African Journal of Advanced Pure and Applied Sciences (AJAPAS)
Online ISSN: 2957-644X
Volume 5, Issue 3, 2026
Page No: 458-462
Website: *https://aaasjournals.com/index.php/ajapas/index*
ISI 2025: 1.126 **SJIFactor 2024: 6.752** **معامل التأثير العربي: 1.62**

Machine Learning versus Multiple Linear Regression for Predicting Concrete Compressive Strength: A Group-Aware Validation Study on a Pooled 431-Record Database

Ali Hasan [*1], Wasfi Albadry [2], Mohammed Almisteeri [3], Mohammed Zobi [4]
[1,3] Faculty of Engineering and Petroleum, University of Benghazi, Benghazi, Libya
[2] Faculty of Engineering, University of Benghazi, Benghazi, Libya
[4] College of Science and Technology, Qaminis, Libya

عنوان البحث: مقارنة بين التعلم الآلي والانحدار الخطي المتعدد للتنبؤ بمقاومة الانضغاط للخرسانة: دراسة تحقق مراعية للمجموعات على قاعدة بيانات مجمعة تضم 431 سجلاً

علي حسن [*1]، وصفي البدري [2]، محمد المستيري [3]، محمد زوبي [4]
[1،3] كلية الهندسة والنفط، جامعة بنغازي، بنغازي، ليبيا
[2] كلية الهندسة، جامعة بنغازي، بنغازي، ليبيا
[4] كلية العلوم والتقنية، قمينس، ليبيا

[*]*Corresponding author: Ali.hasan@uob.edu.ly*

Received: June 25, 2026	Accepted: September 13, 2026	Published: September 24, 2026

Abstract:
Multiple linear regression (MLR) is widely used for early prediction of concrete compressive strength because it is transparent and accessible to practising engineers, but recent literature suggests that machine-learning (ML) methods achieve higher accuracy on large, heterogeneous datasets. This study re-examines that question on a pooled, multi-source database of 431 concrete mixture records, comparing three MLR specifications against Random Forest, Gradient Boosting, and a neural network (multilayer perceptron). A key methodological issue is addressed explicitly: 96 unique mix designs in the database are each tested at multiple curing ages, so a naive random train/validation split places near-duplicate records of the same mix design on both sides of the split, inflating apparent ML accuracy through data leakage. A group-aware split that holds out entire mix designs was therefore used, yielding 379 calibration records (73 mix designs) and 52 independent validation records (23 previously unseen mix designs). Under this rigorous validation scheme, the linear models generalised poorly to unseen mix designs (validation R^2 between -0.30 and 0.18, mean absolute percentage error 32-49%), while Gradient Boosting achieved the best performance (validation $R^2 = 0.894$, MAPE = 12.4%), followed by Random Forest ($R^2 = 0.848$, MAPE = 16.5%); the neural network was intermediate ($R^2 = 0.672$, MAPE = 21.5%). The water-cement ratio and curing age were the dominant predictors in both ensemble models, consistent with established concrete technology. The results indicate that, once evaluated under a leakage-free, mix-design-independent validation protocol, ensemble machine-learning methods provide a substantially more reliable prediction of concrete strength than classical multiple linear regression on this type of pooled dataset, while also demonstrating that validation methodology itself materially affects the conclusions drawn from such comparisons.

Keywords: concrete compressive strength, machine learning, random forest, gradient boosting, neural network, multiple linear regression, data leakage, group-aware cross-validation.

الملخص
يُستخدم الانحدار الخطي المتعدد (MLR) على نطاق واسع للتنبؤ المبكر بمقاومة الانضغاط للخرسانة لشفافيته وسهولة استخدامه من قبل المهندسين الممارسين، إلا أن الأدبيات الحديثة تشير إلى أن طرق التعلم الآلي (ML) تحقق دقة أعلى على مجموعات البيانات الكبيرة والمتنوعة. تعيد هذه الدراسة فحص هذه المسألة على قاعدة بيانات مجمعة من مصادر

متعددة تحتوي على 431 سجلاً لخلطات خرسانية، مع مقارنة ثلاثة نماذج للانحدار الخطي المتعدد مقابل الغابات العشوائية (Random Forest)، والتدرج المعزز (Gradient Boosting)، والشبكة العصبية (Multilayer Perceptron). تم التعامل بوضوح مع مسألة منهجية رئيسية: 96 تصميم خلطة فريد في قاعدة البيانات تم اختبار كل منها عند أعمار معالجة متعددة، مما يعني أن التقسيم العشوائي التقليدي يضع سجلات متطابقة تقريباً في مجموعتي التدريب والتحقق، مما يضخم دقة التعلم الآلي بسبب تسرب البيانات. لذلك تم استخدام تقسيم مراع للمجموعات يستبعد خلطات كاملة، مما نتج عنه 379 سجلاً للعايرة (73 خلطة) و52 سجلاً للمصادقة المستقلة (23 خلطةً غير معروضة مسبقاً). تحت هذا المخطط الصارم للمصادقة، كان تعميم النماذج الخطية ضعيفاً على الخلطات غير المعروضة (معامل التحديد للتحقق R^2 بين -0.30 و0.18)، بينما حققت نموذج التدرج المعزز أفضل أداء ($R^2 = 0.894$).

الكلمات المفتاحية: مقاومة الانضغطات للخرسانة، التعلم الآلي، الغابات العشوائية، التدرج المعزز، الشبكات العصبية، الانحدار الخطي المتعدد، تسرب البيانات.

1. Introduction

Multiple linear regression (MLR) performed in general-purpose statistical software such as SPSS remains a popular tool for early-age prediction of concrete compressive strength because it produces transparent, hand-computable equations that do not require programming expertise. A companion study [1] developed and validated three MLR specifications on a pooled, literature-compiled database of 431 concrete mixture records, achieving calibration coefficients of determination (R^2) in the range 0.55-0.65 and independent-validation mean absolute percentage errors (MAPE) of 32-38%. A subsequent literature review [2] situated these results within a broader body of work and noted a recurring pattern: large, heterogeneous, multi-source databases tend to favour machine-learning (ML) methods over classical regression, whereas small, single-source datasets often report very high regression R^2 values that may not generalise.

The present study tests that hypothesis directly on the same 431-record database used in [1], comparing the original MLR specifications against three widely used ML techniques: Random Forest, Gradient Boosting, and a feed-forward neural network. In the course of preparing the calibration/validation split, a structural feature of the database was identified that has direct methodological consequences: many records share identical cement, water, water-cement ratio, and aggregate proportions and differ only in curing age, because the original sources tested the same mix design at several ages (typically 3, 7, 28, 90, and 180 days). A conventional random split at the level of individual records therefore risks placing near-duplicate feature vectors of the same underlying mix design on both sides of the calibration/validation boundary. Because ML models such as Random Forest can effectively memorise a specific combination of mixture proportions, this form of data leakage can artificially inflate validation accuracy for ML methods without a corresponding improvement in genuine predictive ability on new, unseen mix designs.

The specific objectives of this study are therefore to: (1) identify and quantify this data-leakage risk in the pooled database; (2) implement a group-aware calibration/validation split that holds out entire mix designs rather than individual records; (3) compare the predictive accuracy of MLR against Random Forest, Gradient Boosting, and a neural network under this rigorous validation protocol; and (4) interpret the resulting differences in the context of the mixture variables identified as dominant predictors in [1] and [2].

2. Materials and Database

The database used in this study is the same 431-record, multi-source pooled dataset compiled and described in the companion study [1], comprising cement content (C), water content (W), water-cement ratio (W/C), fine aggregate or sand content (FA), coarse aggregate or gravel content (CA), curing age, and measured compressive strength. On inspection, the 431 records reduce to 96 unique combinations of (C, W/C, W, FA, CA); the remaining variation arises from the same mix design being tested at multiple curing ages, and, for a subset of mix designs, from apparent repeat records at the same age. Two mix designs alone account for 260 of the 431 records (60%), reflecting uneven representation of certain mixture proportions across the pooled literature sources; the remaining 94 mix designs contribute between 1 and 10 records each.

3. Methodology

3.1 Group-aware calibration/validation split

To avoid the data-leakage risk described in Section 1, each unique combination of (C, W/C, W, FA, CA) was assigned a mix-design group identifier. Records were then partitioned so that all records belonging to a given mix design fall entirely within either the calibration set or the validation set, never both. Validation groups were selected by random sampling of mix designs (excluding the two dominant groups noted in Section 2, to avoid an oversized validation partition), targeting a validation-set size comparable to the 50-record hold-out used in the companion MLR study [1]. This produced a calibration set of 379 records spanning 73 mix designs and an independent validation set of 52 records spanning 23 previously unseen mix designs.

3.2 Linear regression models

For comparability with [1], three MLR specifications were re-fitted on the group-aware calibration set using ordinary least squares: Model 1 (all six predictors: C, W/C, W, FA, CA, Age), Model 2 (five predictors, excluding W/C), and Model 3 (four predictors, excluding C and W).

3.3 Machine-learning models

Three ML techniques were fitted on the same six predictors (C, W/C, W, FA, CA, Age) used in Model 1, to ensure a fair comparison against the strongest linear specification: (i) Random Forest, an ensemble of de-correlated regression trees [3]; (ii) Gradient Boosting, an additive ensemble of shallow regression trees fitted sequentially to the residuals of the preceding trees [4]; and (iii) a multilayer perceptron (MLP) neural network with standardised input features. Hyperparameters for all three ML models (number of trees, tree depth, minimum leaf size, learning rate, and, for the MLP, hidden-layer architecture and L2 regularisation strength) were selected by grid search using five-fold group-aware cross-validation within the calibration set (GroupKFold, grouped by mix design), so that no fold ever separated records of the same mix design between training and testing during hyperparameter tuning. All models were implemented in scikit-learn [10].

3.4 Evaluation metrics

Model performance was assessed on both the calibration set (in-sample fit, R^2) and the independent validation set of previously unseen mix designs, using the Pearson correlation coefficient (R), the coefficient of determination (R^2), the mean absolute error (MAE), the root-mean-square error (RMSE), and the mean absolute percentage error (MAPE), consistent with the metrics reported in the companion MLR study [1].

4. Results

4.1 Calibration and validation performance

Table 1. Calibration and independent-validation performance of all six models (validation set: 52 records, 23 unseen mix designs)

Model	Calibration R^2	Validation R^2	Validation R	MAE (MPa)	RMSE (MPa)	MAPE (%)
MLR Model 1 (6 var.)	0.631	-0.304	0.798	13.49	16.27	49.3
MLR Model 2 (5 var., no W/C)	0.555	0.177	0.787	10.13	12.92	32.2
MLR Model 3 (4 var., no C & W)	0.597	-0.198	0.803	12.87	15.59	46.6
Random Forest	0.969	0.848	0.928	4.82	5.55	16.5
Gradient Boosting	0.981	0.894	0.949	3.55	4.63	12.4
Neural Network (MLP)	0.873	0.672	0.822	6.40	8.16	21.5

All three linear models achieved calibration R^2 in the range 0.55-0.63, consistent with the companion study [1], but their performance on unseen mix designs was poor: Models 1 and 3 produced negative validation R^2, meaning their predictions were less accurate than simply using the mean measured strength of the validation set, despite reasonably strong linear correlation ($R \approx 0.80$) between predicted and measured values — indicating a systematic bias (offset) rather than a lack of ranking ability. Model 2 generalised somewhat better (validation $R^2 = 0.177$, MAPE = 32.2%) but remained far less accurate than any of the three ML models. Gradient Boosting achieved the best overall performance (validation $R^2 = 0.894$, MAPE = 12.4%), followed closely by Random Forest ($R^2 = 0.848$, MAPE = 16.5%); the neural network trailed both ensemble methods ($R^2 = 0.672$, MAPE = 21.5%), plausibly reflecting the comparatively small number of independent mix designs (73) available for training a more parameter-heavy model.

4.2 Predicted versus measured strength

Figure 1 illustrates the pattern underlying Table 1: Model 1 systematically under-predicts strength across most of the validation range, with predictions clustering well below the line of equality, whereas the Random Forest and Gradient Boosting predictions track the line of equality closely across the full 15-80 MPa range represented in the validation set. The neural network shows a similar qualitative pattern to the ensemble models but with greater scatter, particularly for validation mixtures below 30 MPa.

4.3 Feature importance

Both ensemble models identified the water-cement ratio as the dominant predictor (relative importance 0.57 and 0.54 for Random Forest and Gradient Boosting respectively), followed by curing age (0.26 and 0.25) and sand content (0.12 and 0.10), with cement content, water content, and gravel content each contributing less than 0.06. This ordering is consistent with the strong bivariate correlation between the water-cement ratio and strength reported in the companion study [1] ($r = -0.657$) and with the wider literature synthesised in [2], and confirms that the water-cement ratio's importance is not an artefact of the linear-regression collinearity discussed in [1], since it remains dominant in tree-based models that are not subject to the same multicollinearity concerns.

5. Discussion

The comparison in Table 1 supports the hypothesis, drawn from the wider literature synthesis in [2], that ensemble machine-learning methods outperform classical multiple linear regression on large, pooled, heterogeneous concrete-strength databases. The magnitude of the difference observed here, however, is considerably larger than the moderate gap reported by some large-database comparative studies [5, 6]: under the group-aware validation protocol used in this study, the linear models were essentially unable to generalise to unseen mix designs (negative or near-zero R^2), while Gradient Boosting explained close to 90% of the validation-set variance.

A second, methodological, contribution of this study is the demonstration that validation strategy materially affects the conclusions drawn from an ML-versus-regression comparison on this type of dataset. An initial, record-level random split (not reported in full here) produced validation R^2 values above 0.95 for the ensemble models, a result driven substantially by data leakage from near-duplicate mix designs shared between the calibration and validation partitions. Once the split was made group-aware at the level of the underlying mix design, ML validation accuracy fell to the more moderate, but far more credible, values reported in Table 1.

6. Conclusions

On the pooled, 431-record concrete-strength database compiled in the companion study, a substantial fraction of records (335 of 431) share mixture proportions identical to at least one other record, differing only in curing age; a record-level random calibration/validation split therefore risks data leakage for machine-learning models trained on this data. Under a group-aware split that holds out entire, previously unseen mix designs for validation, ensemble tree-based methods (Gradient Boosting and Random Forest) offer a meaningfully more accurate early-age strength prediction than SPSS-based multiple linear regression.

Compliance with ethical standards

Disclosure of conflict of interest

The authors declare that they have no conflict of interest.

References

[1] Al-Tahouri, D. H., Kernaf, E. S., & Al-Zwai, A. Statistical prediction of concrete compressive strength using multiple linear and nonlinear regression models developed in SPSS. (Unpublished study; source of the 431-record database used in the present work).

[2] Al-Zwai, A. Statistical and machine-learning approaches for predicting concrete compressive strength: a review of methods, predictors, and practical performance. (Companion literature review).

[3] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

[4] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of Statistics, 29(5), 1189-1232.

[5] DeRousseau, M. A., Laftchiev, E., Kasprzyk, J. R., Balaji, R., & Srubar III, W. V. (2019). Machine learning methods for predicting the field compressive strength of concrete (Technical Report No. TR2019-096). Mitsubishi Electric Research Laboratories.

[6] Young, B. A., Hall, A., Pilon, L., Gupta, P., & Sant, G. (2018). 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, 379-388.

[7] Bansal, A., et al. (2023). Comparative analysis of concrete compressive strength prediction using machine learning techniques with hyperparameter optimisation.

[8] Assegie, T. A., et al. (2022). Evaluation of regression models for concrete compressive strength prediction using the UCI database.
[9] Neville, A. M. (2011). Properties of Concrete (5th ed.). Pearson.
[10] Pedregosa, F., et al. (2011). Scikit-learn: machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.
[11] Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of IJCAI, 14(2), 1137-1145.
[12] Roberts, D. R., et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography, 40(8), 913-929.

