Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Predicting 28-Day Compressive Strength of Concrete from Field Data: Uncertainty Quantification and Seasonal Interaction Effects

**Hamed Enayati**

hamed.enayati@mazust.ac.ir

University of Science and Technology of Mazandaran

**Mousa Nazari**

University of Science and Technology of Mazandaran

**Mahdi kiani**

University of Science and Technology of Mazandaran

**Article**

**Keywords:** Concrete compressive strength prediction, field data, uncertainty quantification, conformal prediction, SHAP interaction analysis, ensemble learning

**Posted Date:** November 17th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11442269/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# Predicting 28-Day Compressive Strength of Concrete from Field Data: Uncertainty Quantification and Seasonal Interaction Effects

**Hamed Enayati**
Assistant Professor, University of Science and Technology of Mazandaran
**Mousa Nazari**
Assistant Professor, University of Science and Technology of Mazandaran
**Mahdi kiani**
Assistant Professor, University of Science and Technology of Mazandaran

***Corresponding Author: Hamed Enayati (hamed.enayati@mazust.ac.ir)***

## Abstract

Compressive strength at 28 days remains the standard criterion for concrete acceptance, yet it is unknown until four weeks after placement, well after decisions such as formwork removal must already be made. Most predictive models in the literature are developed and validated on laboratory-mixed specimens, leaving open how well they generalize to field-collected data, where records are incomplete and the early-age conditions of the specimens vary with site and season. This study addresses that gap using 1,018 field records from construction projects in Mazandaran province (2023–2025), evaluated under two realistic scenarios: with and without an early-age (7-day) strength measurement. A stacking ensemble combining XGBoost, SVR, and MLP was benchmarked against several meta-learner variants; formal statistical testing (Wilcoxon signed-rank and Diebold–Mariano tests) showed no significant improvement from more elaborate meta-learners over a simple linear combination once base learners were properly tuned. Conformal prediction was therefore used to quantify predictive uncertainty directly: interval widths nearly doubled in the absence of the 7-day test (11.5 versus 22.4 MPa at 90% confidence and 14.8 versus 27.1 MPa at 95%, standard CV+), with direct implications for quality-control decisions. Segment-wise error analysis showed that errors were largest at the extremes of the strength range when the 7-day test was unavailable, with no detectable seasonal difference, while SHAP interaction analysis revealed a season-dependent modulation of both early-age strength and slump effects that is physically consistent with the temperature sensitivity of cement hydration, although temperature itself was not measured.

**Keywords:** Concrete compressive strength prediction; field data; uncertainty quantification; conformal prediction; SHAP interaction analysis; ensemble learning.

## 1. Introduction

Concrete remains the most widely used construction material worldwide, owing to its availability, versatility, and relatively low cost. Among its mechanical properties, compressive strength is the most direct indicator of structural performance and safety, and the 28-day value is the reference

against which acceptance decisions are made. Because that value is not available until four weeks after placement, engineers have long sought reliable ways to anticipate it from information collected earlier, most commonly a 7-day strength test, so that construction scheduling, formwork removal, and resource allocation need not wait on the full curing period.

Early attempts to establish this link relied on empirical relationships and linear regression between early-age and 28-day strength [4, 6, 40]. Yeh's compressive-strength dataset, which records mixture proportions alongside curing age, became the standard benchmark for this line of work and remains widely used for comparing methods [3]. Oluokun et al.'s early study found that existing equations relating 7-day to 28-day strength could be inaccurate by as much as 13%, and showed that early-age strength prediction requires the heat of hydration to be accounted for [40]; later site-oriented studies proposed age-based regression models that predict 28-day strength from 3- and 7-day results [4, 6], although these models use early-age strength alone, without accompanying mixture or site information.

As it became clear that the strength-development process is nonlinear and depends on several interacting variables simultaneously, attention shifted toward machine learning as a way of capturing these relationships without imposing a fixed functional form [1, 3]. Neural networks were among the first tools applied in this direction: Nguyen et al. combined early-age measurements with mixture properties in a feedforward neural network and found early-age data substantially improved prediction accuracy over mix proportions alone [39], while Palanisamy et al. showed that careful hyperparameter tuning of neural architectures further reduced prediction error [5]. Tree-based ensemble methods followed a similar trajectory: Li and Song reported that XGBoost, using curing age and admixture data, outperformed both regression and neural approaches for high-performance concrete [7]; Zhang et al. reached a similar conclusion with random forest, AdaBoost, XGBoost and LightGBM while additionally emphasizing interpretability [15]; and Su et al. found Random Forest and AdaBoost more reliable than neural networks specifically when fiber content varied, pointing to the value of ensemble flexibility under material heterogeneity [9]. Comparative studies across model families (Sah et al. across neural networks, support vector machines, and tree-based methods [10]; Ahmad et al. across ANN, SVM, and XGBoost [14]) converge on two conclusions relevant here: that no single algorithm dominates independently of dataset and feature quality, and that 7-day strength consistently ranks among the most informative predictors once it is available.

A more recent thread of this literature has focused on interpretability rather than raw accuracy. Elshaarawy et al. combined gradient boosting with feature selection and confirmed the persistent importance of 7-day strength as a predictor [11]. Pandey et al., using machine learning models and sensitivity analysis, investigated the nonlinear relationships between concrete mix design parameters and its compressive strength. In contrast, Fu et al. used explainable machine learning methods, including SHAP analysis, to assess the relative importance of input variables in predicting concrete compressive strength [17, 18]. Abioye et al. extended explainable machine learning to reinforced and high-performance concrete, finding that tree-based models offer a

favorable balance of accuracy and transparency [19]. Systematic reviews of this body of work [12, 13] describe a clear shift toward ensemble and interpretable methods, but also flag recurring limitations: small sample sizes, inconsistent reporting, and, most relevant to the present study, a near-total reliance on laboratory-prepared specimens rather than field-collected data. Open benchmark resources, including Kaggle's expanded versions of Yeh's dataset, have made algorithmic comparison easier but have not addressed this last limitation, since they inherit the same laboratory provenance and typically omit curing and site-condition detail [16].

## Uncertainty Quantification in Concrete Strength Prediction

Predictive uncertainty in concrete strength has two components: aleatoric variability that more data cannot reduce, arising from material heterogeneity, batching, sampling and testing, and epistemic uncertainty arising from limited data and model misspecification [41]. Methods differ in what they deliver, either a full predictive distribution (probabilistic prediction) or a prediction interval with a stated coverage level, and their quality is judged jointly by calibration, meaning whether the stated coverage is achieved, and sharpness, meaning how narrow the intervals are [42]. The approaches applied or applicable to concrete strength fall into four groups, compared below.

Bayesian and probabilistic approaches return a predictive distribution by construction. Gaussian process regression [43] has been used to predict concrete strength trajectories across curing ages [44] and the strength development of alkali-activated concrete, where adding early-age strength measurements improved later-age predictions, increased the coverage of the prediction intervals and reduced their width [45]; an open-source Gaussian process model combined with Bayesian optimization has also been proposed for strength forecasting and mix design [46]. These models give input-dependent uncertainty estimates, but their intervals rest on kernel, prior and noise assumptions, and exact inference scales poorly with sample size.

Uncertainty-aware variants of standard learners offer a middle route. Monte Carlo dropout reinterprets dropout training as approximate Bayesian inference [47], deep ensembles use the disagreement among independently trained networks [48], and natural-gradient boosting fits a tree ensemble that outputs the parameters of a predictive distribution [49]. They retain the flexibility of neural and tree-based models, but their intervals depend on the assumed output distribution, the ensemble design or the quality of the approximation, and none of them guarantees the nominal coverage in finite samples.

Quantile regression estimates conditional quantiles directly by minimizing the pinball loss [50], and combined with tree ensembles, it gives quantile regression forests [51]. Because the quantiles are learned from the data, the resulting intervals adapt to heteroscedasticity, but their coverage can deviate from the nominal level when the model is misspecified or the sample is small. Conformalized quantile regression addresses this by calibrating quantile-based intervals on held-out residuals [52].

Conformal prediction wraps any point predictor and produces intervals with finite-sample, distribution-free marginal coverage under exchangeability [36, 53]. Split conformal prediction uses a calibration set, whereas the jackknife+ and CV+ variants reuse cross-validation residuals so that no data are set aside [54]; [55] gives an accessible introduction. Applications to concrete strength are very recent. Tamuly and Nava combined eight machine-learning models with six conformal variants to construct prediction intervals for concrete compressive strength [56]. Mobasheri et al. combined four ensemble models with CV+, a CV+ variant labeled CV+_Gamma, and conformalized quantile regression on more than 3,000 self-consolidating concrete samples, and reported that CV+ gave narrower but constant intervals, whereas CV+_Gamma and conformalized quantile regression gave adaptive intervals, which exposes a trade-off among precision, adaptability and coverage reliability [57].

Taken together, Bayesian methods provide distributions but depend on modeling assumptions, quantile-based methods adapt to the input but carry no coverage guarantee on their own, and conformal methods guarantee marginal coverage but, in their basic form, produce intervals of constant width. Within the sources reviewed here, none quantifies how prediction intervals change when early-age strength is withheld as an input on field-collected construction data; the closest evidence is the study of alkali-activated concrete in which early-age strength measurements increased interval coverage and narrowed the intervals [45]. This is the gap addressed by the uncertainty analysis in the present study.

Two gaps stand out once this literature is read as a whole rather than study by study. First, models are almost never trained on data collected directly from active construction sites, where mix records are less complete and early-age curing is not held constant, so it remains unclear how much of the reported laboratory accuracy survives the transition to field conditions. Second, and this applies even to the interpretability-focused studies described earlier, most models are reported as single-point predictors: a number is produced, and its accuracy over a held-out set is reported, but the associated predictive uncertainty is rarely quantified, and the recent studies that do quantify it do not examine how it changes with the information available on site, although this information matters as much to an engineering decision as the number itself. The present study addresses both gaps directly using field data collected from construction projects rather than laboratory mixes, and by treating predictive uncertainty as a primary outcome of interest rather than an afterthought.

Table 1 situates the present study relative to a representative set of prior works. Most rely on laboratory or public benchmark data, evaluate performance only when full input information (including 7-day strength) is available, and report accuracy without an accompanying measure of confidence. In contrast, the framework developed here is trained and evaluated on field-collected data under two realistic information scenarios, subjected to formal statistical comparison of ensemble configurations, and paired with conformal prediction to quantify how much each prediction can be trusted, in addition to being benchmarked on the independent UCI dataset to examine whether the modeling behavior holds beyond the local dataset.

Table 1. Comparison of recent studies on concrete compressive strength prediction and the present framework [7–11, 15, 20]

| Study | Dataset Type | Field Data | Prediction w/o 7-Day | Uncertainty Quantified | Explainability | Independent Benchmark |
|---|---|---|---|---|---|---|
| Li and Song (2022) | Laboratory | ✗ | ✗ | ✗ | ✗ | ✗ |
| Pham et al. (2020) | Laboratory | ✗ | ✗ | ✗ | ✗ | ✗ |
| Su et al. (2024) | Experimental | ✗ | ✗ | ✗ | ✗ | ✗ |
| Zhang et al. (2024) | Laboratory | ✗ | ✗ | ✗ | ✓ | ✗ |
| Sah and Hong (2024) | Public Dataset | ✗ | ✗ | ✗ | ✗ | ✗ |
| Elshaarawy et al. (2025) | Public Dataset | ✗ | ✗ | ✗ | Limited | ✗ |
| Dhengare et al. (2025) | Public Dataset | ✗ | ✗ | ✗ | ✗ | ✗ |
| Present Study | Field + Benchmark | ✓ | ✓ | ✓ (Conformal) | ✓ (SHAP) | ✓ (UCI) |

## 2. Problem Statement

The 28-day compressive strength, $S_{28}$, is the reference quantity against which concrete is accepted, yet it cannot be measured until four weeks after placement, well after operational decisions such as formwork removal or early loading must be made. In practice, only a limited set of covariates is available at the time such decisions are required: fresh-concrete slump, cement content, the design (characteristic) strength class, placement location, casting month, and, when performed, an early-age strength test at 7 days. Formally, let $x_i \in \mathbb{R}^d$ denote the covariate vector for specimen $i$,

$$x_i = (Slump_i, Placement_i, Cement_i, f'c,_i, Month_i, [S_{7,i}]) \quad (1)$$

where $S_{7,i}$ is included only under the with-7-day scenario. The prediction task is to learn a function $f: \mathbb{R}^d \rightarrow \mathbb{R}$ such that

$$S_{28,i} = f(x_i) \approx S_{28,i} \quad (2)$$

evaluated separately for each scenario, using field-collected rather than laboratory-prepared specimens.

Two aspects of this problem are treated here as primary rather than secondary. First, most models reported in the literature are trained and validated on laboratory mixes, where water-cement ratio, aggregate composition, and curing regime are tightly controlled and precisely recorded; whether reported accuracy transfers to field conditions, where several of these variables are unrecorded and

the early-age conditions of the specimens (the first 24 h or more at the site) are subject to seasonal variation, has not been systematically established. Second, existing models are almost universally presented as point predictors $f(x_i)$, without an accompanying estimate of predictive uncertainty. For a decision such as formwork removal, however, the relevant quantity is not $S_{28,i}$ alone but an interval $C(x_i)$ satisfying

$$P(S_{28,i} \in C(x_i)) \geq 1 - \alpha \quad (3)$$

for a chosen confidence level 1−α; this interval determines whether a prediction is reliable enough to act on, and is particularly relevant when $S_{7,i}$ is unavailable and $S_{28,i}$ is correspondingly less constrained.

Accordingly, this study does not treat ensemble architecture as the primary source of expected improvement. This assumption is tested directly against field data through formal statistical comparison of meta-learner variants (Section 3.3), and where no significant improvement is found, this is reported rather than omitted; the focus then shifts to characterizing $C(x_i)$ and to identifying the covariate regions (strength class, slump range, and season) under which model predictions are least reliable.

## Research Gap and Contributions

Building on the problem framed above, this study makes five contributions. Unlike the innovations claimed in much of the surrounding literature, several of these are empirical rather than architectural: they report the empirical findings obtained when standard assumptions were tested against field data, rather than assuming a more elaborate model would necessarily perform better.

1. A systematic evaluation of how the availability of early-age strength data (with versus without 7-day strength) affects both the accuracy and the predictive uncertainty of 28-day strength prediction, based on 1,018 field-collected construction records and using a hybrid Stacking framework (XGBoost, SVR, MLP) as the modeling vehicle.
2. A formal comparison of five meta-learner variants (Linear, Ridge, ElasticNet, Bayesian Ridge, and a shallow XGBoost), tested with Wilcoxon signed-rank and Diebold–Mariano statistical tests rather than point-estimate comparison alone; the resulting null finding, that no meta-learner variant significantly outperforms a simple linear combination, is reported as a substantive result.
3. Uncertainty quantification via conformal prediction (split conformal, standard CV+, and a conservative Jackknife+ variant), providing calibrated prediction intervals rather than point estimates, and directly measuring how much predictive reliability is lost when the 7-day test is unavailable.
4. A segment-wise error analysis (strength class, slump range, and casting season) that locates where errors concentrate and finds no detectable difference in error between the warm and

cold halves of the year, complementing aggregate accuracy metrics with condition-specific ones.

5. A SHAP interaction analysis that reveals a season-dependent modulation of early-age strength effects, offering a physically interpretable finding without requiring any additional physical or climatic input.

## 3. Methodology

This study develops and evaluates a hybrid, interpretable machine-learning framework for predicting 28-day compressive strength from field data. The framework proceeds in four stages: data preprocessing, base-learner training, meta-learner comparison within a Stacking architecture, and a validation layer that adds uncertainty quantification and statistical testing on top of standard accuracy metrics. Figure 1 summarizes the overall workflow, from raw input variables through the hybrid Stacking model to the extended validation and interpretation layer developed in this study.

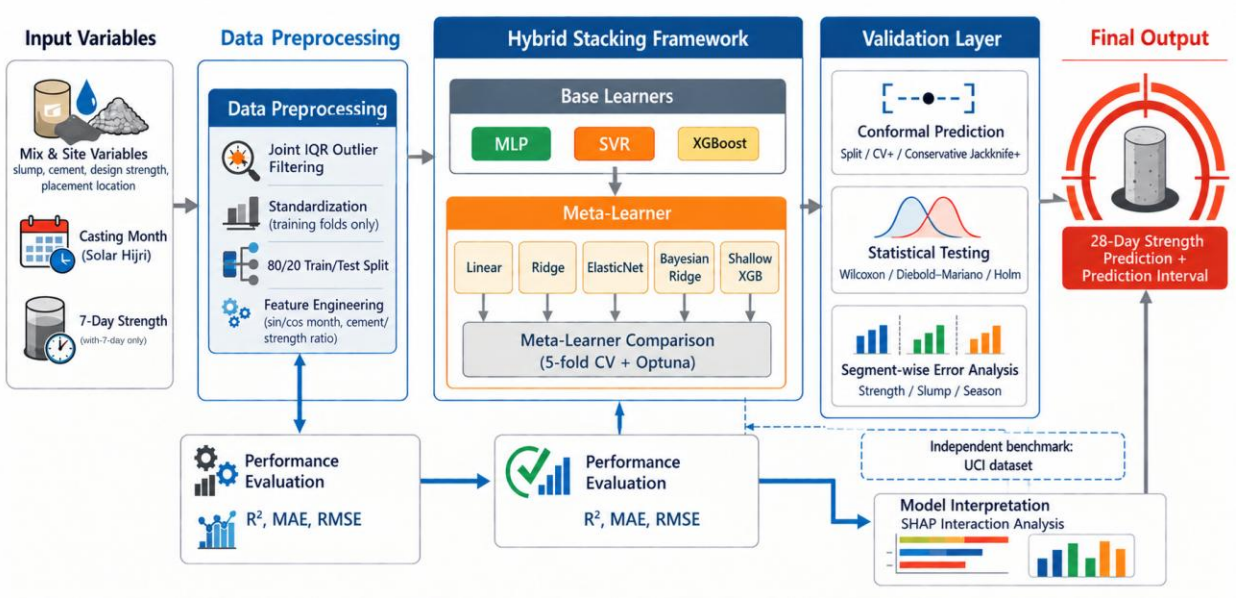



Figure 1. Overall workflow of the proposed framework: data preprocessing, the hybrid Stacking model (base learners and meta-learner comparison), the extended validation layer (cross-validation, conformal prediction, and statistical testing), and SHAP-based model interpretation.

### 3.1. Data Preprocessing

Outlier Detection. Values that deviate substantially from a feature's overall distribution were identified using the Interquartile Range (IQR) method, which requires no distributional assumption [21, 58]. The first and third quartiles $(Q_1, Q_3)$ were computed for slump, 28-day strength and 7-day strength, and the interquartile range defined as

$$IQR = Q_3 - Q_1 \quad (4)$$

Samples falling below $Q_1 - 1.5 \times IQR$ or above $Q_3 + 1.5 \times IQR$ were removed. A record was removed if it fell outside the fences of any of these variables (joint filtering): [4.5, 16.5] for slump, [5.7, 58.5] MPa for 28-day strength and [3.3, 46.5] MPa for 7-day strength. Cement content, characteristic strength, casting month and placement location were not filtered, because they are design or administrative variables rather than measurements (cement content takes only seven distinct values, with 350 kg/m³ in about 83% of the retained records). The filter was applied to the full dataset before the 80/20 split, and both field scenarios use the same common sample of records, so that differences between them are not caused by sample composition (Table 3).

Feature Engineering. Two derived features were constructed [2]. Casting month, being cyclical (Esfand, month 12, and Farvardin, month 1, are climatically adjacent despite being numerically distant), was encoded as

$$sin_month = sin(2\pi \cdot Month/12),\ cos_month = cos(2\pi \cdot Month/12) \quad (5)$$

and a cement-to-characteristic-strength ratio was defined as

$$cement_to_strength_ratio = Cement\ Content\ /\ Characteristic\ Strength \quad (6)$$

As reported in Section 5.2, interaction values involving raw cement content were small in the with-7-day scenario (at most 0.017), whereas without the 7-day test the cement/strength ratio entered several of the strongest interaction pairs.

## 3.2. Base Learners

Three base learners were selected to capture complementary aspects of the strength-development relationship: XGBoost for tree-based nonlinear interactions, SVR for margin-based robustness to noise, and MLP for unconstrained nonlinear function approximation.

XGBoost optimizes a regularized objective of the form

$$obj = \Sigma_i\, l(y_i, \hat{y}_i) + \Sigma_k\, \Omega(f_k) \quad (7)$$

where $l(\cdot)$ is a loss function (mean squared error here) and $\Omega(\cdot)$ penalizes tree complexity [22, 23, 32]. SVR solves

$$min_{w,b}\ ½\|\omega\|^2 + C\Sigma_i\xi_i,\ subject\ to\ |y_i - (\omega^T\varphi(x_i) + b)| \le \varepsilon + \xi_i \quad (8)$$

with an RBF kernel used to model nonlinear relationships between mix features and strength [24, 25, 29]. The MLP had one or two hidden layers (8–64 units each); its activation function (ReLU or tanh), L2 penalty and initial learning rate were tuned by Optuna, and it was trained with Adam under squared-error loss for up to 2,000 iterations with early stopping; no dropout was used [26, 27, 31]; the selected architecture was a single hidden layer of 45, 62 and 29 units for the with-7-day, without-7-day and UCI datasets, respectively. Hyperparameters for all three base learners were tuned on the 80% training split by Bayesian optimization (Optuna, TPE sampler, seed 42; 40

trials per base learner and 25 per meta-learner) minimizing 5-fold cross-validated RMSE (non-nested), so cross-validation scores are mildly optimistic; the test split was never used for tuning.

### 3.3. Stacking Ensemble and Meta-Learner Comparison

The base learners' out-of-fold predictions were combined by a second-level meta-learner,

$$\hat{y}_final = \alpha_1 \hat{y}_XGBoost + \alpha_2 \hat{y}_SVR + \alpha_3 \hat{y}_MLP \quad (9)$$

with coefficients $\alpha_i$ fitted by the meta-learner (for the linear variants; the shallow XGBoost meta-learner is nonlinear). Rather than fixing this meta-learner to Linear Regression as in the initial design, five variants were compared under identical base-learner configurations and cross-validation folds: Linear Regression, Ridge, ElasticNet, Bayesian Ridge [30], and a shallow XGBoost ($max_depth \leq 3$, to limit overfitting given the sample size). Each variant was additionally compared against each base learner used alone, giving a full ablation of the ensemble's added value. All comparisons used the same 5-fold cross-validation on the training split, with hyperparameters for every base learner and every meta-learner selected via Optuna. The meta-learner received only the three base-learner predictions, without passthrough of the original features.

### 3.4. Uncertainty Quantification via Conformal Prediction

To move beyond point predictions, prediction intervals were constructed using conformal prediction, which provides distribution-free coverage guarantees [36]. Two variants were used. Split conformal prediction divides the training set into a fitting subset and a calibration subset, computes calibration residuals $|y_i - \hat{y}_i|$, and sets the half-width of the interval to the empirical $(1-\alpha)$ quantile of these residuals,

$$C(x) = [\hat{y}(x) - \hat{q}, \hat{y}(x) + \hat{q}], \quad \hat{q} = Quantile(1-\alpha)\{|y_i - \hat{y}_i| : i \in calibration\ set\} \quad (10)$$

CV+ instead refits the pipeline across K folds, retains each fold's out-of-fold residuals, and constructs the interval from the $\alpha$ and $1-\alpha$ quantiles of the fold-specific predictions minus and plus these residuals; a more conservative Jackknife+ variant takes the $\alpha/2$ and $1-\alpha/2$ quantiles instead. Both variants were evaluated at 90% and 95% nominal confidence, and assessed by empirical coverage (the fraction of true values falling inside the interval) and average interval width. In the split variant, 30% of the training split was held out for calibration and the model was refit on the remaining 70%; in the Jackknife+/CV+ variant, K = 5 folds were used and two constructions were evaluated: the standard CV+ interval (the primary method) and a conservative variant that takes the $\alpha/2$ and $1-\alpha/2$ quantiles of the lower and upper bounds, which over-covers by construction. Intervals are centered on the mean of the five fold-model predictions, and coverage is reported with Wilson 95% confidence intervals.

## 3.5. Statistical Significance Testing

Differences in performance among the best stacking configuration, the best single base learner, and the original Linear Regression baseline were tested formally rather than inferred from point estimates alone, using out-of-fold predictions from the same 5-fold cross-validation on the training split (n = 728 paired records per field scenario; because the folds share training data, the pairs are not fully independent and the p-values should be read as approximate). The Wilcoxon signed-rank test was applied to paired absolute errors [37], and the Diebold–Mariano test, with the Harvey–Leybourne–Newbold small-sample correction, was applied to squared prediction errors [38]:

$$DM = \bar{d} / \sqrt{(Var(d)/n)}, \; d_i = e^2_{1,i} - e^2_{2,i} \tag{11}$$

The DM statistic used a one-step horizon (h = 1), squared-error loss (with an absolute-error variant as a sensitivity check), the HLN correction, and a t-distribution with n − 1 degrees of freedom; Wilcoxon effect sizes are rank-biserial correlations, and Holm's procedure was applied across the six comparisons.

## 3.6. Model Configuration and Experimental Setup

Table 2 consolidates the final configuration used across all experiments reported in Section 4, bringing together the hyperparameter search spaces, the five meta-learner variants, the uncertainty-quantification methods, and the validation settings described in the preceding subsections into a single reference.

Table 2. Model configuration and experimental settings

| Component | Key Settings |
|---|---|
| XGBoost (base) | Tuned via Optuna (TPE, 40 trials): n_estimators ∈ [100,500], max_depth ∈ [2,8], learning_rate ∈ [0.01,0.3] (log), subsample and colsample_bytree ∈ [0.6,1], reg_alpha and reg_lambda ∈ [1e-3,10] (log) |
| SVR (base) | Kernel = RBF; tuned via Optuna (40 trials): C ∈ [1e-2,1e3], ε ∈ [1e-3,1], γ ∈ {scale, auto} |
| MLP (base) | 1–2 hidden layers (8–64 units each); activation ∈ {ReLU, tanh}; L2 penalty α ∈ [1e-5,1e-1] and initial learning rate ∈ [1e-4,1e-1] tuned via Optuna (40 trials); Adam, squared-error loss, early stopping (max 2,000 iterations); no dropout |
| Meta-learners compared | Linear Regression (baseline), Ridge (α ∈ [1e-3,100]), ElasticNet (α ∈ [1e-3,10], l1_ratio ∈ [0,1]), Bayesian Ridge ($\alpha_1$, $\alpha_2$, $\lambda_1$, $\lambda_2$ ∈ [1e-7,1e-1]), shallow XGBoost (n_estimators ∈ [50,300], max_depth ∈ [1,3]); 25 Optuna trials each |
| Uncertainty quantification | Split Conformal; CV+ (standard) and conservative Jackknife+, at 90% and 95% confidence |
| Significance testing | Wilcoxon signed-rank; Diebold–Mariano (HLN-corrected; squared and absolute loss); Holm correction (6 comparisons) |

| Component | Key Settings |
|---|---|
| Dataset splitting | 80% training / 20% testing (random, unstratified; 182 field and 183 UCI test records) |
| Cross-validation | 5-fold (repeated 5×5 for the feature-engineering ablation); fixed random seed (42) throughout; model selection by CV RMSE |

Outlier filtering was applied to the full dataset before the 80/20 random split (random_state = 42, unstratified), so the IQR fences were computed from all records, including those later assigned to the test set; because 28-day strength is itself a filtered variable, the test set is also filtered on the target, which is not possible at deployment and may make the reported test accuracy slightly optimistic. Placement location was one-hot encoded (six categories) and numeric features were standardized, with both transformations fitted on training folds only. The test split (182 records per field scenario; 183 for UCI) was not used for hyperparameter tuning or meta-learner fitting; all tuning used 5-fold cross-validation on the training split, and optimized parameters were applied to the test set without re-tuning. The best stacking configuration and the best single model for each dataset were selected by cross-validated RMSE on the training split, not by test-set performance. Within the Stacking architecture, base-learner predictions used for meta-learner training were generated out-of-fold to prevent information leakage between levels.

## 4. Results

This section reports results in the order the Methodology section introduced them: the experimental dataset (Section 4.1), predictive performance across meta-learner variants (Section 4.2), formal statistical testing of the resulting differences (Section 4.3), uncertainty quantification via conformal prediction (Section 4.4), the feature-engineering ablation (Section 4.5), condition-specific error analysis (Section 4.6), and generalization to the independent UCI benchmark (Section 4.7). Results are reported here without extended interpretation; their engineering and physical implications are discussed jointly in Section 5.

### 4.1. Experimental Dataset

A total of 1,018 records were collected from concrete laboratories in several cities in Mazandaran province between 2023 and 2025. Each record contains the fresh-concrete slump, cement content, design (characteristic) compressive strength, placement location, casting month, and the 28-day and 7-day compressive strengths, the latter missing for only five records. Casting month follows the Solar Hijri calendar (1 = Farvardin), so that month 12 (Esfand) adjoins month 1 in the cyclical encoding. Slump was measured on site by trained technicians.

Specimens were taken from the concrete at the discharge point of the pump, at the placement site. Each sampling event typically produced six 150 mm cubes, and each cube was compacted in three layers with 25 blows per layer. The cubes stayed in their molds at the site for at least 24 h. They were then taken to the laboratory, demolded, and cured in a water tank until testing at 7 or 28 days.

Usually, two cubes were tested at 7 days and three at 28 days, and the mean result, evaluated under the relevant ACI provisions, was recorded as the 7-day or 28-day strength. Both strengths were then converted to standard cylindrical equivalents. What makes these data field data is the sampling at the placement site and the first day or longer of exposure to ambient conditions; once the cubes reach the laboratory, curing follows standard procedure.

Of the 1,018 raw records, 12 were excluded because at least one required field (month, 7-day strength, or 28-day strength) was missing. The remaining 1,006 records form the common sample for both scenarios. Four pairs of records are identical in all seven fields and were retained. Only one such pair straddles the training and test split, so duplicates can affect at most one of the 182 test records. Outlier filtering is summarized in Table 3. It removed 93 records for slump, 3 for 28-day strength, and none for 7-day strength (fences: slump [4.5, 16.5], 28-day strength [5.7, 58.5] MPa, 7-day strength [3.3, 46.5] MPa). Cement content was not filtered.

The UCI Concrete Compressive Strength dataset [35] was used for independent generalization testing. It was processed with the same modeling and validation procedures, using dataset-specific preprocessing. After 25 duplicates were removed from the original 1,030 records, the IQR filter removed 59 records with curing ages above 129.5 days (180–365 days), so the retained data cover ages of 1–120 days.

Table 3. Sample sizes after IQR-based outlier filtering

| Dataset | Records before filtering | Records removed | Records used |
|---|---|---|---|
| Field (both scenarios, common sample) | 1,006 | 96 | 910 |
| UCI (independent benchmark) | 1,005 | 94 | 911 |

*Note. IQR filtering was applied jointly to slump, 28-day strength, and 7-day strength for the field data, and to all eight inputs and the target for UCI. Both field scenarios share the same 1,006-record common sample.*

## 4.2. Predictive Performance Across Meta-Learner Variants

Table 4 summarizes test-set performance ($R^2$, MAE and RMSE [28]) for the stacking configuration selected by cross-validated RMSE, the original Linear Regression baseline, and the best single base learner, for each dataset. Full results across all five meta-learner variants and all three single base learners are available; only the relevant extremes are reported here for brevity.

Table 4. Best stacking configuration versus baseline and best single model

| Dataset | Configuration | $R^2$ | MAE | RMSE | RMSE vs. baseline |
|---|---|---|---|---|---|
| Field (with 7-day) | Best stacking (ElasticNet) | 0.846 | 2.440 | 3.197 | −0.23% |
| Field (with 7-day) | Baseline (Linear) | 0.846 | 2.446 | 3.205 | N/A |
| Field (with 7-day) | Best single (XGBoost) | 0.835 | 2.543 | 3.314 | +3.4% |

| | | | | | |
|---|---|---|---|---|---|
| Field (without 7-day) | Best stacking (Bayesian Ridge) | 0.391 | 5.043 | 6.365 | −0.17% |
| Field (without 7-day) | Baseline (Linear) | 0.389 | 5.043 | 6.376 | N/A |
| Field (without 7-day) | Best single (XGBoost) | 0.378 | 5.100 | 6.434 | +0.9% |
| UCI (independent) | Best stacking (ElasticNet) | 0.947 | 2.524 | 3.717 | ≈0% |
| UCI (independent) | Baseline (Linear) | 0.947 | 2.525 | 3.717 | N/A |
| UCI (independent) | Best single (XGBoost) | 0.946 | 2.553 | 3.770 | +1.4% |

Models were selected by cross-validated RMSE on the training split rather than by test performance. With 7-day strength, the selected stacking configuration (ElasticNet meta-learner) reached a test RMSE of 3.197 MPa ($R^2 = 0.846$), 0.23% below the Linear Regression baseline (3.205 MPa); without 7-day strength, the selected configuration (Bayesian Ridge) reached 6.365 MPa ($R^2 = 0.391$), 0.17% below the baseline (6.376 MPa); and on UCI the selected configuration (ElasticNet) matched the baseline (3.717 MPa). The best single base learner (XGBoost in all three cases) was 3.4%, 0.9% and 1.4% worse than the baseline, respectively. The four linear meta-learners were within 0.3% of one another on every dataset. The shallow nonlinear XGBoost meta-learner was worse than the linear ones on both field datasets and was the worst stacking variant under cross-validation on all three datasets, although on the UCI test split it had the lowest RMSE (3.687 MPa). Cross-validated RMSE (mean ± SD over folds) was 3.501 ± 0.129, 3.509 ± 0.122 and 3.574 ± 0.090 for the selected stacking, baseline and best single model with 7-day strength; 6.706 ± 0.231, 6.709 ± 0.229 and 6.723 ± 0.194 without it; and 4.488 ± 0.479, 4.488 ± 0.480 and 4.454 ± 0.452 on UCI, where the single XGBoost model therefore had the lowest cross-validated RMSE. The margins among meta-learners are thus negligible, whereas the advantage of stacking over the best single learner is small, not consistent across evaluation schemes and, as Section 4.3 shows, statistically fragile.

## 4.3. Statistical Significance of Model Comparisons

Table 5 reports Wilcoxon signed-rank and Diebold–Mariano test results for the comparisons implied by Table 4, testing directly whether the small margins observed there are distinguishable from random variation or fall within it.

Table 5. Paired tests on out-of-fold errors (n = 728 per scenario): raw p-values with Holm-adjusted values in parentheses (six comparisons), and rank-biserial effect size r (positive: the first model has smaller absolute errors)

| Dataset | Comparison | Wilcoxon p (Holm) | Effect size r | DM, squared loss p (Holm) | DM, absolute loss p (Holm) |
|---|---|---|---|---|---|
| with 7-day | Stacking (ElasticNet) vs. single XGBoost | 0.2063 (1.000) | +0.054 | 0.0171 (0.085) | 0.1367 (0.706) |
| with 7-day | Stacking (ElasticNet) vs. baseline (Linear) | 0.8010 (1.000) | +0.011 | 0.3786 (1.000) | 0.4325 (1.000) |

| | | | | | |
|---|---|---|---|---|---|
| with 7-day | Single XGBoost vs. baseline (Linear) | 0.0945 (0.567) | −0.072 | 0.0055 (0.033) | 0.1177 (0.706) |
| without 7-day | Stacking (Bayesian Ridge) vs. single XGBoost | 0.7942 (1.000) | +0.011 | 0.6503 (1.000) | 0.6424 (1.000) |
| without 7-day | Stacking (Bayesian Ridge) vs. baseline (Linear) | 0.2424 (1.000) | +0.050 | 0.2604 (1.000) | 0.1297 (0.706) |
| without 7-day | Single XGBoost vs. baseline (Linear) | 0.8206 (1.000) | −0.010 | 0.7086 (1.000) | 0.7333 (1.000) |

Tests were run on out-of-fold errors from the training split (n = 728 per field scenario) for six comparisons across the two scenarios, and p-values were adjusted with Holm's procedure (Table 5). No alternative stacking configuration differed from the Linear Regression baseline (adjusted $p \geq 0.71$ in every test and scenario), and no stacking-versus-single comparison remained significant after adjustment; the closest was the Diebold–Mariano test on squared errors with the 7-day test (raw p = 0.017, adjusted p = 0.085). The only comparison that remained significant was the Diebold–Mariano test on squared errors for the best single learner (XGBoost) against the baseline in the with-7-day scenario (raw p = 0.0055, adjusted p = 0.033), in which the baseline stacking model had the smaller errors. The corresponding Wilcoxon test (raw p = 0.0945) and the Diebold–Mariano test on absolute errors (raw p = 0.118) were not significant, which is consistent with, though it does not prove, a difference driven by a small number of large errors. All Wilcoxon effect sizes were negligible ($|r| < 0.08$). Taken together with Table 4, these results do not support the premise that a more elaborate meta-learner materially improves prediction accuracy once base learners are properly tuned, and they show only a small and fragile advantage of stacking over the best single learner; this is reported as a finding rather than treated as a shortcoming to be engineered away.

## 4.4. Uncertainty Quantification via Conformal Prediction

Table 6 reports empirical coverage and average interval width for the split-conformal, standard CV+ and conservative Jackknife+ intervals, at 90% and 95% nominal confidence, for each field scenario.

Table 6. Conformal prediction coverage and average interval width for the selected stacking model, test set (n = 182)

| Dataset | Method | Confidence | Coverage [Wilson 95% CI] | Avg. width (MPa) |
|---|---|---|---|---|
| with 7-day | Split conformal | 90% | 0.934 [0.888, 0.962] | 11.31 |
| with 7-day | CV+ (standard) | 90% | 0.912 [0.862, 0.945] | 11.50 |
| with 7-day | Jackknife+ (conservative) | 90% | 0.967 [0.930, 0.985] | 14.63 |
| with 7-day | Split conformal | 95% | 0.962 [0.923, 0.981] | 13.73 |
| with 7-day | CV+ (standard) | 95% | 0.967 [0.930, 0.985] | 14.78 |

| | | | | |
|---|---|---|---|---|
| with 7-day | Jackknife+ (conservative) | 95% | 0.989 [0.961, 0.997] | 17.53 |
| without 7-day | Split conformal | 90% | 0.934 [0.888, 0.962] | 23.12 |
| without 7-day | CV+ (standard) | 90% | 0.929 [0.882, 0.958] | 22.39 |
| without 7-day | Jackknife+ (conservative) | 90% | 0.962 [0.923, 0.981] | 26.83 |
| without 7-day | Split conformal | 95% | 0.978 [0.945, 0.991] | 28.93 |
| without 7-day | CV+ (standard) | 95% | 0.967 [0.930, 0.985] | 27.11 |
| without 7-day | Jackknife+ (conservative) | 95% | 0.989 [0.961, 0.997] | 31.65 |

With 182 test records, the standard error of an empirical coverage value is about 2.2 percentage points, and Wilson 95% intervals are reported for each coverage estimate. At 90% nominal confidence in the with-7-day scenario, split conformal prediction achieved a coverage of 0.934 with an average width of 11.31 MPa; the conservative Jackknife+ variant over-covered significantly (0.967, Wilson 95% interval 0.930–0.985, width 14.63 MPa); and standard CV+ was consistent with the nominal level (0.912, interval 0.862–0.945) with a width of 11.50 MPa, 21% narrower than the conservative variant. The without-7-day scenario shows the same pattern (conservative variant: coverage 0.962, width 26.83 MPa; standard CV+: coverage 0.929, width 22.39 MPa). The practically important result is the difference between scenarios: the CV+ interval widens from 11.50 to 22.39 MPa at 90% confidence, a ratio of about 1.95, and at 95% confidence from 14.78 to 27.11 MPa (ratio 1.83; CV+ coverage 0.967 in both scenarios, split conformal 0.962 and 0.978, conservative variant 0.989). The Linear Regression baseline gave nearly identical results (CV+ widths of 11.64 and 22.37 MPa at 90%). Figure 2 shows this directly, plotting intervals against actual 28-day strength for both scenarios; the substantially wider error bars in the without-7-day panel are visible without any statistical test.

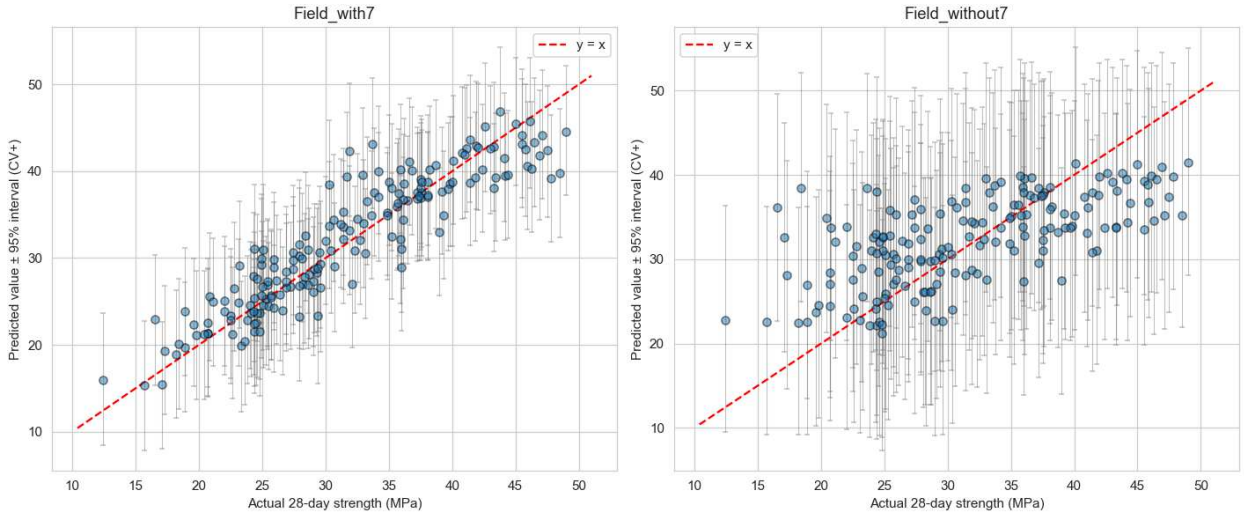



Figure 2. Standard CV+ 95% prediction intervals versus true 28-day strength, with (left) and without (right) 7-day strength as input.

## 4.5. Feature Engineering Ablation

Table 7 reports cross-validated performance (repeated 5×5 cross-validation) under four feature configurations, isolating the individual and combined contribution of cyclical month encoding and the cement-to-strength ratio introduced in Section 3.1.

Table 7. Effect of feature engineering choices on cross-validated performance (repeated 5×5 cross-validation: mean $R^2$, MAE and RMSE ± SD across folds, and RMSE difference from the full configuration with 95% CI)

| Scenario | Feature configuration | $R^2$ | MAE | RMSE ± SD | ΔRMSE vs. full [95% CI] |
|---|---|---|---|---|---|
| Field (with 7-day) | No engineered features (raw month) | 0.831 | 2.651 | 3.520 ± 0.270 | +0.019 [−0.059, 0.097] |
| Field (with 7-day) | + cyclical month only | 0.832 | 2.647 | 3.512 ± 0.296 | +0.011 [−0.024, 0.046] |
| Field (with 7-day) | + cement/strength ratio only | 0.831 | 2.648 | 3.521 ± 0.285 | +0.020 [−0.059, 0.099] |
| Field (with 7-day) | + both (used in this study) | 0.833 | 2.638 | 3.501 ± 0.297 | reference |
| Field (without 7-day) | No engineered features (raw month) | 0.386 | 5.316 | 6.729 ± 0.377 | +0.025 [−0.050, 0.101] |
| Field (without 7-day) | + cyclical month only | 0.386 | 5.329 | 6.727 ± 0.386 | +0.023 [−0.053, 0.099] |
| Field (without 7-day) | + cement/strength ratio only | 0.389 | 5.287 | 6.712 ± 0.386 | +0.007 [−0.061, 0.075] |
| Field (without 7-day) | + both (used in this study) | 0.390 | 5.303 | 6.704 ± 0.384 | reference |

Table 7 shows no detectable benefit of either engineered feature. The full configuration had the lowest mean RMSE in both scenarios, but in every comparison the 95% confidence interval of the RMSE difference (repeated 5×5 cross-validation) included zero; the point estimates differed by at most 0.025 MPa, far below the fold-to-fold standard deviation of 0.27–0.39 MPa, and the intervals extended between about 0.02 and 0.10 MPa on either side. The differences in Table 7 should therefore not be read as evidence that either engineered feature improves accuracy. What the ablation supports is the absence of a detectable penalty, together with an interpretability advantage of the cyclical month encoding, which treats Esfand and Farvardin as neighboring months and allows the seasonal structure examined in Section 5.2 to be expressed directly.

## 4.6. Error Analysis by Operating Condition

Table 8 reports mean absolute error and root-mean-squared error on the test set, segmented by strength class, slump range, and casting season, to characterize where the aggregate accuracy reported in Table 4 breaks down; the MAE column includes 95% confidence intervals. Strength segments are tertiles of the true 28-day strength in the test set, slump segments are split at the test-set median slump, and seasons are defined by casting month in the Solar Hijri calendar (months 1–6, Farvardin to Shahrivar, as the warm half of the year and months 7–12 as the cold half).

Table 8. Segment-wise error analysis (test set)

| Dataset | Segment type | Segment | N | MAE [95% CI] | RMSE |
|---|---|---|---|---|---|
| with 7-day | Strength | Low | 62 | 2.09 [1.69, 2.55] | 2.69 |
| with 7-day | Strength | Medium | 60 | 2.68 [2.14, 3.29] | 3.54 |
| with 7-day | Strength | High | 60 | 2.56 [2.05, 3.12] | 3.31 |
| with 7-day | Slump | High | 77 | 2.37 [1.93, 2.83] | 3.10 |
| with 7-day | Slump | Low | 105 | 2.49 [2.09, 2.91] | 3.26 |
| with 7-day | Season | Cold | 89 | 2.48 [2.05, 2.91] | 3.23 |
| with 7-day | Season | Warm | 93 | 2.40 [2.02, 2.82] | 3.16 |
| without 7-day | Strength | Low | 62 | 6.20 [5.09, 7.50] | 7.81 |
| without 7-day | Strength | Medium | 60 | 3.37 [2.78, 4.00] | 4.17 |
| without 7-day | Strength | High | 60 | 5.51 [4.64, 6.42] | 6.52 |
| without 7-day | Slump | High | 77 | 4.58 [3.74, 5.44] | 5.99 |
| without 7-day | Slump | Low | 105 | 5.38 [4.66, 6.10] | 6.63 |
| without 7-day | Season | Cold | 89 | 5.15 [4.31, 6.07] | 6.59 |
| without 7-day | Season | Warm | 93 | 4.94 [4.21, 5.68] | 6.15 |

Seasonal differences in error were negligible and their confidence intervals overlapped almost completely (MAE 2.48 MPa in the cold half of the year versus 2.40 MPa in the warm half with the 7-day test, and 5.15 versus 4.94 MPa without it). With the 7-day test, the confidence intervals of the three strength segments also overlapped (MAE 2.09, 2.68 and 2.56 MPa for low, medium and high strength). Without it, the error was U-shaped across strength classes (MAE 6.20, 3.37 and 5.51 MPa), with the medium class showing lower error than both extremes (non-overlapping confidence intervals), whereas the difference between the two extremes was not distinguishable

from zero. Low-slump concrete had a higher MAE than high-slump concrete only without the 7-day test (5.38 versus 4.58 MPa). None of the six pairwise contrasts (high minus low strength, low minus high slump, and cold minus warm season, in each scenario) had a 95% confidence interval excluding zero; for season, the contrasts were +0.08 MPa [−0.52, 0.68] with the 7-day test and +0.22 MPa [−0.89, 1.30] without it. Because the strength segments are defined from the true outcome, errors in the tails are mechanically inflated (regression to the mean), and the U-shape should not by itself be read as systematic under- or over-prediction; Figure 3 shows residuals that increase with true strength in both scenarios, as expected when predictions are shrunk toward the mean.

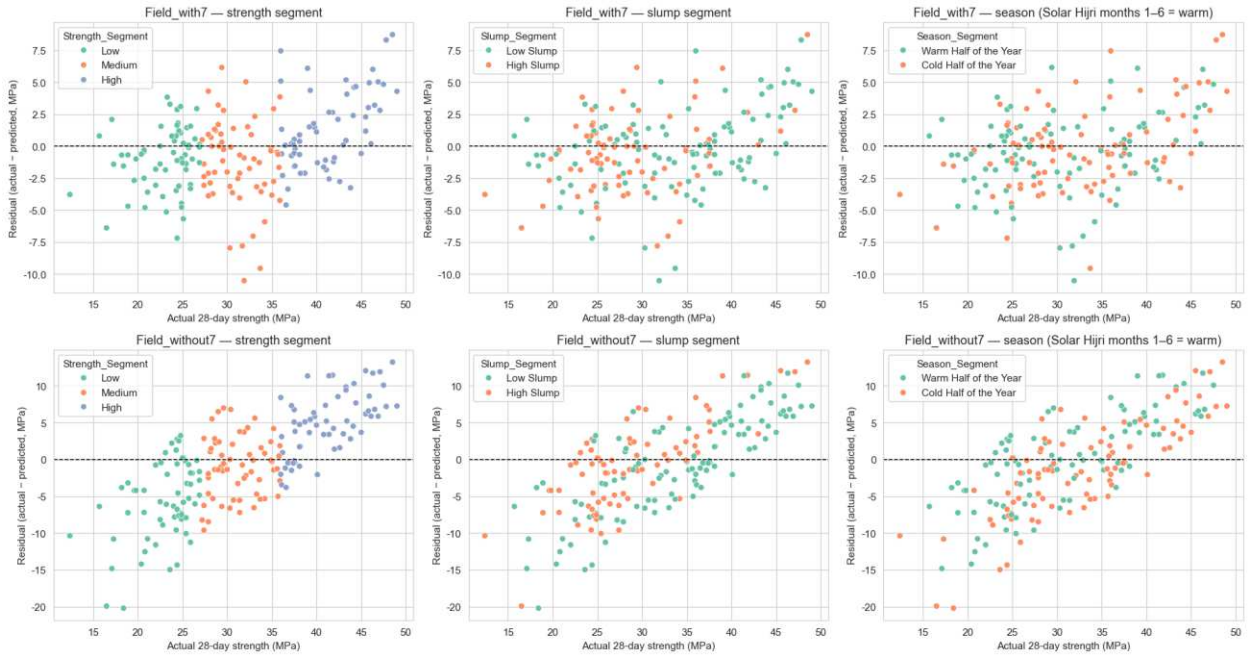



Figure 3. Test-set residuals segmented by strength class, slump range, and casting season, with (top) and without (bottom) 7-day strength as input.

## 4.7. Generalization to the UCI Benchmark

The UCI Concrete Compressive Strength dataset was processed using the same modeling and validation procedures (the same five meta-learner variants and the same cross-validation protocol), with dataset-specific feature preprocessing, since it has a substantially different feature set: it records cement, blast-furnace slag, fly ash, water, superplasticizer, and coarse and fine aggregate contents together with curing age, of which only cement is also present in the field data, while omitting slump, placement location, casting month, characteristic strength, and 7-day strength, the first three of which are central to the field analysis in Sections 4.5 and 4.6. Its target is compressive strength at ages from 1 to 365 days (age is an input feature), not the 28-day strength predicted in the field data. This makes it a test of whether the behavior of the modeling procedure, rather than any property specific to the field dataset, is reproduced on a second dataset; it is not a validation of the field-trained model.

On the held-out test split, the stacking configuration selected by cross-validation (ElasticNet meta-learner) matched the Linear Regression baseline (RMSE 3.717 MPa), whereas the best single model (XGBoost) was 1.4% worse (3.770 MPa); under cross-validation, however, the single XGBoost model had the lower RMSE (4.454 versus 4.488). The UCI results therefore show the same qualitative pattern as the field data: alternative meta-learners produce no meaningful gain over a linear combination, and any advantage of stacking over the best single learner is small and depends on the evaluation scheme. This consistency across two datasets supports, but does not establish, the interpretation of the finding in Section 4.3 as more than an artifact of the field dataset.

At the same time, the absolute accuracy gap between UCI and the field data is informative in its own right. $R^2$ on UCI (about 0.95) substantially exceeds both field scenarios (Table 4), which is consistent with UCI's laboratory provenance and, above all, with its richer inputs: it records the full mixture composition and curing age, whereas the field dataset carries only routinely available site information and the measurement variability of active construction sites. The comparison therefore cannot separate the effect of laboratory control from the effect of input information, and the two targets also differ (strength at 1–120 days versus 28-day strength). The gap is not a shortcoming of the modeling approach, but it should be read as an illustration of the accuracy penalty when only routinely available information is used, not as a pure laboratory-versus-field effect.

## 5. Discussion

The results reported in Section 4 raise three questions that merit direct discussion: why more elaborate meta-learners failed to improve on a simple linear combination (Section 5.1); what the SHAP interaction analysis reveals once this null result is taken at face value rather than engineered away (Section 5.2); and what these findings, together with the uncertainty and error analyses of Sections 4.4 and 4.6, jointly imply for engineering practice (Section 5.3).

### 5.1. Interpreting the Meta-Learner Comparison

The absence of a statistically robust advantage for any alternative meta-learner over Linear Regression (all linear meta-learners fell within 0.3% of one another, and the shallow nonlinear meta-learner was worse) warrants explicit interpretation rather than being treated merely as a lack of improvement. It suggests that, once XGBoost, SVR, and MLP are each individually well-tuned, their out-of-fold predictions are sufficiently correlated that a simple linear combination already captures most of the achievable gain from combining them; additional flexibility in the combination rule has little left to exploit. This has a practical implication beyond this dataset: for field-collected concrete data of comparable size (hundreds to low thousands of records), the marginal return on meta-learner sophistication may be smaller than the ensemble-learning literature, largely developed on larger benchmark datasets, would suggest.

### 5.2. SHAP Interaction Analysis: Seasonal Modulation Effects

Interpretability analysis in this study is deliberately scoped to the XGBoost base learner rather than extended to SVR and MLP. TreeExplainer computes exact Shapley values for tree ensembles in polynomial time; SVR and MLP are not tree-based and would instead require an approximate method such as KernelExplainer, whose values are computed differently and are not directly comparable in scale or precision to TreeExplainer's exact decomposition. Rather than report interaction values from two different explainer families side by side, SHAP interaction analysis was performed on the tuned XGBoost base learner using TreeExplainer, which yields interaction values that sum to the exact prediction decomposition for tree ensembles [33, 34]. Interaction values were computed on a random sample of 300 training records for the XGBoost base learner refit on the training split with its tuned hyperparameters, and Table 9 reports mean absolute off-diagonal interaction entries. As a preliminary check, interaction values involving raw cement content were small in the with-7-day scenario (at most 0.017), and the symmetry of the interaction matrix was verified; the cement/strength ratio, in contrast, entered several of the strongest pairs without the 7-day test (Table 9b).

The interaction pairs that do carry weight are informative. In the with-7-day scenario, the three strongest pairs are cos_month × 7-day strength (mean |interaction| = 0.215), sin_month × 7-day strength (0.108) and sin_month × cos_month (0.106; two components of a single variable, so this pair reflects the cyclical encoding rather than an interaction between distinct variables), followed by Slump × 7-day strength (0.043); the remaining pairs are at most 0.039 (Table 9). This indicates that the predictive value of the 7-day test is not constant across the year: the relationship between early and late strength appears to be modulated by casting season, plausibly reflecting the known sensitivity of cement hydration rate to curing temperature. In the without-7-day scenario, where this early signal is unavailable, the strongest pair is Slump × cos_month (0.268; Table 9b and Figure 4, left panel), followed by Slump × cement/strength ratio (0.224), characteristic strength × ratio (0.190) and Placement Location_1 × Slump (0.179). Because cement content varies in the filtered data, the ratio enters five of the eight strongest pairs, which suggests that it acts jointly with slump and design strength class rather than as an independent cement signal. For the slump–season interaction, a diagnostic plot showed that the interaction values cluster by casting month rather than varying continuously, with the sign depending on slump level (negative in months 4–8, Tir to Aban, and positive in month 12, Esfand, at low slump), so the pattern is better described as a month-specific effect than as a bimodal distribution; it would be difficult to detect from aggregate accuracy metrics alone.

Table 9. Mean absolute SHAP interaction values, with-7-day scenario (eight strongest pairs)

| Feature pair | Mean \|interaction\| |
|---|---|
| cos_month × Day Strength 7 | 0.215 |
| sin_month × Day Strength 7 | 0.108 |
| sin_month × cos_month | 0.106 |

| | |
|---|---|
| Slump × Day Strength 7 | 0.043 |
| Placement Location_1 × cos_month | 0.039 |
| Slump × cos_month | 0.035 |
| cos_month × cement/strength ratio | 0.033 |
| Placement Location_2 × Day Strength 7 | 0.030 |

Table 9b. Mean absolute SHAP interaction values, without-7-day scenario (eight strongest pairs)

| Feature pair | Mean \|interaction\| |
|---|---|
| Slump × cos_month | 0.268 |
| Slump × cement/strength ratio | 0.224 |
| Characteristic Strength × cement/strength ratio | 0.190 |
| Placement Location_1 × Slump | 0.179 |
| Placement Location_1 × cement/strength ratio | 0.160 |
| cos_month × cement/strength ratio | 0.148 |
| Slump × sin_month | 0.135 |
| sin_month × cement/strength ratio | 0.117 |

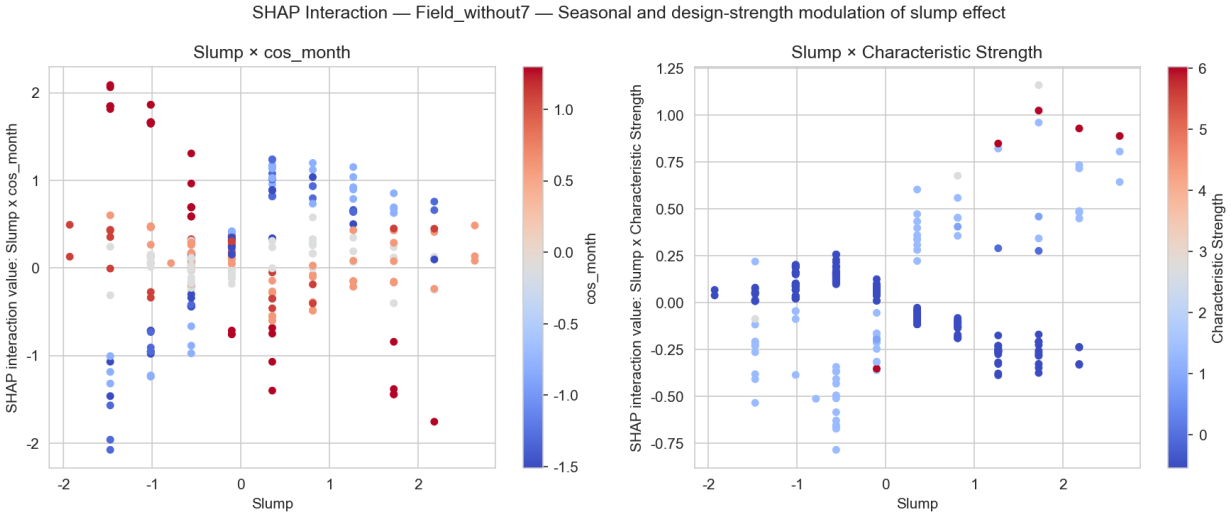



Figure 4. SHAP interaction values, without-7-day scenario: Slump × cos_month (left) and Slump × Characteristic Strength (right).

These findings reveal a season-dependent interaction pattern that is physically consistent with the temperature sensitivity of cement hydration, although temperature itself was not directly measured in the present dataset; the pattern emerges from a calendar variable already present in routine site records, without requiring any additional physical measurement or climatic input, an outcome consistent with, though more modest than, what a dedicated physics-informed model would aim

to achieve. Because the specimens remain at the site for at least 24 h before standard water curing (Section 4.1), the seasonal signal most plausibly originates in the ambient conditions of this initial period and in the temperature of the fresh concrete at placement, rather than in the later curing regime; neither was measured.

### 5.3. Engineering and Structural Implications

Taken together, these results support a more appropriately calibrated interpretation of the proposed framework than a single accuracy figure would suggest. When a 7-day test is available, the framework provides both a reasonably accurate point estimate and a prediction interval narrow enough to inform routine decisions such as formwork timing. When it is not, the point estimate remains usable as a rough guide, but the corresponding interval, nearly twice as wide, signals that this estimate carries substantially less information, and decisions relying on it should incorporate a wider safety margin or additional inspection rather than treating it as equivalent to a 7-day-informed prediction. The segment-wise results add a further caution: without the 7-day test, errors are largest at the two extremes of the strength range, whereas no detectable difference in error was found between the warm and cold halves of the year. None of this is intended to substitute for standard testing or code-mandated procedures; the framework's role is to narrow uncertainty during the interval before 28-day results are available, and to make explicit, rather than implicit, how much that narrowing can be trusted under different site conditions.

## 6. Limitations of the Study

The field data used here were collected exclusively from projects in northern Iran, under a moderate and humid climate; the seasonal interaction effects reported in Section 5.2, in particular, may not generalize to markedly different climates without re-estimation. As with any field-collected dataset, the records reflect a degree of measurement variability arising from differences in site practice and testing conditions that a laboratory dataset would not exhibit, and this should be kept in mind when interpreting reported accuracy. The feature set was deliberately restricted to variables that are routinely available on most construction sites; richer information (detailed curing conditions, ambient temperature and humidity during the initial on-site period, concrete temperature at placement, or admixture chemistry) was not collected, and its absence is a plausible contributor to the comparatively low accuracy observed in the without-7-day scenario. Each strength value is the mean of two (7-day) or three (28-day) cubes, so testing variability is embedded in both the target and the 7-day input and sets a floor on achievable accuracy; moreover, records may be clustered by project, laboratory or casting date, and because the train/test split was random, correlated records may fall on both sides of the split, which could make the reported accuracy slightly optimistic. In addition, cement content varied little within the dataset (350 $kg/m^3$ in about 83% of the retained records), so the data cannot support conclusions about the effect of cement content on strength. Finally, the meta-learner and statistical-testing results reported here reflect a single dataset of moderate size (910 records per scenario); while the qualitative conclusion, that ensemble sophistication offers limited marginal benefit once base learners are well tuned, is

supported, with caveats, by the independent UCI benchmark, its magnitude should not be assumed to hold for substantially larger or differently structured datasets.

## 7. Conclusion

This study developed a hybrid, interpretable machine-learning framework for predicting 28-day compressive strength from field-collected construction data, evaluated with and without an early-age (7-day) strength measurement. Contrary to the premise that a more elaborate ensemble combination rule would be the main source of improvement, formal statistical testing showed no significant advantage for any of several alternative meta-learners over a simple linear combination, once base learners were properly tuned, a finding reported here rather than obscured. The more consequential contribution is the direct quantification of predictive uncertainty: conformal prediction intervals nearly double in width when the 7-day test is unavailable, and segment-wise analysis shows that, without the 7-day test, errors are largest at the extremes of the strength range, while no difference in error between the warm and cold halves of the year was detectable. SHAP interaction analysis further revealed that casting season interacts substantially with both early-age strength and slump, a pattern physically consistent with the temperature sensitivity of cement hydration, although temperature itself was not directly measured, identified here using only a calendar variable without any additional physical or climatic measurement. Results on the independent UCI benchmark show a similar pattern of marginal meta-learner gains, which is consistent with, but does not establish, the generality of this finding. Together, these results reframe the practical contribution of this work: rather than a marginally more accurate predictor, the proposed framework offers engineers a calibrated sense of how much a given prediction can be trusted, and under which site conditions that trust should be discounted.

## Data Availability Statement

The UCI Concrete Compressive Strength dataset used for the independent benchmark is publicly available from the UCI Machine Learning Repository [35]. The field dataset was provided by the concrete testing laboratories and is available from the corresponding author on reasonable request, subject to the permission of the data owners.

## Disclosure Statement

No potential conflict of interest was reported by the author(s).

## Declaration of Generative AI in Scientific Writing

After preparation of the first draft, large language models were used to assist with statistical analysis, data visualization, and improving the readability of the manuscript. The authors carefully reviewed and verified all reported results and take full responsibility for the content.

## Research Funding

This research did not receive funding.

## Data Availability Statement

The UCI Concrete Compressive Strength dataset used for independent benchmarking is publicly available from the UCI Machine Learning Repository at https://archive.ics.uci.edu/dataset/165/concrete+compressive+strength. The field dataset used in this study was provided by concrete testing laboratories and is available from the corresponding author upon reasonable request, subject to the permission of the data owners.

## References

[1] W. B. Inqiad, M. S. Siddique, S. S. Alarifi, M. J. Butt, T. Najeh, and Y. Gamil, "Comparative analysis of various machine learning algorithms to predict 28-day compressive strength of Self-compacting concrete," Heliyon, vol. 9, no. 11, 2023.

[2] A. Mumuni and F. Mumuni, "Automated data processing and feature engineering for deep learning and big data applications: a survey," Journal of Information and Intelligence, vol. 3, no. 2, pp. 113-153, 2025.

[3] I.-C. Yeh, "Modeling of strength of high-performance concrete using artificial neural networks," Cement and Concrete Research, vol. 28, no. 12, pp. 1797-1808, 1998.

[4] M. M. Hasan and A. Kabir, "Prediction of compressive strength of concrete from early age test result," in 4th Annual Paper Meet and 1st Civil Engineering Congress, 2011, vol. 2011, pp. 978-984.

[5] A. K. Palanisamy et al., "Prediction of early age compressive strength of concrete using machine learning," Scientific Reports, 2025.

[6] M. Hasan and A. Kabir, "Early age tests to predict 28 days compressive strength of concrete," Caspian Journal of Applied Sciences Research, vol. 2, pp. 234-241, 2013.

[7] Q.-F. Li and Z.-M. Song, "High-performance concrete strength prediction based on ensemble learning," Construction and Building Materials, vol. 324, p. 126694, 2022.

[8] A.-D. Pham, N.-T. Ngo, Q.-T. Nguyen, and N.-S. Truong, "Hybrid machine learning for predicting strength of sustainable concrete," Soft Computing, vol. 24, no. 19, 2020.

[9] N. Su, S. Guo, C. Shi, and D. Zhu, "Predictions of mechanical properties of Fiber Reinforced Concrete using ensemble learning models," Journal of Building Engineering, vol. 98, p. 110990, 2024.

[10] A. K. Sah and Y.-M. Hong, "Performance comparison of machine learning models for concrete compressive strength prediction," Materials, vol. 17, no. 9, p. 2075, 2024.

[11] M. K. Elshaarawy, A. K. Hamed, and M. M. Alsaadawi, "Hybrid gradient boosting models for concrete compressive strength classification and prediction," International Journal of Machine Learning and Cybernetics, vol. 16, no. 11, pp. 8857-8891, 2025.

[12] K. Khan, W. Ahmad, M. N. Amin, and A. Ahmad, "A systematic review of the research development on the application of machine learning for concrete," Materials, vol. 15, no. 13, p. 4512, 2022.

[13] S. I. Hassan, S. A. Syed, S. W. Ali, H. Zahid, S. Tariq, and M. M. Alam, "Systematic literature review on the application of machine learning for the prediction of properties of different types of concrete," PeerJ Computer Science, vol. 10, p. e1853, 2024.

[14] A. Ahmad, K. A. Ostrowski, M. Maślak, F. Farooq, I. Mehmood, and A. Nafees, "Comparative study of supervised machine learning algorithms for predicting the compressive strength of concrete at high temperature," Materials, vol. 14, no. 15, p. 4222, 2021.

[15] Y. Zhang, W. Ren, Y. Chen, Y. Mi, J. Lei, and L. Sun, "Predicting the compressive strength of high-performance concrete using an interpretable machine learning model," Scientific Reports, vol. 14, no. 1, p. 28346, 2024.

[16] Kaggle. 2023. Concrete Compressive Strength Dataset. [online] Available at: https://www.kaggle.com/datasets.

[17] S. Pandey, S. Paudel, K. Devkota, K. Kshetri, and P. G. Asteris, "Machine learning unveils the complex nonlinearity of concrete materials' uniaxial compressive strength," International Journal of Construction Management, vol. 25, no. 6, pp. 635-649, 2025.

[18] H. Fu, X. Zhou, P. Xu, and D. Sun, "Prediction of Compressive Strength of Concrete Using Explainable Machine Learning Models," Materials, vol. 18, no. 21, p. 5009, 2025.

[19] S. O. Abioye, Y. O. Babatunde, O. A. Abikoye, A. N. Shaibu, and B. J. Bankole, "Optimized machine learning algorithms with SHAP analysis for predicting compressive strength in high-performance concrete," AI in Civil Engineering, vol. 4, no. 1, p. 16, 2025.

[20] S. Dhengare, U. Waghe, G. Yenurkar, and A. Shyamala, "A comprehensive model for concrete strength prediction using advanced learning techniques," Discover Applied Sciences, vol. 7, no. 6, p. 551, 2025.

[21] F. R. Ododo and R. R. Sadiq, "A Review of Outlier Detection Techniques in Cybersecurity: A Machine Learning Perspective," Journal of Science Innovation and Technology Research, 2025.

[22] T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794, 2016.

[23] S. Hakkal and A. Ait Lahcen, "XGBoost to enhance learner performance prediction," Computers and Education: Artificial Intelligence, vol. 7, p. 100254, 2024.

[24] H. Xue, Q. Yang, and S. Chen, "SVM: Support vector machines," in The Top Ten Algorithms in Data Mining: Chapman and Hall/CRC, 2009, pp. 51-74.

[25] D. Valero-Carreras, J. Alcaraz, and M. Landete, "Comparing two SVM models through different metrics based on the confusion matrix," Computers & Operations Research, vol. 152, p. 106131, 2023.

[26] M.-C. Popescu, V. E. Balas, L. Perescu-Popescu, and N. Mastorakis, "Multilayer perceptron and neural networks," WSEAS Transactions on Circuits and Systems, vol. 8, no. 7, pp. 579-588, 2009.

[27] S.-C. Wang, "Artificial neural network," in Interdisciplinary Computing in Java Programming: Springer, 2003, pp. 81-100.

[28] A. Botchkarev, "Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology," arXiv preprint arXiv:1809.03006, 2018.

[29] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," Statistics and Computing, vol. 14, no. 3, pp. 199-222, 2004.

[30] C. M. Bishop and N. M. Nasrabadi, Pattern Recognition and Machine Learning (no. 4). Springer, 2006.

[31] Y. Bengio, I. Goodfellow, and A. Courville, Deep Learning. MIT Press, 2017.

[32] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," Annals of Statistics, pp. 1189-1232, 2001.

[33] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," Advances in Neural Information Processing Systems, vol. 30, 2017.

[34] C. Molnar, Interpretable Machine Learning. Lulu.com, 2020.

[35] I.-C. Yeh. Concrete Compressive Strength. UCI Machine Learning Repository, doi: https://doi.org/10.24432/C5PK67.

[36] V. Vovk, A. Gammerman, and G. Shafer, Algorithmic Learning in a Random World. Springer, 2005.

[37] F. Wilcoxon, "Individual comparisons by ranking methods," Biometrics Bulletin, vol. 1, no. 6, pp. 80-83, 1945.

[38] F. X. Diebold and R. S. Mariano, "Comparing predictive accuracy," Journal of Business & Economic Statistics, vol. 13, no. 3, pp. 253-263, 1995.

[39] T.-A. Nguyen, H.-B. Ly, H.-V. T. Mai, and V. Q. Tran, "Prediction of later-age concrete compressive strength using feedforward neural network," Advances in Materials Science and Engineering, vol. 2020, Article ID 9682740, 2020.

[40] F. A. Oluokun, E. G. Burdette, and J. H. Deatherage, "Early-age concrete strength prediction by maturity—another look," ACI Materials Journal, vol. 88, no. 5, pp. 584-586, 1991.

[41] E. Hüllermeier and W. Waegeman, "Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods," Machine Learning, vol. 110, pp. 457-506, 2021.

[42] T. Gneiting, F. Balabdaoui, and A. E. Raftery, "Probabilistic forecasts, calibration and sharpness," Journal of the Royal Statistical Society: Series B, vol. 69, no. 2, pp. 243-268, 2007.

[43] C. E. Rasmussen and C. K. I. Williams, Gaussian Processes for Machine Learning. MIT Press, 2006.

[44] K. Severson, O. Pfeiffer, J. Chen, K. Gong, J. Gregory, R. Goodwin, and E. Olivetti, "Amortized inference of Gaussian process hyperparameters for improved concrete strength trajectory prediction," NeurIPS 2021 Workshop on Tackling Climate Change with Machine Learning, 2021.

[45] C. L. Chan and M. Zhang, "Data-driven prediction of strength development of alkali-activated concrete with uncertainty quantification using interpretable amortised Gaussian process," Developments in the Built Environment, 2026, doi: 10.1016/j.dibe.2026.100971.

[46] B. Baten et al., "BOxCrete: A Bayesian optimization open-source AI model for concrete strength forecasting and mix optimization," arXiv:2603.21525, 2026.

[47] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: representing model uncertainty in deep learning," in Proc. 33rd International Conference on Machine Learning, PMLR vol. 48, pp. 1050-1059, 2016.

[48] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in Advances in Neural Information Processing Systems, vol. 30, 2017.

[49] T. Duan, A. Anand, D. Y. Ding, K. K. Thai, S. Basu, A. Ng, and A. Schuler, "NGBoost: natural gradient boosting for probabilistic prediction," in Proc. 37th International Conference on Machine Learning, PMLR vol. 119, pp. 2690-2700, 2020.

[50] R. Koenker and G. Bassett, "Regression quantiles," Econometrica, vol. 46, no. 1, pp. 33-50, 1978.

[51] N. Meinshausen, "Quantile regression forests," Journal of Machine Learning Research, vol. 7, pp. 983-999, 2006.

[52] Y. Romano, E. Patterson, and E. J. Candès, "Conformalized quantile regression," in Advances in Neural Information Processing Systems, vol. 32, pp. 3543-3553, 2019.

[53] J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman, "Distribution-free predictive inference for regression," Journal of the American Statistical Association, vol. 113, no. 523, pp. 1094-1111, 2018.

[54] R. F. Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani, "Predictive inference with the jackknife+," The Annals of Statistics, vol. 49, no. 1, pp. 486-507, 2021.

[55] A. N. Angelopoulos and S. Bates, "Conformal prediction: a gentle introduction," Foundations and Trends in Machine Learning, vol. 16, no. 4, pp. 494-591, 2023.

[56] P. Tamuly and V. Nava, "Machine learning based conformal predictors for uncertainty-aware compressive strength estimation of concrete," Construction and Building Materials, vol. 487, 2025.

[57] F. Mobasheri, M. Hosseinpoor, A. Yahia, and F. Pourkamali-Anaraki, "Predicting the compressive strength of self-consolidating concrete using machine learning and conformal inference," Computer Modeling in Engineering & Sciences, vol. 145, no. 3, pp. 3309-3347, 2025.

[58] J. W. Tukey, Exploratory Data Analysis. Addison-Wesley, 1977.