JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Predicting Compressive Strength of Concrete using Machine Learning Algorithm: A Comparative Analysis

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

[1]Department of Civil Engineering, University Institute of Technology, RGPV, Bhopal, India (Orcid ID: https://orcid.org/0009-0009-7346-4328)
[2] RGPV, Bhopal, India (Orcid ID: https://orcid.org/0009-0005-5008-9626)
[3] Department of Civil Engineering, University Institute of Technology, RGPV, Bhopal, India (Orcid ID: https://orcid.org/0000-0001-7279-8830)

Abstract: This study explores the application of various machine learning algorithms in predicting the compressive strength of concrete. The models used for analysis include Decision Tree, Random Forest, Bagging, Gradient Boosting, and Neural Networks. A comparative evaluation is conducted based on key performance metrics such as R^2, Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). The results demonstrate that Gradient Boosting outperforms other models, achieving an R^2 of 0.898 and the lowest MSE. The study further discusses the feature importance of concrete mix constituents and proposes potential improvements for predictive modeling for compressive strength of concrete.

Keywords: Gradient Boosting, Compressive Strength, Concrete Strength Prediction, Machine Learning in Concrete, Predictive Modeling

1. Introduction

The performance of concrete's durability and structural integrity depend largely on its compressive strength. The conventional laboratory procedures for testing deliver accurate results but they create delays and enhance construction costs which slow down the development. Artificial intelligence (AI) and machine learning (ML) technologies have appeared on the scene as an interesting improvement opportunity [1]. The analysis of expansive concrete mix databases combined with their associated strengths allows ML models to perform efficient strength predictions. Real-time decision optimized mix design solutions together with improved quality control, decreased costs and sustainability benefits characterize the construction industry's operations through this technology..

The importance of predicting concrete's compressive strength is vital for structural safety, preventing failures and optimizing material use to minimize waste and construction costs. The accurate prediction also accelerates project planning and promotes sustainability by reducing cement consumption and carbon emissions.

Accurately predicting compressive strength is challenging due to material variability and the critical influence of the water-cement ratio. Variations in the quality and composition of concrete mix components introduce uncertainty, while even slight changes in the water-cement ratio can significantly impact strength development [2]. Curing conditions significantly influence concrete's final strength. Temperature, humidity, and curing duration affect hydration, impacting strength development. Traditional models often fail to accurately capture the complex, nonlinear interactions between mix components and environmental factors. Concrete strength development is time-dependent, increasing over time. Machine learning models can address the challenge of predicting long-term strength from early-age results by analysing large datasets and identifying strength development patterns.

Machine learning helps predict compressive strength in a more effective way than traditional methods. It learns from past data to understand complex relationships. This means it can accurately predict the strength of new mix designs. Common techniques include decision trees (offer interpretable results), random forests (enhances accuracy), bagging (improves stability), gradient boosting (achieves high accuracy), and neural networks (handle complex predictions).These models enable real-time prediction, outperform traditional methods, generalize across diverse mix designs, and provide significant value to the construction industry.

This research uses machine learning to predict concrete compressive strength, a key structural engineering parameter. It applies and compares Decision Trees, Random Forest, Bagging, Gradient Boosting, and Neural Networks. A comprehensive dataset of concrete mixes and their strengths is collected and preprocessed, addressing missing data, outliers, and feature scaling. Model performance is evaluated using R^2, MSE, MAE, RMSE, RMSLE, and MAPE. Feature importance analysis identifies key concrete mix components influencing strength. Findings are discussed for concrete quality control and construction.

2. Related Work

Table 1 gives the previous researches based on prediction of concrete using various machine learning techniques. The table describes the prediction method used for various types of concrete and also describes the best technique found from their research to predict the compressive strength of concrete.

Table 1. Previous related researches for predicting the compressive strength of concrete

Author Name	Type of Concrete	Machine Learning Techniques Used	Best Technique
Mustapha et al. [1]	Quaternary Blend Concrete	Gradient Boosting, LightGBM, XGBoost, CatBoost	CatBoost
Ahmed et al. [2]	Recycled Aggregate Concrete	Gradient Boosting Machine (GBM), LightGBM, XGBoost, Categorical Gradient Boost (CGB), and HistGradientBoosting (HGB)	Categorical Gradient Boost (CGB)
Kumar et al. [3]	Self-compacting concrete (SCC)	Gradient Boosting (GB), Gradient Boosting-Particle Swarm Optimization, Gradient Boosting-Bayesian Optimization, and Gradient Boosting-Differential Evolution (GB-DE)	Gradient Boosting-Differential Evolution (GB-DE)
Rathakrishnan et al. [4]	High-Performance Concrete	Light Gradient Boosting Machine, CatBoost Regressor, Gradient Boosting Regressor (GBR), Adaboost Regressor, and Extreme Gradient Boosting	Gradient Boosting Regressor (GBR)
de-Prado-Gil et al. [5]	Recycled Aggregate Concrete (RAC)	Random Forest (RF), K-Nearest Neighbor (KNN), Extremely Randomized Trees (ERT), Extreme Gradient Boosting (XGB), Gradient Boosting (GB), Light Gradient Boosting Machine (LGBM), Category Boosting (CB), and the generalized additive models: Inverse Gaussian (GAM1) and Poisson (GAM2)	Gradient Boosting (GB)
Kaloop et al. [6]	High-performance concrete (HPC)	MARS-GBM, -MARS-KRR, and MARS-GPR	MARS-GBM
Malhotra et al. [7]	Various concrete mixes	Support Vector Regression (SVR), XGBoost, Gradient Boosting Regression (GBR)	Support Vector Regressor (SVR)
Paudel et al. [8]	Fly Ash Cement Concrete	Multiple Linear Regressor, Support Vector Regressor, AdaBoost Regressor, Random Forest Regression, XGBoost Regressor, and Bagging Regressor	XGBoost Regressor
Chou et al. [9]	Ready Mix Concrete	Artificial Neural Network, support vector machines, classification and regression tree (CART) and Extreme Gradient Boosting	Extreme Gradient Boosting
Nguyen-Sy et al. [10]	High-Performance Concrete	SVM, ANN, XGBoost	XGBoost
Hoang et al. [11]	High-Performance Concrete	DFP-XGBoost, an ANN and SVR	DFP-XGBoost
Liu et al. [12]	High/ultra-high performance concrete, self-compacting concrete and self-healing concrete	Ridge Regression (RR), Least Absolute Shrinkage and Selection Operator (Lasso), Elastic Net Regression (EnR), Support Vector Regression (SVR) and K-Nearest Neighbor (KNN), Gradient Boosting Decision Tree (GBDT), eXtreme Gradient Boosting (XGBoost) and Random Forest (RF)	GS-XGBoost
Feng et al. [13]	High-Performance Concrete	AdaBoost, Artificial Neural Network, Support Vector Machine	AdaBoost

3. Theory

The theory underlying the prediction of concrete compressive strength relies primarily on empirical relationships and statistical modelling techniques. The compressive strength of concrete is influenced by several critical factors, such as the water-cement ratio, cement content, aggregate properties, admixtures and curing conditions.

To assess the accuracy and reliability of prediction models, the following statistical metrics are applied:

Mean Squared Error (MSE):

$$MSE = (1/n) \times \Sigma(y_i - \hat{y}_i)^2 \quad (1)$$

It calculates the average squared differences between actual and predicted strength values. Lower values indicate better accuracy.

Mean Absolute Error (MAE):

$$MAE = (1/n) \times \Sigma|y_i - \hat{y}_i| \quad (2)$$

It computes the mean of absolute differences between actual and predicted values, offering a direct measure of prediction errors.

Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{[(1/n) \times \Sigma(y_i - \hat{y}_i)^2]} \quad (3)$$

It represents the standard deviation of prediction errors, providing insight into the dispersion of errors.

Root Mean Squared Logarithmic Error (RMSLE):

$$RMSLE = \sqrt{[(1/n) \times \Sigma(\log(y_i + 1) - \log(\hat{y}_i + 1))^2]} \quad (4)$$

It gives more weight to underpredictions than overpredictions, making it a balanced metric for error assessment.

Coefficient of Determination (R^2):

$$R^2 = 1 - [\ \Sigma(y_i - \hat{y}_i)^2 / \Sigma(y_i - \bar{y})^2\ ] \quad (5)$$

It measures how well the model explains the variance in compressive strength values.

Mean Absolute Percentage Error (MAPE):

$$MAPE = (100/n) \times \Sigma|\ (y_i - \hat{y}_i) / y_i\ | \quad (6)$$

It expresses prediction errors as a percentage, helping to assess relative model performance.

These statistical calculations provide the theoretical framework and metrics necessary for evaluating the performance of machine learning models in predicting concrete compressive strength.

3.1 Machine Learning Models

In the present research five machine learning models are evaluated for predicting concrete compressive strength, and

their performance were compared due to their unique advantages.

1 Decision Tree Regression

Decision Tree Regression minimizes variance through recursive data partitioning. Predictions are the average target value within each leaf node:

$$\hat{y} = (1/n) \Sigma y_i, \quad (7)$$

where n represents the node's sample size.

2 Random Forest

Random Forest is a powerful machine learning method. It improves decision trees by combining several trees which are trained on random parts of the data. Instead of relying on one decision tree, Random Forest builds a group of trees. Each tree makes its own prediction, and the final result is found by averaging all these predictions. This approach reduces overfitting and improves accuracy. The prediction formula is:

$$\hat{y} = (1/M) \Sigma (m=1 \text{ to } M) T_m(x) \quad (8)$$

where M is the number of trees, and $T_m(x)$ represents the prediction from the m^{th} tree. By aggregating multiple predictions, Random Forest offers better generalization to new data compared to a single tree.

3 Bagging Regressor

Bagging is a short form for Bootstrap Aggregating, is another ensemble technique. It improves prediction stability by training multiple models on different bootstrapped subsets of the dataset. The idea behind bagging is that by averaging multiple weak learners, the variance in predictions is reduced, leading to a more reliable and accurate model. The prediction is calculated as:

$$\hat{y} = (1/B) \Sigma (b=1 \text{ to } B) f_\beta(x) \quad (9)$$

where $f_\beta(x)$ is the output of the b^{th} base model. Bagging is particularly effective in reducing variance and making models more robust to noise in data, making it a valuable method for predicting concrete strength.

4 Gradient Boosting

Gradient Boosting is a strong method used in machine learning. It creates models one after the other. Each new model tries to fix the errors made by the previous ones. Unlike bagging, which trains models separately, Gradient Boosting improves each new tree based on the mistakes of the earlier trees. This helps to gradually make predictions more accurate. The process follows a specific update rule:

$$F_m(x) = F_{m-1}(x) + \gamma h_m(x) \quad (10)$$

where γ is the learning rate that controls the contribution of each weak learner, and $h_m(x)$ represents the new model trained to reduce the residual errors. Gradient Boosting is widely used for its ability to handle complex patterns and improve predictive performance over multiple iterations.

5 Neural Network

Neural Networks are deep learning models inspired by the functioning of the human brain. They consist of interconnected layers of artificial neurons, where each neuron applies a mathematical transformation to the input and passes the output to the next layer. The prediction formula for a single neuron is:

$$y = \sigma(Wx + b) \quad (11)$$

where W represents the weight matrix, x is the input feature vector, b is the bias term, and σ is the activation function, which introduces non-linearity into the model. Neural Networks are particularly effective for capturing complex relationships in data and have demonstrated superior performance in various regression and classification tasks, including predicting compressive strength in concrete mixes.

3.2 Model Training

To ensure a fair evaluation of each model's predictive capability, the dataset was divided into two sets: 80% for training and 20% for testing. This split ensures that models are trained on a substantial portion of the data while preserving an independent set for performance validation.

4. Experimental Method

Compressive strength prediction concrete requires a systematic approach that begins with collecting relevant data, preprocessing it for consistency, selecting appropriate machine learning models, and evaluating their performance using well-defined metrics. This section provides a detailed overview of the methodology used in this study, ensuring transparency and reproducibility of the research findings.

4.1 Methodology adopted

The methodology delineates a comprehensive strategy for utilizing machine learning to predict concrete compressive strength, ensuring that the proposed models are not only precise but also pragmatic for real-world implementation as shown in Figure 1. By incorporating machine learning into construction and material testing workflows, these models have the capability to enhance quality control, diminish the need for time-consuming laboratory tests, and promote more streamlined, cost-effective, and sustainable construction practices.

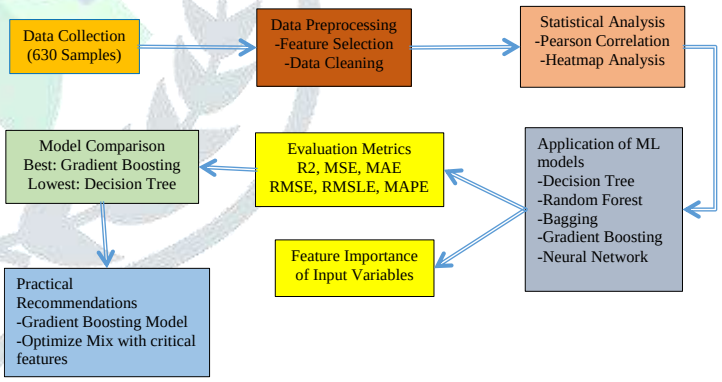


Figure 1 Methodology flowchart

4.1 Data Collection and Processing

This study utilizes a dataset of 630 concrete mix designs, each defined by the proportions of cement, blast furnace slag (BFS)), fly ash, water, superplasticizer, coarse aggregate, fine aggregate, and curing age, and their corresponding compressive strength. The dataset represents diverse, real-world mix designs. Table 2 describes the detailed statistics of various variables used in the dataset.

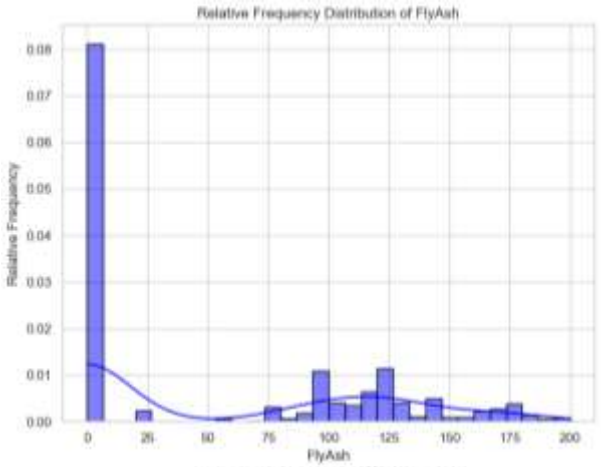
Table 2: Detailed statistics of variables in the dataset

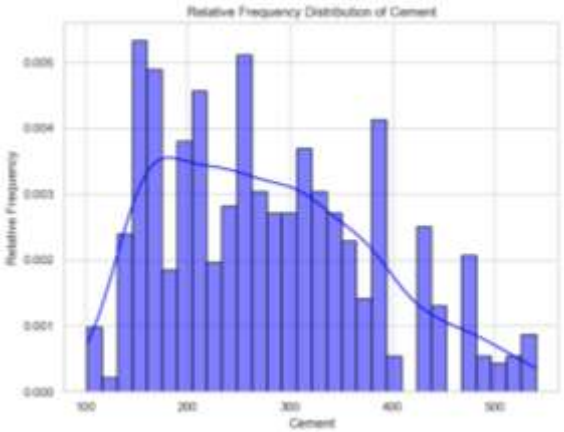
	Cement	BFS	Fly ash	Water	Superplasticizer	Coarse Aggregate	Fine Aggregate	Curing age (days)	Strength (MPa)
COUNT	630	630	630	630	630	630	630	630	630

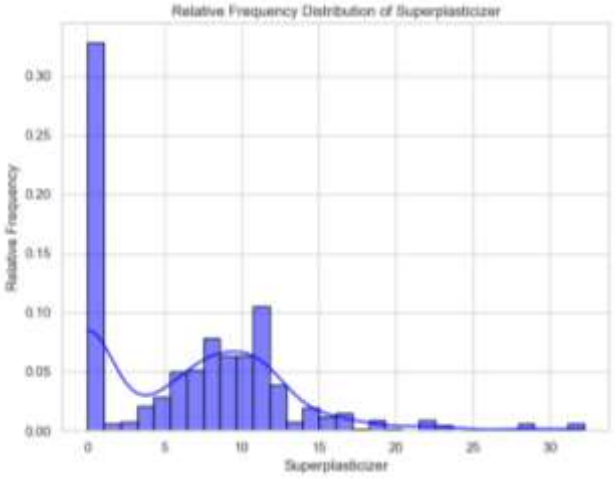
MIN	102	0	0	121.8	0	801	594	1	2.33
25%	192.37	0	0	0	0	932	734.3	7	23.23
50%	266	24	0	182.5	7	968	780.1	28	33.945
75%	350	142.5	118.3	192	10.725	1029.4	827.75	56	46.352
MAX	540	359.4	200.1	247	32.2	1145	992.6	365	82.60
MEAN	278.88	74.56	55.34	180.49	6.53	971.61	776.6	45.73	35.72
STDDEV	102.16	85.18	64.49	21.79	6.17	78.01	81.2	67.019	17.11

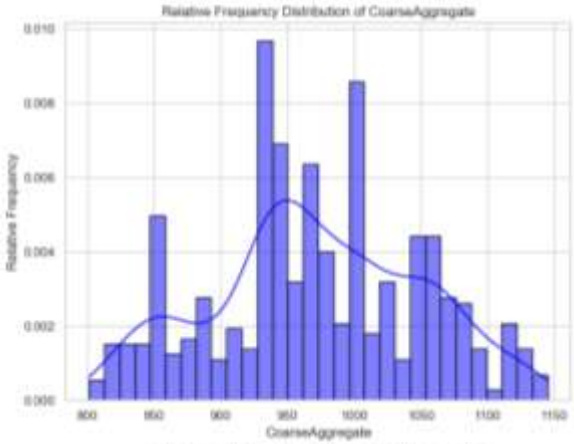
4.1.1 Relative Frequency Distribution of Input Variables

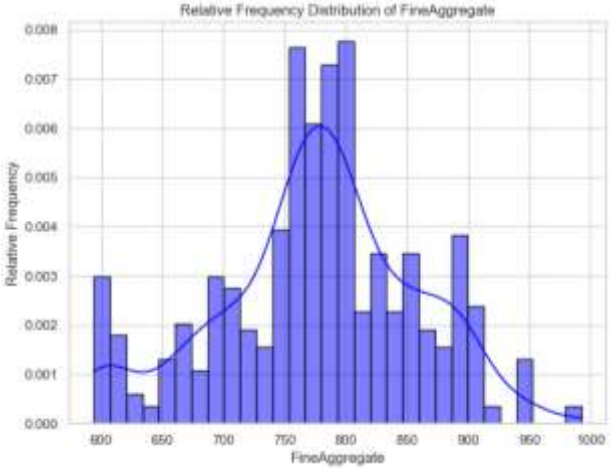
The relative frequency distribution helps understand the frequency and distribution of input variables as shown in Figure 2. It reveals how frequently certain values appear within the dataset, aiding in identifying prevalent ranges for each feature.

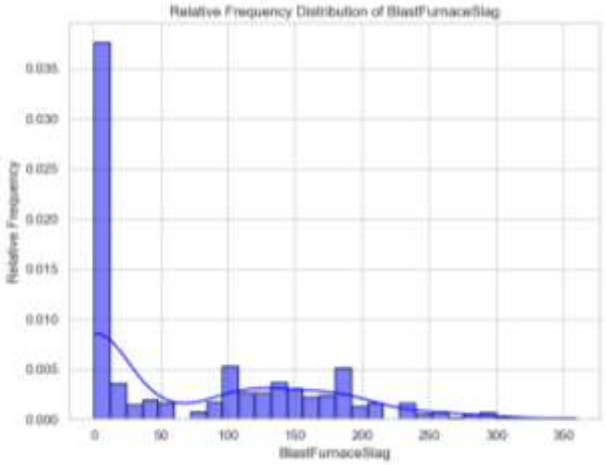












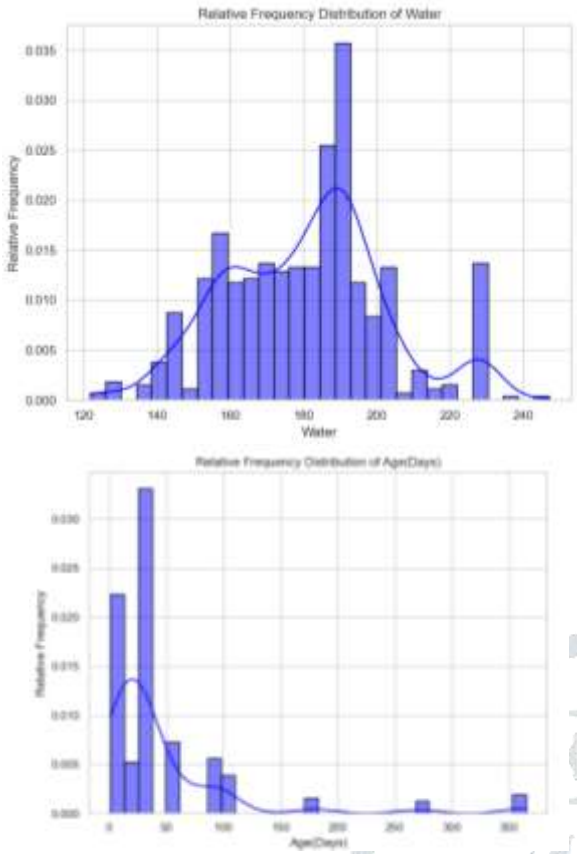


Figure 2: Relative Frequency Distribution of input parameters

4.1.2 Correlation coefficient analysis (Pearson's Coefficient coefficient)

Correlation analysis with Pearson's coefficient helps us understand the linear relationships between variables as shown in Figure 3. A high correlation coefficient shows a strong relationship, which makes it easier to choose important predictors. Pearson correlation heatmap visually confirmed cement content and age as highly correlated with concrete strength, reinforcing findings from the feature importance analysis.

Figure 3: Pearson's Correlation Heatmap

5. Results and Discussions

Each machine learning model's performance was assessed using several key evaluation metrics, such as the coefficient of determination (R^2), Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Root Mean Squared Log Error (RMSLE), and Mean Absolute Percentage Error (MAPE).

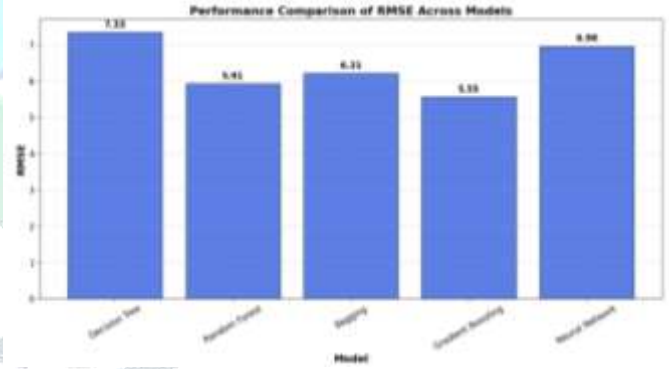
5.1 Comparative analysis between models

The each ML model was evaluated based on performance metrics like R^2 (coefficient of determination), MSE (mean squared error), MAE (mean absolute error), RMSE (root mean squared error), RMSLE (root mean squared log error), and MAPE (mean absolute percentage error) as shown in Table 3 and in Figure 4.

- Gradient Boosting stood out clearly, achieving the highest R^2 value of 0.898, indicating excellent accuracy in predicting concrete strength. It also had the lowest errors across all categories (MSE: 30.838, MAE: 4.025, RMSE: 5.553, RMSLE: 0.176, MAPE: 13.872), making it the top-performing model in this study.
- Random Forest came in second, performing significantly well with an R^2 of 0.884 and low error metrics (MSE: 34.978, RMSE: 5.914), which suggests it's also a robust model for practical applications.
- Bagging and Neural Networks showed good but slightly lower performance (R^2 of 0.873 and 0.84 respectively). Neural Networks, despite their complex nature, didn't surpass Random forest or Gradient boosting in this scenario.
- Decision Tree had the lowest performance among all models tested, with an R^2 of 0.822, suggesting it's less suitable compared to the others for accurate predictions in this context.

Table 3: Statistical Properties of ML results

Model	R2	MSE	MAE	RMSE	RMSLE	MAPE
Decision Tree	0.822	53.795	5.099	7.334	0.244	17.996
Random Forest	0.884	34.978	4.377	5.914	0.192	15.237
Bagging	0.873	38.512	4.542	6.206	0.201	15.845
Gradient Boosting	0.898	30.838	4.025	5.553	0.176	13.872
Neural Network	0.84	48.389	5.186	6.956	0.204	16.479



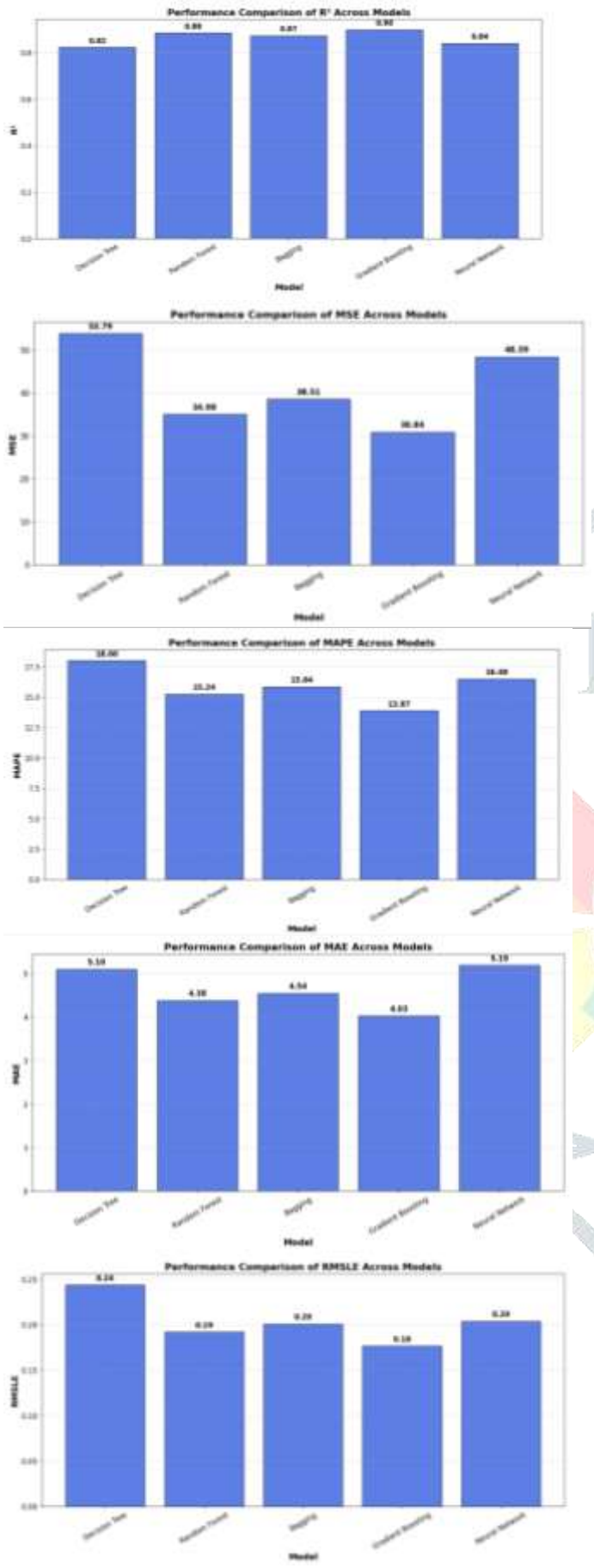


Figure 4: Statistical performance comparison of models

5.2 Feature importance analysis

The feature importance analysis provided insights into which concrete composition factors significantly impact the compressive strength as given in Table 4 and Figure 5:

- Curing age (days) emerged as the most influential feature with an importance score of 0.36. This makes sense practically, as concrete typically gains strength over time.
- Cement also had a high impact (0.271), implies that more cement generally indicates stronger compressive strength of concrete.
- Other components like water (0.114) and superplasticizer (0.099) also showed meaningful influence, highlighting their roles in determining concrete's compressive strength.
- BFS (0.079), fine aggregate (0.043), fly ash (0.018), and coarse aggregate (0.017) are less influential individually but are still relevant factors in optimizing concrete mixes and also its compressive strength.

Table 4: Feature importance of input parameters

Feature	Importance
Curing age (days)	0.36
Cement	0.271
Water	0.114
Superplasticizer	0.099
Blast furnace slag	0.079
Fine aggregate	0.042
Flyash	0.018
Coarse aggregate	0.017

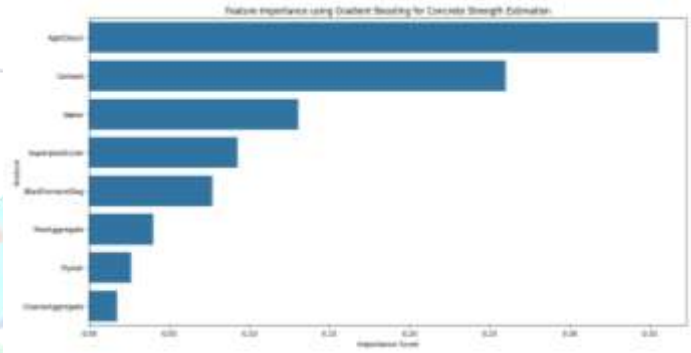


Figure 5: Feature importance score of input parameters

6. Conclusions and Future Scopes

Machine learning techniques have proven to be very effective at predicting the compressive strength of concrete. By analysing crucial factors such as the water-cement ratio, cement content, aggregate properties, and curing age, these models deliver highly accurate and reliable results in real-time offering a practical and efficient alternative to traditional laboratory testing. To measure and enhance the performance of these models, various evaluation metrics such as Root Mean Squared Logarithmic Error (RMSLE), Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), coefficient of determination (R^2) and Mean Absolute Percentage Error (MAPE) are utilized. Gradient Boosting turns out to be the most effective among all the machine learning algorithms adopted in the present analysed. Based on the feature importance analysis of concrete composition the curing age of concrete and cement content are the major factors in optimizing concrete mixes and also its compressive strength. The predicted compressive strength of concrete helps in developing optimized concrete mix designs and also promotes sustainability as it reduces resource consumption.

Looking ahead, more research can improve the accuracy and usability of predictions. Expanding existing datasets to include a wider variety of concrete mixes and different environmental conditions can help these models work better. Hybrid models or deep learning can also be helpful in determining the parameters and how much feature importance

of each parameter carry which affect the compressive strength of concrete.

Conflict of Interest

All authors are requested to disclose any actual or potential conflict of interest including any financial, personal or other relationships with other people or organizations that could inappropriately influence, or be perceived to influence, their work. Otherwise, Authors declare that they do not have any conflict of interest.

Funding Source

none

Authors' Contributions

Author-1 researched literature and conceived the study. Author-2 involved in protocol development, gaining ethical approval, patient recruitment, and data analysis. Author-3 wrote the first draft of the manuscript. All authors reviewed and edited the manuscript and approved the final version of the manuscript.

References

[1] I. B. Mustapha, M. Abdulkareem, T. M. Jassam, A. H. AlAteah, K. A. Alawi Al-Sodani, M.M. H. Al-Tholaia, H. Nabus, S. C. Alih, Z. Abdulkareem, A. Ganiyu "Comparative analysis of gradient-boosting ensembles for estimation of compressive strength of quaternary blend concrete," *Int. J. Concrete Struct. Mater*., vol. **18**, no. **1**, pp. **1-20**, **2024**.

[2] M. R. Kaloop, D. Kumar, P. Samui, J. W. Hu, D. Kim, "Compressive strength prediction of recycled concrete using gradient boosting machine," *Ain Shams Eng. J.*, vol. **15, 2024**.

[3] R. Kumar, S. Kumar, B. Rai, P. Samui, "Development of hybrid gradient boosting models for predicting the compressive strength of high-volume fly ash self-compacting concrete with silica fume," *Structures*, vol. **66**, **2024**.

[4] V. Rathakrishnan, S. B. Beddu, A. N. Ahmed, "Predicting compressive strength of high-performance concrete with high volume ground granulated blast-furnace slag replacement using boosting machine learning algorithms," *Sci. Rep.*, vol. **12**, **2022**.

[5] J. de-Prado-Gil, C. Palencia, N. Silva-Monteiro, R. Martínez-García, "To predict the compressive strength of self-compacting concrete with recycled aggregates utilizing ensemble machine learning models," *Case Stud. Constr. Mater*., vol. **16**, **2022**.

[6] M. R. Kaloop, D. Kumar, P. Samui, J. W. Hu, D. Kim, "Compressive strength prediction of high-performance concrete using gradient tree boosting machine," *Constr. Build. Mater.*, vol. **264**, **2020**.

[7] K. Malhotra, D. Mishra, C. S. Tumrate, "Prediction of concrete compressive strength employing machine learning techniques," *Mater. Today Proc*., **2023**.

[8] S. Paudel, A. Pudasaini, R. K. Shrestha, E. Kharel, "Compressive strength of concrete material using machine learning techniques," *Clean Eng. Technol*., vol. **15**, **2023**.

[9] J. Chou, L. Chen, C. Liu, "Forensic-based investigation-optimized extreme gradient boosting system for predicting compressive strength of ready-mixed concrete," *J. Comput. Des. Eng*., vol. **10**, no. **1**, pp. **425–445, 2023**.

[10] T. Nguyen-Sy, J. Wakim, Q. D. To, M. N. Vu, T. D. Nguyen, T. T. Nguyen, "Predicting the compressive strength of concrete from its compositions and age using the extreme gradient boosting method," *Constr. Build. Mater.*, vol. **260**, **2020**.

[11] N. Hoang, V. Tran, X. Tran, "Predicting compressive strength of high-performance concrete using hybridization of nature-inspired metaheuristic and gradient boosting machine," *Mathematics*, vol. **12**, no. **8**, **2024**.

[12] K. Liu, L. Zhang, W. Wang, G. Zhang, L. Xu, D. Fan, Y. Li, "Development of compressive strength prediction platform for concrete materials based on machine learning techniques," *J. Build. Eng.*, vol. **80**, **107977**, **2023**.

[13] D. Feng, Z. Liu, X. Wang, Y. Chen, J. Chang, D. Wei, Z. Jiang, "Machine learning-based compressive strength prediction for concrete: An adaptive boosting approach," *Constr. Build. Mater*., vol. **230**, **117000**, **2020**.