buildings

MDPI

Article

Data-Driven Compressive Strength Screening of Slag and Fly Ash Concrete Mixtures Before Trial Batching

Zhengjun Wang [1], Cheng Zhang [1], Shai Du [1], Xiaoyan Guo [1] and Xinyue Hou [2,*]

[1] School of Water Conservancy and Electric Power, Heilongjiang University, Harbin 150080, China
[2] Heilongjiang Province River and Lake Chief System Support Center, Harbin 150000, China
* Correspondence: 2242177@s.hlju.edu.cn

Abstract

Replacing part of the cement with blast-furnace slag or fly ash lowers the clinker content of concrete, but such mixtures often gain early strength more slowly depending on binder composition. On site, slower early strength can delay formwork removal, and in mix design the uncertain strength gain is managed with extra trial batches and higher cement contents. Stacked ensembles predict compressive strength accurately but usually apply one weight vector to all mixtures. This study tested whether regime-specific weights offer better predictions. K-means clustering divided 1030 results of the UCI dataset into four regimes of binder composition and water-to-cement ratio, and the weights of random forest, multilayer perceptron and long short-term memory learners tuned by the sparrow search algorithm were selected for each regime on out-of-fold predictions. This cluster-conditioned stacking (CCS) was compared with six other combination rules on 20 splits grouped by mixture. CCS reached a mean test root-mean-square error (RMSE) of 5.76 MPa, significantly lower than equal weighting (6.27 MPa) but not significantly different from global regression weights (5.78 MPa) or random forest alone (5.77 MPa). Screening with the model before trial batching can avoid some failed batches, each costing about a month of retesting.

Keywords: blast-furnace slag; fly ash; compressive strength; mix design; trial batching; cluster-conditioned stacking

1. Introduction

Compressive strength is the property through which a concrete mixture enters structural design. The design compressive strength used in resistance checks and the elastic modulus used in structural analysis are both derived from it [1,2], and the selected mixture has to be shown to reach the specified strength class. This is normally verified by compression tests after 28 days of standard curing, so each trial batch adds about a month to mix qualification, and a mixture that falls short must be redesigned and tested again [3]. Because the strength of a new mixture is uncertain, it is proportioned for a target mean strength above the specified value [4], and this margin is obtained mainly through a higher cement content. Global cement production exceeds 4 billion metric tons per year and accounts for roughly 8% of anthropogenic CO_2 emissions [5]; one of the most effective routes to lower emissions is to replace part of the clinker with supplementary cementitious materials (SCMs) such as ground granulated blast-furnace slag and fly ash [6]. Mixtures with high SCM contents usually gain strength more slowly at early ages, and their strength at a given age depends more strongly on binder composition and curing than that of plain Portland cement concrete [7]. These are therefore the mixtures whose strength is hardest to

Check for updates

Academic Editor: Antonio Caggiano

Received: 9 October 2026
Revised: 26 October 2026
Accepted: 27 October 2026
Published: 30 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

anticipate, and an excessive cement margin erodes the CO_2 saving they are meant to deliver. On site, formwork can be removed only after the concrete has reached a specified early-age strength, so slower strength development can lengthen the construction cycle [8]. Reliable prediction of compressive strength from mixture proportions can reduce the number of trial batches and allow candidate mixtures to be screened against the specified class at the design stage [3,9]. Such a predictor is also the core of computational mix optimization, in which strength is traded against cost and embodied CO_2 [10,11], so its accuracy limits the quality of the optimized mixtures. Abrams' law [12] describes the decrease in strength with increasing water-to-cement ratio (w/c), but it does not account for the nonlinear effects of mineral and chemical admixtures and their interactions.

Slag and fly ash are the most widely used SCMs, and current research on them follows two directions. In cement-based concrete, the supply of conventional SCMs, fly ash in particular, is becoming limited as coal-fired power plants close, and new sources are being qualified with reactivity tests that are judged by their correlation with the 28-day relative strength [13,14]. In alkali-activated or geopolymer concrete, slag and fly ash are the main precursors, and the strength depends on the activator, the chemistry of the precursors and the curing temperature, and on the mixture proportions [15,16]. For both binder types, the strength at a given age depends on more variables than the proportions that most strength datasets record.

Machine learning has become the main tool for such predictions [17]. Since Yeh [18] modeled the strength of high-performance concrete with artificial neural networks on the data that later formed the UCI concrete dataset, the field has progressed from single models to tree-based ensembles [19–21] and deep architectures [22], and it has been extended to low-carbon concretes such as ternary blended nano-engineered concrete with bio-inspired optimization [23] and alkali-activated concrete [24]. Stacking, which combines heterogeneous base learners through a meta-level rule, is among the most accurate approaches in recent studies [25–28]. Reported accuracy, however, depends strongly on the data split, the input features and the evaluation protocol. On the UCI dataset, for example, the hybrid ensemble of Asteris et al. [25] and the long short-term memory (LSTM) network of Latif [29] were evaluated with different inputs and partitions, so their accuracy values cannot be compared directly. A new combination rule can therefore be judged fairly only when all rules are evaluated on identical splits and identical base-model predictions.

Studies on particular concrete types show how the composition affects the predictability of strength. For high-performance concrete with a high volume of slag, gradient boosting gave the most accurate predictions among five boosting algorithms [30], and a partial dependence analysis of slag concrete with recycled coarse aggregate identified the water-to-binder ratio, the age, and the slag percentage as the governing inputs [31]. For ternary blends containing silica fume, the CaO and SiO_2 contents of the binder were among the most important inputs [32]. Fibers add inputs of their own, such as the fiber type, length, and aspect ratio. This is illustrated by models developed for lightweight fiber-reinforced concrete [33], steel fiber-reinforced concrete (for which the effect of the fibers on compressive strength is still debated [34]), and fiber-reinforced self-compacting concrete [35].

For geopolymer concrete, ensemble models were developed for slag-based mixtures, in which the specimen age was the most influential input [36], and explainable deep learning models of slag- and fly ash-based mixtures linked the predicted strength to the precursor chemistry and curing conditions [15,16] or were used for inverse mix design with an estimate of the global warming potential [37]. These studies indicate that a dataset recording only mixture masses, as the UCI dataset does (Section 2.2), describes slag and fly ash concretes incompletely, and that an accuracy reached for one family of mixtures cannot be assumed for another.

In these studies, as in most others, one model or one combination rule is applied to the whole sample space. Stacking models use a single global weight vector, whether it is obtained by equal averaging, error-based normalization, ordinary least squares, or least absolute shrinkage and selection operator (LASSO) regression. This implies that the relative accuracy of the base learners is the same for every mixture, an assumption that is difficult to justify for heterogeneous concrete data. The UCI dataset combines plain Portland cement mixtures with mixtures containing large amounts of slag or fly ash, tested at ages from 1 to 365 days. The hydration kinetics of SCM-rich binders differ substantially from those of Portland cement [7], and the accuracy of strength models for concretes with SCMs varies with both the algorithm and the optimization strategy [38]. A learner that is reliable for one family of mixtures may therefore be unreliable for another, and a compromise weight vector may then leave part of the achievable accuracy unused. Because slag and fly ash contents are a main source of this heterogeneity (Section 2.2), the global-weight assumption would be most restrictive for the mixtures that sustainable concrete development seeks to promote. This matters in mix design because adjacent strength classes in the normal-strength range differ by only 5 MPa [1], so even a fraction of a megapascal in prediction error affects how reliably a mixture can be assigned to its class.

Combination rules that depend on the region of the input space are well established in machine learning, from mixtures of local experts [39] to dynamic ensemble selection [40]. Reinforcement learning has been used to adapt the weights of forecasting ensembles in wind-power, load, and other time-series forecasting [41–43]. Clustering has also been used to improve concrete strength prediction, either by fitting a separate regression model to each cluster [44] or by structuring hierarchical classification and regression models [45]. To the best of our knowledge, however, clusters have not been used to condition the combination weights of a stacked model for concrete strength prediction, and such region-dependent weighting has not been evaluated under controlled conditions in which the data splits and base-model predictions are held fixed.

This study proposes cluster-conditioned stacking (CCS) to remove the global-weight assumption. K-means clustering partitions the standardized mixture space into regimes defined by binder composition and w/c, and the weight vector of each regime is selected from a discrete simplex of 66 candidate combinations by evaluating every candidate on out-of-fold predictions. The base learners are a random forest (RF), a multilayer perceptron (MLP), and an LSTM network, all tuned with the sparrow search algorithm (SSA) under a common protocol. SSA was chosen because it has been reported to converge faster than classical metaheuristics such as particle swarm optimization in benchmark tests [46], and an improved SSA has been used to tune neural networks for concrete strength prediction [47]. The shortened search used here and its consequences are described in Sections 2.7 and 4.3. The three learners have different inductive biases. RF partitions the input space with decision trees, the MLP learns a smooth global mapping, and the LSTM, used here with a single input step, acts as a gated nonlinear learner whose errors differ from those of the other two (Section 2.6). This diversity gives a regime-dependent rule room to improve on a global one. The hypothesis tested is that weights conditioned on the mixture regime predict strength more accurately than any global weight vector because the relative accuracy of the base learners changes with mixture composition. To test this hypothesis, CCS is compared with five global-weight strategies, including one global weight vector chosen by the same search, and with a constrained least-squares variant of CCS. All of them combine the same base-model predictions on 20 data splits in which the results of a mixture are kept in one subset. SHapley Additive exPlanations (SHAP) and partial dependence plots (PDP) [48,49] are used to check that the learned input effects agree with established mix-

design relationships, and the prediction error is finally expressed in terms of strength-class screening in mix design and the design quantities that depend on compressive strength.

The main contributions of this study are as follows.

(1) A cluster-conditioned weighting scheme for stacked concrete-strength models, in which the weights of the SSA-tuned base learners are selected separately for each mixture regime instead of being fixed globally, together with a test of whether this improves on global weighting. On 20 splits grouped by mixture, CCS was significantly more accurate than equal weighting but not more accurate than global weights fitted to the same predictions or than RF alone; at this data size, and with the learners trained here, the regime dependence of the learners was too small to outweigh the error of estimating a separate weight vector for each regime.

(2) A controlled evaluation protocol in which all weighting strategies share the same splits and the same base-model predictions, the weights are selected on out-of-fold predictions, and all results of a mixture are kept in one subset, so that the measured differences can be attributed to the weighting rule alone. Compared with random splits, this grouping raised the test RMSE by about 1 MPa, which shows how far random splitting overstates the accuracy for new mixtures.

(3) An explainable, mix-design-oriented interpretation of the model. SHAP and PDP show that the learned effects follow established w/c, curing-age, and slag relationships. The prediction error is translated into prediction intervals and into its effect on strength-class screening, on the strength margin required at the screening stage, and on the design strength and elastic modulus of concrete.

2. Materials and Methods

2.1. Research Design and Material System

This study addresses a design-stage problem in concrete construction: judging, before trial batches are cast, whether a candidate low-clinker mixture is likely to reach its specified strength class. The material system was chosen for this problem. Ground granulated blast-furnace slag and fly ash are among the most widely used supplementary cementitious materials for replacing part of the cement in structural concrete [6,7], and they are the two contained in the dataset used here (Section 2.2). Their contribution to strength develops differently from that of cement. Slag is latently hydraulic and reacts once it is activated by the calcium hydroxide and alkalis released by cement hydration, whereas fly ash reacts pozzolanically with calcium hydroxide [7]. Both therefore contribute to strength later than cement does, and their effect depends on the replacement level, the remaining cement content, and the curing age. For such mixtures, the strength at a given age is harder to anticipate from the water-to-cement ratio alone, and the uncertainty is usually covered by additional trial batches or a higher cement content (Section 1).

Three questions follow from this problem. The first is whether the relative accuracy of different base learners changes between regimes of mixture composition strongly enough for regime-dependent weights to predict strength more accurately than a single global weight vector. The second is whether the effects learned by the model agree with established mix-design relationships for slag and fly ash concretes. The third is what the remaining prediction error means for strength-class screening and for the design quantities derived from compressive strength. The study was organized to answer them in this order. A public dataset was used so that the results can be reproduced and related to earlier studies (Section 2.2). The data splits, base learners, and base-model predictions were held fixed while only the weighting rule was varied (Sections 2.3–2.8). The learned effects were examined with SHAP and partial dependence analyses (Sections 2.9 and 2.10), and the

prediction error was expressed as prediction intervals (Section 2.11) and in terms of strength-class screening and the design inputs of EN 1992-1-1 and ACI 318-19 (Section 4.2).

The scope of the study follows from the data. The dataset records mixture proportions, curing age, and compressive strength, but not the chemical composition or fineness of the binders or the curing temperature; strengths measured on specimens of different sizes and shapes had been converted to 15 cm cylinders when the data were compiled (Section 2.2). The unrecorded factors therefore enter the model only through their correlation with the recorded inputs; their consequences for the range of application are discussed in Section 4.3.

All analyses were implemented in Python 3.12.3 (Python Software Foundation, Wilmington, DE, USA) with the open-source packages NumPy 2.5.4, pandas 3.1.0, scikit-learn 1.9.1, SciPy 1.18.1, TensorFlow 2.21.0 (Google LLC, Mountain View, CA, USA), SHAP 0.54.1, and Matplotlib 3.11.2. No physical materials, reagents, or laboratory instruments were used in this computational study.

2.2. Dataset Description

This study used the UCI concrete compressive strength dataset [18], which contains 1030 compressive strength results between 2.33 and 82.60 MPa. Many mixtures were tested at more than one age, so these results are obtained from 428 distinct mixture proportions. The dataset is one of the most widely used benchmarks in concrete strength prediction, and its fixed feature definitions allow the results to be related to earlier studies. Its mixtures range from plain Portland cement concrete to concretes containing up to 359.40 kg/m^3 of slag or 200.10 kg/m^3 of fly ash (Table 1), so it covers both conventional and SCM-rich compositions. Of the 428 distinct mixtures, 88 use cement as the only binder, 124 combine cement with slag, 94 with fly ash, and 122 with both (Figure 1a); slag and fly ash together make up as much as 74% of the binder. According to Yeh [18], the data were assembled from 17 different sources, which used specimens of different sizes and shapes, and the strengths were converted to those of 15 cm cylinders following accepted guidelines. The type of cement, the class of fly ash, the fineness of the slag, and the type of superplasticizer are not recorded for the individual results [18]. The results cannot be grouped by material type, and the consequences of this gap are discussed in Section 4.3.

Table 1. Descriptive statistics of all variables in the dataset.

Variable	*N*	Mean	Std. Dev.	Min	25th	Median	75th	Max
w/c	1030	0.748	0.314	0.267	0.533	0.676	0.935	1.882
Blast-furnace slag (kg/m^3)	1030	73.90	86.28	0.00	0.00	22.00	142.95	359.40
Fly ash (kg/m^3)	1030	54.19	64.00	0.00	0.00	0.00	118.30	200.10
Superplasticizer (kg/m^3)	1030	6.20	5.97	0.00	0.00	6.40	10.20	32.20
Coarse aggregate (kg/m^3)	1030	972.92	77.75	801.00	932.00	968.00	1029.40	1145.00
Fine aggregate (kg/m^3)	1030	773.58	80.18	594.00	730.95	779.50	824.00	992.60
Age (days)	1030	45.66	63.17	1	7	28	56	365
Strength (MPa)	1030	35.82	16.71	2.33	23.71	34.44	46.14	82.60

The original dataset contains eight input features and one output variable. The inputs are the contents of cement, water, blast-furnace slag, fly ash, superplasticizer, coarse aggregate and fine aggregate (all in kg/m^3), and the curing age (days); the output is compressive strength (MPa). Cement and water contents were replaced by their ratio, w/c, so that the variable that governs strength in mix design enters the model explicitly [50,51]. This reduced the number of inputs from eight to seven. Because slag and fly ash are not included in the denominator, w/c here is the water-to-cement ratio and not the water-to-binder ratio used in many design provisions.

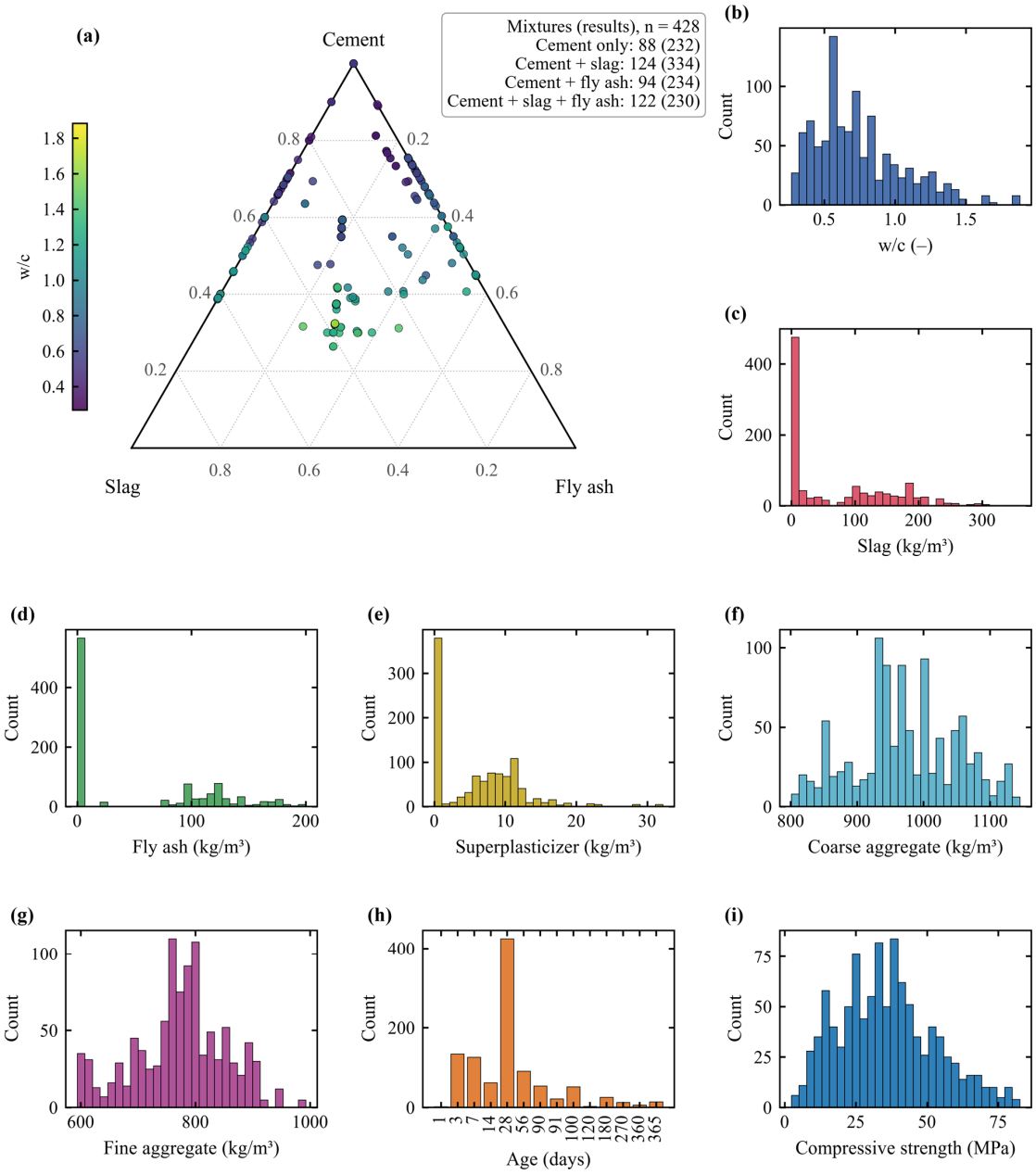


Figure 1. Composition and distributions of the UCI concrete dataset. (**a**) Binder composition of the 428 distinct mixtures, given as the mass fractions of cement, slag, and fly ash in the binder and colored by w/c; the tick labels on the left, bottom, and right edges give the fractions of cement, slag, and fly ash, respectively, and the box gives the number of mixtures and, in parentheses, of strength results in each binder group. (**b**–**h**) Distributions of the seven inputs for the 1030 strength results, with (**h**) giving the number of results at each test age, and (**i**) distribution of compressive strength.

Table 1 summarizes the descriptive statistics of the seven input features and the output variable in the dataset.

Two features of the distributions are relevant to the modeling. First, slag and fly ash are zero-inflated, with medians of 22.00 and 0.00 kg/m^3, so the dataset combines mixtures with and without these constituents. This compositional heterogeneity is the reason for conditioning the ensemble weights on mixture regimes. Second, age is right-skewed (mean 45.66 days, median 28 days), and most samples were tested between 1 and 56 days, the range of main interest in practice.

The histograms in Figure 1b–i show the same characteristics. The w/c ratio has a multimodal distribution, slag and fly ash are concentrated at zero, age clusters around the standard testing intervals (3, 7, 14, 28, and 56 days), and the aggregate contents vary within comparatively narrow ranges without zero values.

Figure 2 presents the Pearson correlation matrix (a) and the scatter plots of strength against each input (b–h). Four patterns are relevant for the modeling. (i) w/c has the strongest linear association with strength ($r = -0.501$), consistent with its role as the primary mix-design parameter. (ii) Superplasticizer ($r = 0.366$) and age ($r = 0.329$) follow. The age panel shows a saturating trend beyond about 56 days, so the linear coefficient understates the effect of age. (iii) Superplasticizer is negatively correlated with w/c ($r = -0.223$), so part of its association with strength may reflect the lower w/c of superplasticized mixtures. (iv) Slag and fly ash are negatively correlated with each other ($r = -0.324$) and only weakly correlated with strength ($r = 0.135$ and -0.106), yet in their scatter panels the strength varies widely at any given dosage. These patterns support the use of nonlinear ensemble models and motivate the SHAP analysis in Section 3.4.

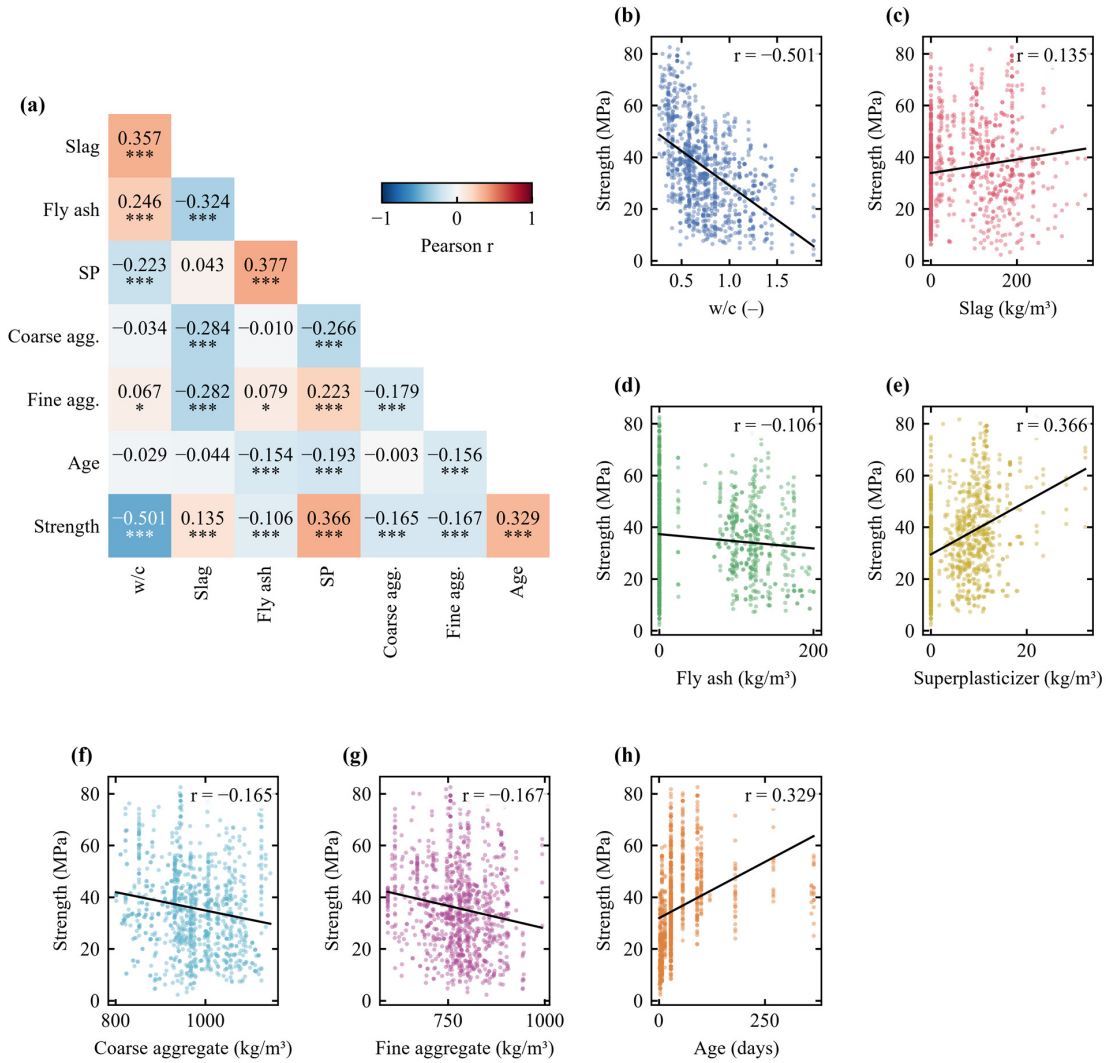


Figure 2. (**a**) Pearson correlation coefficients between all variables with their significance levels (* $p < 0.05$, *** $p < 0.001$). (**b**–**h**) Compressive strength plotted against each of the seven inputs, with the least-squares line and the correlation coefficient r ($n = 1030$). SP = superplasticizer.

2.3. Data Preprocessing and Splitting Strategy

2.3.1. Standardization

Because the input features differ widely in scale (w/c ranges from 0.267 to 1.882, whereas coarse aggregate ranges from 801 to 1145 kg/m^3), Z-score standardization (Equation (1)) was applied to all of them. Standardization parameters (μ_j, σ_j) were computed on the training set and applied to the validation and test sets to prevent data leakage [52].

$$x_j^* = \frac{x_j - \mu_j}{\sigma_j} \quad (1)$$

where x_j is the raw value of feature j and x_j^* its standardized counterpart. Although RF is insensitive to feature scaling, standardized features were uniformly used across all three base learners to maintain input consistency. Compressive strength was used in MPa without transformation.

2.3.2. Data Splitting

Because the 1030 results come from only 428 mixtures, a random division of the results places results of the same mixture, tested at different ages, in both the training and the test sets. The main evaluation therefore used splits grouped by mixture. In each of 20 grouped splits (seeds 1 to 20), the mixtures were taken in random order and assigned as a whole to the test set until it contained at least 10% of the results and then to the validation set until it contained another 10%; the remaining mixtures formed the training set. Because the mixtures have different numbers of results, the subsets contained on average 820.6 training, 104.2 validation, and 105.3 test results, and no test mixture had results in the training or validation set. The ratio of 8:1:1 leaves about 80% of the results for fitting the base learners during the search and about 90% for the final refit described below, which matters because their test error was still decreasing when the training set was enlarged to its full size (Section 3.1), while each test set still contains about 105 results from about 42 mixtures. The variation between splits is captured by repeating the division 20 times rather than by a larger test share. Five random splits of the individual results (seeds 1, 5, 10, 15, and 20) were kept as a secondary protocol to show the effect of grouping; in them, on average, 64 of the 90 mixtures in a test set also had results in the training set. A further grouped split (seed = 42) served as the development split for the single-split examples and the interpretation analyses in Section 3.

Within each split, the training set was used to fit the base learners and the validation set to evaluate the SSA fitness function (Section 2.7). The combination weights were not selected on the validation set. Instead, the development set, that is, the training and validation results, was divided into five folds of similar size, each consisting of whole mixtures. The three base learners were refitted with the SSA-selected hyperparameters on four folds and predicted the fifth, which gave out-of-fold predictions for all development results, and these predictions were used to select the weights of every strategy [53]. The base learners were then refitted on the whole development set to predict the test set, which was used only for the final evaluation (Section 2.8.2). Within a split, all strategies therefore combined the same out-of-fold and test predictions, so that differences between them are caused by the weighting rule alone. As a control, the weights were also selected on the validation predictions of learners fitted on the training set only, as in a conventional hold-out design; this control is referred to as design V and the out-of-fold design as design O.

Paired differences in test RMSE between CCS and the other strategies over the 20 grouped splits were tested with the two-sided Wilcoxon signed-rank test [54] and with the corrected resampled t-test of Nadeau and Bengio [55], which allows for the overlap of the training sets of different splits. The p-values of the seven comparisons with CCS were

adjusted with the Holm procedure [56], and the seven combination rules were compared jointly with the Friedman test [54].

2.4. Evaluation Metrics

Four regression metrics were used, namely the mean absolute error (MAE), the mean absolute percentage error (MAPE), the root-mean-square error (RMSE), and the coefficient of determination (R^2), as defined in Equations (2)–(5). MAE and RMSE are expressed in MPa, and RMSE gives greater weight to large residuals than MAE. MAPE is dimensionless but is inflated by low-strength samples. R^2 is given for comparison with earlier studies, whereas RMSE, which expresses the error in the units used in mix and structural design, is used as the primary metric.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2}$$

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100 \tag{3}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{5}$$

where n is the number of samples, y_i and $\hat{y}_i$ are the measured and predicted strengths of sample i, and $\bar{y}$ is the mean measured strength.

2.5. Overall Model Framework

Figure 3 shows the proposed framework. Its upper part contains the six modules used for model development, namely dataset preparation and feature engineering, data preprocessing and splitting, SSA hyperparameter optimization, training of the heterogeneous base learners, K-means clustering of mixture regimes, and cluster-conditioned stacking by exhaustive weight selection. The essential difference from conventional stacking lies in Module 6, where a weight vector is learned for each cluster instead of a single vector for all samples, so that the base learners are combined differently in different regions of the mixture space. The middle part shows how a new mixture is predicted, and the lower part (Modules 7 to 9) shows the controlled comparison, the model interpretation, and the use of the predictions in design. Figure 4 gives the training and prediction procedures of Module 6, and Sections 2.6–2.11 describe the individual methods.

2.6. Base Learners

Three heterogeneous models, RF, MLP and LSTM, were selected as base learners. RF is based on bagging-ensembled decision trees and captures feature interactions through recursive binary splits without requiring explicit interaction terms. The MLP is a fully connected network with two hidden layers, rectified linear unit (ReLU) activation, the Adam optimizer, L2 regularization, and early stopping, and it learns a smooth global nonlinear mapping. Two hidden layers were used as a common compromise between flexibility and the risk of overfitting for a dataset of about 1000 results; the effect of the depth is examined in Section 3.1. Because the dataset is tabular and has no temporal ordering, each seven-dimensional feature vector was supplied to the LSTM as a single-step input (sequence length = 1). The LSTM therefore does not exploit temporal recurrence and serves as a gated nonlinear learner whose forget-gate parameterization differs from the ReLU

architecture of the MLP [57]. It was retained because its errors differ from those of RF and the MLP, which is the property that a weighted combination exploits (Section 3.3).

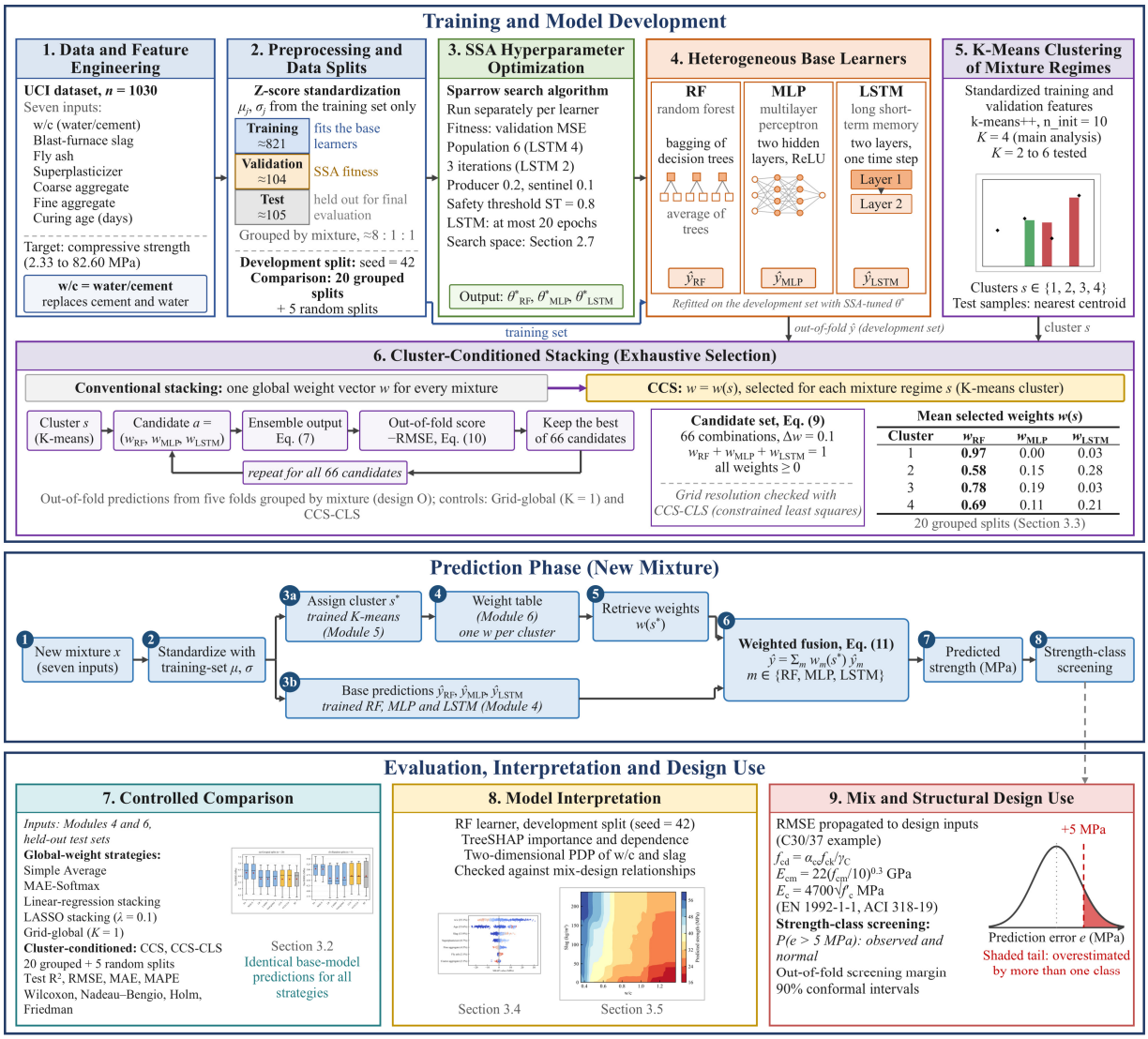


Figure 3. Framework of the proposed cluster-conditioned stacking (CCS) model, comprising model development (Modules 1–6), the prediction of a new mixture (Steps 1–8), and evaluation, interpretation, and design use (Modules 7–9). Solid arrows indicate the flow of data and results, with labels giving the transferred quantity; the purple arrow in Module 6 contrasts conventional stacking with CCS; and the dashed arrow indicates that the predicted strength is used for strength-class screening in Module 9. Box colors only distinguish the modules. The chart in Module 5 is a schematic illustration of the clustering, and the thumbnails in Modules 7 and 8 are reduced previews of the results reported in Sections 3.2, 3.4 and 3.5. UCI, University of California, Irvine; μ and σ, training-set mean and standard deviation; θ^*, SSA-tuned hyperparameters; MSE, mean squared error; ReLU, rectified linear unit; $\hat{y}$, predicted strength; K, number of clusters; n_init, number of k-means initializations; s and s^*, cluster index and cluster assigned to the new mixture; $w(s)$, weight vector of cluster s; Δw, grid step of the weights; design O, weights selected on out-of-fold predictions; LASSO, least absolute shrinkage and selection operator; λ, regularization coefficient; R^2, coefficient of determination; RMSE, MAE and MAPE, root-mean-square, mean absolute and mean absolute percentage errors; SHAP, SHapley Additive exPlanations; PDP, partial dependence plot; e, prediction error. The design quantities in Module 9 are defined in Section 4.2.

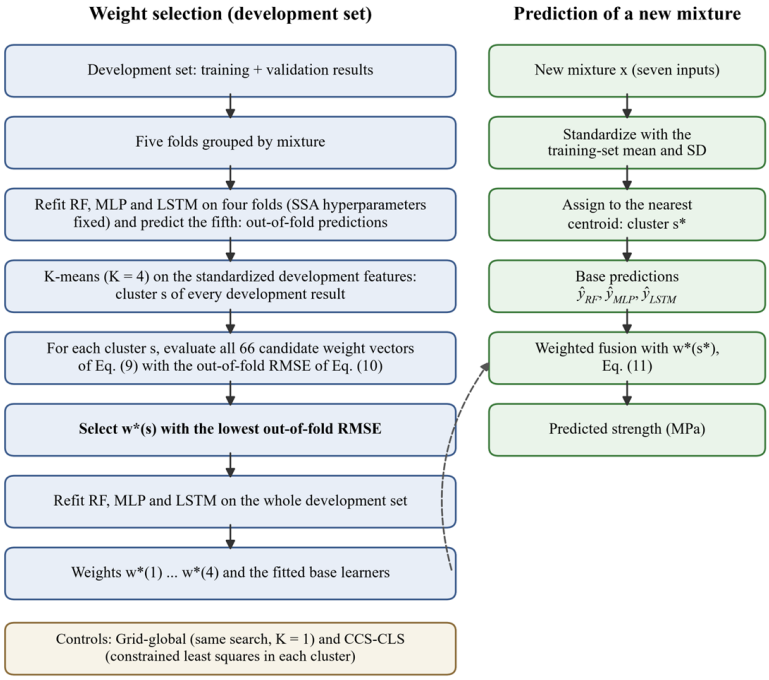


Figure 4. Selection of the cluster-conditioned weights on the out-of-fold predictions of the development set (left) and their use in the prediction of a new mixture (right). Solid arrows indicate the sequence of computational steps; the dashed arrow transfers the selected cluster-specific weights to the prediction phase. SD, standard deviation.

2.7. SSA Hyperparameter Optimization

SSA represents a population of sparrows with three roles, namely producers, scavengers and sentinels, which together balance exploration and exploitation [46]. It is used here as a hyperparameter tuner, and details of the algorithm are given in the survey by Gharehchopogh et al. [58]. SSA was applied separately to each base learner, with the validation mean squared error (MSE) as the fitness function to be minimized over the search space in Table 2. The fitness function is defined as

$$f(\theta) = \mathrm{MSE}_{\mathrm{val}} = \frac{1}{N_{\mathrm{val}}} \sum_{i=1}^{N_{\mathrm{val}}} (y_i - \hat{y}_i(\theta))^2 \tag{6}$$

where θ denotes the hyperparameter vector to be optimized, N_{val} is the number of validation set samples, and y_i and $\hat{y}_i(\theta)$ represent the true value and the model prediction for the i-th validation sample, respectively.

Table 2. Hyperparameter search space for the three base learners.

Base Learner	Hyperparameter	Lower Bound	Upper Bound	Description
RF	n_estimators	50	300	Number of decision trees
RF	max_depth	3	25	Maximum tree depth
RF	min_samples_split	2	10	Minimum samples for splitting
RF	min_samples_leaf	1	5	Minimum samples per leaf node
MLP	h_1	16	128	Number of neurons in the first hidden layer
MLP	h_2	8	64	Number of neurons in the second hidden layer
MLP	$\log_{10}(\mathrm{lr})$	−4	−2	Learning rate (log scale)
MLP	$\log_{10}(\lambda)$	−5	−2	Regularization coefficient (log scale)
LSTM	u_1	16	64	Number of hidden units in the first LSTM layer
LSTM	u_2	8	32	Number of hidden units in the second LSTM layer
LSTM	$\log_{10}(\mathrm{lr})$	−4	−2	Learning rate (log scale)
LSTM	dropout	0.05	0.4	Dropout ratio

SSA was run with a population of six individuals and three iterations for RF and the MLP and with a population of four individuals and two iterations for the LSTM, with a producer proportion of 0.2, a sentinel proportion of 0.1, and a safety threshold ST = 0.8. The LSTM was trained for at most 20 epochs with a batch size of 50. A search of 30 iterations had been planned, but the evaluation protocol of Section 2.3.2 requires the base learners to be tuned and refitted on 26 splits, including five grouped folds per split, and the budget was shortened so that the complete set of experiments could be run within the available computing time.

The early-stopping rule (no improvement in three consecutive iterations) cannot take effect within this budget, and the search cannot be regarded as converged; its effect on the accuracy of the MLP and LSTM is discussed in Sections 3.1 and 4.3.

Log-scale encoding (optimizing $\log_{10}(\mathrm{lr})$ rather than lr) was used for the learning rate and regularization coefficient of MLP and LSTM so that values spanning several orders of magnitude were sampled uniformly. Integer-valued hyperparameters (e.g., n_estimators, h_1) were rounded before fitness evaluation.

2.8. Stacking Ensemble Framework

Stacking combines the outputs of base learners through a meta-strategy. Here, the final prediction is the weighted combination of the RF, MLP, and LSTM outputs

$$\hat{y} = w_{\mathrm{RF}} \cdot \hat{y}_{\mathrm{RF}} + w_{\mathrm{MLP}} \cdot \hat{y}_{\mathrm{MLP}} + w_{\mathrm{LSTM}} \cdot \hat{y}_{\mathrm{LSTM}} \quad (7)$$

For Simple Average, MAE-Softmax, Grid-global, and the two cluster-conditioned rules, the weights were constrained to be non-negative and to sum to one. Linear-regression and LASSO stacking were fitted as standard regression meta-models without this constraint. Seven combination rules were compared (Table 3): five global-weight strategies, the proposed CCS, and a constrained least-squares variant of CCS (CCS-CLS). Wherever weights were estimated, they were estimated from the out-of-fold predictions of the development set (Section 2.3.2).

Table 3. Combination rules compared in this study.

Rule	Weights	How the Weights are Obtained
Simple Average (SA)	Global	Equal weights of 1/3
MAE-Softmax (MAE-S)	Global	Softmax of the reciprocal out-of-fold MAE
Linear-regression stacking (LR)	Global	Ordinary least squares, unconstrained
LASSO stacking (LASSO)	Global	L1-regularized least squares ($\lambda = 0.1$), unconstrained
Grid-global	Global ($K = 1$)	Lowest out-of-fold RMSE among the 66 candidates of Equation (9)
CCS	One vector per cluster ($K = 4$)	Lowest out-of-fold RMSE among the 66 candidates, separately in each cluster
CCS-CLS	One vector per cluster ($K = 4$)	Constrained least squares in each cluster (weights ≥ 0, summing to one)

Where weights are estimated, the out-of-fold predictions of the development set are used (Section 2.3.2).

2.8.1. Global-Weight Strategies

The five global-weight strategies were (1) equal weighting (Simple Average), (2) MAE-Softmax weighting, which uses Softmax-normalized reciprocals of the out-of-fold MAE as weights, (3) linear-regression stacking, which fits the weights by ordinary least squares, (4) LASSO stacking, which adds L1 regularization ($\lambda = 0.1$) to the linear regression, and (5) Grid-global, which selects one weight vector for all samples from the same 66 candidates as CCS by minimizing the out-of-fold RMSE, that is, CCS with a single cluster ($K = 1$). All five apply the same weights to every sample [26,27].

2.8.2. Cluster-Conditioned Weight Selection

(1) *Mixture regimes defined by K-means clustering*

K-means clustering was used to partition the samples into K clusters, each of which is treated as one mixture regime. The objective function of K-means is

$$J = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2 \tag{8}$$

where C_k denotes the k-th cluster and μ_k is its centroid. $K = 4$ was used for the main analysis; the four clusters can be described by binder composition and w/c and contain, on average, 185 to 295 development results (Section 3.3). The silhouette coefficient, averaged over the grouped splits, rose steadily from 0.21 at $K = 2$ to 0.29 at $K = 6$ and did not single out any number of clusters; values below 0.3 show that the clusters overlap, as expected for mixtures whose composition varies continuously. CCS does not require sharply separated clusters, because a cluster only needs to group mixtures for which the same weight vector works well, and the sensitivity of the results to K is reported in Section 3.3. K-means was run with k-means++ initialization and n_init = 10 on the standardized features of the development set, and test samples were assigned to the nearest centroid.

(2) *Candidate weight vectors*

Each candidate is a weight combination of the three base learners, $a = (w_{\mathrm{RF}}, w_{\mathrm{MLP}}, w_{\mathrm{LSTM}})$, subject to the normalization constraint $w_{\mathrm{RF}} + w_{\mathrm{MLP}} + w_{\mathrm{LSTM}} = 1$. The weights were discretized at step size $\Delta w = 0.1$, and the candidate set is

$$A = \{(w_1, w_2, w_3) | w_i \in \{0,\ 0.1,\ 0.2,\ \ldots,\ 1.0\}, w_1 + w_2 + w_3 = 1\} \tag{9}$$

Through enumeration, a total of $|A| = 66$ valid weight combinations were obtained.

(3) *Selection criterion*

For each cluster s and candidate a, the score is the negative RMSE of the ensemble predictions for the development samples in that cluster, so that more accurate candidates receive higher scores

$$r(s,\ a) = -\mathrm{RMSE}_{s,a} = -\sqrt{\frac{1}{|S_s|} \sum_{i \in S_s} \left(y_i - \hat{y}_{i,a}\right)^2} \tag{10}$$

where S_s denotes the set of development samples in cluster s, and $\hat{y}_{i,a}$ is the ensemble prediction for the i-th sample under candidate a (i.e., the weight combination $(w_{\mathrm{RF}}, w_{\mathrm{MLP}}, w_{\mathrm{LSTM}})$). RMSE was preferred to MAE because it penalizes large errors more strongly, so that the selection favors weight vectors that avoid large errors on individual mixtures. All 66 candidates were evaluated in every cluster, and the candidate with the highest score, that is, the lowest out-of-fold RMSE, was selected. Because the selection for a cluster is a one-step decision with a fixed criterion, a reinforcement-learning search such as tabular Q-learning [43] offers no advantage here, whereas evaluating all candidates guarantees the optimum on the grid at negligible cost. As a check on the grid resolution, the weights of each cluster were also obtained by constrained least squares (CCS-CLS), which minimizes the squared out-of-fold error of Equation (7) subject to non-negative weights that sum to one [53].

(4) *Weight allocation procedure in the prediction phase*

After the selection, each cluster has one weight vector. The prediction procedure for a new test sample proceeds as follows.

Step 1 (Cluster assignment). The test sample is fed into the trained K-means model to determine its assigned cluster s^*.

Step 2 (Weight look-up). The weight vector selected for cluster s^* is retrieved, giving the weight combination $(w^*_{\mathrm{RF}}, w^*_{\mathrm{MLP}}, w^*_{\mathrm{LSTM}})$.

Step 3 (Weighted fusion). The predictive outputs of the three base learners are fused with the selected weights

$$\hat{y}_{\mathrm{final}} = w^*_{\mathrm{RF}} \cdot \hat{y}_{\mathrm{RF}} + w^*_{\mathrm{MLP}} \cdot \hat{y}_{\mathrm{MLP}} + w^*_{\mathrm{LSTM}} \cdot \hat{y}_{\mathrm{LSTM}} \tag{11}$$

At inference, CCS adds only a nearest-centroid assignment and a table look-up to the cost of the base learners, so its prediction cost is practically that of conventional stacking. The resulting weighting function is piecewise constant over the mixture space and allows the contribution of each base learner to change from one mixture regime to another.

2.9. SHAP Interpretability Analysis

SHAP [59] attributes each prediction to the input features on the basis of Shapley values from cooperative game theory. The SHAP value of feature j for sample x is defined as

$$\phi_j(x) = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(d - |S| - 1)!}{d!} \left[f_{S \cup \{j\}}(x) - f_S(x) \right] \tag{12}$$

where $F = \{1, 2, \ldots, d\}$ is the index set of all d features, S is any feature subset not containing feature j, $f_S(x)$ represents the expected prediction using only subset S for sample x, and the factorial term $|S|!(d - |S| - 1)!/d!$ is the normalized weighting factor over all possible permutation orderings.

SHAP values were computed for the RF learner with TreeSHAP, which gives exact attributions for tree ensembles at low computational cost. RF was the most accurate single learner (Section 3.1) and received the largest weight in every cluster (Section 3.3), so its attributions show which inputs drive the predictions that CCS combines. The SHAP and PDP results are used to check that the learned relationships agree with established knowledge of concrete strength. SHAP values are reported in MPa.

2.10. Partial Dependence Analysis

The PDP [60] shows the average effect of one or two features on the prediction by marginalizing over the remaining features. For a feature subset X_S, the partial dependence function is defined as

$$\hat{f}_S(X_S) = \frac{1}{n} \sum_{i=1}^{n} f\left(X_S, x_i^C\right) \tag{13}$$

where $x_i{}^C$ denotes the observed values of the i-th training sample on the complementary feature set. PDP assumes that the target features are independent of the others, and strongly correlated features can lead to extrapolation. For this reason, the PDP results were cross-checked against the SHAP dependence plots.

2.11. Prediction Intervals

Prediction intervals for individual mixtures were constructed from the absolute out-of-fold residuals of CCS on the development set, following the conformal approach of Lei et al. [61]. The half-width of the 90% interval is the empirical quantile of these n residuals at level $\lceil 0.9(n + 1) \rceil / n$; it was computed once for all clusters together and once for each cluster separately, with the global value used for clusters with fewer than 10 development results. Because the same out-of-fold predictions were also used to select the weights, the residuals do not form an independent calibration set, and the finite-sample coverage guarantee of split-conformal prediction does not strictly apply; the coverage was therefore checked on the test sets (Section 4.2). The screening margin for a 10% risk of overestimating the strength was taken as the 90th percentile of the out-of-fold overestimation, that is, of the

predicted minus the measured strength. For comparison, a Gaussian process regression (GPR) model, which gives predictive intervals directly and has been used to quantify the uncertainty of strength predictions for sustainable construction composites [62], was fitted to the development set of each split with a constant-times-RBF kernel, one length scale per input and a white-noise term; its 90% interval is the predictive mean ± 1.645 predictive standard deviations.

3. Results

3.1. Performance of the SSA-Tuned Base Learners

Table 4 gives the test performance of the three base learners over the 20 grouped splits.

Table 4. Test performance of the three base learners over the 20 grouped splits.

Model	MAE (MPa)	RMSE (MPa)	R^2	MAPE (%)
RF	4.26 ± 0.57	5.77 ± 0.77	0.876 ± 0.030	14.3 ± 1.5
MLP	5.40 ± 0.77	7.41 ± 1.39	0.788 ± 0.092	18.5 ± 3.5
LSTM	5.35 ± 0.69	7.19 ± 0.90	0.809 ± 0.041	18.0 ± 3.0

Values are means ± standard deviations over the 20 grouped splits (design O).

RF was the most accurate learner, with a mean test RMSE of 5.77 MPa, followed by the LSTM (7.19 MPa) and the MLP (7.41 MPa), whose accuracy also varied most between splits (standard deviation of R^2 0.092). Part of this gap is a consequence of the shortened search (Section 2.7). On the development split, all four LSTM hyperparameters ended at a bound of the search space (16 and 8 units, a learning rate of 0.01 and a dropout ratio of 0.05), which indicates that the search and the 20 training epochs did not allow the LSTM to reach its best accuracy, and the large variation in the MLP suggests the same for this network. A longer search would probably make both networks more accurate.

Figure 5 shows learning curves of the three base learners and of their simple average on five grouped splits, for which the learners were trained on 20% to 100% of the training mixtures, each with all its results, and evaluated on the same test sets. The test RMSE of RF rose from 6.0 MPa with the full training set to 7.9 MPa with 20% of the mixtures (158 to 226 results), and the errors of the MLP and LSTM rose much more. The RF error was still decreasing between 80% and 100% of the training set (6.5 and 6.0 MPa), so the learners had not reached a plateau, and more data would probably improve the predictions further. For the MLP and LSTM, the training and test errors were of similar size at small training sets, which indicates underfitting rather than overfitting and is consistent with the shortened search.

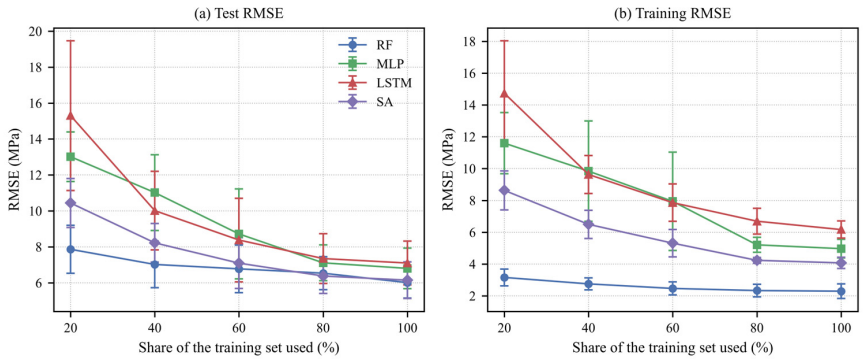


Figure 5. Learning curves of RF, the MLP, the LSTM, and their simple average (SA): (**a**) test RMSE and (**b**) training RMSE against the share of the training set used, as means ± standard deviations over five grouped splits.

Two further checks concerned the capacity of the learners (Supplementary Tables S2 and S3). For the MLP, networks with one, two, and three hidden layers were tuned with the same search on five grouped splits. The three-layer network had the lowest mean validation and test RMSE (5.61 and 6.13 MPa, compared with 6.18 and 6.80 MPa for two layers and 6.40 and 6.60 MPa for one layer), with a similar search time and, because narrower layers were selected, slightly fewer parameters, although these differences were smaller than the standard deviations between splits (Supplementary Table S2). Under the present budget, the two-layer network is neither the most accurate nor the most economical choice. It was kept so that all results in this paper refer to the same learners. For RF, the mean test RMSE fell from 6.25 MPa with 10 trees to 6.01 MPa with 100 trees and changed by at most 0.06 MPa between 100 and 500 trees, while the fitting time grew from 0.14 s to 0.60 s, which supports the search range of 50 to 300 trees in Table 2.

3.2. Cluster-Conditioned Versus Global Weighting

Figure 6 shows the out-of-fold predictions of CCS for the 927 development results and its predictions for the 103 test results of the development split. Both sets scatter around the line of equality, but above about 65 MPa most points lie below it, so the highest strengths tend to be underestimated.

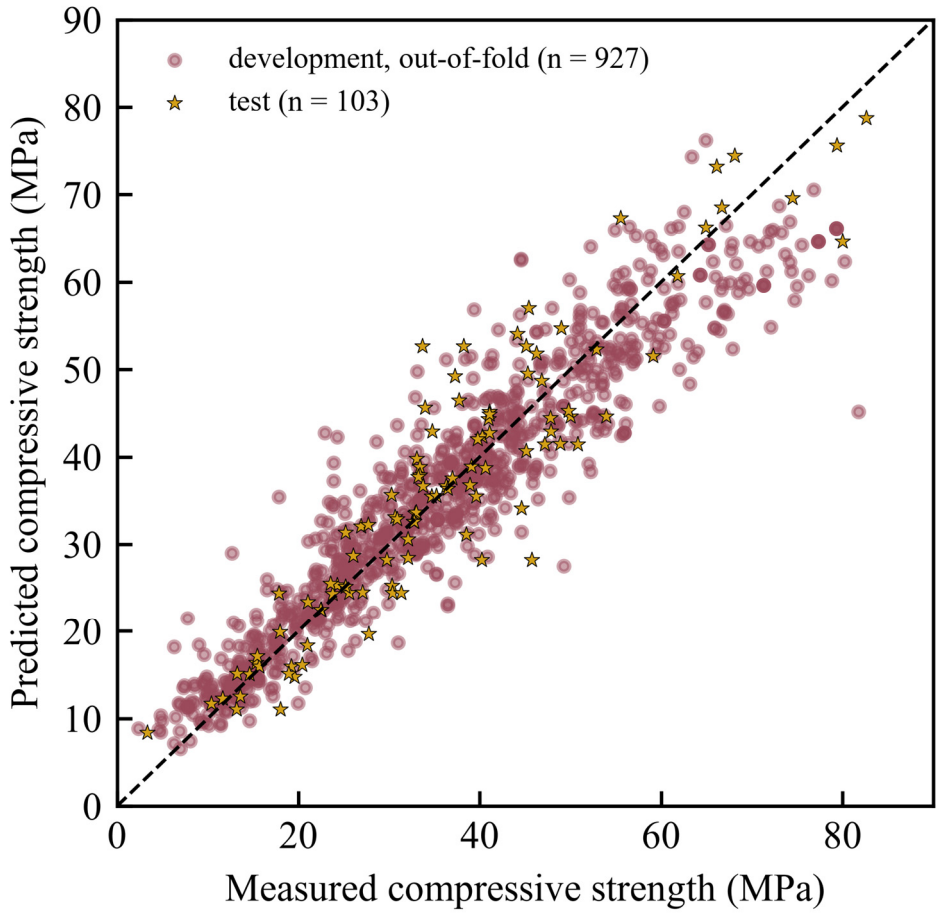


Figure 6. Predicted versus measured compressive strength of CCS for the development split (seed = 42): out-of-fold predictions for the 927 development results and predictions for the 103 test results.

Table 5 compares the seven combination rules and RF alone on the 20 grouped splits and, for reference, on the five random splits. On the grouped splits, CCS reached a mean test R^2 of 0.876 ± 0.036 and a mean RMSE of 5.76 ± 0.83 MPa (Figure 7). The two rules that

weight the learners equally or by their MAE were clearly less accurate (6.27 and 6.24 MPa) because they gave the weaker MLP and LSTM a large share of the prediction. All rules that fitted the weights to the data reached 5.75 to 5.80 MPa, and RF alone reached 5.77 MPa. On the random splits, every rule appeared more accurate, with RMSE values of 4.60 to 5.04 MPa, and CCS (4.71 MPa) ranked behind linear-regression and LASSO stacking (4.60 MPa) and Grid-global (4.62 MPa). In these splits, on average, 64 of the 90 mixtures in a test set also had results at other ages in the training set, so random splitting underestimated the error for new mixtures by about 1 MPa.

Table 5. Mean test performance of the combination rules and of RF alone on the 20 grouped splits and the five random splits (design O).

Method	Grouped Splits (20)				Random Splits (5)	
	R^2	RMSE (MPa)	MAE (MPa)	MAPE (%)	R^2	RMSE (MPa)
Simple Average	0.853 ± 0.041	6.27 ± 0.86	4.64 ± 0.64	15.5 ± 2.7	0.899 ± 0.026	5.04 ± 0.45
MAE-Softmax	0.854 ± 0.041	6.24 ± 0.85	4.62 ± 0.64	15.5 ± 2.6	0.899 ± 0.026	5.02 ± 0.45
Linear-regression stacking	0.874 ± 0.038	5.78 ± 0.77	4.26 ± 0.57	14.1 ± 1.7	0.915 ± 0.026	4.60 ± 0.67
LASSO stacking	0.875 ± 0.038	5.78 ± 0.77	4.26 ± 0.57	14.1 ± 1.7	0.915 ± 0.026	4.60 ± 0.67
Grid-global	0.875 ± 0.032	5.80 ± 0.77	4.27 ± 0.57	14.4 ± 1.8	0.914 ± 0.028	4.62 ± 0.69
CCS	0.876 ± 0.036	5.76 ± 0.83	4.24 ± 0.60	14.5 ± 2.2	0.912 ± 0.026	4.71 ± 0.65
CCS-CLS	0.876 ± 0.036	5.75 ± 0.83	4.23 ± 0.60	14.5 ± 2.2	0.912 ± 0.026	4.71 ± 0.66
RF alone	0.876 ± 0.030	5.77 ± 0.77	4.26 ± 0.57	14.3 ± 1.5	0.911 ± 0.030	4.70 ± 0.77

Values are means ± standard deviations over the 20 grouped splits (seeds 1–20) and the five random splits (seeds 1, 5, 10, 15, and 20). The values for each split are given in Supplementary Tables S1a–S1c.

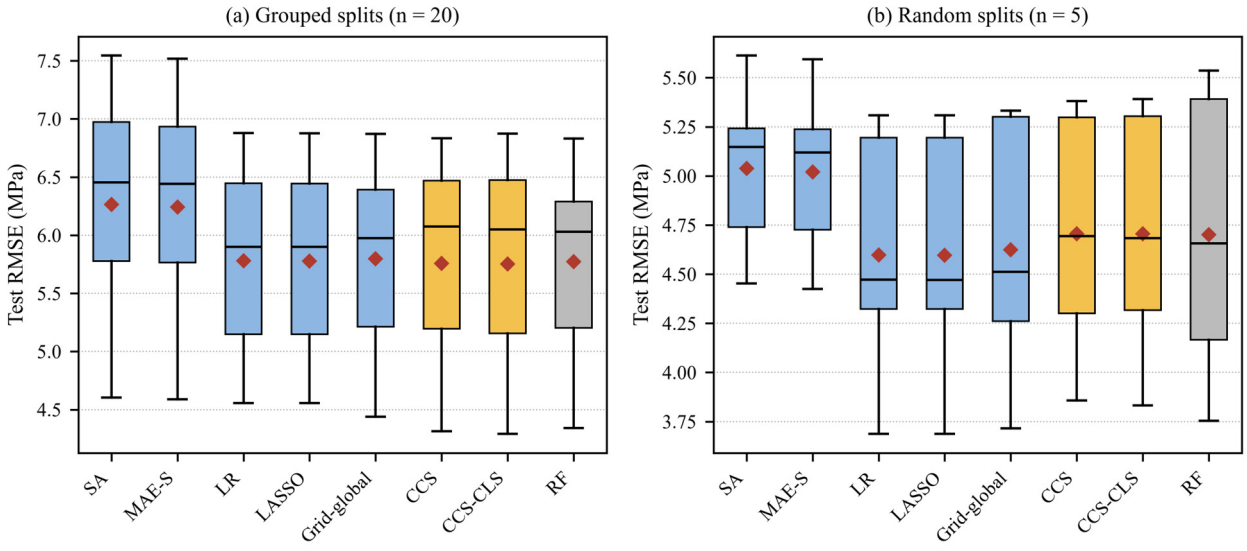


Figure 7. Test RMSE of the combination rules and of RF alone on (**a**) the 20 grouped splits and (**b**) the five random splits (design O). Boxes show interquartile ranges, horizontal lines medians, diamonds means, and whiskers full ranges. Blue boxes indicate global-weight combination rules, yellow boxes indicate the cluster-conditioned rules (CCS and CCS-CLS), and the gray boxes indicate RF alone. Rule abbreviations are as in Table 3.

Table 6 gives the paired comparison of CCS with the other rules on the grouped splits. CCS was more accurate than Simple Average and MAE-Softmax on 19 of the 20 splits, by 0.51 and 0.48 MPa on average, and both differences remained significant after Holm correction. Against the rules that fitted global weights to the data and against RF alone, the mean difference was 0.01 to 0.04 MPa, CCS was better on 11 to 13 of the 20 splits, and none of the differences was significant even without correction (Wilcoxon $p \geq 0.29$, Nadeau–Bengio $p \geq 0.59$). Among the seven rules, CCS had the lowest test RMSE on five of the 20 grouped splits and ranked between second and fifth on the others (Supplementary Table S1b). CCS-

CLS differed from CCS by less than 0.01 MPa, so the grid step of 0.1 did not limit the accuracy. The Friedman test indicated differences among the seven rules ($p < 0.001$), but the mean ranks of the five rules with fitted weights lay close together, between 2.75 (CCS-CLS) and 3.65 (Grid-global), compared with 5.65 for MAE-Softmax and 6.75 for Simple Average. The hypothesis that regime-dependent weights predict strength more accurately than any global weight vector is therefore not supported by these data.

Table 6. Paired comparison of CCS with the other combination rules and with RF alone over the 20 grouped splits (design O).

Compared with	ΔRMSE (MPa)	CCS Better (of 20)	Wilcoxon p (Holm)	Nadeau–Bengio p (Holm)
Simple Average	−0.51 ± 0.37	19	<0.001 (<0.001)	0.003 (0.023)
MAE-Softmax	−0.48 ± 0.36	19	<0.001 (<0.001)	0.004 (0.023)
Linear-regression stacking	−0.02 ± 0.18	11	0.596 (1.000)	0.782 (1.000)
LASSO stacking	−0.02 ± 0.18	11	0.622 (1.000)	0.797 (1.000)
Grid-global	−0.04 ± 0.17	13	0.294 (1.000)	0.598 (1.000)
CCS-CLS	+0.01 ± 0.03	8	0.452 (1.000)	0.632 (1.000)
RF alone	−0.01 ± 0.31	12	0.701 (1.000)	0.912 (1.000)

ΔRMSE: test RMSE of CCS minus that of the compared rule (mean ± standard deviation); negative values favor CCS. CCS better: number of splits on which CCS had the lower test RMSE. p-values: two-sided Wilcoxon signed-rank test [54] and corrected resampled t-test [55] on the paired test RMSE values, with Holm-adjusted values for the seven comparisons in parentheses [56]. Friedman test over the seven combination rules: $p < 0.001$.

When the weights were selected on the validation predictions instead (design V), CCS had a higher mean test RMSE than Grid-global (6.22 versus 6.04 MPa on the grouped splits) and was better on only eight of the 20 splits, although the paired difference was not significant (Wilcoxon $p = 0.13$); Grid-global had the best mean rank among the seven rules (Friedman $p = 0.015$). With four clusters, each cluster vector was then selected on about 26 validation results on average, and the selected weights were affected by the sampling noise of these small samples. Selecting the weights on the out-of-fold predictions of all development results (design O) removed this disadvantage but did not turn it into a gain.

3.3. Cluster-Specific Weight Selection

To compare the clusters across splits, the four clusters of each grouped split were matched one-to-one to a reference clustering of all 1030 results ($K = 4$) by minimizing the total squared distance between centroids; the reference clustering was used only for this matching. Table 7 gives their composition and the selected weights. The clusters were distinguished by binder composition and w/c and not by curing age, since all four had a median age of 28 days. Cluster 1 contained plain cement mixtures (median w/c 0.58), Cluster 2 fly ash mixtures (median fly ash content 122 kg/m^3, w/c 0.77), Cluster 3 slag mixtures with a low w/c (slag 116 kg/m^3, w/c 0.45), which had the highest median strength (49.9 MPa), and Cluster 4 slag mixtures with a high w/c (slag 182 kg/m^3, w/c 1.14).

Figure 8 shows the mean weights and their standard deviations over the 20 grouped splits. RF received the largest weight in every cluster, from 0.58 in the fly ash cluster to 0.97 in the plain cement cluster, where the other two learners were practically excluded. The LSTM received a noticeable weight only in the fly ash cluster (0.28) and in the high-w/c slag cluster (0.21), and the MLP received at most 0.19. The weights of the neural learners in these clusters, however, varied considerably between splits (standard deviations of 0.09 to 0.16). On the development split, the selected vectors (RF:MLP:LSTM) were 1.0:0.0:0.0 for the plain cement cluster, 0.6:0.1:0.3 for the fly ash cluster, 0.7:0.3:0.0 for the low-w/c slag cluster and 0.5:0.1:0.4 for the high-w/c slag cluster.

Table 7. Composition of the four clusters and their selected weights over the 20 grouped splits (design O).

Cluster	n (dev.)	n (Test)	w/c	Slag (kg/m³)	Fly Ash (kg/m³)	Age (Days)	Strength (MPa)	w_{RF}	w_{MLP}	w_{LSTM}
1 Plain cement	234	25	0.58	0	0	28	32.5	0.97 ± 0.04	0.00 ± 0.00	0.03 ± 0.04
2 Fly ash	295	35	0.77	0	122	28	32.5	0.58 ± 0.09	0.15 ± 0.11	0.28 ± 0.12
3 Slag, low w/c	185	20	0.45	116	0	28	49.9	0.78 ± 0.11	0.19 ± 0.12	0.03 ± 0.08
4 Slag, high w/c	210	25	1.14	182	0	28	28.5	0.69 ± 0.12	0.11 ± 0.09	0.21 ± 0.16

Composition values are medians of the per-split cluster values; numbers of results and weights are means (weights ± standard deviations) over the 20 splits. n (dev.) and n (test): numbers of development and test results in the cluster.

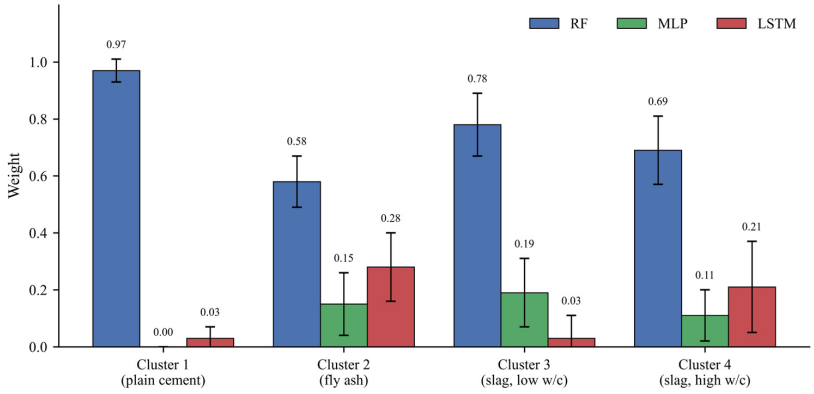


Figure 8. Weights selected for the four clusters, as means ± standard deviations over the 20 grouped splits (design O). Error bars are truncated at zero because the weights cannot be negative.

The number of clusters had little effect on the accuracy. With $K = 2$ to 6, the mean test RMSE on the grouped splits ranged from 5.76 to 5.79 MPa and the mean R^2 from 0.875 to 0.876, and none of the variants differed significantly from $K = 4$ (Wilcoxon $p \geq 0.29$).

3.4. SHAP Interpretation of the RF Learner

Figure 9 summarizes the SHAP analysis of the RF learner on the development split. Each point is the SHAP value of one result, and the share of each input in the total mean absolute SHAP value is given next to its name. The w/c ratio had the largest share (35.1%), followed closely by curing age (33.6%) and by slag content (13.9%). Together, w/c and age accounted for 68.7% of the total attribution, whereas superplasticizer, fine aggregate, fly ash, and coarse aggregate accounted for 8.7%, 4.5%, 2.1%, and 2.1%, respectively. The sample-level values show the expected directions. Low w/c values received positive SHAP values of up to about 30 MPa and high w/c values negative ones of down to about −20 MPa; the earliest ages received values of down to about −24 MPa and the oldest ages values of about 10 to 15 MPa. Slag showed a positive effect at higher contents with considerable scatter, which indicates that its contribution depends on the other mixture variables.

Figure 10 shows the SHAP dependence plots of six inputs. The w/c attribution falls from about 17 to 31 MPa at the lowest w/c of about 0.3 to zero near w/c = 0.6 and to about −10 to −18 MPa above w/c = 1.1, beyond which it changes little. The age attribution rises from about −15 to −24 MPa at 1 to 3 days through zero at 28 days to about 8 to 15 MPa at 56 days and later, with little change beyond 90 days, which matches the familiar shape of strength development. Both sign changes lie near the median values of the inputs (0.676 and 28 days, Table 1). This is expected because SHAP values measure the effect of a feature relative to the average prediction, so the sign changes are reference points of the fitted model and not material thresholds. The slag attribution rises from −1 to −7 MPa without slag to about 2 to 7 MPa at 150 to 250 kg/m³, with values of up to 13 MPa for some mixtures, and the fly ash attribution rises from about zero without fly ash to about 1 to

3 MPa at 150 to 175 kg/m^3. For fly ash, this positive trend contrasts with its weak negative correlation with strength ($r = -0.106$, Figure 2). The reason for this difference is discussed in Section 4.1. The superplasticizer attribution rises from between about −5 and 0 MPa without superplasticizer to values of up to about 8 MPa at 10 to 14 kg/m^3 and is about 1 to 3 MPa at the few higher dosages. The fine aggregate attribution decreases from positive values of up to about 6 MPa below 700 kg/m^3 to negative values of down to about −5 MPa above 850 kg/m^3, a smaller effect than those of w/c, age, and slag.

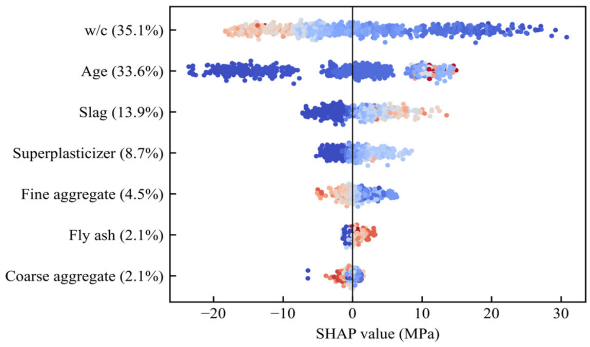


Figure 9. SHAP summary of the RF learner on the development split. Each point is the SHAP value of one result, colored from blue (low feature value) to red (high feature value); the share of each input in the total mean absolute SHAP value is given in parentheses. SHAP values are in MPa.

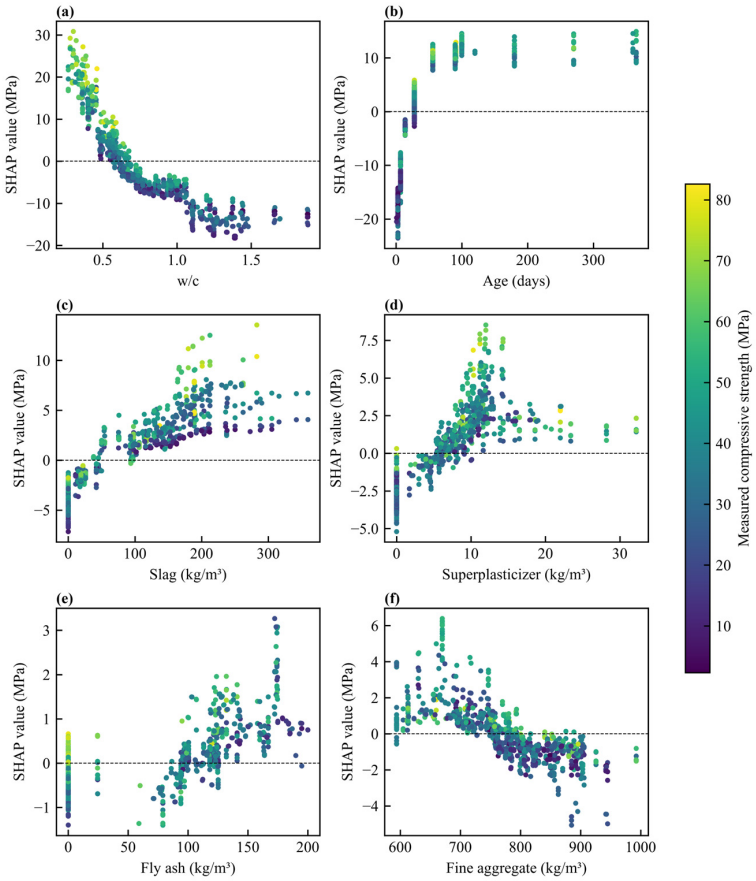


Figure 10. SHAP dependence plots of six inputs for the RF learner on the development split: (**a**) w/c; (**b**) age; (**c**) slag; (**d**) superplasticizer; (**e**) fly ash; and (**f**) fine aggregate. Point color indicates the measured compressive strength (color bar). SHAP values are in MPa; dashed lines mark zero.

3.5. Partial Dependence of w/c and Slag

Figure 11 shows the joint partial dependence of the RF prediction on w/c and slag content. The predictions range from about 18 to 58 MPa. The highest occur at w/c below about 0.45 combined with more than about 165 kg/m^3 of slag, and the lowest at w/c above about 1.1 combined with less than about 40 kg/m^3 of slag. The contours are steep along the w/c axis, so w/c changes the prediction more than slag does, in agreement with their SHAP shares of 35.1% and 13.9%. A large part of the slag effect occurs between zero and about 50 kg/m^3, where the prediction rises by one to two contour intervals (4 to 8 MPa); at high w/c it continues to rise up to about 190 kg/m^3. Raising the slag content from zero to about 200 kg/m^3 increases the prediction by about 8 MPa at w/c = 0.4 and 0.8 and by about 12 MPa at w/c = 1.2, that is, by roughly 15%, 30%, and 65% of the strength without slag. The predicted benefit of slag is therefore largest, especially in relative terms, in mixtures with high w/c, for reasons discussed in Section 4.1.

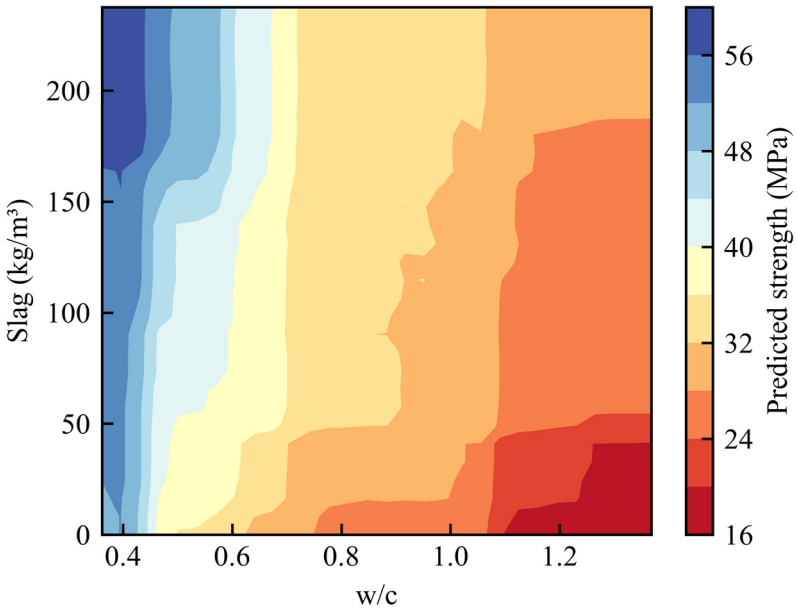


Figure 11. Partial dependence of the RF prediction on w/c and slag content (development split, seed = 42). The contour interval is 4 MPa.

4. Discussion

4.1. Comparison with Previous Studies

The ranking of the inputs in Figure 9 reflects the central role of the water-to-cement ratio and curing time in mix design [12] and agrees with earlier SHAP analyses of the same dataset, which also placed the water-to-cement ratio and age at the top [21,63]. Studies on slag concrete reached similar rankings with partial dependence analysis [31]. The age dependence learned by the model (Figure 10) has a shape similar to that of the strength-development function of EN 1992-1-1 [1], which gives 0.46 to 0.60 of the 28-day strength at 3 days, 0.68 to 0.78 at 7 days, and 1.12 to 1.18 at 90 days for slow- and normal-hardening cements (coefficient s = 0.38 and 0.25), so that most of the strength gain after 3 days is reached by 56 days. The learned effects of slag and fly ash follow from the definition of w/c used here and from the reactions of the two materials. Because slag and fly ash are not part of the denominator of w/c, adding either of them at a fixed w/c increases the binder content and lowers the water-to-binder ratio, and the relative reduction is largest when the cement content is low, as in high-w/c mixtures. This is consistent with the positive SHAP contributions of both materials at a given w/c (Figure 10), although fly ash is weakly and

negatively correlated with strength (Figure 2a), and with the larger predicted benefit of slag at high w/c in the partial dependence surface (Figure 11). The latent hydraulic reaction of slag, which contributes to strength mainly at later ages [7], adds to this effect. The model therefore credits slag and fly ash with a strength contribution at a given w/c instead of treating them as inert diluents of the cement, which is the behavior needed for screening low-clinker mixtures.

The clustering studies cited in Section 1 used the clusters to fit separate regression models [44] or to structure hierarchical classification and regression models [45], and both reported more accurate predictions with clustering than without it. CCS uses the clusters differently: all base learners are trained on all development results, and the clusters only determine how their predictions are combined, so that no learner is trained on a fraction of the data. The weights selected for the four clusters differed (Table 7 and Figure 8), which indicates that the relative accuracy of the base learners changes with the mixture regime, but CCS was not more accurate than global weights fitted to the same predictions or than RF alone (Table 6). The forecast combination literature offers an explanation: weights estimated from limited data often fail to outperform simple rules because the estimation error of the weights offsets the expected gain [64,65]. The present results show both sides of this effect. Selected on about 26 validation results per cluster, the cluster weights gave a higher mean error than one global vector (design V, Section 3.2); selected on the out-of-fold predictions of all development results, they matched it but did not improve on it. Unlike in the classic combination puzzle, the simple average was clearly worse here because the three learners differed strongly in accuracy; the gain from fitted weights came mainly from reducing the share of the weaker MLP and LSTM. With better-tuned neural learners the learners would differ less and their errors might be more complementary, so the present data cannot rule out a benefit of regime-dependent weighting (Section 4.3).

Published models for the UCI dataset differ in their inputs, data partitions, and evaluation protocols (Section 1) [21,25,29,63], so their reported accuracies cannot be compared directly with each other or with the present results. The absolute accuracy reached here, a mean test RMSE of 5.76 MPa on splits grouped by mixture and 4.71 MPa on random splits (Table 5), is not used to rank CCS against them. The difference between the two protocols is itself relevant to such comparisons. In random splits, most test mixtures also have results at other ages in the training set. These splits therefore measure the accuracy for known mixtures at new test ages, whereas mix design requires the accuracy for new mixtures. The contribution of this study lies instead in a controlled answer to a narrower question, namely whether conditioning the combination weights of a stacked model on the mixture regime improves the prediction of slag and fly ash concrete strength when everything else is held fixed. For the data and learners used here, it did not: a global weight vector fitted to the data, or RF alone, was as accurate, whereas equal weighting was not.

4.2. Implications for Mix Design and Structural Design

A strength predictor is used at two points of the design process. In mix design it screens candidate mixtures against the specified strength class before trial batches are cast, and in structural design the predicted strength supplies the design compressive strength and the elastic modulus. Table 8 relates the mean test RMSE of CCS on the grouped splits (5.76 MPa) to both uses for a C30/37 concrete (f_{ck} = 30 MPa) according to EN 1992-1-1 [1] and ACI 318-19 [2]. Adjacent strength classes in the normal-strength range differ by 5 MPa in f_{ck} [1]. If the errors were normally distributed, the probability that a mixture is overestimated by more than one class step, which is the non-conservative case in screening, would be 19.3%. The observed frequency was lower: 14.8% of the test results on the grouped splits were overestimated by more than 5 MPa. For screening, the margin was therefore set from

the out-of-fold errors instead of from an assumed distribution (Section 2.11). The 90th percentile of the out-of-fold overestimation was 6.2 MPa; with this margin, 11.5% of the test results were still overestimated by more than the margin, close to the intended 10%.

Table 8. Effect of a prediction error of one RMSE (CCS, grouped splits) on strength-class screening and on the design inputs of a C30/37 concrete (f_{ck} = 30 MPa, f_{cm} = 38 MPa).

Quantity	Expression or Basis	Nominal Value	CCS (RMSE = 5.76 MPa)
Design compressive strength f_{cd} (EN 1992-1-1)	$\alpha_{cc} f_{ck}/\gamma_C$, α_{cc} = 1.0, γ_C = 1.5	20.0 MPa	16.2 to 23.8 MPa (±19.2%)
Secant modulus E_{cm} (EN 1992-1-1)	$22(f_{cm}/10)^{0.3}$ GPa	32.8 GPa	31.3 to 34.3 GPa (−4.8%/+4.3%)
Elastic modulus E_c (ACI 318-19)	$4700\sqrt{f'_c}$ MPa	25.7 GPa	23.1 to 28.1 GPa (−10.1%/+9.2%)
Share overestimated by more than 5 MPa	Normal errors, $1 - \Phi(5/\text{RMSE})$	n/a	19.3%
Share overestimated by more than 5 MPa	Observed, grouped test sets	n/a	14.8%
Margin for a 10% risk of overestimation	90th percentile of out-of-fold overestimation	n/a	6.2 MPa (observed share 11.5%)
90% prediction interval	Conformal, out-of-fold residuals	n/a	±9.5 MPa (coverage 88.7%)

f'_c in ACI 318-19 is taken equal to f_{ck}. The strengths in the dataset refer to 15 cm cylinders (Section 2.2), the specimen type for which f_{ck} is defined in EN 1992-1-1. Observed values are means over the test sets of the 20 grouped splits (design O).

The prediction intervals lead to the same conclusion for individual mixtures. The 90% conformal intervals had a half-width of 9.5 MPa and covered 88.7% of the test results on the grouped splits, slightly below the nominal 90%, and 93.8% on the random splits. Intervals computed separately for each cluster were practically as wide as the global ones (mean width 18.8 versus 18.9 MPa) and covered 87.9%. The GPR model was less accurate than CCS (RMSE 6.73 MPa), and its nominal 90% intervals covered only 83.7% of the test results. An interval width of about 19 MPa, almost four strength classes, means that the model can screen out clearly inadequate candidates but cannot replace trial batches for mixtures close to the class limit.

For structural design, an error of one RMSE changes the design compressive strength f_{cd} by ±3.8 MPa (±19.2%). The elastic modulus is much less sensitive because it grows with a fractional power of strength. One RMSE changes E_{cm} by about 5% and the ACI modulus E_c by about 10%. The change in E_{cm} is smaller than the 10% to 30% adjustments that EN 1992-1-1 applies to E_{cm} for different aggregate types [1], so the predicted strength is accurate enough to supply the stiffness used in preliminary structural analysis, whereas resistance checks remain more sensitive to the strength error.

The prediction error also matters for sustainable mix design. In multi-objective optimization of concrete mixtures, the strength model acts as the constraint that separates acceptable from unacceptable mixtures while cement content, cost, or embodied CO_2 is minimized [10,66]. Prediction errors near this constraint either admit mixtures that miss their class or reject low-clinker mixtures that would have reached it, and the second type of error pushes the optimum toward higher cement contents. A smaller error would narrow this band of misclassified candidates and allow lower-clinker mixtures to be selected with the same confidence.

The predictor can also shorten mix qualification (Figure 12). A trial batch that misses its class adds about a month because the mixture has to be redesigned, cast again, and cured for 28 days before it can be retested [3]. If candidate mixtures are first ranked by predicted strength and only those whose prediction exceeds the class limit (the strength that the trial batch has to reach) by the margin discussed above are cast, fewer batches fail, and the casting, curing, and testing of those batches is no longer needed. How much time and labor this saves depends on the number of candidates and on the margin chosen, so no general figure is given here. The selected mixture is still accepted on the basis of standard compression tests.

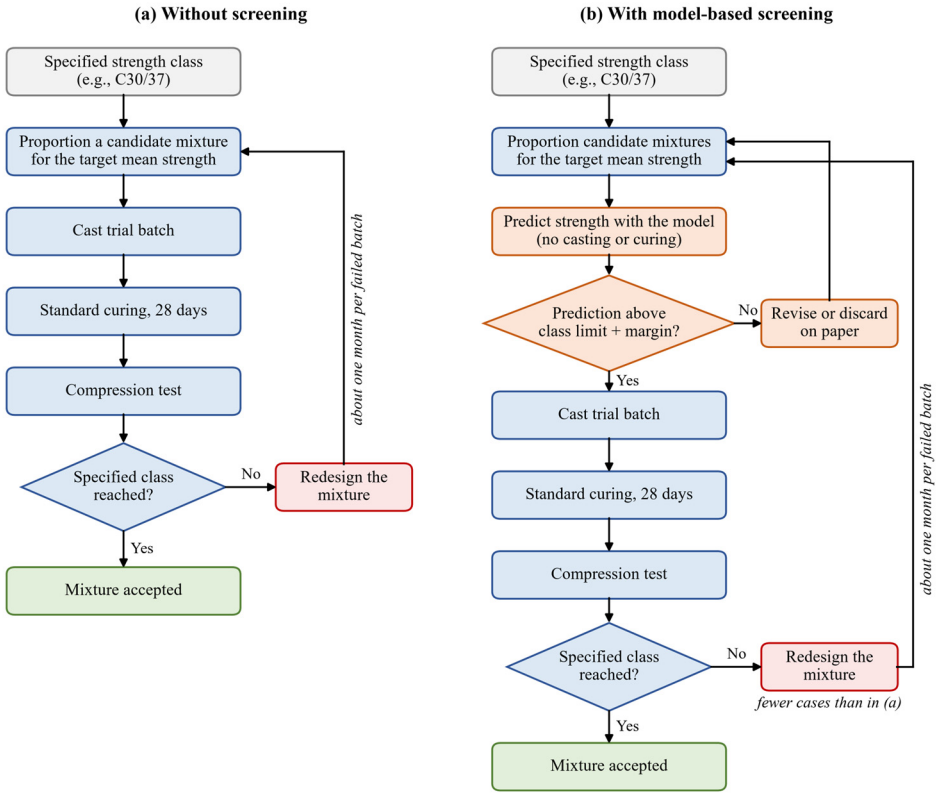


Figure 12. Qualification of a new concrete mixture (**a**) without and (**b**) with model-based screening before trial batching. A trial batch that misses the specified class has to be redesigned, cast again, and cured for 28 days before retesting, which adds about one month. Screening removes candidates whose predicted strength does not exceed the class limit by the chosen margin, so fewer batches fail; the selected mixture is still accepted on the basis of standard compression tests.

Four results of this study can now be applied, within the limits stated in Section 4.3. First, a predictor trained in the way described here can screen candidate slag and fly ash mixtures within the range of Table 1 before trial batches are cast, provided that the margin between the predicted strength and the class limit is set from the observed error; with a margin of about 6 MPa set from the out-of-fold errors, 11.5% of the test results were overestimated by more than the margin (Table 8). Second, such screening can reduce the number of failed trial batches, each of which costs about a month of redesign, casting, and curing (Figure 12). Third, the model credits slag with a strength contribution at a given w/c, which is largest at high w/c (Figure 11). It can therefore estimate the net effect of replacing part of the cement with slag at a constant water content, in which the higher w/c lowers and the added slag raises the predicted strength. Fourth, the predicted strength is accurate enough to supply the elastic modulus for preliminary structural analysis, whereas resistance checks and the acceptance of the mixture require tested strengths. For all four uses, the overall prediction error matters more than the weighting rule: RF alone or any combination rule with fitted weights would serve, whereas equal weighting of the present learners raised the test RMSE by about 0.5 MPa (Table 6).

4.3. Limitations and Applicability

The results apply to the range of the UCI data and should not be extrapolated beyond it. The dataset contains 1030 results from 428 mixtures with w/c between 0.27 and 1.88, up to 359 kg/m^3 of slag, up to 200 kg/m^3 of fly ash, and test ages from 1 to 365 days (Table 1), and it does not include silica fume, fibers, recycled aggregate, or alkali-activated binders. Within this range, four limitations remain.

First, the raw materials are described only by their masses. The dataset does not specify the cement type, the class of the fly ash, the fineness and chemical composition of the slag, or the type of superplasticizer, so the model cannot distinguish, for example, a low-calcium (Class F) from a high-calcium (Class C) fly ash, which differ in reactivity. Predictions for mixtures whose materials differ from those behind the dataset are therefore extrapolations and should be checked against local trial data. The data were also assembled from 17 sources (Section 2.2), and the splits could be grouped by mixture but not by source, which the dataset does not record. Mixtures from the same laboratory share materials and procedures, so the error for mixtures made with entirely different materials is probably larger than the grouped estimate.

Second, the base learners were tuned with a shortened search (Section 2.7). On the development split, all four LSTM hyperparameters ended at a bound of the search space, the learning curves of the MLP and LSTM indicate underfitting, and a three-layer MLP had a lower mean error than the two-layer network used here (Section 3.1). Both networks were probably weaker than a longer search would make them. Because the weights of all combination rules were fitted to these learners, stronger neural learners could change the weights and possibly the ranking of the rules, so the conclusions on regime-dependent weighting apply to the learners as trained here. The comparison should be repeated with a longer search and a deeper MLP.

Third, the screening margin and the prediction intervals in Section 4.2 are pooled over all mixtures, and the out-of-fold residuals that define them were also used to select the weights. Intervals computed per cluster were practically as wide as the pooled ones, but the error may still depend on the strength level and on the distance of a mixture from the training data, and the coverage of the 90% intervals on the grouped splits (88.7%) was slightly below the nominal value.

Fourth, the model does not include the curing temperature, which strongly affects the early strength of slag and fly ash concretes on site. It can support mix design, but not the decision on formwork removal for a particular placement.

5. Conclusions

This study examined an assumption built into stacked models for the compressive strength of concrete, namely that one global weight vector suits all mixtures, which is least tenable for the heterogeneous slag and fly ash mixtures on which low-clinker concrete relies. In the cluster-conditioned stacking model (CCS), K-means clustering defines regimes of binder composition and w/c, the RF, MLP and LSTM learners are tuned by SSA, and the weights of the learners are selected for each regime from 66 candidate combinations. All combination rules were compared on the same 20 splits grouped by mixture and on the same base-model predictions. The main scientific findings are as follows.

(1) The weights selected for the regimes differed. RF received the largest weight in every regime, from 0.58 in the fly ash regime to 0.97 in the plain cement regime, and the neural learners contributed mainly in the fly ash and high-w/c slag regimes, where their weights varied considerably between splits.

(2) Regime-dependent weights did not improve the accuracy. On 20 splits grouped by mixture, CCS reached a mean test RMSE of 5.76 MPa, compared with 5.78 to 5.80 MPa for global weights fitted to the same predictions and 5.77 MPa for RF alone, and none of these differences was significant. Fitted weights of any kind were more accurate than equal weighting (5.75 to 5.80 versus 6.27 MPa), and the difference was significant for CCS. Cluster weights selected on small validation samples gave a higher mean error than one global vector. Random splits underestimated the error for new mixtures by about 1 MPa.

(3) The learned input effects agree with mix-design practice and binder chemistry. SHAP analysis ranked w/c (35.1%) and curing age (33.6%) as the dominant inputs, slag and fly ash received positive contributions at a given w/c, and the partial dependence surface showed the largest predicted benefit of slag at high w/c, consistent with the lower water-to-binder ratio that slag produces at a fixed w/c and with its latent hydraulic reaction.

For engineering practice, the study leads to the following outcomes.

(4) The prediction error can be expressed directly as a screening margin and as prediction intervals. With a mean RMSE of 5.76 MPa for new mixtures, 14.8% of the test results were overestimated by more than one strength class (5 MPa). With a margin of about 6 MPa set from the out-of-fold errors for a 10% risk, 11.5% of the test results were overestimated by more than the margin, and the 90% prediction intervals had a half-width of about 9.5 MPa and covered 88.7% of the test results.

(5) The predicted strength is accurate enough to supply the elastic modulus for preliminary structural analysis, because one RMSE changes the modulus by only about 5% to 10%, whereas resistance checks remain sensitive to the strength error and require tested strengths.

(6) Screening candidate mixtures with the model before trial batching can avoid some failed batches, each of which costs about a month of redesign, casting and curing (Figure 12), while standard 28-day compression tests remain the basis for accepting the selected mixture.

These conclusions are limited by the data and by the training of the base learners (Section 4.3). They rest on one public dataset whose raw materials are not characterized and on neural learners tuned with a shortened search. Future work should repeat the comparison with fully tuned learners, validate the approach on independent datasets and on other binders and concretes, such as silica fume, recycled aggregate, alkali-activated and fiber-reinforced lightweight concretes, and couple it with multi-objective optimization of strength, cost, and embodied CO_2.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/buildings16214348/s1, Tables S1a–S1c: Test R^2, RMSE and MAE of the combination rules and base learners on each grouped and random split; Table S2: Effect of the number of hidden layers of the MLP; Table S3: Effect of the number of trees of the RF.

Author Contributions: Z.W.: writing—original draft, investigation, data curation, formal analysis, and methodology. C.Z.: resources, project administration, validation, and investigation. S.D.: methodology, supervision, writing—review and editing, and conceptualization. X.G.: software, visualization, investigation, and data curation. X.H.: conceptualization, supervision, funding acquisition, writing—review and editing, and project administration. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Major Water Conservancy Project for Enhancing Grain Production Capacity in Heilongjiang Province (Embankment Protection Engineering) (Section 3), grant number WSLJ-JSFW-2025-02; the Heilongjiang Provincial Key Research and Development Program Guidance Project, grant number GZ20220138; and the Scientific Research Project of Heilongjiang Provincial Department of Ecology and Environment, grant number HST2022GF004.

Data Availability Statement: The data analyzed in this study are openly available in the UCI Machine Learning Repository at https://archive.ics.uci.edu/dataset/165/concrete+compressive+strength (accessed on 8 October 2026). The code used for the analyses is available from the corresponding author on request, and the per-split results are given in Supplementary Tables S1a–S1c.

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

References

1. *EN 1992-1-1:2004*; Eurocode 2: Design of Concrete Structures. Part 1-1: General Rules and Rules for Buildings. European Committee for Standardization: Brussels, Belgium, 2004.
2. *ACI 318-19*; Building Code Requirements for Structural Concrete. American Concrete Institute: Farmington Hills, MI, USA, 2019.
3. Moein, M.M.; Saradar, A.; Rahmati, K.; Mousavinejad, S.H.G.; Bristow, J.; Aramali, V.; Karakouzian, M. Predictive models for concrete properties using machine learning and deep learning approaches: A review. *J. Build. Eng.* **2023**, *63*, 105444. [CrossRef]
4. *ACI 211.1-91*; Standard Practice for Selecting Proportions for Normal, Heavyweight, and Mass Concrete. American Concrete Institute: Farmington Hills, MI, USA, 1991.
5. Cheng, D.; Reiner, D.M.; Yang, F.; Cui, C.; Meng, J.; Shan, Y.; Liu, Y.; Tao, S.; Guan, D. Projecting future carbon emissions from cement production in developing countries. *Nat. Commun.* **2023**, *14*, 8213. [CrossRef] [PubMed]
6. Scrivener, K.L.; John, V.M.; Gartner, E.M. Eco-efficient cements: Potential economically viable solutions for a low-CO_2 cement-based materials industry. *Cem. Concr. Res.* **2018**, *114*, 2–26. [CrossRef]
7. Lothenbach, B.; Scrivener, K.; Hooton, R.D. Supplementary cementitious materials. *Cem. Concr. Res.* **2011**, *41*, 1244–1256. [CrossRef]
8. Lee, T.; Lee, J.; Kim, J.; Choi, H.; Lee, D.-E. Effect of Formwork Removal Time Reduction on Construction Productivity Improvement by Mix Design of Early Strength Concrete. *Appl. Sci.* **2020**, *10*, 7046. [CrossRef]
9. Li, Z.; Yoon, J.; Zhang, R.; Rajabipour, F.; Srubar, W.V., III; Dabo, I.; Radlińska, A. Machine learning in concrete science: Applications, challenges, and best practices. *npj Comput. Mater.* **2022**, *8*, 127. [CrossRef]
10. Zhang, J.; Huang, Y.; Wang, Y.; Ma, G. Multi-objective optimization of concrete mixture proportions using machine learning and metaheuristic algorithms. *Constr. Build. Mater.* **2020**, *253*, 119208. [CrossRef]
11. DeRousseau, M.A.; Kasprzyk, J.R.; Srubar, W.V., III. Computational design optimization of concrete mixtures: A review. *Cem. Concr. Res.* **2018**, *109*, 42–53. [CrossRef]
12. Abrams, D.A. *Design of Concrete Mixtures*; Bulletin No. 1; Structural Materials Research Laboratory, Lewis Institute: Chicago, IL, USA, 1918.
13. Snellings, R.; Suraneni, P.; Skibsted, J. Future and emerging supplementary cementitious materials. *Cem. Concr. Res.* **2023**, *171*, 107199. [CrossRef]
14. Jhatial, A.A.; Nováková, I.; Gjerløw, E. A Review on Emerging Cementitious Materials, Reactivity Evaluation and Treatment Methods. *Buildings* **2023**, *13*, 526. [CrossRef]
15. Yu, Y.; Al-Damad, I.M.A.; Foster, S.; Nezhad, A.A.; Hajimohammadi, A. Compressive strength prediction of fly ash/slag-based geopolymer concrete using EBA-optimised chemistry-informed interpretable deep learning model. *Dev. Built Environ.* **2025**, *23*, 100736. [CrossRef]
16. Hassan, A.; Saleh, R.A.A.; Al-Sameai, H.; de Moura, J.; Alomayri, T.; Zhang, C. Novel hybrid machine learning framework for high-fidelity prediction of fly ash-based geopolymer concrete strength. *Compos. Struct.* **2026**, *378*, 119906. [CrossRef]
17. Kumar, R.; Althaqafi, E.; Patro, S.G.K.; Simic, V.; Babbar, A.; Pamucar, D.; Singh, S.K.; Verma, A. Machine and deep learning methods for concrete strength prediction: A bibliometric and content analysis review of research trends and future directions. *Appl. Soft Comput.* **2024**, *164*, 111956. [CrossRef]
18. Yeh, I.-C. Modeling of strength of high-performance concrete using artificial neural networks. *Cem. Concr. Res.* **1998**, *28*, 1797–1808. [CrossRef]
19. Li, Q.; Song, Z. High-performance concrete strength prediction based on ensemble learning. *Constr. Build. Mater.* **2022**, *324*, 126694. [CrossRef]
20. Arora, H.C.; Bhushan, B.; Kumar, A.; Kumar, P.; Hadzima-Nyarko, M.; Radu, D.; Cazacu, C.E.; Kapoor, N.R. Ensemble learning based compressive strength prediction of concrete structures through real-time non-destructive testing. *Sci. Rep.* **2024**, *14*, 1824. [CrossRef] [PubMed]
21. Zhang, Y.; Ren, W.; Chen, Y.; Mi, Y.; Lei, J.; Sun, L. Predicting the compressive strength of high-performance concrete using an interpretable machine learning model. *Sci. Rep.* **2024**, *14*, 28346. [CrossRef] [PubMed]
22. Ahmad, F.; Alasskar, A.; Samui, P.; Asteris, P.G. Machine learning-based Graphical User Interface for predicting high-performance concrete compressive strength: Comparative analysis of Gradient Boosting Machine, Random Forest, and Deep Neural Network Models. *Front. Struct. Civ. Eng.* **2025**, *19*, 1075–1090. [CrossRef]
23. Vairagade, V.S.; Bahoria, B.V.; Isleem, H.F.; Shelke, N.; Mungle, N.P. Strength and durability predictions of ternary blended nano-engineered high-performance concrete: Application of hybrid machine learning techniques with bio-inspired optimization. *Eng. Appl. Artif. Intell.* **2025**, *148*, 110470. [CrossRef]
24. Shafighfard, T.; Kazemi, F.; Asgarkhani, N.; Yoo, D.-Y. Machine-learning methods for estimating compressive strength of high-performance alkali-activated concrete. *Eng. Appl. Artif. Intell.* **2024**, *136*, 109053. [CrossRef]
25. Asteris, P.G.; Skentou, A.D.; Bardhan, A.; Samui, P.; Pilakoutas, K. Predicting concrete compressive strength using hybrid ensembling of surrogate machine learning models. *Cem. Concr. Res.* **2021**, *145*, 106449. [CrossRef]

26. Li, Q.; Song, Z. Prediction of compressive strength of rice husk ash concrete based on stacking ensemble learning model. *J. Clean. Prod.* **2023**, *382*, 135279. [CrossRef]
27. Taiwo, R.; Yussif, A.-M.; Adegoke, A.H.; Zayed, T. Prediction and deployment of compressive strength of high-performance concrete using ensemble learning techniques. *Constr. Build. Mater.* **2024**, *451*, 138808. [CrossRef]
28. Khan, M.I.; Abbas, Y.M. Intelligent data-driven compressive strength prediction and optimization of reactive powder concrete using multiple ensemble-based machine learning approach. *Constr. Build. Mater.* **2023**, *404*, 133148. [CrossRef]
29. Latif, S.D. Concrete compressive strength prediction modeling utilizing deep learning long short-term memory algorithm for a sustainable environment. *Environ. Sci. Pollut. Res.* **2021**, *28*, 30294–30302. [CrossRef] [PubMed]
30. Rathakrishnan, V.; Bt. Beddu, S.; Ahmed, A.N. Predicting compressive strength of high-performance concrete with high volume ground granulated blast-furnace slag replacement using boosting machine learning algorithms. *Sci. Rep.* **2022**, *12*, 9539. [CrossRef] [PubMed]
31. Al-Taai, S.R.; Azize, N.M.; Thoeny, Z.A.; Imran, H.; Bernardo, L.F.A.; Al-Khafaji, Z. XGBoost Prediction Model Optimized with Bayesian for the Compressive Strength of Eco-Friendly Concrete Containing Ground Granulated Blast Furnace Slag and Recycled Coarse Aggregate. *Appl. Sci.* **2023**, *13*, 8889. [CrossRef]
32. Nagaraju, T.V.; Mantena, S.; Azab, M.; Alisha, S.S.; El Hachem, C.; Adamu, M.; Rama Murthy, P.S. Prediction of high strength ternary blended concrete containing different silica proportions using machine learning approaches. *Results Eng.* **2023**, *17*, 100973. [CrossRef]
33. Stel'makh, S.A.; Shcherban', E.M.; Beskopylny, A.N.; Mailyan, L.R.; Meskhi, B.; Razveeva, I.; Kozhakin, A.; Beskopylny, N. Prediction of Mechanical Properties of Highly Functional Lightweight Fiber-Reinforced Concrete Based on Deep Neural Network and Ensemble Regression Trees Methods. *Materials* **2022**, *15*, 6740. [CrossRef] [PubMed]
34. Pakzad, S.S.; Roshan, N.; Ghalehnovi, M. Comparison of various machine learning algorithms used for compressive strength prediction of steel fiber-reinforced concrete. *Sci. Rep.* **2023**, *13*, 3646. [CrossRef] [PubMed]
35. Mai, H.-V.T.; Nguyen, M.H.; Ly, H.-B. Development of machine learning methods to predict the compressive strength of fiber-reinforced self-compacting concrete and sensitivity analysis. *Constr. Build. Mater.* **2023**, *367*, 130339. [CrossRef]
36. Kına, C.; Tanyıldızı, H.; Türk, K. Forecasting the compressive strength of GGBFS-based geopolymer concrete via ensemble predictive models. *Constr. Build. Mater.* **2023**, *405*, 133299. [CrossRef]
37. Ranasinghe, R.S.S.; Kulasooriya, W.K.V.J.B.; Perera, U.S.; Ekanayake, I.U.; Meddage, D.P.P.; Mohotti, D.; Rathanayake, U. Eco-friendly mix design of slag-ash-based geopolymer concrete using explainable deep learning. *Results Eng.* **2024**, *23*, 102503. [CrossRef]
38. Kumar, S.; Kumar, R.; Sirimontree, S.; Kumar, D.R.; Wipulanusat, W.; Keawsawasvong, S.; Thongchom, C. Smart prediction: Hybrid random forest for high-volume fly ash self-compacting concrete strength. *Front. Struct. Civ. Eng.* **2025**, *19*, 892–918. [CrossRef]
39. Jacobs, R.A.; Jordan, M.I.; Nowlan, S.J.; Hinton, G.E. Adaptive mixtures of local experts. *Neural Comput.* **1991**, *3*, 79–87. [CrossRef] [PubMed]
40. Cruz, R.M.O.; Sabourin, R.; Cavalcanti, G.D.C. Dynamic classifier selection: Recent advances and perspectives. *Inf. Fusion* **2018**, *41*, 195–216. [CrossRef]
41. Li, M.; Yang, M.; Yu, Y.; Shahidehpour, M.; Wen, F. Adaptive weighted combination approach for wind power forecast based on deep deterministic policy gradient method. *IEEE Trans. Power Syst.* **2024**, *39*, 3075–3087. [CrossRef]
42. Fu, Y.; Wu, D.; Boulet, B. Reinforcement learning based dynamic model combination for time series forecasting. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 22 February–1 March 2022; Volume 36, pp. 6639–6647. [CrossRef]
43. Feng, C.; Sun, M.; Zhang, J. Reinforced deterministic and probabilistic load forecasting via Q-learning dynamic model selection. *IEEE Trans. Smart Grid* **2020**, *11*, 1377–1386. [CrossRef]
44. Nagwani, N.K.; Deo, S.V. Estimating the concrete compressive strength using hard clustering and fuzzy clustering based regression techniques. *Sci. World J.* **2014**, *2014*, 381549. [CrossRef] [PubMed]
45. Demetriou, D.; Polydorou, T.; Nicolaides, D.; Petrou, M.F. A clustering machine learning approach for improving concrete compressive strength prediction. *Eng. Rep.* **2024**, *6*, e12934. [CrossRef]
46. Xue, J.; Shen, B. A novel swarm intelligence optimization approach: Sparrow search algorithm. *Syst. Sci. Control Eng.* **2020**, *8*, 22–34. [CrossRef]
47. Li, P.; Zhang, Z.; Gu, J. Prediction of concrete compressive strength based on ISSA-BPNN-AdaBoost. *Materials* **2024**, *17*, 5727. [CrossRef] [PubMed]
48. Ekanayake, I.U.; Meddage, D.P.P.; Rathnayake, U. A novel approach to explain the black-box nature of machine learning in compressive strength predictions of concrete using Shapley additive explanations (SHAP). *Case Stud. Constr. Mater.* **2022**, *16*, e01059. [CrossRef]
49. Kashem, A.; Karim, R.; Malo, S.C.; Das, P.; Datta, S.D.; Alharthai, M. Hybrid data-driven approaches to predicting the compressive strength of ultra-high-performance concrete using SHAP and PDP analyses. *Case Stud. Constr. Mater.* **2024**, *20*, e02991. [CrossRef]

50. Akber, M.Z. Improving the experience of machine learning in compressive strength prediction of industrial concrete considering mixing proportions, engineered ratios and atmospheric features. *Constr. Build. Mater.* **2024**, *444*, 137884. [CrossRef]
51. Li, Z.; Pei, T.; Ying, W.; Srubar, W.V., III; Zhang, R.; Yoon, J.; Ye, H.; Dabo, I.; Radlińska, A. Can domain knowledge benefit machine learning for concrete property prediction? *J. Am. Ceram. Soc.* **2024**, *107*, 1582–1602. [CrossRef]
52. Kapoor, S.; Narayanan, A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* **2023**, *4*, 100804. [CrossRef] [PubMed]
53. Breiman, L. Stacked regressions. *Mach. Learn.* **1996**, *24*, 49–64. [CrossRef]
54. Demšar, J. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.* **2006**, *7*, 1–30.
55. Nadeau, C.; Bengio, Y. Inference for the generalization error. *Mach. Learn.* **2003**, *52*, 239–281. [CrossRef]
56. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **1979**, *6*, 65–70.
57. Gers, F.A.; Schmidhuber, J.; Cummins, F. Learning to forget: Continual prediction with LSTM. *Neural Comput.* **2000**, *12*, 2451–2471. [CrossRef] [PubMed]
58. Gharehchopogh, F.S.; Namazi, M.; Ebrahimi, L.; Abdollahzadeh, B. Advances in sparrow search algorithm: A comprehensive survey. *Arch. Comput. Methods Eng.* **2023**, *30*, 427–455. [CrossRef] [PubMed]
59. Lundberg, S.M.; Lee, S.-I. A unified approach to interpreting model predictions. In Proceedings of the Advances in Neural Information Processing Systems 30 (NeurIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 4765–4774.
60. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
61. Lei, J.; G'Sell, M.; Rinaldo, A.; Tibshirani, R.J.; Wasserman, L. Distribution-free predictive inference for regression. *J. Am. Stat. Assoc.* **2018**, *113*, 1094–1111. [CrossRef]
62. Ahmad, H.; Ejaz, M.F.; Riaz, M.R.; Shahzad, S.; El Kadri, S.; Kmeid, M.G. Machine learning-based prediction of compressive and flexural strength of wheat straw reinforced sustainable gypsum composites. *Sci. Rep.* **2026**, *16*, 15087. [CrossRef] [PubMed]
63. Nikoopayan Tak, M.S.; Feng, Y.; Mahgoub, M. Advanced machine learning techniques for predicting concrete compressive strength. *Infrastructures* **2025**, *10*, 26. [CrossRef]
64. Claeskens, G.; Magnus, J.R.; Vasnev, A.L.; Wang, W. The forecast combination puzzle: A simple theoretical explanation. *Int. J. Forecast.* **2016**, *32*, 754–762. [CrossRef]
65. Wang, X.; Hyndman, R.J.; Li, F.; Kang, Y. Forecast combinations: An over 50-year review. *Int. J. Forecast.* **2023**, *39*, 1518–1547. [CrossRef]
66. Li, E.; Zhang, N.; Xi, B.; Zhou, J.; Gao, X. Compressive strength prediction and optimization design of sustainable concrete based on squirrel search algorithm-extreme gradient boosting technique. *Front. Struct. Civ. Eng.* **2023**, *17*, 1310–1325. [CrossRef]

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