Fusion Journal of Engineering and Sciences

Article

Comparative Evaluation of Machine Learning Models for Predicting Compressive Strength of Concrete Made with Soap Factory Wastewater

Asad Wahab[1]*, Tausif Junaid Khan[1], Touqeer Ali Rind[1], Maaz Khan[1], Muhammad Faarid Shah[1], Muhammad Faisal Javed[1]

1 Department of Civil Engineering (DCvE), Ghulam Ishaq Khan (GIK) Institute of Engineering Sciences and Technology, Topi, Khyber Pakhtunkhwa; gcv2520@giki.edu.pk, tausif.junaid@giki.edu.pk, touqeer.ali@giki.edu.pk, gcv2463@giki.edu.pk, gcv2560@giki.edu.pk, arbabfaisal@giki.edu.pk

* Correspondence: gcv2520@giki.edu.pk

Abstract

The reuse of industrial wastewater in concrete production offers a sustainable alternative to reducing freshwater consumption in construction. This study investigates the prediction of compressive strength of concrete prepared with soap factory wastewater using machine learning techniques. A dataset containing key mix design parameters was analyzed using five machine learning models: Random Forest, Gradient Boosting, Decision Tree, k-Nearest Neighbors (kNN), and Linear Regression. Model performance was evaluated using MSE, RMSE, MAE, MAPE, and R^2 with 10-fold cross-validation. The results show that Random Forest achieved the best predictive performance ($R^2 = 0.883$), followed by Gradient Boosting ($R^2 = 0.872$), while kNN and Linear Regression showed lower accuracy. Feature importance analysis indicates that cement content, water–cement ratio, and curing age are the most influential parameters affecting compressive strength. These findings demonstrate the potential of machine learning for predicting the strength of wastewater-based concrete and supporting sustainable construction practices.

Keywords: Machine learning, Compressive strength, Gradient Boosting, Sustainable construction, Mix design optimization

1. Introduction

One of the most widely used construction materials is concrete due to its dura-bility, versatility and affordability. One of the most significant mechanical properties that dictate the performance of structure and service life is the compressive strength of concrete. A number of mix design parameters play a critical role in this strength such as cement content, water/ cement ratio, proportions of aggregates, dosage of superplasticizer, and curing age [1], [2]. Traditionally, determination of compressive strength is done in the form of experimental testing in the laboratory, which is time consuming, resource consuming and not always feasible in the context of quick mix optimization [3]. As the scale and complexity of the current construction projects continue to increase, there is also an increased

need of efficient and reliable predictive techniques that are capable of estimating the strength of concrete without engaging in cumbersome laboratory experimentation [4].

Over the last several years, machine learning (ML) techniques have received a lot of attention because of their capabilities to represent rather complicated and highly nonlinear correlations between concrete mix components and compressive strength [5]. It has been established in previous studies that existing statistical and linear regression models could not fully model the nonlinear behavior of cement-based materials [6]. To alleviate these limitations, artificial intelligence-related methods like artificial neural network (ANN), support vector machine (SVM), and decision-tree algorithms have been highly studied. Yeh [7] indicated that ANN models are much better predictors of the compressive strength of concrete compared to linear regression models, whereas Chou et al. [8] revealed that SVM performs better in the prediction of strengths.

Random Forest and Gradient Boosting algorithms are ensemble learning algorithms that have displayed better predictive behavior because of their capabilities to decrease over-fitting and enhance generalization [9], [10]. The method of the Random Forest algorithm proposed by Breiman [9] is a randomized and trained set of decision trees that are integrated to increase robustness. Gradient Boosting as a model suggested by Friedman [10], creates sequence of models to correct the past errors in prediction hence it is especially powerful in regression models. Recent studies have used these ensemble methods to model concrete compressive strength and have reported a greater coefficient of determinism (R2) and a smaller error measure than the non-ensembled models like Decision Trees and Linear Regression [11], [12]. The above conclusions indicate that ensemble-based ML models can be successfully used when forecasting the properties of concrete that depend on several interacting variables.

In addition to the development of predictive models, the issue of sustainability has triggered the desire to explore the nature of alternative resources and greener concrete production. The construction sector is a major consumer of freshwater which puts a lot of pressure on the environment, especially in areas that experience water shortage [13]. As such, reuse of factual wastewater in the mixing of concrete has been suggested as a sustainable approach. Experimental experiments have already been conducted to determine the impact of wastewater in textile, sugar, and chemical industries on concrete performance and have shown that with the proper usage of wastewater, compressive strength can be acceptable or even better [14], [15]. Nevertheless, the nature of wastewater can be affecting concrete properties in a complex way, emphasizing the importance of prediction modeling methods.

The wastewater of soap factories is full of surfactants, organic compounds, and dissolved salts that may affect the cement hydration and the microstructural formation. Experimentally, Zoyem et al. [16] showed that compressive strength can be similar using partial replacement of mixing water with soap factory wastewater when con-trolled conditions are considered. Although these findings are encouraging, most of the current research is based on ex-experimental trial-and-error techniques limiting the scalability and

implementation in practice. Furthermore, the use of machine learning methods in using wastewater-based concrete data is sparse in literature.

Hence, there is an obvious research gap in nominating machine learning models with experimental data of industrial wastewater-modified concrete. This research fills this gap by using and comparing five machine learning models, i.e., Gradient Boosting, Random Forest, Decision Tree, k-Nearest Neighbors (kNN), and Linear Regression to predict the compressive strength of concrete containing soap factory wastewater. These models are measured in terms of several statistical metrics that focus on the most efficient predictive method and contribute to the sustainable and data-driven concrete mix design.

2. Methodology

2.1 Data Collection

The dataset used in this study was obtained from a previously published experimental investigation that examined the influence of soap factory wastewater on the compressive strength of concrete (Zoyem et al.). In that study, several concrete mixtures were prepared by partially replacing conventional mixing water with industrial wastewater in different proportions. The experimental program documented key mix design parameters, including cement content, aggregate proportions, water–cement ratio, and curing age. Compressive strength tests were performed at different curing periods to evaluate the mechanical performance of the mixtures. The resulting dataset provides detailed information on mix composition and corresponding compressive strength values, which were used in this study to develop machine learning models for predicting concrete strength and assessing the potential use of industrial wastewater as a sustainable alternative in concrete production.

2.2 Data Preprocessing

Prior to model development, exploratory statistical analysis was carried out to understand the distribution and variability of the dataset variables. Descriptive statistics of the input parameters and target variable are presented in Table 1, including the mean, mode, median, dispersion, minimum, and maximum values.

The cement content shows a mean value of 276.50 kg/m^3 and a median of 266.00 kg/m^3, indicating moderate variability across the concrete mixtures. The water content has a mean value of 182.98 kg/m^3 and relatively low dispersion, suggesting that water content remained comparatively consistent among most samples.

The water–cement ratio shows noticeable variation across the dataset, reflecting differences in mix design proportions and its important role in strength development. Similarly, the superplasticizer content exhibits greater variability than most other input variables, indicating that admixture dosage was not uniform across all mixtures.

The coarse aggregate and fine aggregate contents remain comparatively stable, with mean values of 964.83 kg/m³ and 770.49 kg/m³, respectively. Their relatively low dispersion suggests that aggregate proportions were maintained within a narrower range than cementitious and admixture-related parameters.

The age of testing has a mean value of 44.06 days, while both the mode and median are 28 days, which corresponds to the standard curing duration commonly used for compressive strength evaluation. This indicates that most specimens were tested at 28 days, although additional curing ages were also included to capture longer-term strength development.

The compressive strength values show substantial variation, with a mean of 35.83 MPa, demonstrating that the dataset covers a broad range of concrete performance levels. As shown in Figure 1, the strength distribution is concentrated mainly in the medium-strength range, with fewer observations at very low and very high strength values. This spread provides sufficient variability for training and evaluating machine learning models.

Overall, the statistical analysis indicates that the dataset contains meaningful variation in the principal mix design variables, particularly cement content, water–cement ratio, superplasticizer dosage, and curing age, all of which are known to influence concrete compressive strength. Before model training, the input variables were standardized to ensure fair learning across features with different numerical scales.

A correlation heatmap was generated to visualize relationships among the dataset variables. The analysis in figure 2 indicates that cement content and curing age exhibit positive relationships with compressive strength, while the water–cement ratio tends to show a negative relationship. These trends align with established principles of concrete mix design and confirm that the dataset contains meaningful interactions suitable for machine learning modeling.

Table 1: Descriptive Statistics

Feature Statistics Name	Mean	Mode	Median	Dispersion	Minimum	Maximum
Cement	276.50	362.60	266.00	0.37	102.00	540.00
Water	182.98	192.00	185.70	0.12	121.75	247.00
water/cement	0.76	0.45	0.70	0.42	0.27	0.88
Super-plasticizer	6.41	0.00	6.70	0.90	0.00	32.20
Coarse Aggregate	964.83	932.00	966.80	0.09	708.00	1145.00
Fine Aggregate	770.49	594.00	777.50	0.10	594.00	992.60
Age of testing	44.06	28.00	28.00	1.37	1	365.00
Compressive strength	35.83	33.39	34.67	0.45	2.33	82.60

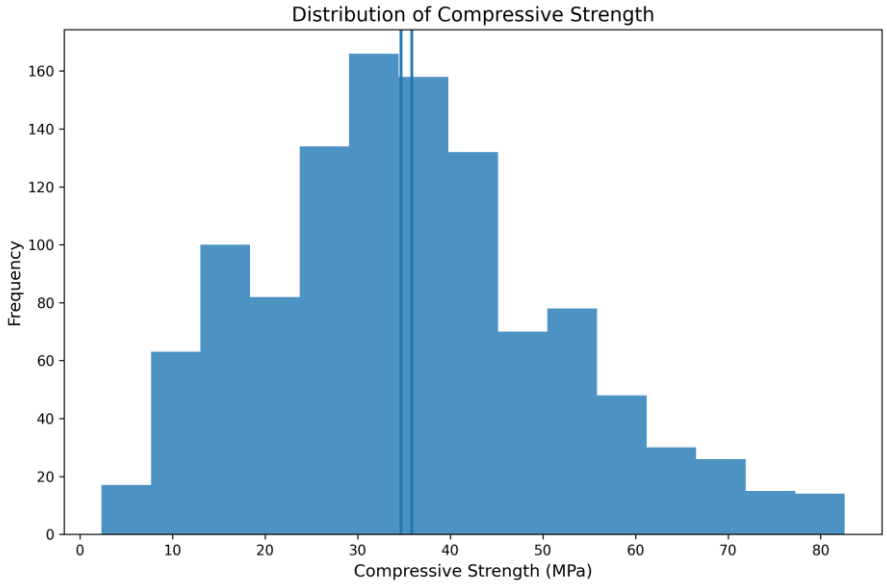

Figure 1: Distribution of compressive strength

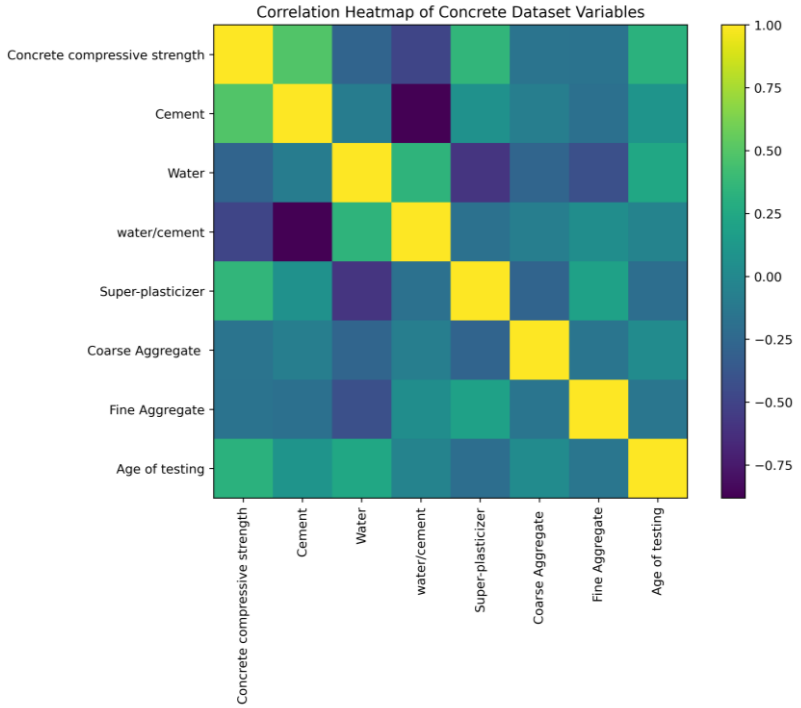

Figure 2: Correlation heatmap showing relationships between input variables and compressive strength.

2.3 Machine Learning Models

2.3.1 Gradient Boosting

Gradient Boosting is an ensemble learning technique that builds multiple weak models (typically decision trees) sequentially, where each new model corrects the errors of the previous ones. It minimizes a loss function using gradient descent [10].

Equation:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$$

2.3.2 Random Forest

Random Forest is an ensemble of multiple decision trees where each tree is trained on a random subset of the data. The final prediction is obtained by averaging (for regression) or majority voting (for classification) [9].

Equation:

$$\hat{y} = \frac{1}{N} \sum_{i=1}^{N} T_i(x)$$

2.3.3 Decision Tree

A Decision Tree splits the data into branches based on feature values, forming a tree-like structure. The split is determined by using criteria like Gini impurity or mean squared error (MSE) for regression [21].

Equation (for regression using MSE):

$$\text{Split} = \arg\min_{s} \sum_{i=1}^{n} (y_i - \bar{y}_s)^2$$

2.3.4 k-Nearest Neighbors (kNN)

kNN is a non-parametric model that predicts a value based on the average (for regression) or majority class (for classification) of its nearest neighbors [22].

Equation:

$$\hat{y} = \frac{1}{k} \sum_{i=1}^{k} y_i$$

2.3.5 Linear Regression

Linear Regression models the relationship between input features and output by fitting a linear equation to the data [6] [23].

Equation:

$$y = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \epsilon$$

2.4 Training and Testing

To evaluate the predictive performance of the machine learning models, a 10-fold cross-validation approach was employed. In this method, the dataset is divided into ten equal subsets. During each iteration, nine subsets are used for model training while the remaining subset is used for validation. This process is repeated ten times so that each subset is used once as a validation set. The performance metrics are then averaged across all folds to obtain a reliable estimate of model accuracy. This cross-validation strategy reduces the risk of overfitting and ensures that the models are evaluated on multiple data partitions, providing a more robust assessment of predictive capability. The procedure was applied consistently to all five models: Gradient Boosting, Random Forest, Decision Tree, k-Nearest Neighbors (kNN), and Linear Regression.

2. Results and Discussion

3.1 Feature Importance Analysis

Feature importance analysis was conducted to determine the relative influence of input variables on compressive strength prediction. The results in figure 3 indicate that cement content and water–cement ratio are the most influential parameters, followed by superplasticizer content and curing age. In contrast, fine and coarse aggregate contents show comparatively lower importance, indicating that variations in these parameters contribute less to the predictive performance of the machine learning models. These findings align with established concrete mix design principles where cement content and water–cement ratio are primary factors governing strength development.

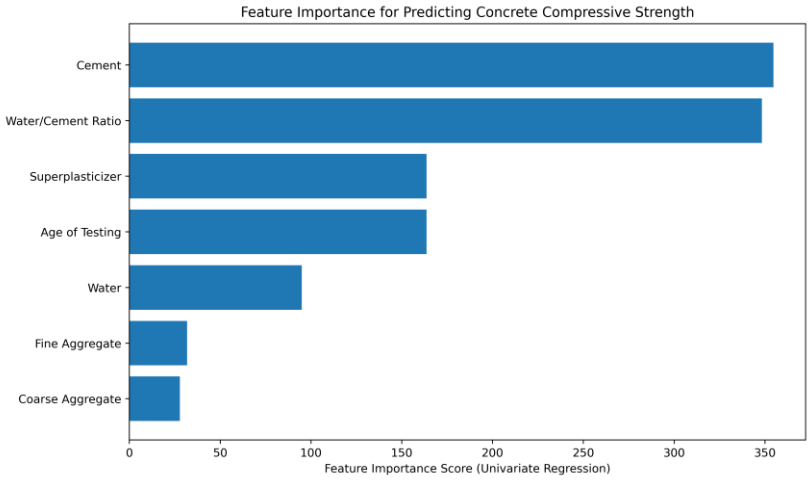

Figure 3: Feature importance ranking of input variables influencing compressive strength prediction based on Univariate Regression

3.2 Model Evaluation

The performance comparison of the five machine learning models used for predicting the compressive strength of concrete is presented in Table 2. Among the evaluated models, Random Forest achieved the best predictive performance, exhibiting the lowest error values with MSE = 30.417, RMSE = 5.515, MAE = 3.954, and MAPE = 0.140, along with the highest coefficient of determination ($R^2 = 0.883$). These results indicate that Random Forest is highly effective in capturing the complex nonlinear relationships between the concrete mix design parameters and compressive strength.

The Gradient Boosting model also demonstrated strong predictive capability, with slightly higher error values (MSE = 33.250, RMSE = 5.766, MAE = 4.240, MAPE = 0.147) and an R^2 value of 0.872, indicating that ensemble learning methods perform well for this prediction task. The Decision Tree model showed moderate performance with an R^2 value of 0.824, although its prediction errors were higher compared to the ensemble-based models.

In contrast, the k-Nearest Neighbors (kNN) and Linear Regression models showed comparatively lower prediction accuracy. The kNN model produced higher error values and an R^2 of 0.718, while Linear Regression performed the worst among all models, with the largest prediction errors and the lowest R^2 value of 0.553. This suggests that simple

linear models are not sufficient to capture the complex nonlinear interactions among the mix design variables affecting compressive strength.

Overall, the results demonstrate that ensemble learning models, particularly Random Forest and Gradient Boosting, outperform simpler machine learning approaches in predicting the compressive strength of wastewater-based concrete mixtures.

Table 2: Models' Performance

Model	MSE	RMSE	MAE	MAPE	R^2
Random Forest	30.417	5.515	3.954	0.140	0.883
Gradient Boosting	33.250	5.766	4.240	0.147	0.872
Decision Tree	45.550	6.749	4.720	0.166	0.824
k-Nearest Neighbors (kNN)	73.166	8.554	6.660	0.255	0.718
Linear Regression	115.831	10.762	8.400	0.318	0.553

3.2 Prediction Performance Analysis

Figure 4 presents the relationship between the actual compressive strength values and the values predicted by the five machine learning models. In each plot, the dashed diagonal line represents the ideal prediction line, where the predicted values perfectly match the actual compressive strength. The closer the data points lie to this line, the more accurate the model predictions.

Among the evaluated models, Random Forest shows the best prediction performance, with an R^2 value of 0.883. The predicted values closely follow the ideal line, indicating a strong agreement between actual and predicted compressive strength values. The Gradient Boosting model also demonstrates high prediction accuracy (R^2 = 0.872), with most data points concentrated near the reference line.

The Decision Tree model shows moderate prediction capability (R^2 = 0.824), although a slightly larger spread of points around the reference line indicates higher prediction errors compared to the ensemble models. In contrast, the k-Nearest Neighbors (kNN) model exhibits greater dispersion (R^2 = 0.718), suggesting weaker predictive performance. The Linear Regression model performs the worst (R^2 = 0.553), as many predicted values deviate from the ideal line, particularly at higher compressive strength levels.

Overall, the graphical analysis confirms that ensemble learning models, particularly Random Forest and Gradient Boosting, provide more accurate predictions of concrete compressive strength compared with simpler machine learning approaches.

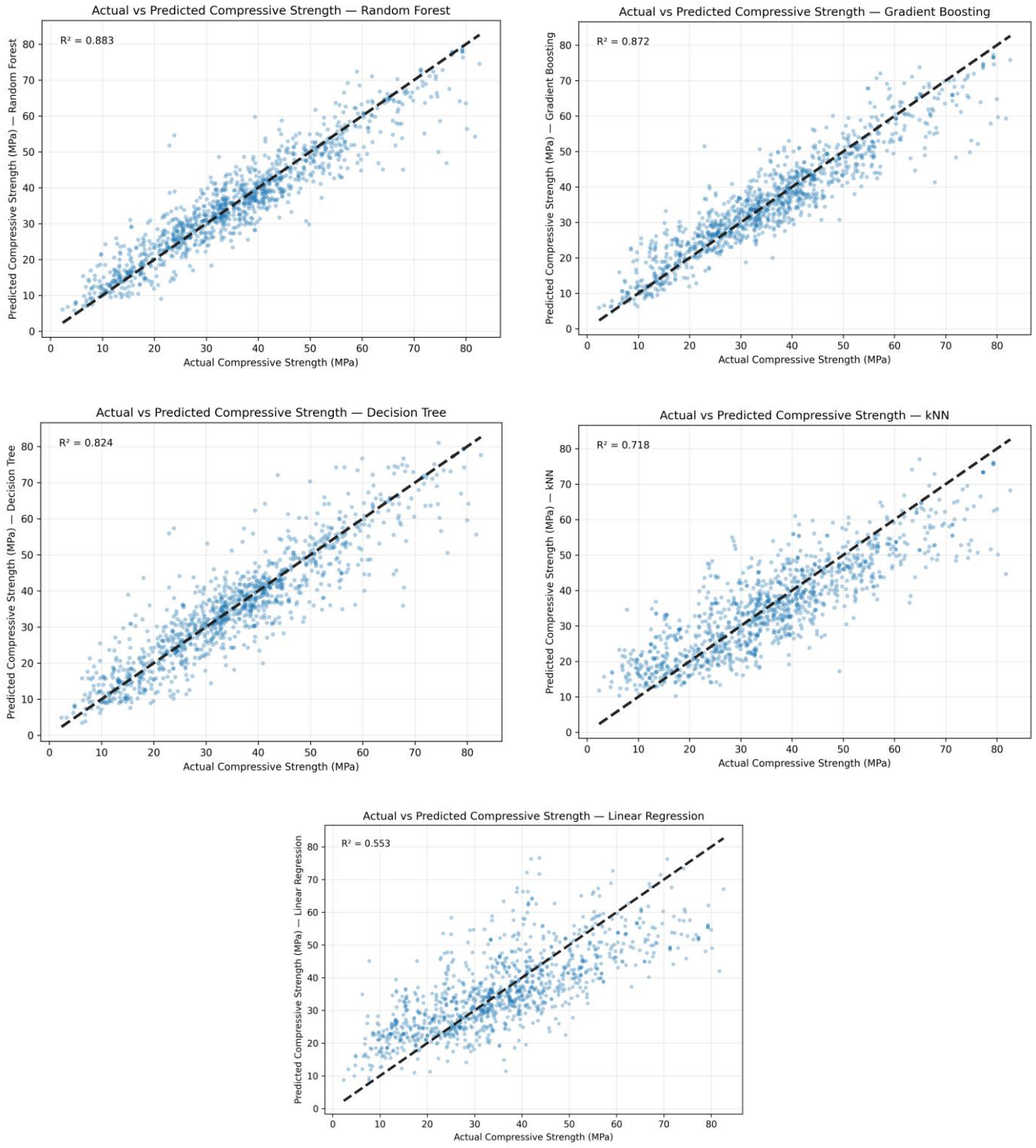

Figure 4: Actual vs predicted compressive strength for the evaluated machine learning models.

4. Conclusions

This study evaluated the ability of five machine learning models—Random Forest, Gradient Boosting, Decision Tree, k-Nearest Neighbors (kNN), and Linear Regression—to predict the compressive strength of concrete produced using soap factory wastewater. The models were assessed using performance metrics including MSE, RMSE, MAE, MAPE, and the coefficient of determination (R^2).

Among the evaluated models, Random Forest achieved the best predictive performance with the highest R^2 value (0.883) and the lowest prediction errors. Gradient Boosting also demonstrated strong performance, while Decision Tree showed moderate accuracy. In contrast, kNN and Linear Regression exhibited lower prediction accuracy, indicating limited ability to capture the nonlinear relationships between mix design parameters and compressive strength.

Overall, the results suggest that ensemble learning models, particularly Random Forest and Gradient Boosting, are effective tools for predicting compressive strength of wastewater-based concrete mixtures, supporting more efficient mix design and sustainable construction practices.

5. Patents

No Patents have resulted from the work reported in this manuscript.

Author Contributions:

The conceptualization, methodology development, data collection, formal analysis, model testing, and data validation, initial drafting and final pape preparation were carried out by Asad Wahab, Tauseef Junaid Khan, Touqeer Ali Rind. Maaz Khan and Muhammad Faarid Shah played a key role in the initial drafting, methodology development, and critical analysis of the results. Muhammad Faisal Javed supported the supervision, manuscript preparation, including reviewing and editing.

All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: Not applicable. This study did not involve humans or animals and therefore did not require ethical approval.

Informed Consent Statement: Not applicable. This study did not involve human participants.

Data Availability Statement: Data will be made available upon request.

Funding: This research received no external funding.

Acknowledgments: The authors gratefully acknowledge Ghulam Ishaq Khan Institute of Engineering Science and Technology (GIKI EST), for providing support and facilities in conducting this research.

Conflicts of Interest: The authors declare that there is no conflict of interest regarding this study and affirm that the work is original, without any form of plagiarism. All sources of information have been properly cited and acknowledged.

Abbreviations

The following abbreviations are used in this manuscript:

ANN	Artificial Neural Network
LR	Linear Regression
MSE	Mean Squared Error
RMSE	Root Mean Squared Error
MAE	Mean Absolute Error

References

[1] A. M. Neville, Properties of Concrete, 5th ed. London, UK: Pearson, 2011.

[2] S. B. Singh, P. Munjal, and N. Thammishetti, "Role of water–cement ratio on strength development of cement mortar," J. Build. Eng., vol. 4, pp. 94–100, 2015.

[3] S. Popovics, Strength and Related Properties of Concrete, New York, NY, USA: Wiley, 1998.

[4] S. C. Chapra and R. P. Canale, Numerical Methods for Engineers, 7th ed. New York, NY, USA: McGraw-Hill, 2015.

[5] M. Behnood and E. M. Golafshani, "Machine learning approaches for concrete strength prediction," Constr. Build. Mater., vol. 270, 2021.
[6] M. Zain and S. M. Abd, "Multiple regression model for compressive strength prediction of concrete," J. Appl. Sci., vol. 9, pp. 155–160, 2009.
[7] I.-C. Yeh, "Modeling of strength of high-performance concrete using artificial neural networks," Cem. Concr. Res., vol. 28, no. 12, pp. 1797–1808, 1998.
[8] J.-S. Chou, C.-F. Tsai, and Y.-H. Pham, "Predicting concrete strength using machine learning techniques," Autom. Constr., vol. 20, pp. 471–479, 2011.
[9] L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.
[10] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Ann. Stat., vol. 29, no. 5, pp. 1189–1232, 2001.
[11] H. Nguyen et al., "Concrete compressive strength prediction using ensemble learning," J. Build. Eng., vol. 32, 2020.
[12] M. Kamath et al., "Machine learning algorithms for high-performance concrete," J. Eng. Des. Technol., 2022.
[13] V. Harshit, A. K. Rizwan, and K. K. Iqbal, "Sustainable use of wastewater in concrete construction," J. Build. Eng., 2021.
[14] A. K. Padmini, K. Ramamurthy, and M. S. Mathews, "Influence of wastewater on properties of concrete," Build. Environ., vol. 44, pp. 1761–1767, 2009.
[15] K. S. Al-Jabri et al., "Reuse of industrial wastewater in concrete," Cem. Concr. Compos., vol. 33, pp. 603–608, 2011.
[16] Z. G. Mathurin et al., "Prediction of compressive strength of concrete made with soap factory wastewater using machine learning," Model. Earth Syst. Environ., vol. 8, pp. 5625–5638, 2022.
[17] ASTM C39/C39M-21, Standard Test Method for Compressive Strength of Cylindrical Concrete Specimens, ASTM International, 2021.
[18] H. Chen et al., "Predicting compressive strength of cement-based materials," PLoS One, 2018.
[19] L. K. A. Sear et al., "Abrams' law and water–cement ratio," Constr. Build. Mater., vol. 10, pp. 221–226, 1996.
[20] F. Khademi and K. Behfarnia, "ANN and regression models for concrete strength prediction," Constr. Build. Mater., 2016.
[21] L. Rokach and O. Maimon, Data Mining with Decision Trees, Singapore: World Scientific, 2014.
[22] T. Cover and P. Hart, "Nearest neighbor pattern classification," IEEE Trans. Inf. Theory, vol. 13, no. 1, pp. 21–27, 1967.
[23] A. Al-Shamiri et al., "Modeling compressive strength of high-strength concrete," Constr. Build. Mater., vol. 208, pp. 204–219, 2019.