babiker A, Ismail Ismail F. Optimized XGBoost-Based Framework for Robust Prediction of the Compressive Strength of Recycled Aggregate Concrete Incorporating Silica Fume, Slag, and Fly Ash. Comput Model Eng Sci 2025;145:3279–307. https://doi.org/10.32604/cmes.2025.074069.

[29] Al-Taai SR, Azize NM, Thoeny ZA, Imran H, Bernardo LFA, Al-Khafaji Z. XGBoost Prediction Model Optimized with Bayesian for the Compressive Strength of Eco-Friendly Concrete Containing Ground Granulated Blast Furnace Slag and Recycled Coarse Aggregate. Appl Sci 2023;13:8889. https://doi.org/10.3390/app13158889.

[30] Yogi B, Majumdar R, Ghosh P, Roy S, Das SK. Data-driven multi-objective optimization of recycled aggregate concrete mixes via LightGBM strength modelling and SHAP interpretability. Discov Artif Intell 2026;6:536. https://doi.org/10.1007/s44163-026-01230-2.

[31] Fan M, Li Y, Shen J, Jin K, Shi J. Multi-objective optimization design of recycled aggregate concrete mixture proportions based on machine learning and NSGA-II algorithm. Adv Eng Softw 2024;192:103631. https://doi.org/10.1016/j.advengsoft.2024.103631.

[32] Fakharian P, Nouri Y, Ghanizadeh AR, Safi Jahanshahi F, Naderpour H, Kheyroddin A. Bond strength prediction of externally bonded reinforcement on groove method (EBROG) using MARS-POA. Compos Struct 2024;349–350:118532. https://doi.org/10.1016/j.compstruct.2024.118532.

[33] Nouri Y, Ghanizadeh AR, Safi Jahanshahi F, Fakharian P. Data-driven prediction of axial compression capacity of GFRP-reinforced concrete column using soft computing methods. J Build Eng 2025;101:111831. https://doi.org/10.1016/J.JOBE.2025.111831.