IMPROVING CONCRETE STRENGTH PREDICTION WITH DOMAIN-INFORMED FEATURE ENGINEERING AND MACHINE LEARNING: A STUDY USING THE UCI DATASET

Md Rakib Hossain*[1]

[1] *MSc Student, King Fahd University of Petroleum and Minerals, e-mail: hossain.rakib.ce@gmail.com*

***Corresponding Author**

ABSTRACT

Concrete compressive strength is a key indicator of structural performance, yet laboratory testing is time-consuming, costly, and sometimes destructive. Using machine learning to predict this parameter could facilitate rapid, non-destructive evaluations and optimise mix design. This study aims to assess whether embedding civil engineering knowledge into feature engineering improves the accuracy and transparency of machine learning models for predicting 28-day compressive strength. A benchmark dataset of 1,030 mixes from the UCI repository was cleaned and expanded with engineered variables such as water–cement ratio, binder content, aggregate-to-binder ratio, and log-transformed age. Ten models, including Linear Regression, Ridge Regression, Random Forest, Gradient Boosting, and Neural Networks, were trained and evaluated using MAE, MSE, RMSE, and R^2. The best results were obtained by ensemble approaches ($R^2 \approx 0.90$; RMSE = 5 MPa), whereas simpler models demonstrated notable improvements upon the inclusion of engineering features (mean R^2 for linear regression rose from 0.62 to 0.83). Feature importance and SHAP analyses revealed that the binder content, water–cement ratio, and log(age) were the most influential variables, according to Abrams' law and hydration theory. These findings demonstrate that domain-informed feature engineering improves the relationship between materials science and machine learning by increasing prediction accuracy and model interpretability.

Keywords: *Concrete Compressive Strength, Feature Engineering, Machine Learning Models, Gradient Boost, SHAP Analysis*

1. INTRODUCTION

Concrete remains the backbone of modern infrastructure, and its compressive strength (CCS) serves as the most critical indicator of structural performance and safety. The ability to accurately predict CCS is fundamental for design optimization, quality assurance, and sustainability evaluations (Yeh 1998). However, laboratory testing for compressive strength is both destructive and time-consuming, requiring standardized curing periods and controlled conditions before testing (Elhishi, Elashry, and El-Metwally 2023a). These constraints make traditional testing impractical when rapid assessment or large-scale optimization of concrete mixes is required.

The compressive strength of concrete is affected by complicated, nonlinear interactions between various mix ingredients, including cement, water, supplementary cementitious materials (SCMs), aggregates, admixtures, and curing time (Zhang et al. 2024a). Classical empirical relations, such as Abrams' law, capture only a subset of this behavior and fail to generalize across other mixture proportions. In recent decades, machine learning (ML) techniques have evolved as effective tools for modeling such nonlinear relationships, offering accurate predictions based on mix design parameters while lowering experimental dependency (Hoang, Nguyen, and Ly 2024).

The Concrete Compressive Strength dataset from the UCI Machine Learning Repository, compiled by Yeh (1998), is an important resource for advancing this research (Concrete Compressive Strength - UCI Machine Learning Repository n.d.). It includes 1,030 concrete mix samples with eight independent variables, such as cement, slag, fly ash, water, superplasticizer, coarse aggregate, fine aggregate, and age, and one dependent variable, compressive strength (MPa). The dataset includes a wide variety of mixture compositions and curing periods, giving it a robust benchmark for data-driven CCS prediction. Its accessibility and simplicity have encouraged numerous comparative studies using diverse machine learning algorithms.

Early research primarily focused on Artificial Neural Networks (ANN) to capture nonlinear relations between inputs and strength. Yeh (1998) first showed that ANN models substantially outperformed multiple linear regression, establishing a foundation for data-driven approaches in concrete research. Later, studies extended this exploration by using Support Vector Machines (SVM), Decision Trees, and Ensemble Models (Kaloop et al. 2020; Rengasamy et al. 2022). These approaches improved accuracy but were largely confined to the eight raw input features, leaving room for incorporating deeper engineering insight through feature engineering.

More recently, researchers have benchmarked numerous models using the UCI dataset. Elhishi et al. (2023) evaluated eight algorithms, Linear, Ridge, Lasso, Decision Tree, Random Forest, XGBoost, SVM, and ANN, and found ensemble techniques like XGBoost achieved the highest performance with an R^2 of about 0.91 and RMSE near 4.37 MPa (Elhishi, Elashry, and El-Metwally 2023b). They also applied SHapley Additive exPlanations (SHAP) to identify influential parameters such as cement, water, and age. Similarly, Zhang et al. (2024) tested twelve regressors, including DeepForest, LightGBM, and CatBoost, concluding that DeepForest produced the best results with comparable accuracy (Zhang et al. 2024b). These studies demonstrate the predictive potential of ML algorithms but rely exclusively on raw features, meaning that physical relationships such as the water–cement ratio (w/c) or binder content are not explicitly represented in the models.

The integration of domain knowledge into data representation remains a major gap in the literature. While ML models can detect correlations, they may fail to infer causal or mechanistic relationships known to engineers. For example, an increase in w/c ratio reduces strength due to higher porosity, and the total binder content directly affects hydration product formation (Chen et al. 2024). Most existing models learn these effects implicitly rather than encoding them through engineered features. As a result, their predictions, although accurate, often behave as “black boxes” that lack transparency for practical engineering use (Ghrici et al. 2025).

Efforts to improve interpretability have gained traction through Explainable Artificial Intelligence (XAI) frameworks. Methods such as Partial Dependence Plots (PDP) and SHAP values allow researchers to visualize how each variable affects predicted strength. Nikoopayan Tak et al. (2025) combined SHAP and PDP with Gradient Boosting to reveal that w/c ratio and curing age were the

most influential variables in their models, achieving R^2=0.94 (Nikoopayan Tak, Feng, and Mahgoub 2025). They also noted that dataset enrichment and explicit feature engineering would further improve accuracy and interpretability. Despite this, few studies have systematically incorporated engineered variables such as aggregate-to-binder ratio, logarithmic age, or interaction terms (e.g., cement × water). Consequently, model performance improvements have often been reported without statistical validation, making it unclear whether they result from genuine learning gains or random variation (Willard et al. 2023).

This research builds upon these developments by embedding civil-engineering knowledge into the machine learning process through carefully designed feature engineering. The approach involves deriving new explanatory variables including water to cement ratio, binder content, aggregate-to-binder ratio, and log-transformed age, that capture essential physical mechanisms governing concrete strength development. Interaction terms such as cement × water and slag × age are also introduced to reflect hydration synergy and delayed pozzolanic activity. By enriching the dataset in this manner, the study aims to enhance both model accuracy and interpretability.

Ten machine learning models, such as Linear Regression, Ridge, Lasso, Support Vector Regression, k-Nearest Neighbours, Decision Tree, Random Forest, Extra Trees, Gradient Boosting, and Multi-Layer Perceptron are trained and validated on both the baseline dataset (original variables) and the engineered dataset (original plus derived features). Their performance is evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Square Error (RMSE), and the Coefficient of Determination (R^2). To ensure that improvements from feature engineering are not incidental, statistical significance is tested using paired t-tests across ten-fold cross-validation scores.

In addition, Explainable AI techniques are applied to interpret results from top-performing models. SHAP summary and dependence plots, along with PDPs, are used to visualize the influence of each feature on predictions. This interpretability analysis verifies whether the machine learning models adhere to fundamental engineering principles, such as the inverse relationship between w/c ratio and compressive strength and the positive impact of curing age.

In summary, this study contributes to the advancement of machine-learning-based CCS prediction in three key ways. First, it introduces a domain-informed feature-engineering framework that integrates materials science knowledge directly into the dataset, thereby improving both accuracy and interpretability. Second, it provides statistical evidence showing that engineered features yield significant performance gains across multiple algorithms. Third, it bridges the gap between data science and civil engineering by ensuring that model predictions reflect established physical laws. By combining engineering intuition with computational intelligence, this research moves toward more transparent, explainable, and scientifically grounded predictions of concrete compressive strength, enabling faster, more reliable, and more sustainable mix design in modern construction.

2. METHODOLOGY

This paper proposes an integrative methodology that combines domain-informed feature engineering with a structured machine learning framework to predict the compressive strength of concrete. Figure 1 presents the workflow followed in the present study, which involves data acquisition, data preprocessing, feature engineering, model training, model evaluation, statistical validation, and model interpretability analysis in order. The method ensures that knowledge of concrete mix design enters the process of predictive modeling directly, rather than through algorithmic learning alone.

2.1 Dataset Description

This research used the Concrete Compressive Strength Dataset obtained from the UCI Machine Learning Repository, originally contributed by Prof. I-Cheng Yeh (Concrete Compressive Strength - UCI Machine Learning Repository n.d.). The dataset consists of 1,030 data records that comprise nine different variables related to a given concrete mixture. These input variables refer to the proportion of different concrete ingredients in kg/m³ and curing age in days. These are the input features, whereas the target variable is compressive strength in MPa. The descriptive analysis of the dataset is presented in Table 1.

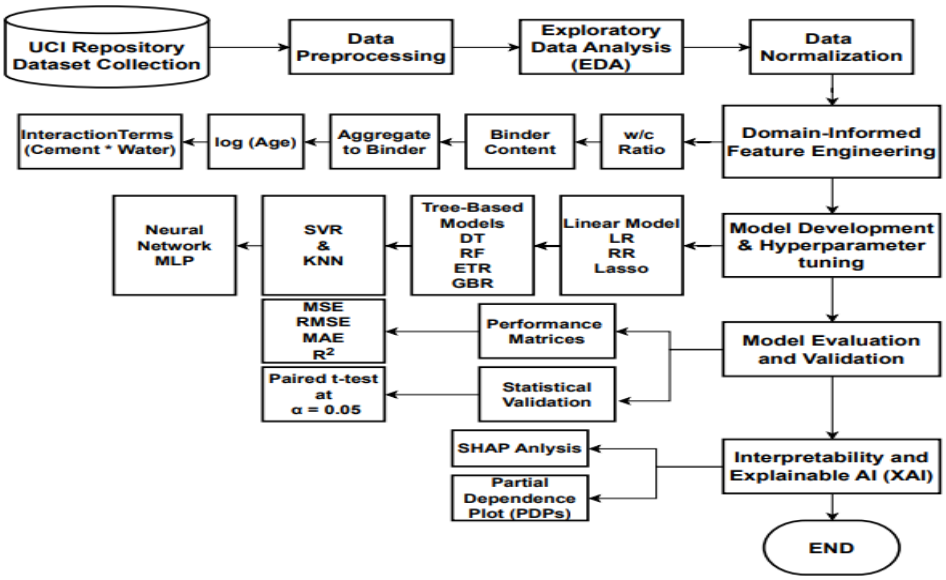


Figure 1: Schematic Diagram of Workflow

Table 1 Statistical analysis of dataset

	cement	slag	Fly ash	water	superplasticizer	Coarse agg	Fine agg	age	Cs mpa
count	1,030	1,030	1,030	1,030	1,030	1,030	1,030	1,030	1,030
mean	281.17	73.89	54.19	181.6	6.20	972.91	773.58	45.66	35.82
std	104.51	86.28	63.99	21.4	5.97	77.75	80.18	63.17	16.71
min	102	0	0	121.8	0	801	594	1	2.33
25%	192.38	0	0	164.9	0	932	730.95	7	23.71
50%	272.90	22	0	185	6.35	968	779.51	28	34.44
75%	350	142.95	118.27	192	10.16	1,029.4	824	56	46.14
max	540	359.4	200.1	247	32.20	1,145	992.6	365	82.60

2.2 Data Preprocessing and Splitting

This dataset preprocessing was performed in Python using the pandas and scikit-learn libraries. First, the data was checked for duplication; three duplicate records were found and removed to maintain data quality. In order to have the same scale among all the features, which is a key preprocessing step for some machine learning models whose performance depends on the magnitude of the features, including but not limited to SVR and MLP, all continuous variables were standardized using the StandardScaler method.

Then, boxplots were used to investigate likely outliers. The outliers found were valid since they represented realistic experimental scenarios, involving either a higher cement content or longer curing time. Therefore, none of the data points was removed from the dataset. Finally, the dataset was randomly split into two parts: 80% for training and 20% for testing, using a fixed random state (42) in

order to be able to reproduce these results. Furthermore, cross-validation was used to make the model performance robust and reliable.

The dataset were divided into training and testing subsets randomly to ensure model validation, a standard approach in ML that helps prevent overfitting and assesses model generalizability. A commonly adopted ratio is taken as 80% of the data were used for training and 20% of samples were allocated for testing.

- Training dataset: These selected sets are used to rain the data models.
- Testing dataset: These datasets are used to check the performance of algorithm based on unseen to the models.

2.3 Feature Engineering

Feature engineering was a key contribution of this study, leveraging domain expertise in cement chemistry and concrete mix design to generate physically meaningful variables that enhance model interpretability and predictive power.

- **Water–Cement Ratio (w/c):**
  w/c ratio = Water/ Cement
  A fundamental strength indicator based on Abrams' Law, where higher w/c ratios correspond to lower compressive strength.
- **Binder Content:**
  Binder = Cement + Fly Ash + Slag
  Represents the total cementitious materials contributing to hydration and strength gain.
- **Aggregate–Binder Ratio (A/B):**
  A/B ratio = (Coarse Aggregate+Fine Aggregate) / Binder
  Indicates the relative volume of aggregates to binder; higher ratios typically reduce strength due to reduced paste content.
- **Logarithmic Age:**
  log Age = log (1 + Age)
  Reflects the nonlinear, logarithmic trend of strength development over curing time.
- **Age Squared (Age²):**
  Captures higher-order effects of curing duration to model nonlinear strength growth.
- **Interaction Terms:**
  Interaction features such as *Cement* × *Water* and *Slag* × *Age* were introduced to capture coupled effects between material composition and curing behavior.

This structured feature set integrates fundamental engineering principles with data-driven modeling to improve both interpretability and predictive accuracy.

2.4 Machine Learning Model

Linear Regression (LR):

Linear Regression models the relationship between input variables X and the target variable y using a straight-line equation:

$$Y = \beta_0 + \sum \beta_i X_i + \varepsilon \quad (1)$$

where β_i denotes model coefficients and ε represents the error term. It serves as a simple yet interpretable baseline for evaluating nonlinear models (Shaaban et al. 2025).

Ridge Regression (RR):

Ridge Regression extends linear regression by including an L2 regularization term to prevent overfitting and mitigate multicollinearity:

$$min_\beta \|y - X\beta\|^2 + \lambda\|\beta\|^2 \quad (2)$$

The penalty parameter λ controls coefficient shrinkage, ensuring more stable predictions when features are correlated .

Lasso Regression (Lasso):

Lasso introduces L1 regularization to enforce sparsity in coefficients, performing both feature selection and regularization simultaneously:

$$min_{\beta}||y - X\beta||^2 + \lambda ||\beta||_1 \quad (3)$$

This model is particularly valuable for identifying the most influential predictors in complex datasets.

Support Vector Regression (SVR):

SVR employs kernel functions to capture nonlinear relationships by mapping data into higher-dimensional spaces. It minimizes the error within an ε-insensitive margin:

$$min_{w,b} \tfrac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \quad (4)$$

subject to | yi − (w · ϕ(xi)+ b) | ≤ε. This makes it effective for smooth nonlinear regression tasks (Wang et al. 2024).

k-Nearest Neighbours (KNN):

KNN is a non-parametric model that predicts outcomes based on the average of the k closest data points in the feature space. It relies on distance metrics such as Euclidean distance:

$$\hat{y} = \frac{1}{k}\sum_{\{i \in N_{k(x)}\}} y_i \quad (5)$$

This approach excels in capturing local trends without assuming a global model structure (Duan 2024).

Decision Tree Regressor (DT):

A Decision Tree recursively partitions data into smaller homogeneous subsets using splitting criteria such as mean squared error (MSE). It forms hierarchical decision rules that can naturally represent nonlinear relationships.

Random Forest Regressor (RF):

Random Forest is an ensemble method combining multiple decision trees trained on random subsets of data and features. The final prediction is the average of all individual trees, reducing variance and overfitting.

Extra Trees Regressor (ET):

Extra Trees, or Extremely Randomized Trees, follow the same ensemble principle as Random Forest but introduce additional randomness in feature selection and split thresholds, further improving generalization and speed (Wang et al. 2025).

Gradient Boosting Regressor (GBR):

GBR builds an additive model in a forward stage-wise fashion by iteratively fitting new trees to the residual errors of previous ones:

$$F_m(x) = F_{(m-1)}(x) + \eta h_m(x) \quad (6)$$

where η is the learning rate. This process gradually minimizes prediction errors, resulting in a strong predictive ensemble (Mustapha et al. 2024).

Multi-Layer Perceptron (MLP):

MLP is a feed-forward neural network composed of input, hidden, and output layers. Each neuron computes weighted sums followed by nonlinear activation functions such as ReLU:

$$y = f(w_2\sigma(w_1x + b_1) + b_2 \quad (7)$$

This structure allows MLPs to approximate complex nonlinear relationships between variables (Ta, Vo, and Nguyen 2024).

2.5 Performance Metrices

To compare model performance comprehensively, four standard regression metrics were employed which is presented in Table

Statistical Indicator	Equation	Acceptable Range
Mean Absolute Error (MAE)	$MAE = \frac{1}{n}\sum(Y_i - Y'_i)$	Greater than 0.65 for an excellent model
Mean Square Error (MAE)	$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - Y'_i)^2$	Closer to zero (0≤MSE≤∞)

Root Mean Squared Error (RMSE)	RMSE = $\sqrt{MSE}$	MAE < RMSE
Coefficient of determination (R^2)	$R^2 = 1 - \frac{RSS}{TSS}$	Close to 1

These metrics collectively quantify absolute, squared, and relative predictive errors.

2.6 Statistical Significance Testing

To evaluate whether improvements from feature engineering were statistically significant, paired t-tests were performed on 10-fold cross-validated R^2 scores between baseline and engineered datasets.

- Null Hypothesis (H_0): Mean R^2 (baseline) = Mean R^2 (engineered).
- Alternative Hypothesis (H_1): Mean R^2 (baseline) < Mean R^2 (engineered).

Significance was tested at $\alpha = 0.05$.

2.7 Interpretability and Explainable AI (XAI)

The feature importance and SHAP were used for interpreting the model predictions. Tree-based models such as Random Forest, Gradient Boosting, and XGBoost provide intrinsic feature importances, while SHAP values offer both global and local interpretability for all models.

- Global Interpretation: SHAP summary plots identified dominant predictors (binder, w/c ratio, log Age).
- Local Interpretation: SHAP dependence plots explained individual predictions, highlighting the direction and magnitude of feature impacts.
- Partial Dependence Plots (PDPs): Showed monotonic relationships: strength decreased with w/c ratio and increased with log Age, verifying physical consistency with Abrams' law and hydration theory.

3. RESULTS AND DISCUSSION

3.1 Correlation Analysis

Correlation analysis showed the first indication of how individual features influenced compressive strength. The baseline correlation heatmap Figure 4 showed that cement and age were positively correlated with strength, while water showed a clear negative correlation. Superplasticizer showed a mild positive correlation. These relationships make sense in the context of the fundamentals of concrete: increased cement content and longer curing time improve hydration and strength, but excess water weakens the matrix. Scatter plots confirmed these trends: strength increased with cement and age, and decreased with water. The log transformed age plots showed the nonlinear development of rapid strength at early ages, slowing as hydration proceeds. Most tellingly, the scatter plot of w/c ratio versus strength confirmed Abrams' law, showing an inverse relationship between w/c ratio and compressive strength.

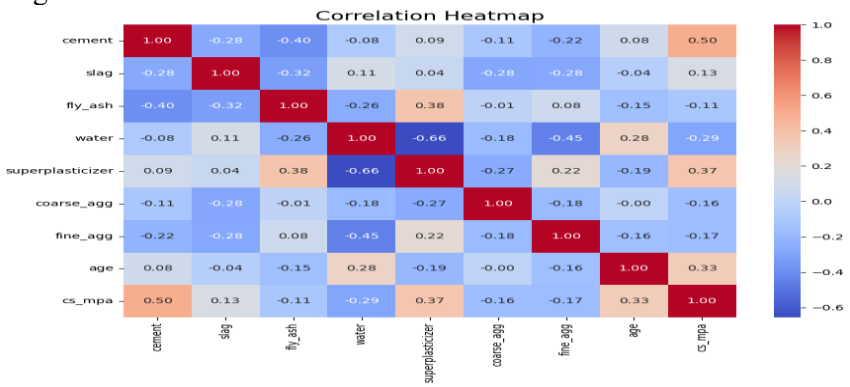
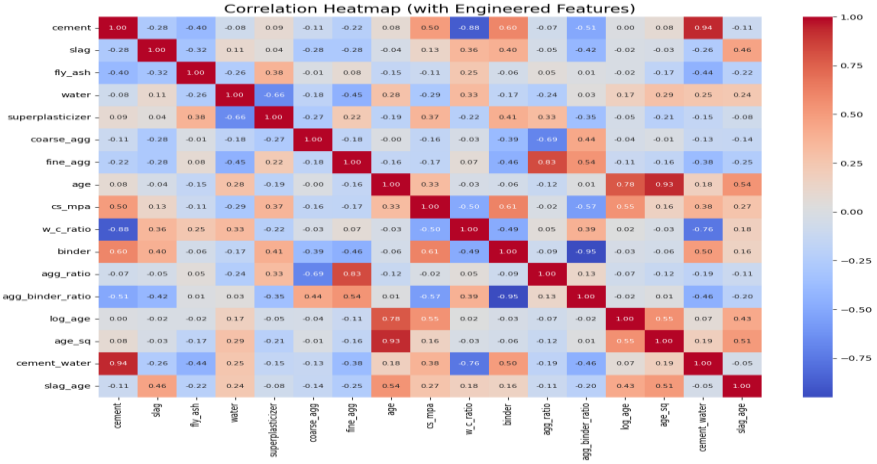


Figure 2: Correlation Heatmap (without Feature Engineering)

Introducing engineered features noticeably strengthened these relationships. The newly added variables included w/c ratio, binder content, aggregate-to-binder ratio, and log(age), which encoded materials-science principles directly into the dataset. It is observed from the engineered correlation heatmap, presented in Figure 7, that binder content and log(age) have the strongest positive correlations of r = +0.61 and r = +0.55 with compressive strength, while aggregate-to-binder ratio and water–cement ratio have strong negative correlations of r = –0.57 and r = –0.50, respectively. These captured the physics of strength development and were far more powerful than their raw counterparts.



3.2 Comparison of Model Results Before and After Feature Engineering

Table 2 compares model performance before and after feature engineering. The evaluation metrics are MAE, MSE, RMSE, and the coefficient of determination, R^2. Feature engineering resulted in some improvement for all models, as the error values were reduced, while the R^2 score increased. Among these algorithms, Gradient Boosting proved to be the most accurate with the highest R^2 value of 0.904, followed by Extra Trees with 0.903 and SVR with 0.894. On the other hand, linear models like Linear, Ridge, and Lasso Regression showed relatively low predictive performance, which thus implies that nonlinear models and ensemble algorithms generalize better for the complex relationships between variables.

Table 2 Model Performance before and after Feature Engineering

	MAE_B	MSE_B	RMSE_B	R^2_B	MAE_Eng	MSE_Eng	RMSE_Eng	R^2_Eng
GB	4.101	30.716	5.542	0.881	3.625	24.79	4.979	0.904
ET	3.411	28.242	5.314	0.89	3.26	24.965	4.997	0.903
SVR	4.205	35.468	5.955	0.862	3.589	27.318	5.227	0.894
RF	3.812	31.076	5.575	0.879	3.587	28.398	5.329	0.89
MLP	4.579	32.905	5.736	0.872	3.9	32.187	5.673	0.875
KNN	7.093	79.578	8.921	0.691	4.951	39.662	6.298	0.846
Lasso	7.752	95.975	9.797	0.628	5.2	42.28	6.502	0.836
RR	7.752	95.97	9.796	0.628	5.208	42.393	6.511	0.835

LR	7.745	95.975	9.797	0.628	5.225	42.523	6.521	0.835
DT	4.461	48.681	6.977	0.811	3.999	44.316	6.657	0.828

3.3 Statistical Validation

Paired t-tests were conducted on the 10-fold cross-validated R^2 scores shown in **Table 3**. These improvements for Linear, Ridge, Lasso, and KNN models were statistically significant at $p < 0.05$, though improvements for ensembling methods such as Gradient Boosting and Random Forest showed smaller, non-significant differences. This shows that feature engineering adds value mainly to the most simple models that cannot learn non-linear interactions automatically and that more advanced algorithms are less reliant on manual feature transformations.

Table 3 Statistical Analysis of Models

Model	Baseline Mean R^2	Engineered Mean R^2	Mean Improvement	t-statistic	p-value
ET	0.786	0.782	-0.004	0.334	0.746
RF	0.731	0.759	0.028	-1.829	0.101
GB	0.75	0.757	0.007	-0.39	0.705
SVR	0.53	0.673	0.142	-1.392	0.197
MLP	0.607	0.663	0.056	-1.687	0.126
Lasso	0.279	0.635	0.357	-3.482	0.007
DT	0.345	0.633	0.288	-1.687	0.126
RR	0.279	0.632	0.353	-3.519	0.007
LR	0.278	0.629	0.351	-3.482	0.007
kNN	0.257	0.536	0.278	-3.73	0.005

3.4 Model Interpretability

Figure 4 SHAP summary plot showing both feature importance and directionality in the prediction of compressive strength. The three most relevant features detected were binder content, water-cement ratio, and log(age). Consequently, increasing binder content or longer curing ages (red points with positions on the right) increased the predicted strength, while a higher water-cement ratio (red points with positions on the left) decreased it. The consistency of these patterns for all samples was a further indication that the model captured established behaviors of the material in question, such as Abrams' law.

The partial dependence plots shown in **Figure 3** give further support for these interpretations. Strength varies monotonically with the water–cement ratio, increasing with curing age and reaching a near plateau after about 90 days during which hydration slows down. These interpretability results demonstrate that machine learning models capture true physical mechanisms of concrete strength development rather than relying solely on data correlations.

The outcomes of this study are that domain informed feature engineering significantly enhances the accuracy and interpretability of machine learning models applied to the prediction of concrete compressive strength. Embedding established engineering principles, such as the influence of w/c ratio, binder content, and curing age, transformed model performance and gave forecast physical meaning. Simpler algorithms like Linear Regression showed dramatic gains, with R^2 improving from 0.62 to 0.83, whereas more advanced models like Random Forest and Gradient Boosting, although yielding smaller statistical improvements, reinforced known material behaviors. Analyses of interpretability via SHAP values and Partial Dependence Plots further supported that these engineered variables acted in concert with well-established laws like Abrams' law, in which higher binder content and longer curing increased strength, and higher w/c ratios decreased it. This dual achievement of better predictive accuracy with simpler models and greater interpretability with complex ones speaks to the contribution of the study in tying together traditional materials science with modern, data-driven modeling.

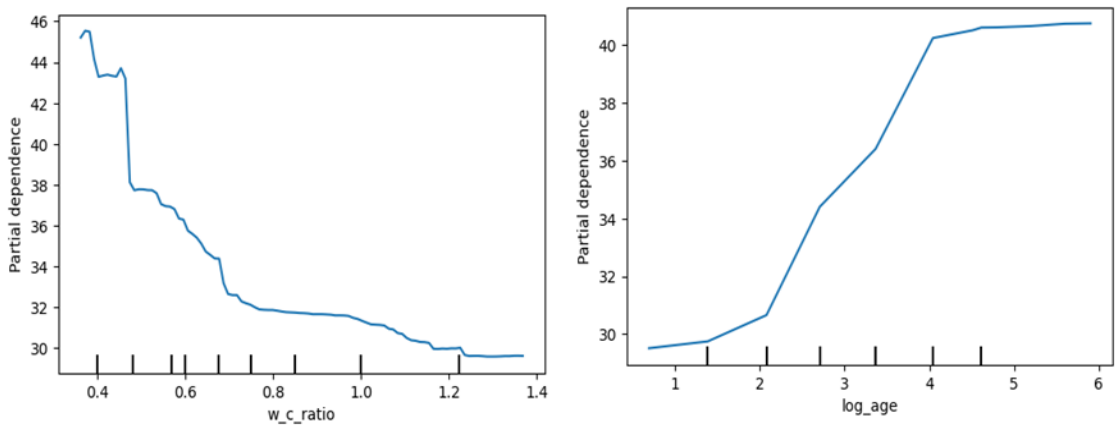


Figure 3 Partial Dependence Plot (PDPs)

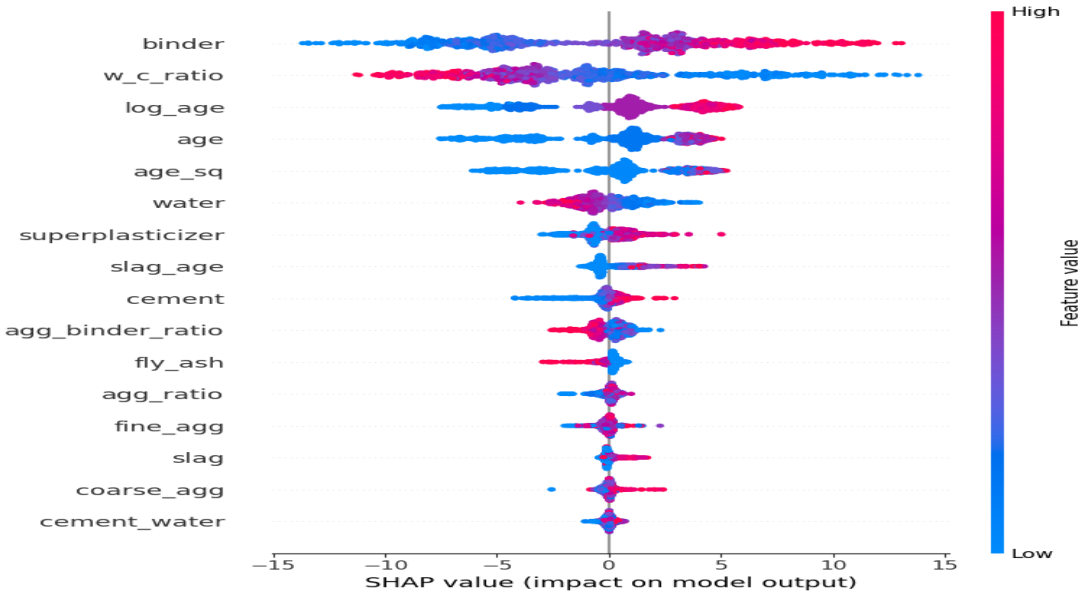


Figure 4 SHAP Analysis

4. CONCLUSIONS

This study successfully validated that the integration of domain-informed feature engineering and machine learning techniques enhances the prediction of CCS using the UCI dataset. By embedding knowledge in civil engineering, especially introducing physically meaningful variables such as w/c

ratio, binder content, aggregate to binder ratio, and logarithmic age, the predictive models captured the underlying material behavior much better than those with raw features alone.

Comparing these results across the ten models showed that feature engineering appeared to substantially improve model accuracy, especially for less complex algorithms, such as Linear, Ridge, and Lasso regressions. How statistically significant these changes in predictive performances are was validated through a paired t-test. Whereas multiple ensemble algorithms like Gradient Boosting and Extra Trees had the top overall accuracies, their improvements were relatively smaller; domain-informed features primarily benefit the models that cannot, by nature, capture nonlinear relationships.

Besides accuracy, interpretability analysis using feature importance, SHAP values, and Partial Dependence Plots showed that the engineered features followed established concrete science principles. The results confirmed that higher binder content and longer curing time improve strength, however higher w/c ratio decrease it, which supports Abrams' law and hydration theory.

Overall, the proposed integration of domain knowledge into data preprocessing enhanced not only the predictive capability but also the transparency and physical consistency of machine learning models. Particularly, Gradient Boosting proved to be the most reliable and accurate algorithm among all tested in the prediction of compressive strength, representing an optimal balance between performance and interpretability.

Future research could extend this approach by incorporating larger and more diverse datasets, exploring hybrid physics informed machine learning models, and developing practical decision-support tools for concrete mix design and optimization in real-world construction scenarios.

REFERENCES

Chen, Xia, Ruiji Sun, Ueli Saluz, Stefano Schiavon, and Philipp Geyer. 2024. "Using Causal Inference to Avoid Fallouts in Data-Driven Parametric Analysis: A Case Study in the Architecture, Engineering, and Construction Industry." Developments in the Built Environment 17. doi:10.1016/j.dibe.2023.100296.

Concrete Compressive Strength - UCI Machine Learning Repository. n.d. Retrieved November 13, 2025. https://archive.ics.uci.edu/dataset/165/concrete+compressive+strength.

Duan, Min. 2024. "Innovative Compressive Strength Prediction for Recycled Aggregate/Concrete Using K-Nearest Neighbors and Meta-Heuristic Optimization Approaches." Journal of Engineering and Applied Science 71(1). doi:10.1186/s44147-023-00348-9.

Elhishi, Sara, Asmaa Mohammed Elashry, and Sara El-Metwally. 2023a. "Unboxing Machine Learning Models for Concrete Strength Prediction Using XAI." Scientific Reports 13(1). doi:10.1038/s41598-023-47169-7.

Elhishi, Sara, Asmaa Mohammed Elashry, and Sara El-Metwally. 2023b. "Unboxing Machine Learning Models for Concrete Strength Prediction Using XAI." Scientific Reports 13(1). doi:10.1038/s41598-023-47169-7.

Ghrici, Ahmed Abdelghafour, Ali Benzaamia, Freha Mezzoudj, Chahreddine Medjahed, and Mohamed Ghrici. 2025. "SHAP-Enhanced Tree-Based Regression for Predicting the Compressive Strength of High Performance Concrete." Construction and Building Materials 495. doi:10.1016/j.conbuildmat.2025.143602.

Hoang, Huong-Giang Thi, Thuy-Anh Nguyen, and Hai-Bang Ly. 2024. "Interpretable Machine Learning Models for Concrete Compressive Strength Prediction." Innovative Infrastructure Solutions 10(1):5. doi:10.1007/s41062-024-01808-8.

Kaloop, Mosbeh R., Deepak Kumar, Pijush Samui, Jong Wan Hu, and Dongwook Kim. 2020. "Compressive Strength Prediction of High-Performance Concrete Using Gradient Tree Boosting Machine." Construction and Building Materials 264. doi:10.1016/j.conbuildmat.2020.120198.

Mustapha, Ismail B., Muyideen Abdulkareem, Taha M. Jassam, Ali H. AlAteah, Khaled A. Alawi Al-Sodani, Mohammed M. H. Al-Tholaia, Hatem Nabus, Sophia C. Alih, Zainab Abdulkareem, and Abideen Ganiyu. 2024. "Comparative Analysis of Gradient-Boosting Ensembles for Estimation of Compressive Strength of Quaternary Blend Concrete." International Journal of Concrete Structures and Materials 18(1). doi:10.1186/s40069-023-00653-w.

Nikoopayan Tak, Mohammad Saleh, Yanxiao Feng, and Mohamed Mahgoub. 2025. "Advanced Machine Learning Techniques for Predicting Concrete Compressive Strength." Infrastructures 10(2). doi:10.3390/infrastructures10020026.

Rengasamy, Divish, Jimiama M. Mase, Aayush Kumar, Benjamin Rothwell, Mercedes Torres Torres, Morgan R. Alexander, David A. Winkler, and Grazziela P. Figueredo. 2022. "Feature Importance in Machine Learning Models: A Fuzzy Information Fusion Approach." Neurocomputing 511:163–74. doi:10.1016/j.neucom.2022.09.053.

Shaaban, Mohammed, Mohamed Amin, S. Selim, and Islam M. Riad. 2025. "Machine Learning Approaches for Forecasting Compressive Strength of High-Strength Concrete." Scientific Reports 15(1). doi:10.1038/s41598-025-10342-1.

Ta, Thuy Trang, Thanh Loc Vo, and Ut Em Nguyen. 2024. Assessing Machine Learning Algorithms for Predicting Compressive Strength of Normal and High-Early Strength Concrete: A Case Study in Binh Thuan, Viet Nam. Vol. 20.

Wang, Jie, Junqi Deng, Siyi Li, Weijie Du, Zengqi Zhang, and Xiaoming Liu. 2025. "Explainable Machine Learning for Multicomponent Concrete: Predictive Modeling and Feature Interaction Insights." Materials 18(19):4456. doi:10.3390/ma18194456.

Wang, Liuyan, Lin Liu, Dong Dai, Bo Liu, and Zhenya Cheng. 2024. "Predictive Modeling of UHPC Compressive Strength: Integration of Support Vector Regression and Arithmetic Optimization Algorithm." Applied Sciences (Switzerland) 14(17). doi:10.3390/app14178083.

Willard, Jared, Xiaowei Jia, Shaoming Xu, Michael Steinbach, and Vipin Kumar. 2023. "Integrating Scientific Knowledge with Machine Learning for Engineering and Environmental Systems." ACM Computing Surveys 55(4). doi:10.1145/3514228.

Yeh, I. C. 1998. MODELING OF STRENGTH OF HIGH-PERFORMANCE CONCRETE USING ARTIFICIAL NEURAL NETWORKS.

Zhang, Wan, Jiangtao Guo, Cuiping Ning, Ruifang Cheng, and Ze Liu. 2024a. "Prediction of Concrete Compressive Strength Using a Deepforest-Based Model." Scientific Reports 14(1). doi:10.1038/s41598-024-69616-9.

Zhang, Wan, Jiangtao Guo, Cuiping Ning, Ruifang Cheng, and Ze Liu. 2024b. "Prediction of Concrete Compressive Strength Using a Deepforest-Based Model." Scientific Reports 14(1). doi:10.1038/s41598-024-69616-9.