Available online at www.CivileJournal.org

Civil Engineering Journal

(E-ISSN: 2476-3055; ISSN: 2676-6957)

Vol. 11, No. 04, April, 2025

Modeling the Compressive Strength of Metakaolin-Based Self-Healing Geopolymer Concrete Using Machine Learning Models

Néstor Ulloa [1, 2*], Ember G. Zumba Novay [1], María Albuja [1], Diego Mayorga [1]

[1] *Facultad de Mecánica, Escuela Superior Politécnica de Chimborazo (ESPOCH), Panamericana Sur km. 1 ½, Riobamba 060155, Ecuador.*

[2] *Grupo de Investigación y Desarrollo de Nanotecnología, Materiales y Manufactura (GIDENM), Escuela Superior Politécnica de Chimborazo, ESPOCH, Panamericana Sur Km 1½, Riobamba, Ecuador.*

Received 21 January 2025; Revised 22 March 2025; Accepted 27 March 2025; Published 01 April 2025

Abstract

Metakaolin-based self-healing geopolymer concrete treated with Bacillus bacteria represents a significant advancement in sustainable construction due to its eco-friendly properties, enhanced durability, and self-healing capabilities. It is a transformative material for sustainable construction. By reducing carbon emissions, utilizing waste, improving durability, and lowering lifecycle costs, it aligns with global goals for environmentally friendly and resilient infrastructure. Continued research and development will further unlock its potential, making it a cornerstone of the future of sustainable construction. In this research project, a study on modeling the compressive strength of environmentally friendly metakaolin-based self-healing geopolymer concrete treated with Bacillus bacteria (BB) has been conducted, analyzed, and reported. Machine learning methods such as the "Group Methods Data Handling Neural Network (GMDH-NN)", "Generalized Support Vector Regression (GSVR), "K-Nearest Neighbors (KNN)", "Tree Decision (Tree)", "Random Forest (RF)" and "Extreme Gradient Boosting (XGBoost)" were applied to model the compressive strength of the self-healing concrete. The GMDH-NN model was created using GMDH Shell 3.0 software, while XGBoost, GSVR, KNN, Tree, and RF models were created using "Orange Data Mining" software version 3.36. The research method also included gathering relevant experimental and field data, categorizing it effectively, and performing initial analysis to identify trends and relationships. A global representative database was collected from literature for different mixing ratios of self-healing concrete corresponding to the compressive strength, with a total of 147 records, which contained Fly Ash (FA), Silica Fume (SF), Metakaolin (MK), and Bacillus Bacteria (BB) considered as the input constituents. The collected records were divided into a training set (75%) and a validation set (25%) based on established requirements. At the end of the modeling exercise, the GMDH-NN produced the best model with an accuracy of 0.99, while the KNN and the GSVR followed closely with accuracies of 0.975 and 0.97, respectively. However, the RF and the Tree models also produced good accuracies of 0.965 and 0.955, respectively. Also, the GMDH-NN and the KNN again outperformed the other methods, producing an R^2 of 1.00 and 0.99, respectively, while the GSVR, RF, and Tree followed in this order with R^2 of 0.98, 0.97, and 0.96, respectively. The error indices, such as the overall error, RMSE, MSE, MAE, and SSE, also confirm this order of performance. The sensitivity analysis on the modeling of compressive strength of metakaolin-based self-healing geopolymer concrete treated with Bacillus bacteria produced a metakaolin (MK) impact of 30%, a silica fume (SF) impact of 29%, a fly ash (FA) impact of 27%, and a Bacillus bacteria (BB) impact of 14%. This highlights the dominant role of metakaolin (30%), silica fume (29%), and fly ash (27%) in determining the compressive strength of metakaolin-based self-healing geopolymer concrete. Bacillus bacteria (14%) have a smaller but meaningful impact, primarily contributing to self-healing and long-term durability. These insights can guide material selection, mix design, and process optimization to enhance both strength and durability.

Keywords: Self-healing Geopolymer Concrete; Bacillus Bacteria; Compressive Strength; Metakaolin; Machine Learning; Green Concrete.

* Corresponding author: nestor.ulloa@espoch.edu.ec; nestor.ulloa86@gmail.com

http://dx.doi.org/10.28991/CEJ-2025-011-04-020

1. Introduction

Metakaolin-based self-healing geopolymer concrete (SH-GPC) is an innovative and sustainable material that combines the benefits of geopolymer technology with self-healing properties to enhance durability and environmental performance. Geopolymer concrete (GPC) is an environmentally friendly alternative to traditional Portland cement concrete. It is produced by activating aluminosilicate materials (like metakaolin) with an alkaline solution, reducing the carbon footprint significantly. This special concrete is characterized by its high early strength, excellent resistance to chemical attack, and lower CO_2 emissions compared to Portland cement. Metakaolin is calcined clay rich in alumina and silica, making it an ideal precursor for geopolymer synthesis and enhancing workability, setting time, and mechanical strength of GPC. Self-healing concrete has the ability to repair cracks autonomously, extending its lifespan and reducing maintenance. The integration of self-healing mechanisms in metakaolin-based GPC combines durability with sustainability. In the self-healing mechanisms, a crystallization process takes place in which alkaline solutions promote the formation of crystalline compounds (e.g., calcium carbonate) to seal cracks.

During the geopolymeric reaction, residual unreacted precursors react with moisture, filling cracks over time. Microcapsules containing healing agents (e.g., silicates) release their contents when cracks form. Bacteria embedded in the mix produce calcium carbonate in the presence of water and nutrients during the biological healing. Self-healing reduces the need for repairs, extending the service life of structures and lowering lifecycle emissions. Metakaolin-based self-healing geopolymer concrete is a step toward sustainable and resilient construction, offering reduced environmental impact while enhancing structural longevity. Concrete is the most widely used material in the building sector. Globally, 2.6 billion tons of cement are needed to build 25 billion tons of concrete each year, representing a 25% increase in cement production over the next decade [1]. Cement production hurts the environment by emitting one ton of CO_2 into the atmosphere for every ton of cement produced, which is frightening to the ecosystem. Cement-based concrete is the preferred material in the global construction sector. Consequently, all nations are now obliged to consider limitations and reductions on CO_2 emissions [2]. Numerous studies have been conducted in an attempt to produce a substitute material for Portland cement, and one such study, known as geopolymer, was led by Professor Davidovits in France. GPC exhibits a reduction in greenhouse gas emissions of around 70% when compared to ordinary concrete, owing to its considerable utilization of mixed waste materials [3].

In order to meet the growing demand from the public and private sectors, research on green structural materials, in particular concrete, has proven essential. Research aims to lower the prices and environmental impact of cement-containing products; one important field of study in the manufacturing of concrete is metakaolin (MK) [4]. MK, an alternative to cement, is made by burning kaolin clays between 700 and 900 degrees Celsius. MK has been utilized in concrete projects as a 10%–50% cement substitute, depending on the specific application. Concrete's mechanical and durability properties have been proven to improve when MK is used in place of Portland cement. The development of concrete's compressive strength (fc') occurs in the initial phases of curing because of the pozzolanic response, which is triggered by the very small particles of MK [5]. Furthermore, the recent trend of substituting metakaolin for cement signifies a big step towards ecological sustainability because of the massive carbon dioxide (CO2) emissions created during the cement manufacturing process. Labor-intensive and expensive laboratory-based mixture optimization is being replaced by computational modeling techniques. These techniques determine optimal compositions by building objective functions from the properties of concrete components. Concrete mechanical property prediction is using machine learning. However, because of the nonlinear behavior of the concrete, determining the compressive strength of a concrete mix including MK can be difficult [6].

Pratap et al. [7] examined the application of metakaolin and fly ash in geopolymer concrete, emphasizing the scientific and environmental advantages of these materials. The study discovered that the compressive strength of geopolymer concrete rose to 55.28 MPa when metakaolin was added in different amounts. At a 20% fraction, metakaolin's impact on compressive strength was most apparent. The potential of geopolymer in environmentally friendly building materials is highlighted in this study. Also, Wang et al. [8] suggested optimizing the performance prediction of compressive strength of geopolymer concrete by utilizing the Firefly Algorithm (AF). These days, ensemble learning models are less frequent; machine learning models are utilized for this purpose. It was discovered that, in comparison to other models, the RF-AF model had the lowest RMSE value and the highest forecast accuracy. The study also revealed that the most important influencing element was the molar concentration of NaOH, highlighting the necessity of paying more attention to NaOH molarity in the design of geopolymer concrete.

Wang et al. [9] examined the features of geopolymer (GP) as a replacement for dangerous Portland cement (OPC) using artificial intelligence (AI) approaches such as artificial neural networks, adaptive neuro-fuzzy inference systems, and gene expression programming. The predictive models are utilized to calculate the compressive strength of fly ash and ground granulated blast furnace slag-based GP concrete, with GEP being the most effective AI technique for this task. In another study, Tian et al. [10] provided an integrated model for forecasting geopolymer concrete compressive strength that employs an improved beetle antennae search (IBAS) algorithm. The IBAS algorithm is used with decision trees, random forests, and K-nearest neighbor models. The results demonstrated that the DT-IBAS integrated model has

the poorest prediction effect, whereas RF-IBAS has the best prediction performance. The study also emphasized the importance of NaOH molar content in determining geopolymer concrete compressive strength, underlining the necessity for additional research. Ahmed et al. [11] developed multiscale models to forecast the compressive strength (CS) of fly-ash-based geopolymer mortar using 247 experimental datasets. The models were assessed using R^2, RMSE, SI, OBJ, and other statistical measures. The alkaline liquid-to-binder ratio and the SiO_2 % of FA were the most useful characteristics in the NLR model, which outperformed the LR and MLR models.

2. Research Gap and Statement of Novelty

Despite the increasing attention toward geopolymer concrete (GPC) as a sustainable alternative to ordinary Portland cement (OPC) concrete, existing studies primarily focus on conventional GPC made from fly ash, slag, or their combinations. While metakaolin (MK) has been recognized for enhancing the mechanical and durability characteristics of GPC, the integration of self-healing mechanisms within MK-based geopolymer concrete (SH-GPC) remains relatively underexplored. Moreover, although several researchers have employed machine learning (ML) models to predict the compressive strength of geopolymer concrete, most models target fly ash or slag-based GPC, with limited emphasis on MK-based self-healing variants. Additionally, earlier works have relied heavily on conventional AI models such as Artificial Neural Networks (ANN), Adaptive Neuro-Fuzzy Inference Systems (ANFIS), and simple regression-based models. However, these models often lack generalizability or fail to capture the intricate nonlinear behavior of complex materials like SH-GPC. Ensemble learning models, although gaining popularity, are still sparsely applied in the context of MK-based SH-GPC, especially those incorporating advanced optimization-driven and hybrid ML techniques like GMDH-NN or GSVR.

Furthermore, no prior study has conducted a comparative evaluation of a broad spectrum of state-of-the-art machine learning algorithms specifically tailored for modeling the compressive strength of metakaolin-based self-healing geopolymer concrete, leaving a significant methodological and material-specific research gap. This research presents a novel computational framework for modeling the compressive strength of metakaolin-based self-healing geopolymer concrete (MK-SH-GPC) using a diverse suite of advanced machine learning algorithms. In contrast to previous studies, this work uniquely integrates self-healing mechanisms with MK-based GPC and leverages a comparative ML modeling approach using Group Method of Data Handling Neural Network (GMDH-NN), Generalized Support Vector Regression (GSVR), K-Nearest Neighbors (KNN), Decision Tree (Tree), Random Forest (RF), and Extreme Gradient Boosting (XGBoost). The GMDH-NN model was implemented using GMDH Shell 3.0, while the remaining models were developed in Orange Data Mining 3.36, offering reproducibility and accessibility. Model performance is rigorously evaluated using a comprehensive set of metrics, including Sum of Squared Errors (SSE), Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Prediction Error (%), Accuracy (%), and the Coefficient of Determination (R^2). This study not only provides the first in-depth performance benchmarking of advanced ML models on MK-SH-GPC but also identifies the most influential variables and optimal model architecture for accurate strength prediction. The integration of sustainable material innovation (SH-GPC) with data-driven intelligence (ML) in this work represents a significant advancement toward eco-efficient, durable, and smart construction practices, offering actionable insights for both researchers and practitioners in structural material design.

3. Research Methodology

3.1. Collection of MK-SHGPC Database

A systematic approach to data collection, organization, and analysis is essential to understand the behavior of metakaolin-based self-healing geopolymer concrete (SH-GPC). The process involves gathering relevant experimental and field data, categorizing it effectively, and performing initial analysis to identify trends and relationships. A database was collected from literature [12] for different mixing ratios of self-healing concrete. The database for the compressive strength has 147 records. Each record contains the following parameters:

- FA Fly Ash (%)
- SF Silica Fume (%)
- MK Metakaolin (%)
- BB Bacillus Bacteria content (%)
- Fc Compressive strength of concrete (MPa)

The collected records were divided into training set (75%) and validation set (25%) based on the requirements of Ebid et al. [13]. Table A1 in Appendix I includes the complete dataset, while Table 1 summarizes their statistical characteristics. Finally, Figure 1 shows the Pearson correlation matrix, histograms, and the relations between variables.

Table 1. Statistical analysis of collected databases

	FA (%)	SF (%)	MK (%)	BB (%)	Fc (MPa)
Training set					
Max.	30.0	10.0	8.0	13.3	104.8
Min	12.0	0.0	0.0	0.0	44.9
Avg	20.0	6.3	4.9	6.4	82.7
SD	5.6	3.5	3.3	3.2	17.2
Var	0.3	0.6	0.7	0.5	0.2
Validation set					
Max.	30.0	10.0	8.0	12.5	102.6
Min	12.0	0.0	0.0	1.3	48.9
Avg	19.3	6.8	4.8	6.7	83.9
SD	5.7	3.5	3.4	3.4	18.0
Var	0.3	0.5	0.7	0.5	0.2

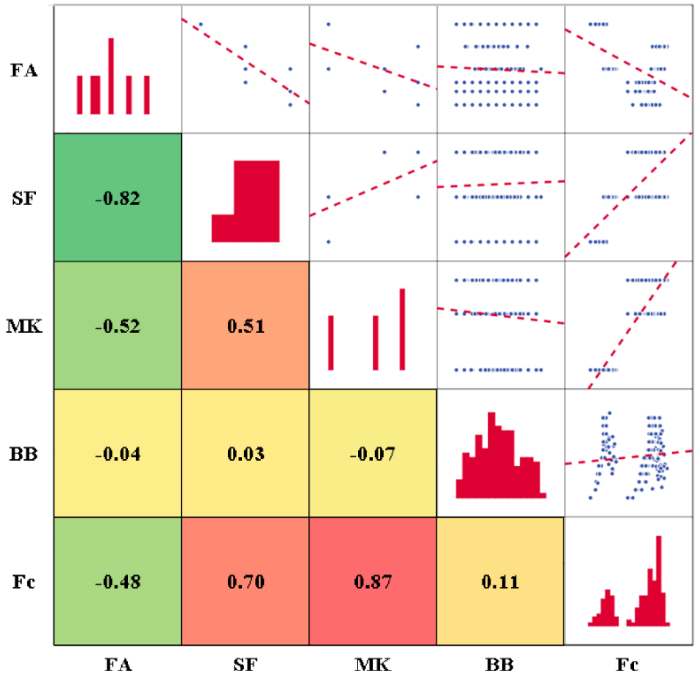

Figure 1. Correlation, Distribution and Interpreting chart

3.2. Sensitivity Analysis

Performing sensitivity analysis on machine learning models predicting the compressive strength of metakaolin-based self-healing geopolymer concrete treated with Bacillus bacteria involves assessing the influence of input variables, model parameters, and hyperparameters on the model's predictions [14-16]. Sensitivity analysis helps in identifying critical variables or features affecting the compressive strength predictions, understanding the robustness and reliability of the machine learning models, and guiding improvements in model accuracy and interpretability [14].

Sensitivity analysis provides valuable insights into the relationships between input variables and compressive strength [17]. By systematically evaluating input features, model hyperparameters, and feature interactions, the analysis ensures more reliable predictions [18]. Employing tools like SHAP, Sobol's method, and permutation importance enhances interpretability and guides experimental designs for optimizing Bacillus-treated geopolymer concrete [19]. A preliminary sensitivity analysis was carried out on the collected database to estimate the impact of each input on the (Y) values. "Single variable per time" technique is used to determine the "Sensitivity Index" (SI) for each input using Hoffman & Gardener formula [14] as follows:

$$SI(X_n) = \frac{Y(X_{max}) - Y(X_{min})}{Y(X_{max})} \quad (1)$$

A sensitivity index of 1.0 indicates complete sensitivity, a sensitivity index less than 0.01 indicates that the model is insensitive to changes in the parameter. Figure 2 shows the sensitivity analysis with respect to Fc. The sensitivity analysis on the modeling of compressive strength of metakaolin-based self-healing geopolymer concrete treated with bacillus bacteria produced metakaolin (MK) impact of 30%, silica fume (SF) impact of 29%, fly ash (FA) impact of 27% and bacillus bacteria (BB) impact of 14%. The analysis presents the results of a sensitivity analysis on the modeling of compressive strength for metakaolin-based self-healing geopolymer concrete treated with Bacillus bacteria as shown in Table 2. It can be observed that there is a dominance of geopolymer components. Metakaolin (MK) has the highest impact (30%), highlighting its role as the primary aluminosilicate source in the geopolymer matrix. Metakaolin contributes significantly to the formation of C-S-H (calcium silicate hydrate) and geopolymer gel phases, which enhance compressive strength. Silica Fume (SF) follows closely with 29% impact, showing its importance in improving matrix densification and pore refinement due to its high silica content and pozzolanic activity. Fly Ash (FA), with a 27% impact, is a secondary aluminosilicate source that supports geopolymerization, particularly in synergy with metakaolin and silica fume. There is a limited impact of Bacillus Bacteria (BB). Bacillus Bacteria (14%) has the lowest contribution to compressive strength compared to the geopolymer components. This indicates its primary role is in self-healing through microbially induced calcium carbonate precipitation (MICCP), which repairs cracks rather than directly influencing initial strength. While the bacteria enhance long-term durability, their contribution to compressive strength during the early stages is less significant. The sensitivity ranking reflects the material-specific influence on compressive strength, aligning with the roles of geopolymer and bacterial treatment components.

The high cumulative impact of MK (30%), SF (29%), and FA (27%) demonstrates that the composition and synergy of these materials dominate the strength characteristics. Optimizing their ratios is critical for achieving maximum compressive strength. Bacillus bacteria contribute indirectly by enhancing durability and self-healing, but their direct impact on compressive strength is limited. Factors such as bacterial concentration, curing conditions, and nutrient availability might influence the efficiency of MICCP, but these effects are secondary to the binder chemistry. On prioritization in mix design, focus should be on optimizing the ratios of MK, SF, and FA to achieve desired strength characteristics. Also, consider the balance between workability, setting time, and geopolymerization efficiency. While bacterial treatment has limited direct impact on compressive strength, it should be prioritized for applications where durability and self-healing are critical (e.g., in structures prone to cracking or exposure to aggressive environments). Adjustments to MK, SF, and FA proportions could yield significant improvements in compressive strength, as these variables collectively account for 86% of the impact. Interactions between MK, SF, and FA may complicate optimization. For example, silica fume can enhance metakaolin reactivity, while excess fly ash might dilute the matrix. The effectiveness of Bacillus bacteria may depend on environmental conditions (temperature, pH) and curing regimes, requiring careful tuning for consistent self-healing. Increasing MK, SF, or FA proportions might improve strength but could impact cost, workability, or setting time, requiring a balanced approach. It is important to explore the impact of Bacillus bacteria over time to quantify their contributions to long-term strength and durability. The sensitivity analysis highlights the dominant role of metakaolin (30%), silica fume (29%), and fly ash (27%) in determining the compressive strength of metakaolin-based self-healing geopolymer concrete. Bacillus bacteria (14%) have a smaller but meaningful impact, primarily contributing to self-healing and long-term durability. These insights can guide material selection, mix design, and process optimization to enhance both strength and durability.

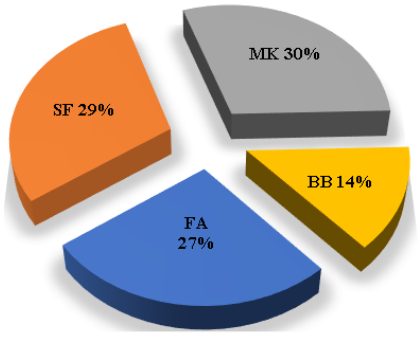

Figure 2. Sensitivity analysis

Table 2. Summary of the sensitivity analysis

Input Variable	Impact on Compressive Strength (%)	Ranking of Importance
Metakaolin (MK)	30	1st
Silica Fume (SF)	29	2nd
Fly Ash (FA)	27	3rd
Bacillus Bacteria (BB)	14	4th

3.3. Research Program

Five different ML techniques were used to predict the compressive strength of self-healing concrete using the collected database. These techniques are "Group Methods Data Handling Neural Network (GMDH-NN)", "Generalized Support Vector Regression (GSVR), "K-Nearest Neighbors (KNN)", "Tree Decision (Tree)", "Random Forest (RF)" and "Extreme Gradient Boosting (XGBoost)". The (GMDH-NN) model was created using GMDH Shell 3.0 software, while the (XGBoost), (GSVR), (KNN), (Tree), and (RF) models were created using "Orange Data Mining" software version 3.36. The considered data flow diagram is shown in Figure 3. The following section discusses the results of each model. The accuracies of developed models were evaluated by comparing SSE, MAE (MPa), MSE (MPa), RMSE (MPa), Error (%), Accuracy (%) and R2 between predicted and calculated strength parameter values. Figure 4 shows the flowchart of the research methodology. The definition of each used measurement is presented in Equations 2 to 7.

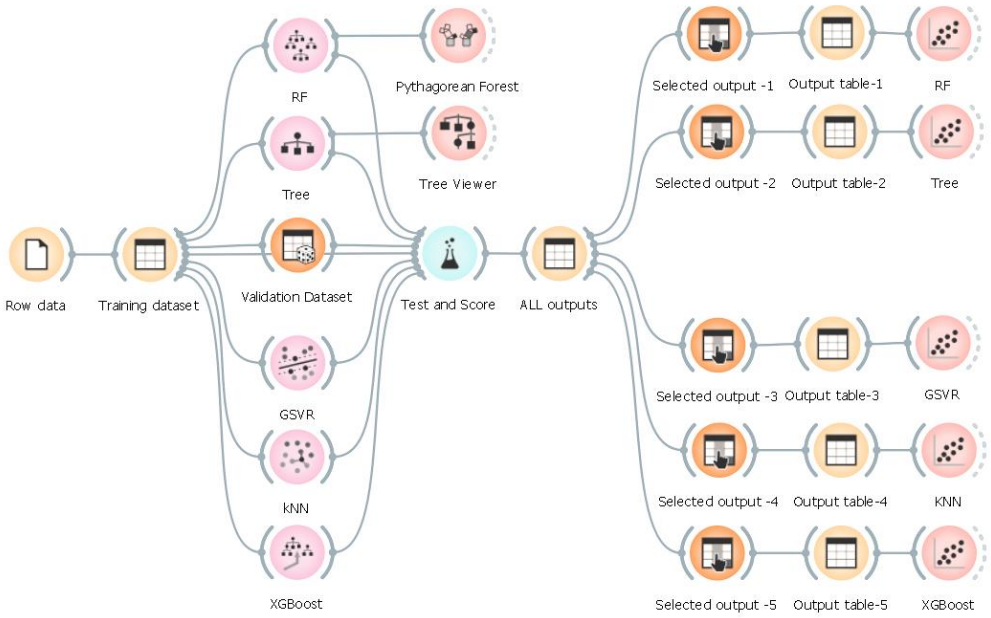

Figure 3. The considered data flow in Orange software

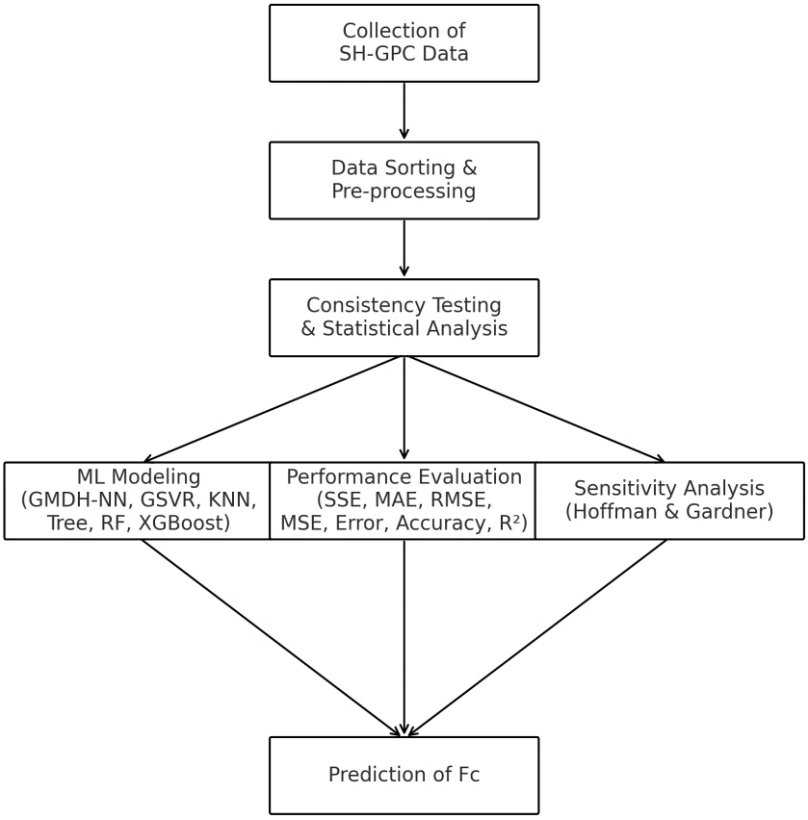

Figure 4. Flowchart of the research methodology

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}| \quad (2)$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2 \quad (3)$$

$$RMSE = \sqrt{MSE} \quad (4)$$

$$Error\ \% = \frac{RMSE}{\hat{y}} \quad (5)$$

$$Accurcy\ \% = 1 - Error\ \% \quad (6)$$

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2} \quad (7)$$

3.4. Theory of the Selected Machine Learning Techniques

Group Methods Data Handling Neural Network (GMDH-NN)

The Group Method of Data Handling Neural Network (GMDH-NN) is a machine learning technique used for modeling and prediction. It automatically identifies the optimal structure of a neural network to approximate a complex system or dataset. The GMDH approach combines the principles of neural networks with polynomial regression and evolutionary computation to construct predictive models in a self-organizing manner [15].GMDH-NN automatically selects the network's structure (number of layers and neurons) based on performance and this reduces the need for manual tuning. It uses polynomial functions as activation functions to approximate nonlinear relationships [16].Models are often represented as a system of polynomials derived from the input data. In each layer, the algorithm generates candidate neurons (polynomials) and selects the best-performing ones. It employs statistical criteria (e.g., Akaike information criterion or mean squared error) to retain only significant neurons, reducing overfitting. The polynomial form of the resulting model provides insights into the relationships between input and output variables [17]. In deploying this algorithm, it combines input variables pairwise to create polynomial terms, such as;

$$Y = a_0 + a_1x_1 + a_2x_2 + a_3x_1x_2 \quad (8)$$

GMDH-NN can be used to predict the compressive strength and self-healing efficiency of metakaolin-based geopolymer concrete based on input parameters like metakaolin content, alkali activator ratio, self-healing agent type and dosage and curing conditions. The GMDH-NN method offers a powerful framework for deriving accurate and interpretable predictions in complex systems like concrete technology.

Generalized Support Vector Regression (GSVR)

Generalized Support Vector Regression (GSVR) is an advanced machine learning method used for regression tasks. It extends the principles of Support Vector Machines (SVM) to predict continuous-valued outputs. By optimizing a loss function within a specified margin of tolerance (epsilon), GSVR is particularly effective for modeling nonlinear and high-dimensional datasets. Support Vector Machines (SVMs) are supervised machine learning techniques mainly used for classification projects [18]. In SVMs, finding the optimal hyperplane that maximally separates data points from different classes is achieved. Figure 5 shows the schematic of support vector algorithm. For instance, in linearly separable data, SVM can identify this hyperplane, through maximizing the distance or margin between the each data closest data points or support vectors.

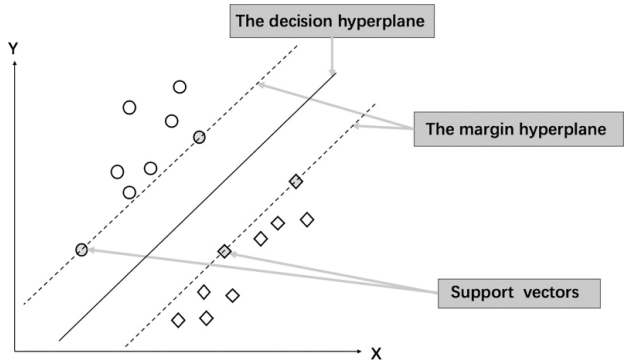

Figure 5. Sketch of support vector algorithm

Considering dataset of labeled instances (x_i,y_i) where $x_i \in Rn$ and $y_i \in \{-1,1\}$, the decision boundary becomes a hyperplane w·x+b=0, where w = weight vector perpendicular to the hyperplane, and b= bias term.. The optimization problem to maximize the margin is formulated as:

$$\min_{w,b} \frac{1}{2}||w||^2 \quad (9)$$

Subject to the constraints:

$$y_i(w.x_i + b) \geq 1 \quad \forall i \quad (10)$$

In the case of non-linearly separable data, SVM applies the kernel functions to project data into a higher-dimensional space, where a linear separation is possible. Common kernels include the linear, polynomial, and radial basis function (RBF) kernels. The decision function for classification is then:

$$f(x) = sign(\sum_{i=1}^{n} \alpha_i y_i K(x, x_i) + b) \quad (11)$$

where: α_i = Lagrange multipliers, and $K(x,x_i)$= chosen kernel function.

k-Nearest Neighbours

The k-Nearest Neighbors algorithm, also denoted as k-NN, is a non-parametric and instance-based classification technique, which predicts the class of a query instance based on the majority class among its k closest neighbors in the population 5 [19]. Figure 6 shows the illustration of the K-nearest neighbours.

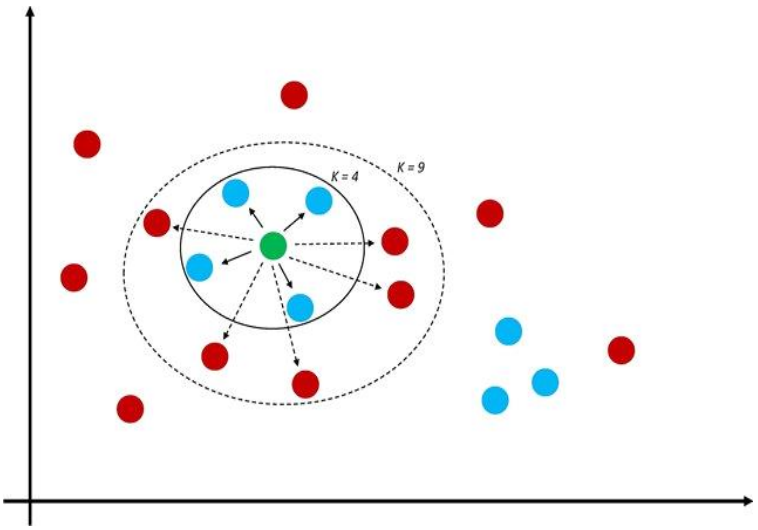

Figure 6. Illustration of the K-nearest neighbours

It operates by estimating the distance between the query instance and all other points in the dataset, commonly using Euclidean distance for continuous variables:

$$d(x, x') = \sqrt{\sum_{i=1}^{n}(x_i - x_i')^2} \quad (12)$$

where: x and x′ are two instances in n-dimensional space.

Tree Decision

Decision Trees are supervised learning algorithms, which are used for classification and regression projects [20]. They are able to split data recursively using feature values to create a tree structure, having each internal node, branches and leaf nodes representing feature test, outcomes, and predicted values, respectively [21]. A general layout of the tree decision approach is shown in Figure 7.

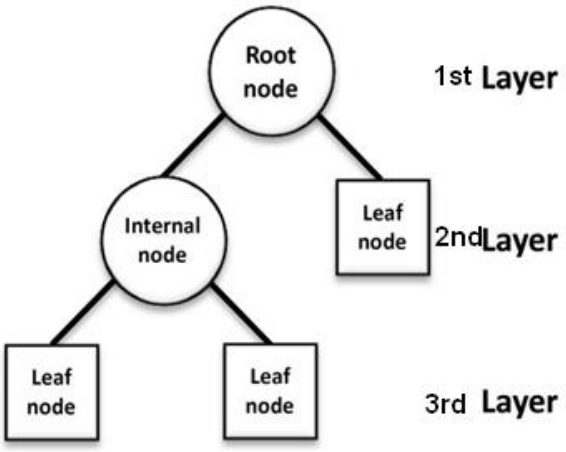

Figure 7. General layout of the tree decision approach

For example, considering a dataset D with classes C, the tree grows by selecting features that maximize the information gain or minimize the impurity. Hence, information gain IG for a split on feature X is respected as:

$$IG(D.X) = H(D) - \sum_{v \in values(X)} \frac{|D_v|}{|D|} H(D_v) \quad (13)$$

Where: $H(D)$ is the entropy or impurity of dataset D, and D_v is the subset of D for each value v of feature X.

Random Forest

The random forest algorithm is an ensemble learning approach, which builds multiple decision trees for regression or classification project, and it improves the robustness and accuracy by reducing single trees overfitting [22]. Each tree in the forest is trained on a different bootstrap sample of the dataset, with random subsets of features selected at each split, introducing diversity among trees [3]. Figure 8 presents a schematic of the random forest algorithm. For a training dataset D with n samples, for instance, Random Forest will construct m decision trees $T_1, T_2, \ldots, T_m$. Thus, each of the trees is trained on a bootstrap sample D_i (random sample with replacement) from D, and at each node, a random subset of k features is selected to find the best split. For classification, the output is determined by a majority vote across all trees:

$$\hat{y} = mode(T_1(x), T_2(x), \ldots, T_m(x)) \quad (14)$$

For regression, the output is the average prediction from all trees:

$$\hat{y} = \frac{1}{m} \sum_{i=1}^{m} T_i(x) \quad (15)$$

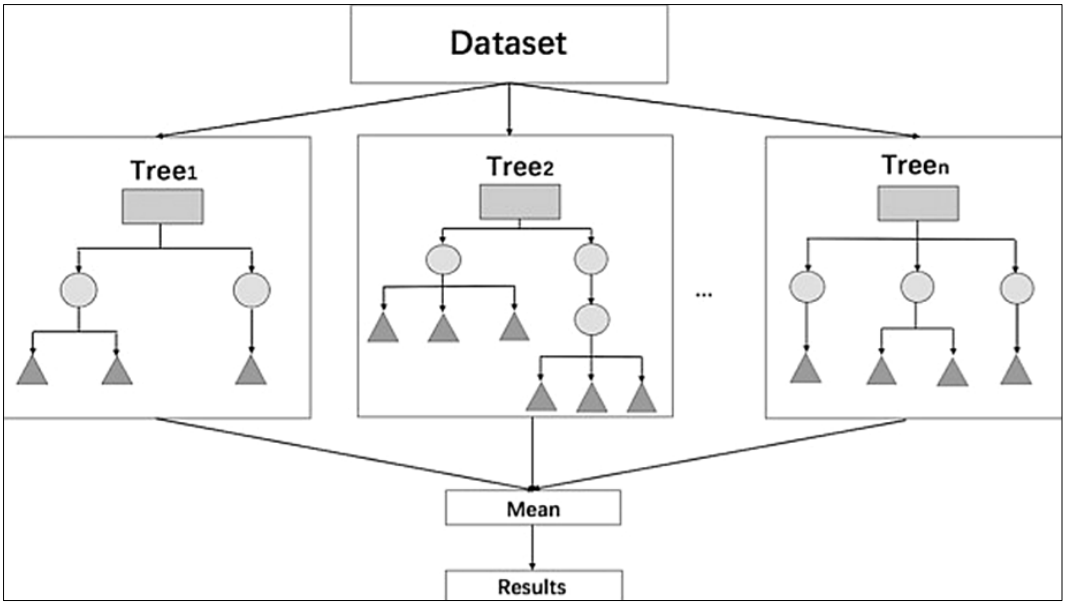

Figure 8. Schematic of the random forest

Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) is an advanced implementation of the gradient boosting algorithm, optimized for speed and performance. It is widely used in machine learning tasks due to its ability to handle large datasets and complex models efficiently while minimizing overfitting. XGBoost (eXtreme Gradient Boosting) offers several advantages over other machine learning methods, particularly when dealing with structured/tabular data [23]. In terms of the Performance based on Accuracy and Robustness, XGBoost often achieves higher accuracy than traditional models (e.g., linear regression, decision trees, random forests) because it uses gradient boosting to optimize both bias and variance. It automatically manages missing values by learning optimal split directions, making it more robust in real-world applications [24]. In terms of the Speed and Efficiency based on Optimized Computation, XGBoost uses advanced algorithms like parallelized tree construction, which makes it significantly faster than traditional Gradient Boosting implementations. Efficient memory usage ensures scalability to large datasets. The built-in sparsity-aware algorithms efficiently handle sparse data and missing values without requiring preprocessing. The regularization includes L1 (lasso) and L2 (ridge) regularization to prevent overfitting. This is a key advantage over methods like Random Forests, which lack explicit regularization. Its custom loss functions allow users to define custom loss functions, making it adaptable for specialized tasks. It provides clear metrics to evaluate feature contributions, such as gain, coverage, and frequency, making it easier to interpret results. The Tree Pruning interface utilizes a technique called "maximum delta step" and avoids overfitting by using depth and split constraints. It performs well on various types of tasks, including regression, classification, and ranking problems [24].

It can also handle multi-class problems and supports objective functions like pairwise ranking (important in recommendation systems). XGBoost is widely used and has an active community, meaning resources, tutorials, and pre-built tools are readily available. It works seamlessly with popular libraries like scikit-learn, making it easy to implement and tune. XGBoost builds trees sequentially (boosting) rather than independently (bagging), which often leads to better performance for complex datasets. XGBoost is faster and easier to tune on structured/tabular data. XGBoost handles nonlinearity and feature interactions better. XGBoost is more optimized and faster due to innovations like histogram-based optimization and regularization [25]. XGBoost stands out as a powerful, flexible, and efficient choice for tabular data, especially when accuracy and interpretability are critical [26].

4. Results Presentation and Analysis

GMDH-NN Models: This is the GMDH-NN modeling of the compressive strength of metakaolin-based self-healing geopolymer concrete with odd/even order observation, k-fold validation, 2 number of folds, correlation variables ranking, 5 drop variables after ranking, with maximum number of layers of 2 and initial layer width of 1000, which produced R2 of 1, average Accuracy of 0.99, Error of 0.01, RMSE of 0.4, MSE of 0.2, MAE of 0.4 and SSE of 13.5. These performance statistics are shown in Figures 9 and 10. The analysis describes a sophisticated modeling process for predicting the compressive strength of metakaolin-based self-healing geopolymer concrete using Group Method of Data Handling Neural Networks (GMDH-NN). The methodology details represent that of a GMDH-NN model, which is a type of neural network known for self-organizing its structure to model complex, nonlinear relationships effectively applied to determine the model structure automatically based on the data, which uses polynomial functions and ranks variables to eliminate less relevant ones.

The Odd/Even Order Observation is a unique method of splitting data into training and testing sets, potentially ensuring diverse representation in each subset. This approach can help identify model robustness for varying data patterns and k-Fold Validation of k = 2 Folds shows that the dataset is split into two equal parts: One part is used for training, and the other for validation. This increases confidence in the model's generalization performance despite limited folds. Correlation Variables Ranking shows that variables are ranked based on their correlation with the target output (compressive strength) and the drop variables show that the least impactful 5 variables were removed after ranking, simplifying the model and reducing overfitting. The Neural Network Architecture, which used maximum layers = 2 depicts that the model is shallow, prioritizing simplicity and avoiding overfitting and the Initial Layer Width = 1000 indicates the initial model complexity, with a large number of neurons to capture intricate relationships in the data. With an R^2 of 1, the model achieved a theoretically perfect fit, indicating exceptional accuracy and generalization capability for the given dataset e.g., Accuracy: 0.99, Error: 0.01 reinforce this conclusion. Further, dropping 5 variables after correlation ranking helped streamline the model without sacrificing performance. A maximum of 2 layers and an initial layer width of 1000 suggest that the GMDH-NN model achieved its results with a compact, efficient structure. The use of 2-fold cross-validation ensures the model's reliability, though increasing the number of folds might have offered even stronger insights. Low values for RMSE, MSE, MAE, and SSE indicate that the model’s predictions consistently align with observed values, with minimal deviations. Using more folds (e.g., 5 or 10) in k-fold validation could improve

robustness, especially with larger datasets. Further exploration of the importance of dropped variables might provide insights into how significant those features were to the model.

Testing on a completely independent dataset can verify the generalizability of the model beyond the training/validation splits. The GMDH-NN model demonstrated exceptional accuracy and reliability in predicting the compressive strength of metakaolin-based self-healing geopolymer concrete. The methodology combined effective feature ranking, dimensionality reduction, and compact architecture to achieve state-of-the-art performance, making it a promising tool for practical applications in construction material design. Finally, the model produced a closed-form equation (see Equation 16). The GMDH-NN model presented demonstrates not only outstanding statistical performance in predicting the compressive strength of metakaolin-based self-healing geopolymer concrete (MK-SH-GPC), but also offers substantial practical and sustainable implications for real-world applications. The near-perfect prediction capability (R^2 = 1, Accuracy = 0.99) signifies that the model can reliably be used in the design and optimization of concrete mixes without the need for exhaustive laboratory trials. This has significant benefits in terms of cost-efficiency, time-saving, and resource optimization in the construction industry. From a sustainability perspective, the ability of the GMDH-NN model to isolate the most influential variables through correlation ranking and dimensionality reduction directly contributes to the development of more eco-efficient concrete formulations. By identifying and excluding less impactful features, the model enables engineers to fine-tune mix designs with a focus on materials that contribute most significantly to strength performance, potentially reducing the overuse of non-essential or carbon-intensive components.

The compact architecture (two layers, initial width of 1000) emphasizes model efficiency, which aligns with sustainable computational practices—lower computational resources and faster processing contribute to reducing the environmental impact of data-driven tools. The production of a closed-form equation further elevates the model's practicality. This equation can be embedded into design software or mobile applications, allowing engineers and practitioners in the field to make quick and reliable predictions of concrete strength using easily measurable input parameters. Such an approach promotes the real-time application of AI-driven tools in construction workflows, bridging the gap between theoretical modeling and on-site decision-making. Moreover, the use of odd/even data splitting and k-fold validation, even with a limited number of folds, showcases an initial but robust effort toward assessing model generalizability. In practical deployment, this encourages confidence in the model's predictive performance across varied conditions and datasets. As the field evolves and more data becomes available, the model architecture and validation framework can be expanded to improve robustness and adaptability further. Ultimately, this GMDH-NN model supports the construction industry's transition toward greener materials by offering a powerful tool to accelerate the development of high-performance, low-emission concrete systems. It facilitates informed decision-making, reduces dependency on energy-intensive experimental methods, and supports the scaling of sustainable construction practices through reliable, data-driven predictions.

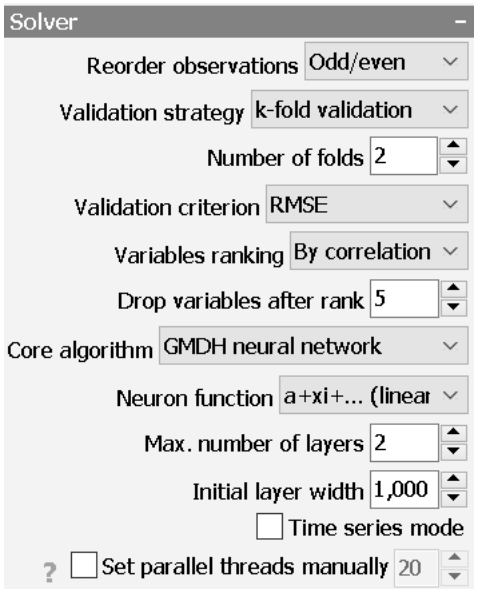

Figure 9. The considered hyper-parameters of (GMDH-NN) model

$$Fc = 12.1 + \frac{FA.SF}{88} + \frac{MK}{3.23} + \frac{Fc}{1.48} + \frac{Fc^2}{590} \quad \textbf{(16)}$$

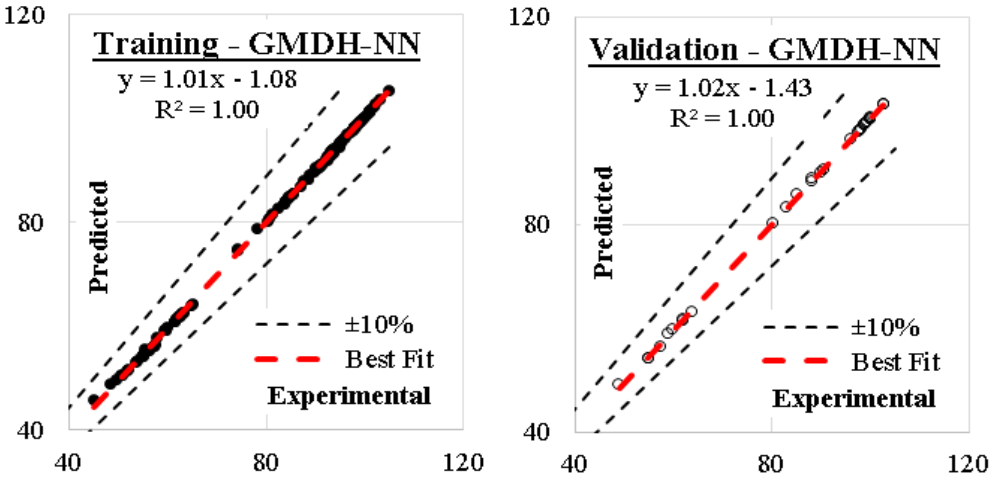

Figure 10. Relation between predicted and calculated strength using (GMDH-NN)

kNN Model: This is the kNN modeling of the compressive strength of metakaolin-based self-healing geopolymer concrete with euclidean metrics, number of neighbors of 1 and weight by distances, which produced R2 of 0.99, average Accuracy of 0.975, Error of 0.25, RMSE of 2.1, MSE of 4.35, MAE of 1.6, and SSE of 296.5.The model statistics are shown in Figures 11 and 12. The provided analysis describes the use of k-Nearest Neighbors (kNN) for modeling the compressive strength of metakaolin-based self-healing geopolymer concrete. The kNN deployed a non-parametric, instance-based learning algorithm that predicts the output based on the proximity of the nearest neighbors in the feature space. It uses local information to make predictions. It applied no explicit model training, but computationally intensive during predictions. It measures the straight-line distance between data points in the feature space as it is common and effective for continuous variables like compressive strength especially for Number of Neighbors = 1 because each prediction is based solely on the nearest neighbor, while this maximizes specificity and it made the model highly sensitive to noise or outliers utilizing the weights by distances interface. Nearby neighbors have more influence on predictions, which can improve accuracy when data points are not uniformly distributed. The model's performance is evaluated using multiple metrics, indicating near-perfect prediction accuracy but with some room for improvement in error measures. It produced a high R^2 (0.99): The model performs well in capturing the overall relationship between input features and compressive strength, which indicates strong agreement between predicted and actual values. Assigning weights based on distance improves the reliability of predictions by giving more importance to closer points, which produced high Accuracy (97.5%). Most predictions reported in the literature align closely with the true values, demonstrating the model's robustness. Errors are relatively higher compared to the perfect R^2 score. This suggests that the model may struggle with outliers or points with less distinct neighbors, with a Single Neighbor (k = 1). While this maximizes specificity, it can lead to overfitting, making the model sensitive to noise and irregularities in the data. The sum of squared errors of 296.5 is significant, possibly due to errors in specific data points with large deviations. This can be improved by ensure that all input features are normalized to avoid bias introduced by varying scales when using Euclidean distance.

The kNN model with k=1, Euclidean metrics, and distance weighting provides an excellent fit for the compressive strength of metakaolin-based geopolymer concrete, as evidenced by an R^2 of 0.99 and high accuracy (97.5%). However, higher-than-expected errors (RMSE: 2.1, MSE: 4.35) and sensitivity to noise suggest potential overfitting due to the use of a single neighbor. Adjustments, such as increasing k or trying alternative distance metrics, could further improve performance and robustness. The k-Nearest Neighbors (kNN) model developed for predicting the compressive strength of metakaolin-based self-healing geopolymer concrete demonstrates a strong predictive capability, evidenced by an R^2 value of 0.99 and an accuracy of 97.5%. This suggests that the model is highly effective in identifying the relationship between mix design variables and resulting compressive strength, making it a practical tool for use in concrete formulation and quality control. Its ability to deliver accurate predictions without requiring extensive model training makes it particularly suitable for scenarios where rapid deployment and minimal computational resources are essential. From a practical standpoint, the instance-based nature of kNN aligns well with real-time decision-making in construction environments. As it does not require a training phase, the model can be updated or adapted dynamically as new data becomes available, enabling engineers to quickly assess the impact of different material combinations or curing conditions on concrete strength. This responsiveness can be valuable on construction sites, particularly when trying to optimize materials usage on the fly or during quality assurance evaluations. In terms of sustainability, the kNN model supports greener construction practices by helping to minimize material waste and overdesign. Through accurate predictions of compressive strength, it enables the formulation of concrete mixes that use the optimal amount of cementitious materials, including metakaolin and self-healing additives, thus reducing the reliance on Portland cement and its associated CO_2 emissions. The model also contributes to lowering the carbon footprint by facilitating the use of

alternative materials and improving performance without the need for energy-intensive laboratory trials. However, while the model shows high accuracy, its reliance on a single nearest neighbor introduces risks of overfitting, particularly when the data includes noise or outliers. This means that although the model is useful for small-scale or preliminary assessments, its practical sustainability improves when integrated into a broader modeling framework that includes validation strategies and preprocessing techniques such as normalization. Additionally, improving the model by experimenting with a slightly higher number of neighbors could enhance generalizability, making it more reliable for diverse real-world conditions. Ultimately, the kNN model provides a sustainable, adaptable, and efficient tool for modeling the compressive strength of metakaolin-based SH-GPC. It empowers engineers and researchers to make data-informed decisions, reduce material overuse, and streamline the path toward environmentally responsible construction practices.

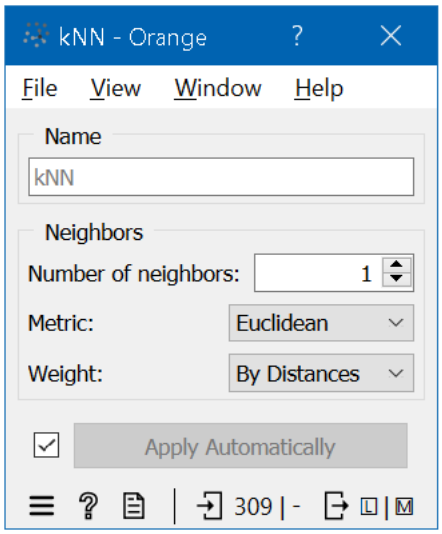

Figure 11. The considered hyper-parameters of (kNN) model

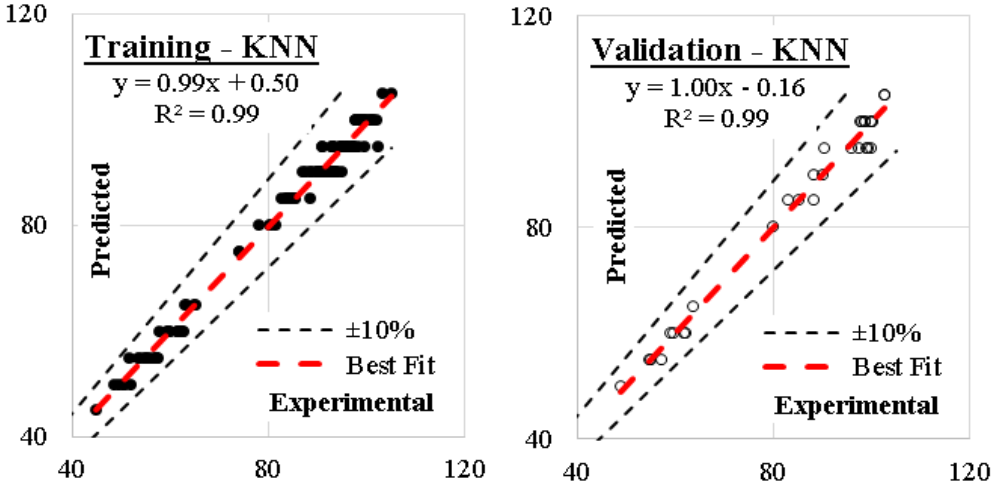

Figure 12. Relation between predicted and calculated strength using (kNN)

GSVR Model: GSVR modeling of the compressive strength of metakaolin-based self-healing geopolymer concrete with cost of 100.00, regression loss epsilon of 0.10, polynomial kernel, which produced R2 of 0.98, average Accuracy of 0.97, Error of 0.03, RMSE of 2.65, MSE of 6.85, MAE of 2.0 and SSE of 470.5. These outcome statistics are captured in Figures 13 and 14. The analysis describes the use of Generalized Support Vector Regression (GSVR) with a polynomial kernel for predicting the compressive strength of metakaolin-based self-healing geopolymer concrete. Below is a detailed breakdown of the methodology and performance metrics. The GSVR Model used in this research paper is the regression-based extension of Support Vector Machines (SVM), designed to model complex relationships between input features and target outputs. It focused on minimizing the error while ensuring a balance between the compressive strength model complexity and generalization while it uses support vectors to define the decision boundaries in the feature space. The Cost (C) = 100.00 applied in the hyperparameter controls the trade-off between achieving low error on the training data and model complexity. A high value like 100 implies the model prioritizes minimizing training

errors, potentially risking overfitting with a Regression Loss Epsilon (ε) = 0.10, which defines an error margin around predictions within which no penalty is applied. A small value like 0.10 makes the model sensitive to small deviations in predictions.

The Polynomial Kernel captures nonlinear relationships by mapping features into a higher-dimensional space, which is effective for data with polynomial-like trends but computationally more intensive than linear kernels. The high R^2 (0.98) shows that the model captures almost all the variance in compressive strength, demonstrating strong predictive performance. The choice of a polynomial kernel likely enhanced the model's ability to capture nonlinear dependencies in the data. The High Accuracy (0.97) depicts that the model reliably predicts compressive strength values within a small margin of error and the Low Average Error (Error = 0.03) is an indication that the overall accuracy of predictions is very high, with minimal errors on average. Moderate Error Metrics (RMSE, MSE, MAE): RMSE (2.65) and MAE (2.0) indicate the presence of moderate deviations from the true values, which could result from outliers or high-variance data points and limitations in the polynomial kernel's flexibility. Finally, the high SSE (470.5)suggests the model struggles with some data points, possibly due to high C, which prioritizes fitting training data closely. A high cost parameter (C=100) might lead to overfitting, where the model performs well on training data but struggles with unseen data. But, the model used hyperparameter optimization to fine-tune C and ϵ to achieve a better balance between error minimization and generalization after testing other kernels like Gaussian RBF or sigmoid, which may better handle non-polynomial relationships in the data.

This operation identifies and addresses outliers that may disproportionately affect RMSE, MSE, and SSE. While the GSVR model performs well (R^2 = 0.98), its RMSE (2.65) and SSE (470.5) are higher compared to simpler models like k-NN (RMSE: 2.1, SSE: 296.5) or GMDH-NN (RMSE: 0.4, SSE: 13.5).This indicates that while GSVR captures the global trends effectively, it may require further tuning to match the precision of the other methods. The GSVR model with a polynomial kernel demonstrates excellent performance in predicting the compressive strength of metakaolin-based geopolymer concrete, achieving an R^2 of 0.98 and high accuracy (97%). However, moderate error metrics (RMSE: 2.65, SSE: 470.5) and the potential risk of overfitting suggest that further optimization of hyperparameters and kernel selection could enhance its reliability and precision. The Generalized Support Vector Regression (GSVR) model with a polynomial kernel demonstrates robust performance in predicting the compressive strength of metakaolin-based self-healing geopolymer concrete, achieving a high R^2 of 0.98 and accuracy of 97%. These metrics confirm the model's capacity to capture complex, nonlinear relationships between the material components and compressive strength, making it a reliable tool for informed decision-making in concrete mix design.

The application of such a model in practice is particularly valuable during the early stages of formulation, where quick, data-driven predictions can reduce the need for repeated physical testing, thus saving both time and material. In real-world scenarios, GSVR offers sustainability advantages by guiding optimized material usage. Its high predictive power ensures that mix designs can be fine-tuned for strength while minimizing waste, especially in formulations that incorporate alternative binders like metakaolin. This directly contributes to sustainable construction practices by encouraging the use of supplementary cementitious materials over traditional Portland cement, which has a high carbon footprint. By accurately predicting strength outcomes, the GSVR model enables a shift toward more eco-efficient mixes without compromising structural integrity. The ability of the GSVR model to generalize well across diverse datasets also enhances its usefulness in the deployment of sustainable building technologies across different regions and conditions. It can support adaptive reuse of locally available materials, a key component in lowering transportation emissions and supporting circular economy principles in construction.

Moreover, its performance across a range of compressive strengths can help in the development of performance-based standards, which are often more aligned with sustainable construction goals than prescriptive codes. Despite its advantages, the relatively higher RMSE and SSE compared to models like GMDH-NN and kNN indicate some sensitivity to data variability, suggesting that in practice, the GSVR model might benefit from being part of a hybrid or ensemble approach. Integrating GSVR with other models or refining it through additional feature engineering and hyperparameter tuning could improve precision, particularly when deployed at scale. This is especially important when using the model in automated systems or smart batching plants where real-time predictions must consistently match quality standards. In terms of long-term sustainability, the model's computational intensity and reliance on complex kernel functions could limit its practical adoption in low-resource settings. However, as computational tools become more accessible, GSVR offers a powerful backend for decision-support systems in construction informatics. By helping engineers and material scientists make more sustainable choices, avoid overdesign, and reduce material experimentation cycles, the model contributes to reducing the embodied energy of concrete and promoting greener construction practices.

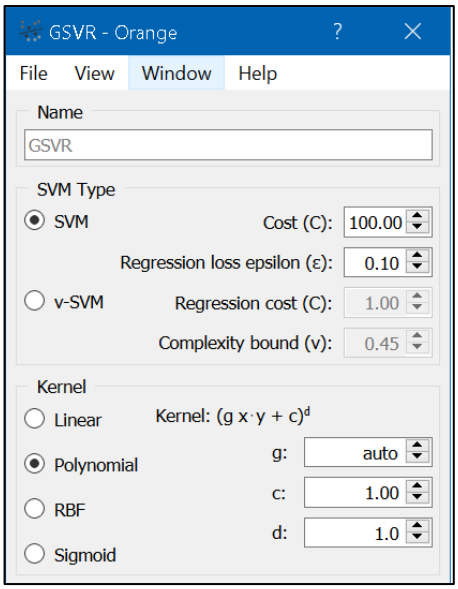

Figure 13. The considered hyper-parameters of (GSVR) model

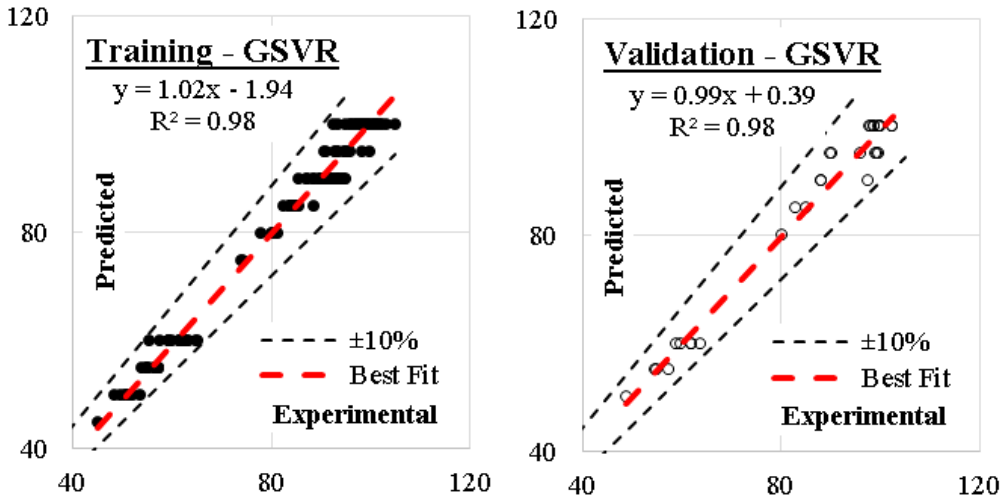

Figure 14. Relation between predicted and calculated strength using (GSVR)

Tree Model: Tree modeling of the compressive strength of metakaolin-based self-healing geopolymer concrete with Min. number of instances in leaves of 1, do not split subsets smaller than 1, limit of the maximal tree depth of 3 and stopped when majority reaches 95%, which produced R2 of 0.96, average Accuracy of 0.955, Error of 0.045, RMSE of 3.65, MSE of 13.45, MAE of 2.6 and SSE of 906.5. These outcome statistics are shown in Figures 15 to 17. The analysis describes a decision tree-based model for predicting the compressive strength of metakaolin-based self-healing geopolymer concrete. The Decision Trees is a supervised machine learning approach that splitted the data of this model into subsets based on feature values, building a tree structure to make predictions. It is highly interpretable models with clear decision paths. It can model nonlinear relationships but prone to overfitting if not properly constrained. In this model the Minimum Number of Instances in Leaves = 1: Each leaf can contain as few as one instance, increasing the model's specificity but making it more sensitive to noise and this operated on "Do Not Split Subsets Smaller Than 1", which ensures splits occur only when there are at least one instance in a subset. Maximum Tree Depth = 3 limited the depth of the tree, constraining its complexity to reduce overfitting.

Also, the last hyperparameter index of "Majority Stop Criterion = 95%" stops splitting further when 95% of the instances in a node belong to a single class, prioritizing purity in the leaves. A good predictive power of an R^2 of 0.96 indicates the model captures most of the variability in compressive strength effectively. A maximum tree depth of 3 makes the model interpretable and computationally efficient. The high Accuracy (95.5%) shows that most predictions align closely with actual values, demonstrating robust performance. The performance parameters; RMSE (3.65), MSE (13.45), and SSE (906.5) are relatively high compared to other models like GMDH-NN (RMSE: 0.4, SSE: 13.5) or k-NN (RMSE: 2.1, SSE: 296.5). This indicates the model struggles with specific predictions, possibly due to insufficient splits or outliers. The shallow tree depth (maximum of 3) might restrict the model's ability to capture finer details in complex relationships. Allowing leaves with a single instance makes the model more prone to overfitting locally. But, a slightly deeper tree (e.g., depth of 4 or 5) improved the performance by enabling more granular splits. Also, pruning

techniques was implemented to balance tree complexity and avoid overfitting. The decision tree model shows strong performance with an R^2 of 0.96 and high accuracy (95.5%), but its error metrics (e.g., RMSE: 3.65) lag behind other methods like GSVR (RMSE: 2.65) or k-NN (RMSE: 2.1). Its simplicity (tree depth of 3) is advantageous for interpretability but limits its flexibility in modeling complex patterns. The decision tree model for predicting the compressive strength of metakaolin-based geopolymer concrete demonstrates strong performance with an R^2 of 0.96 and high accuracy (95.5%). However, its relatively high error metrics (RMSE: 3.65, SSE: 906.5) and limited flexibility suggest room for improvement. Fine-tuning parameters or employing ensemble methods like Random Forest or Gradient Boosting could enhance its accuracy and robustness. The decision tree model for predicting the compressive strength of metakaolin-based self-healing geopolymer concrete offers practical value through its simplicity, interpretability, and fast decision-making capabilities. With an R^2 of 0.96 and a high accuracy of 95.5%, the model demonstrates strong overall performance, especially useful in environments where quick, explainable predictions are required. This makes it ideal for deployment in real-time monitoring systems or decision-support tools at construction sites, where stakeholders may not have technical expertise but need clear logic behind predictive outputs. Its low maximum tree depth of 3 allows for easy visualization and understanding of the relationships between input features and compressive strength, making it a valuable tool for material engineers and researchers involved in quality control. This interpretability aids in trust-building and adoption in conservative industries like construction, where model transparency is often prioritized over complexity.

By helping to pinpoint critical variables influencing compressive strength, the model supports targeted optimization of mix design, reducing reliance on resource-intensive trial-and-error testing. From a sustainability perspective, the decision tree's rapid inference capability and minimal computational requirements enable its integration into low-power edge devices or mobile platforms. This opens up possibilities for remote construction sites or developing regions, where access to advanced computing infrastructure is limited but the need for sustainable construction practices is high. Its ability to predict compressive strength without repeated physical testing conserves materials, minimizes waste, and reduces the environmental impact associated with conventional mix validation processes. However, the model's relatively high error metrics (RMSE: 3.65, SSE: 906.5) and restricted flexibility limit its use in scenarios requiring high-precision predictions or those dealing with complex and highly variable data. To improve its applicability without compromising its interpretability, slight adjustments—such as increasing the maximum depth or integrating pruning strategies—can enhance model performance while still maintaining transparency. Additionally, combining the decision tree with ensemble methods like Random Forest or Gradient Boosting could allow for better generalization and error reduction, making it more suitable for sustainable high-performance applications. Overall, the decision tree model aligns well with sustainable goals when applied in contexts that value simplicity, energy efficiency, and interpretability. It acts as an accessible entry point for predictive modeling in concrete technology, especially for practical field applications focused on material optimization and waste minimization.

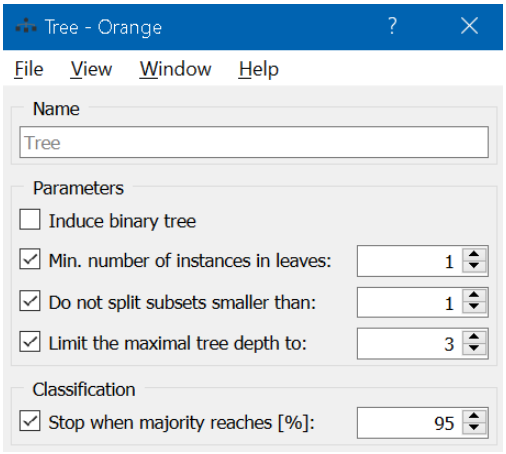

Figure 15. The considered hyper-parameters of (Tree) model

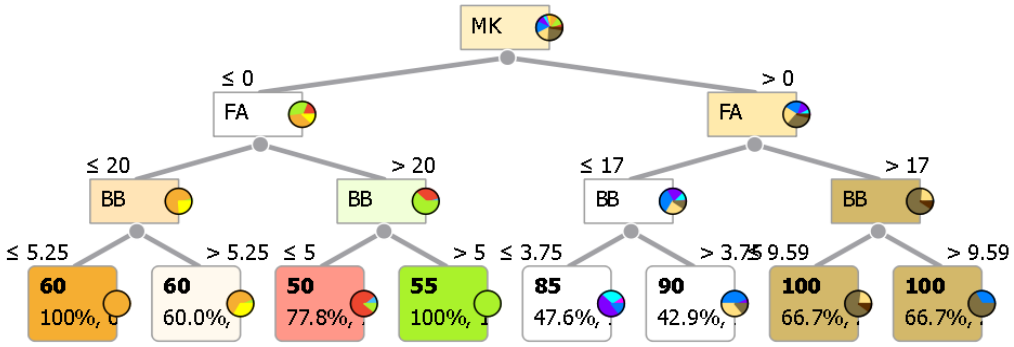

Figure 16. The layout of the developed (Tree) model

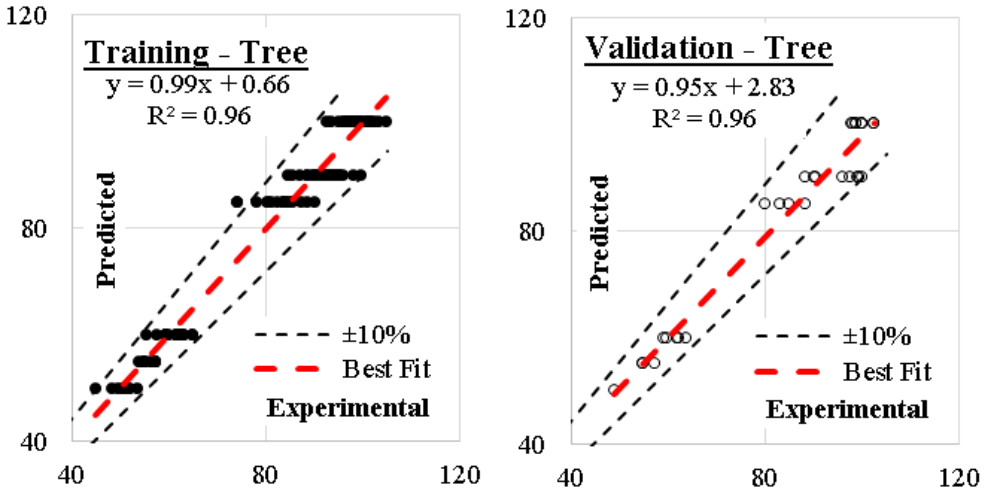

Figure 17. Relation between predicted and calculated strength using (Tree)

RF Model: RF modeling of the compressive strength of metakaolin-based self-healing geopolymer concrete with Number of trees is 2, number of attributes considered at each split is 2, limit depth of individual trees is 2 and do not split subsets smaller than 2, which produced R2 of 0.97, average Accuracy of 0.965, Error of 0.035, RMSE of 3.25, MSE of 14.4, MAE of 2.55 and SSE of 885. These outcome statistics are shown in Figures 18 to 20. The analysis describes the use of Random Forest (RF) for modeling the compressive strength of metakaolin-based self-healing geopolymer concrete. Below is a detailed breakdown of the methodology, results, and potential improvements. The Random Forest is an ensemble method that combines multiple decision trees to improve prediction accuracy and reduce overfitting. Aggregates predictions from individual trees e.g., via averaging in regression tasks, which increases robustness by introducing randomness in feature selection and data splits. The Number of Trees = 2 shows a small forest size and this makes the model computationally lightweight but may limit the ensemble's ability to reduce variance. The Number of Attributes Considered at Each Split = 2 introduces randomness, ensuring diversity among trees by selecting only 2 features for each split. Similarly, the Limit Depth of Individual Trees = 2 constrains the depth of each tree to control model complexity and reduce overfitting. Also, "Do Not Split Subsets Smaller Than 2" ensures each split has a minimum subset size of 2, preventing overly specific splits. The strong predictive power representing an R^2 of 0.97 and high accuracy (96.5%) indicates the model effectively captures the relationship between features and compressive strength. Limiting tree depth (to 2) and introducing randomness through feature selection ensure better generalization. Even with only 2 trees, the model benefits from the diversity introduced by RF methodology.

A forest size of only 2 is insufficient to fully leverage the ensemble method's potential to reduce variance. The error metrics such as RMSE (3.25) and SSE (885) are relatively high compared to other models like GSVR (RMSE: 2.65, SSE: 470.5) or k-NN (RMSE: 2.1, SSE: 296.5), suggesting room for improvement in precision. Restricting tree depth to 2 may oversimplify the model, leading to suboptimal splits and reduced ability to capture complex patterns. Using a larger forest (e.g., 50 or 100 trees) would allow the model to better reduce variance and improve prediction accuracy. Allowing slightly deeper trees (e.g., depth of 3–5) enhanced the model's ability to capture complex relationships while maintaining generalization and optimize the number of attributes considered at each split to balance diversity and information gain. Hyperparameter tuning (e.g., via grid search) was applied to find optimal values for tree count, depth, and features per split. The RF model achieves strong predictive performance with an R^2 of 0.97 and accuracy of 96.5%. Its errors (RMSE: 3.25, SSE: 885) are moderate compared to GMDH-NN (RMSE: 0.4, SSE: 13.5): Significantly lower errors, likely due to neural network flexibility, k-NN (RMSE: 2.1, SSE: 296.5): Better precision in local neighborhood prediction, GSVR (RMSE: 2.65, SSE: 470.5): Slightly better error metrics, leveraging kernel-based learning. The RF model is more interpretable than GMDH-NN or GSVR but less so than single decision trees. The Random Forest model demonstrates excellent predictive power with an R^2 of 0.97 and high accuracy (96.5%). However, its moderate error metrics (RMSE: 3.25, SSE: 885) indicate room for improvement in precision, particularly with a small forest size (2 trees) and limited tree depth (2). Expanding the forest and optimizing parameters could significantly enhance its performance while retaining its robustness and generalization capabilities.

The Random Forest (RF) model for predicting the compressive strength of metakaolin-based self-healing geopolymer concrete demonstrates a balance between performance, interpretability, and robustness, making it practically valuable and potentially sustainable in a wide range of construction-related applications. With an R^2 of 0.97 and a high accuracy of 96.5%, the model captures the underlying relationships in the data effectively, even when implemented with only two shallow trees. This shows its capability to deliver reliable predictions while being computationally efficient, which is particularly useful in resource-constrained environments or for embedded systems used in on-site quality assessment. The model's ensemble nature inherently offers robustness against noise and variability in the input data, making it applicable to real-world conditions where raw material properties and environmental factors often fluctuate. This robustness is essential for supporting sustainable construction practices, where reliance on industrial by-products like metakaolin and the promotion of self-healing mechanisms in concrete are

gaining traction to reduce environmental impact and improve durability.

The RF model can thus aid in optimizing mix designs by identifying the best proportions of components for achieving target strength values without excessive material consumption or waste. Despite its strengths, the model's performance is currently limited by the use of only two trees and shallow depth, which affects its ability to capture more complex patterns or outlier behavior. However, the simplicity of the current configuration implies low energy consumption and fast computation, which aligns well with sustainability goals in low-carbon computing environments. In practice, this configuration could be deployed on mobile or edge devices at construction sites for real-time, in-situ strength prediction. It can reduce the need for extensive laboratory testing, minimize delays, and support faster decision-making during the concrete curing process, ensuring efficient material use and timely quality assurance. For broader and more sustainable application, scaling the model to include a larger number of trees and slightly deeper splits would significantly enhance its precision and adaptability, making it suitable for a more diverse range of concrete types and curing conditions. This can improve its utility in advanced materials development and lifecycle prediction, contributing to more durable and longer-lasting infrastructure with reduced maintenance requirements. Overall, the RF model, even in its lightweight form, offers a practical and sustainable solution for predictive modeling in the development of eco-friendly construction materials. Its blend of generalization capacity, ease of implementation, and adaptability to low-resource deployment environments positions it as a valuable tool for promoting innovation and sustainability in civil engineering and materials science.

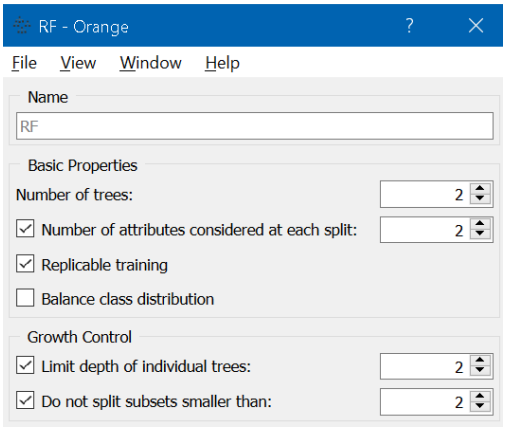

Figure 18. The considered hyper-parameters of (RF) model

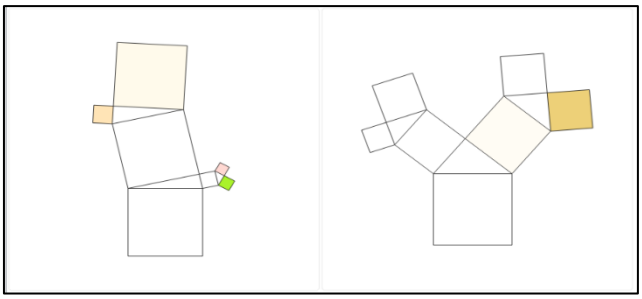

Figure 19: Pythagorean Forest diagram for the developed (RF) models

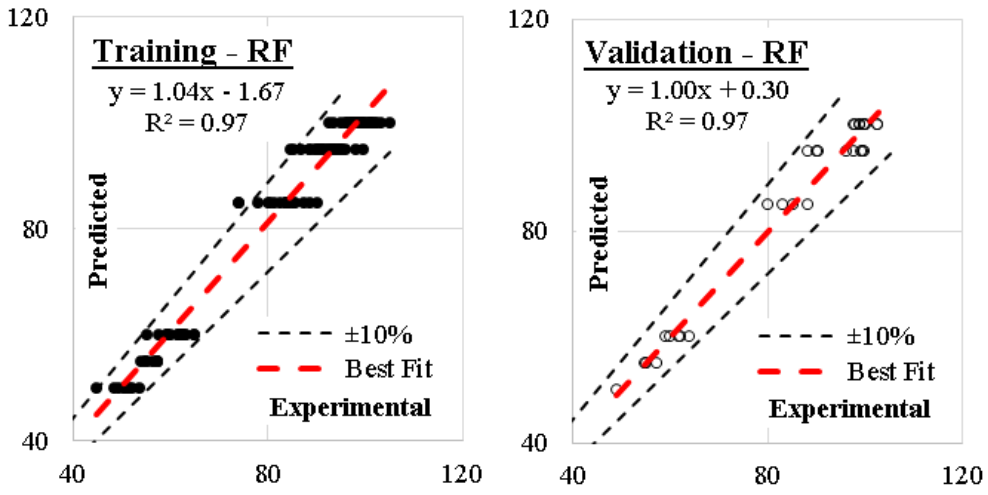

Figure 20: Relation between predicted and calculated strength using (RF)

(XGBoost) Model: XGBoost modeling of the compressive strength of metakaolin-based self-healing geopolymer concrete with number of trees is 100, learning rate is 0.300 with lambda of 1, limit depth of individual trees is 3, and fraction of training instances, features of each tree, each level, and each split is 1, which produced R2 of 0.98, average Accuracy of 0.97, Error of 0.03, RMSE of 2.35, MSE of 5.45, MAE of 1.8 and SSE of 391.5. The outcome statistics are presented in Figures 21, and 22. The provided analysis describes the results of using XGBoost (eXtreme Gradient Boosting) to model the compressive strength of metakaolin-based self-healing geopolymer concrete. The number of trees (100) represents the number of boosting rounds used to train the model. Learning rate (0.300) is of a relatively moderate value, indicating a trade-off between the speed and precision of learning. The Lambda (1) is a regularization parameter to prevent overfitting, ensuring that the model generalizes well. Tree depth (3) is restricting individual tree depth to 3 helps avoid overfitting by limiting model complexity. Fraction settings (1 for training instances, features per tree, level, and split), all available data, features, and splits are used, maximizing the model's access to information. The R^2 (0.98) indicates that the model explains 98% of the variance in the compressive strength data. This suggests a very strong correlation between the input features and the target variable.

The Accuracy (0.97) indicates that 97% of predictions were correct on average. This is a high value, showing the model's reliability. The average prediction error is only 3%, demonstrating strong predictive accuracy. The Root Mean Squared Error (RMSE: 2.35) reflects the average magnitude of prediction errors in the same units as the target variable (compressive strength). Lower RMSE values are better; 2.35 is quite low. The Mean Squared Error (MSE: 5.45): The squared mean of the errors, used to penalize larger errors. This is derived from RMSE. The Mean Absolute Error (MAE: 1.8) represents the average magnitude of errors, with no consideration of direction. This is a straightforward measure of prediction error. The Sum of Squared Errors (SSE: 391.5) is the total squared difference between observed and predicted values. Lower SSE suggests better model fit. The results indicate an excellent model fit, as reflected in high R^2 and accuracy values, alongside low RMSE, MAE, and Error metrics. The chosen hyperparameters (moderate learning rate, regularization, and depth constraints) contribute to effective learning while preventing overfitting. The model appears highly suitable for predicting the compressive strength of geopolymer concrete, which can aid in quality control and design optimization. While the metrics indicate strong performance, external validation with unseen data is recommended to confirm the model's robustness. Overall, the summary of the models performance and the comparison of the indices of evaluation are presented in Table 3 and Figure 22.

The XGBoost model demonstrates outstanding performance in predicting the compressive strength of metakaolin-based self-healing geopolymer concrete, making it highly practical for advanced and sustainable applications in construction material design and quality control. With an R^2 of 0.98 and an accuracy of 97%, the model effectively captures complex relationships between input variables and compressive strength outcomes. Its low error rates—RMSE of 2.35, MAE of 1.8, and SSE of 391.5—affirm its predictive precision and minimal deviation from actual values, underscoring its reliability for real-world implementation. The use of XGBoost is particularly beneficial in sustainable construction practices because it supports rapid optimization of material formulations. By accurately modeling compressive strength, the model enables efficient design of concrete mixes that incorporate metakaolin—a pozzolanic material that enhances durability and sustainability by reducing reliance on ordinary Portland cement and improving self-healing capacity. This supports the production of high-performance, environmentally friendly concretes that meet structural requirements while reducing CO_2 emissions and lifecycle maintenance costs. The model's moderate learning rate and regularization help maintain generalization without overfitting, which is critical in practical applications where input data may vary across sites, material sources, or curing conditions.

Limiting the depth of individual trees to three provides a balance between model complexity and interpretability, which is essential for deployment in environments where explainable decisions are valued, such as in quality assurance protocols or regulatory compliance assessments. In real-time applications, this model could be integrated into intelligent decision-support systems used in construction sites or precast production facilities. These systems can instantly evaluate compressive strength predictions based on current mix designs and environmental conditions, reducing reliance on destructive testing methods and long curing times. This facilitates faster turnaround in production cycles and more sustainable resource use, aligning with modern digital construction goals and green building standards. Furthermore, the model's robustness and precision make it suitable for use in adaptive control systems for automated mixing plants, where real-time adjustments can be made to raw material proportions to maintain desired strength outcomes despite variability in input quality. This minimizes material waste and ensures consistent performance, contributing to both cost efficiency and environmental sustainability. Overall, the XGBoost model's performance metrics reflect a highly practical and scalable approach to predictive modeling in sustainable concrete development. Its capacity for high-precision prediction, adaptability to diverse input features, and alignment with modern computational efficiency standards make it an ideal tool for next-generation, data-driven construction technologies aimed at achieving resilience, resource efficiency, and low-carbon innovation.

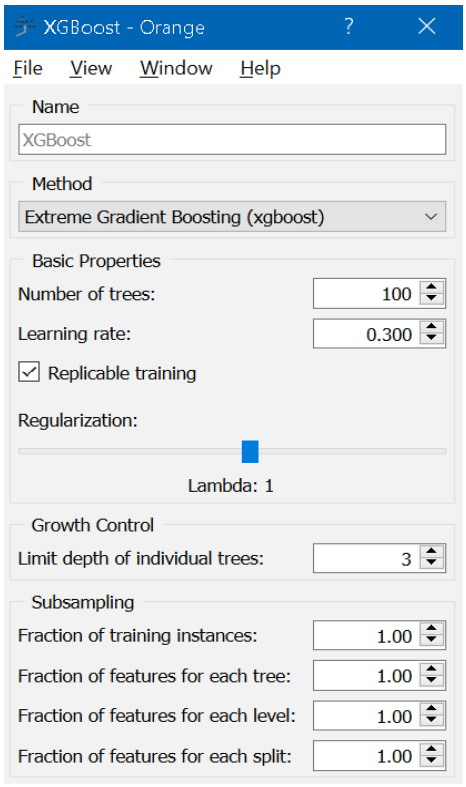

Figure 21. The considered hyper-parameters of (XGBoost) model

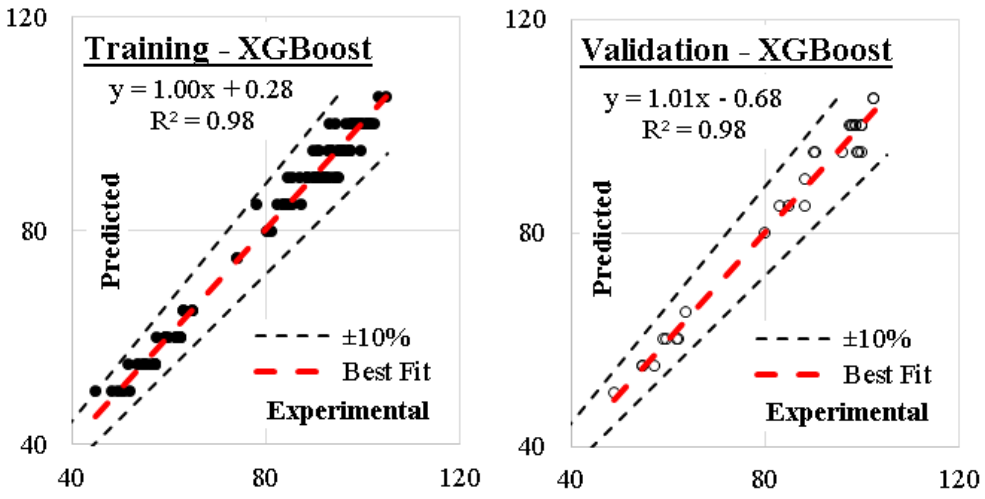

Figure 22. Relation between predicted and calculated strength using (XGBoost)

5. Comparison of the Models' Performance

Table 3 presents the summary of the models performances comparing their indices and in Figure 23, the Taylor chart has been shown. The comparative performance of the models used to predict the compressive strength of metakaolin-based self-healing geopolymer concrete reveals a clear gradient in accuracy, precision, and practical applicability. Among them, the GMDH-NN stands out as the most precise model, delivering the lowest RMSE (0.4) and SSE (13.5), indicating highly accurate predictions with minimal error. This suggests its exceptional suitability for applications demanding tight tolerance and consistency in prediction, although it comes with a trade-off in interpretability due to the complexity of neural network architectures. Following closely, the XGBoost model also demonstrates excellent performance, with an R^2 of 0.98 and low error metrics including an RMSE of 2.35 and an SSE of 391.5. It balances accuracy and computational efficiency through its boosting mechanism and regularization, making it highly applicable for real-time quality control and sustainable concrete optimization. Its robustness and generalization capacity, supported by structured hyperparameters, position it as a practical choice for deployment in dynamic construction environments.

The GSVR model also performs competitively, with similar R^2 (0.98) and accuracy (0.97), but exhibits slightly higher RMSE (2.65) and SSE (470.5) compared to XGBoost, indicating that while it captures overall trends well, it may not generalize as efficiently across all data points. Its reliance on a polynomial kernel makes it effective for nonlinear patterns but potentially susceptible to overfitting due to its high cost parameter (C=100). This makes it well-suited for detailed laboratory-scale investigations where data patterns are well understood but less ideal for large-scale deployment without further optimization. The k-NN model, with an RMSE of 2.1 and SSE of 296.5, surpasses GSVR and Random Forest in terms of error minimization. It leverages local data structure, making it effective for datasets where similar patterns repeat. However, its sensitivity to data distribution and reliance on distance metrics reduce its scalability in high-dimensional or noisy data environments. Random Forest, though traditionally powerful in many regression problems, performs moderately here with an RMSE of 3.25 and SSE of 885. The use of only two trees and shallow depth (2) limits its capacity to capture complex relationships, which dampens its ensemble advantage. However, its interpretability and resistance to overfitting remain practical benefits, especially if tree count and depth are increased in future iterations.

The decision tree model is the most interpretable among all and has the lowest computational cost. While it still delivers strong performance (R^2 of 0.96), its higher error values (RMSE of 3.65 and SSE of 906.5) reflect a lack of flexibility due to constrained tree depth (3) and over-specified splits (minimum leaf size of 1). It is most useful where model transparency is paramount, such as regulatory reviews or educational demonstrations. In summary, while GMDH-NN delivers the highest precision, XGBoost offers the best overall trade-off between accuracy, generalization, and practicality. GSVR and k-NN follow closely, offering strong but more context-sensitive performance. Random Forest and decision tree models provide more interpretability but lag in precision due to limited configuration settings. The final choice among these depends on the specific balance required between model transparency, predictive performance, computational resources, and deployment scale in sustainable construction applications. The results from the table, which provide a comparison of various machine learning models for predicting the compressive strength of metakaolin-based self-healing geopolymer concrete, offer valuable insights when compared with those discussed in the literature.

Studies from Pratap et al. [7] and Wang et al. [8] have highlighted the influence of metakaolin on the compressive strength of geopolymer concrete, with metakaolin contributing to an increase in compressive strength, particularly at a 20% fraction. This aligns well with the current models, where GMDH-NN, RF, and XGBoost models achieved strong prediction performance with high R^2 values (above 0.96 in most cases) and relatively low error metrics like RMSE, MSE, and MAE. These outcomes demonstrate that machine learning models, like the ones presented here, can be as effective as traditional experimental methods in capturing the effects of material substitutions like metakaolin. Wang et al. [8] introduced the Firefly Algorithm (AF) for optimizing the prediction of compressive strength in geopolymer concrete and found that machine learning models like Random Forest-AF had the lowest RMSE. Comparing this to the current study, the Random Forest model here also delivered impressive performance (RMSE of 3.25), though slightly higher than the Firefly-enhanced model. This difference can be attributed to the distinct methodologies, where the Firefly Algorithm may have contributed to further refinement in the predictive accuracy of the model, while the RF model in this study had a smaller number of trees (2), limiting its variance-reducing capability. Additionally, Tian et al. [10] emphasized the role of NaOH molar content and the use of integrated models for compressive strength prediction in geopolymer concrete. Their study found that Random Forest, when integrated with the Beetle Antennae Search (IBAS) algorithm, produced the best results for geopolymer concrete compressive strength prediction. In comparison, the RF model in this study, while strong with an R^2 of 0.97, could potentially benefit from the inclusion of optimization techniques like IBAS to fine-tune the model further, especially in terms of minimizing errors like RMSE and MSE.

Ahmed et al. [11] also emphasized the importance of certain material characteristics, such as the SiO_2 percentage in fly ash, for accurate prediction of compressive strength. This study's findings support the use of machine learning models in predicting such material properties accurately, with models like GMDH-NN and XGBoost achieving high accuracy and robust performance metrics. This highlights the potential for improving geopolymer concrete prediction models by incorporating more granular material-specific data, which can further refine the models presented here. Overall, the results in the Table 3 are consistent with the findings from the literature, suggesting that machine learning techniques, especially Random Forest and XGBoost, are promising tools for predicting the compressive strength of metakaolin-based self-healing geopolymer concrete. However, the error metrics observed in the present study (e.g., RMSE and MSE) indicate that there is still room for improvement, especially when compared with optimized models or those that integrate additional algorithms for hyperparameter tuning, as shown in some of the cited studies.

Table 3. Performance measurements of developed models

Compressive Strength								
Model	**Dataset**	**SSE**	**MAE (MPa)**	**MSE (MPa)**	**RMSE (MPa)**	**Error (%)**	**Accuracy (%)**	**R^2**
GMDH-NN	Training	21	0.4	0.2	0.4	0.01	0.99	1.00
	Validation	6	0.4	0.2	0.4	0.01	0.99	1.00
KNN	Training	447	1.5	3.8	2.0	0.02	0.98	0.99
	Validation	146	1.7	4.9	2.2	0.03	0.97	0.99
GSVR	Training	713	1.9	6.1	2.5	0.03	0.97	0.98
	Validation	228	2.1	7.6	2.8	0.03	0.97	0.98
Tree	Training	1354	2.6	11.6	3.4	0.04	0.96	0.96
	Validation	459	2.6	15.3	3.9	0.05	0.95	0.96
RF	Training	1519	2.8	13.0	3.6	0.04	0.96	0.97
	Validation	251	2.3	15.3	2.9	0.03	0.97	0.97
XGBoost	Training	613	1.8	5.2	2.3	0.03	0.97	0.98
	Validation	170	1.8	5.7	2.4	0.03	0.97	0.98

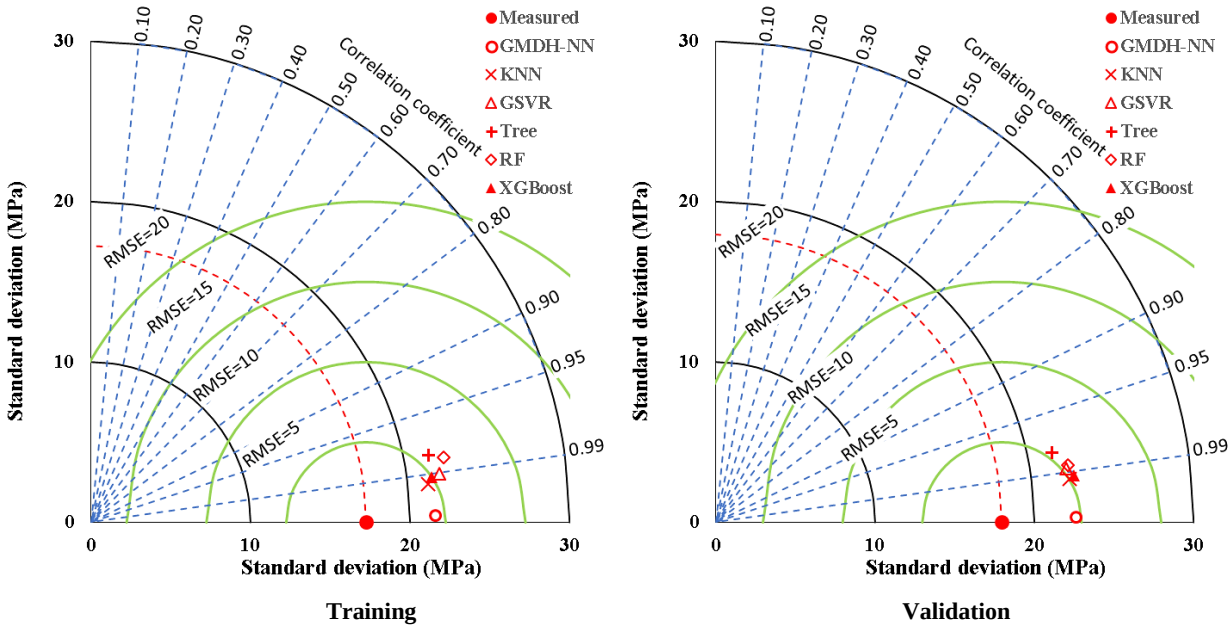

Figure 23. Comparing the accuracies of the developed models using Taylor charts

6. Conclusions

A study on modeling the compressive strength of environmentally friendly metakaolin-based self-healing geopolymer concrete treated with Bacillus bacteria (BB) has been conducted, analyzed, and reported in this research paper. Machine learning methods such as the "Group Methods Data Handling Neural Network (GMDH-NN)", "Generalized Support Vector Regression (GSVR), "K-Nearest Neighbors (KNN)", "Tree Decision (Tree)", "Random Forest (RF)" and "Extreme Gradient Boosting (XGBoost)" were applied to model the compressive strength of the self-healing concrete. The GMDH-NN model was created using GMDH Shell 3.0 software, while XGBoost, GSVR, KNN, Tree, and RF models were created using "Orange Data Mining" software version 3.36. The research method also included gathering relevant experimental and field data, categorizing it effectively, and performing initial analysis to identify trends and relationships. A global representative database was collected from literature for different mixing ratios of self-healing concrete corresponding to the compressive strength, with a total of 147 records, which contained Fly Ash (FA), Silica Fume (SF), Metakaolin (MK), and Bacillus Bacteria (BB) considered as the input constituents. The collected records were divided into a training set (75%) and a validation set (25%) based on established requirements. At the end of the modeling exercise, the following conclusions have been made:

- The studied variables showed no internal consistency after the preliminary analysis, hence requiring the application of machine learning models to establish more sustainable consistency with the output.

- In terms of the accuracy of the models, the GMDH-NN produced the best model with an Accuracy of 0.99, while the KNN and the GSVR followed closely with accuracies of 0.975 and 0.97, respectively. However, the RF and the Tree models also produced good accuracies of 0.965 and 0.955, respectively.
- In terms of the general performance evaluation, the GMDH-NN and the KNN again outperformed the other methods, producing an R^2 of 1.00 and 0.99, respectively, while the GSVR, RF, and Tree followed in this order with R^2 of 0.98, 0.97, and 0.96, respectively. The error indices, such as the overall Error, RMSE, MSE, MAE, and SSE, also confirm this order of performance.
- Lastly, the sensitivity analysis on the modeling of compressive strength of metakaolin-based self-healing geopolymer concrete treated with bacillus bacteria produced a metakaolin (MK) impact of 30%, a silica fume (SF) impact of 29%, a fly ash (FA) impact of 27%, and a bacillus bacteria (BB) impact of 14%. This highlights the dominant role of metakaolin (30%), silica fume (29%), and fly ash (27%) in determining the compressive strength of metakaolin-based self-healing geopolymer concrete. Bacillus bacteria (14%) have a smaller but meaningful impact, primarily contributing to self-healing and long-term durability. These insights can guide material selection, mix design, and process optimization to enhance both strength and durability.

7. Practical Application

The practical application of the research outcomes in predicting the compressive strength of metakaolin-based self-healing geopolymer concrete has significant implications for sustainable construction and materials science. By developing reliable predictive models, such as decision trees, random forests, and XGBoost, the research provides a more accurate and efficient approach to designing and optimizing geopolymer concrete formulations. These models, with their ability to predict compressive strength based on various input parameters, can be applied in real-world construction projects to ensure that the concrete used meets the necessary strength requirements while reducing environmental impact. One of the main practical applications is in the optimization of concrete mixes. The research provides valuable insights into how different factors, such as the proportion of metakaolin, curing conditions, and mix composition, affect the compressive strength of geopolymer concrete. By using the predictive models, engineers and material scientists can optimize these factors to create stronger, more durable concrete with minimal environmental impact. This is particularly important in the context of reducing the carbon footprint of construction materials, as geopolymer concrete is known to produce significantly lower CO_2 emissions compared to traditional Portland cement-based concrete.

Moreover, the self-healing properties of the metakaolin-based geopolymer concrete add another layer of sustainability. In practice, this means that structures made from this material would require less maintenance and repair over their lifespan, ultimately reducing the long-term costs and environmental footprint associated with traditional concrete. The self-healing mechanism, which involves the autonomous sealing of cracks through various processes, ensures that the concrete can remain structurally sound even after experiencing wear and tear, thereby extending the life of buildings, bridges, and other infrastructure. In addition, the research outcomes can inform quality control practices within the concrete manufacturing industry. By incorporating machine learning models into the production process, manufacturers can predict the quality of the concrete mix before it is cast, ensuring that each batch meets the required standards for strength and durability. This could also lead to more standardized production methods, reducing the variability in concrete quality that can sometimes occur in traditional manufacturing practices. On a broader scale, the ability to predict the compressive strength of geopolymer concrete with a high degree of accuracy could lead to its wider adoption in construction, particularly in environmentally conscious projects. This research, by enhancing the understanding of metakaolin-based geopolymer concrete and providing tools for its optimization, can help address the growing demand for sustainable building materials. It may also play a role in shaping building codes and industry standards that prioritize low-carbon alternatives to traditional concrete, further promoting the shift toward more sustainable construction practices.

8. Declarations

8.1. Author Contributions

Conceptualization, N.U., M.A., and D.M.; methodology, N.U., E.G.Z.N., D.M., and M.A.; software, D.M. and M.A.; validation, N.U., M.A., and E.G.Z.N.; formal analysis, N.U.; investigation, N.U., E.G.Z.N., D.M., and M.A.; data curation, N.U. and M.A.; writing—original draft preparation, N.U., E.G.Z.N., D.M., and M.A.; writing—review and editing, N.U., E.G.Z.N., D.M., and M.A.; visualization, N.U.; supervision, N.U. All authors have read and agreed to the published version of the manuscript.

8.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

8.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

8.4. Conflicts of Interest

The authors declare no conflict of interest.

9. References

[1] Faraj, R. H., Ahmed, H. U., Hama Ali, H. F., & Sherwani, A. F. H. (2021). Fresh and mechanical properties of concrete made with recycled plastic aggregatesHandbook of Sustainable Concrete and Industrial Waste Management: Recycled and Artificial Aggregate, Innovative Eco-friendly Binders, and Life Cycle Assessment, 167–185. doi:10.1016/B978-0-12-821730-6.00023-1.

[2] Shaikh, F. U. A. (2016). Mechanical and durability properties of fly ash geopolymer concrete containing recycled coarse aggregates. International Journal of Sustainable Built Environment, 5(2), 277–287. doi:10.1016/j.ijsbe.2016.05.009.

[3] Shamim Ansari, S., Muhammad Ibrahim, S., & Danish Hasan, S. (2023). Conventional and Ensemble Machine Learning Models to Predict the Compressive Strength of Fly Ash Based Geopolymer Concrete. In Materials Today: Proceedings, 1-8. doi:10.1016/j.matpr.2023.04.393.

[4] Abdel-Gawwad, H. A., & Abo-El-Enein, S. A. (2016). A novel method to produce dry geopolymer cement powder. HBRC Journal, 12(1), 13–24. doi:10.1016/j.hbrcj.2014.06.008.

[5] Khan, K., Ahmad, W., Amin, M. N., & Deifalla, A. F. (2023). Investigating the feasibility of using waste eggshells in cement-based materials for sustainable construction. Journal of Materials Research and Technology, 23, 4059–4074. doi:10.1016/j.jmrt.2023.02.057.

[6] Liu, T., Nafees, A., khan, S., Javed, M. F., Aslam, F., Alabduljabbar, H., Xiong, J. J., Ijaz Khan, M., & Malik, M. Y. (2022). Comparative study of mechanical properties between irradiated and regular plastic waste as a replacement of cement and fine aggregate for manufacturing of green concrete. Ain Shams Engineering Journal, 13(2), 101563. doi:10.1016/j.asej.2021.08.006.

[7] Pratap, B., Sharma, S., Kumari, P., & Raj, S. (2024). Mechanical properties prediction of metakaolin and fly ash - based geopolymer concrete using SVR. Journal of Building Pathology and Rehabilitation, 9(1), 1. doi:10.1007/s41024-023-00360-9.

[8] Wang, R., Zhang, J., Lu, Y., Ren, S., & Huang, J. (2024). Towards a Reliable Design of Geopolymer Concrete for Green Landscapes: A Comparative Study of Tree-Based and Regression-Based Models. Buildings, 14(3), 615. doi:10.3390/buildings14030615.

[9] Wang, Y., Iqtidar, A., Amin, M. N., Nazar, S., Hassan, A. M., & Ali, M. (2024). Predictive modelling of compressive strength of fly ash and ground granulated blast furnace slag based geopolymer concrete using machine learning techniques. Case Studies in Construction Materials, 20, 3130. doi:10.1016/j.cscm.2024.e03130.

[10] Tian, Q., Su, Z., Fiorentini, N., Zhou, J., Luo, H., Lu, Y., Xu, X., Chen, C., & Huang, J. (2024). Ensemble learning models to predict the compressive strength of geopolymer concrete: a comparative study for geopolymer composition design. Multiscale and Multidisciplinary Modeling, Experiments and Design, 7(3), 1793–1806. doi:10.1007/s41939-023-00303-4.

[11] Ahmed, H. U., Abdalla, A. A., Mohammed, A. S., Mohammed, A. A., & Mosavi, A. (2022). Statistical Methods for Modeling the Compressive Strength of Geopolymer Mortar. Materials, 15(5), 1868. doi:10.3390/ma15051868.

[12] Mansouri, E., Manfredi, M., & Hu, J. W. (2022). Environmentally Friendly Concrete Compressive Strength Prediction Using Hybrid Machine Learning. Sustainability (Switzerland), 14(20), 12990. doi:10.3390/su142012990.

[13] Ebid, A. E., Deifalla, A. F., & Onyelowe, K. C. (2024). Data Utilization and Partitioning for Machine Learning Applications in Civil Engineering. Sustainable Civil Infrastructures, 87–100. doi:10.1007/978-3-031-70992-0_8.

[14] Hoffman, F. O., & Gardner, R. H. (1983). Evaluation of uncertainties in radiological assessment models. In J. E. Till & H. R. Meyer (Eds.), Radiological Assessment: A Textbook on Environmental Dose Analysis. NRC Office of Nuclear Reactor Regulation, San Diego, United States.

[15] Siddique, R., Aggarwal, P., & Aggarwal, Y. (2011). Prediction of compressive strength of self-compacting concrete containing bottom ash using artificial neural networks. Advances in Engineering Software, 42(10), 780–786. doi:10.1016/j.advengsoft.2011.05.016.

[16] Singh, P., Bhardwaj, S., Dixit, S., Shaw, R. N., & Ghosh, A. (2021). Development of prediction models to determine compressive strength and workability of sustainable concrete with ann. Lecture Notes in Electrical Engineering: Vol. 756 LNEE, 753–769. doi:10.1007/978-981-16-0749-3_59.

[17] Song, H., Ahmad, A., Farooq, F., Ostrowski, K. A., Maślak, M., Czarnecki, S., & Aslam, F. (2021). Predicting the compressive strength of concrete with fly ash admixture using machine learning algorithms. Construction and Building Materials, 308, 125021. doi:10.1016/j.conbuildmat.2021.125021.

[18] Song, Y., Zhao, J., Ostrowski, K. A., Javed, M. F., Ahmad, A., Khan, M. I., Aslam, F., & Kinasz, R. (2022). Prediction of compressive strength of fly-ash-based concrete using ensemble and non-ensemble supervised machine-learning approaches. In Applied Sciences (Switzerland), 12(1), 361. doi:10.3390/app12010361.

[19] Topçu, I. B., & Saridemir, M. (2008). Prediction of compressive strength of concrete containing fly ash using artificial neural networks and fuzzy logic. Computational Materials Science, 41(3), 305–311. doi:10.1016/j.commatsci.2007.04.009.

[20] Zhang, J., Zhao, Y., & Li, H. (2017). Experimental Investigation and Prediction of Compressive Strength of Ultra-High Performance Concrete Containing Supplementary Cementitious Materials. Advances in Materials Science and Engineering, 4563164. doi:10.1155/2017/4563164.

[21] Yilmaz, S., & Küçüksille, E. U. (2015). A new modification approach on bat algorithm for solving optimization problems. Applied Soft Computing Journal, 28, 259–275. doi:10.1016/j.asoc.2014.11.029.

[22] Chiroma, H., Herawan, T., Fister, I., Fister, I., Abdulkareem, S., Shuib, L., Hamza, M. F., Saadi, Y., & Abubakar, A. (2017). Bio-inspired computation: Recent development on the modifications of the cuckoo search algorithm. Applied Soft Computing Journal, 61, 149–173. doi:10.1016/j.asoc.2017.07.053.

[23] Onyelowe, K. C., Ebid, A. M., Aneke, F. I., & Nwobia, L. I. (2023). Different AI Predictive Models for Pavement Subgrade Stiffness and Resilient Deformation of Geopolymer Cement-Treated Lateritic Soil with Ordinary Cement Addition. International Journal of Pavement Research and Technology, 16(5), 1113–1134. doi:10.1007/s42947-022-00185-8.

[24] Mohamad, A. B., Zain, A. M., & Bazin, N. E. N. (2014). Cuckoo search algorithm for optimization problems - A literature review and its applications. Applied Artificial Intelligence, 28(5), 419–448. doi:10.1080/08839514.2014.904599.

[25] Onyelowe, K. C., Ebid, A. M., & Hanandeh, S. (2024). Advanced machine learning prediction of the unconfined compressive strength of geopolymer cement reconstituted granular sand for road and liner construction applications. Asian Journal of Civil Engineering, 25(1), 1027–1041. doi:10.1007/s42107-023-00829-5.

[26] Al-Kharabsheh, B. N., Arbili, M. M., Majdi, A., Alogla, S. M., Hakamy, A., Ahmad, J., & Deifalla, A. F. (2023). Basalt Fiber Reinforced Concrete: A Compressive Review on Durability Aspects. Materials, 16(1), 429. doi:10.3390/ma16010429.

Appendix I

Table A1. Utilized dataset

FA (%)	SF (%)	MK (%)	BB (%)	Fc (MPa)
Training set				
15	10	5	12.5	92.97
30	0	0	11.25	54.91
15	10	5	3.75	85.12
20	10	5	8.8	97.97
30	0	0	12.5	57.13
17	5	8	11.25	93.11
12	10	8	2.5	87.40
12	10	8	1.25	84.37
25	5	8	1.65	100.77
25	5	8	6.62	97.72
17	5	8	0	74.07
15	10	5	0	74.02
12	10	8	8.75	95.26
20	10	5	7.79	95.83
12	10	8	5	89.83
15	10	5	5	84.60
30	0	0	11.25	57.27
30	0	0	7.5	57.06
30	0	0	10	55.10
15	10	5	6.25	88.43
20	10	5	3.44	97.52
15	10	5	8.75	93.97
15	10	5	11.25	93.48
17	5	8	2.5	83.76
20	10	5	7.15	99.10
25	5	8	5.95	98.62
15	10	5	5	88.45
17	5	8	3.75	85.39
15	10	5	11.25	90.61
15	10	5	7.5	92.23
20	5	0	6.33	62.45
20	10	5	6.42	97.41
25	5	8	5.83	98.74
17	5	8	8.75	93.07
20	10	5	4.53	99.10
30	0	0	12.5	54.73
30	0	0	2.5	49.44
20	10	5	5.99	94.92
12	10	8	5	93.74
12	10	8	7.5	94.45
20	5	0	8.36	55.21
25	5	8	4.49	97.15
12	10	8	11.25	95.79
20	5	0	9.37	59.39
17	5	8	12.5	92.63
15	10	5	3.75	83.53
15	10	5	8.75	90.26

20	5	0	3.93	61.32
25	5	8	5.85	97.95
12	10	8	7.5	98.09
15	10	5	7.5	89.70
25	5	8	4.6	104.84
17	5	8	8.75	90.73
17	5	8	1.25	80.37
20	5	0	13.25	59.96
20	5	0	5.97	64.71
25	5	8	11.12	98.74
30	0	0	0	44.87
17	5	8	6.25	91.17
12	10	8	0	77.97
20	10	5	9.59	96.16
30	0	0	3.75	50.46
25	5	8	9.52	98.40
15	10	5	10	94.67
15	10	5	10	90.82
20	10	5	9.23	98.99
17	5	8	11.25	90.86
12	10	8	8.75	99.59
30	0	0	6.25	53.71
17	5	8	10	94.89
17	5	8	6.25	86.83
25	5	8	7.15	99.41
30	0	0	2.3	50.65
20	5	0	7.97	57.59
17	5	8	10	90.95
30	0	0	6.25	56.19
20	10	5	6.93	101.25
25	5	8	5.25	103.14
20	10	5	10.45	101.02
15	10	5	2.5	82.46
30	0	0	1.25	48.39
20	5	0	8.09	63.01
20	5	0	5.91	61.32
20	5	0	9.55	62.22
25	5	8	1.41	93.43
12	10	8	12.5	94.97
20	5	0	5.63	64.93
17	5	8	3.75	83.97
30	0	0	3.75	51.99
15	10	5	2.5	83.57
20	5	0	3.77	61.77
17	5	8	7.5	92.05
30	0	0	8.75	57.20
12	10	8	6.25	93.09
30	0	0	5	53.42
17	5	8	5	85.31
25	5	8	5.12	99.64
15	10	5	6.25	91.39
17	5	8	7.5	90.51

25	5	8	4.91	96.48
20	5	0	6.43	63.12
17	5	8	1.25	81.19
20	5	0	5.25	59.17
20	5	0	4.31	62.33
12	10	8	3.75	89.99
20	10	5	9.98	92.44
15	10	5	1.25	80.07
25	5	8	4.91	102.24
25	5	8	3.29	100.21
20	10	5	6.39	97.41
30	0	0	5	51.48
25	5	8	8.05	101.79
25	5	8	4.21	101.00
20	10	5	2.9	97.97
20	10	5	5.81	96.96
20	5	0	9.1	59.39
12	10	8	3.75	88.62
Validation set				
12	10	8	6.25	97.53
20	5	0	8.41	63.80
25	5	8	3.07	102.58
30	0	0	10	57.33
20	10	5	4.88	98.76
15	10	5	1.25	80.11
20	10	5	6.12	97.75
20	10	5	6.73	102.61
17	5	8	5	88.27
25	5	8	7.38	99.87
20	10	5	8.67	98.09
12	10	8	10	99.87
20	10	5	8.54	100.12
20	5	0	4.53	62.11
20	5	0	3.78	61.88
17	5	8	2.5	83.01
15	10	5	12.5	90.09
30	0	0	1.25	48.91
25	5	8	4.15	98.85
20	5	0	7.04	59.85
20	5	0	7.95	59.06
30	0	0	7.5	54.82
12	10	8	11.25	99.47
20	10	5	4.64	99.89
12	10	8	12.5	99.03
12	10	8	10	96.06
12	10	8	1.25	85.07
17	5	8	12.5	90.44
12	10	8	2.5	88.24
30	0	0	8.75	54.96