



Study-Disjoint Machine Learning for Concrete Compressive Strength

Muhammad Safi Ullah^{1*}, Laiba Qamar²

¹ Lecturer, Department of Civil Engineering, Swedish College of Engineering and Technology, Pakistan

² Research Scholar, Department of Biosciences, University of Wah, Pakistan

e-mail: muhammadsafiullah64@gmail.com

DOI : <https://doi.org/10.55248/gengpi.07.0926.2923>

ABSTRACT:

This study developed and evaluated machine-learning models for predicting 28-day cylinder compressive strength of concrete using a literature-derived dataset of 1,456 mixtures collected from 136 source studies. Eight recorded mixture quantities were used as the original inputs. Source-study identity and cross-study duplicate relationships were incorporated into the validation design to prevent linked observations from appearing on both sides of a train-test split. Seven prediction approaches were assessed using repeated nested grouped cross-validation, with model selection and hyperparameter tuning performed only within the training partitions. ExtraTrees gave the lowest mean outer-fold error among the evaluated model families, with an MAE of 9.758 MPa, RMSE of 12.884 MPa, and study-macro-MAE of 10.392 MPa. Conventional random validation produced substantially more optimistic performance than study-disjoint validation; for ExtraTrees, R^2 decreased from 0.844 to 0.471 and MAE increased from 4.783 to 9.807 MPa. Applicability-domain assessment showed higher errors for observations classified outside the represented training domain. Permutation importance identified the water-to-binder ratio as the strongest model-specific predictor. Group-aware prediction intervals achieved approximately 91% empirical coverage for a nominal 90% level. The results show that study-disjoint validation provides a more rigorous estimate of model transferability across literature-derived concrete datasets than conventional random splitting.

Keywords: Concrete compressive strength; Machine learning; Supplementary cementitious materials; ExtraTrees; Grouped cross-validation; Applicability domain; Permutation importance; Prediction uncertainty

1. Introduction

Concrete compressive strength is a primary parameter used to assess the mechanical performance of concrete and its suitability for structural applications. Its value depends on the proportions and characteristics of cement, water, aggregates, supplementary cementitious materials, chemical admixtures, and curing conditions. As concrete mixtures become more complex, particularly with the use of fly ash, silica fume, and ground granulated blast-furnace slag, conventional empirical relationships become less capable of representing the nonlinear interactions between mixture constituents and strength [1], [2]. Recent reviews therefore identify machine learning as an effective data-driven approach for modelling the mechanical behaviour of modern concrete mixtures [1]-[3].

Machine-learning methods have consequently been applied extensively to concrete compressive-strength prediction. Ahmad et al. [4] evaluated several supervised-learning approaches for concrete containing supplementary cementitious materials and reported that ensemble methods could model the nonlinear relationships between mixture composition and strength. Aslam and Shahab [5] subsequently compared eleven machine-learning techniques for blended concrete containing several supplementary cementitious materials. Recent studies have also used larger literature-derived datasets. Loureiro and Stefani [6] evaluated several regression and machine-learning methods using 1,234 concrete-strength records, while Elshaarawy et al. [7] developed predictive models from 1,030 observations. These studies demonstrate the ability of machine-learning algorithms to extract useful relationships from multivariable concrete datasets.

The increasing size of literature-derived datasets has also allowed more diverse concrete systems to be evaluated. Liu et al. [8], for example, compiled 3,382 compressive-strength records from 208 published studies for machine-learning prediction and interpretation of manufactured-sand concrete. Such datasets provide considerably greater compositional diversity than individual experimental programs. However, the use of data collected from many publications introduces an important validation issue. Observations originating from the same source study can share material sources, testing procedures, specimen preparation, laboratory conditions, and other characteristics that are not explicitly included in the predictor matrix. If individual rows are divided randomly between training and testing sets, records from the same experimental source can occur in both partitions.

This issue is important because the reported accuracy of a machine-learning model depends on the validation problem being tested. Several recent concrete-strength studies have relied on random train-test division or conventional k -fold cross-validation [4]-[7]. These approaches are suitable for estimating interpolation within a dataset, but they do not necessarily represent prediction for mixtures originating from previously unseen source studies. Similar concerns have been identified more broadly in materials informatics. Shimakawa et al. [9] showed that observations from the same experimental group appearing in both training and testing data can inflate apparent model performance. Group-based validation has therefore been recommended when the intended prediction task requires transfer to previously unseen experimental groups or material families [9], [10].

Validation alone is also insufficient for judging whether a model can be used reliably. Concrete machine-learning studies increasingly include feature interpretation to determine which mixture variables carry the strongest predictive information [5], [7], [8]. However, prediction reliability also depends

on whether a new observation lies within a region represented by the training data and on the uncertainty associated with the predicted value. Recent work in concrete and materials machine learning has therefore emphasized careful cross-validation, assessment of model generalization, interpretation of predictor effects, and explicit consideration of observations outside the represented training domain [3], [8], [11].

The present study addresses these issues using a literature-derived dataset containing 1,456 concrete mixtures from 136 source studies. The target variable is 28-day cylinder compressive strength, predicted from eight recorded mixture quantities: cement, water, coarse aggregate, fine aggregate, fly ash, silica fume, ground granulated blast-furnace slag, and superplasticizer. The study first audits source-study identity and cross-study duplicate relationships so that linked observations cannot occur on opposite sides of the validation split. Multiple regression and ensemble-learning approaches are then evaluated using repeated nested grouped cross-validation. A separate random-versus-grouped benchmark is used to quantify the effect of validation strategy. Model performance is further examined using study-level error, prediction uncertainty, applicability-domain assessment, and permutation importance. The objective is therefore not only to identify the best-performing model among the evaluated algorithms, but also to determine how reliably the resulting model transfers to concrete mixtures originating from source studies that were not used for model training.

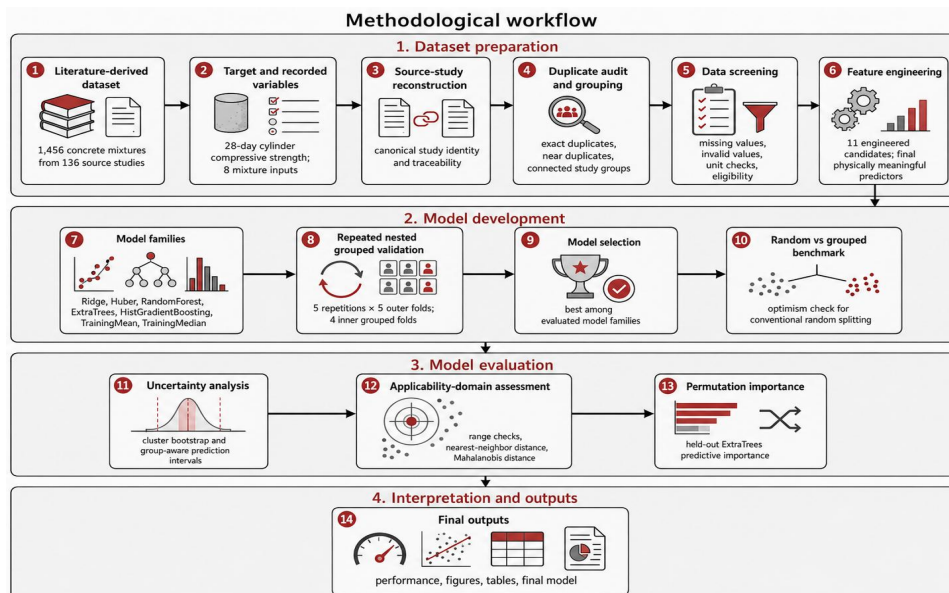
1.1. Synthesis of Literature and Novelty Positioning

Recent studies show that ensemble, boosting, AutoML, and interpretable models can predict concrete compressive strength effectively [12]–[24]. Uncertainty-aware prediction is also emerging [25]–[27]. However, most studies emphasize row-level accuracy, model comparison, or feature interpretation rather than transfer to unseen source studies. Materials-ML research separately stresses nested validation and applicability-domain control [28]–[31]. The present study combines source-study auditing, duplicate-linkage control, repeated nested grouped cross-validation, random-versus-grouped benchmarking, study-level error analysis, uncertainty estimation, and fold-fitted applicability-domain checks to evaluate cross-study generalization.

Theme	Literature synthesis	Remaining gap
Prediction	Ensemble and boosting models are consistently strong [12]–[18]. AutoML extends model search [19].	Mostly row-level validation.
Interpretation	EBM, SHAP, and feature-based interpretation are common [19]–[24].	Interpretation does not test source transfer.
Uncertainty	Recent work adds uncertainty and conformal intervals [25]–[27].	Coverage may weaken under source shift.
Validation / domain	Nested evaluation and applicability-domain control are recommended in materials ML [28]–[31].	Rarely integrated with source-grouped concrete validation.

2. Methodology

The methodology began with a literature-derived dataset comprising 1,456 concrete mixtures collected from 136 source studies. Eight recorded mixture variables were used as predictors of 28-day cylinder compressive strength. Source-study identities were reconstructed, while duplicate and near-duplicate observations were systematically audited. Studies containing linked or highly similar records were grouped to prevent information leakage across training and testing partitions. The dataset was then screened for consistency and transformed to derive physically meaningful engineered features representing key mixture relationships. Seven prediction approaches were evaluated using repeated nested grouped cross-validation, ensuring that observations from the same source group remained within a single validation partition. Hyperparameter optimization was performed exclusively within the training folds to avoid optimistic performance estimates. Model performance was assessed using MAE, RMSE, and R^2 , together with study-level error analysis to evaluate transfer across unseen source studies. Prediction reliability was further examined using uncertainty analysis, applicability-domain assessment, and held-out permutation importance. The random-versus-grouped validation benchmark was also used to quantify the effect of validation strategy on apparent model performance. Based on the complete grouped-validation results, ExtraTrees was identified as the best-performing model family among the evaluated approaches and was therefore selected for subsequent interpretation and reliability assessment.



2.1. Dataset and prediction task

The workbook was constructed from published literature. It contained 1,456 concrete mixtures from 136 source studies. The prediction target was 28-day cylinder compressive strength. Eight recorded mixture quantities were used as raw inputs: cement, water, coarse aggregate, fine aggregate, fly ash (FA), silica fume (SF), ground granulated blast-furnace slag (GGBFS), and superplasticizer (SP). All mixture quantities were expressed in kg/m³. Mix ID and Reference were retained only to trace observations and form validation groups. They were not used as model inputs.

Table 1. Raw variables used for 28-day compressive-strength prediction.

Variable	Workbook field	Unit
Cement	Cement	kg/m ³
Water	Water	kg/m ³
Coarse aggregate	Coarse aggregate	kg/m ³
Fine aggregate	Fine aggregate	kg/m ³
Fly ash	FA	kg/m ³
Silica fume	SF	kg/m ³
GGBFS	GGBFS	kg/m ³
Superplasticizer	SP	kg/m ³
Target	Cylinder compressive strength	MPa

2.2. Source-study grouping and duplicate control

Each publication was treated as one source study. In the workbook, the publication reference appeared at the first row of a study block and the following rows were blank. These blank cells were assigned the most recent nonblank reference until the next publication was reached. DOI and URL formats were then standardized so that one publication had one study ID. Every modelling row was required to have a valid study ID.

Duplicate control was completed before any model was trained. Exact matches were checked using the eight raw predictors with strength, the eight raw predictors without strength, and all available numeric fields. Near-duplicate mixtures were also checked. For this check, engineered variables were centered by their median and divided by their population standard deviation. Records from different studies with Euclidean distance at or below 0.08 were flagged as near duplicates.

If a record from one study was linked to a record from another study, all rows from both studies were placed in the same validation group. The same rule was applied to connected chains of studies. Cross-validation then split these complete validation groups. This kept linked source records on one side of a train-test split.

2.3. Data screening

The eight raw inputs and the target were checked for missing values, the sentinel value -1, negative values, zeros, and extreme ranges. Binder total and aggregate total were required to be greater than zero before ratios were calculated. Rows were flagged for review when water-to-binder ratio was below 0.15 or above 1.20, total SCM-to-binder ratio exceeded 0.95, SP-to-binder ratio exceeded 0.08, total recorded mass was below 1,200 or above 3,000 kg/m³, or compressive strength was below 1 or above 150 MPa. These checks were diagnostic. No row was removed by the pipeline.

The raw predictor vector was also checked for identical inputs associated with different strength values. These records were retained and counted before validation.

2.4. Feature engineering

The raw quantities were converted to totals and ratios that describe binder composition, water demand, and aggregate composition. Binder total (B), total SCM content (S), and aggregate total (A) were calculated as follows:

$B = \text{Cement} + \text{FA} + \text{SF} + \text{GGBFS}$	$\text{SCM}/b = S / B$
$S = \text{FA} + \text{SF} + \text{GGBFS}$	$\text{SF}/b = \text{SF} / B$
$A = \text{Coarse aggregate} + \text{Fine aggregate}$	$\text{GGBFS}/b = \text{GGBFS} / B$
$w/b = \text{Water} / B$	$\text{SP}/b = \text{SP} / B$

Eleven engineered candidates were generated. Pearson and Spearman correlations were calculated to identify redundant variables. An absolute Pearson correlation of 0.98 was used as the redundancy flag. Two predefined feature sets were evaluated. RAW_ONLY contained the eight original quantities. ENGINEERED_PHYSICAL contained binder total, water-to-binder ratio, total SCM-to-binder ratio, SF-to-binder ratio, GGBFS-to-binder ratio, SP-to-binder ratio, aggregate total, and coarse aggregate fraction. The main analysis used ENGINEERED_PHYSICAL.

2.5. Candidate models and hyperparameter tuning

Seven model choices were compared. TrainingMean and TrainingMedian were simple reference predictors. Ridge and Huber were linear models. ExtraTrees, RandomForest, and HistGradientBoosting were nonlinear ensemble models. Ridge and Huber were standardized within their model pipelines. The tree-based models used the engineered values directly. Hyperparameters were selected only inside the inner grouped cross-validation described in Section 2.6.

A fixed pseudo-random seed of 20260903 was used whenever a split or model required random sampling. This fixes the random choices made by the

program, so the same data and code reproduce the same stochastic splits and model sampling.

Table 2. Candidate models and hyperparameter search used inside grouped inner cross-validation.

Model	Configurations evaluated
TrainingMean	Training-set mean
TrainingMedian	Training-set median
Ridge	$\alpha = 1, 10, 100$
Huber	$(\alpha, \epsilon) = (0.001, 1.35)$ or $(0.01, 1.50)$
ExtraTrees	40 trees; max_features = 0.7 or 1.0; min_samples_leaf = 2; max_depth = None
RandomForest	40 trees; max_features = 0.7 or 1.0; min_samples_leaf = 2; max_depth = None
HistGradient Boosting	100 iterations; learning rate=0.05 or 0.08; max leaf nodes=31; min samples leaf=15

2.6. Repeated nested grouped validation and model selection

The main validation used five repetitions of five-fold outer cross-validation. The complete validation groups defined in Section 2.2 were split, not individual rows. In every outer fold, the test groups were absent from the training data. Each outer training partition was then divided into four grouped inner folds for hyperparameter and model selection.

The main inner selection score was study-macro-MAE. MAE was first calculated separately for each source study in the inner validation fold. These study MAE values were then averaged with equal weight. For each outer fold, the best configuration of every model family was fitted on the complete outer training partition and evaluated on the untouched outer test partition. The family with the lowest inner selection score produced the out-of-fold prediction for that outer fold. Family-specific outer-fold results were also retained for direct model comparison.

The program checked source-study overlap and validation-group overlap for every inner and outer split. A split was rejected if either overlap was nonzero. Each observation appeared once in an outer test fold during each repetition. Final model-family selection used mean outer-fold MAE first, followed by study-macro-MAE, MAE variation, and RMSE. The selected family was then tuned with grouped cross-validation on the full dataset and fitted to all 1,456 observations.

2.7. Validation benchmark and performance measures

A separate 10-fold benchmark was used to measure the effect of the split strategy. Ordinary random cross-validation split individual rows. Grouped cross-validation split the complete validation groups from Section 2.2. The benchmark was kept separate from the repeated nested analysis.

Prediction performance was reported with MAE, RMSE, R^2 , median absolute error, and signed bias. For the equations below, y_i is the observed strength, p_i is the predicted strength, and n is the number of observations.

$$MAE = \left(\frac{1}{n}\right) \sum abs(y_i - p_i) \quad \text{eq (1)}$$

$$RMSE = \sqrt{\frac{1}{2} \sum (y_i - p_i)^2} \quad \text{eq (2)}$$

$$R^2 = 1 - \frac{\sum (y_i - p_i)^2}{\sum (y_i - y_{mean})^2} \quad \text{eq(3)}$$

$$Bias = \left(\frac{1}{n}\right) \sum (p_i - y_i) \quad \text{eq(4)}$$

Study-macro-MAE and study-macro RMSE were calculated by first computing the metric within each source study and then taking the unweighted mean across studies.

2.8. Uncertainty analysis

Metric uncertainty was estimated with a study-cluster bootstrap. Source studies were sampled with replacement, and all rows from a sampled study were retained together. One thousand bootstrap samples were generated for each validation repetition. The 2.5th and 97.5th percentiles formed the 95% interval for each metric. Prediction intervals were calculated inside each outer training partition with group-aware split conformal calibration. Twenty percent of the outer-training validation groups were assigned to calibration. The fold-selected model was fitted on the remaining groups. Absolute calibration residuals were calculated, and the finite-sample 90% quantile was used as the interval radius. The outer-test prediction was reported as the point prediction plus and minus this radius.

3. Results and Discussion

3.1. Dataset audit and validation groups

The 1,456 compressive-strength observations ranged from 6.09 to 122.58 MPa. Study reconstruction produced 136 source-study IDs. The cross-study duplicate audit merged linked studies into 130 validation groups. Eleven studies were involved in five connected duplicate components, and the largest component contained three studies. The audit identified three exact cross-study duplicate groups, 41 cross-study raw-predictor duplicate groups, and 64 near-duplicate pairs. In total, 294 rows belonged to connected duplicate components.

Identical recorded input vectors with different strength values were found in 153 rows. These rows formed 66 conflict groups. Twenty-eight groups occurred within one source study and 38 occurred across studies. The largest strength spread within an identical-input group was 26.38 MPa. All rows were retained. The study-based grouping therefore kept both publication identity and cross-study record similarity under control during validation.

Table 3. Dataset and grouping results used in the final validation design.

Item	Result
Modeled observations	1,456
Source studies	136
Validation groups after duplicate linking	130
Studies in duplicate components	11
Rows in duplicate components	294
Exact cross-study duplicate groups	3
Cross-study predictor duplicate groups	41
Cross-study near-duplicate pairs	64
Identical-input conflicting-target rows	153
Identical-input conflict groups	66

3.2. Effect of random and grouped validation

The validation strategy changed the apparent model accuracy substantially. In the benchmark shown in Figure 1, the best random-split result was obtained by HistGradientBoosting with MAE = 4.772 Mpa, RMSE = 6.912 Mpa, and $R^2 = 0.850$. The best grouped result was obtained by ExtraTrees with MAE = 9.807 Mpa, RMSE = 12.969 Mpa, and $R^2 = 0.471$. The grouped test therefore produced much larger prediction errors and a lower coefficient of determination.

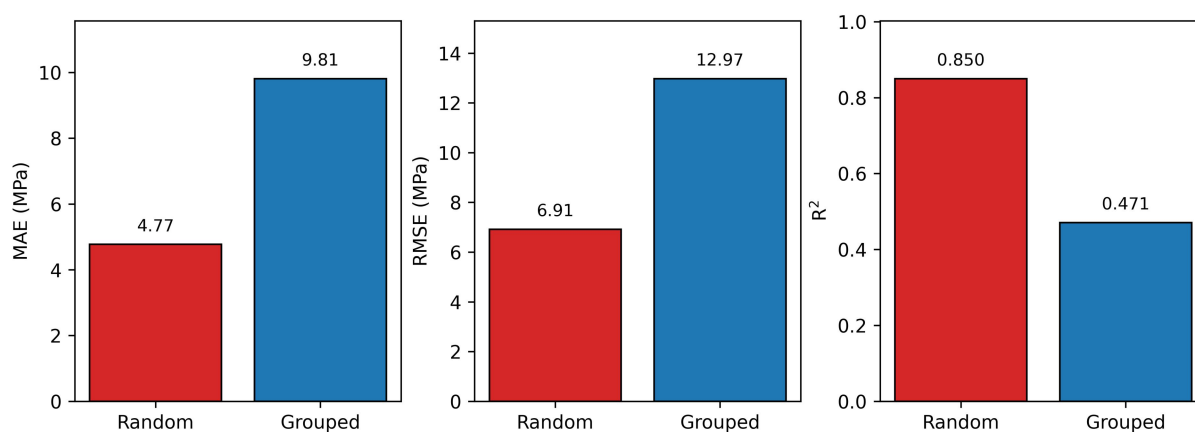


Figure 1. Best random and grouped 10-fold benchmark results. The random benchmark selected HistGradientBoosting, while the grouped benchmark selected ExtraTrees. The same-family ExtraTrees comparison is reported in Table 4.

The split effect remained when the model family was held constant. ExtraTrees gave MAE = 4.783 Mpa, RMSE = 7.052 Mpa, and $R^2 = 0.844$ under random 10-fold validation. Under grouped 10-fold validation, the same family gave MAE = 9.807 Mpa, RMSE = 12.969 Mpa, and $R^2 = 0.471$. MAE therefore increased by 5.025 Mpa, RMSE increased by 5.917 Mpa, and R^2 decreased by 0.373. The size of this same-family change shows that the split strategy was the main source of the performance difference, rather than a change in algorithm alone.

Table 4. Same-family ExtraTrees benchmark under random and grouped 10-fold validation.

Validation	MAE (MPa)	RMSE (MPa)	R^2
Random 10-fold	4.783	7.052	0.844
Grouped 10-fold	9.807	12.969	0.471
Grouped – Random	+5.025	+5.917	–0.373

3.3. Repeated nested out-of-fold prediction

The repeated nested procedure produced 7,280 out-of-fold predictions, equal to 1,456 observations evaluated once in each of five repetitions. ExtraTrees was selected inside 21 of the 25 outer folds. RandomForest and HistGradientBoosting were each selected in two folds. Across all repeated out-of-fold predictions, MAE was 9.754 MPa, RMSE was 12.984 MPa, and R^2 was 0.469. Median absolute error was 7.676 MPa and signed bias was –0.181 MPa.

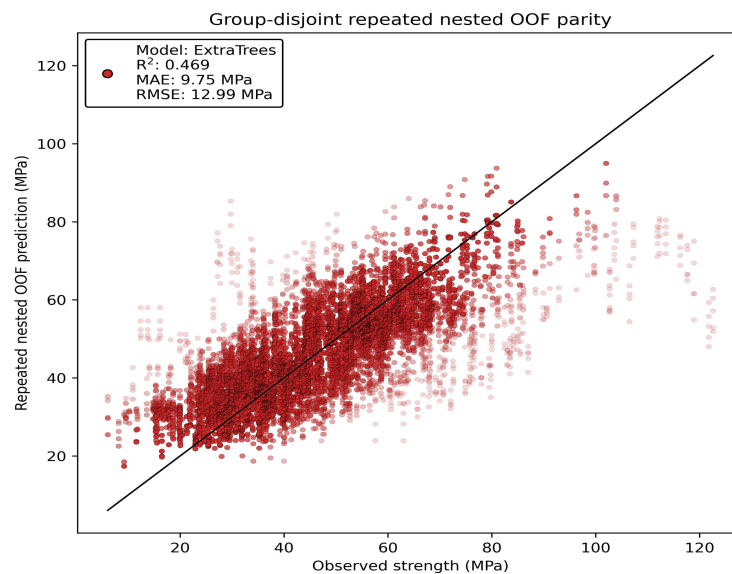


Figure 2. Observed 28-day compressive strength versus repeated nested group-disjoint out-of-fold predictions. The points were generated by fold-specific model selection. ExtraTrees was the final family selected from the family-level comparison.

The prediction cloud followed the 1:1 line through most of the central strength range. Dispersion increased as observed strength increased, and many observations above approximately 90 MPa were predicted below the identity line. The overall bias remained close to zero because positive and negative errors largely balanced across the full dataset. Study-macro-MAE was 10.510 MPa and study-macro RMSE was 11.501 MPa. Study-level MAE exceeded 15 MPa in 23 of the 136 studies and exceeded 20 MPa in 12 studies. The study-macro score therefore captured between-study variation that was not visible from the pooled row-level MAE alone.

Table 5. Repeated nested out-of-fold performance.

Metric	Result
MAE	9.754 MPa
RMSE	12.984 MPa
R ²	0.469
Median absolute error	7.676 MPa
Bias	−0.181 MPa
Study-macro MAE	10.510 MPa
Study-macro RMSE	11.501 MPa

3.4. Model comparison and final model selection

ExtraTrees had the lowest mean outer-fold MAE under the main engineered feature set and study-macro selection criterion. Its mean MAE was 9.758 MPa with a standard deviation of 1.555 MPa. Study-macro MAE was 10.392 MPa, mean RMSE was 12.884 MPa, and mean fold R² was 0.425. ExtraTrees ranked first in 13 of the 25 outer folds and had a mean rank of 1.68.

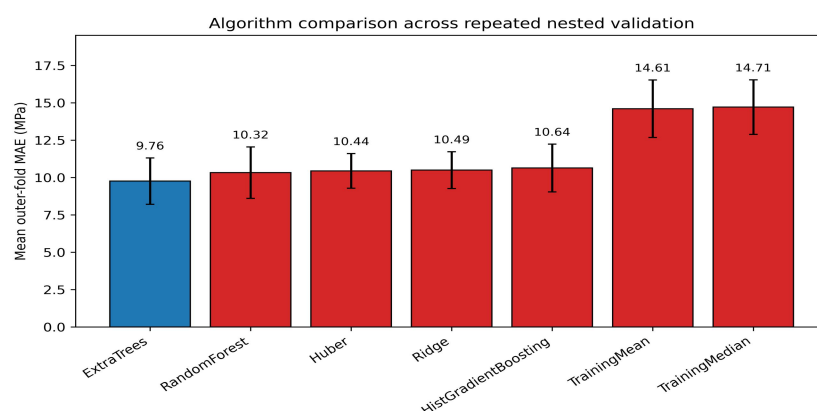


Figure 3. Mean outer-fold MAE for the candidate models across repeated nested grouped validation. Error bars show the standard deviation across outer folds.

RandomForest was the next ensemble model with mean MAE = 10.325 MPa. Huber and Ridge produced 10.442 and 10.494 MPa, respectively. HistGradientBoosting produced 10.636 MPa. The training-mean and training-median reference models remained above 14.6 MPa. ExtraTrees therefore gave the lowest average error and the strongest rank stability among the tested families. It was selected as the final model family. The engineered feature set produced only a small change relative to the eight raw quantities. For ExtraTrees under study-macro selection, RAW_ONLY gave MAE =

9.895 MPa and study-macro MAE = 10.443 MPa. ENGINEERED PHYSICAL gave 9.758 and 10.392 MPa, respectively. The engineered representation therefore reduced mean MAE by 0.137 MPa while expressing the mixture in binder, SCM, water, and aggregate ratios.

Table 6. Candidate-model performance under repeated nested grouped validation.

Model	MAE $\pm$ SD (MPa)	Study-macro MAE	RMSE	Mean R ²	Wins
ExtraTrees	9.758 $\pm$ 1.555	10.392	12.884	0.425	13
RandomForest	10.325 $\pm$ 1.720	10.615	13.514	0.368	4
Huber	10.442 $\pm$ 1.164	11.023	13.677	0.350	4
Ridge	10.494 $\pm$ 1.229	11.353	13.558	0.369	2
HistGradient Boosting	10.636 $\pm$ 1.599	10.931	13.852	0.327	2
Training mean	14.607 $\pm$ 1.925	14.696	17.933	−0.067	0
Training median	14.712 $\pm$ 1.830	14.660	18.108	−0.092	0

3.5. Applicability-domain behavior

The applicability-domain analysis contained the same 7,280 repeated out-of-fold records. Of these, 6,599 were classified as IN_DOMAIN and 681 received OOD_WARNING. Mean absolute error was 9.248 MPa for IN_DOMAIN records and 14.660 MPa for OOD_WARNING records. The corresponding median errors were 7.397 and 11.203 MPa. The warning group therefore had a mean absolute error 5.412 MPa higher than the in-domain group.

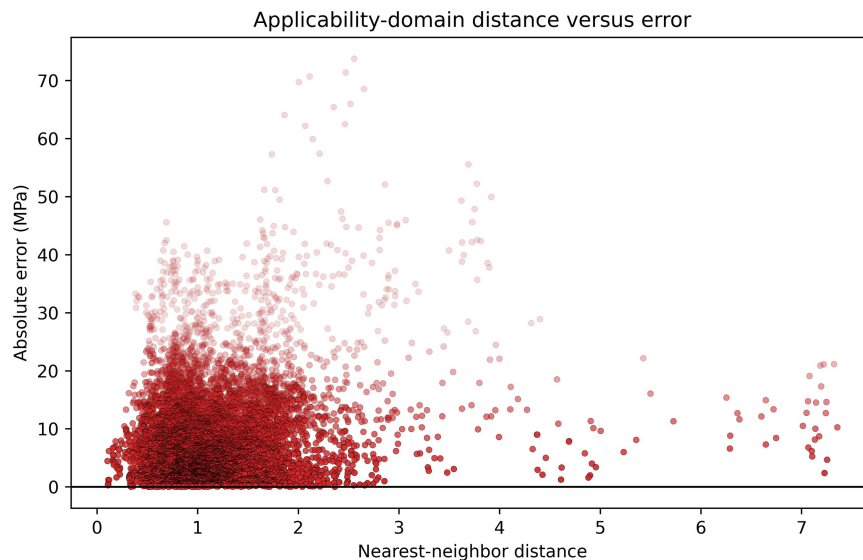


Figure 4. Nearest-neighbor applicability-domain distance versus absolute error for repeated nested group-disjoint out-of-fold predictions.

Nearest-neighbor distance had a weak positive rank association with absolute error (Spearman $\rho = 0.102$). Standardized covariance distance and squared Mahalanobis distance each gave $\rho = 0.165$. The continuous association was therefore weak, but the combined domain rule still separated a subset with higher average error. This is also visible in Figure 4, where large errors occur mainly within the central distance range rather than increasing monotonically with distance.

Table 7. Out-of-fold error by applicability-domain status.

Status	OOF records	Mean absolute error (MPa)	Median absolute error (MPa)
IN_DOMAIN	6,599	9.248	7.397
OOD_WARNING	681	14.660	11.203

3.6. ExtraTrees permutation importance

Water-to-binder ratio produced the largest increase in ExtraTrees MAE when it was permuted. The mean increase was 4.045 MPa with a standard deviation of 1.395 MPa, and it was positive in every outer evaluation. Total SCM-to-binder ratio ranked second at 1.178 MPa, followed by SF-to-binder ratio at 0.965 MPa and binder total at 0.506 MPa.

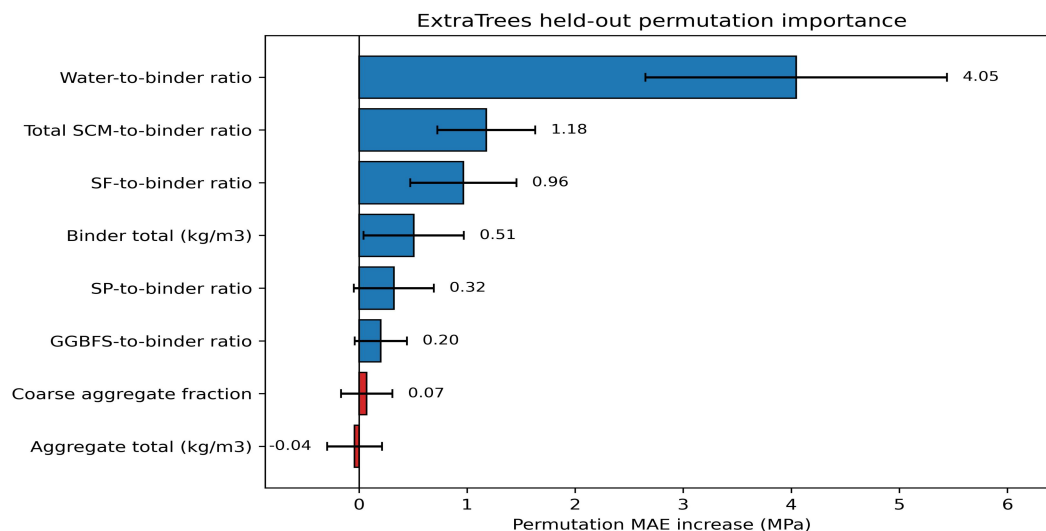


Figure 5. ExtraTrees held-out permutation importance across outer test folds. Importance is the increase in MAE after a predictor is permuted.

SP-to-binder ratio and GGBFS-to-binder ratio produced smaller mean MAE increases of 0.322 and 0.201 MPa. Coarse aggregate fraction was close to zero at 0.070 MPa. Aggregate total was -0.041 MPa and was positive in 40% of outer evaluations. Within the fitted ExtraTrees models, water-to-binder ratio therefore carried the strongest predictive signal, while total aggregate mass added little independent predictive information after the other engineered variables were present.

3.7. Metric and prediction uncertainty

The study-cluster bootstrap gave a 95% interval of 8.469 to 11.331 MPa for MAE and 11.114 to 15.213 MPa for RMSE. The corresponding R^2 interval was 0.344 to 0.563. Median absolute error ranged from 6.605 to 9.242 MPa, while the bias interval ranged from -2.368 to 1.840 MPa. The group-aware split conformal procedure achieved mean empirical coverage of 0.910 for the nominal 90% interval. Mean interval width was 45.512 MPa. The empirical coverage was therefore close to the target level on held-out validation groups.

4. Conclusion

This study evaluated machine-learning prediction of 28-day cylinder compressive strength using 1,456 concrete mixtures from 136 literature sources. Source-study identity and cross-study duplicate relationships were incorporated into the validation design so that linked studies remained together during train-test splitting. In the same-family ExtraTrees benchmark, random 10-fold validation produced MAE = 4.783 MPa and R^2 = 0.844, whereas grouped validation produced MAE = 9.807 MPa and R^2 = 0.471. The change demonstrates that row-wise random splitting can substantially overstate performance when related literature-derived observations are distributed across training and testing. Among the evaluated model families, ExtraTrees achieved the lowest mean outer-fold MAE of 9.758 MPa and a study-macro MAE of 10.392 MPa. Water-to-binder ratio produced the largest model-specific permutation importance, followed by total SCM-to-binder ratio and SF-to-binder ratio. Applicability-domain screening separated a warning group with higher average error than the in-domain group. The group-aware prediction intervals achieved 91.0% empirical coverage for the nominal 90% level, with a mean interval width of 45.512 MPa. Overall, the results support source-disjoint validation as the appropriate basis for reporting performance when concrete-strength models are developed from multiple literature sources.

REFERENCES

- [1] W. Ben Chaabene, M. Flah, and M. L. Nehdi, "Machine learning prediction of mechanical properties of concrete: Critical review," *Construction and Building Materials*, vol. 260, Art. no. 119889, 2020.
- [2] I. Núñez, A. Marani, M. Flah, and M. L. Nehdi, "Estimating compressive strength of modern concrete mixtures using computational intelligence: A systematic review," *Construction and Building Materials*, vol. 310, Art. no. 125279, 2021.
- [3] Z. Li, J. Yoon, R. Zhang, F. Rajabipour, W. V. Srubar III, I. Dabo, and A. Radlińska, "Machine learning in concrete science: Applications, challenges, and best practices," *npj Computational Materials*, vol. 8, Art. no. 127, 2022.
- [4] W. Ahmad, A. Ahmad, K. A. Ostrowski, F. Aslam, P. Joyklad, and P. Zajdel, "Application of advanced machine learning approaches to predict the compressive strength of concrete containing supplementary cementitious materials," *Materials*, vol. 14, no. 19, Art. no. 5762, 2021.
- [5] F. Aslam and M. Z. Shahab, "Supplementary cementitious materials in blended cement concrete: Advancements in predicting compressive strength through machine learning," *Materials Today Communications*, vol. 38, Art. no. 107725, 2024.
- [6] A. A. B. Loureiro and R. Stefani, "Comparing the performance of machine learning models for predicting the compressive strength of concrete," *Discover Civil Engineering*, vol. 1, Art. no. 19, 2024.
- [7] M. K. Elshaarawy, M. M. Alsaadawi, and A. K. Hamed, "Machine learning and interactive GUI for concrete compressive strength prediction," *Scientific Reports*, vol. 14, Art. no. 16694, 2024.
- [8] X. Liu, S. Mei, X. Wang, and X. Li, "Estimation of compressive strength of concrete with manufactured sand and natural sand using interpretable artificial intelligence," *Case Studies in Construction Materials*, vol. 21, Art. no. e03840, 2024.
- [9] H. Shimakawa, A. Kumada, and M. Sato, "Prevention of leakage in machine learning prediction for polymer composite properties," *Journal of Chemical Information and Modeling*, vol. 64, no. 9, pp. 3621–3629, 2024.

- [10] S. Durdy, M. W. Gaultois, V. V. Gusev, D. Bollegala, and M. J. Rosseinsky, "Random projections and kernelised leave one cluster out cross validation: Universal baselines and evaluation tools for supervised machine learning of material properties," *Digital Discovery*, vol. 1, no. 6, pp. 763–778, 2022.
- [11] W. M. Mesfin and H. K. Kim, "Parametric study on global estimation models for compressive strength adopting various machine learning algorithms in concrete," *Engineering Applications of Artificial Intelligence*, vol. 136, Art. no. 108888, 2024.
- [12] H. Nguyen, T. Vu, T. P. Vo, and H.-T. Thai, "Efficient machine learning models for prediction of concrete strengths," *Construction and Building Materials*, vol. 266, Art. no. 120950, 2021. doi: 10.1016/j.conbuildmat.2020.120950.
- [13] T. Nguyen-Sy, J. Wakim, Q.-D. To, M.-N. Vu, T.-D. Nguyen, and T.-T. Nguyen, "Predicting the compressive strength of concrete from its compositions and age using the extreme gradient boosting method," *Construction and Building Materials*, vol. 260, Art. no. 119757, 2020. doi: 10.1016/j.conbuildmat.2020.119757.
- [14] M. R. Kaloop et al., "Compressive strength prediction of high-performance concrete using gradient tree boosting machine," *Construction and Building Materials*, vol. 264, Art. no. 120198, 2020. doi: 10.1016/j.conbuildmat.2020.120198.
- [15] P. G. Asteris, A. D. Skentou, A. Bardhan, P. Samui, and K. Pilakoutas, "Predicting concrete compressive strength using hybrid ensembling of surrogate machine learning models," *Cement and Concrete Research*, vol. 145, Art. no. 106449, 2021. doi: 10.1016/j.cemconres.2021.106449.
- [16] Y. Xu, W. Ahmad, A. Ahmad, K. A. Ostrowski, M. Dudek, F. Aslam, and P. Joyklad, "Computation of high-performance concrete compressive strength using standalone and ensembled machine learning techniques," *Materials*, vol. 14, no. 22, Art. no. 7034, 2021. doi: 10.3390/ma14227034.
- [17] H. S. Ullah, R. A. Khushnood, F. Farooq, J. Ahmad, N. I. Vatin, and D. Y. Z. Ewais, "Prediction of compressive strength of sustainable foam concrete using individual and ensemble machine learning approaches," *Materials*, vol. 15, no. 9, Art. no. 3166, 2022. doi: 10.3390/ma15093166.
- [18] P. Huang, K. Dai, and X. Yu, "Machine learning approach for investigating compressive strength of self-compacting concrete containing supplementary cementitious materials and recycled aggregate," *Journal of Building Engineering*, vol. 79, Art. no. 107904, 2023. doi: 10.1016/j.jobe.2023.107904.
- [19] B. Sun, W. Cui, G. Liu, B. Zhou, and W. Zhao, "A hybrid strategy of AutoML and SHAP for automated and explainable concrete strength prediction," *Case Studies in Construction Materials*, vol. 19, Art. no. e02405, 2023. doi: 10.1016/j.cscm.2023.e02405.
- [20] G. Liu and B. Sun, "Concrete compressive strength prediction using an explainable boosting machine model," *Case Studies in Construction Materials*, vol. 18, Art. no. e01845, 2023. doi: 10.1016/j.cscm.2023.e01845.
- [21] W. Wang, Y. Zhong, G. Liao, Q. Ding, T. Zhang, and X. Li, "Prediction of compressive strength of concrete specimens based on interpretable machine learning," *Materials*, vol. 17, no. 15, Art. no. 3661, 2024. doi: 10.3390/ma17153661.
- [22] Y. Zhang, W. Ren, Y. Chen, Y. Mi, J. Lei, and L. Sun, "Predicting the compressive strength of high-performance concrete using an interpretable machine learning model," *Scientific Reports*, vol. 14, Art. no. 28346, 2024. doi: 10.1038/s41598-024-79502-z.
- [23] W. Zhang, J. Guo, C. Ning, R. Cheng, and Z. Liu, "Prediction of concrete compressive strength using a Deepforest-based model," *Scientific Reports*, vol. 14, Art. no. 18918, 2024. doi: 10.1038/s41598-024-69616-9.
- [24] S. Kumar, R. Kumar, B. Rai, and P. Samui, "Prediction of compressive strength of high-volume fly ash self-compacting concrete with silica fume using machine learning techniques," *Construction and Building Materials*, vol. 438, Art. no. 136933, 2024. doi: 10.1016/j.conbuildmat.2024.136933.
- [25] M. A. Hariri-Ardebili, P. Mahdavi, and F. Pourkamali-Anaraki, "Benchmarking AutoML solutions for concrete strength prediction: Reliability, uncertainty, and dilemma," *Construction and Building Materials*, vol. 423, Art. no. 135782, 2024. doi: 10.1016/j.conbuildmat.2024.135782.
- [26] P. Tamuly and V. Nava, "Machine learning based conformal predictors for uncertainty-aware compressive strength estimation of concrete," *Construction and Building Materials*, vol. 487, Art. no. 141844, 2025. doi: 10.1016/j.conbuildmat.2025.141844.
- [27] A. Thoriya and H. Rangwala, "Calibrated probabilistic prediction of compressive strength in palm oil fuel ash-modified concrete using conformal learning," *Scientific Reports*, 2026. doi: 10.1038/s41598-026-60994-w.
- [28] A. Dunn, Q. Wang, A. Ganose, D. Dopp, and A. Jain, "Benchmarking materials property prediction methods: The Matbench test set and Automatminer reference algorithm," *npj Computational Materials*, vol. 6, Art. no. 138, 2020. doi: 10.1038/s41524-020-00406-3.
- [29] H. Kaneko, "Evaluation and optimization methods for applicability domain methods and their hyperparameters, considering the prediction performance of machine learning models," *ACS Omega*, vol. 9, no. 10, pp. 11453–11458, 2024. doi: 10.1021/acsomega.3c08036.
- [30] K. Boonpalit, J. Kinchagawat, and S. Namuangruk, "Expanding the applicability domain of machine learning model for advancements in electrochemical material discovery," *ChemElectroChem*, vol. 11, no. 10, Art. no. e202300681, 2024. doi: 10.1002/celec.202300681.
- [31] L. E. Schultz, Y. Wang, R. Jacobs, and D. Morgan, "A general approach for determining applicability domain of machine learning models," *npj Computational Materials*, vol. 11, Art. no. 95, 2025. doi: 10.1038/s41524-025-01573-x.