Multi-objective optimization of engineered cementitious composite based on machine learning and generative adversarial network

Authors: Yufei Wang[1,2], Junbo Sun[3], Xiangyu Wang[2*], Shengping Li[1], Hongyu Zhao[3], Bo Huang[4], Yujie Cao[2], Mohamed Saafi[5]

Yufei Wang[1,2], PhD, PhD Candidate
[1] School of Design and the Built Environment, Curtin University, Perth, WA 6102, Australia
[2] School of Civil Engineering and Architecture, East China Jiao Tong University, Nanchang 330013, China.
E-mail address: wangyf0113_suz@163.com

Junbo Sun[3], PhD, Senior Researcher
[3] Institute for Smart City of Chongqing University In Liyang, Chongqing University, Jiangsu 213300, China;
E-mail address: tunneltc@gmail.com

Xiangyu Wang[2*], PhD, Professor
[2] School of Civil Engineering and Architecture, East China Jiao Tong University, Nanchang 330013, China.
E-mail address: Xiangyu.Wang@curtin.edu.au

Shengping Li[1], PhD, Lecturer
[1] School of Design and the Built Environment, Curtin University, Perth, WA 6102, Australia;
E-mail address: Shengping.Li@curtin.edu.au

Hongyu Zhao[3], PhD, PhD Candidate
[3] Institute for Smart City of Chongqing University In Liyang, Chongqing University, Jiangsu 213300, China;
E-mail address: 20211601069@cqu.edu.cn

Bo Huang [4], PhD, Lecturer
[4] School of Civil Engineering, Hunan University of Science and Technology, Xiangtan, 411201, China;
E-mail address: Bohuang@hnust.edu.cn

Yujie Cao [2], Undergraduate
[2] School of Civil Engineering and Architecture, East China Jiao Tong University, Nanchang 330013, China.
E-mail address: yujie_cao2003@163.com

Mohamed Saafi[5], PhD, Professor
[5] Department of Engineering, Lancaster University, Lancaster, LA1 4YR, UK
E-mail address: m.saafi@lancaster.ac.uk

*Corresponding Author: Xiangyu Wang, Xiangyu.Wang@curtin.edu.au

Abstract: This study aims to establish a novel framework for mixture design optimization of engineered cementitious composite (ECC) by first collecting two original datasets of ECC's tensile stress and strain from the extensive and credible literature. The datasets comprise a wide range of variables including cementitious ingredients, 9 types of fiber and characteristics, admixtures, and experimental conditions. The data augmentation is then performed using a tuned constraints-modified Conditional Tabular Generative Adversarial Network (Tuned-CTGAN) to increase the model accuracy and generalizability. The fitness functions of tensile stress and strain are established based on four machine learning models with the hyperparameters tuned by the Hunger Games search (HGS) algorithm. After the data augmentation, the values of R^2 in their test sets are increased from 0.874 to 0.925 and from 0.772 to 0.889, respectively. Subsequently, the third objective function (cost) is computed by polynomials and four classes of constraints (Min-max, volume, ratio, and fiber) are set up to define the variable's search space. A non-dominated sorting genetic algorithm based on reference-point strategy (NSGA-III) is introduced to optimize the mixture proportions of ECC by simultaneously optimizing tensile stress, tensile strain, and cost. This paper combines the results of data augmentation, model prediction, and multi-objective optimization for complex ECC design, which aims to provide a basis for practical application.

Keywords: engineered cementitious composite; machine learning; generative adversarial network; multi-objective optimization

1. Introduction

Engineered Cementitious Composite (ECC) is a strain-hardening cementitious material recognized for its high tensile strength and ductility. When fibers are evenly dispersed in the composite matrix and their volume is kept below 2%, ECC exhibits strain-hardening behavior with multi-cracks under 100μm. ECC's tensile strength typically ranges from 4 to 20 MPa, with a tensile strain capacity between 3 and 12% [2]. The tensile properties of ECC are significantly influenced by the mixture design and fiber types. The choice of fibers such as polyethylene fiber (PE), polyvinyl alcohol fiber (PVA), polypropylene fiber (PP), and steel fiber (SF) and their specific properties (strength, diameter, length, Young's modulus, oil coating) significantly affect ECC's performance. For instance, PVA-ECC can reach a tensile strain of 5%, while PE-ECC may achieve up to 13% [2, 3]. Determining the optimal mix for ECC to meet specific tensile strength and strain poses a major challenge, mainly due to the dependence on trial-and-error methods. Moreover, classical micro-mechanical design theories for ECC can exhibit whether a mix exhibits strain hardening, while it fails to provide targeted parameters [4, 5].

Nowadays, the rise of artificial intelligence (AI) and Machine Learning (ML) has revolutionized the design process. Advanced ML models, such as neural networks (ANN), support vector regression (SVR), and tree-based models, are now frequently used to accurately predict concrete properties [6-8]. For instance, Shariati, Armaghani, et al. (2021) [9] effectively used ANN and Extreme Learning Machine (ELM) to predict the strength of concrete containing additives, while Feng et al. (2020) [10] leveraged an adaptive boosting algorithm for greater predictive accuracy. The eXtreme Gradient Boosting (XGBoost) algorithm, developed by Chen and Guestrin, represents a significant advancement in the realm of gradient boosting machines [11]. It is particularly effective in mitigating overfitting while enhancing computational efficiency. Hunger Games Search (HGS) is an advanced meta-heuristic algorithm inspired by the concept of 'Hunger-Driven Motivational State Competition', as first introduced by Yang, Chen et al. [12]. In this study, the HGS algorithm is selected for ML hyperparameter adjustment.

ML models in ECC design also suffer from limited generalization due to small databases, as evident in Huynh et al.'s 2020 study on predicting geopolymer concrete strength [13]. A potential solution is data augmentation using Generative Adversarial Networks (GAN), an advanced algorithm based on the theory of a two-player game (discriminator D and the generator G) [14]. While GANs have been extensively used in computer vision, particularly for image generation, their application in augmenting concrete datasets is still limited [15, 16]. Conditional Tabular GAN (CTGAN), a GAN-based model, is particularly designed for synthesizing tabular data. CTGAN addresses data imbalance in databases by using variational Gaussian mixture models for continuous features and conditional generation for discrete ones. Therefore, the generalization and accuracy of ML models are predicted to be improved by introducing a Tuned-CTGAN model which is specifically designed for the ECC dataset. To optimize multiple objectives under highly nonlinear constraints, a non-dominated sorting genetic algorithm based on reference-point strategy (NSGA-III) was first introduced by Deb and Jain (2014) [17]. This study proposed NSGA-III to optimize the mixture proportions of ECC by simultaneously optimizing tensile stress, tensile strain, and cost.

This study introduces a novel framework integrating Tuned-CTGAN for ECC data augmentation, HGS optimized ML models for performance prediction, and multi-objective optimization via NSGA-III for ECC design optimization. Through enhancing model accuracy and enabling precise ECC mixture formulation, this study is expected to contribute to the application of AI techniques to ECC, providing valuable insights and practical tools for optimizing ECC mixtures with improved performance and cost-effectiveness.

2. Literature review and comparative analysis

The adoption of AI techniques for predicting and optimizing the properties of ECC has been significantly developed. This literature review synthesizes key contributions between 2018-2024, highlighting the diverse approaches and outcomes in this domain, as shown in Table 1. It covers a diverse range of fiber materials, such as PVA, PE, and steel fibers, each selected for their potential to enhance specific aspects of ECC performance. The focus of ECC research has predominantly been on enhancing tensile strength and ductility [18-20]. The additional ECC properties such as self-healing capacity (SC), freeze-thaw resistance (FR), and chloride ion permeability (CP) are also gaining attention [21, 22].

Recent research has explored a variety of ML algorithms, including Random Forest (RF), Support Vector Machine (SVM), and eXtreme Gradient Boosting (XGBoost). For instance, Uddin, Shanmugasundaram [23] employed multiple ML techniques to predict the compressive strength and tensile strain of ECC. The capability of ECC for self-healing was modeled using ensemble ML algorithms by Alabduljabbar, Khan [24], showcasing the potential of AI in enhancing the durability aspects of ECC. The inclusion of hybrid and advanced ML models has also been noteworthy. Tanyildizi [25] utilized hybrid deep learning models to predict the compressive strength of nano-silica-modified ECC under high temperatures. Moreover, invertible neural networks (INNs) was applied by Yu, Weng [5] for the performance-based design of ECC mixtures, illustrating a novel approach to ECC formulation that aligns with specific mechanical and sustainability requirements. However, only one group of mixture design can be obtained by one output. Besides, response surface methodology was also used to assess the importance of variables and optimize the mixture design of ECC [26-28].

However, few studies focused on the data treatment before applying the ML or DL methods. According to the authors' review, Mahjoubi, Barhemat [19] utilized a decision tree algorithm based on the isolated forest method to remove anomalous data. Guo, Meng [29] conducted the Principal Component Analysis and Semi-empirical model to enlarge the database. A commonly used method in other publications is data normalization, which indicates that the data treatment research is limited. Therefore, to fill this gap, this study attempted to introduce a tuned-CTGAN model specifically designed for ECC to enlarge the raw database, enabling the training of more generalized and robust models. This approach is novel within the ECC optimization domain and allows for a better understanding of ECC's complex behavior. Besides, the majority of the research concentrates on a limited range of fiber types, predominantly PVA and steel fibers, without extensive analysis across a broader spectrum. This study investigated 9 different fiber types and collected around 400 data points for both tensile strength and tensile strain.

In summary, while existing studies have laid a solid foundation in the application of AI for enhancing ECC's properties, this research introduces data augmentation method, optimization strategies, various fiber types, and large database that further the capabilities of AI in this field.

Table 1. The literature review and comparative analysis of recent ECC researches

Fiber type of ECC	Investigated properties and data points	Data treatment	ML prediction	Optimizati on design	Refer ences	Year

9 fiber types (PP, PVA, etc.)	TS (429), TSt (392)	Tuned-CTGAN	HGS-XGBoost, Gradient Boosting Regressor, RF, SVR	NSGA-III	This study	2024
PVA, PE, PP, basal, glass fiber	CS (180), TSt (105)	-	RF, SVM, XGBoost, light gradient boosting machine, CatBoost, natural gradient boosting, MEMD-ADE-SVM	-	[23]	2024
PVA fiber	CS (100)	-	AE-ELM, AE-DT	-	[25]	2024
Not specified	SC (617)	-	AR, DT, BR	-	[24]	2023
PE fiber	TS (129), TSt (129)	-	ANN	INNs	[5]	2023
Not specified	SC (619)	-	SVM, XGBoost, RF	-	[30]	2023
Steel fibers (twisted, hooked, and smooth fibers)	TS (103), TSt (103)	-	ANN	-	[31]	2023
Steel fibers (straight, hooked-end, and spiral geometric)	Pull-out force and slip (382)	-	ANN	-	[32]	2023
Not specified	CS (182), TS (97), FS (50)	-	GEP	-	[20]	2023
PVA fiber	CS, TS, TSt, FS, EM, PR, DS (13)	-	RSM	RSM	[27]	2023
Polyacrylonitrile fiber	FR, CP (36)	-	RSM	NSGA-III	[28]	2023

PVA fiber	SC (617)	-	LR, ANN, CART, SVM, ensemble methods (bagging, AdaBoost, and stacking)	-	[21]	2023
Not specified	SC (617)	-	BR, SR	-	[22]	2023
PVA Fiber	CS (89)	-	ANN	-	[33]	2022
PVA fiber	CS (79), TS (36)	-	ANN	-	[34]	2022
PVA fiber	CS (151), FS (76), TS-TSt (44)	-	ANN	-	[35]	2022
PVA fiber	CS, FS, TS, TC (16)	-	RSM	RSM	[26]	2021
PVA, PP, PE, and steel fiber	CS (264), TS (244), TSt (237)	A decision tree algorithm based on the isolated forest method to remove anomalous data	SVM, AR, XGBoost	UNSGA-III and NSGA-III	[19]	2021
10 fiber types (PP, PVA, etc.)	TS (284), TSt (293), FS (189), FSt (166), CS (313), etc.	-	FDNN	-	[36]	2021
Chopped fiber	CS (387), TS (387), TSt (387)	Principal Component Analysis + Semi-empirical model	ANN, SVM, CART, XGBoost	-	[29]	2021
Steel fiber	CS (220), FS (220)	-	9 ML (RF, AR, etc.)	-	[37]	2021
PVA fiber	TS (36), TSt (36)	-	ANN	Numerical analysis	[18]	2019

PVA, steel fiber	CS (24), TS (44), FS (47), TSc (54)	-	ANN	-	[38]	2018

Note: the abbreviation is as follows:

Algorithms: Random forest (RF), support vector machine (SVM), eXtreme Gradient Boosting (XGBoost), categorical gradient boosting (CatBoost), AdaBoost regressor (AR), decision tree (DT), bagging regressor (BR), Artificial neural networks (ANN), gene expression programming (GEP), response surface methodology (RSM), linear regression (LR), classification and regression tree (CART), stacking regressor (SR), forest deep neural network (FDNN), autoencoder (AE), extreme learning machine (ELM), multivariate empirical modal decomposition (MEMD), adaptive differential evolution (ADE), Invertible neural networks (INNs), Unified NSGA-III (UNSGA-III).

Properties: compressive strength (CS), tensile strength (TS), tensile strain (TSt), flexural stress (FS), flexural strain (FSt), self-healing capacity (SC), elastic modulus (EM), Poisson's ratio (PR), drying shrinkage (DS), Freeze−thaw resistance (FR), chloride ion permeability (CP).

3. Database description

To accurately predict the generalized properties of ECC, it is imperative to assemble a meticulously curated comprehensive database. This database must encompass a broad spectrum of key variables, reflecting a general distribution. Its establishment is grounded in the utilization of existing datasets coupled with an extensive review of contemporary literature. This process adheres to stringent criteria to ensure the reliability and relevance of the data included [3, 36, 39-92]:

(1) The literature references the use of Ordinary Portland Cement in producing ECC;

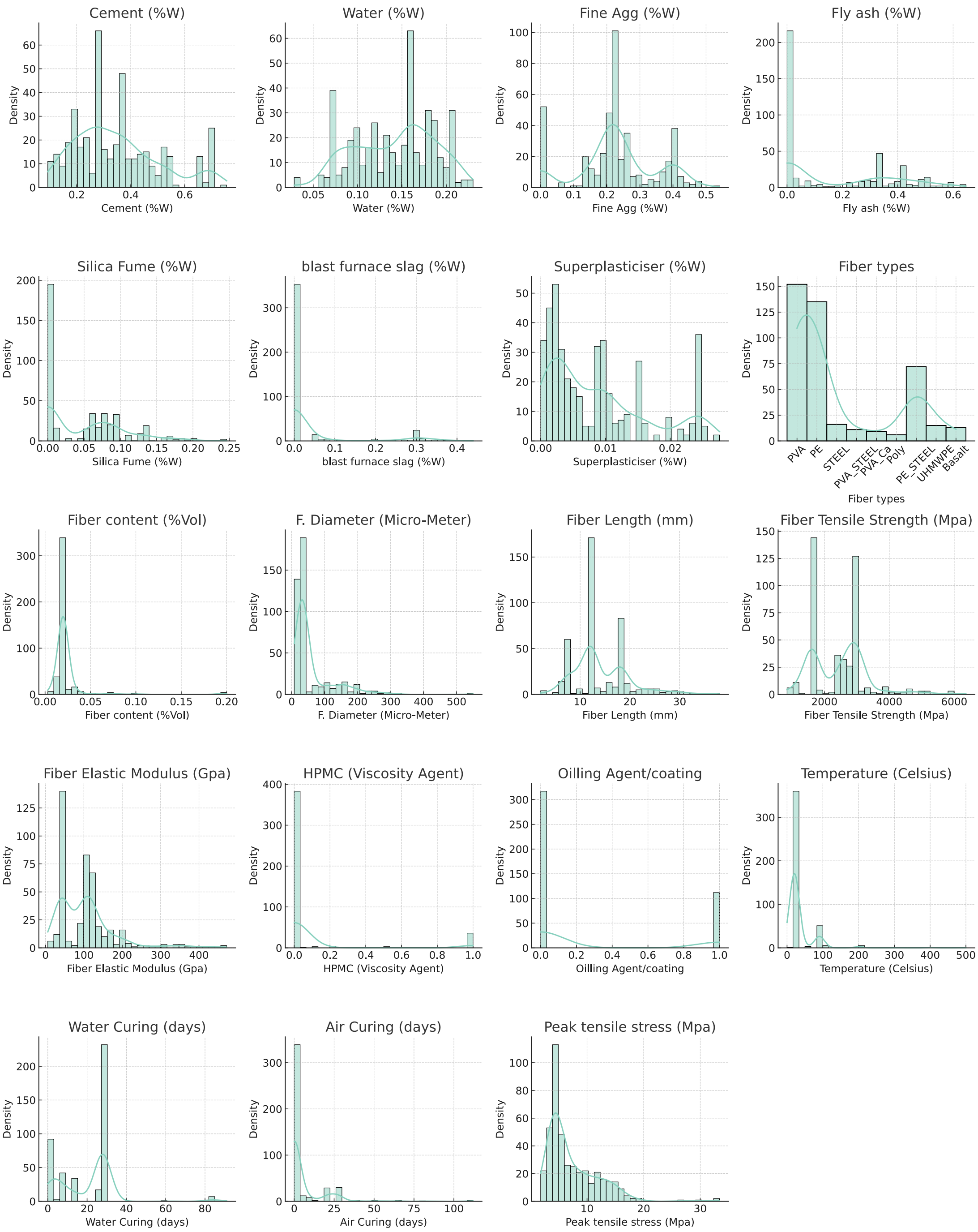
(2) For fine or coarse aggregates, natural aggregates with proper size distribution are chosen over types like lightweight aggregate;

(3) To ensure data reliability, references are drawn from authoritative international journals;

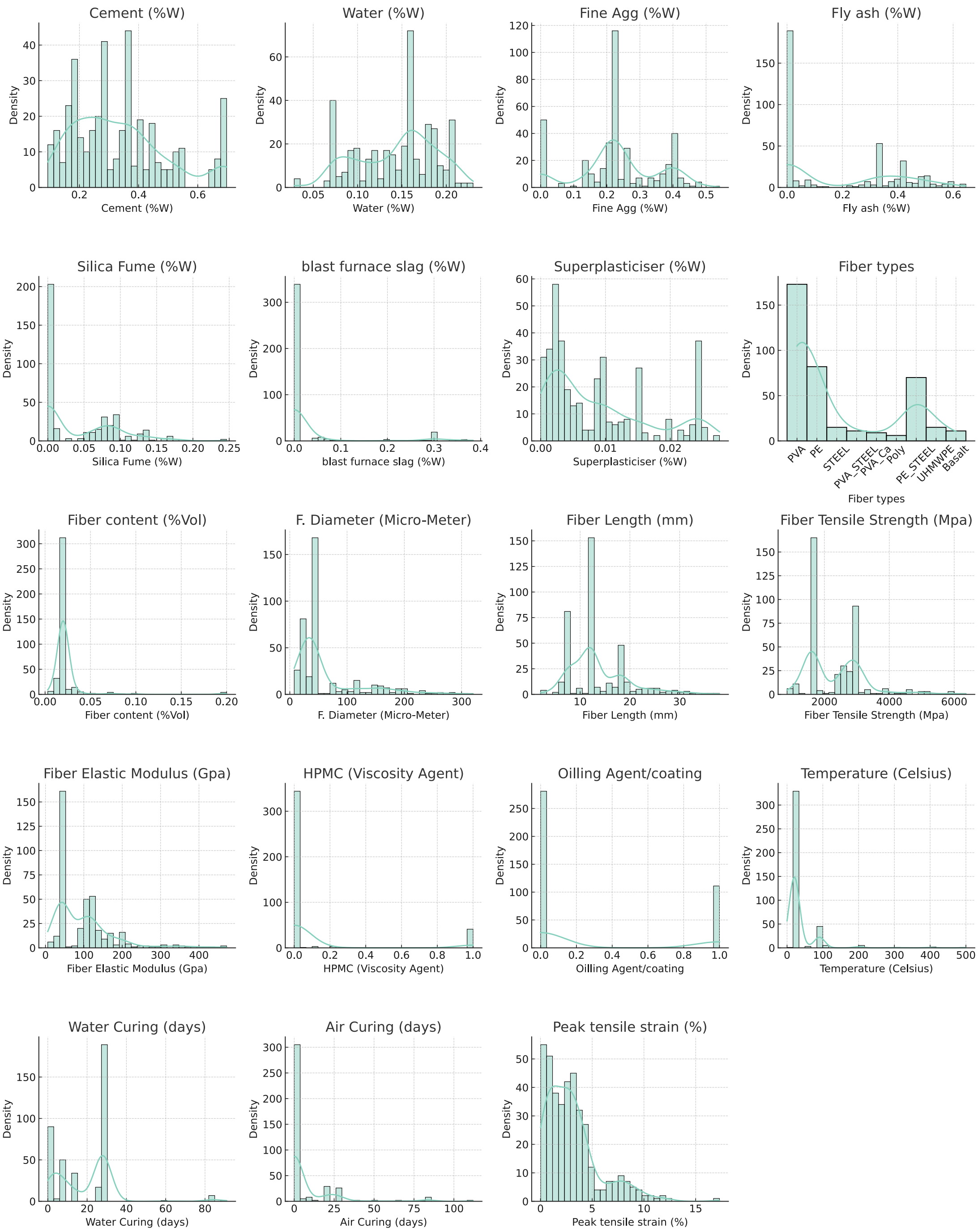
(4) The characteristics of fibers and other admixtures are clearly stated.

The database comprehensively records 18 ECC features and 2 outputs, namely tensile stress and strain. The features fall into four main categories: main ingredients of mortar normalized to a total weight of 1, including elements like cement, water, and fly ash; fiber characteristics such as fiber type, volume proportion, and diameter; admixtures including superplasticizer, viscosity agent, and oiling agent; and experimental conditions covering aspects like temperature, water curing days, and air curing days. The distribution plots of these features and output are depicted in Figure 1.

The database encompasses a wide array of fiber reinforcements, such as PE, PVA, PP, basalt, steel, ultra-high molecular weight polyethylene fiber (UHMWPE), and hybrid fibers like PE-steel, PVA-calcium carbonate whiskers, and PVA-steel. Each of these selected features has been verified to significantly impact the strength and ductility of ECC [2]. The outputs of the dataset are the peak tensile stress (MPa) and peak tensile strain (%). A statistical analysis of ECC stress and strain datasets is summarized in Appendix 1 and Appendix 2.


Cement (%W)
Water (%W)
Fine Agg (%W)
Fly ash (%W)
Silica Fume (%W)
blast furnace slag (%W)
Superplasticiser (%W)
Fiber types
PVA
PE
STEEL
PVA_STEEL
PVA_Ca
Poly
PE_STEEL
UHMWPE
Basalt
Fiber content (%Vol)
F. Diameter (Micro-Meter)
Fiber Length (mm)
Fiber Tensile Strength (Mpa)
Fiber Elastic Modulus (Gpa)
HPMC (Viscosity Agent)
Oilling Agent/coating
Temperature (Celsius)
Water Curing (days)
Air Curing (days)
Peak tensile stress (Mpa)
Density

(a)



(b)

Figure 1. The distribution plots of each feature and output in (a) the tensile stress dataset and (b) the tensile strain dataset.

4. Methodologies

The methodology employed in this study consists of four progressive parts: ECC database collection, data augmentation, ML modelling, and optimization process. The database collection was demonstrated in Section 2 with 429 and 392 original tensile stress and tensile strain samples, respectively. The ECC database augmentation is achieved by combining the collected original data mentioned and the unsupervised learning algorithm - Tuned CTGAN. The primary objective of this part is to create more data to establish a robust machine learning model. The third part involves the development of the prediction model, wherein HGS based ML models are trained for ECC properties' prediction using both the original and synthesized database. The last part focuses on the mixture design optimization of ECC. Four classes of constraints (min-max, volume, ratio, and fiber) were established to define the search space. Then, a cost function was defined and NSGA-III was introduced to tackle the problem. The above-mentioned framework is expected to be valuable in accurately predicting ECC properties and reversely optimizing mixture design, as exhibited in Figure 2.

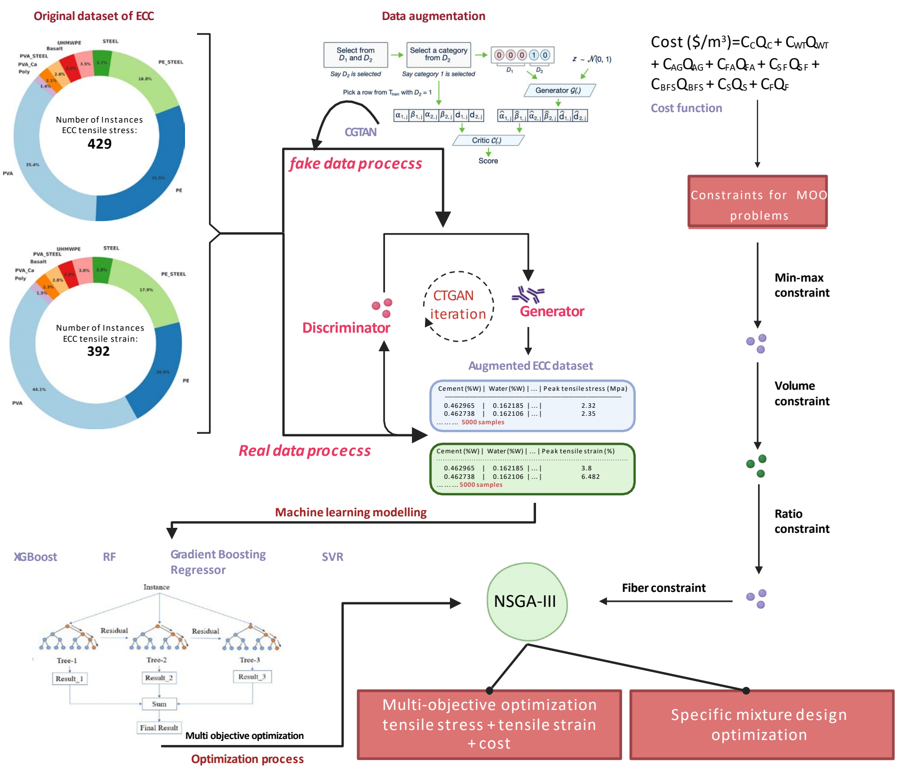


Figure 2. The flowchart of the ECC optimization framework comprising data collection, data augmentation, ML modelling, and optimization process

4.1 Data augmentation

4.1.1 Concepts of Tuned-CTGAN

Introduced by Xu, Skoularidou, et al. (2019) [93], Conditional Tabular GAN (CTGAN) is a GAN-based architecture tailored for synthesizing tabular data. It effectively tackles a major challenge in augmenting tabular data for continuous columns: handling non-Gaussian and multimodal distributions, an issue not adequately addressed by traditional GANs like TableGAN. While traditional GANs typically employ min-max normalization to scale continuous values to a range from -1 to 1, CTGAN adopts a modality-specific normalization technique. This approach transforms continuous values into a bounded vector by independently modelling each continuous column using a Variational Gaussian Mixture (VGM) model.

$$\mathbb{P}_{C_i}(c_{i,j}) = \sum_{k=1}^{m_i} \mu_k \mathcal{N}(c_{i,j}; \eta_k, \phi_k) \tag{1}$$

where C_i and m_i are the ith continuous column and the mode number in the VGM; in each mode, μ_k, η_k, and ϕ_k are the respective weight, mean, and standard deviation. For each value $c_{i,j}$ in C_i, the model calculates the probability of $c_{i,j}$ from each mode in the VGM as:

$$\rho_k = \mu_k \mathcal{N}(c_{i,j}; \eta_k, \phi_k), \qquad k = 1,2,\dots,m \tag{2}$$

It is assumed that the mode k with the highest probability is selected. This is followed by a normalization process which can be written as follows:

$$\alpha_{i,j} = \frac{c_{i,j} - \eta_k}{4\phi_k} \tag{3}$$

$$\beta_{i,j} = [0,\cdots,0,1,0,\cdots,0] \tag{4}$$

where $\alpha_{i,j}$ is a normalized scalar within the range [-1, 1], and $\beta_{i,j}$ is a one-hot encoding with the kth element set to 1 corresponding to the mode k. Thus, each value in a continuous column is represented by a combination of scalar $\alpha_{i,j}$ for the normalized value and a one-hot vector $\beta_{i,j}$ to indicate the mode. This normalized data is then used as the input for the model.

When augmenting data in discrete columns of a table, an imbalance present in these columns in the original dataset may result in a skewed distribution of discrete outputs, commonly known as mode collapse. To effectively counter this issue, two distinct strategies are implemented: the introduction of a conditional generator and the application of sampling training. Both are strategically designed to promote a balanced representation of all categories within the discrete columns. In the case of the conditional generator, it is specifically trained to synthesize data reflecting the conditions associated with a specific value k^* in a specific column D_{i^*}. To represent this condition, a masking vector m_i is defined in the following manner:

$$m_i^{(k)} = \begin{cases} 1 & if\ i = i^*\ and\ k = k^* \\ 0 & otherwise \end{cases} \tag{5}$$

Equation 6 represents the $cond$ vector which is introduced to concatenate different masking vectors for all discrete columns $D_1, \dots, D_{N_d}$.

$$cond = m_1 \oplus m_2 \oplus \cdots \oplus m_{N_d} \tag{6}$$

The $cond$ vector is designed to facilitate the conditioning of specific column values through one-hot encoding. Following this, the conditional generator G operates by taking random noise along with the $cond$ as its inputs. This setup compels G to adhere to the specified conditions, achieving this through the minimization of conditional loss, specifically cross-entropy.

4.1.2 Establishment of Tuned-CTGAN

In the Tuned-CTGAN framework, the architecture integrates a conditional generator G and a discriminator D. G is structured with two fully connected layers, each layer being augmented with batch normalization and $ReLU$ activation for enhanced performance. Following these layers, a mixed activation function is utilized to create synthetic row representations. Specifically, the continuous column values α_i are processed through a $Tanh$ activation, whereas the discrete column values d_i and mode indicators β_i undergo activation via $Gumbel\ Softmax$. The architecture also includes an $embedding\_dim$, a critical hyperparameter set at 32 to enrich the synthetic data generation with diversity. Furthermore, to optimize the training efficiency of Tuned-CTGAN, the batch size, number of epochs, and learning rate are carefully set to 50, 2000, and 2e-6, respectively.

Within the discriminator D of the Tuned-CTGAN model, a dual-layer fully connected structure is implemented. These layers are succeeded by a dropout layer, strategically designed to sporadically deactivate certain nodes, thereby addressing potential overfitting challenges. $LeakyReLU$ is selected as the activation function, enhancing the model's capability to differentiate between data types. This is followed by another fully connected layer tasked with scoring the current batch. The discriminator's layer sizes, denoted as $discriminator\_dim$, are set to (512, 512), where the higher dimensionality of 512 plays a pivotal role in accurately distinguishing between real and synthesized data. Additionally, the discriminator's learning rate is calibrated at 0.0005. Figure 3 visually depicts the Tuned-CTGAN's architecture and its training procedure. Notably, the $discriminator\_steps$ parameter is set to three, indicating that for every single update of the generator, the discriminator undergoes three update cycles. During data generation, the fiber parameters' boundaries are set based on the original data. These fiber parameters are generated according to the normal distribution. This approach ensures a precise correlation between fiber types and their traits, improving the synthetic data's reliability. The pseudocode is shown in Figure 4.

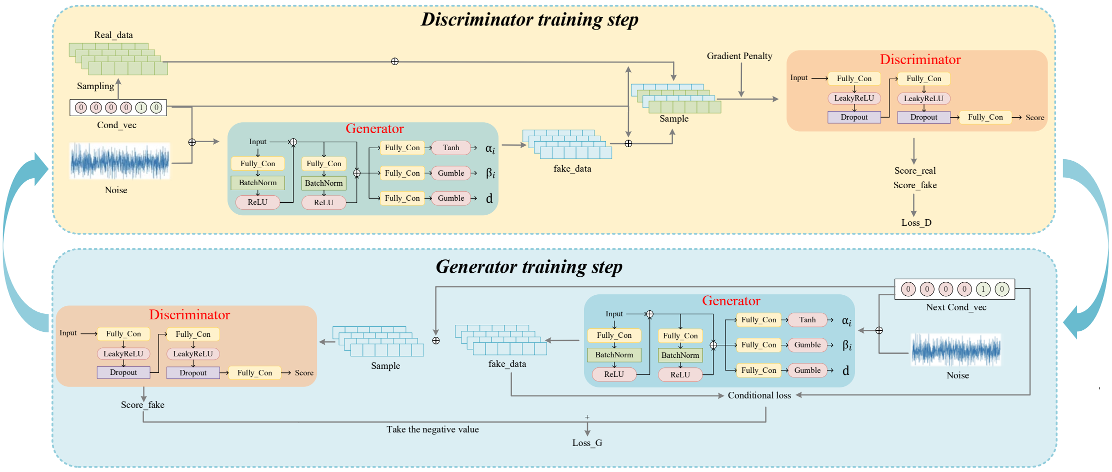


Figure 3. The training process and architecture of the Tuned-CTGAN model

```
Algorithm 1 Fiber Constraint for Synthetic Data Generation
Require: desired_strength, desired_modulus
Initialize dictionaries: desired_strength, desired_modulus
function IsValid(column_names, data)
    Extract fiber_type(ft), fiber_tensile_strength, fiber_elastic_modulus from data
    Initialize valid_rows as False array
    for ft in desired_strength/modulus.dictionaries() do
        Check if values are within range for strength and modulus
        Update valid_rows using OR operation
    end for
    return valid_rows
end function
function Transform(column_names, data)
    Extract fiber_type, fiber_tensile_strength, fiber_elastic_modulus from data
    for ft in desired_strength/modulus.dictionaries() do
        Calculate avg_strength and std_strength
        Generate new strengths using normal distribution
        Update fiber_tensile_strength for these rows
        Calculate avg_modulus and std_modulus
        Generate new modulus values using normal distribution
        Update fiber_elastic_modulus for these rows
    end for
    return data
end function
Create custom constraint class using IsValid, Transform
```

Figure 4. The pseudocode of fiber constraint

4.1.3 Evaluation metrics for CTGAN

Evaluating the quality of data generated by GANs poses a challenge due to the variability in outcomes produced by different evaluation metrics. This study proposes a comprehensive approach combining visual, statistical, and ML based metrics for assessing the quality of tabular data generation. The visual and statistical metrics are detailed below, while the ML based metrics are thoroughly elaborated in Section 4.2.

(1) Visual based metrics: This approach involves using visualization techniques to intuitively compare real and synthetic data. In this study, three methods are introduced for visual evaluation: Distribution plot, Cumulative Sum, and Correlation table. The Distribution plot is used to assess the statistical properties and similarities of real and synthetic datasets. Cumulative Sum offers a column-by-column visual comparison, focusing on the distribution similarities. Lastly, the Correlation table is employed to analyze the interrelationships between table columns, thereby gauging the generator's efficacy in accurately modelling these relationships.

(2) Statistical based metrics: In this research, two statistical tests, KSTest and CSTest are utilized for evaluation [94]. KSTest uses the empirical Cumulative Distribution Function (CDF) and the Kolmogorov–Smirnov test to assess continuous features' distributions, mainly focusing on the maximal divergence between the observed and expected CDF values. On the other hand, CSTest applies the Chi-squared test to evaluate discrete columns' distributions, with its p-value indicating the probability that values from two columns originate from the same distribution.

4.2 Machine learning model establishment

4.2.1 XGBOOST

The technique of XGBoost is built on the “boosting” idea, which combines the prediction of weak learners with additive training approaches to build a powerful learner. Boosting tree algorithms are based on the decision tree, which is known as the classification and regression tree (CART). The structure of XGBoost is depicted in Figure 5. For regression tasks, CART divides the dataset into two subsets at each level according to the boundary for one variable until reaching the tree's maximum depth set by users. It can be described as below:

$$R_1(j,s) = \{x \mid x^j \leq s\} \; and \; R_2(j,s) = \{x \mid x^j \geq s\} \tag{7}$$

Mean squared error of each leaf node is calculated:

$$MSE_{node} = \sum_{i \in node} (\hat{y}_{node} - y^{(i)})^2 \tag{8}$$

$$\hat{y}_{node} = \frac{1}{m_{node}} \sum_{i \in node} y^{(i)} \tag{9}$$

where m_{node} is the number of instances in one node. The cost function for regression of CART can be expressed as:

$$\mathrm{J}(\mathrm{k}, t_k) = \frac{m_{left}}{m} MSE_{left} + \frac{m_{right}}{m} MSE_{right} \tag{10}$$

The algorithm will search for the best solutions for boundaries of variables to minimize the cost function. The prediction is then the average target value of all instances in one subset. The mechanism of XGBoost is to keep adding and training new trees to fit residual errors of the last iteration as shown in Figure 5. A predicted value is assigned to each instance by adding all corresponding leaves' scores together:

$$\hat{y} = \phi(x_i) = \sum_{k=i}^{K} f_k(x_i) \text{ and } f_k(x_i) = w_{q(x)}, f_k \in \mathcal{L} \tag{11}$$

where K is the quantity of trees; $f_k(x_i)$ means the outcome of input x_i for the kth tree; $w_{q(x)}$ is the score for each leaf node; $q(x)$ denotes the number of leaf nodes; $\mathcal{L}$ represents an assemble of all corresponding functions f_k.

The objective function of XGBoost contains two parts: the training error and the regularization, written as:

$$\mathrm{obj}(\theta) = \sum \mathrm{L}(\theta) + \sum \Omega(\theta) \tag{12}$$

where L is the loss function measuring the deviation of the predicted values from the actual values. Ω is the regularization function measuring the complexity of the training model in order to avoid overfitting.

$$\Omega(\theta) = \gamma \mathrm{T} + \frac{1}{2}\lambda \parallel \omega \parallel^2 \tag{13}$$

where T represents the total number of leaf nodes and ω is the score of each leaf node. γ and λ are controlling factors employed to avoid overfitting.

When a new tree is created to fit residual errors of last iteration, the predicted score for the tth tree can be expressed as:

$$\widehat{y_i}^{(t)} = \widehat{y_i}^{(t-1)} + f_t(x_i) \tag{14}$$

The objective function is thus written as:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \widehat{y_i}^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \tag{15}$$

An appropriate function, f_t, is replaced with the second-order Taylor polynomial of $f_t = 0$. Accordingly, the objective function can be approximated as:

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n} \left[l(y_i, \widehat{y_i}^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \tag{16}$$

where g_i is the first-order derivative and h_i denotes the second-order derivative:

$$g_i = \partial_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right) \; and \; h_i = \partial^2_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right) \tag{17}$$

Since previous $(t-1)$ trees' residual errors (y) have minimal influence on the modification of the objective function, Equation 15 is then simplified as:

$$\widetilde{\mathcal{L}^{(t)}} = \sum_{i=1}^{n} \left[\mathrm{g}_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \tag{18}$$

As each instance will finally be classified into one leaf node, all instances that belong to the same leaf node can be reassembled as:

$$\mathrm{obj}^{(t)} \approx \sum_{j=1}^{T} [(\sum_{i \in I_j} g_i) w_j + \frac{1}{2} (\sum_{i \in I_j} h_i + \lambda) w_j^2] + \gamma \tag{19}$$

Therefore, the optimum w and objective function obj are derived as:

$$w_j^* = -\frac{G_j}{\mathrm{H}_j + \lambda} \text{ and } obj = -\frac{1}{2} \sum_{j=1}^{t} \frac{G_j^2}{H_j + \lambda} + \gamma \mathrm{T} \tag{20}$$

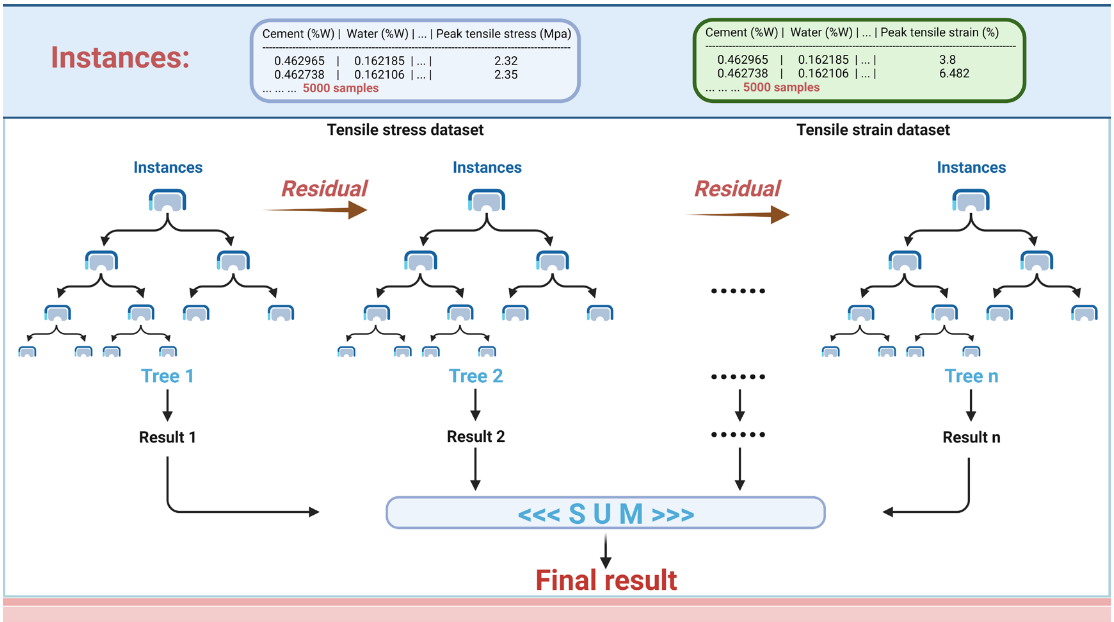


Figure 5. The structure of XGBoost

4.2.2 Other ML algorithms

SVR is normally used to solve regression problems by constructing a model that attempts to fit the error within a certain threshold. Its main strategy involves finding a hyperplane in a high-dimensional space that best fits the data, with the flexibility to manage both linear and non-linear models through kernel tricks [95]. Gradient Boosting Regressor is another ML algorithm that builds sequential models, typically decision trees, where each tree corrects errors made by the previous ones [96]. It focuses on minimizing a loss function, iteratively adding trees that predict the residuals or errors of prior trees to improve accuracy. This method excels in predictive accuracy even with non-linear data and efficiently handles missing values. Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the majority vote or average prediction of the trees for classification and regression tasks, respectively [97]. It effectively handles large datasets and high-dimensional spaces, offering robust performance and mitigating overfitting through its ensemble approach, making it highly versatile across diverse applications. To compare the HGS-XGBoost performance with other ML models, SVR, Gradient Boosting Regressor, and RF were also introduced in this study.

4.2.3 Hunger games search

Hunger Games search (HGS) is an advanced meta-heuristic algorithm that leverages strategies to discover diverse solution spaces to refine solutions towards optimality. It was first introduced by Yang, Chen et al. [12]. Each group of hyperparameters represents each individual and the fitness function is the loss function (RMSE) of the XGBoost. The objective is to identify the best individual that minimizes the RMSE when employing this group of hyperparameters for model training. "Hunger-Driven Motivational State Competition" is the searching strategy of HGS. The key advantage of HGS lies in its ability to efficiently navigate the search space and converge rapidly to the optimal solution, thereby reducing computational time and resource usage significantly. Considering the sources of food and the motions of species, a logical game can be set up between various animals with the aim of winning the game and acquiring the food. Equation 21 illustrates the foraging behaviours of individuals based on self-dependent spirit ($Game_1$) or collaborative communication ($Game_2$ and $Game_3$).

$$\overrightarrow{X(t+1)} = \begin{cases} Game_1{:}\overrightarrow{X(t)} \cdot (1 + randn(1)), r_1 < l \\ Game_2{:}\overrightarrow{W_1} \cdot \overrightarrow{X_b} + \vec{R} \cdot \overrightarrow{W_2} \cdot \left|\overrightarrow{X_b} - \overrightarrow{X(t)}\right|, r_1 > l, r_2 > E \\ Game_3{:}\overrightarrow{W_1} \cdot \overrightarrow{X_b} - \vec{R} \cdot \overrightarrow{W_2} \cdot \left|\overrightarrow{X_b} - \overrightarrow{X(t)}\right|, r_1 > l, r_2 < E \end{cases} \quad (21)$$

$$E = \operatorname{sech}(|F(i) - BF|), i = 1, 2, 3, \ldots, n \quad (22)$$

$$\vec{R} = 2 \times shrink \times rand - shrink \quad (23)$$

$$shrink = 2 \times (1 - \frac{t}{T}) \quad (24)$$

where $\vec{R}$ is in the range of [-a, a]; r_1 and r_2 are random numbers between 0 to 1; $randn(1)$ is a random number following a standard normal distribution; l is set as 0.03 in this study; t denotes the current iteration and T is the maximum iterations; $\overrightarrow{W_1}$ and $\overrightarrow{W_2}$ are two hunger weights; $\overrightarrow{X_b}$ and $\overrightarrow{X(t)}$ represent the locations of the best and other individuals, respectively; sech is a hyperbolic function; $F(i)$ denotes the fitness value of each individual and BF is the current best fitness value.

Regarding the hunger role, $\overrightarrow{W_1}$ and $\overrightarrow{W_2}$ can be calculated as follows:

$$\overrightarrow{W_1(i)} = \begin{cases} hungry(i) \cdot \frac{N}{sum(hungry)} \times r_4, r_3 < l \\ 1, r_3 > l \end{cases} \quad (25)$$

$$\overrightarrow{W_2(i)} = (1 - \exp(-|hungry(i) - sum(hungry)|)) \times r_5 \times 2 \quad (26)$$

$$hungry(i) = \begin{cases} 0, F(i) == BF \\ hungry(i) + H, F(i) != BF \end{cases} \quad (27)$$

where $hungry(i)$ denotes the hunger degree of each individual; N is the total number of individuals; r_3, r_4, and r_5 are three random numbers from 0 to 1; $F(i)$ is the fitness value of each individual. Equation 27 means that the hunger degree is 0 for the best individual and a new hunger H will be added for other individuals.

$$TH = \frac{F(i) - BF}{WF - BF} \times r_6 \times 2 \times (UB - LB) \quad (28)$$

$$H = \begin{cases} LH \times (1 + r), TH < LH \\ TH, TH \geq LH \end{cases} \tag{29}$$

where WF is the worst fitness value in the current iteration; UB and LB are the upper and lower bounds of the feature space; LH is determined as the lower bound of the H (100 in this study). The pseudocode of HGS is represented in Figure 6.

Algorithm 1 Pseudo-code of Hunger Games Search (HGS)

```
Initialize parameters N, T, l, D, sum(hungry)
Initialize positions of Individuals X_i for i = 1, 2, ..., N
while t ≤ T do
    Calculate fitness of all Individuals
    Update BF, WF, X_b, BF
    Calculate hungry by Eq. (27)
    Calculate W_1 by Eq. (25)
    Calculate W_2 by Eq. (26)
    for each Individual do
        Calculate E by Eq. (22)
        Update R by Eq. (23)
        Update positions by Eq. (21)
    end for
    t = t + 1
end while
return BF, X_b
```

Figure 6. The pseudocode of HGS

4.2.4 Model establishment

The ML models' hyperparameters are tuned by HGS in this study. Since the final utilized model is XGBoost, the model establishment process of XGBoost is described instead of the other ML models. When training XGBoost, some hyperparameters are crucial to the model predictive performance comprising *booster*, *objective*, *max_depth*, *eta*, *min_child_weight*, *subsample*, *colsample_bytree*, *learning_rate*, *num_parallel_tree*, *and n_estimators*. In this study, 'regression: linear' was determined for the hyperparameter objective because it was a typical regression task. After the trial experiment, the hyperparameters of *booster*, *eta*, *min_child_weight*, *subsample*, *colsample_bytree*, *num_parallel_tree* were set as 'gbtree', '0.1', '5', '0.7', '1', '1', respectively. Besides, *max_depth*, *learning_rate*, and *n_estimators* were automatically optimized using HGS algorithm with the initial boundaries of [1, 100], [0.01, 1], and [1, 100], respectively. This is because these three hyperparameters are most important which significantly affect the model complexity, convergence efficiency, and overfitting/underfitting balance.

Once determining the hyperparameters, the XGBoost model was generated with the optimal values of *max_depth*, *learning_rate*, and *n_estimators* after 20 iterations on the training set (80% of the database). The fiber type, which is the eighth feature of the dataset, is transferred to one binary attribute using the One-hot-encoding function. Thus, the 18 features of the initial dataset were increased to 26 features. Subsequently, the optimal XGBoost model was tested on the test set (20% of the database) to ascertain its predictive efficacy. The XGBoost framework calculates the importance weight of each feature. If a certain feature is deemed to have minimal predictive significance, it is excluded in the subsequent training cycle.

4.2.5 Performance evaluation

In this study, three accompanying evaluating indicators aim to evaluate the precision of the ML model: correlation coefficient (R), mean absolute error (MAE), and root mean square error (RMSE). These indicators are calculated as follows [98]:

$$R = \frac{\sum_{i=1}^{n}(y_i^* - \overline{y^*})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(y_i^* - \overline{y^*})^2}\sqrt{\sum_{i=1}^{n}(y_i^* - \bar{y})^2}} \tag{30}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i^* - y_i| \tag{31}$$

$$RSME = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i^* - y_i)^2} \tag{32}$$

where n is the n groups of data samples; y_i^* and y_i are the predicted and actual results; $\overline{y^*}$ and $\bar{y}$ illustrate the mean values of the predicted and actual results.

4.3 Multi-objective optimization

4.3.1 Objective function establishment

After the establishment of two XGBoost models, they are adopted as the objective functions for ECC's tensile stress and tensile strain, respectively. The third objective function (cost) is computed by polynomials as follows:

$$Cost(\$/m^3) = C_C Q_C + C_{WT} Q_{WT} + C_{AG} Q_{AG} + C_{FA} Q_{FA} + C_{SF} Q_{SF} + C_{BFS} Q_{BFS} + C_S Q_S + C_F Q_F \quad (33)$$

In Equation 33, Q_C, Q_{WT}, Q_{AG}, Q_{FA}, Q_{SF}, Q_{ca}, Q_{BFS}, Q_S and Q_F denote the unit weight (kg/m^3) of cement, water, fine aggregate, fly ash, silica fume, blast furnace slag, superplasticiser, and fiber, respectively. Besides, C means the unit price ($/kg) of each raw material of ECC, which is checked from Alibaba in China and summarised in Table 2. A simplified assumption was employed, utilizing the average cost of fiber, despite the recognition that fiber costs are influenced by parameters such as diameter, tensile strength, and elastic modulus, etc [99]. The fiber type, which is the eighth feature of the dataset, is transferred to one binary attribute using One-hot-encoding function. Therefore, the unit weight and cost of the specific fiber will be determined upon the fiber type selected.

Table 2. The unit weight and cost of each variable of ECC

Raw materials	Notation	Density (kg/m^3)	Notation	Unit price ($/kg)
Cement	U_C	3100	C_C	0.59
Water	U_{WT}	1000	C_{WT}	0.0005

Fine Aggregate	U_{AG}	2600	C_{AG}	0.009
Fly ash	U_{FA}	2300	C_{FA}	0.027
Silica Fume	U_{FA}	2220	C_{SF}	0.82
blast furnace slag	U_{BFS}	2900	C_{BFS}	0.062
Superplasticiser	U_S	1100	C_S	1.07
Fiber type:				
Basalt		2670		3.42
PE		970		15.04
PE_Steel		4410		7.86
PVA		1300		3.2
PVA_Ca	U_F	1400	C_F	3.415
PVA_Steel		4575		1.94
PP		910		3.9
Steel		7850		0.68
UHMWPE		950		31.44

4.3.2 Constraints

In this study, the following four classes of constraints (Min-max, volume, ratio, and fiber) were set up to define the search space in the MOO problem.

- Min-max constraint

Constraints were introduced to specify the range within which each input variable can vary to avoid a covariate shift. The ranges of input features were limited to the minimum and maximum values according to the datasets, as follows:

$$d_{imin} \leq d_i \leq d_{imax} \tag{34}$$

where d_{imin} and d_{imax} represent the lowest and highest value of the *i*th feature (Appendix 1 and 2).

- Volume constraint

In the ECC matrix, the cumulative volume of each component, coupled with inevitable air content for each blend, approximately equates to one cubic meter. Therefore, the design domain should be restricted to exclude mixes deviating significantly from a volume of one cubic meter. This was realized through a constraint that governs the volume of each mix design, termed the "volume constraint", as follows:

$$V_m = \left(\frac{Q_C}{U_C} + \frac{Q_{WT}}{U_{WT}} + \frac{Q_{AG}}{U_{AG}} + \frac{Q_{FA}}{U_{FA}} + \frac{Q_{SF}}{U_{SF}} + \frac{Q_{BFS}}{U_{BFS}} + \frac{Q_S}{U_S} + \frac{Q_F}{U_F}\right) \in [\alpha_{min}, \alpha_{max}] \quad (35)$$

where U_C, U_{WT}, U_{AG}, U_{FA}, U_{SF}, U_{BFS}, U_S and U_F are the density of cement, water, fine aggregate, fly ash, silica fume, blast furnace slag, superplasticiser, and fiber, respectively, specified in Table 2; α_{min} and α_{max} were set to 0.96 and 1.04, considering the variations in air content and disparities in ingredient densities across different studies.

- Ratio constraint

To ensure consistency within the dataset and prevent undue variations in properties, two ratio constraints were established. These constraints are crucial to mitigate potential complications arising from extreme water-to-binder (W/B) ratios that lie outside the raw database's scope, potentially affecting the concrete's fresh state. Consequently, the water to binder ratio and the fine aggregate to binder ratio were delineated as follows.

$$0.131 \le \frac{Water}{Cement + Fly\ ash + Silica\ fume + Blast\ furnace\ slag} \le 0.568 \quad (36)$$

$$0 \le \frac{Fine\ aggregate}{Cement + Fly\ ash + Silica\ fume + Blast\ furnace\ slag} \le 1.667 \quad (37)$$

- Fiber constraint

The type, diameter, length, tensile strength, and elastic modulus of fiber are crucial to the properties of ECC, simultaneously affecting the cost owing to totally different characteristics. They cannot be considered as unrelated features during the MOO solving process, otherwise the characteristics fail to match the specific fiber. Therefore, this study introduced the boundary constraints corresponding to each fiber category, as shown in Table 3. In this case, the diameter, length, tensile strength, and elastic modulus were constrained according to the one binary attribute of the feature 'fiber type' using the One-hot-encoding function.

Table 3. The fiber constraints of ECC

Fiber type	F. Diameter (Micro-Meter)		Fiber Length (mm)		Fiber Tensile Strength (MPa)		Fiber Elastic Modulus (GPa)	
	Min	Max	Min	Max	Min	Max	Min	Max
Basalt	13	13	6	12	2600	2600	85	85
PE	12	78	6	38	2400	5200	66	158
PE_STEEL	54	279	8.9	34.5	1772	6360	115.8	379
PVA	35	40	8	12	900	1620	23	43
PVA_Ca	7.667	28.25	2.061	14.6	1929.6	4010.4	137.04	472.11
PVA_STEEL	60	192.7	7.68	16.3	1633.3	2608	82.1	155.12
PP	12	19.5	6	12	350	3700	4	6
STEEL	150	550	6	30	2428	5000	200	360
UHMWPE	18	25	12	13	3000	3000	100	114

4.3.3 Construction of optimizer

Regarding multi-objective optimization problems, an optimizer can seek within a set of compromise options in the objective space called the Pareto front to obtain the optimal solution [100]. The Pareto front implies the condition that one objective fails to be improved without worsening the other objectives. If A represents the group of feasible solutions and $\boldsymbol{x}^* \in A$ is one of the Pareto optimal solutions, there is no existence of $x \in A$ satisfy that:

$$f_k(\boldsymbol{x}) \leq f_k(\boldsymbol{x}^*)\ for\ k = 1,2,3,\ldots,t\ and \tag{38}$$

$$f_k(\boldsymbol{x}) < f_k(\boldsymbol{x}^*)\ for\ at\ least\ one\ k \tag{39}$$

Deb and Jain [17] introduced NSGA-III by replacing the crowding distance to a reference point-based selection approach in NSGA II, significantly improving the convergence speed and population diversity. Therefore, this study proposed NSGA-III to acquire the above-mentioned Pareto front in a larger objective space. The procedure of the establishment of NSGA-III started by randomly generating an initial population (P_t) of size N. The iteration process is described as follows:

1. At tth iteration, an offspring population, denoted as Q_t, is derived from the parent population P_t. This derivation employs techniques comprising simulated binary crossover (SBX), random selection (RS), and polynomial mutation (PM), as described in [17]. The population size for both P_t and Q_t remains consistent at N.

2. The populations P_t and Q_t are amalgamated to form a new population, denoted as $M_t = P_t \cup Q_t$, which possesses a size of $2N$. Subsequently, the top N individuals are chosen for the succeeding generation through a non-dominated sorting mechanism. This process categorizes the elements of M_t into distinct non-domination levels, represented as L_z.

3. A new population G_t is constructed by sequentially incorporating elements from M_t across L_z levels until the population size either matches or surpasses N for the first time, where the last level is called level z. The individuals from level z to the end are excluded from consideration.

4. If the element size of G_t precisely matches N, all the individuals in the last level z will be chosen for P_{t+1}, followed by the next iteration. If the size of G_t exceeds N, the elements in the last level z will be sorted and selected according to the reference point-based selection approach. By this approach, the size of P_{t+1} maintains N and the population diversity is ascertained.

Figure 7 depicts the environment selection of NSGA-III, showing non-dominated sorting and reference point-based selection in each iteration.

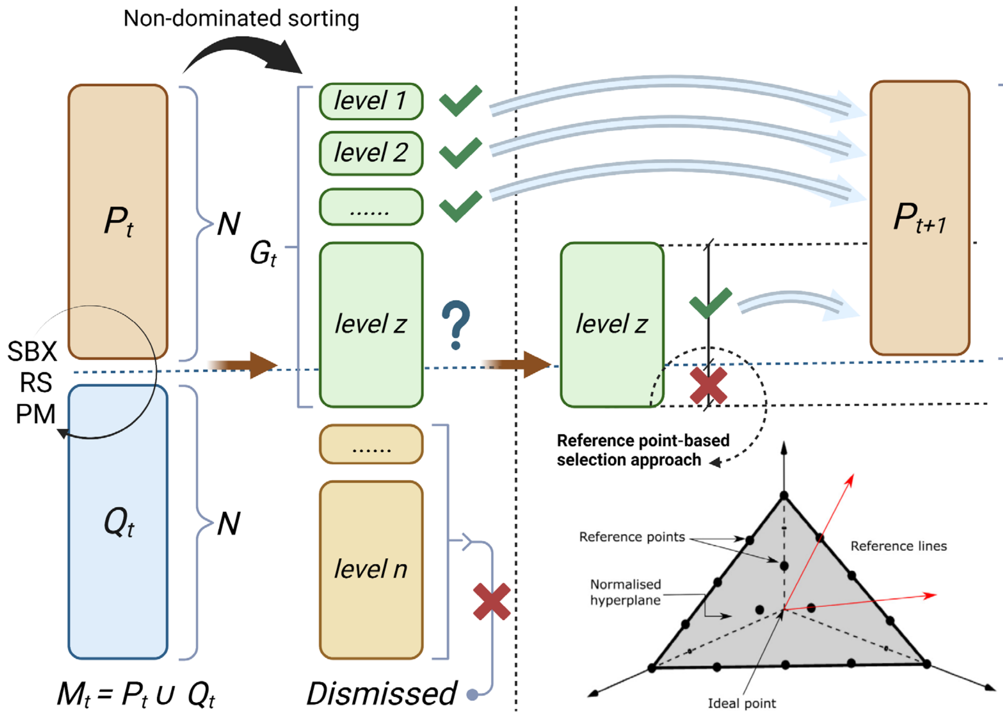


Figure 7. The environment selection of NSGA-III

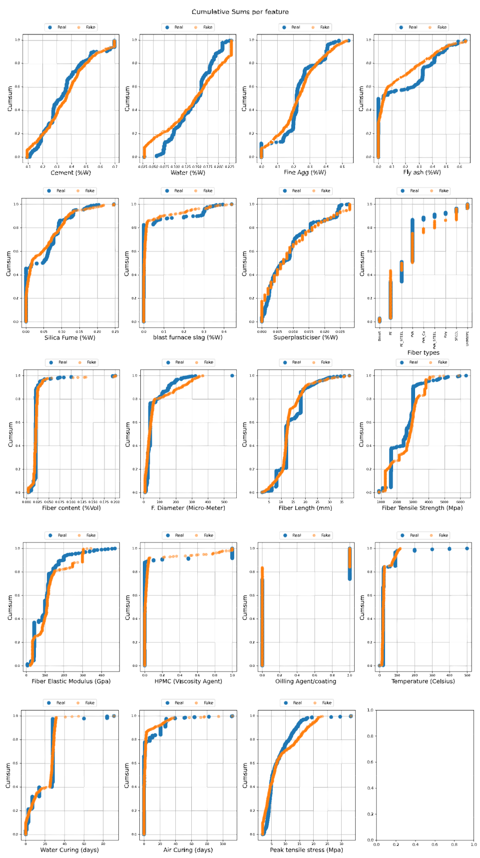
5. Results and discussion

5.1 Data augmentation

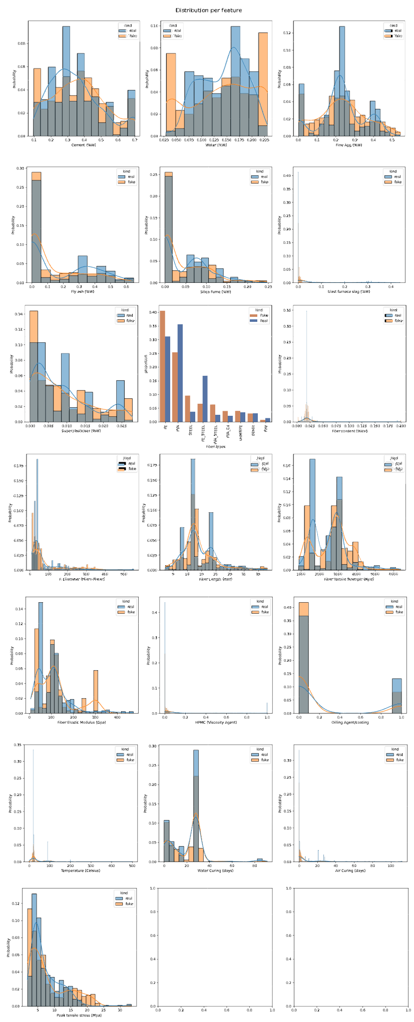
The Tuned-CTGAN augmentation technique was used to increase the size of the original tensile stress and strain datasets from 429 and 392 data records to 5000 each. The performance of the Tuned-CTGAN augmentation on the tensile stress and strain databases is depicted in Figures 8 and 9, respectively, through three visualization plots: the cumulative sum plot, distribution plot, and correlation table.

The cumulative plot for various features reveals a notable similarity between the original and synthesized datasets. For instance, the cumulative plot for the 'cement' feature exemplifies this similarity, as depicted in Figures 8a and 9a. While the curves representing the original (in blue) and the generated data (in red) are not identical, their close proximity reflects the performance in capturing the statistical characteristics of the original dataset. This phenomenon can be observed in the cumulative plots of the other variables. Meanwhile, the distribution plots of the variables also demonstrate a similarity in the probabilistic distribution of values between the original and synthesized datasets. This means that values frequently observed in the original dataset are generated with higher probability in the synthesized data, which is exhibited in Figures 8b and 9b. It indicates that the data generation process effectively captures and reproduces the underlying statistical tendencies of the original data. Besides, the distribution of the fiber type (the single discrete variable) is well simulated by the Tuned-CTGAN model. For instance, PE and PVA are the two most common categories in the original database, and this is also reflected in the generated fake dataset. However, the distribution of the fiber types is not strictly identical to the original database, mainly due to the characteristics of Tuned-CTGAN. As mentioned earlier, a conditional generator and sampling training applied in Tuned-CTGAN are used to tackle the problem of mode collapse. This technology simultaneously partly neglects the actual distribution of discrete features, increasing the sampling randomness and ultimately leading to no identical distribution. Additionally, the distribution plot and cumulative sum of the other fiber type related variables comprising fiber diameter, length, tensile strength, and elastic modulus are mainly determined by the fiber type. This leads to the same phenomenon of no identical distribution as that of fiber type.

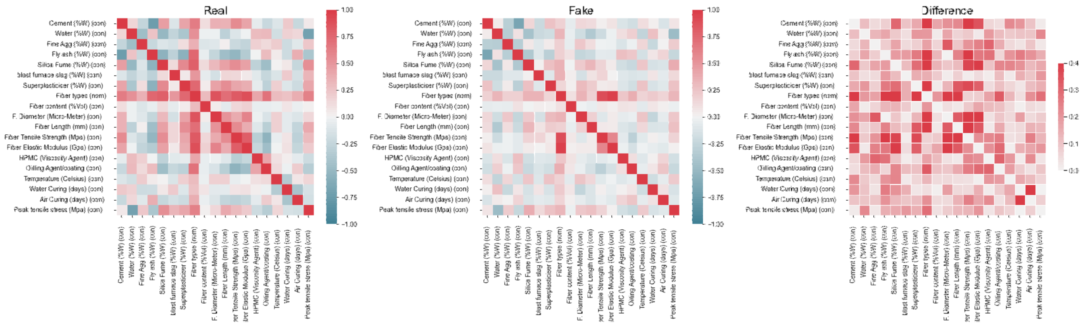
The correlation table evaluates the association between each column of the table as shown in Figures 8c and 9c. Apart from the relatively strong interdependency between fiber types and their associated parameters, dependencies among other parameters are comparatively weaker (below 0.25). This observation aligns with expectations, as the characteristics of fibers inherently determine their parameters. However, this strong interdependency may result in multicollinearity, reducing the predictive performance and influencing design accuracy. To reduce the influence caused by multicollinearity, the generator's output is specifically directed by imposing constraints based on the selected fiber type in the training of Tuned-CTGAN. Meanwhile, these constraints are also imposed during the NSGA-III optimization process as mentioned in the previous section. In addition, the correlation tables between the original database and synthetic database depict a similar pattern with a relatively minor difference (the maximum of 0.3) for both the tensile stress and strain database. Therefore, the generator’s performance in accurately modelling the relationships between columns has been assessed according to the visualization process.



(a)

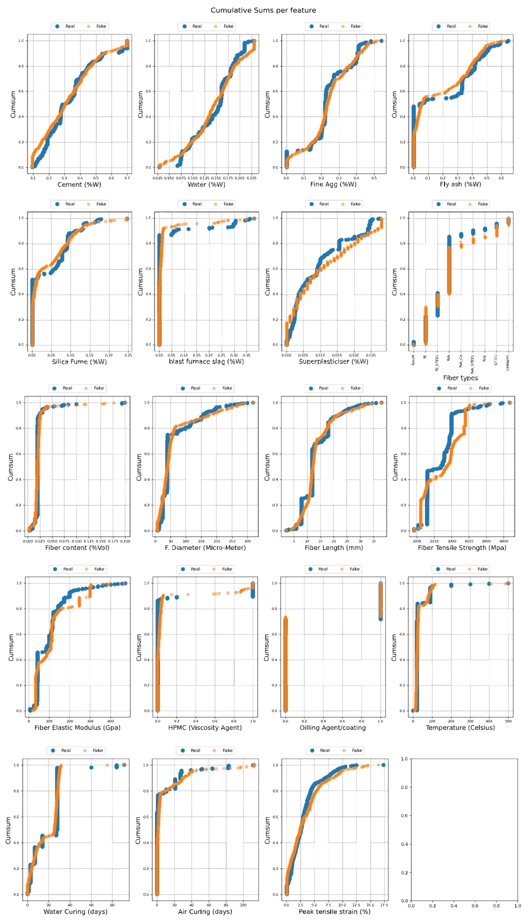


(b)

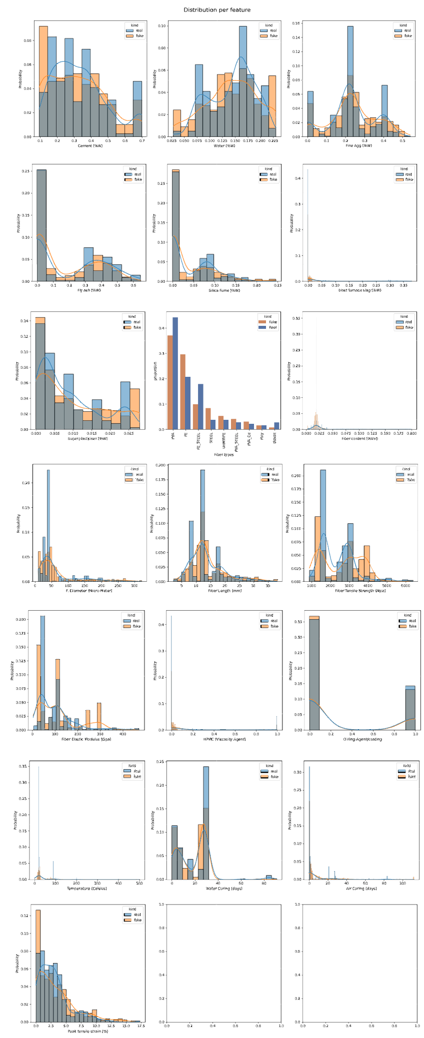


(c)

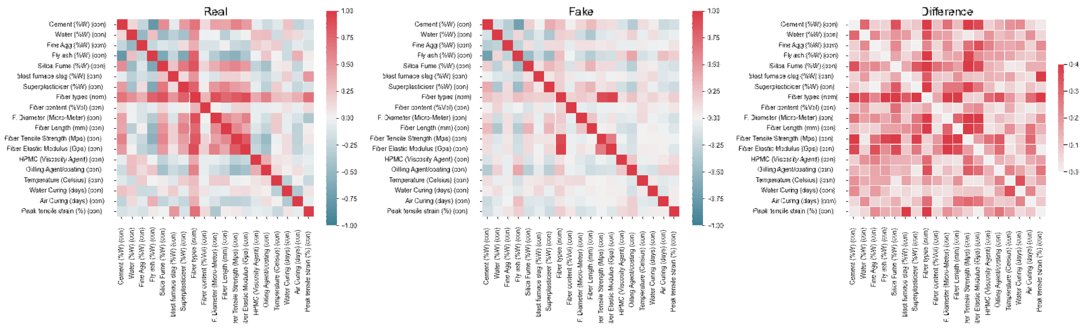
Figure 8. The cumulative (a), distribution (b), and correlation plots (c) for eighteen features before and after Tuned-CTGAN augmentation based on the dataset of tensile stress



(a)



(b)



(c)

Figure 9. The cumulative (a), distribution (b), and correlation plots (c) for eighteen features before and after Tuned-CTGAN augmentation based on the dataset of tensile strain

The statistical based metrics (e.g., KSTest and CSTest) of the three data augmentation methods for both tensile stress and strain datasets are depicted in Table 4. The KSTest examines the distributions of continuous features, while the CSTest is used to evaluate distributions of discrete columns with the values depicted in Figure 10. For the KSTest, the results are expressed as $1 - calculated\ value$, so that a result closer to 1 indicates a slighter discrepancy between the real and generated distributions. In Table 4, the KSTest values of most variables in both stress and strain datasets are over 0.7, indicating a relatively reliable outcome using Tuned-CTGAN data augmentation. The average metric values of 18 continuous variables and one output are 0.7795 and 0.7764 on stress and strain datasets, respectively. These results explained the reliable data augmentation performance of Tuned-CTGAN, further verifying the conclusions acquired from the visualization plots.

Table 4. The statistical based metrics based on Tuned-CTGAN

Variables and outcome	Tensile stress dataset	Tensile strain dataset
Cement (C)	0.8892	0.8680
Water (C)	0.8122	0.8955
Fine Agg (C)	0.8687	0.8884
Fly ash (C)	0.8030	0.7800
Silica Fume (C)	0.7923	0.7836
blast furnace slag (C)	0.6250	0.5918
Superplasticiser (C)	0.8825	0.8664

Fiber type (D)	0.8773	0.9243
Fiber content (C)	0.6721	0.6677
F. Diameter (C)	0.7962	0.7530
F. Length (C)	0.7667	0.8047
F. Tensile Strength (C)	0.7722	0.6685
F. Elastic Modulus (C)	0.8299	0.6913
HPMC (C)	0.5763	0.5752
Oiling Agent (C)	0.9085	0.9687
Temperature (C)	0.6549	0.6752
Water Curing (C)	0.7169	0.7428
Air Curing (C)	0.7159	0.7276
Peak tensile stress (C)	0.8510	-
Peak tensile strain (C)	-	0.8785
Average value	0.7795	0.7764

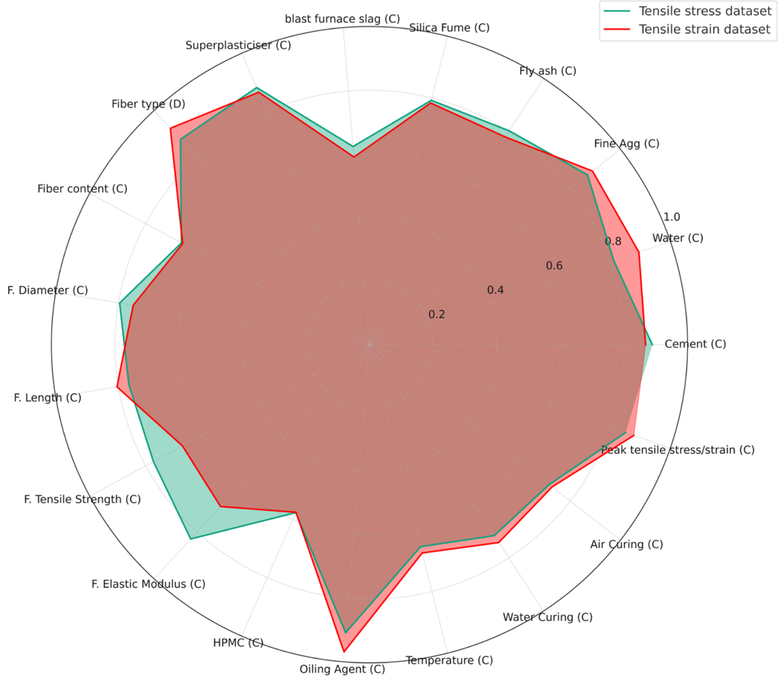


Figure 10. The radar maps of Tuned-CTGAN on tensile stress and strain datasets

5.2 Model prediction

Figures 11 and 12 illustrate the scatter plots of four ML models before and after data augmentation with the x-axis denoting the actual values and the y-axis reflecting the predicted values. The diagonal line displays a comparison between the actual and predicted tensile stress or strain of ECC. The closeness of the data points to the diagonal line is indicative of the accuracy of the predictions. A data point located near the diagonal line indicates a minor difference between the real and predicted values, demonstrating the accuracy of the forecasting model. Besides, the corresponding evaluation indexes of four ML models before and after data augmentation are summarized in Tables 5 and 6, respectively.

Before the data augmentation, the XGBoost model indicates the highest predictive accuracy in the stress dataset with the respective MAE, R^2, and MSE values of 1.189, 0.874, and 1.425, as shown in Table 5. Its performance in accurately predicting the ECC stress can be verified in Figure 11a showing few outliers. The Gradient Boosting Regressor and RF also demonstrate high predictive accuracy, achieving R^2 values of 0.847 and 0.846, respectively. In the strain dataset, RF model possesses the highest predictive accuracy (R^2=0.805), followed by the XGBoost (R^2=0.772) and Gradient Boosting Regressor (R^2=0.757), as shown in Table 6. Thus, the tensile strain prediction is observed to be more complex and challenging compared to the prediction of tensile stress. The SVR models exhibit poor generalization ability in predicting both stress and strain due to the low values of R^2.

The obtained results are derived from the optimized ML models using the HGS algorithm. After hyperparameters' tune, max_depth, learning_rate, and n_estimators of XGBoost with the initial boundaries of [1, 100], [0.01, 1], and [1, 100], respectively are determined as 44, 0.856, and 47 (stress dataset) and 39, 0.793, and 25 (strain dataset). Besides, n_estimators and max_depth of Gradient Boosting Regressor in training stress and strain are 78, 7 and 70, 7, respectively. Regarding the RF model, the two parameters are set as 76, 11 and 94, 13, respectively. Besides, the kernel function used in the SVR model is the Radial Basis Function.

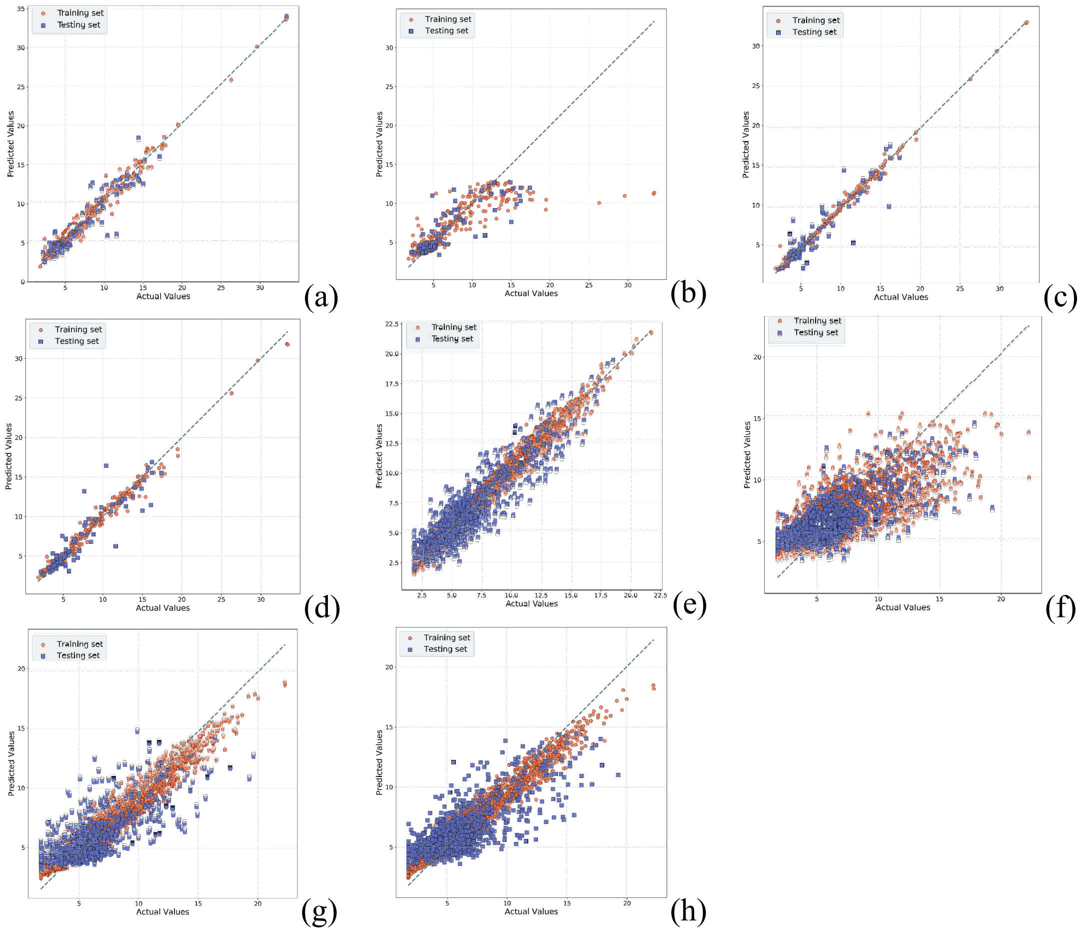


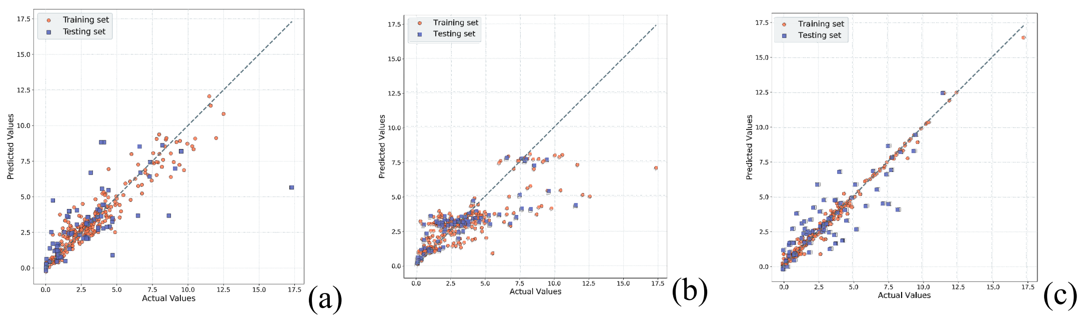
Figure 11. Training and testing accuracy of established stress prediction models before augmentation: (a) XGBoost (b) SVR (c) Gradient Boosting Regressor (d) RF and after augmentation: (e) XGBoost (f) SVR (g) Gradient Boosting Regressor (h) RF

Table 5. The evaluation matric of varying ML models before and after tensile stress data augmentation

	XGBoost	SVR	Gradient Boosting Regressor	RF
Before data augmentation				
RMSE (trainset/testset)	1.667/1.425	2.871/2.216	0.352/1.534	0.593/1.539
MAE	0.998/1.189	1.374/1.442	0.180/0.990	0.389/0.951
R^2	0.881/0.874	0.627/0.680	0.994/0.847	0.984/0.846
After data augmentation				
RMSE	0.587/0.829	1.924/2.086	0.926/2.077	0.794/1.990
MAE	0.545/0.737	1.378/1.558	0.742/1.579	0.620/1.527
R^2	0.966/0.925	0.566/0.499	0.895/0.756	0.926/0.544

After the data augmentation, the tensile stress and strain data points are increased to 5,000. In Figures 11a and 12a, the XGBoost model exhibited significantly improved performance, especially in the strain dataset. The values of R^2 in the stress and strain test sets are increased from 0.874 to 0.925 and from 0.772 to 0.889, respectively. The other two evaluation indexes (MAE and MSE), are correspondingly reduced which further verifies the efficacy of data augmentation. However, the evaluation indexes of other ML models are worse after the data augmentation, with SVR and RF being especially impacted. The possible reason is their poor capacity to resist overfitting when handling complex datasets characterized by high feature dimensions and large volumes of data. In contrast, XGBoost applies its regularization and optimization techniques to overcome the overfitting problems.

It is acknowledged that the reliability of the proposed design hinges on the precision of the established ML models. Inaccuracies in the predictive capability can lead to considerable errors. The high values of R^2 of XGBoost models in this study (0.925 and 0.889) on both tensile stress and strain datasets demonstrate the relatively reliable predictive ability. In conclusion, XGBoost is finally established based on CTGAN data augmentation and HGS hyperparameters' optimization, which is used as the objective function in multi-objective optimization tasks.



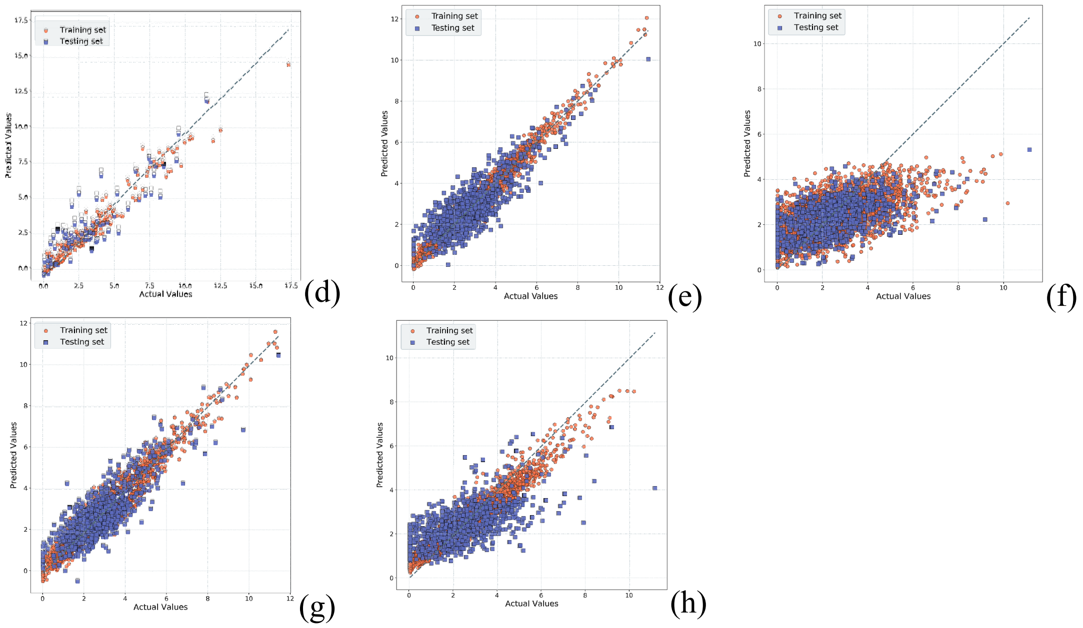


Figure 12. Training and testing accuracy of established strain prediction models before augmentation: (a) XGBoost (b) SVR (c) Gradient Boosting Regressor (d) RF and after augmentation: (e) XGBoost (f) SVR (g) Gradient Boosting Regressor (h) RF

Table 6. The evaluation matric of varying ML models before and after tensile strain data augmentation

	XGBoost	SVR	Gradient Boosting Regressor	RF
Before data augmentation				
RMSE (trainset/testset)	1.612/1.414	1.513/1.628	0.286/1.296	0.528/1.160
MAE	0.912/0.845	0.884/1.065	0.169/0.912	0.338/0.888
R^2	0.833/0.772	0.660/0.616	0.912/0.757	0.958/0.805
After data augmentation				
RMSE	0.581/0.631	1.183/1.302	0.595/1.012	0.493/1.212
MAE	0.640/0.546	0.877/1.006	0.482/0.920	0.398/0.958
R^2	0.936/0.889	0.411/0.292	0.850/0.726	0.897/0.386

5.3 Multi-objective optimization

When implementing NSGA-III for multi-objective optimization, the water and air curing ages are respectively set as 28 and 0 days to guarantee the same curing method and time. Besides, although the fiber content in the original database ranged from 0.003 to 0.3, the actual dosage is usually below 0.03 [92]. When fiber content is low, the bridging stress is inadequate for supporting widespread crack propagation due to the scarcity of fibers within the mortar matrix. When the tensile stress across a given interface length falls beneath the matrix's cracking strength, this results in a reduction of interface number that is unfavorable for crack spread [40]. In contrast, exceeding a fiber content of 0.03 compromises the matrix's workability, causing issues like fiber clustering and a greater occurrence of sizable voids. These problems significantly detract from the composite's tensile strength [91]. Therefore, the upper limit of fiber content in this study is set as 0.03.

Figure 13 depicts the multi-objective optimization (maximizing peak tensile stress, maximizing peak tensile strain, and minimizing cost) by NSGA-III for ECC incorporating each kind of fiber. In these figures, a total of 100 Pareto points are represented by dots and are widely distributed across the feasible objective space, showcasing a reasonable range of stress, strain, and cost values. The original dataset points are denoted by small crosses in the figures. By implementing constraints (Min-max, volume, ratio, and fiber constraints), the obtained results are more practical and reasonable. This distribution serves as a testament to the efficacy of NSGA-III in addressing the complexities of multi-objective optimization problems. Within each plot, three points correspond to the maximum tensile strain, minimum cost, and maximum tensile stress among the Pareto points, where their corresponding mixture designs are presented in Tables 7 to 9.

Compared to the other fibers, the largest tensile strain is observed in PE incorporated ECC with the largest strain value of 10.51%. The contents of cement, water, fine aggregate, fly ash, silica fume, slag, and superplasticiser are 0.23, 0.19, 0.25, 0.25, 0.04, 0.06, and 0.01, respectively. The aggregate to binder ratio is 0.16 and the water to binder ratio is 0.32. The tensile strain of 10.51% is higher than PVA-ECC

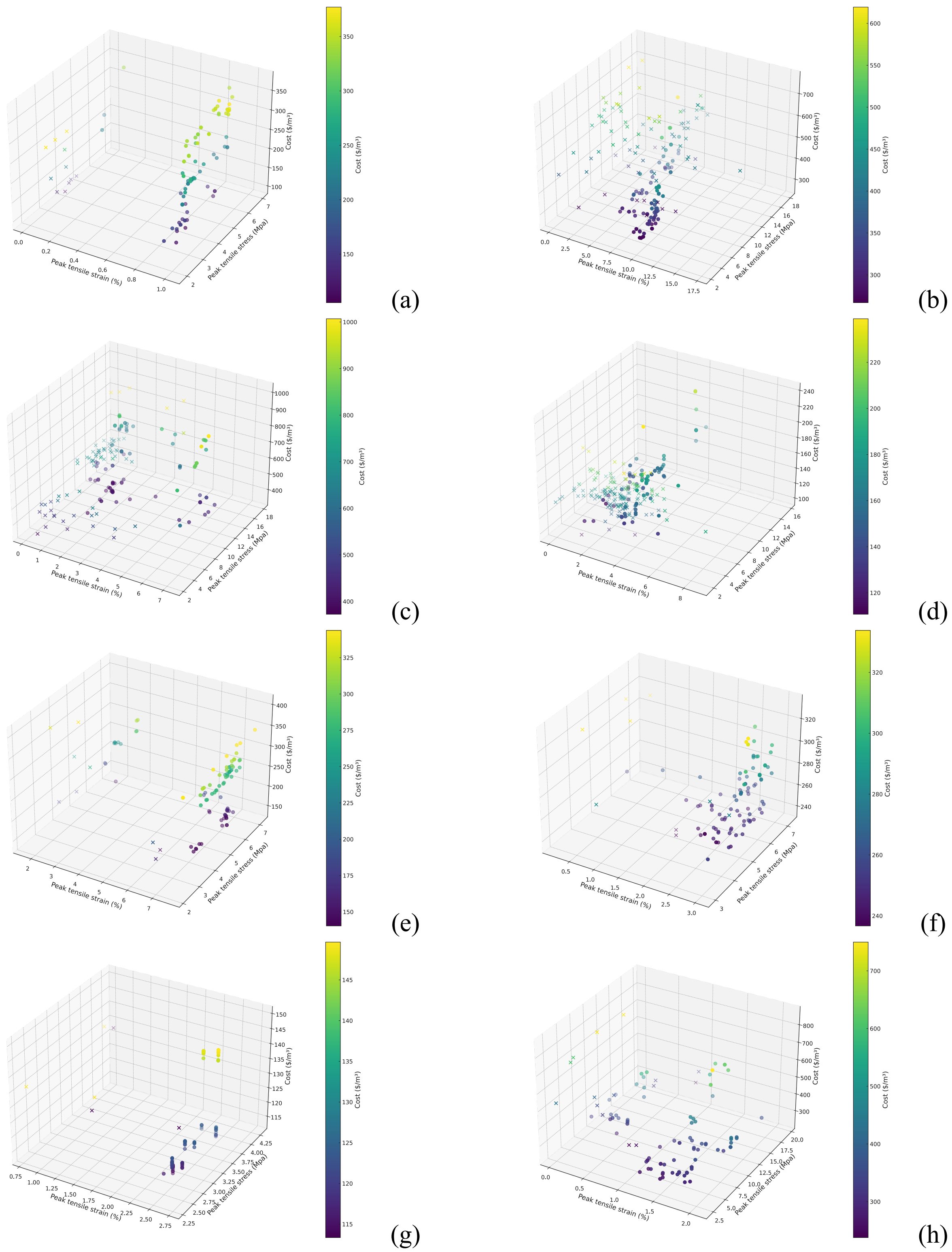
(6.45%) and is significantly higher than PP-ECC (2.73%) or Steel-ECC (2.08%). PE fibers are distinguished by their exceptional ductility and ability to withstand significant deformation without rupture [40]. This characteristic permits ECC incorporating PE fibers to demonstrate strain-hardening behavior when subjected to tensile stresses, thus improving tensile strain capacity. The inherent flexibility and toughness of PE fibers enhance crack bridging capacities, enabling a more effective distribution of stress and preserving structural integrity under loading conditions, resulting in a superior tensile strain performance relative to ECC reinforced with PP, steel, or PVA fibers [39]. While PP fibers exhibit flexibility, they lack the comprehensive elongation performance and adhesion properties with the cement matrix exhibited by PE fibers. Conversely, steel fibers offer reduced elongation capabilities despite their high tensile strength, constraining the composite's ability to deform prior to failure. PVA fiber suffers from the poor dispersion which affects its strain capacity and durability under tensile loading.

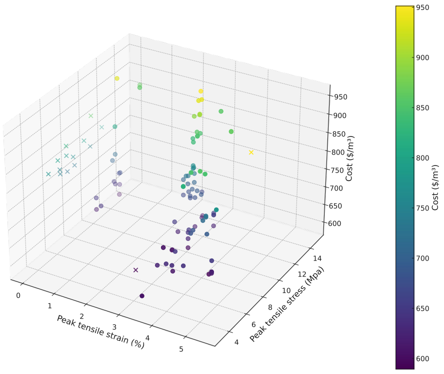
Apart from PE-ECC, the sum of binder content and values of water to cement ratio in ECC incorporated with other fibers are all higher than 0.6 and below 0.3, respectively, to achieve the highest peak tensile strain. It indicates that the high binder content and low water to cement ratio are essential in ECC which is consistent with the previous researches. When the binder content is low, the strength of the matrix decreases notably, resulting in weaker bonding strength at the fiber-matrix interface. At specific fiber concentrations, fibers may entirely detach from the matrix under tensile load, eliminating the potential for strain hardening. Additionally, maintaining a low water to cement ratio is crucial, primarily to ensure the mortar matrix remains dense.

PE, steel, PE-Steel hybrid, and UHMWPE fibers are four types of reinforcement that enable ECC composites to reach peak tensile stress levels of approximately 15MPa. In contrast, the stress observed in ECC composites reinforced with other fibers is found to be approximately half that of ECC composites reinforced with PE, Steel, or UHMWPE fibers. This is mainly due to the properties of fiber and the fiber-matrix structure. Steel fibers possess high tensile strength and stiffness, significantly enhancing the composite's load-bearing capacity to facilitate the efficient stress transfer. PE and UHMWPE fibers exhibit excellent tensile strength and elongation at break, which are crucial for the composite's toughness, bond capabilities, and crack resistance. The incorporation of PE-Steel hybrid fibers leverages the high stiffness and strength of steel with the flexibility and toughness of PE, resulting in a composite that exhibits both high peak stress and enhanced ductility [89]. The largest peak tensile stress is found as 15.25MPa for ECC incorporated with 1.5%vol PE_STEEL. The diameter, length, tensile strength, and elastic modulus of PE_STEEL are 127 micro-meter, 20mm, 5481MPa, and 285GPa, respectively. However, these values are theoretical estimates, given that PE_STEEL is a composite material consisting of PE fiber and Steel fiber. It is assumed that these properties of a hybrid fiber can be estimated by averaging the characteristics of the individual fibers according to their proportions in the mixture.

Within the 100 Pareto points, the mixture designs corresponding to the lowest cost are summarized in Table 10. The peak tensile strain and peak tensile stress are reduced with the reduction of cost. The lowest cost can be found as 105 $\$/m^3$ in Basalt-ECC with a fiber content of 0.4%. The corresponding tensile stress and strain are 3.79MPa and 0.89%, respectively. This variability in achievable properties allows for a range of options to be considered and selected by the decision-maker based on specific project requirements. Compared to Pareto points within multi-objective optimization, direct optimization for specific mixture proportions is more practical in actual situations. In generic multi-objective optimization tasks, the Pareto points are randomly generated until the end condition is reached. For specific engineering applications, this may result in numerous useless results that are not superior to the existing mixture design at all these three objectives. Therefore, this study conducts directed optimization for two samples of PE and PV (two typical fiber types). This can be achieved by setting up three additional objective constraints, namely the lower limit of tensile stress and tensile strain and the upper limit of cost. The actual and the optimized mixture designs are shown in Table 10. For instance, the strain, cost, and stress values of the chosen PVA-ECC sample are 3.7%, 134.45\$/m^3, and 4.8MPa, respectively. After optimization, these objectives can be enhanced to 4.69%-122.58\$/m^3-7.53 MPa, 3.82%-123.89\$/m^3-5.41 MPa, and 3.74%-115.42\$/m^3-7.22 MPa according to the established AI models and constraints. Similarly, this direct optimization can be found in PE-ECC samples. In cases of optimizing certain properties while keeping a strain or stress value constant, the performance bounds can be adjusted strictly.

However, it is noted that the optimized designs represent local optima rather than global optima. This implies the existence of alternative mixture designs that could potentially yield equivalent or superior objective values. Besides, the optimized mixture designs derived in this study can offer guidance for practical projects to a certain extent. However, given that the data originate from literature, the model bias caused by incomplete data is inevitable. Therefore, in practical engineering applications, actual validation is required to ensure the reliability of the proposed designs. For attaining more precise and globally optimal design mixtures, it is essential to compile a more comprehensive database reflecting local materials and environmental conditions.


Peak tensile strain (%)
Peak tensile stress (Mpa)
Cost ($/m³)
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)



(i)

Figure 13. Multi-objective mixture designs of ECC for each fiber type from (a) to (i)

Table 7. The multi-optimization of strain-cost-stress for each fiber type (Max peak tensile strain)

Cement (%W)	Water (%W)	Fine Agg (%W)	Fly ash (%W)	Silica Fume (%W)	Slag (%W)	Superplasticiser (%W)	Fiber types	Fiber content (100% Vol)	F. Diameter (Micro-Meter)	Fiber Length (mm)	Fiber Tensile Strength (Mpa)	Fiber Elastic Modulus (Gpa)	HPMC (Viscosity Agent)	Oiling Agent/coating	Temperature (Celsius)	Water Curing (days)	Air Curing (days)	Peak tensile strain (%)	Cost （$/m3）	Peak tensile stress (Mpa)
0.53	0.19	0.11	0.09	0.01	0.04	0.00	Basalt	0.018	13	10	2600	85	0	0	20	28	0	1.02	261.88	2.50
0.23	0.19	0.25	0.25	0.04	0.06	0.01	PE	0.012	47	23	3645	135	0	0	20	28	0	10.51	295.95	6.05
0.37	0.13	0.04	0.00	0.19	0.29	0.01	PE_STEEL	0.016	82	34	6248	152	0	0	20	28	0	7.24	1006.98	8.32
0.46	0.19	0.14	0.22	0.01	0.00	0.00	PVA	0.023	40	11	1340	36	0	0	20	28	0	6.45	183.83	4.25
0.13	0.17	0.20	0.46	0.00	0.00	0.00	PVA_Ca	0.020	22	14	2309	323	0	0	20	28	0	7.28	142.16	6.03
0.18	0.19	0.13	0.44	0.03	0.00	0.00	PVA_STEEL	0.026	99	8	2306	97	0	0	20	28	0	3.05	328.67	5.16
0.19	0.18	0.07	0.55	0.00	0.00	0.00	Poly	0.025	12	10	850	6	0	0	20	28	0	2.73	149.65	3.33
0.59	0.12	0.01	0.05	0.06	0.13	0.03	STEEL	0.025	234	11	3310	377	0	0	20	28	0	2.08	402.91	9.61
0.45	0.18	0.12	0.18	0.04	0.05	0.01	UHMWPE	0.021	21	13	3000	101	0	0	20	28	0	4.56	778.37	7.44

Table 8. The multi-optimization of strain-cost-stress for each fiber type (Min Cost)

Cement (%W)	Water (%W)	Fine Agg (%W)	Fly ash (%W)	Silica Fume (%W)	Slag (%W)	Superplasticiser (%W)	Fiber types	Fiber content (100% Vol)	F. Diameter (Micro-Meter)	Fiber Length (mm)	Fiber Tensile Strength (Mpa)	Fiber Elastic Modulus (Gpa)	HPMC (Viscosity Agent)	Oiling Agent/coating	Temperature (Celsius)	Water Curing (days)	Air Curing (days)	Peak tensile strain (%)	Cost （$/m3）	Peak tensile stress (Mpa)
0.49	0.20	0.27	0.03	0.00	0.02	0.00	Basalt	0.004	13	8	2600	85	0	0	20	28	0	0.89	105.48	3.79
0.23	0.19	0.23	0.25	0.02	0.07	0.02	PE	0.010	28	36	4649	137	0	0	20	28	0	9.38	267.05	3.53
0.48	0.11	0.34	0.00	0.02	0.03	0.01	PE_STEEL	0.007	221	33	5956	234	0	0	20	28	0	1.68	372.50	11.63
0.19	0.17	0.27	0.37	0.01	0.00	0.00	PVA	0.013	39	11	1429	36	0	0	20	28	0	1.27	110.70	6.54
0.13	0.17	0.21	0.48	0.00	0.00	0.00	PVA_Ca	0.020	28	12	2687	275	0	0	20	28	0	7.02	140.31	6.42
0.23	0.18	0.16	0.42	0.00	0.00	0.00	PVA_STEEL	0.020	138	11	2183	102	0	0	20	28	0	2.22	236.69	5.23
0.19	0.19	0.06	0.57	0.00	0.00	0.00	Poly	0.016	12	10	850	6	0	0	20	28	0	2.45	113.35	2.81
0.45	0.25	0.01	0.05	0.05	0.16	0.01	STEEL	0.010	308	27	4626	326	0	0	20	28	0	1.38	237.77	4.68
0.50	0.18	0.17	0.10	0.01	0.05	0.00	UHMWPE	0.016	18	13	3000	102	0	0	20	28	0	4.31	589.40	7.67

Table 9. The multi-optimization of strain-cost-stress for each fiber type (Max peak tensile stress)

Cement (%W)	Water (%W)	Fine Agg (%W)	Fly ash (%W)	Silica Fume (%W)	Slag (%W)	Superplasticiser (%W)	Fiber types	Fiber content (100% Vol)	F. Diameter (Micro-Meter)	Fiber Length (mm)	Fiber Tensile Strength (Mpa)	Fiber Elastic Modulus (Gpa)	HPMC (Viscosity Agent)	Oiling Agent/coating	Temperature (Celsius)	Water Curing (days)	Air Curing (days)	Peak tensile strain (%)	Cost （$/m3）	Peak tensile stress (Mpa)
0.45	0.19	0.21	0.07	0.06	0.01	0.00	Basalt	0.020	13	8	2600	85	0	0	20	28	0	0.84	343.83	6.99
0.29	0.12	0.20	0.06	0.19	0.09	0.02	PE	0.012	44	8	3811	73	0	0	20	28	0	6.16	619.40	15.70
0.55	0.12	0.24	0.00	0.03	0.02	0.03	PE_STEEL	0.015	127	20	5481	285	0	0	20	28	0	3.91	712.26	15.25
0.27	0.12	0.27	0.28	0.02	0.00	0.00	PVA	0.013	35	8	1432	30	0	0	20	28	0	1.97	150.78	8.71
0.17	0.17	0.21	0.47	0.00	0.00	0.00	PVA_Ca	0.028	14	14	2034	313	0	0	20	28	0	1.67	183.82	6.58
0.25	0.16	0.24	0.35	0.01	0.00	0.00	PVA_STEEL	0.020	115	11	2326	94	0	0	20	28	0	2.51	255.13	7.24
0.19	0.18	0.07	0.55	0.00	0.00	0.00	Poly	0.025	12	10	850	6	0	0	20	28	0	2.73	149.65	3.33
0.41	0.12	0.14	0.05	0.05	0.16	0.03	STEEL	0.022	307	12	4602	383	0	0	20	28	0	2.04	359.37	15.80
0.26	0.10	0.31	0.15	0.14	0.01	0.01	UHMWPE	0.018	24	13	3000	113	0	0	20	28	0	0.19	864.94	14.87

Table 10. The direct optimization of strain-cost-stress for specific mixture design

Cement (%W)	Water (%W)	Fine Agg (%W)	Fly ash (%W)	Silica Fume (%W)	Slag (%W)	Superplasticiser (%W)	Fiber types	Fiber content (100% Vol)	F. Diameter (Micro-Meter)	Fiber Length (mm)	Fiber Tensile Strength (Mpa)	Fiber Elastic Modulus (Gpa)	HPMC (Viscosity Agent)	Oiling Agent/coating	Temperature (Celsius)	Water Curing (days)	Air Curing (days)	Peak tensile strain (%)	Cost（$/m3）	Peak tensile stress (Mpa)
0.16	**0.16**	**0.23**	**0.45**	**0.00**	**0.00**	**0.00**	**PVA**	**0.020**	**39**	**12**	**1620**	**43**	**0**	**0**	**20**	**28**	**0**	**3.70**	**134.45**	**4.80**
0.22	0.12	0.31	0.31	0.00	0.00	0.00	PVA	0.014	40	10	1152	28	0	0	20	28	0	4.69	122.58	7.53
0.24	0.18	0.46	0.13	0.01	0.00	0.00	PVA	0.015	36	9	1155	26	0	0	20	28	0	3.82	123.89	5.41
0.10	0.10	0.41	0.35	0.00	0.00	0.01	PVA	0.011	39	10	1240	33	0	0	20	28	0	3.74	115.42	7.22
0.37	**0.17**	**0.45**	**0.00**	**0.00**	**0.00**	**0.01**	**PVA**	**0.025**	**39**	**12**	**1620**	**43**	**1**	**1**	**20**	**28**	**0**	**4.59**	**182.94**	**5.00**
0.17	0.11	0.32	0.42	0.00	0.00	0.00	PVA	0.019	40	10	1238	43	1	1	20	28	0	4.93	136.58	6.52
0.21	0.09	0.36	0.30	0.01	0.00	0.00	PVA	0.020	40	10	1410	42	1	1	20	28	0	4.61	151.58	6.68
0.24	0.10	0.30	0.31	0.02	0.00	0.00	PVA	0.013	40	10	1609	27	1	1	20	28	0	4.80	143.20	8.87
0.70	**0.21**	**0.00**	**0.00**	**0.08**	**0.00**	**0.02**	**PE**	**0.018**	**70**	**22**	**4698**	**142**	**0**	**0**	**20**	**28**	**0**	**3.50**	**509.28**	**3.10**
0.32	0.15	0.15	0.29	0.06	0.04	0.00	PE	0.020	49	21	3405	122	0	0	20	28	0	5.80	462.16	5.31
0.23	0.15	0.08	0.23	0.13	0.18	0.00	PE	0.015	61	24	4562	68	0	0	20	28	0	6.95	501.08	3.22
0.32	0.15	0.15	0.30	0.06	0.04	0.00	PE	0.015	49	21	2822	122	0	0	20	28	0	6.09	389.84	4.47
0.53	**0.12**	**0.20**	**0.00**	**0.13**	**0.00**	**0.01**	**PE**	**0.020**	**24**	**18**	**3000**	**120**	**0**	**0**	**23**	**28**	**0**	**7.00**	**633.40**	**7.10**
0.44	0.12	0.12	0.12	0.09	0.11	0.00	PE	0.014	40	32	2528	114	0	0	20	28	0	7.96	449.44	7.88
0.27	0.13	0.23	0.16	0.08	0.13	0.02	PE	0.016	24	28	3689	79	0	0	20	28	0	8.80	462.35	9.16
0.29	0.11	0.13	0.07	0.06	0.30	0.02	PE	0.010	61	13	3078	123	0	0	20	28	0	7.35	400.36	10.55

6. Conclusion

ECC is a strain-hardening cementitious material with high tensile strength and ductility mainly due to the fiber micromechanical reinforcement. This study proposes a novel MOO framework to achieve a reliable and smart mixture design of ECC. The main conclusions are drawn as follows:

(1) Data augmentation is performed using a Tuned-CTGAN on both tensile stress and strain datasets. As evidenced by its high KSTest and CSTest values for 18 features, the reliability of virtual data generation has been verified.

(2) The HGS algorithm is applied to automatedly tune the hyperparameters of ML models. The XGBoost and RF show the highest predictive accuracy in tensile stress and strain datasets before the data augmentation, respectively.

(3) After the data augmentation, the evaluation indexes of SVR and RF become worse. In contrast, the values of R^2 of XGBoost in stress and strain test sets are increased from 0.874 to 0.925 and from 0.772 to 0.889, respectively.

(4) The NSGA-III successfully generates the Pareto front for multi-objective (tensile stress, tensile strain, and cost) with four classes of constraints (Min-max, volume, ratio, and fiber). It depends on the decision-maker to select the multi-objective

optimization solutions. Future research will conduct cost-based optimization, incorporating a cost formula that takes into account fiber parameters and regions.

Future researches: Since the observed phenomenon showed a negative influence of data augmentation on SVR, RF, and GBR, the possible reason is they are not suitable to tackle big database tasks. The efficacy of data augmentation will be investigated on deep learning algorithms in the future.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT-4 in order to improve readability and language. After using this tool/service, the authors reviewed and edited the content as needed and took full responsibility for the content of the publication.

Reference

[1] LI VICTOR C, *Engineered cementitious composites (ECC): bendable concrete for sustainable and resilient infrastructure*. 2019: Springer.

[2] YU KEQUAN, LI LINGZHI, YU JIANGTAO, WANG YICHAO, YE JUNHONG and XU QINGFENG, *Direct tensile properties of engineered cementitious composites: A review.* Construction and Building Materials, 2018. **165**: p. 346-362.

[3] YU KE-QUAN, YU JIANG-TAO, DAI JIAN-GUO, LU ZHOU-DAO and SHAH SURENDRA P., *Development of ultra-high performance engineered cementitious composites using polyethylene (PE) fibers.* Construction and Building Materials, 2018. **158**: p. 217-227 DOI: 10.1016/j.conbuildmat.2017.10.040.

[4] SUN JUNBO, WANG YUFEI, YAO XUPEI, REN ZHENHUA, ZHANG GENBAO, ZHANG CHAO, CHEN XIANGHONG, MA WEI, and WANG XIANGYU, *Machine-Learning-Aided Prediction of Flexural Strength and ASR Expansion for Waste Glass Cementitious Composite.* Applied Sciences, 2021. **11**(15): p. 6686.

[5] YU JIE, WENG YIWEI, YU JIANGTAO, CHEN WENGUANG, LU SHUAINAN and YU KEQUAN, *Generative AI for performance-based design of engineered cementitious composite.* Composites Part B: Engineering, 2023. **266**: p. 110993.

[6] FENG WANHUI, WANG YUFEI, SUN JUNBO, TANG YUNCHAO, WU DONGXIAO, JIANG ZHIWEI, WANG JIANQUN, and WANG XIANGYU,

Prediction of thermo-mechanical properties of rubber-modified recycled aggregate concrete. Construction and Building Materials, 2022. **318**: p. 125970.

[7] SUN JUNBO, MA YONGZHI, LI JIANXIN, ZHANG JUNFEI, REN ZHENHUA and WANG XIANGYU, *Machine learning-aided design and prediction of cementitious composites containing graphite and slag powder.* Journal of Building Engineering, 2021: p. 102544.

[8] SUN JUNBO, ZHANG JUNFEI, GU YUNFAN, HUANG YIMIAO, SUN YUANTIAN and MA GUOWEI, *Prediction of permeability and unconfined compressive strength of pervious concrete using evolved support vector regression.* Construction and Building Materials, 2019. **207**: p. 440-449 DOI: https://doi.org/10.1016/j.conbuildmat.2019.02.117.

[9] SHARIATI MAHDI, ARMAGHANI DANIAL JAHED, KHANDELWAL MANOJ, ZHOU JIAN and KHORAMI MAJID, *Assessment of longstanding effects of fly ash and silica fume on the compressive strength of concrete using extreme learning machine and artificial neural network.* Journal of Advanced Engineering and Computation, 2021. **5**(1): p. 50-74.

[10] FENG DE-CHENG, LIU ZHEN-TAO, WANG XIAO-DAN, CHEN YIN, CHANG JIA-QI, WEI DONG-FANG, and JIANG ZHONG-MING, *Machine learning-based compressive strength prediction for concrete: An adaptive boosting approach.* Construction and Building Materials, 2020. **230**: p. 117000.

[11] CHEN TIANQI and GUESTRIN CARLOS. *Xgboost: A scalable tree boosting system*. in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016.

[12] YANG YUTAO, CHEN HUILING, HEIDARI ALI ASGHAR and GANDOMI AMIR H, *Hunger games search: Visions, conception, implementation, deep analysis, perspectives, and towards performance shifts.* Expert Systems with Applications, 2021. **177**: p. 114864.

[13] HUYNH AN THAO, NGUYEN QUANG DANG, XUAN QUI LIEU, MAGEE BRYAN, CHUNG TAECHOONG, TRAN KIET TUAN, and NGUYEN KHOA TAN, *A machine learning-assisted numerical predictor for compressive strength of geopolymer concrete based on experimental data and sensitivity analysis.* Applied Sciences, 2020. **10**(21): p. 7726.

[14] CRESWELL ANTONIA, WHITE TOM, DUMOULIN VINCENT, ARULKUMARAN KAI, SENGUPTA BISWA and BHARATH ANIL A, *Generative adversarial networks: An overview.* IEEE signal processing magazine, 2018. **35**(1): p. 53-65.

[15] KARRAS TERO, AILA TIMO, LAINE SAMULI and LEHTINEN JAAKKO, *Progressive growing of gans for improved quality, stability, and variation.* arXiv preprint arXiv:1710.10196, 2017.

[16] ZHANG ZHIXIN, PAN XUHUA, JIANG SHUHAO and ZHAO PEIJUN, *High-quality face image generation based on generative adversarial networks.*

Journal of Visual Communication and Image Representation, 2020. **71**: p. 102719.

[17] DEB KALYANMOY and JAIN HIMANSHU, *An evolutionary many-objective optimization algorithm using reference-point-based nondominated sorting approach, part I: solving problems with box constraints.* IEEE transactions on evolutionary computation, 2013. **18**(4): p. 577-601.

[18] NATEGHI-A FARIBORZ and AHMADI MOHAMMAD HOSSEIN, *Prediction of engineered cementitious composite material properties using artificial neural network.* International Journal of Engineering, 2019. **32**(11): p. 1534-1542.

[19] MAHJOUBI SOROUSH, BARHEMAT ROJYAR, GUO PENGWEI, MENG WEINA and BAO YI, *Prediction and multi-objective optimization of mechanical, economical, and environmental properties for strain-hardening cementitious composites (SHCC) based on automated machine learning and metaheuristic algorithms.* Journal of Cleaner Production, 2021. **329**: p. 129665.

[20] KHAN YASAR, ZAFAR ADEEL, REHMAN MUHAMMAD FAISAL, JAVED MUHAMMAD FAISAL, IFTIKHAR BAWAR and GAMIL YASER, *Bio-inspired based meta-heuristic approach for predicting the strength of fiber-reinforced based strain hardening cementitious composites.* Heliyon, 2023. **9**(11).

[21] CHEN GUANGWEI, TANG WAICHING, CHEN SHUO, WANG SHANYONG and CUI HONGZHI, *Prediction of self-healing of engineered cementitious composite using machine learning approaches.* Applied Sciences, 2022. **12**(7): p. 3605.

[22] ASLAM FAHID, ALYOUSEF RAYED, AWAN HAMAD HASSAN and JAVED MUHAMMAD FAISAL. *Forecasting the self-healing capacity of engineered cementitious composites using bagging regressor and stacking regressor*. in *Structures*. 2023. Elsevier.

[23] UDDIN MD NASIR, SHANMUGASUNDARAM N, PRAVEENKUMAR S and LI LING-ZHI, *Prediction of compressive strength and tensile strain of engineered cementitious composite using machine learning.* International Journal of Mechanics and Materials in Design, 2024: p. 1-46.

[24] ALABDULJABBAR HISHAM, KHAN KAFFAYATULLAH, AWAN HAMAD HASSAN, ALYOUSEF RAYED, MOHAMED ABDELIAZIM MUSTAFA and ELDIN SAYED M, *Modeling the capacity of engineered cementitious composites for self-healing using AI-based ensemble techniques.* Case Studies in Construction Materials, 2023. **18**: p. e01805.

[25] TANYILDIZI HARUN, *Prediction of compressive strength of nano-silica modified engineering cementitious composites exposed to high temperatures using hybrid deep learning models.* Expert Systems with Applications, 2024. **241**: p. 122474.

[26] ABDULKADIR ISYAKA, MOHAMMED BASHAR S, LIEW MS and WAHAB MMA, *Modelling and optimization of the mechanical properties of engineered cementitious composite containing crumb rubber pretreated with graphene oxide using response surface methodology.* Construction and Building Materials, 2021. **310**: p. 125259.

[27] BHEEL NARAINDAS, MOHAMMED BASHAR S, ABDULKADIR ISYAKA, LIEW MS and ZAWAWI NOOR AMILA WAN ABDULLAH, *Effects of Graphene Oxide on the Properties of Engineered Cementitious Composites: Multi-Objective Optimization Technique Using RSM.* Buildings, 2023. **13**(8): p. 2018.

[28] WANG XINGYU, CUI FENGKUN, CUI LONG and JIANG DI, *Research on a Multi-Objective Optimization Design for the Durability of High-Performance Fiber-Reinforced Concrete Based on a Hybrid Algorithm.* Coatings, 2023. **13**(12): p. 2054.

[29] GUO PENGWEI, MENG WEINA, XU MINGFENG, LI VICTOR C and BAO YI, *Predicting mechanical properties of high-performance fiber-reinforced cementitious composites by integrating micromechanics and machine learning.* Materials, 2021. **14**(12): p. 3143.

[30] YUAN XIONGZHOU, CAO QINGYU, AMIN MUHAMMAD NASIR, AHMAD AYAZ, AHMAD WAQAS, ALTHOEY FADI, and DEIFALLA AHMED FAROUK, *Predicting the crack width of the engineered cementitious materials via standard machine learning algorithms.* Journal of Materials Research and Technology, 2023. **24**: p. 6187-6200.

[31] TRAN NGOC THANH, NGUYEN TAN KHOA, NGUYEN DUY-LIEM and LE QUANG HUY, *Assessment of fracture energy of strain-hardening fiber-reinforced cementitious composite using experiment and machine learning technique.* Structural Concrete, 2023. **24**(3): p. 4185-4198.

[32] HEMMATIAN ABOLFAZL, JALALI MEYSAM, NADERPOUR HOSEIN and NEHDI MONCEF L, *Machine learning prediction of fiber pull-out and bond-slip in fiber-reinforced cementitious composites.* Journal of Building Engineering, 2023. **63**: p. 105474.

[33] SHANMUGASUNDARAM N, PRAVEENKUMAR S, GAYATHIRI K and DIVYA S, *Prediction on compressive strength of engineered cementitious composites using machine learning approach.* Construction and Building Materials, 2022. **342**: p. 127933.

[34] UDDIN MD NASIR, LI LING-ZHI, AHMED ASIF and ALMAJHALI KHALIL YAHYA MOHAMMED, *Prediction of PVA fiber effect in Engineered Composite cement (ECC) by Artificial neural Network (ANN).* Materials Today: Proceedings, 2022. **65**: p. 537-542.

[35] MORSY ALAA M, ABD ELMOATY M and HARRAZ ABDELRHMAN B, *Predicting mechanical properties of Engineering Cementitious Composite*

reinforced with PVA using artificial neural network. Case Studies in Construction Materials, 2022. **16**: p. e00998.

[36] ALTAYEB MOHAMEDELMUJTABA, WANG XIN and MUSA TAHA HUSSEIN, *An ensemble method for predicting the mechanical properties of strain hardening cementitious composites.* Construction and Building Materials, 2021. **286**: p. 122807.

[37] KANG MIN-CHANG, YOO DOO-YEOL and GUPTA RISHI, *Machine learning-based prediction for compressive and flexural strengths of steel fiber-reinforced concrete.* Construction and Building Materials, 2021. **266**: p. 121117.

[38] SHI LEIFENG, LIN STK, LU YE, YE LIN and ZHANG YX, *Artificial neural network based mechanical and electrical property prediction of engineered cementitious composites.* Construction and Building Materials, 2018. **174**: p. 667-674.

[39] ZHU JI-XIANG, XU LING-YU, HUANG BO-TAO, WENG KE-FAN and DAI JIAN-GUO, *Recent developments in Engineered/Strain-Hardening Cementitious Composites (ECC/SHCC) with high and ultra-high strength.* Construction and Building Materials, 2022. **342**: p. 127956.

[40] YOO DOO-YEOL and BANTHIA NEMKUMAR, *High-performance strain-hardening cementitious composites with tensile strain capacity exceeding 4%: A review.* Cement and Concrete Composites, 2022. **125**: p. 104325.

[41] YOO DOO-YEOL, OH TAEKGEUN, KANG MIN-CHANG, KIM MIN-JAE and CHOI HONG-JOON, *Enhanced tensile ductility and sustainability of high-strength strain-hardening cementitious composites using waste cement kiln dust and oxidized polyethylene fibers.* Cement and Concrete Composites, 2021. **120** DOI: 10.1016/j.cemconcomp.2021.104030.

[42] YU KE-QUAN, LU ZHOU-DAO, DAI JIAN-GUO and SHAH SURENDRA P., *Direct Tensile Properties and Stress–Strain Model of UHP-ECC.* Journal of Materials in Civil Engineering, 2020. **32**(1) DOI: 10.1061/(asce)mt.1943-5533.0002975.

[43] YU KEQUAN, GUO YINGYING, ZHANG Y. X. and SOE KHIN, *Magnesium oxychloride cement-based strain-hardening cementitious composite: Mechanical property and water resistance.* Construction and Building Materials, 2020. **261** DOI: 10.1016/j.conbuildmat.2020.119970.

[44] NGUYỄN HUY HOàNG, CHOI JEONG-IL, PARK SE-EON, CHA SANG LYUL, HUH JUNGWON and LEE BANG YEON, *Autogenous healing of high strength engineered cementitious composites (ECC) using calcium-containing binders.* Construction and Building Materials, 2020. **265** DOI: 10.1016/j.conbuildmat.2020.120857.

[45] ZHANG ZHIGANG, YANG FAN, LIU JIN-CHENG and WANG SHUPING, *Eco-friendly high strength, high ductility engineered cementitious composites (ECC) with substitution of fly ash by rice husk ash.* Cement and Concrete Research, 2020. **137** DOI: 10.1016/j.cemconres.2020.106200.

[46] JI JING, SONG HUAYU, JIANG LIANGQIN, REN HONGGUO, ZHANG YUNFENG and LIU YINGCHUN, *Tensile Performance of High Ductility Cementitious Composites With Recycled Powder From C&D Waste.* Frontiers in Materials, 2021. **8** DOI: 10.3389/fmats.2021.673752.

[47] YOO DOO-YEOL, OH TAEKGEUN and CHUN BOOKI, *Highly ductile ultra-rapid-hardening mortar containing oxidized polyethylene fibers.* Construction and Building Materials, 2021. **277** DOI: 10.1016/j.conbuildmat.2021.122317.

[48] HUANG BO-TAO, WU JIA-QI, YU JING, DAI JIAN-GUO, LEUNG CHRISTOPHER K. Y. and LI VICTOR C., *Seawater sea-sand engineered/strain-hardening cementitious composites (ECC/SHCC): Assessment and modeling of crack characteristics.* Cement and Concrete Research, 2021. **140** DOI: 10.1016/j.cemconres.2020.106292.

[49] ZHANG ZHIGANG, YUVARAJ ANANYA, DI JIN and QIAN SHUNZHI, *Matrix design of light weight, high strength, high ductility ECC.* Construction and Building Materials, 2019. **210**: p. 188-197 DOI: 10.1016/j.conbuildmat.2019.03.159.

[50] YU KEQUAN, DING YAO and ZHANG Y. X., *Size effects on tensile properties and compressive strength of engineered cementitious composites.* Cement and Concrete Composites, 2020. **113** DOI: 10.1016/j.cemconcomp.2020.103691.

[51] RANADE RAVI, LI VICTOR C, STULTS MICHAEL D, HEARD WILLIAM F and RUSHING TODD S, *Composite Properties of High-Strength, High-Ductility Concrete.* ACI Materials Journal, 2013. **110**(4).

[52] OH TAEKGEUN, KIM MIN-JAE, BANTHIA NEMKUMAR and YOO DOO-YEOL, *Influence of curing conditions on mechanical and microstructural properties of ultra-high-performance strain-hardening cementitious composites with strain capacity up to 17.3%.* Developments in the Built Environment, 2023. **14** DOI: 10.1016/j.dibe.2023.100150.

[53] LUO JIANJUN, CAI ZIWEI, YU KEQUAN, ZHU WENJUN and LU ZHOUDAO, *Temperature impact on the micro-structures and mechanical properties of high-strength engineered cementitious composites.* Construction and Building Materials, 2019. **226**: p. 686-698 DOI: 10.1016/j.conbuildmat.2019.07.322.

[54] LU ZEYU, YAO JIE and LEUNG CHRISTOPHER K. Y., *Using graphene oxide to strengthen the bond between PE fiber and matrix to improve the strain hardening behavior of SHCC.* Cement and Concrete Research, 2019. **126** DOI: 10.1016/j.cemconres.2019.105899.

[55] ZHOU YINGWU, ZHENG SHUYUE, HUANG XIAOXU, XI BIN, HUANG ZHENYU and GUO MENGHUAN, *Performance enhancement of green high-ductility engineered cementitious composites by nano-silica incorporation.* Construction and Building Materials, 2021. **281** DOI: 10.1016/j.conbuildmat.2021.122618.

[56] DING YAO, YU KEQUAN, MAO WEIHAO and ZHANG SHUO, *Performance‐enhancement of high‐strength strain‐hardening cementitious composite by nano‐particles: Mechanism and property characterization.* Structural Concrete, 2021. **23**(4): p. 2061-2075 DOI: 10.1002/suco.202000637.

[57] DING YAO, LIU JIE-PENG and BAI YU-LEI, *Linkage of multi-scale performances of nano-CaCO3 modified ultra-high performance engineered cementitious composites (UHP-ECC).* Construction and Building Materials, 2020. **234** DOI: 10.1016/j.conbuildmat.2019.117418.

[58] YU KE-QUAN, DAI JIAN-GUO, LU ZHOU-DAO and POON CHI-SUN, *Rate-dependent tensile properties of ultra-high performance engineered cementitious composites (UHP-ECC).* Cement and Concrete Composites, 2018. **93**: p. 218-234 DOI: 10.1016/j.cemconcomp.2018.07.016.

[59] WANG JIANQUN, LIU HENG, SUN JUNBO, HUANG BO, WANG YUFEI, ZHAO HONGYU, SAAFI MOHAMED, and WANG XIANGYU, *Research on Concrete Early Shrinkage Characteristics Based on Machine Learning Algorithms for Multi-objective Optimization.* Journal of Building Engineering, 2024: p. 109415.

[60] KIM MIN-JAE, CHOI HONG-JOON, SHIN WONSIK, OH TAEKGEUN and YOO DOO-YEOL, *Development of impact resistant high-strength strain-hardening cementitious composites (HS-SHCC) superior to reactive powder concrete (RPC) under flexure.* Journal of Building Engineering, 2021. **44** DOI: 10.1016/j.jobe.2021.102652.

[61] ZHOU YINGWU, XI BIN, SUI LILI, ZHENG SHUYUE, XING FENG and LI L., *Development of high strain-hardening lightweight engineered cementitious composites: Design and performance.* Cement and Concrete Composites, 2019. **104** DOI: 10.1016/j.cemconcomp.2019.103370.

[62] LI LINZHI, CAI ZIWEI, YU KEQUAN, ZHANG Y. X. and DING YAO, *Performance-based design of all-grade strain hardening cementitious composites with compressive strengths from 40 MPa to 120 MPa.* Cement and Concrete Composites, 2019. **97**: p. 202-217 DOI: 10.1016/j.cemconcomp.2019.01.001.

[63] DING YAO, YU KE-QUAN, YU JIANG-TAO and XU SHI-LANG, *Structural behaviors of ultra-high performance engineered cementitious composites (UHP-ECC) beams subjected to bending-experimental study.* Construction and Building Materials, 2018. **177**: p. 102-115 DOI: 10.1016/j.conbuildmat.2018.05.122.

[64] SUI L., ZHONG Q., YU K., XING F., LI P. and ZHOU Y., *Flexural Fatigue Properties of Ultra-High Performance Engineered Cementitious Composites (UHP-ECC) Reinforced by Polymer Fibers.* Polymers (Basel), 2018. **10**(8) DOI: 10.3390/polym10080892.

[65] SHETA AHMED, MA XING, ZHUGE YAN, ELGAWADY MOHAMED A., MILLS JULIE E., SINGH ASHESH, and ABD-ELAAL EL-SAYED,

Structural performance of novel thin-walled composite cold-formed steel/PE-ECC beams. Thin-Walled Structures, 2021. **162** DOI: 10.1016/j.tws.2021.107586.

[66] ZHOU YINGWU, GONG GUOQIANG, HUANG YUANJUN, CHEN CHI, HUANG DONGSHENG, CHEN ZIQIAN, and GUO MENGHUAN, *Feasibility of incorporating recycled fine aggregate in high performance green lightweight engineered cementitious composites.* Journal of Cleaner Production, 2021. **280** DOI: 10.1016/j.jclepro.2020.124445.

[67] YU KEQUAN, WANG YICHAO, YU JIANGTAO and XU SHILANG, *A strain-hardening cementitious composites with the tensile capacity up to 8%.* Construction and Building Materials, 2017. **137**: p. 410-419 DOI: 10.1016/j.conbuildmat.2017.01.060.

[68] YU KE-QUAN, ZHU WEN-JUN, DING YAO, LU ZHOU-DAO, YU JIANG-TAO and XIAO JIAN-ZHUANG, *Micro-structural and mechanical properties of ultra-high performance engineered cementitious composites (UHP-ECC) incorporation of recycled fine powder (RFP).* Cement and Concrete Research, 2019. **124** DOI: 10.1016/j.cemconres.2019.105813.

[69] OH TAEKGEUN, KIM MIN-JAE, KIM SOONHO, LEE SEUNG KYUN, KANG MIN-CHANG and YOO DOO-YEOL, *Electrical and mechanical properties of high-strength strain-hardening cementitious composites containing silvered polyethylene fibers.* Journal of Building Engineering, 2022. **46** DOI: 10.1016/j.jobe.2021.103719.

[70] LEI DONG-YI, GUO LI-PING, CHEN BO, CUROSU IURIE and MECHTCHERINE VIKTOR, *The connection between microscopic and macroscopic properties of ultra-high strength and ultra-high ductility cementitious composites (UHS-UHDCC).* Composites Part B: Engineering, 2019. **164**: p. 144-157 DOI: 10.1016/j.compositesb.2018.11.062.

[71] ZHOU YINGWU, ZHONG QIANLI, XING FENG, SUI LILI, HUANG ZHENYU and GUO MENGHUAN, *Influence of cyclic loading on the tensile fracture characteristics of ultra-high performance engineered cementitious composites.* Construction and Building Materials, 2020. **240** DOI: 10.1016/j.conbuildmat.2019.117937.

[72] TANG YUNCHAO, WANG YUFEI, WU DONGXIAO, CHEN MENGCHENG, PANG LAN, SUN JUNBO, FENG WANHUI, and WANG XIANGYU, *Exploring temperature-resilient recycled aggregate concrete with waste rubber: An experimental and multi-objective optimization analysis.* Reviews on Advanced Materials Science, 2023. **62**(1): p. 20230347.

[73] XU LING-YU, HUANG BO-TAO, LI VICTOR C. and DAI JIAN-GUO, *High-strength high-ductility Engineered/Strain-Hardening Cementitious Composites (ECC/SHCC) incorporating geopolymer fine aggregates.* Cement and Concrete Composites, 2022. **125** DOI: 10.1016/j.cemconcomp.2021.104296.

[74] WU JIAN-DONG, GUO LI-PING, CAO YUAN-ZHANG and LYU BANG-CHENG, *Mechanical and fiber/matrix interfacial behavior of ultra-high-strength and high-ductility cementitious composites incorporating waste glass powder.* Cement and Concrete Composites, 2022. **126** DOI: 10.1016/j.cemconcomp.2021.104371.

[75] WU JIAN-DONG, GUO LI-PING and QIN YAO-YI, *Preparation and characterization of ultra-high-strength and ultra-high-ductility cementitious composites incorporating waste clay brick powder.* Journal of Cleaner Production, 2021. **312** DOI: 10.1016/j.jclepro.2021.127813.

[76] HUANG BO-TAO, ZHU JI-XIANG, WENG KE-FAN, LI VICTOR C. and DAI JIAN-GUO, *Ultra-high-strength engineered/strain-hardening cementitious composites (ECC/SHCC): Material design and effect of fiber hybridization.* Cement and Concrete Composites, 2022. **129** DOI: 10.1016/j.cemconcomp.2022.104464.

[77] HUANG BO-TAO, WU JIA-QI, YU JING, DAI JIAN-GUO and LEUNG CHRISTOPHER K. Y., *High-strength seawater sea-sand Engineered Cementitious Composites (SS-ECC): Mechanical performance and probabilistic modeling.* Cement and Concrete Composites, 2020. **114** DOI: 10.1016/j.cemconcomp.2020.103740.

[78] XU LING-YU, HUANG BO-TAO and DAI JIAN-GUO, *Development of engineered cementitious composites (ECC) using artificial fine aggregates.* Construction and Building Materials, 2021. **305** DOI: 10.1016/j.conbuildmat.2021.124742.

[79] XU SHI-LANG, XU HONG-LIN, HUANG BO-TAO, LI QING-HUA, YU KE-QUAN and YU JIANG-TAO, *Development of ultrahigh-strength ultrahigh-toughness cementitious composites (UHS-UHTCC) using polyethylene and steel fibers.* Composites Communications, 2022. **29** DOI: 10.1016/j.coco.2021.100992.

[80] KIM MIN-JAE, YOO DOO-YEOL and YOON YOUNG-SOO, *Effects of geometry and hybrid ratio of steel and polyethylene fibers on the mechanical performance of ultra-high-performance fiber-reinforced cementitious composites.* Journal of Materials Research and Technology, 2019. **8**(2): p. 1835-1848 DOI: 10.1016/j.jmrt.2019.01.001.

[81] YAO XIAOFEI, LYU XIN, SUN JUNBO, WANG BOLIN, WANG YUFEI, YANG MIN, WEI YAO, ELCHALAKANI MOHAMED, LI DANQI, and WANG XIANGYU, *AI-based performance prediction for 3D-printed concrete considering anisotropy and steam curing condition.* Construction and Building Materials, 2023. **375**: p. 130898.

[82] HE SHAN, QIU JISHEN, LI JUNXIA and YANG EN-HUA, *Strain hardening ultra-high performance concrete (SHUHPC) incorporating CNF-coated polyethylene fibers.* Cement and Concrete Research, 2017. **98**: p. 50-60 DOI: 10.1016/j.cemconres.2017.04.003.

[83] YUAN TIAN-FENG, LEE JIN-YOUNG and YOON YOUNG-SOO, *Enhancing the tensile capacity of no-slump high-strength high-ductility concrete.* Cement and Concrete Composites, 2020. **106** DOI: 10.1016/j.cemconcomp.2019.103458.

[84] TANG YUNCHAO, WANG YUFEI, WU DONGXIAO, LIU ZHONGHE, ZHANG HEXIN, ZHU MING, CHEN ZHENG, SUN JUNBO, and WANG XIANGYU, *An experimental investigation and machine learning-based prediction for seismic performance of steel tubular column filled with recycled aggregate concrete.* Reviews on Advanced Materials Science, 2022. **61**(1): p. 849-872.

[85] LIU TIANAN, BAI RUIXIANG, CHEN ZHITAO, LI YAZHAO and YANG YINGZI, *Tailoring of polyethylene fiber surface by coating silane coupling agent for strain hardening cementitious composite.* Construction and Building Materials, 2021. **278** DOI: 10.1016/j.conbuildmat.2021.122263.

[86] WEI JIAYING, WU CHANG, CHEN YIXIN and LEUNG CHRISTOPHER K. Y., *Shear strengthening of reinforced concrete beams with high strength strain-hardening cementitious composites (HS-SHCC).* Materials and Structures, 2020. **53**(4) DOI: 10.1617/s11527-020-01537-1.

[87] SUN JUNBO, WANG XIANGYU, ZHANG JUNFEI, XIAO FAN, SUN YUANTIAN, REN ZHENHUA, ZHANG GENBAO, LIU SHUKUI, and WANG YUFEI, *Multi-objective optimisation of a graphite-slag conductive composite applying a BAS-SVR based model.* Journal of Building Engineering, 2021. **44**: p. 103223.

[88] KAMAL AHMED, KUNIEDA MINORU, UEDA NAOSHI and NAKAMURA HIKARU, *Evaluation of crack opening performance of a repair material with strain hardening behavior.* Cement and Concrete Composites, 2008. **30**(10): p. 863-871 DOI: 10.1016/j.cemconcomp.2008.08.003.

[89] ZHOU YINGWU, XI BIN, YU KEQUAN, SUI LILI and XING FENG, *Mechanical properties of hybrid ultra-high performance engineered cementitous composites incorporating steel and polyethylene fibers.* Materials, 2018. **11**(8): p. 1448.

[90] SUN JUNBO, LIU SHUKUI, MA ZHANGUO, QIAN HAIMIN, WANG YUFEI, AL-AZZANI HISHAM, and WANG XIANGYU, *Mechanical properties prediction of lightweight coal gangue shotcrete.* Journal of Building Engineering, 2023. **80**: p. 108088.

[91] LI YAZHAO, GUAN XINCHUN, ZHANG CHENCHEN and LIU TIANAN, *Development of high-strength and high-ductility ECC with saturated multiple cracking based on the flaw effect of coarse river sand.* Journal of Materials in Civil Engineering, 2020. **32**(11): p. 04020317.

[92] LIU TIANAN, YANG YINGZI, CHEN ZHITAO, LI YAZHAO and BAI RUIXIANG, *Optimization of fiber volume fraction to enhance reinforcing*

efficiency in hybrid fiber reinforced strain hardening cementitious composite. Cement and Concrete Composites, 2020. **113**: p. 103704.

[93] XU LEI, SKOULARIDOU MARIA, CUESTA-INFANTE ALFREDO and VEERAMACHANENI KALYAN, *Modeling tabular data using conditional gan.* Advances in neural information processing systems, 2019. **32**.

[94] BOUROU STAVROULA, EL SAER ANDREAS, VELIVASSAKI TERPSICHORI-HELEN, VOULKIDIS ARTEMIS and ZAHARIADIS THEODORE, *A review of tabular data synthesis using GANs on an IDS dataset.* Information, 2021. **12**(09): p. 375.

[95] AWAD MARIETTE, KHANNA RAHUL, AWAD MARIETTE and KHANNA RAHUL, *Support vector regression.* Efficient learning machines: Theories, concepts, and applications for engineers and system designers, 2015: p. 67-80.

[96] PRETTENHOFER PETER and LOUPPE GILLES. *Gradient boosted regression trees in scikit-learn*. in *PyData 2014*. 2014.

[97] RIGATTI STEVEN J, *Random forest.* Journal of Insurance Medicine, 2017. **47**(1): p. 31-39.

[98] CHOU JUI-SHENG and TSAI CHIH-FONG, *Concrete compressive strength analysis using a combined classification and regression technique.* Automation in Construction, 2012. **24**: p. 52-60.

[99] NASSIF NADIA, AL-SADOON ZAID A, HAMAD KHALED and ALTOUBAT SALAH, *Cost-based optimization of shear capacity in fiber reinforced concrete beams using machine learning.* Struct Eng Mech, 2022. **83**(5): p. 671-680.

[100] BAYKASOĞLU ADIL, ÖZTAŞ AHMET and ÖZBAY ERDOĞAN, *Prediction and multi-objective optimization of high-strength concrete parameters via soft computing approaches.* Expert Systems with Applications, 2009. **36**(3): p. 6145-6155.

Appendix

Appendix 1. Statistical analysis of variables for the original 429 ECC stress data

Variables	Minimum	Maximum	Mean	Std Dev	Skewness	Kurtosis
Cement (%wt)	0.093	0.755	0.347	0.157	0.726	-0.142
Water (%wt)	0.029	0.230	0.142	0.045	-0.230	-0.916
Fine Agg (%wt)	0.0	0.541	0.228	0.119	-0.159	-0.115
Fly ash (%wt)	0.0	0.646	0.170	0.201	0.631	-1.144
Silica Fume (%wt)	0.0	0.246	0.048	0.054	0.887	0.172
blast furnace slag (%wt)	0.0	0.439	0.040	0.100	2.373	4.044
Superplasticiser (%wt)	0.0	0.028	0.009	0.008	0.897	-0.333
Fiber content (100%vol)	0.003	0.200	0.023	0.019	7.393	61.218
F. Diameter (micro-meter)	7.667	550.0	59.391	62.945	2.716	10.630
F. Length (mm)	2.061	38.0	14.191	5.660	1.065	1.707
F. Tensile Strength (MPa)	850.0	6360.0	2450.852	885.831	0.965	2.062
F. Elastic Modulus (GPa)	6.0	472.11	104.209	74.905	1.970	5.361
HPMC	0.0	1.0	0.090	0.280	2.910	6.610
Oiling Agent	0.0	1.0	0.261	0.440	1.092	-0.812
Temperature (Celsius)	0.0	500.0	35.634	46.911	5.590	41.424
Water Curing (days)	0.0	91.0	20.186	14.505	1.287	5.359
Air Curing (days)	0.0	112.0	5.375	13.061	4.110	24.052
Peak tensile stress (MPa)	1.8	33.4	7.360	4.575	1.834	5.505

Appendix 2. Statistical analysis of variables for the original 392 ECC strain data

Variables	Minimum	Maximum	Mean	Std Dev	Skewness	Kurtosis
Cement (%wt)	0.093	0.698	0.337	0.160	0.754	-0.136
Water (%wt)	0.029	0.230	0.145	0.044	-0.373	-0.792
Fine Agg (%wt)	0.0	0.541	0.231	0.123	-0.195	-0.283
Fly ash (%wt)	0.0	0.646	0.191	0.210	0.427	-1.445
Silica Fume (%wt)	0.0	0.246	0.042	0.052	0.970	0.269
blast furnace slag (%wt)	0.0	0.384	0.028	0.085	3.010	7.562
Superplasticiser (%wt)	0.0	0.028	0.009	0.008	0.882	-0.493
Fiber content (100%vol)	0.003	0.200	0.023	0.020	7.117	56.303
F. Diameter (micro-meter)	7.667	320.0	61.875	59.667	1.964	3.279
F. Length (mm)	2.061	38.0	13.643	5.927	1.245	1.805
F. Tensile Strength (MPa)	850.0	6360.0	2342.169	915.499	1.211	2.300
F. Elastic Modulus (GPa)	6.0	472.11	99.956	78.215	1.941	4.836
HPMC	0.0	1.0	0.107	0.306	2.576	4.691
Oiling Agent	0.0	1.0	0.283	0.451	0.966	-1.072
Temperature (Celsius)	0.0	500.0	35.872	48.463	5.517	39.609
Water Curing (days)	0.0	91.0	19.209	14.912	1.493	5.530
Air Curing (days)	0.0	112.0	7.168	17.291	3.467	13.527

Peak tensile strain (%)	0.017	17.3	3.030	2.619	1.530	3.095