Journal of Applied Research and Technology

www.jart.icat.unam.mx

ICAT Instituto de Ciencias Aplicadas y Tecnología

Journal of Applied Research and Technology 23 (2025) 659-671

Original

Mathematical Model and Machine Learning Techniques to Predict the Compressive Strength of Groundnut Shell Ash Blended Sandcrete

N. Sathiparan*[1] and P. Jeyananthan[2]

[1]*Department of Civil Engineering, Faculty of Engineering, University of Jaffna, Sri Lanka*
[2]*Department of Computer Engineering, Faculty of Engineering, University of Jaffna, Sri Lanka*

Received: 09 09 2024; Accepted: 02 25 2025
Available: 12 31 2025

Abstract: This study uses machine-learning (ML) methodologies to introduce predictive models for the compressive strength of sandcrete mixed with groundnut shell ash (GSA). The models were developed utilizing 140 datasets acquired from published articles. The datasets contained several input variables: aggregate-to-binder ratio, peanut shell ash concentration, and curing time. The output feature was the compressive strength of the sandcrete. Four mathematical and machine-learning models were used to predict the compressive strength of peanut shell ash-blended sandcrete. Based on analyses of several models, the boosted decision tree model outperformed others in predicting compressive strength. The sensitivity analysis outcomes of the boosted decision-tree model show that the aggregate-to-binder ratio was the most significant factor in determining compressive strength. Overall, the boosted decision-tree model achieved an R^2 of 0.965 during testing, indicating excellent predictive accuracy. Additionally, it was found that using 10% to 30% GSA as a cement substitute optimally enhances sandcrete strength. These findings contribute to the understanding of sustainable construction materials and support the practical application of GSA in construction.

Keywords: **sandcrete, machine learning, compressive strength, groundnut shell ash, SHAP analysis.**

*Corresponding author.
E-mail address: sakthi@eng.jfn.ac.lk (N. Sathiparan).
Peer Review under the responsibility of Universidad Nacional Autónoma de México.

1. Introduction

The use of cement in construction materials is a major concern due to its significant contribution to greenhouse gas emissions and associated environmental impacts. The cement production process is responsible for around 8% of the world's CO_2 emissions while consuming substantial energy and natural resources (Seevaratnam et al., 2020). Cement is known to have adverse environmental impacts, including topsoil erosion, increased surface runoff, and air and water pollution. Therefore, minimizing the use of cement in construction materials cannot be overemphasized in the pursuit of sustainability and climate change mitigation (Tan et al., 2024).

One potential strategy to reduce cement use in construction materials is to substitute some cement with supplementary cementitious materials (SCMs) (Adegbemileke et al., 2024; Sathiparan et al., 2022). SCMs have pozzolanic or hydraulic properties, allowing them to react with calcium hydroxide and waste to develop cementitious compounds (McCarthy & Dyer, 2019). SCMs have been used to improve the mechanical properties, durability, workability, and appearance of construction materials while reducing their economic and environmental footprint. Some examples of SCMs include fly ash, natural pozzolans, silica fume, slag, and agro-based cementitious materials (Kumar et al., 2021; Hafez et al., 2024). However, the availability and quality of these resources are often limited by declining coal use and the diversity of industrial processes (Millward-Hopkins et al., 2018).

In recent years, agro-based cementitious materials have been considered possible replacements or enhancements for traditional SCMs (Mayooran et al., 2017; Poorveekan et al., 2021). A notable advantage of agro-based SCMs over conventional SCMs is their improved sustainability, environmental friendliness, and local accessibility (Juenger et al., 2019). Agro-based cementitious materials are produced using agricultural waste such as rice husk, coconut husk, bagasse, olive waste, and similar materials. Typically, these waste materials are incinerated or disposed of in ways that result in the emission of pollutants into the atmosphere and the degradation of soil quality.

Cement substitutes can effectively reduce waste generation, and greenhouse gas emissions related to cement manufacture. Also, using agro-based cementitious materials can improve the mechanical characteristics and durability of construction materials by increasing their strength, workability, resistance to chemical degradation, and aesthetic appeal (Nilimaa, 2023). Agro-based cementitious materials are more readily available and cost-effective than conventional SCMs such as silica fume, fly ash, and slag (Blesson & Rao, 2023). Locally sourced agro-based cementitious materials can be produced from abundant, renewable sources, thereby reducing transport costs and energy consumption (Ortega et al., 2022). It is, therefore, evident that agro-based SCMs have the potential to substitute or complement conventional SCMs in various construction materials.

Groundnut shells are by-products of peanut processing, an important agricultural commodity grown in several countries, particularly in Asia and Africa. In 2020, global groundnut production in shells amounted to 54 million tonnes, representing a remarkable 8% increase from 2019 levels (Mohd Zaini et al., 2023). Groundnut shells have several applications, including use as a fuel source, incorporation as a filler in calf feed, use in the production of rigid particle board, and conversion into activated carbon, among others. Groundnut shell ash (GSA) is a by-product of the incineration of discarded peanut shells. The material consists primarily of silica, followed by lesser amounts of aluminum, iron, alkali, and alkaline earth oxides (Buari et al., 2019). Pozzolans can replace cement in several applications, such as the manufacture of cement-bonded particle boards, high-performance concrete, cement-sand blocks, and whiteware bodies.

The utilization of GSA as a cement substitute has the potential to improve the strength and durability of cementitious materials in demanding environmental conditions. It also has the potential to reduce both the financial cost and the environmental effects associated with cement production (Sathiparan et al., 2023a). The optimal amount of ground granulated GSA to replace cement depends on the specific type and quantity of other materials used in the construction. Several studies indicate that GSA replacement in the 10% to 30% range can yield good results (Sathiparan et al., 2023a). As a partial substitution for cement, GSA affects the strength properties of cement mortar or concrete. The impact of using GSA as a substitute, the proportion of aggregate-to-cement, and the curing condition influence the $f_{C\text{-}GS}$. Therefore, examining these factors and developing a methodology to predict the $f_{C\text{-}GS}$ is essential. However, no attempt has been made thus far to create a predictive model for the $f_{C\text{-}GS}$.

There has been increasing attention among engineers and researchers to using machine learning (ML) methods to predict material properties (Feng et al., 2020; Marani & Nehdi, 2020; Sathiparan & Jeyananthan, 2023a, b). The characteristics of GSA blended sandcrete

are highly responsive to mixing ratios and are influenced by several factors. Hence, ML algorithms are the most suitable choice for forecasting these properties. There is a proposal to utilize more refined approaches to reduce dependence on lab experimentation. In addition, engineers must have the necessary equations or tools to precisely predict experimental outcomes (Sathiparan et al., 2023c; Subramaniam et al., 2024). The ML method can offer alternative methodologies and solutions for linear and non-linear cases where traditional mathematical equations may fail to accurately capture the relationships among the parameters in a particular problem (Gao et al., 2019; Wijekoon et al., 2024).

The primary aim of the present study was to utilize mathematical equations and ML methods to predict the $f_{C\text{-}GS}$. A statistical evaluation was conducted to assess the model's accuracy for predicting the $f_{C\text{-}GS}$. This evaluation used four basic mathematical models and four ML methods. The mathematical models utilized in this study included linear regression, full quadratic model, non-linear regression, and multilinear regression. Moreover, the investigation uses machine learning models such as boosted decision-tree regression, artificial neural networks, support vector, and random forest regressions. These models offer a means to improve the precision of forecasting the $f_{C\text{-}GS}$.

2. Methods

The study's methodology comprised several steps, visually depicted in Figure 1 as a flowchart. The fundamental processes included gathering information and creating a database on GSA-blended sandcrete, based on existing published literature. The data collected was then randomly split into two groups for training machine learning (ML) models. The datasets were split into two groups: the training group, which contained around two-thirds of the datasets, and the testing group, which included the remaining one-third. Predictive models were constructed using mathematical (LR, FQ, NLR, and MLR) and ML (ANN, RFR, BDT, and SVR) methods. The recommended models were evaluated and constructed using several performance indicators and SHAP analysis.

2.1 Data Collection

The database was created using existing literature, as shown in Table 1. A total of 140 items were obtained from the investigation. The datasets used in this analysis were obtained from tests that followed globally accepted standards.

The compressive-strength values in this study ranged from 0.26 to 34.04 MPa. This broad range can be attributed to several factors, including variations in the aggregate-to-binder ratio, differences in GSA content, and varying curing periods. Such disparities can lead to differences in the hydration process and the sandcrete's microstructure, affecting its overall strength. For instance, lower compressive-strength values may indicate insufficient binder content or inadequate curing time, while higher values may reflect optimal mixing ratios and effective ash utilization. The data collected for ML models was separated into two sets using the RAND function. The model was constructed using the initial subset comprising 93 datasets and encompassing approximately 66% of the data. The remaining dataset, around 34% of the whole dataset, was used for validation.

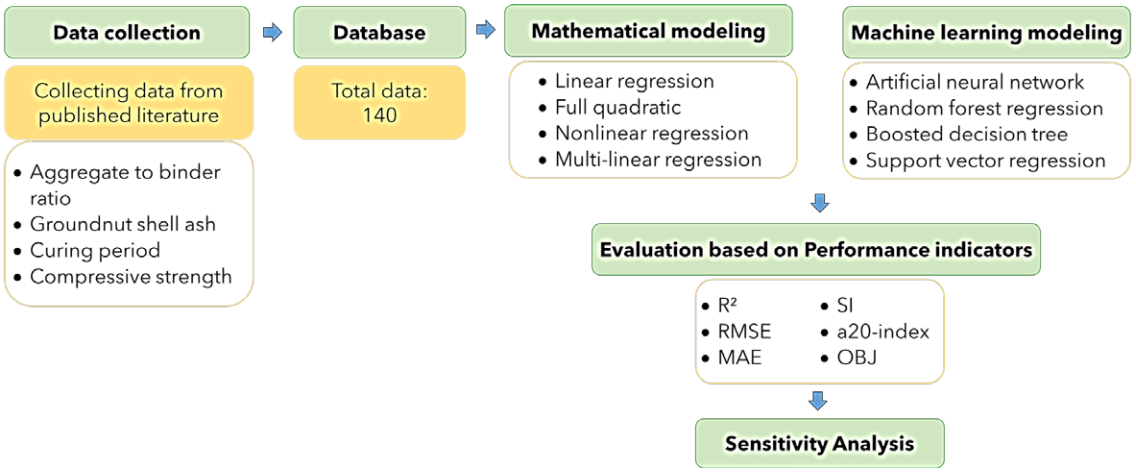

Figure 1. Flowchart for methodology.

Table 1. Summary of the dataset gathered from published literature.

Reference	A/B	GSA	Curing period	CS (MPa)	No. of Data
(Ketkukah & Ndububa, 2006)	5	0, 2, 4, 6, 8, 10	7, 14, 21, 28	1.80-3.45	24
(Mahmoud et al., 2012)	8	0, 10, 20, 30, 40, 50	7, 14, 21, 28	0.26-4.50	24
(Narayana Moorthi et al. 2015)	4	0, 15, 20, 30, 40	7	5.01-17.05	15
(Ogork & Uche, 2014)	3	0, 5, 10, 20, 30, 40	7, 28, 60, 90	11.6-35.1	24
(Fernando et al., 2018)	6	0, 5, 10, 15, 20, 25	7, 14, 28	0.40-1.77	18
(Nicholas, 2019)	3	0, 10, 20, 30, 40	7, 14, 28	1.62-34.04	20
(Sathiparan et al., 2023b)	6	0, 10, 20, 30, 40	28, 56, 90	1.82-4.32	15
Overall	**3-8**	**0-40**	**7-90**	**0.26-34.04**	**140**

A/B: aggregate-to-binder ratio; GSA: ground nutshell ash; CS: compressive strength.

2.2 Mathematical Modeling

2.2.1 Linear Regression Model (LR)

LR is a statistical technique that enables modeling the relationship between a dependent variable and independent variables. It is a method for finding the best-fitting straight line that describes the relationship between the variables, as shown in Eq. (1). Here, α0- α4 are the model parameters.

$$f_{CGS} = \alpha_0 + \alpha_2(A/B) + \alpha_2(GSA) + \alpha_3 t \quad (1)$$

2.2.2 Full Quadratic (FQ) Model

Equation (2) is a comprehensive quadratic formula that relates $f_{C\text{-}GS}$ to the first- and second-order of each independent parameter and the relationship among these independent components [8]. The model parameters are denoted β1 to β9. The present model is a complex mixture of mathematical expressions [9].

$$f_{CGS} = \beta_0 + \beta_1(A/B) + \beta_2(GSA) + \beta_3 t + \beta_4(A/B)^2 + \beta_5(GSA)^2 + \beta_6(t)^2 + \beta_7(A/B)(GSA) + \beta_8(GSA)(t) + \beta_9(t)(A/B) \quad (2)$$

2.2.3 Nonlinear Regression (NLR) Model

The NLR model is a statistical model that defines the relationship between a dependent variable and independent variables using a nonlinear function, as shown in Eq. (3). The model parameters are denoted α1-α_6. Unlike linear regression, which assumes that the parameters are linear and additive, nonlinear regression allows the estimation of models with complex and curved geometries (Al-Harthy et al., 2003). Nonlinear regression can provide greater accuracy and flexibility than linear regression.

$$f_{CGS} = \alpha_1(A/B)^{\alpha 2} + \alpha_2(GSA)^{\alpha 4} + \alpha_5(t)^{\alpha 6} \quad (3)$$

2.2.4 Multi-linear Regression (MLR) Model

MLR can be used as an alternative to the conventional multiple linear regression method. When the predictor variable consists of more than two elements, it is recommended to use an MLR model. It can also serve as a conceptualization of the concepts of predictor and independent variables. Equation (4) is the mathematical expression that captures the combined influence of multiple factors on $f_{C\text{-}GS}$. The model parameters are denoted as a, b, c, and d. However, the MLR model has the disadvantage of failing to provide predictions when the GSA content is zero.

$$f_{CGS} = a(A/B)^b(GSA)^c(t)^d \quad (4)$$

2.3 Machine Learning Modeling

2.3.1 Artificial Neural Network

An ANN is a technique that emulates the human brain's cognitive processes. The system consists of a series of interconnected units, known as artificial neurons, that process and transmit information. Artificial neural networks can learn from data and perform tasks such as classification, regression, clustering, and anomaly detection.

When constructing an artificial neural network, making informed decisions about the number of layers and nodes is essential. The layers can be divided into three categories: input, hidden, and output. The input layer is responsible for obtaining the data, while the hidden layer performs various transformations. Finally, the output

layer is accountable for producing the results. The nodes represent the computational units that perform computations within each layer. There is no generally accepted guideline for determining the optimal number of layers and nodes in an artificial neural network. In the present study, to avoid the pitfalls of overfitting, the number of layers was limited to one and the number of nodes to three, as shown in Fig. 2.

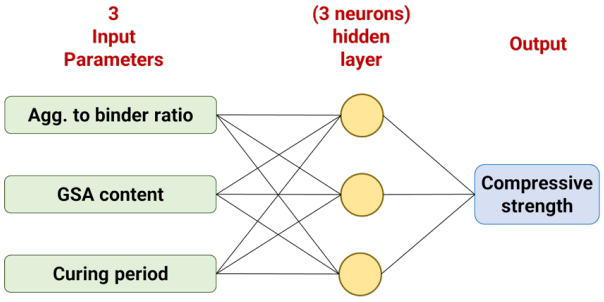

Figure 2. Outline of the ANN model.

2.3.2 Random Forest Regression

Random forest regression (RFR) is an ML method that utilizes an ensemble of decision trees to forecast a continuous output variable, as shown in Fig. 3. It has distinct characteristics compared with other ML approaches. In particular, it can effectively process high-dimensional and nonlinear data using appropriate kernel and activation functions (Albahra et al., 2023).

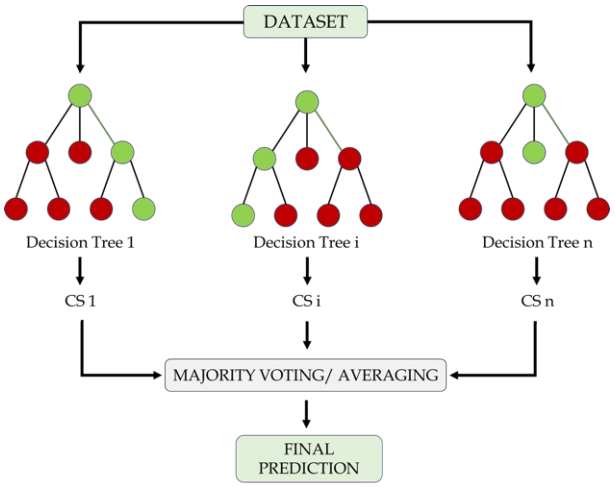

Figure 3. Outline of the RFR model.

It also enables estimating variable importance and performing feature selection by quantifying each variable's contribution to prediction accuracy. The number of trees in the forest plays a substantial role in determining the model's complexity and diversity. Increasing the number of trees in a model can reduce variability and improve prediction accuracy. However, increasing the number of trees also increases computational complexity and the likelihood of overfitting. The splitting criteria and the impurity measure are two key factors that determine how each node in the tree is split.

2.3.3 Boosted Tree Regression

Boosted tree regression (BTR) utilizes an ensemble of decision trees to predict a continuous output variable. The key difference between BTR and RFR lies in their respective approaches to tree construction. BTR builds trees sequentially, with each new tree fitted to the residuals from the previous trees. RFR, on the other hand, builds forests separately, using bootstrap sampling and random feature selection (Esteban et al., 2019).

2.3.4 Support Vector Regression

Support vector regression (SVR) uses the basic concepts of support vector machines (SVMs) to solve regression tasks. SVMs are a family of supervised learning algorithms for classification and outlier identification. They do this by identifying a hyperplane that separates the data into distinct classes or regions. SVR extends this concept by seeking to identify a hyperplane that effectively captures the data points and minimizes deviations within a specified tolerance. SVR is classified as a nonparametric method because it makes no assumptions about the underlying data distribution (Whittingham & Ashenden, 2021). However, the approach relies on kernel functions to transform the data into a higher-dimensional space, enabling the identification of a linear hyperplane. Kernel functions are mathematical functions that quantify the similarity between two data points. The choice of kernel functions is based on the specific characteristics of the data.

2.4 Performance Indicators

Several metrics defined in Eqs. (5) – (10) are utilized to evaluate the developed models. When an optimal prediction model is used, the a20 index values are expected to converge to unity. The a20 index, when fully developed, will benefit from a clear and specific technical interpretation. The proposed method quantifies the samples that meet specific requirements, namely those within 20% of the reported experimental values. Equations (5) to (10) are used to calculate each requirement.

$$R^2 = \left(\frac{\sum_{x=1}^{n} \left(P_x - \bar{P}\right)\left(E_x - \bar{E}\right)}{\sqrt{\sum_{x=1}^{n} \left(P_x - \bar{P}\right)^2} \sqrt{\sum_{x=1}^{n} \left(E_x - \bar{E}\right)^2}} \right)^2 \quad (5)$$

$$RMSE = \sqrt{\frac{\sum_{x=1}^{n} (E_x - P_x)^2}{N}} \quad (6)$$

$$MAE = \frac{\sum_{x=1}^{n} (|E_x - P_x|)}{N} \quad (7)$$

$$SI = \frac{RMSE}{\bar{E}} \quad (8)$$

$$\mathrm{a20}_{index} = \frac{N_{20}}{N} \quad (9)$$

$$OBJ = \left(\frac{n_{tr}}{N} \times \frac{RMSE_{tr} + MAE_{tr}}{R_{tr}^2 + 1}\right) + \left(\frac{n_{te}}{N} \times \frac{RMSE_{te} + MAE_{te}}{R_{te}^2 + 1}\right) \quad (10)$$

Where Px: expected value, $\bar{P}$: average of expected value, Ex: experimental value, $\bar{E}$: average of observed value, N: total number of datasets, N_{20}: total count of expected/practical value between 0.8 and 1.2, n_{tr}: number of the training dataset, n_{te}: number of the test dataset.

3. Results and Discussion

3.1 Statistical Analysis

The data were subjected to statistical analysis to determine the distribution of the dependent variable, $f_{C\text{-}GS}$, across the independent variables: aggregate-to-binder ratio, GSA content, and curing time. Figure 4 shows the correlation between $f_{C\text{-}GS}$ and the specified independent variables, along with a frequency histogram. A summary of the statistical parameters obtained is shown in Table 2.

Table 2. Summary of the statistical parameters.

Parameter	A/B	GSA	Curing period	$f_{C\text{-}GS}$
Minimum	3	0	7	0
Maximum	8	50	90	35
Mean	5	17	25	9
Standard deviation	1.8	14.4	23.6	9.3
Variance	3.165	208.8	559.6	87.91
Kurtosis	-1.06	-0.83	2.10	0.50
Skewness	0.38	0.56	1.72	1.22

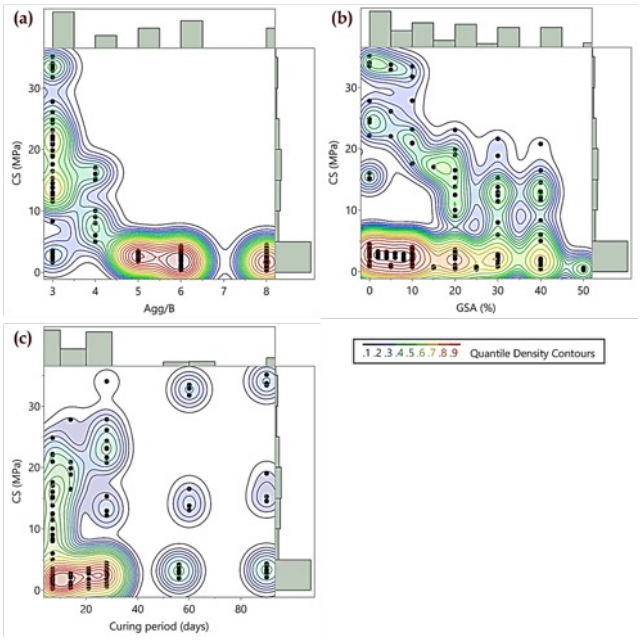

Figure 4. Marginal plots for the $f_{C\text{-}GS}$ with (a) aggregate-to-binder ratio, (b) GSA content (%), and (c) curing period (days).

In the context of kurtosis, a negative number indicates a distribution with shorter tails, while a positive value shows a distribution with longer tails. Skewness is a statistical measure that offers information about the distribution of a variable, indicating whether it is skewed to the right or left. Positive skewness indicates a rightward skew, while negative skewness suggests a leftward skew.

3.2 Prediction by Mathematical Models

Figure 5 illustrates the assessment between predicted and measured $f_{C\text{-}GS}$ values for all four mathematical models. Table 3 presents a concise summary of indicators for the different models.

3.2.1 Linear Regression Model (LR)

Figure 5(a) presents the associations among the predicted and observed $f_{C\text{-}GS}$ for the results obtained using the LR model. The LR model's prediction of $f_{C\text{-}GS}$ is given in Eq. (11). The training data yielded R2 and RMSE values of 0.583 and 6.05 MPa, respectively, whereas the test data yielded R^2 and RMSE values of 0.683 and 5.23 MPa, respectively. The training data set has a margin of error of ±20 %. This means that only 9% of the data lie between 0.80 and 1.2 for the predicted-to-observed fC-GS ratio.

$$f_{CGS} = 27.11 + 3.87(A/B) + 0.04(GSA) + 0.07(t) \quad (11)$$

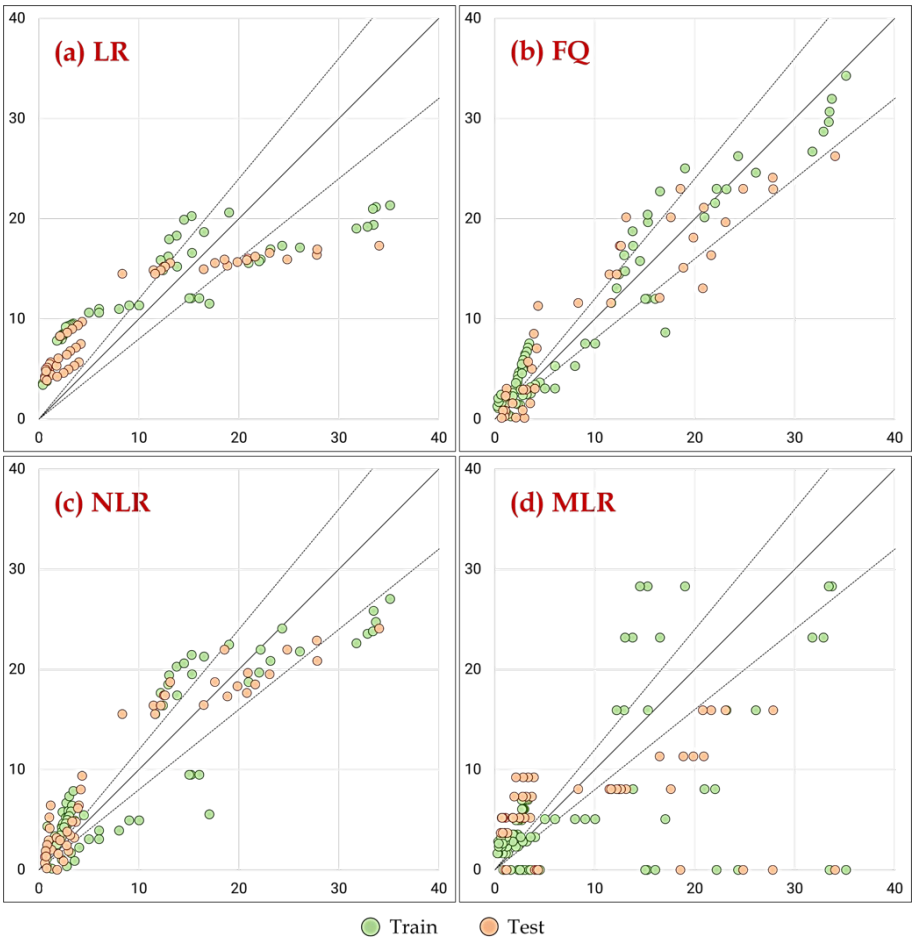

Figure 5. Predicted against measured $f_{C\text{-}GS}$ for different mathematical models.

Table 3. Performance indicators for different models used.

Model	Train					Test					OBJ	Rank
	R^2	RMSE	MAE	SI	a20	R^2	RMSE	MAE	SI	a20		
LR	0.583	6.05	5.51	0.72	0.09	0.683	5.23	4.45	0.60	0.11	6.78	7
FQ	0.907	2.85	2.25	0.34	0.27	0.862	3.45	2.78	0.40	0.32	2.90	5
NLR	0.786	4.33	3.50	0.51	0.07	0.864	3.42	2.65	0.41	0.34	4.00	6
MLR	0.206	8.35	5.25	0.99	0.14	0.005	9.26	6.45	1.42	0.02	12.74	8
ANN	0.942	2.21	1.40	0.25	0.56	0.958	1.99	1.44	0.24	0.53	1.82	2
RFR	0.934	2.35	1.52	0.27	0.52	0.941	2.35	1.64	0.28	0.53	2.02	4
BDT	0.947	2.12	1.48	0.24	0.50	0.965	1.80	1.33	0.22	0.55	1.76	1
SVR	0.941	2.22	1.49	0.26	0.41	0.938	2.41	1.57	0.29	0.47	1.96	3

3.2.2 Full Quadratic (FQ) Model

The FQ model is widely regarded as highly successful because of its intricate mathematical formulation. The model was derived using mathematical parameters, including constants, linear terms, variable product terms/interactions, and quadratic parameters. Equation (12) expresses the mathematical expression for the FQ model to estimate the $f_{C\text{-}GS}$. Figure 5(b) describes the relationship between the predicted and observed $f_{C\text{-}GS}$ for the FQ model. The training data exhibited R^2 and RMSE values of 0.907 and 2.85 MPa, respectively. Similarly, the test data demonstrated R^2 and RMSE values of 0.862 and 3.45 MPa, respectively. The training data set has a margin of error of ±20 %. This means that only 27% of the data lie between 0.80 to 1.2 for the ratio of predicted to observed $f_{C\text{-}GS}$.

$$f_{CGS} = 74.12 - 22.25(A/B) - 0.45(GSA) + 0.16(t) + 1.61(A/B)^2 - 0.00002(GSA)^2 - 0.0003(t)^2 + 0.06(A/B)(GSA) - 0.002(GSA)(t) + 0.001(t)(A/B) \quad (12)$$

3.2.3 Nonlinear Regression (NLR) Model

The results of the NLR model are presented in Eq. (13). Figure 5(c) displays the association between the predicted and measured $f_{C\text{-}GS}$. The training data set exhibits an R^2 coefficient of determination of 0.786 and an RMSE of 4.33 MPa. Furthermore, the test dataset has an R^2 value of 0.864 and an RMSE of 3.42 MPa. The results indicate that the NLR model performs worse than the FQ model but better than the LR and MLR models. The error range in the training data set is from -20 % to 20%. This suggests that around 7% of the data falls between 0.8 and 1.2 for the predicted/observed $f_{C\text{-}GS}$ ratio.

$$f_{CGS} = 1075.72(A/B)^{-3.67} - 1.03(GSA)^{0.50} + 1.29(t)^{0.40} \quad (13)$$

3.2.4 Multilinear Regression (MLR) Model

The MLR model is often considered less successful than other models, mainly because of its simple mathematical formulation. The multiple-linear regression (MLR) model formulae include constant terms and terms raised to the power of constant variables. Equation (14) comprehensively represents the variables involved and their interrelationships. Figure 5(d) shows the relationship between predicted and observed $f_{C\text{-}GS}$. The training data gave R^2 and RMSE values of 0.206 and 8.35 MPa, respectively. Similarly, the test data gave R^2 and RMSE values of 0.005 and 9.26 MPa, respectively. The error range in the training dataset is from -20 % to 20%. This specifies that around 80% of the data is between 0.80 and 1.20 for the ratio of expected to observed.

$$f_{CGS} = 18.26(A/B)^{-1.62}(GSA)^{1.04E-11}(t)^{0.49} \quad (14)$$

3.3 Prediction by Machine Learning Models

Figure 6 compares projected and measured $f_{C\text{-}GS}$ values for the four machine-learning models. All four ML models outperform mathematical models in performance indicators. The BDT model outperforms the other ML models in training and test datasets with higher R^2 and lower RMSE, MAE, and SI. However, the predicted observed $f_{C\text{-}GS}$ ratio for the BDT model, 50% of the training data set, is between 0.8 and 1.2, less than the ANN and RFR models. However, the BDT model outperforms other models as 55% of the data in the testing data set falls within the range of 0.8 and 1.2. The accuracy of the BDT and ANN models in forecasting the $f_{C\text{-}GS}$ is relatively high and classified as 1 and 2, respectively. It is followed by SVR and RFR models.

3.4 Performance of the Models

Figure 7 shows the differences between the predicted values of the examined mathematical and ML models. The error is calculated by comparing the expected and observed $f_{C\text{-}GS}$ values. ML models have a narrower range of error distributions than mathematical models.

In the mathematical models, the FQ model shows a narrow error distribution of 15.9 MPa, whereas the MLR model has the widest error distribution of 47.5 MPa. In the ML models, the BDT model shows the narrowest error distribution at 11.3 MPa, followed by the RFR, ANN, and SVR models at 12.4, 14.9, and 15.4 MPa, respectively. For all mathematical and ML models, the average error is negative, indicating that the models underpredict the fC-GS in most cases. Furthermore, all models exhibit negative skewness. The MLR exhibits the most pronounced skewness, with a score of -1.47, while the BDT model shows a slightly lower skewness of -0.34. Using various statistical and graphical methodologies can significantly improve the assessment of prediction models. Employing diverse indicators and illustrations to assess the efficiency of prediction models can provide a more comprehensive analysis.

Figure 8 shows the Taylor diagram, a statistical tool used to assess the accuracy of machine-learning models. It allows easy comparison of models based on their correlation, standard deviation, and RMSE values. The diagram promotes objectivity and clarity in model evaluation and illustrates the relative performance of each model in predicting $f_{C\text{-}GS}$, compared to a reference model (Taylor, 2001). The closeness of the pentagram to the reference point signifies the model's accuracy. The BDT

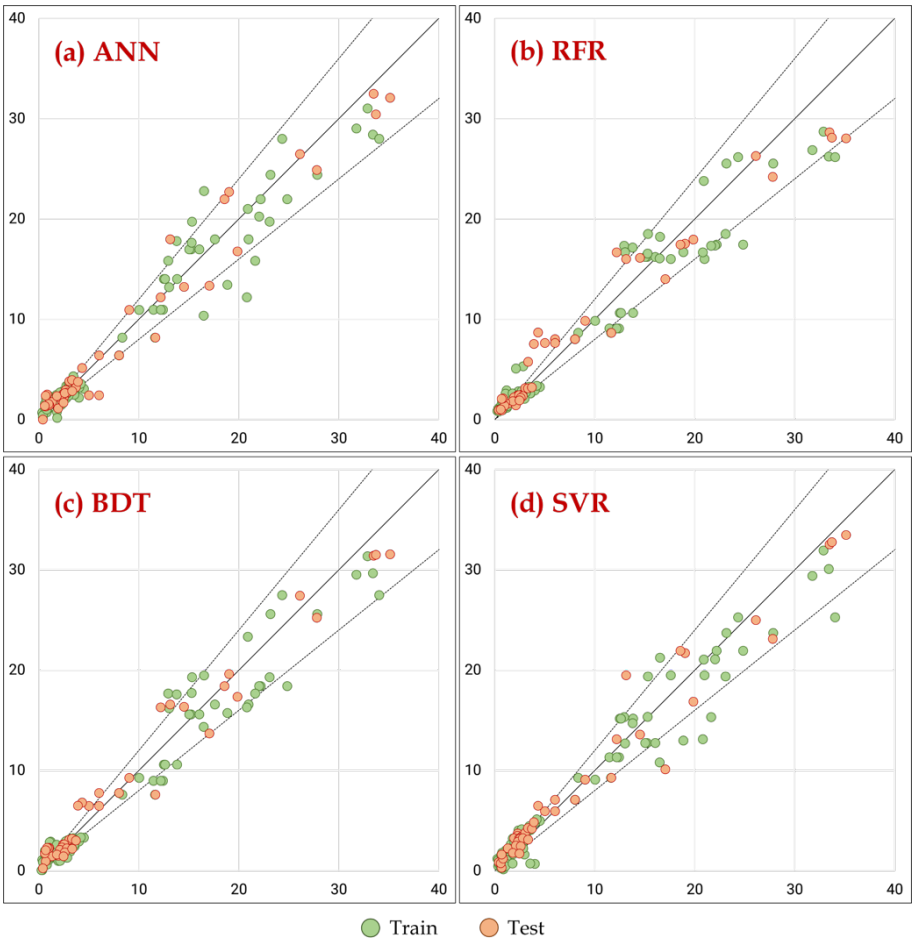

Figure 6. Predicted vs. measured $f_{C\text{-}GS}$ using different machine-learning models.

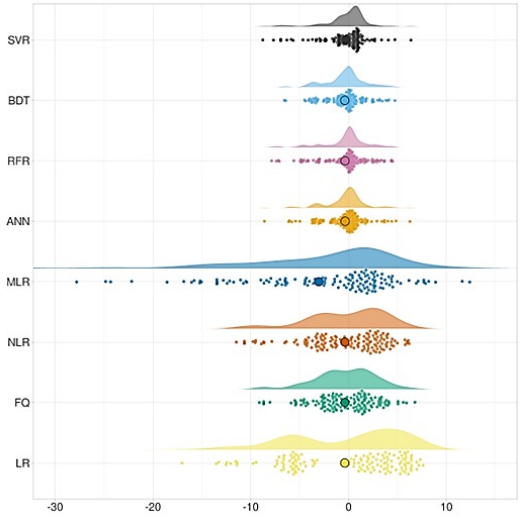

Figure 7. Error distribution in predicted $f_{C\text{-}GS}$ for different mathematical and ML models.

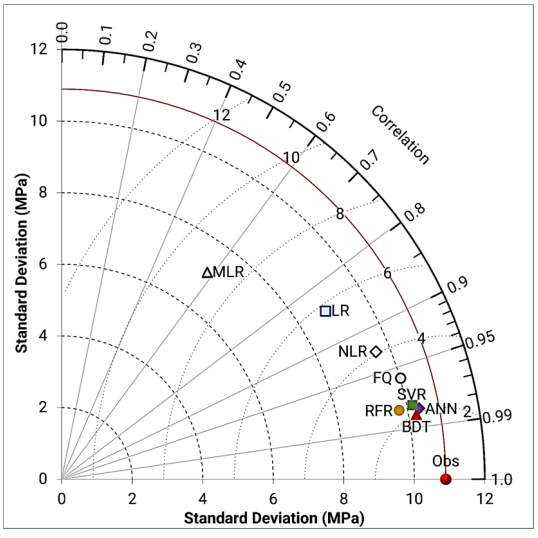

Figure 8. Taylor diagram of mathematical and ML models (The red point signifies the reference for measured).

model demonstrates the highest accuracy, while the MLR models exhibit the lowest accuracy. Based on these measures, the mathematical and ML models can be classified in descending order as follows.: BDT > ANN > SVR > RFR > FQ > NLR > LR > MLR. The results indicate a robust association with the previously set values of the indicator.

3.5 Sensitivity Analysis

Utilizing SHAP (Shapley Additive exPlanations) analysis is immensely advantageous for assessing complicated ML models that encompass various factors. (Tran et al., 2022; Zhang et al., 2022). We chose to use the results of the BDT model, which showed exceptional accuracy in predicting $f_{C\text{-}GS}$, to gain a deeper understanding of the results through SHAP analysis.

Figure 9 displays the mean SHAP values for input variables relative to $f_{C\text{-}GS}$ predictions. These forecasts result from the BDT model. The results reveal that the aggregate-to-binder ratio has the highest SHAP value, indicating its significant impact on predicting $f_{C\text{-}GS}$.

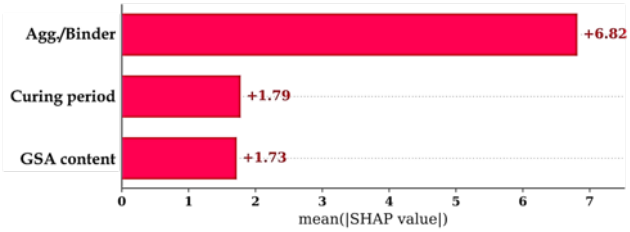

Figure 9. Mean SHAP values.

The SHAP summary graphs in Fig. 10 depict the predictions of $f_{C\text{-}GS}$ generated using the BDT model. The color represents the spectrum of feature values, while the horizontal axis represents the SHAP value, or the feature's influence, on the expected $f_{C\text{-}GS}$. The presence of a red dot indicates a notably increased feature value, corresponding to a high SHAP score. The study reveals a significant finding: a high SHAP value of 14, suggesting that the range of aggregate-to-binder ratio investigated can increase $f_{C\text{-}GS}$ by 14 MPa above the average value. Conversely, a SHAP value of -10 on the extreme left (negative) indicates that reducing the aggregate-to-binder ratio could decrease $f_{C\text{-}GS}$ by 10 MPa below the average value. The findings of the SHAP analysis suggest that employing a game-theory-based methodology to compute SHAP values may enhance the understanding of the proposed hybrid machine-learning models. In addition, the data indicate that the models have reasonable and satisfactory predicted accuracies.

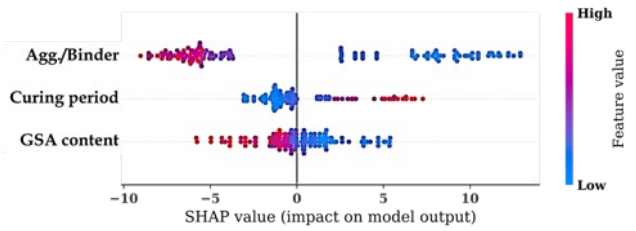

Figure 10. SHAP summary plot.

4. Conclusions

To obtain a precise and reliable model for forecasting the $f_{C\text{-}GS}$, 140 data samples for GSA blended sandcrete with varying aggregate-to-binder ratios, GSA content, and curing periods were collected from published literature. Based on the analysis of the collected data and the outcomes obtained from various mathematical and machine-learning methodologies, the following conclusions can be drawn:

- According to many evaluation parameters, including R^2, RMSE, and SI, the BDT model demonstrated superior accuracy and performance in predicting the $f_{C\text{-}GS}$. The model achieved the highest R^2 values of 0.947 and 0.965 on the training and test datasets, respectively. The training phase yielded the lowest root-mean-square error (RMSE) and structural index (SI) values of 2.12 MPa and 0.24, respectively. During the testing phase, the lowest RMSE and SI were observed at 1.80 MPa and 0.22, respectively.
- Considering the MAE and a20 values, the ANN model achieved the highest rank for training, while the BDT model achieved the highest rank for testing. The MAE and a20 values were 1.40 MPa and 0.56 for the ANN model's training dataset, and 1.33 MPa and 0.55 for the BDT model's testing dataset, respectively. The lowest OBJ function value of 1.76 was observed for the BDT model.
- The sensitivity analysis demonstrates that the aggregate-to-binder ratio is the most significant parameter in determining the $f_{C\text{-}GS}$.
- Utilizing numerous models facilitates the process of validating and cross-verifying outcomes. By comparing the predictions and outcomes of various models, professionals in the construction field can evaluate their precision and reliability across different situations. This practice enhances the trustworthiness of the findings.
- This study demonstrates that machine-learning approaches yield superior predictive performance for

relevant material properties, with the Boosted decision-tree (BDT) model emerging as the most effective. Conversely, when employing traditional mathematical modeling techniques, the full quadratic (FQ) model stands out as the best option.

This study rigorously assesses the $f_{C\text{-}GS}$ prediction, thereby improving the current understanding and real-world applications in this field. It is important to remember that increasing the data used to train the machine-learning model can improve performance. Therefore, upholding an extensive data compilation is crucial. Employing accurate predictive modeling can help researchers and designers select the appropriate mix parameters for constructing sustainable sandcrete with desired characteristics.

Abbreviations

A/B	Aggregate-to-binder ratio
ANN	Artificial neural network
BTR	Boosted tree regression
CO_2	Carbon dioxide
CS	Compressive strength
$f_{C\text{-}GS}$	Compressive strength of GSA blended sandcrete
FQ	Full quadratic
GSA	Groundnut shell ash
LR	Linear regression
MAE	Mean absolute error
ML	Machine learning
MLR	Multi-linear regression
NLR	Non-linear regression
OBJ	Objective function value
RFR	Random forest regression
RMSE	Root mean squared error
R^2	Coefficient of determination
SCM	Supplementary cementitious material
SHAP	SHapley Additive exPlanations
SI	Scatter index
SVR	Support vector regression
t	Curing period

Conflict of interest

The authors have no conflict of interest to declare

Funding

The authors received no specific funding for this work

References

Adegbemileke, S. A., Osuji, S. O., & Ogirigbo, O. R. (2024). An assessment of the pozzolanic potential and mechanical properties of Nigerian calcined clays for sustainable ternary cement blends. *Sustainable Structures*, *4*(3), 1-15.

Al-Harthy, A. S., Taha, R., & Al-Maamary, F. (2003). Effect of cement kiln dust (CKD) on mortar and concrete mixtures. *Construction and Building Materials*, *17*(5), 353-360. https://doi.org/10.1016/S0950-0618(02)00120-4

Albahra, S., Gorbett, T., Robertson, S., D'Aleo, G., Kumar, S. V. S., Ockunzzi, S., ... & Rashidi, H. H. (2023, March). Artificial intelligence and machine learning overview in pathology & laboratory medicine: A general review of data preprocessing and basic supervised concepts. In *Seminars in Diagnostic Pathology* (Vol. 40, No. 2, pp. 71-87). WB Saunders. https://doi.org/10.1053/j.semdp.2023.02.002

Blesson, S., & Rao, A. U. (2023). Agro-industrial-based wastes as supplementary cementitious or alkali-activated binder material: a comprehensive review. *Innovative Infrastructure Solutions*, *8*(4), 125. https://doi.org/10.1007/s41062-023-01096-8

F, T. A., Olutoge, F. A., Ayinnuola, G. M., Okeyinka, O. M., & Adeleke, J. S. (2019). Short term durability study of groundnut shell ash blended self consolidating high performance concrete in sulphate and acid environments. *Asian Journal of Civil Engineering*, *20*(5), 649-658. https://doi.org/10.1007/s42107-019-00131-3

Esteban, J., McRoberts, R. E., Fernández-Landa, A., Tomé, J. L., & Næsset, E. (2019). Estimating forest volume and biomass and their changes using random forests and remotely sensed data. *Remote Sensing*, *11*(16), 1944. https://doi.org/10.3390/rs11161944

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. *Construction and Building Materials*, *230*, 117000. https://doi.org/10.1016/j.conbuildmat.2019.117000

Fernando, P. R., Hatangala, H. A. Y. N., Karunagaran, S., & Dissanayake, D. M. J. C. (2018). Evaluates some engineering properties of innovative sustainable cement blocks as a partial replacement of groundnut shell ash (GSA). *Acta Scientific Agriculture (ISSN: 2581-365X)*, *2*(7).

Gao, W., Karbasi, M., Derakhsh, A. M., & Jalili, A. (2019). Development of a novel soft-computing framework for the

simulation aims: a case study. *Engineering with Computers*, *35*(1), 315-322.
https://doi.org/10.1007/s00366-018-0601-y

Hafez, R. A., Ftah, R. O. A. A., & Abdelsamie, K. (2024). The influence of nucleus dates waste and ceramic wastes in sustainable concrete. *Sustainable Structures*, *4*(2).

Juenger, M. C., Snellings, R., & Bernal, S. A. (2019). Supplementary cementitious materials: New sources, characterization, and performance insights. *Cement and Concrete Research*, *122*, 257-273.
https://doi.org/10.1016/j.cemconres.2019.05.008

Ketkukah, T. S., & Ndububa, E. E. (2006). Ground nut husk ash (GHA) as a partial replacement of cement in mortar. *Nigerian Journal of technology*, *25*(2), 84-90.

Kumar, R., Goyal, S., & Srivastava, A. (2021). A comprehensive study on the influence of supplementary cementitious materials on physico-mechanical, microstructural and durability properties of low carbon cement composites. *Powder Technology*, *394*, 645-668.
https://doi.org/10.1016/j.powtec.2021.08.081

Narayana Moorthi, V., Muthu Mariappan, P., & Kuppusamy, K. A. (2015). A study on Strength of Cement Mortar with Partial Replacement of Groundnut Shell Ash. Singaporean Journal of Scientific Research, 7(1), 380-384

Mahmoud, H., Belel, Z. A., & Nwakaire, C. (2012). Groundnut shell ash as a partial replacement of cement in sandcrete blocks production. *International Journal of Development and sustainability*, *1*(3), 1026-1032.

Marani, A., & Nehdi, M. L. (2020). Machine learning prediction of compressive strength for phase change materials integrated cementitious composites. *Construction and Building Materials*, *265*, 120286.
https://doi.org/10.1016/j.conbuildmat.2020.120286

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. *Journal of Building Engineering*, *13*, 137-145.
https://doi.org/10.1016/j.jobe.2017.07.011

McCarthy, M. J., & Dyer, T. D. (2019). Pozzolanas and pozzolanic materials. *Lea's Chemistry of Cement and Concrete*, *5*, 363-467.

Millward-Hopkins, J., Zwirner, O., Purnell, P., Velis, C. A., Iacovidou, E., & Brown, A. (2018). Resource recovery and low carbon transitions: The hidden impacts of substituting cement with imported 'waste'materials from coal and steel production. *Global Environmental Change*, *53*, 146-156.
https://doi.org/10.1016/j.gloenvcha.2018.09.003

Mohd Zaini, N. A., Azizan, N. A. Z., Abd Rahim, M. H., Jamaludin, A. A., Raposo, A., Raseetha, S., ... & Wan-Mohtar, W. A. A. Q. I. (2023). A narrative action on the battle against hunger using mushroom, peanut, and soybean-based wastes. *Frontiers in Public Health*, *11*, 1175509.
https://doi.org/10.3389/fpubh.2023.1175509

Nicholas, A. (2019) An assessment of the utilization of ground nut shell ash and fly ash as a partial replacement of cement in plaster. B.Sc.Thesis, Kyambogo University.

Nilimaa, J. (2023). Smart materials and technologies for sustainable concrete construction. *Developments in the Built Environment*, *15*, 100177.
https://doi.org/10.1016/j.dibe.2023.100177

Ogork, E.N., & Uche, O.K. (2014) Effects of groundnut husk ash (GHA) in cement paste and mortar. *International Journal of Civil Engineering and Technology*, 5(10), 88-95.

Ortega, F., Versino, F., López, O. V., & García, M. A. (2022). Biobased composites from agro-industrial wastes and by-products. *Emergent Materials*, *5*(3), 873-921.
https://doi.org/10.1007/s42247-021-00319-x

Poorveekan, K., Ath, K. M. S., 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. *Construction and Building Materials*, *277*, 122371.
https://doi.org/10.1016/j.conbuildmat.2021.122371

Sathiparan, N., Anburuvel, A., & Selvam, V. V. (2023a). Utilization of agro-waste groundnut shell and its derivatives in sustainable construction and building materials–A review. *Journal of Building Engineering*, *66*, 105866.
https://doi.org/10.1016/j.jobe.2023.105866

Sathiparan, N., Anburuvel, A., Selvam, V.V. and Vithurshan, P.A., (2023b), Potential use of groundnut shell ash in sustainable stabilized earth blocks. Construction and Building Materials 393, 132058.
https://doi.org/10.1016/j.conbuildmat.2023.132058

Sathiparan, N., Jaasim, J. H. M., & Banujan, B. (2022). Sustainable production of cement masonry blocks with the combined use of fly ash and quarry waste. *Materialia*, *26*, 101621.
https://doi.org/10.1016/j.mtla.2022.101621

Sathiparan, N. and Jeyananthan, P., (2023a), Predicting compressive strength of cement-stabilized earth blocks using machine learning models incorporating cement content, ultrasonic pulse velocity, and electrical resistivity. Nondestructive Testing and Evaluation, 1-25.
https://doi.org/10.1080/10589759.2023.2240940

Sathiparan, N., and Jeyananthan, P., (2023b), Prediction of masonry prism strength using machine learning technique: Effect of dimension and strength parameters. Materials Today Communications 35, 106282.
https://doi.org/10.1016/j.mtcomm.2023.106282

Sathiparan, N., Jeyananthan, P., and Subramaniam, D.N., (2023c), 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 Journal of Civil Engineering 25, 495-509.
https://doi.org/10.1007/s42107-023-00790-3

Seevaratnam, V., Uthayakumar, D., & Sathiparan, N. (2020). Influence of rice husk ash on characteristics of earth cement blocks. *MRS Advances*, *5*(54-55), 2793-2805.
https://doi.org/10.1557/adv.2020.294

Subramaniam, D. N., Jeyananthan, P., & Sathiparan, N. (2024). Soft computing techniques to predict the electrical resistivity of pervious concrete. *Asian Journal of Civil Engineering*, *25*(1), 711-722.
https://doi.org/10.1007/s42107-023-00806-y

Tan, Y. Y., Awang, H., & Kaus, N. M. (2024). Integration of fly ash and ground granulated blast furnace slag into palm oil fuel ash based geopolymer concrete: A review. *Sustain. Struct*, *4*(2), 000050.
https://doi.org/10.54113/j.sust.2024.000050

Taylor, K. E. (2001). Summarizing multiple aspects of model performance in a single diagram. *Journal of geophysical research: atmospheres*, *106*(D7), 7183-7192.
https://doi.org/10.1029/2000JD900719

Tran, V. Q., Dang, V. Q., & Ho, L. S. (2022). Evaluating compressive strength of concrete made with recycled concrete aggregates using machine learning approach. *Construction and Building Materials*, *323*, 126578.
https://doi.org/10.1016/j.conbuildmat.2022.126578

Whittingham, H., & Ashenden, S. K. (2021). Hit discovery. In *The era of artificial intelligence, machine learning, and data science in the pharmaceutical industry* (pp. 81-102). Academic Press.
https://doi.org/10.1016/B978-0-12-820045-2.00006-4

Wijekoon, S. H., Shajeefpiranath, T., Subramaniam, D. N., & Sathiparan, N. (2024). 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 journal of civil engineering*, *25*(1), 67-79.
https://doi.org/10.1007/s42107-023-00757-4

Zhang, J., Niu, W., Yang, Y., Hou, D., & Dong, B. (2022). Machine learning prediction models for compressive strength of calcined sludge-cement composites. *Construction and Building Materials*, *346*, 128442.
https://doi.org/10.1016/j.conbuildmat.2022.128442