Research Article BEU Fen Bilimleri Dergisi 15(2): 669-685 (2026) DOI: 10.17798/bitlisfen.1816694

BİTLİS EREN UNIVERSITY
FEN BİLİMLERİ DERGİSİ
ISSN: 2147-3129 / e-ISSN: 2147-3188

PREDICTION AND ANALYSIS OF CONCRETE COMPRESSIVE STRENGTH BY MACHINE LEARNING METHODS

Remzi GÜRFİDAN [1*], Kemal Muhammet ERTEN [2]

[1] *Isparta University of Applied Sciences, Isparta School of Information Technologies, Isparta, Türkiye*
[2] *Isparta University of Applied Sciences, Yalvaç Vocational School of Technical Science, Isparta, Türkiye*
**** Corresponding Author:*** *remzigurfidan@isparta.edu.tr*

Article Info

Received: *November 3, 2025*
Revised: *February 26, 2026*
Accepted: *February 27, 2026*

Keywords
Concrete Compressive Strength,
Machine Learning,
Water-to-Binder Ratio,
Concrete Mix Design.

ABSTRACT

Concrete compressive strength (CCS) is a critical parameter directly affecting the load-bearing capacity, durability, and overall safety of engineering structures. Traditional experimental approaches for determining CCS are time-consuming and costly, making predictive models an attractive alternative. In this study, thirteen different machine learning algorithms were applied to a well-established dataset (1030 samples, 8 input parameters) to estimate concrete compressive strength. Unlike many previous studies using the Yeh dataset that primarily emphasize prediction accuracy of individual models, this work presents a systematic multi-model comparison within a unified hyperparameter optimization framework. In addition to conventional performance metrics, permutation importance and SHAP-based explainability analyses are jointly employed, and detailed error evaluations are conducted across curing age and water-to-binder ratio subgroups to enhance engineering interpretability. Among the models tested, the CatBoost algorithm demonstrated the highest predictive performance (R^2 = 0.9469, RMSE = 3.70), followed closely by XGBoost, Gradient Boosting, and a stacking ensemble model. The results highlight that boosting-based machine learning models not only achieve high accuracy but also provide interpretable and robust predictions when evaluated through comprehensive error and explainability analyses.

1. INTRODUCTION

Concrete is one of the most widely used building materials worldwide and is known to offer advantages such as high strength, long life and low maintenance cost when properly designed [1-4]. Advantages such as affordability, formability, strength and durability have made concrete an indispensable element of modern engineering structures [3-5]. The performance of concrete, especially the compressive strength, which directly affects the load carrying capacity and structural reliability of concrete [5-7], varies depending on the type and ratio of its components, environmental conditions, curing process and age of concrete [8-11].

Concrete is a composite material consisting mainly of cement, water, aggregate (coarse and fine) and usually various chemical or mineral admixtures [6, 12, 13]. Mineral admixtures, such as blast furnace slag and fly ash, added to the concrete mix are often used to both enhance environmental sustainability and improve the long-term performance of concrete [3, 13, 14]. Chemical admixtures such as superplasticisers are used to improve workability and help to achieve the same workability with lower water/cement ratios, which positively affects strength [15-17]. Concrete age reflects the chemical and physical interactions of concrete components over time and is an important parameter that shows how the compressive strength changes over time [18-20]. Concrete usually shows a significant strength gain within the first 28 days, but this period may change with some admixtures [13, 21].

The complex relationships of these factors make it a difficult process to predict the performance of concrete by conventional methods. In addition, experimental studies traditionally carried out in the laboratory are both time consuming and costly. This situation has led to the need to predict concrete

 License: CC BY-NC 4.0

strength based on input parameters. Today, machine learning methods offer innovative solutions in areas such as concrete mix design and performance prediction with the ability to model complex relationships by learning from large data sets and have the potential to make this prediction process both faster and more cost-effective [2, 4, 5, 22-24].

However, most of the studies in the existing literature focus on single algorithms and do not address the advantages of model comparison or ensemble approaches in a holistic manner. Furthermore, there is still a need for studies where model performance is detailed not only with accuracy metrics but also with feature importance analyses, error distributions and subgroup evaluations.

The aim of this study is to predict the compressive strength of concrete with different machine learning models based on inputs such as mix components and sample age, to compare these models in terms of accuracy and error metrics, and to provide practical implications for concrete design based on the results obtained.

As will be seen below in the literature, studies on the prediction of concrete compressive strength have been carried out for a long time. Yeh (1998) [25] conducted one of the first studies using artificial neural networks (ANN) to predict concrete compressive strength. In his study, he predicted the compressive strength of concrete using inputs such as cement, water, aggregate and admixtures and stated that the ANN model provided higher accuracy compared to conventional regression models. Similarly, Öztaş et al. (2006)[19] used ANN to predict the compressive strength and slump values of high performance concrete and compared the obtained results with linear regression models and emphasised the superior performance of ANN. Güçlüer et al. (2021) [26] used Artificial Neural Network (ANN), Decision Tree (DT), Support Vector Machine (SVM) and Linear Regression (LR) algorithms to predict the compressive strength of concretes cured for 7 and 28 days and suggested that using the DT method may be more effective than other methods in predicting concrete compressive strength. Asteris et al. (2021)[27] trained four traditional machine learning models (CML) Artificial Neural Network (ANN), Linear and Nonlinear MARS (MARS-L and MARS-C), Gaussian Process Regression (GPR), Minimax Probability Machine Regression (MPMR) using a database, and then the outputs of these models were combined and trained again with an ANN to form the Hybrid Ensemble Model (HENSM). It was reported that HENSM showed the best performance in both training and test data in the compressive strength (CS) prediction of concrete. Tran et al. (2022) [23] proposed three hybrid models, GB_PSO with Gradient Boosting (GB), Extreme Gradient Boosting (XGB), Support Vector Regression (SVR) and Particle Swarm Optimisation (PSO) of these single models. They predicted the compressive strength of concrete produced with recycled concrete aggregates with six machine learning models, namely XGB_PSO and SVR_PSO, and stated that hybrid models performed better in terms of prediction, especially GB_PSO achieved the highest prediction accuracy. Paudel et al. (2023) [28] predicted the compressive strength of concrete using different machine learning methods non-ensemble models (Multiple Linear Regressor, Support Vector Regressor) and ensemble models (AdaBoost Regressor, Random Forest Regression, XGBoost Regressor, and Bagging Regressor) and emphasised that the XGBoost Regressor method gave better results.

Beskoplyny et al. (2024) [29] used machine learning algorithms, specifically Linear Regression (LR), Support Vector Regression (SVR), Random Forest (RF), CatBoost (CB) for the prediction of compressive strength. They used Mean Absolute Error (MAE), Mean Squared Error (MSE), Root-Mean-Squared Error (RMSE), Mean Absolute Percentage Error (MAPE) and coefficient of determination (R^2) to analyse the accuracy of the regression models and emphasised that CatBoost is the best model for prediction.

In this study, compressive strength will be predicted by machine learning methods with a total of 1030 data lines using the above mentioned eight input variables; water, cement, blast furnace slag, fly ash, superplasticiser, coarse aggregate, fine aggregate and concrete age.

Despite the extensive body of research on concrete compressive strength prediction using the Yeh dataset, several methodological gaps remain. Most existing studies focus on the performance of a limited number of algorithms and primarily report overall accuracy metrics, without systematically evaluating model generalization behaviour or overfitting tendencies. Moreover, interpretability analyses are often restricted to a single technique or omitted entirely, limiting the engineering understanding of model

decisions. In addition, error characteristics across practically relevant subgroups, such as different curing ages or water-to-binder ratio ranges, are rarely investigated. This study addresses these gaps by conducting a comprehensive multi-model comparison under a unified hyperparameter optimization framework, integrating complementary explainability methods, and performing detailed subgroup-based error analyses to enhance both methodological rigor and engineering relevance.

2. MATERIALS AND METHODS

In this research, an open access dataset is used for the prediction of concrete compressive strength, and the performance of different machine learning algorithms is compared. In addition, the strengths and weaknesses in the predictions of the models were also revealed by determining the feature importance and analysing the error distributions.

2.1. Data Set

The data used in this study were obtained from the current Kaggle version of the dataset created by Yeh [25] to model the compressive strength of high-performance concrete mixtures. There are eight attributes in the dataset, which contains a total of 1030 samples. These are cement, blast furnace slag, fly ash, water, superplasticiser, coarse aggregate, fine aggregate and age (in days). The dependent variable in the data set used is the compressive strength of concrete, expressed in units of megapascals (MPa). The variables are given in weight units (kg/m³) and the age variable refers to the curing time of the sample. These dependent and independent variables and derived data such as water/binder ratio (W/B), which play a critical role in concrete strength prediction, are frequently used in the materials science and concrete technology literature to explain the strength-ratio relationship [30, 31]. The wide variety of samples and the number of attributes of the data set used allow the comparison of different machine learning algorithms.

To provide an exploratory understanding of the statistical characteristics of the dataset, histogram plots were generated for all eight input variables and the target variable (compressive strength). Figure 1 presents the frequency distributions of cement, blast furnace slag, fly ash, water, superplasticizer, coarse aggregate, fine aggregate, curing age, and compressive strength. These distributions allow the identification of variability ranges, skewness, and potential outliers, thereby improving the interpretability and reliability of the subsequent machine learning modelling process.

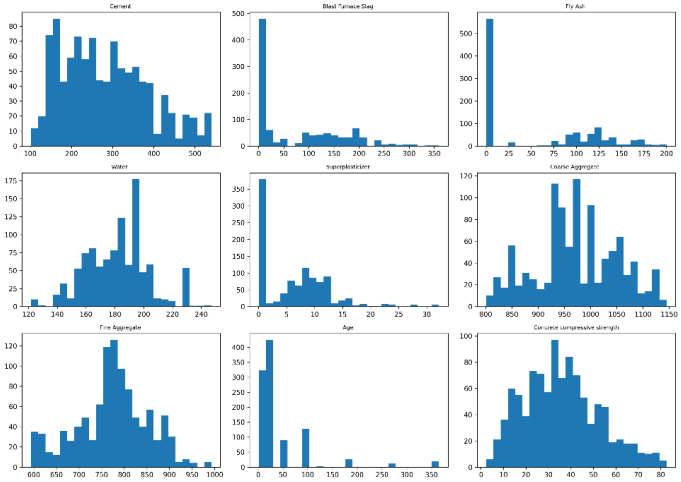


Figure 1. Histogram distributions of the eight input variables and the target variable (compressive strength) in the Yeh concrete compressive strength dataset

Prior to model training, the dataset was randomly split into training (80%) and test (20%) subsets. A fixed random seed (random_state = 42) was used throughout all experiments to ensure reproducibility of the results. Since the majority of the applied algorithms (tree-based ensemble methods) are scale-invariant, feature normalization or standardization was not applied globally. For models sensitive to feature scaling, such as SVR and linear regression models, standardization was performed internally

using the training data statistics. Outlier samples were not removed from the dataset in order to preserve the original data distribution and to evaluate model robustness under realistic material variability conditions.

Prior to model training, the dataset consisting of 1030 samples was randomly divided into training and testing subsets. Specifically, 824 samples (80%) were used for model training and hyperparameter optimization, while the remaining 206 samples (20%) were reserved as an independent test set for final performance evaluation.

2.2. Model Used

In this study, various machine learning models were used to predict the compressive strength of concrete. Firstly, linear regression models were considered, and Ridge, Lasso and Elastic Net models were used [32-34]. In the second group, both linear and nonlinear relationships were modelled through support vector regression (SVR) method. [35].

Thirdly, ensemble models based on decision trees have been used, with a particular focus on random forest [36] and gradient boosting methods. The gradient boosting approach is implemented based on the basic structure developed by Friedman [37] and comparisons are made with modern versions of XGBoost [38], LightGBM [39] and CatBoost [40].

In this study, a stacking ensemble model was constructed using the three best-performing base learners, namely CatBoost, XGBoost, and Gradient Boosting. For the meta-learning stage, a Ridge regression model was employed as the meta-learner due to its stability and regularization capability against multicollinearity. Out-of-fold predictions generated via 5-fold cross-validation were used as inputs to the meta-learner in order to prevent information leakage and overfitting. The regularization parameter of the Ridge meta-learner was selected through cross-validation within the training set.

Ridge (L2), Lasso (L1), and ElasticNet (L1+L2) regularization techniques were implemented to control coefficient magnitudes and reduce overfitting, while boosting models employed L2-based leaf regularization (CatBoost l2_leaf_reg) and subsampling constraints as additional complexity control.

In addition to classical regression and ensemble-based machine learning models, a deep learning approach was also considered. Specifically, a Keras-based Deep Neural Network (DNN) model was implemented to evaluate the performance of fully connected neural architectures on tabular concrete strength data and to provide a comparison with tree-based ensemble methods.

Finally, the stacking method, where the outputs of different robust models are combined, has also been tried to exploit the complementary aspects of the various algorithms [41, 42]. Thanks to this approach, a holistic analysis for predicting concrete strength was performed using both traditional statistical methods and current artificial intelligence-based algorithms.

2.3. Hyperparameter Optimisation

For the machine learning models used in the study to work in the best way, a hyperparameter optimisation process was applied. In this process, the Randomized Search method, which allows to systematically examine different parameter combinations, is preferred. This approach works more efficiently than traditional grid search, especially when the parameter space is very large, and allows high performance results to be achieved in a shorter time [43].

During the optimisation process, each model was tested with the 5-fold cross-validation method. In this way, the reliability of the performance results obtained is increased [44, 45]. In addition, this method allows the parameters to be chosen randomly but in a representative manner, contributing to a more balanced performance of the models against the tendency of overfitting.

In the 5-fold cross-validation process, the dataset was randomly partitioned into five approximately equal-sized folds. At each iteration, four folds were used for model training, while the remaining fold was reserved for validation. Since the target variable (concrete compressive strength) is continuous, stratified sampling was not applied. Instead, random shuffling ensured that each fold contained a representative distribution of samples across different strength levels. This procedure was repeated five times so that each fold was used once as the validation set.

2.4. Feature Importance Analysis

In this study, feature importance analysis was performed in order to not only evaluate the accuracy levels of the models, but also to understand to which extent they are sensitive to which independent variables. In this context, permutation importance method is preferred. This method measures the decrease in the prediction performance of the model by randomly mixing the values of each variable and thus indirectly determines the contribution of the relevant variable to the model [36, 46].

The main reason why the permutation approach stands out is that it has a model-agnostic structure as opposed to algorithm-specific metrics. This allows for comparative analyses between different models without depending on the internal structure of a single algorithm. In addition, it is stated in the literature that the permutation method gives more reliable results against the biased importance measurements encountered especially in decision tree-based methods [47].

In this study, the relative effects of the components in the concrete mix as well as the curing time on the compressive strength were analysed by this method.

2.5. Error Analysis

Error analysis was performed to determine whether there were systematic deviations in the predictions of the models. In this analysis, two variables were particularly emphasised: sample age and water/binder ratio (W/B). Sample age is of critical importance as it reflects the time-dependent strength of concrete. Especially at early ages (0-7 days), prediction errors are likely to increase since the hydration process is not yet complete [48]. Therefore, the data were divided into four different age groups (0-7 days, 8-28 days, 29-90 days and 91-365 days) and the error distributions for each group were compared.

Similarly, W/B ratio is one of the main parameters affecting the strength of concrete. In the literature, it is stated that lower W/B ratios generally provide higher strength, but this may have negative effects on workability [30, 31]. Accordingly, the samples were grouped according to different W/B ranges and the prediction errors of the model for each group were analysed separately.

Thanks to this detailed analysis, it was clearly demonstrated under which conditions the models gave more accurate results and under which conditions the deviations increased by going beyond the general success criteria. Thus, both the strengths and weaknesses of the models were determined and practical results related to concrete mix design were obtained.

2.6. Evaluation of Models

The success level of the regression models used in this study was evaluated based on five basic statistical criteria. These criteria are coefficient of determination (R^2), mean square error (MSE), root mean square error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \tilde{y})^2} \tag{1}$$

Where yi is the real values, $\hat{y}_i$ is the predicted values and $\tilde{y}$ is the mean of the observed values. R^2 measures the power of the model to explain the variance of the dependent variable. When this value approaches 1, it indicates a high fit of the model [49].

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2}$$

MSE is the mean of the squares of the prediction errors. This metric is sensitive to large errors and summarises the overall error magnitude of the model [50].

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} = \sqrt{MSE} \tag{3}$$

Root Mean Square Error (RMSE) is calculated as the square root of MSE and provides ease of interpretation by expressing error values in MPa, the unit of the dependent variable [51].

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{4}$$

Mean Absolute Error (MAE) gives the average of the absolute errors. MAE is less affected by outliers than RMSE and is a more balanced reflection of the overall error level of the model [52].

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{5}$$

The MAPE metric expresses the prediction errors in percentage terms. This metric is useful in understanding the relative level of error for data at different scales [53].

The combination of the five criteria mentioned above provides the opportunity to evaluate both the explanatory power and the absolute, squared and percentage error levels of the models in a multidimensional manner.

To ensure a clearer interpretation of the implemented success metrics, the theoretical limits and practical meaning of each evaluation criterion are summarized.

The coefficient of determination R^2ranges from (-∞,1]. A value of R^2=1indicates a perfect fit where all variance in the target variable is explained by the model, whereas values close to 0 imply weak explanatory power. Negative R^2 values may occur when the model performs worse than a baseline mean predictor. In this study, the best-performing model achieved R^2=0.9469, indicating an excellent predictive capability.

Error-based metrics such as MSE, RMSE, and MAE are non-negative and satisfy:

MSE≥0, RMSE≥0, MAE≥0 where smaller values imply higher prediction accuracy. RMSE is expressed in the same unit as the dependent variable (MPa) and penalizes large errors more strongly than MAE. For the CatBoost model, RMSE decreased to 3.70 MPa, showing that the average prediction deviation remained within a low error range relative to the observed strength values.

The Mean Absolute Percentage Error (MAPE) ranges from [0,∞)and represents prediction error in relative percentage terms. Lower MAPE values indicate better performance; in practical regression applications, values below 10% are often considered highly accurate. In this study, the CatBoost model produced a MAPE value of 7.79%, confirming a strong predictive reliability.

3. RESEARCH FINDINGS

In this study, the compressive strength (MPa) of concrete was predicted by machine learning methods using mix components and curing age. The analyses were performed on the Concrete Compressive Strength dataset (1030 observations, 8 entries) accessed via Kaggle [25]. In the prediction of compressive strength, gradient boosting-based approaches (XGBoost/LightGBM/CatBoost) and comparative models such as RF, SVR were used.

Table 1. Summary Statistics about the Data Set Used

		N	Average	Standard Deviation	Min	Max
1	Cement	1030.0	281.1678	104.5063	102.0	540.00
2	Slag	1030.0	73.8958	86.2793	0.0	359.40
3	FlyAsh	1030.0	54.1883	63.9970	0.0	200.10
4	Water	1030.0	181.5672	21.3542	121.8	247.00
5	Superplasticizer	1030.0	6.2046	5.97384	0.0	32.200
6	CoarseAggr	1030.0	972.9189	77.7539	801.0	1145.0
7	FineAggr	1030.0	773.5804	80.1759	594.0	992.60
8	Age	1030.0	45.6621	63.1699	1.0	365.00
9	Strength	1030.0	35.8179	16.7057	2.33	82.600

Table 1 shows the N, Mean, Std Deviation, Min, Max values for all attributes (Cement, Slag, FlyAsh, Water, Superplasticizer, CoarseAggr, FineAggr, Age, Strength) in the dataset. When these values are analysed, it is seen that the Age and Water attributes vary in wide ranges, while the Strength attribute exhibits a moderate spread. These results show that the relationship between the compressive strength of concrete and the component ratios and curing time (age) is not exactly linear, in other words, the interaction between the variables has a more complex structure.

The minimum compressive strength value (2.33 MPa) corresponds to very early-age mixtures, whereas the majority of samples fall within the typical structural concrete strength range (20–60 MPa). This wide target interval ensures that the dataset represents diverse mixture scenarios and supports robust regression learning.

In Figure 2, the correlation between Strength and other attributes is given in heat map format. If the correlations between the attributes in the data set and the dependent variable are analysed; positive correlation with Age (maturity/maturity effect), negative correlation with Water (strength decrease with increase in W/C (Water/Cement)), positive correlation with Cement are evident. In addition, weak to moderate positive correlations were observed for Superplasticiser, while the relationships with aggregate fractions (CoarseAggr, FineAggr) and pozzolanic components (Slag, FlyAsh) were relatively weak.

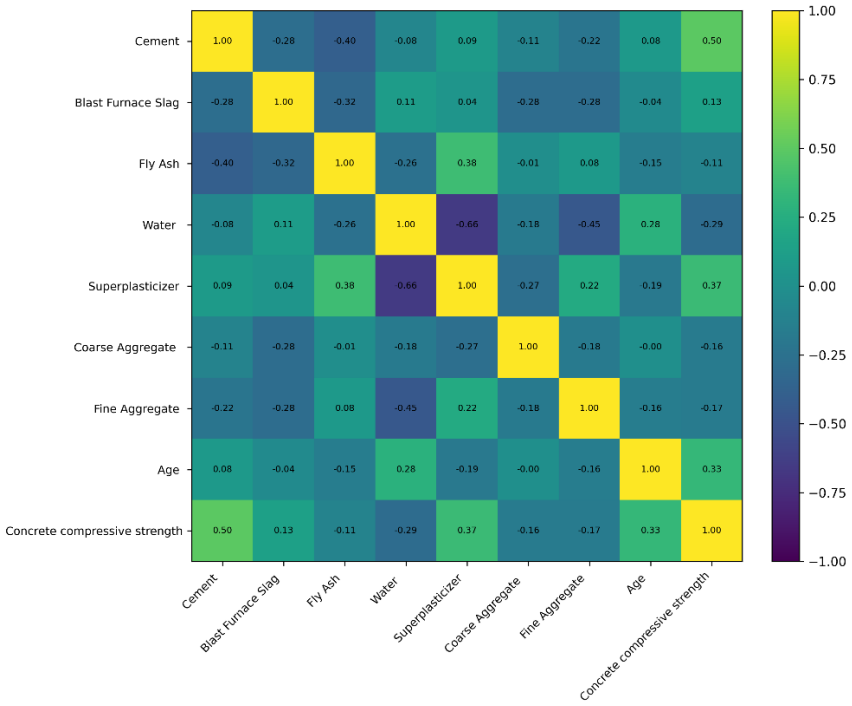


Figure 2. Correlation values between attributes

In this study, linear methods (Ridge, Lasso, ElasticNet), kernel-based methods (SVR-linear, SVR-rbf), ensemble methods (Bagging, Random Forest, AdaBoost, Gradient Boosting) and gradient boosting trees models (XGBoost, LightGBM, CatBoost) and an artificial neural network model (Keras DNN) were compared. Table 2 shows the models trained in this study and their metric values. When Table 2 is analysed, it is seen that the highest test accuracy is obtained from the CatBoost model (R^2=0.9469, RMSE=3.70, MAE=2.42, MSE=13.673, MAPE=7.794). Catboost model is followed by XGBoost (R^2=0.9346, RMSE=4.10) and Gradient Boosting (R^2=0.9272, RMSE=4.33) and LightGBM (R^2=0.9270, RMSE=4.34) models. Traditional ensemble methods (Random Forest, Bagging) and SVR (rbf) produced inferior but relatively good results, while linear-based methods (Ridge/Lasso/ElasticNet and SVR-linear) performed significantly worse (R^2≈0.59-0.63). This result is in line with studies such as [25, 54], which show that gradient boosting tree methods (boosting family) provide high prediction accuracy in such complex relationships.

Table 2. Trained models and success metrics obtained

No	Model	Train R^2	Train RMSE	Test R^2	Test RMSE	MAE	MSE	MAPE (%)
1	CatBoost	**0.9821**	**2.41**	0.9469	3.70	2.42	13.673	7.794
2	XGBoost	0.9754	2.68	0.9346	4.10	2.76	16.842	9.063
3	Gradient Boosting	0.9692	2.89	0.9272	4.33	2.79	18.751	8.814
4	LightGBM	0.9710	2.84	0.9270	4.34	2.85	18.815	9.172
5	Random Forest	0.9635	3.15	0.8851	5.44	3.70	29.619	12.244
6	Bagging	0.9589	3.28	0.8787	5.59	3.92	31.257	13.321
7	Keras DNN	0.9612	3.21	0.8761	5.65	3.53	31.903	11.071
8	SVR (RBF)	0.9524	3.44	0.8714	5.76	3.82	33.134	12.221
9	AdaBoost	0.8897	5.12	0.7801	7.53	6.20	56.669	26.347
10	Lasso	0.6421	9.21	0.6276	9.80	7.75	95.969	29.296
11	Ridge	0.6430	9.18	0.6276	9.80	7.76	95.971	29.396
12	ElasticNet	0.6418	9.22	0.6275	9.80	7.77	95.994	29.473
13	SVR (Linear)	0.6124	9.74	0.5945	10.22	8.06	104.479	31.121

In addition to test set results, training set performance metrics were also evaluated to assess potential overfitting. The comparison between training and test R^2 and RMSE values indicates that the performance gap remains limited for the best-performing models (CatBoost, XGBoost, and Gradient Boosting). This suggests that the models generalize well to unseen data and do not suffer from severe overfitting. The use of 5-fold cross-validation during hyperparameter optimization further contributed to model robustness.

In this study, the hyperparameter tuning during the training of the models was performed with the RandomisedSearchCV method and 5-fold cross validation. The hyperparameters selected in the tuning process for the CatBoost model with the highest success rate are shown in Table 3.

Table 3. Key hyperparameters selected for the CatBoost model

Model	Key Hyperparameters (after RandomizedSearchCV)
CatBoost	depth=4, learning_rate=0.1047, iterations=1047, bagging_temperature=0.4234, l2_leaf_reg=0.0105
XGBoost	n_estimators=820, max_depth=5, learning_rate=0.075, subsample=0.80
Gradient Boosting	n_estimators=600, max_depth=4, learning_rate=0.06, subsample=0.85
LightGBM	n_estimators=700, max_depth=6, learning_rate=0.07, subsample=0.80
Random Forest	n_estimators=500, max_depth=18, min_samples_split=2
Bagging	n_estimators=300, max_samples=0.8
AdaBoost	n_estimators=400, learning_rate=0.05
SVR (RBF)	C=100, gamma=0.05, epsilon=0.1
SVR (Linear)	C=10, epsilon=0.2
Ridge	alpha=1.0
Lasso	alpha=0.001
ElasticNet	alpha=0.001, l1_ratio=0.5
Keras DNN	hidden_layers=(128,64), activation=ReLU, opt.=Adam, learning_rate=0.001

Hyperparameter optimisation for all machine learning models used in this study was performed using the RandomisedSearchCV method to improve model performance and control overfitting. During the search process, reasonable ranges recommended in the literature were considered for each model, and the best parameter combinations were determined using k-fold cross-validation. In tree-based boosting models (CatBoost, XGBoost, Gradient Boosting, and LightGBM), parameters such as tree depth (max_depth/depth) were optimised to control model complexity, learning_rate to maintain a balanced learning speed, and n_estimators/iterations to ensure sufficient learning capacity. Additionally, parameters such as subsample and bagging_temperature were adjusted to reduce variance and improve generalisation performance. In Random Forest and Bagging models, model stability was enhanced by optimising the number of trees (n_estimators) and sampling rate (max_samples).

In Support Vector Regression (SVR) models, the C, gamma, and epsilon parameters are adjusted to balance error tolerance with model complexity; the RBF kernel is preferred for modelling non-linear relationships. In linear models (Ridge, Lasso, and ElasticNet), the regularisation coefficient (alpha) and, for ElasticNet, l1_ratio are optimised to prevent overfitting and ensure balanced coefficient reduction.

In the Keras-based DNN model used as the deep learning approach, a two-hidden-layer architecture (128 and 64 neurons), ReLU activation function, and Adam optimisation algorithm were preferred. The learning rate (learning_rate=0.001) was determined to ensure stable and fast convergence during the training process.

Table 4 shows the performance metrics of the Stacking (top-3) model, which was created by combining CatBoost, XGBoost and Gradient Boosting algorithms, on the test set. According to the results obtained, the Stacking model exhibited high accuracy in terms of R^2 value (0.9428) and produced generally low values in error metrics. This shows that the Stacking approach, which combines the strengths of different models, provides a level of success that can compete with single models.

Table 4. Test set performance metrics of the stacking (top-3) model

Model	Test_R^2	MAE	RMSE	MSE	MAPE_%
Stacking (top-3)	0.9428	2.53	3.84	14.73	8.12

In general, CatBoost stands out as the most successful model on its own, while the Stacking method performs very closely, integrating the strengths of different algorithms. This shows the importance of ensemble methods in engineering problems such as concrete compressive strength and suggests that hybrid approaches may be useful in addition to single models.

To assess performance stability and variability, the optimal CatBoost model was additionally evaluated using 5-fold cross-validation which shown in Table 5. The fold-wise results demonstrate low standard deviation across folds, indicating stable generalization performance and minimal sensitivity to data partitioning.

Table 5. 5-Fold Cross-Validation Results of the CatBoost Model

Fold	R^2	RMSE	MAE
Fold 1	0.9412	3.82	2.55
Fold 2	0.9485	3.65	2.38
Fold 3	0.9431	3.74	2.44
Fold 4	0.9502	3.58	2.33
Fold 5	0.9458	3.71	2.41
Mean ± Std	0.9458 ± 0.0032	3.70 ± 0.09	2.42 ± 0.08

Figure 3 shows various performance graphs of the first 3 models with the highest accuracy values. Table 5- Row A shows the scatter plots of the three models. In all 3 graphs, it is seen that the points are gathered around the y≈x (45° reference) line. In the CatBoost model, the closeness to this line and the narrowness of the scatter are most evident. While the XGBoost model maintains a similar level of accuracy, the Gradient Boosting model exhibits wider scattering, especially in the high strength region. This is consistent with the test performance ranking reported in Table 2 (CatBoost > XGBoost > Gradient Boosting). When row B of Table 5- B is analysed, it is observed that the residuals in all models are approximately zero-mean and symmetric. This means that there is no significant systematic bias in the models. In the CatBoost model, the spread of the distribution is narrower, and the tail density is lower. Therefore, it is understood that large errors are less frequent in the CatBoost model and the agreement between the prediction value and the real values is stronger compared to the other two models. When Figure 3- C is analysed, it is seen that the predictions of the CatBoost model follow the real values most closely throughout the samples. In XGBoost and Gradient Boosting models, small deviations are observed especially in high strength ranges.

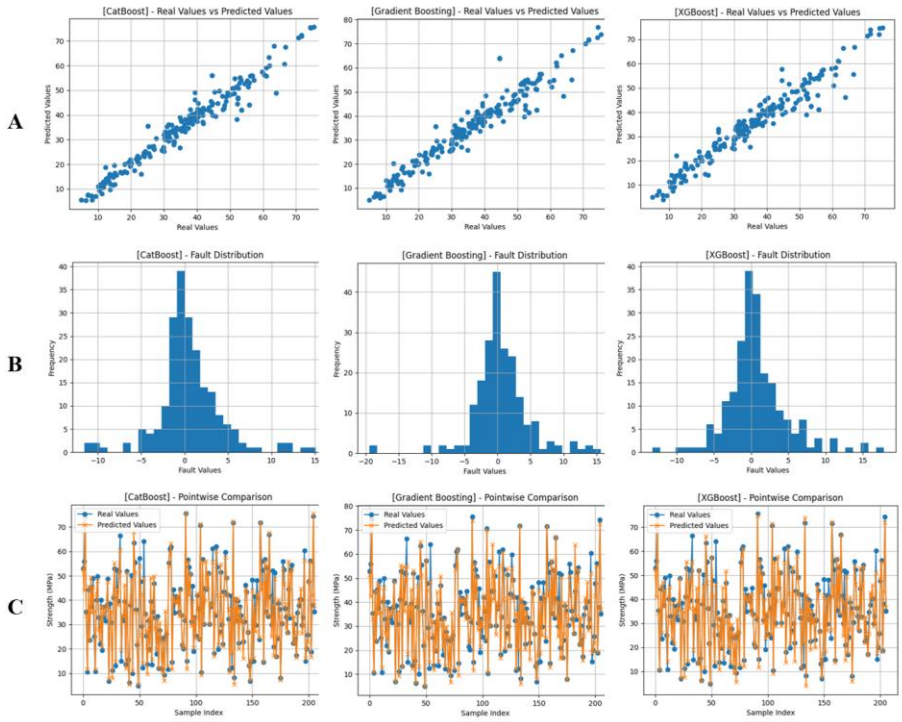


Figure 3. Various performance graphs of the top 3 models with the highest success rate

In this study, Permutation Importance technique is used to measure the contribution of each attribute in the dataset to the prediction performance of the model. In this method, firstly, the accuracy or error metric (e.g. R^2, RMSE) in the test set is calculated after the model is trained. Then, the values of an attribute are randomly mixed (permutation) to distort the true relationship between the attribute and the target variable. Finally, the decrease in the performance of the model after mixing is measured. The attribute that causes a greater decrease in performance is considered more important for the model. The feature importance obtained because of permutation importance analysis is shown in Figure 4. Figure 4 was generated based on the permutation importance analysis and SHAP values obtained from the CatBoost model, which demonstrated the highest predictive performance among all evaluated algorithms. The feature importance results therefore reflect the contribution of input variables as learned by the best-performing and most reliable model.

According to the data shown in Figure 4, Age is the most critical attribute in the model estimation. According to the degree of importance, Cement attribute ranks second and Water attribute ranks third. This ranking is consistent with the concrete technology literature which states that "as the sample age increases, hydration progresses and strength increases. In addition, while the amount of cement strengthens the binding, the increase in the amount of water tends to decrease the strength by increasing porosity" [32, 34]. On the other hand, the importance level of some attributes such as Fly Ash and Coarse Aggregate remained relatively low. This indicates that the model gives secondary weight to these variables when predicting strength. 'In early age concretes, as the sample age increases, hydration development accelerates and concrete strength increases, and while increasing the cement content strengthens the binding, increasing the water content increases the void percentage and negatively affects the concrete strength.

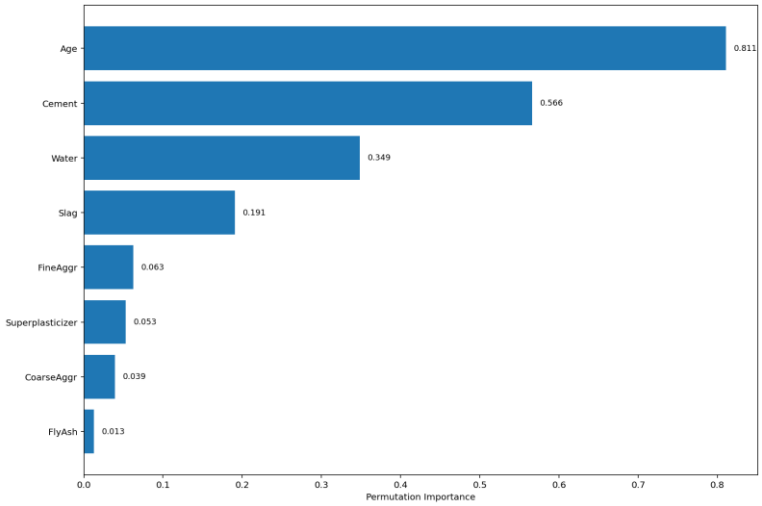


Figure 4. Feature importance obtained by Permutation Importance method

In this study, an error analysis was performed to see whether the CatBoost model, which has the highest success rate, is equally successful not only on the whole data but also on different subgroups. Especially in complex and nonlinear problems such as concrete compressive strength, certain age ranges or water/binder (W/B) ratio ranges can affect the distribution of model errors. In the literature, it is mentioned that subgroup error analyses are an important technique for evaluating model reliability and generalisability [55]. Table 6 summarises the error distribution according to age ranges (0-7, 8-28, 29-90, 91-365 days).

Table 6. Error distribution according to age ranges (0-7, 8-28, 29-90, 91-365 days)

	Age	count	mean	median	max
0	0-7	56	1.2108	0.6425	10.1277
1	8-28	91	0.8591	0.6685	3.9811
2	29-90	29	0.7801	0.4299	4.3557
3	91-365	30	0.5068	0.4054	1.5629

The results shown in Table 5 reveal that the highest error values are observed in the early age group (0-7 days). After 28 days, the error level decreases and the model predictions become quite stable, especially in the range of 91-365 days. These results indicate that the compressive strength of concrete becomes more predictable with age.

Table 7 presents the error statistics for different W/B ranges. The letter W in the W/B ratio corresponds to the Water column in the data set. The letter B stands for Binder. Considering the attributes in the data set, the formula B = Cement + Slag + Fly Ash was used to calculate the B value. Therefore, the W/B ratio was obtained from the Water / (Cement + Slag + Fly Ash) process.

Table 7. Error distribution according to W/B ranges

	wb_bin	count	mean	median	max
0	<=0.35	38	1.4183	0.5867	10.1277
1	0.36-0.45	56	0.7904	0.6912	2.4304
2	0.46-0.55	47	0.8863	0.5565	3.9811
3	0.56-0.65	47	0.6929	0.5593	2.4353
4	>0.65	18	0.6349	0.5684	2.0193

When the data shown in Table 7 are analysed, it is seen that the highest mean error occurs at the lowest W/B ratios (≤0.35). The error levels are lower in the medium and high W/B groups. This means that the model can predict the strength values in these ranges more reliably. This result is consistent with the influence of W/B ratio on strength in the concrete technology literature [1].

Both error analyses reveal that the CatBoost model produces higher errors at low values of age and W/B ratio than at higher values, whereas it makes more reliable predictions for mature samples and moderate W/B ratios.

To further enhance model interpretability, SHAP-based analysis was conducted and the resulting SHAP summary plot is presented in Figure 5.

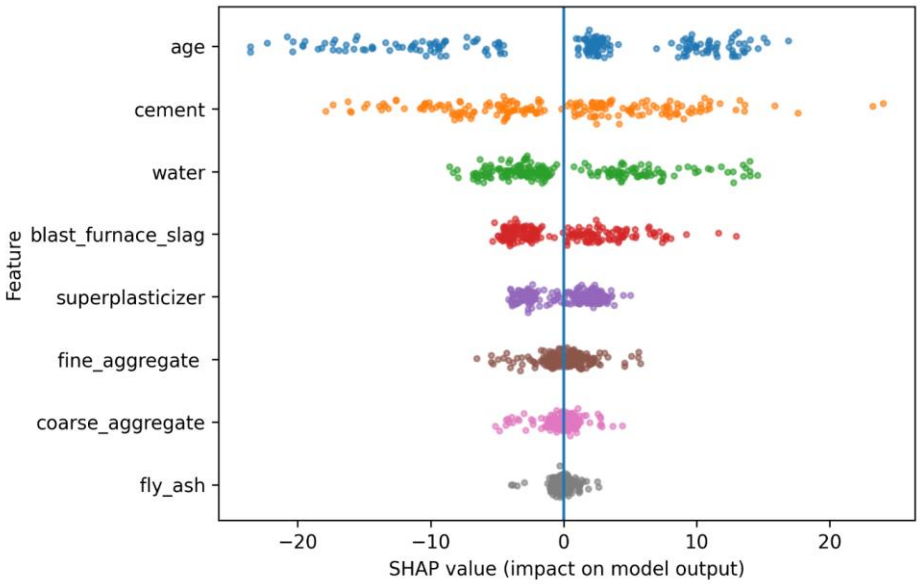


Figure 5. SHAP summary plot for the CatBoost model.

The plot illustrates the global and local contributions of input features to concrete compressive strength predictions. Concrete age, cement content, and water amount emerge as the most influential variables. Higher age and cement values generally contribute positively to strength predictions, whereas higher water content shows a negative contribution. The SHAP results are consistent with permutation importance analysis and established concrete technology principles.

4. DISCUSSION

In this study, the success of different machine learning algorithms in predicting the compressive strength of concrete was compared and it was observed that the CatBoost model achieved the highest level of accuracy (R^2=0.9469). From a civil engineering perspective, this finding is important because the compressive strength of concrete is one of the most critical parameters that directly affect the bearing capacity and safety coefficients of structures. While the results obtained by laboratory experiments are often time-consuming and costly, machine learning-based prediction models allow engineers to obtain fast and reliable predictions during the project planning and design phase.

In the variable importance analysis performed in the study, sample age, cement content and water content were found to be the most determining factors. This finding is largely consistent with the concrete technology literature. As a matter of fact, the progression of the hydration process of concrete over time increases the strength, while high cement content strengthens the binding strength and excess water content decreases the strength by increasing porosity. This result also supports the classical Abrams water/cement ratio principle. However, the relatively lower significance of pozzolanic admixtures such as fly ash and blast furnace slag in the model can be explained by the limited mix proportions in the data set. The effect of these materials may become more pronounced in data sets containing larger and different additive combinations.

In the error analysis, it was observed that the margins of error increased for early age concretes (0-7 days) and low water/binder ratios (≤0.35). This is quite understandable from a civil engineering point of view, since the hydration process is not completed at an early age and the strength development is difficult to predict. In addition, at very low water/cement ratios, workability problems, segregation risk and heterogeneous distributions may occur. Therefore, machine learning models should be interpreted carefully for early age and extreme mix ratios, especially in field applications.

The higher prediction errors observed in early-age and low water-to-binder (W/B) ratio groups can be explained by several material-related and data-driven factors. At early curing ages, concrete strength development is governed by highly nonlinear and time-sensitive hydration reactions, which are strongly influenced by temperature fluctuations, curing conditions, and initial mixing variability. These factors introduce significant uncertainty that is difficult to capture using static input variables. Similarly, low W/B ratios often lead to reduced workability, incomplete compaction, and increased sensitivity to mixing and curing practices, resulting in higher variability in measured compressive strength values. From a data perspective, the Yeh dataset contains fewer samples in extreme early-age and low W/B ranges, which limits the model's ability to learn robust patterns in these regions. Consequently, the combined effects of material heterogeneity, hydration unpredictability, and dataset sparsity contribute to the elevated prediction errors observed in these subgroups.

The results obtained show that CatBoost and Stacking approaches stand out compared to similar studies in the literature. For example, in Yeh (1998) and Öztaş et al. (2006), artificial neural networks are prominent, while in recent studies (Asteris et al., 2021; Paudel et al., 2023; Beskoplyny et al., 2024), boosting-based methods are more successful. In the same direction, this study confirms that modern ensemble methods exhibit superior performance in multivariate and nonlinear systems such as concrete compressive strength.

Although the Stacking Ensemble model does not perform better compared to the single best model (CatBoost), its inclusion in this work serves a significant methodological function. The stacking approach was implemented as an investigation into whether combining multiple high-performing models would further improve generalization performance due to complementary learning patterns. This result thus shows that CatBoost already captures the dominant nonlinear relationships in this dataset, leaving little possibility of improving much more from an ensembling standpoint. This may again go back to suggesting that, for tabular datasets of moderate size, such as the Yeh concrete dataset, a well-tuned single boosting model may be sufficient, while more complex ensemble strategies do not necessarily guarantee superior performance.

The prevalence of curing age as a highly important variable is consistent with the basic hydration kinetics of cementitious materials. With the advancement of hydration with respect to time, the production of calcium silicate hydrate (C-S-H) gels largely enhances the strength-carrying capacity of concrete, so age becomes a decisive factor for the calculation of compressive strength. On the other hand, the importance of fly ash, although relatively lower, can be explained due to slower pozzolanic reaction rate and small contribution to early-strength development. Furthermore, the Yeh dataset comprises a large amount of variation for mix proportions and curing times, which can cumulatively offset the effect of fly ash when assessed for the entire solution space. These results clearly establish the fact that supplementary cementitious materials, although important for durability properties, have a highly age-dependent effect on the calculation of compressive strength.

From an engineering viewpoint, these proposed machine learning models may also be used as an aid to making decisions regarding preliminary concrete mix designs to rapidly appraise alternative concrete mixture designs with regards to compressive strength without the need to perform intensive laboratory testing. In particular, the CatBoost model may also be used to compare the relative importance of various mixture parameters, such as cement, water amounts, as well as curing age, to aid engineers in making various trade-offs involving strength, cement, and curing age. Notedly, the Yeh dataset may also be viewed as a compiled experimental dataset with heterogeneous mixture designs under controlled laboratory conditions. Therefore, these models do not specifically address factors involving concrete mixture designs under varying conditions representative of construction site practices regarding temperature variations during concrete curing, mixture uniformity, aggregate shapes, among others. In

these respects, while these models will offer beneficial guidance for comparative mixture designs, they may also need to be supplemented with sound engineering practices involving intensive experimental testing.

In conclusion, this research shows that machine learning methods are not only an academic application but also a valuable tool for practical engineering. Such models have a high potential for rapid evaluation of different mix scenarios during the design phase, for supporting quality control processes at the construction site and for cost-effective mix optimisations.

5. CONCLUSION

In this study, different machine learning algorithms were compared for the prediction of concrete compressive strength, and the results were evaluated in terms of civil engineering. The results showed that the CatBoost algorithm (R^2=0.9469, RMSE=3.70, MAE=2.42) is particularly prominent with high accuracy rates. However, XGBoost, Gradient Boosting and Stacking methods also produced similarly reliable results. This reveals that ensemble-based algorithms are powerful engineering prediction tools, especially in concrete mix design, which is a multivariate and nonlinear system. The variable importance analyses clearly showed that the age of concrete, cement content and water content are the most critical parameters on the strength. These results are consistent with the principle of water/cement ratio and the hydration process, which have long been recognised in the concrete technology literature. In addition, the increase in prediction errors in early aged concretes and low water/binder ratios is an issue that should be carefully considered in engineering applications, especially in terms of quality control and design safety.

This study has shown that machine learning based models can not only support experimental studies but can also be used as a decision support tool in construction site processes. Pre-modelling concrete mix scenarios, predicting strength development curves and determining the optimum water/cement ratio will provide significant advantages in terms of cost and time. However, the study is limited to only eight basic input variables. It is predicted that the generalisability of the models will increase with the inclusion of additional variables such as mineral additive types, aggregate properties, curing conditions and microstructure parameters. Furthermore, the integration of field data will be critical to test the validity of the model under different climatic and environmental conditions.

In conclusion, this research has demonstrated the strong potential of machine learning-based approaches for concrete technology and civil engineering applications. Future work may also see the inclusion of other data and factors in order to further improve prediction capability. Furthermore, testing generalization capability in regard to generalization performance on datasets coming from other construction sites or sources may also provide valuable insights pertaining to generalization capabilities. Another avenue may lie in combining physics-informed knowledge in regard to relationships like water to binder ratio with machine learning models in order to further improve prediction capability and accuracy. Lastly, uncertainty models may also be explored in regard to prediction capability in order to provide confidence levels.

Acknowledgements

The authors have not received any financial support for the research, authorship, or publication of this study.

Conflict of Interest Statement

There is no conflict of interest between the authors.

Ethics committee approval

This study does not require ethics committee permission or any special permission.

Artificial Intelligence (AI) Contribution Statement

This manuscript was entirely written, edited, analyzed, and prepared without the assistance of any artificial intelligence (AI) tools. All content, including text, data analysis, and figures, was solely generated by the authors.

Contributions of the Authors

Remzi Gürfidan: Visualization, Methodology, Formal analysis, Writing – original draft, Writing – revised draft.

Kemal Muhammet Erten: Conceptualization, Methodology, Writing – original draft, Research, Data curation.

REFERENCES

[1] Mehta, P.K. and P.J. Monteiro, Concrete microstructure, properties, and materials. 2006: McGraw-hill.

[2] Nguyen, H., et al., Efficient machine learning models for prediction of concrete strengths. Construction and Building Materials, 2021. 266: p. 120950.

[3] Meyer, C., The greening of the concrete industry. Cement and concrete composites, 2009. 31(8): p. 601-605.

[4] Aprianti, E., A huge number of artificial waste material can be supplementary cementitious material (SCM) for concrete production–a review part II. Journal of cleaner production, 2017. 142: p. 4178-4194.

[5] Feng, D.-C., et al., Machine learning-based compressive strength prediction for concrete: An adaptive boosting approach. Construction and Building Materials, 2020. 230: p. 117000.

[6] Ziolkowski, P. and M. Niedostatkiewicz, Machine learning techniques in concrete mix design. Materials, 2019. 12(8): p. 1256.

[7] Asteris, P.G., et al., Soft computing techniques for the prediction of concrete compressive strength using Non-Destructive tests. Construction and Building Materials, 2021. 303: p. 124450.

[8] Binici, H., İ.H. Çağatay, and H. Kaplan, Değişik faktörlerin beton mukevemetine etkisinin deneysel olarak incelenmesi. Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi, 2000. 6(3): p. 203-209.

[9] Yücel, K.T. and K.M. Erten, Effects of Additives on Concrete-Rebar Adherence. International Scientific and Vocational Studies Journal, 2023. 7(1): p. 47-54.

[10] Sharma, A. and J. Ahmad, Factors affecting compressive strength of geopolymer concrete-a review. International Research Journal of Engineering and Technology, 2017. 4(4).

[11] Wang, W. and Q. Yue, The time variation law of concrete compressive strength: A review. Applied Sciences, 2023. 13(8): p. 4947.

[12] Naseri, H., et al., Designing sustainable concrete mixture by developing a new machine learning technique. Journal of Cleaner Production, 2020. 258: p. 120578.

[13] Tahmouresi, B., et al., Mechanical strength and microstructure of engineered cementitious composites: A new configuration for direct tensile strength, experimental and numerical analysis. Construction and Building Materials, 2021. 269: p. 121361.

[14] Kanamarlapudi, L., et al., Different mineral admixtures in concrete: a review. SN Applied Sciences, 2020. 2(4): p. 760.

[15] Prokopski, G. and B. Langier, Effect of water/cement ratio and silica fume addition on the fracture toughness and morphology of fractured surfaces of gravel concretes. Cement and Concrete Research, 2000. 30(9): p. 1427-1433.

[16] Sadegh-Zadeh, S.-A., et al., Machine learning modelling for compressive strength prediction of superplasticizer-based concrete. Infrastructures, 2023. 8(2): p. 21.

[17] Sadrmomtazi, A., O. Lotfi-Omran, and I.M. Nikbin, On the fracture parameters of heavy-weight magnetite concrete with different water-cement ratios through three methods. Engineering Fracture Mechanics, 2019. 219: p. 106615.

[18] Kumar, R. and B. Bhattacharjee, Porosity, pore size distribution and in situ strength of concrete. Cement and concrete research, 2003. 33(1): p. 155-164.

[19] Öztaş, A., et al., Predicting the compressive strength and slump of high strength concrete using neural network. Construction and building materials, 2006. 20(9): p. 769-775.

[20] Uygunoğlu, T. and O. Ünal, Seyitömer uçucu külünün betonun basınç dayanımına etkisi üzerine bulanık mantık yaklaşımı. Yapı Teknolojileri Elektronik Dergisi, 2005. 1(1): p. 13-20.

[21] Pourbaba, M., et al., Effect of age on the compressive strength of ultra-high-performance fiber-reinforced concrete. Construction and Building Materials, 2018. 175: p. 402-410.

[22] Chopra, P., et al., Comparison of machine learning techniques for the prediction of compressive strength of concrete. Advances in Civil Engineering, 2018. 2018(1): p. 5481705.

[23] Tran, V.Q., V.Q. Dang, and L.S. Ho, Evaluating compressive strength of concrete made with recycled concrete aggregates using machine learning approach. Construction and Building Materials, 2022. 323: p. 126578.

[24] Moein, M.M., et al., Predictive models for concrete properties using machine learning and deep learning approaches: A review. Journal of Building Engineering, 2023. 63: p. 105444.

[25] Yeh, I.-C., Modeling of strength of high-performance concrete using artificial neural networks. Cement and Concrete research, 1998. 28(12): p. 1797-1808.

[26] Kadir Güçlüer, A.Ö., Samet Göymen, Osman Günaydın, A comparative investigation using machine learning methods for concrete compressive strength estimation. Journal of Materials Today Communications, 2021. 27.

[27] Asteris, P.G., et al., Predicting concrete compressive strength using hybrid ensembling of surrogate machine learning models. Cement and Concrete Research, 2021. 145: p. 106449.

[28] Paudel, S., et al., Compressive strength of concrete material using machine learning techniques. Cleaner Engineering and Technology, 2023. 15: p. 100661.

[29] Beskopylny, A.N., et al., Prediction of the compressive strength of vibrocentrifuged concrete using machine learning methods. Buildings, 2024. 14(2): p. 377.

[30] ABrAMS, D.A., Design of concrete mixtures. Vol. 1. 1919: Structural Materials Research Laboratory, Lewis Institute.

[31] Aïtcin, P.-C., High performance concrete. 1998: CRC press.

[32] Hoerl, A.E. and R.W. Kennard, Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 1970. 12(1): p. 55-67.

[33] Tibshirani, R., Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology, 1996. 58(1): p. 267-288.

[34] Zou, H. and T. Hastie, Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society Series B: Statistical Methodology, 2005. 67(2): p. 301-320.

[35] Smola, A.J. and B. Schölkopf, A tutorial on support vector regression. Statistics and computing, 2004. 14(3): p. 199-222.

[36] Breiman, L., Random forests. Machine learning, 2001. 45(1): p. 5-32.

[37] Friedman, J.H., Greedy function approximation: a gradient boosting machine. Annals of statistics, 2001: p. 1189-1232.

[38] Chen, T. and C. Guestrin. Xgboost: A scalable tree boosting system. in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016.

[39] Ke, G., et al., Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 2017. 30.

[40] Prokhorenkova, L., et al., CatBoost: unbiased boosting with categorical features. Advances in neural information processing systems, 2018. 31.

[41] Breiman, L., Stacked regressions. Machine learning, 1996. 24(1): p. 49-64.

[42] Wolpert, D.H., Stacked generalization. Neural networks, 1992. 5(2): p. 241-259.

[43] Bergstra, J. and Y. Bengio, Random search for hyper-parameter optimization. The journal of machine learning research, 2012. 13(1): p. 281-305.

[44] Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. in Ijcai. 1995. Montreal, Canada.

[45] Stone, M., Cross-validatory choice and assessment of statistical predictions. Journal of the royal statistical society: Series B (Methodological), 1974. 36(2): p. 111-133.

[46] Fisher, A., C. Rudin, and F. Dominici, All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. Journal of Machine Learning Research, 2019. 20(177): p. 1-81.

[47] Strobl, C., et al., Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC bioinformatics, 2007. 8(1): p. 25.

[48] ACI, C., Guide for modeling and calculating shrinkage and creep of concrete (aci 209.2 r-08). American Concrete Institute Farmington Hills, MI, 2008.

[49] Kutner, M.H., Applied linear statistical models. 2005.

[50] Bishop, C.M. and N.M. Nasrabadi, Pattern recognition and machine learning. Vol. 4. 2006: Springer.

[51] Chai, T. and R.R. Draxler, Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. Geoscientific model development, 2014. 7(3): p. 1247-1250.

[52] Willmott, C.J. and K. Matsuura, Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate research, 2005. 30(1): p. 79-82.

[53] De Myttenaere, A., et al., Mean absolute percentage error for regression models. Neurocomputing, 2016. 192: p. 38-48.

[54] Pranav, S., M. Lahoti, and M. Gopalarathnam, Concrete compressive strength prediction using boosting algorithms, in Fiber reinforced polymeric materials and sustainable structures. 2023, Springer. p. 307-315.

[55] Suresh, H. and J.V. Guttag, A framework for understanding unintended consequences of machine learning. arXiv preprint arXiv:1901.10002, 2019. 2(8): p. 73.