International Journal of Latest Research in Engineering and Technology (IJLRET)
ISSN: 2454-5031
www.ijlret.com || Volume 11 - Issue 04 || April 2025 || PP. 29-48

Prediction of the Coefficient of Thermal Expansion and Other Characteristics of Concrete by Using Machine Learning

Sang Marjan[1], Fawad Ullah[1*], Muhammad Ameer Hamza[1], Muhammad Usama[2], Ubaid Ullah[3], Muhammad Yousaf Khan[1]

[1]*Department of Civil Engineering, COMSATS University Islamabad, Abbottabad Campus,*

Abbottabad, Pakistan

[2]*Department of Geodesy and Geoinformatics, Hamburg, Germany*

[3]*Department of Civil Engineering, University of Wah, Pakistan*

[*]*Corresponding Author*

Abstract: Predicting concrete's thermal expansion coefficient (CTE) is important for improving its performance in different environmental conditions. Traditional methods often fail to capture the complex relationships between concrete composition and thermal properties. Machine learning offers a better way to solve this problem by using data-driven models. This study used 192 datasets that included information on various concrete mixtures and their properties. Both simple and advanced machine learning models were tested to predict CTE. Simple methods, like Multiple Linear Regression (MLR) and Support Vector Regression (SVR), provided basic predictions. Advanced methods, such as AdaBoost, XGBoost, Bagging, and Random Forest, were used to handle the complex and non-linear nature of the data. These models were evaluated using methods like cross-validation and feature importance analysis to check their accuracy and reliability. When comparing the results, the study found that simpler models like MLR could only give limited insights. In contrast, advanced models, especially Random Forest, performed much better. The Random Forest model was highly accurate, identifying key factors affecting CTE and correctly predicting patterns in the data. It also reduced the number of experiments needed for accurate predictions, saving time and resources. This research shows how machine learning can be used to better predict and understand concrete properties. It highlights Random Forest as a powerful tool for accurately modeling CTE and other mechanical properties of concrete while reducing costs and improving efficiency

Keywords: Coefficient of Thermal Expansion, Concrete Performance, Machine Learning Models, Multiple Linear Regression, Support Vector Regression

1. Introduction

Concerning the environment, the widespread application of concrete in construction has a significant impact. The demand for concrete increases in parallel with the construction and maintenance of infrastructure and structures, thereby contributing to resource depletion and carbon emissions [1]– [3]. In light of this difficulty, using recycled aggregates in manufacturing concrete has surfaced as a potentially effective resolution. Produced from pulverized concrete from demolished structures, these aggregates provide an environmentally preferable substitute for conventional materials [4]– [6]. It is essential, nevertheless, to guarantee their mechanical properties, including strength and durability, to maintain structural integrity [7]– [9]. It is possible to forecast these properties, in addition to variables such as thermal coefficients, by employing advanced modeling methods such as machine learning. By comprehending how various factors affect the performance of recycled aggregates and analyzing massive datasets, it is possible to maximize their application in construction while minimizing their ecological footprint [10]– [12].

The significance of the Concrete Coefficient of Thermal Expansion (CTE) in pavement design is emphasized in the Mechanistic-Empirical Pavement Design Guide (MEPDG) published by the AASHTO, which is an organization that represents state highway and transportation officials [13], [14] It is of paramount importance in a multitude of facets, including joint faulting, slab cracking, and surface irregularity [15]. The

MEPDG of AASHTO specifies three tiers of inputs for the design of concrete CTE. Level-1 testing requires concrete CTE to be evaluated on-site or per project using the exact materials intended for the undertaking [16]– [18]. This method provides the greatest level of dependability but at the expense of increased financial expenditures. The mean CTE values of mixture constituents (coarse aggregate, fine aggregate, and cement paste), weighted by their respective volumetric proportions, are employed as Level-2 design input [19]– [21]. These values may be acquired from databases or determined experimentally. Level-3 design permits the utilization of standard concrete CTE, which is commonly acquired from national databases while taking into account particular coarse aggregate varieties. Although Levels 2 and 3 require less effort to implement, their reliability is intrinsically inferior to that of Level 1. Furthermore, there is a growing trend among transportation agencies to create their own concrete CTE databases, which are customized to project sites and utilize standard mix designs and materials [22], [23]. While these values might not exact Level-1 inputs, they are expected to provide more precise outcomes compared to the averages of Level-2 or Level-3. The emergence of local databases highlights the increasing demand for methodologies that improve the predictive capabilities of CTE data and facilitate concrete CTE forecasting [24], [25].

These methodologies possess the capacity to optimize database construction while enhancing the accuracy of CTE forecasts for particular projects that depend on said databases. Regardless of being in its early stages, research on predicting concrete CTE has attempted to utilize linear regression or mechanical models to achieve this objective. A study was conducted by Yang and Kim in which they employed analysis of variance (ANOVA) to assess the significance of variables related to concrete mix design [26]– [29]. Following this, a linear equation was derived using these variables to forecast CTE. Nevertheless, it is critical to mention that the current study emphasizes the significance of evaluating the predictive effectiveness of these models via cross-validation, an aspect that was neglected in the study by Yang and Kim [30]. To date, there has been a scarcity of research investigating the application of machine learning techniques for predicting concrete CTE or assessing models that go beyond linear regressions. On the other hand, neural networks have found widespread implementation in the prediction of compressive strength (CS) of concrete, regardless of a shortage of research utilizing machine learning techniques [31]– [34].

In addition, machine learning has been widely implemented in numerous fields of civil engineering , which presents an innovative neural network architecture that integrates attention mechanisms to improve the accuracy of predictions, , which examines the integration of convolutional neural networks (CNNs) and long short-term memory (LSTM) networks for time-series prediction of concrete strength, and , which investigates the application of ensemble neural networks, are among the notable papers in the domain of neural network-based concrete compressive strength prediction [35]–[37].

Furthermore, an examination of various neural network architectures for the prediction of concrete strength is undertaken in a comparative analysis. This analysis emphasizes the superior predictive accuracy and computational efficiency of particular architectures. Furthermore, to optimize the prediction of concrete strength, provides an exhaustive examination of feature selection techniques and neural networks. Machine learning methodologies have been applied to a variety of civil engineering tasks in a broader sense [38]– [42]. These tasks include the evaluation of bridge performance, inspection of structural health, and assessment of pavement condition [42]– [44]. Collectively, these inquiries highlight the capacity of machine learning methods to enhance decision-making and predictive capacities in a variety of civil engineering domains [45] – [47].

A range of artificial intelligence (AI) modeling methods, in addition to empirical models, have been utilized to facilitate continuous advancements, specifically in the design of plastic concrete [48], [49], where critical mechanical parameters include compressive strength (fc') and split tensile strength (fsts) [50]. Computational modeling methodologies present a viable substitute for the arduous and complex procedure of mixture optimization conducted in a laboratory setting [51] – [53]. By utilizing optimization techniques and objective functions derived from concrete constituents and their properties, these methods determine the optimal concrete mixtures. In the past, objective functions for linear and nonlinear models were formulated individually [54], [55]. However, the establishment of precise relationships between input parameters and concrete attributes has been a formidable challenge, owing to the substantial nonlinear correlations between these variables. As an outcome, scholars have resorted to employing machine learning (ML) methodologies to forecast the properties of concrete [56].

Previous research has applied a range of machine learning (ML) methods to predict concrete properties, such as elasticity modulus (fc'), fsts, and fc' [57]. The most frequently used techniques include genetic engineering programming (GEP) [3], support vector machines (SVM) [9], multilayer perceptron neural networks (MLPNN) [58], and ANN [59]. MLPNN, which falls under the category of ANN, provides a nonlinear modeling strategy that can capture intricate input-output relationships. During training, SVM models attempt to discover a global solution to reduce the structural risk of model complexity [60]. Classification problems are typically resolved using decision trees, in which each leaf node corresponds to a classification result and internal

nodes represent dataset features. GEP is a contemporary approach within the field of computational intelligence that distinguishes itself by employing sophisticated genetic algorithms and expression trees to articulate nonlinear associations [61]. Deep learning (DL) algorithms have robust architectures that enable more accurate predictions than traditional machine learning (ML) techniques. This is largely attributable to the massive amount of data that has been amassed in the past decade, which has propelled the adoption of DL [62]. Ensemble methodologies, including random forest (RF), have exhibited remarkable efficacy in addressing intricate problems with exceptional accuracy. As an illustration, the RF method was utilized to forecast the mechanical characteristics of high-strength concrete, exhibiting a remarkable coefficient of determination (R2) of 0.96, surpassing the performance of alternative models [63].

Scholarly investigations highlight the efficacy of ensemble learning techniques, such as Bagging and Boosting, in improving predictions of concrete properties, including compressive strength and plastic concrete (PC) durability. This study aims to predict the coefficient of thermal expansion (CTE) and other critical characteristics of concrete using machine learning models, including Multiple Linear Regression (MLR), Support Vector Regression (SVR), AdaBoost, XGBoost, Bagging, and Random Forest. Using a dataset of 192 concrete mixtures, the models were evaluated through cross-validation and feature importance analysis. Results show that advanced models, particularly Random Forest, outperform simpler models, offering accurate predictions while reducing experimental costs and providing insights into concrete's mechanical and thermal properties.

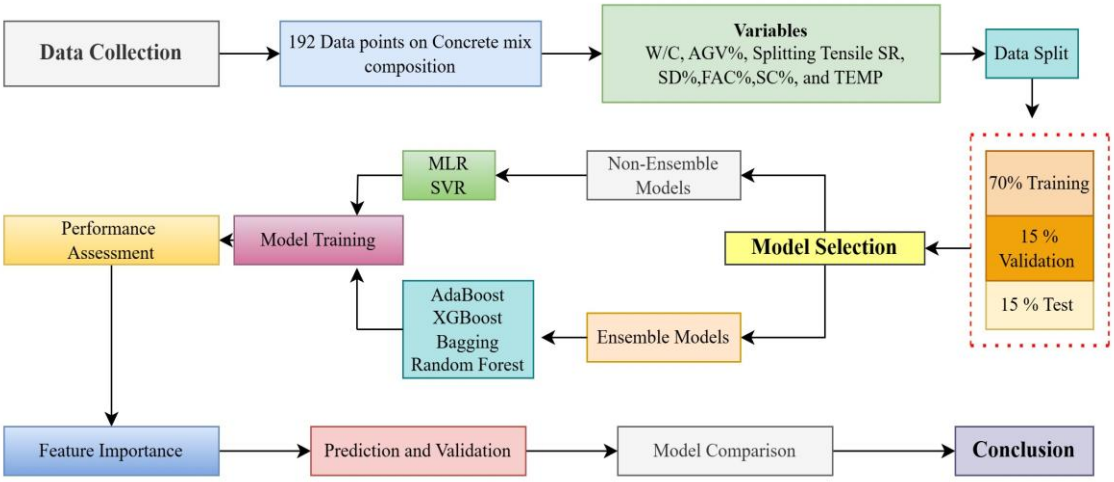

Fig 1. Block diagram of the research

2. Concrete Database Collection

Before performing machine learning calculations, it is critical to ensure that the datasets are thoroughly organized. The datasets comprise a wide array of properties that are essential for analysis, including characteristics of recycled aggregate and mix proportions [64]–[66], The dataset comprises 192 data points and offers comprehensive details on concrete mix composition, including variables such as water-cement ratio W/C, aggregate gradation void percentage AGV%, splitting tensile strength ratio SR, standard deviation percentage SD%, fine aggregate content percentage FAC%, and slump cone percentage SC% and temperature as shown in Table 1. It was subdivided into three subsets: Seventy percent of the data (134 points) constituted the training set for the development of machine learning models, fifteen percent (29 points) comprised the validation set for hyperparameter optimization and overfitting reduction, while the remaining fifteen percent (29 points) served as the test set to evaluate model performance on new data.

3. Methods

3.1. Multivariable linear regression

In this study, the linear correlation between multiple independent variables and a dependent variable was modeled using multivariate linear regression (MVLR) [67]. The primary objective was to determine the coefficients that most accurately represent the linear equation describing the relationship between these variables. Multiple independent variables were used to predict the values of the dependent variable via MVLR. Within the framework of the study, random forests, which are less complex and nonlinear, were compared to MVLR as a baseline. The regression algorithm provided by the Python Scikit-Learn toolbox was used to implement these techniques, and the sum of squared differences was minimized by calculating and adjusting the coefficients.

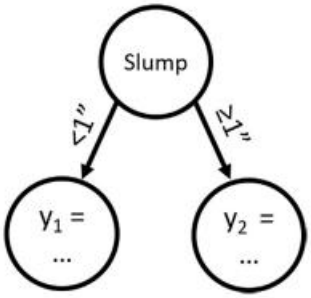

Fig. 2. A basic DT, where the slump of the concrete mix is being used to predict a concrete property [49]

3.2. Random Forest

Decision trees are the fundamental tool for utilizing data in decision-making processes and serve as the foundation for random forest methods [68]. Each tree in the forest provides its forecast, and the overall forecast is derived from the aggregated results of all tree forecasts. The tree employs these junctions and arcs to select the most appropriate regression equation for predicting the concrete property based on the slump value [67]. Figure 1 depicts a tree where each junction represents a question about the dataset and each arc represents a possible response to that question. In the provided example, the base junction determines the next passage step by analyzing the concrete slump of the dataset [69]. If the slump of a concrete mixture is less than 100, the data value would follow Figure 1's left branch. If the decline is greater than 100, however, the data value will follow the right branch. The accompanying image depicts the structure of a decision tree developed for this study [69].

The Scikit-Learn software toolkit was used to implement both the random forest and decision tree techniques in this study [70]. During the phase of model training, the structure of the decision tree is determined based on the features. [71]. Minor deviations from the default values for these variables had negligible effects on the predictive performance of the model in this study [72]. It is essential to note that decision trees can sometimes memorize the training data, resulting in fore-castings that may not accurately reflect the actual data patterns [73]. Random forests reduce the impact of below-average individual trees by averaging the results of many trees. In random forests, additional hyperparameters include the number of trees in the forest and the use of scaling for tree sampling. In this study, Scikit-Learn was used to implement the random forest model. For the majority of hyperparameters, the default values were used, except the split criterion and the number of trees in the forest [74].

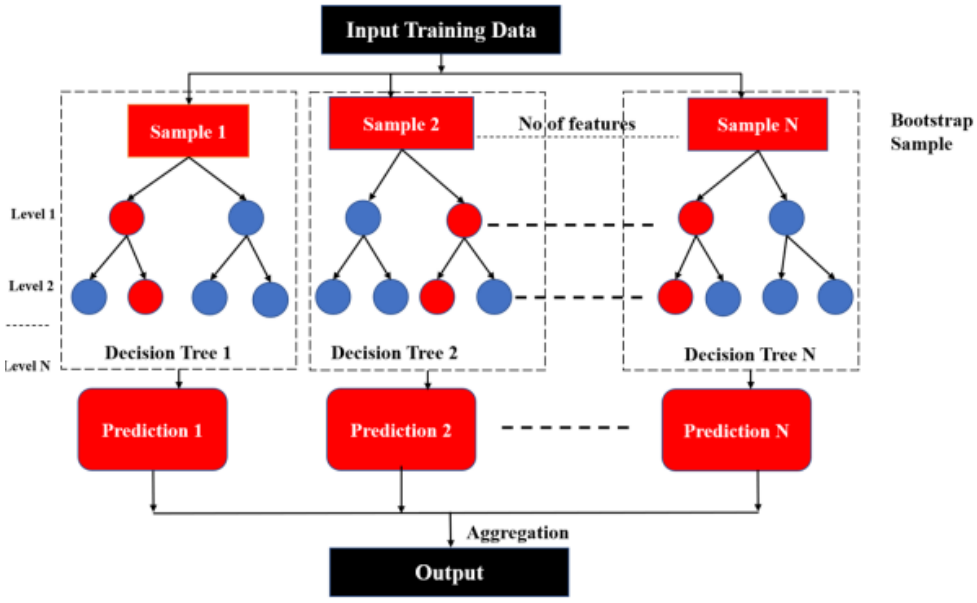

Fig 3. Random Forest working tree [75][76]

3.3. Machine learning model development

A mathematical model known as Artificial Neural Networks (ANN) is specifically engineered to emulate the structure and operation of biological systems [75]-[77]. Consequently, ANN demonstrates several characteristics comparable to those of biological brain systems [78]. These include self-learning and adaptation, computational parallelism, nonlinear mapping, and robustness/fault tolerance. The prospective applications of artificial neural networks' robust nonlinear processing capability continue to inspire researchers [79]. Out of the various structures considered, the backpropagation neural network is the most developed and extensively employed multilayer feedforward network architecture. This enables the network to acquire an accurate mapping between input and output, even in the absence of prior knowledge regarding particular mapping relationships. Fundamentally, Support Vector Machines utilize an inner product function-enabled nonlinear transformation to establish a correspondence between the input space and a higher-dimensional space. [80].

Out of the various structures considered, the backpropagation neural network is the most developed and extensively employed multilayer feedforward network architecture. This enables the network to acquire an accurate mapping between input and output, even in the absence of prior knowledge regarding particular mapping relationships. To optimize and achieve the desired solution, PSO utilizes information exchange among particles; it is an intelligent optimization algorithm based on populations.

3.4. Performance Assessment

The training set, the validation set, and the testing set were constituted of these subsets. Following this, the validation set was utilized to assess the convergence and development of the model throughout the training process, enabling the optimization of hyperparameters to enhance the overall effectiveness of the model. A comparison analysis was performed in a study by Asteris et al. [82] to evaluate the predictive capabilities of four machine learning models and identify any discrepancies. The purpose of these metrics was to quantify the degree of concordance between the predicted values and the empirical observations. The initial evaluation of the first two parameters was conducted based on a single sample examination, whereas the subsequent assessment of the remaining five parameters was determined by the degree of algorithmic fitting across multiple subsets. Furthermore, the integration of these statistical parameters into a composite performance index was demonstrated to be possible, as demonstrated in the study conducted by Cook et al [83].

3.5. Partial Dependence plot

Interpretability of a model refers to the capacity to comprehend the fundamental operations and results of the model [84]. Machine learning is frequently regarded as an opaque operation, which requires the creation of interpretable methods for visually examining the process on a local and global scale. An example of a widely used interpretable technique on a global scale is the Partial Dependence Plot, which prioritizes the importance of features according to operational outcomes and illustrates the marginal impact of one or two features on the model's predictions [85].

$$f_{xS}(xS) = E_{xC}[f_{xS}(xS, xC)] = \int f_{xS}(xS, xC)dP(xC) \quad (9)$$

To perform partial dependency analysis, the feature distribution within the set is combined with the output of the machine learning model. The function resulting from this amalgamation establishes the correlation between the features present in set C and the prediction outcomes that are obtained. Using this marginalization process, it is possible to derive a function that is exclusively dependent on the features present in set S, thus encapsulating their interaction [86].

3.6. Shapely additive explanations

When implemented in the domain of predictive modeling, the Shapley value can be used to evaluate the combined influence of every individual prediction technique on the overall accuracy of the predictions. It functions as an indicator of the level of effort put forth by each participant in a cooperative game, and the results are frequently perceived as equitable and rational by all participants [87]. When implemented in the domain of predictive modeling, the Shapley value can be used to evaluate the combined influence of every individual prediction technique on the overall accuracy of the predictions [88].

$$g(z') = \varphi_0 + \sum_{j=1}^{M} \varphi_j z'_j \quad (10)$$

The value assigned to each feature provides clarification regarding its specific impact on the overall outcomes of the predictions, thus emphasizing the discrepancy between the mean prediction made by the model and the actual prediction from the database. A SHAP value of a positive sign signifies that the feature has a positive influence on the predicted value. Conversely, a value of a negative sign signifies a diminishing effect on the contribution [89].

3.7. Cross Validation

The passage shows how cross-validation can be used to assess the performance of a machine-learning model using evaluation data that was not used for training. Specifically, the study employed k-fold cross-validation, in which the data is divided at random into k subsets or folds [90]. These folds consist of (k-1) training folds, which are used to train the model, and 1 validation fold, which is used to evaluate the accuracy of the model. RMSE (Root Mean Square Error) and R2 (Coefficient of Determination) values are calculated for each validation cycle. The average of these values across all rounds of cross-validation provides a single measure of RMSE and R2 for the entire process. On unseen validation data, the CV-RMSE metric is used to evaluate the performance of the machine learning model. Using cross-validation to divide the data into training and validation sets yields the CV-RMSE, which represents the squared mean error. Noting that the CV-RMSE may differ from the RMSE obtained from a direct fit is essential, as overfitting can occur when a model is trained on a small dataset. When a model closely matches the training data, resulting in low training error but high validation error, this is known as overfitting. In certain instances, the RMSE of the direct fit may be less than the CV-RMSE [91].

3.8. Prediction of CTE through level 1-3 model approach

In typical roadway design, three separate levels of calculation are required to forecast the thermal expansion coefficient (CTE) for a particular concrete formulation. At Level 1, the concrete mixture's CTE is determined empirically using a standard test procedure by AASHTO T336 specifications. In Level 2 calculations, the CTE values of the coarse aggregate's main mineral component and the cement paste are linearly blended, based on their respective volume proportions. In addition, the Level 3 prediction of concrete CTE is based on historical data collected on the CTE values of the primary minerals present in the coarse

aggregate [1]. The cement paste CTE ranges between 18 and 20 10(-6) C(-1) when the water-to-binder or cement ratio is held constant at 0.4 for the dataset utilized in this investigation. Quartz, dolomite, and basalt are the primary mineral components of the coarse aggregate, with CTE values of 9.3 10(-6) C(-1), 8.9 10(-6) C(-1), and 7.8 10(-6) C(-1), respectively. The linear combination used for Level 2 determination is CTE_mix = a_Paste * CTE_paste + a_CA * CTE_CA + a_FA * CTE_FA, where a_Paste is the volume percentage of cement or binder grout, a_CA is the volume percentage of the primary mineral composition in the coarse aggregate, and a_FA is the volume percentage of the main minerals in the fine aggregate [1], [15].

3.9. Background on ML algorithms

Mathematical models utilized in machine learning (ML) incorporate principles that originate from the fundamental biological structure of the human brain. These models demonstrate the ability to undergo training, computation, and adaptation in response to the datasets that are supplied to them. Thus, they are capable of producing forecasts using datasets that have never been analyzed before [34], [58], [92]. In general, a machine learning model is composed of input and output layers, which are designated with the names of the input and output variables they signify, respectively. Following this, the validation set was utilized to assess the convergence and development of the model throughout the training process, enabling the optimization of hyperparameters to enhance the overall effectiveness of the model. In general, a machine learning model is composed of input and output layers, which are designated with the names of the input and output variables they signify, respectively. To formulate predictions for novel datasets, the model implements a multitude of neurons functioning as processing units a network of neurons that are interconnected and process information [93].

In the course of this inquiry, the strength of concrete was characterized by utilizing both individual machine learning (ML) models and ensemble models that integrated feeble learners. This subject will be further expounded upon in subsequent sections of this research paper.

3.9.1. Non- ensemble models

The current investigation employed non-ensemble machine learning models, namely Support Vector Regression (SVR) and Multiple Linear Regression (MLR), for analysis and experimentation [94].

3.9.1.1. Multiple linear regressions (MLR)

Regression models function as mathematical tools utilized to determine the degree of correlation and interrelationship between input and output variables. When estimating linear regression (LR) models, the utilization of least squares methodologies is commonplace. On the other hand, as described by Neter et al.[95], Multiple Linear Regression (MLR) examines the correlation between a set of independent variables and the dependent response, producing results that suggest a linear relationship. Equation (1) represents the formulation of MLR.

$$Y = a0 + \sum m\, j=1\ aj\ X \tag{1}$$

The mathematical expression for the output Y of a model is denoted by equation (1), where Y represents the resultant output, Xj represents the input variables used in the algorithms, and a0, a1, a2,..., represents the partial regression coefficients that are associated with each input variable. The equation illustrates the weighted contribution of each input variable to the output, as represented by its corresponding coefficient, to summarize the linear relationship between the two [96].

3.9.1.2. SVR

When applied to regression tasks, the support vector machine (SVM) model is recognized as support vector regression (SVR), a prominent data mining technique. SVR, which was first introduced in 1995 by Cortes and Vapnik [97] has since been extensively implemented in tasks involving classification, prediction, and regression. SVR, functioning as a supervised learning algorithm, demonstrates exceptional proficiency in input-output analysis, thereby efficiently tackling the obstacles that are inherent in nonlinear regression, limited datasets, and high-dimensional input spaces. Using nonlinear transformations, the input space is transformed into a space of a higher dimension as part of this methodological procedure. Following that, nonlinear activation operations are implemented on this expanded space, which enhances the accuracy of the matching procedure due to the unique characteristics of the input parameters derived from the original dataset [98]. The linear function, denoted as f(x,w), can be approximated in this transformed space using the formula specified in Equation (2).

$$f(x, w) = \sum_{i=1}^{n} w_i g_i(x) + b \quad (2)$$

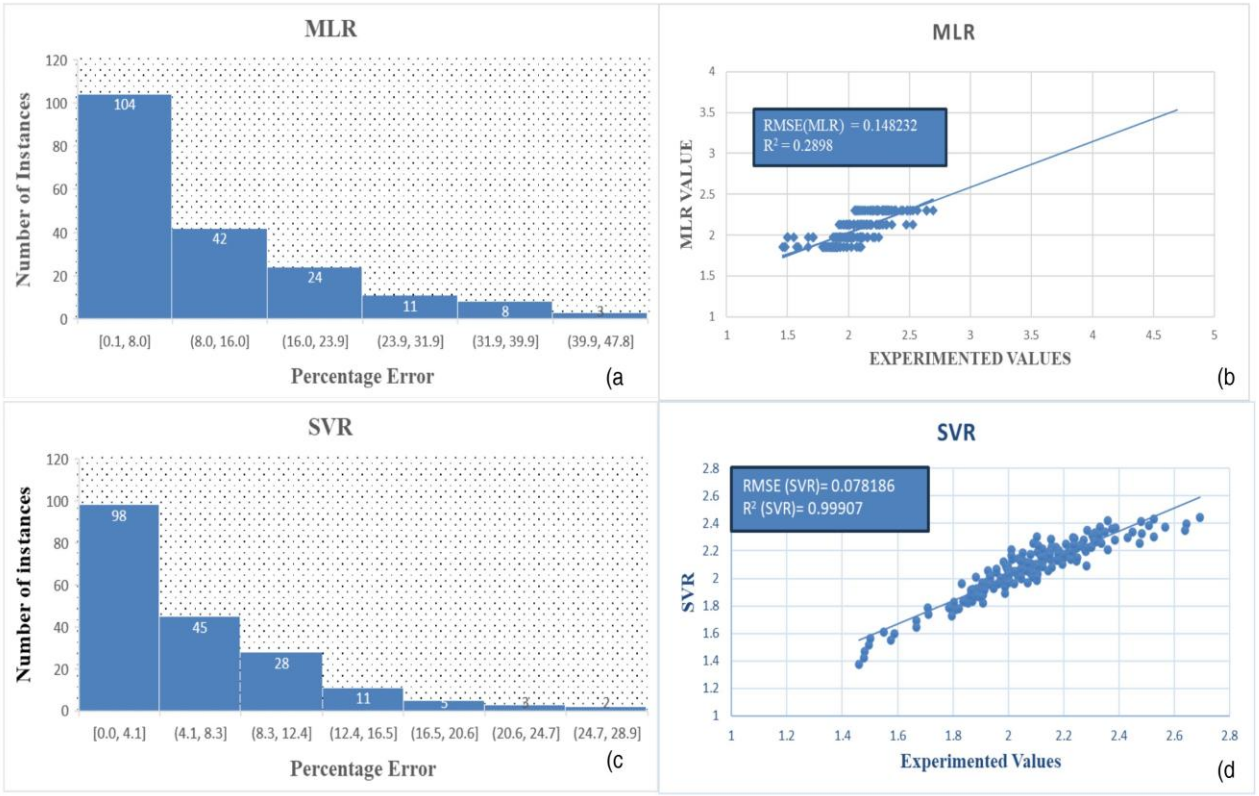

Fig.4 a)Number of Percentage Errors in MLR Method, b)MLR Scatter Chart between Predicted Values and Experimented Values, c)Number of Percentage Errors in SVR Method, d)Scatter Chart Between Predicted Values and Experimented Values

3.9.2. Ensemble models

As defined by Dietterich [99], Ensemble models are a category of learning algorithms in which the weighted aggregation of the predictions from a collection of classifiers is utilized to combine the outcomes. At the outset, Bayesian averaging functioned as the fundamental ensemble technique, establishing the path for succeeding algorithmic approaches in the field of machine learning, such as boosting, error-correcting output coding, and bagging. Four ensemble machine-learning models were utilized in the present inquiry: AdaBoost, XGBoost, Bagging, and Random Forest.

The binary dataset illustrated in Figure 4 consists of circles colored black and white. Its primary function is to train a model that can precisely classify the data points. At the outset, Model 1, functioning as a weak learner, is trained on the dataset, albeit with specific data points exhibiting classification errors. Following this, boosting is utilized to improve the performance of Model 1 through the incorporation of Model 2 and, later, Model 3, to rectify the deficiencies that were identified in the antecedent models. Nevertheless, an overabundance of model ensemble additions could result in overfitting; therefore, a meticulous assessment of the hyperparameter governing the number of models included in the boosting procedure is required.

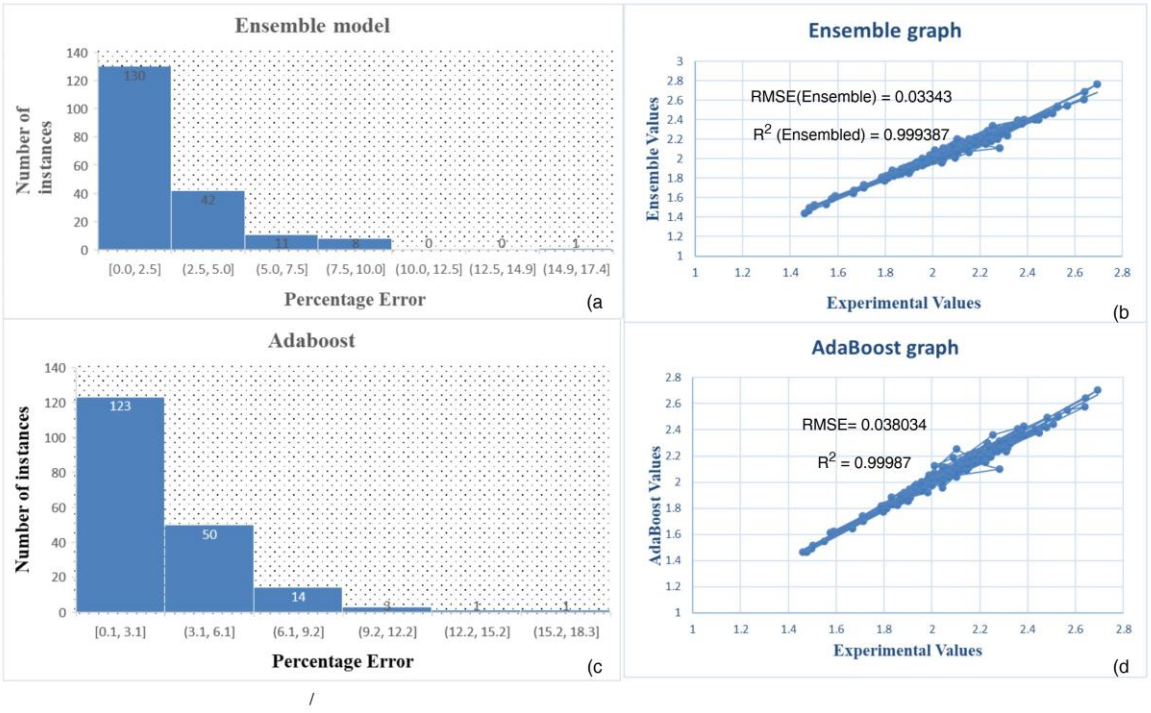

Fig.5 a)Number of Percentage Errors in Ensemble Model, b)Fig 8. Scatter Chart Between Predicted and Experimental Values, c) Number of Percentage Errors in AdaBoost, d)Scatter Chart between Predicted and Experimental Values

3.9.2.1. AdaBoost

Boosting, first introduced by Schapire and Freund in 1990 [100], is a machine-learning algorithm that is widely recognized and utilized .Expanding upon this seminal research, Freund and Schapire (1995) proposed AdaBoost, also referred to as adaptive boosting. It aimed to combine weak classifiers obtained during the training process into a robust collective. To achieve this, the training set was iteratively modified to optimize the training process for the weak classifier Chengsheng et al., [101].

3.9.2.2. Random Forest regression:

The Random Forest (RF) model is highly regarded for its exceptional performance when it comes to solving problems related to both classification and regression. Functioning within the domain of supervised learning, RF accomplishes regression tasks utilizing an ensemble learning approach. The structure is comprised of a collection of decision tree models organized within the bagging framework. Each tree node is represented by a leaf node, which does not contain any offspring nodes. Every node in this algorithmic framework is determined by a criterion that is constructed using the input features that are supplied. Using the established criteria, the algorithm assesses the veracity of statements as they advance from the root to the leaf node. The RF algorithm was developed by L. Breiman in 2001 [102].To improve its effectiveness, Breiman incorporated the principles of bootstrap aggregation [103] and random feature selection [104].

3.9.2.3. XGBoost

XGBoost is a renowned and extensively implemented scalable machine learning algorithm that was developed with tree boosting in mind. The scalability of the system is supported by various crucial characteristics, including the ability to utilize parallel and distributed computation, an advanced tree-learning algorithm that is adept at processing sparse data, and algorithmic enhancements, as explained by Chen and Guestrin [105]. In the context of the present investigation, XGBoost was selected based on its plethora of pertinent characteristics. The proposed enhancements encompass provisions for regularization, utilization of second-order gradients to expedite convergence, techniques for sparsity-aware split-finding, integration of stochastic gradient descent to introduce diversity and alleviate overfitting, and mechanisms for shrinkage aimed

at further mitigating overfitting. Additionally, XGBoost is equipped with a suite of system-level functionalities, such as cache optimization and parallelization, ensuring scalability and efficiency across diverse computational environments.

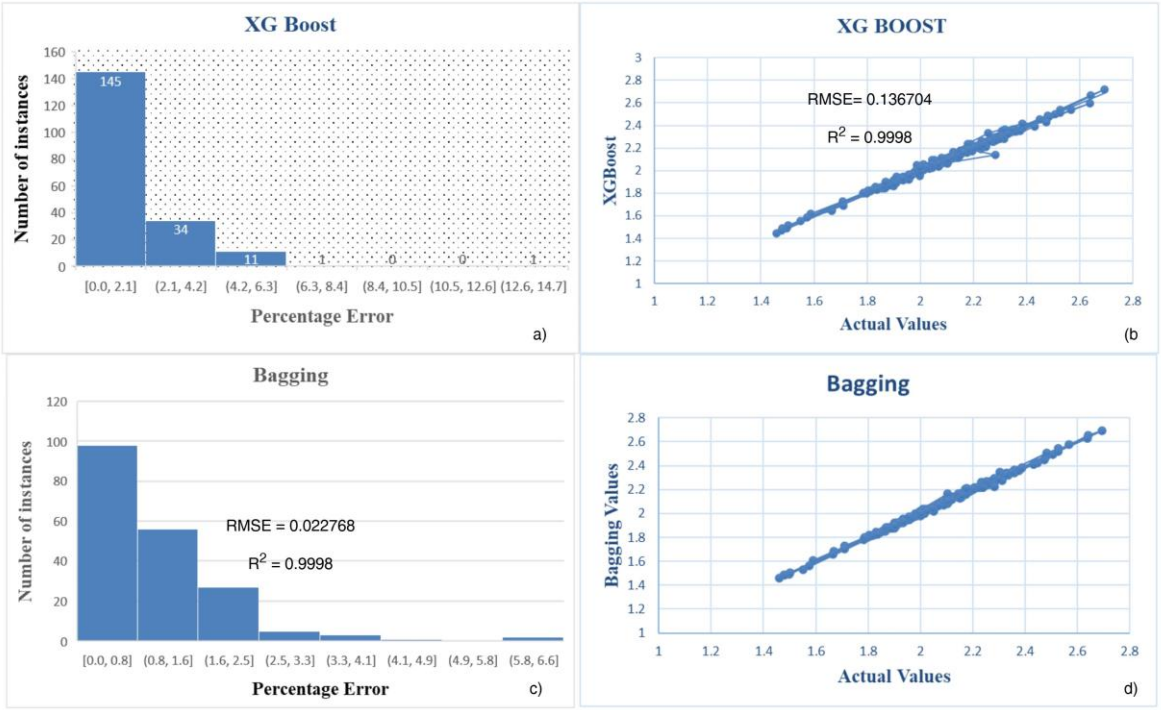

Fig.6 a)Number of Percentage Errors in XGBoost, b)Scatter Chart between Predicted and Experimental Values, c)Number of Percentage Errors in Bagging, d)Scatter Chart between Predicted and Experimental Values

3.9.2.4. Bagging

By employing bootstrap sampling and sampling with replacement, the bagging technique is capable of producing multiple datasets from a single input. The iterative procedure results in unique samples, which are then employed to train a multitude of models. Regression tasks involve the aggregation of the predictions of multiple models to derive the final model, whereas classification tasks involve the determination of the mode. Bootstrap sampling is prone to producing redundant data instances, which are estimated to comprise 63% of the unique data. Bootstrap Aggregating, also referred to as Bagging, was first proposed by Breiman in 1996 and is currently recognized as one of the most extensively utilized ensemble methodologies. By arbitrarily selecting subsets from the initial dataset and aggregating or voting the predictions of each regressor, this method determines the final forecast [106].

Meta-estimators of this nature effectively reduce the variability introduced by black-box estimation methods through the integration of a randomization process into their formulation of predictions. Furthermore, Bagging (BGR) exhibits improved performance, especially when faced with intricate algorithms such as completely developed decision trees, by effectively mitigating concerns related to overfitting [107]. The methodology employed in this study is illustrated in Figure 4, which provides a schematic diagram of the specific procedural steps involved.

- For this research undertaking, a comprehensive collection of 192 datasets was obtained from a well-established research database. Following that, a thorough examination of the statistical properties of the datasets was undertaken through exhaustive data analysis, which utilized techniques like correlation matrix analysis and histograms.

- The analysis utilized both the ensemble and non-ensemble machine learning (ML) models, which are

illustrated in Figure 4. The results obtained from this analysis were subsequently evaluated utilizing a range of performance metrics.

- The dataset was subsequently partitioned into subsets for training and testing, after which k-fold cross-validation was implemented. In addition, performance indices were employed to evaluate the model's accuracy, and a sensitivity analysis was carried out.

- In the end, parametric and sensitivity analyses were performed to further examine the properties of the data and determine the contribution or significance of each input characteristic to the corresponding output.

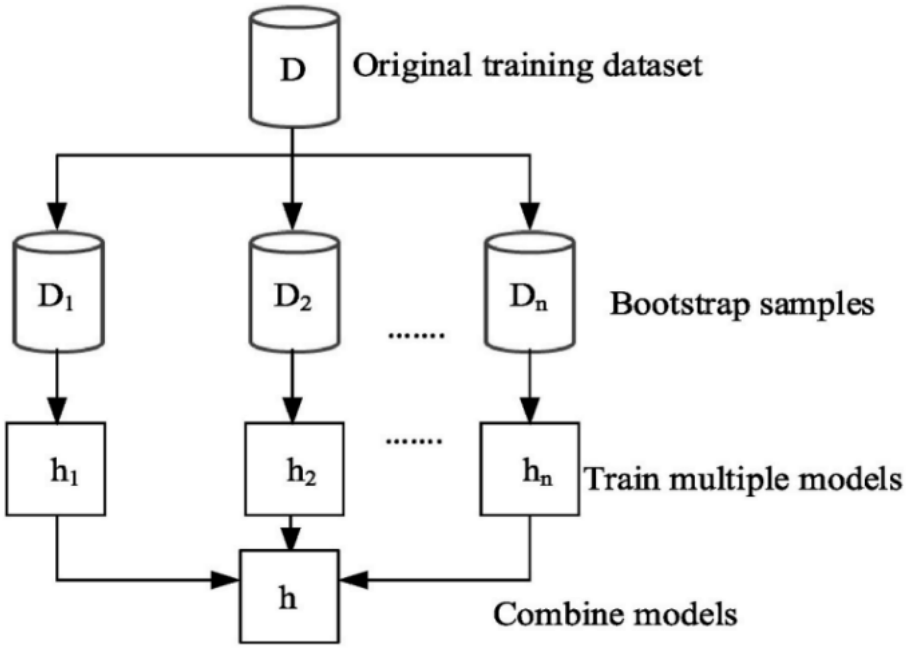

Fig 7. Generic Bragging Framework

Table 1 Database of Six Inputs Used In The MVLR and Random Forest Models and It Output

S.No	W/C	AGV%	SR	SD%	FAC%	SC%	TEMP	EXP_TCC	MLR	SVR	Ensemble	adaboost	RF	Xgboost	Bagging
1	40	30	35	0	0	0	-10	1.5	1.98	1.51	1.51	1.49	2.01	1.49	1.49
2	40	60	35	100	0	0	20	2.11	2.13	2.19	2.07	2.09	2.22	2.1	2.1
3	30	60	35	0	0	0	-30	2.25	1.98	2.13	2.25	2.25	2.08	2.23	2.25
4	40	60	35	75	0	0	20	1.98	2.13	2.12	1.98	2.02	2.12	1.99	1.99
5	40	60	35	100	0	0	-30	2.36	2.3	2.42	2.4	2.41	2.45	2.37	2.34
6	30	60	35	50	0	0	20	2.08	2.13	2.04	2.06	2.05	2.03	2.07	2.07
7	30	60	35	75	0	0	-20	2.23	2.3	2.29	2.29	2.3	2.31	2.23	2.27
8	30	60	35	0	0	0	20	1.99	1.86	1.89	2.01	2.02	1.85	2.01	1.98
9	30	60	35	25	0	0	-30	2.28	2.3	2.2	2.31	2.28	2.18	2.3	2.28
10	40	30	35	0	0	0	20	1.46	1.86	1.38	1.44	1.46	1.87	1.45	1.46

11	40	60	30	0	0	0	-10	2.11	1.98	2.1	2.13	2.13	2	2.11	2.11
12	40	50	35	0	0	0	20	1.8	1.86	1.73	1.78	1.77	1.85	1.8	1.8
13	40	40	35	0	0	0	10	1.59	1.86	1.6	1.62	1.62	1.91	1.62	1.61
14	40	60	40	0	0	0	-30	2.11	1.98	2.08	2.09	2.08	2.07	2.1	2.1
15	40	50	35	0	0	0	0	1.86	1.86	1.82	1.84	1.82	1.95	1.84	1.85
16	40	60	45	0	0	0	-20	1.98	1.98	1.97	2	1.99	2.02	1.98	1.97
17	30	60	35	75	0	0	20	2.08	2.13	2.11	2.06	2.08	2.12	2.09	2.08
18	30	60	35	100	0	0	-30	2.48	2.3	2.41	2.46	2.42	2.45	2.49	2.48
19	40	60	35	0	0	0	-30	2.15	1.98	2.14	2.07	2.11	2.08	2.18	2.16
20	40	60	35	0	0	0	-20	2.09	1.98	2.09	2.01	2.05	2.04	2.1	2.1
21	40	60	40	0	0	0	-10	2	1.98	1.98	1.98	1.97	1.98	1.99	1.98
22	40	60	40	0	0	0	0	1.91	1.86	1.94	1.93	1.92	1.93	1.95	1.92
23	60	60	35	0	0	0	20	1.89	1.86	1.92	1.89	1.91	1.85	1.88	1.89
24	50	60	35	0	0	0	20	1.87	1.86	1.91	1.86	1.86	1.85	1.85	1.85
25	50	60	35	25	0	0	-30	2.12	2.3	2.22	2.18	2.16	2.18	2.12	2.13
26	40	60	35	0	0	20	-20	2	1.98	2.01	1.99	1.99	2.02	1.99	1.99
27	40	60	35	0	30	0	20	1.78	1.86	1.78	1.79	1.8	1.82	1.8	1.78
28	40	60	35	0	20	0	20	1.8	1.86	1.82	1.82	1.82	1.83	1.83	1.82
29	60	60	35	100	0	0	20	2.36	2.13	2.21	2.37	2.35	2.22	2.34	2.36
30	40	60	35	0	10	0	-30	2.09	1.98	2.1	2.1	2.09	2.08	2.09	2.09
31	60	60	35	75	0	0	10	2.23	2.13	2.19	2.27	2.29	2.17	2.26	2.24
32	60	60	35	75	0	0	20	2.23	2.13	2.14	2.2	2.23	2.12	2.19	2.21
33	60	60	35	100	0	0	-30	2.69	2.3	2.44	2.77	2.7	2.46	2.72	2.69
34	60	60	35	100	0	0	-20	2.64	2.3	2.4	2.69	2.64	2.41	2.67	2.65
35	60	60	35	50	0	0	-20	2.33	2.3	2.25	2.34	2.34	2.22	2.34	2.32
36	60	60	35	50	0	0	-10	2.28	2.3	2.21	2.27	2.29	2.17	2.29	2.27
37	60	60	35	50	0	0	0	2.19	2.13	2.16	2.22	2.23	2.13	2.24	2.19
38	60	60	35	50	0	0	10	2.12	2.13	2.11	2.16	2.17	2.08	2.16	2.12
39	50	60	35	25	0	0	-20	2.08	2.3	2.17	2.13	2.11	2.13	2.11	2.09
40	50	60	35	25	0	0	-10	2.05	2.3	2.12	2.08	2.05	2.08	2.09	2.04
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
183	50	60	35	50	0	0	-20	2.24	2.3	2.24	2.21	2.21	2.22	2.23	2.22
184	50	60	35	100	0	0	-30	2.53	2.3	2.43	2.54	2.51	2.46	2.54	2.52
185	60	60	35	25	0	0	-30	2.3	2.3	2.23	2.28	2.26	2.18	2.28	2.29
186	60	60	35	75	0	0	-30	2.57	2.3	2.37	2.55	2.55	2.36	2.54	2.57
187	60	60	35	100	0	0	-10	2.64	2.3	2.35	2.61	2.57	2.36	2.6	2.63
188	40	60	35	0	20	0	0	1.87	1.86	1.92	1.9	1.9	1.92	1.9	1.88
189	40	60	35	0	0	10	-30	2.1	1.98	2.09	2.08	2.09	2.08	2.1	2.11
190	40	60	35	0	0	10	-20	2.02	1.98	2.05	2.03	2.04	2.03	2.03	2.02
191	40	60	35	0	0	20	0	1.91	1.86	1.91	1.91	1.92	1.92	1.91	1.92
192	40	60	35	0	0	20	10	1.9	1.86	1.87	1.89	1.89	1.88	1.89	1.9

4. The comparison of CTE models' outcomes and discussion

Meta- In data analysis, when the dependent variable's scale (in this case, CTE) is large and the range of values is broad, it is customary to normalize the RMSE for improved comparison and interpretation of the error metric. Normalization involves dividing the RMSE by the standard deviation of the CTE data, which converts the error metric from an absolute value to one that is relative to the variation in the data. This normalization procedure enables more accurate model comparisons and facilitates comprehension of the significance of the findings. In this investigation, the normalization factor was the standard deviation of the CTE data, which was r = 0.45 * 10(-6) C(-1).

Table 2 Performance Metrics of Models

Models	**CV-RMSE** ($*10^{-6}$ C^{-1})	**RMSE/δ** (δ =0.45$*10^{-6}$ C^{-1})	**R^2**
MLR	0.148232	0.329404444	0.28984
SVR	0.078186	0.173746667	0.9991
Ensemble model	0.03343	0.074288889	0.99939
Adaboost	0.038034	0.08452	0.99989
XG Boost	0.136704	0.303786667	0.1791
Bagging	0.022768	0.050595556	0.99988
RF Models	0.014098	0.031328889	0.99987

4.1. Level-2 and 3 approximations

The outcomes of the study indicate that the level-2 and level-3 models perform poorly in predicting the precise CTE values for the given dataset. Although the level-2 model allows for greater variability than the level-3 approximation, it generates a wider range of predicted values, which reduces its predictive ability. In contrast, the predictions generated by the level-3 model closely align with three distinct regression lines.

However, it is essential to note that both models produce negative R2 values, indicating a poor fit and that the RMSE values for both models exceed the variability of the observed CTE values. This suggests that the predictive ability of these models is limited, and it is not possible to accurately predict how CTE would vary in response to changes in the concrete mix using the given data set. Therefore, using the dataset's mean value as a predictor would be more accurate than relying on either level model.

4.2. Multivariate linear regression

Across all evaluation metrics, the MVLR model outperformed both the level-2 and level-3 forecasts. Even though the average RMSE/r values were above 1, the average R2 values for the cross-validation procedure with 2 and 10 folds were negative, indicating inadequate model performance. This suggests that although the level-2 and level-3 methods possess comparatively stronger predictive capabilities overall, the MVLR model surpasses both approaches regarding predictive precision. Notably, numerous data points were forecasted incorrectly across all tests, indicating that the MVLR model did not adequately represent certain non-linear properties of concrete CTE for these mixtures. According to the evidence, it is necessary to employ a nonlinear regression model, such as a random forest, to more accurately represent the complexity of the data.

4.3. Random forest

The results of this study indicate that the model can effectively forecast new data, particularly for comparable systems. In addition to making CTE predictions, random forests permit feature relevance analysis. Consistent with Yang and Kim's [108] research, the principal mineral content of coarse and fine aggregate was identified as a crucial factor in predicting concrete CTE. As shown in Table 2, the random forest model outmatches simpler CTE estimation techniques such as level-1, level-2, and MVLR. Only the random forest model is capable of predicting patterns in the CTE data archive, according to the study. This emphasizes the significance of employing nonlinear modeling techniques, which are prevalent in machine learning, for accurate CTE modeling. By employing the random forest model instead of the simpler MVLR model, this study significantly enhanced the accuracy of its predictions. With RMSE/r and R2 values of 0.78 and 0.39 for 2- and 10-fold cross-validation, respectively, the model can predict CTE patterns. Variations in R2 and RMSE/r values indicate that the model's efficacy is bolstered by the significant predictive power across all folds. The study found that coarse and fine aggregate varieties were the most important predictors of concrete CTE, and the random forest model facilitated the evaluation of feature significance. The findings demonstrate the importance of utilizing non-linear regression techniques, such as random forests, for accurate and efficient CTE modeling.

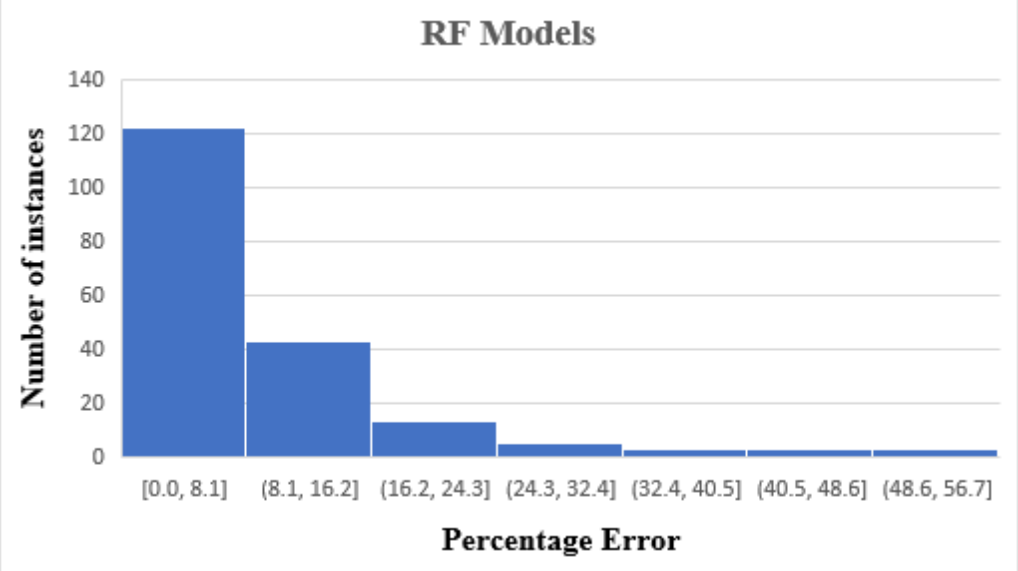

Fig 8. Number of Percentage Errors in Random Forest

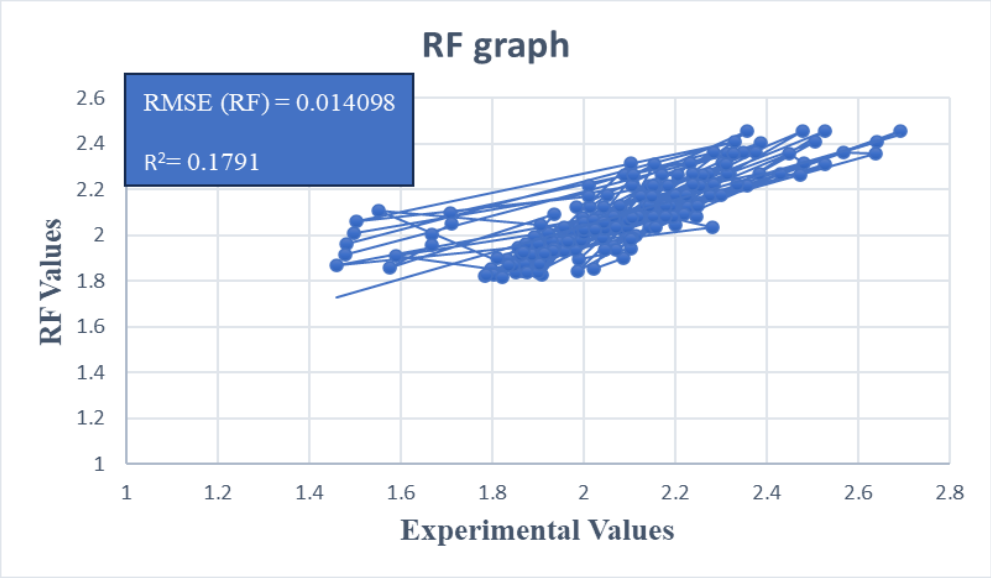

Fig 9. Scatter Chart between Predicted and Experimental Values

5. Lowering the number of CTE tests required through the use of machine learning

The relationship between the extent of the training dataset and the average RMSE/r of the testing set is illustrated in Figure 4. The results indicate that as more training data points are utilized, the average RMSE/r of the testing set decreases, indicating an advance in the model's predictive ability. However, beyond a certain threshold, increasing the number of training data points does not significantly improve prediction performance. Using approximately 150 training data points causes the average RMSE/r of the testing set to plateau in this study. This insightful knowledge can be used to devise a cost-effective experimental plan for attaining the desired predictive ability. By approximating the number of experiments required to achieve a particular level of predictive ability using the machine learning model, one can reduce the cost and duration of experimental testing, thereby saving both time and money. Before employing such a model, extensive validation and testing would be required to ensure that the results obtained in this study apply to various test matrices and mixture designs. Moreover, since the model's efficacy depends on the quality and representativeness of the training data, it is essential to ensure that the training data are accurate and inclusive of all potential scenarios.

6. Models for various concrete properties using machine learning

Remarkably, the random forest model used in this study was able to predict not only the coefficient of thermal expansion (CTE), but also other mechanical qualities of concrete, such as compressive strength (CS), elastic modulus, rupture modulus, indirect tensile strength, and Poisson's constant. Four distinct aging times were predicted: 7, 14, 28, and 90 days. The Additional Information comprises the specific results of these projections. The data generated by the random forest model exhibited a range of 27 MPa, with a root mean square error (RMSE) of 1.9 MPa, leading to a deviation-to-range ratio of 0.07. In contrast, prior research yielded a deviation-to-range ratio of 0.05 and RMSE values of 3.8 MPa for data with a range of 82 MPa. This finding demonstrates that the methodologies used in this study are accurate and consistent with those employed in prior research.

7. Conclusions

The random forest method of artificial intelligence determined the necessary number of CTE prediction experiments. The study utilized a portion of the available data to guide the model, reserving the remaining data for evaluating the model's predictive capabilities. Using 55 out of 110 data points (50% of the dataset) to guide the model resulted in a considerably lower RMSE/r for various mechanical properties of concrete. Notably, the model accurately estimated compressive strength (CS), Young's modulus, flexural strength, tensile fracture resistance, CTE, and Poisson's constant. In previous research, the random forest model obtained an RMSE of

1.9 MPa when predicting CS after 90 days of aging. In the future, the adoption of similar models of computational intelligence could lower the number of experiments necessary to characterize concrete characteristics for a particular assessment matrix. This research demonstrates how machine learning techniques can accurately predict the mechanical properties of concrete, such as tensile fracture strength, CTE, CS, Young's modulus, and Poisson's ratio. The random forest model was the most precise of the models evaluated in this study, able to detect trends in the data that other methods could not. In addition, the study demonstrated how the application of machine learning can lower the number of readings required for accurate concrete property predictions, thereby potentially reducing costs and augmenting efficiency when establishing concrete property databases. This study demonstrates the immense potential of machine learning in predicting and understanding the mechanical properties of concrete.

References

[1]. V. Nilsen, L. T. Pham, M. Hibbard, A. Klager, S. M. Cramer, and D. Morgan, "Prediction of concrete coefficient of thermal expansion and other properties using machine learning," Constr. Build. Mater., vol. 220, pp. 587–595, 2019.
[2]. V. Q. Tran, "Machine learning approach for investigating chloride diffusion coefficient of concrete containing supplementary cementitious materials," Constr. Build. Mater., vol. 328, p. 127103, 2022.
[3]. H. Gardezi et al., "Predictive modeling of rutting depth in modified asphalt mixes using gene-expression programming (GEP): A sustainable use of RAP, fly ash, and plastic waste," Constr. Build. Mater., vol. 443, p. 137809, 2024.
[4]. B. P. Koya, S. Aneja, R. Gupta, and C. Valeo, "Comparative analysis of different machine learning algorithms to predict mechanical properties of concrete," Mech. Adv. Mater. Struct., vol. 29, no. 25, pp. 4032–4043, 2022.
[5]. B. Ghorbani, A. Arulrajah, G. Narsilio, S. Horpibulsuk, and M. W. Bo, "Thermal and mechanical properties of demolition wastes in geothermal pavements by experimental and machine learning techniques," Constr. Build. Mater., vol. 280, p. 122499, 2021.
[6]. M. Usama et al., "Predictive modelling of compression strength of waste GP/FA blended expansive soils using multi-expression programming," Constr. Build. Mater., vol. 392, p. 131956, 2023.
[7]. X. Zhang, N. Zhao, and C. He, "The superior mechanical and physical properties of nanocarbon reinforced bulk composites achieved by architecture design–a review," Prog. Mater. Sci., vol. 113, p. 100672, 2020.
[8]. Y. Xu, B. Jamhiri, and S. A. Memon, "On the Recent Trends in Expansive Soil Stabilization Using Calcium-Based Stabilizer Materials (CSMs): A Comprehensive Review," 2020.
[9]. M. Iqbal, K. Elbaz, D. Zhang, L. Hu, and F. E. Jalal, "Prediction of residual tensile strength of glass fiber reinforced polymer bars in harsh alkaline concrete environment using fuzzy metaheuristic models," J. Ocean Eng. Sci., vol. 8, no. 5, pp. 546–558, 2023.
[10]. A. Ahmad, K. A. Ostrowski, M. Maślak, F. Farooq, I. Mehmood, and A. Nafees, "Comparative study of supervised machine learning algorithms for predicting the compressive strength of concrete at high temperature," Materials (Basel)., vol. 14, no. 15, p. 4222, 2021.
[11]. M. Rahmati and V. Toufigh, "Evaluation of geopolymer concrete at high temperatures: An experimental study using machine learning," J. Clean. Prod., vol. 372, p. 133608, 2022.
[12]. H. Gardezi, X. Li, and Y. Huang, "Machine learning-based landslide velocity prediction model: incorporating multi-expression programming and discrete element modeling," 2024.
[13]. G. Sabih, "Effects of Coefficient of Thermal Expansion on Unbonded Concrete Overlay Design and Performance." The University of New Mexico, 2019.
[14]. B. Keshavarzi, Prediction of thermal cracking in asphalt pavements using simplified viscoelastic continuum damage theory. North Carolina State University, 2019.
[15]. G. Sabih and R. A. Tarefder, "Predicting Long-Term Coefficient of Thermal Expansion of Paving Concrete," Transp. Res. Rec., vol. 2674, no. 9, pp. 792–798, 2020.
[16]. S. S. Kim, S. A. Durham, M. G. Chorzepa, D. Wing, and C. Banks, "Development of Concrete Material Property Database for Pavement ME Input," Georgia. Department of Transportation. Office of Performance-Based …, 2021.
[17]. A. A. Aguib, "Flexible pavement design AASHTO 1993 versus mechanistic-empirical pavement design," 2021.
[18]. S. Islam, A. Sufian, M. Hossain, N. Velasquez Jr, and R. Barrett, "Practical issues in implementation of mechanistic empirical design for concrete pavements," J. Transp. Eng. Part B Pavements, vol. 145, no. 3, p. 4019020, 2019.

[19]. M. Malik, S. K. Bhattacharyya, and S. V Barai, "Thermal and mechanical properties of concrete and its constituents at elevated temperatures: A review," Constr. Build. Mater., vol. 270, p. 121398, 2021.
[20]. A. Mateos, J. Harvey, J. Bolander, R. Wu, J. Paniagua, and F. Paniagua, "Field evaluation of the impact of environmental conditions on concrete moisture-related shrinkage and coefficient of thermal expansion," Constr. Build. Mater., vol. 225, pp. 348–357, 2019.
[21]. A.-L. Beaucour, P. Pliya, F. Faleschini, R. Njinwoua, C. Pellegrino, and A. Noumowé, "Influence of elevated temperature on properties of radiation shielding concrete with electric arc furnace slag as coarse aggregate," Constr. Build. Mater., vol. 256, p. 119385, 2020.
[22]. N. Makul, "Advanced smart concrete-A review of current progress, benefits and challenges," J. Clean. Prod., vol. 274, p. 122899, 2020.
[23]. B. Shafei, P. E. P. Taylor, and P. E. S. B. Nia, "Underwater Concrete Pours and Non-Segregating Concrete," 2024.
[24]. Z. Dong, W. Quan, X. Ma, X. Li, and J. Zhou, "Asymptotic homogenization of effective thermal-elastic properties of concrete considering its three-dimensional mesostructure," Comput. Struct., vol. 279, p. 106970, 2023.
[25]. T. Hielscher, S. Khalil, N. Virgona, and S. A. Hadigheh, "A neural network based digital twin model for the structural health monitoring of reinforced concrete bridges," in Structures, Elsevier, 2023, p. 105248.
[26]. W.-H. Chen et al., "A comprehensive review of thermoelectric generation optimization by statistical approach: Taguchi method, analysis of variance (ANOVA), and response surface methodology (RSM)," Renew. Sustain. Energy Rev., vol. 169, p. 112917, 2022.
[27]. Y. H. Lee et al., "Gait speed and handgrip strength as predictors of all-cause mortality and cardiovascular events in hemodialysis patients," BMC Nephrol., vol. 21, pp. 1–11, 2020.
[28]. H. M. Deberneh and I. Kim, "Prediction of type 2 diabetes based on machine learning algorithm," Int. J. Environ. Res. Public Health, vol. 18, no. 6, p. 3317, 2021.
[29]. M. Iqbal, D. Zhang, M. I. Khan, M. Zahid, and F. E. Jalal, "Effects of rebar size and volume fraction of glass fibers on tensile strength retention of GFRP rebars in alkaline environment via RSM and SHAP analyses," J. Mater. Civ. Eng., vol. 35, no. 9, p. 4023318, 2023.
[30]. J.-J. Zhu, M. Yang, and Z. J. Ren, "Machine learning in environmental research: common pitfalls and best practices," Environ. Sci. Technol., vol. 57, no. 46, pp. 17671–17689, 2023.
[31]. J. Zhang, Y. Cao, L. Xia, D. Zhang, W. Xu, and Y. Liu, "Intelligent prediction of the frost resistance of high-performance concrete: a machine learning method," J. Civ. Eng. Manag., vol. 29, no. 6, pp. 516–529, 2023.
[32]. F. E. Jalal, Y. Xu, M. Iqbal, M. F. Javed, and B. Jamhiri, "Predictive modeling of swell-strength of expansive soils using artificial intelligence approaches: ANN, ANFIS and GEP," J. Environ. Manage., vol. 289, no. December 2020, p. 112420, 2021, doi: 10.1016/j.jenvman.2021.112420.
[33]. F. E. Jalal, M. Iqbal, W. A. Khan, A. Jamal, K. Onyelowe, and Lekhraj, "ANN-based swarm intelligence for predicting expansive soil swell pressure and compression strength," Sci. Rep., vol. 14, no. 1, p. 14597, 2024.
[34]. F. E. Jalal, Y. Xu, M. Iqbal, B. Jamhiri, and M. F. Javed, "Predicting the compaction characteristics of expansive soils using two genetic programming-based algorithms," Transp. Geotech., vol. 30, no. June, p. 100608, 2021, doi: 10.1016/j.trgeo.2021.100608.
[35]. M. A. Jalal, L. Mihaylova, and R. K. Moore, "An End-to-End Deep Neural Network for Facial Emotion Classification," in 2019 22th International Conference on Information Fusion (FUSION), IEEE, 2019, pp. 1–7.
[36]. A. I. Jabbooree, L. M. Khanli, P. Salehpour, and S. Pourbahrami, "A novel facial expression recognition algorithm using geometry β–skeleton in fusion based on deep CNN," Image Vis. Comput., vol. 134, p. 104677, 2023.
[37]. D. O. Oyewola, E. G. Dada, S. Misra, and R. Damaševičius, "Detecting cassava mosaic disease using a deep residual convolutional neural network with distinct block processing," PeerJ Comput. Sci., vol. 7, p. e352, 2021.
[38]. A. T. G. Tapeh and M. Z. Naser, "Artificial intelligence, machine learning, and deep learning in structural engineering: a scientometrics review of trends and best practices," Arch. Comput. Methods Eng., vol. 30, no. 1, pp. 115–159, 2023.
[39]. S. R. Vadyala, S. N. Betgeri, J. C. Matthews, and E. Matthews, "A review of physics-based machine learning in civil engineering," Results Eng., vol. 13, p. 100316, 2022.
[40]. A. Kaveh, "Applications of artificial neural networks and machine learning in civil engineering," Stud. Comput. Intell., vol. 1168, p. 472, 2024.

[41]. K.-K. Phoon and W. Zhang, "Future of machine learning in geotechnics," Georisk Assess. Manag. Risk Eng. Syst. Geohazards, vol. 17, no. 1, pp. 7–22, 2023.
[42]. Y. Huang and J. Fu, "Review on application of artificial intelligence in civil engineering," Comput. Model. Eng. Sci., vol. 121, no. 3, pp. 845–875, 2019.
[43]. H. Sun, H. V Burton, and H. Huang, "Machine learning applications for building structural design and performance assessment: State-of-the-art review," J. Build. Eng., vol. 33, p. 101816, 2021.
[44]. Y. Xie, M. Ebad Sichani, J. E. Padgett, and R. DesRoches, "The promise of implementing machine learning in earthquake engineering: A state-of-the-art review," Earthq. Spectra, vol. 36, no. 4, pp. 1769–1801, 2020.
[45]. Z. Li et al., "Machine learning in concrete science: applications, challenges, and best practices," npj Comput. Mater., vol. 8, no. 1, p. 127, 2022.
[46]. Y. Pan and L. Zhang, "Roles of artificial intelligence in construction engineering and management: A critical review and future trends," Autom. Constr., vol. 122, p. 103517, Feb. 2021, doi: 10.1016/j.autcon.2020.103517.
[47]. A. Zabin, V. A. González, Y. Zou, and R. Amor, "Applications of machine learning to BIM: A systematic literature review," Adv. Eng. Informatics, vol. 51, p. 101474, 2022.
[48]. M. M. Moein et al., "Predictive models for concrete properties using machine learning and deep learning approaches: A review," J. Build. Eng., vol. 63, p. 105444, 2023.
[49]. B. Zou et al., "Artificial intelligence-based optimized models for predicting the slump and compressive strength of sustainable alkali-derived concrete," Constr. Build. Mater., vol. 409, p. 134092, 2023.
[50]. A. Nafees et al., "Predictive modeling of mechanical properties of silica fume-based green concrete using artificial intelligence approaches: MLPNN, ANFIS, and GEP," Materials (Basel)., vol. 14, no. 24, p. 7531, 2021.
[51]. S. Kar and J. Leszczynski, "Exploration of computational approaches to predict the toxicity of chemical mixtures," Toxics, vol. 7, no. 1, p. 15, 2019.
[52]. C. H. Mehta, R. Narayan, and U. Y. Nayak, "Computational modeling for formulation design," Drug Discov. Today, vol. 24, no. 3, pp. 781–788, 2019.
[53]. K. C. Onyelowe, F. E. Jalal, M. Iqbal, Z. U. Rehman, and K. Ibe, "Intelligent modeling of unconfined compressive strength (UCS) of hybrid cement-modified unsaturated soil with nanostructured quarry fines inclusion," Innov. Infrastruct. Solut., vol. 7, no. 1, p. 98, 2022.
[54]. B. A. Salami et al., "Estimating compressive strength of lightweight foamed concrete using neural, genetic and ensemble machine learning approaches," Cem. Concr. Compos., vol. 133, p. 104721, 2022.
[55]. M. F. Iqbal et al., "Sustainable utilization of foundry waste: Forecasting mechanical properties of foundry sand based concrete using multi-expression programming," Sci. Total Environ., vol. 780, p. 146524, 2021.
[56]. A. Nafees et al., "Forecasting the mechanical properties of plastic concrete employing experimental data using machine learning algorithms: DT, MLPNN, SVM, and RF," Polymers (Basel)., vol. 14, no. 8, p. 1583, 2022.
[57]. A. Nafees et al., "Plastic concrete mechanical properties prediction based on experimental data," Case Stud. Constr. Mater., vol. 18, p. e01831, 2023.
[58]. K. Khan et al., "Estimating flexural strength of FRP reinforced beam using artificial neural network and random forest prediction models," Polymers (Basel)., vol. 14, no. 11, p. 2270, 2022.
[59]. P. Verma, V. K. Awasthi, and S. K. Sahu, "Classification of coronary artery disease using multilayer perceptron neural network," Int. J. Appl. Evol. Comput., vol. 12, no. 3, pp. 35–43, 2021.
[60]. S. I. Abba et al., "Integrating feature extraction approaches with hybrid emotional neural networks for water quality index modeling," Appl. Soft Comput., vol. 114, p. 108036, 2022.
[61]. B. A. Malik, F. E. Jalal, M. Iqbal, and S. Nabi, "Estimating the deformation of micropile stabilized footings by GEP approach," Innov. Infrastruct. Solut., vol. 8, no. 6, p. 167, 2023.
[62]. K. Khan, M. Iqbal, F. E. Jalal, M. N. Amin, M. W. Alam, and A. Bardhan, "Hybrid ANN models for durability of GFRP rebars in alkaline concrete environment using three swarm-based optimization algorithms," Constr. Build. Mater., vol. 352, p. 128862, 2022.
[63]. Y. Ayub, Y. Hu, J. Ren, W. Shen, and C. K. M. Lee, "Hydrogen prediction in poultry litter gasification process based on hybrid data-driven deep learning with multilevel factorial design and process simulation: A surrogate model," Eng. Appl. Artif. Intell., vol. 126, p. 107018, 2023.
[64]. Y. Peng and C. Unluer, "Modeling the mechanical properties of recycled aggregate concrete using hybrid machine learning algorithms," Resour. Conserv. Recycl., vol. 190, p. 106812, 2023.

[65]. E. Golafshani, N. Khodadadi, T. Ngo, A. Nanni, and A. Behnood, "Modelling the compressive strength of geopolymer recycled aggregate concrete using ensemble machine learning," Adv. Eng. Softw., vol. 191, p. 103611, 2024.
[66]. K. Liu, J. Zheng, S. Dong, W. Xie, and X. Zhang, "Mixture optimization of mechanical, economical, and environmental objectives for sustainable recycled aggregate concrete based on machine learning and metaheuristic algorithms," J. Build. Eng., vol. 63, p. 105570, 2023.
[67]. D. Maulud and A. M. Abdulazeez, "A review on linear regression comprehensive in machine learning," J. Appl. Sci. Technol. Trends, vol. 1, no. 2, pp. 140–147, 2020.
[68]. Z. Sun, G. Wang, P. Li, H. Wang, M. Zhang, and X. Liang, "An improved random forest based on the classification accuracy and correlation measurement of decision trees," Expert Syst. Appl., vol. 237, p. 121549, 2024.
[69]. S. I. Abba et al., "Nitrate concentrations tracking from multi-aquifer groundwater vulnerability zones: Insight from machine learning and spatial mapping," Process Saf. Environ. Prot., vol. 184, pp. 1143–1157, 2024.
[70]. D. Stiawan, M. A. S. Arifin, M. Y. Idris, and R. Budiarto, "IoT botnet malware classification Using Weka Tool and scikit-learn machine learning," in 2020 7th International Conference on Electrical Engineering, Computer Sciences and Informatics (EECSI), IEEE, 2020, pp. 15–20.
[71]. T. O. Hodson, "Root mean square error (RMSE) or mean absolute error (MAE): When to use them or not," Geosci. Model Dev. Discuss., vol. 2022, pp. 1–10, 2022.
[72]. T. Amr, Hands-On Machine Learning with scikit-learn and Scientific Python Toolkits: A practical guide to implementing supervised and unsupervised machine learning algorithms in Python. Packt Publishing Ltd, 2020.
[73]. B. Johnston and I. Mathur, Applied supervised learning with Python: use scikit-learn to build predictive models from real-world datasets and prepare yourself for the future of machine learning. Packt Publishing Ltd, 2019.
[74]. B. J. Maiseli, "Optimum design of chamfer masks using symmetric mean absolute percentage error," EURASIP J. Image Video Process., vol. 2019, no. 1, p. 74, 2019.
[75]. X. Gao, J. Wen, and C. Zhang, "An improved random forest algorithm for predicting employee turnover," Math. Probl. Eng., vol. 2019, no. 1, p. 4140707, 2019.
[76]. R. Genuer, J.-M. Poggi, R. Genuer, and J.-M. Poggi, Random forests. Springer, 2020.
[77]. G. Zhang, "Research on safety simulation model and algorithm of dynamic system based on artificial neural network.," Soft Comput. Fusion Found. Methodol. Appl., vol. 26, no. 15, 2022.
[78]. P. Saikia, R. D. Baruah, S. K. Singh, and P. K. Chaudhuri, "Artificial Neural Networks in the domain of reservoir characterization: A review from shallow to deep models," Comput. Geosci., vol. 135, p. 104357, 2020.
[79]. O. I. Abiodun et al., "Comprehensive review of artificial neural network applications to pattern recognition," IEEE access, vol. 7, pp. 158820–158846, 2019.
[80]. D. Zhang, "Support vector machine," in Fundamentals of Image Data Mining: Analysis, Features, Classification and Retrieval, Springer, 2021, pp. 201–228.
[81]. R. M. Balabin and E. I. Lomakina, "Support vector machine regression (SVR/LS-SVM)—an alternative to neural networks (ANN) for analytical chemistry? Comparison of nonlinear methods on near infrared (NIR) spectroscopy data," Analyst, vol. 136, no. 8, pp. 1703–1712, 2011.
[82]. P. G. Asteris, M. Koopialipoor, D. J. Armaghani, E. A. Kotsonis, and P. B. Lourenço, "Prediction of cement-based mortars compressive strength using machine learning techniques," Neural Comput. Appl., vol. 33, no. 19, pp. 13089–13121, 2021.
[83]. R. Cook, J. Lapeyre, H. Ma, and A. Kumar, "Prediction of compressive strength of concrete: critical comparison of performance of a hybrid machine learning model with standalone models," J. Mater. Civ. Eng., vol. 31, no. 11, p. 4019255, 2019.
[84]. X. Li et al., "Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond," Knowl. Inf. Syst., vol. 64, no. 12, pp. 3197–3234, 2022.
[85]. M. Hasanipanah, R. A. Abdullah, M. Iqbal, and H.-B. Ly, "Predicting rubberized concrete compressive strength using machine learning: a feature importance and partial dependence analysis," J. Sci. Transp. Technol., vol. 3, no. 1, pp. 26–43, 2023.
[86]. M. Rajczakowska, M. Szeląg, K. Habermehl-Cwirzen, H. Hedlund, and A. Cwirzen, "Interpretable machine learning for prediction of post-fire self-healing of concrete," Materials (Basel)., vol. 16, no. 3, p. 1273, 2023.

[87]. S. Nazar et al., "Estimation of strength, rheological parameters, and impact of raw constituents of alkali-activated mortar using machine learning and SHapely Additive exPlanations (SHAP)," Constr. Build. Mater., vol. 377, p. 131014, 2023.
[88]. M. N. Amin, W. Ahmad, K. Khan, S. Nazar, A. M. A. Arab, and A. F. Deifalla, "Evaluating the relevance of eggshell and glass powder for cement-based materials using machine learning and SHapley Additive exPlanations (SHAP) analysis," Case Stud. Constr. Mater., vol. 19, p. e02278, 2023.
[89]. X. Zheng et al., "A data-driven approach to predict the compressive strength of alkali-activated materials and correlation of influencing parameters using SHapley Additive exPlanations (SHAP) analysis," J. Mater. Res. Technol., vol. 25, pp. 4074–4093, 2023.
[90]. A. Ahmad et al., "Prediction of geopolymer concrete compressive strength using novel machine learning algorithms," Polymers (Basel)., vol. 13, no. 19, p. 3389, 2021.
[91]. M. H. Nguyen and H.-B. Ly, "Development of machine learning methods to predict the compressive strength of fiber-reinforced self-compacting concrete and sensitivity analysis," Constr. Build. Mater., vol. 367, p. 130339, 2023.
[92]. K. Khan, B. A. Salami, M. Iqbal, M. N. Amin, F. Ahmed, and F. E. Jalal, "Compressive strength estimation of fly ash/slag based green concrete by deploying artificial intelligence models," Materials (Basel)., vol. 15, no. 10, p. 3722, 2022.
[93]. M. N. Amin et al., "Prediction of strength and CBR characteristics of chemically stabilized coal gangue: ANN and random forest tree approach," Materials (Basel)., vol. 15, no. 12, p. 4330, 2022.
[94]. Y. Song et al., "Prediction of compressive strength of fly-ash-based concrete using ensemble and non-ensemble supervised machine-learning approaches," Appl. Sci., vol. 12, no. 1, p. 361, 2021.
[95]. J. Neter, M. H. Kutner, C. J. Nachtsheim, and W. Wasserman, "Applied linear statistical models," 1996.
[96]. T. Nel, C. E. Clarke, and A. G. Hardie, "Evaluation of simple and multivariate linear regression models for exchangeable base cation conversion between seven measurement techniques on South African soils," Geoderma Reg., vol. 30, p. e00571, 2022.
[97]. M. Cortés-Carmona, G. Jiménez-Estévez, and J. Guevara-Cedeño, "Support vector machines for on-line security analysis of power systems," in 2008 IEEE/PES Transmission and Distribution Conference and Exposition: Latin America, IEEE, 2008, pp. 1–6.
[98]. Y. Wang, W. Liao, H. Shen, Z. Jiang, and J. Zhou, "Some notes on the basic concepts of support vector machines," J. Comput. Sci., vol. 82, p. 102390, 2024.
[99]. M. Valipour and J. Dietrich, "Developing ensemble mean models of satellite remote sensing, climate reanalysis, and land surface models," Theor. Appl. Climatol., vol. 150, no. 3, pp. 909–926, 2022.
[100]. R. E. Schapire and Y. Singer, "Improved boosting algorithms using confidence-rated predictions," in Proceedings of the eleventh annual conference on Computational learning theory, 1998, pp. 80–91.
[101]. C. Li, Y. Huang, H. Li, and X. Zhang, "A weak supervision machine vision detection method based on artificial defect simulation," Knowledge-Based Syst., vol. 208, p. 106466, 2020.
[102]. L. Breiman, "Random forests," Mach. Learn., vol. 45, pp. 5–32, 2001.
[103]. T.-H. Lee, A. Ullah, and R. Wang, "Bootstrap aggregating and random forest," Macroecon. Forecast. era big data Theory Pract., pp. 389–429, 2020.
[104]. C. Lai, M. J. T. Reinders, and L. Wessels, "Random subspace method for multivariate feature selection," Pattern Recognit. Lett., vol. 27, no. 10, pp. 1067–1076, 2006.
[105]. T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785–794.
[106]. L. Breiman, "Bagging predictors," Mach. Learn., vol. 24, pp. 123–140, 1996.
[107]. L. Cavaleri, M. S. Barkhordari, C. C. Repapis, D. J. Armaghani, D. V. Ulrikh, and P. G. Asteris, "Convolution-based ensemble learning algorithms to estimate the bond strength of the corroded reinforced concrete," Constr. Build. Mater., vol. 359, p. 129504, 2022.
[108]. K.-H. Yang, H.-Z. Hwang, and S. Lee, "Effects of water-binder ratio and fine aggregate–total aggregate ratio on the properties of Hwangtoh-based alkali-activated concrete," J. Mater. Civ. Eng., vol. 22, no. 9, pp. 887–896, 2010.