

Cite this:
DOI: [10.56748/ejse.261000](https://doi.org/10.56748/ejse.261000)

Received Date: 17 April 2026
Accepted Date: 31 August 2026

1443-9255
<https://ejsei.com/ejse>
Copyright: © The Author(s).
Published by Electronic Journals
for Science and Engineering
International (EJSEI).
This is an open access article
under the CC BY license.
<https://creativecommons.org/licenses/by/4.0/>



Explainable Machine Learning for Compressive Stress Prediction in Cement-Based Construction Materials: A Systematic Benchmarking and SHAP Interpretability Study

M. Mallikarjuna Rao ^{a*}, C. Manjunath ^b, G. Malli Reddy ^c, D. Venkatesh ^d, Shaik Rahamat Basha ^e, D. V. Bhaskar ^f

^a Department of Computer Science and Design, PVKK Institute of Technology, Anantapuramu, India.

^b Department of Computer Applications, PVKK Institute of Technology, Anantapuramu, India.

^c Department of Computer Science and Engineering, PVKK Institute of Technology, Anantapuramu, India.

^d Department of Computer Science and Engineering, Ananthalakshmi Institute of Technology, Anantapuramu, A.P., India.

^e Department of Computer Science and Engineering, St. Peter's Engineering College, Hyderabad, Telangana, India.

^f Department of Computer Applications, Siddaganga Institute of Technology, Tumkur, Karnataka, India.

*Corresponding author: mallikari.mkrao@gmail.com

Abstract

Accurate prediction of the compressive strength of cement-based materials remains a central problem in structural engineering, because the property depends on strongly nonlinear interactions among mixture proportions, supplementary cementitious materials and curing age. Recent systematic reviews of the machine learning literature in this area report three recurring weaknesses: model comparisons are seldom made under one shared data split and are rarely validated statistically; raw mix-design variables are usually passed to the learner without encoding established concrete science; and analysis code is almost never released, so published accuracies cannot be reproduced or placed on a common footing. This study addresses all three on the University of California, Irvine (UCI) Concrete Compressive Strength dataset ($n = 1,030$). Nine regression models, from ordinary least squares to gradient-boosted ensembles, were compared under one identical ten-fold cross-validation (CV) protocol on a fixed 80/20 stratified split, with all model selection confined to the training partition. The eight raw mix-design variables were augmented with eight physics-informed features derived from Abrams' strength law and cement hydration kinetics, among them the water-to-cement ratio, the water-to-binder ratio and log-transformed curing age. Hyperparameters of the two gradient-boosted learners were tuned by Bayesian optimisation in Optuna over 100 trials each. Extreme gradient boosting (XGBoost) gave the best held-out result, with a coefficient of determination (R^2) of 0.9567, a root mean squared error (RMSE) of 3.46 MPa and a mean absolute error (MAE) of 2.35 MPa. The four gradient-boosted variants, tuned and untuned XGBoost and light gradient boosting machine (LightGBM), could not be separated from one another by Wilcoxon signed-rank tests, and all four were significantly better than the kernel, bagging and linear baselines ($p = 0.002$). Bayesian tuning raised the cross-validated score of both boosters yet lowered their held-out score; this negative result is reported in full because it carries a concrete rule for practitioners. SHapley Additive exPlanations (SHAP) computed for both tuned boosters identified curing age, the water-to-binder ratio and total binder content as the dominant predictors, in agreement with physical theory, and the two models returned the same six leading features with closely matching global importance (Spearman $\rho = 0.829$). The complete pipeline is released as open source under a single fixed random seed.

Keywords

Compressive strength prediction, Cement concrete, Machine learning, XGBoost, LightGBM, SHAP explainability, Bayesian optimization, Physics-informed feature engineering

1. Introduction

Compressive strength is the mechanical property that governs the structural performance and durability of cement-based composites (Mehta and Monteiro, 2014). In routine practice it is determined by destructive testing of cylinders or cubes at 28 days (ASTM International, 2021), a procedure that is slow, consumes material and labour, and gives no early feedback while a mix design is still being adjusted. A reliable predictive model that maps mix proportions and cure age to strength therefore has clear practical value for the construction industry.

Mapping itself is difficult. Strength depends nonlinearly on cement content, the water-to-binder ratio, supplementary cementitious materials (SCMs) such as blast furnace slag and fly ash, aggregate composition, chemical admixtures and curing age (Neville, 2011). Classical empirical relations, most notably Abrams' water-cement law (Abrams, 1918), capture the first-order dependence of strength on the water-to-cement ratio, but they do not account for SCM synergies, superplasticiser interactions, or the roughly logarithmic gain of strength with time that follows from hydration kinetics (Powers, 1958).

Machine learning (ML) has consequently attracted sustained attention in this area. Yeh (1998) showed that a feedforward artificial neural network (ANN) trained on what is now the UCI benchmark dataset clearly outperformed multiple linear regression ($R^2 = 0.92$ against 0.62). Chou et al. (2011) reported a random forest reaching R^2 of about 0.93 on similar data, and later multi-nation analyses confirmed the advantage of nonlinear

learners (Chou et al., 2014). Hybrid and boosted formulations followed, including hybrid intelligence models for lightweight aggregate concretes (Zhang et al., 2020), adaptive boosting (Feng et al., 2020) and tuned gradient boosting machines (GBMs) for self-compacting concrete (Wang et al., 2024), with the ensemble frameworks XGBoost (Chen and Guestrin, 2016) and LightGBM (Ke et al., 2017) now the most common choices. A meta-analysis by Nunez et al. (2021) covering 3,450 experimental data points found that gradient-boosted trees consistently gave the lowest RMSE across diverse concrete compositions.

Two persistent weaknesses run through this literature. First, published studies rarely share identical train-test splits or CV protocols, so headline accuracies are not directly comparable, a concern raised repeatedly in critical reviews (Ben Chaabene et al., 2020; Nunez et al., 2021). The systematic review by Pengyu et al. (2025), which synthesised 85 studies published between 2015 and 2024 on ML for stress evolution in cement-based materials, identified this absence of statistically validated cross-model benchmarking as a key gap. That review reported that fewer than one study in five applied any significance test to its model comparisons, and that the reported accuracy of a given algorithm family varied more across papers than it did across algorithms within a paper, which is what one expects when protocols rather than models are driving the numbers.

The same review identified a second gap: most studies feed raw mix-design variables to the learner and ignore the physics of the material. The water-to-cement ratio governs gel porosity through Abrams' law, the water-to-binder ratio generalises it to blended cements (Neville, 2011),

and the strength-time relation is close to log-linear (Powers, 1958), yet these transformations are seldom encoded as model inputs (Pengyu et al., 2025; Ben Chaabene et al., 2020). Where they are, the benefit is measurable. Jiang et al. (2022) improved XGBoost R^2 from 0.931 to 0.947 on 2,100 specimens simply by adding the water-to-cement ratio, total binder content and SCM fraction as features, and that study remains one of the few in the comparison set of Section 3.8 to encode any domain relation at all.

A third limitation concerns interpretability and reproducibility. Accurate black-box predictions are of limited use to a practicing engineer without insight into which variables drive them, and interpretability is increasingly demanded in regulated, high-stakes settings (Rudin, 2019). SHAP values (Lundberg and Lee, 2017), computed exactly for tree ensembles by the TreeSHAP algorithm (Lundberg et al., 2020), attribute each prediction to individual features in a locally accurate and globally consistent way, but their application to concrete strength modelling is still sparse; examples include interpretability analyses for geopolymers concretes (Shahmansouri et al., 2021) and for steel fibre reinforced concrete beams (Rahman et al., 2021). On the tuning side, Bayesian optimisation with tree-structured Parzen estimators (TPE), as implemented in Optuna (Akiba et al., 2019), has repeatedly beaten grid and random search for boosted models (Turner et al., 2021; Zou et al., 2022), yet many concrete studies still rely on manual or grid tuning, and very few release code that would allow their numbers to be reproduced.

The novelty of the present work lies in closing these gaps within a single, openly released framework applied to the standard UCI benchmark. To the best of our knowledge, no previous study on this dataset combines all of the following: (i) a benchmark of nine regressors, from linear baselines to gradient-boosted ensembles, evaluated under one identical ten-fold CV protocol and one fixed stratified split, with model selection confined to the training partition; (ii) eight physics-informed engineered features grounded in Abrams' law, hydration kinetics and mix-design practice; (iii) Bayesian hyperparameter optimisation with 100 Optuna trials per tuned model; (iv) a global and local SHAP explainability analysis carried out on out-of-sample data for two independently tuned boosters, with the agreement between them quantified rather than

assumed; and (v) non-parametric statistical validation of every pairwise model comparison with Wilcoxon signed-rank tests. In addition, the full pipeline is published so that every table and figure can be regenerated from a single command, which directly addresses the reproducibility concern raised in the review literature.

The remainder of the paper is organised as follows. Section 2 describes the dataset, the physics-informed feature engineering, the models, the tuning procedure, the evaluation protocol and the explainability analysis. Section 3 presents and discusses the results, covering exploratory analysis, the CV benchmark, hyperparameter optimisation, test-set performance, SHAP analysis, residual diagnostics, learning curves, comparison with prior literature and limitations of the dataset. Section 4 details data availability and reproducibility. Section 5 concludes and outlines future work.

2. Materials and Methods

2.1 Dataset

The UCI Concrete Compressive Strength dataset, originally compiled by Yeh (1998), contains $n = 1,030$ experimental observations of cement-based mixtures with eight input variables and one continuous target, the compressive strength (CS) in megapascals (MPa). The eight predictors and their physical units are: cement content (kg/m^3), blast furnace slag (kg/m^3), fly ash (kg/m^3), water content (kg/m^3), superplasticiser dosage (kg/m^3), coarse aggregate (kg/m^3), fine aggregate (kg/m^3) and curing age (days). There are no missing values. The target ranges from 2.33 MPa to 82.60 MPa (mean 35.82 MPa, standard deviation (SD) 16.71 MPa), covering normal-strength, high-strength and low water-to-binder ratio mixes.

Feature distributions are summarised in Fig. 1 and Table 1. Slag and fly ash are strongly right skewed, with modes at zero because a large share of specimens contains no SCM. Water content is close to Gaussian and centered at 181.6 kg/m^3 . Age is highly right skewed, running from 1 to 365 days with a modal value of 28 days.

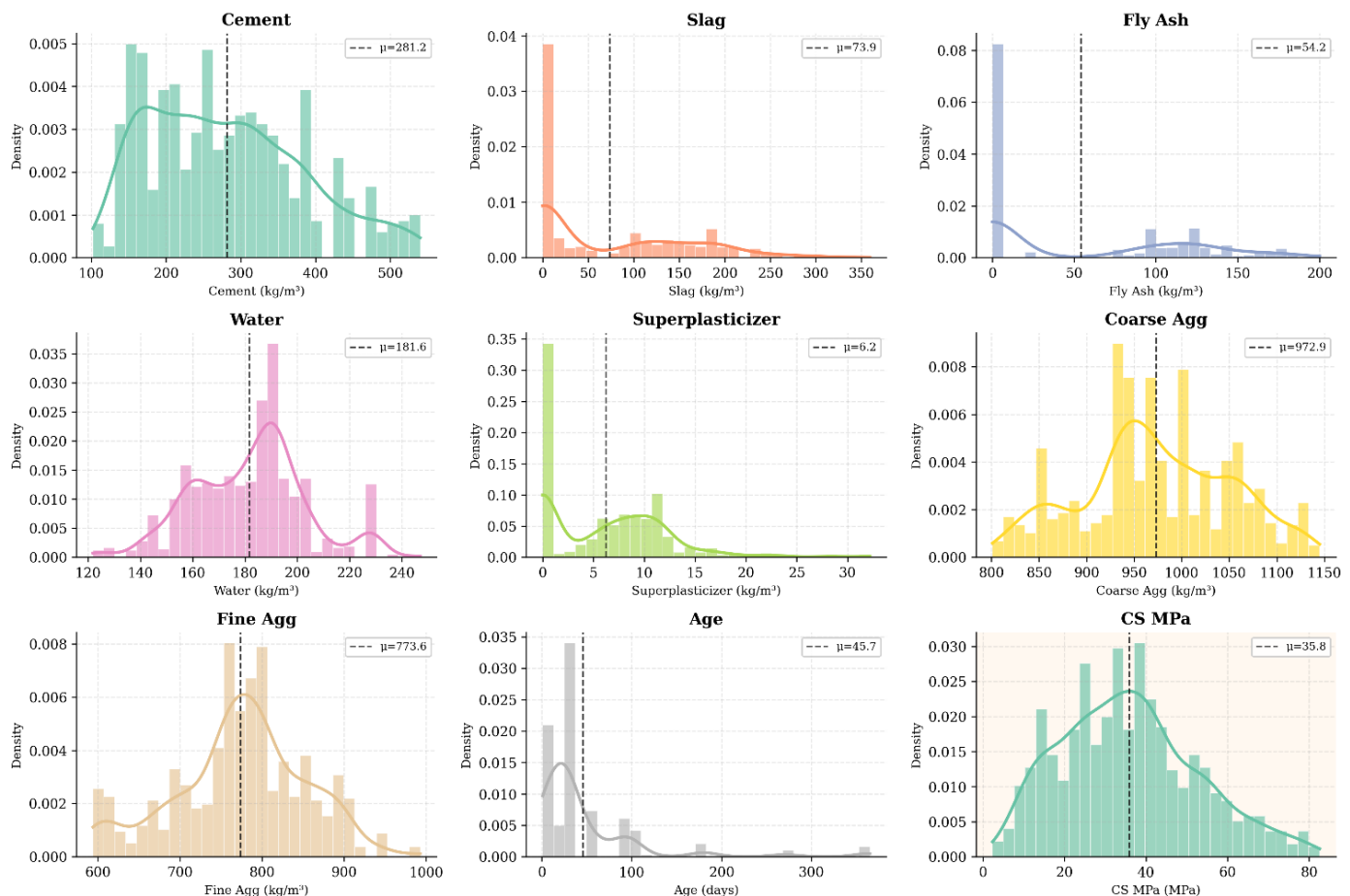


Fig. 1 Marginal distributions of the eight mix-design input variables and the compressive strength target ($n = 1,030$). Dashed vertical lines mark feature means; kernel density overlays are shown in matching colours.

Table 1. Descriptive statistics of the UCI Concrete Compressive Strength dataset (n = 1,030).

Variable	Unit	Mean	SD	Min	Median	Max	Skew
Cement	kg/m ³	281.2	104.5	102.0	272.9	540.0	0.51
Slag	kg/m ³	73.9	86.3	0.0	22.0	359.4	1.32
Fly Ash	kg/m ³	54.2	64.0	0.0	0.0	200.1	1.07
Water	kg/m ³	181.6	21.4	121.8	185.0	247.0	0.07
Superplasticiser	kg/m ³	6.2	5.97	0.0	6.4	32.2	1.09
Coarse Agg.	kg/m ³	972.9	77.8	801.0	968.0	1145.0	-0.01
Fine Agg.	kg/m ³	773.6	80.2	594.0	779.5	992.6	-0.24
Age	days	45.7	63.2	1	28	365	2.71
CS (target)	MPa	35.82	16.71	2.33	34.45	82.60	0.42

SD = standard deviation.

2.2 Physics-Informed Feature Engineering

Eight additional features were derived from the raw mix-design variables, each grounded in established concrete science (Abrams, 1918; Powers, 1958; Neville, 2011). They encode domain knowledge that a regressor cannot easily recover from 1,030 raw samples on its own:

Water-to-cement ratio (WC_Ratio): $w/c = \text{Water} / \text{Cement}$. The primary determinant of gel porosity and capillary pore volume, inversely related to strength through Abrams' law (Abrams, 1918).

Total binder content (Binder): $\text{Binder} = \text{Cement} + \text{Slag} + \text{Fly Ash}$ (kg/m³), the full cementitious contribution to the paste matrix.

Water-to-binder ratio (WB_Ratio): $w/b = \text{Water} / \text{Binder}$, the stronger predictor for blended cement systems where SCMs contribute to long-term strength development (Neville, 2011).

SCM fraction (SCM_Frac): $(\text{Slag} + \text{Fly Ash}) / \text{Binder}$, the degree of cement replacement by supplementary materials.

Aggregate-to-binder ratio (Agg_Binder): $(\text{Coarse Agg.} + \text{Fine Agg.}) / \text{Binder}$, the volumetric balance between the inert aggregate skeleton and the reactive binder paste.

Log-transformed age (Log_Age): $\ln(1 + \text{Age})$, which linearises the approximately log-linear strength-time relation arising from hydration kinetics (Powers, 1958).

Superplasticiser efficiency (SP_Eff): $\text{Superplasticiser} / \text{Water}$, the admixture dosage per unit water, capturing its water-reducing effect.

Fine-to-coarse aggregate ratio (FA_CA): $\text{Fine Agg.} / \text{Coarse Agg.}$, which encodes grading information relevant to void content and the interfacial transition zone.

The final feature matrix contains 16 predictors, eight original and eight engineered. Feature scaling was applied only to the five models whose loss surface or kernel depends on the scale of the inputs: ordinary least squares linear regression, Ridge, Lasso, ElasticNet and support vector regression (SVR). For each of these five, a standard scaler was chained ahead of the estimator inside a single pipeline object, so the scaler is refitted on the training portion of every CV fold and never sees the validation fold or the held-out test set. The four tree ensembles, random forest, extra trees, XGBoost and LightGBM, split on ordered feature values and are invariant to monotone rescaling, so they were trained on the unscaled 16-column matrix. No transformation was applied to the target for any model.

2.3 Machine Learning Models

Nine regressors were evaluated, spanning the complexity range from linear baselines to gradient-boosted ensembles: (i) ordinary least squares linear regression; (ii) Ridge regression ($\alpha = 1.0$); (iii) Lasso regression ($\alpha = 0.5$); (iv) ElasticNet ($\alpha = 0.5$, $l1_ratio = 0.5$); (v) SVR with a radial basis function (RBF) kernel ($C = 100$, γ set to the scikit-learn 'scale' default, $\epsilon = 0.1$); (vi) random forest (300 trees); (vii) extra trees (300 trees); (viii) XGBoost (Chen and Guestrin, 2016) with 500 trees, learning rate 0.05, maximum depth 6, and subsample and column-subsample fractions of 0.8; and (ix) LightGBM (Ke et al., 2017) with 500 trees, 63 leaves, learning rate 0.05 and subsample fraction 0.8. The five scale-sensitive models were wrapped in the scaling pipeline described in Section 2.2. All nine used default or lightly hand-set parameters in this baseline phase. Bayesian optimisation was then applied to XGBoost and LightGBM as described next, and the tuned configurations are reported alongside their untuned counterparts throughout, so that the effect of tuning can be read directly.

2.4 Hyperparameter Optimisation

Bayesian optimisation was run in Optuna (Akiba et al., 2019) with the TPE sampler, using 100 independent trials per model and a fixed sampler seed. The objective was the mean R^2 from an inner five-fold CV on the training set ($n = 824$); the held-out test set played no part in the search. The XGBoost search space covered nine parameters: $n_estimators \in [200, 1200]$, $max_depth \in [3, 10]$, $learning_rate \in [0.005, 0.25]$ on a log scale, $subsample \in [0.5, 1.0]$, $colsample_bytree \in [0.5, 1.0]$, $min_child_weight \in [1, 12]$, $gamma \in [0.0, 0.6]$, $reg_alpha \in [10^{-8}, 1.0]$ and $reg_lambda \in [10^{-8}, 5.0]$, the last two on a log scale. An analogous nine-parameter space was

defined for LightGBM, substituting $num_leaves \in [20, 200]$ and $min_child_samples \in [5, 60]$ for max_depth and min_child_weight .

2.5 Experimental Protocol and Evaluation

The dataset was split into a training set (80%, $n = 824$) and a held-out test set (20%, $n = 206$). The split was stratified on decile bins of the target so that the strength distribution of the two partitions matches, and the seed was fixed at 42. All model selection, hyperparameter tuning and CV used the training partition only; the test set was evaluated once, after every modelling decision had been made. Reported CV figures are the mean and SD of ten-fold CV over the 824 training specimens, using one shuffled fold assignment shared by all eleven model configurations, so that the paired comparisons of Section 3.2 are made fold by fold on identical data.

Four metrics were used to quantify performance. Writing y_i for the observed strength of specimen i , $\hat{y}_i$ for the corresponding prediction, $\bar{y}$ for the mean observed strength and n for the number of evaluated specimens, they are defined as follows:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \quad (1)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_i (y_i - \hat{y}_i)^2} \quad (2)$$

$$MAE = \frac{1}{n} \sum_i |y_i - \hat{y}_i| \quad (3)$$

$$MAPE = \frac{100}{n} \sum_i |y_i - \hat{y}_i| / y_i \quad (4)$$

The coefficient of determination R^2 measures the fraction of the variance in strength explained by the model, RMSE (in MPa) penalises large errors quadratically, MAE (in MPa) gives the average error magnitude, and the mean absolute percentage error MAPE expresses the error relative to the observed strength in percent.

Wilcoxon signed-rank tests (Wilcoxon, 1945) on the paired ten-fold CV R^2 scores were used to test whether the best model differed from each of the other configurations at $\alpha = 0.05$ (Table 2). The tests are two-sided, since the direction of any difference is not known in advance, and the reference model is the configuration with the highest test R^2 rather than a model chosen beforehand. With ten paired folds the smallest two-sided p-value attainable is 0.00195, so all significant results in Table 2 are reported at that resolution and no claim of $p < 0.001$ is made. A significant result indicates that the two vectors of fold scores differ; the direction is read from the accompanying mean CV R^2 column.

2.6 SHAP Explainability Analysis and Choice of the Explained Model

SHAP values were computed with the TreeSHAP algorithm (Lundberg et al., 2020) on the held-out test set ($n = 206$), so that the explanations describe out-of-sample behaviour. Global importance is measured by the mean absolute SHAP value, mean $|\phi|$, reported in MPa; beeswarm plots, global bar charts and dependence plots for the four most influential features characterise both the magnitude and the direction of each contribution.

Which model should carry the explainability analysis deserves an explicit answer, because the four gradient-boosted configurations in Table 2 differ by less than 0.006 in test R^2 and the Wilcoxon tests of Section 3.2 do not separate any of them from the others (p between 0.193 and 0.695). Accuracy alone therefore does not single out one model as the one that must be explained, and picking whichever variant happens to top the test column by a few thousandths would make the interpretation depend on sampling noise in a 206-specimen partition.

Two considerations governed the approach taken. First, TreeSHAP returns exact Shapley values for both boosters, so neither is preferred on grounds of estimator fidelity. Second, and more decisively, a claim about which features drive concrete strength should not rest on a single learner. TreeSHAP was therefore computed for both Optuna-tuned boosters over the same test specimens, and the agreement between their global importance rankings was measured with Spearman's rank correlation coefficient. The XGBoost attributions are presented as the primary analysis in Figs 7 to 9, the LightGBM attributions as a cross-check in Fig. 10, and the agreement statistic is reported in Section 3.5. This turns the transfer of interpretive conclusions from one model to the other into a measured quantity rather than an assumption. The tuned configurations

were used for both because they are the ones produced by the documented optimisation procedure of Section 2.4; the untuned baselines share the same 16-feature input space and differ only in hyperparameters, and the released pipeline allows attributions to be regenerated for any configuration.

3. Results and Discussion

3.1 Exploratory Data Analysis

Pearson correlation analysis of the raw variables (Fig. 2) shows that cement content ($r = 0.50$) and superplasticiser dosage ($r = 0.37$) have the strongest positive linear associations with compressive strength, while water content ($r = -0.29$) and fine aggregate ($r = -0.17$) are negatively associated. Curing age correlates moderately ($r = 0.33$). The mutual correlation between superplasticiser and water content ($r = -0.66$) reflects their joint optimisation in high-performance mix designs.

Fig. 2 covers the eight raw variables only, since it is computed before the transformations of Section 2.2 are applied. The corresponding correlations for the engineered features are therefore given here. They are water-to-binder ratio -0.62 , total binder content 0.61 , aggregate-to-binder ratio -0.57 , log-transformed age 0.55 , water-to-cement ratio -0.50 , superplasticiser efficiency 0.38 , SCM fraction -0.11 and fine-to-coarse aggregate ratio -0.02 . Five of the eight engineered features therefore show a stronger linear association with strength than the strongest raw predictor, and the age correlation rises from 0.33 to 0.55 under the logarithmic transformation, which supports the theoretical motivation for that feature.

Fig. 3 shows the evolution of compressive strength with curing age, stratified by cement content tertile. All three groups follow the characteristic logarithmic gain trajectory, and at extended curing periods the high-cement mixes (above 350 kg/m^3) reach a mean strength of 58.3 MPa against 31.7 MPa for the low-cement mixes (below 255 kg/m^3). The pattern agrees with the hydration kinetics framework summarised by Neville (2011) and gives empirical justification for the Log_Age and WB_Ratio features introduced in Section 2.2.

3.2 Cross-Validation Benchmark

Among the nine untuned baselines, XGBoost reached the highest mean CV R^2 of 0.9331 (SD 0.0291), followed by LightGBM (0.9325 , SD 0.0267) and extra trees (0.9119 , SD 0.0343). Every tree ensemble, and SVR as well, exceeded the 0.90 threshold, while the linear regressors plateaued near

0.83 , which confirms the inherently nonlinear dependence of strength on mix proportions. ElasticNet performed worst (0.7780 , SD 0.0270), because the combined L1 and L2 penalties shrink the small number of highly predictive features. Fig. 4 shows the fold-level distributions behind these means: the boxplots of R^2 and the corresponding RMSE distributions place the four boosted configurations clearly above the remaining models, with a comparable spread across folds.

The Wilcoxon tests separate the models into two clear groups. The four gradient-boosted configurations, tuned and untuned XGBoost and LightGBM, are statistically indistinguishable from one another on the paired fold scores ($p = 0.193$ to 0.695). Every other model differs significantly from the reference at the resolution the test allows ($p = 0.002$), including extra trees and random forest, whose CV means sit 0.021 and 0.024 below the reference. The practical reading is that the choice within the boosted family is not resolved by this dataset, whereas the gap between the boosted family and everything else is real.

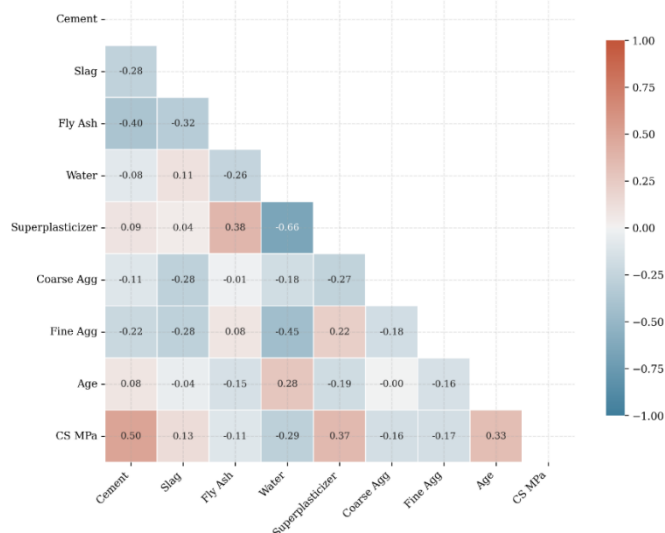


Fig. 2 Pearson correlation matrix for the eight raw mix-design variables and the compressive strength target.

Table 2. Ten-fold cross-validation and test-set performance of all model configurations, sorted by test R^2 . Bold indicates the best value in each metric column.

Model	CV R^2 (mean $\pm$ SD)	Test R^2	RMSE (MPa)	MAE (MPa)	MAPE (%)	Wilcoxon p	Sig.
XGBoost*	0.9331 ± 0.0291	0.9567	3.46	2.35	8.03	—	reference
LightGBM	0.9325 ± 0.0267	0.9561	3.48	2.41	7.79	0.695	ns
LightGBM (Optuna)	0.9372 $\pm$ 0.0286	0.9522	3.63	2.48	8.18	0.193	ns
XGBoost (Optuna)	0.9358 ± 0.0261	0.9514	3.67	2.60	9.11	0.375	ns
Extra Trees	0.9119 ± 0.0343	0.9421	4.00	2.63	9.79	0.002	*
Random Forest	0.9087 ± 0.0285	0.9360	4.21	3.08	11.50	0.002	*
SVR	0.9040 ± 0.0263	0.9120	4.93	3.16	9.65	0.002	*
Ridge	0.8280 ± 0.0227	0.8388	6.67	5.04	20.47	0.002	*
Linear Reg.	0.8282 ± 0.0219	0.8387	6.68	5.07	20.47	0.002	*
Lasso	0.8071 ± 0.0315	0.8067	7.31	5.50	21.17	0.002	*
ElasticNet	0.7780 ± 0.0270	0.7810	7.78	6.07	22.46	0.002	*

* Reference model for the statistical tests, selected as the configuration with the highest test R^2 . Wilcoxon signed-rank test, two-sided, on the paired ten-fold CV R^2 scores of the reference against each other configuration. ns = not significant at $\alpha = 0.05$; an asterisk marks $p < 0.05$. With ten paired folds the smallest attainable two-sided p-value is 0.00195 . CV = cross-validation; SD = standard deviation; RMSE = root mean squared error; MAE = mean absolute error; MAPE = mean absolute percentage error.

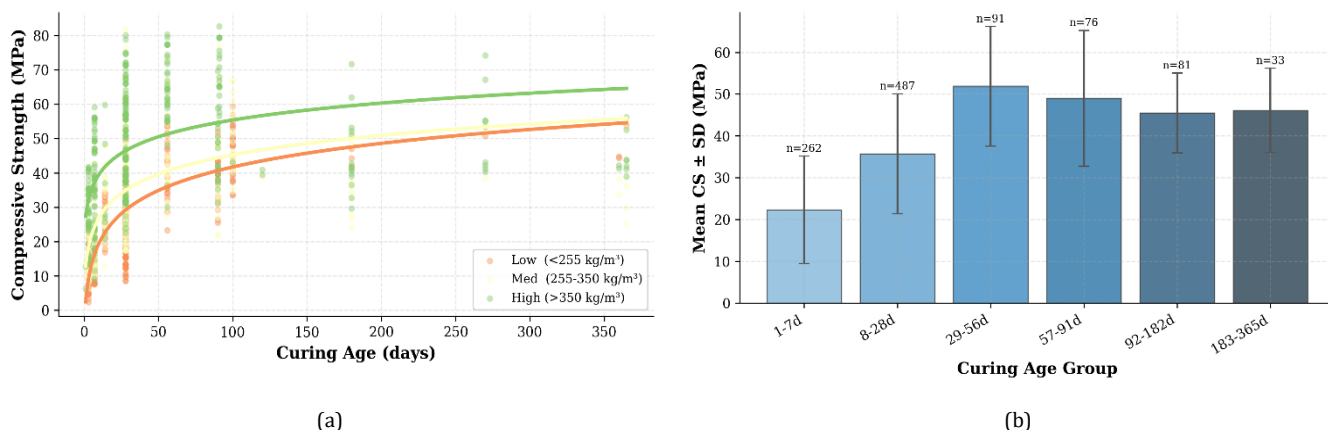


Fig. 3 Compressive strength evolution with curing age: (a) scatter plot with logarithmic fits stratified by cement content tertile; (b) mean compressive strength ± 1 SD by age group, with sample sizes annotated above the bars.

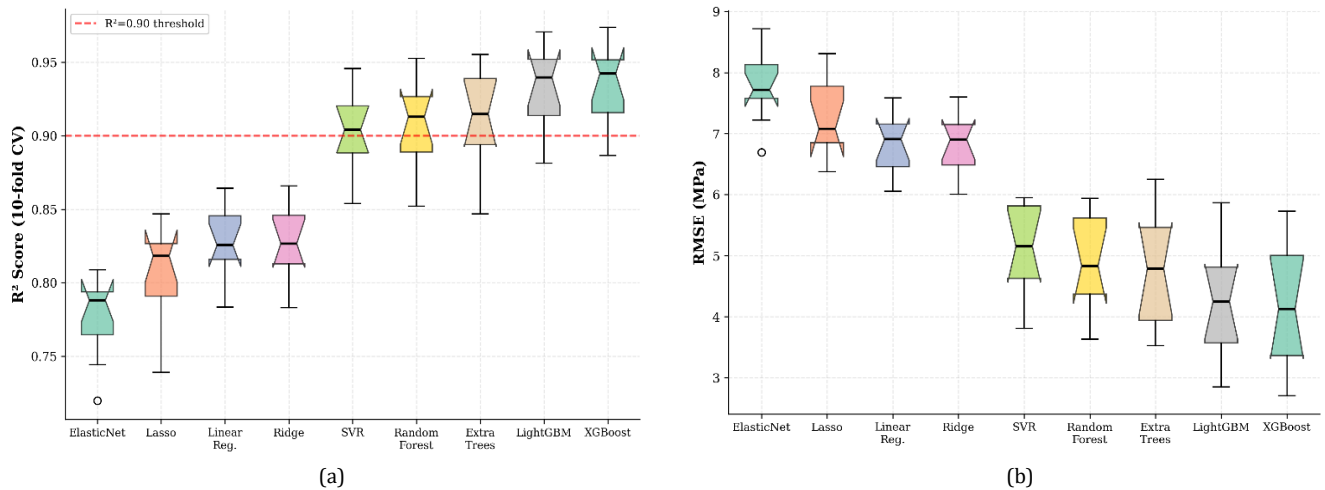


Fig. 4 Ten-fold cross-validation performance on the training partition ($n = 824$): (a) notched boxplots of R^2 scores across folds, sorted by median; (b) the corresponding RMSE distributions.

3.3 Bayesian Hyperparameter Optimisation

Over 100 trials, Optuna raised the inner five-fold CV objective of XGBoost to 0.9356 at trial 72 and that of LightGBM to 0.9354 at trial 75. Re-evaluated under the common ten-fold protocol of Table 2, the tuned configurations score 0.9358 and 0.9372 respectively, above their untuned counterparts by 0.0027 and 0.0047. The convergence plots in Fig. 5 show rapid improvement within the first 20 to 30 trials followed by incremental refinement. The best XGBoost configuration favoured shallow trees ($\text{max_depth} = 3$) in a large ensemble ($\text{n_estimators} = 1,110$) with a moderate learning rate (0.064), aggressive column subsampling ($\text{colsample_bytree} = 0.530$) and light regularisation ($\text{reg_lambda} = 0.141$, $\text{reg_alpha} = 4.6 \times 10^{-4}$), which is consistent with the usual bias-variance trade-off for gradient boosting on datasets of this size.

The held-out results reverse this order. On the test set the tuned XGBoost fell from 0.9567 to 0.9514 and the tuned LightGBM from 0.9561 to 0.9522, so for both boosters the search bought cross-validated accuracy at the cost of held-out accuracy. Three mechanisms plausibly combine to produce this. First, selecting the best of 100 trials by inner CV score is subject to a winner's curse: with $n = 824$ and five inner folds, the standard error of a fold-mean R^2 is large enough that the top-ranked trial is favoured partly by noise, and its advantage does not transfer to unseen data (Turner et al., 2021). A nine-dimensional search space offers many opportunities for such adaptive overfitting. Second, the test set contains only 206 specimens, so its R^2 estimates carry sampling variability of the same order as the observed gaps of 0.004 to 0.005; the reversal is consistent with configurations of nearly equal true accuracy whose ranking is unstable across splits. Third, the untuned settings already sit close to the region the sampler converges on, so the remaining search space contains little genuine improvement to find, and what the sampler does find is largely fold specific.

We report this negative result in full, rather than presenting only the strongest configuration, because it carries a practical message that is easy to state as a rule. Report the tuned and the untuned configuration on the same held-out partition and adopt the tuned model only if it wins there. Where Bayesian optimisation raises the cross-validated score but lowers the held-out score, as it did for both boosters here, revert to the best-performing untuned baseline: the apparent gain belongs to the selection procedure and not to the model. A second and weaker signal is worth watching alongside it. If a paired Wilcoxon test cannot separate the tuned from the untuned configuration on the fold scores, as it could not here ($p = 0.375$ for XGBoost and $p = 0.193$ for LightGBM), then the tuning has produced no measurable improvement even inside CV, and the simpler configuration should be preferred on grounds of parsimony, training cost and ease of documentation in a quality-control setting. An engineer working with a few hundred to a few thousand specimens should budget for this check rather than assume that a longer search yields a better field model.

3.4 Test-Set Performance

On the held-out test set ($n = 206$), XGBoost with the baseline settings of Section 2.3 achieved the best overall result: $R^2 = 0.9567$, RMSE = 3.46 MPa, MAE = 2.35 MPa and MAPE = 8.03%. Baseline LightGBM followed at $R^2 = 0.9561$ with the lowest MAPE of the whole benchmark (7.79%), and the two tuned boosters came next at 0.9522 and 0.9514. The four are separated by 0.0053 in R^2 and by 0.21 MPa in RMSE, which is well inside

the sampling variability of a 206-specimen partition, and the statistical tests of Section 3.2 do not distinguish them. Any of the four is a defensible operational choice, and the ranking among them should not be over-read.

The predicted against actual scatter (Fig. 6) sits tightly on the identity line across the full-strength range, from 2.33 to 82.60 MPa, with no systematic bias. Scatter increases above about 65 MPa, which is attributable to the underrepresentation of ultra-high-strength specimens; fewer than 5% of samples exceed 70 MPa. Extra trees and random forest reach 0.9421 and 0.9360, SVR 0.9120, and the linear baselines converge to about 0.83, the ceiling imposed by their inability to represent nonlinear interactions. Expressed in engineering terms, the gap between the boosted family and the linear family is roughly 3.2 MPa of RMSE, which is the difference between an error band that is comfortably inside a typical design margin and one that is not.

3.5 SHAP Explainability Analysis

SHAP analysis of the tuned XGBoost model shows that the water-to-binder ratio (mean $|\phi| = 4.95$ MPa) and curing age (4.79 MPa) are the two most influential predictors globally (Fig. 7), together accounting for 38% of the total mean absolute SHAP contribution across all 16 features; adding log-transformed age (3.06 MPa) brings the leading three to 51%. This agrees directly with established concrete science: the w/b ratio governs capillary porosity through Abrams' law (Abrams, 1918), and strength gain is time-dependent through hydration (Neville, 2011). The beeswarm plot (Fig. 8) shows that high curing ages produce large positive SHAP contributions while high w/b values drive strongly negative ones, confirming monotonically opposed relationships between these predictors and strength.

Water content (1.96 MPa), the water-to-cement ratio (1.86 MPa) and total binder content (1.45 MPa) form a secondary tier, and the four raw aggregate and admixture variables sit at the bottom of the ranking with mean $|\phi|$ below 0.65 MPa. Three of the five highest-ranked features, and four of the highest six, are engineered rather than raw, which supports the physics-informed strategy of Section 2.2 and the observation by Jiang et al. (2022) that domain-knowledge features outperform raw mix-design variables in SHAP contribution magnitude.

The dependence plots (Fig. 9) show the expected nonlinear patterns. SHAP values for Age rise steeply between 1 and about 60 days, corresponding to the primary hydration phase, and then plateau. SHAP values for WB_Ratio fall sharply once the ratio exceeds about 0.45, consistent with Abrams' relation, and the colouring by cement content shows that at a given w/b the higher-cement mixes receive the larger positive contributions. The Log_Age panel recovers the same age effect on a linearised axis, and the Water panel shows a monotone decline with a break near 180 kg/m^3 .

The cross-check on the tuned LightGBM model, described in Section 2.6, gives a Spearman rank correlation of $\rho = 0.829$ ($p < 0.001$) between the two global importance vectors. More informative than the coefficient itself is the composition of the leading group: the two models select exactly the same six features as most important, namely the water-to-binder ratio, curing age, log-transformed age, water content, the water-to-cement ratio and total binder content, and they differ only in the internal ordering of that set. LightGBM places curing age first (5.88 MPa) ahead of the water-to-binder ratio (5.07 MPa), reversing the XGBoost order, and promotes total binder content from sixth to third while demoting log-transformed age from third to sixth. Four of those six features are engineered in both rankings.

These swaps are what collinearity predicts. Age and Log_Age are deterministic transforms of one another, and WC_Ratio, WB_Ratio and Binder are algebraically linked through the binder definition, so the credit that TreeSHAP assigns within each of these groups depends on which member a given tree happens to split on first. The disagreements are concentrated in the low-importance tail, where the aggregate-to-binder

ratio falls from rank 8 in XGBoost to rank 15 in LightGBM and coarse aggregate rises from 16 to 12; at mean $|\phi|$ below 1 MPa these ranks are not stable and carry little interpretive weight. The conclusions drawn in this section, that hydration time and the water-to-binder ratio dominate and that engineered features occupy most of the leading positions, held under both models. Fig. 10 shows the two importance profiles side by side.

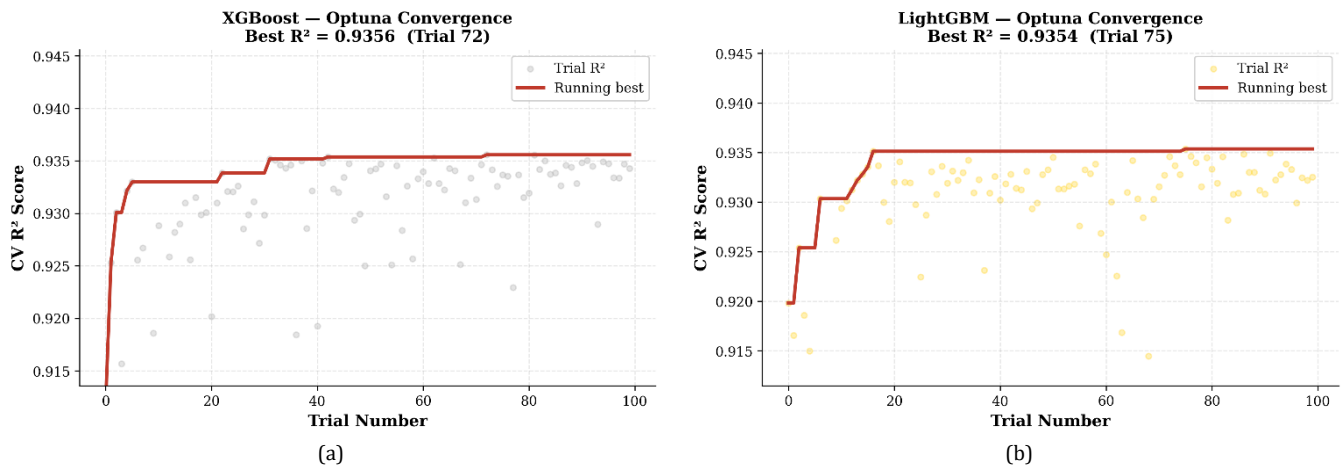


Fig. 5 Bayesian hyperparameter optimisation convergence for (a) XGBoost and (b) LightGBM over 100 Optuna trials on the training partition.

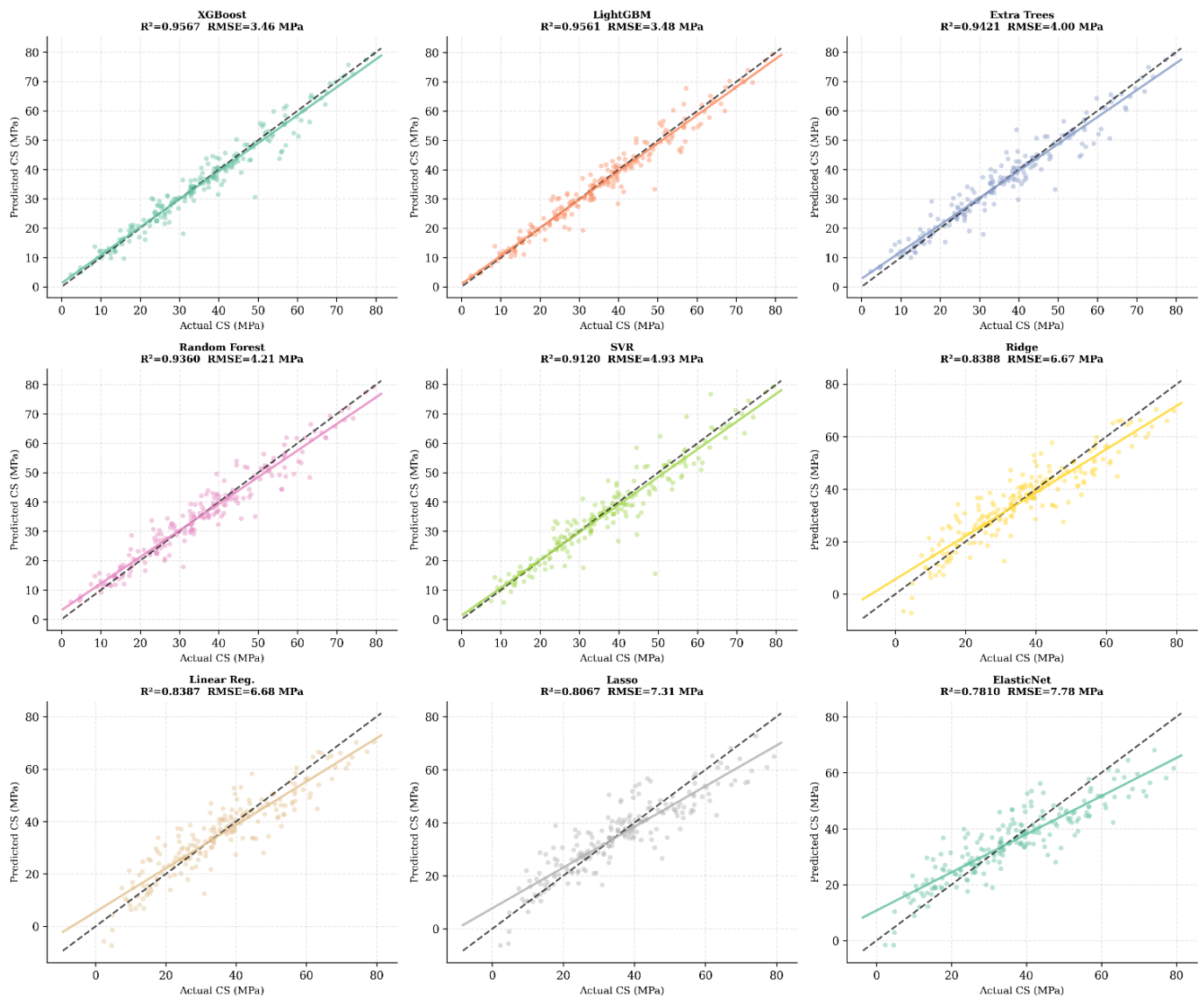


Fig. 6 Predicted against actual compressive strength for all nine baseline models on the held-out test set ($n = 206$).

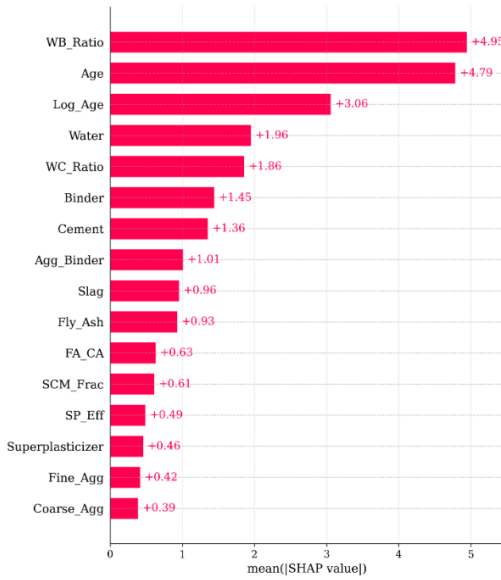


Fig. 7 SHAP global feature importance: mean absolute SHAP values (MPa) for all 16 features, tuned XGBoost model, test set, $n = 206$

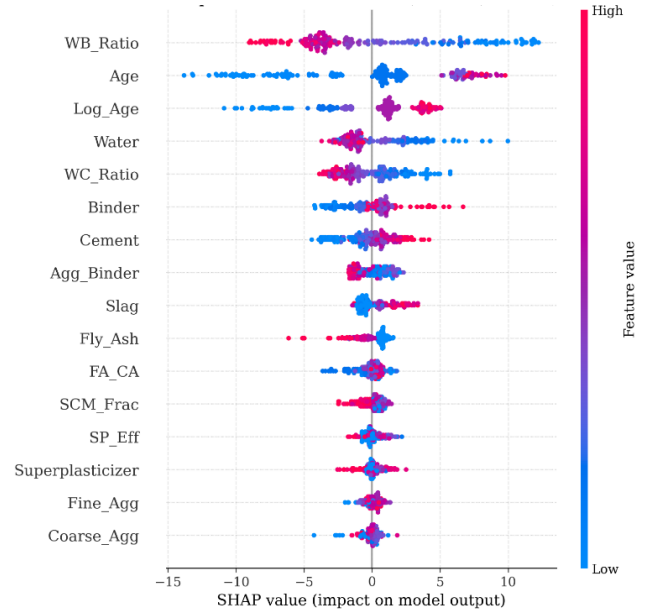


Fig. 8 SHAP beeswarm plot for the tuned XGBoost model

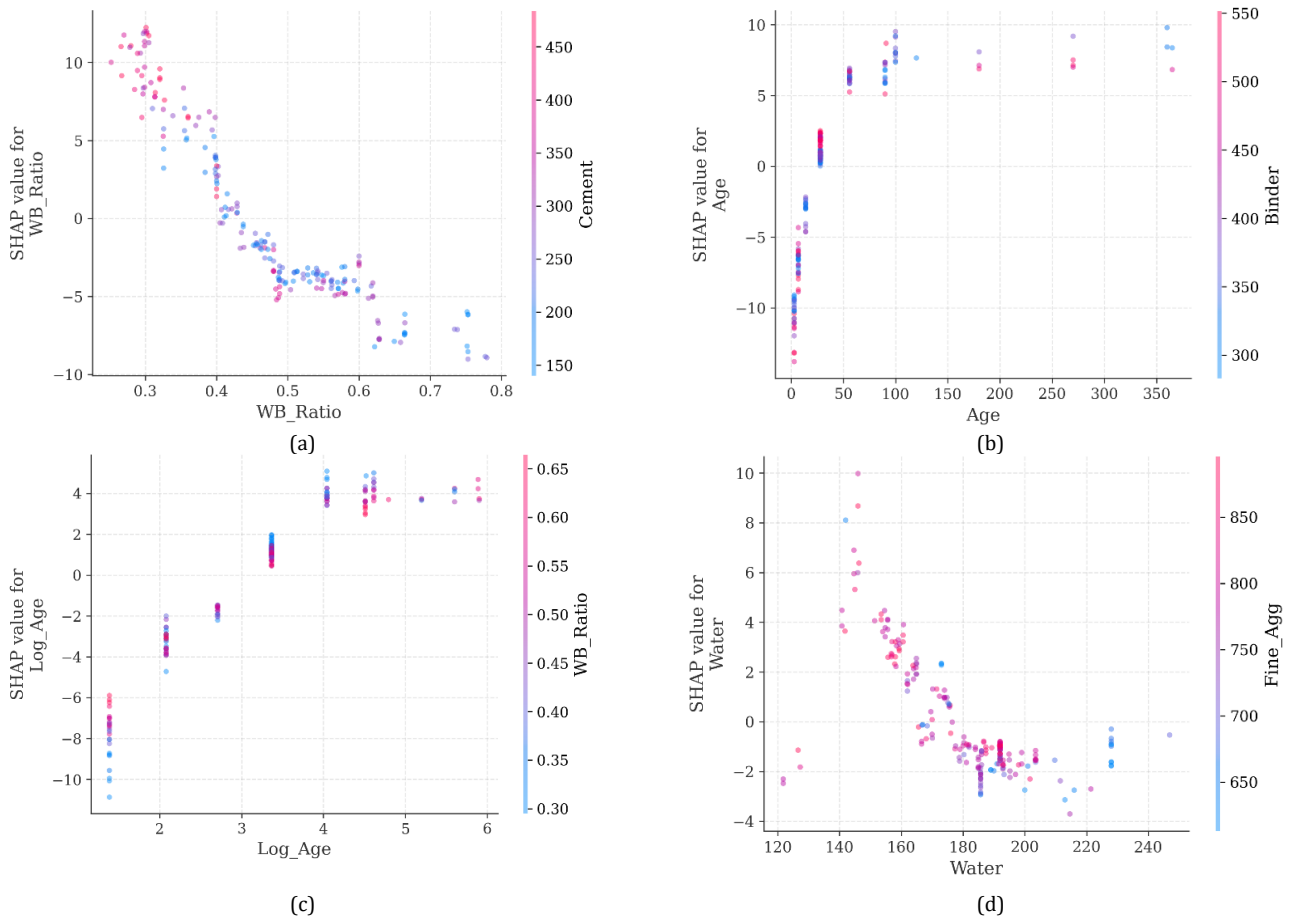


Fig. 9 SHAP dependence plots for the four most influential features of the tuned XGBoost model: (a) WB_Ratio; (b) Age; (c) Log_Age; (d) Water. Colour encodes the automatically selected interaction feature.

3.1 Residual Analysis

Residual diagnostics for the best held-out model appear on Fig. 11. The mean residual is 0.39 MPa, so the model under-predicts strength very slightly on average, an offset of about 1% of the mean observed strength and roughly one ninth of the RMSE. The residual standard deviation is 3.44 MPa, giving a $\pm 2\sigma$ band of about ± 6.9 MPa.

The residual distribution is unimodal and centered close to zero, but it is not normal: a Kolmogorov-Smirnov (KS) test on the standardised residuals rejects normality (KS statistic 0.126, $p = 0.003$). The quantile-quantile (Q-Q) plot in panel B makes the source visible. The central body of the distribution tracks the theoretical quantiles closely, with a correlation of 0.952, and the departure comes almost entirely from a short right tail of specimens under-predicted by 10 to 19 MPa. We report this

rather than describing the residuals as approximately normal, because it has a consequence for practice: prediction intervals built on a Gaussian assumption will be slightly optimistic in the upper tail, and an engineer using the model for acceptance decisions on high-strength mixes should either widen the interval or derive it empirically from the residual quantiles.

The residuals against predicted values in panel C show no systematic heteroscedasticity; the scatter band stays approximately constant across the prediction range, so the constant-variance assumption needed for interval estimation is otherwise satisfied. No curvature appears in the residuals against actual values in panel D, indicating that the dominant nonlinearities have been captured and that the remaining error is close to unstructured.

Figure 10: The two tuned-boost global SHAP importance profiles side by side (test set, $n = 206$; Spearman $\rho = 0.829$)

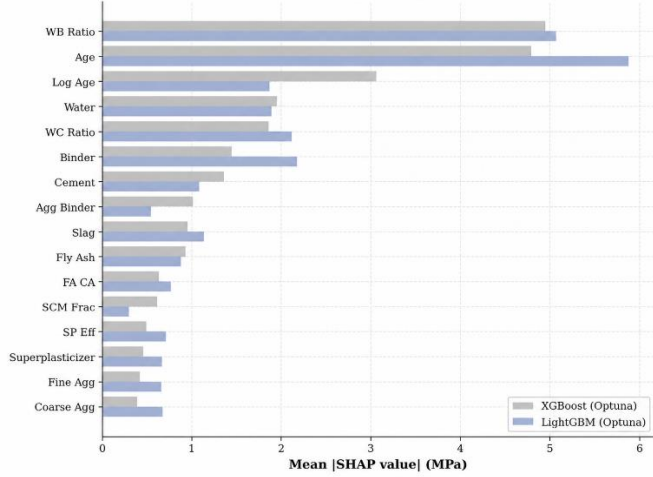
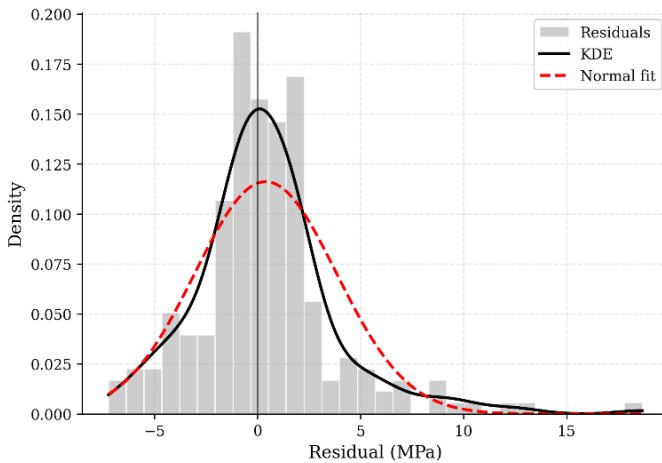


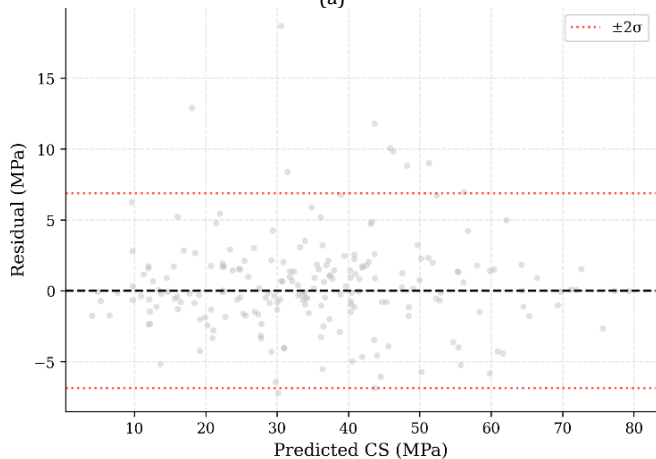
Fig.10 Two tuned-boost global SHAP importance profiles side by side (test set, $n = 206$; Spearman $\rho = 0.829$)

3.2 Learning Curves

Learning curves computed by five-fold CV on the training partition (Fig. 12) show the validation score rising steeply from $R^2 \approx 0.77$ at 50 training samples, crossing 0.90 at roughly 330 samples, and then flattening to about 0.93 as the training set grows towards its full size. The training curve remains close to 1.0 throughout, and the persistent gap between the two curves is characteristic of moderately complex gradient-boosted trees under the regularisation imposed by the tuned hyperparameters. The converging trajectory at large n suggests that 1,030 specimens are broadly sufficient for this 16-feature space, although the curve is not fully saturated, and larger datasets with better coverage of ultra-high-strength and blended compositions would likely bring further improvement (Nunez et al., 2021).



(a)



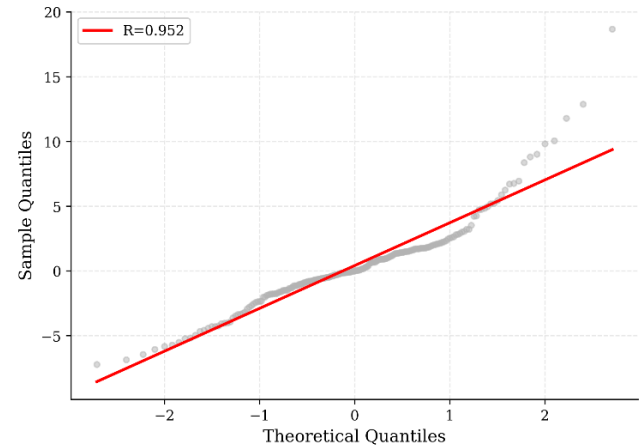
(c)

3.3 Comparison with Prior Literature, and Limitations of the Benchmark

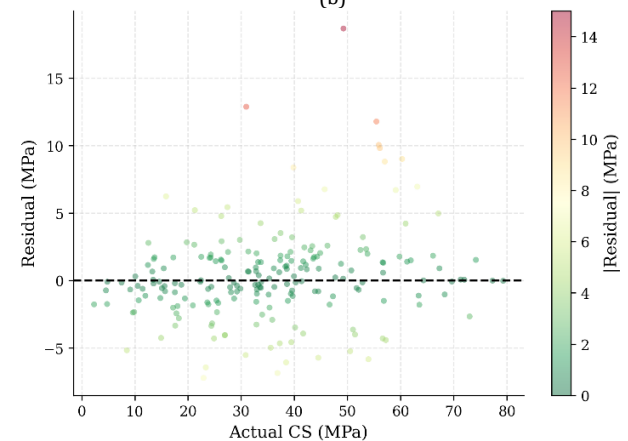
Table 3 places the present results alongside published ML studies of compressive strength prediction on the UCI benchmark or comparable datasets. Accuracy figures alone are a weak basis for comparison, because the listed studies differ in dataset size, split protocol and validation rigour, so Table 3 also records five methodological dimensions: the hyperparameter optimisation (HPO) strategy, whether physics-informed features were used, whether explainability analysis was performed, whether model comparisons were statistically tested, and whether code was released for reproduction.

The feature engineering column is the one most directly relevant to the contribution claimed here. Of the five prior studies listed, only Jiang et al. (2022) encoded any domain relation in the inputs, adding three quantities, the water-to-cement ratio, total binder content and SCM fraction, to the raw mix variables and reporting a rise in XGBoost R^2 from 0.931 to 0.947 on 2,100 specimens. The remaining four fed raw mix-design variables to the learner. The present study extends that line from three engineered quantities to eight, covering Abrams' law, the blended-cement generalisation, hydration kinetics, SCM replacement level, the aggregate-to-paste balance, admixture efficiency and aggregate grading, and it is the only entry that combines physics-informed inputs with SHAP attribution, formal significance testing of every pairwise comparison, and released code. That combination, rather than the headline accuracy, is what the paper contributes.

Two of the listed studies report higher accuracy than the classical references on substantially larger datasets (Jiang et al., 2022, with 2,100 specimens; Nunez et al., 2021, with 3,450), where more training data alone is expected to lift performance, and no two of the listed studies used the same split. The R^2 column should therefore be read as indicative rather than directly comparable. The configurations reported here sit at the upper end of the values published for the 1,030-specimen benchmark, ahead of the classical references (Yeh, 1998; Chou et al., 2011), but that figure comes from one stratified split, and Section 3.4 has already shown that differences of a few thousandths in R^2 on a 206-specimen test set are not meaningful. We make no claim to an accuracy record.



(b)



(d)

Fig. 11 Residual analysis for the best held-out model (XGBoost, $R^2 = 0.9567$, $RMSE = 3.46$ MPa): (a) residual distribution with kernel density and normal fit; (b) normal Q-Q plot; (c) residuals against predicted values with $\pm 2\sigma$ bands; (d) residuals against actual values coloured by absolute residual.

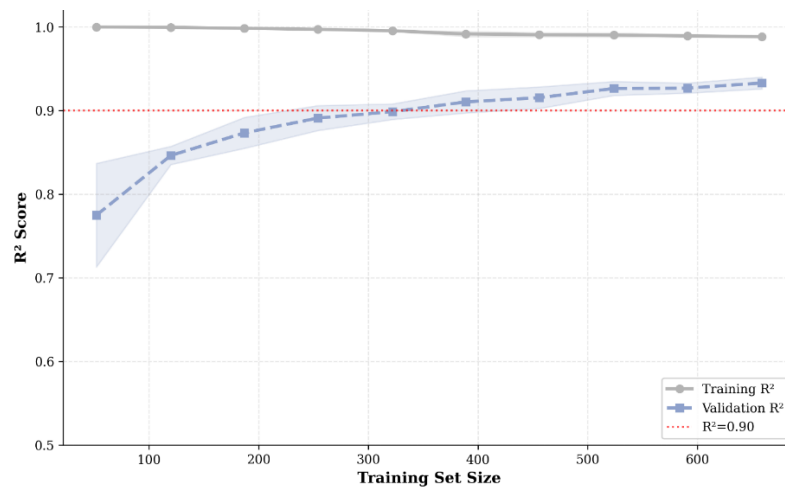


Fig. 12 Learning curves for the tuned XGBoost configuration from five-fold cross-validation on the training partition ($n = 824$). Shaded regions show ± 1 SD across folds; the dashed line marks $R^2 = 0.90$.

Table 3. Published ML models for compressive strength prediction on the UCI Concrete dataset or equivalent benchmarks, compared to accuracy and five methodological dimensions.

Study	Model	n	R^2	RMSE	HPO	Physics-informed features	SHAP	Stat. test	Code
Yeh (1998)	ANN	1,030	0.920	5.58	No	No	No	No	No
Chou et al. (2011)	Random forest	1,030	0.931	4.44	Manual	No	No	No	No
Jiang et al. (2022)	XGBoost	2,100	0.947	4.01	Grid	Yes (3)	Partial	No	No
Shahmansouri et al. (2021)	GBM	740	0.936	4.82	Random	No	Yes	No	No
Nunez et al. (2021)	LightGBM	3,450	0.951	3.87	Bayesian	No	No	No	No
Present study	XGBoost, untuned	1,030	0.9567	3.46	None	Yes (8)	Yes	Yes	Yes
Present study	LightGBM, tuned	1,030	0.9522	3.63	Bayesian	Yes (8)	Yes	Yes	Yes

RMSE in MPa. HPO = hyperparameter optimisation strategy. ANN = artificial neural network, two hidden layers in Yeh (1998). GBM = gradient boosting machine. Physics-informed features give the number of domain-derived inputs added to the raw mix variables; Jiang et al. (2022) added the water-to-cement ratio, total binder content and the SCM fraction. Partial = feature importance reported without Shapley attribution. Stat. test = formal significance testing of model comparisons. Code = analysis code released. Present-study values are on the fixed 80/20 stratified split described in Section 2.5. R^2 values across studies are not directly comparable owing to differing datasets and splits.

Age and scope of the benchmark dataset. The UCI Concrete Compressive Strength dataset is a seminal benchmark, and it is the right basis for a comparability study, precisely because so many published models have been fitted to it. It is also more than twenty-five years old, and its 1,030 specimens describe the ordinary Portland cement systems, slag and fly ash replacements, and polycarboxylate-era superplasticiser dosages that were current when Yeh (1998) compiled it. Several classes of material that matter commercially fall outside it. Ultra-high-performance concrete (UHPC) operates at water-to-binder ratios near 0.20 with silica fumes, quartz flour and steel fibres, and derives much of its strength from particle packing and fibre bridging rather than from paste chemistry alone (Richard and Cheyrezy, 1995). Carbon-negative and alkali-activated binders replace clinker chemistry entirely, so the hydration model that underlies the age transformation used here does not apply to them in the same form (Scrivener et al., 2018). Multi-component SCM systems combining calcined clay, limestone and slag develop strength through reaction pathways that a single aggregated binder term cannot distinguish (Juenger and Siddique, 2015). None of these appear in the dataset, and the maximum recorded strength of 82.60 MPa lies below the range where UHPC begins.

Do the engineered features generalise? Our position is that the features are divided into two groups with different prospects. The ratio features that express mass balance, the water-to-binder ratio, the aggregate-to-binder ratio, the SCM fraction, superplasticiser efficiency and the fine-to-coarse aggregate ratio, are definitional rather than empirical: they remain computable and physically meaningful for any mixture for which the constituent masses are known, including UHPC and alkali-activated systems, and we expect them to transfer as feature definitions. What need not transfer are the fitted relationships between those features and strength. Abrams' law flattens at very low water-to-binder ratios where free water is insufficient for complete hydration and strength becomes packing-limited, so a model trained on the range 0.25 to 0.90 seen here should not be extrapolated to 0.20. The logarithmic age transformation encodes Portland cement hydration kinetics specifically and would need re-derivation for alkali-activated binders, whose strength-time curves have a different shape. The aggregated binder term treats slag and fly ash as interchangeable with cement on a mass basis, which is a reasonable

approximation for the two SCMs present in this dataset and a poor one for calcined clay or limestone blends.

The honest scope statement is therefore that the quantitative results of this paper, every R^2 , RMSE and SHAP attribution reported above, are strictly limited to the composition space of the UCI dataset: Portland cement systems with slag and fly ash, water-to-binder ratios above roughly 0.25, and strengths up to about 83 MPa. The feature definitions and the methodological framework, the shared split, the paired statistical testing, the tuned-against-untuned comparison and the multi-model SHAP agreement check, are portable to modern material classes, but a model for those classes would have to be refitted on data that contains them. Extending the benchmark in this direction, rather than pushing the R^2 on this dataset higher, is where we see the useful next step.

4. Data Availability and Reproducibility

The UCI Concrete Compressive Strength dataset used in this study is publicly available at the UCI Machine Learning Repository (<https://doi.org/10.24432/C5PK67>) and is redistributed with the project code for convenience. The complete analysis pipeline, covering data loading, feature engineering, model training, Optuna optimisation, statistical testing, SHAP computation and the generation of every figure and table, is released under an open-source license in the project repository. All stochastic components, the train-test split, the CV fold assignment, model initialisation and the Optuna sampler, are controlled by a single fixed seed (seed = 42), so every number reported in this paper can be regenerated exactly with one command. The two protocol choices that most affect the results, stratification of the split and restriction of cross-validation to the training partition, are exposed as explicit flags at the top of the pipeline so that a reader can reproduce either setting or see the difference for themselves. The software environment is pinned in a requirements file; the principal versions are Python 3.11, scikit-learn 1.4, XGBoost 2.0 (Chen and Guestrin, 2023), LightGBM 4.3 (Microsoft Corporation, 2024), Optuna 3.5 and SHAP 0.44. Readers who wish to verify the results, or to extend the benchmark with additional models, can therefore do so without re-implementing any part of the study.

5. Conclusions

This study sets out to address the methodological gaps identified in the Introduction: the lack of statistically validated cross-model benchmarking, the neglect of physics-informed inputs, and limited attention to interpretability and reproducibility in ML studies of concrete strength. Four conclusions follow.

Benchmark outcome - Under one identical stratified split and one shared CV protocol, XGBoost achieved the best held-out performance (test $R^2 = 0.9567$, RMSE = 3.46 MPa, MAE = 2.35 MPa). The four gradient-

boosted configurations could not be separated from one another statistically, while all four differed significantly from the kernel, bagging and linear baselines (Wilcoxon, $p = 0.002$). Linear models plateau roughly twelve R^2 points below the boosted ensembles, confirming that strength is an inherently nonlinear function of mix design.

Physics-informed features work - Engineered features occupy three of the five and four of the six highest positions in global SHAP importance, led by the water-to-binder ratio, log-transformed age and total binder content, and the dependence plots recover the logarithmic strength gain with age and the sharp decline above a water-to-binder ratio of about 0.45 that concrete theory predicts. Both tuned boosters return the same six leading features (Spearman $\rho = 0.829$), so the finding is not an artefact of one learner.

Tuning is not free - Bayesian search improved the cross-validated score of both boosters and reduced the held-out score of both. On datasets of this size, a tuned model should be re-validated against its untuned baseline on held-out data, and the untuned baseline should be preferred when the tuned configuration fails that check.

A reproducible reference point - Residuals of the best model are centred close to zero with no heteroscedasticity, though a short right tail causes a formal departure from normality that matters for interval estimation on high-strength mixes. Because the entire pipeline is released with a fixed seed and explicit protocol flags, the benchmark can serve as a verified starting point against which future models are compared on identical footing.

The quantitative findings are bounded by the composition space of the UCI dataset, as set out in Section 3.8. Future work will extend the framework to fibre-reinforced concrete, geopolymers and alkali-activated binders and UHPC, where the feature definitions carry over but the fitted relations must be re-derived; incorporate microstructural descriptors such as mercury intrusion porosimetry and X-ray diffraction through multi-modal feature fusion; develop a physics-informed neural network that enforces Abrams' law as a constraint in the loss function; explore federated learning for cross-laboratory dataset aggregation without data sharing; and deploy the model as a real-time digital twin for structural health monitoring and adaptive mix design.

References

- Abrams, D.A. (1918) Design of Concrete Mixtures. Structural Materials Research Laboratory Bulletin 1. Chicago, IL: Lewis Institute.
- Akiba, T., Sano, S., Yanase, T., Ohta, T. and Koyama, M. (2019) 'Optuna: a next-generation hyperparameter optimization framework', Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Anchorage, AK, USA, 4-8 August, pp. 2623-2631. <https://doi.org/10.1145/3292500.3330701>
- ASTM International (2021) ASTM C39/C39M-21: Standard Test Method for Compressive Strength of Cylindrical Concrete Specimens. West Conshohocken, PA: ASTM International.
- Ben Chaabene, W., Flah, M. and Nehdi, M.L. (2020) 'Machine learning prediction of mechanical properties of concrete: critical review', Construction and Building Materials, 260, 119889. Available at: <https://doi.org/10.1016/j.conbuildmat.2020.119889>
- Chen, T. and Guestrin, C. (2016) 'XGBoost: a scalable tree boosting system', Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13-17 August, pp. 785-794. <https://doi.org/10.1145/2939672.2939785>
- Chen, T. and Guestrin, C. (2023) XGBoost Documentation, Version 2.0. Available at: <https://xgboost.readthedocs.io> (Accessed: 10 April 2025).
- Chou, J.S., Chiu, C.K., Farfoura, M. and Al-Taharwa, I. (2011) 'Optimizing the prediction accuracy of concrete compressive strength based on a comparison of data-mining techniques', Journal of Computing in Civil Engineering, 25(3), pp. 242-253. Available at: [https://doi.org/10.1061/\(ASCE\)CP.1943-5487.0000088](https://doi.org/10.1061/(ASCE)CP.1943-5487.0000088)
- Chou, J.S., Tsai, C.F., Pham, A.D. and Lu, Y.H. (2014) 'Machine learning in concrete strength simulations: multi-nation data analytics', Construction and Building Materials, 73, pp. 771-780. Available at: <https://doi.org/10.1016/j.conbuildmat.2014.09.054>
- Feng, D.C., Liu, Z.T., Wang, X.D., Chen, Y., Chang, J.Q., Wei, D.F. and Jiang, Z.M. (2020) 'Machine learning-based compressive strength prediction for concrete: an adaptive boosting approach', Construction and Building Materials, 230, 117000. Available at: <https://doi.org/10.1016/j.conbuildmat.2019.117000>
- Jiang, Y. et al. (2022) 'A gradient boosting regressor with physics-informed feature engineering for concrete compressive strength prediction', Materials, 15(24), 8953. Available at: <https://doi.org/10.3390/ma15248953>
- Juenger, M.C.G. and Siddique, R. (2015) 'Recent advances in understanding the role of supplementary cementitious materials in concrete', Cement and Concrete Research, 78, pp. 71-80. Available at: <https://doi.org/10.1016/j.cemconres.2015.03.018>
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.Y. (2017) 'LightGBM: a highly efficient gradient boosting decision tree', Advances in Neural Information Processing Systems, 30, pp. 3146-3154.
- Lundberg, S.M. and Lee, S.I. (2017) 'A unified approach to interpreting model predictions', Advances in Neural Information Processing Systems, 30, pp. 4765-4774.
- Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N. and Lee, S.I. (2020) 'From local explanations to global understanding with explainable AI for trees', Nature Machine Intelligence, 2(1), pp. 56-67. Available at: <https://doi.org/10.1038/s42256-019-0138-9>
- Mehta, P.K. and Monteiro, P.J.M. (2014) Concrete: Microstructure, Properties, and Materials, 4th edn. New York, NY: McGraw-Hill.
- Microsoft Corporation (2024) LightGBM: Light Gradient Boosting Machine, Version 4.3. Available at: <https://lightgbm.readthedocs.io> (Accessed: 10 April 2025).
- Neville, A.M. (2011) Properties of Concrete, 5th edn. Harlow: Pearson Education.
- Nunez, I. et al. (2021) 'Estimating compressive strength of modern concrete mixtures using computational intelligence: a systematic review', Construction and Building Materials, 310, pp. 125-148. Available at: <https://doi.org/10.1016/j.conbuildmat.2021.125279>
- Pengyu, C. et al. (2025) 'Machine learning for modeling stress evolution in cement-based materials: a systematic review', Acta Polytechnica Hungarica, 22(12). Available at: <https://doi.org/10.12700/APH.22.12.2025.12.7>
- Powers, T.C. (1958) The Physical Structure and Engineering Properties of Concrete. Research Department Bulletin 90. Skokie, IL: Portland Cement Association.
- Rahman, J., Ahmed, K.S., Khan, N.I., Islam, K. and Mangalathu, S. (2021) 'Data-driven shear strength prediction of steel fiber reinforced concrete beams using machine learning approach', Engineering Structures, 233, 111743. <https://doi.org/10.1016/j.engstruct.2020.111743>
- Richard, P. and Cheyrezy, M. (1995) 'Composition of reactive powder concretes', Cement and Concrete Research, 25(7), pp. 1501-1511. Available at: [https://doi.org/10.1016/0008-8846\(95\)00144-2](https://doi.org/10.1016/0008-8846(95)00144-2)
- Rudin, C. (2019) 'Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead', Nature Machine Intelligence, 1(5), pp. 206-215. Available at: <https://doi.org/10.1038/s42256-019-0048-x>
- Scrivener, K.L., John, V.M. and Gartner, E.M. (2018) 'Eco-efficient cements: potential economically viable solutions for a low-CO₂ cement-based materials industry', Cement and Concrete Research, 114, pp. 2-26. Available at: <https://doi.org/10.1016/j.cemconres.2018.03.015>
- Shahmansouri, A.A. et al. (2021) 'Artificial neural network model to predict the compressive strength of eco-friendly geopolymer concrete incorporating silica fume and natural zeolite', Journal of Cleaner Production, 279, 123697. Available at: <https://doi.org/10.1016/j.jclepro.2020.123697>
- Turner, R., Eriksson, D., McCourt, M., Kiili, J., Laaksonen, E., Xu, Z. and Guyon, I. (2021) 'Bayesian optimization is superior to random search for machine learning hyperparameter tuning: analysis of the black-box optimization challenge 2020', Proceedings of the NeurIPS 2020 Competition Track, PMLR, vol. 133, pp. 3-26.
- Wang, X. et al. (2024) 'Predicting the compressive strength of self-compacting concrete using an optimised gradient boosting machine with SHAP analysis', Case Studies in Construction Materials, 20, e02890. Available at: <https://doi.org/10.1016/j.cscm.2024.e02890>
- Wilcoxon, F. (1945) 'Individual comparisons by ranking methods', Biometrics Bulletin, 1(6), pp. 80-83. Available at: <https://doi.org/10.2307/3001968>
- Yeh, I.C. (1998) 'Modeling of strength of high-performance concrete using artificial neural networks', Cement and Concrete Research, 28(12), pp. 1797-1808. Available at: [https://doi.org/10.1016/S0008-8846\(98\)00165-3](https://doi.org/10.1016/S0008-8846(98)00165-3)
- Zhang, J. et al. (2020) 'Predicting uniaxial compressive strength of oil palm shell concrete using a hybrid artificial intelligence model', Journal of Building Engineering, 30, 101282. Available at: <https://doi.org/10.1016/j.jobe.2020.101282>
- Zou, Y. et al. (2022) 'Evaluation of artificial intelligence methods to estimate the compressive strength of geopolymers', Gels, 8(5), 271. Available at: <https://doi.org/10.3390/gels8050271>

Disclaimer

The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of EJSEI and/or the editor(s). EJSEI 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.