Sustainable Structures
ISSN: 2789-3111 (Print); ISSN: 2789-312X (Online)
http://www.sustain-dpl.com/journal/sust
DOI: 10.54113/j.sust.2026.000110

ORIGINAL ARTICLE

Innovative use of rice husk ash in sustainable cement mortar: insights from machine learning on oxide influence

Navaratnarajah Sathiparan[a,*]

[a] Department of Civil Engineering, Faculty of Engineering, University of Jaffna, Jaffna 40000, Sri Lanka
[*]Corresponding Author: Sathiparan. Email: sakthi@eng.jfn.ac.lk

Abstract: This study explores the use of rice husk ash (RHA) as a supplementary material in cement mortar, focusing on its impact on compressive strength through machine learning predictive modelling. The methodology involved a systematic literature review to compile a comprehensive dataset of 692 records from 20 published sources. A range of machine learning techniques, comprising Linear Regression, Artificial Neural Networks, Random Forest Regression, and Extreme Gradient Boosting, were utilized to assessment compressive strength based on important variables such as the ratios of aggregate-to-binder, RHA-to-binder, water-to-binder, along with the curing time and chemical composition. Results designate that RHA significantly contributes to the pozzolanic activity of cement mortar, with optimal RHA-to-binder ratios enhancing compressive strength without compromising workability. The machine learning models exhibited high predictive precision, with R^2 values more than 0.95 and low RMSE values across the tested datasets. Sensitivity analysis has shown that the aggregate-to-binder ratio, SiO_2 content and curing period were the important parameters affecting compressive strength, emphasizing the significance of precise formulation. The study recommends further research to validate the predictive models through experimental studies, broaden the dataset for enhanced generalizability, and explore additional chemical compositions of RHA.

Keywords: cement mortar, machine learning, oxide, compressive strength

1 Introduction

Cement mortar is a blend of cement, sand, and water, plays a important role in construction, serving as a binding agent for materials such as bricks and stones. The composition typically includes cement as the primary binding agent, sand to provide bulk and fill voids, and water to activate the cement and facilitate hardening [1]. The purposes of cement mortar extend beyond mere binding; it is essential for creating smooth surface finishes through plastering, providing structural support in load-bearing applications, and enhancing aesthetic appeal. Additionally, specific formulations can offer waterproofing benefits, preventing water infiltration into structures [2]. The need for higher cement content in mortar often stems from the desire to enhance strength, durability, and workability. Increased cement can improve compressive strength and resistance to environmental factors, making the mix more straightforward to apply [3]. However, an imbalance in the mix, with excessive cement, can lead to issues such as cracking and shrinkage, underscoring the importance of precise formulation [4].

Despite its advantages, using cement in mortar raises significant environmental concerns. Burning fossil fuels for energy within the cement manufacturing process discharges substantial amounts of CO_2

Received: 24 December 2025; Received in revised form: 4 July 2026; Accepted: 8 July 2026

into the atmosphere [5, 6]. Additionally, extracting raw materials for cement can lead to environment devastation and biodiversity loss, although the manufacturing process demands considerable water resources, potentially straining local supplies [7]. The waste materials generation during cement production further contributes to environmental degradation if not managed through recycling or other sustainable practices [8]. Therefore, while cement mortar is indispensable in modern construction, addressing its environmental impact is crucial for promoting sustainable building practices [9].

Researchers and engineers are exploring alternatives to traditional cement to address these issues. Options such as supplementary binders like fly ash, and natural pozzolans can enhance the sustainability of cement mortar although reducing its carbon footprint [10, 11]. Rice husk ash (RHA) is one such material that can be used as a cement replacement in mortar. The RHA is an agricultural byproduct produced from the burning of rice husks, which are the outer covers of rice grains. As a pozzolanic material, RHA has got attentiveness in the construction industry for its possible to replace traditional cement in concrete mixtures [12]. This sustainable approach addresses the environmental concerns related with cement manufacture and improves the performance characteristics of cement-based materials. Utilizing RHA in cement mortar promotes waste reduction and encourages the recycling of agricultural byproducts. This practice helps to lessening the carbon footprint related with cement manufacture [13]. While the benefits of using RHA as a cement substitution are significant, there are also challenges to consider. The inconsistency in the RHA’s properties, varying on the source and burning conditions, can affect its performance in cement mortar. Additionally, achieving the optimal replacement ratio requires careful experimentation to balance workability, strength, and durability [14].

Table 1. Summary of study on compressive strength prediction of RHA blended cementitious materials

Ref.	Materials	Input variables	ML model used	No of dataset
[15]	Cement mortar	Agg/B, RHA/B, W/B, age	LR, ANN, KNN, SVR, XGB	795
[16]	Concrete	cement, RHA, F.Agg, C.Agg, water, age, SP	DT, RFR, SVR	1212
[17]	Concrete	cement, RHA, Agg, water, age, SP	GEP, RFR	192
[18]	Concrete	cement, RHA, Agg, water, age, SP	RFR, XGB	192
[19]	Concrete	cement, RHA, F.Agg, C.Agg, water, age, SP	SVR	291
[20]	Concrete	RHA, F.Agg, C.Agg, W/C, SP	RFR, LightGBM, XGB, Ridge Regression	348
[21]	Concrete	cement, RHA, Agg, water, SP	DT, AdaBoost	192
[22]	Concrete	cement, RHA, Agg, water, age, SP	ANN, GEP	192
[23]	Concrete	cement, RHA, F.Agg, C.Agg, water, age, SP	MEP	192
[24]	Concrete	cement, RHA, F.Agg, C.Agg, water, age, SP	MLR, TB, RFR, BT, SVR, ANN, GPR	909
[25]	Concrete	cement, RHA, Agg, water, age, SP	XGB, LiightGBM, RFR	1404
[26]	Concrete	cement, RHA, F.Agg, C.Agg, water, age, SP	CatBoost, GBM, CNN, GRU	1212

Agg: Aggregate, F.Agg: fine aggregate, C.Agg: coarse aggregate, Agg/B: aggregate-to-binder ratio, RHA/B: RHA to binder ration, W/B: water to binder ratio, W/C: water to cement ratio, SP: superplasticizer, DT: Decision Tree, RFR: Random Forest Regression, GEP: Gene Expression Programming, LR: linear Regression, SVR: Support Vector Regression, MEP: Multi-Expression Programming, MLR: Multiple Linear Regression, TB: TreeBagger, GBR: Gaussian process regression, GBM: Gradient Boosting Machines, GRU: Gated Recurrent Unit, CNN: Convolutional Neural Network

For several reasons, forecasting the strength of RHA mixed cement mortar is crucial. First and foremost, compressive strength is a critical indicator of the material's load-bearing capacity, which directly impacts its structural integrity in various applications, including masonry blocks [27]. Accurate predictions allow engineers and architects to improve mixed designs, balancing the beneficial properties of RHA with the need for sufficient strength, thereby minimizing material costs without compromising performance [20]. Moreover, as the construction industry increasingly embraces sustainability, understanding how RHA influences compressive strength becomes vital for reducing reliance on traditional cement and this approach effectively reduces the environmental effect of cement mortar [28].

Several findings have explored the forecast of compressive strength of cementitious materials with RHA, focusing on various aspects like mix design, curing period, and effects of RHA content, as brief in **Table 1**. Different methods, such as machine learning models, have been utilized to estimate strength by analyzing various mix parameters. These studies demonstrate that advanced modelling techniques can successfully detent the complex interactions among material properties and the resulting cement mortar strength. However, a critical oversight in the published literature is the absence of consideration for the RHA’s chemical composition, specifically the oxides present in the ash. These oxides show an important part in influencing the pozzolanic activity and performance of RHA in cement-based materials. The current models may not fully capture the complexities influencing compressive strength by neglecting this aspect.

The authors note a growing trend in the usage of ML models, like as ANN, random forest and XGB, for forecasting the strength of RHA-blended cement mortar. These models are increasingly favored for their high accuracy and robustness. Furthermore, the frequent inclusion of mix design factors like the aggregate-to-binder (Agg/B) ratio and water-to-binder (W/B) ratio underscores their important influence on the strength of the mortar. A key gap identified in the literature is the limited focus on the chemical composition of RHA, specifically oxides like SiO_2 and CaO, which influence the pozzolanic activity of RHA. Many studies have concentrated mainly on mix design parameters while overlooking the role of RHA's chemical properties. Another gap identified is the lack of comprehensive datasets incorporating varying environmental conditions, which limits the generalizability and applicability of existing predictive models. By integrating chemical composition (oxide analysis) with ML, the model can detect the more complex, non-linear interactions among the RHA’s chemical properties and the resulting mortar strength. This allows for a more accurate forecast of compressive strength, considering material properties and mix design parameters. The integration of oxide analysis addresses a gap in the existing literature, where the chemical composition of RHA has not been fully incorporated into predictive models. Addressing this gap could improve the precision of strength predictions and indication to more efficient utilization of RHA in sustainable construction practices. Therefore, the present study integrates the analysis of RHA's oxide composition alongside traditional mix design parameters to develop a more comprehensive understanding of its effects on compressive strength by machine learning (ML) models.

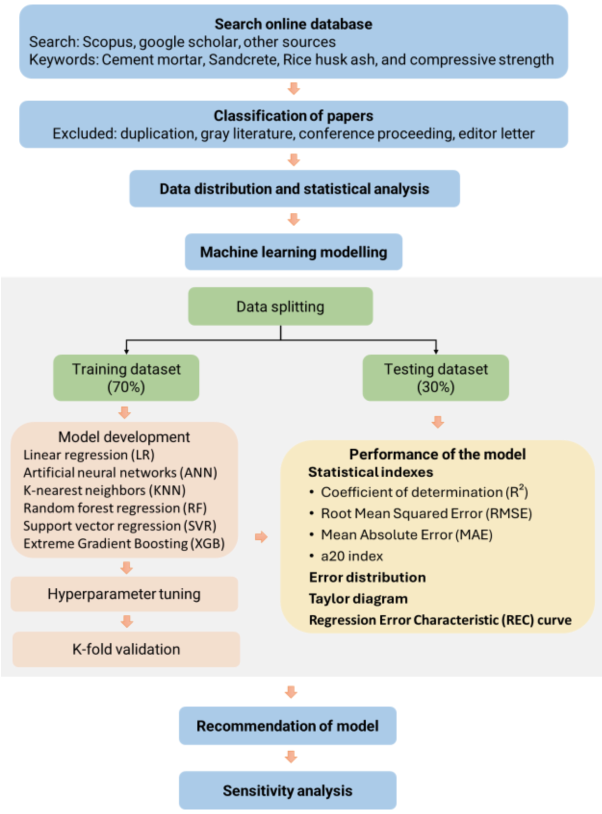


Fig. 1. Methodology outline for prediction on compressive strength of cement mortar incorporating RHA

2 Methods

The methodology outlined in **Fig. 1** establishes a complete methodology for evaluating how well different ML models can forecast the strength of cement mortar that includes RHA. The process begins with a systematic literature review, identifying relevant studies through searches in online databases like Scopus and Google Scholar. Then, the research progresses to data distribution and statistical analysis. This step involves organizing the gathered data to prepare for ML modelling. The dataset is then divided into training (70%) and testing (30%) subsets, crucial for evaluating the model's performance and generalization ability.

The process involves building and refining a suite of ML models. This includes established algorithms like Artificial Neural Networks (ANN), K-Nearest Neighbors (KNN), and Linear Regression (LR), alongside more advanced methods such as Random Forest Regression (RF), Support Vector Regression (SVR), and Extreme Gradient Boosting (XGB). For each model, hyperparameter tuning is undertaken to maximize its performance. To guarantee robustness and reliability, K-fold cross-validation is retained, thoroughly testing each model on diverse subsets of the data.

Model effectiveness is gauged using several key statistical measures: the Coefficient of Determination (R^2), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the a20 index. Furthermore, the methodology includes a detailed examination of error distribution using tools like the Taylor diagram and Regression Error Characteristic (REC) curve. These tools offer valuable insights into the model's forecasting abilities and the nature of its errors. The final component of the methodology involves conducting a sensitivity analysis to determine how variations in input parameters affect the model's predictions.

2.1 Data collection

The data collection is done by systematic literature review through searches in online databases such as Scopus and Google Scholar. Key terms, including "cement mortar", "sandcrete", "rice husk ash," and "compressive strength," are employed to ensure a focused search. During this phase, papers that are duplicates, grey literature, conference proceedings, or editor letters are excluded to maintain the integrity of the data set. This study's data was meticulously collected from 20 published literature sources [29-48], resulting in a comprehensive dataset comprising 692 individual records. Table 2 provides an overview of the dataset collected from previously published literature.

This extensive collection of data highlights the importance of empirical study in understanding the properties of cement mortar when incorporating RHA as a supplementary material. The dataset includes key input variables known to influence concrete's compressive strength significantly. These variables are:

- Agg/B: This ratio is important as it defines the relative proportions of aggregates and binder materials in the cement mortar mix, which can affect strength and durability.
- RHA/B: This specific ratio assesses the proportion of RHA used relative to the total binder content, providing insights into the effectiveness of RHA as a partial substitute for traditional binders.
- W/B: This ratio plays a critical role in cement mortar performance, as it impacts the hydration activity and the resulting microstructure of the hardened mortar.
- Curing Period: The duration of curing is essential for achieving optimal strength development, as it directly impacts the hydration and, consequently, the properties of the cement mortar.
- Calcium Oxide (CaO): This chemical component contributes to the pozzolanic reactions within the concrete matrix, enhancing overall strength.
- Silica Oxide (SiO_2): A key ingredient in pozzolanic materials, silica oxide enhances the binding properties of cement mortar.
- Aluminium Oxide (Al_2O_3): This oxide is important in influencing cement mortar's setting time and strength development.
- Equivalent Alkali (Na_2O Eq.): This parameter is relevant for assessing the potential for alkali-silica reaction, which can adversely affect the strength and durability.

The output variable is compressive strength, a fundamental property that determines the load-bearing capability of cement mortar. The study aims to develop predictive models to inform best practices in cement mortar formulation, particularly when utilizing RHA, by analyzing the association amongst the input variables and compressive strength.

Table 2. Overview of the dataset gathered from existing research studies

Reference	Specimen size used	Agg/B	RHA/B	W/B	Curing period (days)	Compressive strength (MPa)
[29]	70.7 mm cube	3.00	0 - 0.20	0.56	7 - 90	23.8 – 43.0
[30]	50 mm cube	1.96 – 2.00	0 - 0.20	0.75 - 0.84	3 - 90	4.7 – 39.0
[31]	50 mm cube	2.75	0 - 0.20	0.48	7 - 90	31.2 - 64.4
[32]	50 mm cube	1.69 - 3.01	0 - 0.40	0.25 - 0.55	7 - 90	0.1 – 65.0
[33]	215×105×65 mm block	5.41	0 - 0.20	0.60 - 0.88	7 - 28	1.4 - 14.5
[34]	70.7 mm cube	3.00 – 5.00	0 - 0.50	0.50 – 1.00	3 - 90	2.1 - 28.7
[35]	80×40×40 mm block	2.00	0 - 0.20	0.50	7 - 28	6.4 - 29.2
[36]	450 x 225 x 150 mm block	6.00 – 8.00	0 - 0.60	0.60	7 - 90	0.5 - 4.3
[37]	450 x 150 x 150 mm block	8.00	0 - 0.50	0.50 - 0.58	1 - 28	0.1 - 4.6
[38]	50 mm cube	2.25	0 - 0.50	0.50	7 - 90	16.5 - 56.7
[39]	70.7 mm cube	3.00	0 - 0.20	0.33 - 0.39	3 - 28	18.5 - 36.7
[40]	80×40×40 mm block	3.00	0 - 0.70	0.48 - 0.70	1 - 360	8.0 - 68.2
[41]	50 mm cube	2.25	0 - 0.20	0.49 - 0.58	7 - 56	18.8 - 33.4
[42]	160×40×40 mm block	3.00	0 - 0.25	0.50	2 - 90	24.4 - 66.1
[43]	50 mm cube	2.36 - 2.75	0 - 0.40	0.48	7 - 28	1.8 - 8.7
[44]	305×203×152 mm block	8.00	0 - 0.50	0.45 - 0.55	3 - 28	0.8 - 7.0
[45]	80×40×40 mm block	3.00	0 - 0.30	0.60	3 - 56	0.6 - 57.5
[46]	160×40×40 mm block	3.00	0 - 0.10	0.50	28 - 90	37.5 - 44.4
[47]	50 mm cube	2.75	0 - 0.40	0.50	7 - 90	33.0 - 62.5
[48]	50 mm cube	2.75	0 - 0.40	0.50	7 - 90	33.0 - 63.7

2.2 Machine learning models

2.2.1 Linear Regression

Linear regression is a basic statistical and machine learning method employed to model the relationships among various variables. It assumes a linear relationship expressed in a mathematical formulation. The relationship can be expressed as:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \varepsilon \tag{1}$$

where $x_1, x_2, \ldots, x_n$: Independent variables; $\beta_1, \beta_2, \ldots, \beta_n$: Coefficients for each independent variable

The main goal of linear regression is to identify the line (or hyperplane in the case of multiple regression) that most accurately represents the observed data by reducing the total of squared differences between predicted and actual values.

2.2.2 Artificial Neural Network (ANN)

Based on the structure of the human brain, ANNs are ML models adept at recognizing complex patterns [49]. They are made up of linked layers of neurons as shown in **Fig. 2** [50]. Each neuron processes inputs through weighted connections and an activation function (Eqs. (2) and (3)).

$$z = w_1 x_1 + w_2 x_2 + ... + w_n x_n + b \tag{2}$$

$$a = f(z) \tag{3}$$

where x_i is Input features; w_i is weights; b is bias term; *f* is activation function (e.g., Linear, Gaussian, TanH)

Despite their advantages of the capability to model complex relationships and flexibility in architecture, ANNs also face challenges. They need substantial data volumes to train effectively, can be prone to overfitting, and are computationally intensive, especially for deep networks [51]. Overall, ANNs are powerful tools in ML, adept at solving a wide range of problems by learning from data.

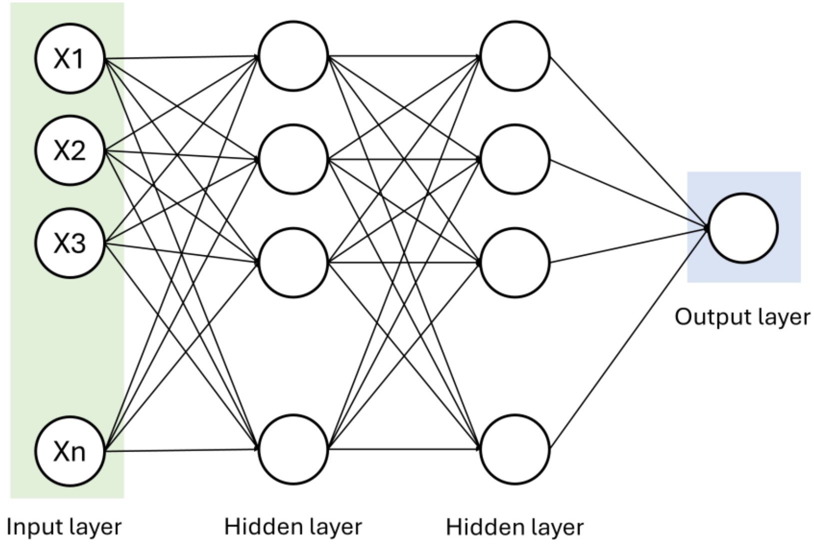


Fig. 2. The architecture of a simple ANN model

2.2.3 K-Nearest Neighbors (KNN)

KNN is a machine learning method that makes classifications or predictions by examining the closeness of data points within the feature space [52]. The KNN algorithm finds the k closest data points to a specified point in the training dataset. For classification tasks, the algorithm determines the predicted class label by selecting the most frequent class among the k closest neighbors. In regression tasks, the forecast is made by averaging the values of the k nearest neighbors, as presented in **Fig. 3**. To determine proximity, KNN typically employs distance metrics, with Euclidean and Manhattan distances being the common [53].

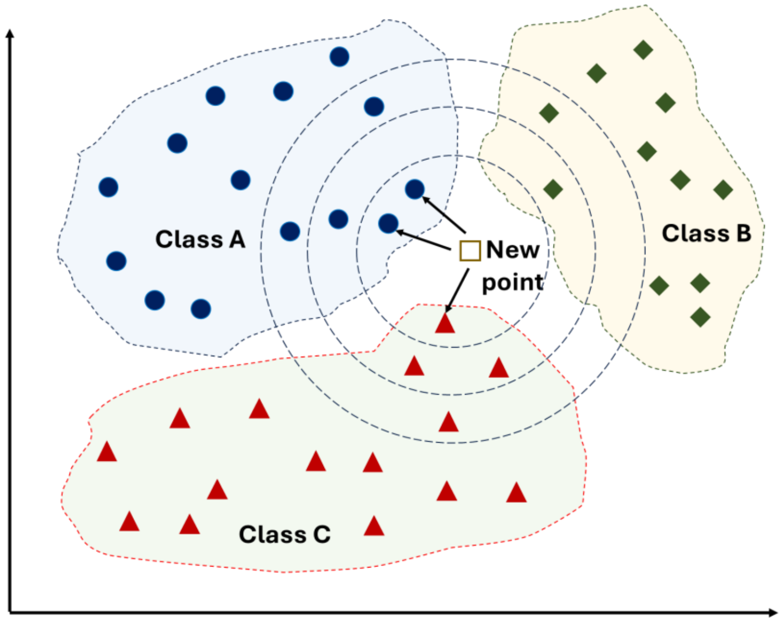


Fig. 3. The architecture of a simple K-Nearest Neighbors (KNN) model

The algorithm follows a straightforward process: first, it selects the number of neighbors k; next, for a given test data point, it calculates the distance to all training data points; then, it sorts these distances and selects the k closest points. Finally, it assigns a class label based on majority voting among the neighbors for classification or computes the average for regression tasks [54]. KNN's non-parametric nature allows it to operate without assuming any underlying data distribution, making it flexible for various datasets. However, KNN has challenges, including computational intensity for large datasets, as it requires distance calculations for all training points.

2.2.4 Random Forest Regression (RF)

RF is an ensemble method that builds multiple decision trees during the training phase [55]. By merging the forecasts from multiple trees, RF effectively minimizes overfitting and improves the model's capability to generalize [56]. The algorithm starts by generating several decision trees, each trained on a arbitrary subset of the training data, with sampling done through replacement (a method known as bootstrap aggregating or bagging). Furthermore, when splitting a node in a tree, RF selects a random subdivision of features, which presents additional variety between the trees and reduces correlation [57]. This random selection aids in producing a robust model less sensitive to noise in the data, as presented in **Fig. 4**.

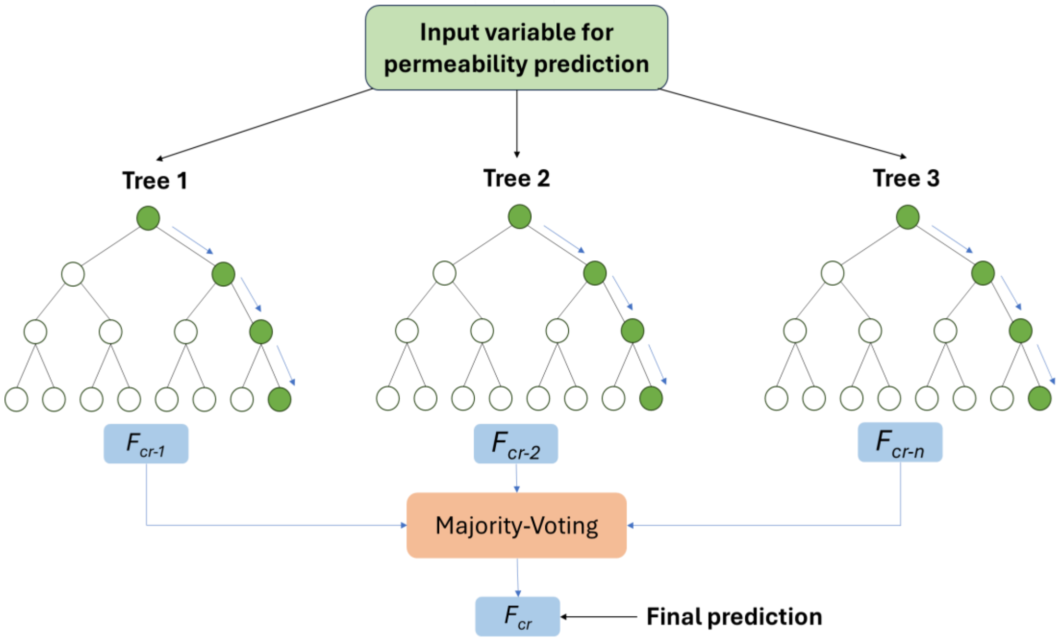


Fig. 4. The architecture of a simple RF Regression model

Random Forest has several advantages, including its high accuracy, robustness to outliers, and ability to handle large datasets with numerous features. However, the model can be computationally intensive and less interpretable compared to individual decision trees. Achieving optimal performance necessitates careful tuning of hyperparameters, including the number of trees and maximum tree depth [58].

2.2.5 Support Vector Regression (SVR)

SVR is a method based on Support Vector Machines designed for predicting continuous outcomes [59]. SVR aims to discover a function that estimates the target variable by determining a hyperplane in a high-dimensional space, focusing on maintaining a tolerance margin around this hyperplane. The primary objective is to minimize the forecast error while warranting that most data points fall within a specified margin, controlled by a factor known as epsilon (ϵ) [60].

By employing kernel functions, SVR maps input features into a higher-dimensional space, enabling the algorithm to detect complex associations within the data effectively [61]. The model optimizes by minimizing a loss function that includes the sum of distances for outliers and a regularization term to prevent overfitting, as illustrated in **Fig. 5**.

SVR effectively models non-linear relationships and demonstrates resilience to overfitting, especially in high-dimensional environments. However, it requires careful hyperparameter tuning (kernel type, regularization) and processing extensive datasets may demand significant computational resources [62].

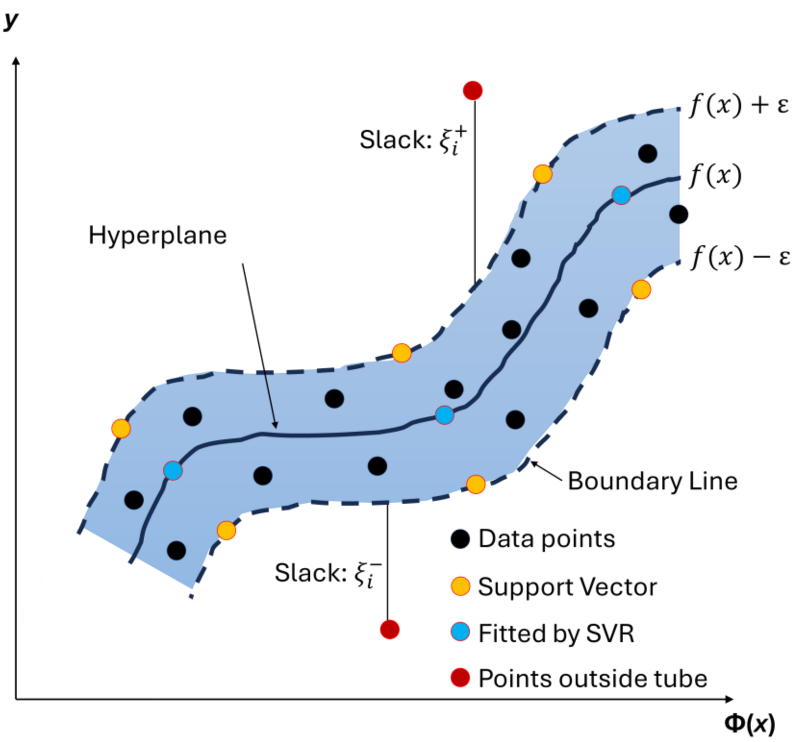


Fig. 5. The architecture of a simple SVR model

2.2.6 Extreme Gradient Boosting (XGB)

A highly effective and optimized machine learning algorithm, XGB is engineered for supervised learning. It exhibits exceptional proficiency in scenarios requiring either classification or regression outcomes. Fundamentally, it's a refined form of gradient boosting, designed to accelerate and enhance the efficacy of conventional boosting approaches [63]. XGB is recognized for its capability to handle huge datasets, manage missing values, and provide high predictive precision, making it a popular choice in data science competitions and real-world applications [64].

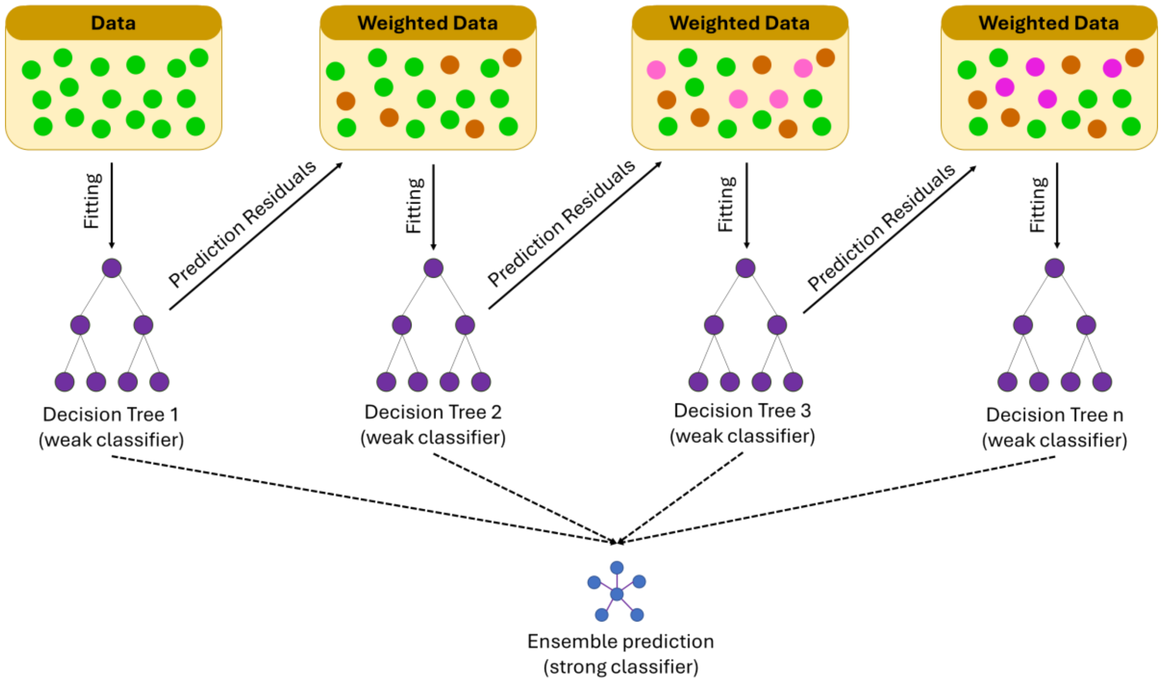


Fig. 6. The architecture of a simple XGB model

The core of XGB is its use of decision trees as base learners, which are built sequentially. Each new tree is constructed to precise the errors created by the previously trained trees, focusing on the residuals of the predictions as shown in **Fig. 6**. This sequential approach allows XGB to minimize a loss function through gradient descent, where the algorithm optimizes both the training loss and regularization terms to prevent overfitting [65]. XGB incorporates features such as tree pruning, which improves model efficiency, and a unique approach to handling missing values by learning the best direction to take when encountering missing data.

One of the key advantages of XGB is its speed and scalability, which stems from its parallel processing capabilities and optimized memory usage. Despite its strengths, XGB can be sensitive to hyperparameter tuning, requiring careful selection of parameters alike learning rate, maximum depth of trees, and the number of estimators to achieve optimal performance[66].

2.3 Data split and validation strategy

2.3.1 Data Split Strategy

To evaluate the models fairly, we divided the dataset into two distinct parts: 70% was assigned for training, allowing the models to learn from that segment, while the left-over 30% was kept for independent testing on previously unobserved data. This strategy ensures the model's capability to simplify to new data is properly assessed.

2.3.2 Cross-Validation:

To effectively minimize the chance of overfitting, we implemented K-fold cross-validation during model development. This method includes splitting the training data into K unique sets. The model is iteratively trained K times, with a diverse subset reserved for validation each time. Consolidating the results from all folds yields a stable and accurate gauge of the model's true performance.

2.3.3 Hyperparameter Tuning and Regularization:

Hyperparameter tuning was carried out to optimize the complexity of the models, reducing the risk of overfitting. For both RF and XGB, this involved selecting the optimal number of trees, depth, and other hyperparameters to balance bias and variance. Furthermore, regularization techniques such as pruning in RF and early stopping in XGB were used to prevent the models from fitting too attentively to the training data.

2.4 Performance indicators

Several statistical performance indexes are generally used to assess the precision of ML models. These metrics offer precious insights into the model's predictive precision and error measurement. The most widely used metrics are R^2, RMSE, MAE, and the a20 index.

The coefficient of determination (R^2) measures the proportion of variance in the dependent variable explained by the model. Mathematically, it is represented as Eq. (4).

$$R^2 = \left(\frac{\sum_{i=1}^{n} \left(P_i - \bar{P}\right)\left(E_i - \bar{E}\right)}{\sqrt{\sum_{i=1}^{n} \left(P_i - \bar{P}\right)^2} \sqrt{\sum_{i=1}^{n} \left(E_i - \bar{E}\right)^2}} \right)^2 \tag{4}$$

where Ei and Pi are the actual and forecast values; $\bar{E}$ and $\bar{P}$ are the mean of actual and forecast values.

The R^2 values provide insights into model fit. A value near 1 signifies that the model successfully accounts for the variability in the data, whereas a value closer to 0 indicates a weak fit.

RMSE is calculated as Eq. (5).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(E_i - P_i\right)^2}{N}} \tag{5}$$

This measure calculates the mean value of the errors, clearly indicating how much the predictions deviate from the actual values. A smaller RMSE value indicates improved model performance.

MAE quantifies the average size of the difference among predicted and actual values, as shown in Eq. (6).

$$MAE = \frac{\sum_{i=1}^{n}\left(\left|E_i - P_i\right|\right)}{N} \tag{6}$$

In contrast to RMSE, MAE gives equal weight to all errors, providing a more interpretable quantity of average prediction error. It is instrumental when comparing models or when errors are not normally distributed.

The a20 index is defined as Eq. (7).

$$a20_{index} = \frac{N_{20}}{N} \tag{7}$$

N20 is the number of points within the 20% error line and n is the total amount of data. This index evaluates the percentage of predictions that lie within a defined margin of error (e.g., 20% of the actual values). It is a precious metric for assessing the accuracy of a model, especially in purposes where relative error is more significant than absolute error.

Along with conventional statistical performance metrics, advanced indicators like the Taylor diagram, prediction error distribution, and Regression Error Characteristic (REC) curves are used to assess ML models. The Taylor diagram is a visual tool that contrasts the performance of different models against observed data. It incorporates correlation and standard deviation into a single plot, visually summarizing how well the models replicate the observed variability. In a Taylor diagram, the x-axis signifies the standard deviation of predictions, while the y-axis indicates the correlation coefficient with the observed data. The distance from the origin denotes the RMSE. Models that lie closer to the reference point demonstrate better performance [67]. This diagram is particularly useful for visualizing the trade-offs between accuracy and variability in model predictions, making it a powerful tool for model evaluation in contexts such as climate modeling and environmental science.

Prediction error distribution offers a greater empathetic of the errors made by a model. Instead of summarizing errors with a single statistic, this approach visualizes the distribution of errors across a range of predicted values. By analyzing the frequency and magnitude of errors, practitioners can identify systematic biases in their models, such as underestimation or overestimation of specific ranges. This information is critical for refining model performance and ensuring that the model meets the specific needs of its application. Additionally, visualizing the distribution of errors can help uncover patterns that may not be apparent through traditional metrics, enabling a more nuanced understanding of model reliability.

The REC curve is another valuable tool for assessing model performance, particularly regarding its robustness across different error thresholds. The REC curve graphs the percentage of predictions within a given error boundary relative to the error threshold [68]. This visualization enables the evaluation across various acceptable error thresholds. A model consistently achieving high accuracy within a specified threshold will illustrate a steep curve, indicating strong performance.

3 Results and Discussion

3.1 Data distribution and correlation analysis

The chemical composition of RHA used in the present study is illustrated in **Fig 7**. From the data provided, silicon dioxide (SiO_2) is the most abundant component, ranging from approximately 54.25% to 94.97%. This high silica content contributes to the pozzolanic properties of RHA, crafting it valuable in cement-based purposes. Calcium oxide (CaO), present in moderate quantities (1.1% to 11.31%), can enhance the binding properties of RHA when utilized in construction materials. Alumina (Al_2O_3) and is also present in moderate amounts (ranging from 0.03% to 22.8%. Iron oxide (Fe2O3) and magnesium oxide (MgO) are also present in smaller amounts (Fe_2O_3 < 3%; MgO < 1.8%). The equivalent alkali content in rice RHA is primarily determined by the presence of sodium oxide (Na_2O) and potassium oxide (K_2O). These compounds are significant because they can influence the reactivity and properties of RHA when used in construction materials. The data shows varying amounts of Na_2O and K_2O across different samples of RHA. The Na_2O content ranges are in low values (< 1.7%) and K_2O content varies significantly, with some samples having up to 7.5%. When utilizing RHA in construction, it is crucial to consider the alkali content to ensure compatibility with other materials and to meet performance specifications.

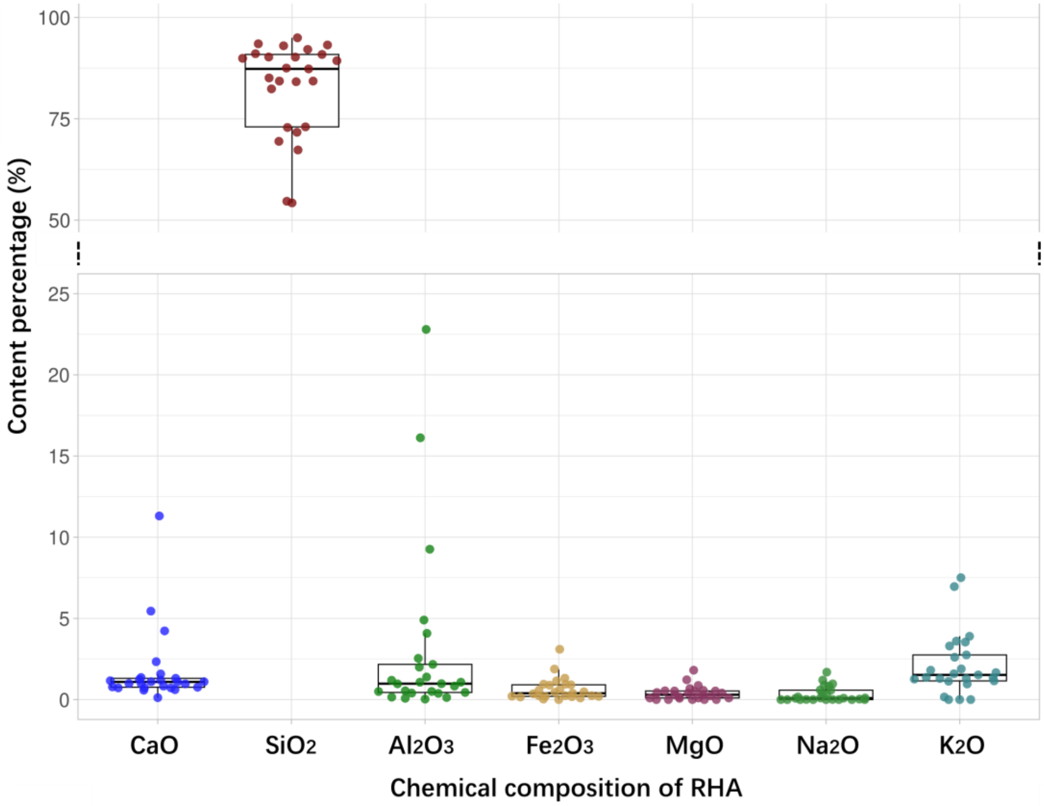
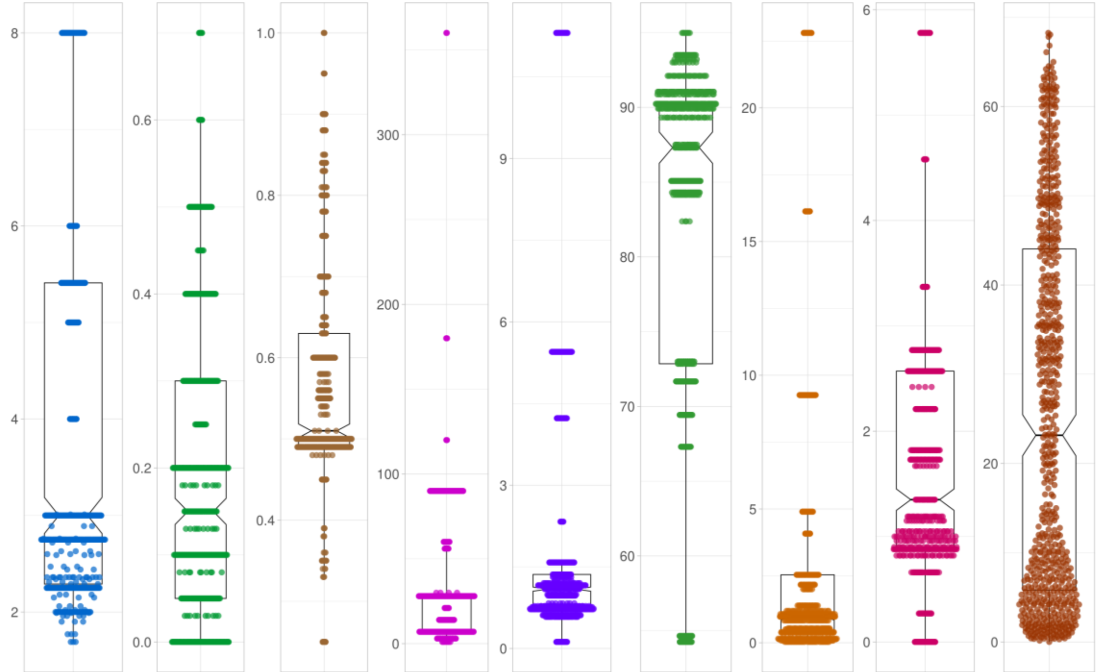


Fig. 7. Chemical composition variation of RHA from various sources

The variability in chemical composition among different samples suggests that RHA can be tailored for specific applications; for insistence, a greater SiO_2 content preferred for applications requiring superior strength. However, this variability may also challenge standardizing RHA for commercial use, highlighting the necessity for thorough analyses to ensure consistent quality.

The forecast of strength in cement mortar is significantly influenced by various input variables, including the Agg/B, RHA/B, W/B, curing period, and the chemical composition of the binder, specifically the percentages of CaO, SiO_2, Al_2O_3, and Na_2O equivalent. **Fig. 8** summarizes the data distribution matrix for the input and output variables. For example, an Agg/B ratio ranging from 1.7:1 to 8:1 can optimize the balance between strength and workability, while RHA/B ratios of 0% to 70% have been shown to enhance pozzolanic activity without compromising structural integrity. The W/B ratio is critical, with values typically between 0.25 and 1.0 correlating with increased strength due to reduced porosity, although excessively low water content can hinder workability. The curing period also plays a significant role; 7 to 28 days is often optimal for maximizing compressive strength, as extended curing allows for more extensive hydration of the cement. Additionally, the chemical compositions of CaO (typically 0.1% to 11.3%), SiO_2 (54.2% to 95.0%), Al_2O_3 (0.03% to 22.8%), and Na_2O equivalent (up to 5.8%) are vital for assessing the reactivity of the binder. For instance, higher

CaO content promotes hydration, while the presence of SiO_2 and Al_2O_3 is crucial in forming calcium silicate hydrates, significantly contributing to strength development.



	Agg/B	RHA/B	W/B	Curing period (days)	CaO (%)	SiO_2 (%)	Al_2O_3 (%)	Na_2O Eq. (%)	Comp. strength (MPa)
Minimum	1.69	0.00	0.25	1.00	0.12	54.25	0.03	0.00	0.15
Maximum	8.00	0.70	1.00	360.00	11.31	94.97	22.80	5.78	68.25
Average	3.77	0.19	0.57	34.25	1.82	82.30	2.84	1.67	25.72
Std	1.91	0.16	0.14	39.25	2.35	11.48	4.92	1.27	20.49
Skewness	1.18	0.80	0.69	3.22	2.91	-1.19	2.84	1.44	0.35
Kurtosis	0.09	-0.08	0.59	19.79	8.35	0.28	7.85	2.62	-1.28

Fig. 8. Range of values for input variables used

Statistical analysis of the data can reveal important insights into the distribution of results. For instance, skewness and kurtosis measures can help assess the normality of the data distribution. A skewness value close to zero indicates a symmetric distribution of compressive strength values, while positive skewness suggests a tail on the right side, indicating that most values are concentrated on the lower end. Conversely, negative skewness would indicate a concentration on the higher end. Kurtosis measures the "tailedness" of the distribution; a kurtosis value more significant than three indicates a distribution with heavier tails than a normal distribution, this could indicate the existence of outliers or extreme values in the strength data.

Fig. 9 analyzes the correlation between input variables themselves and strength. The Agg/B exhibits a strong negative correlation with strength (-0.66), demonstrating that an increase in Agg/B ratio significantly diminishes strength. Similarly, the RHA/B shows a moderate negative correlation (-0.31), implying that higher RHA content may slightly reduce compressive strength, though the effect is less pronounced compared to Agg/B. The W/B ratio also displays a notable negative correlation (-0.52), reinforcing the conventional understanding that excessive water can weaken cement mortar. On the other hand, the curing period shows a positive correlation (0.36), emphasizing the significance of longer curing times in improving compressive strength due to the binder's hydration process.

Examining the chemical composition, CaO shows a negative correlation (-0.17), suggesting that increased CaO content may adversely affect strength, potentially due to its impact on hydration chemistry. On the other hand, SiO_2 has a strong positive correlation (0.54), indicating that higher SiO_2 levels contribute significantly to strength, likely through the formation of beneficial silicate hydrates. The Al_2O_3 presents a moderate negative correlation (-0.34), suggesting that elevated levels could hinder strength development, while Na_2O Eq. displays a negligible negative correlation (-0.1), signifying minimal impact on strength.

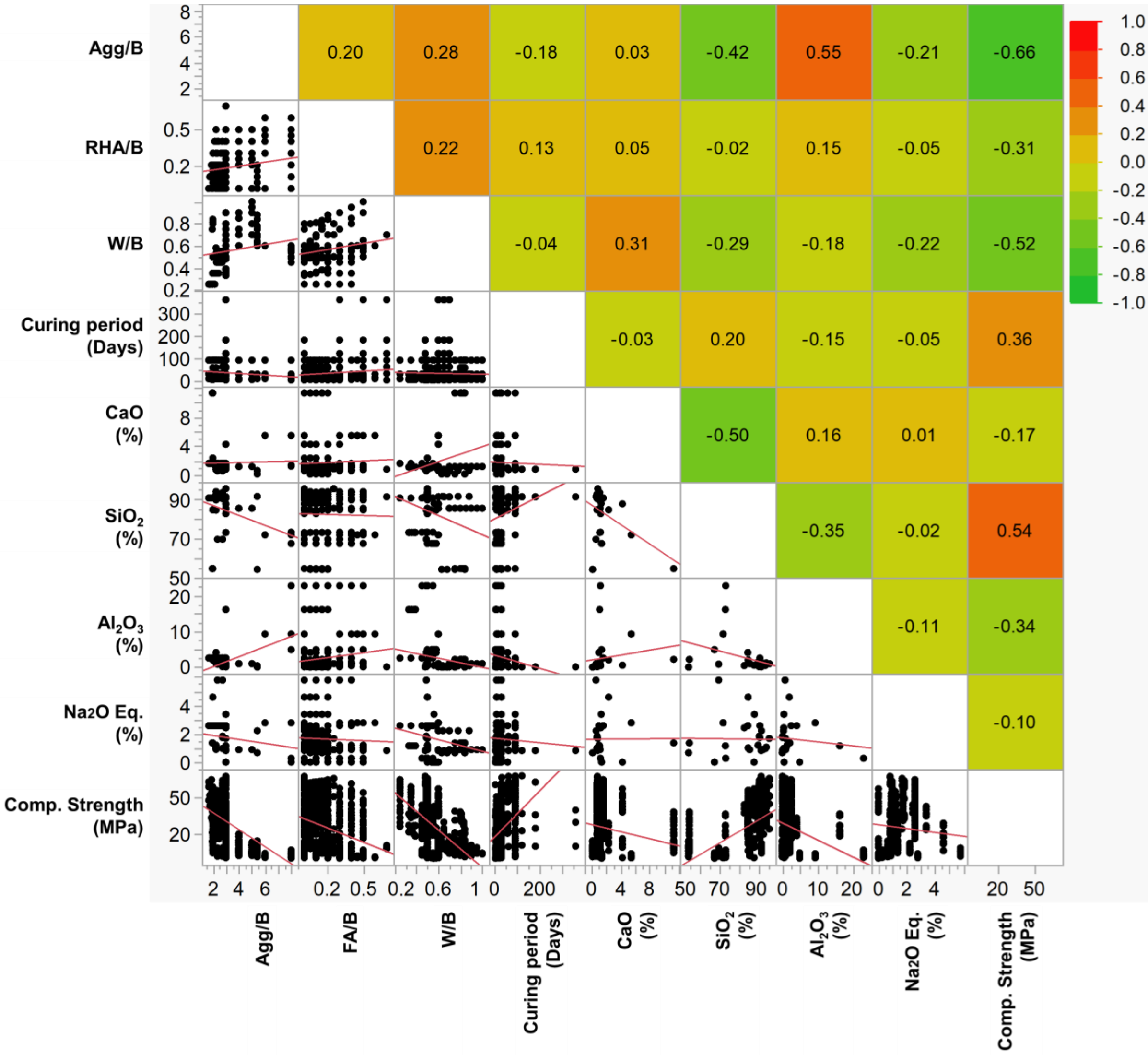


Fig. 9. The relationships between variables and correlation coefficients between variables

3.2 Hyperparameter tuning and k-fold validation

The hyperparameter tuning results for ML models expected at forecasting the strength are summarized in **Table 3**. For the ANN, the selection of the TanH activation function reflects a preference for non-linear transformations, effectively capturing complex associations within the data. The choice of two layers indicates a moderate complexity that balances learning capacity and the risk of overfitting. Additionally, the nodes per layer were optimized within a range of 1-16, allowing for effective learning while maintaining manageability, and a learning rate of 0.1 suggests a focus on rapid convergence, albeit with a need for careful monitoring to prevent overshooting. In the KNN model, the choice of k=2 indicates a sensitivity to local variations in the dataset, which could enhance the model's ability to capture subtle changes in compressive strength. The RF model, utilizing 50 trees, enhances robustness and stability in predictions, while parameters such as eight terms sampled per split and a maximum of 10 splits reflect a balanced approach to model complexity and generalization.

The SVR employs a Radial Basis Function, which adeptly handles non-linear relationships, supported by finely tuned parameters for cost and gamma (4.38068 and 0.46392, respectively). This careful tuning illustrates a well-structured approach to balancing the trade-off between margin width and classification accuracy. The XGB model features a maximum depth of 3 to prevent overfitting while allowing for sufficient model complexity. Parameters such as a subsample rate of 0.5308 and a colsample by tree rate of 0.5273 indicate a meticulous consideration of data sampling strategies, which enhances model generalization.

Table 3. Optimization of hyperparameters for predictive models

Model	Hyperparameters	Selected value	Analyzed Values
ANN	Activation Functions	TanH	{Linear, Gaussian, TanH}
	Number of layers	2	{1, 2}
	Nodes per layer	[5,15]	{1-16}
	Learning rate	0.1	{0.01, 0.05, 0.1, 0.25, 0.5}
KNN	k	2	{1-10}
RF	Number of trees	50	{1-50}
	Number of terms sampled per split	8	{1-8}
	Maximum number of splits per tree	10	10
	Minimum splits per tree	1	1
	Minimum size split	1	1
SVR	Kernel	Radial basis function	{linear, Radial basis function}
	Cost	4.38068	{0.01-5}
	Gamma	0.46392	{0.001-0.5}
XGB	Max_depth	3	{1-5}
	Subsample	0.5308	{0.5-1}
	Colsample_by tree	0.5273	{0.5-1}
	Min_child_weight	2.0910	{1-3}
	Alpha	0.2248	{0-0.5}
	Lambda	1.5570	{0-2}
	Learning rate	0.1674	{0.05-0.2}
	Iterations	21	{1-50}

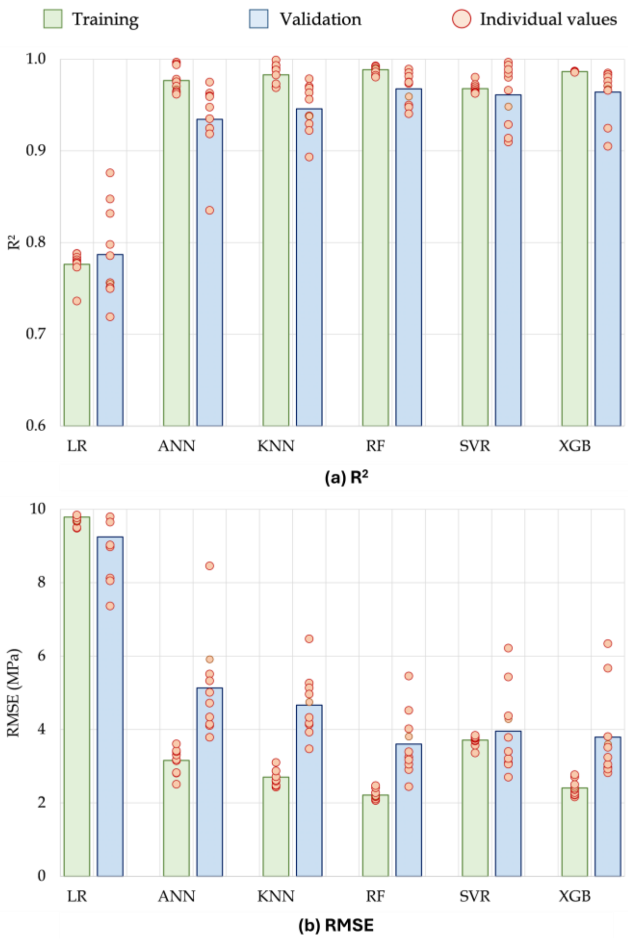


Fig. 10. R^2 and RMSE values for k-fold validation of ML models

The results of the R^2 and RMSE from the k-fold validation are illustrated in **Fig. 10**. The Linear Regression model exhibits moderate R^2 values, with 0.776 for the testing dataset and 0.787 for the testing dataset, suggesting that while it captures some variability in the data, it may not fully account for the complex relationships inherent in the dataset. In contrast, the ANN demonstrates exceptional performance, achieving R^2 values of 0.977 for the testing dataset and 0.934 for the testing dataset, indicative of its ability to capture intricate patterns effectively. Similarly, the K-Nearest Neighbors (KNN) model shows strong results with an R^2 of 0.983 for the testing dataset and 0.9458 for the testing dataset, benefiting from its instance-based learning approach. The RF model stands out with the highest R^2 of 0.989 for training and 0.968 for testing datasets, underscoring its robustness and capacity to model complex relationships through ensemble learning. The SVR maintains solid performance with an R^2 of 0.9679 for training and 0.961 for testing datasets, effectively capturing non-linear patterns while balancing bias and variance. The XGB also performs exceptionally well, with R^2 of 0.986 for the training dataset and 0.964 for the testing dataset, reinforcing its capability to handle complex data structures.

Regarding RMSE values, which indicate the average prediction error, LR shows relatively high RMSE values (9.785 MPa and 9.244 MPa for training and testing datasets, respectively), aligning with its moderate R^2 performance. The ANN achieves a low RMSE of 3.158 MPa for training and 5.131 for testing datasets, demonstrating substantial predictive accuracy, though the discrepancy suggests potential overfitting. KNN records an RMSE of 2.701 MPa for the training dataset, while RF achieves the lowest RMSE of 2.214 MPa for the testing dataset, reinforcing its reliability. SVR's test RMSE of 3.709 MPa indicates reasonable accuracy, while XGB's test RMSE of 2.405 MPa showcases its efficiency and precision in predictions. The results demonstrate that RF and XGB offer the most favourable balance of high R^2 values and low RMSE, indicating robust predictive capabilities for compressive strength. Conversely, LR's moderate performance suggests that more complex models are better suited for capturing the intricacies of this prediction task.

3.3 Performance of the models

Fig. 11 illustrates the predicted versus measured compressive strength, visually assessing how closely each model's predictions align with actual measurements. **Table 4** summarizes key performance indicators, including R^2, RMSE MAE and the a20 index for training and testing datasets.

The LR model achieves R^2 of 0.78 for training and 0.79 for testing, indicating moderate explanatory power. However, its RMSE of 9.78 MPa (train) and 9.24 MPa (test) are relatively high, suggesting substantial prediction errors. The a20 index of 0.36 (train) and 0.37 (test) further highlights that only a tiny proportion of predictions fall within the acceptable 20% error margin, underscoring the model's limitations in practical accuracy. In contrast, the ANN demonstrates superior performance, with training and testing R^2 of 0.98 and 0.93, respectively. The RMSE are significantly lower at 3.16 MPa (train) and 5.13 MPa (test), indicating more accurate predictions. The a20 index of 0.73 (train) and 0.70 (test) suggests that a considerable proportion of forecasts fall within the 20% margin, showcasing the model's practical applicability.

The RF model achieves the highest R^2 of 0.99 (train) and 0.97 (test), reflecting excellent predictive capability. Its RMSE of 2.21 MPa (train) and 3.60 MPa (test) is the lowest among all models, demonstrating high accuracy. The a20 index of 0.81 (train) and 0.78 (test) specifies that a significant percentage of forecasts are within the acceptable error range, emphasizing its robustness. Similarly, the KNN model shows strong performance, with R^2 of 0.98 (train) and 0.95 (test) and RMSE of 2.70 MPa (train) and 4.66 MPa (test), reflecting good predictive accuracy. Its a20 index of 0.89 (train) and 0.66 (test) indicates a high proportion of forecasts within the 20% error margin, particularly in training.

The SVR exhibits strong performance with R^2 of 0.97 (train) and 0.96 (test) and RMSE of 3.71 MPa (train) and 3.95 MPa (test), suggesting reasonable accuracy. The a20 index of 0.70 (train) and 0.71 (test) reflects good practical accuracy. The XGB delivers R^2 of 0.99 (train) and 0.96 (test), demonstrating excellent predictive capability. With RMSE of 2.41 MPa (train) and 3.79 MPa (test), it ranks among the more accurate models. The a20 index of 0.76 (train) and 0.73 (test) indicates a strong proportion of predictions within the 20% error margin, supporting its reliability. The results highlight that while traditional models like LR may struggle with accuracy, more sophisticated models such as

RF and XGB demonstrate exceptional performance characterized by high R², low RMSE, and substantial a20 indices.

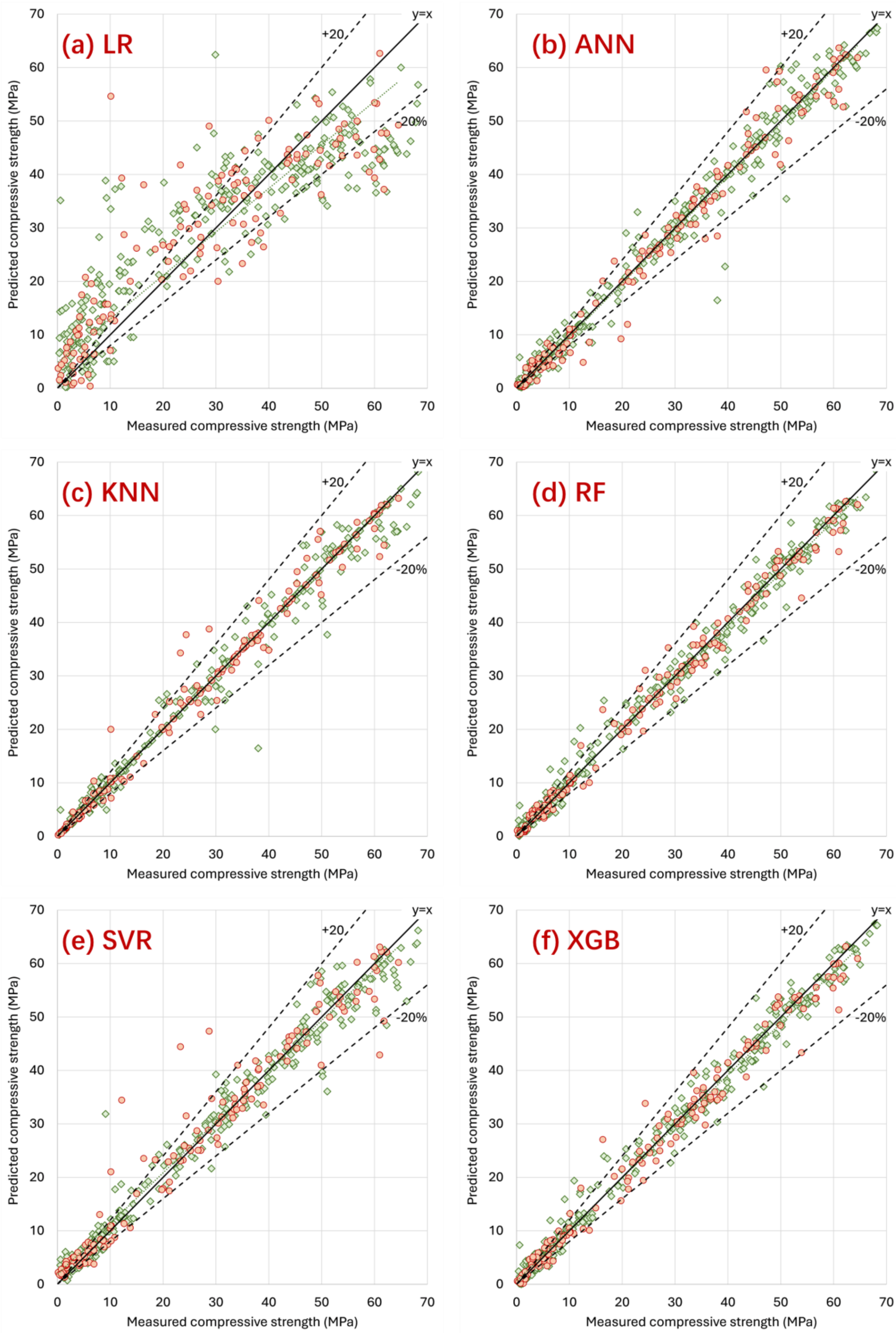
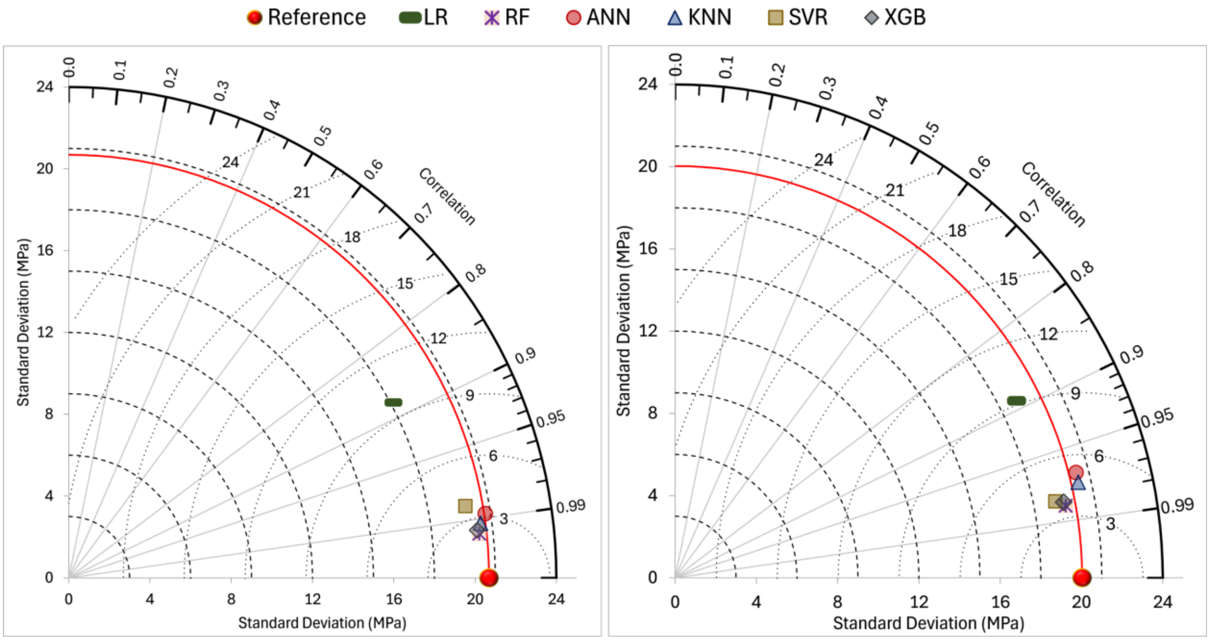


Fig. 11. The performance of various ML models (predicted vs. actual values)

Table 4. Performance indicators for the prediction of compressive strength

ML model	Train				Test			
	R2	RMSE (MPa)	MAE (MPa)	a20	R2	RMSE (MPa)	MAE (MPa)	a20
LR	0.78	9.78	7.64	0.36	0.79	9.24	7.47	0.37
ANN	0.98	3.16	2.01	0.73	0.93	5.13	3.11	0.70
RF	0.99	2.21	1.52	0.81	0.97	3.60	2.33	0.78
KNN	0.98	2.70	1.36	0.89	0.95	4.66	3.06	0.66
SVR	0.97	3.71	2.40	0.70	0.96	3.95	2.66	0.71
XGB	0.99	2.41	1.70	0.76	0.96	3.79	2.51	0.73



(a) Training dataset **(b) Testing dataset**

Fig. 12. Taylor diagram for (a) Training dataset and (b) Testing dataset

The Taylor graph presented in **Fig. 12** delivers valuable understandings into the performance of different ML models for assessing strength. The results indicate that all the models, except linear regression, exhibit strong correlation coefficients, ranging from 0.96 to 0.99 for both training and testing datasets. These high correlation values suggest that these models closely align with the observed data, demonstrating their effectiveness in capturing the underlying patterns in compressive strength predictions.

In contrast, the Linear Regression model, while showing low performance, has lower correlation values of 0.88 (training) and 0.89 (testing). This specifies that LR may not fully account for the complexities of the data, which is reflected in its higher standard deviations of 18.15 (training) and 18.89 (testing), competed to the other models. The standard deviations for the ML model, particularly RF and ANN, are relatively close to the observed standard deviations, suggesting that these models provide accurate predictions and maintain low variability in their outputs. The results underscore the robustness of RF, ANN, KNN, SVR, and XGB, which consistently demonstrate high correlation and low standard deviation variation from observed value, indicating their reliability for realistic purposes in forecasting compressive strength.

Based on the performance indicators, RF and XGB models are the most appropriate for forecasting the strength of RHA-incorporated cement mortar. Both models demonstrated high R^2 values (0.989 for RF and 0.986 for XGB) and low RMSE values, indicating robust predictive capabilities. Among these, RF offers a better balance between interpretability and performance, with an RMSE of 2.21 MPa (train) and 3.60 MPa (test). XGB, while similarly high-performing, may be more complex due to its sensitivity to hyperparameter tuning. Thus, for practical applications, RF is recommended due to its simplicity and excellent performance, although XGB could also be a viable option for more complex scenarios or when computational resources are available.

3.4 Error distribution

Fig. 13 provides a detailed assessment of prediction errors for various ML models that estimate strength. For the training subset, the LR exhibits a minimum error of -25.52 MPa and a maximum error of 44.51 MPa, indicating a broad range of forecast inaccuracies. Its mean prediction error is close to zero at -0.08 MPa, suggesting that the model is nearly unbiased, but the substantial range of errors highlights its susceptibility to significant deviations from actual values. The skewness of 0.31 specifies a slight positive skew, suggesting that the model sometimes overestimates compressive strength.

In contrast, the RF, ANN, KNN, SVR, and XGB models demonstrate notably lower maximum errors. For instance, RF has a maximum error of 8.87 MPa, while ANN, KNN, SVR, and XGB show maximum errors between 10.77 and 12.34 MPa. The mean prediction errors for these models are also relatively close to zero, reflecting a balanced predictive capacity. Notably, the ANN model has a mean error of -0.24 MPa, indicating a slight tendency to underestimate compressive strength. The skewness values for RF (-0.06) and XGB (0.15) are close to zero, suggesting a more symmetric distribution of errors, whereas ANN (-0.94) and KNN (-0.91) exhibit significant negative skewness, indicating a tendency towards underestimation.

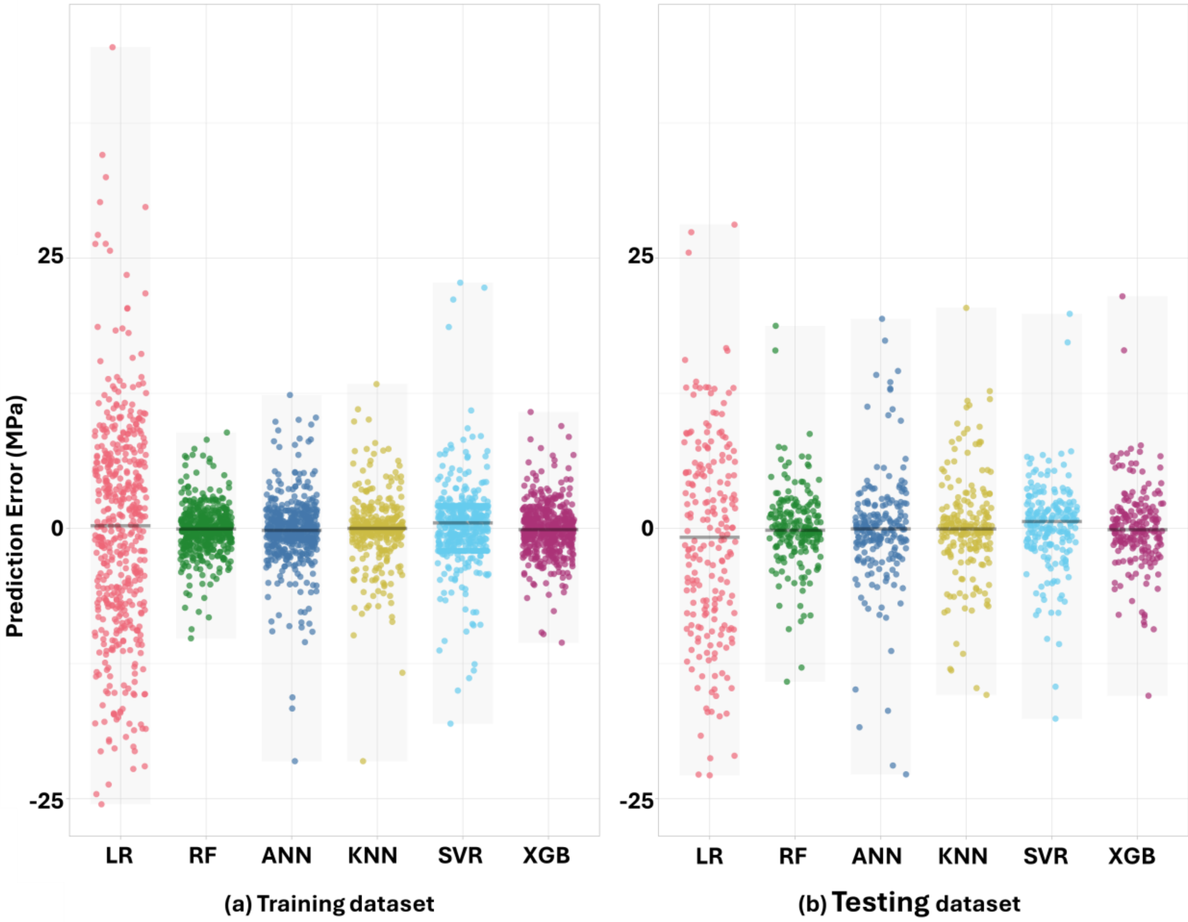
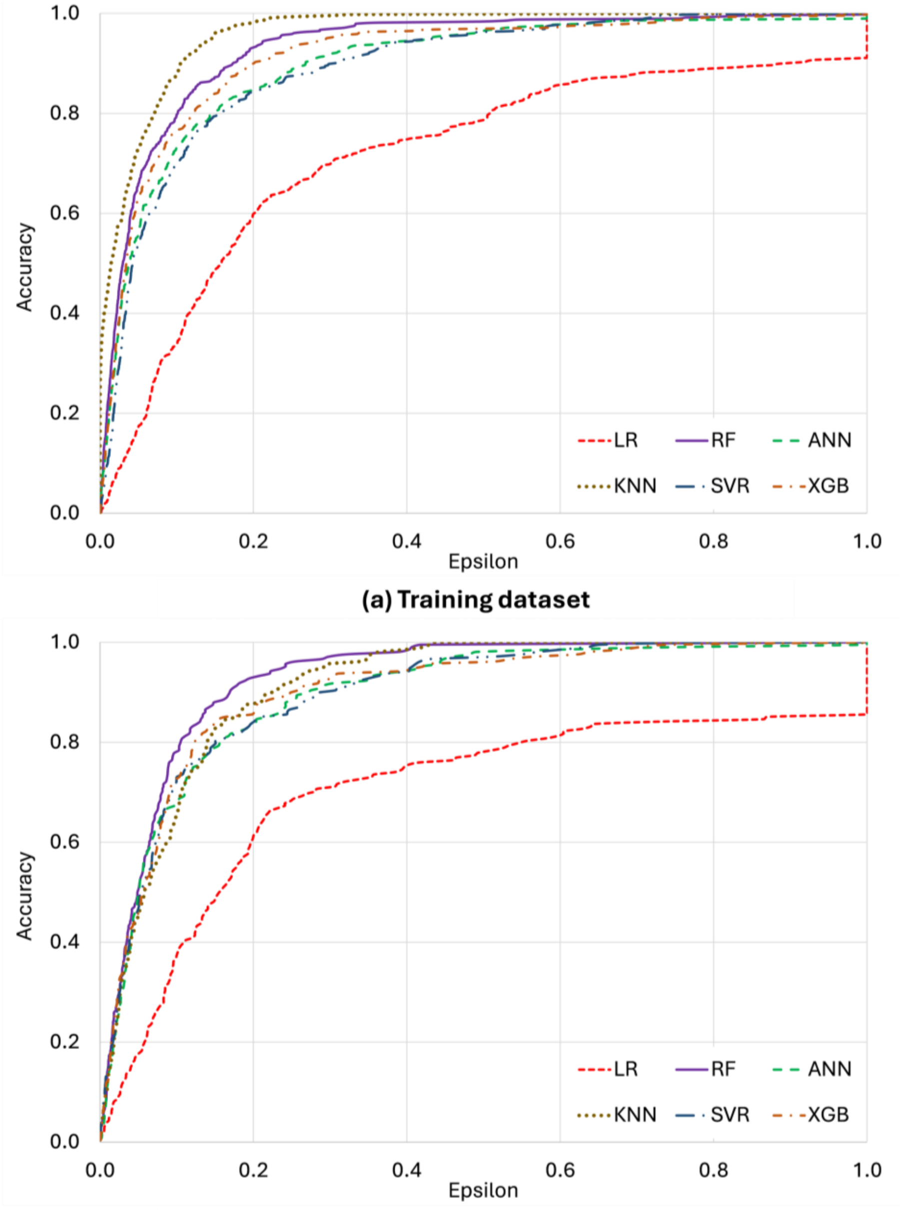


Fig. 13. Prediction error distribution

In the testing dataset, similar trends are observed. LR again shows a wide range of errors, with a minimum of -22.84 MPa and a maximum of 28.10 MPa, leading to a mean error of -0.92 MPa. This reinforces the model's inconsistency and potential for substantial overestimations. The skewness of 0.14 specifies a minor positive skew, suggesting a tendency towards overestimation. The RF model maintains a low maximum error of 18.74 MPa, with a mean error of 0.08 MPa, indicating its dependability in forecasting strength. The ANN and KNN models also show strong performance, with maximum errors of 19.39 MPa and 20.40 MPa, respectively, and mean errors close to zero, reflecting their effectiveness. The SVR model exhibits a mean error of 0.27 MPa with a maximum of 19.85 MPa, while XGB shows a mean error of 0.11 MPa and a maximum error of 21.47 MPa, indicating reliable predictive capabilities.

The statistical indices for prediction error highlight the strengths and weaknesses of each model in estimating compressive strength. While traditional models like LR struggle with high variability and significant prediction errors, the advanced ML models (RF, ANN, KNN, SVR, and XGB) demonstrate greater consistency and accuracy, characterized by lower maximum errors and mean close to zero.

Fig. 14 illustrates the REC curves, which comprehensively evaluate the performance of different ML models. The Area Under the Curve (AUC) of these curves acts as a key indicator of model performance, with higher AUC values reflecting a greater ability to reduce prediction errors.



(a) Training dataset

(b) Testing dataset

Fig. 14. The Regression Error Characteristic (REC) curve

In the training dataset, the KNN establishes the highest AUC value of 0.96, demonstrating its capability to accurately capture the connection between the predictors and compressive strength, leading to minimal regression errors. The RF and XGB models closely follow, with AUC of 0.93 and 0.92, respectively. The results specify that both models demonstrate robust predictive capabilities, effectively managing bias and variance trade-offs. The ANN and SVR models achieve respectable AUC of 0.90 and 0.89, respectively, demonstrating their competence in forecasting compressive strength, albeit with slightly less effectiveness than KNN, RF, and XGB. However, the LR model shows the lowest AUC value of 0.72, indicating a relatively poor performance in minimizing prediction errors, which aligns with its earlier challenges in identifying intricate patterns within the data.

In the testing dataset, the performance trends are somewhat consistent with the training results. KNN again leads with an AUC of 0.91, demonstrating its continued effectiveness in predicting compressive strength. RF maintains its robust performance with an AUC of 0.93, while ANN and SVR exhibit AUC values of 0.90, indicating reliable predictive capabilities. XGB also remains competitive with an AUC of 0.90, reinforcing its robustness. LR continues to underperform in the testing dataset with an AUC of 0.70, reaffirming its limitations in this context. The lower AUC value for LR suggests that it struggles to balance true positive and false positive rates, resulting in less effective predictions than the more advanced models. The AUC values derived from the REC curves underscore the effectiveness of advanced ML models (KNN, RF, XGB, ANN, and SVR) in predicting compressive strength while minimizing regression errors. These models' consistently high AUC values indicate their robustness and reliability in practical applications.

3.5 Sensitivity analysis

The SHAP analysis results in **Fig. 15** and **Fig. 16** specify precious perceptions into the sensitivity of input variables affecting the strength of cement mortar. The mean absolute values for the input variables indicate their relative importance in the prediction model, with Agg/B showing the highest mean absolute value (9.858), followed by SiO2 (4.865) and Curing period (4.808). This suggests that these variables play a critical role in affecting compressive strength.

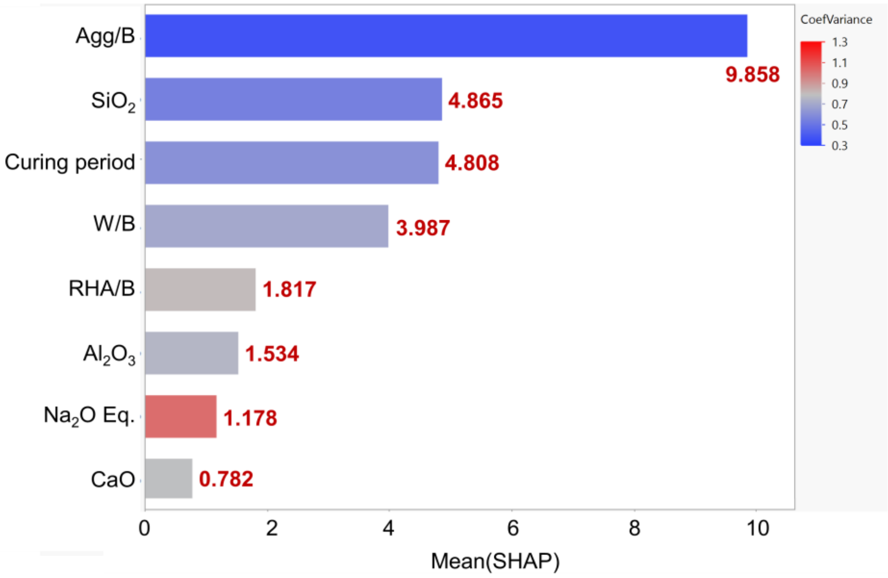


Fig. 15. Mean SHAP value for the input variable derived from the RF model

The analysis shows a notable SHAP value of -20 for high Agg/B ratios, indicating that higher values of this ratio adversely affect compressive strength. In contrast, lower Agg/B ratios result in a positive SHAP value (+15), underscoring the significance of achieving an optimal balance in the Agg/B ratio for improved strength. The SHAP value (+12) for extended curing periods proposes that longer curing times contribute positively to compressive strength. The decrease in SHAP value (-13) with reduced curing periods reinforces the significance of adequate curing in cementitious materials. The SiO_2 exhibits mixed behaviour, with negative SHAP values for lower and mid-range concentrations and a positive trend for higher concentrations. This indicates that while too little SiO_2 may be detrimental, optimal levels can enhance strength, potentially due to its role in the pozzolanic reaction. Higher W/B ratios lead to negative SHAP values (-18), suggesting increased porosity and reduced strength. In contrast, lower W/B ratios positively influence strength (+10), emphasizing the importance of water management in mix design. Like W/B, RHA/B ratios show adverse effects at higher values and benefits at lower ratios. This suggests that while RHA can enhance strength judiciously, excessive amounts may compromise structural integrity.

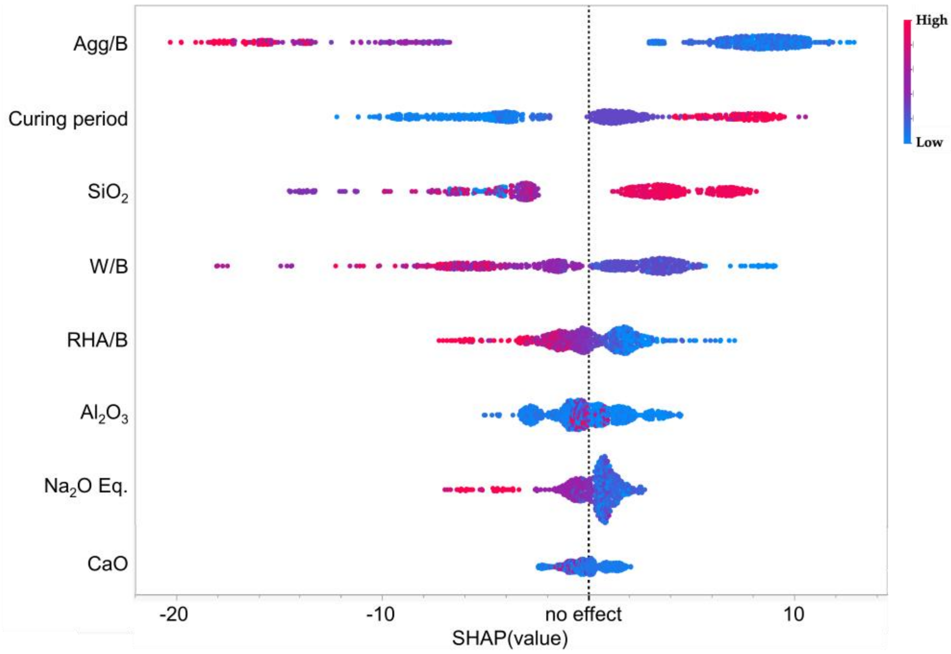


Fig 16. Summary plot of input variables generated from the RF model using SHAP values

Although the SHAP analysis emphasizes the impact of primary oxides like SiO_2 and CaO, the effect of Al_2O_3 is inconsistent, with SHAP values varying around zero for medium to high concentrations. This variability likely stems from its interactions with other chemical components in the mix, which complicates the understanding of its influence on compressive strength. The observed fluctuations in Al_2O_3's impact suggest that its effects are not independent and may depend on the specific formulation of the mix. To gain a more comprehensive understanding, further studies should emphasis on isolating the effects of Al_2O_3 and examining its interaction with other oxides in greater detail. Higher values of Na_2O equivalent correlate with negative SHAP values, while lower values have a positive impact. This may relate to the detrimental effects of excessive sodium in alkali-silica reactions. CaO shows a narrow range of SHAP values (-2 to 2), suggesting that its effect on compressive strength is minimal associated to other variables. The SHAP analysis effectively demonstrates the complex relations among input variables and the strength of cement mortar. Notably, optimizing the Agg/B and W/B ratios, alongside adequate curing periods, emerges as critical factors in enhancing strength.

3.6 Limitations of the model and future research direction

While this research offers precious perceptions into using RHA as an additive in cement mortar and employs ML models to predict compressive strength, several limitations warrant consideration. Future studies should focus on evaluating additional factors beyond compressive strength, such as workability, setting time, durability, and other mechanical properties like tensile and flexural strength. Workability tests and setting time measurements are essential to ensure RHA-blended mortars meet practical construction needs, while durability tests, including resistance to chemical attacks and water absorption, are crucial for assessing long-term performance. A broader chemical analysis of RHA and its interactions in cement mortar could provide deeper insights into its pozzolanic activity, while larger, diverse datasets would improve the generalizability of predictive models. Experimental validation, including field studies and sustainability assessments, should be prioritized to confirm the applicability and environmental benefits of RHA in cement-based materials. Exploring advanced ML techniques could further enhance model accuracy, ensuring RHA's optimal use in sustainable construction practices. The following points outline the primary limitations identified in this research:

- Limited Dataset: The dataset, sourced from existing literature, is acknowledged for potentially limiting the practical relevance and broad applicability of the conclusions. While it provides useful perceptions into the link among RHA properties and strength, the lack of experimental validation undermines the reproducibility and real-world applicability of the findings.

- Focus on Specific Oxides: The analysis primarily considers select chemical components of RHA (e.g., SiO_2, CaO, Al_2O_3). Other potentially influential oxides and their interactions may not have been sufficiently explored, thus restricting the understanding of their combined effects on compressive strength.
- Model Generalizability: It is suggested that future research include experimental studies in diverse locations and under varying environmental conditions to ensure the models' generalizability and robustness. This would help refine the models and ensure their applicability across different contexts in the construction industry.
- Further research is necessary to leverage RHA's benefits and address the identified limitations. Exploring new avenues will refine the predictive models and advance the understanding of RHA's role in sustainable construction practices. The following points outline the future research direction:
- Broader Chemical Analysis: Future studies should develop the investigation to include a broader range of chemical compositions and oxides in RHA, examining their synergistic effects on the properties of cement mortar.
- Larger and Diverse Datasets: Collecting and analyzing a more extensive and diverse dataset, including various sources of RHA and different environmental conditions, could enhance predictive models' robustness and applicability in real-world scenarios.
- Experimental Validation: Further studies should prioritize experimentally validating the predictive models created in this research. This process would require performing lab experiments to evaluate the compressive strength, workability, and setting time of cement mortars incorporating RHA, considering various RHA sources and environmental factors. Such experimental validation would provide empirical data to confirm the predictions and enhance the practical relevance of the models.
- Exploration of Additional ML Techniques: Future research could investigate more sophisticated ML methods, like ensemble or deep learning techniques, to better capture the complex relationships within the data.
- Sustainability Assessment: Investigating the environmental impacts of using RHA in cement mortar, including life cycle assessments, could provide a more comprehensive understanding of its sustainability benefits than traditional materials.
- Field Studies: Conducting field studies to evaluate the performance of RHA-blended cement mortar in real construction assignments would provide practical insights and help validate laboratory findings in real-world applications.

By overcoming these limitations and following the proposed research paths, further research can offer a deeper understanding of RHA's role in sustainable construction.

4 Conclusions

This study emphasizes the considerable prospective of RHA as an additive in cement mortar, particularly its capacity to boost compressive strength through ML predictive models. Key observations reveal that the chemical composition of RHA, especially the presence of silica oxide (SiO_2), is vital in influencing its pozzolanic activity and the overall effectiveness of the mortar.

The research demonstrates that varying the ratios of Agg/B, RHA/B, and W/B significantly influences strength. Specifically, an optimal RHA/B ratio can enhance strength without compromising workability, while excessive use of aggregates adversely affects performance. The study also indicates that longer curing periods are positively linked to strength development, underscoring the importance of adequate curing to attain the desired material properties.

Moreover, applying advanced ML techniques effectively captures the complex interactions among various input parameters, enabling accurate predictions of compressive strength. Nevertheless, the study highlights significant limitations, including the variability in the chemical composition of RHA and crucial for enhancing the model's generalizability is the inclusion of larger datasets.

These findings not only advance the understanding of RHA's potential in sustainable cement mortar formulations but also highlight the role of ML in optimizing material properties for more sustainable construction practices. As ML continues to evolve, its integration into the development of mortar standards could lead to more precise, data-driven guidelines, enabling engineers to design stronger, more durable, and environmentally friendly materials. Future studies should aim to verify these predictions through experimental investigations, enhance the dataset, and examine other chemical compositions to optimize the usage of RHA in cement-based materials. The construction industry can shift on more ecological and eco-friendly exercises by focusing on these key areas.

Abbreviations

Agg/B	Aggregate-to-binder ratio
RHA	Rice husk ash
W/B	Water-to-binder ratio
CaO	Calcium oxide
SiO_2	Silicon dioxide
Al_2O_3	Aluminium Oxide
Na_2O Eq.	Equivalent Alkali
ANN	Artificial Neural Networks
KNN	K-Nearest Neighbors
RF	Random Forest Regression
SVR	Support Vector Regression
XGB	eXtreme Gradient Boosting
R^2	Coefficient of determination
RMSE	Root Mean Squared Error
MAE	Mean Absolute Error
SHAP	SHapley Additive exPlanations
REC	Regression Error Characteristic
AUC	The Area Under the REC Curve

Acknowledgement

The author would like to express their sincere appreciation to the Department of Civil Engineering, University of Jaffna, for their exceptional assistance.

Funding Statement

The author did not receive any dedicated funding.

CRediT authorship contribution statement

Sathiparan: Conceptualization, Investigation, Formal analysis, Writing – original draft. Writing – review & editing.

Conflicts of Interest

The author affirm that there are no conflicts of interest to disclose in relation to this study.

References

[1] Niaounakis M. Building and Construction Applications, In: Niaounakis M (Eds.), Biopolymers: Applications and Trends, Oxford: William Andrew Publishing, New York 2015; 445-505.

[2] Ben Ghida D. Organic stabilization in earthen plaster: Eco-compatible architecture and ancient techniques in Tata Somba homes. Frontiers of Architectural Research 2024; 13(3): 625-638. https://doi.org/10.1016/j.foar.2024.02.004.

[3] Mohamed OA, Zuaiter HA, Jawa MM. Carbonation and chloride penetration resistance of sustainable structural concrete with alkali-activated and ordinary Portland cement binders: a critical review. Sustainable Structures 2024; 5(2): 000075. https://doi.org/10.54113/j.sust.2025.000075.

[4] Claisse PA. Mortars and grouts. In: Claisse PA (Eds.), Civil Engineering Materials, Butterworth-Heinemann, Boston 2016; 303-311.

[5] Barbhuiya S, Kanavaris F, Das BB, Idrees M. Decarbonising cement and concrete production: Strategies, challenges and pathways for sustainable development. Journal of Building Engineering 2024; 86: 108861. https://doi.org/10.1016/j.jobe.2024.108861.

[6] Tan YY, Hanizam A, Kaus NHM. Integration of fly ash and ground granulated blast furnace slag into palm oil fuel ash based geopolymer concrete: a review. Sustainable Structures 2024; 4(2): 000050. https://doi.org/10.54113/j.sust.2024.000050.

[7] Villagrán-Zaccardi YA, Marsh ATM, Sosa ME, Zega CJ, De Belie N, Bernal SA. Complete re-utilization of waste concretes–Valorisation pathways and research needs. Resources, Conservation and Recycling 2022; 177: 105955. https://doi.org/10.1016/j.resconrec.2021.105955.

[8] Sathiparan N, Subramaniam DN. Optimizing fly ash and rice husk ash as cement replacements on the mechanical characteristics of pervious concrete. Sustainable Structures 2025; 5(1): 000065. https://doi.org/10.54113/j.sust.2025.000065.

[9] Chen L, Yang M, Chen Z, Xie Z, Huang L, Osman AI, Farghali M, Sandanayake M, Liu E, Ahn YH et al. Conversion of waste into sustainable construction materials: A review of recent developments and prospects. Materials Today Sustainability 2024; 27: 100930. https://doi.org/10.1016/j.mtsust.2024.100930.

[10] Snellings R, Suraneni P, Skibsted J. Future and emerging supplementary cementitious materials. Cement and Concrete Research 2023; 171: 107199. https://doi.org/10.1016/j.cemconres.2023.107199.

[11] Hafez RDA, Ftah ROA-A, Abdelsamie K. The influence of nucleus dates waste and ceramic wastes in sustainable concrete. Sustainable Structures 2024; 4(2): 000048. https://doi.org/10.54113/j.sust.2024.000048.

[12] Kumar Das S, Adediran A, Rodrigue Kaze C, Mohammed Mustakim S, Leklou N. Production, characteristics, and utilization of rice husk ash in alkali activated materials: An overview of fresh and hardened state properties. Construction and Building Materials 2022; 345: 128341. https://doi.org/10.1016/j.conbuildmat.2022.128341.

[13] Tayeh BA, Alyousef R, Alabduljabbar H, Alaskar A. Recycling of rice husk waste for a sustainable concrete: A critical review. Journal of Cleaner Production 2021; 312: 127734. https://doi.org/10.1016/j.jclepro.2021.127734.

[14] Indumathi M, Nakkeeran G, Roy D, Gupta SK, Alaneme GU. Innovative approaches to sustainable construction: a detailed study of rice husk ash as an eco-friendly substitute in cement production. Discover Applied Sciences 2024; 6(11): 597. https://doi.org/10.1007/s42452-024-06314-1.

[15] Sathiparan N. Prediction model for compressive strength of rice husk ash blended sandcrete blocks using a machine learning models. Asian Journal of Civil Engineering 2024; 25(6): 4745-4758. https://doi.org/10.1007/s42107-024-01077-x.

[16] Amin MN, Khan K, Abu Arab AM, Farooq F, Eldin SM, Javed MF. Prediction of sustainable concrete utilizing rice husk ash (RHA) as supplementary cementitious material (SCM): Optimization and hyper-tuning. Journal of Materials Research and Technology 2023; 25: 1495-1536. https://doi.org/10.1016/j.jmrt.2023.06.006.

[17] Iftikhar B, Alih SC, Vafaei M, Elkotb MA, Shutaywi M, Javed MF, Deebani W, Khan MI, Aslam F. Predictive modeling of compressive strength of sustainable rice husk ash concrete: Ensemble learner optimization and comparison. Journal of Cleaner Production 2022; 348: 131285. https://doi.org/10.1016/j.jclepro.2022.131285.

[18] Li Q, Song Z. Prediction of compressive strength of rice husk ash concrete based on stacking ensemble learning model. Journal of Cleaner Production 2023; 382: 135279. https://doi.org/10.1016/j.jclepro.2022.135279.

[19] Huang Y, Lei Y, Luo X, Fu C. Prediction of compressive strength of rice husk ash concrete: A comparison of different metaheuristic algorithms for optimizing support vector regression. Case Studies in Construction Materials 2023; 18: e02201. https://doi.org/10.1016/j.cscm.2023.e02201.

[20] Alyami M, Nassar R-U-D, Khan M, Hammad AWA, Alabduljabbar H, Nawaz R, Fawad M, Gamil Y. Estimating compressive strength of concrete containing rice husk ash using interpretable machine learning-based models. Case Studies in Construction Materials 2024; 20: e02901. https://doi.org/10.1016/j.cscm.2024.e02901.

[21] Nasir Amin M, Iftikhar B, Khan K, Faisal Javed M, Mohammad AbuArab A, Faisal Rehman M. Prediction model for rice husk ash concrete using AI approach: Boosting and bagging algorithms. Structures 2023; 50: 745-757. https://doi.org/10.1016/j.istruc.2023.02.080.

[22] Odeh R, Alawadi R, Tarawneh A, Alghossoon A, Al-Mazaidh ReA, Amerah H. Estimating Rice Husk Ash Concrete Compressive Strength Using Hybrid Machine Learning Methodology. Engineered Science 2024; 29: 1111. https://doi.org/10.30919/es1111.

[23] Amin MN, Khan K, Javed MF, Ewais DY, Qadir MG, Faraz MI, Alam MW, Alabdullah AA, Imran M. Forecasting Compressive Strength of RHA Based Concrete Using Multi-Expression Programming. Materials

2022; 15(11): 3808. https://doi.org/10.3390/ma15113808.
[24] Kovačević M, Hadzima-Nyarko M, Grubeša IN, Radu D, Lozančić S. Application of Artificial Intelligence Methods for Predicting the Compressive Strength of Green Concretes with Rice Husk Ash. Mathematics 2024; 12(1): 66. https://doi.org/10.3390/math12010066.
[25] Kashem A, Karim R, Das P, Datta SD, Alharthai M. Compressive strength prediction of sustainable concrete incorporating rice husk ash (RHA) using hybrid machine learning algorithms and parametric analyses. Case Studies in Construction Materials 2024; 20: e03030. https://doi.org/10.1016/j.cscm.2024.e03030.
[26] Paul S, Das P, Kashem A, Islam N. Sustainable of rice husk ash concrete compressive strength prediction utilizing artificial intelligence techniques. Asian Journal of Civil Engineering 2024; 25(2): 1349-1364. https://doi.org/10.1007/s42107-023-00847-3.
[27] Iqtidar A, Bahadur Khan N, Kashif-ur-Rehman S, Faisal Javed M, Aslam F, Alyousef R, Alabduljabbar H, Mosavi A. Prediction of Compressive Strength of Rice Husk Ash Concrete through Different Machine Learning Processes. Crystals 2021; 11(4): 352. https://doi.org/10.3390/cryst11040352.
[28] Barbhuiya S, Bhusan Das B, Adak D. Roadmap to a net-zero carbon cement sector: Strategies, innovations and policy imperatives. Journal of Environmental Management 2024; 359: 121052. https://doi.org/10.1016/j.jenvman.2024.121052.
[29] Alwash JJH: Use of rice husk ash in cement mortar. Journal of University of Babylon 2013, 21(2):582-590.
[30] Gowda MR, Narasimhan MC, Karisiddappa N. Development and Study of the Strength of Self-Compacting Mortar Mixes Using Local Materials. Journal of Materials in Civil Engineering 2011; 23(5): 526-532. https://doi.org/10.1061/(ASCE)MT.1943-5533.0000202.
[31] Jamil M, Khan MNN, Karim MR, Kaish ABMA, Zain MFM. Physical and chemical contributions of Rice Husk Ash on the properties of mortar. Construction and Building Materials 2016; 128: 185-198. https://doi.org/10.1016/j.conbuildmat.2016.10.029.
[32] Kondraivendhan B. Strength and flow behavior of rice husk ash blended cement paste and mortar. Asian Journal of Civil Engineering 2013; 14(3): 405-416.
[33] Mayooran S, Ragavan S, Sathiparan N. 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 2017; 13: 137-145. https://doi.org/10.1016/j.jobe.2017.07.011.
[34] Nandurkar BP, Pande AM. Compressive strength of mortars containing fly ash and rice husk ash. International Journal of Civil Engineering and Technology 2018; 9(7): 1012-1020.
[35] Nguyen CV. Properties of locally-sourced rice husk ash (RHA)-blended mortar. Journal of Hunan University (Natural Sciences) 2013; 48(4): 63-75.
[36] Okpala DC. Some engineering properties of sandcrete blocks containing rice husk ash. Building and Environment 1993; 28(3): 235-241. https://doi.org/10.1016/0360-1323(93)90029-3.
[37] Oyetola EB, Abdullahi M. The use of rice husk ash in low - cost sandcrete block production. Leonardo Electronic Journal of Practices and Technologies 2006; 8: 58-70.
[38] Rasoul BI, Gunzel FK, Rafiq MI. Effect of rice husk ash properties on the early age and long-term strength of mortar. In: Journal of University of Babylon, 12-14, June 2017, Maastricht.
[39] Shinge KS, Warad BB, Rathod SBR, Pendhari SP. Partial replacement of cement in mortar by using red mud and rice husk ash. International Journal of Scientific & Engineering Research 2015; 6(9): 1269-1272.
[40] Shukla A, Kishore K, Gupta N. Mechanical properties of cement mortar with lime & rice hush ash. In: IOP Conf Series: Materials Science and Engineering, 12-14, June 2017; Mathura, India.
[41] Subashi De Silva GHMJ, Vishvalingam S, Etampawala T. Effect of waste rice husk ash from rice husk fuelled brick kilns on strength, durability and thermal performances of mortar. Construction and Building Materials 2021; 268: 121794. https://doi.org/10.1016/j.conbuildmat.2020.121794.
[42] Thiedeitz M, Ostermaier B, Kränkel T. Rice husk ash as an additive in mortar – Contribution to microstructural, strength and durability performance. Resources, Conservation and Recycling 2022; 184: 106389. https://doi.org/10.1016/j.resconrec.2022.106389.
[43] Tuleun LZ, Adedeji AA. Behavior of low-carbon rice husk ash in cement mortar under compressive loading. Journal of Civil Engineering and Materials Application 2024; 8(1): 1-16.
[44] Khan W, Fahim M, Zaman S, Khan SW, Badrashi YI, Khan F. Use of Rice Husk Ash as Partial Replacement of Cement in Sandcrete Blocks. Adv Sci Technol Res J 2021; 15(2): 101-107. https://doi.org/10.12913/22998624/133470.
[45] Hien TT. Experimental study on influence of rice husk ash on mortar compressive strength at different temperatures. DTU Journal of Science and Technology 2022; 3(52): 51-59.
[46] Baeza F, Payá J, Galao O, Saval JM, Garcés P. Blending of industrial waste from different sources as partial substitution of Portland cement in pastes and mortars. Construction and Building Materials 2014; 66: 645-653. https://doi.org/10.1016/j.conbuildmat.2014.05.089.
[47] Chindaprasirt P, Rukzon S. Strength, porosity and corrosion resistance of ternary blend Portland cement, rice husk ash and fly ash mortar. Construction and Building Materials 2008; 22(8): 1601-1606. https://doi.org/10

.1016/j.conbuildmat.2007.06.010.
[48] Rukzon S, Chindaprasirt P. Use of disposed waste ash from landfills to replace Portland cement. Waste Management & Research 2009; 27(6): 588-594. https://doi.org/10.1177/0734242X09103189.
[49] Peter SC, Dhanjal JK, Malik V, Radhakrishnan N, Jayakanthan M, Sundar D. Quantitative structure-activity relationship (QSAR): Modeling approaches to biological applications, In: Ranganathan S, Gribskov M, Nakai K, Schönbach C (Eds.), Encyclopedia of Bioinformatics and Computational Biology, Academic Press, Oxford. 2019; 661-676.
[50] Montesinos López OA, Montesinos López A, Crossa J. Fundamentals of Artificial Neural Networks and Deep Learning. In: Montesinos López OA, Montesinos López A, Crossa J (Eds.), Multivariate Statistical Machine Learning Methods for Genomic Prediction, Springer International Publishing, New York, 2022: 379-425.
[51] Ahmed SF, Alam MSB, Hassan M, Rozbu MR, Ishtiak T, Rafa N, Mofijur M, Shawkat Ali ABM, Gandomi AH. Deep learning modelling techniques: current progress, applications, advantages, and challenges. Artificial Intelligence Review 2023; 56(11): 13521-13617. 10.1007/s10462-023-10466-8.
[52] Shi Y, Yang K, Yang Z, Zhou Y. Chapter Two - Primer on artificial intelligence, In: Shi Y, Yang K, Yang Z, Zhou Y (Eds.), Mobile Edge Artificial Intelligence, Academic Press, Massachusetts, 2022; 7-36.
[53] Çetin Aİ, Büyüklü AH. A new approach to K-nearest neighbors distance metrics on sovereign country credit rating. Kuwait Journal of Science 2025; 52(1): 100324. https://doi.org/10.1016/j.kjs.2024.100324.
[54] Mallik M, Panja AK, Chowdhury C. Paving the way with machine learning for seamless indoor–outdoor positioning: A survey. Information Fusion 2023; 94: 126-151. https://doi.org/10.1016/j.inffus.2023.01.023.
[55] Rezaei N, Jabbari P. Random forests in R, In: Rezaei N, Jabbari P (Eds.) Immunoinformatics of Cancers, Academic Press, Massachusetts, 2022; 169-179.
[56] Mansour Y, Schain M. Learning with Maximum-Entropy Distributions. Machine Learning 2001; 45(2):123-145. https://doi.org/10.1023/A:1010950718922.
[57] Adnan MN. Improving the random forest algorithm by randomly varying the size of the bootstrap samples. In: Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration (IEEE IRI 2014), 13-15, Aug. 2014; 303-308.
[58] Salman HA, Kalakech A, Steiti A. Random Forest Algorithm Overview. Babylonian Journal of Machine Learning 2024; 2024: 69-79. https://doi.org/10.58496/BJML/2024/007.
[59] Comito C, Pizzuti C. Artificial intelligence for forecasting and diagnosing COVID-19 pandemic: A focused review. Artificial Intelligence in Medicine 2022; 128: 102286. https://doi.org/10.1016/j.artmed.2022.102286.
[60] Smola AJ, Schölkopf B. A tutorial on support vector regression. Statistics and Computing 2004; 14(3): 199-222. https://doi.org/10.1023/B:STCO.0000035301.49549.88.
[61] Negri RG, Dutra LV, Sant'Anna SJS. An innovative support vector machine based method for contextual image classification. ISPRS Journal of Photogrammetry and Remote Sensing 2014; 87: 241-248. https://doi.org/10.1016/j.isprsjprs.2013.11.004.
[62] Aldhafferi N. Android Malware Detection Using Support Vector Regression for Dynamic Feature Analysis. Information 2024; 15(10): 658. https://doi.org/10.3390/info15100658.
[63] Sathiparan N, Jeyananthan P, Subramaniam DN. Effect of rice husk ash on compressive strength of sustainable pervious concrete and prediction model using machine learning algorithms. Sustainable Structures 2025; 5(3): 000080. https://doi.org/10.54113/j.sust.2025.000080.
[64] Tarwidi D, Pudjaprasetya SR, Adytia D, Apri M. An optimized XGBoost-based machine learning method for predicting wave run-up on a sloping beach. MethodsX 2023; 10: 102119. https://doi.org/10.1016/j.mex.2023.102119.
[65] Sagi O, Rokach L. Approximating XGBoost with an interpretable decision tree. Information Sciences 2021; 572: 522-542. https://doi.org/10.1016/j.ins.2021.05.055.
[66] Wahab S, Salami BA, AlAteah AH, Al-Tholaia MMH, Alahmari TS. Exploring the interrelationships between composition, rheology, and compressive strength of self-compacting concrete: An exploration of explainable boosting algorithms. Case Studies in Construction Materials 2024; 20: e03084. https://doi.org/10.1016/j.cscm.2024.e03084.
[67] Izzaddin A, Langousis A, Totaro V, Yaseen M, Iacobellis V. A new diagram for performance evaluation of complex models. Stochastic Environmental Research and Risk Assessment 2024; 38(6): 2261-2281. https://doi.org/10.1007/s00477-024-02678-3.
[68] Torgo L. Regression error characteristic surfaces. In: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, Association for Computing Machinery, 21-24, Aug 2005. 697–702.