SvedbergOpen
DISSEMINATION OF KNOWLEDGE
International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/
Research Paper Open Access

Machine Learning-Based Prediction of Compressive Strength of Recycled Aggregate Concrete

Subhash Kumar[1*], Dr. Pappu Kumar[2], Dr. Ajay Kumar[3], Dr. Md Daniyal[4]

[1*]Research Scholar, Department of Civil Engineering, Sandip University, Madhubani, Bihar India - 847235. Email: subhash.sk2025@gmail.com , ORCID: 0009-0003-7471-8138
[2]Assistant Professor, Department of Civil Engineering, Sandip University, Madhubani, Bihar India - 847235. Email: pappusobh@gmail.com , ORCID: 0000-0001-8641-2302
[3]Assistant Professor, Department of Civil Engineering, Sandip University, Madhubani, Bihar India - 847235. Email: ajaykumar8mail@gmail.com , ORCID: 0009-0002-8418-0135
[4]Assistant Professor, Department of Civil Engineering, Sandip University, Madhubani, Bihar India - 847235. Email: daniyalzhcet@gmail.com , ORCID: 0000-0002-4070-3574

ABSTRACT
Recycled aggregate concrete (RAC) provides a promising route for reducing the consumption of natural aggregates and diverting construction and demolition waste from disposal. However, the prediction of RAC compressive strength remains challenging because it is governed by the combined effects of mixture proportions, recycled aggregate characteristics, parent concrete quality and aggregate morphology. The present study develops a comprehensive author-curated database of 637 RAC observations and investigates the applicability of two ensemble machine-learning algorithms, Random Forest (RF) and Extreme Gradient Boosting (XGBoost), for predicting compressive strength. Twelve input parameters were considered: effective water-to-cement ratio, aggregate-to-cement ratio, RCA replacement ratio, parent concrete strength, maximum RCA size, maximum natural aggregate size, RCA bulk density, natural aggregate bulk density, RCA water absorption, natural aggregate water absorption, RCA Los Angeles abrasion and natural aggregate Los Angeles abrasion. Compressive strength was used as the sole output variable. Pearson correlation analysis showed that effective water-to-cement ratio had the strongest negative correlation with compressive strength ($\gamma=-0.4835$), followed by maximum RCA size ($r=-0.3883$) and aggregate-to-cement ratio ($\gamma=-0.3390$). An 80:20 training-testing strategy was adopted. RF achieved testing values of $R^2=0.8827$, $R^2=0.7791$, RMSE = 6.4922 MPa and MAE = 4.8381 MPa, whereas XGBoost achieved $R^2=0.9084$, $R^2=0.8239$, RMSE = 5.7973 MPa and MAE = 4.1539 MPa. XGBoost consequently demonstrated superior predictive performance across all principal evaluation criteria. The findings indicate that ensemble tree-based models can effectively capture the nonlinear relationships governing RAC compressive strength and provide a useful computational tool for preliminary structural-material assessment.

Keywords: Recycled aggregate concrete; compressive strength; recycled concrete aggregate; machine learning; Random Forest; XGBoost

1. INTRODUCTION

The construction sector is one of the major consumers of natural mineral resources and simultaneously generates substantial quantities of construction and demolition waste. Recycling concrete waste into recycled concrete aggregate (RCA) and subsequently incorporating RCA into new concrete can reduce natural aggregate consumption and contribute to circular construction practices [1–4]. Recent reviews have highlighted the increasing relevance of recycled aggregate concrete (RAC) for sustainable infrastructure while also identifying variability in recycled aggregate quality as one of the principal barriers to wider structural application [5–8].

The mechanical properties of RAC are influenced by several characteristics of the recycled aggregate, including density, water absorption, abrasion resistance and the quality of the residual mortar attached to the aggregate surface [9–11]. In comparison with natural aggregates, RCA generally exhibits greater porosity and water absorption and lower density, which can affect both fresh and hardened concrete properties [12–14]. The quality of the original or parent concrete from which RCA is produced is also important because it affects the physical and mechanical characteristics of the resulting recycled aggregate [15].

Compressive strength is one of the most important mechanical properties used for assessing concrete quality and structural suitability. Conventional determination of compressive strength requires specimen preparation, curing and laboratory testing. Although such testing remains essential for experimental validation, it is time-consuming and does not directly provide a generalized predictive relationship between mixture variables and strength.

Machine learning (ML) offers an alternative data-driven approach for describing complex nonlinear relationships between concrete constituents and engineering properties. Several recent studies have demonstrated the applicability of ML to RAC strength prediction. For example, RF, gradient boosting and other ensemble methods have been used for RAC compressive-strength prediction [16–19].

A comprehensive study by Zhang et al. developed ML models for RAC compressive-strength prediction and used twelve RAC-related factors, including effective water-cement ratio, aggregate-cement ratio, RCA replacement ratio, parent concrete strength and aggregate characteristics [20]. Other investigations have developed databases containing several hundred RAC observations and demonstrated that ML models can provide improved prediction compared with conventional empirical approaches [21,22].

More recently, Golafshani et al. developed a large RAC database and applied multiple ML algorithms to sustainable mix design, while Zhao et al. critically reviewed ML applications for RAC properties and emphasized the importance of input-variable selection, model choice and hyperparameter selection [23,24]. Recent work has further extended RAC prediction toward geopolymer RAC, self-compacting RAC, fiber-reinforced RAC and experimentally validated ML-based mix design [25–28].

Despite these developments, there remains a need for relatively compact and practically interpretable prediction frameworks that focus on a physically meaningful set of RCA, aggregate and mixture characteristics. Therefore, the present study aims to develop and systematically curate a comprehensive database of 637 RAC observations and investigate the influence of twelve mixture and aggregate-related parameters on compressive strength. Based on the developed database, Random Forest (RF) and Extreme Gradient Boosting (XGBoost) regression models are developed using the same training and testing observations to ensure a consistent and reliable comparison. The predictive performance of the models is assessed using the Correlation Coefficient (R), Coefficient of Determination (R^2), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Nash–Sutcliffe Efficiency (NSE), Average Absolute Deviation (AAD), and Scatter Index (SI). Furthermore, feature-importance analysis is performed to identify the relative contribution of the selected RAC mixture and aggregate characteristics to compressive-strength prediction. The study ultimately seeks to identify the more suitable ensemble-learning approach for accurate and reliable prediction of RAC compressive strength and to provide an interpretable machine-learning framework for structural concrete applications.

2. Methodology

The methodology adopted in the present study was developed to predict the compressive strength of recycled aggregate concrete (RAC) using two ensemble machine-learning algorithms, namely Random Forest (RF) and Extreme Gradient Boosting (XGBoost). The overall methodology comprised database development, selection of relevant input and output variables, data preprocessing, statistical screening, correlation analysis, dataset partitioning, machine-learning model development, hyperparameter optimization, predictive performance evaluation, and feature-importance analysis. The complete database-development and machine-learning workflow is presented in Figure 1.

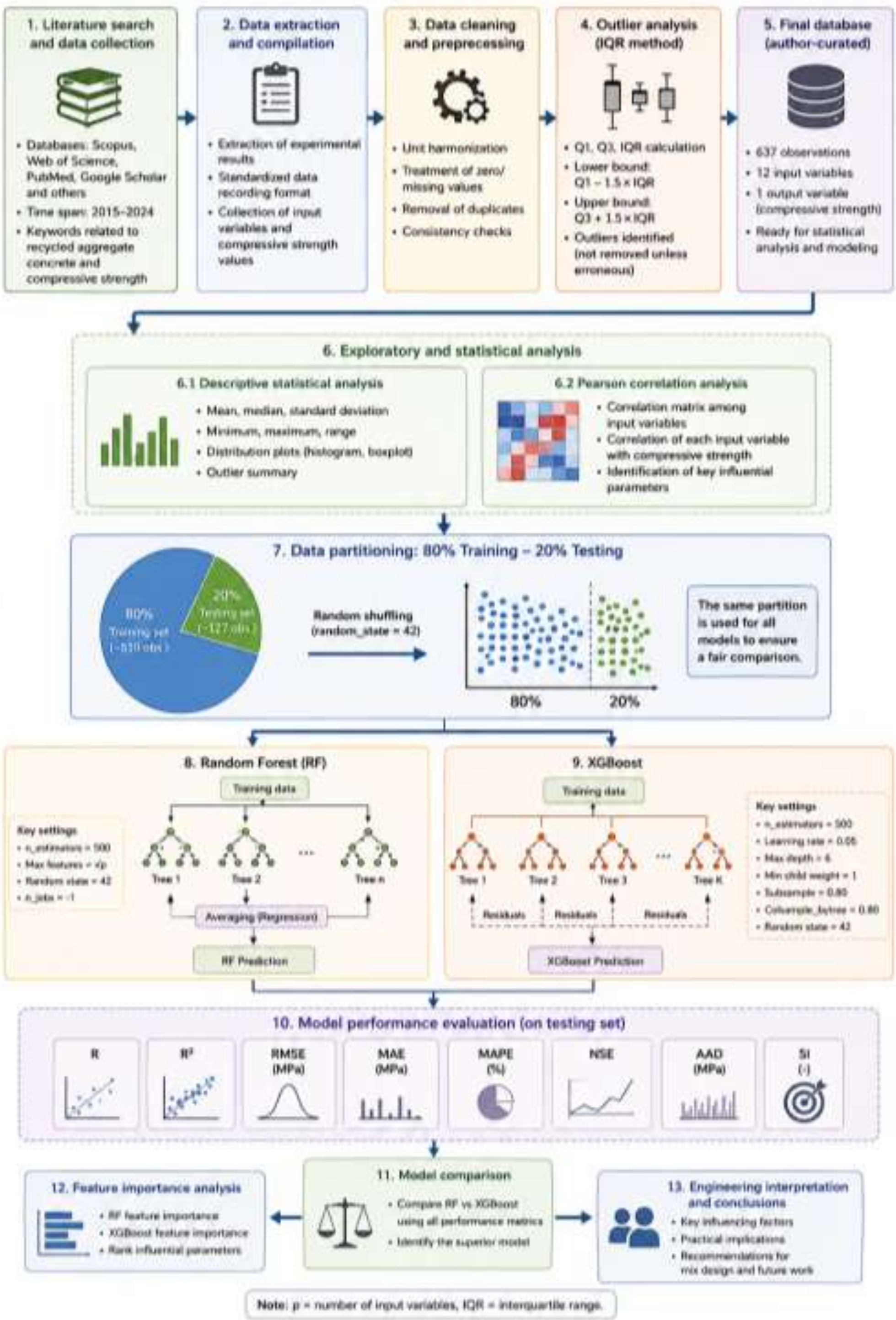

Figure 1. Database-development and machine-learning workflow

2.1 Database Development and Selection of Variables

Database comprising 637 observations of recycled aggregate concrete was developed and systematically curated for the present machine-learning investigation. The database was structured to establish the relationship between concrete mixture parameters, recycled aggregate characteristics, natural aggregate properties, parent concrete strength, and the resulting compressive strength of RAC.

The database included observations covering different effective water-to-cement ratios, aggregate-to-cement ratios, RCA replacement levels, aggregate sizes, aggregate densities, water absorption characteristics, abrasion properties, and parent concrete strengths. The observations were converted into a consistent numerical format and harmonized with respect to units before being used for statistical and machine-learning analysis.

Twelve parameters were selected as independent input variables based on their engineering relevance to RAC compressive-strength behaviour. These variables were effective water-to-cement ratio, aggregate-to-cement ratio, RCA replacement percentage, parent concrete strength, maximum RCA size, maximum natural aggregate size, RCA bulk density, natural aggregate bulk density, RCA water absorption, natural aggregate water absorption, RCA Los Angeles abrasion, and natural aggregate Los Angeles abrasion. The compressive strength of RAC (MPa) was retained as the only dependent/output variable. The complete input-output configuration adopted in the present study is provided in Table 1.

Table 1. Input and output variables used for RAC compressive-strength prediction

No.	Variable	Symbol	Unit	Type
1	Effective water-to-cement ratio	X1	–	Input
2	Aggregate-to-cement ratio	X2	–	Input
3	RCA replacement	X3	%	Input
4	Parent concrete strength	X4	MPa	Input
5	Maximum RCA size	X5	mm	Input
6	Maximum natural aggregate size	X6	mm	Input
7	RCA bulk density	X7	kg/m^3	Input
8	Natural aggregate bulk density	X8	kg/m^3	Input
9	RCA water absorption	X9	%	Input
10	Natural aggregate water absorption	X10	%	Input
11	RCA Los Angeles abrasion	X11	%	Input
12	Natural aggregate Los Angeles abrasion	X12	%	Input
13	Compressive strength	Y	MPa	Output

The relationship to be learned by the machine-learning models can therefore be expressed as:

$$\mathbf{Y = f\,(X1, X2, X3, \ldots, X12)}$$

where Y represents the compressive strength of RAC and X1– X12 represent the selected input variables. This input-output configuration was subsequently used consistently for the Random Forest and XGBoost models.

2.2 Data Preprocessing and Quality Control

Data preprocessing was conducted before model development to ensure that the database was numerically consistent and suitable for statistical analysis and machine-learning applications. The preprocessing procedure included numerical conversion, unit harmonization, examination of missing and zero values, consistency checking, and statistical screening.

All selected variables were converted into numerical format. Measurements were represented using the units specified in Table 1. The resulting dataset therefore contained only numerical predictor and response variables, which allowed the same database structure to be used by both RF and XGBoost.

Particular attention was given to zero values. A zero value was not automatically treated as a valid measurement because zero does not represent a physically meaningful value for several aggregate properties, such as bulk density, water absorption, abrasion resistance, or parent concrete strength. Where a zero represented an unavailable or unreported value rather than an actual physical measurement, it was suitably replaced by a non-zero value during preprocessing. This procedure prevented the machine-learning models from interpreting unavailable measurements as genuine zero-valued engineering properties.

The database was also examined for missing and duplicate observations. Records were checked for numerical consistency before being incorporated into the final dataset. Following the preprocessing and quality-control procedure, 637 observations were retained for subsequent analysis.

2.3 Statistical Analysis and Outlier Screening

Descriptive statistical analysis was performed to characterize the distribution and variability of the input and output variables. The first quartile (Q1), third quartile (Q3), interquartile range (IQR), lower boundary, and upper boundary were determined for each numerical variable.

Observations outside these boundaries were identified as potential statistical outliers. However, an observation was not automatically considered erroneous merely because it fell outside the IQR limits. Since

RAC is a heterogeneous structural material, genuine variations in mixture composition and aggregate properties may produce statistically extreme but physically plausible observations. Therefore, IQR analysis was used primarily for statistical screening and characterization, while physically meaningful observations were retained. Table 2. Presents the statistical characteristics and IQR-based screening of the RAC database.

Table 2. Statistical characteristics and IQR-based screening of the RAC database

Variable	Q1	Q3	IQR	Lower bound	Upper bound	Outliers
Effective w/c	0.41	0.55	0.14	0.2	0.76	9
Aggregate/cement ratio	2.5	3.3	0.8	1.3	4.5	37
RCA replacement (%)	50	100	50	−25	175	0
Parent concrete strength (MPa)	41.4	41.4	0	41.4	41.4	56
Maximum RCA size (mm)	19	25	6	10	34	2
Maximum NA size (mm)	20	25	5	12.5	32.5	27
RCA bulk density (kg/m^3)	2360	2430	70	2255	2535	93
NA bulk density (kg/m^3)	2620	2652	32	2572	2700	179
RCA water absorption (%)	4.7	5.4	0.7	3.65	6.45	165
NA water absorption (%)	0.9	1.1	0.2	0.6	1.4	154
RCA LA abrasion	34.3	34.3	0	34.3	34.3	126
NA LA abrasion	24.8	24.8	0	24.8	24.8	98
Compressive strength (MPa)	33.1	50	16.9	7.75	75.35	13

The effective ratio showed a central distribution between 0.41 and 0.55, with an interquartile range of 0.14. The RCA replacement percentage extended from partial replacement to complete replacement, while the compressive-strength values covered a broad range from approximately 7.75 to 75.35 MPa. Such variation in the response variable is particularly important for regression modeling because it enables the models to learn the relationship between different RAC mixture and aggregate characteristics and the resulting compressive strength.

The statistical screening also demonstrated considerable variation in several RCA-related parameters. RCA bulk density, RCA water absorption, and RCA Los Angeles abrasion exhibited relatively wide distributions. Similar variability was observed in the corresponding natural aggregate properties. This variation indicates that the developed database incorporates RAC mixtures with substantially different aggregate characteristics rather than being restricted to a narrow range of material properties.

2.4 Correlation Analysis

Pearson correlation analysis was performed to quantify the linear association between each input variable and RAC compressive strength. A correlation heatmap was generated to provide a visual representation of the interrelationships among all input variables and the compressive-strength output as shown in Figure 2. The numerical Pearson coefficients were displayed within the heatmap to facilitate direct quantitative interpretation.

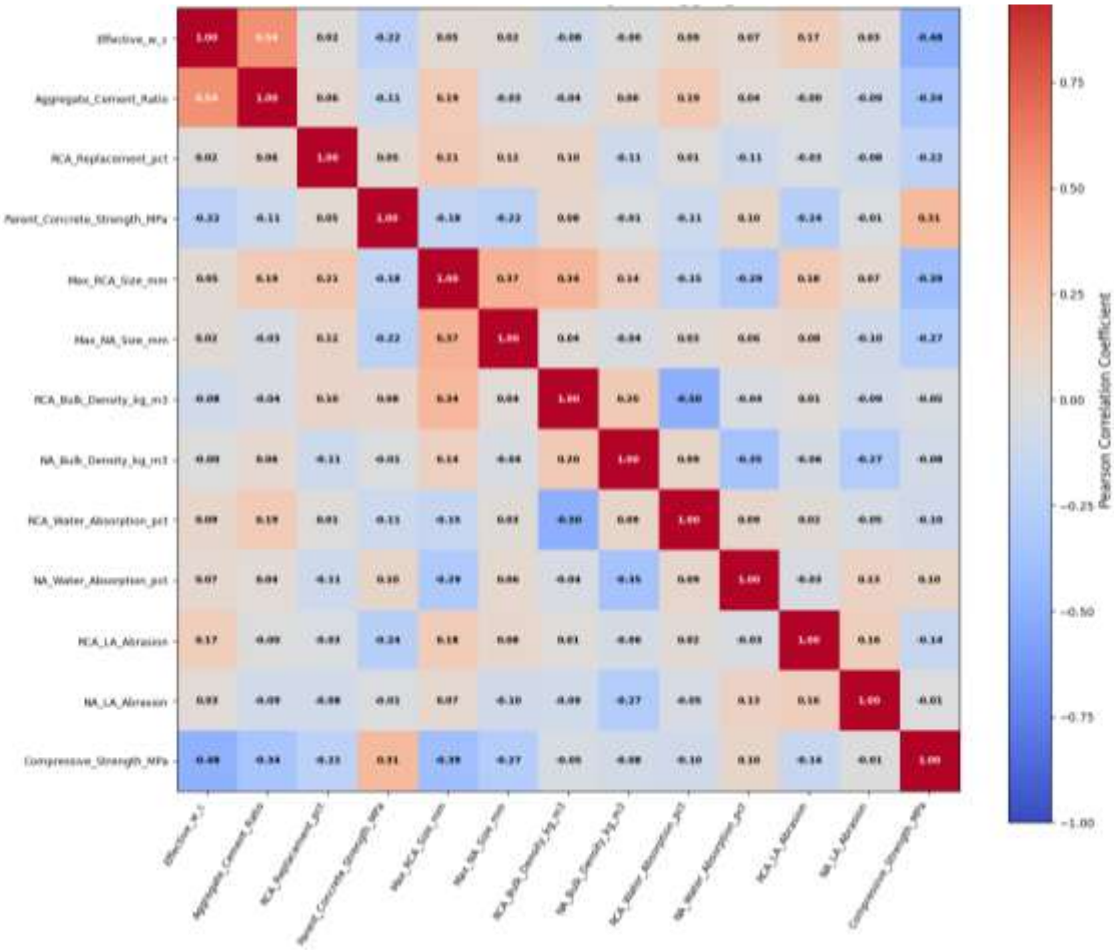

Figure 2. Pearson correlation heatmap of the input variables and compressive strength of RAC

The effective w/c ratio exhibited the strongest correlation with compressive strength among the selected input variables, with a Pearson correlation coefficient of $R=-0.4835$. This indicates a moderate negative linear association, suggesting that higher effective w/c ratios are generally associated with lower compressive-strength values within the developed database.

Maximum RCA size and aggregate-to-cement ratio exhibited correlation coefficients of −0.3883 and −0.3390, respectively. RCA replacement percentage showed a comparatively weaker negative correlation of −0.2177. Parent concrete strength exhibited a positive correlation of 0.3130, indicating that higher parent concrete strength was generally associated with higher RAC compressive strength.

The remaining variables exhibited relatively weaker linear correlations. RCA Los Angeles abrasion, RCA water absorption, RCA bulk density, natural aggregate bulk density, and natural aggregate water absorption showed correlation coefficients of −0.1446, −0.0985, −0.0453, −0.0768, and 0.0978, respectively. The natural aggregate Los Angeles abrasion value exhibited a nearly negligible linear correlation of −0.0063.

These results demonstrate that Pearson correlation alone does not fully describe the predictive importance of the input variables. In particular, machine-learning models can capture nonlinear relationships and interactions that may not be reflected by individual linear correlation coefficients. Therefore, the subsequent RF and XGBoost analyses provide a more comprehensive assessment of the predictive relationships within the RAC database.

2.5 Dataset Partitioning

The final database containing 637 observations was divided into training and testing subsets using an 80:20 ratio. Approximately 80% of the observations (510 records) were used for model training, whereas the remaining 20% (127 records) were reserved for independent testing.

A fixed random state of 42 was adopted to ensure reproducibility of the data partition. Importantly, the same training and testing subsets were used for both RF and XGBoost, thereby ensuring a fair comparison between the two algorithms. The testing observations were not used during model fitting or hyperparameter optimization and were retained exclusively for the final evaluation of model generalization.

2.6 Random Forest Regression

Random Forest (RF) is an ensemble-learning algorithm that combines the predictions of multiple decision trees to obtain a robust regression output. Each tree is developed using randomly selected training observations and predictor variables, while the final prediction is obtained by averaging the individual tree predictions:

$$\hat{Y} = \frac{1}{N}\sum_{i=1}^{N} T_i\,(X)$$

where Y is the predicted compressive strength, N is the number of decision trees, Ti X represents the prediction of the ith decision tree, and X denotes the input feature vector.

RF was selected because of its ability to model nonlinear relationships and interactions among concrete mixture parameters and aggregate characteristics without requiring a predefined functional relationship between the predictors and RAC compressive strength.

The RF model was optimized by varying the number of estimators, maximum tree depth, maximum number of features considered for splitting, minimum samples required for a split, and minimum samples required at a leaf node. A five-fold cross-validation procedure was used for hyperparameter optimization.

The optimized RF configuration consisted of 600 estimators, unlimited maximum depth, maximum feature fraction of 1.0, minimum samples per leaf of 1, and minimum samples per split of 2.

2.7 Extreme Gradient Boosting Regression

Extreme Gradient Boosting (XGBoost) is an ensemble regression technique based on the sequential construction of decision trees. Each successive tree is developed to reduce the residual errors produced by the existing ensemble, thereby progressively improving prediction accuracy.

The general XGBoost objective function can be expressed as:

$$L = \sum_{i=1}^{n} l\left(Y_i, \hat{Y}_i\right) + \sum_{j} \Omega\left(f_j\right)$$

where Yi, and Yi represent the observed and predicted compressive strengths, respectively, l(·) denotes the loss function, fj represents the jth regression tree, and Ωfj represents the regularization term associated with the tree. The regularization component controls model complexity and helps reduce overfitting.

The XGBoost model was optimized with respect to the number of estimators, learning rate, maximum tree depth, minimum child weight, subsampling ratio, and column-sampling ratio. Five-fold cross-validation was used for hyperparameter optimization.

The optimized XGBoost configuration consisted of 200 estimators, a learning rate of 0.1, maximum tree depth of 2, minimum child weight of 1, subsampling ratio of 0.8, and column-sampling ratio of 1.0.

2.8 Model Performance Evaluation

The predictive performance of the developed Random Forest (RF) and Extreme Gradient Boosting (XGBoost) models was evaluated using widely accepted statistical performance indicators. These metrics were selected to assess correlation, goodness of fit, magnitude of prediction errors, relative prediction error, predictive efficiency, and normalized error between the observed and model-predicted compressive strengths.

The evaluation metrics included the Correlation Coefficient (R), Coefficient of Determination (R^2), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Nash–Sutcliffe Efficiency (NSE), Average Absolute Deviation (AAD), and Scatter Index (SI).

2.9 Feature Importance Analysis

Feature importance analysis was performed to determine the relative contribution of the twelve input variables to the prediction of RAC compressive strength.

For RF, feature importance was obtained from the contribution of individual predictors to impurity reduction across the ensemble of decision trees. For XGBoost, feature importance was determined from the contribution of individual predictors to improvements in the boosting objective.

The feature-importance results were subsequently used to identify the dominant predictors and to provide an engineering interpretation of the machine-learning models. The analysis also provides an additional basis for comparing the predictive behavior of RF and XGBoost beyond conventional statistical performance metrics.

3. 3. Results and Discussion

3.1 Random Forest Prediction Performance

The predictive performance of the optimized Random Forest (RF) model was evaluated using the independent testing dataset. The analysis was conducted through four complementary aspects: actual versus predicted compressive strength, regression fit, residual analysis, and RF feature importance. These analyses provide both quantitative and graphical assessment of the predictive capability and interpretability of the RF model.

3.1.1 Actual versus Predicted Compressive Strength

The comparison between the observed (actual) and RF-predicted compressive strength values was performed for both the training and testing datasets. The actual-versus-predicted plots provide a direct

visual assessment of the agreement between the experimentally reported and model-predicted values.
For an ideal prediction, the data points should be distributed closely around the 1:1 line, indicating that the predicted compressive strength is equal to the observed value. The RF model produced a strong agreement between actual and predicted values for the training dataset, as reflected by the Correlation Coefficient (R) of 0.986666 and Coefficient of Determination (R^2) of 0.968443. The corresponding RMSE and MAE were 2.629394 MPa and 1.897056 MPa, respectively.
For the independent testing dataset, the RF model achieved an R of 0.882703 and an R^2 of 0.779092, with an RMSE of 6.492188 MPa and MAE of 4.838119 MPa. Although a reduction in predictive performance was observed when moving from the training to the testing dataset, the predicted values remained reasonably associated with the actual compressive-strength observations.

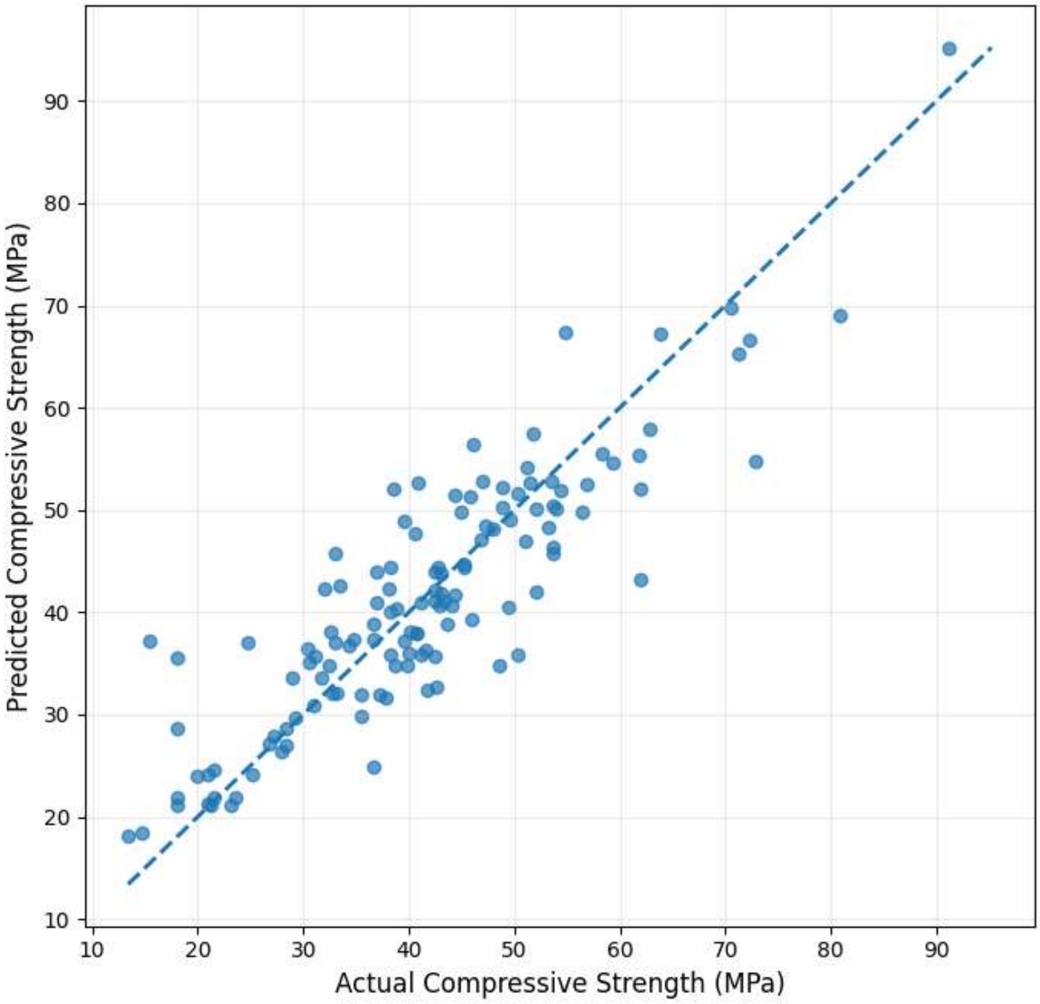

Figure 3. Actual versus RF-predicted compressive strength

3.1.2 Regression Fit Analysis

Regression-fit analysis was performed to further quantify the relationship between the actual and RF-predicted compressive-strength values. The fitted regression line was compared with the ideal 1:1 reference line. A regression line close to the 1:1 line indicates better agreement between observed and predicted values.
The training dataset produced a strong regression relationship, with R = 0.986666 and R^2 = 0.968443. This indicates that approximately 96.84% of the variation in the observed compressive strength was explained by the RF predictions for the training observations.
For the testing dataset, the RF model obtained R = 0.882703 and R^2 = 0.779092, indicating that approximately 77.91% of the variation in the observed compressive strength was explained by the model predictions for previously unseen observations.
The regression-fit results therefore demonstrate that the RF model captured the nonlinear relationship between the selected RAC input variables and compressive strength. However, the reduction in from training to testing also indicates that the model performance should primarily be judged from the independent testing results rather than from training accuracy alone.

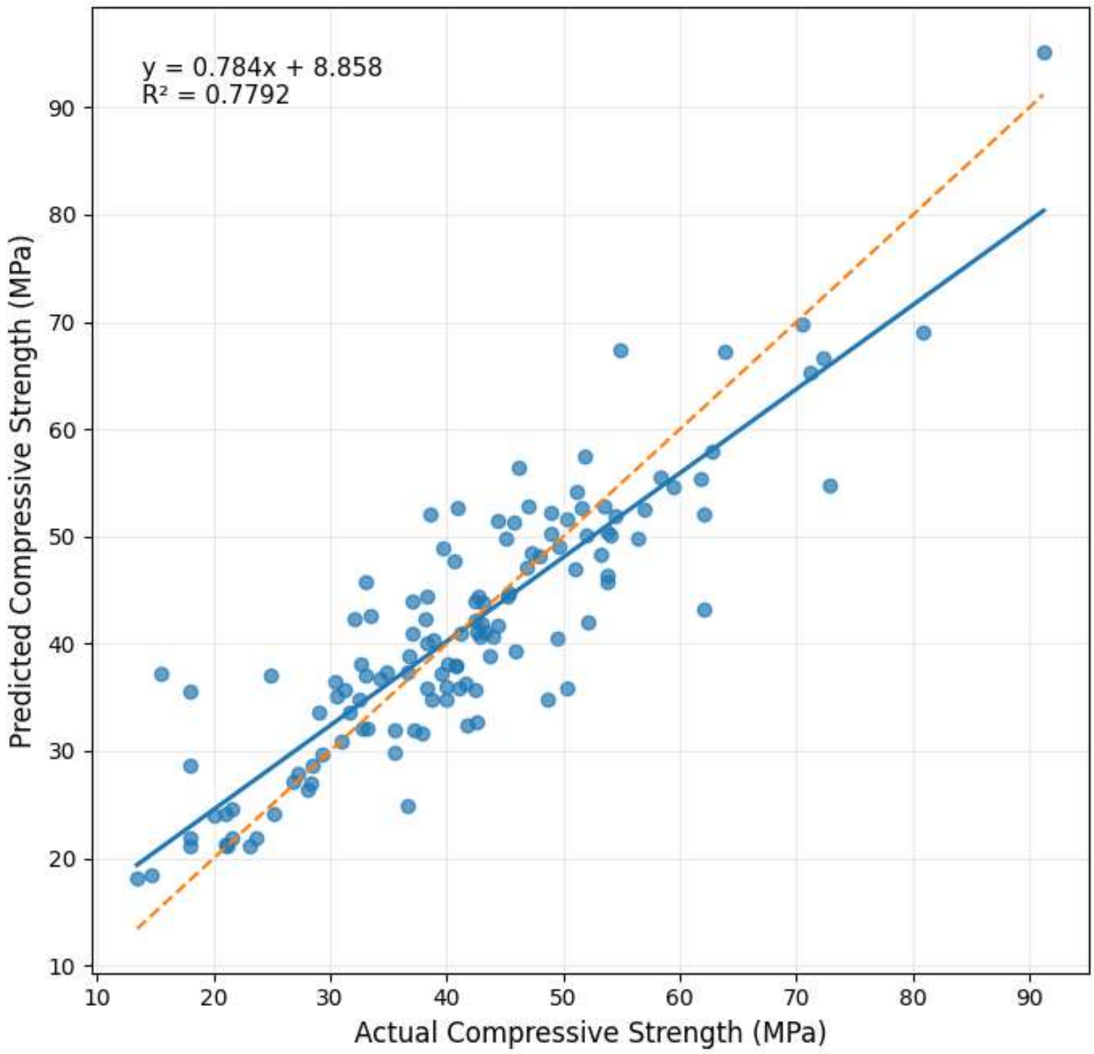

Figure 4. Regression-fit analysis between actual and RF-predicted compressive strength

3.1.3 Residual Analysis

Residual analysis was performed to examine the distribution and magnitude of prediction errors generated by the RF model. For a well-performing regression model, the residuals should be distributed around zero without a pronounced systematic pattern. The RF model produced a lower magnitude of error for the training observations, as demonstrated by an RMSE of 2.629394 MPa, MAE of 1.897056 MPa, and AAD of 1.897056 MPa.

For the testing dataset, the RMSE increased to 6.492188 MPa, while the MAE and AAD were 4.838119 MPa. The corresponding MAPE was 13.295055% and the Scatter Index (SI) was 0.156680. These values indicate that the prediction deviations were greater for the unseen testing observations than for the training observations.

The residual analysis complements the and results by providing information on the magnitude of individual prediction errors. The testing residuals should therefore be examined together with the actual-versus-predicted and regression-fit plots to assess whether any systematic bias exists across the compressive-strength range.

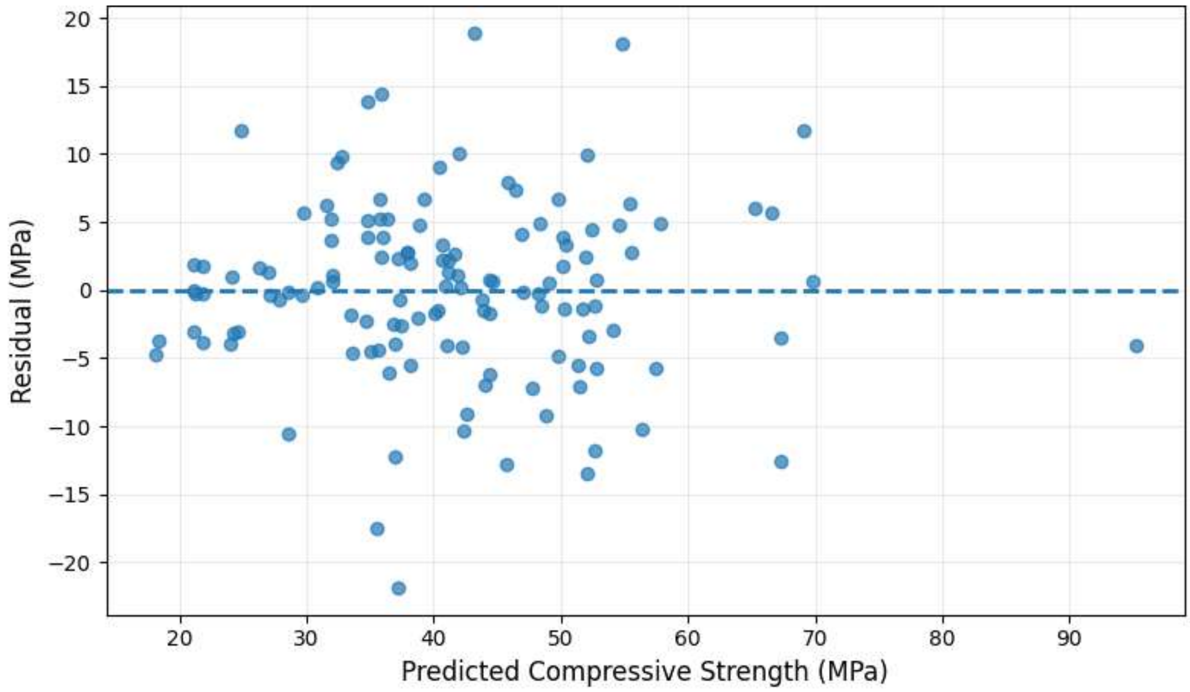

Figure 5. Residual analysis of the optimized RF model

3.3.4 Random Forest Feature Importance

The relative contribution of the input variables to the prediction of RAC compressive strength was evaluated using the feature-importance function of the optimized Random Forest (RF) model. The results are presented in Figure 6. The feature-importance values represent the relative contribution of each predictor within the RF ensemble, with the total importance normalized to 100%.

The effective ratio was identified as the most influential input variable, contributing 26.9338% to the RF prediction. This finding is consistent with the Pearson correlation analysis, where effective exhibited the strongest linear correlation with compressive strength (). The relatively high RF importance indicates that the effective water-to-cement ratio plays a substantial role in differentiating compressive-strength responses within the developed RAC database.

The maximum RCA size was the second most influential variable, with an importance of 16.8556%, followed by the maximum natural aggregate size, which contributed 14.0967%. Together, these results indicate that aggregate size characteristics constitute important predictors of RAC compressive strength in the developed RF model.

The aggregate-to-cement ratio ranked fourth, with an importance of 10.1133%. RCA Los Angeles abrasion contributed 7.7091%, indicating a comparatively important contribution of RCA mechanical characteristics to the prediction. RCA bulk density and RCA water absorption contributed 4.7878% and 4.4436%, respectively.

The remaining variables showed comparatively lower contributions. Natural aggregate water absorption and natural aggregate bulk density accounted for 4.1361% and 3.8842%, respectively. RCA replacement percentage contributed 2.9325%, while parent concrete strength contributed 2.8893%. Natural aggregate Los Angeles abrasion exhibited the lowest RF feature importance, at 1.2180%.

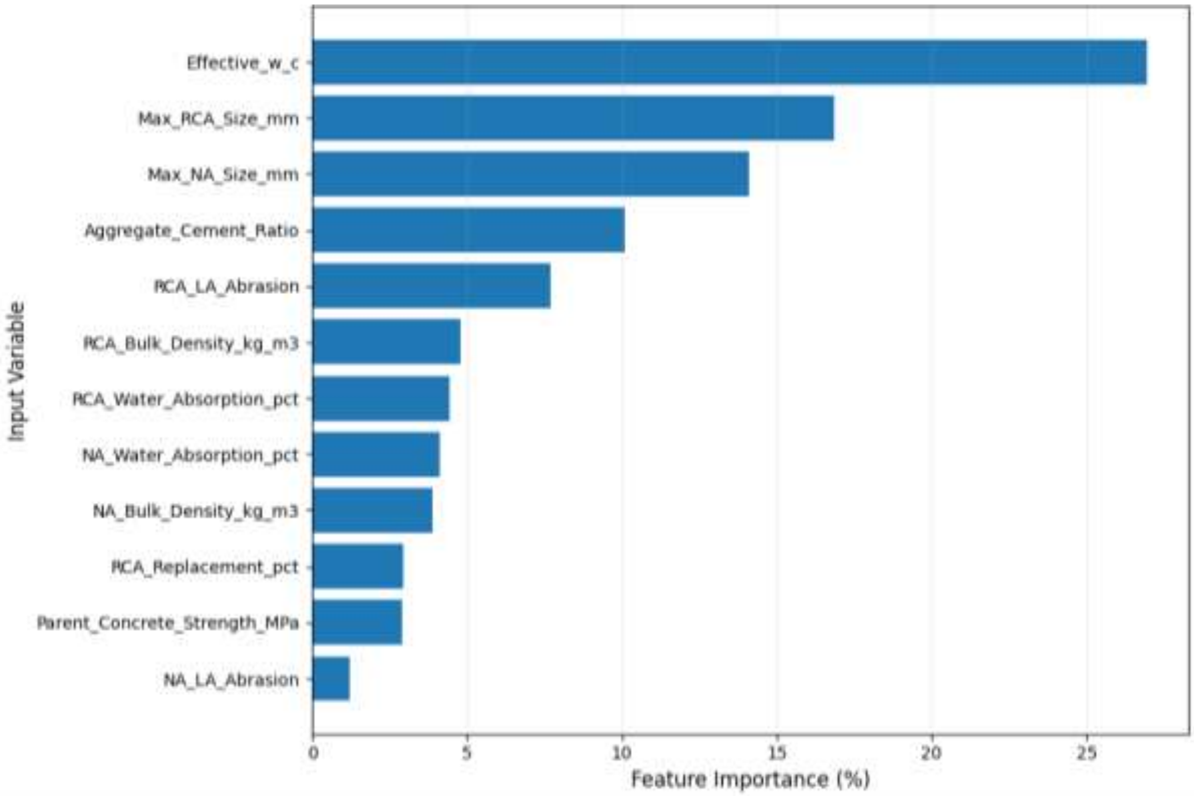

Figure 6. Feature importance of input variables in the optimized Random Forest model.

3.4 Extreme Gradient Boosting (XGBoost) Model Performance

The optimized Extreme Gradient Boosting (XGBoost) model was evaluated using the same training and testing datasets adopted for the Random Forest (RF) model. The analysis considered the actual-versus-predicted relationship, regression fit, residual characteristics, and relative importance of the input variables. The results obtained from these analyses provide both quantitative evidence of model accuracy and an interpretation of the variables governing the prediction of RAC compressive strength.

3.4.1 Actual versus Predicted Compressive Strength

Figure 7 presents the comparison between the observed and XGBoost-predicted compressive-strength values for the training and testing datasets. The closeness of the data points to the 1:1 line indicates the agreement between the observed and predicted values.

For the training dataset, XGBoost produced a very strong agreement, with R = 0.998614 and R^2 = 0.997179. The model also produced a relatively low RMSE of 0.786190 MPa and MAE of 0.510694 MPa. The corresponding MAPE, NSE, AAD, and SI values were 1.355478%, 0.997179, 0.510694 MPa, and 0.018325, respectively. The close clustering of the training observations around the 1:1 line is therefore consistent with the high statistical performance of the model.

For the independent testing dataset, the predicted values remained strongly correlated with the observed compressive strengths, producing R = 0.908395 and R^2 = 0.823852. The corresponding RMSE and MAE were 5.797290 MPa and 4.153936 MPa, respectively. The testing MAPE, NSE, AAD, and SI were 11.314059%, 0.823852, 4.153936 MPa, and 0.139910, respectively.

The reduction in performance from training to testing is expected because the testing observations were not used during model fitting. Nevertheless, the testing results demonstrate that the XGBoost model retained a substantial predictive capability when applied to unseen RAC observations.

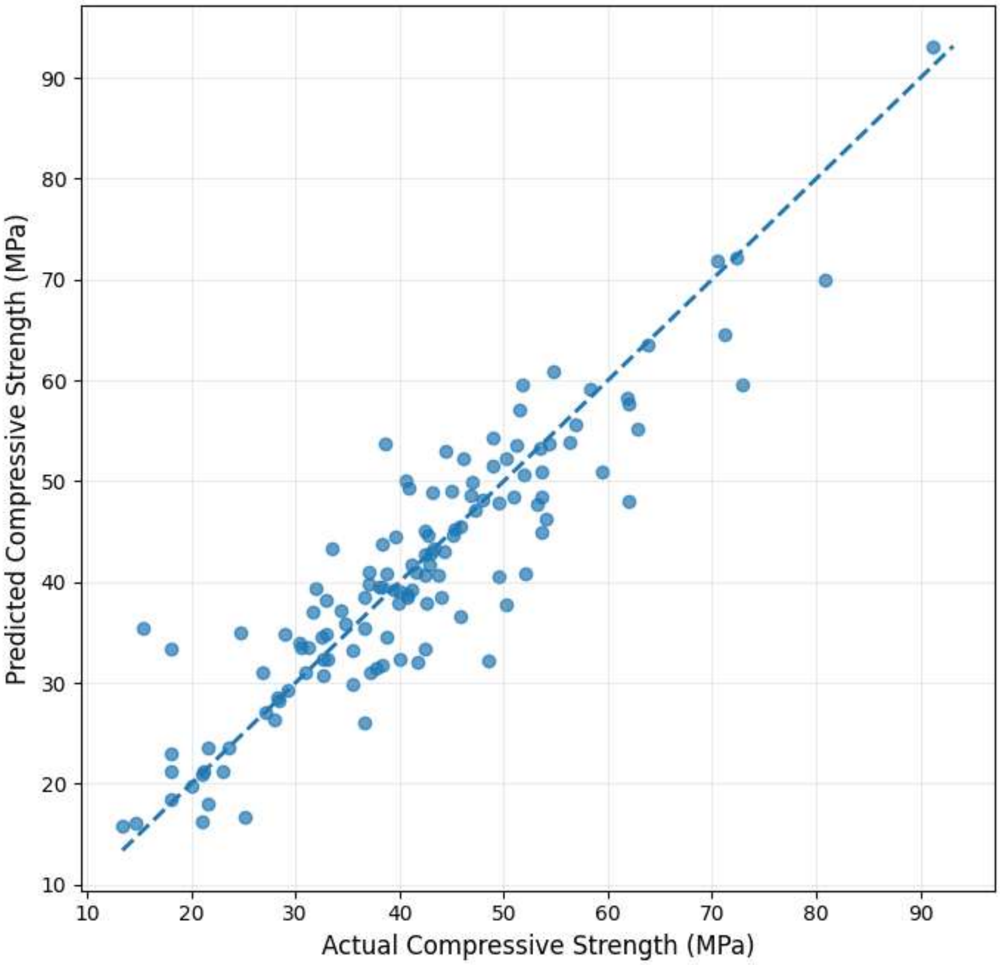

Figure 7. Actual versus XGBoost-predicted compressive strength

3.4.2 Regression Fit Analysis

The regression-fit analysis provides a quantitative assessment of the relationship between observed and XGBoost-predicted compressive strength. Figure 8 presents the regression relationship together with the ideal 1:1 line.

The training regression produced R^2 = 0.997179, indicating that approximately 99.72% of the variation in the observed compressive strength was explained by the XGBoost predictions. The corresponding correlation coefficient of R = 0.998614 demonstrates a very strong positive association between the observed and predicted values.

For the testing dataset, the model achieved R^2 = 0.823852 and R = 0.908395. Thus, approximately 82.39% of the variation in the observed compressive strength was explained by the XGBoost model for the independent testing observations. The testing regression relationship was therefore weaker than the training relationship but remained strong enough to demonstrate useful predictive capability.

The testing value obtained with XGBoost was also higher than that obtained with RF (0.823852 versus 0.779092). This indicates that XGBoost explained a greater proportion of the variation in the unseen compressive-strength observations.

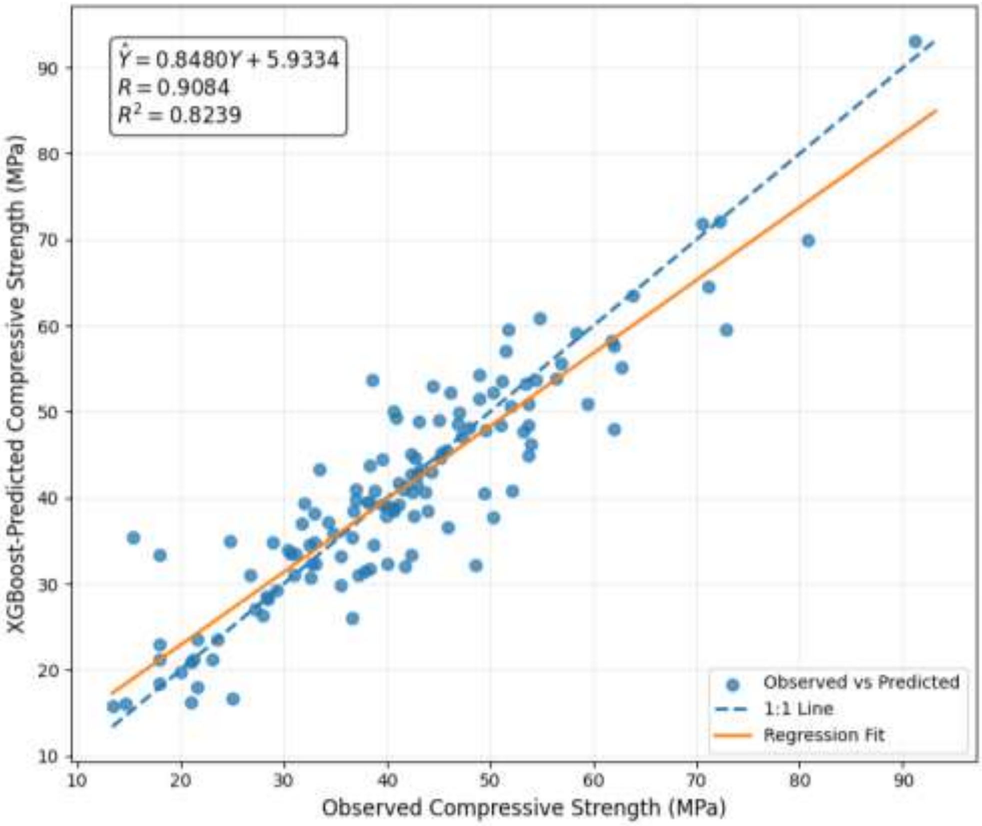

Figure 8. Regression-fit analysis between observed and XGBoost-predicted compressive strength

3.4.3 Residual Analysis

Residual analysis was performed to examine the prediction errors generated by the XGBoost model. The residual for each observation was calculated as the difference between the observed and predicted compressive strength:

where is the residual error, is the observed compressive strength, and is the XGBoost-predicted compressive strength.

The training dataset exhibited relatively small residual errors, as demonstrated by the RMSE of 0.786190 MPa, MAE of 0.510694 MPa, and AAD of 0.510694 MPa. The low SI of 0.018325 further indicates that the prediction errors were small relative to the magnitude of the observed compressive-strength values.

For the testing dataset, the error magnitude increased, with an RMSE of 5.797290 MPa, MAE of 4.153936 MPa, and AAD of 4.153936 MPa. The testing MAPE was 11.314059%, while the SI was 0.139910. The increase in residual magnitude from training to testing reflects the reduction in predictive accuracy when the model encounters previously unseen observations.

However, the testing residual indicators remain consistent with the relatively high and values. The NSE of 0.823852 further confirms that the XGBoost model provided substantial predictive efficiency for the independent testing dataset. Therefore, the residual analysis supports the conclusion that the model was capable of capturing the principal nonlinear relationships present in the RAC database.

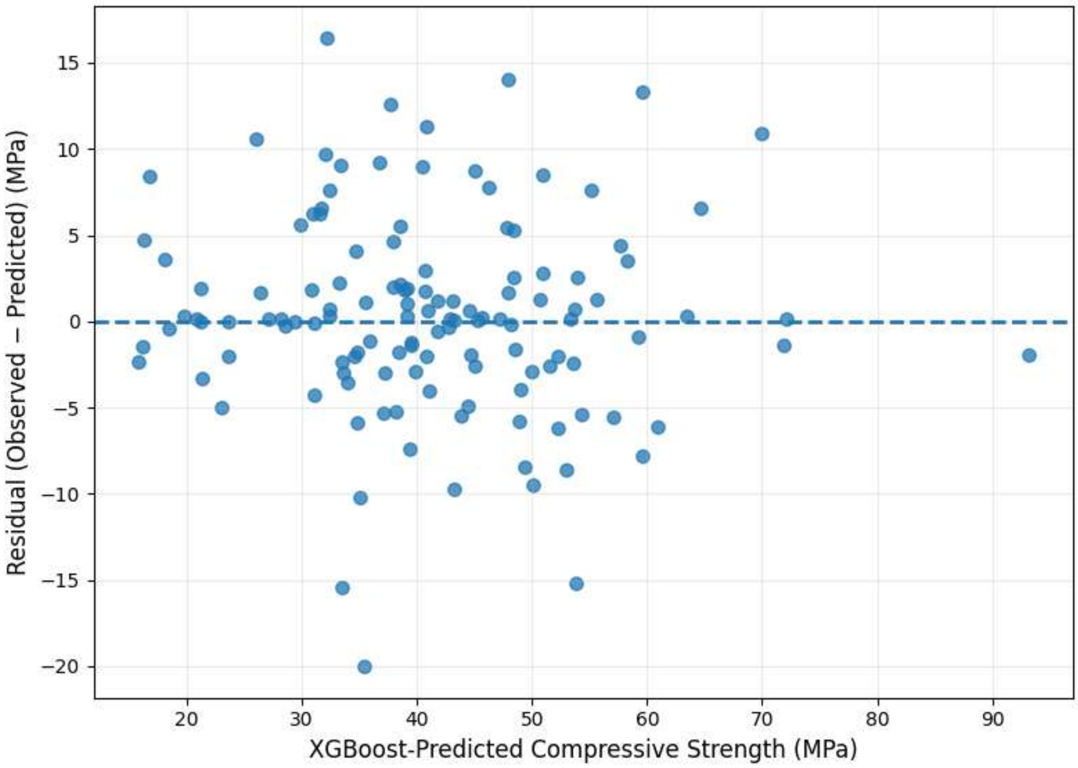

Figure 9. Residual analysis of the optimized XGBoost model

3.4.4 XGBoost Feature Importance

The relative contribution of the twelve input variables was evaluated using the feature-importance values obtained from the optimized XGBoost model. The results are presented in Figure 10.

RCA Los Angeles abrasion was identified as the most influential variable, with an importance of 25.4497%. This was followed by maximum RCA size (17.4634%) and maximum natural aggregate size (12.8175%). The effective ratio ranked fourth, contributing 8.6078%.

The dominance of RCA Los Angeles abrasion indicates that the mechanical characteristics of recycled coarse aggregate made a substantial contribution to the XGBoost prediction. Maximum RCA size and maximum natural aggregate size also exhibited high importance, demonstrating the relevance of aggregate-size characteristics in predicting RAC compressive strength.

The NA bulk density contributed 6.7299%, followed by NA water absorption (5.6360%) and parent concrete strength (5.6295%). RCA water absorption and RCA bulk density contributed 4.5290% and 4.0878%, respectively. The remaining variables—NA Los Angeles abrasion, aggregate-to-cement ratio, and RCA replacement percentage—showed comparatively lower importance values of 3.6710%, 2.9295%, and 2.4489%, respectively.

The four most influential variables—RCA LA abrasion, maximum RCA size, maximum NA size, and effective —accounted for approximately 64.34% of the total XGBoost feature importance. This indicates that the model relied predominantly on aggregate mechanical characteristics, aggregate size, and the effective water-to-cement relationship when predicting RAC compressive strength.

A comparison with the RF model provides an additional insight. RF identified effective as its most influential predictor (26.9338%), whereas XGBoost identified RCA LA abrasion as the dominant predictor (25.4497%). Despite this difference, both models ranked maximum RCA size and maximum natural aggregate size among the leading predictors. The consistency in these aggregate-size variables indicates their importance within the developed database, while the different rankings of the remaining variables demonstrate that RF and XGBoost capture nonlinear relationships and interactions differently.

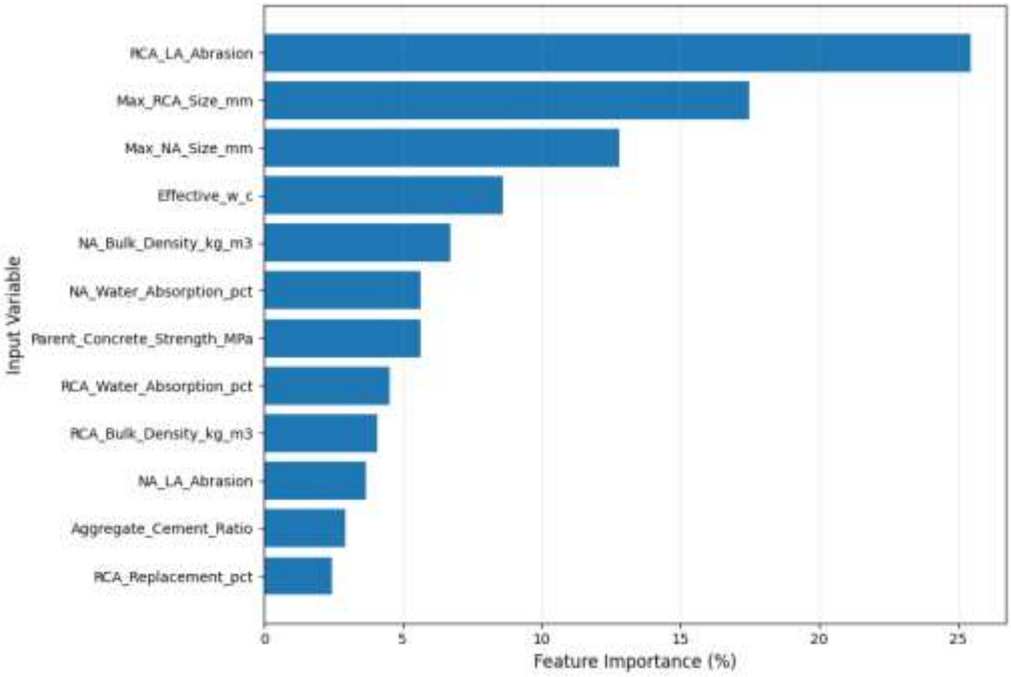

Figure 10. Feature importance of input variables in the optimized XGBoost model

4. Conclusions

The present study developed a machine-learning framework for predicting the compressive strength of recycled aggregate concrete (RAC) using a comprehensive database of 637 observations and twelve mixture- and aggregate-related input variables. Random Forest (RF) and Extreme Gradient Boosting (XGBoost) regression models were developed and evaluated using the same training and independent testing datasets. The principal conclusions are as follows:

1.The developed RAC database exhibited substantial variability in mixture proportions, aggregate characteristics, parent concrete strength, and compressive strength, providing a suitable basis for developing data-driven prediction models.

2.Effective w/c ratio showed the strongest Pearson correlation with compressive strength, with a correlation coefficient of −0.4835, followed by maximum RCA size and aggregate-to-cement ratio. Parent concrete strength exhibited a positive correlation of 0.3130.

3.The optimized RF model achieved R = 0.986666 and for the training dataset and R^2 = 0.882703 and for the testing dataset. The testing RMSE, MAE, MAPE, NSE, AAD, and SI were 6.492188 MPa, 4.838119 MPa, 13.295055%, 0.779092, 4.838119 MPa, and 0.156680, respectively.

4.The optimized XGBoost model demonstrated superior predictive performance compared with RF. It achieved R = 0.998614 and for training and R^2 = 0.908395 and for testing. For the testing dataset, XGBoost produced an RMSE of 5.797290 MPa, MAE of 4.153936 MPa, MAPE of 11.314059%, NSE of 0.823852, AAD of 4.153936 MPa, and SI of 0.139910.

5.The comparative results demonstrate that XGBoost provided better generalization than RF for the present RAC database. XGBoost achieved higher R, R^2 , and NSE values and lower RMSE, MAE, MAPE, AAD, and SI values on the independent testing dataset.

6.RF feature-importance analysis identified effective w/c ratio (26.9338%), maximum RCA size (16.8556%), maximum natural aggregate size (14.0967%), and aggregate-to-cement ratio (10.1133%) as the four most influential variables. Together, these variables accounted for approximately 68.00% of the total RF feature importance.

7.XGBoost feature-importance analysis identified RCA Los Angeles abrasion (25.4497%), maximum RCA size (17.4634%), maximum natural aggregate size (12.8175%), and effective ratio (8.6078%) as the four dominant predictors, collectively accounting for approximately 64.34% of the total feature importance.

8.Although the relative ranking of individual variables differed between RF and XGBoost, both models identified maximum RCA size and maximum natural aggregate size among the most influential predictors. This consistency highlights the importance of aggregate characteristics in predicting RAC compressive strength.

9.Overall, the results demonstrate that ensemble machine-learning techniques, particularly XGBoost, can provide an effective approach for predicting RAC compressive strength from readily available mixture and aggregate characteristics. The developed framework can support preliminary strength estimation and facilitate data-driven assessment of RAC mixtures in structural engineering applications.

Authors' Contributions

Subhash Kumar: Data curation, database development, data preprocessing, machine-learning modelling, model validation, statistical analysis, visualization, and initial manuscript drafting. Pappu Kumar: Conceptualization, research methodology, supervision, interpretation of results, and critical evaluation of the study.

Acknowledgements

The authors sincerely acknowledge the Department of Civil Engineering, Sandip University, Madhubani, India, for providing the academic environment and institutional support necessary for conducting this research. The authors also acknowledge the computational resources and facilities used for data analysis and machine-learning model development.

References

1. Wang B, Yan L, Fu Q, Kasal B. A comprehensive review on recycled aggregate and recycled aggregate concrete. Resources, Conservation and Recycling. 2021;171:105565. doi:10.1016/j.resconrec.2021.105565.
2. Makul N, Fediuk R, Amran M, et al. Design strategy for recycled aggregate concrete: A review of status and future perspectives. Crystals. 2021;11:695. doi:10.3390/cryst11060695.
3. Bu L, Du G, Hou Q. Prediction of the compressive strength of recycled aggregate concrete based on artificial neural network. Materials. 2021;14:3921. doi:10.3390/ma14143921.
4. Vasanthalin CP, Kavitha NC. Prediction of compressive strength of recycled aggregate concrete using artificial neural network and cuckoo search method. Materials Today: Proceedings. 2021;46:8480–8488. doi:10.1016/j.matpr.2021.03.500.
5. Piccinali A, Diotti A, Plizzari G, Sorlini S. Impact of recycled aggregate on the mechanical and environmental properties of concrete: A review. Materials. 2022;15:1818. doi:10.3390/ma15051818.
6. Influence of quality of recycled aggregates on the mechanical properties of recycled aggregate concretes: An overview. Construction and Building Materials. 2022;328:127071. doi:10.1016/j.conbuildmat.2022.127071.
7. Influence of the moisture states of aggregate recycled from waste concrete on the performance of the prepared recycled aggregate concrete: A review. Construction and Building Materials. 2022;326:126891. doi:10.1016/j.conbuildmat.2022.126891.
8. Salgado FA, Silva FA. Recycled aggregates from construction and demolition waste towards an application on structural concrete: A review. Journal of Building Engineering. 2022;52:104452. doi:10.1016/j.jobe.2022.104452.
9. Evaluating compressive strength of concrete made with recycled concrete aggregates using machine learning approach. Construction and Building Materials. 2022;323:126578. doi:10.1016/j.conbuildmat.2022.126578.
10. Compressive strength of concrete with recycled aggregate; a machine learning-based evaluation. Cleaner Materials. 2022;3:100044. doi:10.1016/j.clema.2022.100044.
11. Comparison of prediction models based on machine learning for the compressive strength estimation of recycled aggregate concrete. Materials. 2022;15:3430. doi:10.3390/ma15103430.
12. To predict the compressive strength of self-compacting concrete with recycled aggregates utilizing ensemble machine learning models. Case Studies in Construction Materials. 2022;16:e01046. doi:10.1016/j.cscm.2022.e01046.
13. Munir MJ, Kazmi SMS, Wu YF, Lin X, Ahmad MR. Development of a novel compressive strength design equation for natural and recycled aggregate concrete through advanced computational modeling. Journal of Building Engineering. 2022;55:104690. doi:10.1016/j.jobe.2022.104690.
14. Liu C, Wang Y, Gao X, et al. Review of the strengthening methods and mechanical properties of recycled aggregate concrete. Crystals. 2022;12:1321. doi:10.3390/cryst12091321.
15. Modeling the mechanical properties of recycled aggregate concrete using hybrid machine learning algorithms. Resources, Conservation and Recycling. 2023;190:106812. doi:10.1016/j.resconrec.2022.106812.
16. Mechanical performance of recycled aggregate concrete in green civil engineering: Review. Case Studies in Construction Materials. 2023;19:e02384. doi:10.1016/j.cscm.2023.e02384.
17. Zhang X, Dai C, Li W, Chen Y. Prediction of compressive strength of recycled aggregate concrete using machine learning and Bayesian optimization methods. Frontiers in Earth Science. 2023;11:1112105. doi:10.3389/feart.2023.1112105.
18. Performance of recycled aggregate concrete with supplementary cementitious materials (fly ash, GBFS, silica fume, and metakaolin): Mechanical properties, pore structure, and water absorption. Construction and Building Materials. 2023;368:130455.
19. Zhao Z, Liu Y, Lu Y, Ji C, Lin C, Yao L, Pu Z, de Brito J. Prediction of properties of recycled aggregate

concrete using machine learning models: A critical review. Journal of Building Engineering. 2024;90:109516. doi:10.1016/j.jobe.2024.109516.

20. Mohammadi Golafshani E, Kim T, Behnood A, Ngo T, Kashani A. Sustainable mix design of recycled aggregate concrete using artificial intelligence. Journal of Cleaner Production. 2024;442:140994.
21. Yang S, Sun J, Zhifeng X. Prediction on compressive strength of recycled aggregate self-compacting concrete by machine learning method. Journal of Building Engineering. 2024;88:109055. doi:10.1016/j.jobe.2024.109055.
22. Golafshani E, Khodadadi N, Ngo T, Nanni A, Behnood A. Modelling the compressive strength of geopolymer recycled aggregate concrete using ensemble machine learning. Advances in Engineering Software. 2024;191:103611. doi:10.1016/j.advengsoft.2024.103611.
23. Parent concrete strength effects on the quality of recycled aggregate concrete: A review. Heliyon. 2024;10:e26212. doi:10.1016/j.heliyon.2024.e26212.
24. Zhao Z, Liu Y, Lu Y, et al. Prediction of properties of recycled aggregate concrete using machine learning models: A critical review. Journal of Building Engineering. 2024;90:109516. doi:10.1016/j.jobe.2024.109516.
25. Mandal A, Shiuly A. Exploring mechanical characteristics of recycled concrete aggregates from demolition waste: Advancements, challenges, and future directions for sustainable construction: A review. Discover Civil Engineering. 2025;2:33. doi:10.1007/s44290-025-00185-0.
26. Zhang B, Pan L, Chang X, et al. Sustainable mix design and carbon emission analysis of recycled aggregate concrete based on machine learning and big data methods. Journal of Cleaner Production. 2025;489:144734.
27. Peiris D, Gunasekara C, Law DW, et al. Impact of treatment methods on recycled concrete aggregate performance: A comprehensive review. Environmental Science and Pollution Research. 2025;32:14405–14438.
28. Antunes A, Costa H, Do Carmo R, Júlio E. A comprehensive review of sustainable use of construction and demolition waste as recycled aggregates in the production of concrete—Properties, mix design and on-site applications. Construction and Building Materials. 2025;482:141733. doi:10.1016/j.conbuildmat.2025.141733.