Heuristic algorithm-based semi-empirical formulas for estimating the compressive strength of the normal and high performance concrete

Ngoc-Hien Nguyen[a], Thuc P. Vo[b,*], Seunghye Lee[c,**], Panagiotis G. Asteris[d]

[a] *CIRTech Institute, Ho Chi Minh City University of Technology (HUTECH), Ho Chi Minh City, Vietnam*
[b] *School of Engineering and Mathematical Sciences, La Trobe University, Bundoora, VIC 3086, Australia*
[c] *Deep Learning Architecture Research Center, Sejong University*
209, Neungdong-ro, Gwangjin-gu, Seoul 05006, Republic of Korea
[d] *Computational Mechanics Laboratory, School of Pedagogical and Technological Education,*
Heraklion, 14121 Athens, Greece

Abstract

It is a big challenge to design mixture proportions of the high-performance concrete due to highly nonhomogeneous relationships and coherent among many components. Although machine learning (ML) algorithms have been employed effectively to solve this problem, they are black box models and do not show an explicit relation between the compressive strength and mixture proportions. In order to overcome this inherent weakness, this paper proposes general semi-empirical formulas involving nondimensionalization and optimisation techniques. The optimisation process employs the Nelder-Mead simplex algorithm and takes into account the behaviour of uncertain variables, which may occur in experimental data. Successful compressive strength predictions of five datasets with high accuracy in compared to available ML models indicate that the proposed framework has the universal capacity, which can be used for various datasets. Furthermore, the explicit relation of semi-empirical formulas may be a useful tool for engineers and researchers in this area for the prediction purposes.

Keywords: High-performance concrete (HPC), semi-empirical formulas, nondimensionalisation and optimisation techniques, uncertain variables, noise model

---

1. Introduction

High-performance concrete (HPC) materials have been widely used in long-span bridges, high-rise buildings, dams, etc. These materials are often included blast-furnace (BF) slag, fly ash (FLA), silica fume (SF) and other supplementary substance such as super-plasticizer [1, 2]. The proportioning of each component can be tailored to meet imposed target strengths and performance [3]. Due to highly nonhomogeneous mixture, it is difficult to select mixture proportions and then predict the concrete compressive strength (CCS). The use of machine learning and statistical approaches to reduce the error between predicted and experimental data has received significant attention. Over the last two decades, various machine learning algorithms have been employed to propose an accurate and effective models for the HPC's strength. The two most popular ones are related to Neural Networks (NN) with single layer [4], multi-layer [5, 6] or combination with Monte Carlo stochastic sampling [7] as well as Artificial Neural Networks (ANNs) with fuzzy-ARTMAP type [8], multi-layer [9, 10] and with modified firefly algorithm [11]. Notably, Yeh [12, 13] presented the list of experimental data

---

[*] Corresponding author
[**] Corresponding author
Email addresses: `t.vo@latrobe.edu.au` (Thuc P. Vo), `seunghye@sejong.ac.kr` (Seunghye Lee)

of HPC with 1030/1133 samples, in which mixture proportions had eight input variables, and one output variable as CCS. Thanks for his contribution, these datasets have been widely used by many researchers in this area. There are also different ML models that can be used to predict strengths of HPC which include support vector machine (SVM) [14, 15], ensemble computational techniques such as random forest (RF) [16], adaptive boosting [17], gradient boosting (GB) [18] and boosting smooth transition regression trees [19] and data-mining [20]. Besides, some authors combined ANNS and fuzzy logic [21, 22], ANNs and regression analysis [23] or used various models such as linear regression, ANNs and SVM [24, 25]. Young et al. [26] used NN, GB, RF and SVM models to predict the CCS of more than 10,000 samples based on actual mixtures and considered industrial importance. More details related to ML models can be found in recent publications ([27], [28]). Although they have advantage benefits such as high accuracy, easy application and robustness, etc, they are black box models and do not provide details how input variables are being combined to make predictions. An explicit relation between the CCS and input variables cannot be found thus it is not easy to use. Therefore, many complicated mathematical models have been developed to propose it. Yeh and Lien [29] presented genetic operation tree model, which combined of the operation tree and genetic algorithm to calculate the CCS. These gene expression programmings (GEP) were applied to derive various models for the strength prediction of the HPC by Lim et al. [30], Tsai and Lin [31], Gandomi and Alavi [32] and Mousavi et al. [33, 34]. Due to advantage of ANNs, some researchers combined them with GEP ([35], [36]) or with fuzzy logic ([37]). It should be noted that explicit equations in these papers are quite complicated using cumbersome mathematics formulations, which are sometime difficult to use. Based on the correlation coefficients of proportioning of each component, CCS can also be predicted using linear, non-linear and metaheuristic regression methods [38, 39]. Bharatkumar et al. [40] investigated the effects of water content and mineral admixture of the of HPC to modify mix design procedure. Bhanja and Sengupta [41] developed relationship between the 28-day CCS of SF concrete with water-to-cement and SF replacement ratio. Namyong et al. [42] presented the regression equation for CCS of in-situ normal concrete (18-27MPa) based on cement, water-cement and fine/coarse aggregate. Videla and Gaedicke [43] combined hyperbolic function for strength evolution and exponential one for mixture design parameters for CCS of portland BF cement HPC. Zain and Abd [44] developed a multiple non-linear regression model to predict the strength of the HPC. It should be noted that these formulas in references [40–44] only use for one specific concrete material. Besides, the experiment data may contain some errors in mixture proportions and testing process ([26]), thus it is important to consider uncertain variables in the prediction model. The present study focuses on proposing a general semi-empirical framework that can predict accurately the CCS for different normal and HPCs with various mixture proportions and take into account the uncertain variables. These two main contributions are briefly highlighted as follows.

Firstly, this paper uses nondimensionalization and optimisation techniques to solve this complicated problem. Four dimensionless variables, which are weight of volume ratios and/or linear combination of existing components, is carefully selected to make sure that they are consistent with the physical terms. The optimal set of semi-empirical coefficients are searched using the means of optimisation techniques. The Nelder–Mead simplex algorithm, which is heuristic search method, is employed to find the optimum of cost function in a multidimensional space.

Secondly, the proposed method takes into account the behaviour of uncertain variables. While making laboratory-produced concrete samples, due to variance within proportion, mixture, and testing process, the experiment data may contain errors. Thus, a simple white Gaussian noise model is employed, in which the noise is assumed to be an independent, identically distributed with zero mean, and variance.

The proposed formulas are validated with five popular existing datasets, which have different mixture proportions. The obtained results are reasonably comparable with previous ones, especially for those using GEP and ANNs. The highest coefficient of determination (R^2) and a20-index, are for Dataset 3 with value of 0.9555 and 0.9310 and the lowest ones are for Dataset 1 with value of 0.8567 and 0.7527, respectively. The complex relationship between mixture proportion and CCS of the HPC can be predicted accurately by using the semi-empirical formulas, which provide a better understanding of how predictions are made.

The remaining of this study is outlined as follows. Section 2 provides a details of research methodology. Results of the predictive models and discussion are presented in Section 3. Section 4 draws some limitations and proposes future works. Finally, some concluding remarks are given in Section 5.

2. Research Methodology

This section focuses on two main contributions, which are nondimensionalization and optimisation techniques, of the proposed method. Besides, the behaviour of uncertain variables is also mentioned here. A general dataset of the HPC normally consists of several input variables including cement (C), water (W), blast-furnace slag (BF), fly ash (FLA), silica fume (SF), super-plasticizer (SP), high rate water reducing agent (HRWRA), air entraining agent content, which is the amount of air as a percentage of concrete volume (AE (%)), coarse aggregate (CA), fine aggregate (FA), Age of testing (Age), etc. Based on the correlation coefficients, the effect of each input variable to the CCS as output variable is different. In polynomial regression approach, one either uses all the variables to build the model which usually leads to a very long and too complicated polynomial equation or usually selects high correlation variables and ignore the others that leads to a low accuracy model. In this paper, a new set of dimensionless variables, $\alpha, \beta, \gamma, \delta$, which are ratios defined in terms of weight of volume and/or linear combination of existing variables, is carefully selected to make sure that they are consistent with the physical terms as below:

$$\alpha = \frac{\mathrm{C}}{\mathrm{W}}, \tag{1a}$$

$$\beta = \frac{\mathrm{FA}}{a_1\mathrm{CA}} + a_2, \tag{1b}$$

$$\gamma = \frac{a_3\mathrm{C} + a_4\mathrm{BF} + a_5\mathrm{SF} + a_6\mathrm{FLA}}{a_7\mathrm{W} + a_8\mathrm{SP} + a_9\mathrm{HRWRA}} + a_{10}\mathrm{AE}, \tag{1c}$$

$$\delta = \frac{(1 + \mathrm{Age})}{1\mathrm{day}}. \tag{1d}$$

It should be noted that in Eq. (1d), Age can be 1, 14, 28, 56, 90 and 365 days and 1 day is divided to make δ become dimensionless variable.

The proposed semi-empirical formula for the CCS is a linear combination of highly nonlinear terms of the dimensionless variables, as follows

$$y_p = f_{CCS} = f(\mathbf{a}, \mathbf{b}; \mathbf{\Theta}) = b_1 + b_2\alpha^{b_3} + b_4\beta^{b_5} + b_6\left(\log\gamma\right)^{b_7} + b_8\delta^{b_9}. \tag{2}$$

Here $\mathbf{\Theta} = (\alpha, \beta, \gamma, \delta)$, is the dimensionless variables, $\mathbf{a} = a_i, i = 1 \ldots 10$, and $\mathbf{b} = b_j, j = 1 \ldots 9$ are semi-empirical coefficients. The sum-square-error cost function, $J(\mathbf{a}, \mathbf{b})$, is defined as:

$$J(\mathbf{a}, \mathbf{b}; \mathbf{\Theta}) = \sum (y - f_{CCS})^2, \tag{3}$$

where y is the CCS from experimental data.

The behaviour of uncertain variables is introduced by adding noise in the experimental data. Here, a simple white Gaussian noise model [45, 46], i.e. $\varepsilon \sim N(\mu, \sigma_n^2 I)$ is employed, in which the noise is assumed to be an independent, identically distributed with zero mean ($\mu = 0$), and variance (σ_n^2). For ease of discussion and consistent, the noise variance is defined as a fraction of mean over standard deviation of experimental data:

$$\sigma_n^2 = \frac{\mathrm{mean}(y)}{\mathrm{std}(y)}. \tag{4}$$

Thus, the 'real' data with noise is generated synthetically as follows

$$y_{real} = y + \varepsilon. \tag{5}$$

Hence, the sum-square-error cost function becomes

$$J(\mathbf{a}, \mathbf{b}; \boldsymbol{\Theta}) = \sum (y_{real} - f_{CCS})^2. \tag{6}$$

Note that f_{CCS} is a linear combination of various highly nonlinear terms. Therefore, the optimal set of semi-empirical coefficients are searched using the means of optimisation techniques. In this paper, the Nelder–Mead simplex algorithm, which is heuristic search method, is employed to find the optimum of cost function in a multidimensional space. Based on a direct search method, it is often used for nonlinear optimisation problems in which the derivatives of variables might not be found (for more details, refer to references [47, 48]). In particular, the minimum of cost function is needed to find:

$$\min_{\mathbf{a},\mathbf{b}} J(\mathbf{a}, \mathbf{b}; \boldsymbol{\Theta}), \tag{7}$$

using *'fminsearch'* toolbox in MATLAB, with a suitable initial search condition.

The procedure of this approach consists of two stages: training and testing. A dataset is first collected and cleaning. The cleaned dataset is then divided into a portion of 80% – 20% for train and test sets. The training process is performed by adding 50% noise to the ideal target y. The testing one is taken place by introducing 4 levels of noise, from level-0 to level-3, corresponding to 0%, 10%, 50% and 100% noise, which is added to both training and testing sets.

The performance of the model is evaluated using four standard criteria, namely

Coefficient of determination (R^2)

$$\mathrm{R}^2 = \left(\frac{n \sum y y_p - (\sum y)(\sum y_p)}{\sqrt{n(\sum y^2) - (\sum y)^2} \sqrt{n(\sum y_p^2) - (\sum y_p)^2}} \right)^2, \tag{8}$$

Mean Absolute Error (MAE)

$$\mathrm{MAE} = \frac{1}{n} \sum_{i=1}^{n} \left| y^i - y_p^i \right|, \tag{9}$$

Root Mean Squared Error (RMSE)

$$\mathrm{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(y^i - y_p^i \right)^2}, \tag{10}$$

where y and y_p are the true and predicted values, and n indicates the number of data samples.

The a20-index, which has the advantage of a physical engineering meaning proposed by [49]:

$$\text{a20-index} = \frac{\text{n20}}{\text{n}}, \tag{11}$$

where n20 is the number of data samples, whose ratios of experimental and predicted value are in the range of 0.80 and 1.20. As a perfect model, a20-index is close to be the unity. It illustrates the number of samples that have the predicted values within a deviation of ±20% comparing to experimental ones.

The procedure of this approach is summarised in Algorithm: Semi-empirical Procedure.

Algorithm: Semi-empirical Procedure

Given a dataset S with n samples.

1. Cleaning data
   - Checking duplicate and remove if any
   - Checking NA/N and impute with mean value if any
   - Checking outliers using IQR criteria and remove if any
   - Randomly split the cleaned data into training (80%) and testing (20%).
2. Training
   - Carefully check the availability of input variables to define $\alpha, \beta, \gamma, \delta$ and f_{CCS} as mentioned in Eqs. (1) and (2), respectively.
   - If a variable does not exist, the corresponding coefficients can be set to zero.
   - Adding white noise into the training set.
   - Solving the sum-square-error cost in Eqs. (6) and (7) to find the semi-empirical coefficients $\mathbf{a}, \mathbf{b}$.
3. Testing
   - Checking the MAE, RMSE, R^2 and a20-index of the training set.
   - Adding some noise levels to the training and testing sets and checking MAE, RMSE, R^2, and a20-index.
   - If the criteria is not satisfied, return to step 2 and modify the proposed formula and go on.

3. Numerical examples

In this section, the proposed prediction formulas are applied for some reliable datasets with various input variables, which are from eight to ten, to show the universal capacity of the present approach. The performance is evaluated using four standard criteria, R^2, MAE, RMSE and a20-index, which illustrate a strong correlation between the predicted and experimental data. The obtained results are compared with those available in the literature to validate the semi-empirical formulas.

3.1. Datasets of the HPC

Five datasets with concrete compressive strength (CCS) as the output variable and various input variables are considered. Dataset 1 consists of 1030 testing results, which are collected by Yeh [12] for different mixtures with eight input variables including C, W, BF, FLA, SP, CA, FA, and Age. Figure 1 shows the correlation heat-map and outliers[1] of Dataset 1. It is clear that there is no really dominant variables that have strong correlation with the CCS. The Age variable has extremely outliers that can affect to the model performance. Dataset 2 with a total of 200 samples is collected by Videla and Gaedicke [43]. The CCS can be predicted by nine input variables, namely, C, W, SF, SP, HRWRA, AE, CA, FA and Age. It is clear from Figure 2 that only the Age has strong correlation with CCS. There is no extremely outliers that can affect to the performance of our model. Dataset 3 consists of 144 testing results, which are collected by Lam et al. [50] and later Pala et al. [3] with eight input variables, namely, TCM, W, $\mathrm{FLA_R}$, $\mathrm{SF_R}$, HRWRA, FA, CA, and Age. Figure 3 shows that most of the all variables have strong correlation with the CCS except for SF. There is no extremely outliers that can affect to the performance of our model. Dataset 4 is combined of Datasets 2 and 3 to show the versatility of the proposed formula. This dataset has ten input variables, C, W, SF, SP, HRWRA, AE, CA, FA and Age. It is noticed that the range of AE is small, however, its correlation is weak positive linear relationship to the CCS as illustrated in Figure 4. Finally, Dataset 5 with a total of 104 samples is collected by Lim et al. [30] with seven input variables, W/B, W, fine aggregate to total aggregate ratio (S/A), $\mathrm{FLA_R}$, AE and SP. Figure 5 shows that most of the variables have strong correlation with the CCS except for FLA, W, and S/A. The statistical information about input and output variables of each dataset can be found in Table 1.

3.2. Results and discussions

By using Algorithm: Semi-empirical Procedure, after the cleaning step, there are 892, 195, 144, 339 and 101 cleaned samples, which are randomly split to 714, 156, 115, 271 and 81 for training and then 178, 39, 29, 68 and 20 for testing for Datasets 1–5, respectively. It is from Eqs. (1) and (2) that four dimensionless variables, $\alpha, \beta, \gamma, \delta$ and corresponding coefficients $\mathbf{a}, \mathbf{b}$ of semi-empirical formulas are found. These values for each dataset are given in Tables 2 and 3.

Table 4 presents the four criteria to evaluate the performance of training and testing sets. It should be noted that to train the model, the noise level-2 is added to the target y. The reason to use this level is to ensure that the model is able to represent the behaviour of the target variable at both end (0% and 100%) of the noise. The consistency between training and testing results indicates that there is no overfitting in the training process. It can be seen from this table that the proposed semi-empirical formulas perform outstanding with five datasets with the lowest values of $\mathrm{R}^2 = 0.8316$ and a20-index $= 0.6758$ for the testing set of Dataset 1 with noise at level-3, which implies that a significant correlation between the predicted and experiment data ([51]). Among five datasets, the best performance with the highest values of $\mathrm{R}^2 = 0.9555$ and a20-index $= 0.9310$ for the testing set of Dataset 3 with noise at level-1. Because of combining between Datasets 2 and 3, the performance of Dataset 4 has slightly reduced to $\mathrm{R}^2 = 0.8996$ and a20-index $= 0.7971$ for noise at level-1. Besides, it is clear that the model with noise level-1 shows outstanding performance thus it is selected to do the interpretative analysis.

Table 5 shows the validation of semi-empirical formulas with the previous studies [3, 25, 28, 30–32, 34, 43], which used different approach for each dataset. While Lim et al. [30], Tsai and

---

[1] It would be notice that the proposed approach does not apply any transformation data techniques. However, range of values of the box-plots seems to be hard to observe due to the difference in value ranges of inputs. Hence, all the box-plot figures are applied log-scaled x-axis to increase the visibility.

Lin [31], Gandomi and Alavi [34], Mousavi et al. [32] and Videla and Gaedicke [43] derived explicit equations for compressive strength of HPC, Pala et al. [3], Chou and Pham [25] and Asteris et al. [28] solved this problem by using ANNs. It can be seen that the obtained results with noise level-1 are reasonably comparable with previous ones, especially for those using GEP and Multi-GGP. The results for Dataset 1 agree very well with those from Asteris et al. [28] in terms of R^2 and a20-index. Although R^2 from previous studies using ANNs is slightly higher than that of this one, the explicit relation of semi-empirical formula could be benefit in practice. Due to interpretable models, they provide a better understanding of how predictions are made. The correlation plot between the actual and predicted output for Datasets 1-5 shows an excellent performance and is very close to the ideal line as shown in Figures 6–10. It can be concluded that the present novel approach is simple and suitable various datasets with very high accuracy.

Before conducting parameter study, it is necessary to explore the feature importance of the dataset. Feature importance scores can be used to determine the highest/lowest interest in design mixture proportions and helps to understand which input variables need to pay more attention than others. They can be calculated using linear models and decision trees i.e., CART, Random Forest, XGBoost. Figure 11 shows the feature importance of Dataset 1, 2, 3 and 5 extracted from XGBoost Feature Importance [52]. The dash-line indicates 70% important level. For Dataset 1, it can be seen that C and Age are the two most significant effects, while CA appears to be the least important input variables. Surprisingly when there is only one feature important for Dataset 2, 3, and 5. In particular, Age is the most important variable in Dataset 2, although HRWRA can also worth to consider. TCM is the only candidate for Dataset 3. While SP is the important factor in Dataset 5, though W/B can also play a important role on the outcome of the CCS.

Figure 12 presents the CCS with respect to W/B ratio for different days with a specific mixture of Dataset 1 (BF/B = 0%, FLA/B = 0%, SP/B = 0% and CA/FA = 1.1, 1.3 and 1.8). As expected in the left figure, smooth curves follow the same trends for various days (7, 14, 28, 56 and 90) and already pass over some points in Dataset 1, which verifies the accuracy of semi-empirical formula. The response of the proposed model is further investigated for various days (91, 180, 270 and 360), where only a few data points are available. The plot in the right figure confirms again that the overfitting problems, which usually happens due to an insufficient data, are not occurring here since all curves follow the exactly same way.

Effects of BF and FLA on the CCS with respect to W/B ratio are plotted in Figure 13. The dash-line is for the predicted CCS with BF/B = 0% (left) and FLA/B = 0% (right) while the solid-line is for BF/B = 30% (left) and FLA/B = 30% (right). It should be mentioned that due to the group of CCS curves very near to each other, the dots which represent the experimental data are not added to the plots to avoid confusion. It can be seen that the CCS increases with the increase of SP and decreases when BF and/or FLA are used.

For a specific mixture of Dataset 2 (W/B = 30%, SF/B = 10%, SP/B = 0.5%, AE = 2%, HRWRA/B = 1.8% and CA/FA = 1.7, 1.9 and 2.1), the CCS increases completely with increased CA/FA ratio as illustrated in Figure 14. Effects of SP and FLA on the CCS with respect to W/B ratio for different days (7, 28 and 56) of Dataset 3 are shown in Figure 15 with (FLA/B = 0, 20%, SP/B = 0, 5%, CA/FA = 2). It implies that SP enhances the CCS and when FLA is used, the CCS decreases. Finally, it should be noted again that all smooth curves in Figures 12–15 from Datasets 1-3 always go through at least few points in the datasets, which confirms the accuracy of the semi-empirical formula proposed in this paper.

4. Limitations and future work

The proposed semi-empirical equations are applied in the tenth-dimensional space defined by the ten parameters which effect the development of the compressive strength of the normal and high performance concrete. In fact, the derived optimum semi-empirical equations are applicable for parameters, whose values range between the lowest and highest ones as presented in Table 1. In other words, the proposed semi-empirical equations learns the inputs-output relationship in the optimum sense, and can generalise their predictions to new data once achieving the optimal coefficients. However, if the unseen data is out of bound of the learning model, the inaccurate predictions are expected. For example, for Dataset 1, due to experimental data for W/B ratios over 0.6 and SP/B over 4% are sparse, it is proposed to limit the use of the proposed equations only for W/B ratios between 0.30 and 0.60 while regarding the SP/B, the proposed equations are valid for values less than 4%. It means that more experiments in certain area are needed to achieve the best equations for this dataset. For Datasets 2 and 3, similar suggestions are recommended for W/B ratios between 0.25 and 0.37, and 0.3 and 0.5, respectively.

Besides, performing nondimensionalisation and optimisation techniques to obtain semi-empirical equations of compressive concrete strength are mathematically straightforward. However, selecting of correct/accurate dimensionless variables is not a trivial task and requires considerably analysis/judgement and experience. For each extra nondimensional variable added to the formula, the computation and complexity of the solution will increase. Optimal the set of dimensionless variables is therefore necessary and will be carried to future work.

Furthermore, it should be mentioned that the accuracy of the proposed model depends on quality of datasets, which relates to number, variability of inputs and meaningful to the prediction. Therefore, feature engineering and feature selection will be an essential tool for future study, which helps preparing a suitable and reliable input variables and thus enhancing the overall performance of the predictor model.

5. Conclusion

In this paper, optimum semi-empirical framework is proposed for to predict the compressive strength of the normal and high performance concrete various types datasets from eight to ten parameters. Four dimensionless variables, which are weighted ratios and/or linear combination of existing ones, is carefully selected and then Nelder-Mead simplex algorithm is employed as optimisation process. The technique shows significant advantage in reducing the number of independent variables and can deal with various mixture proportions. In order to consider some errors within mixture proportions and testing process, the behaviour of uncertain variables is included by adding noise effect using a simple white Gaussian noise model. The performance of five datasets proved that the semi-empirical formulas are capable of predicting accurately the data with certain noises. Although the results are slightly higher error than those from available machine learning models, their explicit relation may be benefit for the prediction purposes. Due to interpretable models, they provide a better understanding of how predictions are made. Successful predictions of the compressive strength of five datasets indicate the proposed optimum semi-empirical equations are useful tool for researchers, engineers, and for supporting the teaching and interpretation of the behaviour concrete materials as they reveal their strongly non-linear nature in the tenth-dimensional space.

Acknowledgement

This research was supported by a grant (NRF-2018R1C1B6004751) from NRF (National Research Foundation of Korea) funded by MEST (Ministry of Education and Science Technology) of Korean government.

References

[1] A. Neville, P.-C. Aïtcin, High performance concrete—An overview, Materials and Structures 31 (2) (1998) 111–117.

[2] C. K. Y. Leung, Concrete as a Building Material, in: Encyclopedia of Materials: Science and Technology, Elsevier, 2001, pp. 1471–1479.

[3] M. Pala, E. Özbay, A. Öztaş, M. I. Yuce, Appraisal of long-term effects of fly ash and silica fume on compressive strength of concrete by neural networks, Construction and Building Materials 21 (2) (2007) 384–394. `doi:10.1016/j.conbuildmat.2005.08.009`.

[4] S. Lai, M. Serra, Concrete strength prediction by means of neural network, Construction and Building Materials 11 (2) (1997) 93–98. `doi:https://doi.org/10.1016/S0950-0618(97)00007-X`.
URL `https://www.sciencedirect.com/science/article/pii/S095006189700007X`

[5] H.-G. Ni, J.-Z. Wang, Prediction of compressive strength of concrete by neural networks, Cement and Concrete Research 30 (8) (2000) 1245–1250. `doi:https://doi.org/10.1016/S0008-8846(00)00345-8`.
URL `https://www.sciencedirect.com/science/article/pii/S0008884600003458`

[6] A. Öztaş, M. Pala, E. Özbay, E. Kanca, N. Çag˘lar, M. A. Bhatti, Predicting the compressive strength and slump of high strength concrete using neural network, Construction and Building Materials 20 (9) (2006) 769–775. `doi:https://doi.org/10.1016/j.conbuildmat.2005.01.054`.
URL `https://www.sciencedirect.com/science/article/pii/S0950061805000942`

[7] M. Słoński, A comparison of model selection methods for compressive strength prediction of high-performance concrete using neural networks, Computers and Structures 88 (21) (2010) 1248–1253. `doi:https://doi.org/10.1016/j.compstruc.2010.07.003`.
URL `https://www.sciencedirect.com/science/article/pii/S0045794910001598`

[8] K. Janusz, R. Janusz, D. Artur, Hpc strength prediction using artificial neural network, Journal of Computing in Civil Engineering 9 (4) (1995) 279–284. `doi:10.1061/(ASCE)0887-3801(1995)9:4(279)`.
URL `https://doi.org/10.1061/(ASCE)0887-3801(1995)9:4(279)`

[9] S.-C. Lee, Prediction of concrete strength using artificial neural networks, Engineering Structures 25 (7) (2003) 849–857. `doi:https://doi.org/10.1016/S0141-0296(03)00004-X`.
URL `https://www.sciencedirect.com/science/article/pii/S014102960300004X`

[10] M. Sarıdemir, Prediction of compressive strength of concretes containing metakaolin and silica fume by artificial neural networks, Advances in Engineering Software 40 (5) (2009) 350–355. `doi:https://doi.org/10.1016/j.advengsoft.2008.05.002`.
URL `https://www.sciencedirect.com/science/article/pii/S0965997808001087`

[11] D.-K. Bui, T. Nguyen, J.-S. Chou, H. Nguyen-Xuan, T. D. Ngo, A modified firefly algorithm-artificial neural network expert system for predicting compressive and tensile strength of high-performance concrete, Construction and Building Materials 180 (2018) 320–333.

[12] I.-C. Yeh, Modeling of strength of high-performance concrete using artificial neural networks, Cement and Concrete Research 28 (12) (1998) 1797–1808. `doi:10.1016/s0008-8846(98)00165-3`.

[13] I.-C. Yeh, Analysis of strength of concrete using design of experiments and neural networks, Journal of Materials in Civil Engineering 18 (4) (2006) 597–604. `doi:10.1061/(ASCE)0899-1561(2006)18:4(597)`.
URL `https://doi.org/10.1061/(ASCE)0899-1561(2006)18:4(597)`

[14] K. Yan, C. Shi, Prediction of elastic modulus of normal and high strength concrete by support vector machine, Construction and Building Materials 24 (8) (2010) 1479–1485. `doi:https://doi.org/10.1016/j.conbuildmat.2010.01.006`.
URL `https://www.sciencedirect.com/science/article/pii/S0950061810000188`

[15] M. H. Rafiei, W. H. Khushefati, R. Demirboga, H. Adeli, Supervised Deep Restricted Boltzmann Machine for Estimation of Concrete, Materials Journal 114 (2) (2017) 237–244.

[16] Q. Han, C. Gui, J. Xu, G. Lacidogna, A generalized method to predict the compressive strength of high-performance concrete by improved random forest algorithm, Construction and Building Materials 226 (2019) 734–742. `doi:https://doi.org/10.1016/j.conbuildmat.2019.07.315`.
URL `https://www.sciencedirect.com/science/article/pii/S0950061819319890`

[17] D.-C. Feng, Z.-T. Liu, X.-D. Wang, Y. Chen, J.-Q. Chang, D.-F. Wei, Z.-M. Jiang, Machine learning-based compressive strength prediction for concrete: An adaptive boosting approach, Construction and Building

Materials 230 (2020) 117000. doi:https://doi.org/10.1016/j.conbuildmat.2019.117000.
URL https://www.sciencedirect.com/science/article/pii/S0950061819324420

[18] H. I. Erdal, O. Karakurt, E. Namli, High performance concrete compressive strength forecasting using ensemble models based on discrete wavelet transform, Engineering Applications of Artificial Intelligence 26 (4) (2013) 1246–1254.

[19] U. Anyaoha, A. Zaji, Z. Liu, Soft computing in estimating the compressive strength for high-performance concrete via concrete composition appraisal, Construction and Building Materials 257 (2020) 119472. doi:https://doi.org/10.1016/j.conbuildmat.2020.119472.
URL https://www.sciencedirect.com/science/article/pii/S095006182031477X

[20] C. Jui-Sheng, C. Chien-Kuo, F. Mahmoud, A.-T. Ismail, Optimizing the prediction accuracy of concrete compressive strength based on a comparison of data-mining techniques, Journal of Computing in Civil Engineering 25 (3) (2011) 242–253. doi:10.1061/(ASCE)CP.1943-5487.0000088.
URL https://doi.org/10.1061/(ASCE)CP.1943-5487.0000088

[21] İlker Bekir Topçu, M. Sarıdemir, Prediction of compressive strength of concrete containing fly ash using artificial neural networks and fuzzy logic, Computational Materials Science 41 (3) (2008) 305–311. doi:https://doi.org/10.1016/j.commatsci.2007.04.009.
URL https://www.sciencedirect.com/science/article/pii/S0927025607001085

[22] M. Fazel Zarandi, I. Türksen, J. Sobhani, A. Ramezanianpour, Fuzzy polynomial neural networks for approximation of the compressive strength of concrete, Applied Soft Computing 8 (1) (2008) 488–498. doi:https://doi.org/10.1016/j.asoc.2007.02.010.
URL https://www.sciencedirect.com/science/article/pii/S1568494607000348

[23] U. Atici, Prediction of the strength of mineral admixture concrete using multivariable regression analysis and an artificial neural network, Expert Systems with Applications 38 (8) (2011) 9609–9618. doi:https://doi.org/10.1016/j.eswa.2011.01.156.
URL https://www.sciencedirect.com/science/article/pii/S0957417411001898

[24] J.-S. Chou, C.-F. Tsai, Concrete compressive strength analysis using a combined classification and regression technique, Automation in Construction 24 (2012) 52–60. doi:https://doi.org/10.1016/j.autcon.2012.02.001.
URL https://www.sciencedirect.com/science/article/pii/S0926580512000167

[25] J.-S. Chou, A.-D. Pham, Enhanced artificial intelligence for ensemble approach to predicting high performance concrete compressive strength, Construction and Building Materials 49 (2013) 554–563. doi:https://doi.org/10.1016/j.conbuildmat.2013.08.078.
URL https://www.sciencedirect.com/science/article/pii/S0950061813008088

[26] B. A. Young, A. Hall, L. Pilon, P. Gupta, G. Sant, Can the compressive strength of concrete be estimated from knowledge of the mixture proportions?: New insights from statistical analysis and machine learning methods, Cement and Concrete Research 115 (2019) 379–388. doi:https://doi.org/10.1016/j.cemconres.2018.09.006.
URL https://www.sciencedirect.com/science/article/pii/S0008884617313807

[27] H. Nguyen, T. Vu, T. P. Vo, H.-T. Thai, Efficient machine learning models for prediction of concrete strengths, Construction and Building Materials 266 (2021) 120950. doi:https://doi.org/10.1016/j.conbuildmat.2020.120950.
URL https://www.sciencedirect.com/science/article/pii/S095006182032955X

[28] P. G. Asteris, A. D. Skentou, A. Bardhan, P. Samui, K. Pilakoutas, Predicting concrete compressive strength using hybrid ensembling of surrogate machine learning models, Cement and Concrete Research 145 (2021) 106449. doi:https://doi.org/10.1016/j.cemconres.2021.106449.
URL https://www.sciencedirect.com/science/article/pii/S0008884621000983

[29] I.-C. Yeh, L.-C. Lien, Knowledge discovery of concrete material using Genetic Operation Trees, Expert Systems with Applications 36 (3, Part 2) (2009) 5807–5812.

[30] C.-H. Lim, Y.-S. Yoon, J.-H. Kim, Genetic algorithm in mix proportioning of high-performance concrete, Cement and Concrete Research 34 (3) (2004) 409–420. doi:10.1016/j.cemconres.2003.08.018.

[31] H.-C. Tsai, Y.-H. Lin, Predicting high-strength concrete parameters using weighted genetic programming, Engineering with Computers 27 (4) (2011) 347–355.
URL https://doi.org/10.1007/s00366-011-0208-z

[32] A. H. Gandomi, A. H. Alavi, A new multi-gene genetic programming approach to nonlinear system modeling. part i: materials and structural engineering problems, Neural Computing and Applications 21 (1) (2012) 171–187.
URL https://doi.org/10.1007/s00521-011-0734-z

[33] S. M. Mousavi, A. H. Gandomi, A. H. Alavi, M. Vesalimahmood, Modeling of compressive strength of hpc mixes using a combined algorithm of genetic programming and orthogonal least squares, Structural Engineering and Mechanics 36 (2) (2010) 225–241. doi:10.12989/SEM.2010.36.2.225.

URL https://doi.org/10.12989/SEM.2010.36.2.225

[34] S. M. Mousavi, P. Aminian, A. H. Gandomi, A. H. Alavi, H. Bolandi, A new predictive model for compressive strength of hpc using gene expression programming, Advances in Engineering Software 45 (1) (2012) 105–114. doi:https://doi.org/10.1016/j.advengsoft.2011.09.014.
URL https://www.sciencedirect.com/science/article/pii/S0965997811002535

[35] A. Baykasoğlu, T. Dereli, S. Tanış, Prediction of cement strength using soft computing techniques, Cement and Concrete Research 34 (11) (2004) 2083–2090. doi:https://doi.org/10.1016/j.cemconres.2004.03.028.
URL https://www.sciencedirect.com/science/article/pii/S0008884604001164

[36] P. Chopra, R. K. Sharma, M. Kumar, Prediction of compressive strength of concrete using artificial neural network and genetic programming, Advances in Materials Science and Engineering 2016 (2016) 7648467. doi:10.1155/2016/7648467.
URL https://doi.org/10.1155/2016/7648467

[37] S. Akkurt, G. Tayfur, S. Can, Fuzzy logic model for the prediction of cement compressive strength, Cement and Concrete Research 34 (8) (2004) 1429–1433. doi:https://doi.org/10.1016/j.cemconres.2004.01.020.
URL https://www.sciencedirect.com/science/article/pii/S0008884604000444

[38] C. Jui-Sheng, K. Chong Wai, B. Dac-Khuong, Nature-inspired metaheuristic regression system: Programming and implementation for civil engineering applications, Journal of Computing in Civil Engineering 30 (5) (2016) 04016007. doi:10.1061/(ASCE)CP.1943-5487.0000561.
URL https://doi.org/10.1061/(ASCE)CP.1943-5487.0000561

[39] T. Le-Duc, Q.-H. Nguyen, H. Nguyen-Xuan, Balancing composite motion optimization, Information Sciences 520 (2020) 250–270. doi:10.1016/j.ins.2020.02.013.

[40] B. Bharatkumar, R. Narayanan, B. Raghuprasad, D. Ramachandramurthy, Mix proportioning of high performance concrete, Cement and Concrete Composites 23 (1) (2001) 71–80, special Issue on Theme Analysis. doi:https://doi.org/10.1016/S0958-9465(00)00071-8.
URL https://www.sciencedirect.com/science/article/pii/S0958946500000718

[41] S. Bhanja, B. Sengupta, Investigations on the compressive strength of silica fume concrete using statistical methods, Cement and Concrete Research 32 (9) (2002) 1391–1394. doi:https://doi.org/10.1016/S0008-8846(02)00787-1.
URL https://www.sciencedirect.com/science/article/pii/S0008884602007871

[42] J. Namyong, Y. Sangchun, C. Hongbum, Prediction of compressive strength of in-situ concrete based on mixture proportions, Journal of Asian Architecture and Building Engineering 3 (1) (2004) 9–16. arXiv:https://doi.org/10.3130/jaabe.3.9, doi:10.3130/jaabe.3.9.
URL https://doi.org/10.3130/jaabe.3.9

[43] C. Videla, C. Gaedicke, Modeling portland blast-furnace slag cement high-performance concrete, ACI Materials Journal 101 (5) (2004) 365–375. doi:10.14359/13422.

[44] M. F. M. Zain, S. M. Abd, Multiple regression model for compressive strength prediction of high performance concrete, Journal of Applied Sciences 9 (2009) 155–160.

[45] J. Nocedal, S. J. Wright, Numerical Optimization, 2nd Edition, Springer New York, 2006. doi:10.1007/978-0-387-40065-5.

[46] C. E. Rasmussen, C. K. I. Williams, Gaussian Processes for Machine Learning, The MIT Press, 2005. doi:10.7551/mitpress/3206.001.0001.

[47] J. A. Nelder, R. Mead, A Simplex Method for Function Minimization, The Computer Journal 7 (4) (1965) 308–313. arXiv:https://academic.oup.com/comjnl/article-pdf/7/4/308/1013182/7-4-308.pdf, doi:10.1093/comjnl/7.4.308.
URL https://doi.org/10.1093/comjnl/7.4.308

[48] J. C. Lagarias, J. A. Reeds, M. H. Wright, P. E. Wright, Convergence properties of the nelder–mead simplex method in low dimensions, SIAM Journal on Optimization 9 (1) (1998) 112–147. doi:10.1137/s1052623496303470.

[49] P. G. Asteris, V. G. Mokos, Concrete compressive strength using artificial neural networks, Neural Computing and Applications 32 (15) (2020) 11807–11826. doi:10.1007/s00521-019-04663-2.
URL https://doi.org/10.1007/s00521-019-04663-2

[50] L. Lam, Y. Wong, C. Poon, Effect of fly ash and silica fume on compressive and fracture behaviors of concrete, Cement and Concrete Research 28 (2) (1998) 271–283. doi:10.1016/s0008-8846(97)00269-x.

[51] G. N. Smith, Probability and statistics in civil engineering: An introduction, London: Collins, 1986.

[52] T. Chen, C. Guestrin, XGBoost, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2016. doi:10.1145/2939672.2939785.

List of Figures

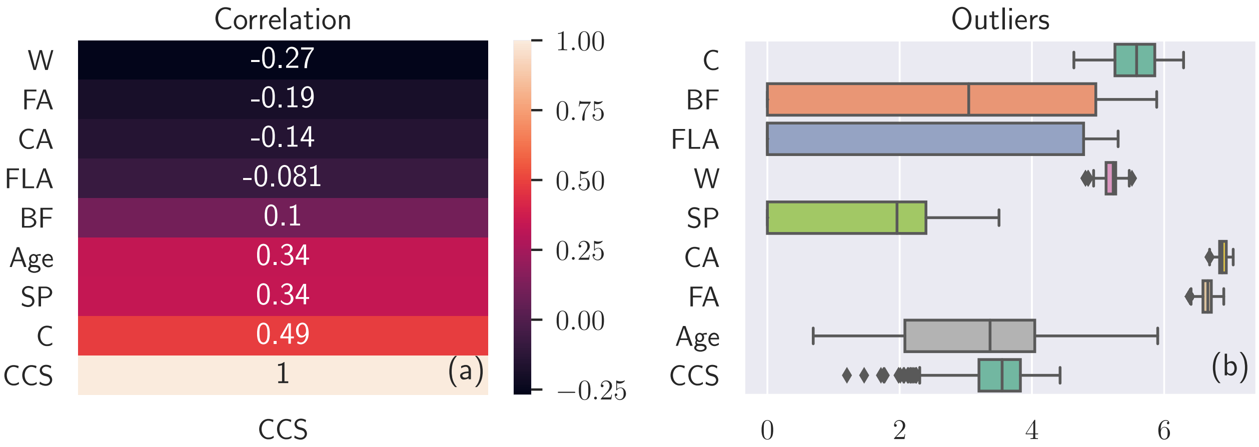

Figure 1: Correlation heat-map and boxplot outliers of features for Dataset 1 with 1030 samples.

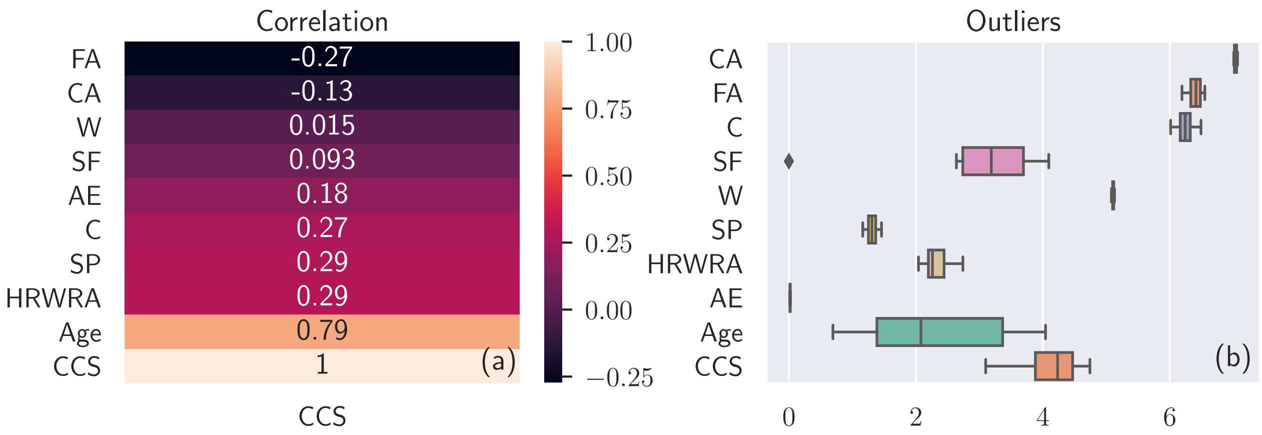

Figure 2: Correlation heat-map and boxplot outliers of features for Dataset 2 with 200 samples.

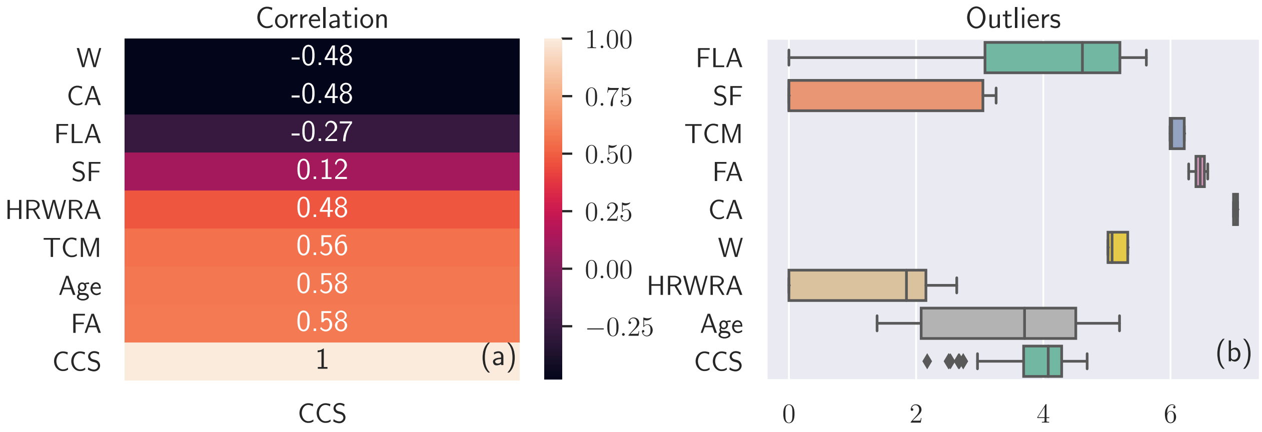

Figure 3: Correlation heat-map and boxplot outliers of features for Dataset 3 with 144 samples.

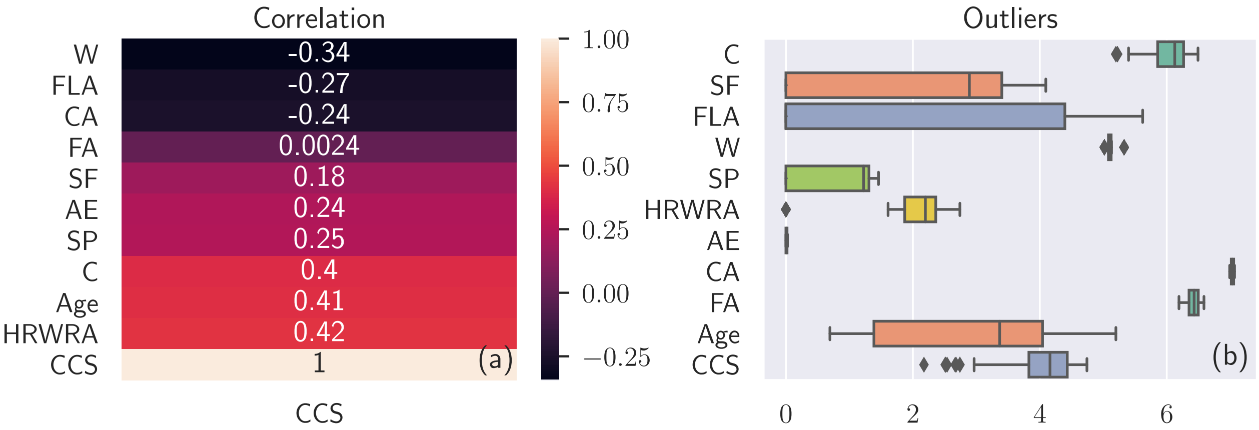

Figure 4: Correlation heat-map and boxplot outliers of features for Dataset 4 with 344 samples.

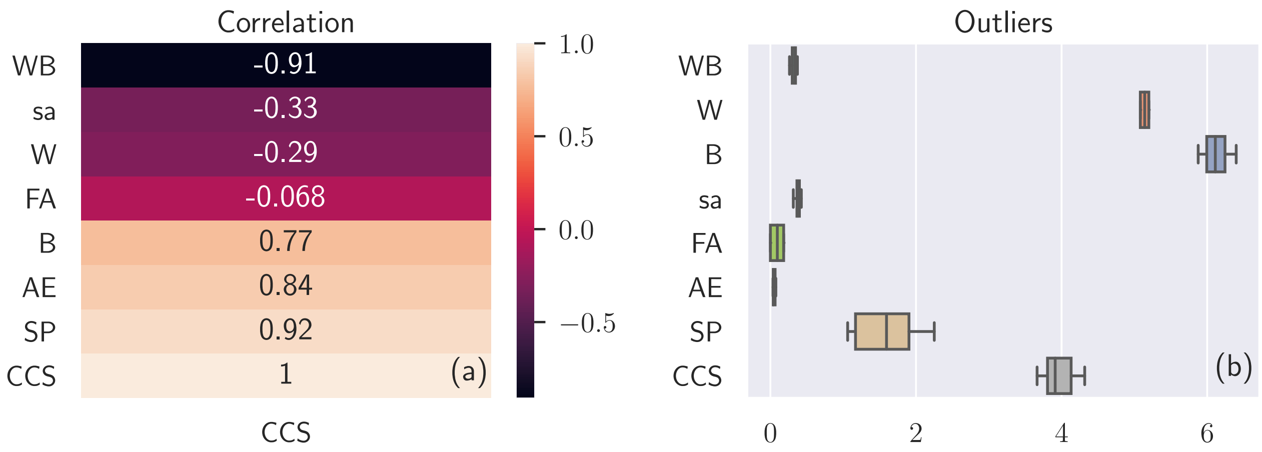

Figure 5: Correlation heat-map and boxplot outliers of features for Dataset 5 with 104 samples.

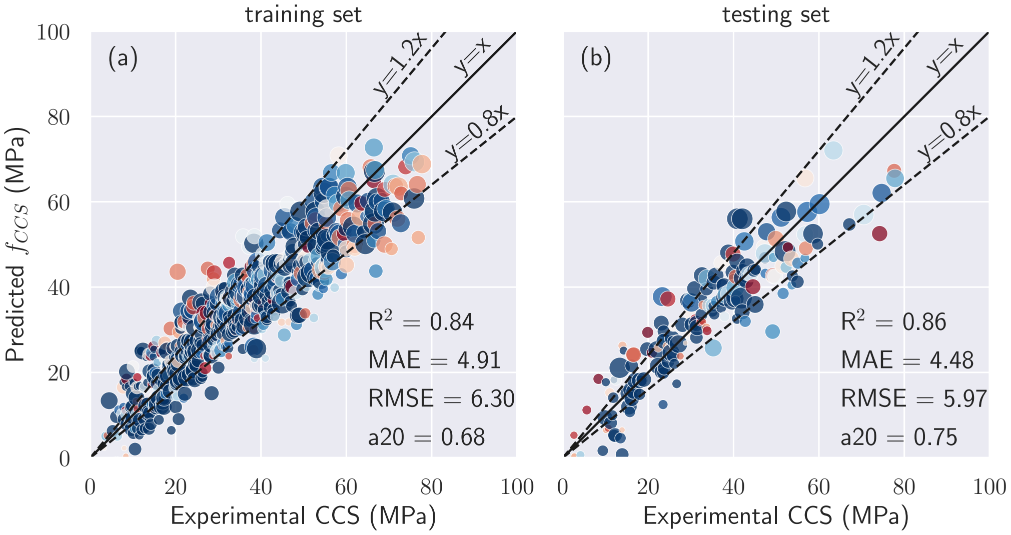

Figure 6: Predicted versus experimental data for Dataset 1 with 10% noise. The size of the dots illustrates C values, while the color represents W from small (light color) to large values (darker color).

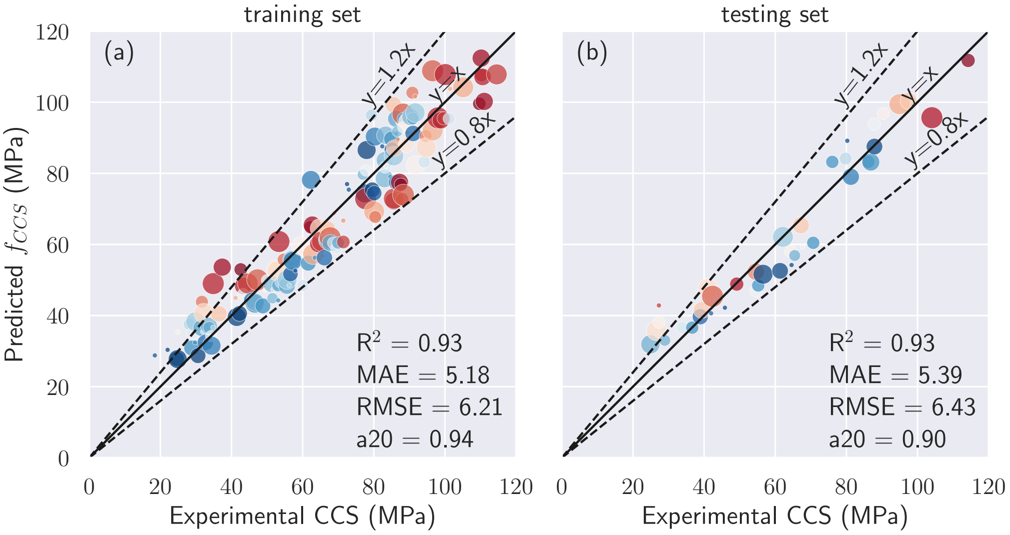

Figure 7: Predicted versus experimental data for Dataset 2 with 10% noise. The size of the dots illustrates C values, while the color represents W from small (light color) to large values (darker color).

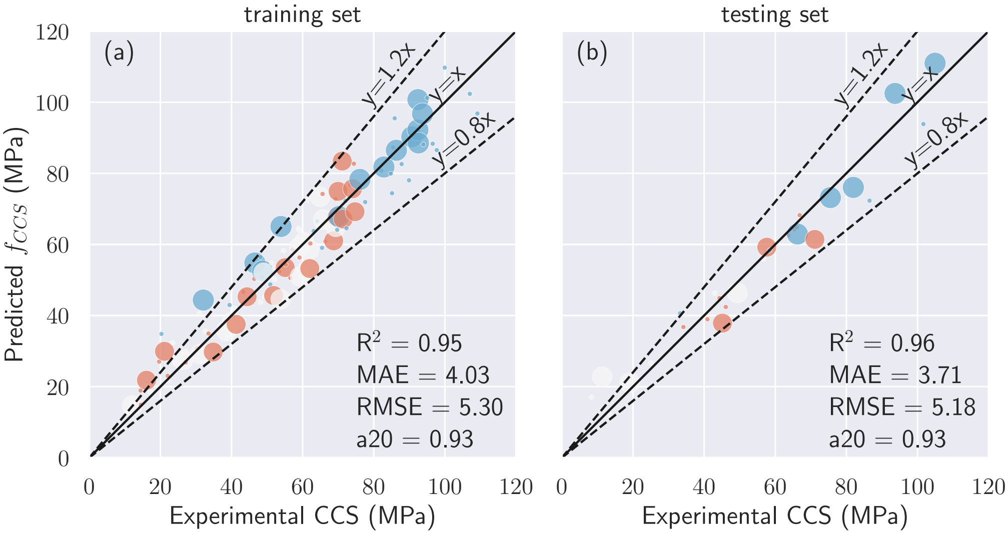

Figure 8: Predicted versus experimental data for Dataset 3 with 10% noise. The size of the dots illustrates TCM values, while the color represents W from small (light color) to large values (darker color).

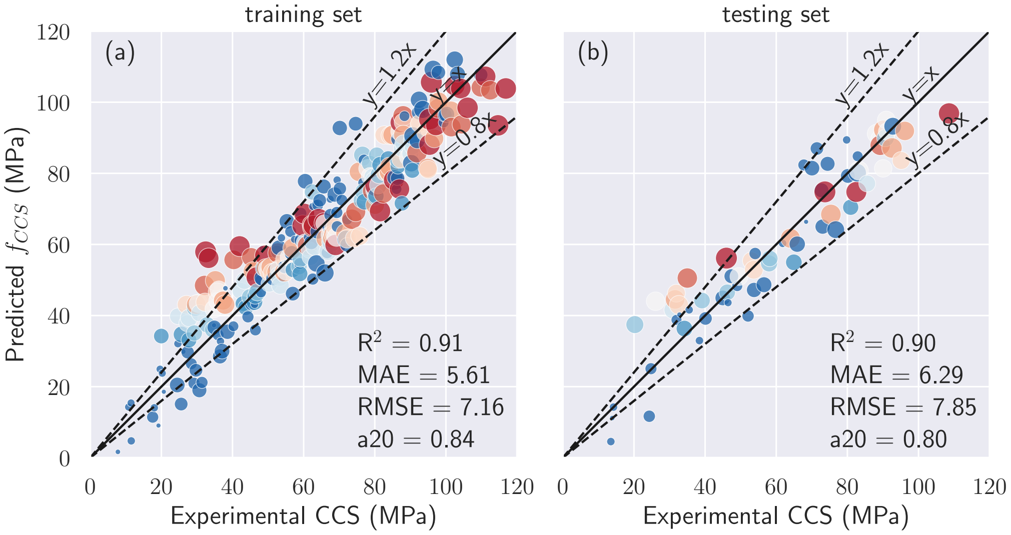

Figure 9: Predicted versus experimental data for Dataset 4 with 10% noise. The size of the dots illustrates C values, while the color represents W from small (light color) to large values (darker color).

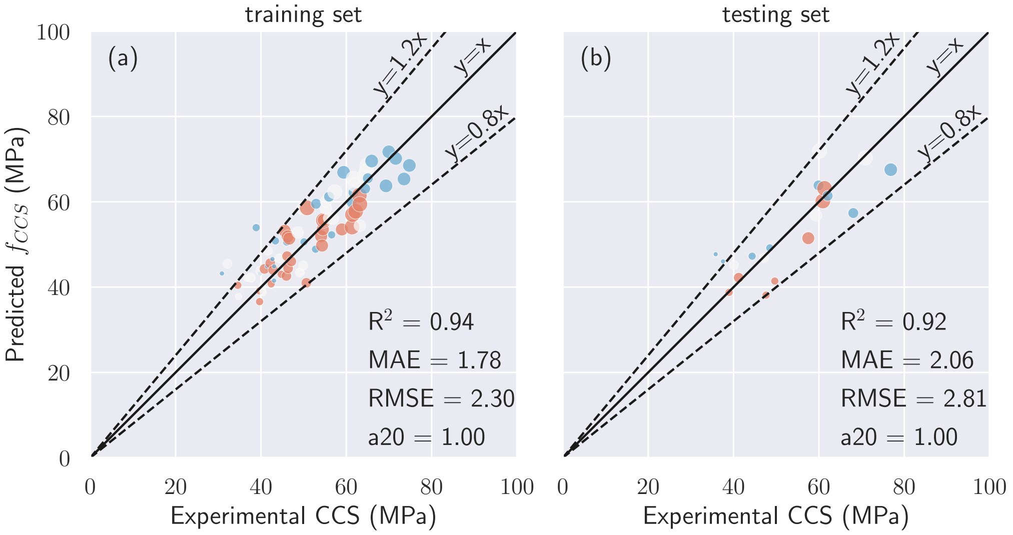

Figure 10: Predicted versus experimental data for Dataset 5 with 10% noise. The size of the dots illustrates B values, while the color represents W from small (light color) to large values (darker color).

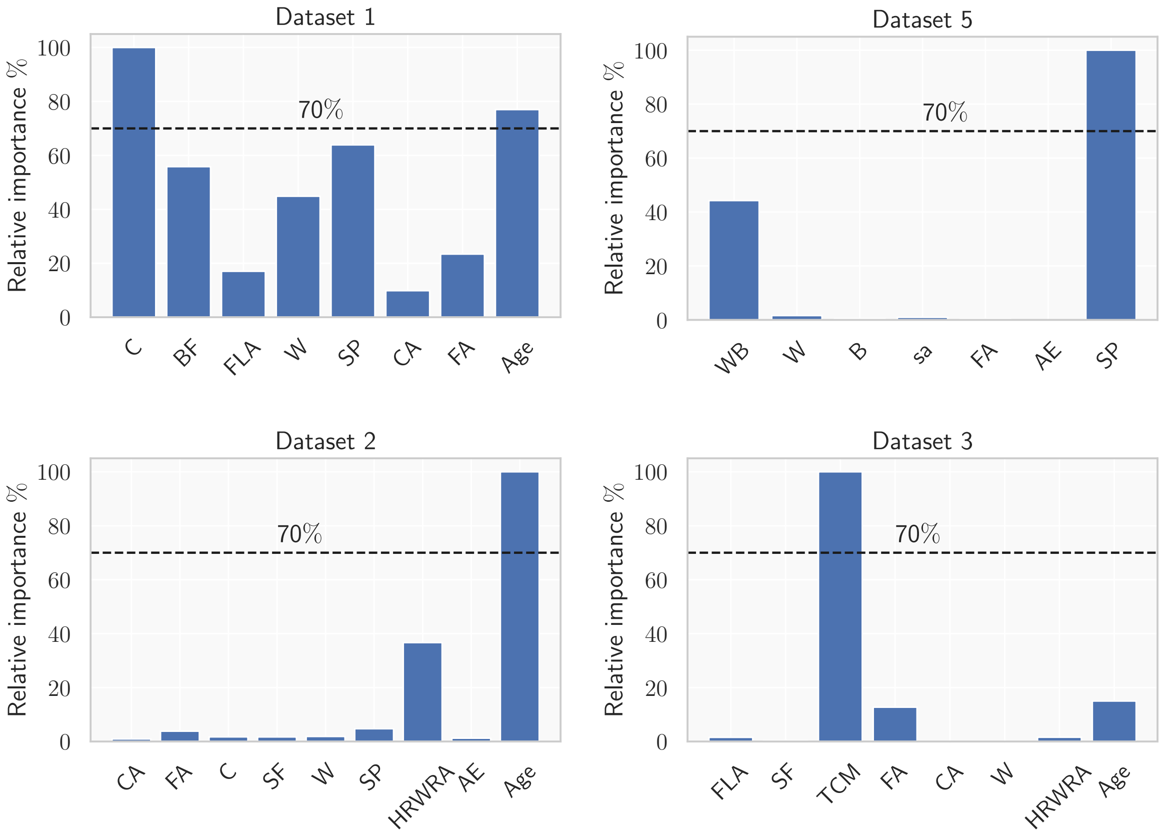

Figure 11: Feature importance of Dataset 1, 2, 3, and 5 extracted from XGBoost library.

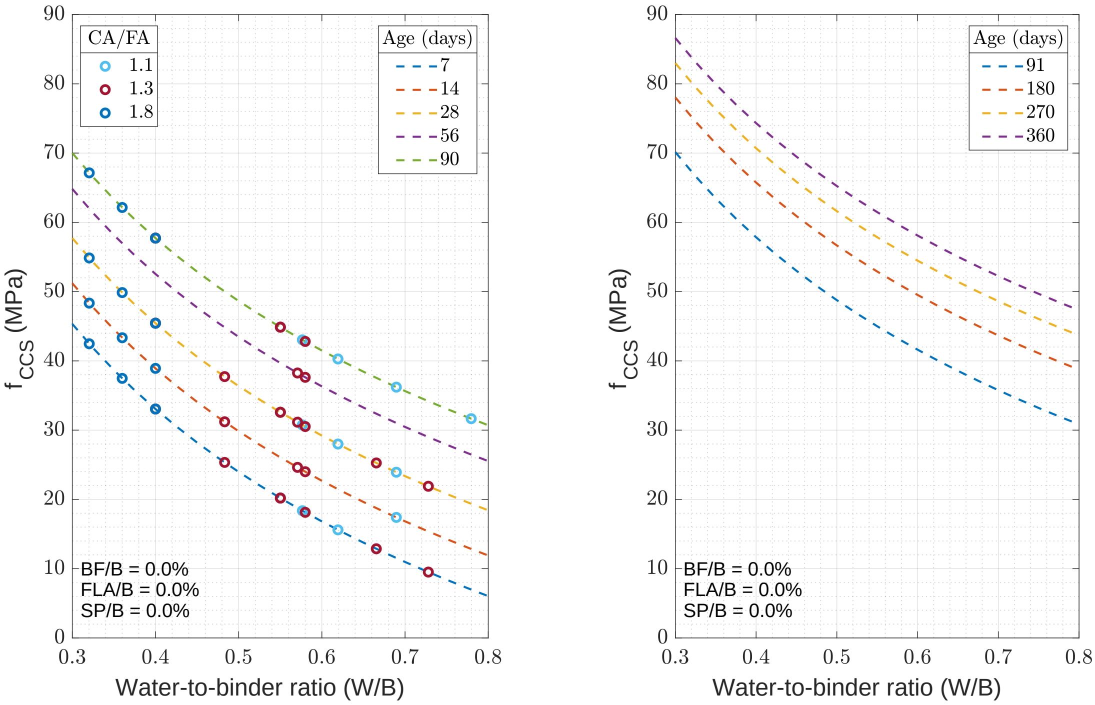

Figure 12: Compressive strength with respect to W/B ratio for different days of Dataset 1 with 1030 samples. The dash-line is for the predicted data while the dots represent the experimental one.

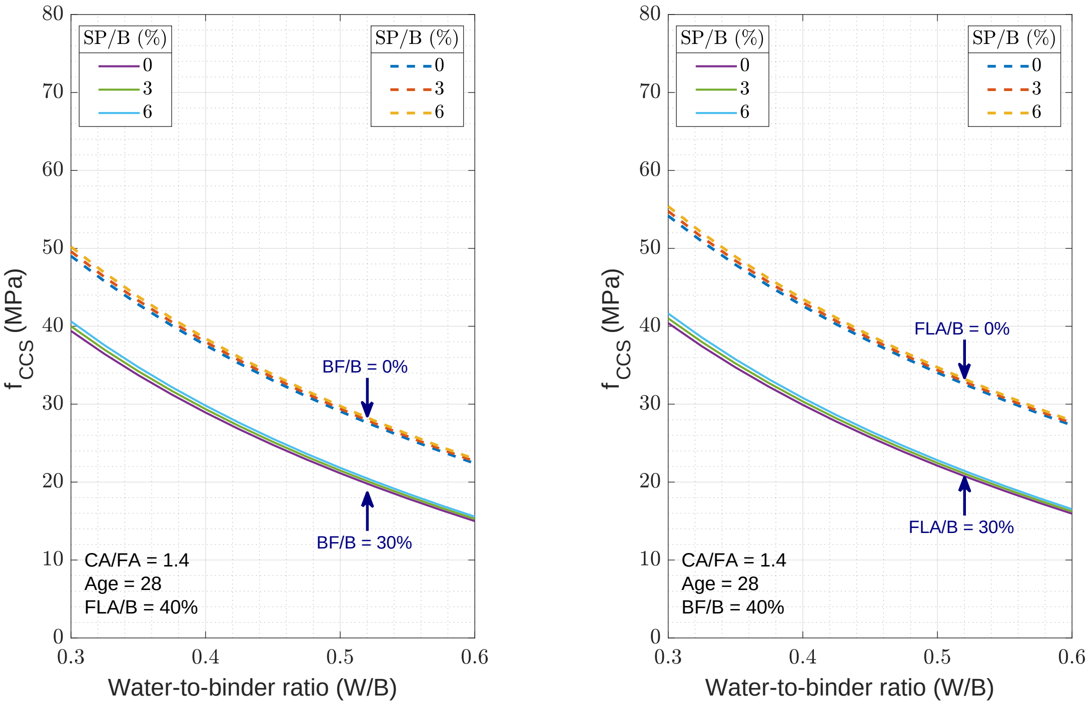

Figure 13: Compressive strength with respect to W/B ratio of Dataset 1 with 1030 samples plotting with different BF/B (left) and FLA/B (right).

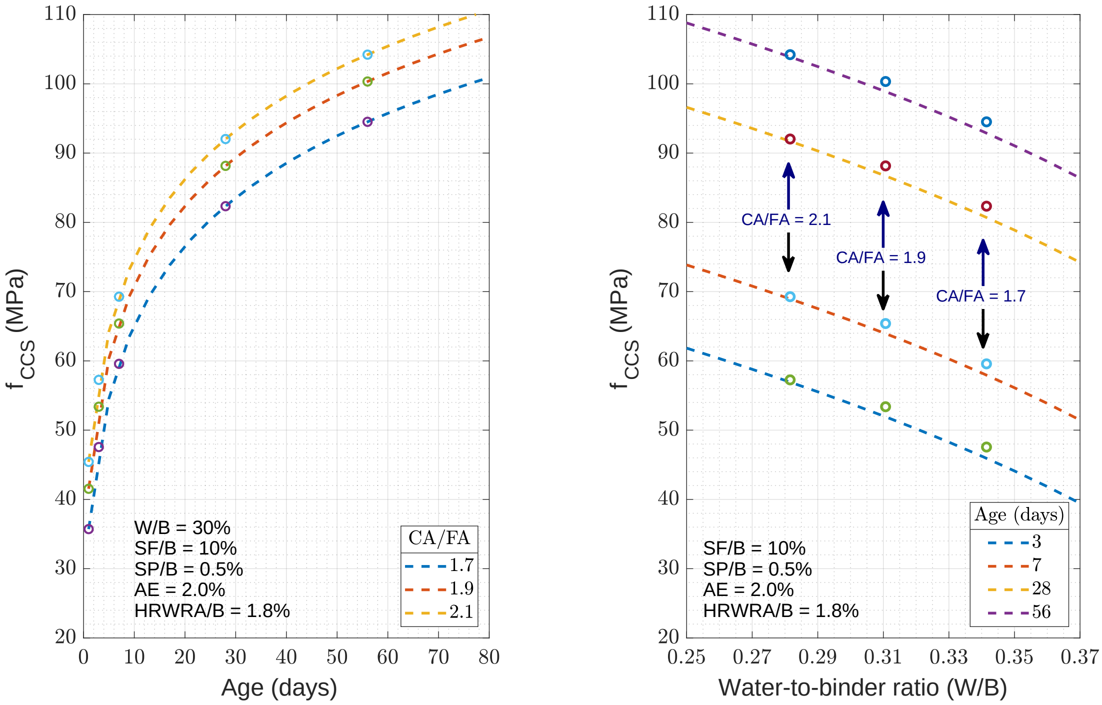

Figure 14: Compressive strength with respect to Age and W/B ratio for Dataset 2 with 200 samples.

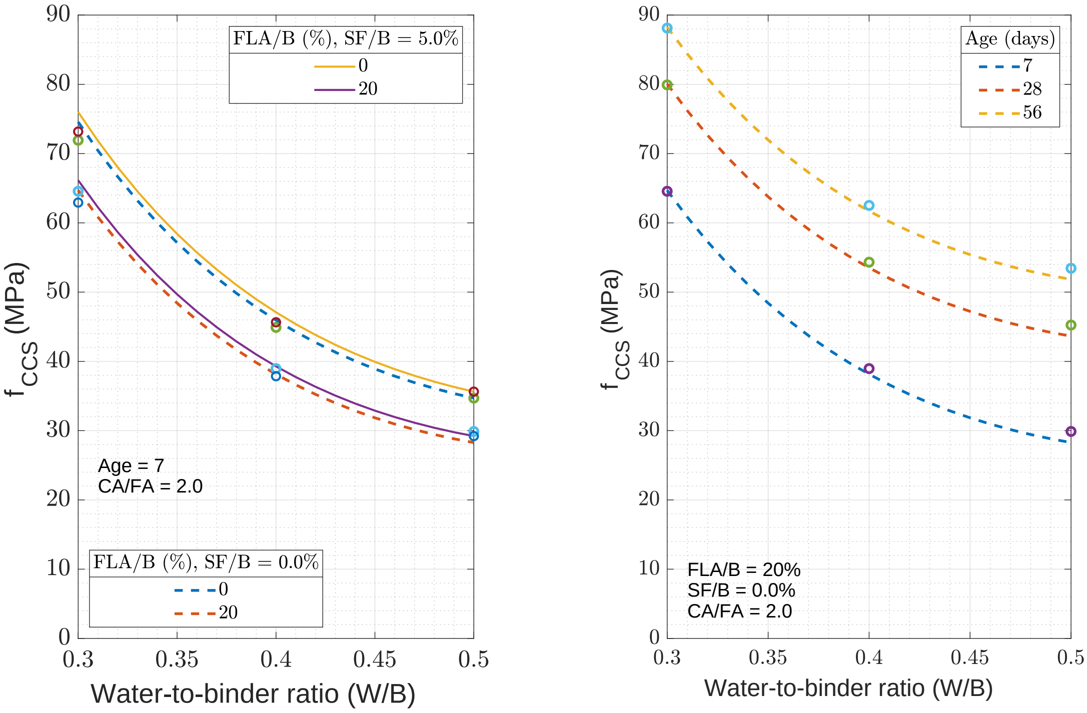

Figure 15: Compressive strength with respect to W/B ratio for Dataset 3 with 144 samples. The dash-line is for the predicted CCS while the dots represent the experimental data.

List of Tables

Table 1: Statistical input variables of the datasets.

Attribute	Unit	Minimum	Maximum	Mean	SD*
Dataset 1: 1030 samples ([12])					
Cement (C)	kg/m^3	102.0	540.0	281.2	104.5
Water content (W)	kg/m^3	121.8	247.0	181.6	21.4
Blast-furnace slag (BF)	kg/m^3	0.0	359.4	73.9	86.3
Fly ash (FLA)	kg/m^3	0.0	200.1	54.2	64.0
Super-plasticizer (SP)	kg/m^3	0.0	32.2	6.2	6.0
Coarse aggregate (CA)	kg/m^3	801.0	1145.0	972.9	77.8
Fine aggregate (FA)	kg/m^3	594.0	992.6	773.6	80.2
Age of testing (Age)	day	1.0	365.0	45.7	63.2
Binder (B)=C+ BF+ FLA	kg/m^3	200.0	640.0	409.2	92.8
W/B	%	23.5	90.0	46.9	12.7
Concrete compressive strength (CCS)	MPa	2.3	82.6	35.8	16.7
Dataset 2: 200 samples [43]					
Cement (C)	kg/m^3	1105.0	1173.0	1136.2	21.4
Water content (W)	kg/m^3	160	168	164.7	2.1
Silica fume (SF)	kg/m^3	0.0	59	24.9	18.2
Super-plasticizer (SP)	kg/m^3	2.2	3.3	2.7	0.3
High rate water reducing agent (HRWRA)	kg/m^3	6.7	14.5	9.3	1.9
Air entraining agent content (AE)	%	1.3	2.5	1.8	0.3
Coarse aggregate (CA	kg/m^3	1105.0	1173.0	1136.2	21.4
Fine aggregate (FA)	kg/m^3	488.0	700.0	602.5	59.1
Age of testing (Age)	day	1.0	56.0	18.9	20.9
Binder (B)=C+ SF	kg/m^3	446.0	661.0	542.5	64.7
W/B	%	25.0	37.1	30.8	3.7
Concrete compressive strength (CCS)	MPa	21.2	113.7	66.8	23.6
Dataset 3: 144 samples [3, 50]					
Total cementitious material (TCM, B)	kg/m^3	400.0	500.0	436.7	45.1
Water content (W)	lt/m^3	150.0	205.0	171.7	24.0
Fly ash replacement (FLA$_\mathrm{R}$)	%	0.0	55.0	25.0	19.1
Silica fume replacement (SF$_\mathrm{R}$)	%	0.0	5.0	1.9	2.4
High rate water reducing agent (HRWRA)	lt/m^3	0.0	13.0	4.9	4.0
Fine aggregate (FA)	kg/m^3	536.0	724.0	639.4	54.9
Coarse aggregate (CA)	kg/m^3	1086.0	1157.0	1125.0	29.5
Age of samples (Age)	day	3.0	180.0	60.7	61.3
W/B	%	30.0	50.0	40.0	8.2
Concrete compressive strength (CCS)	MPa	7.8	107.8	56.6	23.8
Dataset 4: 344 samples [3, 43, 50]					
Cement (C)	kg/m^3	180.0	659.0	434.6	123.5
Water content (W)	kg/m^3	150	205	167.6	16
Silica fume (SF)	kg/m^3	0.0	59	17.9	17.6
Super-plasticizer (SP)	kg/m^3	0	3.3	1.6	1.4
High rate water reducing agent (HRWRA)	kg/m^3	0	14.5	7.4	3.7
Air entraining agent content (AE)	%	1.3	2.5	1.8	0.3
Coarse aggregate (CA	kg/m^3	1186.0	1173.0	1131.5	25.7
Fine aggregate (FA)	kg/m^3	488.0	724.0	617.9	36.4
Age of testing (Age)	day	1.0	180	36.4	47.4
Binder (B)=C + SF + FLA	kg/m^3	400.0	661.0	498.2	77.5
W/B	%	25	50.0	34.6	7.5
Concrete compressive strength (CCS)	MPa	21.2	113.7	66.8	23.6
Dataset 5: 104 samples [30]					
Water-to-binder (W/B)	%	30.0	45.0	37.6	5.6
Water content (W)	kg/m^3	160.0	180.0	170.0	8.2
Fine aggregate to total aggregate (s/a)	%	37.0	53.0	46.0	3.6
Fly ash replacement (FLA$_\mathrm{R}$)	%	0.0	20.0	10.1	8.3
Air entraining agent content (AE)	%	0.0	0.1	0.1	0.0
Super-plasticizer content (SP)	kg/m^3	1.9	8.5	4.5	2.3
Concrete compressive strength (CCS)	MPa	38.0	74.0	52.7	9.4

Table 2: Input variables and four dimensionless variables, $\alpha, \beta, \gamma, \delta$ for five datasets.

Dataset	Input variables	α	β	α	δ
1	C, W, BF, FLA, SP, CA, FA, Age	$\frac{\mathrm{C}}{\mathrm{W}}$	$\frac{\mathrm{FA}}{a_1\mathrm{CA}} + a_2$	$\frac{a_3\mathrm{C} + a_4\mathrm{BF} + a_6\mathrm{FLA}}{a_7\mathrm{W} + a_8\mathrm{SP}}$	$\frac{(1+\mathrm{Age})}{1\mathrm{day}}$
2	C, W, SF, SP, HRWRA AE, CA, FA, Age	$\frac{\mathrm{C}}{\mathrm{W}}$	$\frac{\mathrm{FA}}{a_1\mathrm{CA}} + a_2$	$\frac{a_3\mathrm{C} + a_5\mathrm{SF}}{a_7\mathrm{W} + a_8\mathrm{SP} + a_9\mathrm{HRWRA}} + a_{10}\mathrm{AE}$	$\frac{(1+\mathrm{Age})}{1\mathrm{day}}$
3	TCM, W, FLA$_\mathrm{R}$, SF$_\mathrm{R}$, HRWRA, CA, FA, Age	$\frac{\mathrm{TCM}}{\mathrm{W}}$	$\frac{\mathrm{FA}}{a_1\mathrm{CA}} + a_2$	$\frac{\mathrm{TCM}(a_3 + a_5\mathrm{SF}_R + a_6\mathrm{FLA}_R)}{a_7\mathrm{W} + a_9\mathrm{HRWRA}}$	$\frac{(1+\mathrm{Age})}{1\mathrm{day}}$
4	C, W, SF, FLA, HRWRA, SP, AE, CA, FA, Age	$\frac{\mathrm{C}}{\mathrm{W}}$	$\frac{\mathrm{FA}}{a_1\mathrm{CA}} + a_2$	$\frac{a_3\mathrm{C} + a_5\mathrm{SF} + a_6\mathrm{FLA}}{a_7\mathrm{W} + a_8\mathrm{SP} + a_9\mathrm{HRWRA}} + a_{10}\mathrm{AE}$	$\frac{(1+\mathrm{Age})}{1\mathrm{day}}$
5	W/B, W, s/a, FLA, AE, SP	$\frac{\mathrm{B}}{\mathrm{W}}$	$a_1 s/a + a_2$	$\frac{\mathrm{B}(a_3 + a_6\mathrm{FLA}_R)}{a_7\mathrm{W} + a_8\mathrm{SP}} + a_{10}\mathrm{AE}$	1

Table 3: Coefficients $\mathbf{a}, \mathbf{b}$ of the semi-empirical equations.

Dataset	i	1	2	3	4	5	6	7	8	9	10
1	a	3.6×10^{7}	378.07	12.38	40.62	0	23.94	0.02	-7.4×10^{-3}	0	0
	b	-68.92	91.04	0.23	-2.73	0.72	3.56	1.52	89.62	0.09	0
2	a	-1.27	2.05	0.12	0	0.02	0	-6.33	0.28	0.49	-1.41
	b	-698.02	-106.07	0.44	-22.74	-306.81	101.82	0.7	836.89	0.02	0
3	a	-2.36×10^{3}	2388.23	3.70	0	0.19	-0.32	0.42	0	0.04	0
	b	-276.89	-474.81	0.50	-1.65	-16.55	6.33	3.63	666.93	0.02	0
4	a	-5.7×10^{8}	0.02	1.12	0	0.21	0.09	0.40	-0.49	-1.42	-16.52
	b	-58.61	-0.11	4.58	-0.40	-1.52	4.13	3.37	215.08	0.06	0
5	a	1.28	2.63	0.70	0	0	-0.17	0.29	-1.79	0	-7.64
	b	-29.17	-6.39e-09	17.99	2.35	2.64	2.38	3.76	0	0	0

Table 4: Performance criteria of five datasets with various noise levels.

		R^2		RMSE		MAE		a20-index	
Dataset	Noise (%)	train	test	train	test	train	test	train	test
1	0	0.8444	0.8564	6.2934	5.9786	4.9052	4.4937	0.6776	0.7417
	10	0.8443	0.8567	6.2954	5.9681	4.9058	4.4822	0.6831	0.7527
	50	0.8410	0.8539	6.3800	6.0413	4.9856	4.4903	0.6913	0.7197
	100	0.8289	0.8316	6.6753	6.4774	5.2255	4.9599	0.6762	0.6758
2	0	0.9289	0.9241	6.2006	6.4748	5.1728	5.4531	0.9359	0.8974
	10	0.9287	0.9256	6.2051	6.4263	5.1778	5.3946	0.9359	0.8974
	50	0.9236	0.9176	6.4309	6.7391	5.3266	5.7546	0.9294	0.8974
	100	0.9138	0.9291	6.8443	6.3219	5.6216	5.2241	0.9230	0.8974
3	0	0.9493	0.9544	5.2843	5.2392	4.0145	3.7656	0.9304	0.9310
	10	0.9492	0.9555	5.2976	5.1757	4.0316	3.7117	0.9304	0.9310
	50	0.9515	0.9540	5.1927	5.2147	4.0223	3.8016	0.9217	0.8620
	100	0.9429	0.9518	5.6255	5.5808	4.4103	4.2241	0.9130	0.9310
4	0	0.9109	0.8995	5.5953	6.2951	7.1402	7.8478	0.8401	0.8115
	10	0.9103	0.8996	5.6105	6.2914	7.1587	7.8545	0.8363	0.7971
	50	0.9085	0.8895	5.6584	6.7261	7.2457	8.2731	0.8472	0.7826
	100	0.9029	0.8983	5.8904	6.0330	7.4492	7.7707	0.8254	0.8405
5	0	0.94535	0.9261	1.7454	1.9211	2.2631	2.6879	1.0000	1.0000
	10	0.94146	0.9191	1.7771	2.0597	2.3025	2.8073	1.0000	1.0000
	50	0.87593	0.7774	2.6783	3.8272	3.4739	4.7797	1.0000	1.0000
	100	0.78156	0.6987	4.0305	4.8212	5.0136	6.3677	0.9506	0.9000

Table 5: Comparison of the performance of semi-empirical formula in four datasets.

Dataset	Reference	R^2	RMSE	MAE	a20-index
1	Present	0.8567	5.968	4.482	0.752
	Gandomi and Alavi [34] (GEP[a])	0.8354	-	5.190	-
	Mousavi et al. [32] (Multi-GGP[b])	0.8046	7.310	5.480	-
	Asteris et al. [28] (GPR[c])	0.8858	-	-	0.757
	Chou and Pham [25] (ANNs[d])	0.8649	6.329	4.421	-
2	Present	0.9256	6.426	5.394	0.897
	Chou and Pham [25] (ANNs[d])	0.9584	4.783	3.660	-
	Videla and Gaedicke [43] (Hyp-Exp[e])	0.9600	-	5.000	-
3	Present	0.9555	5.175	3.711	0.931
	Pala et al. [3] (ANNs[e])	0.9980	-	-	-
	Chou and Pham [25] (ANNs[e])	0.9860	5.867	4.992	-
5	Present	0.9191	2.059	2.807	1.000
	Tsai and Lin [31] (WGP[f])	0.9570	2.180	-	-
	Lim et al. [30] (MR[g])	0.9530	-	-	-
	Chou and Pham [25] (ANNs[e])	0.9741	1.548	1.156	-

[a] GEP: Gene Expression Programming.

[b] Multi-GGP: Multi-gene Genetic Programming.

[c] GPR: Gaussian Process Regression.

[d] ANNs: Artificial Neural Networks.

[e] Hyp-Exp: Combination of hyperbolic and exponential equation

[f] WGP: Weighted genetic programming

[g] MR: Multi Regression modeling