SSRG International Journal of Civil Engineering
ISSN: 2348-8352/ https://doi.org/10.14445/23488352/IJCE-V11I3P107

Volume 11 Issue 3, 80-93, March 2024

Original Article

Development of A Machine Learning-Based Model for Predicting Compressive Strength of Ordinary Portland Cement

Tesfahiwet Kesete[1], Christopher Kanali[2], Victoria Okumu[3], Gidewon Tekeste[4]

[1]*Department of Civil and Construction Engineering, Pan African University Institute for Basic Sciences, Technology and Innovation (PAUSTI) Hosted at Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya.*
[2]*Department of Agricultural and Biosystems Engineering, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya.*
[3]*Department of Civil Engineering, Multimedia University, Nairobi, Kenya.*
[4]*Department of Civil Engineering, LNEC & University of Aveiro (UA), Lisbon, Portugal.*

[1]*Corresponding Author : kesete.tesfahiwet@students.jkuat.ac.ke*

Received: 14 January 2024 Revised: 16 February 2024 Accepted: 10 March 2024 Published: 31 March 2024

Abstract - *Cement is a vital construction material with widespread use in the construction industry, acting as a binding agent for various construction materials. The compressive strength of cement, which measures its binding force and ability to withstand compression, is a crucial factor in manufacturing cement and constructing concrete-based structures. Traditionally, costly laboratory tests have been employed to determine cement's compressive strength. However, with the complexity of material engineering, this approach has become inefficient, leading to resource and time losses. Establishing a logical connection between cement's chemical composition, physical characteristics, and compressive strength is also challenging due to its heterogeneous properties and nonlinear behaviour. However, to address these issues, with the evolution of machine learning and its efficient modelling techniques, different modelling techniques are prepared to study its behaviour and satisfy the desired performance. This paper aims to demonstrate the effectiveness of different shallow supervised machine learning techniques such as Multivariant linear regressions, Decision Tree (DT), Nonlinear regression, and ensemble Random Forests (RF) and apply Principal Component Analysis (PCA) to develop a compressive strength prediction model to overcome the disadvantages of traditional approaches (experimental analysis) in estimating the compressive strength of cement. Finally, the study has compared the results obtained by these different Machine Learning (ML) techniques and provided a general conclusion.*

Keywords - *Cement, Cement compressive strength, Machine Learning model, Nonlinear regression, Principal Component Analysis.*

1. Introduction

1.1. Background of the Study

In the contemporary construction landscape, Ordinary Portland cement stands as the predominant building material. Reports indicate that there has been a significant increase in cement production over the years. In 1990, the global production of cement was approximately 1.2 billion tonnes. However, it is projected that by the year 2050, the global production of cement will increase substantially to reach approximately 5.8 billion tonnes [1].

This suggests a notable growth trend in cement production over the decades, reflecting the increasing demand for cement in various construction and infrastructure projects worldwide. Many types of cement are artificially manufactured (e.g., ordinary Portland cement, low-heat cement, high alumina cement, expensive cement, waterproof cement, hi-bond cement, etc.). They are being used under certain conditions due to their special properties [2].

In the building industry, however, regular Portland cement is the most widely used type of cement for making non-speciality grouts, mortar, stucco, and concrete paste. Currently, various types of cement with distinct strength grades based on the compressive strength obtained after 28 days of setting, such as 32.5, 42.5, and 52.5 MPa, are manufactured and utilized in diverse construction projects under similar construction and curing conditions [3].

In order to estimate the compressive strength using empirical formulas and a variety of statistical and mathematical techniques are needed, inputs such as material

compositions have been established [4-6]. In addition to this, most empirical models require the prior assumption or knowledge of the underlying mathematical and physical models [7]. But in comparison to clever computer algorithms like ML models, these statistical and mathematical models are less precise and less dependable for making future predictions [8].

In machine learning, the goal is to build computer systems that learn from the data that is already available. This allows for the efficient determination of the relationship between the response (dependent) variable(s) and the input features in a complex system for the purpose of future forecasting. Also, it can calculate the relationship between the input variables and the response parameter(s) [7].

Experimental procedures for measuring cement compressive performance are typically time-consuming, expensive, and labour-intensive. To address this issue, researchers have explored machine learning models for predicting cement compressive strength. [9] Examine the following three kernel-based models: Gaussian Process Regression (GPR), Relevance Vector Machine (RVM), and Support Vector Regression (SVR) using input variables such as C3S (%), SO3 (%), Alkali (%), and Blaine (cm^2/g), with the output being the 28-day cement compressive strength (N/mm^2).

The results demonstrate that these models perform similarly to Artificial Neural Networks (ANN) but provide superior empirical performance and capacity for generalization, overcoming some of the limitations of ANN in predicting cement compressive strength. While this study does not address all factors affecting cement compressive strength, it underscores the potential of machine learning techniques to forecast material qualities, reducing the need for costly and time-consuming trial tests.

By merging several separate base learners or base machine learning models, ensemble machine learning algorithms technique improve prediction performance [10]. By utilizing the advantages of each model while minimizing its shortcomings [11] these foundational models, which may come from the same or separate classes used to provide forecasts that are more accurate.

Furthermore, machine learning models' performance and capacity for generalization are assessed using the K-fold cross-validation technique. With this strategy, the performance estimate is more stable because the dataset is divided into k equal-sized folds and every data point is used exactly once for validation [12].

In order to reduce the unpredictability associated with single train-test divides, performance measures collected from each iteration are averaged to create the final assessment [13]. When developing a machine learning regression model, the database description offers accurate and comprehensive details on the dataset that is used to train and evaluate the model. [14] Typically, database description in the machine learning field of study contains information on the number of samples or instances, features or qualities, and the type of target variable being forecasted are common details included.

It may also contain details about the features' data types, including null, missing, and category values, as well as information about any features that are ordinal, categorical, or numerical. Apart from these specifics, the description of the database could also encompass details regarding the procedure followed for gathering the data, the sampling strategy employed, the sources from which the data were acquired, and any actions taken for preprocessing the data, like scaling or normalization.

In addition, it describes any difficulties or biases that exist in the dataset and the methods used to lessen them. Moreover, any feature selection strategies or transformations used to enhance model performance and lower computing complexity are highlighted in the database description. Ensuring the integrity and quality of the machine learning regression model creation process is largely dependent on having a thorough description of the database.

The process of developing machine learning regression models heavily relies on data normalization sometimes referred to as feature scaling. Input feature numerical values are converted to a common scale, usually ranging from 0 to 1 or centered around a mean of 0 and a standard deviation of 1. Because of this standardization, features with bigger scales are kept from controlling the learning process, and all features are guaranteed to have an equal influence on the model.

[15] Two widely used methods for data normalization are Min-Max scaling, which scales values to fall inside a specified range, and Z-score normalization, which normalizes findings to have a mean of 0 and a standard deviation of 1. By improving convergence speed, enhancing model interpretability, and preventing bias towards certain features [16], data normalization is essential for building accurate and robust regression models.

The physical and chemical characteristics of ordinary Portland cement largely dictate its compressive performance. The primary factors influencing Portland cement's compressive strength are its chemical makeup and physical properties [17].

The four main constituents of cement are dicalcium silicate (C2S), tricalcium silicate (C3S), tricalcium aluminate (C3A), and tetracalcium alumino-ferrite (C4AF). Other factors that are analyzed include free lime (CaO), magnesia (MgO), free silica (SO3), Loss on Ignition (LOI), Insoluble

Residue (IR), Lime Saturation Factor (LSF), Silica Modulus (SM), and Alumina Ratio (AR) moreover, setting time, fineness, and soundness. The compressive strength of cement is the most important and has the highest value of all its mechanical qualities.

2. Methodology

2.1. Study Frame Work

The comprehensive workflow for this study is shown in Figure 1 below. In order to construct machine learning modelling, it starts with preparatory processes such as reliable data collection, descriptive statistics, data mining, and model evaluation.

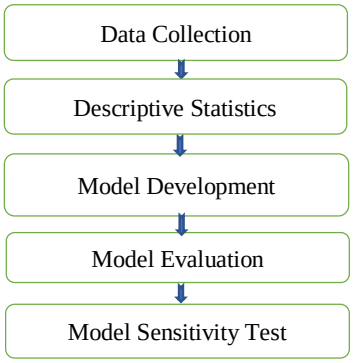

Fig. 1 Workflow schematic diagram

2.2. Dataset Information

Accurate prediction results hinge on the synergy between an effective model evaluation and reliable data. In previous studies, scholars focused more on developing mathematical and machine learning models to predict the compressive strength of cement but often ignored the importance of a reliable database to address all the factors affecting the compressive strength of cement. A reliable and comprehensive database serves as the foundation for verifying the accuracy of the model.

In this study, data was collected from previous experimental tests that were gathered from the Gedem cement factory in Eritrea from 2008 to 2022 and established a large and reliable database as a dataset for predicting the compressive strength of cement having chemical, physical and mechanical properties.

Therefore, the newly compiled dataset, consisting of 2217 samples with 17 attributes, each input attribute contributes to the dependent variable, namely cement compressive strength. Within this database, all sixteen parameters serve as independent variables, while cement compressive strength acts as the dependent variable.

2.3. Descriptive Statistics and Feature Engineering

Before starting to develop a machine learning model, the generally recommended procedure is to study the behaviour of the dataset and explore the summary of the descriptive statistics of the variables [18]. A good knowledge of the behaviour of the dataset (data exploratory analysis) helps in the upcoming modelling steps.

Based on the exploratory analysis, some sort of data mining is performed, such as dealing with missing values, duplicate values, outliers, skewness and data normalization.As a result, during the data collection and exploratory analysis phases, input variables are tightly screened.

Specifically, datasets with sixteen input variables (i.e., none of which are null) were chosen simultaneously. 2217 test trails total, randomly split between the training and testing sets in the database.

2.4. Modelling Techniques

Following data observation and adjustment, different shallow supervised machine learning regression techniques are employed based on the input data type and study objectives to predict the compressive strength of cement. Specifically, this study explores 11 machine learning algorithms across 4 techniques: Multivariate Linear Regression, Fractional polynomial or Nonlinear Regression, Decision Tree Regression and Ensemble Models.

Additionally, Principal Component Analysis (PCA) is conducted to project the original predictors and reduce their dimensionality, typically retaining eigenvectors with the least eigenvalue or vectors with the least variance [19]. Subsequently, the aforementioned regression techniques are applied to the reduced-dimensional data.

The ultimate goal of training a predictive model is to ensure its ability to generalize effectively to unseen data, thereby enabling accurate predictions based on internally adjusted parameters derived from training and validation. Leveraging Python, a versatile high-level programming language, the selected machine learning algorithm has been implemented using well-known libraries such as Scikit-learn [20].

The dataset has been divided into training and test sets, with the former comprising 80% and the latter 20% of the total data [18]. Within the training set, a portion (20%) has been further allocated for validation to facilitate parameter tuning. Initially, base models were trained using default parameter values, and their performance was assessed against actual values.

Next, k-fold cross-validation was used to optimize the algorithm parameters in order to increase the performance of each model. The evaluation results that follow show that this process significantly improved each model. Ultimately, a comparative analysis is conducted between the performance of various machine learning models.

Table 1. Dataset parameters

No.	Name of the Attribute	Data Type	Unit	Variable Category	Number of Attributes
1	C3S	Numerical	Percentage	Input variable	2217
2	C2S	Numerical	Percentage	Input variable	2217
3	C3A	Numerical	Percentage	Input variable	2217
4	C4AF	Numerical	Percentage	Input variable	2217
5	MgO	Numerical	Percentage	Input variable	2217
6	SO3	Numerical	Percentage	Input variable	2217
7	Alkalies	Numerical	Percentage	Input variable	2217
8	SM	Numerical	Percentage	Input variable	2217
9	IM	Numerical	Percentage	Input variable	2217
10	fCaO	Numerical	Percentage	Input variable	2217
11	LOI	Numerical	Percentage	Input variable	2217
12	IR	Numerical	Percentage	Input variable	2217
13	Fineness	Numerical	Percentage	Input variable	2217
14	IST	Numerical	Percentage	Input variable	2217
15	FST	Numerical	Percentage	Input variable	2217
16	Soundness	Numerical	Percentage	Input variable	2217
17	Cement Strength on the 28th day of curing	Numerical	Mega Pascal	Output variable	2217

2.5. Model Evaluation

A regression model's performance is dependent on how well it makes predictions, which are evaluated based on the error rate between actual and predicted values. A robust regression model exhibits minimal discrepancies between actual and predicted values while maintaining impartiality.

To assess model performance in this study, four evaluation metrics have been chosen: Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Absolute Squared Errors (RMSE), and R-squared (R^2) or coefficient of determination. Moreover, the K-Fold Cross-Validation (CV) technique has been utilizing.

1. Mean Squared Error (MSE): Lower MSE values indicate better performance, as they reflect smaller errors between predicted and actual values. Comparing MSE across models helps identify the model with the least error on average.
2. Mean Absolute Error (MAE): Similar to MSE, lower MAE values indicate better performance, with each error being measured independently of magnitude. Models with lower MAE values are preferred.
3. Root Mean Squared Error (RMSE): An indicator of the average magnitude of mistakes is provided by RMSE. Models with lower RMSE values have smaller errors on average and are considered better performers [21].
4. Coefficient of Determination (R2): R2 values range from 0 to 1, with higher values indicating a better fit in the model to the data. A higher R2 value suggests that a larger proportion of the variance in the dependent variable is explained by the independent variables. [22]. Comparing R2 across models helps identify the model with the best fit to the data.
5. Kfold Cross-validation with Standard Deviation: Kfold Cross-validation provides a more reliable indication of the model's capacity for generalization by evaluating its performance across several data subsets. The standard deviation of performance metrics across folds indicates the variability of model performance, with lower standard deviation values suggesting more consistent performance [21].

Generally, an elevated R-squared value on the training dataset indicates the model's ability to account for a significant portion of the variance in the dependent variable

using the features present in the training data [22]. However, excessively high R-squared values may signal overfitting, wherein the model fits the training data too closely and may struggle to generalize to new, unseen data. Consequently, evaluating model performance on a separate testing dataset, which the model has not been exposed to during training, is imperative.

The R-squared value on the testing dataset offers insights into the model's generalization capability, with higher values indicating accurate predictions on previously unseen data, thereby signifying robust generalization. Also, utilizing the K-Fold Cross-Validation (CV) method to all samples for both training and testing inputs during model training. The findings of the CV evaluation give a more thorough picture of how well the model performs in practical situations. For the final model performance comparison and ideal model selection, a 20-fold CV technique has been utilized. Therefore, at last each model is evaluated, and their performances are compared to identify the most effective solution model.

2.6. Software Used

The MS-Excel tool serves to collect, extract, and systematically organize the cement compressive strength data. Subsequently, Jupyter Notebook, coupled with the Python programming language, facilitates preprocessing, statistical analysis and application of machine learning modelling techniques.

3. Results and Discussion

3.1. Descriptive Statistics and Feature Engineering

Table 2 provides a statistical overview of the dataset used in this study, presenting the minimum and maximum values, standard deviation, and average values of both input and output variables. Utilizing the proposed expression within the specified parameter range is essential for developing the most effective machine learning predictive model. The statistical analysis indicates that the dataset encompasses a wide range of ingredients, with the standard deviation reflecting the distribution of data around the mean values.

The compressive strength range of all the gathered datasets spans from 42.52 to 67.86 MPa, indicating that the proposed model is suitable for estimating the compressive strength of samples falling within this range. The table demonstrates that the variables included in the database encompass a diverse range, validating the reliability of the dataset. Consequently, with this extensive dataset, the suggested model can predict compressive strength with accuracy.

Table 2. Statistical description of parameters

	Count	Mean	Std.	Min.	25%	50%	75%	Max.
C3S	2217	41.832296	11.255896	10.3	34.28	42.37	49.43	69.63
C2S	2217	30.970095	9.711784	7.24	24.34	30.22	37.16	61.13
C3A	2217	5.931299	1.080159	3.43	5.05	5.94	6.81	8.76
C4AF	2217	14.186784	1.767174	11.35	12.57	14.19	15.77	16.99
MgO	2217	1.814777	0.807886	0.12	1.24	1.94	2.38	4.63
SO3	2217	1.162963	0.885565	0.02	0.42	0.95	1.64	3.4
Alkalies	2217	0.795291	0.535222	0.01	0.4	0.61	1.18	2.07
SM	2217	2.213157	0.147098	1.81	2.09	2.2	2.33	2.65
IM	2217	1.137384	0.148285	0.87	1.01	1.12	1.26	1.49
fCaO	2217	2.70313	1.433114	0.2	1.47	2.71	3.93	5.2
LOI	2217	2.493667	1.267004	0.3	1.4	2.47	3.59	4.7
IR	2217	2.358755	1.079656	0.39	1.43	2.36	3.29	4.28
Fineness	2217	369.105918	39.14649	301.47	335.22	369.04	402.73	437.52
IST	2217	125.755016	37.837629	60.15	92.92	125.98	158.47	191.25
FST	2217	240.875313	98.658078	90	159.28	227.73	313.2	450
Soundness	2217	3.843424	2.055417	0.3	2.06	3.85	5.61	7.42
Strength	2217	56.857763	6.138749	42.52	53.55	57.08	61.26	67.86

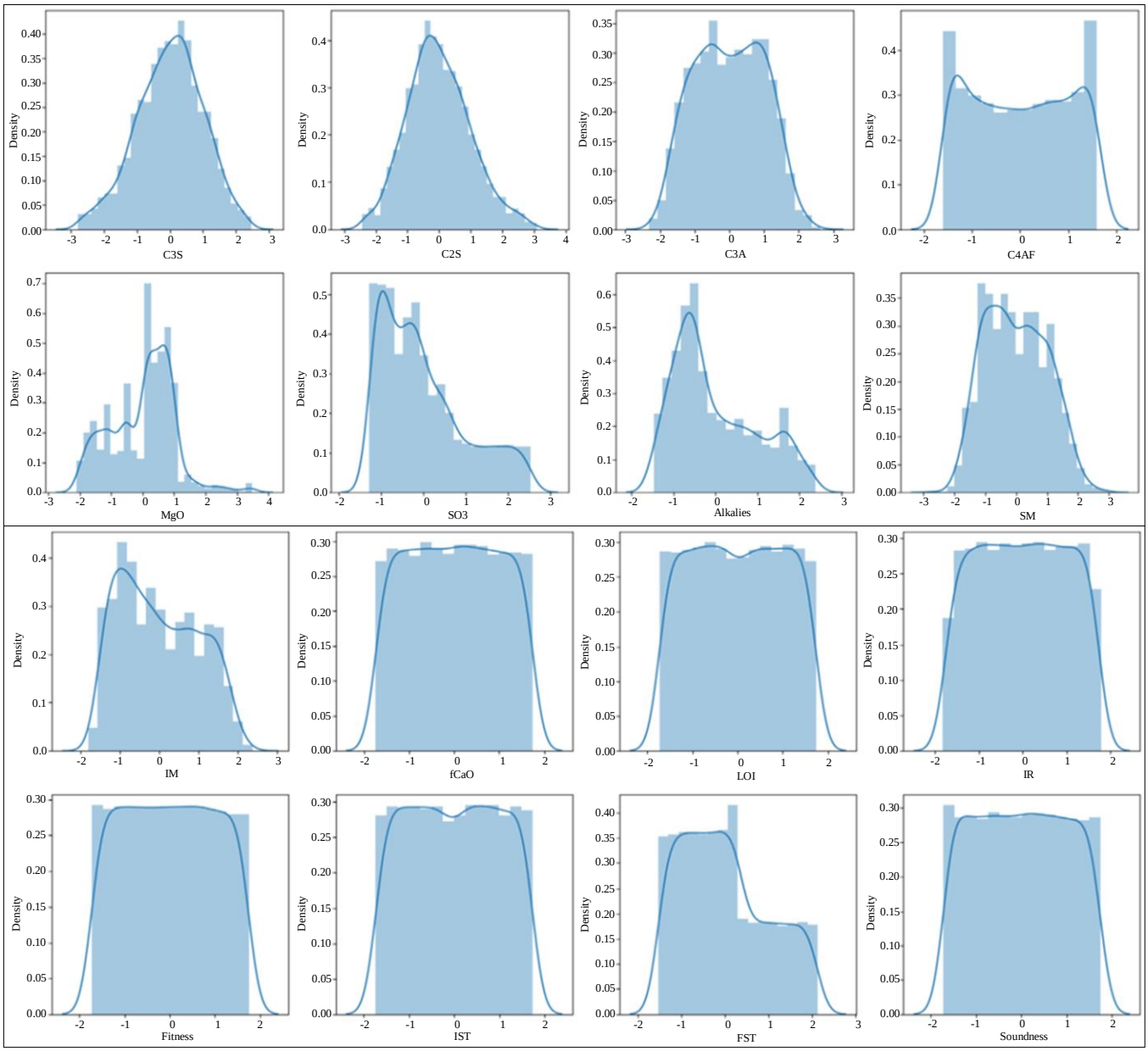

Fig. 2 Input variables relative frequency distribution

The relative frequency distribution of all the seventeen attributes was examined by the probability distribution. Some notes can be taken whether they follow a normal distribution, skewed distribution, or other distribution shapes. After the data treatment has been made, the variables have normal distribution except SO3 and Alkalies have slight skewness.

Moreover, the data spread or variability is assessed by examining the range Interquartile Range (IQR) and standard deviation. Therefore, from the Gaussian distribution graph shown in Figure 2 standard scale is applied to ensure that all features contribute equally to the learning process and enhance the performance, stability, and interpretability of the machine learning model.

3.1.1. Correlation Analysis

Correlation analysis involves assessing the closeness between correlated variables to understand their relationship. High correlation among input parameters can hinder model efficiency and complicate the interpretation of input parameter effects on output parameters. Therefore, prior to training machine learning models, it is crucial to analyze the correlation between cement compressive strength and independent variables, as depicted in Figure 4.

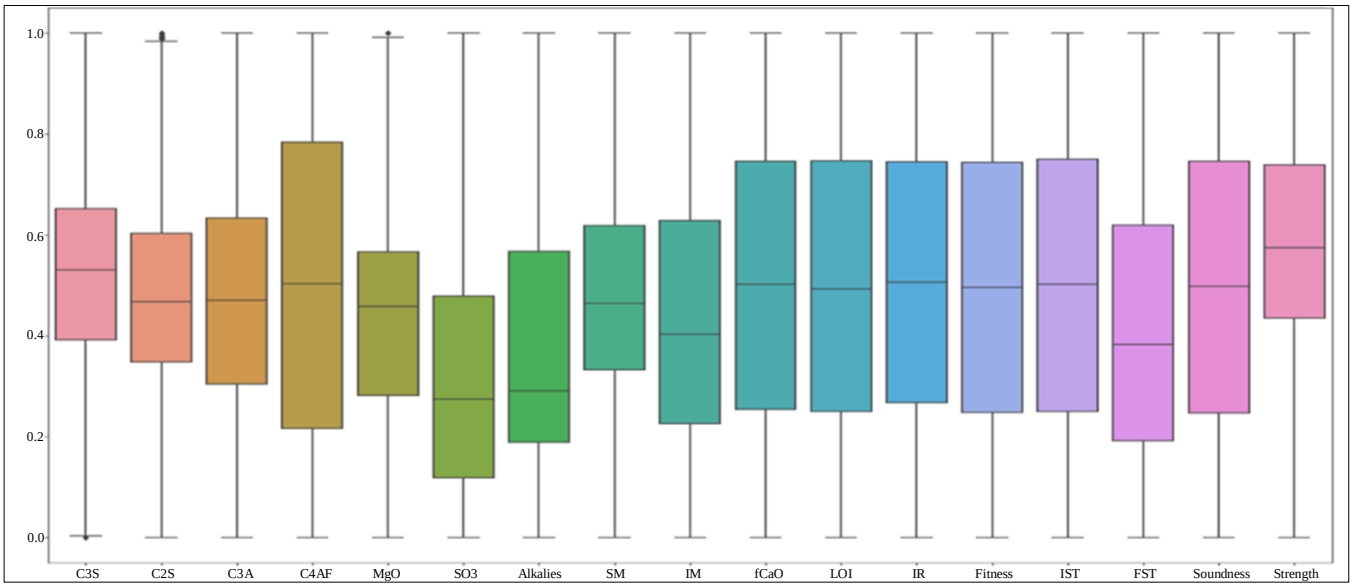

Fig. 3 Variable outlier box-plot graph

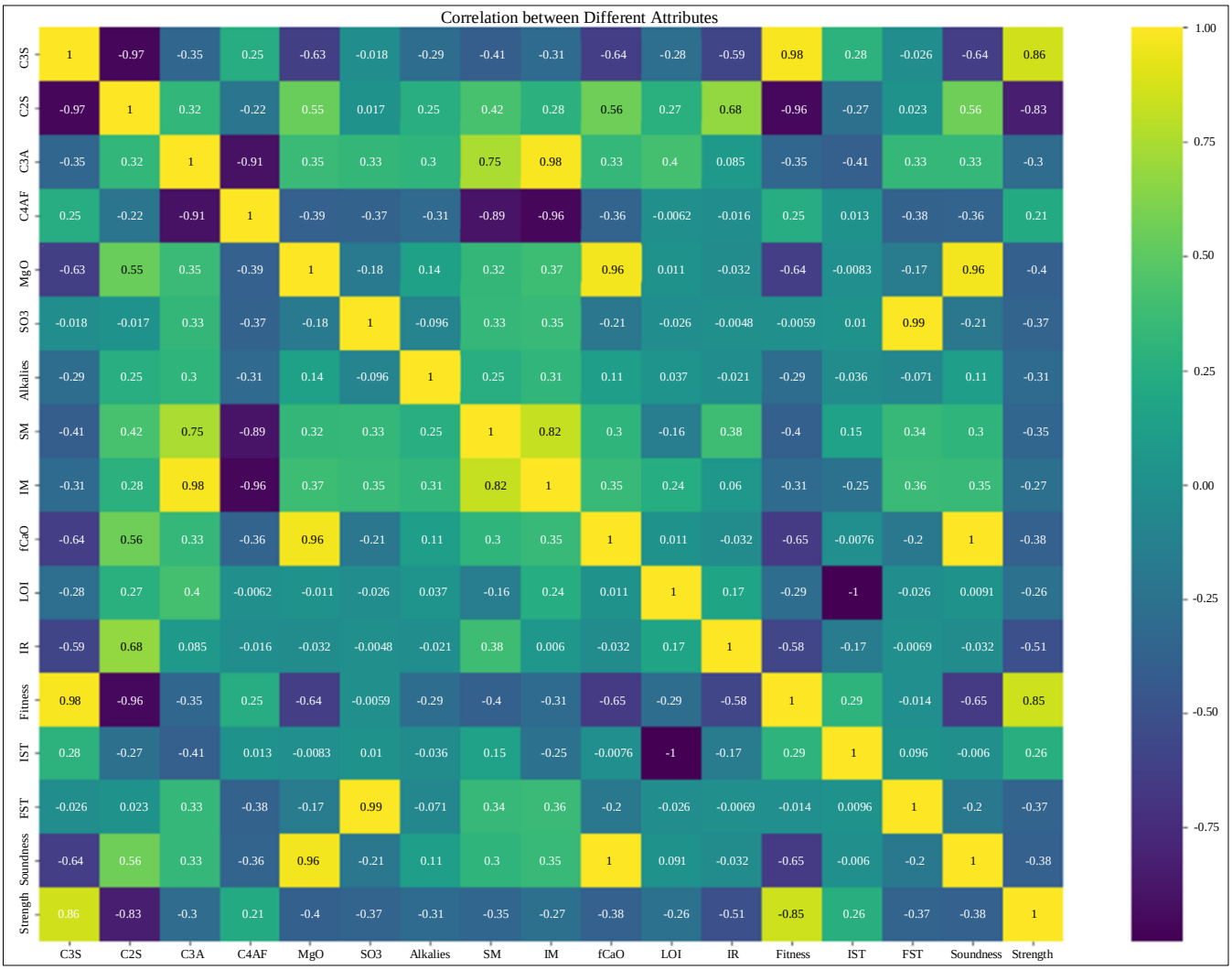

Fig. 4 Variable relationships

The analysis reveals that certain input parameters exhibit high correlation coefficients which are greater than 0.6, suggesting multicollinearity issues, which can destabilize regression models. To address this issue, the authors employ Principal Component Analysis (PCA) to mitigate multicollinearity effects.

PCA transforms a correlated set of original variables into orthogonal principal components, which are linear combinations of the original variables, each capturing a unique source of variance in the data. By retaining a subset of principal components that explain the majority of variance, PCA reduces dimensionality while preserving information.

This process helps mitigate multicollinearity by creating orthogonal features, less correlated with each other, thus enhancing the stability and interpretability of regression models; 7 to 8 of the input variables can represent up to 99% of the data after Principal Component Analysis (PCA), as shown in Figure 5.

The correlation analysis and associated graph reveal the direction of relationships between each pair of variables and compressive strength. Positive correlations indicate that variables such as C3S, C4AF, Fineness, and Initial Setting Time (IST) are positively associated with compressive strength, meaning that as their values increase, so does compressive strength. Conversely, negative correlations suggest that increasing certain variables leads to a decrease in compressive strength.

In addition to the visual representation, numerical values are provided within each cell of the heatmap, as shown in Figure 4, indicating the exact correlation coefficient between the two variables. These numerical values allow for precise quantification of the strength and direction of the correlations, facilitating more detailed analysis.

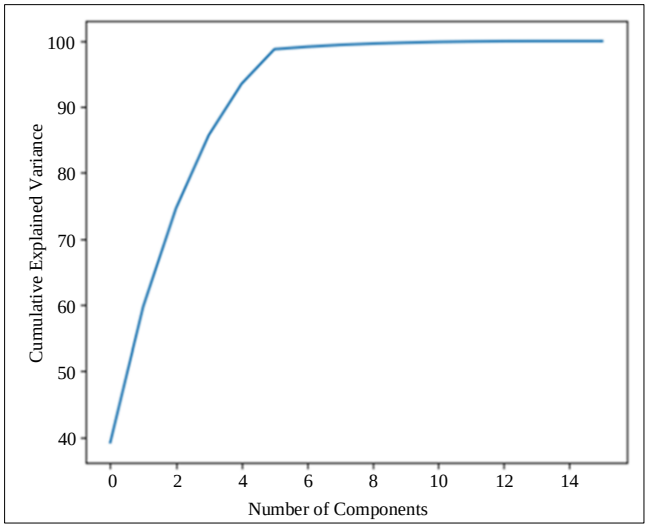

Fig. 5 Principal Component Analysis (PCA)

The correlation analysis and correlation graph can also provide valuable insights into the relationships between variables, which can help in selecting appropriate machine learning algorithms for training. The relationship or correlation between variables is not strictly linear, as shown in Figure 6; selecting an appropriate machine learning algorithm becomes crucial for accurate modelling.

In such cases, nonlinear regression algorithms are more suitable for capturing the complex patterns in the data. Fractional-polynomial or nonlinear regression, Decision trees, and Ensemble models are suitable fit models for nonlinear regression algorithms that can effectively capture nonlinear relationships between variables. These algorithms are capable of capturing complex interactions and nonlinearities in the data, making them well-suited for situations where linear models may not adequately represent the underlying patterns.

By considering the nature of the data and the complexity of the relationships between variables, one can choose the most appropriate nonlinear regression algorithm to build predictive models that accurately capture the underlying patterns in the data. Additionally, ensemble methods like random forests and gradient boosting can offer robustness and generalization capabilities, making them suitable choices for complex, nonlinear datasets.

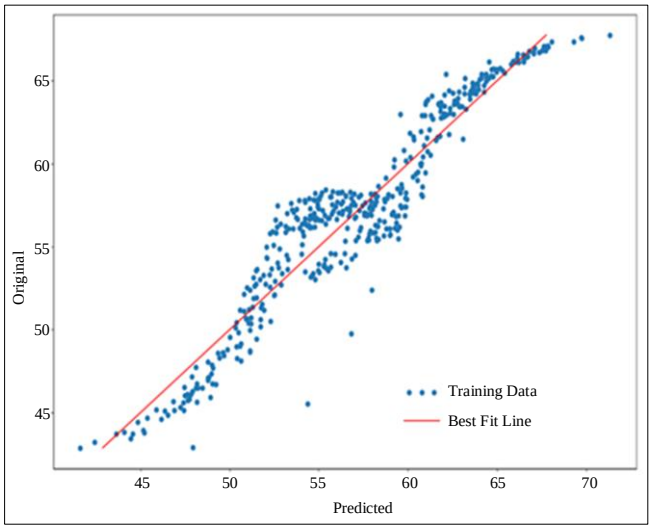

Fig. 6 Model correlation graph

The correlation analysis graph of the numerical data, as illustrated in Figure 6, thus demonstrates that the target feature (Cement compressive strength) has a nonlinear association with the other features. Thus, suitable ML methods are chosen based on the correlation findings of the input features.

3.2. Model Train and Evaluation

In the process of developing and evaluating the regression algorithms listed in Table 3, the dataset is split into 80% training and 20% testing sets to facilitate model training and

evaluation. During the training phase, each algorithm is trained on the training dataset using default hyperparameters or optimized hyperparameters. Once trained, the models are evaluated on the testing dataset using performance metrics of Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared to assess predictive accuracy and goodness of fit.

Additionally, the k-fold cross-validation technique has been employed to ensure the robustness of the models and reduce the risk of overfitting. After carefully evaluating the performance metrics of these various machine learning models, it was evident that both the nonlinear or fractional polynomial regression model and the ensemble models outperformed others. However, the considerable disparity between the R2 values of the training and testing datasets for ensemble models indicates overfitting, rendering them less suitable as generalized models.

Consequently, the nonlinear regression or fractional polynomial model was selected as a more generalized and reliable option. This model exhibited the lowest error rates, signifying higher accuracy in predicting the target variable and demonstrated the best fit to the data by effectively capturing underlying patterns and relationships.

Moreover, its consistent performance across different subsets of the data further affirmed its reliability and robustness in predictive tasks. Therefore, based on these findings, the nonlinear or fractional polynomial regression model emerged as the most suitable choice for the predictive modelling task at hand.

Table 3. Performance evaluation summaries

No.	Machine Learning Algorithm	MSE	MAE	RMSE	R2_Training	R2_Testing	KFold	Std.
1.1	Linear Regression	3.6762	1.5023	1.9173	91.6754	89.9184	90.6946	1.831
1.2	Ridge Regression	3.7113	1.5041	1.9265	91.6559	89.8221	90.6219	2.019
1.3	Lasso Regression	7.5786	2.2542	2.7529	81.4307	79.2166	86.1212	1.9079
1.4	Elastic Net Regression	7.3044	2.2449	2.7027	81.2382	79.9686	87.2671	1.7887
2	Fractional Polynomial / Nonlinear Regression	3.3594	1.4805	1.8329	91.347	90.6126	90.6946	1.831
3	Decision Tree Regression	12.4442	2.6849	3.5276	70.0518	65.2263	66.3041	5.6148
4.1	Bagging Regressor	2.2271	1.0892	1.4923	98.8814	93.7767	94.1696	1.4144
4.2	Random forest Regression	1.8872	1.0247	1.3737	99.2862	94.7266	95.1128	0.9547
4.3	Ada Boost Regressor	4.6214	1.8249	2.1497	88.9059	87.0861	87.1072	1.3058
4.4	Gradient Boosting Regression	2.2834	1.1898	1.5111	96.5587	93.6192	94.1289	1.1259
4.5	XGB Regressor	2.1975	1.1354	I.4824	99.914	93.8593	94.8331	1.1877

3.2.1. Variable Importance Evaluation

According to the variable importance analysis shown in Figure 7, the machine learning method provides a practical means of predicting the compressive strength of regular ordinary Portland cement. This approach can be used to understand the significance of physical properties and chemical composition in relation to cement compressive strength.

In this study, the machine learning method was applied to assess the significance of the sixteen input parameters on cement compressive strength, as depicted in Figure 7 and Table 4. Notably, the influence scores of C3S, Fineness, C2S, IR, MgO, fCaO, Soundness, FST, SO3, SM, Alkalies, C3A, IM, IST, LOI, and C4AF on cement compressive strength are 29.578, 27.040, 24.185, 3.688, 1.971, 1.845, 1.837, 1.639, 1.637, 1.483, 1.137, 1.073, 0.843, 0.786, 0.754, and 0.505, respectively, showing a decreasing degree of influence.

The most significant factor is C3S, followed by fineness, while C4AF has the least influence. When creating typical Portland cement of a high grade, engineers can find some guidance from the analysis of the impact of cement's physical and chemical compositions on the material's compressive strength.

In order to produce cement of higher compressive strength, the engineers can pay more attention to the cement C3S from chemical composition and the fineness from the physical properties or manufacturing process.

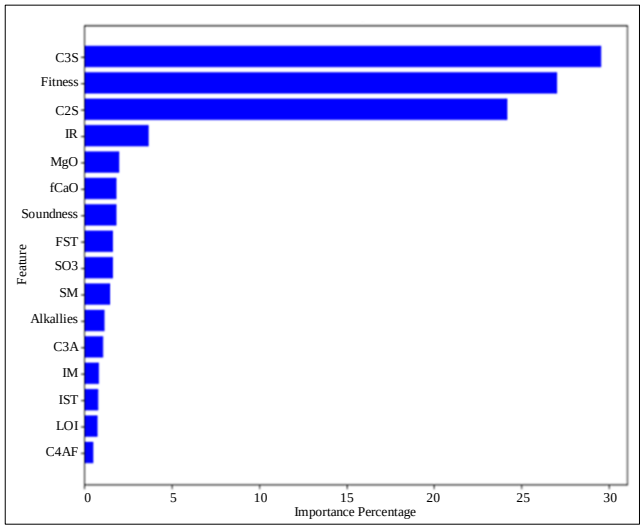

Fig. 7 Variable importance evaluations

Table 4. Variable importance evaluations

Feature	Importance %
C3S	29.578
Fineness	27.04
C2S	24.185
IR	3.688
MgO	1.971
fCaO	1.845
Soundness	1.837
FST	1.639
SO3	1.637
SM	1.483
Alkalies	1.137
C3A	1.073
IM	0.843
IST	0.786
LOI	0.754
C4AF	0.505

3.2.2. Sensitivity Test of Nonlinear Regression Machine Learning Model

The purpose of every regression technique is to fit functions that minimize the residuals between the function and the data when the sum of their squares is calculated. Least-squares regressions are the name given to such techniques. In cases where a dependent variable is not linearly related to the independent variables, nonlinear regression models, such as fractional polynomial regression, are utilized; unlike linear regression, where the relationship between variables follows a straight line, fractional polynomial regression allows for more flexible and curved relationships between the dependent and independent variables.

The general form of the fractional polynomial regression equation can be represented as:

$$f(t) = a0+a1*X^{a2}+ a3* X2^{a4}+ a5* X3^{a6}....+ am* X4^{n} \quad (1)$$

Where a0, a1, a3..., am are the regression coefficients and a2, a4, a6..., an are the powers to which the independent variables are raised.

This type of regression model is more suitable when the relationship between the input variables and the output variable is nonlinear or when the data points do not follow a straight-line pattern. By allowing for nonlinear relationships, fractional polynomial regression can capture more complex patterns in the data and provide more accurate predictions.

In this study, fractional polynomial regression was found to be very suitable for predicting the compressive strength of Portland cement from factors affecting this strength, such as C_3S, fineness (S_s), C_2S, IR, MgO, Free CaO, soundness, FST, SO_3, Alkalies, C_3A, IM, IST, LSF, LOI, and C_4AF and the mathematical expression of the model used was:

$$f(t) = a0 +a1*C_3S^{a2} + a3* S_s^{a4} + a5* C_2S^{a6} + a7* IR^{a8}+ a9*MgO^{a10}+.....+ a31* C_4AF^{a32} \quad (2)$$

The sensitivity test is also conducted to assess the robustness and stability of a model with respect to changes or variations in its input variables. Specifically, in the case of the nonlinear regression model developed for predicting cement compressive strength, a sensitivity test involves systematically altering the values of the sixteen input variables within a specified range and observing the corresponding changes in the predicted output (compressive strength).

This allows for an evaluation of how sensitive the model's predictions are to variations in the input variables, helping to identify which variables have the greatest impact on the predicted outcome and how changes in those variables affect the model's performance.

Sensitivity tests are valuable for understanding the reliability and accuracy of the model across different scenarios and can inform decisions regarding model refinement and optimization. Table 5 gives the nonlinear regression coefficient and exponents of the machine learning prediction model for the prediction of 28 days compressive strength, as well as the value of the coefficient of correlation and standard error of the estimate corresponding to each set of input variables used in each model.

Table 5. Regression coefficients and Exponents for the 28-day compressive strength Nonlinear machine learning prediction model

Variable	Coefficients & Exponents	Model 1	Model 2	Model 3	Model 4	Model 5
	a0	56.858	56.858	56.858	56.858	56.858
C_3S (%)	a1	0.125	-0.023	-0.023	-0.023	-0.023
	a2	1.133	0.977	0.977	0.977	0.977
S_s (m^2/kg)	a3	0.247	-0.124	-0.124	-0.124	-0.102
	a4	1.281	0.883	0.883	0.883	0.903
C_2S (%)	a5	-0.231	0.183	0.183	0.185	0.069
	a6	0.794	1.201	1.201	1.204	1.072
IR (%)	a7	-0.838	-0.069	-0.070	-0.088	0.200
	a8	0.433	0.933	0.933	0.916	1.222
MgO (%)	a9	2.155	-0.571	-0.572	-0.599	-0.208
	a10	8.630	0.565	0.564	-0.550	0.812
Free CaO (%)	a11	1.228	1.340	1.346	0.488	-0.519
	a12	3.415	3.818	3.844	1.629	0.595
Soundness (%)	a13	1.042	0.891	0.797	1.255	-0.222
	a14	2.834	2.438	2.220	3.507	0.801
FST (min)	a15		-2.987	-2.908	-0.196	1.175
	a16		0.050	0.055	0.822	3.237
SO3 (%)	a17		2.247	4.059	1.149	-0.527
	a18		9.455	57.926	3.156	0.590
SM	a19			1.522	-3.574	1.219
	a20			4.580	-0.028	3.385
Alkalies (%)	a21				5.058	-4.263
	a22				157.322	0.014
C3A (%)	a23				5.282	-1.393
	a24				196.806	0.248
IM	a25					1.086
	a26					2.963
IST (min)	a27					-10.873
	a28					0.0001
LOI (%)	a29					6.626
	a30					754.354
C4AF (%)	a31					-13.913
	a32					0.0001
Correlation Coefficient (R)		81.820	88.969	89.428	90.372	90.613
Kfold Validation		81.990	88.841	89.331	90.418	90.695
Standard Deviation		4.540	3.009	2.755	2.309	1.831
Error		±2.0480	±1.5746	±1.5404	±1.4899	±1.4805

To determine which input variable significantly contributes to the equation and provide more accurate strength estimates, it is crucial to focus on evaluating the correlation coefficient and the standard error.

Additionally, it is noteworthy that the correlation coefficient increases as the standard error decreases from model (1) to model (5). Furthermore, there is a substantial reduction in the standard deviation, nearly halving from 4.540 to 1.831.

4. Conclusion

The research findings reported in this study led to the following conclusions being made:

- Correlation analysis reveals that certain input parameters display high correlation coefficients exceeding 0.6, indicating the presence of multicollinearity issues. Therefore, Principal Component Analysis (PCA) has been effectively utilized to reduce the dimensionality of the feature space and address multicollinearity problems, thereby enhancing the robustness of predictive models.
- Nonlinear regression or fractional polynomial regression models have demonstrated superior performance compared to linear regression models, indicating the importance of capturing nonlinear relationships between input variables and compressive strength, exhibiting an outstanding average root mean square value and correlation between the target and output values.
- Ensemble methods such as XGB Regressor have demonstrated exceptional accuracy in predicting compressive strength, achieving a remarkable accuracy of 99.914% on the testing dataset. However, the disparity between this high accuracy on the testing dataset and the slightly lower accuracy of 93.859% on the training dataset suggests potential overfitting, thereby diminishing their suitability for generalization to new data.
- The linear regression model, despite being a commonly used technique, does not adequately capture the underlying patterns in the data for this prediction study. This inadequacy is evident when examining the model's performance metrics and visualizing the regression graph. Notably, the model exhibits signs of overfitting, where it excessively adapts to the noise and fluctuations in the training data, leading to poor generalization of unseen data. As a result, the linear regression model fails to provide accurate predictions and may not be suitable for addressing the complexities inherent in the dataset.
- After comparing the nonlinear machine learning regression models, it is evident that including the variables Silica Modulus (SM), Alkalies, Tricalcium aluminate (C3A), Alumina Ratio (IM), Initial Setting Time (IST), Loss on Ignition (LOI), and Tetracalciumaluminoferrite (C4AF) did not lead to a significant improvement in the correlation coefficient. Therefore, these variables may be excluded from the regression analysis. The correlation coefficient achieved in this case was 88.969, whereas it was 90.613 when these variables were included along with the others. C_3S, Fineness (S_s), C_2S, IR, MgO, Free CaO, and unsoundness are the major factors in the ordinary Portland cement compressive strength machine learning prediction model.
- The current research concludes the critical importance of considering both the chemical composition and physical properties of cement when developing predictive models for compressive strength prediction. By harnessing the power of machine learning algorithms, the study has revealed a more efficient and cost-effective alternative to traditional experimental analysis methods, enabling faster and more accurate predictions of cement compressive strength.
- This advancement not only expedites the process of materials testing but also enhances our understanding of the intricate relationships between input variables and compressive strength, thereby facilitating the optimization of cement manufacturing processes. Moreover, the successful application of machine learning models in predicting cement compressive strength opens new avenues for further advancements in materials science and engineering, offering opportunities to enhance the sustainability, durability and performance of concrete structures.
- In essence, this research contributes to the ongoing evolution of predictive modelling in materials science. It lays the groundwork for future studies aimed at advancing the field and addressing pressing challenges in the construction industry.

4.1. Recommendations

While this study achieved promising results, it is essential to acknowledge its limitations. It is recommended to explore the following aspects.

- Explore advanced machine learning algorithms: Investigate deep learning models and neural networks to enhance the accuracy and robustness of the predictive model. Experiment with different architectures, including optimal numbers of layers and neurons, to effectively capture nonlinear relationships and improve prediction performance.
- Validate the model on independent datasets: Verify the performance of the developed predictive model using independent datasets sourced from diverse geographical locations or manufacturing facilities.
- This validation process will assess the model's generalizability and applicability across different contexts, ensuring reliable predictions in real-world scenarios.
- Collaborate with industry partners: Establish partnerships with industry stakeholders to deploy the predictive model

in practical applications within the cement industry. Validate the model's performance in real-world settings, facilitate technology transfer, and support adoption by industry professionals, ultimately enhancing efficiency and productivity in cement production and usage.

Acknowledgement

The authors express sincere gratitude to the Pan African University of Sciences, Technology, and Innovation (PAUSTI) for their steadfast and invaluable support throughout the duration of this research endeavour.

References

[1] Mohammed S. Imbabi, Collette Carrigan, and Sean McKenna, "Trends and Developments in Green Cement and Concrete Technology," *International Journal of Sustainable Built Environment*, vol. 1, no. 2, pp. 194-216, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[2] S.N. Ghosh, *Advances in Cement Technology: Chemistry, Manufacture and Testing*, Taylor and Francis, pp. 1-828, 2002. [Google Scholar] [Publisher Link]

[3] Farzad Naseri et al., "Experimental Observations and SVM-based Prediction of Properties of Polypropylene Fibres Reinforced Self-Compacting Composites Incorporating Nano-CuO," *Construction and Building Materials*, vol. 143, pp. 589-598, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[4] Xiangyang Xu, Nasim Fallahi, and Hao Yang, "Efficient CUF-Based FEM Analysis of Thin-Wall Structures with Lagrange Polynomial Expansion," *Mechanics of Advanced Materials and Structures*, vol. 29, no. 9, pp. 1316-1337, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[5] E. Vintzileou, and E. Panagiotidou, "An Empirical Model for Predicting the Mechanical Properties of FRP-Confined Concrete," *Construction and Building Materials*, vol. 22, no. 5, pp. 841-854, 2008. [CrossRef] [Google Scholar] [Publisher Link]

[6] Mostafa Jalal, "Soft Computing Techniques for Compressive Strength Prediction of Concrete Cylinders Strengthened by CFRP Composites," *Science and Engineering of Composite Materials*, vol. 22, no. 1, pp. 97-112, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[7] Ian Flood, "Towards the Next Generation of Artificial Neural Networks for Civil Engineering," *Advanced Engineering Informatics*, vol. 22, no. 1, pp. 4-14, 2008. [CrossRef] [Google Scholar] [Publisher Link]

[8] Vanessa Nilsen et al., "Prediction of Concrete Coefficient of Thermal Expansion and Other Properties Using Machine Learning," *Construction and Building Materials*, vol. 220, pp. 587-595, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[9] Mohit Verma, A. Thirumalaiselvi, and J. Rajasankar, "Kernel-based Models for Prediction of Cement Compressive Strength," *Neural Computing and Applications*, vol. 28, pp. 1083-1100, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[10] Xibin Dong et al., "A Survey on Ensemble Learning," *Frontiers of Computer Science*, vol. 14, pp. 241-258, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[11] Cha Zhang, and Yunqian Ma, *Ensemble Machine Learning: Methods and Applications*, Springer, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[12] Tzu-Tsung Wong, and Po-Yang Yeh, "Reliable Accuracy Estimates from K-Fold Cross Validation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 8, pp. 1586-1594, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[13] Isaac Kofi Nti, Owusu Nyarko-Boateng, and Justice Aning, "Performance of Machine Learning Algorithms with Different K Values in K-Fold Cross-Validation," *International Journal of Information Technology and Computer Science*, vol. 13, no. 6, pp. 61-71, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[14] Jacob Berlin, and Amihai Motro, "Database Schema Matching Using Machine Learning with Feature Selection," *International Conference on Advanced Information Systems Engineering*, pp. 452-466, 2002. [CrossRef] [Google Scholar] [Publisher Link]

[15] Peshawa Jammal Muhammad Ali, and Rezhna Hassan Faraj, "*Data Normalization and Standardization: A Technical Report*," Machine Learning Technical Reports, vol. 1, no. 1, pp. 1-6, 2014. [CrossRef] [Google Scholar]

[16] Kelsy Cabello-Solorzano et al., "The Impact of Data Normalization on the Accuracy of Machine Learning Algorithms: A Comparative Analysis," *18th International Conference on Soft Computing Models in Industrial and Environmental Applications*, pp. 344-353, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[17] Zheng Xiong et al., "Evaluating Explorative Prediction Power of Machine Learning Algorithms for Materials Discovery Using K-Fold Forward Cross-Validation," *Computational Materials Science*, vol. 171, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[18] Chris Albon, *Machine Learning with Python Cookbook: Practical Solutions from Preprocessing to Deep Learning*, O'Reilly Media, pp. 1-366, 2018. [Google Scholar] [Publisher Link]

[19] Michael Greenacre et al., "Principal Component Analysis," *Nature Reviews Methods Primers*, vol. 3, no. 22, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[20] Aurélien Géron, *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow*, O'Reilly Media, pp. 1-856, 2022. [Google Scholar] [Publisher Link]

[21] T. Chai, and R.R. Draxler, "Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)?–Arguments against Avoiding RMSE in the Literature," *Geoscientific Model Development*, vol. 7, no. 3, pp. 1247-1250, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[22] Davide Chicco, Matthijs J. Warrens, and Giuseppe Jurman, "The Coefficient of Determination R-Squared is more Informative than SMAPE, MAE, MAPE, MSE and RMSE in Regression Analysis Evaluation," *PeerJ Computer Science*, vol. 7, PP. 1-24, 2021. [CrossRef] [Google Scholar] [Publisher Link]