



Comparative study of machine learning methods to predict compressive strength of high-performance concrete and model validation on experimental data

Tushar Bansal¹ · Visalakshi Talakokula² · T. Jothi Saravanan³

Received: 13 July 2023 / Accepted: 17 July 2023 / Published online: 25 July 2023
© The Author(s), under exclusive licence to Springer Nature Switzerland AG 2023

Abstract

Concrete compressive strength (CCS) is one of the most important parameters to determine the performance of concrete during service conditions. To accurately predict the compressive strength of the entire concrete system makes it a great challenge for a sustainable built environment and future generations since the materials are randomly distributed materials throughout the concrete. In this study, a comparative analysis for predicting the compressive strength of high-performance concrete has been carried out using various machine learning methods. Further, the top-performing models are hyper-parameter optimized to improve the accuracy of the model. To understand the importance of each feature in the trained model, feature selection is done based on the best-performing model. The result indicates that the Gradient Boosted Tree algorithm performs best with 0.94 R^2 , and the most important features for concrete compressive strength prediction are the age of concrete, cement, and water, and the least important feature is coarse aggregate. Hence, the Gradient Boosted Tree algorithm can be used to predict the compressive strength of concrete which helps the contractors to reduce the cost and time in concrete mix designing and prevent the unnecessary wastage of material caused by numerous mixture trials.

Keywords Concrete compressive strength · High-performance concrete · Optimization · Regression · Machine learning

Introduction

Concrete is considered a highly used building construction material worldwide due to its several advantages over other materials (Berodier et al., 2019; Larsen et al., 2019; Shamsutdinova et al., 2019; Yoon & Kim, 2019). In recent times, researchers have put enormous effort into improving concrete sustainability, fresh properties (including rheology,

stability, and setting) and hardened properties (including strength and durability) by substituting the cement with different supplementary cementitious material (SCMs) (Kaplan & Salem Elmekahal, 2021; Sivamani & Renganathan, 2021). Among the various properties of concrete, compressive strength is one of the most widely used mechanical properties of concrete, and it is directly related to the safety of the structures. Strong concrete based on specific compressive strength is still required (Al-Shamiri et al., 2019; Liu & Li, 2019; Yu et al., 2019; Yuan et al., 2019) since insufficient concrete compressive strength can lead to catastrophic civil infrastructure failures. However, as it is known to all, concrete is made up of various materials such as cement, blast furnace slag fly ash, water, superplasticizer, and coarse and fine aggregate, and these materials are randomly distributed throughout the entire concrete system. To accurately predict the compressive strength of this entire concrete system makes it a significant challenge.

Generally, concrete compressive strength (CCS) can be obtained through physical experiments by preparing the concrete cube or cylinder according to the mix design and then cured for the required time. However, this method is

✉ Tushar Bansal
tushar.bennett@gmail.com

Visalakshi Talakokula
basavishali@gmail.com

T. Jothi Saravanan
tjs.saravanan@gmail.com

¹ Department of Civil Engineering, Sharda University, Greater Noida 201310, India

² Department of Civil Engineering, Mahindra University, Hyderabad 500043, India

³ School of Infrastructure, Indian Institute of Technology Bhubaneswar, Odisha 752050, India

undesirable and destructive, time-consuming, requires many mixed trails, and has low working efficiency (Bischoff & Perry, 1991; Shi et al., 2009). Many researchers used the empirical regression method (Bhanja & Sengupta, 2002; Bharatkumar et al., 2001; Zain & Abd, 2009) and numerical simulation (Feng & Li, 2016; Feng et al., 2018, 2019) method to predict the compressive strength of concrete and capture the concrete behaviour but unfortunately, results show a non-linear relation between the compressive strength and concrete mixing parameters; thus it is difficult to predict the accurate compressive strength.

On the other hand, with the advancement and promising results of artificial intelligence (AI) in recent years, numerous researchers used ML algorithms/approaches such as bagged artificial neural networks (BANNs), gradient-boosted artificial neural networks (GBANNs) (Erdal et al., 2013), support vector machines (SVMs) (Latif, 2021), chi-squared automatic interaction detection (CAID), regression trees, linear regression and ARIMA (Bansal et al., 2021, 2022a, 2022b; Chou & Pham, 2013; Kaveh et al., 2021), ensemble models, genetic weighted pyramid operation tree (GWPOT) (Cheng et al., 2014; Kaveh et al., 2008), ensemble decision trees (Erdal, 2013), metaheuristic-optimized least squares support vector regression (Pham et al., 2016), fracture mechanics approach (Shafiei Dastgerdi et al., 2019), compressible packaging model (CPM) (Amario et al., 2017), artificial neural networks (ANNs) (Kaveh & Iranmanesh, 1998; Kaveh & Khalegi, 1998; Kaveh et al., 2023; Kostić & Vasović, 2015; Mohammed et al., 2021; Naderpour & Mirrashid, 2018; Naderpour et al., 2018; Słoiński, 2010; Young et al., 2019), hybrid model (Shishegaran et al., 2021), quadratic polynomial model (Imanzadeh et al., 2018), and mixture optimization model (Miller et al., 2016; Zahiri & Eskandari-Naddaf, 2019; Zhang et al., 2016) to predict the CCS and in other applications. The literature review of some of the models are shown below.

Shafiei Dastgerdi et al. (2019) investigated the impact of different concrete parameters such as w/c ratio, aggregate shape, paste, air void, and fly ash content on the crack resistance of concrete over the concrete railroads by utilizing a two-parameter model (TPM). Test was carried out on twelve three-point prisms at different concrete compressive strengths. Their study shows that the decreasing w/c ratio and increasing aggregate size and volume improved the fracture toughness ratio by 30%, whereas other concrete modules negligibly influenced fracture toughness. Amario et al. (2017) analysed the probability of adopting the compressible packing model (CPM) for the concrete mixture proportion generated with recycled concrete aggregates (RCAs). The aggregate replacement variation from 0 to 100% was taken into account and various structural RCA mixtures were designed for three strength classes. At last, the implemented process was verified experimentally by

carrying out the durability and mechanical tests on chosen mixtures with RCAs content nearer to 60% for three strength classes. Their study shows that CPM has a high correlation with RCAs and that overall durability performance is not influenced by RCA's presence. Young et al. (2019) presented the initial analysis of a large dataset consisting of calculated compressive strengths from original (job-site) mixtures and their respective original mixture proportions. The correlation between the mixture design variables and strength was investigated by applying a predictive model such as ANN. Their methods were well adopted over the laboratory-based dataset based on strength measurements, and the method's performance among the two data sets was differentiated. Their results show that ANN reduces the labor and time intensity, better robustness, quality control, and is cost-efficient and thus proving the superiority of the proposed architecture. However, this method needs many data, specifically for large architecture, and is harder to visualize. Kostić and Vasović (2015) implemented a prediction approach based on ANN for CCS. Three-layer feed-forward neural networks were examined under two, six and nine hidden nodes using four diverse learning methods in their work. The more precise prediction methods having the largest coefficient of determination (R^2) have been attained with six hidden nodes through Levenberg–Marquardt, with nine hidden nodes by means of Broyden–Fletcher Goldfarb–Shannon model and with scaled conjugate gradient and one-step secant models. The analysis has thus shown the improved efficiency of the proposed ANN model over conventional models. To achieve an expected compressive strength, Imanzadeh et al. (2018) introduced the usage of mixture structure as a tool for optimizing the concrete formulation on the raw earth. The experiment was made in terms of comparative analysis over the conventional models. The outcomes have demonstrated that the mixture design technique has considered being an effective tool for developing and optimizing the raw earth concrete formulation. Miller et al. (2016) developed an approach to predict the global warming potential (GWP) and compressive strength based on the water-to-binder(w/b) ratio for concrete mixtures. Their results show a linear correlation among GWP and cement content. However, developing more robust prediction tools is needed in the future and needs to examine multiple design criteria. Zhang et al. (2016) used RCA by replacing the coarse natural aggregate (NA) at different replacement levels. Their outcomes show that asphalt concrete mix design provides lower apparent relative density, higher absorption of water and lower crushing and wearing value with an increase of RCA. The main limitation is it needs further study to be conducted on the RCA from various sources. Zahiri and Eskandari-Naddaf (2019) designed twelve mix structures involving diverse percentages of nano-silica (NS), micro-silica (MS) and polymer fibers in three cement strength classes (CSC) based on the

mixture optimization model. The experimental outcome has shown that every CSC's sensitivity has diverse on MS or NS in concrete compressive strength. Subsequently, in the concrete mix design, the strength classes have a considerable impact on the quantity of NS and MS, whereas in polymer fibers, no considerable impact was made on the compressive strength while accounting for the CSCs. The mixture Optimization method achieved better strength class with the increase of CCS. Many researchers also monitor and predict early-age hydration and compressive strength using a smart sensor such as piezo sensor and machine learning techniques (Saravanan et al., 2015a, b; Bharathi Priya et al., 2018; Bansal & Talakokula, 2020).

Based on the above-mentioned literature, it is concluded that only a single ML prediction model has been used to predict compressive strength. Also, no comparative approach is available that can help the community to choose the best prediction method. This paper presents a comparative study to predict the compressive strength of concrete to overcome this issue by utilizing various ML models. The contribution of this work is summarized below:

- A comparative study and analysis of various ML models (Ordinary Least Square, Ridge Regression, Lasso Regression, ElasticNet, K Nearest Neighbours, CART, Random Forest, AdaBoost, Gradient Tree Boosting and Xtreme Gradient Boost) for the precise prediction of compressive strength of concrete.
- Performance of Hyperparameter optimization algorithm to improve the accuracy of the model.

Dataset

The performance of the machine learning models is generally dependent on the scale of samples in the dataset. To build and compare high-accuracy models, a large number of samples are necessary which can be obtained from previous literature (Erdal et al., 2013). The dataset consists of 1030 samples and nine features.

Features

The features of concrete considered are listed below in Table 1.

Statistical information

Since concrete compressive strength is being predicted it will be the target variable and the rest of the attributes are input variables. Table 2 gives the statistical description of the dataset such as count, minimum/maximum values, mean value, quartile (25%, 50% and 75% values) and standard

Table 1 Features of the dataset

S. no	Name	Data type	Measurement	Description
1	Cement	Quantitative	kg/m ³	Input variable
2	Blast furnace slag	Quantitative	kg/m ³	Input variable
3	Fly ash	Quantitative	kg/m ³	Input variable
4	Water	Quantitative	kg/m ³	Input variable
5	Superplasticizer	Quantitative	kg/m ³	Input variable
6	Coarse aggregate	Quantitative	kg/m ³	Input variable
7	Fine aggregate	Quantitative	kg/m ³	Input variable
8	Age	Quantitative	Day (1–365)	Input variable
9	Concrete compressive strength	Quantitative	MPa	Output variable

deviation. From this table, it is observed that the average compressive strength is found to be 35.82 MPa with a standard deviation of 16.70 MPa in the count of 1030 samples. The cement values ranging between 102 and 540 kg/m³, it is due to the replacement of cement with fly ash and blast furnace slag with the minimum and maximum value ranging between 0–359.4 and 0–200.1 kg/m³, respectively. Age is also one of the most important parameters for the development of compressive strength, here the age values range between 1 and 365 days with a standard deviation of 63.17 days. Feature distribution of the involved parameter is shown in Fig. 1, which can help us for direct observation of the parameters in which the dotted line represents the mean value in all the parameters.

Correlation analysis

The first step in any predictive analytics is data exploration which includes checking for missing values, checking for attribute correlation, and observing the distribution of all features. To build a robust and accurate model, the output feature should be correlated to the input variables. The correlation between the features can be calculated using Pearson's correlation factor, where a higher absolute value of the correlation factor represents higher dependence of the two variables. Correlation analysis also helps us to remove features that are not correlated with the target-dependent variable. It should be noted that we can remove one of the two highly correlated input features or independent variables as they are considered redundant.

Figure 2 shows the correlation heatmap of all the features. From this figure, it can be seen that the cement, superplasticizer and age are the most three important parameters that are best correlated with the compressive strength with absolute values of 0.5, 0.37 and 0.33, respectively, however, fly

Table 2 Statistical description of data set

Feature	Cement	Blast furnace Slag	Fly ash	Water	Super-plasticizer	Coarse aggregate	Fine aggregate	Age	Concrete compressive strength
Count	1030	1030	1030	1030	1030	1030	1030	1030	1030
Mean	281.17	73.90	54.19	181.57	6.20	972.92	773.58	45.66	35.82
Std	104.50	86.28	63.99	21.36	5.97	77.76	80.18	63.17	16.70
Min	102	0	0	121.75	0	801	594	1	2.33
25%	192.38	0	0	164.9	0	932	730.95	7	23.70
50%	272.9	22	0	185	6.35	968	779.51	28	34.44
75%	350	142.95	118.27	192	10.16	1029.4	824	56	46.14
Max	540	359.4	200.1	247	32.2	1145	992.6	365	82.60

ash is the least with absolute value of 0.11. In addition to this, the correlation between the water and superplasticizer shows an absolute value of 0.66, which is very higher since the superplasticizer allows a reduction in water content and increases the strength and workability of concrete.

Methodology

The machine learning model aims to convert the input independent variable into a target dependent variable by automatically learning a mathematical model. Models are trained on training set samples and are tested on unknown samples called testing sets. A Machine Learning model is said to be “underfitting” when it cannot train or perform well on training set samples. The model is said to be “overfitting” if the performance on the training set is considered good, but it underperforms on unknown samples from the testing set. Machine learning engineers commonly apply regularization strategies to handle overfitting scenarios.

Models for study

In this study, we used twelve Machine Learning (ML) algorithms are used to predict the concrete compressive strength. They are,

1. **Ordinary least square** is a simple linear regression approach that aims to map the input features to the output target by learning a linear model with $n + 1$ parameters (where n is the number of features) using a numerical solution.
2. **Ridge regression** is the regularized linear regression approach that aims to train a generalized regression model that can perform well for both training and testing set samples. Ridge regression applies L2 norm penalty (which is a regularization strategy avoid overfitting model) on the parameters which keeps the model from becoming too much non-linear.
3. **Lasso regression** is another variant of regularized linear regression where the L1 norm penalty is applied for the regularization.
4. **ElasticNet** combines the L1 and L2 norm penalties into a linear regression model for more efficient regularization. This model utilizes the best of both norm penalties into an integrated solution.
5. **K-nearest neighbour (KNN)** is a non-parametric model for both classifications as well as regression. KNN model for regression finds the nearest or similar samples and average the neighbours’ target value as a final prediction. This model works on the assumption that similar samples are most likely to have the same output.
6. **Classification and regression trees (CART)** is a rule-based machine learning algorithm resembles tree data structure where every node represents a condition and every leaf contains the prediction value.
7. **Random forest** is a classic example for ensemble learning. It takes advantage of multiple decision trees and their predictions by fusing them into a single prediction which compensates the error of individual regression trees.
8. **AdaBoost** is another ensemble learning model (similar to random forest) which builds multiple small decision trees (weak learners) with a single split stumps. The stumps are learned gradually by focusing more on samples mistakenly predicted by previous stumps learned.
9. **Gradient tree boosting** also utilizes the ensembling of decision trees which are weak learners. Gradient boosting technique allows the trees to trees to reconstruct the residual of the initial trees.
10. **Xtreme gradient boost** is a faster implementation of the gradient boosting technique, which aims to provide scalable, portable and distributed gradient boosting for applications.

Fig. 1 Feature distribution bar charts

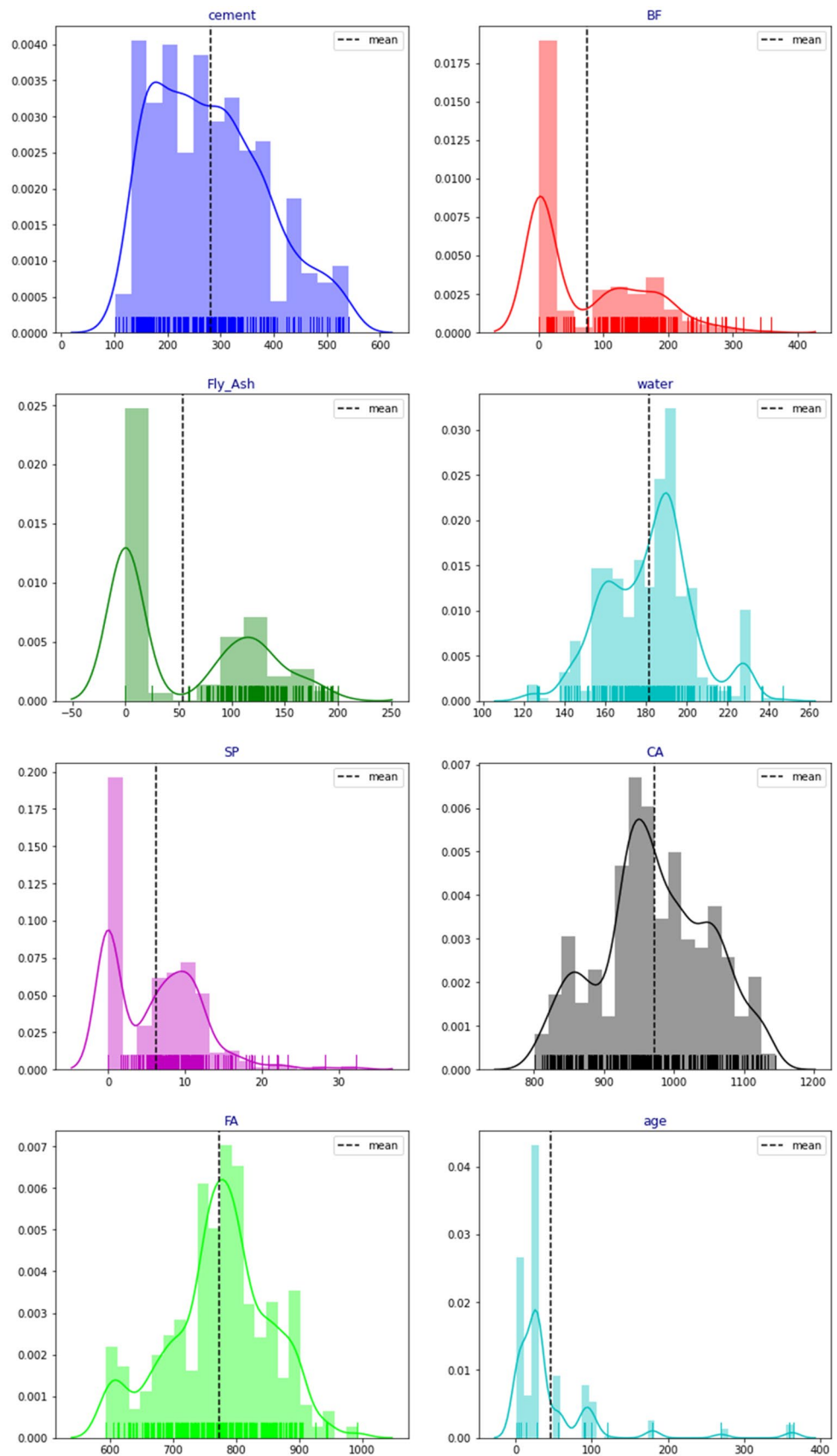
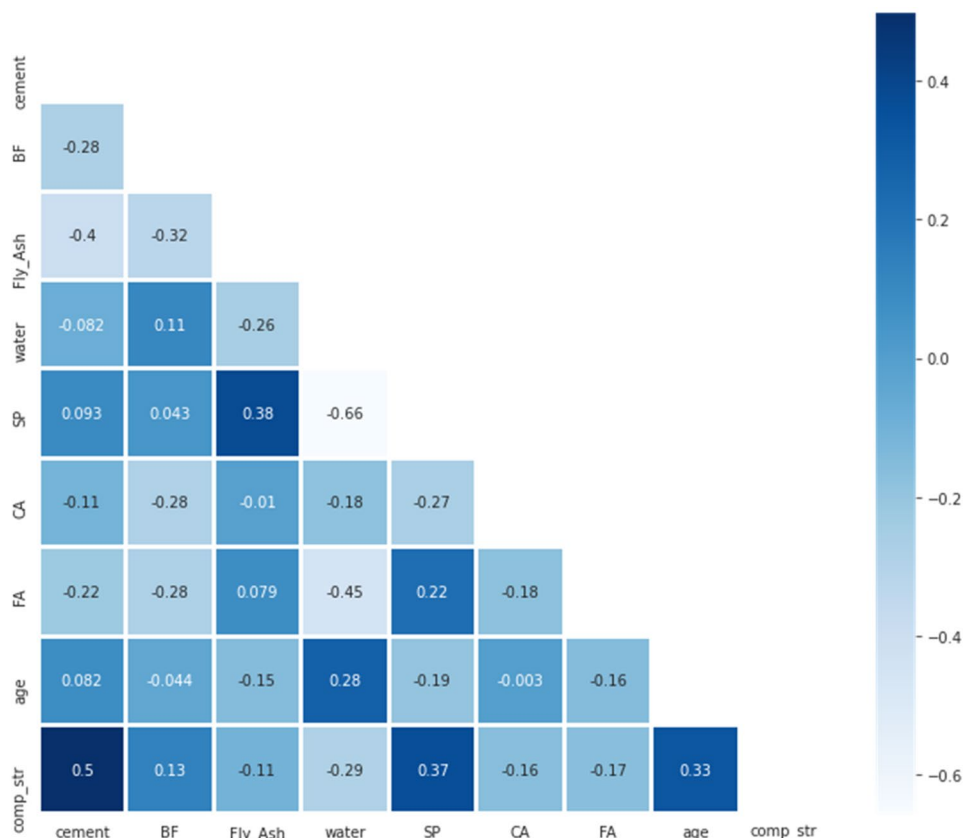
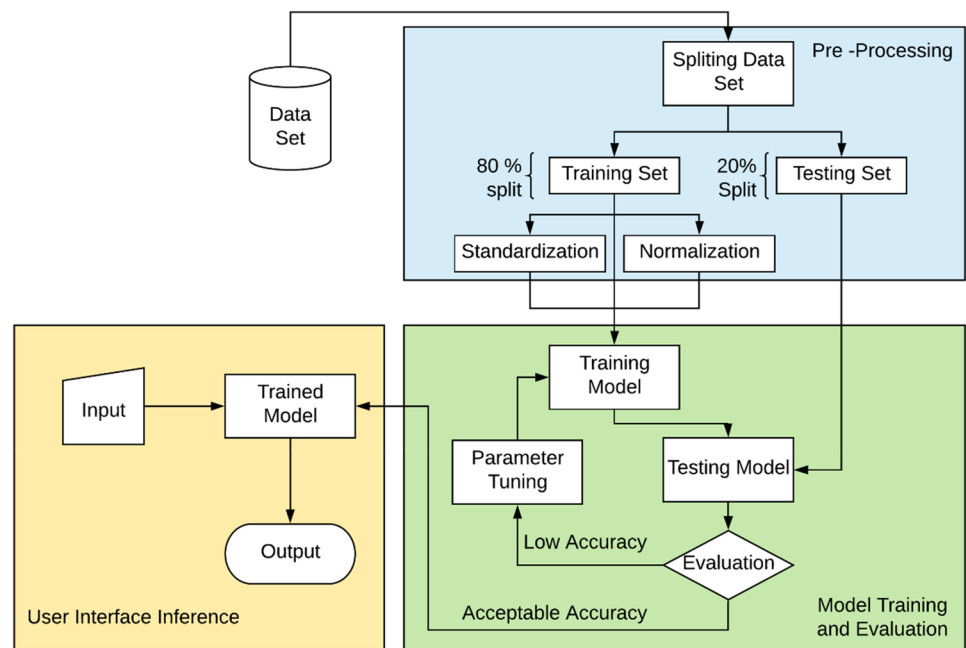


Fig. 2 Pearson correlation between the features represented in the form of a heatmap



11. **MLP regressor** stands for multilayer perceptron regressor, is a densely connected neural network for regression. Neural networks are known for their superior performance in most computational problems and are used for both classification and regression. Neural Networks mimics human brain by having multiple layers of neurons for learning hierarchical features from the input array. In a neural network, each neuron receives array input and outputs a scalar value called activation.
 12. **Support vector regression (SVR)** is a linear model for regression. Unlike the support vector classification (SVC) which aims to maximize the margin of classes through support vectors, SVR gives flexibility for a regressor by ignoring error made for the support vectors inside the boundary.
1. **Dataset Pre-processing:** In this module, a dataset is loaded, and randomly split into training and testing sets. The features of the dataset are further scaled into a smaller range.
 - **Dataset splitting:** Firstly, an unprocessed dataset is checked for missing values, if found then the data entry with the missing value should be either removed or imputed. Here the dataset used in the study has no missing values. The unprocessed dataset has been split into training (80%) and testing (20%) set using random shuffling in which the training set is used to train the models and the testing set is used to evaluate the model performance. The training set contains 834 data entries and the testing set contains 226 data entries. The reason for splitting the dataset is because it tells how the trained model will perform on samples that were not seen by the model yet. After this, a training and testing set is used to train all the algorithms and to evaluate the performance which ensures fair comparison.
 - **Feature scaling:** In addition, the training set has been standardized and normalized by scaling the dataset and a new version of the training set has been obtained which called as normalized dataset and standardized dataset. The models are trained on

Fig. 3 Flowchart of training methodology

this new version of the dataset to determine which scaling if required gives us a higher accuracy in evaluation. The accuracy of an algorithm from the three-training set are compared and the algorithm with appropriate scaling is listed. The top five performing algorithms are then selected for hyperparameter optimization to further improve the accuracy of the model because hyperparameter helps to control the learning process.

2. **Model training and evaluation:** In this module, the pre-processed features are used for training the machine learning models. We also use feature selection to find the best feature subset.

- **Feature selection:** The algorithm with the highest accuracy after optimization is used for the feature selection. Feature selection with just Pearson's correlation factor is not conclusive so three strategies for feature selection are applied. First feature selection uses the correlation between the input features and the output feature. Removing input feature in increasing order to give an understanding of the importance of features in prediction. Second strategy is to remove the input features according to the correlation in between hence removing highly correlated input features (i.e. more than ± 0.5). High interdependence in input features means any one of the inter-dependent features can be used for prediction without affecting the accuracy much and removing a feature decreases the complexity of a model. The third strategy is to remove the feature one by one to

train the models and to use each feature to train the model. Third strategy's model accuracy also give us the importance of each input feature in model prediction. The feature importance of the model can be then compared to the theoretical understanding of concrete ingredients and their effect on compressive strength.

- **Hyperparameter optimization:** The feature subset is further used for tuning the hyperparameters of the machine learning models. This step helps us to find the optimal hyperparameter for each machine learning model.
3. **Model inference using GUI:** In this module, the trained models can be used for inference with new incoming data. Usually, results from multiple machine learning models can be fused into a single prediction using averaging or weighted averaging fusion technique. The real-time data for inference also goes through the same pre-processing steps as the training and testing set samples.

Results and discussion

Evaluation measures

In this study, the performance of the prediction model is evaluated by calculating the following measures such as R^2 , explained variance score (EVS), mean absolute error (MAE), mean squared error (MSE) and maximum residual error (MRE). The expressions of the evaluating measures

Table 3 Unprocessed dataset—concrete compressive strength results

Comparative methods	R^2	EVS	MRE	MAE	MSE
Ordinary least square	0.60	0.60	35.64	8.63	124.13
Ridge regression	0.60	0.60	35.64	8.63	124.13
Lasso regression	0.60	0.60	35.58	8.66	124.47
ElasticNet	0.60	0.60	35.59	8.64	124.30
K nearest neighbours	0.76	0.76	30.89	6.74	76.25
CART	0.88	0.88	32.99	3.83	36.64
Random forest	0.91	0.91	31.78	3.45	27.89
AdaBoost	0.80	0.80	32.36	6.59	63.35
Gradient tree boosting	0.90	0.90	27.84	4.00	31.12
Xtreme gradient boost	0.90	0.90	27.38	3.99	31.23
MLP regressor	0.83	0.83	33.94	5.67	55.35
SVR	0.54	0.54	45.32	8.65	145.62

Table 4 Standardized dataset—concrete compressive strength results

Comparative methods	R^2	EVS	MRE	MAE	MSE
Ordinary least square	0.60	0.60	35.64	8.63	124.13
Ridge regression	0.60	0.60	35.68	8.65	124.30
Lasso regression	0.55	0.55	38.18	9.60	141.38
ElasticNet	0.48	0.48	41.35	10.53	163.60
K nearest neighbours	0.75	0.76	26.41	6.94	77.93
CART	0.88	0.88	32.99	3.84	36.640
Random forest	0.92	0.91	31.78	3.46	27.90
AdaBoost	0.79	0.80	31.83	6.59	63.72
Gradient tree boosting	0.90	0.90	27.84	4.00	31.12
Xtreme gradient boost	0.90	0.90	27.37	3.99	31.23
MLP regressor	0.89	0.89	24.88	4.50	35.54
SVR	0.55	0.55	43.18	8.82	143.01

are shown in Eqs. (1–4). MRE is the maximum error made by the model among all samples.

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \quad (1)$$

$$\text{explained variance } (y, \hat{y}) = 1 - \frac{\text{Var}(y - \hat{y})}{\text{Var}(y)} \quad (2)$$

$$\text{MSE} = \sum_i (y_i - \hat{y}_i)^2 \quad (3)$$

$$\text{MAE} = \sum_i (y_i - \hat{y}_i) \quad (4)$$

Evaluation measures such as MAE, MSE and MRE represent errors made by the model during prediction whereas other measures R^2 score and EVS denote similarity between prediction and target values.

Influence of feature scaling: normalization versus standardization

Comparison of the algorithm starts with exploring and processing the data to make it suitable for training. The model training result with an unprocessed training set is given in Table 3. The result shows that the linear models ordinary least square, regularized linear models ridge, lasso and ElasticNet perform with an accuracy of 0.61 R^2 . Since the features might have non-linear relationships with the target variables, these linear models do not fit well for this scenario. Support vector regression performs poorly with only 0.54 R^2 as SVR is also a kind of linear model. K-nearest neighbour (KNN) algorithm which works best in the dataset, performs better than the linear models with an R^2 score of

0.76. Artificial neural network model multilayer perceptron performs well with 0.83 R^2 by extracting non-linear features at each layer. The best-performing models for unprocessed concrete compressive strength datasets are decision trees. Adaboost algorithm which is an ensemble boosting decision tree model performs with 0.80 R^2 accuracy while classification and regression trees (CART) perform with an accuracy of 0.88 R^2 . Ensemble models with boosting gradient tree boosting and regularized ensemble model Xtreme tree boosting perform with almost the same accuracy with 0.90 R^2 each. Random forest an ensemble model decision trees model with bagging gives an accuracy of 0.91 R^2 . It can be seen that the tree-based and tree ensemble models outperform linear models for regression. It is also widely known that tree-based machine learning models do not need much pre-processing to be done on the training set.

Scaling of machine learning models is generally required as the estimators may perform in a subpar manner if it is not done. The standardization features scaling of the dataset scales the features into smaller ranges with zero mean and unit standard deviation. Table 4 shows the result of model training with a standardized dataset. The accuracy of the ordinary least square method and ridge regression does not change much with a standardized dataset at 0.60 R^2 . KNN also has no improvement with standardization and stays at 0.75 R^2 . All the decision tree models stay at the same accuracy without any improvement in accuracy. The accuracy of Lasso regression and ElasticNet decrease with accuracy of 0.55 R^2 and 0.48 R^2 . SVR has a minor increase in accuracy with 0.55 R^2 . A standardized dataset is most helpful in MLP regressor as it increases accuracy to 0.89 R^2 .

Table 5 shows the performance of models on a normalized dataset. The normalization feature scaling technique scales the feature into a 0–1 range using min–max normalization. The performance of almost all models decreases in the normalized training set. Ordinary least squares and

Table 5 Normalized dataset—concrete compressive strength results

Comparative methods	R^2	EVS	MRE	MAE	MSE
Ordinary least square	0.68	0.69	33.42	7.75	100.71
Ridge regression	0.68	0.45	44.23	11.05	176.09
Lasso regression	0.68	0.0001	45.89	14.76	317.07
ElasticNet	0.68	0.0014	45.89	14.75	316.60
K-nearest neighbours	0.77	0.77	35.58	6.72	75.58
CART	0.82	0.82	51.37	4.67	59.39
Random forest	0.89	0.89	47.82	3.74	34.71
AdaBoost	0.77	0.78	34.50	7.22	72.98
Gradient tree boosting	0.89	0.89	40.47	4.18	36.05
Xtreme gradient boost	0.89	0.89	40.23	4.20	36.84
MLP regressor	0.82	0.82	33.06	5.58	57.16
SVR	0.12	0.12	46.40	13.82	279.72

Table 6 Model accuracy with appropriate scaling

S. no	Model name	R^2	Scaling approach
1	Ordinary least squares	0.68	Normalization
2	Ridge regression	0.68	Normalization ($\alpha = 10^{-6}$)
3	Lasso regression	0.68	Normalization ($\alpha = 10^{-5}$)
4	ElasticNet	0.68	Normalization ($\alpha = 10^{-7}$)
5	KNN	0.77	Normalization
6	CART	0.88	None
7	Random forest	0.90	None
8	AdaBoost	0.80	None
9	Gradient tree boosting	0.90	None
10	Xtreme gradient boost	0.90	None
11	MLP regressor	0.88	Standardization
12	SVR	0.55	Standardization

KNN perform better in normalized datasets. In Table 5, the ordinary least square performs better with a 0.08 increase in accuracy with 0.68 R^2 . While KNN's accuracy increases with 0.02 at 0.77 R^2 . The reduction in regularized linear models is due to high alpha (learning rate). When the learning rate is decreased for Lasso, Ridge and ElasticNet to an alpha of 10^{-5} , 10^{-6} and 10^{-7} , respectively, the accuracy is at 0.68 R^2 . Though further decreasing alpha does not increase the accuracy.

Top performing models

The selection of appropriate scaling for models is done based on the accuracy obtained from the three versions of a dataset (unprocessed vs standardized vs normalized). Table 6 lists the models with the highest accuracy for that specific model from the unprocessed, normalized and standardized datasets. The linear models and KNN perform better with a normalized dataset. MLP regressor and SVR improve their

Table 7 Top 5 performing model after hyper-parameter optimization

S. no	Model	R^2 score
1	Gradient boosted trees	0.94
2	Xtreme gradient boosted trees	0.93
3	Random forest	0.91
4	MLP regressor	0.90
5	CART	0.88

accuracy with standardized datasets. The decision tree models are most accurate for concrete compressive strength datasets. Ensemble models with boosting (Gradient Tree Boosting and Xtreme Gradient Boosting) and bagging (Random Forest) perform better with an accuracy of 0.9 R^2 . Decision trees do not require scaling for model performance is apparent as the scaling of features from -1 to 1 (standardization) to $0-1$ (normalization) decreases the difference between the data entries. The reduced difference between features makes it harder for the regression decision tree to form nodes and set conditions for a split.

Hyperparameter optimization

Table 7 shows the performance of the top five models after hyperparameter optimization. For hyperparameter optimization, we have used both a grid-based hyperparameter search approach as well as random hyperparameter search approach. From this, it is observed that Gradient boosted trees algorithm performs very well for concrete compressive strength prediction with an R^2 value of 0.94 followed by the Xtreme gradient-boosted trees, random forest, MLP regressor and CART algorithm with an R^2 value of 0.93, 0.91, 0.90, and 0.88, respectively. In addition, it is noticed that after hyperparameter optimization, the performance of the top five models is improved. It is due to the fact that hyperparameter optimization identifies a tuple of hyperparameters that yields an optimal model which minimizes a predefined loss function on given independent data (Yuan-Fu, 2019).

Feature selection

The automated feature selection performed on the Gradient Boosted Tree algorithm show that the most important feature for concrete compressive strength prediction is the age of the concrete and then the cement content. The least important feature is coarse aggregate. Hence, it can be concluded that by using Gradient boosted trees algorithm, contactors at the site can predict the compressive strength of concrete by just providing raw material (such as cement, blast furnace slag, fly ash, water, superplasticizer, coarse aggregate, and fine aggregate) quantity in the input only and get the response (compressive strength).

Conclusion

In this study, twelve prominent Machine Learning algorithms on concrete compressive strength datasets are used and compared. Further, the algorithms experimented with scaled datasets after standardization and normalization to determine which scaling or unprocessed dataset works better for regression. Furthermore, the top five models were hyper-parameter optimized. The linear models performed better with a normalized dataset with 0.68 R^2 . Though regularized linear models performed with the same performance as ordinary least squares but did not outperform even after decreasing alpha to 10^{-7} . The top five performing models after hyper-parameter optimization were Gradient Boosted Trees with 0.94 R^2 , Xtreme Gradient Boosted Trees with 0.93 R^2 , Random Forest with 0.91 R^2 , MLP Regressor 0.90 R^2 and CART with 0.88 R^2 . Hence, based on the value of R^2 , it is concluded that Gradient Boosted Trees algorithm can be used to predict the compressive strength of concrete which helps the research community to reduce the cost and time in concrete mix designing and avoid the waste of materials caused by numerous mixture trials.

Acknowledgements The authors would like to express their gratitude to Dr. Sridhar Swaminathan, L&T Edutech, for his invaluable expert advice in software analysis and research work.

Author contributions TB: Writing-original Draft; Writing-review & editing; Conceptualization, Methodology; Data Curation; Software Analysis. VT: Writing-review & editing; Supervision; Visualization. TJS: Supervision; Visualization.

Funding No funding was provided for this research.

Availability of data and materials <https://archive.ics.uci.edu/ml/datasets/concrete+compressive+strength>.

Declarations

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

Competing interests The authors declare that they have no competing interests.

References

- Al-Shamiri, A. K., Kim, J. H., Yuan, T. F., & Yoon, Y. S. (2019). Modeling the compressive strength of high-strength concrete: An extreme learning approach. *Construction and Building Materials*, 208, 204–219. <https://doi.org/10.1016/j.conbuildmat.2019.02.165>
- Amario, M., Rangel, C. S., Pepe, M., & Toledo Filho, R. D. (2017). Optimization of normal and high strength recycled aggregate concrete mixtures by using packing model. *Cement and Concrete Composites*, 84, 83–92. <https://doi.org/10.1016/j.cemconcomp.2017.08.016>
- Bansal, T., Talakokula, V., & Mathiyazhagan, K. (2021). Equivalent structural parameters based non-destructive prediction of sustainable concrete strength using machine learning models via piezo sensor. *Measurement*, 187, 110202. <https://doi.org/10.1016/j.measurement.2021.110202>
- Bansal, T., Talakokula, V., & Sathujoda, P. (2022a). Machine learning-based monitoring and predicting the compressive strength of different blended cementitious systems using embedded piezo-sensor data. *Measurement*, 205, 112204. <https://doi.org/10.1016/j.measurement.2022.112204>
- Bansal, T., Talakokula, V., & Sathujoda, P. (2022b). A machine learning approach for predicting the electro-mechanical impedance data of blended RC structures subjected to chloride laden environment. *Smart Materials and Structures*. <https://doi.org/10.1088/1361-665X/ac3d6f>
- Bansal, T., & Talakokula, V. (2020). Monitoring strength development of cement substituted by limestone calcined clay using different piezo configurations. In *RILEM Bookseries* (Vol. 25, pp. 555–562). Springer. https://doi.org/10.1007/978-981-15-2806-4_62
- Berodier, E., Gibson, L. R., Burns, E., Roberts, L., & Cheung, J. (2019). Robust production of sustainable concrete through the use of admixtures and in-transit concrete management systems. *Cement and Concrete Composites*, 101, 52–66. <https://doi.org/10.1016/j.cemconcomp.2018.01.008>
- Bhanja, S., & Sengupta, B. (2002). Investigations on the compressive strength of silica fume concrete using statistical methods. *Cement and Concrete Research*, 32, 1391–1394. [https://doi.org/10.1016/S0008-8846\(02\)00787-1](https://doi.org/10.1016/S0008-8846(02)00787-1)
- Bharathi Priya, C., Jothi Saravanan, T., Balamonica, K., Gopalakrishnan, N., & Rao, A. R. M. (2018). EMI based monitoring of early-age characteristics of concrete and comparison of serial/parallel multi-sensing technique. *Construction and Building Materials*, 191, 1268–1284. <https://doi.org/10.1016/j.conbuildmat.2018.10.079>
- Bharatkumar, B. H., Narayanan, R., Raghuprasad, B. K., & Ramachandramurthy, D. S. (2001). Mix proportioning of high performance concrete. *Cement and Concrete Composites*, 23, 71–80. [https://doi.org/10.1016/S0958-9465\(00\)00071-8](https://doi.org/10.1016/S0958-9465(00)00071-8)
- Bischoff, P. H., & Perry, S. H. (1991). Compressive behaviour of concrete at high strain rates. *Materials and Structures*, 24, 425–450.
- Cheng, M. Y., Firdausi, P. M., & Prayogo, D. (2014). High-performance concrete compressive strength prediction using genetic weighted pyramid operation tree (GW POT). *Engineering Applications of Artificial Intelligence*, 29, 104–113. <https://doi.org/10.1016/j.engappai.2013.11.014>
- Chou, J. S., & Pham, A. D. (2013). Enhanced artificial intelligence for ensemble approach to predicting high performance concrete compressive strength. *Construction and Building Materials*, 49, 554–563. <https://doi.org/10.1016/j.conbuildmat.2013.08.078>
- Erdal, H. I. (2013). Two-level and hybrid ensembles of decision trees for high performance concrete compressive strength prediction. *Engineering Applications of Artificial Intelligence*, 26, 1689–1697. <https://doi.org/10.1016/j.engappai.2013.03.014>
- Erdal, H. I., Karakurt, O., & Namli, E. (2013). High performance concrete compressive strength forecasting using ensemble models based on discrete wavelet transform. *Engineering Applications of Artificial Intelligence*, 26, 1246–1254. <https://doi.org/10.1016/j.engappai.2012.10.014>
- Feng, D., & Li, J. (2016). Stochastic nonlinear behavior of reinforced concrete frames. II: Numerical simulation. *Journal of Structural Engineering*, 142, 04015163. [https://doi.org/10.1061/\(asce\)st.1943-541x.0001443](https://doi.org/10.1061/(asce)st.1943-541x.0001443)
- Feng, D.-C., Ren, X.-D., & Li, J. (2018). Softened damage-plasticity model for analysis of cracked reinforced concrete structures.

- Journal of Structural Engineering*, 144, 04018044. [https://doi.org/10.1061/\(asce\)st.1943-541x.0002015](https://doi.org/10.1061/(asce)st.1943-541x.0002015)
- Feng, D. C., Wang, Z., & Wu, G. (2019). Progressive collapse performance analysis of precast reinforced concrete structures. *Structural Design of Tall and Special Buildings*, 28, 1–21. <https://doi.org/10.1002/tal.1588>
- Imanzadeh, S., Hibouche, A., Jarno, A., & Taibi, S. (2018). Formulating and optimizing the compressive strength of a raw earth concrete by mixture design. *Construction and Building Materials*, 163, 149–159. <https://doi.org/10.1016/j.conbuildmat.2017.12.088>
- Kaplan, G., & Salem Elmekahal, M. A. (2021). Microstructure and durability properties of lightweight and high-performance sustainable cement-based composites with rice husk ash. *Environmental Science and Pollution Research*. <https://doi.org/10.1007/s11356-021-14489-y>
- Kaveh, A., Eskandari, A., & Movasat, M. (2023). Buckling resistance prediction of high-strength steel columns using metaheuristic-trained artificial neural networks. *Structures*, 56, 104853. <https://doi.org/10.1016/j.istruc.2023.07.043>
- Kaveh, A., Eslamlou, A. D., Javadi, S. M., & Malek, N. G. (2021). Machine learning regression approaches for predicting the ultimate buckling load of variable-stiffness composite cylinders. *Acta Mechanica*, 232, 921–931. <https://doi.org/10.1007/S00707-020-02878-2/FIGURES/9>
- Kaveh, A., Gholipour, Y., & Rahami, H. (2008). Optimal design of transmission towers using genetic algorithm and neural networks. *International Journal of Space Structures*, 23, 1–19. <https://doi.org/10.1260/026635108785342073>
- Kaveh, A., & Iranmanesh, A. (1998). Comparative study of backpropagation and improved counterpropagation neural nets in structural analysis and optimization. *International Journal of Space Structures*, 13, 177–185. <https://doi.org/10.1177/026635119801300401>
- Kaveh, A., & Khalegi, A. (1998). Prediction of strength for concrete specimens using artificial neural networks. In *1st International conference on engineering computational technology/4th international conference on computational structures technology* (pp 165–171).
- Kostić, S., & Vasović, D. (2015). Prediction model for compressive strength of basic concrete mixture using artificial neural networks. *Neural Computing and Applications*, 26, 1005–1024. <https://doi.org/10.1007/s00521-014-1763-1>
- Larsen, O., Naruts, V., & Aleksandrova, O. (2019). Self-compacting concrete with recycled aggregates. *Mater Today Proc*, 19, 2023–2026. <https://doi.org/10.1016/j.matpr.2019.07.065>
- Latif, S. D. (2021). Concrete compressive strength prediction modeling utilizing deep learning long short-term memory algorithm for a sustainable environment. *Environmental Science and Pollution Research*. <https://doi.org/10.1007/s11356-021-12877-y>
- Liu, F., & Li, Q. M. (2019). Strain-rate effect on the compressive strength of brittle materials and its implementation into material strength model. *International Journal of Impact Engineering*, 130, 113–123. <https://doi.org/10.1016/j.ijimpeng.2019.04.006>
- Miller, S. A., Monteiro, P. J. M., Ostertag, C. P., & Horvath, A. (2016). Concrete mixture proportioning for desired strength and reduced global warming potential. *Construction and Building Materials*, 128, 410–421. <https://doi.org/10.1016/j.conbuildmat.2016.10.081>
- Mohammed, A., Rafiq, S., Sihag, P., Kurda, R., & Mahmood, W. (2021). Soft computing techniques: Systematic multiscale models to predict the compressive strength of HVFA concrete based on mix proportions and curing times. *Journal of Building Engineering*, 33, 101851. <https://doi.org/10.1016/j.jobe.2020.101851>
- Naderpour, H., & Mirrashid, M. (2018). An innovative approach for compressive strength estimation of mortars having calcium inosilicate minerals. *Journal of Building Engineering*, 19, 205–215. <https://doi.org/10.1016/j.jobe.2018.05.012>
- Naderpour, H., Rafiean, A. H., & Fakharian, P. (2018). Compressive strength prediction of environmentally friendly concrete using artificial neural networks. *Journal of Building Engineering*, 16, 213–219. <https://doi.org/10.1016/j.jobe.2018.01.007>
- Pham, A.-D., Hoang, N.-D., & Nguyen, Q.-T. (2016). Predicting compressive strength of high-performance concrete using metaheuristic-optimized least squares support vector regression. *Journal of Computing in Civil Engineering*, 30, 06015002. [https://doi.org/10.1061/\(asce\)cp.1943-5487.0000506](https://doi.org/10.1061/(asce)cp.1943-5487.0000506)
- Saravanan, T. J., Balamonica, K., Priya, C. B., Reddy, A. L., & Gopalakrishnan, N. (2015a). Comparative performance of various smart aggregates during strength gain and damage states of concrete. *Smart Materials and Structures*. <https://doi.org/10.1088/0964-1726/24/8/085016>
- Saravanan, T. J., Kalyanasundaram, B., Priya, B. C., Gopalakrishnan, N., & S.G.N Murthy. (2015b). Non-destructive piezo electric based monitoring of strength gain in concrete using smart aggregate. <https://www.researchgate.net/publication/315761987>
- Shafiei Dastgerdi, A., Peterman, R. J., Riding, K., & Beck, B. T. (2019). Effect of concrete mixture components, proportioning, and compressive strength on fracture parameters. *Construction and Building Materials*, 206, 179–192. <https://doi.org/10.1016/j.conbuildmat.2019.02.025>
- Shamsutdinova, G., Hendriks, M. A. N., & Jacobsen, S. (2019). Topography studies of concrete abraded with ice. *Wear*, 430–431, 1–11. <https://doi.org/10.1016/j.wear.2019.04.017>
- Shi, H., Xu, B., & Zhou, X. (2009). Influence of mineral admixtures on compressive strength, gas permeability and carbonation of high performance concrete. *Construction and Building Materials*, 23, 1980–1985. <https://doi.org/10.1016/j.conbuildmat.2008.08.021>
- Shishegaran, A., Varae, H., Rabczuk, T., & Shishegaran, G. (2021). High correlated variables creator machine: prediction of the compressive strength of concrete. *Computers & Structures*. <https://doi.org/10.1016/j.compstruc.2021.106479>
- Sivamani, J., & Renganathan, N. T. (2021). Effect of fine recycled aggregate on the strength and durability properties of concrete modified through two-stage mixing approach. *Environmental Science and Pollution Research*. <https://doi.org/10.1007/s11356-021-14420-5>
- Słoński, M. (2010). A comparison of model selection methods for compressive strength prediction of high-performance concrete using neural networks. *Computers & Structures*, 88, 1248–1253. <https://doi.org/10.1016/j.compstruc.2010.07.003>
- Yoon, J. Y., & Kim, J. H. (2019). Mechanical properties of preplaced lightweight aggregates concrete. *Construction and Building Materials*, 216, 440–449. <https://doi.org/10.1016/j.conbuildmat.2019.05.010>
- Young, B. A., Hall, A., Pilon, L., Gupta, P., & Sant, G. (2019). Can the compressive strength of concrete be estimated from knowledge of the mixture proportions? New insights from statistical analysis and machine learning methods. *Cement and Concrete Research*, 115, 379–388. <https://doi.org/10.1016/j.cemconres.2018.09.006>
- Yu, F., Sun, D., Wang, J., & Hu, M. (2019). Influence of aggregate size on compressive strength of pervious concrete. *Construction and Building Materials*, 209, 463–475. <https://doi.org/10.1016/j.conbuildmat.2019.03.140>
- Yuan, Y., Niu, K., & Wang, Z. (2019). Compressive strength prediction of fibre reinforced polymer composites under lateral disturbance. *Polymer Testing*, 78, 105952. <https://doi.org/10.1016/j.polymertesting.2019.105952>
- Yuan-Fu, Y. (2019). A deep learning model for identification of defect patterns in semiconductor wafer map. In *ASMC (advanced semiconductor manufacturing conference) proceedings 2019-May* (pp. 250–255). <https://doi.org/10.1109/ASMC.2019.8791815>

- Zahiri, F., & Eskandari-Naddaf, H. (2019). Optimizing the compressive strength of concrete containing micro-silica, nano-silica, and polypropylene fibers using extreme vertices mixture design. *Frontiers of Structural and Civil Engineering*, *13*, 821–830. <https://doi.org/10.1007/s11709-019-0518-6>
- Zain, M. F. M., & Abd, S. M. (2009). Multiple regression model for compressive strength prediction of high performance concrete. *Journal of Applied Sciences*, *1*, 155–160.
- Zhang, Z., Wang, K., Liu, H., & Deng, Z. (2016). Key performance properties of asphalt mixtures with recycled concrete aggregate from low strength concrete. *Construction and Building Materials*, *126*, 711–719. <https://doi.org/10.1016/j.conbuildmat.2016.07.009>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.