УДК 691.32:620.17:004.85 DOI 10.36910/775.24153966.2025.84.41

Н.Р. Печончик, О.С. Приходько

Луцький національний технічний університет

ПОРІВНЯЛЬНИЙ АНАЛІЗ МОДЕЛЕЙ МАШИННОГО НАВЧАННЯ ДЛЯ ПРОГНОЗУВАННЯ МІЦНОСТІ БЕТОНУ

У роботі розглянуто задачу прогнозування міцності бетону на стиск за його складом із використанням методів машинного навчання. На основі відкритого набору даних Concrete Compressive Strength виконано порівняльний аналіз лінійної, поліноміальної регресії, штучних нейронних мереж та ансамблевих методів. Показано, що прості лінійні моделі не здатні адекватно описати складні нелінійні процеси гідратації цементу, тоді як ансамблеві методи забезпечують оптимальне співвідношення точності та стійкості. Найкращий стабільний результат отримано за допомогою алгоритму Random Forest із середньою абсолютною похибкою 3.73 МПа. Окрім цього, шляхом лінеаризації мультиплікативної моделі виведено емпіричну степеневу формулу, придатну для експрес-оцінок міцності без використання обчислювальних засобів. Отримані результати підтверджують доцільність застосування ансамблевих і гібридних підходів для інженерних задач прогнозування властивостей будівельних матеріалів.

Ключові слова: *міцність бетону, машинне навчання, регресійний аналіз, Random Forest, нейронні мережі*

N.R. Pechonchyk, O.S. Prykhodko

COMPARATIVE ANALYSIS OF MACHINE LEARNING MODELS FOR PREDICTING CONCRETE COMPRESSIVE STRENGTH

This paper addresses the problem of predicting concrete compressive strength based on mixture composition using machine learning methods. A comparative analysis of linear regression, polynomial regression, artificial neural networks, and ensemble learning techniques was conducted using the publicly available Concrete Compressive Strength dataset. The results show that simple linear models are unable to adequately capture the complex non-linear physicochemical processes of cement hydration, while ensemble-based approaches provide a superior balance between accuracy and robustness. The best stable performance was achieved by the Random Forest algorithm, with a mean absolute error of 3.73 MPa. In addition, an empirical power-law formula was derived through linearization of a multiplicative model, enabling approximate strength estimation without computational resources. The findings confirm the effectiveness of ensemble and hybrid modeling approaches for practical engineering applications in construction materials science.

Keywords**:** *concrete compressive strength, machine learning, regression analysis, Random Forest, neural networks*

Introduction and problem statement. Concrete is the most widely used construction material worldwide, and its compressive strength is a key parameter governing the quality, reliability, and safety of building structures. Traditionally, compressive strength is determined through destructive laboratory testing of standard specimens, typically performed after 28 days of curing. This inherent time lag poses significant challenges for real-time quality control in concrete production and slows down construction processes.

Moreover, concrete compressive strength is a complex, non-linear function of mixture composition–including cement, water, aggregates, and chemical or mineral admixtures–as well as curing age. Classical empirical relationships, such as Abrams' law or Bolomey's formula, often fail to provide sufficient accuracy for modern multi-component concretes incorporating superplasticizers and supplementary cementitious materials. In this context, the application of machine learning (ML) techniques for *in silico* prediction of concrete properties has emerged as a relevant and promising research direction, offering the potential to replace time-consuming experimental testing with rapid computational estimates.

Analysis of the latest research and publications. The task of predicting concrete compressive strength from mixture proportions is a canonical benchmark problem in regression analysis within construction materials science and machine learning. Contemporary research efforts emphasize not only minimizing prediction error but also improving model interpretability and practical applicability for engineering practice.

In foundational works [1–4], the Concrete Compressive Strength dataset has been extensively used to demonstrate advanced modeling techniques. For example, study [1] proposes an ensemble of predictive intervals to assess reliability of predictions, while [2] introduces the SelectiveNet architecture, which can abstain from predictions in cases of low model confidence, addressing safety-critical decision contexts. Research [3] integrates boosting-based ensemble methods into the optimization of mixture compositions, and [4] explores symbolic regression for automatic derivation of analytical formulas.

Despite these advances, the choice of an optimal model that achieves high accuracy, robustness to overfitting, and interpretability remains an open research question.

Recent contributions in the literature extend these themes in several directions. Wang et al. [5] developed an interpretable machine learning framework based on boosted tree models, demonstrating that the XGBoost algorithm achieves high predictive accuracy on experimental concrete datasets and employing SHAP analysis to elucidate feature contributions consistent with material science principles (e.g., age and mix ratios). Similarly, Zhang et al. [6] applied interpretable ML models including Random Forest, AdaBoost, XGBoost, and LightGBM to high-performance concrete, combining Bayesian optimization with SHAP to both optimize prediction and analyze feature influence.

Other studies have explored interpretable and optimized regression frameworks for concrete strength prediction. An interpretable machine learning model for self-compacting concrete incorporating recycled aggregates and supplementary cementitious materials achieved high accuracy and identified key mix variables, demonstrating the utility of hyperparameter-tuned ensemble methods with feature attribution analysis [7]. Research focused on comparative evaluations of diverse ML techniques, including Random Forest, SVM, neural networks, and rule-based models, further confirms the superior performance of ensemble frameworks over basic linear approaches in concrete strength forecasting [8].

Collectively, these studies highlight ongoing efforts to balance accuracy, robustness, and interpretability in machine learning models for concrete compressive strength prediction. The literature indicates a continued trend toward ensemble and explainable models that provide not only strong predictive performance but also engineering insights into the influence of mix components on material properties.

The aim of the study. The aim of the paper is to conduct a comparative analysis of the effectiveness of machine learning models of different architectural complexity (from linear regression to deep neural networks) for the task of predicting concrete strength, as well as to construct an interpreted empirical dependence for express strength assessment.

Experimental methodology. To achieve the objectives of this study, the publicly available Concrete Compressive Strength dataset from the UCI Machine Learning Repository [9] was used. The dataset consists of 1030 observations describing concrete mixtures with varying compositions and curing ages.

Each sample is characterized by eight input variables, including the contents of cement, blast furnace slag, fly ash, water, superplasticizer, coarse aggregate, and fine aggregate (all measured in kg/m³), as well as the age of the concrete in days. The target variable is the compressive strength of concrete, expressed in megapascals (MPa).

Prior to model training, an exploratory data analysis (EDA) was conducted to assess data quality, variable distributions, and potential modeling challenges. Histograms of all input features and the target variable are shown in Figure 1.

The target variable exhibits an approximately normal distribution with a slight right skew and a concentration around 35–40 MPa, indicating a well-balanced range of concrete strengths suitable for regression modeling. In contrast, the input features demonstrate pronounced heterogeneity. Admixture-related variables (blast furnace slag, fly ash, and superplasticizer) are strongly right-skewed and zero-inflated, reflecting the prevalence of conventional concrete mixtures without additives. Cement content spans a wide range (approximately 100–550 kg/m³) with a multimodal structure, while the age variable is discrete, with peaks corresponding to standard curing periods (7, 28, and 90 days).

Overall, the dataset combines continuous, discrete, and sparse variables, forming a complex non-linear feature space. This characteristic motivates the use of non-linear and tree-based ensemble methods, such as Random Forest regression, which are better suited to capturing feature interactions and handling data sparsity than classical linear models.

The modeling workflow included checking the dataset for missing values (none were found), applying standard feature scaling where required particularly for neural networks and splitting the data into training and test sets using an 80/20 ratio with a fixed random seed to ensure reproducibility.

Three classes of regression models were investigated. As baseline approaches, linear regression and second-degree polynomial regression were employed to establish reference performance levels. For deep learning methods, fully connected multilayer perceptron (MLP) networks with varying architectures were explored, ranging from two to seven hidden layers and containing between 128 and 256 neurons per layer. In addition, an ensemble learning approach based on decision trees, namely the Random Forest Regressor, was evaluated due to its robustness to non-linear relationships and heterogeneous feature distributions.

Model performance was assessed using the mean absolute error (MAE), which provides an intuitive measure of prediction accuracy in physical units (MPa) and is therefore well suited for engineering interpretation. Finally, a linearization approach for a multiplicative regression model was applied to derive an empirical power-law relationship between the input variables and concrete compressive strength.

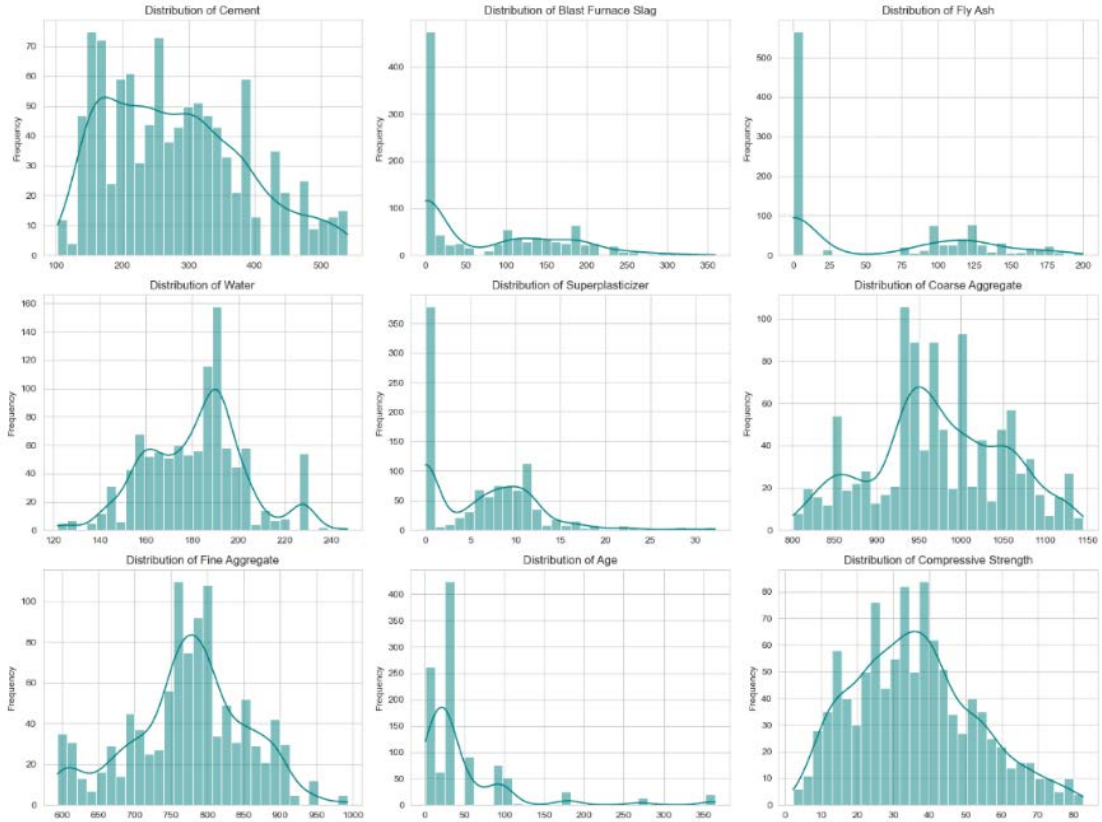

Fig 1. **Exploratory data analysis**

Presentation of the main research material. During the study, multiple regression models were trained and evaluated on a held-out test dataset in order to assess their predictive accuracy and practical applicability.

Linear regression exhibited the poorest performance, which can be attributed to its inability to represent the complex non-linear physicochemical processes governing cement hydration and strength development. Polynomial regression partially addressed this limitation by capturing non-linear interactions between mixture components–most notably the interaction between water and cement content–resulting in a noticeable improvement in prediction accuracy.

Artificial neural networks demonstrated a high modeling potential. Increasing the network depth from two to seven hidden layers led to a systematic improvement in prediction accuracy, with the best results obtained for the most complex architecture. However, the training process proved sensitive to weight initialization and hyperparameter settings, and the resulting models exhibit limited interpretability, effectively functioning as “black boxes” from an engineering perspective.

The most favorable balance between accuracy, robustness, and interpretability was achieved by the Random Forest model. This approach does not require feature scaling, is resilient to outliers, and is capable of capturing non-linear dependencies and feature interactions without explicitly introducing polynomial terms. A visual comparison between the predictions of linear regression and the Random Forest model is presented in Figure 2.

For practical applications where computational resources are unavailable, an empirical formula for estimating compressive strength was derived by logarithmic transformation of the input variables followed by linear regression. The resulting multiplicative power-law model for compressive strength σ□ is expressed as.

$$\sigma_{\mathrm{M}} = 257{,}2635(1 + w^{-0.9702})(1 + c^{0.774})(1 + a^{0.3109})(1 + bf^{0.0659}) \times \\ \times (1 + f^{0.0217})(1 + sp^{0.0715})(1 + fa^{-0.3776})(1 + ca^{-0.0272}).$$

where a denotes the curing age of concrete in days, and w, c, bf, f, sp, fa, and ca represent the masses (kg per 1 m³ of concrete) of water, cement, blast furnace slag, fly ash, superplasticizer, fine aggregate, and coarse aggregate, respectively.

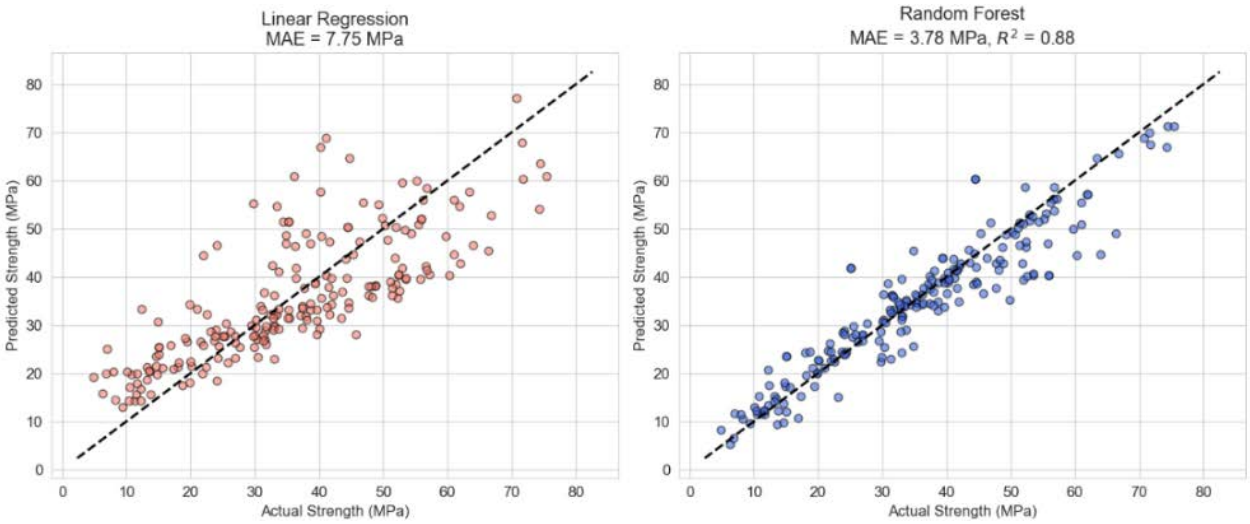

Fig 2. **Comparison of prediction accuracy between linear regression and Random Forest models**

Results. The aggregated performance metrics of the investigated models are summarized in Table 1.

Table 1.

Accuracy metrics of the studied models

Machine learning model	MAE, MPa
Linear regression	7.46
Empirical power-law formula	6.38
Polynomial regression	5,97
Neural network (2 × 128)	4,3
Neural network (7 × 256)	3,3
RandomForest	3,73

The Random Forest model achieved an MAE of 3.73 MPa, representing the best stable performance among the evaluated methods. Although the deep neural network with seven hidden layers achieved a slightly lower MAE of 3.3 MPa, this improvement came at the cost of significantly increased model complexity, computational demands, and reduced interpretability. The empirical power-law formula yielded an MAE of approximately 6.38 MPa, which is acceptable for rapid preliminary assessments.

Figure 2 illustrates a scatter plot comparing predicted and experimental compressive strength values for linear regression and the Random Forest model. The Random Forest predictions cluster more tightly around the ideal prediction line, particularly in the low-to-medium strength range, indicating higher accuracy and reduced systematic bias. In contrast, the linear regression model exhibits a clear trend of underestimating compressive strength, which visually confirms its limited capability to model non-linear material behavior.

Discussion. Feature importance analysis, conducted using the Random Forest model, enabled interpretation of the physical consistency of the learned relationships. Figure 3 presents the distribution of input feature contributions to the prediction of concrete compressive strength.

As shown in the figure, the most influential variables are concrete age and cement content, which fully aligns with established principles of concrete technology and cement hydration theory. This result indicates that the model does not merely memorize the training data but instead captures fundamental mechanisms governing strength development.

Compared to artificial neural networks, which typically operate as black-box models, the Random Forest approach offers a higher degree of transparency by allowing direct assessment of the relative contribution of each input variable. This interpretability is particularly valuable for engineering

applications, where trust in model predictions and consistency with physical knowledge are essential for practical adoption.

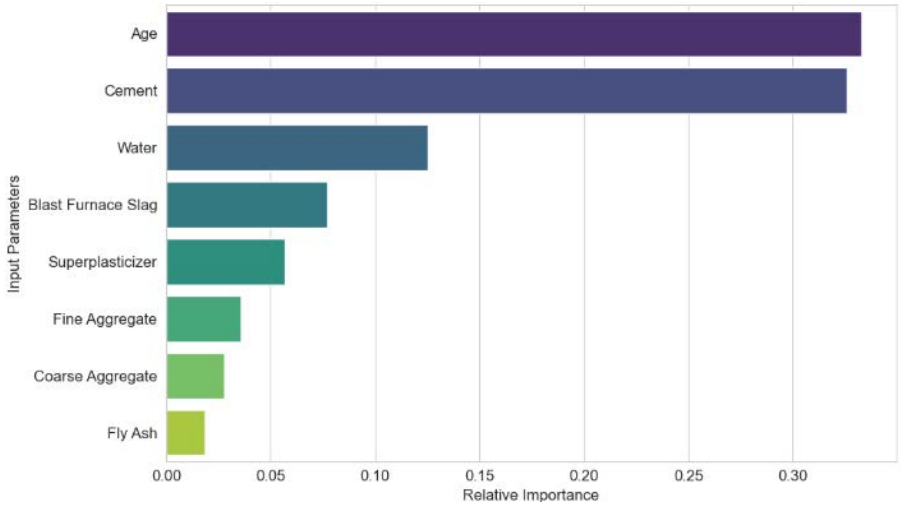

Fig 3. **Feature importance analysis obtained from the Random Forest model**

Conclusions. This study demonstrates that simple linear regression models are insufficient for accurate prediction of concrete compressive strength, yielding mean absolute errors exceeding 7 MPa. Such performance limitations confirm that linear assumptions fail to capture the complex and non-linear relationships inherent in concrete mixture composition and curing processes.

Among the evaluated approaches, ensemble learning methods exhibited the most favorable balance between predictive accuracy and model robustness, achieving an MAE of 3.73 MPa. The Random Forest model consistently outperformed baseline linear and polynomial regressions and approached the predictive performance of deeper neural network architectures, while offering improved stability and reduced sensitivity to hyperparameter selection. In addition to data-driven prediction, an empirical power-law model was derived through linearization of a multiplicative regression formulation. Despite its simplicity, the resulting formula enables approximate manual estimation of compressive strength with an average error of approximately 15%. This level of accuracy may be sufficient for preliminary assessments and rapid decision-making in construction site conditions where computational tools are unavailable.

Feature importance analysis further confirmed the physical plausibility of the learned models. Concrete age and total binder content were identified as the dominant factors influencing compressive strength, in agreement with established material science principles. This alignment between model behavior and domain knowledge supports the interpretability and practical relevance of the proposed approach.

Future Research Directions. While the proposed models demonstrate high predictive accuracy on standardized benchmark data, their industrial application requires adaptation to regional variability in raw materials and production conditions. Future research will focus on incorporating aggregate morphology descriptors (e.g., washed versus unwashed fine aggregates and particle shape characteristics) to improve sensitivity to local material properties. In addition, transfer learning strategies will be investigated to recalibrate the base model using limited historical production data from individual concrete plants, enabling practical deployment without extensive additional laboratory testing.

List of references:

1. Spjuth O., Carrión Brännström R., Carlsson L., Gauraha N. Combining Prediction Intervals on Multi-Source Non-Disclosed Regression Datasets. PMLR. Conference contribution. 2019. Vol. 105. P. 164–177. URL: https://proceedings.mlr.press/v105/spjuth19a.html
2. Geifman Y., El-Yaniv R. SelectiveNet: A Deep Neural Network with an Integrated Reject Option. PMLR. Conference contribution. 2019. Vol. 97. P. 2151–2160. URL: https://proceedings.mlr.press/v97/geifman19a.html
3. Mistry M., Letsios D., Krennrich G., Lee R. M., Misener R. Mixed-Integer Convex Nonlinear Optimization with Gradient-Boosted Trees Embedded. INFORMS Journal on Computing. Journal contribution. 2021. Vol. 33(3). P. 1103–1119. doi:10.48550/arXiv.1803.00952
4. Žegklitz J., Pošík P. Learning Linear Feature Space Transformations in Symbolic Regression. arXiv. Preprint. 2017. URL: https://arxiv.org/abs/1704.05134

5. Wang W., Zhong Y., Liao G., Ding Q., Zhang T., Li X. Prediction of Compressive Strength of Concrete Specimens Based on Interpretable Machine Learning. Materials. Journal contribution. 2024. Vol. 17(15). Article 3661. doi:10.3390/ma17153661
6. Zhang Y., Ren W., Chen Y. et al. Predicting the Compressive Strength of High-Performance Concrete Using an Interpretable Machine Learning Model. Scientific Reports. Journal contribution. 2024. Vol. 14. Article 28346. doi:10.1038/s41598-024-79502-z
7. Miao Q., Gao Z., Zhu K., Guo Z., Sun Q., Zhou L. Interpretable Machine Learning Model for Compressive Strength Prediction of Self-Compacting Concrete with Recycled Concrete Aggregates and SCMs. Journal of Building Engineering. Journal contribution. 2025. Vol. 108. Article 112965. doi:10.1016/j.jobe.2025.112965
8. More S., Kambekar A. Performance Evaluation of Compressive Strength of Concrete Using Different Machine Learning Algorithms. Challenge Journal of Concrete Research Letters. Journal contribution. 2025. Vol. 16(2). P. 60–68. doi:10.20528/cjcrl.2025.02.002
9. Yeh I.-C. Concrete Compressive Strength. UCI Machine Learning Repository. Dataset. 1998. doi:10.24432/C5PK67.

Reviewer: Pasternak Yaroslav, Doctor of Physical and Mathematical Sciences, Professor of the Department of Computer Science and Cybersecurity at Lesya Ukrainka Volyn National University.