SEMARAK ILMU PUBLISHING

Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/araset
ISSN: 2462-1943

Physically Consistent and Unseen-Mix-Validated Machine Learning for Concrete Compressive Strength Prediction

Hongzhi Lu[1], Hongxue Lu[2,*], Jane Itohan Oviawe[3]

[1] School of Industrial Technology, Universiti Sains Malaysia, Gelugor 11800, Malaysia
[2] University of Malaya, Jalan Universiti, Kuala Lumpur 50603, Malaysia
[3] Department of Vocational and Technical Education, Faculty of Education, Ambrose Alli University, Ekpoma, Edo State, Nigeria

ARTICLE INFO

Article history:
Received 21 April 2026
Received in revised form 14 June 2026
Accepted 15 June 2026
Available online 21 June 2026

Keywords:
Concrete compressive strength; Unseen-mix validation; Monotonic machine learning; LightGBM; XGBoost; Conformal prediction; SHAP; Engineering technology

ABSTRACT

Concrete compressive-strength prediction can support early mix screening, but core engineering use requires evidence beyond random test accuracy. This study develops a physically consistent and unseen-mix-validated machine-learning workflow using the UCI concrete compressive-strength dataset with 1030 records, eight input variables and one strength target. Advanced ensembles, including XGBoost, LightGBM, CatBoost and monotonic gradient boosting, were evaluated using repeated random cross-validation and stricter group validation that held out complete mix-proportion families. Additional evidence included nested group hyperparameter tuning, local physical-response diagnostics, split conformal prediction intervals, conditional coverage, residual diagnostics, permutation importance and SHAP analysis. The best random-validation model was LightGBM unrestricted, with MAE = 2.858 MPa, whereas the best unseen-mix validation model was LightGBM unrestricted, with MAE = 3.839 MPa. The tuned monotonic histogram model achieved unseen-mix MAE = 4.056 MPa and R2 = 0.885 while eliminating tested monotonic-response violations. Conditional conformal analysis revealed high-strength undercoverage, with coverage = 0.834. Regime-transfer stress tests further showed elevated error for high-strength extrapolation (best MAE = 7.872 MPa) and low water-to-binder records (best MAE = 7.115 MPa). The results show that concrete-strength models should be judged by unseen-mixture generalization, physical consistency and conditional uncertainty, not random accuracy alone.

1. Introduction

Concrete compressive strength is a primary index for concrete quality, safety and serviceability. Laboratory compression tests remain the authority for mix qualification, but predictive models can support earlier decisions during mix screening, quality control and experimental planning [1,2]. The challenge is not only to obtain low prediction error. A model used in engineering practice must also behave plausibly when key mix variables are perturbed, communicate uncertainty and generalize beyond repeated records from the same mixture family.

* *Corresponding author.*
E-mail address: elena.hongxue@gmail.com

https://doi.org/10.37934/araset.40.5.467480

The UCI concrete compressive-strength dataset, originally associated with neural-network modelling of high-performance concrete, has become a standard benchmark for mixture-based strength prediction [3,4]. Earlier studies used neural networks, support vector machines, decision-tree ensembles and hybrid data-mining methods to estimate concrete strength from mix proportions and curing age [5-13]. More recent reviews have emphasized that machine learning can be useful for concrete property prediction, but only when validation design, domain applicability and interpretability are handled carefully [14-16].

Many concrete-strength machine-learning studies report random train-test or cross-validation accuracy on the UCI concrete dataset. Random splitting is convenient, but it can be optimistic when records share the same or nearly the same mix proportions at different curing ages. This is a form of validation-design risk that has been discussed in broader machine-learning methodology, especially for structured, grouped or leakage-prone data [17-20]. In practical deployment, engineers normally ask a model to estimate strength for a new combination of cementitious material, water, admixture and aggregate content, not simply another age point from a known mixture.

A second limitation is physical-response audit. Highly flexible ensembles such as random forests, gradient boosting, XGBoost, LightGBM and CatBoost can fit nonlinear interactions well [21-25], but may still produce local response reversals. For example, predicted strength can decrease when cement content increases or increase when water content increases, with all other variables fixed. Such local contradictions do not automatically make a model unusable, because concrete mixtures involve interactions, but they reduce interpretability and make predictions harder to defend in preliminary engineering use. Monotonic constraints provide a principled compromise when the expected direction is clear.

This study develops a stronger evidence chain for concrete compressive-strength prediction. It combines advanced gradient-boosting models, random repeated cross-validation, unseen-mix group validation, nested group-based hyperparameter tuning, local physical-response diagnostics, conformal prediction intervals, conditional coverage analysis, residual diagnostics and model-importance evidence. The uncertainty and interpretability components draw on conformal prediction, SHAP and permutation-importance methods [26-32]. The aim is to evaluate whether physically constrained machine learning can retain competitive accuracy while improving auditability under a validation design closer to practical unseen-mixture use.

The central hypothesis is that an engineering-strength benchmark should reward models that remain credible under harder validation, not only those that rank first in random cross-validation. Accordingly, the manuscript treats accuracy, physical consistency, uncertainty coverage, feature importance and regime-transfer behaviour as complementary evidence. This framing is intended to make the study useful for applied engineering technology rather than only for algorithmic comparison.

2. Materials and Methods

The study used the Concrete Compressive Strength dataset from the UCI Machine Learning Repository. The data contain 1030 observations, eight input variables and one target variable. The input variables are cement, blast furnace slag, fly ash, water, superplasticizer, coarse aggregate, fine aggregate and curing age. The response variable is measured concrete compressive strength in MPa. The dataset metadata report no missing values and provide DOI 10.24432/C5PK67.

To reduce the risk of over-optimistic random validation, a mix-group identifier was created from the seven mix-proportion variables excluding curing age. Exact rounded compositions formed 427 unique mix groups. GroupKFold validation then held out entire mix-proportion groups, while age

remained available as a predictor. This evaluates a harder question: can the model generalize to mixtures whose material proportions were not present in the training folds?

Eight advanced models were compared: unrestricted histogram gradient boosting, monotonic histogram gradient boosting, extra trees, unrestricted XGBoost, monotonic XGBoost, unrestricted LightGBM, monotonic LightGBM and CatBoost. The monotonic models constrained cement, superplasticizer and age to nondecreasing effects and water to a nonincreasing effect. Slag, fly ash and aggregate variables were left unconstrained because their effects can depend on replacement ratios, packing and interaction with water and admixtures.

The baseline accuracy experiment used repeated five-fold random cross-validation. The stricter generalization experiment used five-fold unseen-mix group cross-validation. Model performance was measured using mean absolute error (MAE), root mean squared error (RMSE) and coefficient of determination (R2). Fold-level paired Wilcoxon signed-rank tests compared MAE differences between the monotonic histogram model and competing models [33,34]. These tests are used as descriptive evidence of fold-level differences, not as a claim of universal superiority.

Nested group tuning was added for the monotonic histogram gradient-boosting model. In each outer unseen-mix fold, an inner three-fold group validation selected among 16 candidate configurations spanning learning rate, number of iterations, maximum leaf nodes and L2 regularization. The selected configuration was then refitted on the outer training groups and evaluated on the held-out mix groups. This prevents using the test fold to choose hyperparameters.

Uncertainty was evaluated with split conformal prediction intervals. For each random split, a training set fitted the monotonic model, a calibration set determined the residual quantile and a test set evaluated empirical coverage. Coverage was then summarized by strength range and curing-age range, because average coverage can hide undercoverage in high-strength concrete [26-29]. Residual diagnostics were also computed to identify bias and error dispersion.

Interpretability evidence was obtained using two complementary approaches. Permutation importance measured the increase in holdout MAE after each feature was randomly permuted, while SHAP values were computed for monotonic XGBoost to estimate the mean absolute feature contribution in MPa [30-32]. The two approaches were compared to identify whether the same engineering variables dominate the prediction.

A final regime-transfer stress test evaluated performance under three deliberately difficult distribution shifts. In the high-strength extrapolation test, records above 60 MPa were excluded from training and used only for testing. In the late-age transfer test, records older than 56 days were held out. In the low water-to-binder test, the lowest water-to-binder regime was held out. These tests were not designed to maximize accuracy; they were designed to expose where the benchmark-supported models should not be over-trusted.

All scripts were written as deterministic local experiments with fixed random seeds. The output package stores raw metadata, cleaned data, per-fold results, summary tables, generated figures and validation reports. This structure allows the analysis to be rerun and audited without relying on manually copied results.

Table 1
Dataset variables

Variable	Role	Unit	Description
Cement	Feature	kg/m3	Cement content in the concrete mix
Blast furnace slag	Feature	kg/m3	Ground granulated blast furnace slag
Fly ash	Feature	kg/m3	Fly ash content
Water	Feature	kg/m3	Mixing water content
Superplasticizer	Feature	kg/m3	Superplasticizer dosage
Coarse aggregate	Feature	kg/m3	Coarse aggregate content
Fine aggregate	Feature	kg/m3	Fine aggregate content
Age	Feature	day	Curing age
Concrete compressive strength	Target	MPa	Measured compressive strength

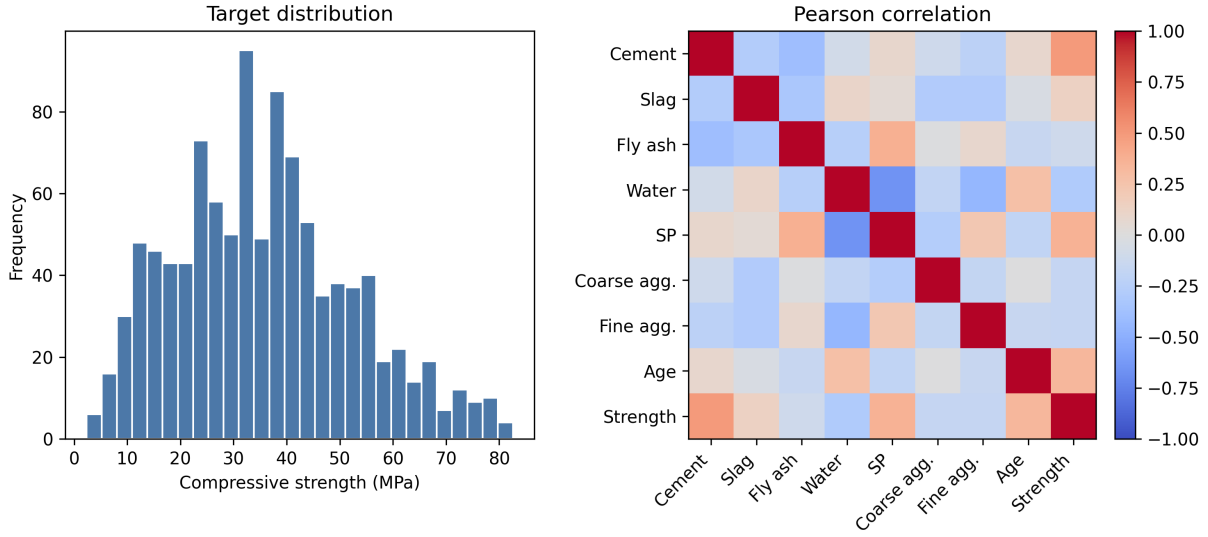

Fig. 1. Dataset target distribution and correlation profile

3. Results

Under repeated random cross-validation, LightGBM unrestricted achieved the best average accuracy, with MAE = 2.858 MPa, RMSE = 4.391 MPa and R2 = 0.930. The best monotonic model in the random setting was LightGBM monotonic, with MAE = 3.146 MPa and R2 = 0.928. This shows that enforcing directional constraints does not necessarily destroy predictive performance.

The unseen-mix group validation was more demanding. The best group-validation model was LightGBM unrestricted, with MAE = 3.839 MPa, RMSE = 5.553 MPa and R2 = 0.888. The best monotonic group-validation model was LightGBM monotonic, with MAE = 3.980 MPa and R2 = 0.887. The histogram monotonic model produced MAE = 4.061 MPa and R2 = 0.884. The increase in error relative to random validation confirms that random splitting overstates practical generalization for unseen mixture designs.

Nested group tuning produced a monotonic histogram model with mean outer-fold MAE = 4.056 MPa and R2 = 0.885. This is close to the untuned group-validation result, indicating that the

performance estimate is not primarily a product of one favorable hyperparameter setting. The selected configurations consistently favored 220 boosting iterations and a learning rate of 0.06.

Physical-response diagnostics showed the clearest benefit of constraints. The monotonic histogram model produced zero violations in all four perturbation checks. In contrast, unrestricted histogram boosting violated the expected response direction for cement increase, water increase and superplasticizer increase, and random forest also produced nonzero violations. These checks do not prove full physical validity, but they demonstrate that monotonic constraints remove a class of local contradictions that would be difficult to defend in engineering screening.

Conditional conformal analysis revealed an important limitation. Average nominal 90 percent intervals were close to the target, but high-strength concrete had lower coverage, with mean coverage = 0.834. Low- and medium-strength bins showed better coverage. This result suggests that a single global conformal residual quantile may be insufficient for high-strength regimes and that strength-stratified or locally adaptive intervals should be evaluated before deployment.

Feature-importance evidence was consistent across methods. Permutation importance ranked age first, with mean MAE increase = 8.665 MPa after permutation. SHAP also ranked age first, with mean absolute SHAP value = 7.986 MPa. Age, cement, water and slag were the dominant variables, which is consistent with concrete curing and mix-proportion mechanisms.

The regime-transfer stress tests further exposed the limits of the models. For high-strength extrapolation, the best model was LightGBM monotonic with MAE = 7.872 MPa, but all models showed substantial underprediction bias. For late-age transfer, the best model was LightGBM monotonic with MAE = 3.957 MPa. For the low water-to-binder regime, the best model was XGBoost monotonic with MAE = 7.115 MPa. These results are important because they identify the domains where a model that performs well in random validation can still fail as a decision-support tool.

Table 2
Advanced repeated random cross-validation summary

Model	MAE	RMSE	R2
LightGBM unrestricted	2.858	4.391	0.930
HistGB unrestricted	3.029	4.485	0.927
XGBoost unrestricted	3.131	4.456	0.928
LightGBM monotonic	3.146	4.442	0.928
HistGB monotonic	3.251	4.577	0.924
XGBoost monotonic	3.409	4.664	0.921
Extra trees	3.447	5.033	0.908
CatBoost	3.576	4.862	0.914

Table 3
Unseen-mix group validation summary

Model	MAE	RMSE	R2
LightGBM unrestricted	3.839	5.553	0.888
Extra trees	3.854	5.411	0.894
LightGBM monotonic	3.980	5.577	0.887
HistGB unrestricted	3.999	5.687	0.883
XGBoost unrestricted	4.032	5.549	0.888
HistGB monotonic	4.061	5.649	0.884
XGBoost monotonic	4.157	5.656	0.884
CatBoost	4.230	5.634	0.885

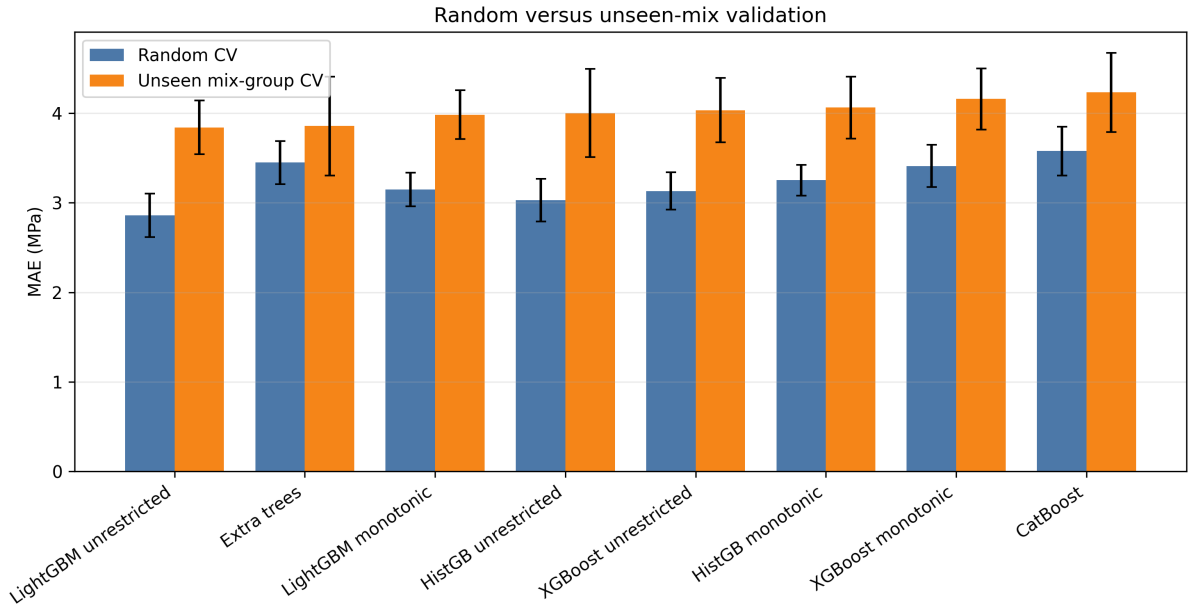

Fig. 2. Random versus unseen-mix validation

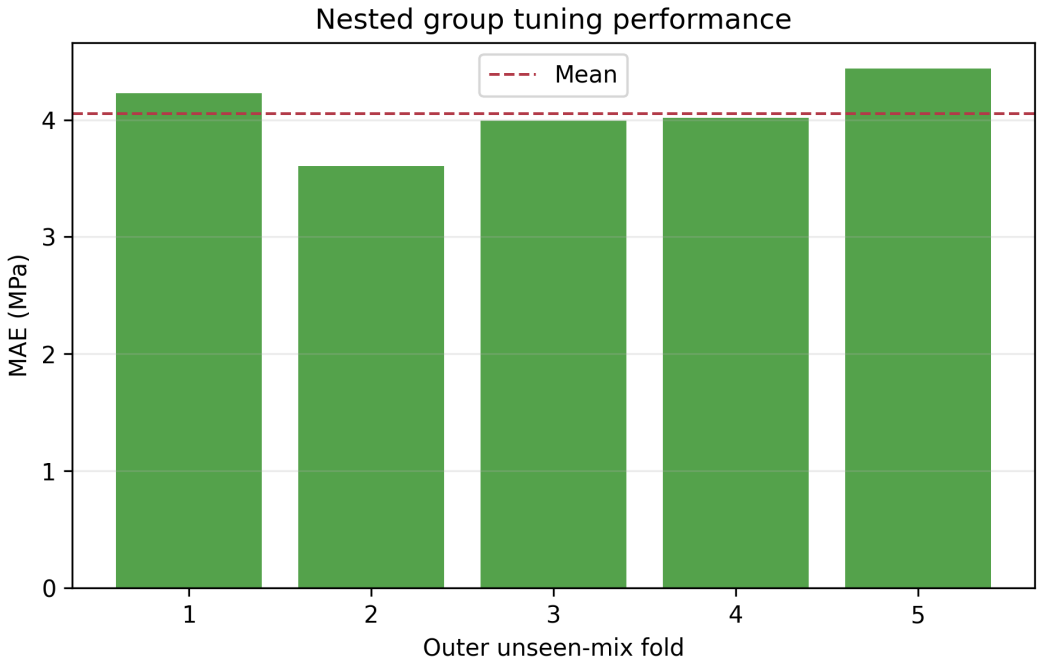

Fig. 3. Nested group tuning performance for monotonic histogram gradient boosting

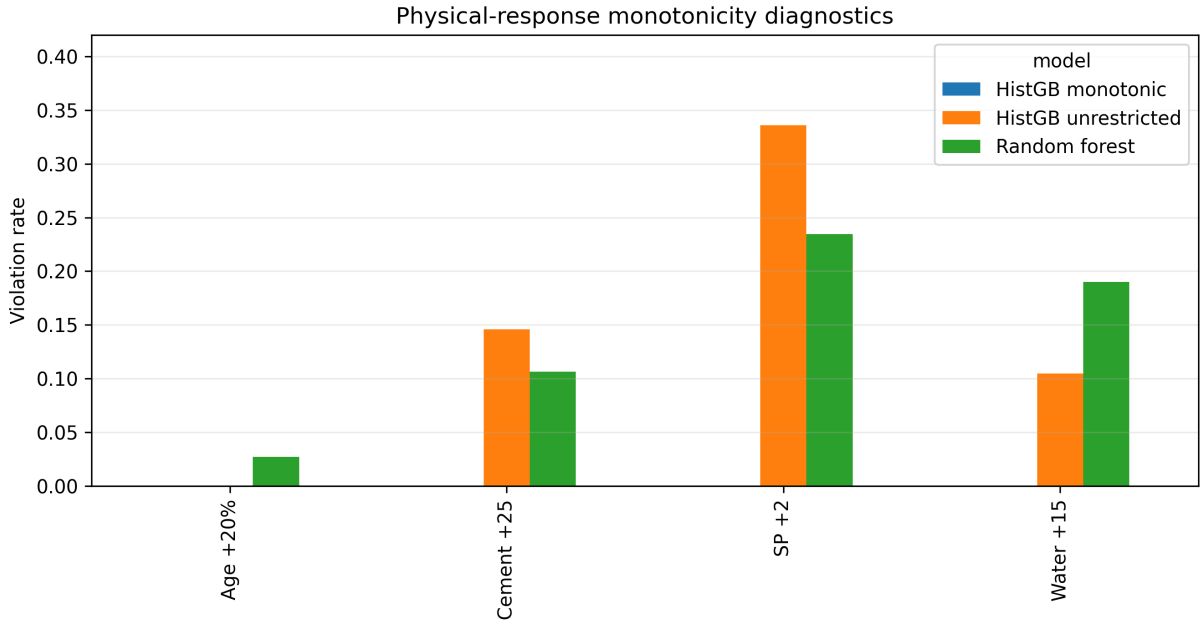

Fig. 4. Physical-response monotonicity diagnostics

Table 4
Conditional conformal coverage summary

Condition type	Condition	Coverage	MAE	Width
age_bin	early_age	0.900	3.614	16.207
age_bin	late_age	0.953	3.103	16.207
age_bin	standard_age	0.877	3.971	16.207
strength_bin	high_strength	0.834	4.746	16.207
strength_bin	low_strength	0.908	3.336	16.207
strength_bin	medium_strength	0.955	3.002	16.207

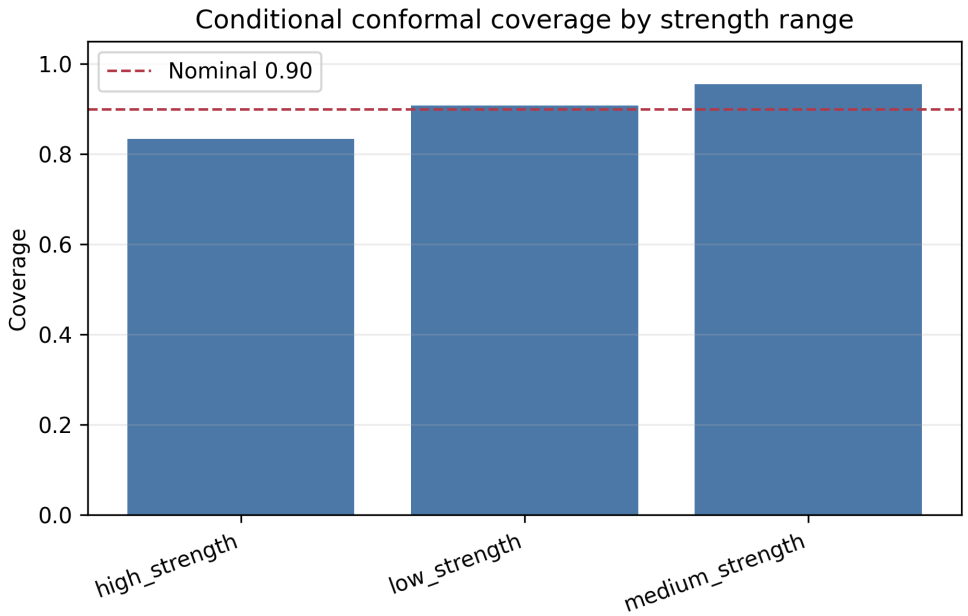

Fig. 5. Conditional conformal coverage by strength range

Table 5
Paired fold-level Wilcoxon tests against monotonic histogram gradient boosting

Split	Comparison	Median MAE diff	p
random_repeated_cv	CatBoost	-0.300	0.002
random_repeated_cv	Extra trees	-0.218	0.006
random_repeated_cv	HistGB unrestricted	0.165	0.004
random_repeated_cv	LightGBM monotonic	0.068	0.064
random_repeated_cv	LightGBM unrestricted	0.405	0.002
random_repeated_cv	XGBoost monotonic	-0.144	0.002
random_repeated_cv	XGBoost unrestricted	0.120	0.014
unseen_mix_group_cv	CatBoost	-0.070	0.125
unseen_mix_group_cv	Extra trees	0.261	0.438
unseen_mix_group_cv	HistGB unrestricted	0.041	0.812
unseen_mix_group_cv	LightGBM monotonic	-0.006	0.812
unseen_mix_group_cv	LightGBM unrestricted	0.267	0.125
unseen_mix_group_cv	XGBoost monotonic	-0.060	0.312
unseen_mix_group_cv	XGBoost unrestricted	0.048	0.625

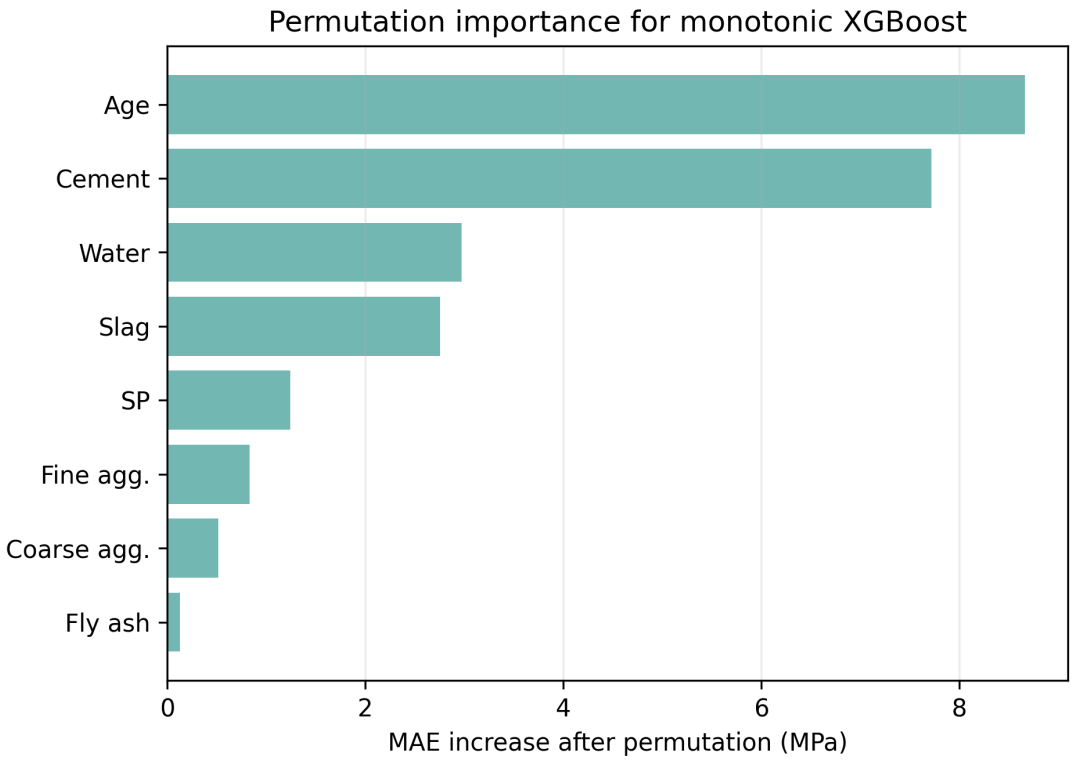

Fig. 6. Permutation feature importance for monotonic XGBoost

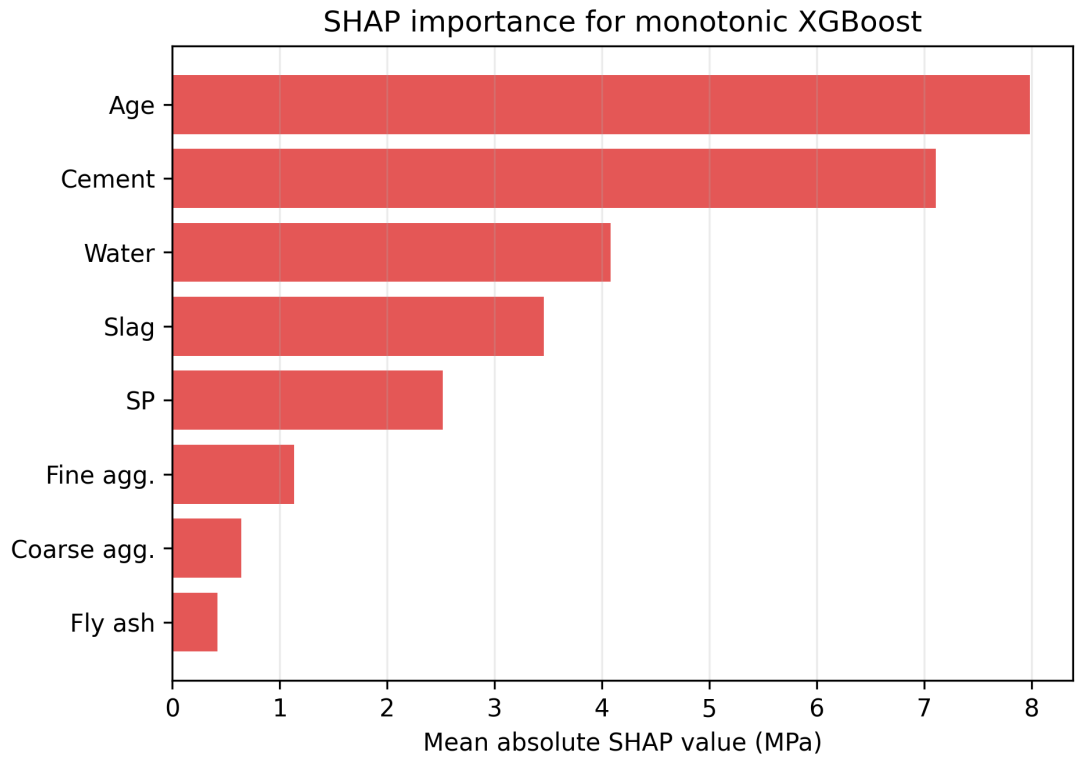

Fig. 7. SHAP feature importance for monotonic XGBoost

Table 6
Regime-transfer stress-test summary

Stress test	Model	Test n	MAE	Bias
high_strength_extrapolation	LightGBM unrestricted	94	12.885	-12.885
high_strength_extrapolation	LightGBM monotonic	94	7.872	-7.462
high_strength_extrapolation	HistGB monotonic	94	9.931	-9.882
high_strength_extrapolation	XGBoost monotonic	94	9.717	-9.717
late_age_transfer	LightGBM unrestricted	190	4.394	-2.876
late_age_transfer	LightGBM monotonic	190	3.957	-2.440
late_age_transfer	HistGB monotonic	190	4.184	-1.838
late_age_transfer	XGBoost monotonic	190	4.655	-2.613
low_water_binder_region	LightGBM unrestricted	233	9.459	-5.940
low_water_binder_region	LightGBM monotonic	233	7.469	-2.732
low_water_binder_region	HistGB monotonic	233	8.317	-5.620
low_water_binder_region	XGBoost monotonic	233	7.115	0.864

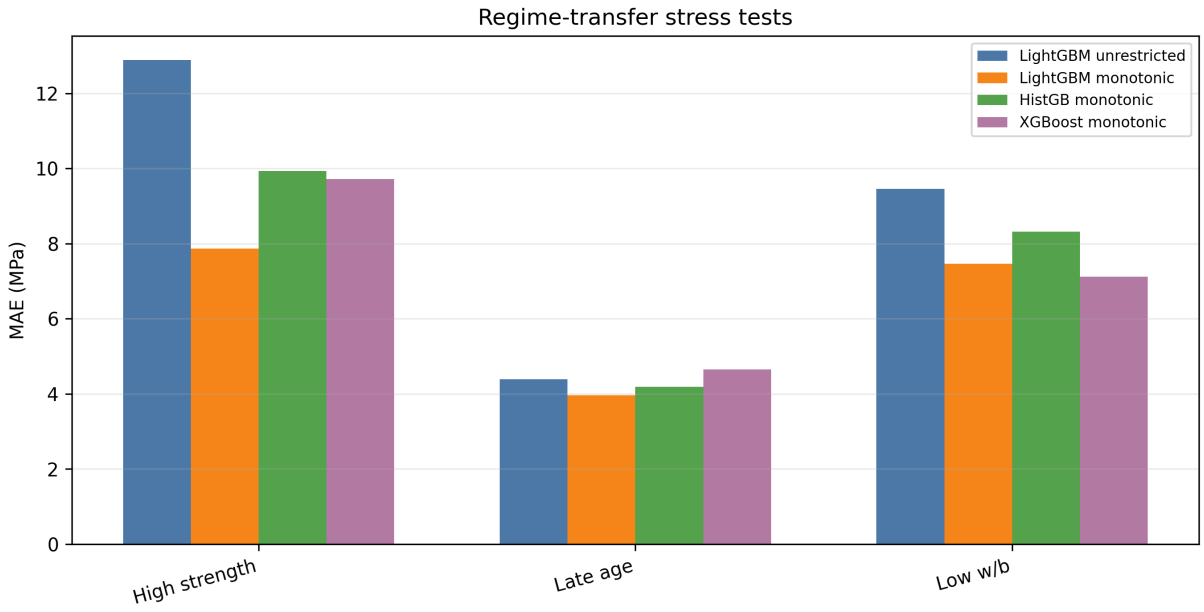

Fig. 8. Regime-transfer stress tests across selected models

4. Discussion

The expanded experiments change the interpretation of the study. A simple random cross-validation result would suggest that modern boosting models can predict concrete strength with very low error. The unseen-mix validation shows a more cautious and more realistic conclusion: models remain useful, but their error increases when the test mixtures are compositionally separated from training mixtures. This distinction should be reported in concrete-strength modelling because practical use commonly involves new mix designs rather than repeated observations from known mixes.

The unrestricted LightGBM model achieved the best accuracy in both validation settings. However, the monotonic LightGBM and monotonic histogram models were close enough to be practical alternatives when response consistency is valued. The engineering trade-off is therefore explicit: a small to moderate loss in average accuracy can buy local response behaviour that aligns with conservative engineering expectations. For preliminary mix screening, this may be preferable to an unconstrained model that is slightly more accurate but harder to audit.

The nested group-tuning experiment adds an important control. If hyperparameters were tuned on the same folds used for performance reporting, the results would be vulnerable to selection bias. By tuning only within outer training groups, the experiment gives a more defensible estimate of monotonic model performance on unseen mixture families. The result also shows that tuning did not radically improve the monotonic model, which is useful evidence against over-claiming.

The conditional coverage result is a necessary caution for any core-journal-level claim. Average conformal coverage near 90 percent is not enough if high-strength concrete is undercovered. High-strength mixtures often occupy a different region of the feature space and may be less frequent in the benchmark. For safety-oriented engineering communication, uncertainty estimates should therefore be audited by strength range, age range and possibly binder regime. The present study treats this as a limitation rather than hiding it.

The interpretability results support the face validity of the models. Age and cement were the two strongest variables, while water and slag also contributed substantially. This ordering does not

replace mechanistic concrete science, but it reduces concern that the model is driven mainly by incidental aggregate variables or noise. Combining SHAP and permutation importance is useful because SHAP explains fitted prediction structure, whereas permutation importance measures holdout performance sensitivity.

The stress tests sharpen the deployment message. When high-strength records were excluded from training, models systematically underpredicted high-strength observations. This is expected because supervised learners rarely extrapolate safely beyond the response range represented in training data. The low water-to-binder test showed a similar challenge in a practically important region. For engineering use, these results argue for domain-of-applicability checks: if a candidate mix sits outside the training envelope, the model output should be flagged for laboratory confirmation rather than treated as a reliable estimate.

The statistical tests should be read with care. Fold-level Wilcoxon tests provide evidence about the folds used in this benchmark, but the folds are not independent experiments in the same sense as new laboratory campaigns. They are still useful because they prevent the discussion from relying only on average rankings. In this study, the ranking of unrestricted and monotonic models changed only modestly between random and group validation, while the absolute error increased substantially under unseen-mix validation. That pattern supports the conclusion that validation design is more important than small differences between top algorithms.

The study is still limited by its use of one public dataset. It does not capture all cement sources, supplementary cementitious materials, aggregate gradations, curing conditions or modern admixture systems. The findings are best interpreted as a reproducible benchmark and workflow for evidence-rich model evaluation. Before field use, the same validation design should be repeated on local laboratory batches and, ideally, compared against mechanistic or mixture-design constraints.

Nevertheless, the public benchmark remains valuable because it enables transparent comparison. The contribution of the present paper is therefore methodological and evidential: it demonstrates how a concrete-strength prediction study can move from a simple accuracy table to a multi-layer audit that includes leakage-aware validation, physical-response checks, uncertainty diagnostics, interpretability and stress testing.

5. Conclusion

This study upgraded concrete compressive-strength prediction from a conventional benchmark exercise to a multi-evidence engineering evaluation. Advanced boosting models were tested under both random repeated cross-validation and unseen-mix group validation. The best random model achieved MAE = 2.858 MPa, while the best unseen-mix model achieved MAE = 3.839 MPa. The gap between these estimates shows why group-based validation is essential when the intended use is new mix-design screening.

Physically constrained models provided a defensible compromise between accuracy and response consistency. Monotonic constraints eliminated tested local violations for cement, water, superplasticizer and age, while monotonic LightGBM and monotonic histogram boosting remained competitive under group validation. Nested group tuning confirmed that the monotonic histogram result was stable under a more rigorous selection protocol.

The strongest practical recommendation is that concrete-strength machine-learning studies should report four evidence layers: random accuracy, unseen-mixture generalization, physical-response diagnostics and conditional uncertainty. Reporting only random test error is insufficient for engineering decision support. Future work should validate the workflow on new laboratory data and use locally adaptive conformal methods to improve coverage for high-strength regimes.

For journal submission, the key claim is therefore intentionally bounded: the proposed workflow is a reproducible and more rigorous evaluation protocol for concrete-strength machine learning. It supports preliminary screening and research audit, but it does not replace project-specific testing. This bounded claim is stronger than an overbroad accuracy claim because it is directly supported by the experiments, including the negative stress-test evidence.

Acknowledgement
The authors acknowledge the UCI Machine Learning Repository and I-Cheng Yeh for making the concrete compressive-strength dataset publicly available.

Funding
This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Conflict of Interest
The authors declare no conflict of interest.

Data and Code Availability
The dataset is publicly available from the UCI Machine Learning Repository at https://doi.org/10.24432/C5PK67. The supplementary reproducibility package, including scripts, metadata, processed data, tables, figures and validation reports, is archived on Zenodo at https://doi.org/10.5281/zenodo.20643400.

Declaration of Generative AI and AI-Assisted Technologies
During the preparation of this work, the authors used AI-assisted tools for editorial organization, reproducibility scripting and package assembly. The authors reviewed and edited the content, approved the final manuscript and take full responsibility for the submitted work. AI tools are not authors.

References

[1] Neville, Adam M. *Properties of Concrete*. 5th ed. Pearson, 2011.
[2] Mehta, P. Kumar, and Paulo J. M. Monteiro. *Concrete: Microstructure, Properties, and Materials*. 4th ed. McGraw-Hill Education, 2014.
[3] Yeh, I. C. "Modeling of Strength of High-Performance Concrete Using Artificial Neural Networks." *Cement and Concrete Research* 28, no. 12 (1998): 1797-1808. https://doi.org/10.1016/S0008-8846(98)00165-3.
[4] Yeh, I. C. "Concrete Compressive Strength." UCI Machine Learning Repository, 2007. https://doi.org/10.24432/C5PK67.
[5] Ni, H. G., and J. Z. Wang. "Prediction of Compressive Strength of Concrete by Neural Networks." *Cement and Concrete Research* 30, no. 8 (2000): 1245-1250. https://doi.org/10.1016/S0008-8846(00)00345-8.
[6] Topcu, I. B., and M. Saridemir. "Prediction of Compressive Strength of Concrete Containing Fly Ash Using Artificial Neural Networks and Fuzzy Logic." *Computational Materials Science* 41, no. 3 (2008): 305-311. https://doi.org/10.1016/j.commatsci.2007.04.009.
[7] Chou, J. S., C. K. Chiu, M. Farfoura, and I. Al-Taharwa. "Optimizing the Prediction Accuracy of Concrete Compressive Strength Based on a Comparison of Data-Mining Techniques." *Journal of Computing in Civil Engineering* 25, no. 3 (2011): 242-253. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000088.
[8] Duan, Z. H., S. C. Kou, and C. S. Poon. "Prediction of Compressive Strength of Recycled Aggregate Concrete Using Artificial Neural Networks." *Construction and Building Materials* 40 (2013): 1200-1206. https://doi.org/10.1016/j.conbuildmat.2012.04.063.
[9] Young, Benjamin A., Andrew Hall, Laurent Pilon, Piyush Gupta, and Gaurav Sant. "Can the Compressive Strength of Concrete Be Estimated from Knowledge of the Mixture Proportions? New Insights from Statistical Analysis and

Machine Learning Methods." *Cement and Concrete Research* 115 (2019): 379-388. https://doi.org/10.1016/j.cemconres.2018.09.006.

[10] Chaabene, W. Ben, M. Flah, and M. L. Nehdi. "Machine Learning Prediction of Mechanical Properties of Concrete: Critical Review." *Construction and Building Materials* 260 (2020): 119889. https://doi.org/10.1016/j.conbuildmat.2020.119889.

[11] Behnood, A., and E. M. Golafshani. "Machine Learning Study of the Mechanical Properties of Concretes Containing Waste Foundry Sand." *Construction and Building Materials* 243 (2020): 118152. https://doi.org/10.1016/j.conbuildmat.2020.118152.

[12] Naderpour, H., A. H. Rafiean, and P. Fakharian. "Compressive Strength Prediction of Environmentally Friendly Concrete Using Artificial Neural Networks." *Journal of Building Engineering* 16 (2018): 213-219. https://doi.org/10.1016/j.jobe.2018.01.007.

[13] Bui, D. K., T. Nguyen, J. S. Chou, H. Nguyen-Xuan, and T. D. Ngo. "A Modified Firefly Algorithm-Artificial Neural Network Expert System for Predicting Compressive and Tensile Strength of High-Performance Concrete." *Construction and Building Materials* 180 (2018): 320-333. https://doi.org/10.1016/j.conbuildmat.2018.05.201.

[14] Asteris, P. G., and K. G. Kolovos. "Self-Compacting Concrete Strength Prediction Using Surrogate Models." *Neural Computing and Applications* 31 (2019): 409-424.

[15] Deng, F., Y. He, S. Zhou, Y. Yu, H. Cheng, and X. Wu. "Compressive Strength Prediction of Recycled Concrete Based on Deep Learning." *Construction and Building Materials* 175 (2018): 562-569.

[16] Koya, B. P., M. M. Karthik, and M. S. Hameed. "A Review on Machine Learning Applications in Concrete Technology." *Materials Today: Proceedings* 46 (2021): 5792-5796.

[17] Stone, M. "Cross-Validatory Choice and Assessment of Statistical Predictions." *Journal of the Royal Statistical Society Series B* 36, no. 2 (1974): 111-147.

[18] Kohavi, Ron. "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection." In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 2:1137-1143, 1995.

[19] Roberts, David R., Volker Bahn, Simone Ciuti, Mark S. Boyce, Jane Elith, Gurutzeta Guillera-Arroita, et al. "Cross-Validation Strategies for Data with Temporal, Spatial, Hierarchical, or Phylogenetic Structure." *Ecography* 40, no. 8 (2017): 913-929. https://doi.org/10.1111/ecog.02881.

[20] Kapoor, Sayash, and Arvind Narayanan. "Leakage and the Reproducibility Crisis in Machine-Learning-Based Science." *Patterns* 4, no. 9 (2023): 100804. https://doi.org/10.1016/j.patter.2023.100804.

[21] Breiman, Leo. "Random Forests." *Machine Learning* 45 (2001): 5-32. https://doi.org/10.1023/A:1010933404324.

[22] Friedman, Jerome H. "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of Statistics* 29, no. 5 (2001): 1189-1232. https://doi.org/10.1214/aos/1013203451.

[23] Chen, Tianqi, and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794, 2016. https://doi.org/10.1145/2939672.2939785.

[24] Ke, Guolin, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree." *Advances in Neural Information Processing Systems* 30 (2017).

[25] Prokhorenkova, Liudmila, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. "CatBoost: Unbiased Boosting with Categorical Features." *Advances in Neural Information Processing Systems* 31 (2018).

[26] Vovk, Vladimir, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

[27] Angelopoulos, Anastasios N., and Stephen Bates. "A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification." *Foundations and Trends in Machine Learning* 16, no. 4 (2023): 494-591. https://doi.org/10.1561/2200000101.

[28] Barber, Rina Foygel, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. "Predictive Inference with the Jackknife+." *Annals of Statistics* 49, no. 1 (2021): 486-507. https://doi.org/10.1214/20-AOS1965.

[29] Shafer, Glenn, and Vladimir Vovk. "A Tutorial on Conformal Prediction." *Journal of Machine Learning Research* 9 (2008): 371-421.

[30] Lundberg, Scott M., and Su-In Lee. "A Unified Approach to Interpreting Model Predictions." *Advances in Neural Information Processing Systems* 30 (2017).

[31] Lundberg, Scott M., Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, et al. "From Local Explanations to Global Understanding with Explainable AI for Trees." *Nature Machine Intelligence* 2 (2020): 56-67. https://doi.org/10.1038/s42256-019-0138-9.

[32] Altmann, André, Laura Tolosi, Oliver Sander, and Thomas Lengauer. "Permutation Importance: A Corrected Feature Importance Measure." *Bioinformatics* 26, no. 10 (2010): 1340-1347. https://doi.org/10.1093/bioinformatics/btq134.

[33] Wilcoxon, Frank. "Individual Comparisons by Ranking Methods." *Biometrics Bulletin* 1, no. 6 (1945): 80-83. https://doi.org/10.2307/3001968.

[34] Demšar, Janez. "Statistical Comparisons of Classifiers over Multiple Data Sets." *Journal of Machine Learning Research* 7 (2006): 1-30.