Hybrid Optimization of CO_2 Emissions and Energyin High-Performance Concrete Using KNN, Elastic Net, and Artificial Rabbits Optimization Models

Xiong Gao[1], Yan Li[2]*, Yansheng Wu[3]
[1]Faculty of Environmental and Chemical Engineering, Kunming Metallurgy College, Kunming City, Yunnan Province, 650033, China
[2]Faculty of Architecture Engineering, Kunming Metallurgy College, Kunming City, Yunnan Province, 650033, China
[3]Logistics and Infrastructure Management Center, Kunming Metallurgy College, Kunming City, Yunnan Province, 650033, China
E-mail: feng_4587@163.com
*Corresponding author

Keywords: machine learning, K-Nearest neighbor, elastic Net regression, artificial rabbits optimization, high-performance concrete

Received: April 10, 2025

In this study, an integrated machine learning framework is proposed to accurately predict and minimize CO_2 emissions and energy consumption in the manufacturing of High-Performance Concrete (HPC). The methodology combines K-Nearest Neighbor (KNN) and Elastic Net Regression (ENR) models with the Artificial Rabbits Optimization (ARO) algorithm for hyperparameter tuning, and employs Recursive Feature Elimination (RFE) to isolate the most influential input variables. A dataset comprising key HPC mix components was curated from experimental sources and subjected to rigorous preprocessing. Among the tested models, the hybrid ENR + ARO (ENAR) model achieved the best performance for energy prediction with an R^2 of 0.986 and RMSE of 52.63 MJ/m³, while the KNN + ARO (KNAR) model yielded the highest accuracy for CO_2 emission prediction with an R^2 of 0.992 and RMSE of 7.57 kg/m³. The application of RFE improved model performance by 12.4% in RMSE reduction for energy prediction and 9.6% for CO_2 estimation, by eliminating redundant features. Cement and superplasticizer content were identified as the most influential predictors. These results provide a reliable and interpretable framework for enhancing the sustainability of concrete production through data-driven mix optimization.

Povzetek: Študija združi KNN, Elastic Net in metahevristiko Artificial Rabbits Optimization z RFE za napovedovanje ter zmanjšanje CO_2 emisij in energije v visokozmogljivem betonu. Najboljša modela (ENAR za energijo, KNAR za CO_2) izpostavita cement in superplastifikator kot ključna dejavnika.

1 Introduction

HPC constitutes a specialized category of concrete characterized by superior performance relative to conventional concrete, as evidenced by selected properties such as durability, service life, and low maintenance requirements [1]. HPC is made up of Portland cement, water, coarse aggregates, and various chemical and mineral admixtures [2]. However, HPC manufacturing utilizes significant amounts of energy and adds a lot to CO_2 emissions, especially from the manufacture of cement, which is responsible for around 8% of CO_2 emissions worldwide [3]. Recently, the environmental impact of HPC has become a focal point, especially when large-scale construction projects increase demand for this material [4], [5]. Heavy machinery employed in performing tasks such as structural simulations, material testing, and optimization is widely used throughout the design and process stages of these projects [6]. Such machines need, by necessity, high-powered systems that consume considerable amounts of energy and produce considerable CO_2 emissions both during their operation and through the lifecycle of manufacture and use [7], [8].

Among other major concerns in the sustainability of construction projects is energy consumption, which is important both ecologically and economically [9]. In general, large computer systems used in high-performance applications for BIM, 3D modeling, and complex simulations consume a huge amount of energy [10]. DVFS and power capping can be two ways of managing power that, once optimized, could reduce energy usage without necessarily affecting computational performance. These methods allow energy resources to be used more efficiently while sustaining the high processing power required for modern construction tasks [11]. The construction industry has to be strengthened with green computing and Energy Efficiency (EE) for it to be sustainable. Only then would modern construction projects result in a minimum environmental impact coupled with economic viability [12].

1.1 Related works

Properties and mixture design of Ultra-High-Performance Concrete (UHPC) have a direct influence on the forecast of CO_2 emissions and energy consumption for manufacturing HPC. Normally, binder content between 800-1000 kg/m³ has the highest impact on the carbon footprint of UHPC. In UHPC, cement hydration is not fully achieved, with only 30–40% of the cement reacting because of a reduced water-to-binder ratio. The remaining unreacted cement acts mainly as an inert filler [13]. The employment of SCMs comprising FA and GGBS results not only in a reduction in cement consumption but also in lowered carbon emissions [14], [15]. That is important because, by replacing the conventional cement with SCMs, during the manufacturing process of UHPC, CO_2 emissions are greatly reduced, thus having a direct influence on the carbon footprint in general. Apart from this, binder content optimization plays a vital role in determining desired mechanical properties; hence, energy consumption is of concern. Such a dense microstructure and high compressive strength through SCM incorporation and superplasticizers increase energy requirements, mainly during the curing phase. For the curing of composites, additional methods like heat or autoclave curing imply greater energy use [16], [17]. On the other hand, the addition of SCMs like 20% FA or 3.2% $nano - CaCCO_3$ has been reported not only to improve mechanical properties but also to have a potential influence on the EE of the manufacturing process by possibly reducing the curing energy required for optimum performance [18], [19].

1.2 Objective

This paper, therefore, aims to develop a new framework for the estimation of feature importance and the optimization of model performance with the multi-objective approach of KNN, ENR, and ARO. This paper designs a new feature importance estimation framework and optimizes model performance for a multi-objective approach that makes use of a combination of KNN, ENR, and ARO techniques. The framework, therefore, performs FS and enhances predictive accuracy by making good use of the complementary strengths of the combined models on a wide range of datasets for the assurance of robustness and completeness when performing ML optimizations. First, RFE is applied as an iterative feature selector that keeps only a subset of the most informative features. RFE eliminates the least important features according to their ranking regarding model performance and iteratively updates the feature importance. This recursive procedure retains only the most relevant features, hence reducing the dimensionality, avoiding overfitting, and improving the efficiency of the model. From the optimization perspective, the ARO novel metaheuristic algorithm inspired by nature has been introduced to optimize hyperparameters of model performance. Indeed, this new ARO metaheuristic mimics rabbit foraging behavior in its process of performing an effective search for near-optimal solutions towards high-dimensional, possibly complicated spaces. Since it optimizes models with numerous interactions among so many variables in the interaction, ARO is certainly the right addition to improve optimization model selection methodologies. The KNNs together with ENR and ARO algorithms systematically model refinement and ensure some combinations of opted features and parameters by a model optimized towards superior predictive result outcomes. In this respect, this study presents a hybrid modeling approach, combining the strengths of KNN, ENR, and ARO, thus building increased model interpretability, performance, and robustness.

2 Methodology

2.1 K-Nearest Neighbor (KNN)

The simplicity, effectiveness, and application of the KNN algorithm are well known. It is related to both RF and ANN in that it can be used in both regression and classification applications. For everyday use, it was fitting since its general concepts were obvious and simple. Both regression and classification problems can train non-linear decision boundaries. This makes it further flexible when these limits are set, and thereby allows the value of K to change. Compared to other algorithms, K-NN does not have to have a separate training phase. In simple words, it uses one hyperparameter, K, which also makes modifying the rest rather easy. The main concept behind KNN involves locating a set of K samples in the calibration data that are similar to the unknown samples, typically using a distance metric. For this purpose, matching groups of samples must be determined. KNN compares the result with a group of K samples and computes the mean value of the response variables to find the classes of the unknown samples [20]. Hence, the effectiveness of the KNN algorithm depends heavily on the choice of K [21]. The three distance functions used in regression projects for calculating the distances between the neighborhood data points are given by Eqs. (1-3).

$$F(e) = \sqrt{\sum_{i=0}^{f}(x_i - y_i)^2} \tag{1}$$

$$F(ma) = \sum_{i=0}^{f}|x_i - y_i| \tag{2}$$

$$F(mi) = \left(\sum_{i=0}^{f}(|x_i - y_i|)^q\right)^{\frac{1}{q}} \tag{3}$$

The Manhattan distance function is denoted by $F(ma)$, the Minkowski distance function by $F(mi)$, and the Euclidean distance function by $F(e)$. The order parameter q is used to calculate the distances between the data points x and y, which are represented by the words x_i and y_i, respectively, for their ith dimensions. The flowchart for the KNN model is demonstrated in Fig. 1.

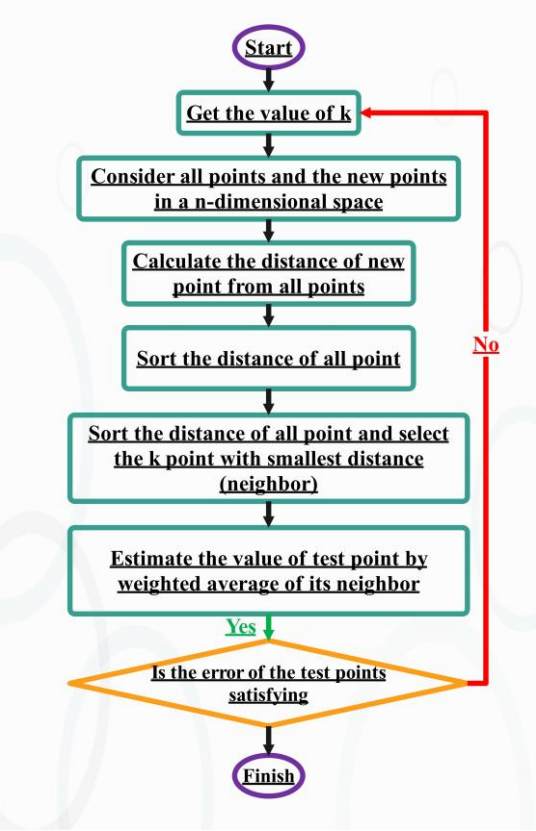


Figure 1: The flowchart of the KNN model

2.2 Elastic Net Regression (ENR)

ENR is a high-performance linear regression technique that combines the best features of L1 (Lasso) and L2 (Ridge) regularization techniques. Its dual regularization technique enhances the model's interpretability and prognostic performance, reduces multicollinearity and overfitting in high-dimensional datasets, and encourages sparsity and stability in coefficient estimates [22], [23].

- **Direct Communication**

$P(y|\beta,\sigma^2) = N(y|X\beta,\sigma^2 I_n)$ this is the probability of the article, where β is a p-vector that contains the regression coefficients. Where predictor variables are included in the $n \times p$ dimensional matrix X. Given that the vector y and the columns of X are considered to be demeaned, the model is specified without an intercept. In this approach, the linear regression parameters are typically estimated as follows:

$$\hat{\beta} = \underset{\beta}{arg\,min}\,(y - X\beta)^T(y - X\beta) + \lambda J(\beta) \quad (4)$$

Considering a nonnegative punishment function J and a regularization value $\lambda > 0$.

$$p(\beta|\lambda,\alpha) \propto exp[-\lambda\,\{\alpha|\beta|^2 + (1-\alpha)|\beta|_1\}] \quad (5)$$

This study offers a completely normalized and explicated version of the prior, extending the Bayesian connection to the ELR approach.

$$p(\beta|\alpha,\lambda,\sigma^2) \propto exp\left[-\frac{\lambda}{2\sigma^2}\,\{\alpha|\beta|^2 + (1-\alpha)|\beta|_1\}\right] \quad (6)$$

For given values of σ^2 and α, the posterior mode corresponds to the naïve elastic net estimate with an overall penalty of λ, and this is expressed by the formulation that states the penalty λ's magnitude is now $2\sigma^2$. This prior is a double-exponential distribution when $\alpha = 0$. When $\alpha \approx 1$, it exhibits the characteristics of a normal distribution. As demonstrated by the integration of Eq. (6), the normalization constant can be expressed in closed form until the evaluation of the univariate standard normal Cumulative Distribution Function (CDF). The appropriate density function on a smaller scale than the previous one.

$$p(\beta|\lambda,\alpha,\sigma^2) = \prod_{j=1}^{p}\left\{(0.5).N^-\left(\beta_j|\frac{1-\alpha}{2\alpha},\frac{\sigma^2}{\lambda\alpha}\right) + (0.5).N^+\left(\beta_j|-\frac{1-\alpha}{2\alpha},\frac{\sigma^2}{\lambda\alpha}\right)\right\}, \quad (7)$$

In truncated normal distributions, N^- and N^+ signifies density functions that have been suitably modified.

$$N^+(t|m,s^2) \equiv \frac{N(t|m,s^2)}{\phi(m/s)}\,1(t \geq 0)\ And\ N^-(t|m,s^2) \equiv \frac{N(t|m,s^2)}{\phi(-m/s)}1(t<0), \quad (8)$$

The tails of a normal distribution always contain the univariate standard normal CDF and $\phi.\beta_j$ since the location parameter for the positive component in Eq. (7) is always negative. The following is an alternative perspective on the previous: Let $Z = \{-1,1\}^p$ be the set of all p-vectors that can have members ± 1 and let $\mathcal{O}_z \subset \mathbb{R}^p$ be the orthant that corresponds to each vector z in Z. $\beta j \geq 0$for $\mathcal{Z}_j = 1$ and $\beta j < 0$for $\mathcal{Z}_j = -1$ if $Z.\beta j \geq 0$. Consequently, the preceding Eq. (10) may be expressed as follows:

$$p(\beta|\lambda,\alpha,\sigma^2) = 2^{-p}\phi\left(\frac{\alpha-1}{2\sigma\sqrt{\alpha/\lambda}}\right)^{-p} \times \sum_{z\in\mathcal{Z}} N\left(\beta|\frac{\alpha-1}{2\alpha}z,\frac{\sigma^2}{\lambda\alpha}I_p\right)1(\beta \in Oz). \quad (10)$$

To show that the prior is piecewise normal, an "orthant normal" prior is constructed by describing each piece over a separate orthant. The following results are obtained when the prior is represented regarding λ_1 and λ_2.

$$p(\beta|\lambda_1,\lambda_2,\sigma^2) = 2^{-p}\phi\left(\frac{-\lambda_1}{2\sigma\sqrt{\lambda_2}}\right)^{-p} \quad (11)$$

$$\times \sum_{z\in\mathcal{Z}} N\left(\beta|-\frac{\lambda_1}{2\lambda_2}z,\frac{\sigma^2}{\lambda_2}I_p\right)1(\beta\ \in\ Oz).$$

From now on, the (λ_1,λ_2) formulation is used unless otherwise noted. The posterior distribution is obtained by multiplying the probability of the regression model by Eq. (11) and using the Bayes theorem.

$$p(\beta|y,\lambda_1,\lambda_2,\sigma^2)=\sum_{z\in\mathcal{Z}}\omega_z N^{[z]}(\beta|\mu_z,\sigma^2R), \tag{12}$$

A weighted sum of the normal distributions represents a multivariate normal orthant integral following the trimming of the 2^P Orthant.

$$N^{[z]}(\beta|m,s)\ \equiv\ \frac{N(\beta|m,s)}{P(z,m,s)}1(\beta\ \in\ Oz),$$
$$where\qquad P(z,m,s)\ =\int_{\mathcal{O}_z}N(t|m,s)dt, \tag{13}$$

Since every component is defined on a different orthogon, the prior and posterior are both multivariate piecewise normal. Its posterior distribution is its collection of parameters.

$$R\ =\ \left(X^TX+\lambda_2\,I_p\right)^{-1}$$
$$and \tag{14}$$
$$\mu_z=\hat{\beta}_R-\frac{,\lambda_1}{2}Rz,$$

The ridge regression estimate in this instance is $\hat{\beta}_R = RX^Ty$, with a penalty of λ_2. The weights for each orthostat are the last components of the posterior.

$$\omega_z=\omega^{-1}\frac{P(z.\mu_z,\sigma^2R)}{N(0|\mu_z,\sigma^2R)},$$
$$where\ \ \omega=\sum_{z\in Z}\frac{P(z,.\mu_z,\sigma^2R)}{N(0|\mu_z,\sigma^2R)}. \tag{15}$$

2.3 Artificial Rabbits Optimization (ARO)

While the hares' natural persistence strategies inspired the endurance algorithm, the ARO algorithm was motivated by them. The model for this method was the bypass exploration tactic used by hares to emerge from their burrows in search of food. So that they are not startled when they are approached, hares dig tunnels near their hiding places. In cases where food is necessary or sufficient for them, they gravitate toward it by nature. A habit known as circumvention foraging occurs when rabbits, when their energy levels are high enough often look for food outside of their burrows. But, in times of low energy, they frequently merely plunge into nearby burrows to find cover [24].

- **Vitality Decrease (Switch between Exploitation and Exploration)**

Depending on their energy level, hares may decide to purposefully conceal or postpone foraging. An energy factor $A(t)$, which mimics the hare's decision-making process, is computed using Eq. (16). The hare chooses not to graze if $A(t)\ >\ 1$. Otherwise, if $A(t)$ equals or is less than 1, random concealment is used.

$$\mathrm{A(t)}=4\left(1-\frac{\mathrm{t}}{\mathrm{T}}\right)\ln\frac{1}{\mathrm{r}} \tag{16}$$

In this case, r is a number between 0 and 1 that is chosen at random.

- **Circumvent Foraging (Exploration)**

According to Eq (16), hares hunt for food at random and distant from their burrows based on the positions of their friends to defend themselves against potential predators: There are several variables in the formula. $\overrightarrow{x_i}(t)$ represents the ith hare's location at time t, whereas $\overrightarrow{S_i}(t+1)$. Represents the candidate's position at time $t\ +$ 1. T is the maximum number of cycles, and L is the hare's speed of movement. The variable n stands for the hare population size, and d signifies the count of variables requiring optimization in the problem. Furthermore, r_1, r_2 and r_3 represent the three random values in the interval (0,1) and n_1 has a standard normal distribution. R is the operating function that mimics the features of a hare running, while c is the mapping vector.

- **Unpredictable Concealment (Exploitation)**

Using Eq. (17), each rabbit burrow produces a collection of d lairs. To find refuge and evade any predators, the rabbit randomly selects various hiding places:

$$\overrightarrow{\mathrm{b_{i,j}}}(\mathrm{t})\ =\overrightarrow{\mathrm{x_i}}(\mathrm{t})+\mathrm{H}\times\mathrm{g}\times\overrightarrow{\mathrm{x_i}}(\mathrm{t}),\mathrm{i}$$
$$=1,\dots,\mathrm{n\ and\ j}=1,\dots,\mathrm{d} \tag{17}$$

$$\mathrm{H}=\frac{\mathrm{T}-\mathrm{t}+1}{\mathrm{T}}\times\mathrm{r_4} \tag{18}$$

$$\mathrm{g(k)}=\begin{cases}1\ \mathrm{if\ k}=\mathrm{j}\\ 0\ \mathrm{else}\end{cases}\quad\mathrm{k}=1,\dots,\mathrm{d} \tag{19}$$

$$\overrightarrow{\mathrm{S_i}}(\mathrm{t}+1)=\overrightarrow{\mathrm{x_j}}(\mathrm{t})+\mathrm{U}$$
$$\times\left(\mathrm{r_4}\times\overrightarrow{\mathrm{b_{i,r}}}(\mathrm{t})-\overrightarrow{\mathrm{x_i}}(\mathrm{t})\right)\mathrm{i}=1,\dots,\mathrm{n} \tag{20}$$

$$\mathrm{gr(k)}=\begin{cases}1\ \mathrm{if\ k}=|\mathrm{r_5}\times\mathrm{d}|\\ 0\ \mathrm{else}\end{cases}\quad\mathrm{k}=1,\dots,\mathrm{d} \tag{21}$$

$$\overrightarrow{\mathrm{b_{i,r}}}(\mathrm{t})=\overrightarrow{\mathrm{x_i}}(\mathrm{t})+\mathrm{H}\times\mathrm{gr}\times\overrightarrow{\mathrm{x_i}}(\mathrm{t}),$$
$$\mathrm{i}=1,\dots,\mathrm{n} \tag{22}$$

$$\overrightarrow{\mathrm{x_i}}(\mathrm{t}+1)=\begin{cases}\overrightarrow{\mathrm{x_i}}(\mathrm{t})\mathrm{f}\left(\overrightarrow{\mathrm{x_i}}(\mathrm{t})\right)\le\mathrm{f}\left(\overrightarrow{\mathrm{s_i}}(\mathrm{t}+1)\right)\\ \overrightarrow{\mathrm{s_i}}(\mathrm{t}+1)\mathrm{f}\left(\overrightarrow{\mathrm{x_i}}(\mathrm{t})\right)>\mathrm{f}\left(\overrightarrow{\mathrm{s_i}}(\mathrm{t}+1)\right)\end{cases} \tag{23}$$
$$\mathrm{k}=1,\dots,\mathrm{d}$$

According to Eq. (22), where H is the hiding factor, jth lair for the ith rabbit and r_4 and r_5 are arbitrary values in the interval (0,1), the arbitrarily selected lair that the ith rabbit would take refuge in is $\overrightarrow{b_{i,r}}(t)$.

2.4 Performance evaluators

To evaluate the predictive performance of the proposed models for CO_2 emissions and energy consumption in HPC manufacturing, the following metrics were employed:

- **Root Mean Square Error (RMSE):** Quantifies the square root of the average of squared differences between predicted and observed

values. RMSE penalizes large errors more than small ones, which is crucial in industrial applications where severe misestimation could lead to significant overuse of materials or excess emissions.

- **Mean Absolute Error (MAE):** Measures the average magnitude of errors in predictions without considering their direction. MAE is easy to interpret in physical units (e.g., MJ/m³ or kg/m³), which makes it particularly valuable for production engineers.
- **Mean Absolute Percentage Error (MAPE):** Indicates the average relative error as a percentage. MAPE is helpful for understanding how large the prediction error is compared to actual values, which supports decision-making for mix design optimization.
- **Nash–Sutcliffe Efficiency (NSE):** Compares the predictive power of a model to the mean of observed values. Values close to 1 indicate excellent predictive accuracy. An NSE > 0.9 is generally considered highly acceptable for environmental and materials modeling.
- **Coefficient of Determination (R^2):** Represents the proportion of variance explained by the model. An R^2 > 0.95 was set as a practical benchmark for high accuracy in this study, based on prior literature in concrete property prediction.

$$R^2 = \left(\frac{\sum_{i=1}^{n}(b_i - \bar{b})(d_i - \bar{d})}{\sqrt{\left[\sum_{i=1}^{n}(b_i - \bar{d})^2\right]\left[\sum_{i=1}^{n}(d_i - d)^2\right]}} \right)^2 \quad (24)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(d_i - b_i)^2} \quad (25)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(P_i - T_i)^2 \quad (26)$$

$$WAPE = \max\left[\frac{|b_i - m_i|}{b_i}\right] \quad (27)$$

$$NSE = 1 - \frac{\sum_{i=1}^{N}(m_i - b_i)^2}{\sum_{i=1}^{N}(b_i - \bar{b})^2} \quad (28)$$

In the context of predicting CO_2 emissions and energy consumption in HPC manufacturing, b_i in this study stands for the expected values for each measurement and p_i for the observed readings for each sample. The actual measured values for each sample are represented by T_i, whereas the word $\bar{b}$ indicates the mean of the anticipated values across all samples. Furthermore, ($\bar{T}$) denotes the mean of all measured values in the sample set, whereas $\bar{m}$ represents the average of the observed values. These symbols are crucial for computing and assessing the precision of the models created to forecast CO_2 emissions and energy usage during the manufacturing of HPC.

2.5 Rationale for model selection

KNN and ENR were selected for their complementary strengths. KNN is a non-parametric algorithm that captures non-linear relationships, while ENR is a linear model that can handle multicollinearity through regularization. The use of both allows the assessment of whether linear or non-linear models are more suitable for the given problem. ARO was integrated to enhance the exploration of hyperparameter space, especially for fine-tuning the models beyond traditional grid or random search methods.

Additionally, Recursive Feature Elimination (RFE) was incorporated to eliminate irrelevant or redundant features, reducing dimensionality and computational overhead while improving generalization.

3 Data description

The data from Kaggle makes a deep simulation of a concrete mixture with varying features that influence the environmental consequences (www.kaggle.com/datasets/taruneshburman/energy-consumption-prediction). This creates more opportunities for exploration in studies seeking to forecast concrete compressive strength with the influence of sustainability as an important factor. Feature descriptions are attached below:

Cement [kg/m^3]: Amount of cement utilized in the mixture. This accounts for so much in concrete strength but leads to high embodied CO_2 emissions due to the very energy-intensive production process.

Water [kg/m^3]: The water used to hydrate it. While an optimum water-to-cement ratio yields a durable concrete mixture, excess water reduces strength and, therefore, impairs sustainability.

Superplasticizer (kg/m^3): It is a chemical additive to the concrete mixture, enhancing workability without adding water. It improves performance but does not contribute much to environmental sustainability.

Coarse Aggregate [kg/m^3]: Is grit or any other similar material in the mixture; this variable influences the concrete strength and ecological impact due to its consumption of raw materials.

Fine Aggregate (kg/m^3): The weight of sand or other similar materials that would affect the structural integrity and sustainability of the mix.

Age (days): Time taken for curing concrete in days. Early-age strength is an essential factor in the construction schedule of concrete, as strength increases with age.

Compressive Strength (MPa): The target variable is the axial load-carrying capacity of the concrete. Measured in MPa, it is predicted based on input features.

Embodied CO_2 (kg): Lists the amount of CO_2 a mixture produced through its making and material usage, showing the carbon footprint.

Energy Consumption: Energy consumption means the total sum of all energy required to produce concrete mixture ingredients in megajoules. It is about features of EE related to the mix.

Resource Consumption (kg): This gives the sum of the mass of all the ingredients consumed in kg for one cubic meter of concrete production kind of resource intensity indicator for the mix.

Table 1 shows the descriptive statistics for all variables included in the dataset: maximum and minimum values, means, and the value of the standard deviation. The correlation between the input and output data is shown in Fig. 2 along with the matching correlation plot. Embodied CO_2 and cement, for instance, shows a substantial connection in the figure, suggesting a considerable link between the two variables.

Table 1: The statistical properties of dataset variables

Category	Variables	Indicators			
		Max	Min	Avg	St. Dev.
Input	Cement	499.9	201.4	347.1	87.6
Input	Water	220.0	150.2	185.5	20.5
Input	Superplasticizer	29.9	0	15.1	8.7
Input	Coarse Aggregate	1099.9	800.2	947.1	85.9
Input	Fine Aggregate	899.3	600	748.2	86
Input	Age	364	1	178.7	105.5
Input	Compressive Strength	615.2	87.8	226.9	109.1
Input	Resource Consumption	2652.76	1857.35	2242.98	150.27
Output	Energy Consumption	2849.19	1009.03	1919.69	436.161
Output	Embodied CO_2	507.878	198.304	350.711	82.7324

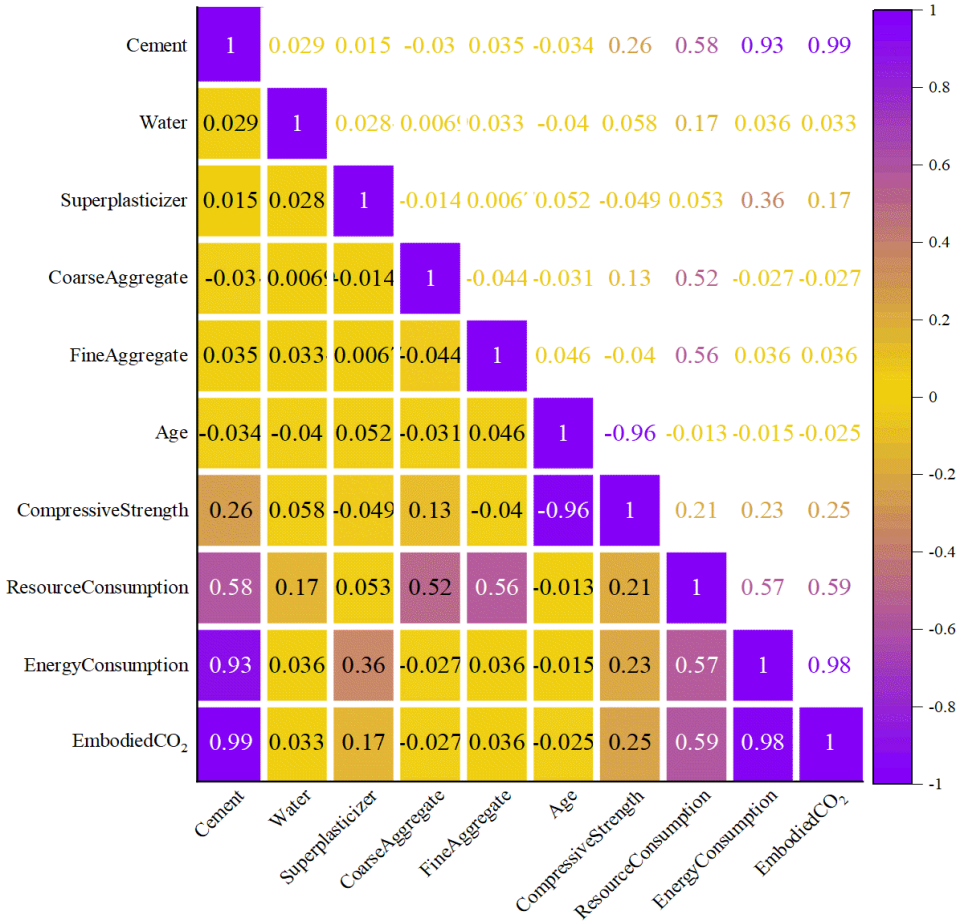
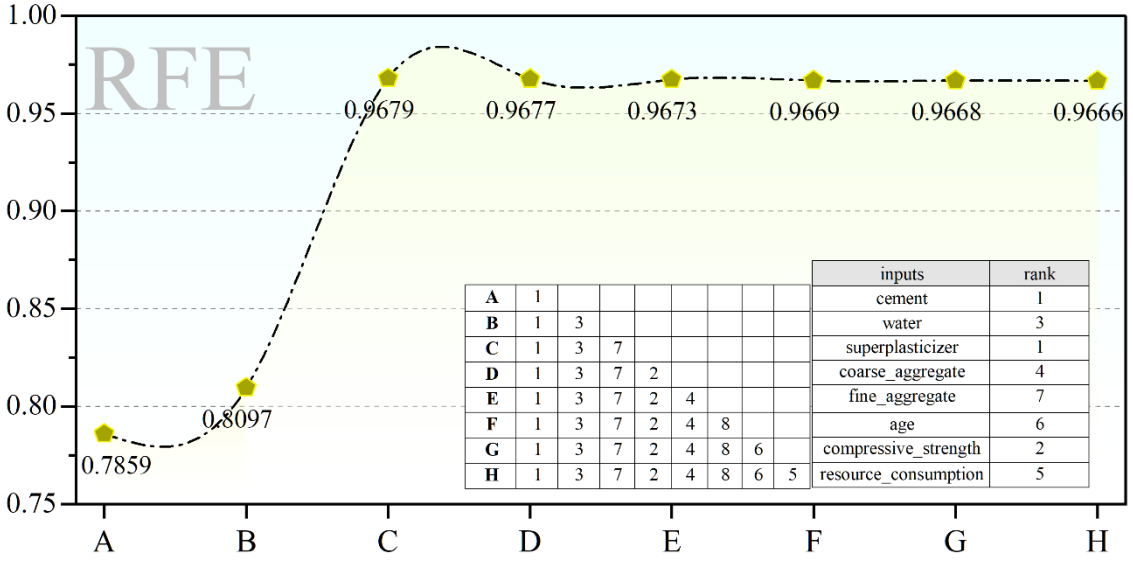


Figure 2: The correlation plot for input and output

3.1 Feature selection analysis for CO_2 emissions and energy consumption in HPC manufacturing: RFE process and key influences

The given analysis in Fig. 3 discusses the FS results for emissions and energy consumption in HPC manufacturing. This plot illustrates the RFE process, ranking features by their importance to model performance. The ranking score is displayed on the vertical axis, while the quantity of considered features is shown on the horizontal axis. The highest score, approximately 0.9679, reflects strong model performance with a specific feature subset. Beyond a certain point, additional features contribute minimally, suggesting that the most relevant information is captured within the top-ranked features. The ranking is topped by cement with rank 1, which signifies that cement has the highest impact on both CO_2 emissions, and energy consumption. Superplasticizer also holds the top rank, showing how critical it is in performance enhancement and resource use. Compressive Strength follows with a rank of 2, indicating it has a high influence. Water ranks third, which gives it considerable importance but less than cement and superplasticizers. Medium influences are given to Coarse Aggregate ranked 4 and Resource Consumption ranked 5, while low influences are accorded to Age and Fine Aggregate ranked 6 and 7, respectively, in this context. The quality of the cement, superplasticizer, and mainly, how green it can be is going to be the priority to reduce the amount of CO_2 emitted. This can be further refined by the optimization of water, compressive strength, and aggregate for performance versus greening balance. Other additional factors concerning resource consumption could also be considered; interaction between several features might also be modeled for better understanding in future work.



A	1							
B	1	3						
C	1	3	7					
D	1	3	7	2				
E	1	3	7	2	4			
F	1	3	7	2	4	8		
G	1	3	7	2	4	8	6	
H	1	3	7	2	4	8	6	5

inputs	rank
cement	1
water	3
superplasticizer	1
coarse_aggregate	4
fine_aggregate	7
age	6
compressive_strength	2
resource_consumption	5

Figure 3: Feature Importance for CO_2 Emissions and Energy Consumption Using RFE

The raw data underwent several preprocessing steps to ensure quality and compatibility with ML models:

- **Missing Values:** Samples with missing target values were discarded. Missing feature values (<5%) were imputed using mean imputation.
- **Normalization:** All features were normalized using min-max scaling to the range [0, 1] to avoid dominance of features with larger magnitudes.
- **Outlier Detection:** Z-score analysis was used to identify extreme outliers, which were reviewed manually and retained only if physically plausible.
- **Data Splitting:** The dataset was split into 70% training and 30% testing sets using stratified random sampling to preserve target distribution balance.

4 Results and discussion

The outcomes of utilizing ML regression models to forecast CO_2 emissions and energy usage in HPC manufacturing are shown in this section. Model performance indicators like RMSE, R^2, MSE, WAPE, and NSE are displayed in the figures and tables. These metrics demonstrate how well the models anticipate the intended results.

- **Analysis**

Fig. 4 compares the performance of two optimization models, namely ENAR and KNAR, in predicting embodied CO_2 emissions, and energy consumption in HPC manufacturing. The RMSE value across 200 iterations is used to assess the convergence trends. The result of the embodied CO_2 emissions for the ENAR model depicts a gradual convergence in terms of reducing the RMSE value to an ending value of 12.588 after roughly 150 iterations. It gives a good, moderate accuracy for the prediction. For the KNAR model, a final lower value of the RMSE was obtained at 7.952, which shows fast and consistent convergence; hence, this proves that the KNAR model is more efficient and more precise in modeling the CO_2 emissions. On the other hand, for energy consumption, ENAR demonstrates a very similar

convergence pattern for an RMSE value flattened to 69.706. It performs medium in developing its accuracy across subsequent iterations. Within this set of models, in turn, comes the performance realized by KNAR, yielding ultimately a better endpoint RMSE result of 83.815. Slower convergence against higher error is indicative of the resulting lower reliability in terms of the point accuracy of the forecasts pertinent to energy consumption. On the whole, KNAR performed better in predicting embodied CO_2 emissions than ENAR, while ENAR outperformed in energy consumption predictions owing to its lower RMSE values. These results further indicate that although KNAR is more robust for modeling CO_2 emissions, ENAR may be more suitable for the optimization of energy consumption in the manufacture of HPC.

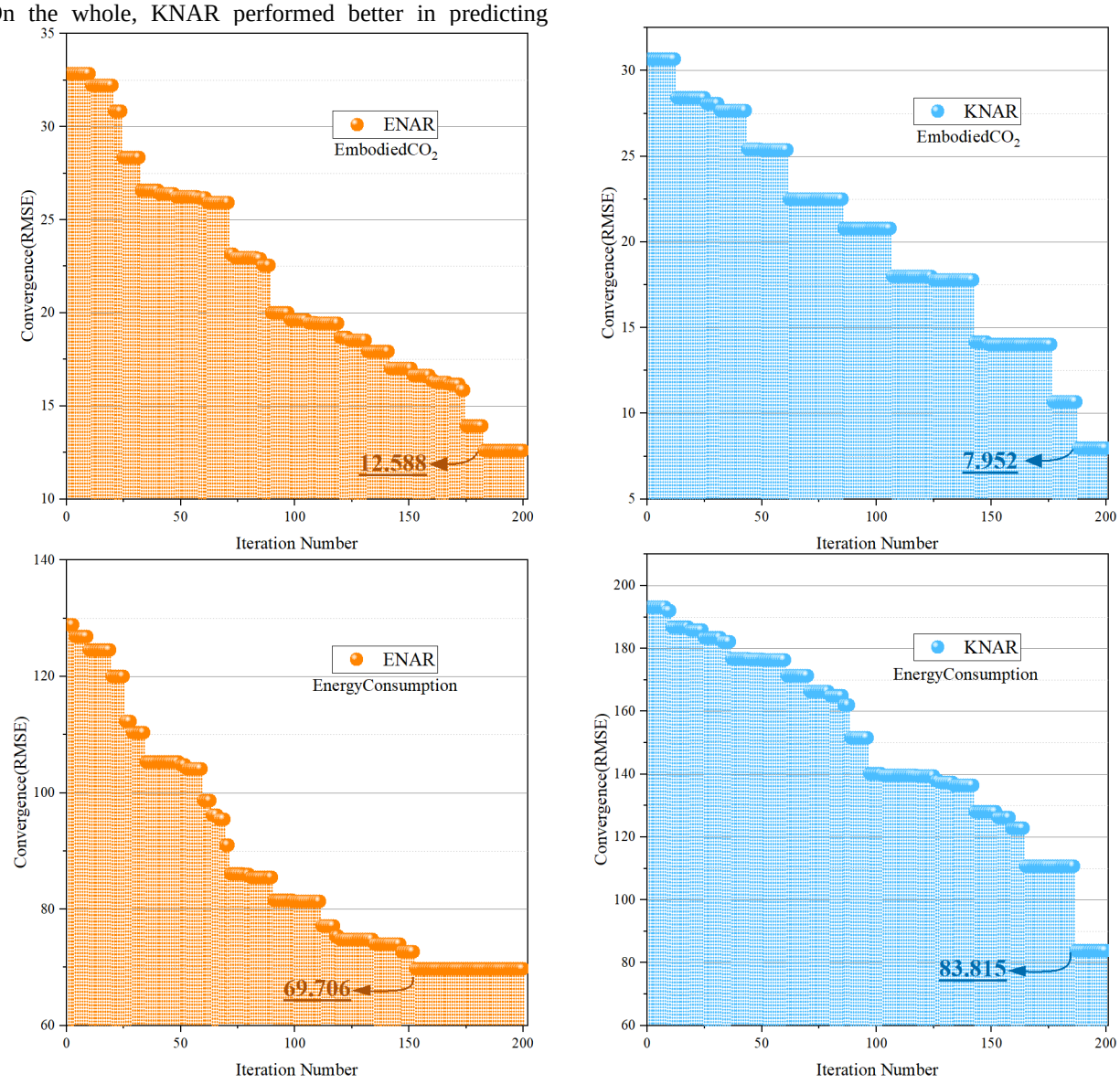


Figure 4: Scatter plot for the convergence Curve of the hybrid models

Energy consumption and embodied CO_2 analysis by Tables 2 and 3 underlines technical performance for the hybrid models of KNN, KNAR, EN, and ENAR through different metrics: the RMSE, R^2, MSE, WAPE, and NSE.

For energy consumption (Table 2): KNN achieves an R^2 value of 0.948 during all phases and an RMSE of 103,498. KNAR shows a notable improvement with an R^2 of 0.968 and an RMSE of 82.629. EN further improves upon this with an R^2 of 0.973 and an RMSE of 71,490. ENAR outperforms all models with the greatest R^2 of 0.986 and the lowest RMSE of 52.626, indicating superior performance in energy consumption prediction.

For embodied CO_2 (Table 3): KNN performs reasonably well with an R^2 of 0.977 and an RMSE of 12.588. KNAR excels in an R^2 of 0.992 and an RMSE of 7.573, showing significant improvement. EN achieves an $R^2$of 0.961 and an RMSE of 16.348, while ENAR follows with an R^2 of 0.975 and an RMSE of 13.083. KNAR outperforms all models in predicting embodied CO_2. In Conclusion, ENAR consistently delivers the best performance for energy consumption prediction, while KNAR is the most accurate model for predicting embodied CO_2, providing the best balance of R^2 and error (RMSE) in both cases.

Table 2: The results of hybrid models for the EN and KNN (Energy Consumption)

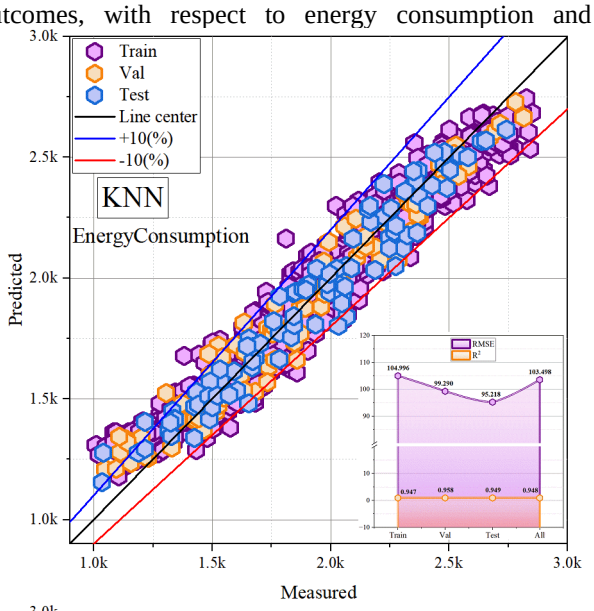
Model	Phase	Index values			
		Train	Validation	Test	All
KNN	RMSE	104.996	99.290	95.218	103.498
	R^2	0.947	0.958	0.949	0.948
	MSE	11024.228	9858.529	9066.412	10711.876
	WAPE	0.044	0.042	0.041	0.044
	NSE	0.942	0.952	0.946	0.944
KNAR	RMSE	83.815	79.345	76.026	82.629
	R^2	0.967	0.974	0.968	0.968
	MSE	7025.029	6295.656	5779.911	6827.580
	WAPE	0.035	0.034	0.033	0.035
	NSE	0.963	0.969	0.966	0.964
EN	RMSE	75.491	56.511	48.207	71.490
	R^2	0.970	0.986	0.987	0.973
	MSE	5698.893	3193.454	2323.951	5110.855
	WAPE	0.030	0.024	0.020	0.028
	NSE	0.970	0.984	0.986	0.973
ENAR	RMSE	55.364	40.752	38.896	52.626
	R^2	0.984	0.993	0.991	0.986
	MSE	3065.210	1660.707	1512.922	2769.531
	WAPE	0.022	0.018	0.017	0.021
	NSE	0.984	0.992	0.991	0.985

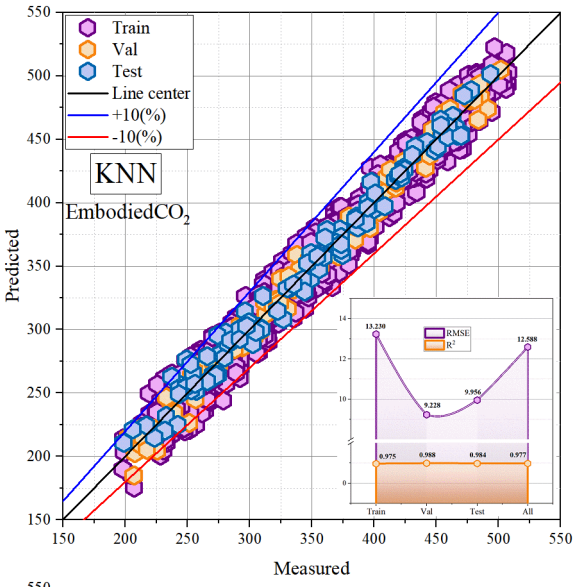
Table 3: The outcomes of hybrid models for the EN and KNN (Embodied CO2)

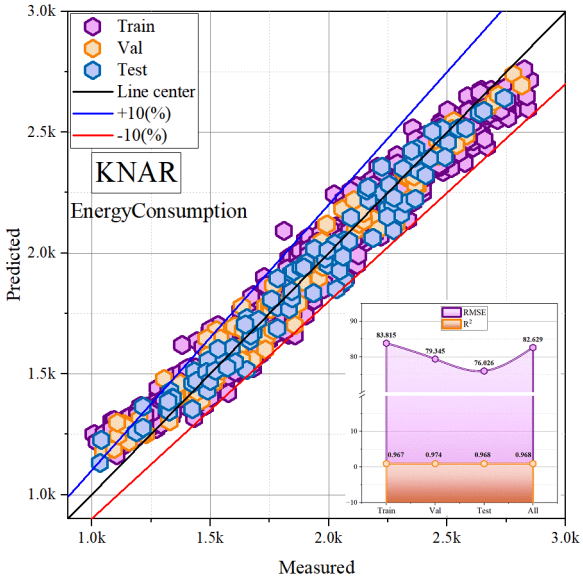
Model	Phase	Index values			
		Train	Validation	Test	All
KNN	RMSE	13.230	9.228	9.956	12.588
	R^2	0.975	0.988	0.984	0.977
	MSE	175.028	85.158	99.119	158.450
	WAPE	0.030	0.021	0.023	0.029
	NSE	0.975	0.988	0.984	0.977
KNAR	RMSE	7.952	5.607	6.012	7.573
	R^2	0.991	0.996	0.994	0.992
	MSE	63.237	31.439	36.148	57.348
	WAPE	0.018	0.013	0.014	0.017
	NSE	0.991	0.996	0.994	0.992
EN	RMSE	15.951	18.338	17.342	16.348
	R^2	0.963	0.953	0.952	0.961
	MSE	254.433	336.276	300.753	267.249
	WAPE	0.038	0.046	0.044	0.040
	NSE	0.963	0.952	0.950	0.961
ENAR	RMSE	12.589	15.124	14.665	13.083
	R^2	0.977	0.968	0.965	0.975
	MSE	158.471	228.733	215.051	171.155
	WAPE	0.030	0.038	0.037	0.032
	NSE	0.977	0.968	0.965	0.975

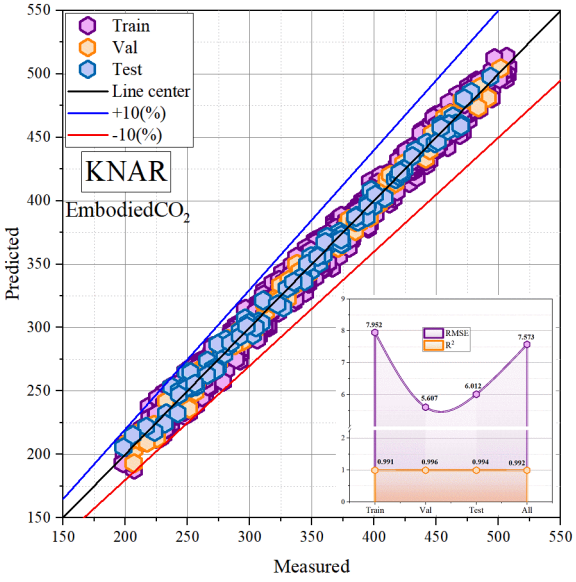
Fig. 5 provides comparisons in train versus test and validation data related to energy use and embodied CO_2, having as the line at the center the perfect prediction, while two dotted lines reflect ± 10% margin, showing the boundaries for the prediction errors. For energy consumption, KNN performed well in moderate clustering around the central line, though it generally featured substantial scatter, especially in the validation and test datasets, with higher prediction errors for extreme values. Even though similar results were obtained with KNN for the embodied CO_2 due to its lower RMSE, the wider scatter limits its accuracy for a proper prediction of consumption. The results show that KNAR offers improved clustering, with reduced scatter and smaller error margins, reflecting a higher robustness across datasets. Furthermore, it consistently remains nearer to the centerline, resulting in enhanced predictive capabilities for both energy usage and embedded CO_2. EN produces

tighter clustering than KNN, especially for lower and moderate energy consumption values, while it lacks the robustness of KNAR. The ENAR produces the best clustering with the least scatter and the highest accuracy in prediction on all datasets. Stability and precision in the outcomes, with respect to energy consumption and embodied CO_2, are superior in the current model when compared with other models. Overall, KNAR and ENAR show the highest reliability and accuracy, hence the best to use in assessing environmental impacts in HPC manufacturing.









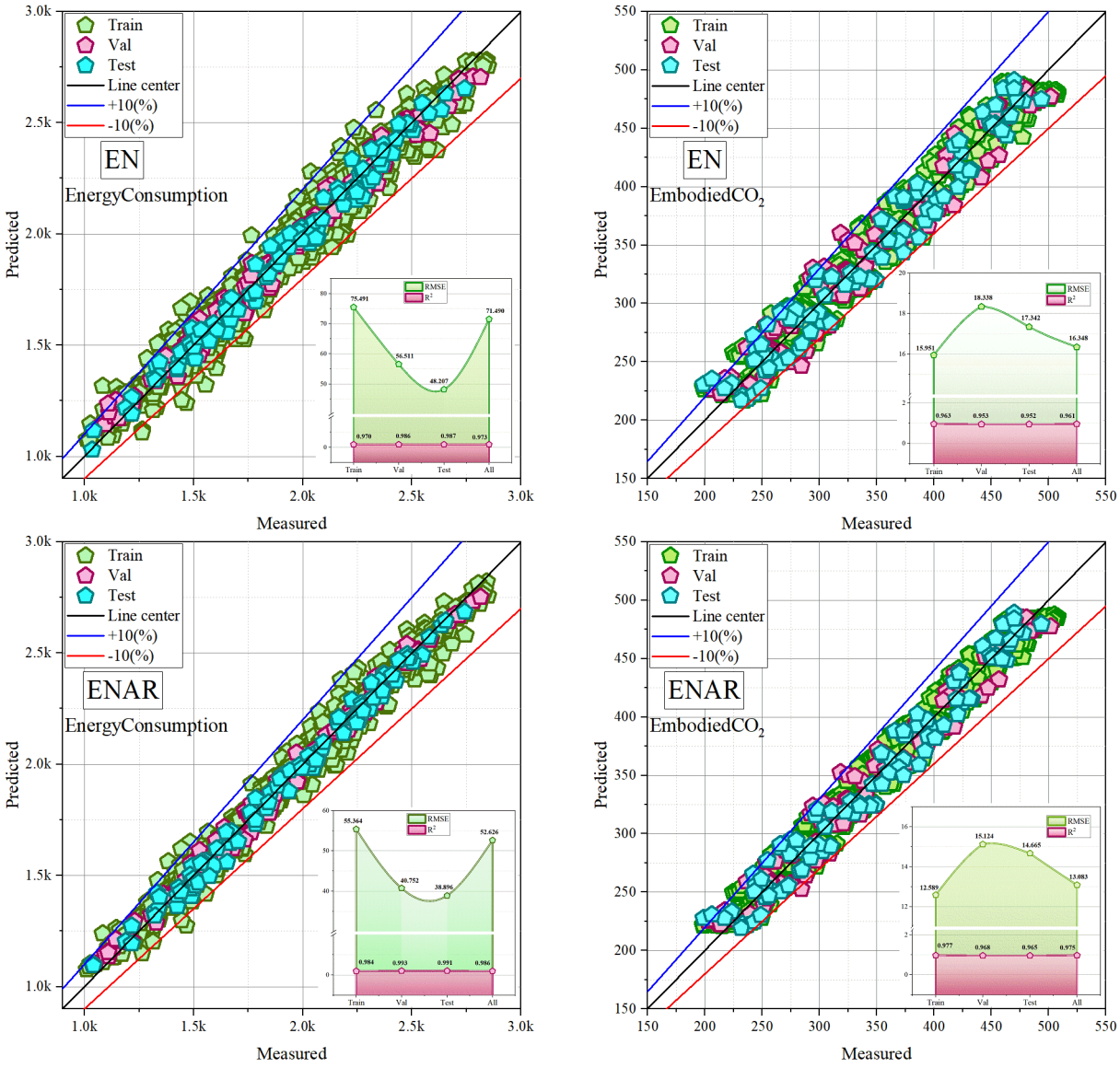


Figure 5: The scatter and line symbol plot for developed hybrid models

Fig. 6 is a violation plot that compares the four ML model errors according to KNN, KNAR, EN, and ENAR across Train Validation and Test in a way that densities may be visualized at different error values.

For Embodied CO_2: KNN is highly variable, with larger error margins, especially when it comes to validation and test datasets, a feature that ascribes to the low reliability of this method. The results indicate that the error distributions for all datasets are tighter, and better stability is displayed, especially at the validation and testing phases by KNAR. EN decreases error variance further than KNN, achieving higher accuracy and stability. Yet, it has a slightly weaker robustness performance than KNAR. ENAR achieves the tightest error distributions with the highest accuracy and robustness on all the datasets.

For Energy Consumption: KNN has a high variability in error, especially in validation and testing datasets, which means it has a poor generalization capability. KNAR has reduced variability, particularly tighter error clustering in the test dataset, which represents an increased reliability in predictive capabilities. EN has tighter error distributions than KNN, especially for smaller values, but is slightly worse than KNAR. ENAR performs the best with the lowest error margins and is consistent on all datasets. Conclusion: ENAR is the most accurate in energy consumption prediction, while KNAR is the best in embodied CO_2 prediction.

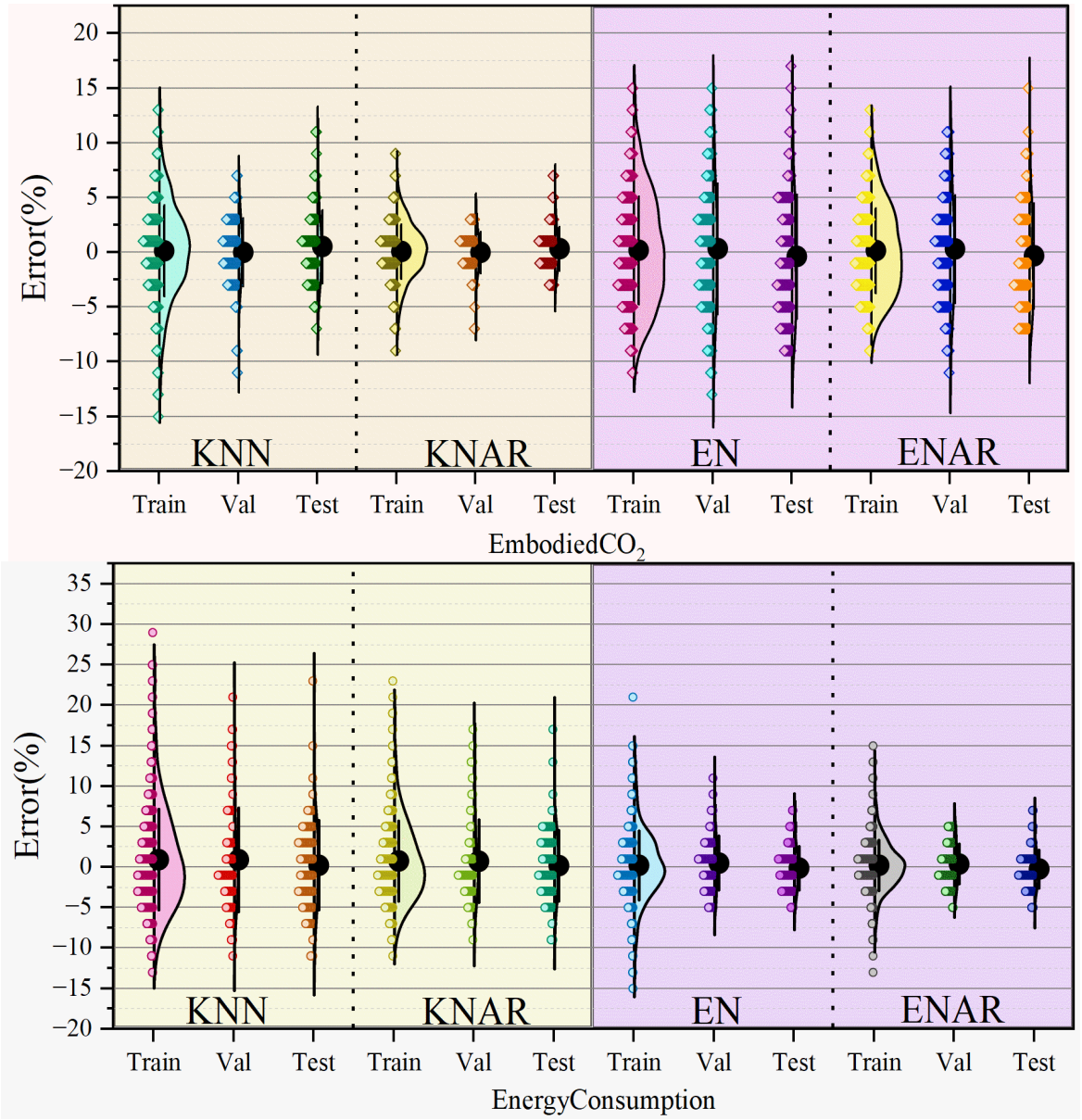


Figure 6: Comparing the errors of the developed models

The Taylor diagram in Fig. 7 provides a way to evaluate model performance against a reference dataset, combining the correlation coefficient, r, and standard deviation, σ, in one diagram. It can be applied to evaluate the accuracy of prediction in HPC manufacturing. In HPC manufacturing, this figure works very well for evaluating forecast accuracy. KNAR performs best in terms of embodied CO_2 since it is closest to the diagram's reference point, indicating a strong correlation and low RMSE. On the other hand, ENAR performs better than ENAR, whereas K-NN and KNAR, which are less accurate, perform worse. Additionally, ENAR outperforms all competing models in energy consumption predictions, obtaining the greatest correlation and lowest RMSE. While EN scores lowest, showing the largest RMSE and the shortest R^2 value, KNN and KNAR perform moderately. While KNAR performs better in estimating embodied CO_2, ENAR is shown to be the most accurate model overall for energy consumption forecasts.

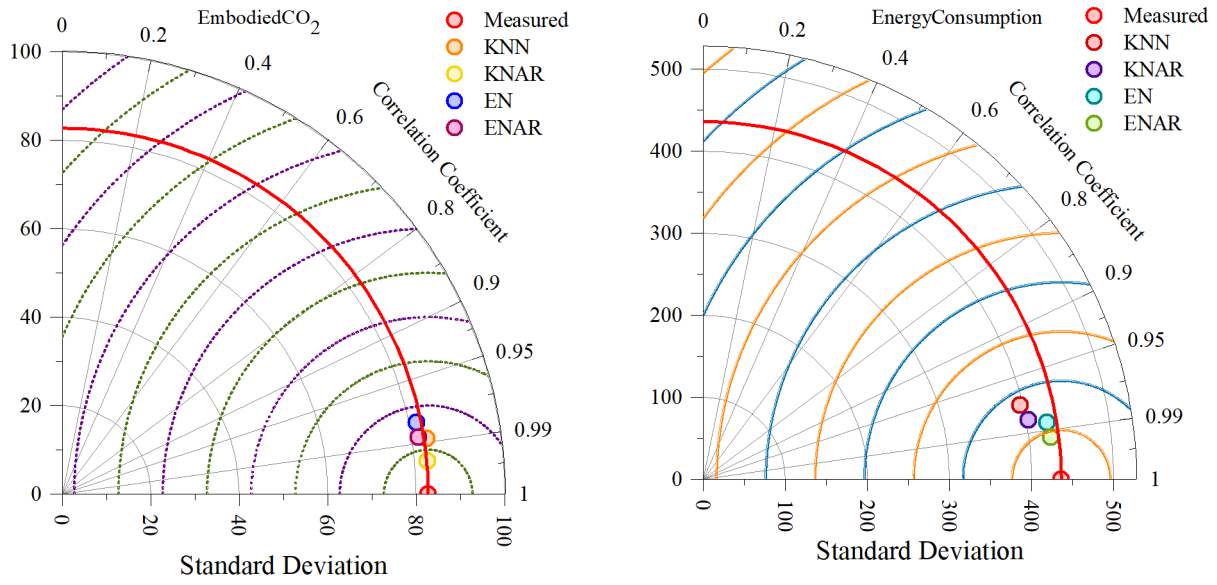


Figure 7: The Taylor diagram of related models with measured value

4.1 Practical use of the model in real-life building and planning

The proposed predictive framework provides a data-driven decision-support tool for real-world applications in concrete production and sustainable construction planning. By accurately forecasting CO_2 emissions and energy consumption based on HPC mix compositions, the model enables:

1. Environmentally Conscious Mix Design: Engineers can identify and minimize high-emission ingredients (e.g., excess cement or superplasticizer) during the early design stage, aligning material selection with carbon footprint targets.
2. Energy-Efficient Manufacturing Planning: By predicting energy demands of specific mix designs, batch plants can optimize their production schedules, reduce energy peaks, and lower operational costs.
3. Compliance with Green Building Standards: The model helps project teams evaluate whether proposed concrete mixes meet environmental criteria defined by certifications such as LEED, BREEAM, or China's Green Building Evaluation Label.
4. Lifecycle Sustainability Assessment: When integrated into Building Information Modeling (BIM) or digital twins, the model contributes to estimating the embedded carbon and energy profiles of structures early in the planning process, aiding urban planners and policymakers in meeting climate mandates.

Ultimately, the model acts as a predictive sustainability lens through which building materials can be assessed, optimized, and selected, without trial-and-error or over-reliance on empirical tables.

5 Conclusion

Most of the manufacturing processes for HPC systems are critical in terms of CO_2 emission and energy consumption, which should be minimized to support the environmentally responsive construction industry. Correct prediction of such factors is highly important for the optimization process of production and in maintaining complete harmony with sustainability considerations and the EE of the process. This work is focused on investigating the most influential ML models and the most influential features that have a great effect on the accuracy of the prediction in CO_2 emissions and energy consumption during the manufacturing process of HPC. ML techniques, including KNN, ENR, and ARO, were used to create predictive models. These models were further refined with optimizers to enhance prediction accuracy and identify the best solutions. Additionally, RFE was employed for FS, ensuring that only the most relevant variables were used to make precise predictions. The best performance for the estimation of energy consumption was given by the ENAR model, yielding an R^2 value of 0.986 and an RMSE of 52.626. The KNAR model performed best in estimating CO_2 emission, giving an R^2 value of 0.992 and an RMSE of 7.573. Therefore, the relevant features that most influenced the performance of the models were cement and superplasticizer. These were important in enhancing the predictive accuracy of energy consumption and CO_2 emissions for the models. The results have shown how important it is to consider the optimization of models along with input features in enhancing sustainability within HPC manufacturing. There are several drawbacks to utilizing ML to anticipate CO_2 emissions and HPC energy use. These include data quality and availability, as erroneous or inadequate datasets might impair model accuracy. Furthermore, it is challenging to generalize forecasts due to the intricacy of HPC systems and their diverse operating environments. Complex interactions between variables may be difficult

for ML models to capture, which might result in overfitting or underfitting. Lastly, scalability and real-time use in dynamic industrial contexts are limited by the computational cost and time needed to train big models on complicated datasets.

Acknowledgement

This work was supported by project of Yunnan Province Education Department Scientific Research Fund Project (No. 2023J1523) and (No. 2025J1423)

References

[1] S. Zhang, "Hybrid RBF-based prediction algorithms for evaluation of the compressive strength of HPC enhanced with blast furnace slag and fly ash," *Journal of Applied Science and Engineering*, Taiwan Association of Engineering and Applied Science, vol. 28, no. 11, pp. 2253–2264, Mar. 2025. DOI: 10.6180/jase.202511_28(11).0016.

[2] A. N. Arasu, N. Muthusamy, B. Natarajan, and R. Parthasaarathi, "Optimization of high performance concrete composites by using nano materials," *Research on Engineering Structures and Materials*, Trans Tech Publications, vol. 9, no. 3, pp. 843–859, 2023. http://dx.doi.org/10.17515/resm2022.602ma1213

[3] S. Wang and Q. Zhang, "Utilization of Machine-Learning-Based model Hybridized with Meta-Heuristic Frameworks for estimation of Unconfined Compressive Strength," *Journal of Applied Science and Engineering*, Taiwan Association of Engineering and Applied Science, vol. 28, no. 8, pp. 1779–1794, Nov. 2024. http://dx.doi.org/10.6180/jase.202508_28(8).0015.

[4] B. Li, R. Basu Roy, D. Wang, S. Samsi, V. Gadepally, and D. Tiwari, "Toward sustainable hpc: Carbon footprint estimation and environmental implications of hpc systems," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, ACM Digital Library, 2023, pp. 1–15. https://doi.org/10.1145/3581784.3607035.

[5] H. H. M. Ali, "Advances in High-Performance Concrete: A Comprehensive Review of Materials, Design, and Applications," *KHWARIZMIA*, vol. 2023, pp. 131–137, Aug. 2023. doi: 10.70470/KHWARIZMIA/2023/013.

[6] L. Hu, "Predicting the Compressive Strength of High-Performance Concrete utilizing Radial Basis Function Model integrating with Metaheuristic Algorithms," *Journal of Applied Science and Engineering*, Taiwan Association of Engineering and Applied Science, vol. 28, no. 8, pp. 1703–1715, Nov. 2024. DOI: 10.6180/jase.202508_28(8).0008.

[7] A. Zahoor, F. Mehr, G. Mao, Y. Yu, and A. Sápi, "The carbon neutrality feasibility of worldwide and in China's transportation sector by E-car and renewable energy sources before 2060," *J Energy Storage*, Elsevier, vol. 61, p. 106696, 2023. https://doi.org/10.1016/j.est.2023.106696.

[8] Y. Zhang and Y. Zhang, "Artificial Intelligence-Driven Models for Predicting Chloride Diffusion in Concrete: A Comparative Regression Analysis," *Journal of Artificial Intelligence and System Modelling*, vol. 03, no. 01, pp. 33–48, 2025. https://doi.org/10.22034/jaism.2025.495993.1092

[9] L. Mu, "Utilization of metaheuristic-based regression analysis to calculate the modified high-performance concrete's compressive strength," *Journal of Applied Science and Engineering*, Taiwan Association of Engineering and Applied Science, vol. 28, no. 8, pp. 1745–1758, Nov. 2024. DOI: 10.6180/jase.202508_28(8).0012.

[10] E. I. Aghimien, L. M. Aghimien, O. O. Petinrin, and D. O. Aghimien, "High-performance computing for computational modelling in built environment-related studies–a scientometric review," *Journal of Engineering, Design and Technology*, Emerald Group Publishing, vol. 19, no. 5, pp. 1138–1157, 2021. https://doi.org/10.1108/JEDT-07-2020-0294.

[11] B. Naeim, A. J. Khiavi, P. Dolatimehr, and B. Sadaghat, "Novel Optimized Support Vector Regression Networks for Estimating Fresh and Hardened Characteristics of SCC," 2024. https://doi.org/10.22034/aeis.2024.483317.1239.

[12] B. Naeim, M. R. Akbarzadeh, and V. Jahangiri, "Machine learning-based prediction of seismic response of elevated steel tanks," *Structures*, Elsevier, vol. 70, p. 107649, 2024. https://doi.org/10.1016/j.istruc.2024.107649.

[13] E. Khajavi, A. R. Taghavi Khanghah, and A. Javadzade Khiavi, "An efficient prediction of punching shear strength in reinforced concrete slabs through boosting methods and metaheuristic algorithms," *Structures*, Elsevier, vol. 74, p. 108519, 2025. https://doi.org/10.1016/j.istruc.2025.108519.

[14] H. Yazıcı, H. Yiğiter, A. Ş. Karabulut, and B. Baradan, "Utilization of fly ash and ground granulated blast furnace slag as an alternative silica source in reactive powder concrete," *Fuel*, Elsevier, vol. 87, no. 12, pp. 2401–2407, 2008. https://doi.org/10.1016/j.fuel.2008.03.005.

[15] N. Van Tuan, G. Ye, K. Van Breugel, and O. Copuroglu, "Hydration and microstructure of ultra high performance concrete incorporating rice husk ash," *Cem Concr Res*, Elsevier, vol. 41, no. 11, pp. 1104–1111, 2011. https://doi.org/10.1016/j.cemconres.2011.06.009.

[16] J. Dils, V. Boel, and G. De Schutter, "Influence of cement type and mixing pressure on air content, rheology and mechanical properties of UHPC," *Constr Build Mater*, Elsevier, vol. 41, pp. 455–

463, 2013. https://doi.org/10.1016/j.conbuildmat.2012.12.050.

[17] S. Abbas, A. M. Soliman, and M. L. Nehdi, "Exploring mechanical and durability properties of ultra-high performance concrete incorporating various steel fiber lengths and dosages," *Constr Build Mater*, Elsevier, vol. 75, pp. 429–441, 2015. https://doi.org/10.1016/j.conbuildmat.2014.11.017

[18] Ç. Yalçınkaya and H. Yazıcı, "Effects of ambient temperature and relative humidity on early-age shrinkage of UHPC with high-volume mineral admixtures," *Constr Build Mater*, Elsevier, vol. 144, pp. 252–259, 2017. https://doi.org/10.1016/j.conbuildmat.2017.03.198.

[19] Z. Wu, C. Shi, K. H. Khayat, and L. Xie, "Effect of SCM and nano-particles on static and dynamic mechanical properties of UHPC," *Constr Build Mater*, Elsevier, vol. 182, pp. 118–125, 2018. https://doi.org/10.1016/j.conbuildmat.2018.06.126.

[20] Y. Akbulut, A. Sengur, Y. Guo, and F. Smarandache, "NS-k-NN: Neutrosophic set-based k-nearest neighbors classifier," *Symmetry (Basel)*, MDPI, vol. 9, no. 9, p. 179, 2017. https://doi.org/10.3390/sym9090179.

[21] Y. Qian, W. Zhou, J. Yan, W. Li, and L. Han, "Comparing machine learning classifiers for object-based land cover classification using very high resolution imagery," *Remote Sens (Basel)*, MDPI, vol. 7, no. 1, pp. 153–168, 2014. https://doi.org/10.3390/rs70100153.

[22] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J R Stat Soc Series B Stat Methodol*, Wiley, vol. 67, no. 2, pp. 301–320, 2005. https://doi.org/10.1111/j.1467-9868.2005.00503.x.

[23] C. Hans, "Elastic net regression modeling with the orthant normal prior," *J Am Stat Assoc*, Taylor & Francis, vol. 106, no. 496, pp. 1383–1393, 2011. https://doi.org/10.1198/jasa.2011.tm09241.

[24] A. J. Riad, H. M. Hasanien, R. A. Turky, and A. H. Yakout, "Identifying the PEM Fuel Cell Parameters Using Artificial Rabbits Optimization Algorithm," *Sustainability*, MDPI, vol. 15, no. 5, p. 4625, 2023. https://doi.org/10.3390/su15054625.