



RESEARCH ARTICLE

OPEN ACCESS

MACHINE LEARNING APPROACHES FOR PREDICTING COMPRESSIVE STRENGTH OF CONCRETE: A COMPARATIVE PERFORMANCE ANALYSIS

Slim Rouabah*¹

¹Department of Electronics and communication, Faculty of Sciences and Technology, Djilali Bounaama University of Khemis Miliana, Khemis Miliana, Ain Defla, Algeria.

¹<https://orcid.org/0009-0008-2295-9470>

E-mail: s.rouabah@univ-dbkm.dz

ARTICLE INFO

Article History

Received on: March 08, 2026

Reviewed on: April 08, 2026

Accepted on: July 01, 2026

Published on: July 31, 2026

Keywords:

Concrete compressive strength,
Machine learning,
Random Forest regression,
Predictive modeling,
Mix design.

ABSTRACT

Accurate prediction of concrete compressive strength is essential for effective mix design, quality control, and structural performance assessment. Conventional empirical models often exhibit limited accuracy due to the complex and nonlinear interactions among concrete constituents. This study investigates the applicability of several machine learning models for predicting the compressive strength of concrete using a publicly available experimental dataset comprising 1030 concrete mixtures. Linear regression was adopted as a baseline model and compared with support vector regression, random forest regression, and artificial neural networks. The performance of machine learning models was meticulously assessed using the coefficient of determination, root mean square error, and mean absolute error. Additionally, the models underwent five-fold cross-validation to evaluate their robustness and generalization capabilities. The results unambiguously demonstrate that machine learning models significantly outperform linear regression models. Cross-validation results confirm the stability and reliability of the developed models. Feature importance analysis reveals that curing age and cement content are the most influential parameters affecting compressive strength, followed by water content, which is consistent with established concrete material behavior. The findings demonstrate that machine learning models, particularly random forest regression, can serve as effective supporting tools for preliminary concrete mix design and performance evaluation.



Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Concrete remains the most extensively used construction material in civil engineering due to its adaptability, durability, and economic efficiency. Among its mechanical properties, compressive strength is the most critical parameter, as it directly governs structural design, quality control, and compliance with design codes. The compressive strength of concrete is influenced by a wide range of parameters, including cement content, water-cement ratio, aggregate proportions, supplementary cementitious materials, chemical admixtures, and curing age. The combined effect of these parameters results in complex and highly nonlinear behavior that is difficult to represent using simple analytical expressions [1], [2]. Traditionally, concrete compressive strength is determined through destructive laboratory testing of standard specimens at designated curing ages. Although this approach provides reliable and standardized results, it is time-consuming, labor-intensive, and requires substantial material resources. Consequently, there is a long-standing interest in developing predictive models that estimate compressive strength from mixture proportions and curing age, thereby supporting efficient mix design, early decision-making, and reduction of experimental effort [3]. Numerous empirical and regression-based models have been proposed to estimate compressive strength. While such models are attractive due to their simplicity, their predictive capability is often limited by linear assumptions and simplified relationships that fail to capture the nonlinear interactions among concrete constituents [4].

These limitations are particularly evident in modern concrete mixtures incorporating supplementary cementitious materials and chemical admixtures, where strength development mechanisms are more complex. In recent years, data-driven approaches have emerged as effective alternatives for modeling concrete properties. Machine learning (ML) techniques are capable of learning complex nonlinear relationships directly from experimental data without requiring predefined functional forms. Artificial neural networks (ANNs), support vector machines (SVMs), and ensemble learning methods have been successfully applied to predict concrete compressive strength with improved accuracy compared to conventional regression approaches [5-8]. The theoretical foundation of support vector-based methods, originally developed within statistical learning theory, enables effective handling of nonlinear relationships and high-dimensional data [9]. Subsequent studies expand this approach to a wider range of concrete types, including high-performance and high-strength concretes, confirming the robustness of neural networks in handling complex mixture interactions [10], [11]. Support vector regression (SVR), based on statistical learning theory [9], has also been widely employed for predicting concrete compressive strength. Compared with ANNs, SVR models offer strong generalization capability and reduced sensitivity to local minima. Several studies report that SVR provides competitive or superior performance relative to ANN models, particularly when kernel functions and hyperparameters are appropriately selected [7], [12]. More recently, ensemble learning techniques such as random forest (RF) regression have gained increasing attention in civil engineering applications. Random forest models aggregate multiple decision trees to enhance prediction accuracy and robustness, while also enabling evaluation of feature importance [8]. Prior studies demonstrate that RF models can achieve high prediction accuracy for concrete compressive strength and provide meaningful insight into the relative influence of mixture constituents, improving model interpretability [13], [14]. Despite their demonstrated predictive capability, the application of ML techniques in concrete engineering requires careful consideration of model robustness, validation strategy, and physical interpretability. High prediction accuracy alone is insufficient for engineering practice; models must also be stable, generalizable, and consistent with established material behavior. Comprehensive training procedures, independent testing, and cross-validation are therefore essential to ensure reliable performance and to avoid overfitting [8], [15].

Publicly available experimental datasets, such as the concrete compressive strength dataset compiled by Yeh [10], have facilitated comparative evaluation of different modeling approaches and contributed to reproducibility in this research area. The present study addresses the need for rigorous comparative evaluation of ML models in concrete compressive strength prediction. Specifically, this research investigates whether advanced ML techniques significantly outperform traditional linear approaches and identifies which modeling strategy offers the optimal balance between accuracy, robustness, and practical applicability for engineering decision-making. The research question is: Which machine learning model provides the most reliable and generalizable predictions of concrete compressive strength when evaluated through rigorous validation procedures? The objectives of this work are to (1) assess the predictive capability of different ML models, including linear regression (LR), support vector regression (SVR), random forest (RF), and artificial neural network (ANN) models; (2) evaluate model robustness through systematic validation procedures; and (3) examine the relative influence of mixture parameters on compressive strength. The hypothesis is that ensemble and kernel-based methods (RF and SVR) achieve superior generalization performance compared to conventional linear regression and basic neural network architectures, provided that proper cross-validation and hyperparameter optimization are employed. The findings aim to support the practical use of data-driven models as auxiliary tools for concrete mix design and performance assessment. The importance of this study lies in its systematic comparison of multiple ML paradigms under consistent validation conditions, addressing a gap in existing literature where studies often emphasize single-model accuracy without comprehensive robustness assessment. A limitation of this study is its reliance on a single publicly available dataset [10], which may not encompass all regional material variations or specialized concrete types; however, this constraint ensures reproducibility and facilitates direct comparison with prior research. The method employed in this work involves training and validating four distinct predictive models—linear regression (LR), support vector regression (SVR), random forest (RF), and artificial neural networks (ANNs)—on a comprehensive experimental dataset obtained from a publicly available source [5], [10]. Model development follows a rigorous procedure incorporating cross-validation, independent testing, and performance evaluation using multiple statistical indicators to ensure reliable and generalizable results.

II. MATERIALS AND METHODS

II.1 DATASET DESCRIPTION

The dataset employed in this study was obtained from the UCI Machine Learning Repository and was originally compiled from experimental investigations on concrete compressive strength [5]. It consists of 1030 concrete mixtures with varying proportions of constituents and curing ages. Each data sample includes eight input variables: cement content, blast furnace slag, fly ash, water, superplasticizer, coarse aggregate, fine aggregate, and curing age. The output variable is the concrete compressive strength expressed in megapascals (MPa). The dataset represents a wide range of mixture proportions and curing conditions, making it suitable for evaluating predictive models for concrete compressive strength. Prior to model development, the dataset was examined for completeness, and no missing values were observed. To develop and evaluate the predictive models, the dataset was randomly divided into training and testing subsets. 80% of the data were used for model training, while the remaining 20% were reserved for independent testing. This separation ensured that model performance was evaluated on unseen data. For machine learning models sensitive to the scale of input variables, namely support vector regression and artificial neural networks, input features were standardized using zero-mean and unit-variance normalization. Linear regression and random forest models were trained using the original, unscaled data, as these models are not significantly affected by feature scaling.

II.2 MACHINE LEARNING MODELS

Four predictive models were developed to estimate the compressive strength of concrete. These models represent increasing levels of complexity and nonlinear learning capability.

II.2.1 Linear Regression

Linear regression was adopted as a baseline model to establish a reference for comparison with nonlinear machine learning techniques. The relationship between the compressive strength and the input variables is expressed as:

$$y = \beta_0 + \sum_{i=1}^n \beta_i x_i + \varepsilon \quad (1)$$

Where y is the predicted compressive strength, x_i represents the input variables, β_i are the regression coefficients, β_0 is the intercept, and ε is the error term. Model parameters were estimated using the ordinary least squares method.

II.2.2 Support Vector Regression

Support vector regression (SVR) is a kernel-based learning method derived from statistical learning theory [9]. SVR aims to find a function that deviates from the actual target values by no more than a predefined tolerance while maintaining model flatness. The SVR function can be expressed as:

$$f(x) = \sum_{i=1}^N (\alpha_i - \alpha_i^*) K(x_i, x) + b \quad (2)$$

Where α_i and α_i^* are Lagrange multipliers, $K(x_i, x)$ is the kernel function, b is the bias term, and N is the number of support vectors. In this study, the radial basis function kernel was employed and defined as:

$$K(x_i, x) = \exp(-\gamma \|x_i - x\|^2) \quad (3)$$

Where γ is the kernel parameter controlling the width of the function.

II.2.3 Random Forest Regression

Random forest regression is an ensemble learning technique that combines multiple decision trees to improve prediction accuracy and robustness [8]. The predicted output of the random forest model is obtained by averaging the predictions of individual trees:

$$\hat{y} = \frac{1}{M} \sum_{j=1}^M T_j(x) \quad (4)$$

Where M is the number of decision trees and $T_j(x)$ represents the prediction of the j -th tree. Feature importance is evaluated based on the contribution of each input variable to the reduction of prediction error across all trees.

II.2.4 Artificial Neural Network

The artificial neural network (ANN) model consists of interconnected neurons organized in layers. The output of a neuron in a hidden or output layer is given by:

$$z_j = \phi\left(\sum_{i=1}^n w_{ij} x_i + b_j\right) \quad (5)$$

Where w_{ij} are the connection weights, b_j is the bias term, and $\phi(\cdot)$ is a nonlinear activation function. The ANN model was trained using an iterative optimization process to minimize prediction error between the predicted and experimental compressive strength values.

II.3 MODEL TRAINING AND VALIDATION

Model training was performed using the training dataset, while predictive performance was evaluated using an independent testing dataset. The accuracy of the developed models was quantified using the coefficient of determination (R^2) in equation (6), root mean square error (RMSE) in equation (7), and mean absolute error (MAE) in equation (8), defined as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2} \quad (6)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2} \quad (7)$$

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i| \quad (8)$$

Where y_i and $\hat{y}_i$ are the experimental and predicted compressive strength values, respectively, $\bar{y}$ is the mean of the experimental values, and N is the number of values. To further assess robustness and generalization capability, five-fold cross-validation was employed. The dataset was divided into five subsets, and each subset was used once for validation while the remaining subsets were used for training. The mean and standard deviation of cross-validated R^2 values were used as indicators of model stability. Feature importance was evaluated using the random forest model by quantifying the contribution of each input variable to reducing prediction error. This analysis provides insight into the relative influence of mixture parameters on compressive strength and supports interpretation of the results from an engineering perspective. All models were implemented using the Python programming language and standard machine learning libraries. Model development, training, validation, and performance evaluation were conducted using reproducible computational procedures under Kaggle platform to ensure transparency and reliability of the results.

III. RESULTS AND DISCUSSION

III.1 PREDICTIVE PERFORMANCE OF MODELS

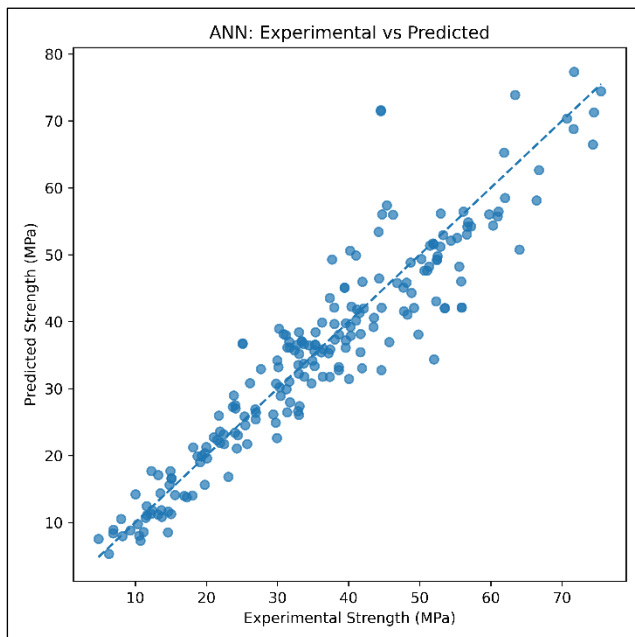
The predictive performance of the developed models is summarized in Tab. 1. Linear regression, used as a baseline model, exhibited limited accuracy with an (R^2) value of 0.63, RMSE of 9.80 MPa, and MAE of 7.75 MPa. These results confirm that linear relationships are insufficient to describe the complex and nonlinear interactions governing concrete compressive strength.

Table 1: Predictive performance, computational cost, and cross-validation results of the developed models. (CV: Cross-Validation, std: Standard deviation).

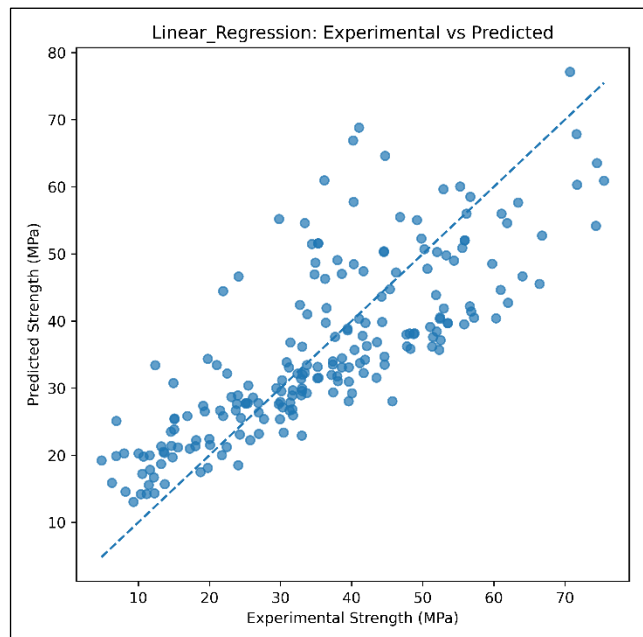
Model	R^2	RMSE	MAE	Training Time (s)	Inference Time (s)	CV R^2 (Mean)	CV R^2 Std
Linear Regression	0.627553	9.796476	7.745559	0.015814	0.002012	0.600696	0.047818
SVR	0.869606	5.796501	4.093966	0.096982	0.008953	0.871918	0.01233
Random Forest	0.881425	5.527568	3.781352	1.110243	0.026125	0.907017	0.016085
ANN	0.875042	5.6744	3.984849	3.028517	0.000387	0.894505	0.0083

Source: Authors, (2026).

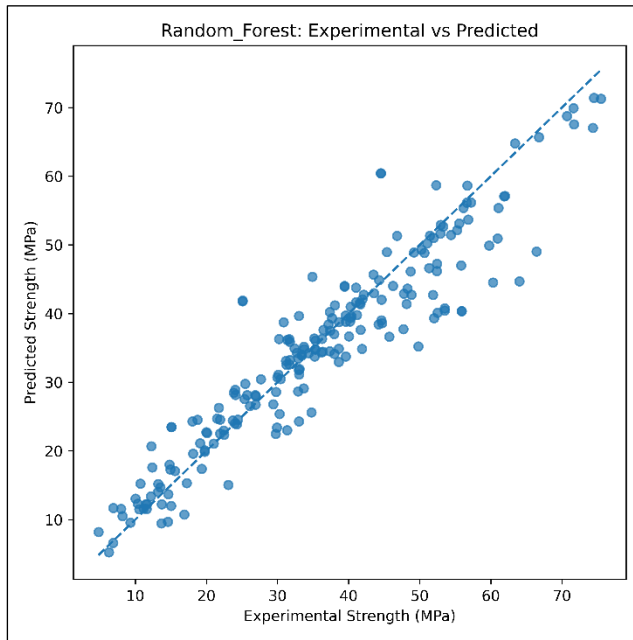
All machine learning models demonstrated a substantial improvement in prediction accuracy compared to linear regression. Support vector regression achieved an (R^2) value of 0.87 with an RMSE of 5.80 MPa, indicating its effectiveness in capturing nonlinear behavior. The artificial neural network model yielded comparable performance with an (R^2) value of 0.88 and an RMSE of 5.67 MPa. Among all models, the random forest regression model provided the best overall performance, achieving the highest coefficient of determination ($R^2 = 0.88$) and the lowest prediction errors (RMSE = 5.53 MPa, MAE = 3.78 MPa). The superior performance of the random forest model can be attributed to its ensemble structure, which effectively reduces variance and improves generalization by aggregating multiple decision trees. These results in Fig. 1 and Tab. 1 demonstrate that machine learning models significantly outperform traditional regression approaches in predicting concrete compressive strength and are capable of accurately modeling the complex relationships between mixture parameters and strength development.



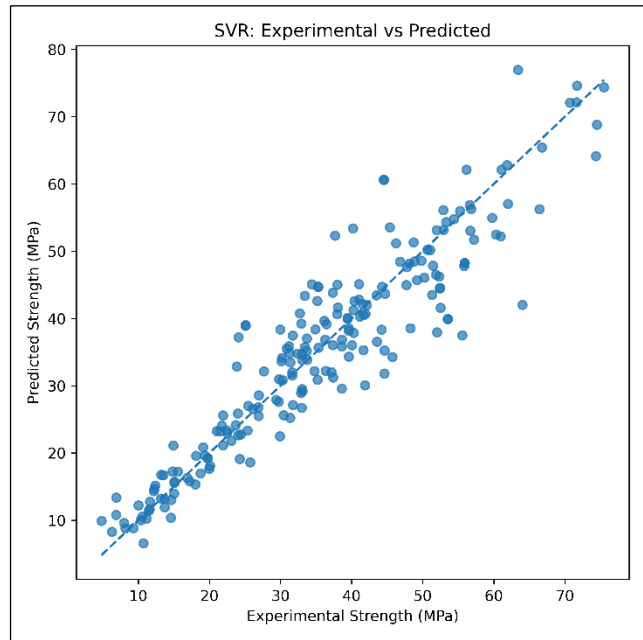
(1): ANN.



(2): Linear regression.



(3): Random forest.



(4): SVR.

Figure 1: Comparison between experimental and predicted compressive strength.

Source: Authors, (2026).

III.2 CROSS-VALIDATION AND MODEL ROBUSTNESS

To evaluate model robustness and generalization capability, five-fold cross-validation was conducted for all models. The mean and standard deviation of the cross-validated (R^2) values are also presented in Tab. 1. The random forest model achieved the highest mean cross-validated (R^2) value of 0.91 with a relatively low standard deviation of 0.016, indicating stable and reliable performance across different data partitions. The artificial neural network model also showed strong robustness, with a mean cross-validated (R^2) of 0.89 and the lowest variability among machine learning models. Support vector regression demonstrated consistent performance with a mean cross-validated (R^2) of 0.87. In contrast, linear regression exhibited both low accuracy and higher variability, further confirming its limited suitability for modeling concrete compressive strength. The cross-validation results indicate that the machine learning models, particularly random forest and artificial neural networks, generalize well and are not overly sensitive to data partitioning.

III.3 COMPUTATIONAL EFFICIENCY

In addition to prediction accuracy, computational efficiency was evaluated by measuring training and inference times. Linear regression required negligible computational effort, as expected. Among the machine learning models, support vector regression and random forest required training times below one second, while the artificial neural network required the longest training time (approximately 3.65 s) due to its iterative optimization process. Inference times for all models were extremely low, indicating that once trained, the models can be used efficiently for real-time or near-real-time strength prediction. These results suggest that the developed machine learning models are computationally feasible for practical engineering applications.

III.4 FEATURE IMPORTANCE ANALYSIS

Feature importance analysis was conducted using the random forest model to identify the relative influence of input variables on concrete compressive strength. The results are illustrated in Tab. 2.

Table 2: Predictive performance of the developed models.

Feature	Age	Cement	Water	furnace slag	superplasticizer	Fine aggregate	Coarse aggregate	Fly ash
Importance	0.333001	0.325946	0.125096	0.07703	0.056764	0.035705	0.027782	0.018677

Source: Authors, (2026).

Curing age was identified as the most influential parameter, accounting for approximately 33% of the total importance. This finding is consistent with established concrete hydration mechanisms, as strength development is strongly governed by the progression of cement hydration over time. Cement content was the second most influential parameter, with a relative importance of approximately 33%, reflecting its direct contribution to the formation of load-bearing hydration products. Water content exhibited a notable influence on compressive strength, accounting for approximately 13% of total importance. This observation aligns with the well-known effect of the water–cement ratio on concrete porosity and strength. Supplementary cementitious materials, such as blast furnace slag and fly ash, showed lower relative importance. This can be attributed to their delayed contribution to strength development, particularly at early curing ages represented in the dataset.

Aggregate contents exhibited comparatively smaller influence on compressive strength, which is consistent with the fact that aggregates primarily affect strength indirectly through packing density and interfacial transition zone characteristics.

III.5 ENGINEERING IMPLICATIONS

The results demonstrate that machine learning models can provide accurate and robust predictions of concrete compressive strength when trained on comprehensive experimental datasets. The strong agreement between feature importance analysis and established concrete material behavior indicates that the developed models are physically meaningful and suitable for engineering applications. The random forest model, in particular, offers a favorable balance between prediction accuracy, robustness, interpretability, and computational efficiency. Therefore, it can be effectively used as a supporting tool for preliminary mix design, quality control, and performance evaluation in concrete engineering practice.

IV. CONCLUSIONS AND LIMITATIONS

IV.1 CONCLUSIONS

This study investigated the applicability of several machine learning models for predicting the compressive strength of concrete using a comprehensive experimental dataset. Linear regression was employed as a baseline model and compared with support vector regression, random forest regression, and artificial neural networks. The results demonstrated that machine learning models significantly outperform linear regression in predicting concrete compressive strength, confirming the nonlinear nature of the relationship between mixture parameters and strength development. Among the evaluated models, random forest regression exhibited the highest prediction accuracy and robustness, achieving an (R^2) value of 0.88 on the testing dataset and a mean cross-validated (R^2) value of 0.91. The artificial neural network and support vector regression models also showed strong predictive capability, with comparable performance and stable generalization behavior. The low variability observed in cross-validation results indicates that the developed models are reliable and not sensitive to data partitioning. Feature importance analysis revealed that curing age and cement content are the most influential parameters affecting compressive strength, followed by water content. These findings are consistent with established concrete material behavior and confirm that the machine learning models capture physically meaningful relationships. The results suggest that data-driven models, particularly random forest regression, can serve as effective supporting tools for preliminary concrete mix design, quality control, and performance assessment.

IV.2 LIMITATIONS

Despite the promising results, several limitations of this study should be acknowledged. The developed models are based on a single publicly available dataset, and their predictive capability is therefore constrained to the range of mixture proportions and curing ages represented in the data. Factors such as curing conditions, aggregate properties, environmental exposure, and specimen geometry were not explicitly considered. Additionally, the models were developed using compressive strength as the sole output variable. Extension of the proposed approach to other concrete properties, such as tensile strength, durability indicators, or long-term performance, would require additional datasets and model adaptation. While machine learning models provide high prediction accuracy, they should be regarded as complementary tools rather than replacements for experimental testing and established design methods. Future research may focus on integrating experimental data from different sources, incorporating additional material and environmental parameters, and exploring hybrid modeling approaches that combine data-driven techniques with physically based models to further enhance reliability and applicability.

V. AUTHOR'S CONTRIBUTION

Conceptualization: Slim Rouabah.

Methodology: Slim Rouabah.

Investigation: Slim Rouabah.

Discussion of results: Slim Rouabah.

Writing – Original Draft: Slim Rouabah.

Writing – Review and Editing: Slim Rouabah.

Resources: Slim Rouabah.

Supervision: Slim Rouabah.

Approval of the final text: Slim Rouabah.

VI. ACKNOWLEDGMENTS

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest/Competing interests: The author declare that he has no conflicts of interest related to this work.

Ethics approval and consent to participate: Not applicable.

Data availability: <https://www.kaggle.com/datasets/elikplim/concrete-compressive-strength-data-set>.

Materials availability: Kaggle cloud platform.

VII. REFERENCES

[1] A. M. Neville, *Properties of Concrete*. Pearson, 2011.

[2] P. K. Mehta and P. J. M. Monteiro, *Concrete: Microstructure, Properties, and Materials*, 4th ed. McGraw-Hill, 2014.

- [3] P.-C. Aïtcin, *High Performance Concrete*, 1st ed. Boca Raton, FL, USA: CRC Press, 1998.
- [4] S. Popovics, "Analysis of the concrete strength versus water–cement ratio relationship," *Mater. Struct.*, vol. 23, no. 3, pp. 193–198, 1990.
- [5] M. Shaaban, M. Amin, S. Selim, *et al.*, "Machine learning approaches for forecasting compressive strength of high-strength concrete," *Sci. Rep.*, vol. 15, Art. no. 25567, 2025, doi: <https://doi.org/10.1038/s41598-025-10342-1>.
- [6] Z. Fei, S. Liang, Y. Cai, and Y. Shen, "Ensemble machine-learning-based prediction models for the compressive strength of recycled powder mortar," *Materials*, vol. 16, no. 2, Art. no. 583, 2023, doi: <https://doi.org/10.3390/ma16020583>.
- [7] A. Mandal, "Predicting compressive strength of concrete using advanced machine learning techniques: A combined dataset approach," *Research Square*, 2024, doi: <https://doi.org/10.21203/rs.3.rs-5401974/v1>.
- [8] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
- [9] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer, 1995.
- [10] I. C. Yeh, "Concrete compressive strength data set," UCI Machine Learning Repository, Univ. California, Irvine, CA, USA. Accessed: Mar. 8, 2026. [Online]. Available: <https://archive.ics.uci.edu/ml/datasets/Concrete+Compressive+Strength>
- [11] Y. T. Altuncı, "A comprehensive study on the estimation of concrete compressive strength using machine learning models," *Buildings*, vol. 14, no. 12, Art. no. 3851, 2024, doi: <https://doi.org/10.3390/buildings14123851>.
- [12] S. Zaman, M. M. Eid, E. A. Mattar, *et al.*, "Predicting concrete compressive strength using optimized deep learning and large language models," *Sci. Rep.*, 2026, doi: <https://doi.org/10.1038/s41598-026-41072-7>.
- [13] M. Z. Akber, G. A. Anwar, W.-K. Chan, and H.-H. Lee, "TPE-xgboost for explainable predictions of concrete compressive strength considering compositions, and mechanical and microstructure properties of testing samples," *Constr. Build. Mater.*, vol. 457, Art. no. 139398, 2024, doi: <https://doi.org/10.1016/j.conbuildmat.2024.139398>.
- [14] K. Xiao, H. Zhang, S. Wei, C. Zhu, J. He, S. Zhu, and X. Yang, "Prediction of compressive strength of high-performance concrete based on multiple machine learning models," *Front. Mater.*, vol. 12, Art. no. 1698248, 2025, doi: <https://doi.org/10.3389/fmats.2025.1698248>.
- [15] J. Zhang, Y. Zhao, and J. Chen, "Prediction of concrete compressive strength using machine learning and deep learning methods," *Struct. Concr.*, vol. 22, no. 1, pp. 342–357, 2021.