International Journal of Applied Mathematics
Volume 38 No. 9s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

ASSESSMENT OF DISTANCE-BASED AND KERNEL-INSPIRED REGRESSION MODELS FOR ESTIMATING THE COMPRESSIVE STRENGTH OF HIGH-STRENGTH CONCRETE

Srajan Kumar Choubey[1]*, Rajesh Bhargava[2], Aruna Rawat[3]

[1*]srajanchoubey9@gmail.com (0009-0009-7346-4328), Research Scholar, Department of Civil Engineering, University Institute of Technology, RGPV, Bhopal, India
[2]rajeshbhargava.bpl@gmail.com, RGPV, Bhopal, India, (0009-0005-5008-9626)
[3]augupta2001@gmail.com (0000-0001-7279-8830), Department of Civil Engineering, University Institute of Technology, RGPV, Bhopal, India

Abstract

Accurate estimation of the compressive strength (CS) of high-strength concrete (HSC) is essential for structural design, quality control, and performance assessment. While machine learning (ML) techniques are widely applied for this purpose, many existing studies emphasise optimisation-driven accuracy improvement, often overlooking model sensitivity, robustness, and error behaviour. This study presents an assessment-oriented investigation of selected regression-based approaches for estimating the compressive strength of HSC. Three representative regression models are examined: K-Nearest Neighbours (KNN), polynomial regression (PR), and Gaussian Process Regression (GPR). A controlled sensitivity analysis is conducted by systematically varying key model hyperparameters, and model stability is evaluated through repeated train–test experiments. Performance is assessed using standard regression metrics, while residual distribution analysis is employed to examine robustness and error dispersion. The results indicate that distance-based and kernel-inspired regression models exhibit distinct behavioural characteristics. KNN regression demonstrates effective local learning when appropriately parameterised, whereas PR shows pronounced sensitivity to model complexity and reduced robustness at higher polynomial degrees. GPR provides comparatively stable performance, with improved residual concentration, highlighting its suitability for assessment-oriented applications. Rather than identifying a single optimal model, the study emphasises the importance of understanding model behaviour under controlled variation and repeated evaluation. The proposed assessment framework offers useful insights into the reliable application of regression-based ML models for HSC strength estimation and supports informed model selection in engineering practice.

Keywords: High-strength concrete estimation; regression models; K-nearest neighbors; Gaussian process regression; sensitivity analysis; residual analysis

1. Introduction

Concrete remains the most utilised material in construction, and its CS is a governing parameter in structural design, safety assessment, and durability evaluation. Accurate estimation of CS is therefore critical not only at the design stage but also during production, quality control, and performance assessment of concrete structures. Traditional empirical and semi-empirical formulations, although simple and interpretable, often fails to capture the complex and nonlinear interactions among mix constituents, curing conditions, and material heterogeneity, particularly in the case of HSC (Yeh, 1998; Young et al., 2018; Imran et al., 2023).

To overcome these limitations, data-driven and ML based approaches have been increasingly adopted for estimating concrete CS (Sobhani et al., 2009; Liu, 2022). The ML based models are capable of approximating highly nonlinear relationships without explicit assumptions regarding functional form (Moradi et al., 2022; Terlumun et al., 2024). A wide range of algorithms utilised, including regression-based methods, support vector machines, ensemble learners, and hybrid

Received: August 11, 2025

Volume 38 No. 9s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

models, highlighting the potential to improve predictive accuracy over conventional approaches (Chou et al., 2014; Erdal, 2013; Güçlüer et al., 2021). The performance of machine learning models for concrete strength prediction is strongly influenced by data characteristics, model complexity, and hyperparameter selection, rather than the choice of a single 'best' algorithm (Behnood and Golafshani, 2018). The ML based models face emphasising challenges related to generalisation, robustness, and reproducibility across datasets (Hadzima-Nyarko et al., 2020). Several studies have focused on benchmarking multiple algorithms across diverse concrete systems, including eco-concretes, high-strength concretes, and sustainable material formulations, with particular attention to robustness and consistency under repeated evaluation (Alyousef et al., 2023; Paudel et al., 2023; Sah et al., 2024; Bentegri et al., 2025). Parallel efforts have explored advanced regression frameworks and probabilistic learning paradigms to better characterise uncertainty and error dispersion in concrete strength prediction (Li et al., 2023).

Among these approaches, kernel-based models such as GPR have attracted renewed attention due to the solid theoretical foundation and ability to provide uncertainty-aware predictions. Recent investigations have demonstrated that GPR can offer competitive accuracy while maintaining improved stability and interpretability when compared to highly complex black-box models (Angiulli et al., 2024). In parallel, distance-based methods such as K-Nearest Neighbours (KNN) continue to serve as transparent and effective baselines for concrete strength estimation, particularly when local data patterns are informative (Asteris et al., 2016).

The study aims to: (i) evaluate the sensitivity of KNN, PR, and GPR models to the governing hyperparameters; (ii) analyse model robustness and error characteristics through residual distribution assessment across repeated runs; and (iii) provide a balanced comparative evaluation based on both average performance and stability indicators. By adopting an assessment-focused framework rather than an optimisation-centric perspective, this work seeks to contribute practical insights into the reliable application of regression-based machine learning models for HSC strength estimation.

2. Literature Review

Related work done in the field are discussed below:

- Zou et al. (2024) proposed an enhanced GPR framework for UHPC strength estimation. Addressing noise and limited-data instability using data-conditioning / signal-processing measures to stabilise GP learning. The key value is probabilistic prediction and improved robustness, not just marginal accuracy gains. This supports using kernel-based regression (GPR) for assessment-oriented strength estimation where reliability matters.
- Li et al. (2024) evaluated multiple ML models for UHPC and applied metaheuristic optimisation for hyperparameter tuning. The approach introduces higher methodological complexity and can reduce transparency. Based on contrasting optimisation gains vs the simplicity and interpretability of assessment-focused models.
- Tak et al. (2025) conducted large-scale benchmarking of many ML algorithms for concrete strength estimation. The evaluation emphasised repeated runs and statistical consistency, highlighting stability alongside mean performance. Showed that simpler regression models can remain competitive under fair, controlled comparison protocols.
- Kumar et al. (2025) studied compressive strength estimation for concrete with biomedical waste ash, emphasizing sustainable mixes. Compared regression/ensemble models and reported performance variability across repeated evaluations. A key insight is the influence of data variability and error dispersion on model reliability. Although the material differs from HSC, the methodology supports robustness-focused comparative assessment in the present study.

Received: August 11, 2025

Volume 38 No. 9s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

3. Materials and Methods

3.1 Dataset description

The present study utilises a publicly available dataset comprising experimental results of HSC mixtures. The dataset consists of a limited yet well-curated number of samples representative of high-performance concrete used in structural applications. Each sample is defined by a set of primary mix constituents, including cement, fine aggregate, coarse aggregate, water, and chemical admixture content, which are known to influence the mechanical response of concrete. The output variable considered in this study is the compressive strength, expressed in MPa. The strength values predominantly fall within the HSC range, confirming the suitability of the dataset for the intended assessment. No additional statistical summarisation is presented here in order to maintain methodological distinction from previous studies based on the same dataset.

3.2 Data preprocessing

Prior to model development, the dataset examined for completeness and consistency. Samples containing missing values in either input or output variables, excluded to avoid introducing uncertainty during model training. As the considered regression models are sensitive to feature scales, all input variables normalised using standard scaling, ensuring zero mean and unit variance.
The dataset randomly divided into training and testing subsets using an 80-20 split ratio. To ensure robustness and reduce the influence of a single favourable data partition, the entire training-testing process repeated ten times using different random seeds. Performance metrics, subsequently averaged over these independent runs, and the corresponding standard deviations reported as indicators of model stability.

3.3 Regression models considered

3.3.1 K-Nearest Neighbors (KNN) regression

K-Nearest Neighbors (KNN) regression is a distance-based, non-parametric learning approach that estimates the target value of a test sample by averaging the responses of its nearest neighbours in the feature space. The primary advantage of KNN lies in its ability to capture local patterns without assuming a predefined functional form. However, the model performance is strongly influenced by the choice of the neighbourhood size, denoted by k. Smaller k values tend to model local variations effectively, whereas larger k values may lead to excessive smoothing and loss of important information. In this study, ‘k’ varied over odd values from 1 to 25 (k = 1, 3, 5, ..., 25) to assess the sensitivity of KNN regression to neighbourhood size.

3.3.2 Polynomial regression

PR extends conventional linear regression by incorporating polynomial terms of the input variables, thereby enabling the modelling of nonlinear relationships. While lower-degree polynomials can improve flexibility without substantial instability, higher-degree polynomials often suffer from numerical ill-conditioning and overfitting, particularly when applied to datasets of limited size. To investigate this behaviour, PR models with degrees 1 to 4, evaluated in this study. The analysis focuses on understanding how global parametric complexity influences estimation stability rather than achieving optimal predictive performance.

3.3.3 Gaussian Process Regression (GPR)

GPR is a probabilistic, kernel-based regression technique that provides both predictive estimates and associated uncertainty measures. Unlike deterministic regression models, GPR defines a distribution over possible functions that fit the data, making it particularly attractive for modelling experimental datasets with inherent variability. The behaviour of a GPR model is governed by the choice of kernel function, which encodes assumptions regarding the smoothness and continuity of the underlying process.

Received: August 11, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

In this work, two commonly used kernel functions considered: the Radial Basis Function (RBF) kernel and the Matérn kernel with smoothness parameter $\nu = 1.5$. The RBF kernel assumes a highly smooth functional relationship, whereas the Matérn kernel allows for rougher response surfaces, which may be more representative of experimental concrete strength data. A categorical comparison of these kernel choices performed to assess their influence on model behaviour.

3.4 Sensitivity and stability assessment framework

The primary objective of this study is to assess the behavioural characteristics of different regression models rather than to perform exhaustive hyperparameter optimisation. Accordingly, a controlled sensitivity analysis conducted for each model by varying its key hyperparameters within reasonable and practically meaningful ranges. For KNN regression, k, systematically varied over k = 1, 3, 5, ..., 25, whereas PR model evaluated across degrees 1–4. For GPR, the sensitivity assessment, limited to a comparison between kernel functions (RBF and Matérn with $\nu = 1.5$), reflecting the categorical nature of this modelling choice.

3.5 Evaluation metrics

Model performance evaluation, using three widely accepted regression metrics: the coefficient of determination (R^2), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). The coefficient of determination provides insight into the proportion of variance explained by the model, while RMSE and MAE quantify the magnitude of prediction errors. RMSE penalises larger errors more strongly, whereas MAE offers a more interpretable measure of average deviation. Together, these metrics enable a balanced assessment of both accuracy and robustness.

4. Results and Discussion

4.1 Sensitivity analysis of regression models

The sensitivity analysis provides insight into how variations in key model hyperparameters influence estimation performance for HSC. Figure 1 represents the training-only cross-validated RMSE obtained for each model as a function of its governing parameters.

For K-Nearest Neighbors (KNN) regression, a clear dependence on the neighborhood size is observed. Lower values of k yield smaller prediction errors, indicating that local learning is effective in capturing the heterogeneous nature of HSC mixtures. As the number of neighbors increases, the model exhibits progressively higher error due to excessive averaging, which smooths out important local variations in the data. This behaviour reflects the classical bias–variance trade-off associated with distance-based models.

PR demonstrates a contrasting pattern. While low-degree polynomials exhibit reasonable performance, increasing the polynomial degree leads to a rapid escalation in error. This trend highlights the numerical instability and over-parameterisation associated with higher-order global PR models, particularly when applied to datasets of limited size. The results indicate that it is highly sensitive to model complexity and lacks robustness beyond simple configurations.

For GPR, a categorical comparison between kernel functions reveals noticeable differences in performance. The Matérn kernel consistently produces lower error than the Radial Basis Function (RBF) kernel. This observation suggests that the Matérn kernel, which permits less smooth functional behaviour, is better suited to modelling the irregular and experimentally derived response of HSC strength.

Received: August 11, 2025

Volume 38 No. 9s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

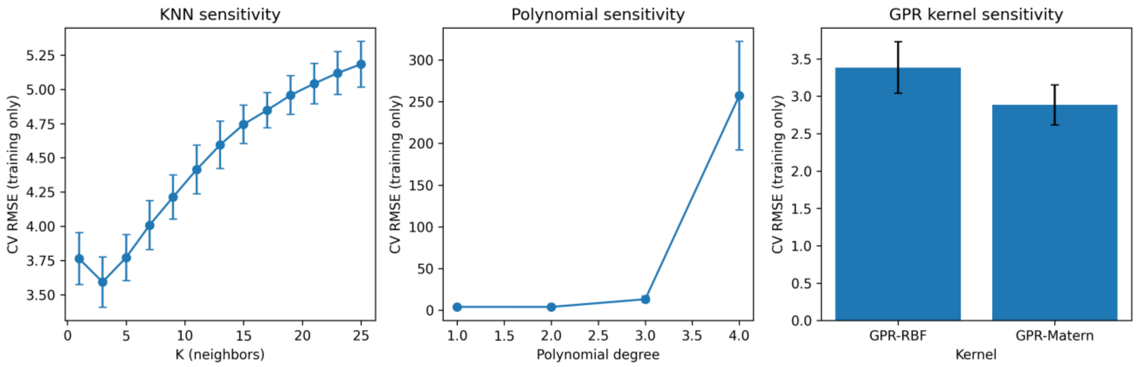

Figure 1. Sensitivity assessment of regression models for HSC.

4.2 Residual behaviour and robustness analysis

Beyond average performance metrics, the distribution of residuals provides valuable insight into model robustness and reliability. Figure 2 illustrates the pooled test-set residual distributions obtained across ten independent runs for all evaluated models.

KNN regression exhibits a relatively compact and symmetric residual distribution centred near zero, indicating stable estimation with limited systematic bias. However, moderate dispersion is observed, reflecting sensitivity to local data density and neighborhood selection.

PR displays a markedly wider residual spread with pronounced tails, signifying poor robustness and a higher likelihood of extreme estimation errors. This behaviour is consistent with the instability observed in the sensitivity analysis and underscores the limitations of global parametric models for complex material-response relationships.

Both GPR variants demonstrate improved residual concentration compared to PR. In particular, the Matérn kernel yields the most compact residual distribution with reduced dispersion, suggesting superior generalisation capability. The RBF kernel, while competitive, shows slightly broader residual spread, likely due to its strong smoothness assumptions.

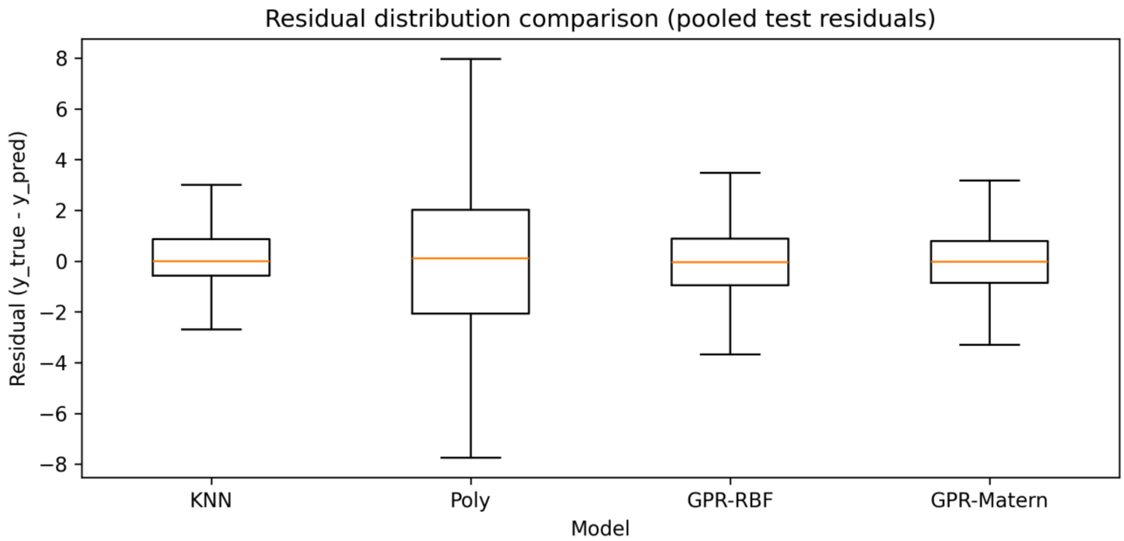

Figure 2. Residual distribution comparison of regression models.

4.3 Comparative performance summary

A quantitative comparison of model performance is summarised in Table 1, where the coefficient of determination (R^2), RMSE, and MAE are reported as mean ± standard deviation across ten independent runs. This presentation enables a balanced assessment of both accuracy and stability.

Volume 38 No. 9s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Among the evaluated models, GPR with the Matérn kernel achieved the best average accuracy (highest mean R^2 and lowest mean RMSE/MAE), but its RMSE and R^2 showed noticeable run-to-run variation. PR produced relatively stable R^2 and RMSE values (lower standard deviations), yet its higher MAE and heavy-tailed residuals indicate less reliable error behaviour. KNN provided a transparent baseline with competitive mean performance and moderate variability. Overall, the combined metric-and-residual evidence suggests that kernel-based probabilistic modelling improves mean accuracy, while stability considerations remain important when selecting models for practical HSC strength estimation.

Model	**R2 (mean ± std)**	**RMSE (mean ± std)**	**MAE (mean ± std)**
KNN	0.9010 ± 0.0495	3.0716 ± 0.8392	1.7784 ± 0.3123
PR	0.9118 ± 0.0242	2.9621 ± 0.4905	2.3851 ± 0.3805
GPR-RBF	0.9053 ± 0.0763	2.9307 ± 1.1350	1.7015 ± 0.3727
GPR-Matern	0.9334 ± 0.0595	2.4325 ± 1.0213	1.4706 ± 0.3097

Table 1. Comparative performance of regression models for HSC (mean ± standard deviation over ten runs).

5. Conclusions

This study assessed three representative regression families, distance-based KNN, global parametric PR, and kernel-based probabilistic GPR for estimating the compressive strength of HSC. The analysis combined controlled sensitivity experiments with pooled residual diagnostics and repeated train-test evaluations to capture both accuracy and stability.

The sensitivity results showed that KNN performs best at small neighbourhood sizes, whereas increasing k increases error due to excessive averaging. Polynomial regression remained viable at low degrees but degraded rapidly at higher degrees, reflecting numerical instability and over-parameterisation. For GPR, the Matérn kernel outperformed the RBF kernel in cross-validated error, suggesting that a less-smooth prior is better suited to the experimentally derived response of HSC.

Residual distributions reinforced these trends: PR exhibited wider spreads and more extreme errors, while GPR particularly with the Matérn kernel produced more concentrated residuals around zero. Across repeated runs, GPR-Matérn achieved the strongest mean performance, although its variability indicates sensitivity to the random split and highlights the value of repeated evaluation.

6. Limitations and Future Scope

Despite the insights gained, the present study is subject to limitations. The analysis is based on a single HSC dataset with a limited number of samples, which may restrict the generalisability of the conclusions. In addition, the study avoided extensive hyperparameter optimisation and feature engineering in order to preserve methodological simplicity and fairness, which may limit the attainable predictive accuracy.

Future research could extend the present framework by incorporating larger and more diverse datasets covering a wider range of concrete types and curing conditions. The integration of uncertainty quantification, physics-informed constraints, or domain-driven feature transformations may further enhance model interpretability and reliability. Additionally, combining the present assessment approach with experimental validation or non-destructive testing data could provide valuable insight into real-world structural applications.

Volume 38 No. 9s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

7. References

[1] Alyousef, R., Rehman, M. F., Khan, M., Fawad, M., Khan, A. U., Hassan, A. M., & Ghamry, N. A. (2023). Machine learning-driven predictive models for compressive strength of steel fiber reinforced concrete subjected to high temperatures. *Case Studies in Construction Materials*, *19*, e02418. https://doi.org/10.1016/j.cscm.2023.e02418

[2] Angiulli, G., Calcagno, S., La Foresta, F., & Versaci, M. (2024). Concrete Compressive strength Prediction using Combined Non-Destructive methods: a calibration procedure using preexisting conversion models based on Gaussian process regression. *Journal of Composites Science*, *8*(8), 300. https://doi.org/10.3390/jcs8080300

[3] Asteris, P., Kolovos, K., Douvika, M., & Roinos, K. (2016). Prediction of self-compacting concrete strength using artificial neural networks. *European Journal of Environmental and Civil Engineering*, *20*(sup1), s102–s122. https://doi.org/10.1080/19648189.2016.1246693

[4] Behnood, A., & Golafshani, E. M. (2018). Predicting the compressive strength of silica fume concrete using hybrid artificial neural network with multi-objective grey wolves. *Journal of Cleaner Production*, *202*, 54–64. https://doi.org/10.1016/j.jclepro.2018.08.065

[5] Bentegri, H., Rabehi, M., Kherfane, S., Nahool, T. A., Rabehi, A., Guermoui, M., Alhussan, A. A., Khafaga, D. S., Eid, M. M., & El-Kenawy, E. M. (2025). Assessment of compressive strength of eco-concrete reinforced using machine learning tools. *Scientific Reports*, *15*(1), 5017. https://doi.org/10.1038/s41598-025-89530-y

[6] Chou, J., Tsai, C., Pham, A., & Lu, Y. (2014). Machine learning in concrete strength simulations: Multi-nation data analytics. *Construction and Building Materials*, *73*, 771–780. https://doi.org/10.1016/j.conbuildmat.2014.09.054

[7] Erdal, H. I. (2013). Two-level and hybrid ensembles of decision trees for high performance concrete compressive strength prediction. *Engineering Applications of Artificial Intelligence*, *26*(7), 1689–1697. https://doi.org/10.1016/j.engappai.2013.03.014

[8] Güçlüer, K., Özbeyaz, A., Göymen, S., & Günaydın, O. (2021). A comparative investigation using machine learning methods for concrete compressive strength estimation. *Materials Today Communications*, *27*, 102278. https://doi.org/10.1016/j.mtcomm.2021.102278

[9] Hadzima-Nyarko, M., Nyarko, E. K., Lu, H., & Zhu, S. (2020). Machine learning approaches for estimation of compressive strength of concrete. *The European Physical Journal Plus*, *135*(8). https://doi.org/10.1140/epjp/s13360-020-00703-2

[10] Imran, M., Khushnood, R. A., & Fawad, M. (2023). A hybrid data-driven and metaheuristic optimization approach for the compressive strength prediction of high-performance concrete. *Case Studies in Construction Materials*, *18*, e01890. https://doi.org/10.1016/j.cscm.2023.e01890

[11] Kumar, R., Karthik, S., Kumar, A., Tantri, A., Shahaji, N., & Sathvik, S. (2025). Machine learning approach for predicting the compressive strength of biomedical waste ash in concrete: a sustainability approach. *Discover Materials*, *5*(1). https://doi.org/10.1007/s43939-025-00223-9

[12] Li, D., Tang, Z., Kang, Q., Zhang, X., & Li, Y. (2023). Machine Learning-Based Method for predicting compressive strength of concrete. *Processes*, *11*(2), 390. https://doi.org/10.3390/pr11020390

[13] Li, Y., Yang, X., Ren, C., Wang, L., & Ning, X. (2024). Predicting the compressive strength of Ultra-High-Performance concrete based on machine learning optimized by Meta-Heuristic algorithm. *Buildings*, *14*(5), 1209. https://doi.org/10.3390/buildings14051209

[14] Liu, Y. (2022). High-Performance concrete strength prediction based on machine learning. *Computational Intelligence and Neuroscience*, *2022*, 1–7. https://doi.org/10.1155/2022/5802217

[15] Moradi, N., Tavana, M. H., Habibi, M. R., Amiri, M., Moradi, M. J., & Farhangi, V. (2022). Predicting the compressive strength of concrete containing binary supplementary cementitious

Received: August 11, 2025

International Journal of Applied Mathematics
Volume 38 No. 9s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

material using machine learning approach. *Materials*, *15*(15), 5336. https://doi.org/10.3390/ma15155336

[16] Paudel, S., Pudasaini, A., Shrestha, R. K., & Kharel, E. (2023). Compressive strength of concrete material using machine learning techniques. *Cleaner Engineering and Technology*, *15*, 100661. https://doi.org/10.1016/j.clet.2023.100661

[17] Sah, A. K., & Hong, Y. (2024). Performance comparison of machine learning models for concrete compressive strength prediction. *Materials*, *17*(9), 2075. https://doi.org/10.3390/ma17092075

[18] Sobhani, J., Najimi, M., Pourkhorshidi, A. R., & Parhizkar, T. (2009). Prediction of the compressive strength of no-slump concrete: A comparative study of regression, neural network and ANFIS models. *Construction and Building Materials*, *24*(5), 709–718. https://doi.org/10.1016/j.conbuildmat.2009.10.037

[19] Tak, M. S. N., Feng, Y., & Mahgoub, M. (2025). Advanced machine learning techniques for predicting concrete compressive strength. *Infrastructures*, *10*(2), 26. https://doi.org/10.3390/infrastructures10020026

[20] Terlumun, S., Onyia, M. E., & Okafor, F. O. (2024). Predicting the compressive strength of self-compacting concrete using artificial intelligence techniques: a review. *AI In Civil Engineering*, *3*(1). https://doi.org/10.1007/s43503-024-00029-3

[21] Yeh, I. (1998). Modeling of strength of high-performance concrete using artificial neural networks. *Cement and Concrete Research*, *28*(12), 1797–1808. https://doi.org/10.1016/s0008-8846(98)00165-3

[22] 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. https://doi.org/10.1016/j.cemconres.2018.09.006

[23] Zou, Z., Peng, B., Xie, L., & Song, S. (2024). Enhanced Gaussian Process model for predicting compressive strength of Ultra-High-Performance Concrete (UHPC). *Materials*, *17*(24), 6140. https://doi.org/10.3390/ma17246140