Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| REVIEW ARTICLE

Interpretable Machine Learning for Concrete Strength Prediction with Uncertainty Quantification

Jafar Sadeghi
Independent Researcher, Boca Raton, Florida, USA
Corresponding Author: Jafar Sadeghi, **E-mail**: jafarsadeghi2021@gmail.com

| ABSTRACT

Predicting concrete compressive strength from mix design and curing age is difficult because the relationship between ingredients, age, and strength is nonlinear. This study compares eight regression models on the University of California, Irvine (UCI)/Yeh dataset of 1,030 concrete mixtures and examines three issues that are often overlooked: statistical comparison between models, the value of physics-informed features, and uncertainty for individual predictions. The evaluated models were linear regression, ridge regression, support vector regression, an artificial neural network, random forest, gradient boosting, XGBoost, and a stacking ensemble. XGBoost gave the best held-out performance (coefficient of determination (R^2) = 0.931, root mean squared error (RMSE) = 4.22 MPa, and mean absolute error (MAE) = 2.79 MPa) and the highest 10-fold cross-validated R^2 (0.944 ± 0.019). Paired tests across cross-validation folds showed a consistent advantage over random forest and gradient boosting ($p < 0.0001$ for both comparisons), although the folds are not fully independent because their training data overlap. An ablation study tested four physics-informed features: water-to-binder ratio, total binder content, aggregate-to-binder ratio, and log-transformed age. These features did not produce a significant improvement in predictive accuracy (paired t-test, $p = 0.74$), suggesting that gradient-boosted trees can learn similar relationships from the raw variables. However, the engineered features remained useful for interpretation. SHapley Additive exPlanations (SHAP) analysis ranked curing age and water-to-binder ratio as the two most important predictors, which agrees with established concrete behavior. Split conformal prediction was also used to estimate uncertainty for individual mixtures. A single 60/20/20 train/calibration/test split produced 91.7% empirical coverage for a nominal 95% interval. Across 100 repeated random splits, mean coverage was 95.06% ± 2.15%, showing that the method was well calibrated on average. Residual analysis showed larger uncertainty in the 40–60 MPa strength range. The results support XGBoost as an accurate and interpretable model for this benchmark while also showing that external validation on an independent dataset is an important next step.

| KEYWORDS

Concrete compressive strength; machine learning; XGBoost; SHAP; conformal prediction; explainable AI

| ARTICLE INFORMATION

ACCEPTED: 15 August 2026 **PUBLISHED:** 10 September 2026 **DOI:** 10.32996/jcsts.2026.8.9.5

1. Introduction

Concrete is the most widely used construction material in the world, and compressive strength is its main quality-control measure. In standard practice, strength is measured at a specified age, most commonly 28 days, by crushing concrete cubes or cylinders. Earlier and later ages are also important for decisions such as formwork removal, prestressing, and long-term performance assessment. This testing method is reliable, but it is slow and costly, and the result is only available after the concrete has been cast and cured. Because strength depends on a nonlinear combination of cement, water, aggregates, admixtures, and curing time, empirical formulas such as Abrams' water-cement ratio law and national mix-design codes capture only part of the behavior and may not work equally well across different mix families, material sources, and curing conditions.

Over the last decade, machine learning (ML) has become a common approach to this prediction problem, because it can model nonlinear interactions between mix constituents without an explicit physical model. Tree-based ensembles (random forest, gradient boosting, Extreme Gradient Boosting (XGBoost), CatBoost) and neural networks often outperform classical regression on the benchmark dataset compiled by Yeh (Yeh, 1998), achieving R^2 values in the 0.90–0.99 range depending on preprocessing, feature set, and validation strategy.

Even with this progress, three practical gaps make it harder to use ML strength predictors in engineering practice. First, most reported models are evaluated mainly using point-prediction accuracy (R^2, RMSE) without showing the uncertainty of each prediction, even though a design engineer needs to know how much a predicted value could reasonably vary before relying on it. Second, the best-performing models are usually tree ensembles or neural networks whose internal decision process is difficult to interpret; without feature-level explanation, it is difficult for engineers to check that a model's behavior is consistent with known concrete science rather than an artifact of the training data. Third, and often overlooked, differences between competing models are rarely tested for statistical significance. A headline R^2 improvement of 0.01–0.02 between two tree ensembles may or may not reflect a real, reproducible advantage rather than random variation from one train/test split.

This study addresses all three gaps within a single, reproducible pipeline. The specific contributions are:

- A physics-informed feature engineering step that augments the eight raw Yeh-dataset variables with the water-to-binder ratio, total binder content, aggregate-to-binder ratio, and log-transformed age (quantities with direct physical meaning in concrete mix design), together with an ablation study that explicitly tests whether these engineered features improve accuracy over the raw variables alone.
- A comparison of eight regression models, including a heterogeneous stacking ensemble, using the same train/test split and 10-fold cross-validation protocol, together with a randomized hyperparameter search reported clearly to support reproducibility.
- Statistical comparison (paired t-tests and Wilcoxon signed-rank tests across cross-validation folds) to assess whether observed differences between models, and between feature sets, show whether an advantage is consistent rather than depending on a particular data split, while recognizing that cross-validation folds share overlapping training data and so are not fully independent.
- Individual-prediction uncertainty quantification using split conformal prediction, evaluated over 100 repeated random splits to separate real calibration problems from variation caused by one data split, together with bootstrap confidence intervals for aggregate model performance and learning-curve analysis.
- A SHAP-based interpretability analysis, together with partial dependence plots and stratified residual analysis, that ranks feature contributions, shows how they affect predictions, and compares the results with established concrete science and against where the model's predictions are least reliable.

The rest of the paper is organized as follows. Section 2 reviews related work on empirical, ML-based, and interpretable concrete strength prediction. Section 3 describes the dataset, feature engineering, candidate models, hyperparameter optimization, evaluation metrics, and statistical testing methodology. Section 4 presents the results of all analyses. Section 5 discusses the findings, limitations, and practical implications. Section 6 concludes the paper.

Research Significance

Existing machine learning studies of concrete strength prediction have mainly focused on point-estimate accuracy, with model comparisons based on a single R^2 value from one train/test split and without formal statistical testing of whether the best-ranked model is truly superior. As a result, published rankings of algorithms may reflect sampling noise rather than real differences in predictive performance, leaving engineers without a clear measure of confidence in a given prediction.

This study addresses that gap by combining a strong gradient-boosted model with (i) paired statistical comparisons across cross-validation folds to assess whether its advantage over competing ensembles is consistent rather than incidental, (ii) an ablation study that tests the specific contribution of physics-informed feature engineering, and (iii) individual-level uncertainty quantification via split conformal prediction, evaluated over 100 repeated random splits rather than a single split, alongside bootstrap-based confidence intervals and stratified residual analysis for aggregate model reliability. Together, these elements provide a practical example of statistically careful ML strength-prediction studies. The findings, in particular that engineered ratios show no detectable predictive benefit for a tree ensemble in this experiment despite retaining strong interpretive value, give useful guidance for future studies deciding whether feature engineering is worth the extra modeling complexity.

2. Literature Review

2.1. Empirical and Statistical Models for Concrete Strength

Concrete mix design has traditionally relied on empirical relationships, most notably Abrams' water–cement ratio law and its many refinements, together with statistical regression models fit to laboratory data. These approaches are still useful because they are

transparent and easy to use in design codes, but they generally assume a fixed functional form (often linear or low-order polynomial) and a small number of input variables, which can limit their accuracy when applied to mixtures containing supplementary cementitious materials, chemical admixtures, or unconventional aggregates (Reddy et al., 2016; Amini et al., 2019). Non-destructive evaluation methods offer a complementary, though generally less accurate, alternative to destructive strength testing and empirical mix-design formulas (Breysse, 2012).

2.2. Machine Learning Approaches: From ANN to Gradient Boosting

The dataset used throughout this literature, and in the present study, was originally compiled and published by Yeh (Yeh, 1998), who applied an early artificial neural network (ANN) to relate concrete mix proportions to compressive strength and demonstrated a substantial improvement over linear regression. Since then, the same 1030-sample dataset has become a widely used benchmark, and a large number of studies have applied a range of more advanced models to it, including support vector regression, random forests, gradient boosting machines, XGBoost, CatBoost, and deep neural networks, typically reporting test R^2 in the range of 0.85–0.98 depending on the split and preprocessing used, consistent with broader reviews of predictive ML models for concrete structures and their life-cycle implications (Dinesh & Prasad, 2024). Ensemble tree methods often outperform single learners: Han et al. (Han et al., 2019) proposed an improved random forest for high-performance concrete strength prediction, while more recent work has systematically compared regularized linear models, kernel methods, tree ensembles, neural networks, and symbolic regression under a unified stacking framework (Liu et al., 2025), showing that gradient-boosted trees (XGBoost, CatBoost) often rank among the best models.

Bayesian and randomized hyperparameter search is now commonly used for tuning these models. Fu et al. (Fu et al., 2025) combined gradient boosting trees, random forest, and a backpropagation neural network with Bayesian hyperparameter optimization for a smaller (223-sample) field dataset, reporting that the optimized CatBoost model outperformed a current national mix-design empirical formula in single-point prediction accuracy, and further validated the model's generalization by applying it, without retraining, to subsets of the same UCI dataset used here.

2.3. Sustainable and Non-Conventional Concretes

Many recent studies extend ML strength prediction to non-conventional and more sustainable concretes, including recycled-aggregate concrete, high-volume fly ash and slag-based concrete, geopolymer concrete, pervious concrete, and manufactured-sand concrete. Golafshani et al. (Golafshani et al., 2024) combined machine learning with optimization algorithms for low-carbon recycled-aggregate concrete mix design, while other studies have modeled compressive strength for glass-powder-substituted eco-friendly concretes (Alcala-Gonzalez et al., 2025) and seawater-mixed sustainable concretes (Al-Shamiri et al., 2020) using ensembles and artificial neural networks, alongside a broader set of recent benchmark comparisons and non-destructive strength-prediction studies on the same or related datasets (Saeheaw, 2025; Nikoopayan Tak & Mahgoub, 2025; Altuncı, 2024; Li et al., 2024). Across these studies, cement content and curing age consistently emerge as dominant predictors regardless of the specific non-conventional material studied, reinforcing the physical basis for the engineered features used in the present study.

2.4. Interpretability and Uncertainty Quantification Gaps

Comprehensive reviews of ML approaches for specific non-conventional concretes, such as fly-ash-based geopolymer concrete (Rathnayaka et al., 2024), also conclude that explainability and standardized evaluation receive less attention than predictive accuracy. A growing number of studies use explainability tools, most commonly SHAP, to identify which mix variables drive model predictions, consistently finding cement content, water content, and curing age among the top contributors. Alavi and Noel (Alavi & Noel, 2025) explicitly addressed uncertainty by reporting prediction intervals for a non-destructive strength evaluation model, an approach that is still uncommon in this field; most studies report only a single point estimate of test accuracy without confidence intervals or repeated-fold significance testing. Reviews of the broader field, including the critical review by Ben Chaabene et al. (Ben Chaabene et al., 2020), have repeatedly identified the lack of interpretability and the lack of rigorous, standardized evaluation protocols as open challenges limiting the practical adoption of ML strength models in engineering design.

This study brings these ideas together: it uses physics-informed feature engineering as its main focus, directly tests whether those engineered features are useful via an ablation study, and combines this with the explainability and uncertainty-quantification elements that recent literature has identified as not studied enough, all within a single, statistically evaluated, reproducible pipeline.

3. Methodology

3.1. Dataset

This study uses the public Concrete Compressive Strength dataset originally donated by I-Cheng Yeh to the University of California, Irvine (UCI) Machine Learning Repository. The dataset contains 1030 concrete mixtures, each described by eight numerical mix-design variables (cement, blast furnace slag, fly ash, water, superplasticizer, coarse aggregate, and fine aggregate content, all in

kg/m³, plus curing age in days) and one continuous target, the age-specific compressive strength in MPa. The dataset contains no missing values. Table 1 summarizes the raw and engineered variables together with their physical units and observed ranges.

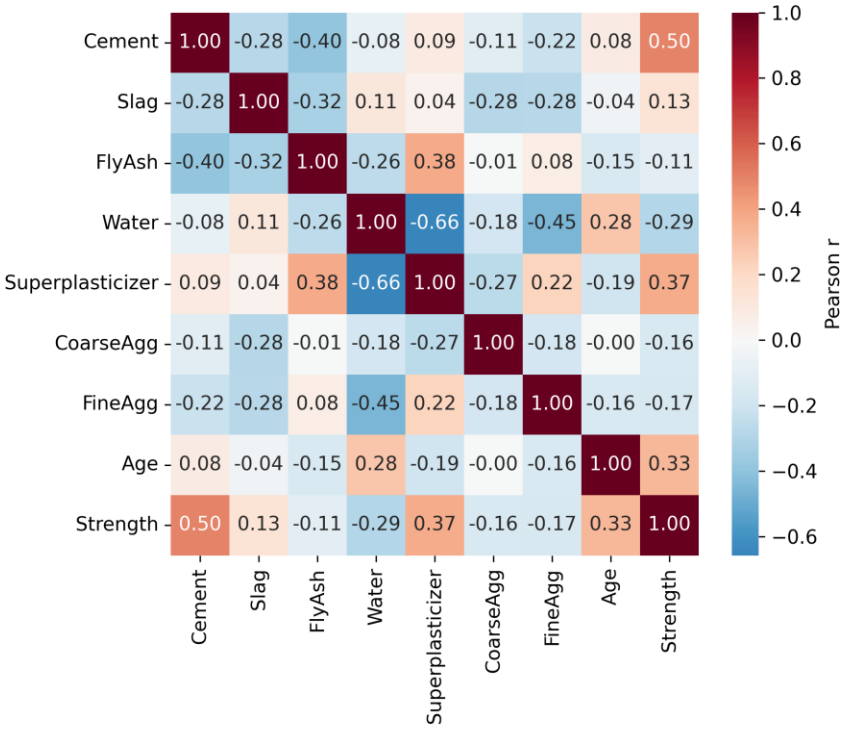

Figure 1. Pearson correlation matrix of the raw mix-design variables and compressive strength.

Table 1. *Raw and engineered variables used in this study. Entries marked with * are engineered features derived from the raw variables.*

Variable	Description	Unit	Range
Cement	Cement content	kg/m³	102–540
Slag	Blast furnace slag content	kg/m³	0–359.4
Fly Ash	Fly ash content	kg/m³	0–200.1
Water	Water content	kg/m³	121.8–247.0
Superplasticizer	Superplasticizer content	kg/m³	0–32.2
Coarse Aggregate	Coarse aggregate content	kg/m³	801–1145
Fine Aggregate	Fine aggregate content	kg/m³	594–992.6
Age	Curing age	days	1–365
Water/Binder ratio*	Water ÷ (cement+slag+fly ash)	–	engineered
Binder total*	Cement+slag+fly ash	kg/m³	engineered
Aggregate/Binder ratio*	(coarse+fine agg.) ÷ binder total	–	engineered
log(1+Age)*	Log-transformed curing age	–	engineered
Strength (target)	Age-specific compressive strength (age 1–365 days)	MPa	2.33–82.6

Figure 1 shows that no pair of raw predictors exceeds |r| = 0.66 (water and superplasticizer), so there is no extremely strong pairwise linear correlation, although this does not rule out multivariate collinearity; this issue matters more for the linear and ridge regression models than for the tree-based ensembles, which are not adversely affected by correlated inputs in the same way. Cement content shows the strongest individual correlation with strength (r = 0.50), followed by superplasticizer (r = 0.37) and age (r = 0.33), consistent with the well-known roles of these variables in concrete science.

3.2. Physics-Informed Feature Engineering

Along with the eight raw variables, four engineered features were created to represent well-known concrete-science relationships that are only implicit in the raw quantities: (i) the water-to-binder ratio, computed as water divided by the sum of cement, slag, and fly ash, which generalizes the classical water–cement ratio to blended-cement systems; (ii) the total binder content (cement + slag + fly ash); (iii) the aggregate-to-binder ratio, computed as total aggregate divided by total binder; and (iv) the natural logarithm of (1 + age), which reflects the well-known diminishing-returns relationship between curing time and strength gain. These four variables were added to the eight raw predictors, giving a 12-dimensional feature vector used for the main modeling pipeline, alongside an 8-dimensional raw-only feature vector used for the ablation study described in Section 3.7.

3.3. Candidate Models

Eight regression models covering linear, kernel, neural-network, and tree-ensemble approaches were benchmarked using the same data split and cross-validation protocol, so the different model types could be compared fairly.

3.3.1. Linear and Ridge Regression

Ordinary least-squares linear regression and its L2-regularized counterpart, ridge regression, are used as simple, transparent baselines. Ridge regression minimizes the sum of squared residuals plus a penalty term $\lambda\Sigma\beta^2_j$ on the regression coefficients, which makes coefficient estimates more stable when predictors are correlated and reduces overfitting relative to unregularized least squares.

3.3.2. Support Vector Regression

Support vector regression (SVR) with a radial basis function (RBF) kernel maps the input features into a higher-dimensional space and fits a function that deviates from each observed target by at most a margin ε, penalizing larger deviations linearly. The RBF kernel allows SVR to capture smooth nonlinear relationships without specifying an explicit polynomial form.

3.3.3. Artificial Neural Network

A feed-forward multilayer perceptron (MLP) with two hidden layers (64 and 32 units, rectified linear unit (ReLU) activation) was trained via backpropagation with the Adam optimizer. The network learns a composition of nonlinear transformations of the input features, and can learn complex nonlinear relationships when enough data are available.

3.3.4. Random Forest

Random forest (Breiman, 2001) is a bagging ensemble of decision trees, each trained on a bootstrap sample of the training data with a random subset of features considered at each split. Predictions are averaged across all trees, which reduces variance relative to a single decision tree while while still modeling nonlinear interactions and variable importance without requiring feature scaling.

3.3.5. Gradient Boosting and XGBoost

Gradient boosting (Friedman, 2001) builds an additive ensemble of shallow decision trees sequentially, with each new tree fit to the negative gradient (residual error) of the current ensemble's loss function. XGBoost (Chen & Guestrin, 2016) is a regularized, optimized implementation of gradient boosting that adds L1/L2 penalties on leaf weights and a tree-complexity penalty to the loss function, together with algorithmic optimizations (approximate split finding, column subsampling, sparsity awareness) that can improve accuracy and training speed relative to classical gradient boosting.

3.3.6. Stacking Ensemble

A heterogeneous stacking ensemble was constructed using random forest, gradient boosting, and XGBoost as base learners, with a ridge-regression meta-learner trained on their out-of-fold predictions. Stacking can combine the strengths of different base learners, although, as shown in Section 4.2, the improvement was very small here because the three tree-based base learners are highly correlated in their predictions.

3.4. Hyperparameter Optimization

For the primary model comparison (Section 4.2), all models were trained with standard, commonly used default-like hyperparameters (documented in the accompanying analysis code) to keep the comparison fair across model families. For the best-performing model (XGBoost), a randomized search was then performed over six hyperparameters using 40 sampled

configurations evaluated by 10-fold cross-validation on the training set only, with the held-out test set never used during search. Table 2 shows the tested values and the selected settings.

Table 2. *Randomized hyperparameter search space for XGBoost and the selected configuration (40 sampled configurations, 10-fold cross-validation on the training set).*

Hyperparameter	Search values tried	Selected value
Number of trees	200, 300, 400, 500, 600, 800	600
Maximum tree depth	3, 4, 5, 6, 8	3
Learning rate	0.01, 0.02, 0.05, 0.08, 0.10	0.10
Row subsample ratio	0.6, 0.7, 0.8, 0.9, 1.0	0.9
Column subsample ratio per tree	0.6, 0.7, 0.8, 0.9, 1.0	1.0
Minimum child weight	1, 3, 5, 7	1

3.5. Model Evaluation Metrics

Three standard regression metrics were used throughout: the coefficient of determination (R^2), which shows how much of the variation in strength is explained by the model; the root mean squared error (RMSE), which penalizes large errors more heavily than small ones; and the mean absolute error (MAE), which shows the average prediction error in the original MPa units. All three metrics were computed both on a single held-out test set (20% of the data) and via 10-fold cross-validation over the full dataset, so the results show both a realistic deployment scenario (train once, predict on new data) and the stability of that performance across different data partitions.

3.6. Interpretability: SHAP Analysis

To better understand the best-performing model, SHapley Additive exPlanations (SHAP) (Lundberg & Lee, 2017) values were computed for the XGBoost model using its exact tree-based explainer. SHAP decomposes each individual prediction into additive contributions from each input feature, grounded in the Shapley value concept from cooperative game theory, and the mean absolute SHAP value across the test set was used as a global importance ranking. This analysis provides a useful check (verifying that the model's learned relationships are consistent with known concrete behavior) and as a practical tool for engineers who need to understand why a given mixture is predicted to reach a certain strength. Partial dependence plots were additionally computed for the four most important features to visualize the functional (not just rank) form of each feature's marginal effect on predicted strength.

3.7. Ablation Study Design

To test whether the physics-informed engineered features (Section 3.2) improve predictive accuracy beyond what the raw eight variables already provide, XGBoost was trained and evaluated twice using the same train/test split and 10-fold cross-validation protocol: once using only the eight raw mix-design variables, and once using the full 12-variable set (raw plus engineered). The per-fold cross-validated R^2 scores from both configurations were then compared using a paired t-test and a Wilcoxon signed-rank test, because both configurations use both configurations are evaluated on the same data partitions.

3.8. Statistical Significance Testing

In addition to the ablation comparison, the same paired-testing approach was applied to compare XGBoost against its two closest competitors, random forest and gradient boosting, using their respective per-fold 10-fold cross-validation R^2 scores. This comparison checks whether XGBoost's higher average cross-validated R^2 (Section 4.2) represents a consistent statistical advantage rather than sampling variability across folds.

3.9. Conformal Prediction Intervals

A bootstrap confidence interval on test R^2 (Section 4.2) describes the reliability of the model's aggregate fit, but it does not tell an engineer how much a single new prediction might be expected to deviate from the true strength of that specific mixture. To provide uncertainty for individual predictions, split conformal prediction (Vovk et al., 2005; Angelopoulos & Bates, 2023) was used, a distribution-free method that adds a calibrated interval around point predictions carrying a finite-sample marginal coverage guarantee under exchangeability of calibration and test observations. In each run, the 1030-sample dataset was partitioned into training (60%), calibration (20%), and test (20%) subsets. XGBoost was fit once on the training subset; nonconformity scores

(absolute residuals) were then computed on the held-out calibration subset, and the (1−α)-quantile of these scores, adjusted for finite-sample coverage, was used as a fixed half-width added to and subtracted from each test-set point prediction to form the prediction interval. This guarantee has an important limitation: split conformal prediction guarantees marginal coverage of at least 1−α in expectation over exchangeable draws of the calibration and test data, not that every individual finite test sample must itself realize an empirical coverage rate at or above 1−α. To measure this sampling variation instead of relying on one split, the entire train/calibration/test partitioning and conformal calibration procedure was repeated over 100 independent random splits (seeds 0–99), and both the single-split result and the distribution of empirical coverage and interval width across all 100 splits are reported in Section 4.10.

3.10. Computational Implementation

All models were built in Python 3 using scikit-learn for linear models, support vector regression, the neural network, random forest, gradient boosting, and stacking; the xgboost library for XGBoost; the shap library for interpretability analysis; and scipy.stats for paired t-tests and Wilcoxon signed-rank tests. A fixed random seed (42) was used throughout for the train/test split, cross-validation folds, and all stochastic model components for reproducibility. The analysis and figure-generation code accompanies this manuscript and, if accepted for publication, will be deposited in a public repository (or the journal's supplementary materials system) to support reproducibility.

4. Results and Findings

4.1. Descriptive Statistics and Correlation Analysis

The 1030-sample dataset covers a wide range of mix designs and curing ages (Table 1), with compressive strength ranging from 2.33 to 82.6 MPa (mean 35.8 MPa, SD 16.7 MPa), covering low-strength to high-performance concretes. This gives broad coverage of the mix-design and strength ranges represented within the Yeh dataset, although, as shown in Section 4.9, prediction uncertainty is not uniform across this range, and as discussed in Section 5, coverage of this dataset's range is not equivalent to demonstrated generalization beyond it.

4.2. Model Comparison

Table 3 reports the held-out test performance and 10-fold cross-validated R^2 for all eight models. The two linear models (linear regression and ridge regression) achieved the weakest performance (test $R^2 \approx 0.83$), confirming that the strength–mix design relationship is strongly nonlinear and cannot be adequately captured by a global linear model even after feature engineering. Support vector regression and the neural network improved substantially over the linear baselines (test R^2 of 0.895 and 0.875, respectively). All three tree-based ensembles outperformed the non-tree models, with XGBoost achieving the best held-out performance (test R^2 = 0.931, RMSE = 4.22 MPa, MAE = 2.79 MPa) and the highest 10-fold cross-validated R^2 (0.944 ± 0.019). The stacking ensemble performed essentially on par with standalone XGBoost (test R^2 = 0.930) but did not surpass it in this configuration, suggesting that at this dataset size the three base learners are sufficiently correlated that stacking mainly reproduces XGBoost's performance rather than adding complementary information.

Table 3. Held-out test performance and 10-fold cross-validated R^2 for all evaluated models (best model highlighted). CV = cross-validation; SD = standard deviation.

Model	Test R^2	Test RMSE (MPa)	Test MAE (MPa)	10-fold CV R^2 (mean ± SD)
Linear Regression	0.829	6.65	5.29	0.822 ± 0.020
Ridge Regression	0.830	6.62	5.26	0.822 ± 0.021
SVR (RBF kernel)	0.895	5.21	3.44	0.909 ± 0.025
ANN (MLP, 64-32)	0.875	5.67	3.72	0.913 ± 0.033
Random Forest	0.911	4.79	3.31	0.922 ± 0.018
Gradient Boosting	0.921	4.52	3.14	0.932 ± 0.020
XGBoost	0.931	4.22	2.79	0.944 ± 0.019
Stacking (RF+GB+XGB)	0.930	4.24	2.79	0.937 ± 0.018

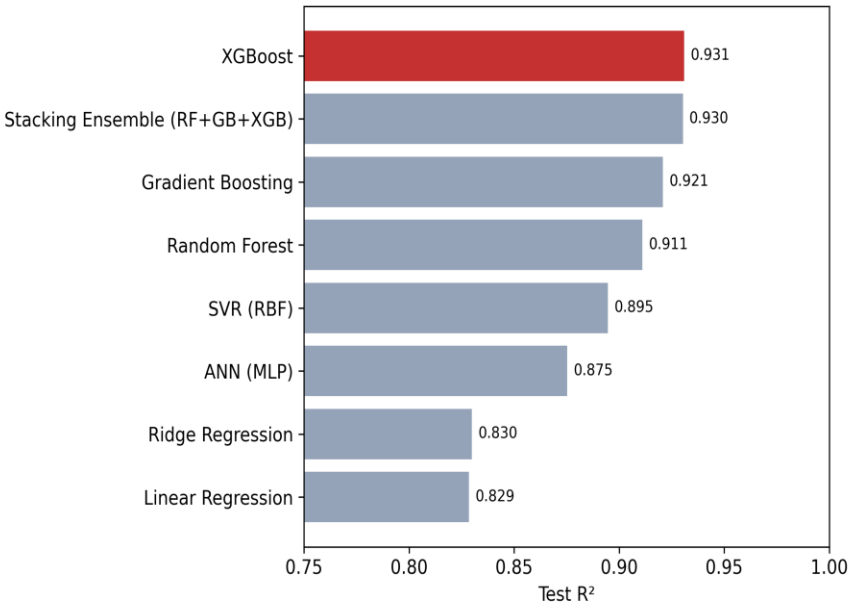

Figure 2. Held-out test R^2 across all evaluated models.

A bootstrap analysis (500 resamples) of the XGBoost test predictions gave a 95% confidence interval of $R^2 \in [0.901, 0.955]$, which indicates that the reported point estimate (0.931) is a stable and reasonably precise estimate of the model's generalization performance on data drawn from the same distribution as the Yeh dataset.

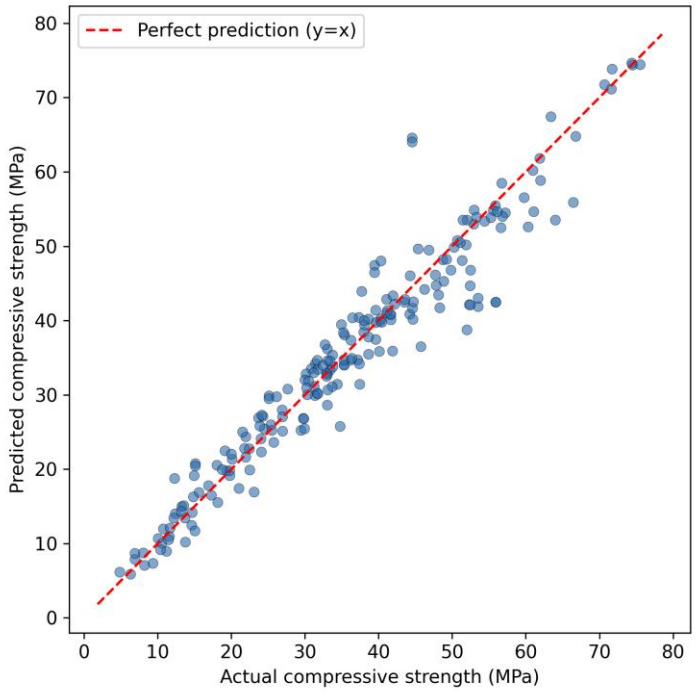

Figure 3. XGBoost predicted vs. actual compressive strength on the held-out test set.

4.3. Statistical Comparison of Model Differences

Paired t-tests across the 10 cross-validation folds showed that XGBoost consistently outperformed its closest tree-ensemble competitors on paired per-fold R^2 scores: XGBoost vs. random forest ($t = 12.48$, $p < 0.0001$) and XGBoost vs. gradient boosting ($t = 8.62$, $p < 0.0001$). Because the same 10 folds were used to evaluate all three models, this paired design isolates the effect of the modeling algorithm from fold-to-fold variation in the data, supporting a consistent XGBoost advantage beyond what the aggregate means in Table 3 alone would suggest. We note, however, that cross-validation folds share substantially overlapping

training data and so are not fully independent; the very small p-values above should therefore be interpreted as evidence of a consistent, replicable ranking across folds rather than as a formally independent hypothesis test, and a corrected resampled test or repeated cross-validation would be needed to treat this comparison as a fully calibrated significance test.

4.4. Hyperparameter Optimization Results

The randomized search (Table 2) selected a configuration with more but shallower trees (600 trees at a maximum depth of 3, versus 500 trees at a maximum depth of 4 in the default configuration used for Section 4.2), a higher learning rate (0.10 vs. 0.05), and near-complete column and row subsampling per boosting round. This tuned configuration achieved a held-out test R^2 of 0.937 (RMSE = 4.04 MPa, MAE = 2.69 MPa), a small but consistent improvement over the default configuration's test R^2 of 0.931, illustrating that hyperparameter tuning provides small rather than major gains once a strong learner (gradient-boosted trees) and a reasonable feature set are already in place. To keep the primary model comparison in Table 3 fair and complexity-matched across all eight model families, the tuned configuration is reported here as a separate, follow-on optimization experiment; it was not substituted back into Table 3 and was not used to establish the baseline model ranking. Unless otherwise stated, all interpretability, residual, ablation, and uncertainty analyses in the remainder of this paper (Sections 4.5–4.10) use the default XGBoost configuration described in Section 3.3.5 and Table 3, not the tuned configuration described here.

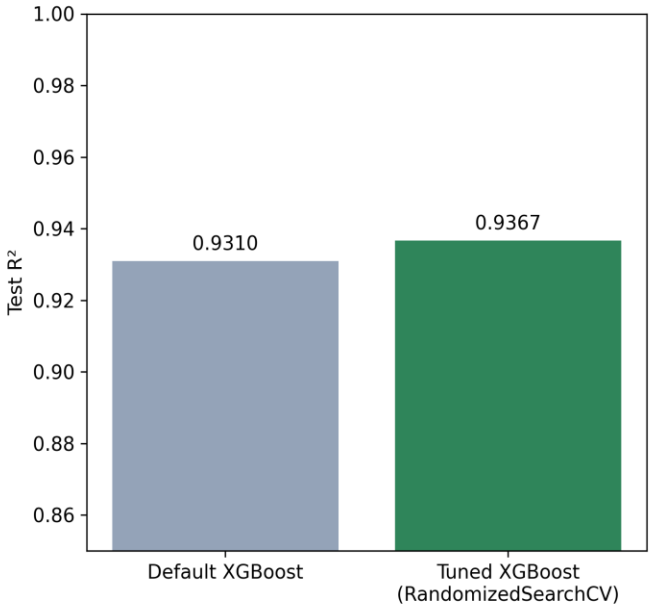

Figure 4. Held-out test R^2 for the default versus randomized-search-tuned XGBoost configuration.

4.5. Ablation Study: Raw vs. Physics-Informed Features

Table 4 compares XGBoost trained on the eight raw mix-design variables against the same model trained on the full 12-variable set including the physics-informed engineered features. Contrary to the initial hypothesis that explicit water-to-binder and aggregate-to-binder ratios would meaningfully improve accuracy, the engineered feature set produced only a marginal change in test R^2 (0.931 vs. 0.930) and RMSE (4.22 vs. 4.25 MPa), with a paired t-test across the 10 cross-validation folds confirming that this difference is not statistically significant ($t = -0.34$, $p = 0.74$; Wilcoxon signed-rank $p = 0.92$). Notably, the engineered-feature model did show a modestly lower test MAE (2.79 vs. 2.96 MPa), suggesting a small reduction in typical-case error even though the difference in R^2/RMSE was not significant.

Table 4. *Ablation study comparing XGBoost trained on raw mix-design variables only versus raw plus physics-informed engineered features.*

Feature set	Test R^2	Test RMSE (MPa)	Test MAE (MPa)	10-fold CV R^2 (mean ± SD)
Raw features only (8)	0.930	4.25	2.96	0.945 ± 0.015
Raw + engineered features (12)	0.931	4.22	2.79	0.944 ± 0.019

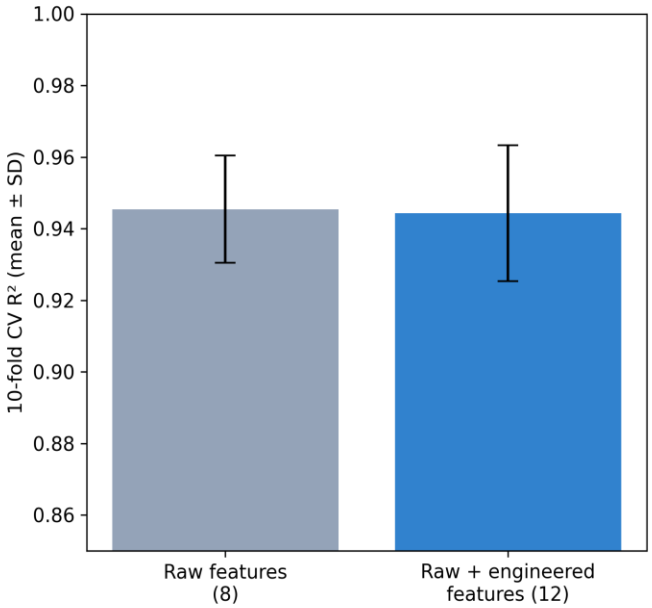

Figure 5. 10-fold cross-validated R^2 for XGBoost trained on raw features only versus raw plus engineered features (error bars show ± 1 SD across folds). SD = standard deviation.

This result is still useful. One possible explanation is that gradient-boosted decision trees construct splits directly on the raw variables and can approximate ratio-like relationships (e.g., a split sequence on both water and binder content) without an explicit ratio feature, which would make tree ensembles less reliant on manual feature engineering than linear or distance-based models. The ablation study only shows that no predictive benefit was detected in this experiment; the mechanism above is a reasonable interpretation rather than something directly demonstrated by the data. The practical value of the engineered features in this study is, based on these results, primarily interpretive (Section 4.7) rather than predictive.

4.6. Learning Curve Analysis

Figure 6 shows the training and 10-fold cross-validated R^2 of the default XGBoost configuration as a function of training set size. The cross-validated R^2 rises steeply from small training sizes and begins to plateau above roughly 500–600 training samples, while the gap between training and validation R^2 remains modest throughout, indicating that the model is not severely overfitting even at the full training-set size. This plateau suggests diminishing returns from additional samples drawn from a similar distribution to the existing Yeh data, although it says nothing about out-of-distribution generalization; additional independent mixtures from sources not represented in the current dataset could still meaningfully improve generalization in ways this within-distribution learning curve cannot reveal.

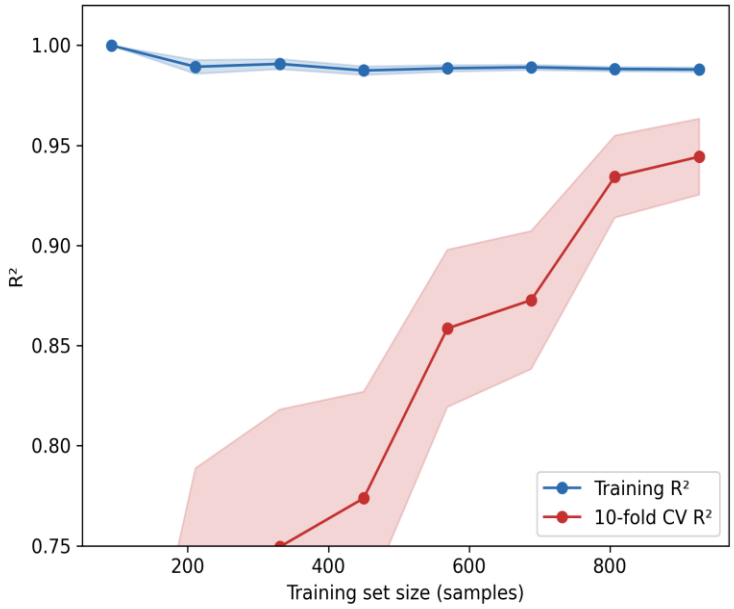

Figure 6. Learning curve for the default XGBoost model: training and 10-fold cross-validated R^2 versus training set size (shaded bands show ± 1 SD across folds). SD = standard deviation.

4.7. Feature Importance and Interpretability

Figure 7 shows the global SHAP feature importance ranking for the XGBoost model. Curing age was the single most influential predictor (mean |SHAP| = 6.56 MPa), followed closely by the engineered water-to-binder ratio (5.11 MPa), cement content (2.32 MPa), and total binder content (1.98 MPa). Water content, log-transformed age, fly ash, and the aggregate-to-binder ratio formed a second tier of moderately important features, while fine aggregate, superplasticizer, slag, and coarse aggregate contributed the least on average. This ranking is consistent with established concrete science: strength gain with curing time and the water-to-binder ratio are the two classical first-order determinants of concrete strength, and their emergence as the top two ML-derived features (ahead of the raw cement or water content taken individually) supports the interpretive, if not predictive, value of the engineered features introduced in Section 3.2.

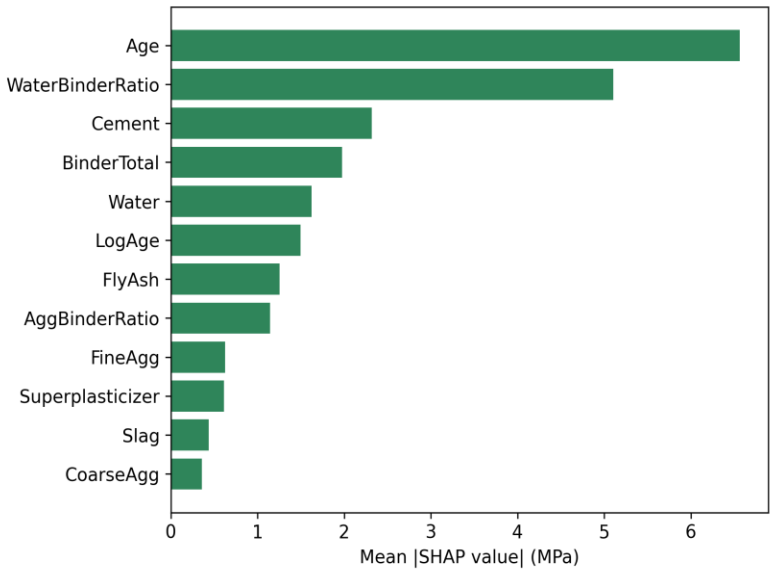

Figure 7. Global SHAP feature importance for the XGBoost model.

4.8. Partial Dependence Analysis

Figure 8 shows partial dependence plots for the four most important features. The partial dependence on age rises steeply up to roughly 28–90 days before flattening, mirroring the well-known diminishing-returns curve of concrete strength gain with curing time and providing a direct visual counterpart to the SHAP ranking. The partial dependence on the water-to-binder ratio is monotonically decreasing, exactly as classical concrete theory predicts, while cement content and total binder content both show a generally increasing, mildly nonlinear relationship with predicted strength. The consistency of these functional forms with established concrete behavior provides a useful qualitative sanity check on the model, although it does not by itself establish causal relationships or rule out dataset-specific associations. This caveat applies with particular force to the engineered ratio features, since water-to-binder ratio and the other engineered variables are deterministic functions of the raw inputs rather than independently measured quantities.

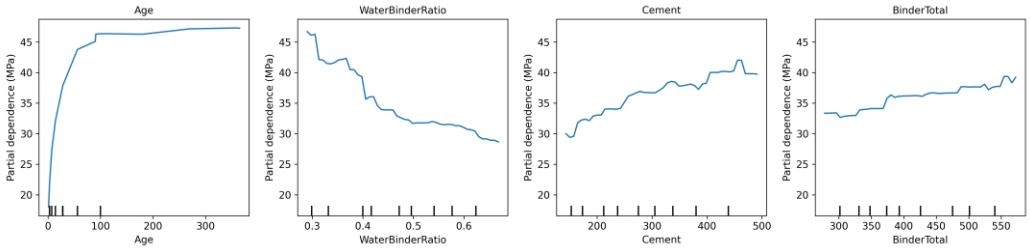

Figure 8. Partial dependence plots for curing age, water-to-binder ratio, cement content, and total binder content.

4.9. Residual Analysis

Figure 9a plots residuals (actual minus predicted strength) against predicted strength for the held-out test set; Figure 9b shows the corresponding residual histogram. The residuals are approximately centered on zero and roughly symmetric, but their spread is not constant across the strength range: Table 5 shows that residual standard deviation grows from 2.23 MPa for mixtures below 20 MPa to 5.80 MPa for mixtures in the 40–60 MPa range, before narrowing slightly above 60 MPa. A similar, though weaker, pattern appears when residuals are stratified by curing age (Table 6), with somewhat larger residual spread at intermediate ages (8–28 days) than at very early or very late ages. Two clear outliers, both under-predicted by roughly 20 MPa, are visible in Figure 9a. Rather than speculate about these, we inspected the underlying rows directly: both have essentially identical mix proportions (cement ≈ 313 kg/m^3, slag ≈ 145 kg/m^3, no fly ash, water ≈ 127 kg/m^3, superplasticizer 8 kg/m^3, age 28 days) and an identical reported actual strength of 44.52 MPa, and both share an unusually low water-to-binder ratio of approximately 0.28, close to the 5th percentile of the dataset (0.29). Given the strong inverse partial dependence of predicted strength on water-to-binder ratio (Figure 8), the model predicts a substantially higher strength (≈ 64 MPa) for both, consistent with the classical expectation that very low water-to-binder ratios yield high strength; the gap with the lower measured value could reflect a genuine practical limitation at this low ratio (e.g., workability or compaction difficulty) that is under-represented elsewhere in the dataset, or, since the two rows are near-duplicates with an identical reported strength, may reflect the same underlying mixture recorded more than once, which would connect directly to the mixture-duplication and leakage risk discussed in Section 5.

Table 5. *Residual statistics on the held-out test set, stratified by actual compressive strength range.*

Strength range	n	Mean residual (MPa)	SD of residual (MPa)
< 20 MPa	41	−0.67	2.23
20–40 MPa	86	−0.41	2.96
40–60 MPa	64	+1.81	5.80
> 60 MPa	15	+2.34	4.45

Table 6. *Residual statistics on the held-out test set, stratified by curing age.*

Curing age	n	Mean residual (MPa)	SD of residual (MPa)
≤ 7 days	56	+0.85	3.97
8–28 days	91	+0.28	5.01

29–90 days	29	+0.39	2.69
> 90 days	30	+0.11	3.14

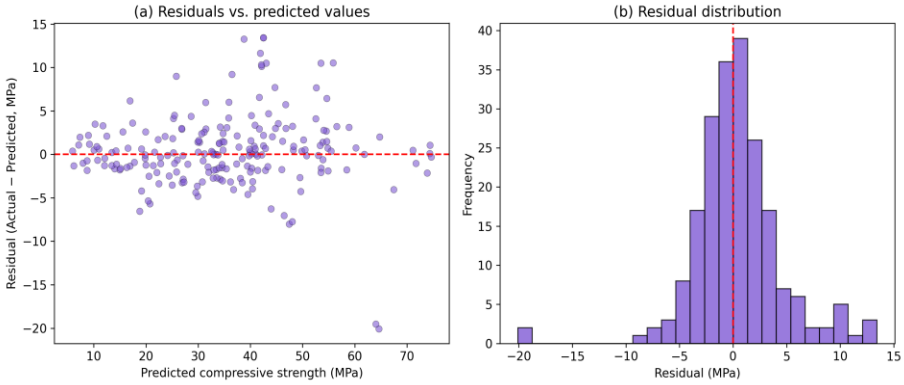

Figure 9. (a) Residuals vs. predicted compressive strength and (b) residual distribution for the XGBoost model on the held-out test set.

4.10. Conformal Prediction Intervals

Table 7 reports the empirical coverage and average width of the 90% and 95% split conformal prediction intervals described in Section 3.9. On the single split used elsewhere in this paper (n = 206 test observations), the 95% interval achieved 91.7% empirical coverage with an average width of 16.7 MPa, and the 90% interval achieved 84.0% empirical coverage with an average width of 12.3 MPa. Both fall below their nominal targets on this particular split. The 95% and 90% conformal intervals achieved empirical coverages of 91.7% and 84.0%, respectively, on this finite held-out test subset; these realized coverage rates fall below their nominal levels but do not, by themselves, contradict the finite-sample marginal coverage guarantee of split conformal prediction, which is defined over exchangeable future observations rather than requiring the realized coverage proportion of every individual finite test sample to meet or exceed the nominal level.

To assess whether this single-split shortfall reflected genuine miscalibration or ordinary sampling variability, the entire train/calibration/test partitioning and conformal calibration procedure was repeated over 100 independent random splits. Averaged across these 100 splits, empirical coverage was 95.06% ± 2.15% for the nominal 95% interval and 90.59% ± 2.62% for the nominal 90% interval, both closely matching their nominal targets on average, with the original single-split coverages (91.7% and 84.0%) falling within the observed range across splits (95% interval: 88.8–98.1%; 90% interval: 82.5–96.1%). These results indicate that the original single-split undercoverage is consistent with ordinary sampling variability rather than providing evidence of systematic miscalibration in the conformal procedure, and they illustrate why single-split coverage figures should always be interpreted alongside their variability across repeated splits rather than in isolation.

Table 7. *Empirical coverage and average width of split conformal prediction intervals: single-split result and summary across 100 repeated random splits.*

Nominal coverage	Single-split coverage	Single-split width (MPa)	Mean coverage ± SD (100 splits)	Mean width ± SD (MPa)
90%	84.0%	12.27	90.59% ± 2.62%	13.86 ± 1.46
95%	91.7%	16.65	95.06% ± 2.15%	18.63 ± 2.16

Table 8 illustrates the practical form of the single-split intervals for five representative test-set mixtures spanning the predicted strength range, each reported as a point prediction together with its 95% conformal prediction interval and the actual observed strength. Unlike the aggregate R^2 confidence interval in Section 4.2, which describes how reliable the model's overall fit is, these intervals give an engineer a direct, per-mixture statement, for example "predicted strength 39.5 MPa, 95% interval 31.2–47.8 MPa," that can be compared against a design strength requirement before a decision is made.

Table 8. *Example point predictions with 95% conformal prediction intervals (single split) for five representative held-out mixtures.*

Predicted (MPa)	95% PI lower (MPa)	95% PI upper (MPa)	Actual (MPa)
5.0	−3.4	13.3	4.8
22.6	14.3	31.0	21.9
32.0	23.7	40.3	30.4
39.5	31.2	47.8	38.1
50.5	42.2	58.9	53.8

The symmetric split-conformal formulation can produce a negative lower bound for very-low-strength predictions, as illustrated by the first example in Table 8. Although such values are mathematically possible with this unconstrained interval method, negative compressive strength is not physically meaningful. We report this value as produced by the method rather than silently clipping it to zero, since clipping without adjusting the underlying procedure would alter the interval's coverage properties in an unprincipled way. Future work could incorporate a non-negative transformation or bounded conformal approach to enforce physically admissible prediction intervals while preserving appropriate coverage properties.

Because the conformal half-width is fixed across the whole test range in this basic split-conformal implementation, it reflects average error rather than local error; Section 4.9 showed that residual spread is itself larger in the 40–60 MPa range, so intervals in that range are, if anything, somewhat too narrow relative to the local error, while intervals for low-strength mixtures are correspondingly conservative (and, as noted above, can extend below zero). A locally adaptive (e.g., quantile-regression-based or bounded) conformal method would address both of these limitations and is noted as a direction for future refinement in Section 5.

5. Discussion

The results show, on a well-known benchmark dataset, that gradient-boosted tree ensembles perform much better than both linear models and single neural networks for concrete strength prediction, and, unlike much of the prior literature, this study shows via paired statistical comparison that XGBoost's advantage over random forest and gradient boosting is consistent across cross-validation folds rather than an artifact of a single data split (with the caveat, noted in Section 4.3, that folds are not fully independent). This distinction matters in practice: a design team choosing between two ML tools based on a 1–2 percentage-point difference in reported R^2 from a single paper cannot know, without such testing, whether that difference would replicate on a new dataset or a new split of the same data.

The null result of the ablation study is one of the more useful practical findings of this study. Physics-informed feature engineering is often presented in the literature as a simple way to add domain knowledge into ML pipelines, and the water-to-binder ratio in particular is treated as a commonly used derived feature in concrete ML studies. The present results show that, for a sufficiently expressive tree ensemble, this engineering step does not produce a clear accuracy gain in this experiment. This suggests that XGBoost may already approximate equivalent nonlinear combinations of the raw variables through its tree-splitting process, though the ablation design used here establishes the absence of a measurable predictive benefit rather than directly proving that specific mechanism. This does not make the engineered features useless: because SHAP consistently and prominently ranks the water-to-binder ratio as the second most influential feature (Section 4.7), and because its partial dependence exactly reproduces the classical strength–ratio relationship (Section 4.8), the engineered feature still serves as a compact, physically interpretable lens through which to communicate the model's behavior to practicing engineers, even though it is not doing predictive work that the raw variables could not already do. Studies using linear, distance-based, or shallow models (where the ablation result may differ) would be a natural place to test whether this finding generalizes.

The residual analysis in Section 4.9 shows a limitation that is not clear from aggregate R^2/RMSE alone: prediction uncertainty is heteroscedastic, concentrated in the 40–60 MPa range where residual variance is roughly 2.5× that of the lowest-strength range. A physical explanation is possible: this mid-to-high strength range likely contains the widest diversity of admixture combinations (superplasticizers, fly ash substitution) in the Yeh dataset, each interacting with strength in ways that are harder to pin down with a finite sample. This has a direct practical implication: point predictions in this range should be treated with wider engineering margins than the aggregate RMSE of 4.22 MPa would suggest on its own.

Three additional limitations are important when interpreting these results. First, and most importantly, the Yeh dataset aggregates data from multiple original experimental studies without a documented grouping variable identifying which rows originate from

the same physical mixture tested at different ages; several rows in the dataset are near-duplicates of each other differing mainly in age, consistent with repeated testing of the same mix. As recent literature has explicitly warned, this raises a real possibility of information leakage between train and test sets if repeated mixtures are split across both, which would inflate every R^2, RMSE, and coverage figure reported in this paper relative to performance on a genuinely independent mixture. Because no reliable mixture identifier is available in the public version of the dataset used here, this study reports standard random-split and k-fold results, consistent with the great majority of prior work on this dataset, but this is a real limitation and not a formality: high R^2 on the Yeh dataset, as reported here and throughout this literature, does not by itself demonstrate generalization to a genuinely unseen concrete mixture, and should not be read as such. Future work should repeat the analysis with mixture-grouped cross-validation (e.g., GroupKFold on a verified Mix ID) if and when such metadata becomes available, and, more importantly, should validate the trained model against an independently collected dataset with no relationship to the Yeh data. We regard this external validation, not further tuning of the present model, as the single most valuable next step for this line of work. Second, the fixed-width split conformal intervals reported in Section 4.10 are designed to provide finite-sample marginal coverage under exchangeability, which the 100-repeated-split analysis confirmed holds on average across splits, but they are not locally adaptive; because residual spread itself grows with strength magnitude (Section 4.9), a conformalized quantile regression or other locally adaptive conformal method would likely produce tighter intervals at low strength and more honestly wide intervals in the 40–60 MPa range, and is a natural refinement of the uncertainty quantification presented here. Third, the dataset is restricted to the eight variables originally recorded by Yeh and does not include other factors known to affect strength (e.g., aggregate gradation, cement fineness, or curing temperature and humidity), so the model's applicability is bounded by the same variable set as the underlying data.

In practice, the combination of high accuracy, model selection supported by paired cross-validation comparisons, per-mixture conformal prediction intervals, and transparent feature attribution makes the proposed framework more suitable for use in early-stage mix-design screening tools than a black-box model alone, since engineers can see both a point prediction and a calibrated (if not yet locally adaptive) range for a specific new mixture, and can verify that the model's reasoning is consistent with known concrete behavior before trusting it, while keeping firmly in mind the leakage caveat above when deciding how much weight that prediction should carry for a mixture genuinely unlike anything in the Yeh dataset.

6. Conclusion

This study developed an interpretable, statistically evaluated machine learning framework for predicting concrete compressive strength across a wide range of curing ages (1–365 days) from mix-design variables, with individual-level uncertainty quantified via conformal prediction. Combining physics-informed engineered features (water-to-binder ratio, binder content, aggregate-to-binder ratio, log-age) with a comparison of eight regression models on the 1030-sample Yeh/UCI dataset, the XGBoost model achieved the best performance (test $R^2 = 0.931$, RMSE = 4.22 MPa; 10-fold CV $R^2 = 0.944 \pm 0.019$; tuned test $R^2 = 0.937$). Paired statistical comparison showed that this advantage over random forest and gradient boosting was consistent ($p < 0.0001$ in both cases), and split conformal prediction produced 95% prediction intervals for individual mixtures, with 91.7% empirical coverage on a single split, found to be consistent with the nominal target (mean 95.06% ± 2.15%) across 100 repeated splits, a more directly usable uncertainty statement for a single new mix design than an aggregate confidence interval on R^2 alone. An ablation study found no statistically significant accuracy benefit from the physics-informed engineered features for this tree-based model ($p = 0.74$), even though those same features dominate the model's SHAP-based interpretability ranking and reproduce textbook concrete-strength relationships in partial dependence plots. This indicates that their primary value in this pipeline is interpretive rather than predictive. Residual analysis further revealed that prediction uncertainty is not uniform across the strength range, being largest for mixtures in the 40–60 MPa band. Because the present split-conformal intervals use a fixed width, they do not adapt to this heteroscedasticity, motivating locally adaptive conformal methods in future work. By explicitly quantifying both aggregate and individual-level predictive uncertainty, testing model and feature-engineering choices for statistical significance, and grounding feature importance in known concrete behavior, the proposed framework addresses practical gaps in reliability, interpretability, and statistical rigor that make it harder to use ML-based strength prediction in engineering practice. The dataset's lack of a verifiable mixture identifier means mixture-grouped, leakage-safe validation was not possible here; together with external validation on an independently collected dataset, this is the clearest and most important direction for future work to strengthen generalization claims beyond the Yeh/UCI benchmark. Extending the same physics-informed feature engineering, ablation, and conformal-prediction methodology to non-conventional concretes (recycled aggregate, high-volume fly ash, geopolymer) is a natural further direction.

Statements and Declarations

Author Contributions

Author Contributions: The specific contributions of Jafar Sadeghi must be confirmed and approved by author before submission.

Funding

This research received no external funding. The article processing charge, if applicable, will be funded by the author.

Data Availability Statement

The dataset analyzed in this study is publicly available from the UCI Machine Learning Repository at https://archive.ics.uci.edu/dataset/165/concrete+compressive+strength (originally donated by I-Cheng Yeh). Analysis and figure-generation code accompanies this submission and will be made available in a public repository upon acceptance.

Conflicts of Interest

The author declares no conflict of interest.

Acknowledgments

The author has no acknowledgments to declare.

References

[1]. Al-Shamiri, A. K., Yuan, T.-F., & Kim, J. H. (2020). Non-tuned machine learning approach for predicting the compressive strength of high-performance concrete. Materials, 13, 1023.
[2]. Alavi, S. A., & Noel, M. (2025). Uncertainty and prediction intervals of new machine learning approach for non-destructive evaluation of concrete compressive strength. Buildings, 15, 544.
[3]. Alcala-Gonzalez, D., Mateo, L. F., Quijano, M. Á., Más-López, M. I., & García-del-Toro, E. M. (2025). Compressive strength-based classification of eco-friendly concretes using machine learning models. Materials, 18, 5344.
[4]. Altuncı, Y. T. (2024). A comprehensive study on the estimation of concrete compressive strength using machine learning models. Buildings, 14, 3851.
[5]. Amini, K., Vosoughi, P., Ceylan, H., & Taylor, P. (2019). Effect of mixture proportions on concrete performance. Construction and Building Materials, 212, 77-84.
[6]. Angelopoulos, A. N., & Bates, S. (2023). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. Foundations and Trends in Machine Learning, 16, 494-591.
[7]. Ben Chaabene, W., Flah, M., & Nehdi, M. L. (2020). Machine learning prediction of mechanical properties of concrete: Critical review. Construction and Building Materials, 260, 119889.
[8]. Breiman, L. (2001). Random forests. Machine Learning, 45, 5-32.
[9]. Breysse, D. (2012). Nondestructive evaluation of concrete strength: An historical review and a new perspective by combining NDT methods. Construction and Building Materials, 33, 139-163.
[10]. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794). ACM.
[11]. Dinesh, A., & Prasad, B. R. (2024). Predictive models in machine learning for strength and life cycle assessment of concrete structures. Automation in Construction, 162, 105412.
[12]. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29, 1189-1232.
[13]. Fu, H., Zhou, X., Xu, P., & Sun, D. (2025). Prediction of compressive strength of concrete using explainable machine learning models. Materials, 18, 5009.
[14]. Golafshani, E. M., Behnood, A., Kim, T., Ngo, T., & Kashani, A. (2024). A framework for low-carbon mix design of recycled aggregate concrete with supplementary cementitious materials using machine learning and optimization algorithms. Structures, 61, 106143.
[15]. Han, Q., Gui, C., Xu, J., & Lacidogna, G. (2019). A generalized method to predict the compressive strength of high-performance concrete by improved random forest algorithm. Construction and Building Materials, 226, 734-742.
[16]. Li, H., Chung, H., Li, Z., & Li, W. (2024). Compressive strength prediction of fly ash-based concrete using single and hybrid machine learning models. Buildings, 14, 3299.
[17]. Liu, J., Guan, D., & Liu, X. (2025). Comparative performance analysis of machine learning models for compressive strength prediction in concrete mix design. Math. Comput. Appl., 30, 128.
[18]. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems 30 (pp. 4765-4774). Curran Associates.
[19]. Nikoopayan Tak, M. S., & Mahgoub, M. (2025). Advanced machine learning techniques for predicting concrete compressive strength. Infrastructures, 10, 26.
[20]. Rathnayaka, M., Karunasinghe, D., Gunasekara, C., Wijesundara, K., Lokuge, W., & Law, D. W. (2024). Machine learning approaches to predict compressive strength of fly ash-based geopolymer concrete: A comprehensive review. Construction and Building Materials, 419, 135519.
[21]. Reddy, M. S., Dinakar, P., & Rao, B. H. (2016). A review of the influence of source material's oxide composition on the compressive strength of geopolymer concrete. Microporous and Mesoporous Materials, 234, 12-23.
[22]. Saeheaw, T. (2025). Data-driven insights into concrete flow and strength: Advancing smart material design using machine learning strategies. Buildings, 15, 2244.
[23]. Vovk, V., Gammerman, A., & Shafer, G. (2005). Algorithmic learning in a random world (2nd ed.). Springer.
[24]. Yeh, I.-C. (1998). Modeling of strength of high-performance concrete using artificial neural networks. Cement and Concrete Research, 28, 1797-1808.