Journal of Science and Transport Technology
Journal homepage: https://jstt.vn/index.php/en

Estimating the compressive strength of self-compacting concrete with fiber using an extreme gradient boosting model

Indra Prakash[1], Thanh-Nhan Phan[2], Hai-Van Thi Mai[2,*]
[1]DDG(R) Geological Survey of India, Gandhinagar, Gujarat 382010, India.
[2]University of Transport Technology, Hanoi 100000, Vietnam.

Article info
Type of article:
Original research paper

DOI:
https://doi.org/10.58845/jstt.utt.2023.en.3.1.12-25

***Corresponding author:**
E-mail address:
vanmth@utt.edu.vn

Received: 02/02/2023
Revised: 18/03/2023
Accepted: 20/3/2023

Abstract: Self-compacting concrete reinforced with fiber (SCCRF) is extensively utilized in the construction and transportation industries due to its numerous advantages, such as ease of building in challenging sites, noise reduction, enhanced tensile strength, bending strength, and decreased structural cracking. Traditional methods for assessing the compressive strength of SCCRF are generally time-consuming and expensive, necessitating the development of a model to forecast compressive strength. This research aimed to predict the CS of SCCRF using the Extreme Gradient Boosting (XGB) machine learning technique. The research uses the grid search method to optimize the XGB model's hyperparameters. A database of 387 samples is collected in this work, which is also an enormous dataset compared to those utilized in previous studies. An excellent result (R^2_{max} = 0.97798 for the testing dataset) proves that the proposed XGB model has excellent predictive power. Finally, Shapley Additive exPlanations (SHAP) analysis is conducted to understand the effect of each input variable on the predicted CS of SCCRF. The results show that the samples' age and cement content are the most critical factors affecting the CS. As a result, the proposed XGB model is a valuable tool for helping materials engineers have the right orientation in the design of SCCRF components to achieve the required compressive strength.

Keywords: Compressive strength (CS), Self-compacting concrete reinforced with fiber (SCCRF), Extreme Gradient Boosting (XGB).

1. Introduction

Self-compacting concrete reinforced with fiber (SCCRF) is a mixture of self-compacting cement concrete and fibers (such as carbon, steel, polypropylene (PP), polyester (PE), and glass) [1-3]. These fibers are tiny, short, randomly dispersed throughout the concrete, and make up around 1-3 percent of the overall volume. Depending on the qualities of various fibers, SCCRF possesses a variety of different outstanding advantages. Some research [4-8] implies that steel-reinforced self-compacting concrete will increase tensile and flexural strength, and decrease structural deformation. SCCRF using polymer fibers aids in reducing breaking and cracking and is inexpensive [9]. Moreover, SCCRF also has all the advantages of conventional SCC, including the ability to self-compact under its weight without needing a compaction mechanism, making it well-suited for projects with difficult construction sites. Due to the benefits mentioned earlier, SCCRF is a commonly utilized material in the construction and

transportation industries, particularly in challenging environments such as high-rise buildings, bridge girders, and road pavements. To efficiently employ SCCRF, it is vital to identify the material's mechanical and physical properties, in which compressive strength (CS) is an important attribute.

Experimental methods are often utilized to determine the CS of concrete and SCCRF in particular. However, the downsides of these approaches are the time-consuming casting of samples, the need for intensive testing equipment, and the results depending on the skill level of technicians [10],[11]. Moreover, the CS of SCCRF may be indirectly measured by the ultrasonic pulse velocity index [12-14]. The link between compressive strength and ultrasonic pulse velocity is quite sensitive, making the estimation of CS using this approach not particularly precise. Some variables, including the type of fiber, the type of cement, the ratio of water to cement, aggregate content, concrete age, and fiber content, impact its strength [12], [15]. Therefore, it is required to study and build a numerical simulation tool to predict SCCRF's CS quickly and cost-effectively.

In recent decades, machine learning (ML) approaches utilizing available experimental data to construct predictive models for material characteristics have been widely adopted [16-19]. However, the number of research employing ML models to predict the CS of SCCRF is limited [20], [21], with only seven studies [12],[20-25]. These investigations all utilize a relatively small quantity of data, with the most significant data set including 189 data [24]. In addition, the vast majority of research employs an artificial neural network (ANN) approach [26-31], and Support Vector Machine (SVM) [25]. While Extreme Gradient Boosting (XGB) is a powerful model [32] that has been applied to solve many complex problems, no research uses the XGB model to predict CS of SCCRF use. XGB is a supervised machine learning algorithm with many advantages, such as no need to normalize the database, can handle null data values, high execution speed, and easily handle big data sets. Due to the numerous hyperparameters, however, the XGB model is challenging to adjust. Overfitting can occur if the hyperparameters are not chosen appropriately.

The goal of this paper is to create a robust, high-precision model based on the XGB algorithm. A dataset of 387 samples was gathered to develop the XGB model. This is the biggest dataset of all accessible ML research on CS of SCCRF, according to the authors. In addition to the input database, the prediction performance of the XGB model depends on the model's selection of hyper-parameters. This study focuses on optimizing the hyperparameters of the XGB model by a grid search to find an optimal XGB predictive model. In addition, the effect of input parameters on the SCCRF's CS is studied using Shapley Additive exExplanations (SHAP) values technique.

2. The database utilized for research

The research used a large dataset, including 387 samples. All these data are collected from 11 international publications. [26],[30],[31],[33-40]. The database includes seventeen input parameters (from A_1 to A_{17}) and one output (Y). The names of the variables and their statistical analysis data are described in Table 1.

The majority of variables in the dataset have a wide distribution, including A_1, A_2, A_3, A_4, A_5, A_7, A_8, A_9, A_{14}, A_{15}, A_{16}, and A_{17}. Specifically, A_1 has a minimum value of 220 kg/m^3 and a maximum value of 754 kg/m^3. A_2 is dispersed between 0 and 1311.9 kg/m3, and A_3 is distributed between 0.83 and 1220 kg/m^3. The range of values for A_4 is mostly between 137 and 239 (kg/m^3), and the range for A_9 is between 0 and 288.9 kg/m^3. Finally, the output (Y) is mainly in the range of 30 to 95 MPa. Next, Fig. 1 illustrates the analysis results of the correlation between inputs and outputs. The level of correlation may be split based on the value of the Pearson correlation index (r_s). As observed, based on the values of r_s, the correlation between

the input variables and the output is rather low. Only several exceptional cases are determined, with a few high correlations between pairs of input variables, such as A_8 (Steel fiber) with A_2 (Coarse aggregate) and A_8 (Steel fiber) with A_{15} (Superplasticizer), respectively, r_s = 0.88 and 0.84. The interdependence between the input variables is low. Thus, in this study, all input variables contribute to the training and development of the machine learning model.

Table 1. Statistical analysis of the input and output parameters

Name	Symbol	Unit	Mean	Std	Min	25%	50%	75%	Max
Cement	A_1	kg/m^3	416.633	99.98	220	363.5	405	440	754
Coarse aggregate	A_2	kg/m^3	725.978	203.36	0	722	772	800	1311.9
Fine aggregate	A_3	kg/m^3	909.838	129.25	0.83	826	932	955	1220
Water	A_4	kg/m^3	172.405	22.40	137.2	158	162	191.5	239
Fly ash	A_5	kg/m^3	40.413	86.46	0	0	0	0	306
Glass fiber	A_6	kg/m^3	0.632	1.74	0	0	0	0	7.95
Polypropylene fiber	A_7	kg/m^3	1.413	2.79	0	0	0	1.4	12
Steel fiber	A_8	kg/m^3	13.594	38.29	0	0	0	0	156
Limestone	A_9	kg/m^3	101.406	136.31	0	0	0	288.9	288.9
Basalt powder	A_{10}	kg/m^3	0.853	10.44	0	0	0	0	165
Marble powder	A_{11}	kg/m^3	0.853	10.44	0	0	0	0	165
Nano silica	A_{12}	kg/m^3	10.473	19.28	0	0	0	16.5	90
Nano CuO	A_{13}	kg/m^3	0.571	2.41	0	0	0	0	13.8
Metakaolin	A_{14}	kg/m^3	2.791	13.30	0	0	0	0	90
Superplasticizer	A_{15}	kg/m^3	8.785	6.92	0	4.5	7	9.18	33
Viscosity modifying admixture	A_{16}	l/m^3	0.166	0.28	0	0	0	0.42	0.9
Age of samples	A_{17}	day	36.566	28.83	1	28	28	28	90
Compressive strength	Y	MPa	65.919	20.25	28.24	53.835	65.61	77.475	159.91

Std=Standard deviation

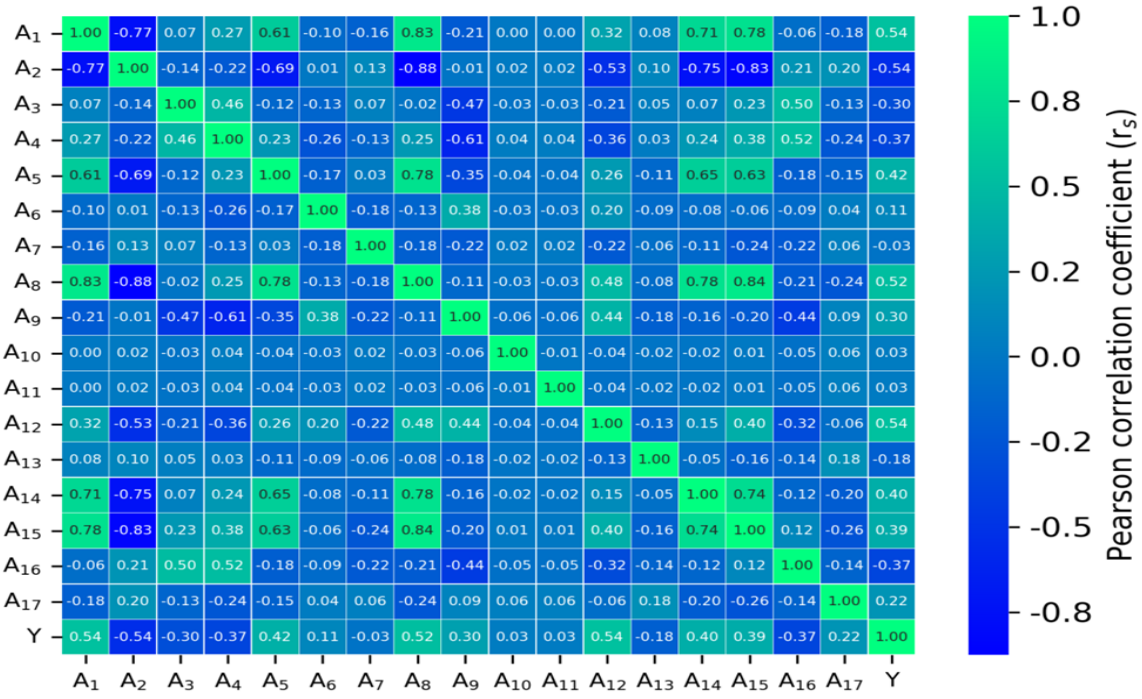

Fig. 1. Correlation matrix between variables.

3. Methods

3.1. Extreme Gradient Boosting (XGB)

The XGB algorithm was developed from the GBM algorithm and added by Chen, Tianqi, and Tong He [41]. The advantage of XGB is its ability to efficiently build boost trees that work in tandem with each other. XGB is applied to both regression and classification problems. The essence of this algorithm is to optimize the value of the objective function and do it based on the slope enhancement framework. Thanks to parallel reinforcement trees, XGB can solve complex problems quickly, flexibly, and accurately.

3.2. Grid search for Hyperparameter optimization

Machine learning models can be used in different fields with different data sets. The machine model's hyperparameters must be adjusted to be suitable for different problems. These values can influence model training, so tuning the hyperparameters to improve the prediction performance is essential. The essence of this process is that the important hyperparameters of the model are initialized and optimized until the suggested objective function reaches the minimum or maximum values [42]. Some commonly used hyperparameter optimization methods include grid search, random search, sequential hyperparameter optimization, etc. The hyperparameters of the XGB model are optimized in this study using the grid search approach.

The predictive performance of the XGB model is dependent on many hyperparameters, of which a group of important hyperparameters is chosen for optimization. The value of the important hyperparameters is changed in a certain range called meshes, the remaining hyperparameters of the model take the default value. This approach exhaustively investigates all parameter combinations by looking for meshes in the multidimensional domain iteratively across the whole sample size. To identify which combination yields the highest accuracy, all combinations are evaluated. The evaluation, in this study, is based on the coefficient of determination (R^2) and standard deviation (Std) criteria. These values are calculated by averaging the results of 5 cross-validations (CV) to evaluate the trained model.

3.3. Evaluate the model's predictive performance

The predictive capability of the machine learning model was assessed using four statistical metrics (R^2, MAE, RMSE, and MAPE). Where R^2 is the coefficient of determination, RMSE stands for root Means square error, MAE stands for Mean absolute error, and MAPE stands for Mean absolute percentage error. The more accurate the model, the higher R^2, the lower the MAE, RMSE, and MAPE, and vice versa. R^2 value ranges from 0 to 1. The model is ideal when R^2 = 1. Following is how these four indicators are determined:

$$MAE=\frac{1}{n}\sum_{i=1}^{n}\left|q_i - q'_i\right| \quad (1)$$

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(q - q'_i)^2} \quad (2)$$

$$R^2 =1-\frac{\sum_{i=1}^{n}(q_i - q'_i)^2}{\sum_{i=1}^{n}(q_i - \bar{q})^2} \quad (3)$$

$$MAPE=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{q_i - q'_i}{q_i}\right| \times 100\% \quad (4)$$

where n infers the number of samples, q_i and q'_i are the actual and predicted outputs, respectively, and $\bar{q}$ is the average value of the q_i

3.4. Shapley Additive exExplanations

SHAP is a frequently employed approach in the field of machine learning for interpreting the model's predictions and determining the effect of input parameters on output parameters. Shapley created Shap value from a game theory perspective [43]. Where the dataset's feature values act as coalition members. This technique predicts how each attribute contributes to the projected value and explains the forecast. SHAP

values are the numerical values allocated to each player in every possible player combination. It assigns a value to each predictor for the regression issue based on all potential predictive models. This technique enables the SHAP value to produce a prediction result that is more similar to the real model. It should be noted that SHAP analysis is only one of the approaches to understanding the impact of input features on a model's output. It is a model-agnostic method with a solid theoretical foundation that provides local and global explanations.

4. Methodology flowchart

The best XGB model for predicting CS of SCCRF is developed by the following four steps: (1) Data collection, (2) Training model and optimizing the model's hyperparameters, (3) Testing model, (4) Evaluating the effect of input parameters. The detailed step-by-step is as follows:

Step 1: Data collection

The database includes 387 experimental results, collected from 11 documents. The database is randomly divided into two sets: the training dataset for model training (accounting for 70%) and the test dataset for model testing (accounting for 30%).

The dataset consists of 387 samples divided into 2 sets. The training dataset has 271 samples and the test dataset has 116 samples. The random separation of the dataset into two distinct parts enables the most objective and precise evaluation of the prediction ability of the machine learning model because the testing dataset is completely unknown to the ML model during the training part. In addition, all data values are standardized to a range of 0 to 1 during model construction to reduce simulation-generated errors.

Step 2: Training model

In this step, the XGB algorithm is trained using the training dataset. Using grid search, the hyperparameters of the XGB model are tuned. After refining the parameters, the predictive performance of the models is assessed and compared in order to select the XGB model with the most significant predictive performance. In order to avoid overfitting and enhance the prediction ability of the machine learning model, the 5-fold cross-validation approach is implemented during model construction. This step is repeated until the models are successfully trained (the tolerance criterion is met).

Step 3: Testing model

After step 2, the aforementioned optimal XGB models are evaluated with the test dataset. Four statistical metrics were used to evaluate the prediction accuracy of the model. The significance of these statistics is presented in section 3.3. Based on the acquired evaluation index values, the XGB_1 model is chosen as the best model. In the following stage, this model is used to predict the CS of SCCRF and analyze the effect of input parameters on CS.

Step 4: Evaluating the effect of input parameters

The recommended XGB model is used in the last step to assess the impact of input parameters on the CS of SCCRF using the Shap value method. The detailed methodological chat of the study is presented in Fig. 2.

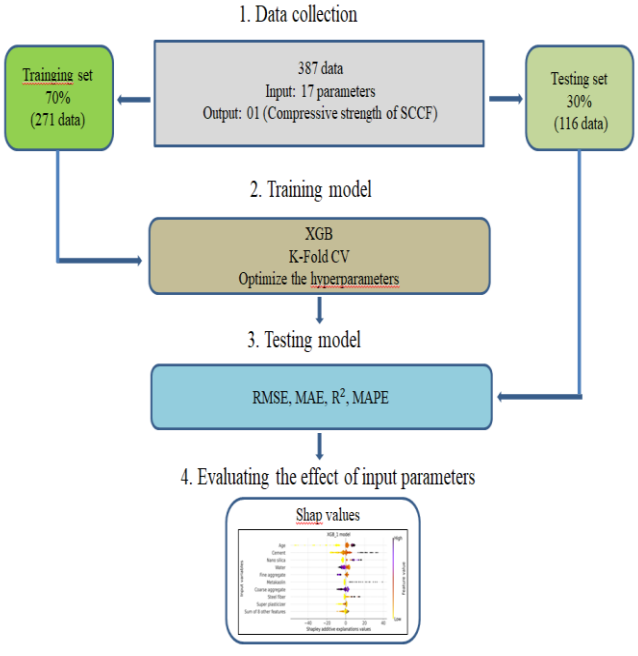

Fig. 2. Methodology flowchart

5. Results of research

The predictive ability of a machine learning model depends on various elements, two of which are the selection of a suitable algorithm and the construction of an exemplary model architecture. The XGB algorithm is chosen for investigation. The model's hyperparameters are optimized using the grid search technique to find an efficient XGB model structure for SCCRF's CS prediction. The grid search method compares all the XGB model structures generated by parameter combinations. Next, the predictability of the five best XGB models is compared using the four criteria for 3 sets (training dataset, validation dataset, and testing dataset). The XGB model with the most optimum hyperparameters and the highest predictive performance is chosen to forecast the CS of SCCRF and the impact of the input parameters.

5.1. Grid search for optimizing the parameters of XGB model

This section describes the procedure for optimizing the hyperparameters of the XGB model. The predictive performance of the XGB model is contingent upon many parameters, some of which are significant hyperparameters, such as 'max_depth', 'n_estimators', 'learing_rate', 'min_child_weight', and 'subsample'. Where 'max_depth' is the maximum depth of the individual regression estimators. 'n_estimators' is the number of boosting stages to perform. 'learing_rate' shrinks the contribution of each tree. 'min_child_weight' is the minimum number of samples required to split an internal node and 'subsample' is the fraction of samples to be used for fitting the individual base learners. Important hyperparameters are parameters whose value modifications substantially impact the predictive performance of the model. Hyperparameters are optimized by selecting meshes to survey for important parameters. The grid values chosen for the survey are shown in Table 2.

Table 2. The selected hyperparameters of XGB model and studied values

Hyperparameters	Grid values
max_depth	(3, 4, 5, 6, 7)
n_estimators	(100, 200, 400, 600,800,1000, 1400, 1600)
learing_rate	(0.1, 0.3, 0.4)
min_child_weight	(1, 3, 5)
subsample	(0.1, 0.5, 1)

The remaining parameters that have little effect on model performance are given default values. The survey findings acquired the R^2, standard deviation of all XGB models for loops, in decreasing order, such as 0.96341282, 0.96341282, 0.96341282,..., 0.70417289, 0.70417289, 0.70417289 (with a total of 1080 models). The results also show that the above values correspond to 1048, 328, 688,... 48, 408, and 768 runs. Specifically, R^2 = 0.96341282 corresponds to the 1048^{th} simulation, and R^2 = 0.70417289 corresponds to the 768^{th} simulation. Simultaneously, the output is an Excel file including all models corresponding to 1080 simulations, all parameters related to those models, the mean_test_scores (R^2) and std_test_score, rank test score... for all 1080 models. Table 2 displays only the five models with the highest rank test score.

Table 3. The results of using Grid search method for optimizing the parameters of the XGB model

Index	learing_rate	max_depth	min_child_ weight	n_estimators	subsample	Mean_ test_score	Std_test_sc ore	Rank_ test_score	Model
112	0.1	4	3	1000	0.5	0.96331	0.01501	4	XGB_4
328	0.1	7	3	1000	0.5	0.96341	0.01477	1	XGB_1
472	0.3	4	3	1000	0.5	0.96331	0.01501	4	XGB_5
688	0.3	7	3	1000	0.5	0.96341	0.01477	1	XGB_2
1048	0.4	7	3	1000	0.5	0.96341	0.01477	1	XGB_3

The results show that, based on R2 and std of the validation data set, it is possible to identify 5 XGB models with the best predictive power, namely XGB_1, XGB_2, XGB_3, XGB_4, XGB_5 with corresponding hyperparameters (Table 3). In order to select an optimal XGB model, the forecasting performance of these five models will be compared using four metrics: MAE, RMSE, R2, and MAE during the training, validation, and testing stages.

5.2. Comparison of predictive ability of XGB models

In the preceding section, the evaluation of the predictive performance of the new XGB models is based solely on the validation dataset's R^2 value. Consequently, the five models with the highest predictability are chosen, namely XGB_1, XGB_2, XGB_3, XGB_4, and XGB_5. Next, the forecast performance of these five models is compared using all four metrics over the training, validation, and testing stages.

Fig. 3 demonstrates that the predictive performance of all five models is excellent and nearly identical, as evidenced by the modest errors of MAE, RMSE, and MAPE and the high R^2 ($R^2_{testing}$ > 0.976). All five models are optimized for hyperparameters and chosen as the ideal models in the preceding phase, which explains why the findings are excellent. However, three models, namely XGB_1, XGB_2, and XGB_3, are more accurate than the two models XGB_4 and XGB_5 (see Table 4). In all three stages of training, validation, and testing, models XGB_1, XGB_2, and XGB_3 have smaller RMSE, MAE, MAPE values, and higher R^2 than the other 2 models. Therefore, it is possible to choose 1 out of 3 models, XGB_1, XGB_2, and XGB_3, which is the most optimal for predicting the CS of SCCRF. Here, the proposed XGB_1 model is SCCRF's CS predictive model with outstanding predictive results ($RMSE_{testing}$ = 3.22784 MPa, $MAE_{testing}$ = 2.19278 MPa, $R^2_{testing}$ = 0.97798 and $MAPE_{testing}$ = 0.03486). In the following part, detailed predicted results from model XGB_1 are presented.

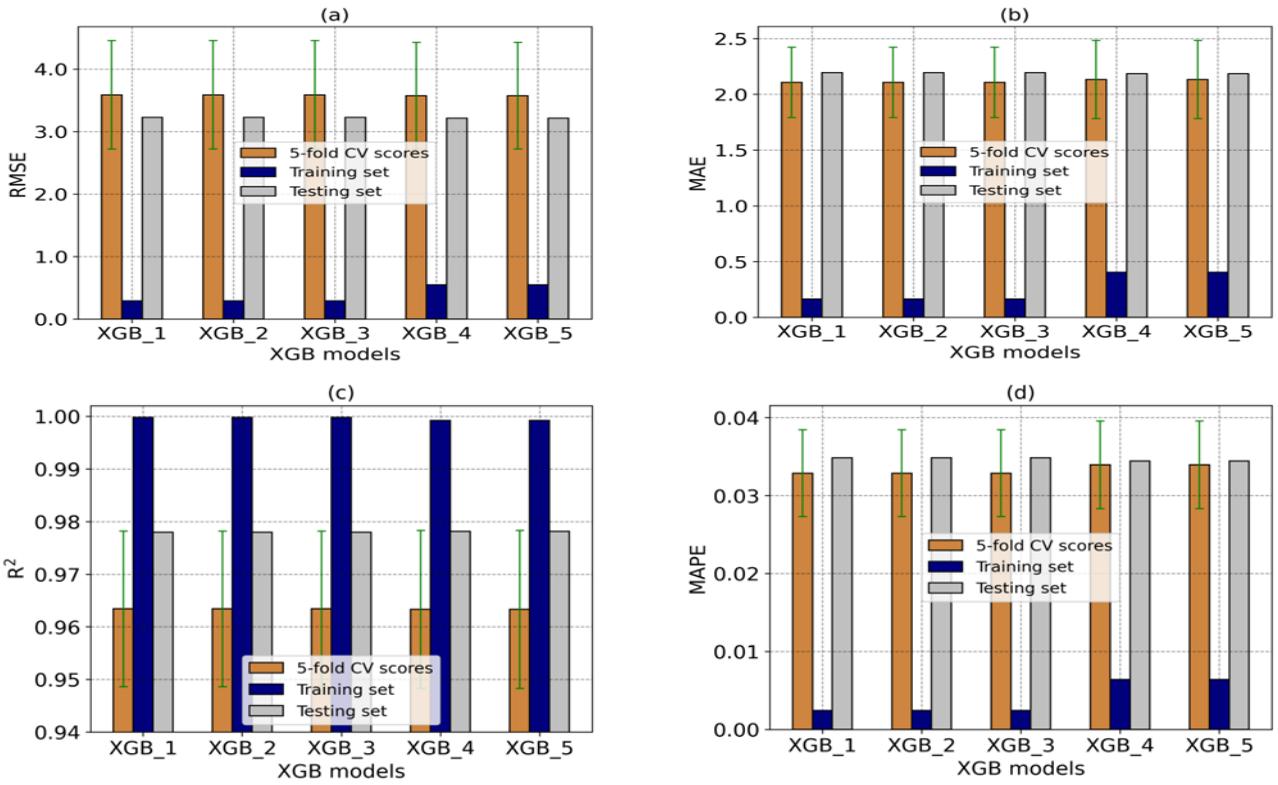

Fig. 3. Compare the predictive ability of 5 models that have been optimized for parameters: XGB_1, XGB_2, XGB_3, XGB_4, and XGB_5

Table 4. The prediction results of five XGB models

	Metrics	XGB_1	XGB_2	XGB_3	XGB_4	XGB_5
Validation	RMSE	3.58562	3.58562	3.58562	3.57254	3.57254
	MAE	2.10675	2.10675	2.10675	2.13313	2.13313
	R^2	0.96341	0.96341	0.96341	0.96331	0.96331
	MAPE	0.03287	0.03287	0.03287	0.03394	0.03394
Training	RMSE	0.29037	0.29037	0.29037	0.54700	0.54700
	MAE	0.16544	0.16544	0.16544	0.40509	0.40509
	R^2	0.99978	0.99978	0.99978	0.99922	0.99922
	MAPE	0.00243	0.00243	0.00243	0.00642	0.00642
Testing	RMSE	3.22784	3.22784	3.22784	3.23459	3.23459
	MAE	2.19278	2.19278	2.19278	2.19447	2.19447
	R^2	0.97798	0.97798	0.97798	0.97616	0.97616
	MAPE	0.03486	0.03486	0.03486	0.03545	0.03545

5.3. Representative results

In this part, to evaluate the effect of optimizing hyperparameters of the XGB model, the prediction results of two models, XGB_1 (the best model) and XGB_0 (model with default parameters), are compared using the regression chart (Fig 4). The results demonstrate that the predictive power of the two models at the training stage is roughly equivalent (Figs 4a, c), with models XGB_0 and XGB_1 expressing $R^2_{training}$ = 0.99999 and $R^2_{training}$ = 0.99978, respectively. The prediction abilities of the two models are different during the test stage, though. The XGB_1 model has more predictive ability than the XGB_0 model, as seen by its $R^2_{testing}$ = 0.97798, which is higher than that of the XGB_ 0 model ($R^2_{testing}$ =0.95383). Moreover, for the test data set of model XGB_1, the errors RMSE, MAE, and MAPE are all lower than those of model XGB_0 for the testing dataset. Thus, once again, it can be confirmed that optimizing the parameters of the XGB model by the grid search method is effective and that XGB_1 is the model with the best predictive ability.

It can be seen that model XGB_1's training ability is nearly ideal, with $R^2_{training}$ = 0.99978. At the testing stage (Fig. 4b), the majority of samples exhibit extremely close predicted results to the regression line (test R^2 = 0.97798), demonstrating the XGB_1 model's excellent predictive ability.

Fig. 5 illustrates the error between the actual and expected CS values based on the ratio of targets to outputs (targets/outputs). The closer this ratio is to 1, the more precise the XGB_1 predictive model is, and vice versa. At the training stage (Fig. 5a), the majority of samples have a target/outputs ratio of 1, indicating that the predicted outcomes closely match the actual results. The majority of samples at the test stage (Fig. 5b) have a target/outputs ratio of 1. Nonetheless, there are a few instances with targets/outputs ratios different than 1, but only between 0.85 and 1.15. This error is entirely acceptable. In addition, the number of erroneous samples is negligible compared to the total of 116 samples in the test data set. It is important to note that although the training performance is excellent, the testing performance, as indicated by the R^2 value, is also high at 0.97798. Along with lower RMSE, MAE and MAPE values, this high R^2 value suggests that the model is generalizing well to the test dataset and is not merely memorizing the training data. In conclusion, the high R^2 and low RMSE, MAE and MAPE values on the testing dataset indicate that the model is likely generalizing well and not overfitting to the training data. Therefore, the optimal model

XGB_1's prediction findings are credible.

In addition, the difference between the actual compressive strength and the strength predicted by the XGB_1 model for each sample is depicted in Fig. 6 for the training set (Fig. 6a) and the testing set (Fig. 6b). The graph's vertical axis shows the sample count, while the graph's horizontal axis shows the error value. The smaller the error number, the closer the predicted CS value is to the actual CS value, or the greater the predictive accuracy of the XGB_1 model. Most samples exhibit minimal errors, specifically [-0.5 ÷ 0.5] MPa for the training set (Fig. 6a) and [-5 ÷ 5] MPa for the testing set (Fig. 6b). Therefore, the proposed XGB_1 model is an excellent predictor for CS of SCCRF.

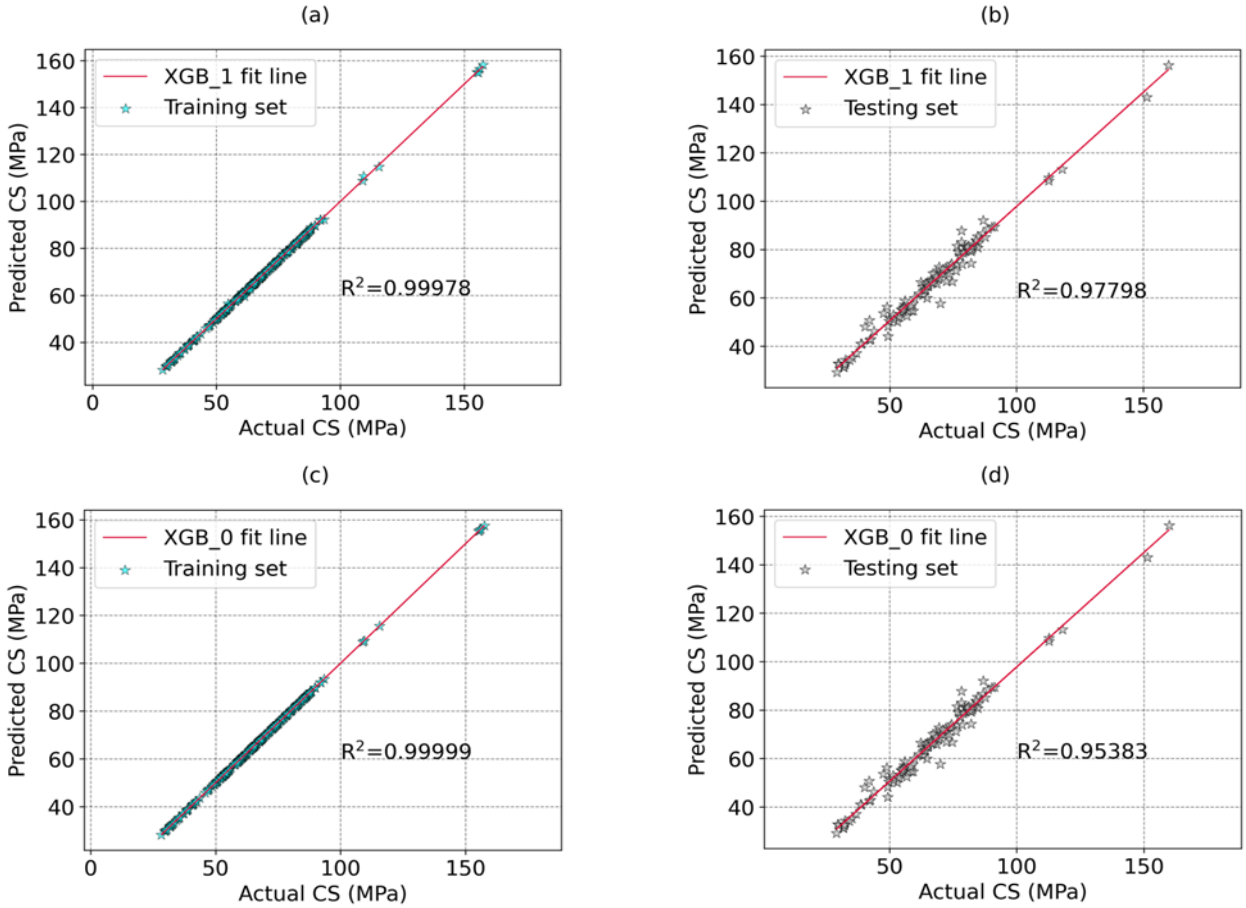

Fig. 4. Regression plots for the representative results of the XGB_1 (best model) and XGB_0 (default XGB model) for: (a,c) training dataset, (b,d) testing dataset

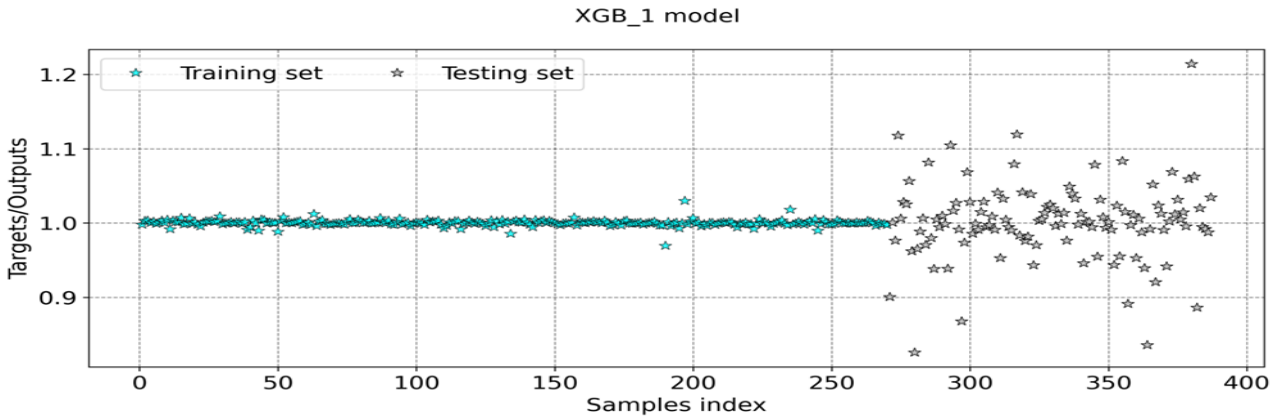

Fig. 5. Comparison between targets and outputs of the XGB_1 model

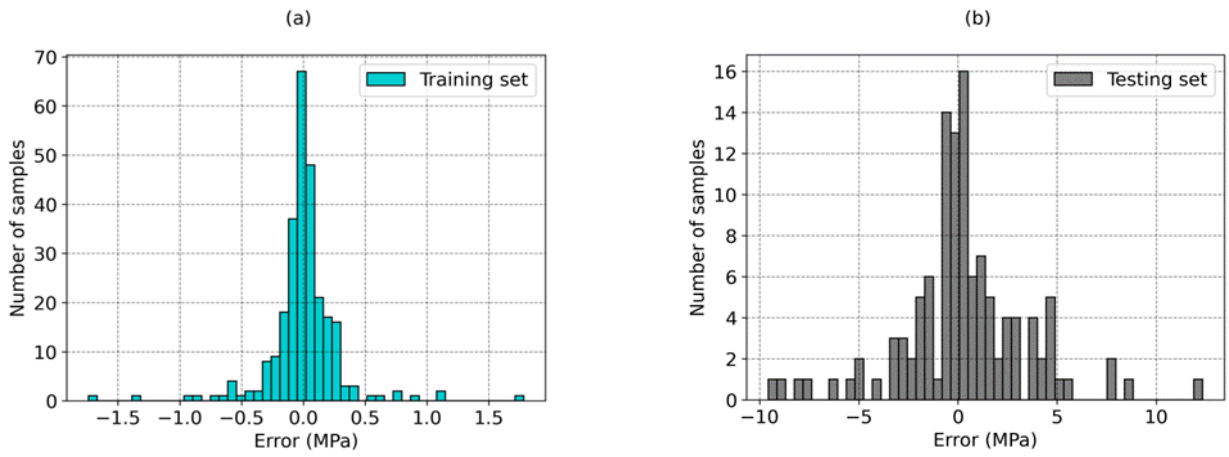

Fig. 6. Comparison between predicted and experimental compressive strength of the representative results for the XGB_1 model

5.4. Result of SHAP analysis

This section evaluates and simulates the importance and impact of input factors on compressive strength using the SHAP value (Fig. 7). The names of the input parameters are represented on the vertical axis, and the effect level decreases from top to bottom. The input variables' Shap values are shown on the horizontal axis; the higher and more positive the Shap value, the more an input parameter will impact an output parameter. Conversely, the more negatively that parameter affects the output parameter, the smaller the Shap value. As the color changes from yellow to black, it shows that the input parameters positively impact CS.

As observed in Fig. 7, the Age of samples has the greatest impact on CS. Cement, Nano silica, Water, Fine Aggregate, Metakaolin, Coarse Aggregate, Steel Fiber, and Superplasticizer are the next input factors that have an impact on CS, listed in descending order. The remaining 8 variables have little effect on CS. In addition, the findings of Shap value analysis also show that the input parameters positively impact CS, including the Age of samples, Cement, Nano silica, and Steel fiber. In other words, the CS increases as the age of sample, cement content, nano silica content, and steel fiber content increase, and vice versa. This finding is completely consistent with the experimental results found in the study [44]. On the contrary, the input parameters negatively influence the CS of SCCRF, such as Water, Fine aggregate. That is, when the content of Water, Fine aggregate increases, the CS of SCCRF will decrease. Several studies [45], [46] have reached similar conclusions. Moreover, the Shap values plot shows that the remaining input parameters have a complex impact on the SCCRF's CS.

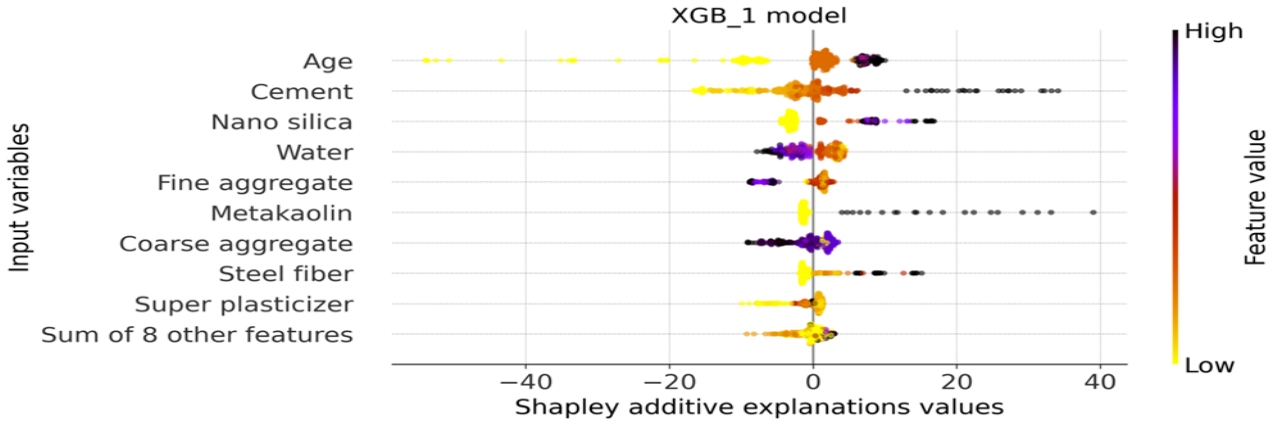

Fig. 7. Shap values analysis of the input variables

6. Conclusions and discussion

Machine learning models, of which XGB is a potent model, have been widely used in the fields of transportation and construction in recent decades, particularly in predicting the mechanical properties of concrete. This study's goal is to create the best XGB model with the right hyperparameters for predicting the CS of SCCRF. Research results indicate:

- According to the author's understanding, with 387 gathered samples, this is the biggest data set compared to other research that uses machine learning models to predict SCCRF's CS

- The prediction results of five models XGB_1, XGB_2, XGB_3, XGB_4, and XGB_5, are highly accurate and almost identical, as evidenced by small MAE, RMSE, and MAPE errors and a high R^2 ($R^2_{testing}$ > 0.976), indicating that the optimization of hyperparameters of the XGB model by grid search is highly effective.

- As evidenced by the $RMSE_{testing}$=3.22784 MPa, $MAE_{testing}$ =2.19278 MPa, $R^2_{testing}$= 0.97798 và $MAPE_{testing}$=0.03486, the XGB_1 model can accurately estimate the CS of SCCRF from its actual data. Therefore, XGB_1 is proposed as SCCRF's CS predictive model quickly, with high reliability, saving cost and time.

- Age of samples had the highest impact on SCCRF compressive strength estimation, followed by Cement, Nano silica, Water, Fine aggregate, Metakaolin, Coarse aggregate, Steel fiber, and Superplasticizer, and the remaining 8 variables have little influence on the compressive strength, as depicted by SHAP analysis.

- The SHAP plot showed that Age, Cement, Nano silica, and Steel fiber positively influence SCCRF compressive strength. Water, Fine aggregate negatively influence the CS of SCCRF. The remaining parameters have a complex influence.

This work also exhibits several limitations that need to be discussed for further studies. First, grid search may not be the most efficient method when dealing with a large grid limit, especially for models like XGB with important hyperparameters such as n_estimators and learning_rate that can vary in a wide range with fine steps. Instead of using grid search, more efficient optimization algorithms can be employed for hyperparameter tuning in future research, such as: (i) Random search: Rather than exhaustively searching the entire parameter space, random search samples a random subset of hyperparameter combinations, leading to a faster and more efficient search process, (ii) Discrete optimization algorithms: Genetic algorithms (GA) and particle swarm optimization (PSO) are examples of metaheuristic optimization techniques that can efficiently explore the search space, making them suitable alternatives for hyperparameter tuning when dealing with large grid limits. Second, 5-fold cross-validation is considered sufficient for this study because of its many benefits, such as reduced computational cost, adequate model evaluation, widely used. However, the use of 10-fold cross-validation should be considered in future studies to further confirm the objectivity of the optimal results of the grid search method and ensure a more comprehensive evaluation of the model.

Data Availability Statement: The raw/processed data required to reproduce these findings cannot be shared at this time as the data also forms part of an ongoing study.

Funding: This research was funded by the University of Transport Technology, Thanh Xuan, Hanoi, Vietnam (UTT), under grant number DTTD 2022-07.

References

[1] A. Boz, A. Sezer, T. Ozdemir, G. HIZAL, and O. DOLMACI. (2018). Mechanical properties of lime-treated clay reinforced with different types of randomly distributed fibers. *Arabian Journal of Geosciences*, 11, 122.

[2] S. Fallah and M. Nematzadeh. (2017). Mechanical properties and durability of high-

strength concrete containing macro-polymeric and polypropylene fibers with nano-silica and silica fume. *Construction and Building materials*, 132, 170-187.

[3] A.C. Bhogayata and N.K. Arora. (2017). Fresh and strength properties of concrete reinforced with metalized plastic waste fibers. *Construction and Building Materials*, 146, 455-463.

[4] R.F. Zollo. (1997). Fiber-reinforced concrete: an overview after 30 years of development. *Cement and Concrete Composites*, 19(2), 107-122,

[5] J. Gao, W. Sun, and K. Morino. (1997). Mechanical properties of steel fiber-reinforced, high-strength, lightweight concrete. *Cement and Concrete Composites*, 19(4), 307-313.

[6] C.X. Qian and P. Stroeven. (2000). Development of hybrid polypropylene-steel fibre-reinforced concrete. *Cement and Concrete Research*, 30(1), 63-69.

[7] P.S. Song and S. Hwang. (2004). Mechanical properties of high-strength steel fiber-reinforced concrete. *Construction and Building Materials*, 18(9), 669-673.

[8] P. Balaguru and H. Najm. (2004). High-Performance Fiber-Reinforced Concrete Mixture Proportions with High Fiber Volume Fractions. *Materials Journal*, 101(4), 281-286.

[9] M. Hsie, C. Tu, and P.S. Song. (2008). Mechanical properties of polypropylene hybrid fiber-reinforced concrete. *Materials Science and Engineering: A*, 494(1), 153-157.

[10] T.T. Nguyen, T.N. Dao, S. Aaleti, K. Hossain, and K.J. Fridley. (2019). Numerical Model for Creep Behavior of Axially Loaded CLT Panels. *Journal of Structural Engineering*, 145(1), 04018224.

[11] H.V. Long. (2017). A sampling method for investigating self-healing property of oncrete damaged by the drying shrinkage. *Journal of Science and Technology in Civil Engineering (STCE) - HUCE*, 11(6), 97-103.

[12] H. Mashhadban, S. Soleimani Kutanaei, and M. Sayarinejad. (2016). Prediction and modeling of mechanical properties in fiber reinforced self-compacting concrete using particle swarm optimization algorithm and artificial neural network. *Construction and Building Materials*, 119, 277-287.

[13] V. Revilla-Cuesta, M. Skaf, R. Serrano-López, and V. Ortega-López. (2021). Models for compressive strength estimation through non-destructive testing of highly self-compacting concrete containing recycled concrete aggregate and slag-based binder. *Construction and Building Materials*, 280, 122454.

[14] M. Abed and J. de Brito. (2020). Evaluation of high-performance self-compacting concrete using alternative materials and exposed to elevated temperatures by non-destructive testing. *Journal of Building Engineering*, 32, 101720.

[15] A. Sadeghi Nik and O. Lotfi Omran. (2013). Estimation of compressive strength of self-compacted concrete with fibers consisting nano-SiO2 using ultrasonic pulse velocity. *Construction and Building Materials*, 44, 654-662.

[16] M. Xue and C. Zhu. (2009). A study and application on machine learning of artificial intellligence. *2009 International Joint Conference on Artificial Intelligence*, pp 272-274.

[17] D. Mhlanga. (2021). Financial inclusion in emerging economies: The application of machine learning and artificial intelligence in credit risk assessment. *International Journal of Financial Studies*, 9(3), 39.

[18] S. Guikema. (2020). Artificial intelligence for natural hazards risk analysis: Potential, challenges, and research needs. *Risk Analysis*, 40(6), 1117-1123.

[19] A. Alimadadi, S. Aryal, I. Manandhar, P.B. Munroe, B. Joe, and X. Cheng. (2020). Artificial

intelligence and machine learning to fight COVID-19. *Physiological genomics*, 52(4). American Physiological Society Bethesda, MD, pp. 200-202.

[20] T.T. Nguyen, H.D. Pham, T.T. Pham, and H.H. Vu. (2020). Compressive Strength Evaluation of Fiber-Reinforced High-Strength Self-Compacting Concrete with Artificial Intelligence. *Advances in Civil Engineering*, 6, 1-12, e3012139.

[21] L.V.P. Meesaraganda, P. Saha, and N. Tarafder. (2019). Artificial Neural Network for Strength Prediction of Fibers' Self-compacting Concrete. *Soft Computing for Problem Solving*, *Springer Singapore*, pp 15-24.

[22] H. Tavakoli, O. Lotfi-Omran, M. Shiade, and S. Soleimani Kutanaei. (2014). Prediction of combined effects of fibers and nano-silica on the mechanical properties of self-compacting concrete using artificial neural network. *Latin American Journal of Solids and Structures*, 11(11), 1906-1923.

[23] P. Saha, P. M.l.v, and P. R. Kumar. (2017). Predicting strength of SCC using artificial neural network and multivariable regression analysis. *Computers and Concrete*, 20(1), 31-38.

[24] M. Uysal, H. Tanyildizi. (2012). Estimation of compressive strength of self compacting concrete containing polypropylene fiber and mineral additives exposed to high temperature using artificial neural network. *Construction and Building Materials*, 27(1), 404-414.

[25] F. Naseri, F. Jafari, E. Mohseni, W. Tang, A. Feizbakhsh, and M. Khatibinia. (2017). Experimental observations and SVM-based prediction of properties of polypropylene fibres reinforced self-compacting composites incorporating nano-CuO. *Construction and Building Materials*, 143, 589-598.

[26] H. Mashhadban, S.S. Kutanaei, and M.A. Sayarinejad. (2016). Prediction and modeling of mechanical properties in fiber reinforced self-compacting concrete using particle swarm optimization algorithm and artificial neural network. *Construction and Building Materials*, 119, 277-287.

[27] T.T. Nguyen, H.D. Pham, T.T. Pham, and H.H. Vu. (2020). Compressive Strength Evaluation of Fiber-Reinforced High-Strength Self-Compacting Concrete with Artificial Intelligence. *Advances in Civil Engineering*, 6, e3012139.

[28] L.V.P. Meesaraganda, P. Saha, and N. Tarafder. (2019). Artificial Neural Network for Strength Prediction of Fibers' Self-compacting Concrete. *Soft Computing for Problem Solving*, *Springer Singapore*, 15-24.

[29] H. Tavakoli, O. Lotfi-Omran, M. Shiade, and S. Soleimani Kutanaei. (2014). Prediction of combined effects of fibers and nano-silica on the mechanical properties of self-compacting concrete using artificial neural network. *Latin American Journal of Solids and Structures*, 11(11), 1906-1923.

[30] P. Saha, P. M.l.v, and P.R. Kumar. (2017). Predicting strength of SCC using artificial neural network and multivariable regression analysis. *Computers and Concrete*, 20(1), 31-38.

[31] M. Uysal, H. Tanyildizi. (2012). Estimation of compressive strength of self compacting concrete containing polypropylene fiber and mineral additives exposed to high temperature using artificial neural network. *Construction and Building Materials*, 27(1), 404-414.

[32] D. Nielsen. (2016). Tree Boosting With XGBoost - Why Does XGBoost Win 'Every' Machine Learning Competition? *Master thesis, NTNU*. https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/2433761

[33] F. Naseri, F. Jafari, E. Mohseni, W. Tang, A. Feizbakhsh, and M. Khatibinia. (2017). Experimental observations and SVM-based prediction of properties of polypropylene fibres reinforced self-compacting composites

incorporating nano-CuO. *Construction and Building Materials*, 143, 589-598.

[34] N. Majain, A. B. A. Rahman, R. N. Mohamed, and A. Adnan. (2019). Effect of steel fibers on self-compacting concrete slump flow and compressive strength. *IOP Conference Series: Materials Science and Engineering*, 513(1), 012007.

[35] S.J. Begum and P.J.D. Anjaneyulu. (2018). A Study on Effect of Steel Fiber in Fly Ash Based Self-Compacting Concrete. International Journal for Innovative Research in Science & Technology, 5(1), 95-99.

[36] M.H. Beigi, J. Berenjian, O.L. Omran, A.S. Nik, and I.M. Nikbin. (2013). An experimental survey on combined effects of fibers and nanosilica on the mechanical, rheological, and durability properties of self-compacting concrete. *Materials & Design*, 50, 1019-1029.

[37] G. Pons, M. Mouret, M. Alcantara and J.L. Granju. (2007). Mechanical behaviour of self-compacting concrete with hybrid fibre reinforcement. *Materials and Structures*, 40(2), 201-210.

[38] Q. Song, R. Yu, X. Wang, S. Rao and Z. Shui. (2018). A novel Self-Compacting Ultra-High Performance Fibre Reinforced Concrete (SCUHPFRC) derived from compounded high-active powders. *Construction and Building Materials*, 158, 883-893.

[39] O. Gencel, W. Brostow, T. Datashvili and M. Thedford. (2011). Workability and Mechanical Performance of Steel Fiber-Reinforced Self-Compacting Concrete with Fly Ash. *Composite Interfaces*, 18(2), 169-184.

[40] O. Gencel, C. Ozel, F. Koksal, G. Martínez-Barrera, W. Brostow, and H. Polat. (2013). Fuzzy Logic Model for Prediction of Properties of Fiber Reinforced Self-compacting Concrete. *Materials Science*, 19(2).

[41] P. Carmona, F. Climent, and A. Momparler. (2019). Predicting failure in the U.S. banking sector: An extreme gradient boosting approach, *International Review of Economics & Finance*, 61, 304-323.

[42] S. Sun, Z. Cao, H. Zhu, and J. Zhao. (2020). A Survey of Optimization Methods from a Machine Learning Perspective. *IEEE Transactions on Cybernetics*, 50(8), 3668-3681.

[43] L.S. Shapley. (1952). A Value for N-Person Games. *RAND Corporation*. https://www.rand.org/pubs/papers/P295.html

[44] B.S. Mohammed and N.J. Azmi. (2014). Strength reduction factors for structural rubbercrete. *Frontiers of Structural and Civil Engineering*, 8(3), 270-281.

[45] A. Oner and S. Akyuz. (2007). An experimental study on optimum usage of GGBS for the compressive strength of concrete. *Cement and Concrete Composites*, 29(6), 505-514.

[46] V.Q. Tran, H.-V.T. Mai, T.-A. Nguyen and H.-B. Ly. (2021). Investigation of ANN architecture for predicting the compressive strength of concrete containing GGBFS. *PLoS One*, 16(12), e0260847.