



Article

Reliable Compressive Strength Prediction of Self-Compacting Concrete with Recycled Coarse Aggregate Using an Interpretable Machine Learning Model (LightGBM) with Uncertainty Quantification

Bojan Milošević ^{1,*} , Nenad Kojić ^{2,*} and Milanko Kragović ²

¹ Faculty of Mechanical and Civil Engineering in Kraljevo, University of Kragujevac, Dositejeva 19, 36000 Kraljevo, Serbia

² Department/School of Information and Communication Technologies, Academy of Technical and Art Applied Studies Belgrade, 24 Starine Novaka Street, 11000 Belgrade, Serbia

* Correspondence: milosevic.b@mfkv.kg.ac.rs (B.M.); nenad.kojic@ict.edu.rs (N.K.)

Abstract

Machine learning is increasingly used to predict the compressive strength of self-compacting concrete with recycled coarse aggregate (SCRCAC), with coefficients of determination of 0.81–0.87 reported in the literature. This paper first shows that the widely used reference dataset of 603 mixtures contains only 504 unique compositions, with 84 groups of identical and 21 contradictory ones, so that identical mixtures leak between the training and test sets. Under an objective, leakage-free evaluation (with the model re-tuned on the deduplicated dataset), the coefficient of determination drops to about 0.73, a correction that applies to all models on this dataset. We then propose an interpretable, hyperparameter-optimized LightGBM model that (i) reaches the level of the best published results under the standard protocol (seed-averaged five-fold cross-validation (CV) $R^2 = 0.813$); (ii) provides a calibrated uncertainty interval for each prediction via split-conformal prediction, achieving an empirical coverage of 0.906 at the 90% nominal level; and (iii) remains fully explainable (SHAP (SHapley Additive exPlanations), partial dependence), with cement as the dominant predictor, followed by water and the mineral admixture. Under a 70/30 protocol averaged over 25 splits, it achieves an $R^2 = 0.794 \pm 0.038$ and a root mean squared error (RMSE) = 6.20 ± 0.48 MPa, exceeding all four machine learning models of the reference study. Aspects in which the reference study retains an advantage are also discussed.

Keywords: self-compacting concrete; recycled aggregate; compressive strength; LightGBM; data leakage; conformal prediction; uncertainty quantification; explainable machine learning; SHAP



Academic Editor: Haifeng Zhao

Received: 17 June 2026

Revised: 21 July 2026

Accepted: 22 July 2026

Published: 23 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

The drive toward sustainable construction encourages replacing natural aggregate with recycled aggregate obtained from construction waste. Concrete production is responsible for a significant share of global carbon dioxide emissions, while the depletion of natural aggregate deposits further burdens the environment [1,2]. Incorporating recycled coarse aggregate into self-compacting concrete simultaneously reduces the amount of disposed waste and the consumption of natural resources, making it an attractive solution within the circular economy [3].

Recycled aggregate, however, introduces pronounced variability into the properties of concrete. Residual old mortar on the grains increases water absorption and porosity and weakens the interfacial transition zone, which makes it harder to reach the target strength [4,5]. In self-compacting concrete, which is sensitive to fines content and superplasticizer dosage, this variability is further pronounced [6]. A reliable estimate of compressive strength from mixture composition, without extensive and costly laboratory testing, has great practical value for quality mix design and quality control.

Classical empirical formulas fail to capture the nonlinear interactions of constituents in such a heterogeneous material, so in recent years machine learning methods have gradually taken over the role of standard tools [1,7]. The most commonly applied models, artificial neural networks (ANNs), act as a “black box”. They yield accurate predictions but do not reveal why. This limits engineers’ trust and their broader use in practice [8,9].

Recent research therefore turns to tree-based models and gradient boosting, which stand out in the literature for their reliability, and complements them with methods of explainable artificial intelligence (XAI), especially SHAP analysis [8,9]. The reliability of the explanations themselves and their relation to prediction uncertainty have also become the subject of systematic study [10]. A detailed review of relevant works is given in Section 2.

The two central contributions of this work are a systematic data-leakage assessment of this widely used benchmark, which lowers the realistic accuracy from about 0.81 to about 0.73, and a calibrated conformal uncertainty interval for every prediction. Neither analysis has been reported in the published literature on this dataset. A complementary study by the same authors, conducted as a separate investigation, examines the same benchmark using a stacking ensemble under a robust multi-split protocol. The data-quality analysis is summarized in both studies so that each remains self-contained, while the modeling approach, the uncertainty quantification, and the conclusions are distinct. Starting from the reference study by Jagadesh et al. [4], which compared a neural network with several machine learning models (Random Forest, Extra Trees, XGBoost, LightGBM), the aim of this work is to propose a single, hyperparameter-optimized gradient-boosting model (LightGBM) and its systematic interpretation, shifting the emphasis from merely maximizing accuracy toward reliability. The contribution is manifold: (1) detecting and quantifying data leakage in a widely used dataset (repeated and contradictory compositions), it is found that the reported accuracy on this dataset is inflated (for all models in the literature, including the model developed in this work), which yields a leakage-free evaluation; (2) each prediction is equipped with a calibrated uncertainty interval through conformal prediction; (3) an interpretable LightGBM is developed that, under the standard protocol, reaches the level of the best published results, with full explainability (SHAP, partial dependence); and (4) a comparative analysis is carried out against the reference and the most recent studies on the same dataset under an identical protocol. In this way, instead of a single point estimate from one split, a transparent and uncertainty-equipped estimate is obtained.

2. Literature Review

Research on predicting the compressive strength of recycled-aggregate concrete using machine learning can be viewed along three thematic axes: the choice and architecture of the model, the drive toward explainability, and the gradual shift toward gradient-boosting methods. Most earlier works rely on artificial neural networks (ANNs), whose ability to map nonlinear dependencies has enabled them to outperform classical regression in modeling the compressive strength of concrete [1,2].

The first thematic axis concerns the network architecture and the training algorithm. De-Prado-Gil et al. [6] compared the Levenberg–Marquardt, Bayesian regularization, and scaled conjugate gradient algorithms, identifying Bayesian regularization as the most

resistant to overfitting for recycled material. The same algorithm-comparison approach was applied by Nikmehr et al. [11] to a geopolymer self-compacting concrete (SCC) with coated aggregate, achieving very high accuracy on a strictly controlled database. Boudali et al. [5] modeled the influence of fine recycled powder as a micro-filler, while de-Prado-Gil et al. [12] introduced a partitioning of the database by target strength and derived explicit engineering equations. The reference study [4] used in this work belongs to this axis, with an extension to a comparison with machine learning models.

The second axis is explainability (XAI). Recent works abandon the “black box” and introduce SHAP analysis to quantify the contribution of each constituent. For example, Yang, Chen, and Xu [8] and Almutairi [9] show that SHAP makes the model physically rational and singles out cement, water, and superplasticizer as the dominant factors. The third axis is the shift toward trees and gradient boosting, where several studies find that tree ensembles and boosting outperform ANNs in consistency and interpretability. For instance, Yang, Sun, and Xu [13] for Random Forest, Almutairi [9] for XGBoost, and Alyaseen et al. [7] report weaker generalization of ANNs compared with alternative models. Comprehensive reviews [3] confirm these trends. The most recent studies extend them to interpretable gradient-boosting models with SHAP analysis, applied to predicting the compressive strength of conventional concrete [14]. Explainable models with metaheuristic optimization have been applied to the properties of self-compacting concrete [15], while hybrid generative–ensemble approaches have been used for recycled-aggregate concrete [16].

The research presented in this work is most similar to the recently published paper by Abood et al. [17], who, on the same dataset from Jagadesh et al. [4], apply a single gradient-boosting model optimized by Bayesian optimization (BO-GBRT) together with SHAP analysis, reporting a five-fold cross-validation $R^2 = 0.806$ and a test $R^2 = 0.868$ on a single 90/10 split. Their approach, however, remains a single-model point prediction and considers neither prediction uncertainty nor the repetition of identical compositions in the dataset. The research in this work starts from the same family of models but goes further toward systematic explainability, conformal uncertainty intervals, and a leakage-aware evaluation, as detailed in Section 5.

Uncertainty quantification, which is central to this work, has only recently been introduced into this field in the research by Dong and Zhang [18], who developed a hybrid deep learning framework (an FT-Transformer combined with a masked multilayer perceptron and two-level stacking) for predicting the compressive strength of self-compacting concrete with recycled aggregate. The framework was equipped with split-conformal intervals and SHAP interpretation, and was utilized on a dataset of 289 mixtures described by 18 variables (including curing age and the water-to-binder ratio), with particular robustness to missing data. The study by Dong and Zhang [18] confirms the timeliness of conformal prediction in this field and is complementary to this work. Here, the conformal framework is applied to the broader reference benchmark dataset [4], with a considerably poorer mixture description (eight variables), using a single fully interpretable gradient-boosting model, with the emphasis shifted to data quality (repeated and contradictory compositions) and robust evaluation across multiple splits.

Despite the clear trend toward boosting and explainability, a single, hyperparameter-optimized leaf-wise gradient-boosting model (LightGBM), accompanied by a systematic explainability analysis (SHAP and partial dependence) and a robust evaluation across multiple splits, has not been sufficiently investigated for self-compacting concrete with recycled coarse aggregate. Existing works mostly use Random Forest or XGBoost. More broadly, even the most recent works on this dataset remain single-point predictions, without uncertainty quantification or a data-quality check. This paper fills these gaps and proposes an interpretable LightGBM model, complemented by conformal uncertainty intervals and

a leakage-aware evaluation, thereby achieving a favorable balance of accuracy, simplicity, explainability, and reliability.

3. Materials and Data

In this research, the publicly available dataset presented in the reference study [4] was used. The dataset comprises 603 mixtures of self-compacting concrete with recycled coarse aggregate, which the authors of the reference study collected from a large number of published experimental studies (61 sources over 14 years). Each mixture is described by eight composition components and the compressive strength measured at the age of 28 days (f_{ck}). The variables and their units are given in Table 1. All specimens in this benchmark are tested at the same age of 28 days, so the curing age is constant across the dataset and cannot act as an input variable. Moreover, the compressive strength determined at the age of 28 days is the standard reference metric defined by structural design codes worldwide for concrete classification and quality control, which makes it the primary focus of this benchmark. Datasets with a variable curing age, such as the dataset of Dong and Zhang [18] with eighteen variables, permit that extension, which is noted as a direction for future work (Section 6.3). Since the data come from different laboratories and recipes, the dataset is diverse but also noisy, which poses a realistic challenge for a predictive model.

Table 1. Description of variables.

Symbol	Description	Unit
Cement	Cement content	kg/m ³
Water	Water content	kg/m ³
FA	Natural fine aggregate	kg/m ³
RFA	Recycled fine aggregate	kg/m ³
CA	Natural coarse aggregate	kg/m ³
RCA	Recycled coarse aggregate	kg/m ³
MA	Mineral admixture	kg/m ³
CHA	Chemical admixture (superplasticizer)	kg/m ³
f_{ck}	28-day compressive strength (output)	MPa

Table 2 presents the descriptive statistics of the input and output variables. The range of the compressive strength (7.2–81.9 MPa) and of cement content (78–635 kg/m³) confirms the diversity of the mixtures, while the very large standard deviations of the recycled fractions (RFA, RCA) reflect different degrees of replacement of natural aggregate by recycled aggregate. Zeros for the aggregate components indicate their absence in a given mixture, and the largest deviation is shown by the chemical admixture, with a maximum of 320 kg/m³, far above the mean (about 6 kg/m³). These extreme admixture values were retained in the modeling. Only a single mixture lies above the 99th percentile, and gradient-boosted trees split on thresholds rather than distances, which makes them robust to such outliers. Discarding physically real mixtures without evidence of error would introduce selection bias. A sensitivity check confirms this choice: removing the extreme-admixture row or winsorizing it at the 99th percentile changes the seed-averaged five-fold R^2 by at most 0.004 (from 0.813 to 0.810 and 0.813, respectively).

Table 2. Descriptive statistics of the input and output variables.

Variable	Mean	Std.	Min.	Max.	Unit
Cement	362.22	107.94	78.0	635.0	kg/m ³
Water	183.72	26.98	105.0	277.0	kg/m ³
FA	684.09	295.08	0.0	1200.0	kg/m ³

Table 2. *Cont.*

Variable	Mean	Std.	Min.	Max.	Unit
RFA	157.91	268.91	0.0	1200.0	kg/m ³
CA	426.75	332.45	0.0	1170.0	kg/m ³
RCA	334.46	323.59	0.0	1115.0	kg/m ³
MA	153.74	109.58	0.0	449.0	kg/m ³
CHA	6.05	14.62	0.0	320.0	kg/m ³
f_{ck}	43.62	13.77	7.17	81.9	MPa

A uniqueness check of the compositions reveals that behind the 603 rows lie 504 distinct compositions. In total, 84 groups of mixtures share an identical composition, and in 21 of them the same composition corresponds to different measured strengths. Two mixtures were treated as identical only when all eight input variables (Table 1) coincided exactly, at the numerical precision reported in the source dataset, with no numerical tolerance applied. Groups of identical compositions associated with different measured strengths were labeled contradictory. The data thus contain identical and mutually inconsistent records, which set an upper bound on the attainable accuracy (irreducible noise) and, if not removed before splitting, can lead to information leakage between the training and test sets (data leakage) and hence to an overestimation of model performance. To objectively prevent this effect, performance is reported in this work in two ways. First, on the original dataset of 603 rows, for direct comparability with the reference study [4]. Second, on a deduplicated dataset, in which identical compositions are merged and the 21 contradictory groups (same composition, different measured strength) are assigned the mean value, yielding 504 unique compositions on which the model is re-tuned and evaluated (Section 5). In this way, the reported result does not rely on a potentially inflated value. Instead, an objective, leakage-free estimate is provided alongside it. Such an analysis is in the spirit of recent findings that careful scrutiny of widely used benchmark datasets can reveal their hidden limitations and redundancy, and thereby refine the realistically attainable performance [19]. From a physical standpoint, the contradictory records are expected: the eight compositional variables do not capture every factor that governs the measured strength, including the properties and origin of the recycled aggregate (residual mortar content, absorption), the curing regime, and the specimen and testing procedure. Nominally identical mixtures, in several cases drawn from different source studies, therefore genuinely yield different strengths, reflecting laboratory-to-laboratory variation and the usual experimental scatter of concrete. Occasional transcription errors in the original compilation cannot be excluded either. The contradictions are thus mainly a consequence of an incomplete mixture description rather than of necessarily erroneous data, and they form part of the irreducible noise.

Compressive strength ranges from 7.2 to 81.9 MPa (mean 43.6 MPa). The distributions of the input variables, shown in standardized form in Figure 1, exhibit pronounced asymmetry, the presence of zeros for recycled aggregates (mixtures without that component), and a few extreme values, particularly for the chemical admixture.

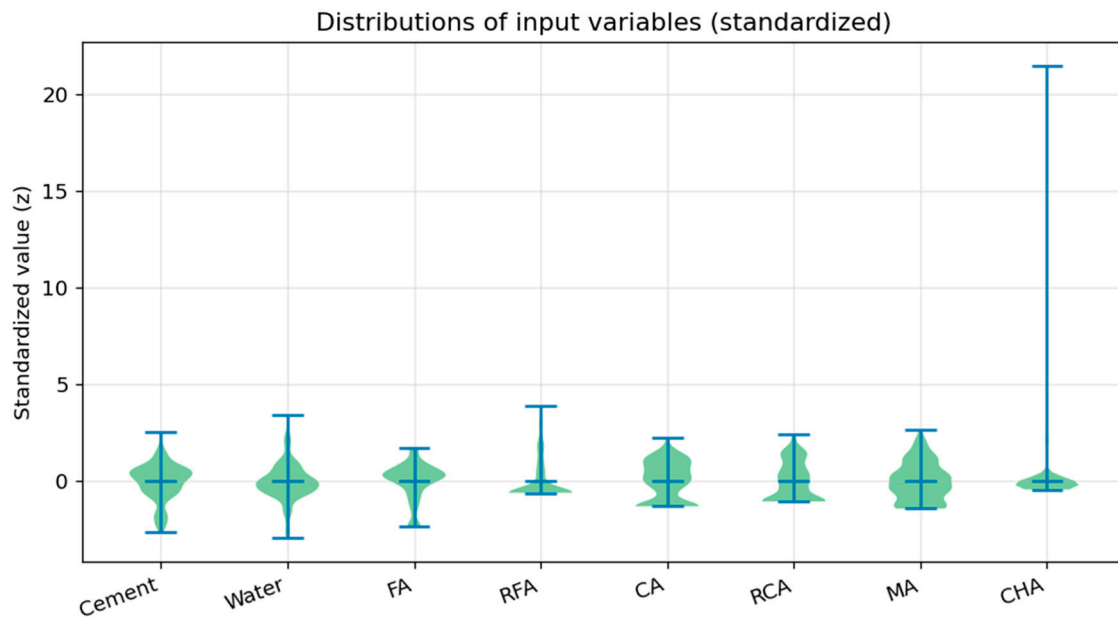


Figure 1. Distributions of the standardized input variables (violin plot); asymmetry and extreme values of some admixtures are visible (all 603 mixtures).

4. Methodology

4.1. Underlying Techniques and Theoretical Background

The proposed model belongs to the field of artificial intelligence (AI) and its subfield of machine learning (ML). Within supervised learning, where a known output is associated with each input example, predicting continuous compressive strength is a regression problem. For tabular data with nonlinear interactions, ensembles based on decision trees, and gradient boosting in particular, are today the standard, often outperforming neural networks on datasets of moderate size. Current research compares boosting with modern deep architectures on real tabular problems [20] and also combines it with neural components into hybrid models [21]. This work uses one such model, LightGBM [22].

The basic element is the regression decision tree [23], which, through recursive binary splits, partitions the input space into disjoint regions, predicting a constant in each leaf. The variable and the split threshold are chosen by minimizing the sum of squared deviations in the resulting subsets (L , R):

$$\text{SSE}(s) = \sum_{i \in L} (y_i - \bar{y}_L)^2 + \sum_{i \in R} (y_i - \bar{y}_R)^2, \quad (1)$$

and the leaf prediction equals the average of the target values within it:

$$\hat{y}_{R_m} = \frac{1}{|R_m|} \sum_{i \in R_m} y_i, \quad (2)$$

where L and R denote the sets of samples that fall, respectively, to the left and right of the split threshold on the chosen variable, and $\bar{y}_L$ and $\bar{y}_R$ are the mean measured strengths within those two subsets. A single tree has high variance, which motivates combining them into ensembles.

Gradient boosting [24] builds an additive model sequentially, with each successive tree correcting the errors of the previous ones:

$$F_M(x) = F_0(x) + \sum_{m=1}^M \nu \gamma_m h_m(x), \quad (3)$$

where F_0 is the initial constant, h_m the m -th tree, γ_m its coefficient, and ν the learning rate. The model is initialized with the constant that minimizes the loss function L :

$$F_0(x) = \arg \min_{\gamma} \sum_{i=1}^n L(y_i, \gamma). \quad (4)$$

At each iteration, the pseudo-residuals are computed as the negative gradient of the loss with respect to the current prediction:

$$r_{im} = - \left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)} \right]_{F=F_{m-1}}. \quad (5)$$

The new tree is fit to these residuals, and the model is updated with scaling by the learning rate ν (a smaller ν gives slower but more stable learning):

$$F_m(x) = F_{m-1}(x) + \nu \gamma_m h_m(x), 0 < \nu \leq 1. \quad (6)$$

Modern implementations (XGBoost [25], LightGBM [22], CatBoost [26]) introduce a second-order Taylor approximation of the objective. For sample i with gradient g_i and Hessian h_i of the loss, the contribution of the m -th tree is approximated as

$$L^{(t)} \simeq \sum_{i=1}^n \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t), \quad (7)$$

from which, for a fixed structure, the optimal leaf weight follows

$$w_j^* = - \frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda}, \quad (8)$$

where λ denotes the L2 regularization and $\Omega(f_t)$ the tree-complexity penalty. The penalty combines an L2 term on the leaf weights (coefficient λ), an L1 term on the leaf weights (coefficient α), and a term proportional to the number of leaves T (coefficient γ). In LightGBM, these map directly to the tuned hyperparameters: λ to `reg_lambda`, α to `reg_alpha`, and T to `num_leaves` (additionally constrained by `min_child_samples`). Their values are listed in Table 3.

Table 3. Tuned hyperparameters of the proposed LightGBM model.

Hyperparameter	Value	Role
n_estimators	770	Number of boosting trees
num_leaves	45	Max. leaves per tree (leaf-wise capacity)
max_depth	10	Max. tree depth
learning_rate	0.037	Shrinkage ν
subsample	0.54	Row subsampling
colsample_bytree	0.76	Column subsampling
min_child_samples	7	Min. samples per leaf
reg_alpha	0.045	L1 regularization
reg_lambda	0.049	L2 regularization

LightGBM [22] introduces several key improvements over classical boosting. First, instead of considering all values of a variable, it uses histogram-based binning. Continuous values are grouped into discrete bins, so that the search for the best split reduces to a pass over the histogram, with the information gain of a split on variable j and threshold d :

$$V_j(d) = \frac{1}{n} \left(\frac{\left(\sum_{x_i \in O_l(j,d)} g_i \right)^2}{n_l(j,d)} + \frac{\left(\sum_{x_i \in O_r(j,d)} g_i \right)^2}{n_r(j,d)} \right), \quad (9)$$

where O_l and O_r are the sets of instances to the left and right of the threshold. For the squared-error loss used here, the Hessian is constant, so this split-gain is presented in the simplified, variance-based form. In the general case, LightGBM uses the full second-order objective of Equations (7) and (8), which retains both the gradient and the Hessian statistics of each leaf. Second, unlike level-wise growth, LightGBM uses leaf-wise growth. At each step it splits the leaf that yields the largest gain:

$$(j^*, d^*) = \arg \max_{j,d} V_j(d), \quad (10)$$

which yields deeper, more accurate trees for the same number of leaves (but requires regularization against overfitting). Third, gradient-based one-side sampling (GOSS) keeps all instances with a large gradient (fraction a) and randomly samples a portion (fraction b) of those with a small gradient, compensating their weight by a factor $(1 - a)/b$ so that the gain estimate remains unbiased:

$$\tilde{V}_j(d) = \frac{1}{n} \left(\frac{\left(\sum_{x_i \in A_l} g_i + \frac{1-a}{b} \sum_{x_i \in B_l} g_i \right)^2}{n_l^j(d)} + \frac{\left(\sum_{x_i \in A_r} g_i + \frac{1-a}{b} \sum_{x_i \in B_r} g_i \right)^2}{n_r^j(d)} \right). \quad (11)$$

Finally, exclusive feature bundling (EFB) bundles sparse, mutually exclusive features into one, reducing dimensionality. The conflict measure of two features is the number of instances in which both are simultaneously nonzero:

$$\text{Conflict}(x_j, x_k) = \sum_{i=1}^n 1[x_{ij} \neq 0 \wedge x_{ik} \neq 0]. \quad (12)$$

The main tunable parameters are the number of leaves (`num_leaves`), maximum depth, learning rate, number of trees, row and column subsampling, the minimum number of samples per leaf (`min_child_samples`), and L1/L2 regularization. The related boosters XGBoost [25] and CatBoost [26] share the same basis but differ in their growth and regularization strategies. Here they are used only as a reference framework.

4.2. Application to the Present Problem

The choice of LightGBM as the single model was guided by the nature of the problem and the aims of the work. The dataset is tabular, of moderate size (603 examples, eight variables), strongly nonlinear, and noisy. For such data, gradient boosting as a rule delivers top accuracy. Unlike a stacking ensemble, a single model was deliberately chosen because the goal is not maximal accuracy at any cost but a favorable balance of accuracy, training speed, simplicity, and explainability. LightGBM is particularly suitable because histogram-based binning and GOSS enable fast repeated validation, while native support for importance measures and compatibility with SHAP analysis make the model interpretable. The eight raw constituents were used directly as inputs. Supplying engineered ratios such as the water-to-cement ratio explicitly was also tested, but it did not improve accuracy. A variant enriched with domain ratios (including the water-to-cement and water-to-binder ratios) gave a seed-averaged five-fold R^2 of 0.802, against 0.813 for the raw inputs. This is expected, because gradient-boosted trees can already reconstruct such ratios through successive splits. The SHAP analysis in Section 5.3 further confirms the primacy of cement, with water among the most influential variables, consistent with the water-to-cement ratio and Abrams' law.

Regarding this type of regression problem, individual boosting typically achieves an R^2 of about 0.74–0.79. With careful tuning, performance in the upper part of that range is expected, at the level of or above the best reproducible model of the reference study. Since a single estimator is used, there are neither two levels nor a serial dependence among the outputs of multiple models. The data flow is simple. The input X of dimension $n \times 8$ (eight numerical variables in kg/m^3) maps directly to a scalar output (the predicted f_{ck} in MPa).

The hyperparameters were optimized with the Optuna library [27] (Tree-structured Parzen Estimator (TPE) search, five-fold cross-validation exclusively on the training set, without access to the test set). The parameters that most affect the bias–variance trade-off were tuned, while the boosting type (gbdt) and the loss function (L2 regression) were kept at their default values, since alternative choices did not bring consistent gains. The specific configuration is given in Table 3. The optimization used 150 TPE trials with a fixed random seed for reproducibility, and the objective was the mean five-fold cross-validation R^2 computed on the training part of the reference split, so the test set was never accessed during tuning. The search ranges were as follows: number of trees 100–1200; number of leaves 7–127; maximum depth 2–12; learning rate 0.005–0.3 on a logarithmic scale; row and column subsampling 0.5–1.0; minimum number of samples per leaf 3–40; and L1 and L2 regularization 0.001–10 on a logarithmic scale. The hyperparameters were optimized once and then reused across all evaluation protocols. The optimization was not repeated within each cross-validation fold, which is legitimate because the objective already uses an internal five-fold split of the training data and never sees the held-out test set. For the deduplicated evaluation, the model was re-tuned on the deduplicated training data (Section 5).

Reasons for the chosen settings were as follows: num_leaves and learning_rate most strongly control the model’s capacity and tendency to overfit, so they were carefully tuned together with the number of trees. Row and column subsampling and the L1/L2 penalties additionally regularize the model on a relatively small dataset. Values of num_leaves well below $2^{\text{max_depth}}$ deliberately limit leaf-wise growth and prevent overfitting.

4.3. Evaluation Protocol

For comparison with the reference study [4], the 70% training/30% test split was retained. To avoid the bias of a single split, performance is reported as the mean $\pm$ standard deviation over 25 independent random splits, with the diagnostics shown on a representative (median R^2) split. This protocol is in line with recommendations for the reliable estimation of the generalization error of regression models [28]. The metrics used are the coefficient of determination, the root mean squared error, and the mean absolute error:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, \quad (13)$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}, \quad (14)$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|. \quad (15)$$

The implementation is in Python 3.11.9 (Python Software Foundation, Wilmington, DE, USA) using the LightGBM 4.6.0 (Microsoft Corporation, Redmond, WA, USA) [22], scikit-learn 1.8.0 [29], SHAP 0.51.0 [30], and Optuna 4.8.0 (Preferred Networks, Inc., Tokyo, Japan) [27] libraries.

4.4. Uncertainty Quantification (Conformal Prediction)

In addition to a point estimate, a measure of prediction uncertainty is important for engineering use, which the reference and related works on this dataset do not provide

(on a related dataset it was recently introduced by a hybrid deep model [18], discussed in Section 6). Uncertainty quantification is also becoming established more broadly in engineering applications of artificial intelligence; for example, through Bayesian neural networks for predicting the response of offshore platforms [31]. Therefore, split-conformal prediction [32] is applied, which, under mild assumptions, yields intervals with guaranteed coverage. The training set is divided into a proper-training part and a calibration part. On the calibration part, the absolute residuals are computed, and their $(1 - \alpha)$ quantile defines the half-width of the interval:

$$\hat{q}_{1-\alpha} = \text{Quantile}_{1-\alpha} \left(\left\{ |y_i - \hat{f}(x_i)| : i \in I_{\text{cal}} \right\} \right), \quad (16)$$

so the prediction interval for a new sample is $[\hat{y} - \hat{q}, \hat{y} + \hat{q}]$. In this way, together with the strength estimate, the engineer also obtains a confidence range, which is crucial for safety-side decisions.

In addition to the basic (symmetric, split) variant, an adaptive variant, conformalized quantile regression (CQR) [33], is also considered, in which, instead of a single half-width, two quantile-regression models are used (for the lower and upper quantiles), so the interval adapts to the local uncertainty and its width varies from sample to sample. Both variants retain guaranteed coverage, and their comparison is presented in Section 5.

5. Results

5.1. Predictive Performance

Under the repeated 70/30 protocol, the proposed LightGBM achieves an $R^2 = 0.794 \pm 0.038$, an RMSE = 6.20 ± 0.48 MPa, a mean squared error (MSE) = 38.7 ± 6.0 MPa², and a mean absolute error (MAE) = 4.00 ± 0.32 MPa. On a representative split, the model achieves an $R^2 = 0.798$ and a correlation coefficient $r = 0.895$ (Figure 2). Table 4 presents the values of the proposed model ($R^2 = 0.794$, RMSE = 6.20 MPa) and of the four machine learning models from the reference study ($R^2 = 0.7249$ – 0.7635 , RMSE = 6.80–7.34 MPa). The comparisons are shown graphically in Figure 3 (R^2) and Figure 4 (the RMSE and MSE errors).

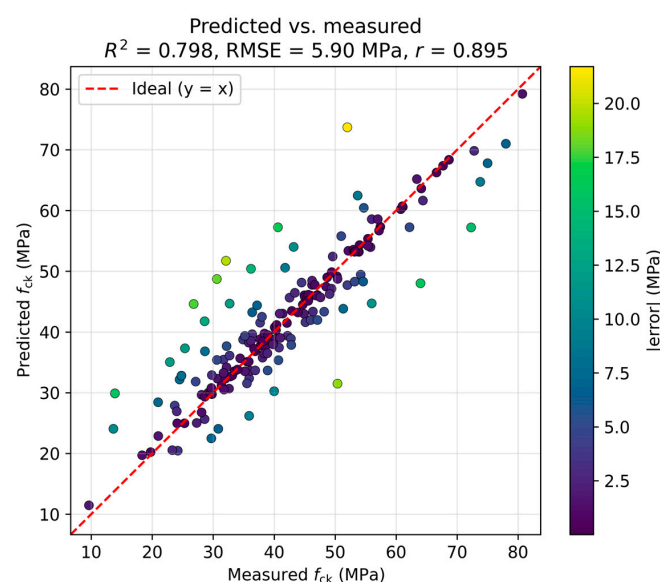
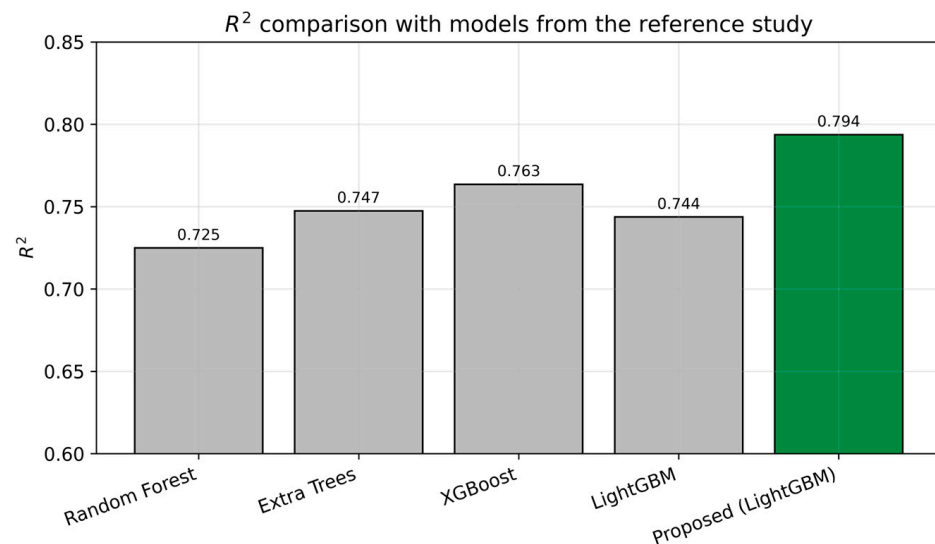
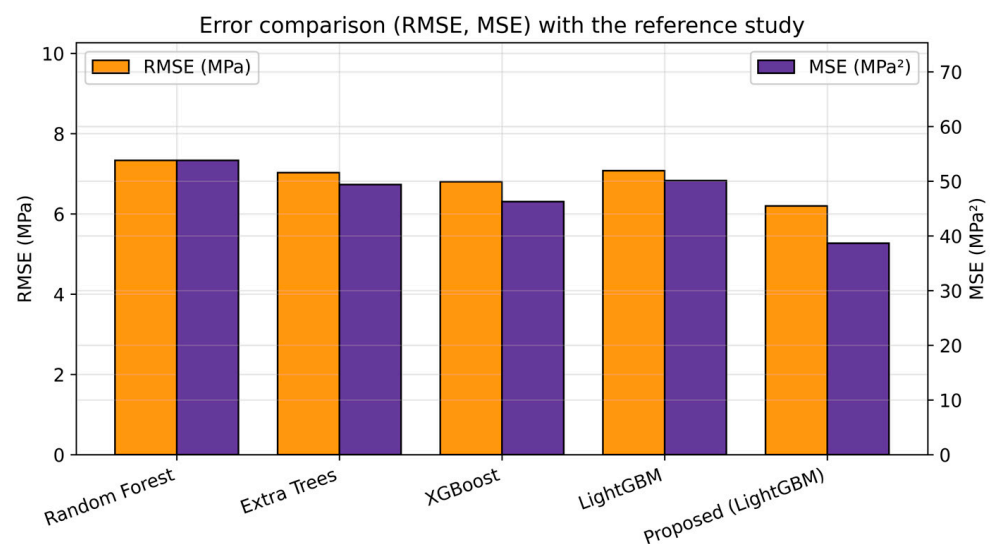


Figure 2. Predicted vs. measured compressive strength (representative median R^2 split of the $25 \times 70/30$ protocol); color encodes absolute prediction error.

Table 4. Comparison with the reference study (proposed model: mean over $25 \times 70/30$). Bold indicates the proposed model.

Model	R^2	RMSE (MPa)	MSE (MPa ²)
Random Forest (reference)	0.7249	7.34	53.80
Extra Trees (reference)	0.7474	7.03	49.40
LightGBM (reference, default)	0.7437	7.08	50.13
XGBoost (reference, best)	0.7635	6.80	46.26
Proposed LightGBM	0.794	6.20	38.7

**Figure 3.** Comparison of the coefficient of determination (R^2) with models from the reference study (proposed model: mean over the $25 \times 70/30$ splits; reference values as reported in [4]).**Figure 4.** Comparison of error (RMSE and MSE) with the reference study (proposed model: mean over the $25 \times 70/30$ splits; reference values as reported in [4]).

5.2. Residual Analysis and Stability

The residuals on the representative split are approximately normally distributed, as shown by the Q–Q plot (Figure 5), without pronounced systematic bias. The learning curve (Figure 6) shows that the validation R^2 stabilizes as the number of samples increases, with a moderate gap between the training and validation curves, indicating controlled overfitting. Across the 25 splits, the error is not uniform over the strength range. The mean

absolute error is lowest in the middle tercile (about 3.3 MPa) and higher in the low and high terciles (about 4.2 and 4.5 MPa), consistent with the slightly lower conformal coverage at the extremes (Section 5.6). By mixture type, recycled-aggregate mixtures ($RCA > 0$) are not predicted less accurately than mixtures with natural coarse aggregate only (mean absolute error 3.9 against 4.1 MPa).

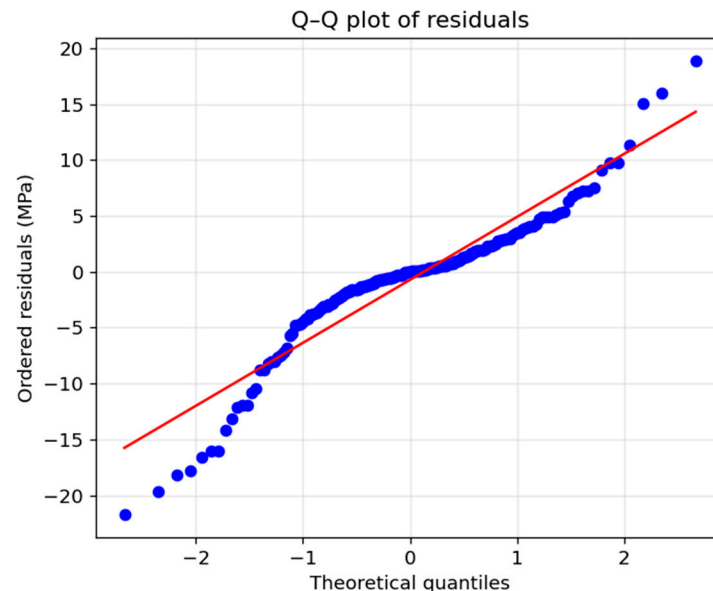


Figure 5. Q–Q plot of residuals against the normal distribution (representative split); the red line denotes the reference normal distribution.

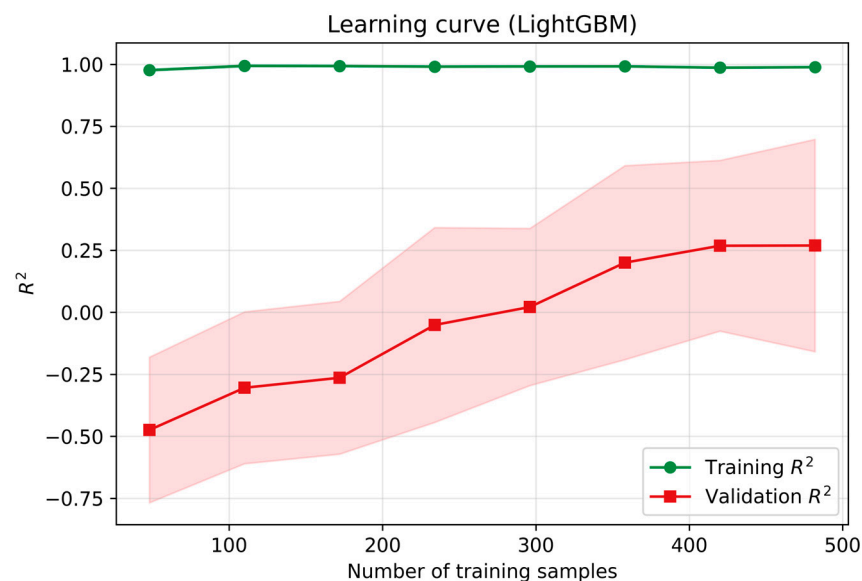


Figure 6. Learning curve of the proposed model (five-fold cross-validation on the full dataset).

5.3. Model Explainability

Unlike a “black box,” the proposed model is fully interpreted. The native importance measure (total gain, Figure 7) and the SHAP analysis [30] (Figure 8) jointly single out cement as the dominant predictor, followed by water and the mineral and chemical admixtures. These rankings are stable across the 25 random splits. Cement ranks first in every split, and the overall ordering is highly consistent (mean pairwise Spearman correlation 0.93 for gain and 0.94 for SHAP), with water and the mineral admixture alternating in the next positions. Cement is unambiguously the most influential variable under both measures.

For the second position, the water content leads by the gain measure while the mineral admixture leads by the SHAP measure, and the two are comparable. This refines the initial description of cement and water alone as jointly dominant. Beyond the clear primacy of cement, water and the mineral admixture contribute comparably. This is consistent with established knowledge about concrete (the water-to-cement ratio, Abrams' law), giving the model a physically meaningful interpretation. Partial dependence (Figure 9) shows the expected monotonic trends: compressive strength increases with cement content and decreases with water content. Because partial dependence is an average over the empirical distribution, in sparsely populated regions (for example, the extreme admixture range) the curves represent an extrapolation and should be read with caution. The rug marks added to Figure 9 indicate the data density along the horizontal axis of each panel, so the trends are reliable mainly within the well-populated range (see also the distributions in Figure 1).

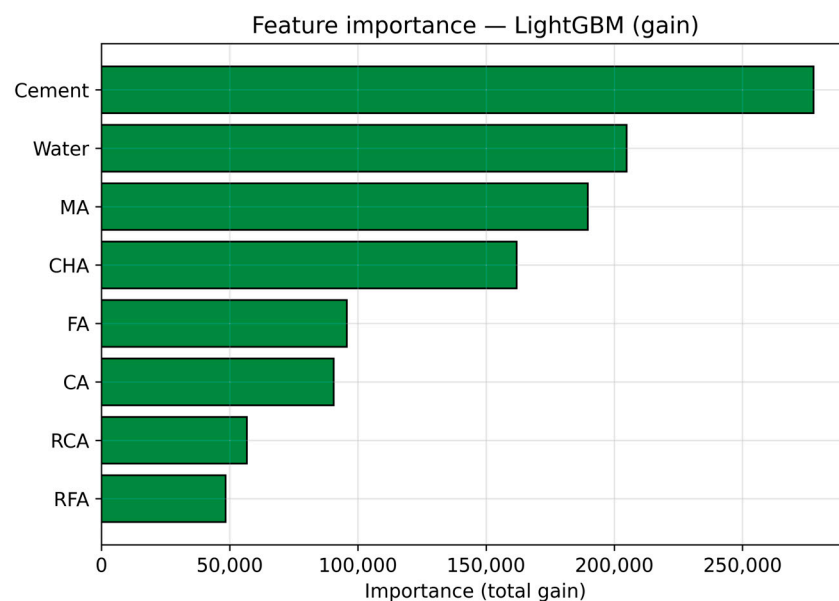


Figure 7. Feature importance of the proposed model (LightGBM total gain; representative split, original 603-row dataset).

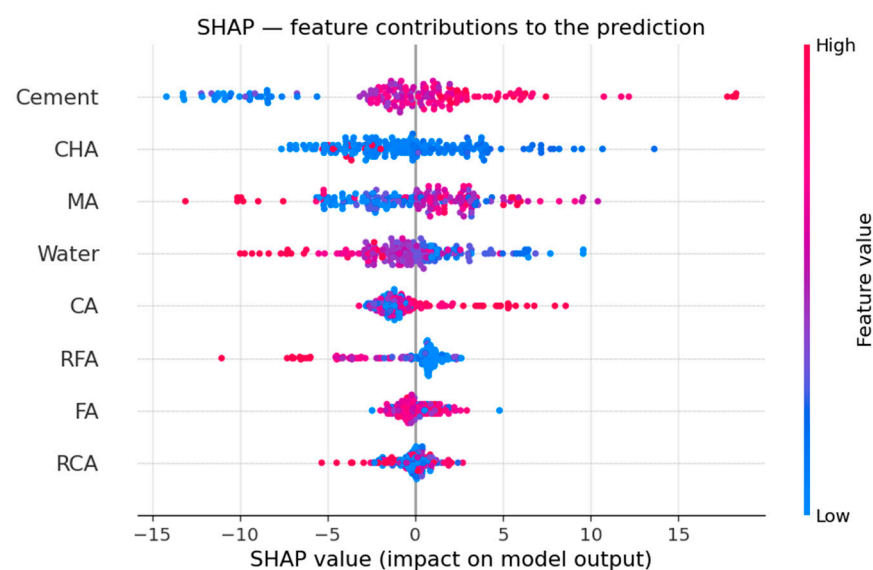


Figure 8. SHAP summary: contribution of each variable to individual predictions (representative split, original 603-row dataset); colour represents the feature value (red = high, blue = low).

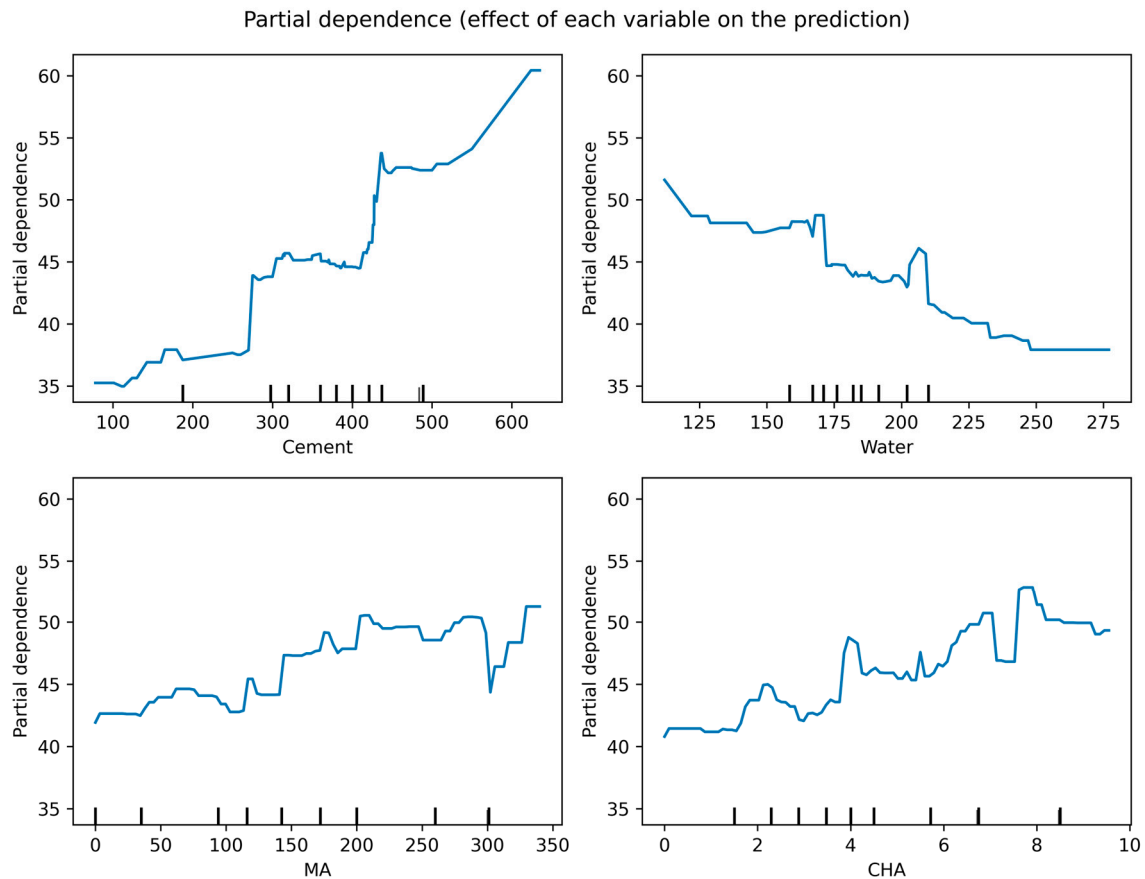


Figure 9. Partial dependence for the four most influential variables (representative split; rug marks show data density).

These findings agree with the results of explainable ML studies from the literature. SHAP analyses [8,9] also single out cement, water, and superplasticizer as the dominant factors, which confirm the physical consistency of the proposed model. As a quantitative measure of engineering reliability, the a20-index was computed: the proportion of predictions whose ratio to the measured value deviates by less than $\pm 20\%$, which for the proposed model on a representative split is 0.85, is within the range of values regarded in the literature as an indicator of a model's readiness for practical use [8].

5.4. Comparison with the Most Recent Work (Abood et al.) and Model Extensions

In the most recent study, Abood et al. [17], conducted on the same dataset presented in the work of Jagadesh et al. [4], present the application of a single gradient-boosting model optimized by Bayesian optimization (BO-GBRT), reporting a five-fold cross-validation $R^2 = 0.806$ (RMSE 6.00 MPa) and, on a single 90/10 split, a test $R^2 = 0.868$. For an objective comparison, the proposed LightGBM was evaluated under the identical five-fold protocol, but averaged over 20 random choices of the fold partition (rather than on the basis of a single partition). The seed-averaged five-fold CV gives an $R^2 = 0.813 \pm 0.017$, with the result exceeding 0.806 in 14/20 partition choices and falling below it in the remainder. Because the reference value comes from a single partition, a two-sample test is not possible. A one-sample test of the 20 seed-averaged values against the constant 0.806 gives $p = 0.07$ (both a t -test and a Wilcoxon signed-rank test), so the small difference is not statistically significant. The 14/20 count is therefore a descriptive indicator rather than a formal test, and the two results are best interpreted as being at the same level on this metric.

The essential contribution relative to [17], however, is not a higher coefficient of determination (which is at parity) but two extensions that the cited approach lacks. The

first extension, and the most important, concerns data quality. When the model is re-tuned and evaluated on the deduplicated dataset (504 unique compositions), the R^2 is 0.73 ± 0.01 , well below the values of about 0.81–0.87 reported on the original dataset (and in all previous works on this dataset). This indicates that those reported values partly originate from the leakage of identical compositions between the training and test sets, and that ~ 0.73 is a more objective estimate of the true performance on this problem (the full sensitivity analysis of the duplicate treatment is given in Section 5.5 and is not repeated here). The second extension is uncertainty quantification. The split-conformal intervals achieve an empirical coverage of 0.906 (nominal 0.90) with an average width of 25.8 MPa. This relatively large width is not a shortcoming of the model but a faithful reflection of the irreducible noise in the data (partly the contradictory records themselves). The model conveys the limit of its reliability rather than concealing it. A concise comparison with the most recent work and the proposed extensions is given in Table 5.

Table 5. Comparison with Abood et al. [17] and the proposed extensions on the same dataset.

Property/Protocol	Abood et al. (BO-GBRT)	Proposed (This Work)
Five-fold CV R^2 (seed-averaged)	0.806 (single)	0.813 ± 0.017
Test R^2 (90/10, reference split)	0.868	0.851
Robust R^2 (70/30, 25 splits)		0.794 ± 0.038
Leakage-free R^2 (re-tuned, unique comp.)	Not assessed	0.73
Uncertainty intervals (conformal)	No	Yes (coverage 0.906)

5.5. Sensitivity of the Leakage Finding to Duplicate Treatment

To verify that the leakage finding does not depend on an arbitrary choice of duplicate handling, the same model (parameters from Table 3) was evaluated by five-fold cross-validation, seed-averaged over 20 partition choices, on three dataset variants: (a) the original dataset of 603 rows; (b) the dataset with contradictory groups removed (rows with the same composition but different measured strength; 549 rows remaining); and (c) a fully deduplicated dataset in which identical compositions are merged into 504 unique ones (the contradictory ones assigned the mean). The results are given in Table 6.

Table 6. Sensitivity of the reported R^2 to the treatment of duplicate compositions.

Treatment of Duplicate Compositions	n	R^2 (Five-Fold, Seed-Averaged)
(a) Original set (as distributed)	603	0.813 ± 0.017
(b) Contradictory groups removed	549	0.811 ± 0.018
(c) Deduplicated to unique compositions	504	0.731 ± 0.012
(d) Group-aware CV (GroupKFold by composition)	603/504 grp	0.733 ± 0.066

Removing only the contradictory groups changes the result negligibly (0.811 versus 0.813), because identical compositions that can simultaneously enter the training and test sets still remain in the dataset. Only full deduplication into unique compositions removes this leakage and reveals a more realistic 0.731. This confirms that the increase in reported accuracy originates precisely from the repeated compositions (and not only from the contradictory measurements) and that ~ 0.73 is the real level of performance on this problem. This conclusion does not depend on the choice of model and applies to all studies on the same dataset.

The strictest and methodologically most correct check is group-aware cross-validation (GroupKFold), in which all mixtures of the same composition remain in the same fold, so identical compositions are never simultaneously in the training and test sets (row

(d) in Table 6). It gives an $R^2 = 0.733 \pm 0.066$, which matches the full deduplication and independently confirms that the apparent ~ 0.81 arises precisely from the leakage of identical compositions between the training and test sets. The larger standard deviation of this variant (0.066 against 0.012–0.018 for the others) does not indicate an unstable estimate. It is the dispersion across the five folds of a single deterministic group partition, whereas the other standard deviations are computed across twenty seed-averaged repetitions, so the two figures measure different quantities and are not directly comparable. The grouping constraint also produces folds of unequal size and heterogeneous source composition and precludes stratification by strength, which further widens the fold-to-fold spread. Crucially, the mean value (0.733) coincides with the independent deduplication estimate (0.731), so the conclusion that the true performance is about 0.73 rests on the agreement of two separate checks rather than on the variance of any single one.

5.6. Calibration and Adaptive Uncertainty Intervals

The reliability of the conformal intervals was verified by calibration. For nominal confidence levels of 80%, 90%, and 95%, the empirical coverage (the proportion of true values within the interval) and the average interval width were measured, averaged over 20 random splits. The results are presented in Table 7, and a graphical representation in Figure 10.

Table 7. Calibration of split-conformal prediction intervals.

Nominal Coverage	Empirical Coverage	Mean Interval Width (MPa)
80%	0.792	17.1
90%	0.906	25.8
95%	0.945	33.6

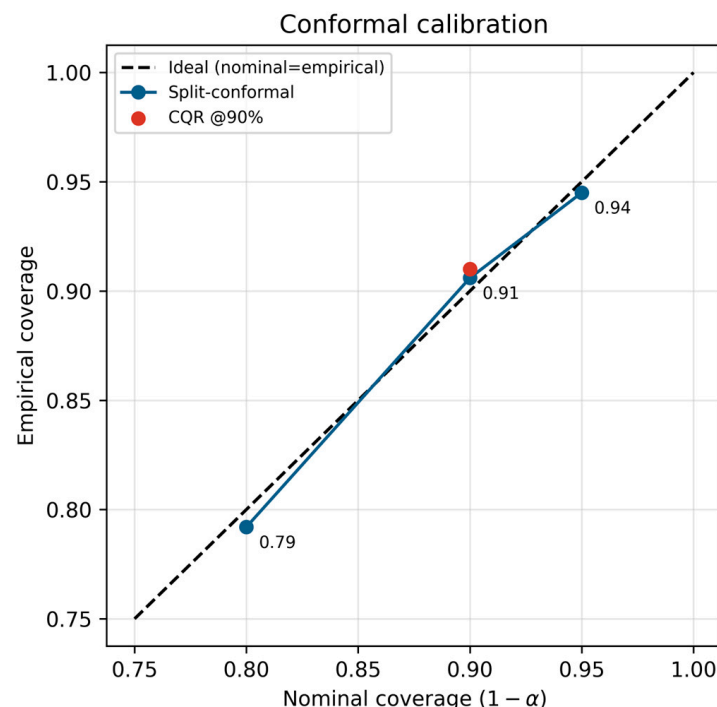


Figure 10. Conformal calibration: empirical vs. nominal coverage (split-conformal at 80/90/95% and CQR at 90%), averaged over 20 random splits; points near the diagonal indicate correct calibration.

The empirical coverage at all three levels almost matches the nominal one (at 90% it is 0.906), confirming that the intervals are correctly calibrated. The model neither under-

nor over-estimates its own uncertainty. Figure 11 illustrates the intervals on one test split. Points inside the band are covered, while the few outside it correspond to the expected percentage of misses.

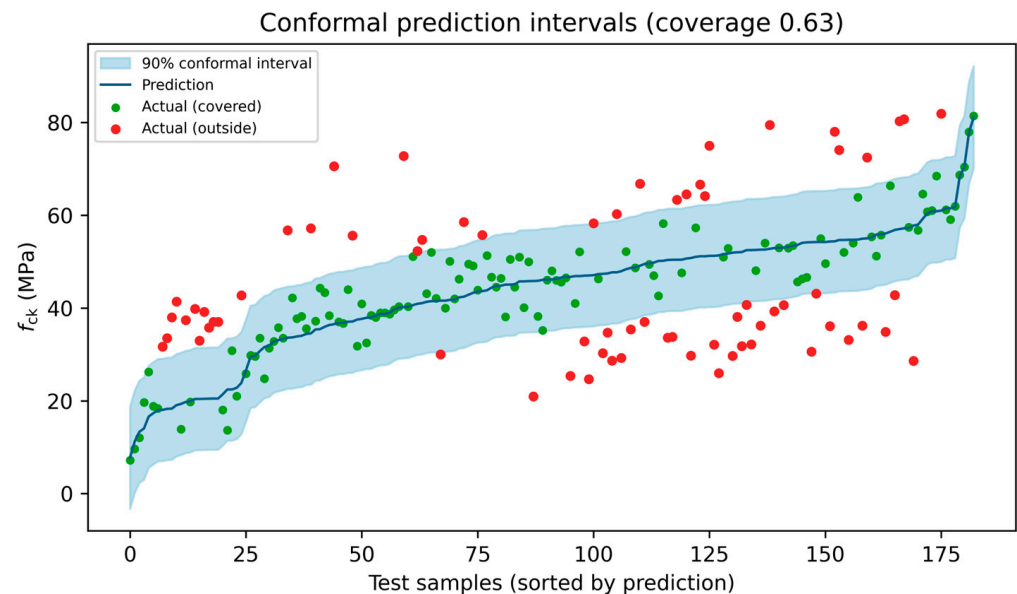


Figure 11. 90% conformal prediction intervals on one test split (samples sorted by predicted value): green points are covered; red are outside the interval.

In addition to the average coverage, the conditional coverage was also examined (the proportion of covered true values across compressive-strength ranges—low, medium, and high terciles). The result is shown in Figure 12. Coverage is close to the nominal 90% in all three ranges (low 0.89, medium 0.96, high 0.87), with slight under-coverage in the extreme ranges, where the data are sparser—a known property of the symmetric (split) variant—which is precisely what motivates the adaptive approach described below.

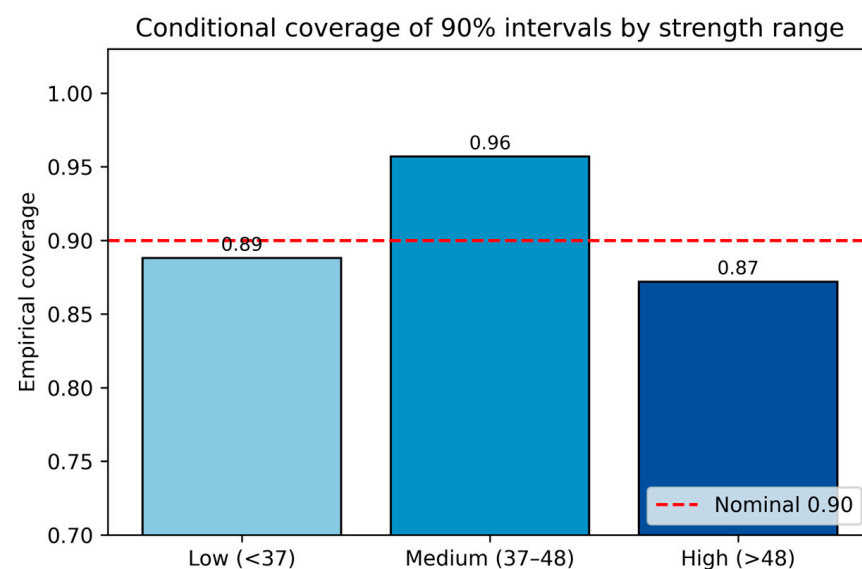


Figure 12. Conditional coverage of the 90% intervals across low/medium/high strength ranges, averaged over 20 random splits (dashed line: nominal 0.90); bars are shaded by strength range (low, medium, high).

In addition to the symmetric (split) variant, the adaptive conformalized quantile regression (CQR) [33] was also examined. At the 90% level, CQR achieves a coverage

of 0.910 with an average width of 29.8 MPa, against 25.8 MPa for the split variant. CQR produces intervals whose width varies locally (narrower where the prediction is more reliable), with the same guaranteed coverage. On this noisy dataset, the average width is not smaller than that of the symmetric variant, but the adaptivity more faithfully follows the distribution of uncertainty. The dominant driver of the width can be identified. Removing the repeated and contradictory records does not narrow the intervals. On the deduplicated dataset the 90% width is 28.4 MPa, slightly wider than the 25.8 MPa on the full dataset (owing to the smaller calibration set), so duplicated records are not the source of the width. The width instead tracks the irreducible prediction error, while the extreme strength ranges, which are more sparsely sampled, show slightly lower coverage (Figure 12). The width is therefore governed mainly by measurement noise and source heterogeneity, with data sparsity affecting the extremes and duplication playing no widening role. The relatively large widths of both variants are a consequence of the irreducible noise in the data (partly the contradictory records), and faithfully convey the real limit of reliability rather than concealing it. The width should be read against the prediction error. For an approximately normal error with the conformal residual spread observed here (a proper-training RMSE of about 7.5 MPa), a two-sided 90% interval already spans about $2 \times 1.645 \times 7.5 = 24.7$ MPa, so the measured 25.8 MPa is close to this theoretical minimum and reflects the residual scatter rather than any conservatism of the method. A width of 25.8 MPa corresponds to roughly one third of the observed strength range (7.2–81.9 MPa). For mix design, the intervals remain useful. The guaranteed coverage supports safety-side decisions based on the lower bound, and a user who needs tighter bounds can adopt a lower nominal level (the 80% interval narrows to 17.1 MPa).

6. Discussion

The proposed single LightGBM model achieves an $R^2 = 0.794$ and an RMSE = 6.20 MPa, values above all four machine learning models from the reference study [4], using a single model. It also offers fast training, a small number of components, and full explainability of predictions through SHAP analysis, which for engineering use is often as important as accuracy itself.

6.1. Aspects in Which the Reference Study Retains an Advantage

For the sake of objectivity, the aspects in which the reference study retains an advantage should also be noted. First, Jagadeesh et al. [4] report that their neural network achieves an $R^2 = 0.86$ on the test set, which is higher than the R^2 (0.794) of the model presented in this work. Although that result comes from a single 70/30 split and is difficult to reproduce, on paper it represents a higher reported performance and serves as a benchmark to aim for. Second, the neural network from the reference study is architecturally very simple (a single hidden layer with nine neurons). Third, more complex stacking ensembles can yield a slightly higher R^2 than a single model. The aim of this work was not to achieve maximal accuracy but the best balance of accuracy, simplicity, and explainability.

6.2. Relation to a Recent Conformal-Prediction Work on a Related Dataset

Methodologically closest to this work is the recently published paper by Dong and Zhang [18], which, together with predicting the compressive strength of self-compacting concrete with recycled aggregate, also provides split-conformal intervals and SHAP interpretation, but on a different dataset (289 mixtures, 18 variables) and with a hybrid deep architecture. On their dataset, they report high performance (test $R^2 = 0.940$, RMSE = 4.22 MPa) and an average 90% interval width of 8.32 MPa, with a measured coverage of 94.8%. These values are not directly comparable with those obtained in this

work, since they come from different datasets. The dataset in [18] contains more informative variables (e.g., curing age and the water-to-binder ratio) and a different noise structure. Both the accuracy and the interval width are therefore, to a large extent, a property of the dataset rather than of the model alone. In terms of calibration, the measured coverage in [18] (94.8% at a nominal 90%) indicates somewhat more conservative intervals, whereas in this work the coverage almost matches the nominal one (90.6%). On the other hand, the framework in [18] has a clear advantage in handling incomplete data (a drop of only about 3% in the R^2 at 15% missing values), which this work does not address because the reference dataset is complete. The two studies are therefore complementary. The work [18] pushes the boundary of model usability under incomplete data and a richer mixture description, while the research in this work pushes the boundary of evaluation reliability through a data-quality check (leakage of identical compositions), evaluation over 25 splits, and exact TreeSHAP interpretation of a simple model. Taken together, the results suggest that conformal prediction is maturing into a standard tool for mix design, but also that reported interval widths should always be interpreted in the context of the specific dataset.

6.3. Practical Implications and Limitations

The agreement of the importance measures and the SHAP analysis with the physical laws of concrete increases confidence in the model and enables its use as a tool for guiding mix design. Limitations include the relatively small dataset (603 mixtures) collected from heterogeneous sources, the presence of extreme values of some admixtures, and the fact that the conclusions are limited to the range of compositions represented in the data. External validation on a fully independent dataset was not performed because no public benchmark with the same eight-variable description is available. The most closely related dataset [18] uses a different, eighteen-variable description and a different noise structure, so it is not directly compatible. As an internal surrogate, a leave-one-source-out analysis was carried out, in which the model was trained on all source studies but one and evaluated on the held-out study. The pooled coefficient of determination then drops to about 0.04, and most individual sources yield negative values, which shows that the reported performance holds within the distribution of the benchmark sources and should not be extrapolated to an entirely new source without recalibration. Future research could therefore include enlarging the dataset and external validation on independent mixtures.

7. Conclusions

This work has shown that the reported machine learning accuracies on this widely used dataset partly originate from data leakage (repeated and contradictory compositions), and that an objective leakage-free estimate (with re-tuning on the deduplicated dataset) yields an R^2 of about 0.73 instead of ~ 0.81 . The conducted analysis singles out this lower value as a more realistic indicator of performance on this problem. With that correction, an interpretable, optimized LightGBM model equipped with calibrated conformal uncertainty intervals was proposed. Under an identical 70/30 protocol (averaged over 25 splits), the model achieves an $R^2 = 0.794 \pm 0.038$ and an RMSE = 6.20 ± 0.48 MPa, with R^2 and RMSE values exceeding all four machine learning models from the reference study ($R^2 = 0.7249\text{--}0.7635$). Beyond accuracy, the contribution of this work lies in explainability, as SHAP and partial dependence confirm that the model learns physically meaningful relationships, with cement as the dominant predictor, followed by water and the mineral admixture. Compared with the most recent work on the same dataset (BO-GBRT) [17], the proposed model under an identical protocol attains a result at the level of the reference (seed-averaged five-fold CV $R^2 = 0.813$ versus 0.806), and, unlike the single-partition protocol of [17], quantifies the variability across partition choices (± 0.017). Because the

reference reports a single partition, a formal variance comparison is not possible and no lower-variance claim is made (Section 5.4). The proposed model additionally augments the analysis with conformal uncertainty intervals (empirical coverage 0.906 at the 90% nominal level) and a leakage-aware evaluation that reveals the repeated compositions in the dataset. The complementary recent work with conformal intervals on a related dataset [18] further confirms that uncertainty quantification is becoming a standard in this field. The proposed approach offers a favorable balance of accuracy, simplicity, interpretability, and reliability, is suitable for practical use, and is in line with the trend in the recent literature [8,9] that recognizes interpretable gradient-boosting models as a reliable and physically consistent solution for predicting the compressive strength of recycled-aggregate concrete. Future work will focus on external validation on independent datasets with richer mixture descriptions, including the curing age, and on applying the leakage-aware evaluation protocol to other benchmark datasets in this field.

Author Contributions: Conceptualization, N.K. and B.M.; methodology, N.K.; software, N.K. and M.K.; validation, N.K., B.M. and M.K.; formal analysis, N.K. and M.K.; investigation, B.M.; resources, N.K.; data curation, B.M. and M.K.; writing—original draft preparation, N.K. and B.M.; writing—review and editing, N.K., B.M. and M.K.; visualization, N.K. and M.K.; supervision, N.K.; project administration, B.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Ministry of Science, Technological Development, and Innovation of the Republic of Serbia, under the Agreement on Financing the Scientific Research Work of Teaching Staff at the Faculty of Mechanical Engineering and Civil Engineering in Kraljevo, University of Kragujevac—Registration number: 451-03-34/2026-03/200108, dated 5 February 2026. The conducted research is related to the goals of sustainable cities and communities and industry, innovation and infrastructure in accordance with the universal strategy of the United Nations (Agenda 2030).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset analyzed in this study was obtained from the Supporting Information of the reference study [4] and is publicly available. The code and the trained model that support the findings are available from the corresponding author upon reasonable request.

Acknowledgments: During the preparation of this manuscript, the authors used the AI assistant Claude (Anthropic; Claude Opus 4.8) for literature-search support and for language and consistency checking. All scientific content, analyses, and conclusions were produced and verified by the authors, who take full responsibility for the content of the publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Duan, Z.H.; Kou, S.C.; Poon, C.S. Prediction of Compressive Strength of Recycled Aggregate Concrete Using Artificial Neural Networks. *Constr. Build. Mater.* **2013**, *40*, 1200–1206. [\[CrossRef\]](#)
2. Suescum-Morales, D.; Salas-Morera, L.; Jiménez, J.R.; García-Hernández, L. A Novel Artificial Neural Network to Predict Compressive Strength of Recycled Aggregate Concrete. *Appl. Sci.* **2021**, *11*, 11077. [\[CrossRef\]](#)
3. Nguyen, T.-D.; Cherif, R.; Mahieux, P.-Y.; Lux, J.; Ait-Mokhtar, A.; Bastidas-Arteaga, E. Artificial Intelligence Algorithms for Prediction and Sensitivity Analysis of Mechanical Properties of Recycled Aggregate Concrete: A Review. *J. Build. Eng.* **2023**, *66*, 105929. [\[CrossRef\]](#)
4. Jagadesh, P.; Khan, A.H.; Priya, B.S.; Asheeka, A.; Zoubir, Z.; Magbool, H.M.; Alam, S.; Bakather, O.Y. Artificial Neural Network, Machine Learning Modelling of Compressive Strength of Recycled Coarse Aggregate Based Self-Compacting Concrete. *PLoS ONE* **2024**, *19*, e0303101. [\[CrossRef\]](#) [\[PubMed\]](#)

5. Boudali, S.; Abdulsalam, B.; Rafiean, A.H.; Poncet, S.; Soliman, A.; ElSafty, A. Influence of Fine Recycled Concrete Powder on the Compressive Strength of Self-Compacting Concrete (SCC) Using Artificial Neural Network. *Sustainability* **2021**, *13*, 3111. [\[CrossRef\]](#)
6. de-Prado-Gil, J.; Palencia, C.; Jagadesh, P.; Martínez-García, R. A Study on the Prediction of Compressive Strength of Self-Compacting Recycled Aggregate Concrete Utilizing Novel Computational Approaches. *Materials* **2022**, *15*, 5232. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Alyaseen, A.; Poddar, A.; Kumar, N.; Tajjour, S.; Prasad, C.V.S.R.; Alahmad, H.; Sihag, P. High-Performance Self-Compacting Concrete with Recycled Coarse Aggregate: Soft-Computing Analysis of Compressive Strength. *J. Build. Eng.* **2023**, *77*, 107527. [\[CrossRef\]](#)
8. Yang, S.; Chen, T.; Xu, Z. Estimation on Compressive Strength of Recycled Aggregate Self-Compacting Concrete Using Interpretable Machine Learning-Based Models. *Eng. Comput.* **2024**, *41*, 2727–2773. [\[CrossRef\]](#)
9. Almutairi, A. Explainable Machine Learning-Based Prediction of Compressive Strength in Sustainable Recycled Aggregate Self-Compacting Concrete Using SHAP Analysis. *Sustainability* **2025**, *17*, 11334. [\[CrossRef\]](#)
10. Chiaburu, T.; Haußer, F.; Bießmann, F. Uncertainty in XAI: Human Perception and Modeling Approaches. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 1170–1192. [\[CrossRef\]](#)
11. Nikmehr, B.; Kafle, B.; Al-Ameri, R. Performance Assessment of One-Part Self-Compacted Geopolymer Concrete Containing Recycled Concrete Aggregate: A Critical Comparison Using Artificial Neural Network (ANN) and Linear Regression Models. *Recycling* **2024**, *9*, 73. [\[CrossRef\]](#)
12. de-Prado-Gil, J.; Martínez-García, R.; Jagadesh, P.; Juan-Valdés, A.; González-Alonso, M.-I.; Palencia, C. To Determine the Compressive Strength of Self-Compacting Recycled Aggregate Concrete Using Artificial Neural Network (ANN). *Ain Shams Eng. J.* **2024**, *15*, 102548. [\[CrossRef\]](#)
13. Yang, S.; Sun, J.; Xu, Z. Prediction on Compressive Strength of Recycled Aggregate Self-Compacting Concrete by Machine Learning Method. *J. Build. Eng.* **2024**, *88*, 109055. [\[CrossRef\]](#)
14. Fu, H.; Zhou, X.; Xu, P.; Sun, D. Prediction of Compressive Strength of Concrete Using Explainable Machine Learning Models. *Materials* **2025**, *18*, 5009. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Zhang, J.; Wang, Z.; Shen, S.; Sheng, S.; He, H.; He, C. Hybrid Explainable Machine Learning Models with Metaheuristic Optimization for Performance Prediction of Self-Compacting Concrete. *Buildings* **2026**, *16*, 225. [\[CrossRef\]](#)
16. Awoyera, P.O.; Simwanda, L.; Vasić, M.V.; Mydin, M.A.O.; Imoh, U.U.; Fadugba, O.G.; Asiz, A.; Movahedi Rad, M. Hybrid Generative–Ensemble Approach for Predicting Recycled Aggregate Concrete Strength Properties. *Sci. Rep.* **2026**, *16*, 15205. [\[CrossRef\]](#) [\[PubMed\]](#)
17. Abood, E.A.; Thoeny, Z.A.; Azize, N.M.; Imran, H.; Kamchoom, V.; Onyelowe, K.C.; Arunachalam, K.P. Self Compacting Concrete with Recycled Aggregate Compressive Strength Prediction Based on Gradient Boosting Regression Tree with Bayesian Optimization Hybrid Model. *Sci. Rep.* **2025**, *15*, 28175. [\[CrossRef\]](#) [\[PubMed\]](#)
18. Dong, S.; Zhang, Z. Hybrid Deep Learning with Conformal Prediction for Recycled Aggregate Self-Compacting Concrete Strength Prediction. *Buildings* **2025**, *15*, 4419. [\[CrossRef\]](#)
19. Stadler, S.; Betancourt, C.; Roscher, R. Explainable Machine Learning Reveals Capabilities, Redundancy, and Limitations of a Geospatial Air Quality Benchmark Dataset. *Mach. Learn. Knowl. Extr.* **2022**, *4*, 150–171. [\[CrossRef\]](#)
20. Potts, R.B.; Hackney, R.; Leontidis, G. Tabular Machine Learning Methods for Predicting Gas Turbine Emissions. *Mach. Learn. Knowl. Extr.* **2023**, *5*, 1055–1075. [\[CrossRef\]](#)
21. Airlangga, G.; Liu, A. A Hybrid Gradient Boosting and Neural Network Model for Predicting Urban Happiness: Integrating Ensemble Learning with Deep Representation for Enhanced Accuracy. *Mach. Learn. Knowl. Extr.* **2025**, *7*, 4. [\[CrossRef\]](#)
22. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems 30*; Curran Associates: Red Hook, NY, USA, 2017; pp. 3146–3154.
23. Breiman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.J. *Classification and Regression Trees*; Wadsworth: Belmont, CA, USA, 1984. [\[CrossRef\]](#)
24. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [\[CrossRef\]](#)
25. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. [\[CrossRef\]](#)
26. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: Unbiased Boosting with Categorical Features. In *Advances in Neural Information Processing Systems 31*; Curran Associates: Red Hook, NY, USA, 2018; pp. 6638–6648.
27. Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A Next-Generation Hyperparameter Optimization Framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 2623–2631. [\[CrossRef\]](#)
28. Emmert-Streib, F.; Dehmer, M. Evaluation of Regression Models: Model Assessment, Model Selection and Generalization Error. *Mach. Learn. Knowl. Extr.* **2019**, *1*, 521–551. [\[CrossRef\]](#)

29. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
30. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*; Curran Associates: Red Hook, NY, USA, 2017; pp. 4765–4774.
31. Li, S.; Chen, J.-W. Enhanced Motion Prediction of a Semi-Submersible Platform Using Bayesian Neural Network and Field Monitoring Data. *AI Eng.* **2026**, *1*, 2. [[CrossRef](#)]
32. Lei, J.; G'Sell, M.; Rinaldo, A.; Tibshirani, R.J.; Wasserman, L. Distribution-Free Predictive Inference for Regression. *J. Am. Stat. Assoc.* **2018**, *113*, 1094–1111. [[CrossRef](#)]
33. Romano, Y.; Patterson, E.; Candès, E.J. Conformalized Quantile Regression. In *Advances in Neural Information Processing Systems 32*; Curran Associates: Red Hook, NY, USA, 2019; pp. 3543–3553.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.