Journal of Engineering and Applied Science

RESEARCH Open Access

Soft computing techniques to predict the compressive strength of groundnut shell ash-blended concrete

Navaratnarajah Sathiparan[1*] and Pratheeba Jeyananthan[2]

*Correspondence: sakthi@eng.jfn.ac.lk

[1] Department of Civil Engineering, Faculty of Engineering, University of Jaffna, Ariviyal Nagar, Kilinochchi, Jaffna, Sri Lanka
[2] Department of Computer Engineering, Faculty of Engineering, University of Jaffna, Jaffna, Sri Lanka

Abstract

Using groundnut shell ash (GSA) as a component in concrete mixtures is a viable approach to achieving sustainability in building practices. This particular kind of concrete has the potential to effectively mitigate the issues associated with high levels of CO_2 emissions and embodied energy, which are primarily attributed to the excessive utilization of cement in conventional construction materials. When GSA is utilized as a partial replacement for cement, the strength characteristics of concrete are influenced not only by the quantity of GSA replacement but also by several other factors, including cement content, water-to-cement ratio, coarse aggregate content, fine aggregate content, and curing length. This work demonstrates a predictive model for the compressive strength (*CS*) of GSA mixed concrete using ML methods. The models were constructed with 297 datasets obtained from published literature. These datasets included various input variables such as cement content, GSA content, fine aggregate content, coarse aggregate content, water need, and curing duration. The output variable included in the models was the *CS* of concrete. In this study, a set of seven machine learning algorithms was utilized as statistical assessment tools to identify the most precise and reliable model for predicting the *CS* of GSA mixed concrete. These techniques included linear regression, full quadratic model, artificial neural network, boosted decision tree regression, random forest regression, K nearest neighbors, and support vector regression. The present study evaluated several machine learning models, and it was shown that the random forest regression model had superior performance in forecasting the *CS* of GSA mixed concrete. The train data's R^2 is 0.91, with RMSE of 2.48 MPa. Similarly, for the test data, the R^2 value is 0.89, with an RMSE of 2.42 MPa. The sensitivity analysis results of the random forest regression model indicate that the cement content primarily drives the material's *CS*. Subsequently, the curing period and GSA content significantly impact the *CS*. This work systematically evaluates the *CS* of GSA mixed concrete, contributing to the existing body of knowledge and practical implementation in this domain.

Keywords: Concrete, Groundnut shell ash, Compressive strength, Machine learning, SHAP analysis

Introduction

Cement manufacturing is a significant contributor to the emission of CO_2 into the atmosphere. It is accountable for around 8% of global anthropogenic CO_2 emissions. Cement production totals over 4 billion tonnes [48], and each tonne of cement releases approximately 900 kg of CO_2 [7]. The use of fossil fuels to generate heat to initiate the cement manufacturing procedure, along with the thermal breakdown of calcium carbonate during the clinker manufacturing process, leads to significant carbon dioxide emissions. A total of 30–40% of the energy used in this process comes from fuel combustion, while 60–70% comes from decarbonization [10, 16, 17, 60]. Despite its high CO_2 emissions, cement is necessary for most building materials, including concrete.

The construction industry is increasingly faced with the need to develop alternative cementitious materials that may serve as viable alternatives for cement in building applications [57]. This is due to the urgent need to reduce CO_2 emissions and embodied energy since these factors play a crucial role in mitigating global warming in the long term [42]. Building materials often use a diverse range of supplementary cementitious materials [8, 9]. These materials encompass metakaolin [55], silica fume [20], volcanic pozzolanas [22], granulated blast furnace slag [38], and limestone [61]. Utilizing these industrial waste by-products as a financially viable substitute for cement does not impair the mechanical properties and long-lasting nature of the construction materials. Nevertheless, it is expected that the accessibility of these industrial by-products will diminish. Furthermore, it should be noted that the accessibility of these resources is limited, especially in less developed nations [9]. The construction sector is very interested in using agro-waste as a cement substitute. Agro-wastes have been widely used as cement substitutes, including sugarcane bagasse ash [21], rice husk ash [26, 34, 53], and sawdust ash [5]. The existing research literature indicates that construction materials, including agro-waste, have been found to meet the minimal standards given in regional building codes. Moreover, the use of these agricultural by-products in the manufacturing of building materials results in a decrease in ecological harm [59]. Most agricultural waste is unprocessed, unused, and often indiscriminately burned, dumped, or landfilled [45].

One of these agricultural wastes is groundnut shells, a by-product of groundnut (peanuts) manufacturing. Global peanut production was peak at around 47 million tonnes in 2020. China was the largest producer, accounting for 40% (or 18 million tonnes) of world peanut production [49, 52]. About 21–29% of the weight of the peanut is in the shell [11, 13]. Thus, the peanut industry generates about 11 million tons of peanut shell waste yearly [40]. In addition, a significant quantity of peanut shells is utilized as biomass for energy. However, a greater volume of discarded peanut shells is disposed away with ordinary waste. Using groundnut shells and their derivatives as construction materials is a viable solution to mitigate the environmental challenges linked to cement consumption and the management of groundnut shell waste. Numerous research has been conducted on the usage of GSA as a potential alternative to cement in concrete and cement mortar. Additionally, GSA has been investigated as a stabilizing agent for soil, road foundation, and masonry blocks. Furthermore, its application as a precursor in the development of geopolymer materials has also been explored.

When GSA is utilized as a partial substitute for cement, the strength characteristics of concrete are influenced not only by the quantity of GSA replacement but also by several other factors, including cement content, W/C ratio, coarse aggregate content, fine aggregate content, and curing length. Therefore, it is crucial to examine the impact of these factors on the *CS* of GSA blended concrete and put forward a method for forecasting the *CS* of GSA blended concrete. Nevertheless, a prediction model for the *CS* of concrete with GSA has not yet been developed.

In recent years, engineers and academics have become increasingly interested in using ML techniques to predict the characteristics of building materials [14, 33, 50, 51]. The properties of GSA mixed concrete are sensitive to the mixing proportions and are influenced by many variables, making ML approaches the best option for predicting these properties. There is a suggestion to use more advanced techniques to minimize reliance on laboratory testing. Additionally, engineers should be equipped with essential tools and mathematical equations to predict the outcomes of tests [49, 56]. ML techniques may be used to provide alternate approaches and resolutions for both linear and nonlinear scenarios, whereby mathematical models are unsuccessful in precisely defining the interdependencies among the variables implicated in a given issue [15, 62].

The primary objective of the current study is to use ML approaches to forecast the *CS* of GSA mixed concrete. Consequently, mixed design elements are utilized to develop predictive models for *CS*, enabling the utilization of these models in the construction industry without the need for previous theoretical comprehension. To determine the most precise and dependable model for predicting the *CS* of GSA blended concrete, a statistical evaluation was conducted using seven distinct machine-learning techniques. These techniques included linear regression, full quadratic model, artificial neural network, boosted decision tree regression, random forest regression, k-nearest neighbors, and support vector regression. The proposed models provide a means to enhance the accuracy of predicting the *CS* of GSA mixed concrete.

Methods

The approach used in the present work encompasses a sequence of procedures, visually shown as a flowchart in Fig. 1. The primary procedures include the following actions:

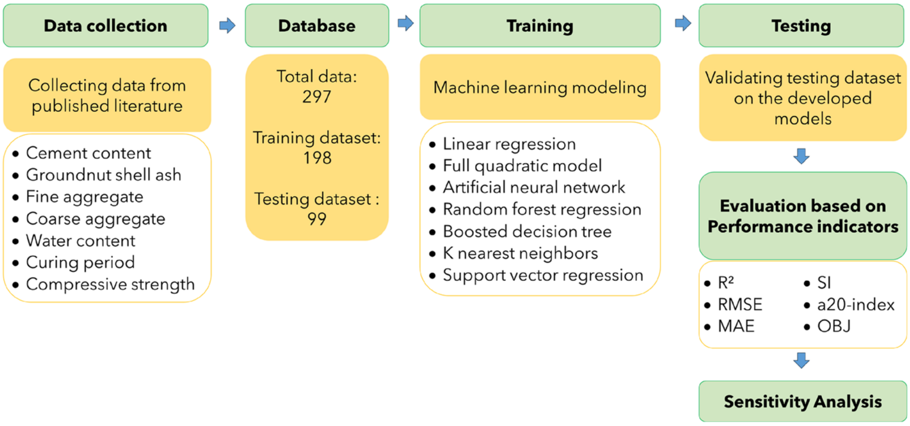


Fig. 1 Methodology of research

- Generating and gathering information about GSA mixed concrete from existing scholarly sources.
- The predictors for the models include cement content (CC), groundnut shell ash content (GSA), fine aggregate (FA), coarse aggregate (CA), quantity of water content (WC), and curing duration (t). The target value of the models is the *CS* of concrete.
- The acquired data is randomly divided into two groups, with two-thirds of the total data assigned to the training group and the remaining one-third allocated to the testing group.
- Developing predictive models using LR, FQ, ANN, BDT, RFR, KNN, and SVR machine learning algorithms.
- Assessing the proposed models based on the following performance indicators: R^2, RMSE, MAE, SI, a-20 index, and OBJ.
- The present study uses SHAP analysis to perform sensitivity analysis to identify the crucial parameter for accurately forecasting the *CS* of GSA mixed concrete.

Data collection

The database used in this work was constructed using seventeen previously published experimental investigations, as shown in Table 1. These studies were employed to create 297 experimental datasets for the database. All datasets included in this study were generated via samples that adhered to internationally recognized standards for casting and testing. The parameters used in this database were as follows:

- Cement content, CC (in kg/m^3)
- Groundnut shell ash content, GSA (in kg/m^3)
- Fine aggregate content, FA (in kg/m^3)
- Coarse aggregate content, CA (in kg/m^3)
- Water content, WC (in kg/m^3)
- Curing period, t (*in days*)
- Compressive strength, CS (*in MPa*)

The collected data were partitioned into two groups with the RAND function. The first subset of 198 data sets accounted for about two-thirds of the total data and was used for model development. The remaining dataset, which accounted for one-third of the total data, was utilized to verify the models constructed based on the first group.

Machine learning modeling

Linear regression

LR is a supervised ML approach that is utilized to determine the linear association between a dependent parameter and a set of independent parameters. The model postulates a linear association between the input parameters and the only output parameter. The purpose of the technique is to recognize the optimal linear equation that can effectively forecast the value of the dependent variable by using the independent variables. The linear regression (LR) model, as stated in Eq. (1), was used to predict the *CS* of GSA blended concrete.

Table 1 The experimental data set collected from published literature

Ref.	Cement (kg/m^3)	GSA (kg/m^3)	FA (kg/m^3)	CA (kg/m^3)	Water (kg/m^3)	Curing period (days)	Strength (MPa)	No. of data
Abro, Kumar et al. [2]	264–310	0–46.5	620	1240	155	3, 28	13.2–36.9	12
Alabadan, Olutoye et al. [4]	155–310	0–155	620	1240	171	7, 14, 21, 28	2.3–31.4	24
Buari, Olutoge et al. [8]	276–460	0–184	650	1068	170	7, 14, 28	9.8–49.0	15
Dharani and Selvan [12]	263–376	0–113	636	1103	177	7, 28	11.1–29.0	12
Kanchidurai, Nanthini et al. [28]	314–392	0–78.4	477	1353	216	7, 28	19.9–29.8	10
Karthikeyan, Saravanan et al. [29]	309–425	0–106	510	1234	234	7, 14	5.5-29.1	12
Krishnan and Nizar [31]	326–383	0–57.5	549	1184	192	7, 14, 28	7.2–30.1	21
Ige, Anifowose et al. [18]	248–310	0–62	620	1240	186	7, 14, 28	10.5–24.3	15
Ikumapayi, Arum et al. [19]	260–310	0–49.6	620	1240	186	7, 28	12.6–16.3	14
Lakshmi and Sagar [32]	202–310	0–109	620	1240	186	7, 14, 28	5.9–23.2	24
Mujedu and Adebara [35]	77.5–310	0–233	620	1240	171	7, 28	1.8–26.7	24
Nwofor and Sule [36]	186–310	0–124	620	1240	171	7, 14, 21, 28	2.0–25.5	20
Ogork, Uche et al. [37]	191–318	0–127	705	1252	175	7, 28, 60, 90	5.5–31.1	24
Pandi, Ganesan et al. [39]	289–385	0–96.3	577	1154	193	28, 56, 90	18.2–26.0	18
Raheem, Oladiran et al. [44]	264–368	0–73.6	620–846	956–1240	171–202	7, 14, 21, 28	8.2–35.1	28
Samuel [46]	353–470	0–118	705	940	282	3, 7, 21, 28	4.8–26.3	24
Overall	**77.5–470**	**0–233**	**477–846**	**940–1353**	**155–282**	**3–90**	**1.79–49.0**	**297**

$$CS = \alpha_0 + \alpha_1(CC) + \alpha_2(GSA) + \alpha_3(FA) + \alpha_4(CA) + \alpha_5(WC) + \alpha_6(t) \tag{1}$$

where a_0 to a_6 are model parameters.

Full quadratic (FQ) model

The full quadratic regression model is a kind of regression analysis that represents the association between the independent and the dependent parameters as a polynomial of degree two in the independent parameters. Polynomial regression is a kind of linear regression that involves using a polynomial equation to represent data that demonstrates a nonlinear relationship among the dependent and independent parameters. Equation 2 introduces a complete quadratic formula that provides a relationship between *CS* and the first and second orders of each independent parameter [27].

$$\begin{aligned} CS = {} & \beta_0 + \beta_1(CC) + \beta_2(GSA) + \beta_3(FA) + \beta_4(CA) + \beta_5(WC) + \beta_6(t) + \beta_7(CC)^2 \\ & + \beta_8(GSA)^2 + \beta_9(FA)^2 + \beta_{10}(CA)^2 + \beta_{11}(WC)^2 2 + \beta_{12}(t)^2 + \beta_{13}(CC)(GSA) \\ & + \beta_{14}(CC)(FA) + \beta_{15}(CC)(CA) + \beta_{16}(CC)(WC) + \beta_{17}(CC)(t) + \beta_{18}(GSA)(FA) \\ & + \beta_{19}(GSA)(CA) + \beta_{20}(GSA)(WC) + \beta_{21}(\mathrm{GSB})(t) + \beta_{22}(FA)(CA) \\ & + \beta_{23}(FA)(WC) + \beta_{24}(FA)(t) + \beta_{25}(CA)(WC) + \beta_{26}(CA)(t) + \beta_{27}(WC)(t) \end{aligned} \tag{2}$$

where β_0 to β_{27} are model parameters.

Artificial neural network (ANN) model

ANNs are computer models that draw inspiration from the functioning of biological neural networks. ANNs consist of linked processing nodes, often called neurons, which can acquire knowledge and identify patterns within incoming data. ANNs are used for pattern recognition, data classification, and making predictions. ANNs have self-learning proficiencies and can provide better results as more data is available [24, 25]. The output of each layer is calculated by taking the sum of its inputs and applying a nonlinear function to it. Given the absence of a standardized approach for constructing the network architecture, the number of hidden layers and neurons was determined by implementing a parameter optimization technique [23]. To mitigate the issue of overfitting, a decision was made to maintain simplicity in the model architecture by using a solitary, hidden layer with three neurons. After several tests and cross-validation, these values were chosen.

Random forest regression

RFR is a kind of ensemble ML technique that creates several decision trees during the training phase. This method is used in regression problems, whereby the resultant prediction is the mean or average of the individual trees [6]. The RFR method is a widely used ML technique that aggregates the predictions of several decision trees to get a unified outcome. The acceptance of this tool has been driven by its user-friendly interface and versatile functionality, which enables it to address classification and regression tasks effectively.

Boosted decision tree

The BDT is an ML methodology that integrates numerous decision trees to enhance the precision of predictive outcomes [47]. It works by training each new tree to emphasize the training instances that were previously mis-modeled. This is done by fitting the residual of the trees that preceded it. Compared with random forest regression with a boosted decision tree, the main difference between the two methods is that in boosting, each tree is dependent on prior trees, while in random forests, each tree is independent of the others [41].

K-nearest neighbors

KNN is a nonparametric approach used in supervised learning to address classification and regression challenges [30]. The input comprises the k-nearest training instances within a given data collection. The main difference between KNN and ANN is that KNN is a simple algorithm that relies on the proximity of data points to make predictions, while ANNs are more complex models that can learn to recognize patterns in data through training.

Support vector regression

SVR is a supervised ML algorithm specifically designed to address regression problems. SVR is a computational technique that aims to identify a mathematical function that effectively models the association between input and output variables, minimizing the overall error [30]. Additionally, SVR permits some flexibility within a predefined range, allowing for some departure. The input data is transformed by SVR into a high-dimensional feature space, allowing a linear model to be fitted using kernel functions. SVR is robust to outliers and can handle nonlinear and high-dimensional data.

Performance indicators

The evaluation of the created models encompasses several metrics, including the R^2, RMSE, MAE, scatter index, a20-index, and OBJ. It is anticipated that the values of the a20-index will be equal to one for an ideal prediction model. The a20-index, as developed, has the benefit of possessing a tangible engineering interpretation. It quantifies the sample count that meets anticipated values within a 20% deviation from experimental values. Equations 3, 4, 5, 6, 7 and 8 are used to calculate each specified criterion.

$$R^2 = \left(\frac{\sum_i \left(P_i - \overline{P}\right)\left(E_i - \overline{E}\right)}{\sqrt{\sum_i \left(P_i - \overline{P}\right)^2} \sqrt{\sum_i \left(E_i - \overline{E}\right)^2}} \right)^2 \tag{3}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (E_i - P_i)^2}{N}} \tag{4}$$

$$MAE = \frac{\sum_{i=1}^{n} (|E_i - P_i|)}{N} \tag{5}$$

$$SI = \frac{RMSE}{\overline{E}} \tag{6}$$

$$a20_{index} = \frac{N20}{N} \tag{7}$$

$$OBJ = \left(\frac{n_{tr}}{N} \times \frac{RMSE_{tr} + MAE_{tr}}{R_{tr}^2 + 1} \right) + \left(\frac{n_{te}}{N} \times \frac{RMSE_{te} + MAE_{te}}{R_{te}^2 + 1} \right) \tag{8}$$

where

Pi: Predicted *CS*

Ei: Experimental *CS*

$\overline{P}$: Mean of predicted *CS*

E: Mean of experimental *CS*

N: Total number of dataset

$N20$: Total number of predicted to the measured data of *CS* ratio ranged from 0.8 to 1.2

n_{tr}: Number of the training dataset

n_{te}: Number of the test dataset

The R^2 value and the a-20 index typically range from zero to one, with a value of 1 being regarded as the optimal outcome. The RMSE, MAE, and OBJ values range from zero to infinity. It is advisable to minimize these values, with zero being the optimal outcome. Additionally, if the value of the SI metric is less than 0.1, the model may be classified as exhibiting good performance. The SI value ranges from 0.1 to 0.2, 0.2 to 0.3, and more than 0.3, denoting the model's performance as excellent, fair, and bad, respectively [1, 3, 27].

Results and discussion

Statistical analysis

Statistical analyses were conducted to evaluate the connection among the factors listed. The relationship between *CS* and the dependent variables is shown in Fig. 2. The findings suggest a reasonable association exists between the amount of cement used and the GSA (ground slag aggregate) content with respect to *CS*, as seen in Fig. 3. However, the relationship between fine aggregate, coarse aggregate, water, and curing time with *CS* was low. The statistical analysis findings are succinctly presented in Table 2.

Machine learning model results

Figure 4 illustrates GSA blended concrete's predicted vs. measured *CS* values for all seven machine-learning models discussed. Table 3 summarizes the performance indicators for each model.

The LR model is a fundamental mathematical model used to predict the *CS* of concrete. The outcome of the LR model is revealed in Eq. (9). Figure 8a illustrates the correlation between the anticipated and observed *CS*. The training dataset has an R^2 value of 0.608 and an RMSE value of 5.11 MPa. Moreover, the testing dataset exhibited an R^2 value of 0.643 and a RMSE of 4.29 MPa. Based on the obtained R^2 and RMSE data, it can be concluded that the performance of the LR model is unsatisfactory. The outcomes of the LR model are among the least effective, mainly owing to its simplistic mathematical formulation. The error range in the training dataset is −20 to 20%. This indicates that 55% of the data is within the range of 0.8 to 1.2 for the ratio used to estimate *CS*.

$$CS = -37.78 + 0.119(CC) + 0.024(GSA) + 0.016(FA) + 0.023(CA) - 0.117(WC) + 0.135(t) \tag{9}$$

Due to its advanced mathematical formulation, the FQ model is one of the most successful mathematical models. It has been developed using mathematical criteria such as constants, linearity, variable product terms and interactions, and quadratic variables. The formula for the FQ model predicting the *CS* of GSA mixed concrete is shown in Eq. 10. Figure 4b shows the relationship between the predicted and measured *CS* of the FQ model. The R^2 and RMSE for the training data were 0.865 and 3.00 MPa, respectively, while for the test data, they were 0.766 and 3.48 MPa, respectively. For the expected observed *CS* ratio, 69% of the data falls between 0.8 and 1.2 in the training data set, with an error line of −20 to 20%. Although FQ model performs better than LR models, its prediction accuracy is still less than RFR and BDT models.

$$\begin{aligned} CS = {} & 154,115.4 - 176.7(CC) - 175.8(GSA) + 47.4(FA) - 342.8(CA) + 903.7(WC) - 1.5(\mathrm{t}) \\ & + 0.3(\mathrm{CC})^2 + 0.3(GSA)^2 - 0.006(FA)^2 + 0.1(CA)^2 + 0.02(W)^2 - 0.005(t)^2 \\ & + 0.6(CC)(GSA) - 0.2(CC)(FA) + 0.2(CC)(CA) - 0.6(CC)(WC) \\ & + 0.007(CC)(t) - 0.2(GSA)(FA) + 0.2(GSA)(CA) - 0.6(GSA)(WC) \\ & + 0.006(GS\mathrm{B})(t) + 0.1(FA)(CA) - 0.3(FA)(WC) \\ & + 0.001(FA)(t) - 0.4(CA)(WC) + 0.0008(CA)(t) - 0.01(WC)(t) \end{aligned} \tag{10}$$

Other ML models except KNN show better performance indicator values than LR and FQ models. RFR models show R^2 closer to unity and lower RMSE, MAE, and SI values than other ML models. For the predicted to observed *CS* ratio, 81% of the data falls between 0.8 and 1.2 in the training data set, which is 8% higher than the

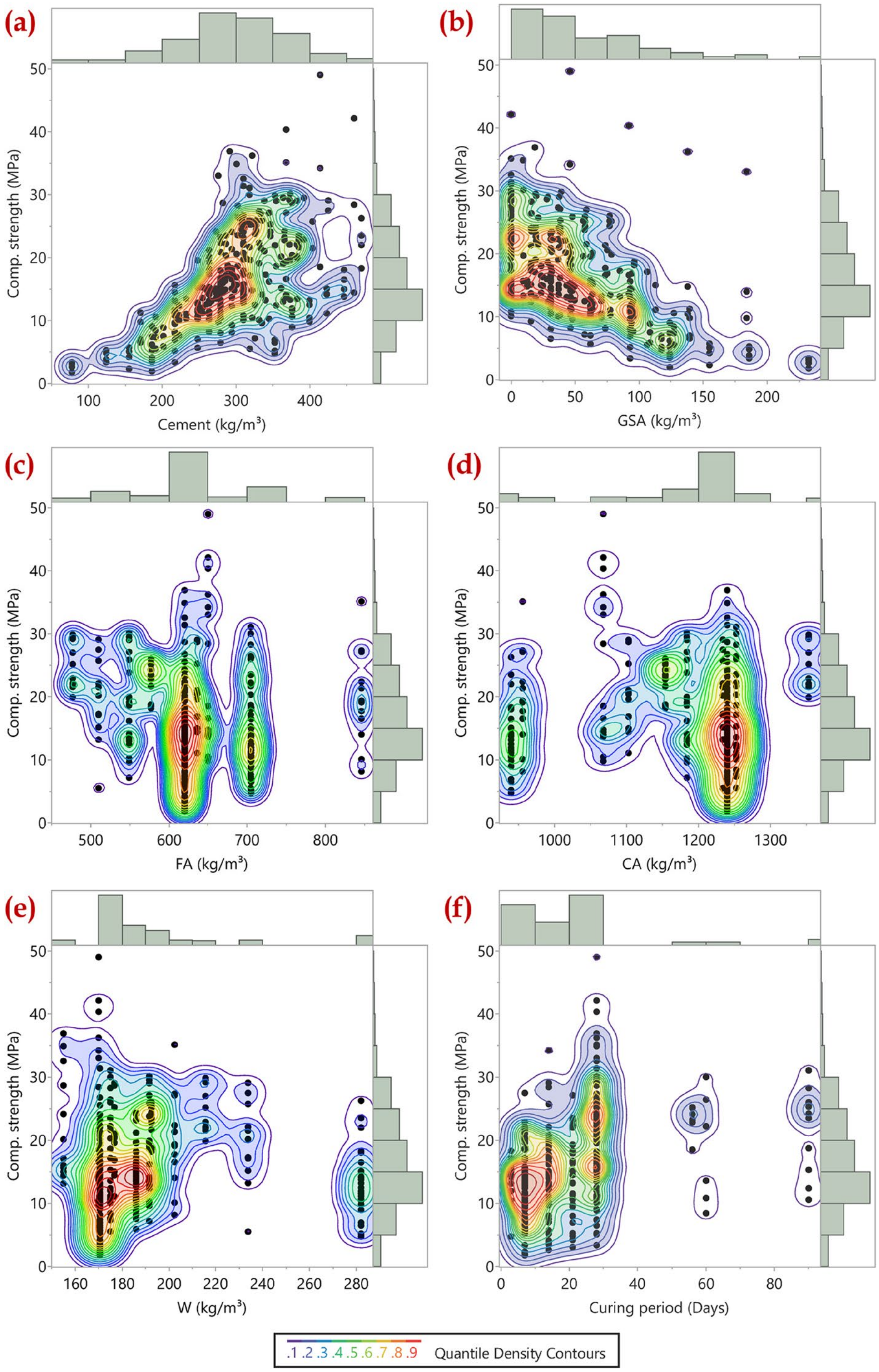


Fig. 2 The variation of *CS* with independent parameters

next-best model (SVR). The precision of the RFR and BDT models in predicting the *CS* of concrete is comparatively good and ranked as 1 and 2, respectively. ANN and FQ models perform closer to each other and are ranked as 3 and 4, respectively. It is followed by SVR, KNN, and LR models. The RFR model has more points inside the 20% error envelope with 81% of the total data, followed by SVR and BDT at 73% and

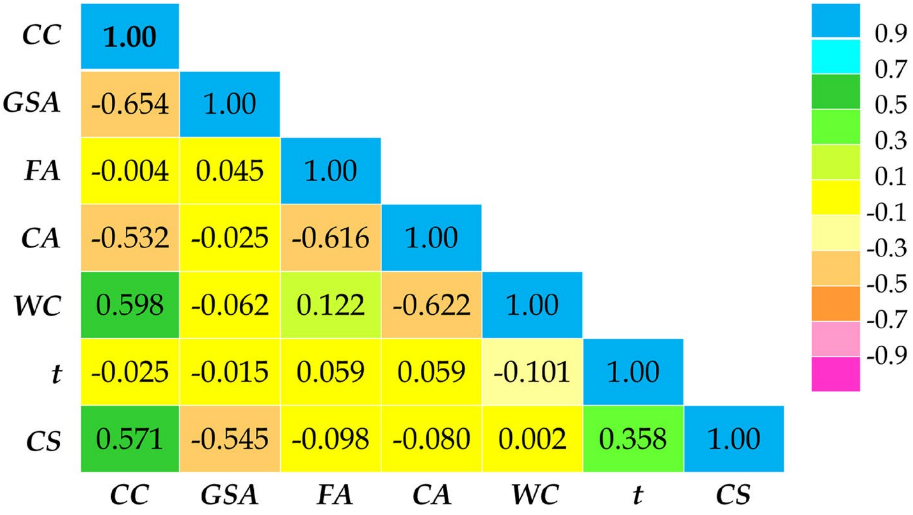


Fig. 3 Correlation matrix graph among the independent and dependent variables and of GSA blended concrete

Table 2 The statistical analysis of the dataset

	CC (kg/m^3)	GSA (kg/m^3)	FA (kg/m^3)	CA (kg/m^3)	WC (kg/m^3)	t (days)	CS (MPa)
Min	78	0	477	940	155	3	2
Max	470	233	846	1353	282	90	49
Mean	298	56	628	1185	190	21	17
Standard deviation (SD)	73.9	49.5	69.7	104.5	31.5	18.5	7.9
Variance (Var)	5474	2458	4877	10959	998.7	342.1	62.27
Kurtosis (Kur)	0.48	1.36	2.36	0.80	3.09	5.60	0.67
Skewness (skew)	−0.22	1.15	0.79	−1.34	1.95	2.23	0.66

72%, respectively. Overall, the RFR model is the best option for predicting the *CS* of GSA blended concrete.

Performance of machine learning models

Figure 5 displays the prediction error for all the examined machine-learning models, calculated as the difference between the expected and observed *CS*. The shown chart demonstrates that many data points have been identified as outliers, suggesting a higher level of inaccuracy. This phenomenon may arise because of inaccuracies in the experimental measurement of *CS* or discrepancies among the laboratory tests conducted in the literature. BDT, FQ, and RFR models show a narrow range of error distribution (highest–lowest error) as 17.56, 18.03, and 18.93 MPa, respectively. LR and KNN models show the most comprehensive range error distribution as 35.95 and 31.07 MPa, respectively.

The BDT and KNN models have errors evenly distributed on both sides. This suggests that the projected values are underestimated in some instances, while in others, they are overstated. However, under other models, the majority of errors are undervalued. Furthermore, except for the BDT model, all other models exhibit negative skewness. The SVR model has the greatest skewness value of −1.54, while the KNN model follows closely with a skewness of −0.86. Using several statistical and

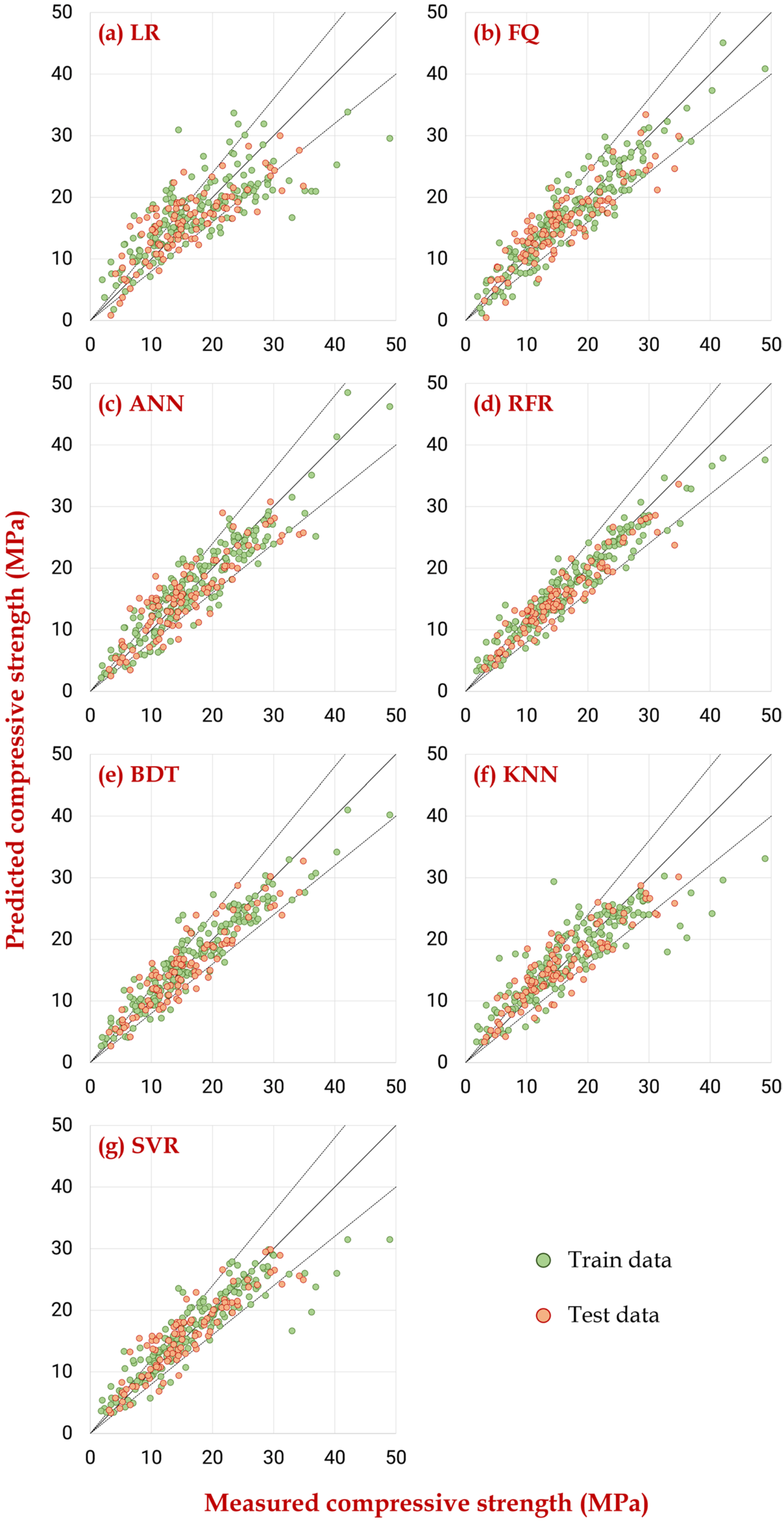


Fig. 4 Predicted vs. measured compressive strength comparison for various ML models

Table 3 Performance indicators for various ML models

	Train					Test					OBJ	Ranking
	R^2	RMSE	MAE	SI	a20	R^2	RMSE	MAE	SI	a20		
LR	0.6083	5.11	3.82	0.30	0.55	0.6435	4.29	3.44	0.27	0.52	5.27	7
FQ	0.8654	3.00	2.34	0.18	0.69	0.7659	3.48	2.82	0.22	0.61	3.10	4
ANN	0.8697	2.95	2.34	0.17	0.69	0.7665	3.47	2.78	0.22	0.60	3.07	3
RFR	0.9079	2.48	1.80	0.15	0.81	0.8866	2.42	1.83	0.15	0.80	2.25	1
BDT	0.8849	2.77	2.12	0.16	0.72	0.8357	2.91	2.33	0.19	0.70	2.68	2
KNN	0.7490	4.09	2.80	0.24	0.64	0.7939	3.26	2.58	0.21	0.64	3.71	6
SVR	0.8020	3.63	2.32	0.21	0.73	0.8197	3.05	2.29	0.20	0.70	3.18	5

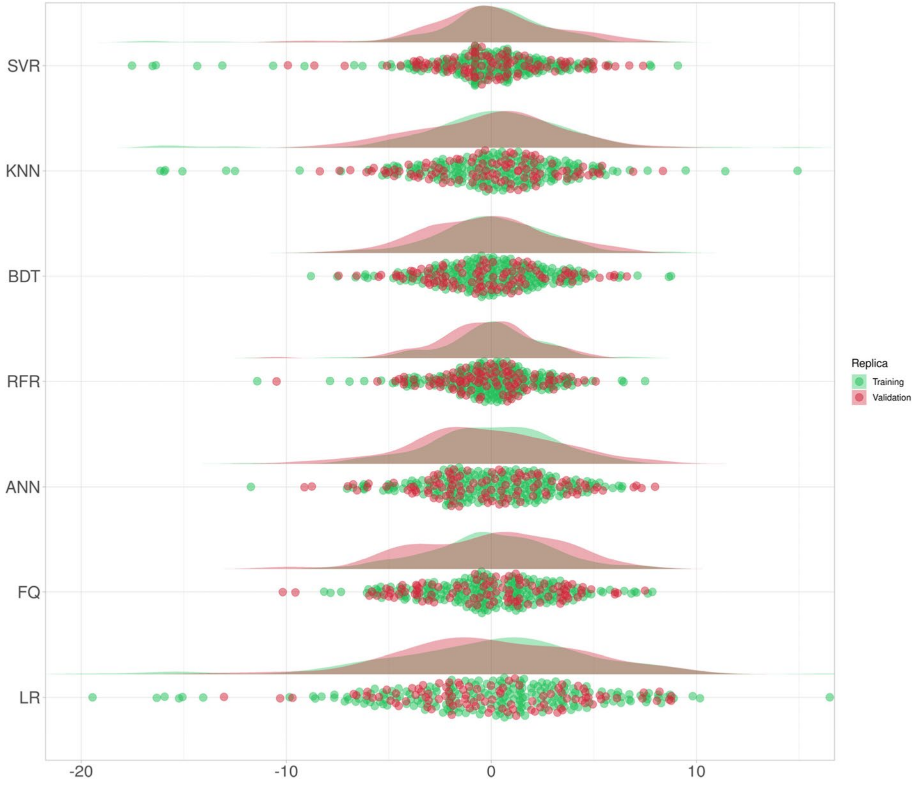
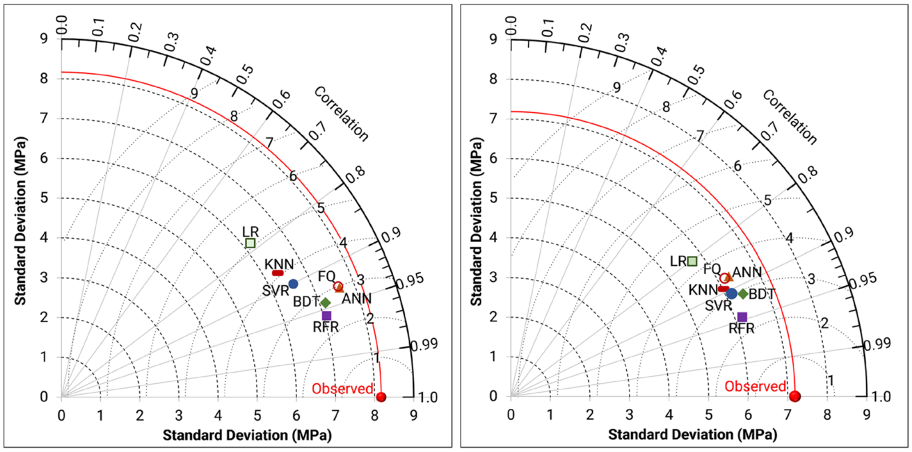


Fig. 5 Error distribution in predicted compressive strength for various ML models

graphical techniques may significantly improve the assessment of prediction models. This implies that using a variety of statistical indicators and graphical illustrations to assess the efficacy of prediction models may provide a more thorough analysis.

Figure 6 depicts the Taylor diagram, a graphical representation utilized to evaluate the predictive performance of ML models. The Taylor diagram, a statistical tool, provides a visual framework for evaluating and comparing several models. The graphic illustrates the degree of alignment between each model and the reference data, as measured by correlation, standard deviation, and RMSE. The diagram can visually represent the comparative proficiency of each model concerning a reference model [58]. The proximity of



(a) Train **(b) Test**

Fig. 6 Taylor diagram of ML models (the red point represents the reference for measured compressive strength)

the pentagram to the reference spot directly correlates with the model's accuracy in forecasting *CS*. Among all ML models, RFR model exhibits the highest level of accuracy, while LR models provide the lowest level of accuracy. According to these criteria, the ML models may be ranked in the following order from highest to lowest performance: RFR > BDT > ANN > FQ > SVR > KNN > LR. The results demonstrate a strong correlation with the previously specified performance indicator values.

Sensitivity analysis

Nonlinear and complicated models such as ANN or RFR sometimes exhibit black box behavior due to their intricate nature [54]. The use of SHAP is quite advantageous in examining intricate machine-learning models encompassing a diverse range of parameters [43, 63]. The decision to use the findings of the random forest regression (RFR) model, which demonstrated superior performance in predicting *CS*, was made to gain insights into the outcomes via applying the SHAP (SHAPley Additive exPlanations) method.

Figure 7 depicts the average SHAP values about various characteristics, which are the independent or input variables, concerning the predictions of *CS*. These predictions are derived from the random forest regression (RFR) model. Based on the findings, it is evident that the cement content exhibits the highest SHAP value, indicating its significant effect on the prediction of *CS*. Concurrently, it was observed that the fine aggregate content exhibited the lowest SHAP value, suggesting a relatively lesser impact on the prediction of *CS*.

Figure 8 displays the SHAP summary plots depicting the predictions of *CS* for concrete using the RFR model. The color gradient represents the spectrum of feature values, while the x-axis denotes the SHAP value or the feature's contribution towards the anticipated *CS*. The red dot represents a notably high feature value, indicating a correspondingly high SHAP score. A notable finding in the current research is identifying an extremely positive SHAP value of 16, indicating that the range of cement content examined can increase *CS* by 16 MPa over the average value. Conversely, a SHAP value of −16

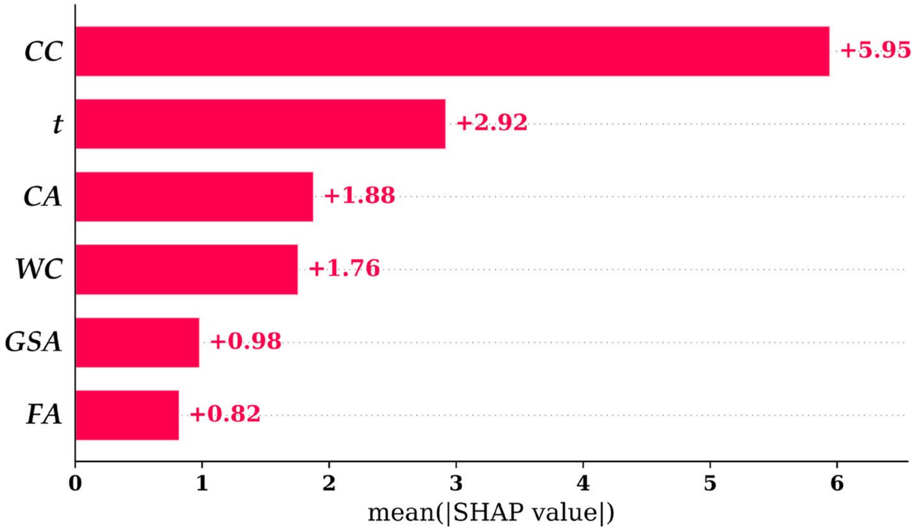


Fig. 7 Mean SHAP values

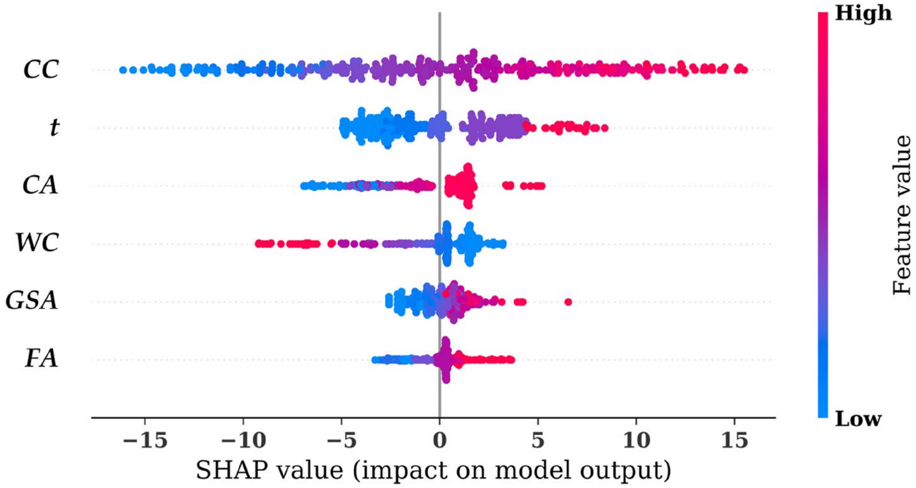


Fig. 8 SHAP summary plot

on the far-left end (negative) indicates that a reduction in cement concentration might result in a loss in *CS* by 16 MPa below the mean value. The results from SHAP indicate that utilizing the game theory approach for calculating SHAP might enhance the understanding of the proposed hybrid ML models. Additionally, these findings demonstrate that the predictive accuracies of the models are both reasonable and satisfactory.

Conclusions

This paper presents a way to use ML algorithms to assess the CS of GSA mixed concrete. To train and assess the models, a total of 297 experimental data sets were collected from published literature. Several baseline predictors were constructed and trained, including linear regression, full quadratic model, artificial neural network, random forest regression, boosted tree regression, K closest neighbors, and support vector regression. The results of the study may potentially provide the following inferences:

- Statistical analysis shows that *CS* is moderately correlated with cement content, GSA content, and curing time. However, it correlates poorly with fine aggregate content, coarse aggregate content, and water content.
- Among the several ML models evaluated in this study, random forest regression (RFR) demonstrated superior performance in predicting *CS*, achieving an R^2 value of 0.91 and RMSE of 2.48 MPa for the training dataset and an R^2 value of 0.89 and an RMSE of 2.42 MPa for the testing dataset.
- The RFR model was first graded concerning various statistical techniques, like MAE, SI, and OBJ. The training dataset's mean absolute error (MAE) and structural integrity (SI) values were recorded as 1.80 MPa and 0.15, respectively. Similarly, for the testing dataset, the MAE and SI values were observed to be 1.83 MPa and 0.15 MPa, respectively. The RFR model exhibited the greatest a-20 index, with 81% and 80% values for the training and test datasets, respectively.
- The results obtained from the feature significance analysis using SHAP demonstrate that the parameter with the greatest influence on the prediction of *CS* is the cement content. This is followed by the curing duration, which exhibit significant relevance in the prediction model.

This study systematically evaluates the predictive capabilities of the *CS* of GSA mixed concrete, contributing to the existing body of knowledge and practical implementation in this domain. It is crucial to bear in mind that augmenting the ML model with more data has the potential to enhance its performance. Consequently, it is vital to maintain a comprehensive data collection. Using precisely predicted model techniques may assist researchers and designers in selecting optimal input variables and making educated selections about the appropriate mix parameters to employ in developing sustainable concrete with desired attributes.

Abbreviations

ANN	Artificial neural network
BTR	Boosted tree regression
CC	Cement content
CA	Coarse aggregate
CS	Compressive strength
FA	Fine aggregate
FQ	Full quadratic
GSA	Groundnut shell ash
LR	Linear regression
KNN	K-nearest neighbors
MAE	Mean absolute error
ML	Machine learning
OBJ	Objective function value
RFR	Random forest regression
RMSE	Root-mean-squared error
R^2	Coefficient of determination
SHAP	SHAPley Additive exPlanations
SI	Scatter index
SVR	Support vector regression
t	Curing period
WC	Quantity of water requirement
W/C	Water to cement ratio

Acknowledgements

None.

Authors' contributions

Study conception and design, NS and PJ; data acquisition, NS; analysis and interpretation of results, PJ; draft manuscript preparation, NS and PJ; and manuscript review and editing, NS.

Funding

Not applicable.

Availability of data and materials

Data can be made available on request by interested parties.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 10 September 2023 Accepted: 9 October 2023
Published online: 06 November 2023

References

1. Abdalla AA, Salih Mohammed A (2022) Theoretical models to evaluate the effect of SiO2 and CaO contents on the long-term compressive strength of cement mortar modified with cement kiln dust (CKD). Arch Civil Mech Eng 22(3):105
2. Abro A, Kumar A, Keerio M, Bheel N (2021) An investigation on compressive strength of concrete blended with groundnut shell ash. Neutron 20(2):123–127
3. Ahmed HU, Abdalla AA, Mohammed AS, Mohammed AA (2022) Mathematical modeling techniques to predict the compressive strength of high-strength concrete incorporated metakaolin with multiple mix proportions. Clean Mater 5:100132
4. Alabadan B, Olutoye M, Zakariya M (2005) Partial replacement of ordinary portland cement (OPC) with bambara groundnut shell ash (BGSA) in concrete. Leonardo Electron J Pract Technol 6:43–48
5. Assiamah S, Agyeman S, Adinkrah-Appiah K, Danso H (2022) Utilization of sawdust ash as cement replacement for landcrete interlocking blocks production and mortarless construction. Case Stud Construct Mater 16:e00945
6. Belgiu M, Drăguţ L (2016) Random forest in remote sensing: a review of applications and future directions. ISPRS J Photogramm Remote Sens 114:24–31
7. Benhelal E, Zahedi G, Shamsaei E, Bahadori A (2013) Global strategies and potentials to curb CO2 emissions in cement industry. J Clean Product 51:142–161
8. Buari TA, Olutoge FA, Ayinnuola GM, Okeyinka OM, Adeleke JS (2019) Short term durability study of groundnut shell ash blended self consolidating high performance concrete in sulphate and acid environments. Asian J Civil Eng 20(5):649–658
9. Chandra Paul S, Mbewe PBK, Kong SY, Šavija B (2019) Agricultural solid waste as source of supplementary cementitious materials in developing countries. Materials (Basel, Switzerland) 12(7):1112
10. Claisse PA (2016) Chapter 17 - Introduction to cement and concrete. In: Claisse PA (ed) Civil Engineering Materials. Butterworth-Heinemann, Boston, pp 155–162
11. Davis, J. P. and L. L. Dean (2016). Chapter 11 - Peanut composition, flavor and nutrition. Peanuts. H. T. Stalker and R. F. Wilson, AOCS Press, Urbana, p 289–345
12. Dharani D, Selvan A (2017) Durability studies on concrete by using groundnut shell ash as mineral admixture. Int J Innov Res SciTechnol 3(10):168–172
13. Duc PA, Dharanipriya P, Velmurugan BK, Shanmugavadivu M (2019) Groundnut shell -a beneficial bio-waste. Biocatal Agric Biotechnol 20:101206
14. Feng D-C, Liu Z-T, Wang X-D, Chen Y, Chang J-Q, Wei D-F, Jiang Z-M (2020) Machine learning-based compressive strength prediction for concrete: an adaptive boosting approach. Construct Build Mater 230:117000
15. Gao W, Karbasi M, Derakhsh AM, Jalili A (2019) Development of a novel soft-computing framework for the simulation aims: a case study. Eng Comput 35(1):315–322
16. Greenspec. (2022). Environmental impacts of concrete. Green building design Retrieved April 21, 2022, fromhttps://www.greenspec.co.uk/building-design/environmental-impacts-of-concrete/
17. Habert G (2014) Chapter 10 - Assessing the environmental impact of conventional and 'green' cement production. In: Pacheco-Torgal F, Cabeza LF, Labrincha J, de Magalhães A (eds) Eco-efficient Construction and Building Materials. Woodhead Publishing, pp 199–238
18. Ige J, Anifowose M, Amototo I, Adeyemi A, Olawuyi M (2017) Influence of groundnut shell ash (GSA) and calcium chloride (CaCl2) on strength of concrete. Int J Eng Tome 15(4):209–214
19. Ikumapayi CM, Arum C, Alaneme KK (2021) Reactivity and hydration behavior in groundnut shell ash based pozzolanic concrete. Mater Today 38:508–513
20. Imam A, Kumar V, Srivastava V (2018) Review study towards effect of silica fume on the fresh and hardened properties of concrete. Adv Concrete Construct 6(2):145–157
21. Jahanzaib Khalil M, Aslam M, Ahmad S (2021) Utilization of sugarcane bagasse ash as cement replacement for the production of sustainable concrete – a review. Construct Build Mater 270:121371

22. Játiva A, Ruales E, Etxeberria M (2021) Volcanic ash as a sustainable binder material: an extensive review. Materials 14(5):1302
23. Jeyananthan P (2022) Prolonged viral shedding prediction on non-hospitalized, uncomplicated SARS-CoV-2 patients using their transcriptome data. Comput Methods Programs Biomed Update 2:100070
24. Jeyananthan P (2023) Role of different types of RNA molecules in the severity prediction of SARS-CoV-2 patients. Pathol Res Pract 242:154311
25. Jeyananthan P (2023) SARS-CoV-2 diagnosis using transcriptome data: a machine learning approach. SN Comput Sci 4(3):218
26. Jittin V, Bahurudeen A, Ajinkya SD (2020) Utilisation of rice husk ash for cleaner production of different construction products. J Clean Product 263:121578
27. Kakasor Ismael Jaf D, Ismael Abdulrahman P, Salih Mohammed A, Kurda R, Qaidi S. M. A, Asteris P. G (2023) "Machine learning techniques and multi-scale models to evaluate the impact of silicon dioxide (SiO2) and calcium oxide (CaO) in fly ash on the compressive strength of green concrete. Construct Build Mater 400:132604
28. Kanchidurai S, Nanthini T, Jai Shankar P (2017) Experimental studies on sisal fibre reinforced concrete with groundnut shell ash. ARPN J Eng Appl Sci 12(21):5914–5920
29. Karthikeyan N, Saravanan M, Deepika M (2018) Performance of groundnut shell ash as partial replacement of cement in concrete. Int J SciResDev 6(4):525–528
30. Kenyhercz v, Passalacqua N. V. (2016) Chapter 9 - Missing data imputation methods and their performance with biodistance analyses. In: Pilloud M. A., Hefner M. A. (eds) Biological Distance Analysis. Academic Press, San Diego, pp 181–194
31. Krishnan C, Nizar N (2016) Groundnut shell ash as partial replacement of cement in concrete. IJRDO-J Mech Civil Eng 2(2):39–48
32. Lakshmi N, Sagar P (2017) Study on partial replacement of groundnut shell ash with cement. Chall J Concrete Res Lett 8(3):84–90
33. Marani A, Nehdi ML (2020) Machine learning prediction of compressive strength for phase change materials integrated cementitious composites. Construct Build Mater 265:120286
34. Mayooran S, Ragavan S, Sathiparan N (2017) Comparative study on open air burnt low- and high-carbon rice husk ash as partial cement replacement in cement block production. J Build Eng 13:137–145
35. Mujedu K, Adebara S (2016) The use of groundnut shell ash as a partial replacement for cement in concrete production. Int J Sci, EngEnviron Technol 1(3):32–39
36. Nwofor TC and Sule S (2012) Stability of groundnut shell ash (GSA)/ordinary portland cement (OPC)concrete in Nigeria. Adv Appl Sci Res 3:2283–2287
37. Ogork E, Uche O, Elinwa A (2014) A study on groundnut husk ash (GHA) - concrete under acid attack. Int J Modern Eng Res 4(7):30–35
38. Özbay E, Erdemir M, Durmuş Hİ (2016) Utilization and efficiency of ground granulated blast furnace slag on concrete properties – a review. Construct Build Mater 105:423–434
39. Pandi K, Ganesan K, Manickavalli M (2018) Studies on the partial replacement of fine aggregate with groundnut shell ash in concrete. Int J Curr Eng Sci Res 5(10):1–5
40. Perea-Moreno, M.-A., F. Manzano-Agugliaro, Q. Hernandez-Escobedo and A.-J. Perea-Moreno (2018). Peanut shell for energy: properties and its potential to respect the environment. Sustainability 10(9)
41. Plaia A, Buscemi S, Fürnkranz J, Mencía EL (2022) Comparing boosting and bagging for decision trees of rankings. J Classif 39(1):78–99
42. Poorveekan K, Ath KMS, Anburuvel A, Sathiparan N (2021) Investigation of the engineering properties of cementless stabilized earth blocks with alkali-activated eggshell and rice husk ash as a binder. Construct Build Mater 277:122371
43. Quan Tran V, Quoc Dang V, Si Ho L (2022) Evaluating compressive strength of concrete made with recycled concrete aggregates using machine learning approach. Construct Build Mater 323:126578
44. Raheem S, Oladiran G, Olutoge F, Odewumi T (2013) Strength properties of groundnut shell ash (GSA) blended concrete. J Civil Eng Construct Technol 4(9):275–284
45. Sadh PK, Duhan S, Duhan JS (2018) Agro-industrial wastes and their utilization using solid state fermentation: a review. Bioresour Bioprocess 5(1):1
46. Samuel V (2020) Groundnut shell ash: a local construction material in concrete production. Fane-Fane Int Multidiscip J 5(1):1–15
47. Sarailidis G, Wagener T, Pianosi F (2023) Integrating scientific knowledge into machine learning using interactive decision trees. Comput Geosci 170:105248
48. Sathiparan N (2021) Utilization prospects of eggshell powder in sustainable construction material – a review. Construct Build Mater 293:123465
49. Sathiparan N, Anburuvel A, Selvam VV (2023) Utilization of agro-waste groundnut shell and its derivatives in sustainable construction and building materials – a review. J Build Eng 66:105866
50. Sathiparan, N. and P. Jeyananthan (2023). Predicting compressive strength of cement-stabilized earth blocks using machine learning models incorporating cement content, ultrasonic pulse velocity, and electrical resistivity. Nondestruct Testing Eval :1–25. https://doi.org/10.1080/10589759.2023.2240940
51. Sathiparan N, Jeyananthan P (2023) Prediction of masonry prism strength using machine learning technique: effect of dimension and strength parameters. Materi Today Commun 35:106282
52. Sathiparan, N., P. Jeyananthan and D. N. Subramaniam (2023). Effect of aggregate size, aggregate to cement ratio and compaction energy on ultrasonic pulse velocity of pervious concrete: prediction by an analytical model and machine learning techniques. Asian J Civil Eng
53. Seevaratnam V, Uthayakumar D, Sathiparan N (2020) Influence of rice husk ash on characteristics of earth cement blocks. MRS Adv 5(54):2793–2805
54. Shah SFA, Chen B, Zahid M, Ahmad MR (2022) Compressive strength prediction of one-part alkali activated material enabled by interpretable machine learning. Construct Build Mater 360:129534

55. Siddique R, Klaus J (2009) Influence of metakaolin on the properties of mortar and concrete: a review. Appl Clay Sci 43(3):392–400
56. Subramaniam, D. N., P. Jeyananthan and N. Sathiparan (2023). Soft computing techniques to predict the electrical resistivity of pervious concrete. Asian J Civil Eng https://doi.org/10.1007/s42107-023-00806-y
57. Sundaralingam K, Peiris A, Anburuvel A, Sathiparan N (2022) Quarry dust as river sand replacement in cement masonry blocks: Effect on mechanical and durability characteristics. Materialia 21:101324
58. Taylor KE (2001) Summarizing multiple aspects of model performance in a single diagram. J Geophys Res 106(D7):7183–7192
59. Thanushan K, Sathiparan N (2022) Mechanical performance and durability of banana fibre and coconut coir reinforced cement stabilized soil blocks. Materialia 21:101309
60. Theconstructor. (2022). "Manufacture of cement- materials and manufacturing process of portlan cement." Building Technology Guide Retrieved April 21, 2022, fromhttps://theconstructor.org/building/manufacture-of-cement/13709/
61. Wang D, Shi C, Farzadnia N, Shi Z, Jia H (2018) A review on effects of limestone powder on the properties of concrete. Construct Build Mater 192:153–166
62. Wijekoon, S. H., T. Shajeefpiranath, D. N. Subramaniam and N. Sathiparan (2023). A mathematical model to predict the porosity and compressive strength of pervious concrete based on the aggregate size, aggregate-to-cement ratio and compaction effort. Asian J Civil Eng
63. Zhang J, Niu W, Yang Y, Hou D, Dong B (2022) Machine learning prediction models for compressive strength of calcined sludge-cement composites. Construct Build Mater 346:128442

Publisher's Note

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