

A generalized method to predict the compressive strength of high-performance concrete by improved random forest algorithm

Original

A generalized method to predict the compressive strength of high-performance concrete by improved random forest algorithm / Han, Q.; Gui, C.; Xu, J.; Lacidogna, G.. - In: CONSTRUCTION AND BUILDING MATERIALS. - ISSN 0950-0618. - STAMPA. - 226:(2019), pp. 734-742. [10.1016/j.conbuildmat.2019.07.315]

Availability:

This version is available at: 11583/2782232 since: 2020-01-18T16:08:19Z

Publisher:

Elsevier Ltd

Published

DOI:10.1016/j.conbuildmat.2019.07.315

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

Elsevier postprint/Author's Accepted Manuscript

© 2019. This manuscript version is made available under the CC-BY-NC-ND 4.0 license
<http://creativecommons.org/licenses/by-nc-nd/4.0/>. The final authenticated version is available online at:
<http://dx.doi.org/10.1016/j.conbuildmat.2019.07.315>

(Article begins on next page)

Manuscript Number: CONBUILDMAT-D-19-00519R1

Title: A generalized method to predict the compressive strength of high-performance concrete by improved random forest algorithm

Article Type: VSI:Damage detection, ICEM18

Keywords: Random forest; High-performance concrete; Compressive strength; Input variable optimization; Parameter determination.

Corresponding Author: Professor Jie Xu, Ph.D.

Corresponding Author's Institution: Tianjin University/Key Laboratory of Coast Civil Structure and Safety of Ministry of Education

First Author: Qinghua Han, Prof.

Order of Authors: Qinghua Han, Prof.; Changqing Gui; Jie Xu, Ph.D.; Giuseppe Lacidogna, Prof.

Abstract: The prediction results of high-performance concrete compressive strength (HPCCS) based on machine learning methods are seriously influenced by input variables and model parameters. This study proposes a method with two stages to select proper variables, simplify parameter settings, and predict HPCCS. The appropriate variables are selected in the first stage by measuring their importance based on random forest, and then are optimized to predict HPCCS in the second stage. The results show that the proposed method was effective for input variable optimization, and could return better predictions than that without variable optimization, provided that the parameters are set within a reasonable range. Compared with previous models, the proposed method shows a strong generalization capacity for HPCCS prediction. We find that the prediction performance of the model is better when the input variables are expressed as absolute mass, and the model performs well when the actual compressive strength of HPC is high.

Dear Editor,

Thank you for your attention.

In this study, a method was proposed to optimize input variables, simplify parameter determination, and predict HPCCS. Some interesting conclusion can be drawn:

- 1) The effect of variable forms on HPCCS prediction was compared, and it was found that input variables in the form of either relative mass or absolute mass have little effect on prediction. We suggested the use of the absolute mass of HPC components as input variables to predict HPCCS.
- 2) The proposed method is effective for optimizing input variables. The model built by the proposed method shows a stronger generalization capacity than that built without input variable optimization.
- 3) Random forest exhibits excellent performance for HPCCS prediction even with default parameter settings, which was confirmed by a comparison with previously published models. Moreover, we confirmed that the prediction of HPCCS is insensitive to parameter settings as long as they are set within a reasonable range.
- 4) In terms of computing expense, we recommend using fewer trees and candidate variables for the predictions.
- 5) The model built by the proposed method was inclined to overestimate the compressive strength of samples with actual strengths of less than 30 MPa, but it could accurately predict the compressive strength of samples with actual strengths greater than 30 MPa.

Thank you and best regards.

Sincerely yours,

Xu Jie

Conflict of interest statement

We declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work, there is no professional or other personal interest of any nature or kind in any product, service and/or company that could be construed as influencing the position presented in, or the review of, the manuscript entitled, "A generalized method to predict the compressive strength of high-performance concrete by improved random forest algorithm" .

Jie XU

On behave of all authors

Dear Reviewers,

Thank you for your useful comments and suggestions on our manuscript. We have modified the manuscript accordingly, and detailed corrections are listed below point by point:

Reviewers' comments:

Reviewer #1

This paper presets a very simple analytical study, which is certainly not of a suitable standard for a journal article. The study lacks the depth and scientific rigor, and its original contribution is extremely limited.

Answer: Yes, this paper seems presents a very simply analytical study based on an open dataset for research, but the work is quite meaningful. The main contribution of this paper can be surmised as three points, which are: 1) This study proposes a two-stage method based on random forest algorithm, which simplifies the work of feature engineering and improves the performance of machine learning algorithm on the high-performance concrete compressive strength prediction task; 2) We have experimentally proved that the random forest algorithm can achieve better prediction performance in the high-performance concrete compressive strength prediction task even when training with the default parameter settings; 3) Through comparative experiments, we find that the absolute mass of concrete components is more suitable than the relative mass of concrete components as the input variable of the model.

There are many similar publications in the related area, and all these publications can be divided into two directions. On one side, researchers apply the same method to different datasets, such as using neural network algorithm to predict the compressive strength of concrete with different components. A serious problem is that researchers cannot guarantee the quality of their datasets, whether in terms of the number of samples or the range of variables in the datasets, which leads to the application of their research results is strictly limited. On the other hand, researchers use different algorithms to compare the prediction performance of different algorithms on the same dataset. For example, researchers may use neural network, decision tree and support vector machines to build the compressive strength models of high-performance

concrete.

For the first direction, it is inevitably for the researchers to design input variables and tune model parameters. The proposed method in this study can evaluate the importance of variables and select the suitable variables for modeling, which simplifies the work of variable design. The result listed in Table 2 in the revised manuscript shows that random forest work well under default parameter settings.

Table 2 Prediction performance of the 10 models

Model	<i>R</i>	MAE (MPa)	RMSE (MPa)	MAPE (%)	SI
a-1	0.9623	3.3350	4.6650	12.0640	0.6591
a-2	0.9625	3.4065	4.7203	13.2777	0.9367
a-3	0.9637	3.2577	4.5281	12.2165	0.4289
a-4	0.9655	3.1055	4.4339	11.7850	0.0595
a-5	0.9662	3.1703	4.4455	11.8262	0.0974
b-1	0.9613	3.2228	4.6267	12.1511	0.5976
b-2	0.9655	3.3147	4.5432	12.5760	0.4464
b-3	0.9644	3.2078	4.4967	12.0204	0.2925
b-4	0.9622	3.1975	4.5481	11.7889	0.4073
b-5	0.9627	3.2150	4.6044	11.6018	0.4173

For the second direction, the researchers still need to select the optimal parameters. The result listed in Table 4 in this study shows that the random forest algorithm achieves the optimal value in many evaluation indicators after parameters optimization.

Table 4 Statistical results for the number of samples in each subgroup, group and the corresponding proportion.

Actual strength (MPa)	Percentage error (%)	Number of samples in subgroup	Number of samples in group	Proportion (%)
[0,30]	$(-\infty, -10)$	174	2018	8.62
	$[-10, 10]$	902		44.70
	$(10, +\infty)$	942		46.68
[30,82.6]	$(-\infty, -10)$	467	3132	14.91
	$[-10, 10]$	2416		77.14
	$(10, +\infty)$	249		7.95
[0,82.6]	$(-\infty, -10)$	641	5150	12.45
	$[-10, 10]$	3318		64.43
	$(10, +\infty)$	1191		23.12

Furthermore, by comparing the effect of input variable representations on model

performance, we find that the model performs well when the input variables are represented as absolute mass forms rather than relative mass forms. Therefore, we recommend using the absolute mass of concrete components to establish the concrete compressive strength model.

Reviewer #2

In the manuscript entitled, "A generalized method to predict the compressive strength of high performance concrete by improved random forest algorithm" authors have done interesting work, nicely planned and well description of the content in the current version of manuscript. In this manuscript, authors successfully used a method with two stages based on random forest (RF) to optimize the input variables, simplify parameter determination to predict high-performance concrete compressive strength (HPCCS) and conclude that optimized RF model works better than other models. But before acceptance in this reputed Journal, I have few minor suggestions:

(1) Describe the Bagging techniques in detail.

Answer: Thanks for high evaluation of this manuscript. The authors reintroduce the bagging method in section 2.1.1 in the revised manuscript. The bagging method can be divided into two parts: bootstrap and aggregation. In the first part, the authors introduce how to generate a new dataset, that is, to sample from the original dataset with playback, ensuring that the size of the new dataset is the same as that of the original dataset. The authors also explain why about 36.8% of the samples in the original dataset do not appear in the new dataset.

The second part is the aggregation operation. For regression tasks, the average method is usually used, that is, the output of multiple predictors is averaged to get the final output. The way to generate the predictor, introduced in section in 2.1.3 in the revised manuscript, does not belong to bagging method.

(2) Figure 1 and 2 is not readable, please improve the quality of the figures.

Answer: Figures 1 and 2 have been improved with high quality in the revised manuscript and also listed below.

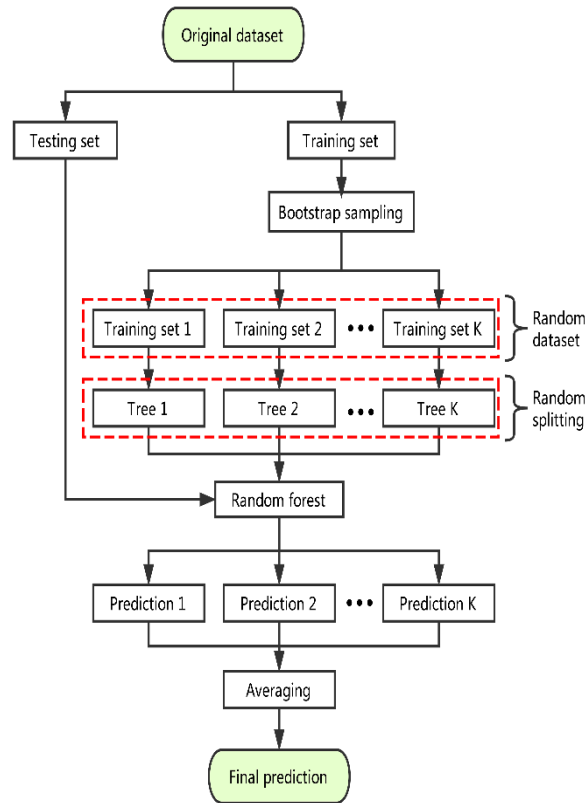


Fig. 1. Schematic of random forest generation and prediction

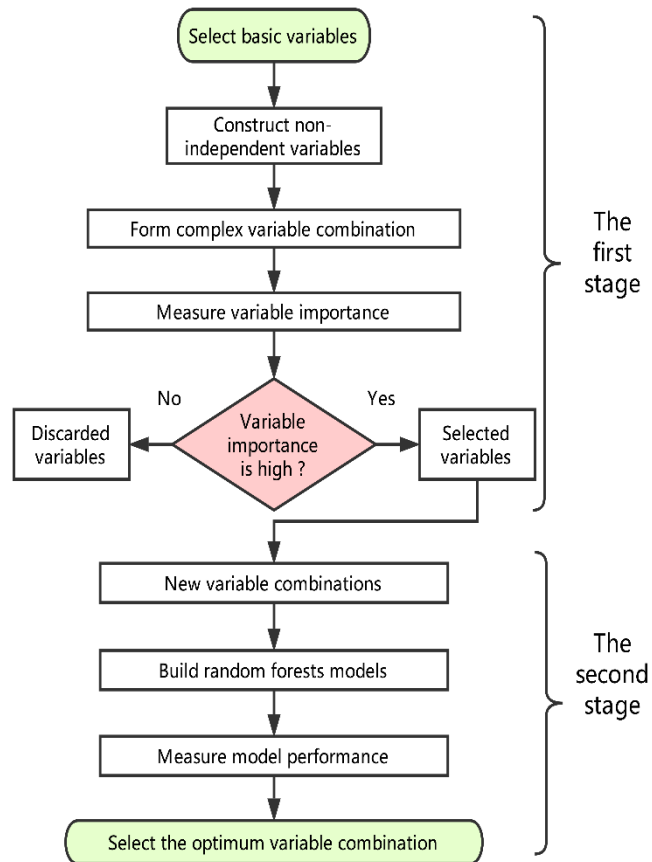


Fig. 2. Flowchart of the proposed method

(3) The authors collected the dataset from researches published between 1998 and 2014, why they didn't used new resources?

Answer: The dataset used in this research is a famous open source dataset, which is often used in the research of concrete compressive strength prediction. The dataset can be downloaded from <http://archive.ics.uci.edu/ml/>.

In order to ensure the comparability of the experimental results, we only collected data from researches using the same dataset with us. In recent years, there have been many researches on the prediction of concrete compressive strength, but we have not found research after 2014 using this dataset in high-level journals. That is why we collected the dataset from researches published between 1998 and 2014.

(4) In Table 2 and 6, what is the significance of the bold values?

Answer: The significance of the bold values in Tables 2 and 6 means “The best result for each performance measure is given in bold type”. We have explained the meaning of bold values in Table 2 in line 297 and the meaning of bold values in Table 6 in line 355-357 in the revised manuscript.

(5) Rewrite the Abstract in another way with summarization of all the finding.

Answer: The abstract has been improved with summarization of all the finding according to the suggestion in the revised manuscript. As it is shown below:

Abstract. The prediction results of high-performance concrete compressive strength (HPCCS) based on machine learning methods are seriously influenced by input variables and model parameters. This study proposes a method with two stages to select proper variables, simplify parameter settings, and predict HPCCS. The appropriate variables are selected in the first stage by measuring their importance based on random forest, and then are optimized to predict HPCCS in the second stage. The results show that the proposed method was effective for input variable optimization, and could return better predictions than that without variable optimization, provided that the parameters are set within a reasonable range. Compared with previous models, the proposed method shows a strong generalization

capacity for HPCCS prediction. We find that the prediction performance of the model is better when the input variables are expressed as absolute mass, and the model performs well when the actual compressive strength of HPC is high.

(6) Please add some recent literature (2018, 2019) in the manuscript.

Answer: Thanks and three related recent literatures were added in the revised manuscript. Two articles listed below on the application of high-performance concrete published in 2019 has been added in the revised manuscript.

Wetzel, A and Middendorf, B. (2019). "Influence of silica fume on properties of fresh and hardened ultra-high performance concrete based on alkali-activated slag. " *CEMENT CONCRETE COMP* **100**: 53-59.

Zhu, H., Wang, Z. J., Xu, J. and Han, Q. H. (2019). "Microporous structures and compressive strength of high-performance rubber concrete with internal curing agent." *CONSTR BUILD MATER* **215**: 128-134.

The third article listed below on predicting the high performance concrete compressive strength by using machine learning method has been added in the revised manuscript.

Bui, DK., Nguyen, T., Chou, J.-S., Nguyen-Xuan, H., and Ngo, TD. (2018). "A modified firefly algorithm-artificial neural network expert system for predicting compressive and tensile strength of high-performance concrete. " *CONSTR BUILD MATER* **180**: 320-333.

Reviewer #3

(1) The study is about using "Random Forest" computational method in predicting the strength of high-performance concrete strength. I believe that this typical paper is best submitted to the journal addressing computing or computational method for engineering application. Hence, I will suggest to the editor to encourage them to submit the paper to another journal of relevance.

Answer: The research in this manuscript seems suitable to the journal addressing computing or computational method for engineering application, while it is also one of the main scope of CBM. The authors have found the corresponding articles (listed

below) using machine learning algorithms to predict the compressive strength of concrete and similar topics were published in CBM, and some of them are cited in our manuscript.

Ayaz, Y., Kocamaz, A. F. and Karakoç, M. B. (2015). "Modeling of compressive strength and UPV of high-volume mineral-admixed concrete using rule-based M5 rule and tree model M5P classifiers." *CONSTR BUILD MATER* **94**: 235-240.

Behnood, A., Behnood, V., Modiri Gharehveran, M. and Alyamac, K. E. (2017). "Prediction of the compressive strength of normal and high-performance concretes using M5P model tree algorithm." *CONSTR BUILD MATER* **142**: 199-207.

Chithra, S., Kumar, S. R. R. S., Chinnaraju, K. and Alfin Ashmita, F. (2016). "A comparative study on the compressive strength prediction models for High Performance Concrete containing nano silica and copper slag using regression analysis and Artificial Neural Networks." *CONSTR BUILD MATER* **114**: 528-535.

Chou, J.-S. and Pham, A.-D. (2013). "Enhanced artificial intelligence for ensemble approach to predicting high performance concrete compressive strength." *CONSTR BUILD MATER* **49**: 554-563.

Chou, J.-S., Tsai, C.-F., Pham, A.-D. and Lu, Y.-H. (2014). "Machine learning in concrete strength simulations: Multi-nation data analytics." *CONSTR BUILD MATER* **73**: 771-780.

Kalman Šipoš, T., Miličević, I. and Siddique, R. (2017). "Model for mix design of brick aggregate concrete based on neural network modelling." *CONSTR BUILD MATER* **148**: 757-769.

Qi, C., Fourie, A. and Chen, Q. (2018). "Neural network and particle swarm optimization for predicting the unconfined compressive strength of cemented paste backfill." *CONSTR BUILD MATER* **159**: 473-478.

Safarzaghan Gilan, S., Bahrami Jovein, H. and Ramezani pour, A. A. (2012). "Hybrid support vector regression – Particle swarm optimization for prediction of compressive strength and RCPT of concretes containing metakaolin." *CONSTR BUILD MATER* **34**: 321-329.

Sonebi, M., Cevik, A., Grünwald, S. and Walraven, J. (2016). "Modelling the fresh properties of self-compacting concrete using support vector machine approach." *CONSTR BUILD MATER* **106**: 55-64.

Othman, H., Marzouk, H. and Sherif, M. (2019). "Effects of variations in compressive

strength and fibre content on dynamic properties of ultra-high performance fibre-reinforced concrete. " *CONSTR BUILD MATER* **195**: 547-556.

Emamian, Seyed Ali and Eskandari-Naddaf, Hamid. (2019). "Effect of porosity on predicting compressive and flexural strength of cement mortar containing micro and nano-silica by ANN and GEP. " *CONSTR BUILD MATER* **218**: 8-27.

In this case, the topic of this manuscript is just in the scope of the CBM, and our manuscript can be published in CBM journal. By the way, the English language, grammar, punctuation, spelling, and overall style has been edited throughout the manuscript by the qualified native English speaking.

The manuscript has been resubmitted to your journal. We look forward to your positive response.

Sincerely,

Jie XU

Highlights.

- An improved random forest method was proposed to predict HPCCS
- Appropriate features for modeling can be obtained by this method
- Satisfactory results with default parameter settings can be obtained
- It performs well when the input variables in absolute mass form
- The prediction accuracy is superior to that of other methods

1 A generalized method to predict the compressive strength of
2 high-performance concrete by improved random forest
3 algorithm

4
5 Qinghua Han^{1, 2}, Changqing Gui², Jie Xu^{1, 2*}, Giuseppe Lacidogna³
6

7 ¹*Key Laboratory of Earthquake Engineering Simulation and Seismic Resilience of China Earthquake*

8 *Administration (Tianjin University), Tianjin 300350, China;*

9 ²*School of Civil Engineering, Tianjin University/Key Laboratory of Coast Civil Structure Safety of China*

10 *Ministry of Education, Tianjin University, Tianjin, 300072, China*

11 ³*Department of Structural, Geotechnical and Building Construction, Politecnico di Torino, Turin, 10129, Italy*

12 *Corresponding author. Tel.: +86-22-27400843; Fax: +86-22-27404319.

13 E-mail addresses: jxu@tju.edu.cn.

14
15 **Abstract.** The prediction results of high-performance concrete compressive strength (HPCCS) based
16 on machine learning methods are seriously influenced by input variables and model parameters. This
17 study proposes a method with two stages to select proper variables, simplify parameter settings, and
18 predict HPCCS. The appropriate variables are selected in the first stage by measuring their importance
19 based on random forest, and then are optimized to predict HPCCS in the second stage. The results
20 show that the proposed method was effective for input variable optimization, and could return better
21 predictions than that without variable optimization, provided that the parameters are set within a
22 reasonable range. Compared with previous models, the proposed method shows a strong generalization
23 capacity for HPCCS prediction. We find that the prediction performance of the model is better when
24 the input variables are expressed as absolute mass, and the model performs well when the actual
25 compressive strength of HPC is high.
26

27 **Keywords:** Random forest; High-performance concrete; Compressive strength; Input variable
28 optimization; Parameter determination.

29
30 **1. Introduction**

31

32 In recent years, the application of high-performance concrete (HPC) has increased markedly in
33 the construction industry (Lim et al. 2004, Chiew et al. 2017, [Wetzel and Middendorf. 2019](#), [Zhu
34 et al. 2019](#)). HPC has many attractive advantages, such as sufficient workability, high strength,
35 and excellent durability. However, chemical admixtures and additional supplementary
36 cementitious materials such as fly ash, blast-furnace slag, silica fume, and superplasticizer are
37 usually necessary to make HPC (Chang et al. 1996, Yeh 1998, Bharatkumar et al. 2001, Lim et al.
38 2004), which can pose a challenge for accurately predicting the compressive strength of HPC.

39 Standard compression tests can determine the actual compressive strength of HPC. However,
40 this is a time-consuming, cumbersome, and costly method for determination of high-performance
41 concrete compressive strength (HPCCS). The empirical formula employed generally introduces
42 various regression coefficients to represent the effects of different added materials. As a result, the
43 prediction ability of this empirical formula is doubtful, as the relationship between the
44 compressive strength of HPC and its components is highly nonlinear.

45 Emerging machine learning techniques provide an opportunity to predict HPCCS accurately.
46 Many machine learning algorithms have been used to predict the compressive strength of HPC in
47 the last two decades, including artificial neural network (Yeh 1998, Sebastia et al. 2003, Chithra et
48 al. 2016, [Bui et al. 2018](#)), support vector machine (Chou and Pham 2015, Sonebi et al. 2016),
49 decision tree (Cheng et al. 2014, Ayaz et al. 2015, Behnood et al. 2017), and ensemble algorithm
50 (Chou and Tsai 2012, Chou and Pham 2013, Erdal 2013, Erdal et al. 2013, Omran et al. 2016).
51 These studies demonstrated that models based on machine learning algorithms can obtain better
52 predictions than those based on regression analysis, and models based on an ensemble algorithm
53 perform best if the base predictors were selected properly (Chou et al. 2014).

54 However, determining the proper base predictors is not an easy task, and numerous
55 experiments are necessary to acquire suitable predictors. Moreover, the influence of input
56 variables and parameter settings on the prediction accuracy should also be considered.

57 On one hand, the model prediction accuracy is related to the input variables and does not
58 necessarily improve with increasing the number of input variables (Matin et al. 2017), and it may
59 be influenced by the variable forms. For the prediction of HPCCS, there are no clear conclusions
60 about what number of input variables and which form of these variables are appropriate. Most
61 studies regarded the absolute mass of the HPC ingredients as input variables, while some studies
62 used the relative mass of the HPC ingredients as input variables (Behnood et al. 2017, Kalman
63 sipos et al. 2017).

64 On the other hand, it is tedious work to determine the proper parameter settings, which have
65 great influence on the model prediction accuracy. Manual tuning requires a great deal of time and
66 attention. Some scholars have summed up some empirical formulas; however, the results of these
67 empirical formulas are often different (Kalman sipos et al. 2017). The application of an

68 optimization algorithm can assist in determining the appropriate parameter settings, which
69 increases the complexity of the model (Safarzadegan Gilan et al. 2012, Chou and Pham 2015, Qi
70 et al. 2018).

71 Assuming that the number of combinations of base predictors, input variables, and parameter
72 settings are *a*, *b*, and *c*, respectively, then the prediction accuracy of the *abc* models should be
73 compared to obtain the model with the strongest generalization capacity. This study aims to
74 establish a convenient but effective method to optimize input variables, simplify parameter
75 determination, and predict HPCCS.

76 Random forest (RF) is one of the most advanced ensemble algorithms, and has the attractive
77 features of variable importance measures (VIMs), few model parameters, and robust resistance to
78 overfitting (Breiman 2001, Auret and Aldrich 2012). As its name implies, the decision tree is the
79 base predictor of RF. Models built using RF can return satisfactory results even with default
80 parameter settings (Svetnik et al. 2004). Utilizing RF allows the number of combinations of base
81 predictors and parameter settings to be reduced to one. Notable applications of RF can be found in
82 the fields such as ecology (Krkač et al. 2016, Dubeau et al. 2017, Fu et al. 2017) and
83 bioinformatics (Hanselmann et al. 2009, Schwarz et al. 2011, Boulesteix et al. 2012), but is has
84 rarely been applied to concrete (Maghrebi et al. 2016, Mohamed et al. 2017, Ozcan et al. 2017,
85 Rao 2017). Mohamed applied the RF algorithm to sustainable self-consolidating concrete
86 compressive strength prediction (Mohamed et al. 2017). Ozcan et al. built a RF model to evaluate
87 the effects of blast furnace slag and waste tire rubber powder on HPCCS (Ozcan et al. 2017). Rao
88 used various algorithms to predict the compressive strength of HPC and found that the RF model
89 had the best performance (Rao 2017).

90 These previous studies all focused on the applicability of RF for HPCCS prediction, but did not
91 mention that RF models can obtain precise predictions with no parameter tuning, which is
92 emphasized and validated in this study. This study uses a RF model to predict HPCCS with default
93 parameter settings, thereby avoiding model parameter tuning. Moreover, this study goes one step
94 further than other recent studies by providing an efficient and understandable approach for
95 optimizing model input variables for HPCCS prediction. The effects of the variable forms and
96 quantity of variables on the model prediction are also considered.

97
98

99 **2. Methods**

100

101 Random forest is a combination of multiple decision trees in which each tree is built by a new
102 training set sampled from the original training set based on the bagging method (Breiman 1996,
103 Breiman 2001). The bagging method and classification and regression tree (CART) method are the
104 basis of RF. Therefore, these two methods are first introduced, and then the concepts of RF are

105 discussed. The proposed method optimizes the model input variables based on RF, which is
106 introduced at the end.

107

108 2.1. Machine learning techniques

109

110 2.1.1 Bagging method

111

112 The bagging method, also known as bootstrap aggregation method, is an ensemble technology
113 of training S predictors separately by resampling S new datasets from the original dataset by
114 sampling with playback. That is, duplicate data is allowed in the datasets trained by these models.

115 This method consists of two steps: bootstrap and aggregation. In the first step, S new
116 independent and identically distributed datasets are generated by resampling the original dataset
117 randomly. The number of samples in each new dataset is the same as that in original dataset. This
118 means that the sample of 36.8% in the original dataset will not appear in each new dataset as

$$119 \lim_{n \rightarrow +\infty} \left(1 - \frac{1}{n}\right)^n = \frac{1}{e}$$

120

(1)

121 where n is the number of samples in original (new) dataset. In the second step, the new datasets
122 are used to train the base predictors independently, and aggregation method is used to obtain the
123 final results by averaging the predictions of each tree predictor (Breiman 1996).

124

125 2.1.2 Classification and regression tree

126

127 The classification and regression tree method was proposed to solve classification and
128 regression problems (Breiman et al. 1984). The CART model is built by a recursive binary
129 partitioning of the input space into several subspaces, and fitting a simple prediction model within
130 each partition (Loh 2011), thus forming several nodes. The splitting criterion for each node except
131 the leaf node is determined according to the purity of the resulting nodes. The mean squared error
132 (MSE) around the mean response of the node is widely used as a measure of node purity for
133 regression. The maximum gain in the MSE is used to select the splitting variable and the
134 segmentation point of each node as follows:

$$135 \Delta \text{MSE}(S, x_j^a) = \text{MSE}(S) - \frac{|S_1|}{|S|} \text{MSE}(S_1) - \frac{|S_2|}{|S|} \text{MSE}(S_2) \quad (2)$$

$$136 \text{MSE}(S) = \frac{1}{|S|} \sum_{i=1}^{|S|} (y_i - \hat{y})^2 \quad (3)$$

$$137 \hat{y} = \frac{1}{|S|} \sum_{i=1}^{|S|} y_i \quad (4)$$

138 where $|s|$ is the number of samples in dataset S that reach the node; S_i is the dataset resulting from
 139 splitting at the node, which falls into a subspace according to the given variable x_j ($j=1, 2, \dots, M$)
 140 and segmentation α ; and y_i is the response value of the i th sample in dataset S .

141 The partitioning will continue until the total maximum **MSE** gain is reached. Once the tree has
 142 been built, the response of any sample can be predicted by following the path to the appropriate
 143 leaf node and averaging the responses in this node.

144

145 2.1.3 Random forest

146

147 Random forest is implemented based on bagging decision trees by employing random split
 148 selection (Breiman 2001). **Fig.1** shows a schematic of the generation and prediction of the RF
 149 model. Each tree in the forest is built by a random training set, and each split within each tree is
 150 created based on a subset of input variables which are selected randomly (Grömping 2009). The
 151 introduction of this randomness increases the diversity of the trees. All of the trees in the forest are
 152 fully-grown binary trees.

153

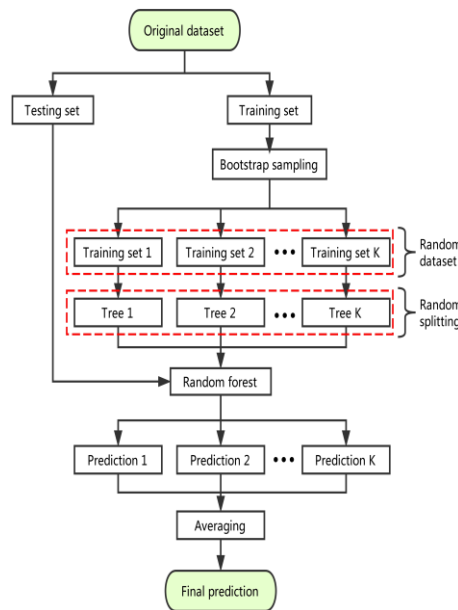


Fig. 1. Schematic of random forest generation and prediction

154

155 Variable importance measures (VIMs) are an inherent product of RF. The basic concept of
 156 VIMs is that if an input variable, x_j , has an impact on the response, the prediction accuracy of the
 157 model will decrease with permutation of the values of variable x_j . As a result, the values of
 158 variables are permuted one at a time and the resulting reduction in prediction accuracy of the new
 159 model is evaluated; the greater the decrease in prediction accuracy, the stronger the association

160 between the permuted variable and the response. Generally, reduction in **MSE** has been used as
 161 the evaluation index. In RF, the out-of-bag (OOB) samples are permuted to measure variable
 162 importance in order to avoid training new forests (Archer and Kimes 2008, Auret and Aldrich
 163 2012). For variable x_j in tree i , the reduction in MSE can be calculated as follows:

$$164 \quad Imp_j^i = \text{MSE}(T_{D(\theta_i)}) - \text{MSE}(T_{D_{\text{OOB}}^j(\theta_i)}) \quad (5)$$

165 where Imp_j^i is the reduction in **MSE** of variable x_j in tree i ; $T_{D(\theta_i)}$ is the i th tree predictor
 166 depending on $D(\theta_i)$, which indicates both bagged samples and random splits in tree i ; $D_{\text{OOB}}^j(\theta_i)$
 167 represents variable x_j in the OOB samples in $D(\theta_i)$ which is permuted. Averaging the results of
 168 all K tree predictors in the forest yields the final MSE reduction of variable x_j :

$$169 \quad Imp_j = \frac{1}{K} \sum_{i=1}^K Imp_j^i \quad (6)$$

170 To express this more intuitively, the relative **MSE** reduction (**RMR**) of each variable was
 171 adopted to measure variable importance. The **RMR** of variable x_j is expressed as:

$$172 \quad RMR_j = \frac{Imp_j}{\sum_{i=1}^M Imp_i} \quad (7)$$

173

174 *2.2 Proposed method*

175

176 The proposed method with two stages inherits the advantages of RF and can be utilized to
 177 optimize model input variables. A flowchart of the proposed method is shown in **Fig. 2**.

178 In the first stage, candidate input variables are selected based on the VIMs of RF. Some
 179 frequently used independent variables are chosen as basic variables first, and then some
 180 non-independent variables are constructed based on these basic variables. Next, the basic variables
 181 are combined with the constructed variables to form a complex variable combination, and the
 182 importance of the input variables in this combination is measured. Finally, the constructed
 183 variables with low **RMR** values are eliminated. The remaining variables are the candidate input
 184 variables. We suggest eliminating the variables whose **RMR** is less than 50% of the average **RMR**.

185 In the second stage, the input variables are optimized and the sample response is predicted.
 186 First, the remaining constructed variables are added to the combination of basic variable to form
 187 new combinations. These new combinations are then used build RF models. The prediction
 188 accuracy of the RF models built with different combinations are compared using the performance
 189 measures introduced in Section 3.3 to select the optimal model. The input variables of this optimal
 190 model are the optimal input variables. Meanwhile, optimal model predictions can be obtained.

191

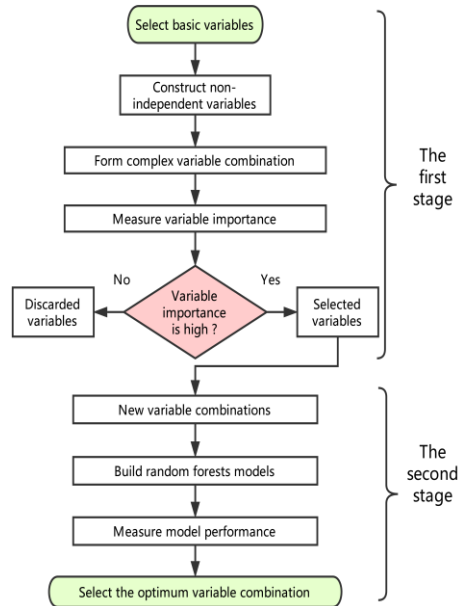


Fig. 2. Flowchart of the proposed method

192

193

194

195

196

197

198

199

3. Materials and modeling setting

200

201

3.1. Dataset

202

203

204

205

206

207

208

209

210

211

The original dataset was collected by Yeh from 17 different sources, and contains 1030 samples made with ordinary Portland cement and cured under normal conditions (Yeh 1998). This dataset can be downloaded from the UCI machine learning repository. All of the specimen types were converted into 15 cm cylinders through accepted methods. This dataset has been used to investigate HPCCS by many researchers and has proven to be robust (Yeh 1998, Chou et al. 2011, Chou and Tsai 2012, Chou and Pham 2013, Erdal 2013, Erdal et al. 2013, Chou et al. 2014). The variables in the original dataset are cement (C), blast furnace slag (BFS), fly ash (Fa), water (W), superplasticizer (SP), coarse aggregate (CA), fine aggregate (FA), age (Age), and concrete compressive strength (CCS). The first seven variables are independent input variables, while the

212 CCS is the response variable. Statistical information about these variables can be found in the
 213 literature (Erdal 2013, Erdal et al. 2013).

214

215 3.2. Input variable combinations

216

217 According to the proposed method, all of the input variables in the original dataset were
 218 designated as basic variables, and five non-independent variables were constructed to form a
 219 complex variable combination. The constructed variables are W/B, BFS/W, Fa/W, CA/B and
 220 CA/FA, where W/B is the water–binder ratio. The ranges of these constructed variables are listed
 221 in **Table 1**.

222

Table 1 Ranges of constructed variables

Constructed variable	Min	Max	Avg	Standard deviation
W/B	0.235	0.900	0.469	0.127
BFS/W	0	1.935	0.407	0.472
Fa/W	0	1.346	0.313	0.376
CA/B	1.284	5.625	2.521	0.680
CA/FA	0.859	1.875	1.274	0.186

223

224 The effect of variable forms on model prediction was considered in this study. Two groups of
 225 variable combinations were constructed: the absolute mass group (group A), and a relative mass
 226 group (group B). The basic variable combinations and complex variable combinations in group A
 227 are referred to as A-1 and A-2, respectively. The corresponding models are named a-1 and a-2,
 228 respectively. The variable combinations and models in group B are named similarly.

229

230 3.3. Performance evaluation

231

232 Four frequently used performance measures were selected, and a synthesis index was designed
 233 to evaluate the generalization capacity of the model.

234

235 3.3.1. Linear correlation coefficient

236

$$R = \frac{n \sum y y' - (\sum y)(\sum y')}{\sqrt{n(\sum y^2) - (\sum y)^2} \sqrt{n(\sum y'^2) - (\sum y')^2}} \quad (8)$$

237 where y is the actual response, y' is the predicted response, and n is the number of samples in the
 238 testing set.

239

240 3.3.2. Mean absolute error

$$241 \quad \text{MAE} = \frac{1}{n} \sum |y - y'| \quad (9)$$

242

243 3.3.3. Root mean squared error

$$244 \quad \text{RMSE} = \sqrt{\frac{1}{n} \sum |y - y'|^2} \quad (10)$$

245

246 3.3.4. Mean absolute percentage error

$$247 \quad \text{MAPE} = \frac{1}{n} \sum \left| \frac{y - y'}{y} \right| \quad (11)$$

248

249 3.3.5. Synthesis index

$$250 \quad \text{SI} = \frac{1}{k} \left(\frac{R_j - R_{\min}}{R_{\max} - R_{\min}} + \sum_{i=2}^k \frac{P_{i,\max} - P_{i,j}}{P_{i,\max} - P_{i,\min}} \right) \quad (12)$$

251 where k is the number of performance measures, and $P_{i,j}$ is the i th performance measure of the
252 j th model except R .

253

254 3.4. Modelling setting

255

256 The number of trees in the forest, n_{tree} , and the number of selected candidate variables when
257 the node is splitting, m_{try} , are two essential parameters which need to be set in the random forest
258 model. The default parameter settings of n_{tree} and m_{try} were 500 and the minimum integer that is
259 greater than \log_2^M , respectively. To obtain precise measurements of variable importance, a large
260 number of trees are needed (Grömping 2009). As a result, n_{tree} was set to 1000 for the variable
261 importance measurement. A total of 70 sets of parameter settings were constructed to verify the
262 robustness of the RF model. The values of n_{tree} range from 100 to 1000 in increments of 100,
263 while m_{try} ranges from 3 to 9 with an increment of 1. Each dataset is divided into two subsets (i.e.,
264 the training set and the testing set). The training set consists of 927 samples selected randomly
265 from the dataset, and the testing set contains the remaining 103 samples. Each RF model was run
266 50 times, and the average was taken as the final result to limit bias due to the random sampling.

267

268

269 4. Results and discussion

270

271 4.1. Results of the first stage

272

273 The results of the variable importance measures of a-2 are shown in **Fig. 3a**, and indicate that
 274 Age, W/B, cement content, and CA/B have the greatest influence on HPCCS, which is consistent
 275 with our cognition. Of these, Age has the most prominent influence with a RMR of 0.3, followed
 276 by W/B with a RMR of 0.25. The average RMR in this combination is approximately 0.08; thus,
 277 BFS/W, Fa/W, and CA/FA were eliminated because their RMR were all less than 0.02, or about
 278 one fourth of the average value.
 279

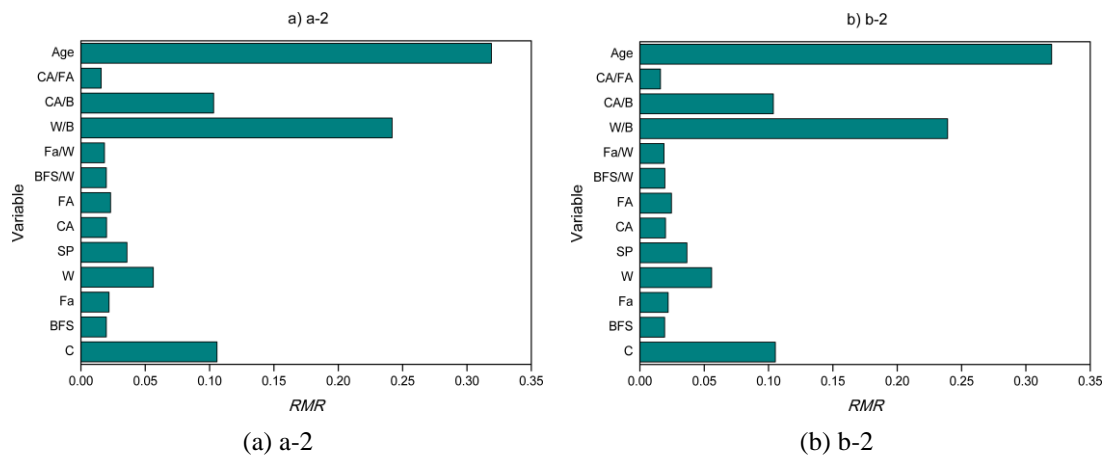


Fig. 3. Results of VIMs (RMR) for combination a-2 and b-2

280 The VIM results for model b-2 are shown in **Fig. 3b**, and are similar to those for model a-2.
 281 Therefore, variable selection results for group B are the same as for group A. It seems that variable
 282 forms have little effect on the result of VIMs.

283 It was assumed that the compressive strength of HPC is affected by its components. This is
 284 why basic variables were not eliminated when selecting candidate variables, even though the RMR
 285 of some basic variables were relatively small.

286

287

288 4.2. Results of the second stage

289

290 4.2.1. New variable combinations

291

292 In group A, three new variable combinations were constructed by adding W/B, CA/B, and W/B
 293 and CA/B to A-1 in turn, which are named A-3, A-4, and A-5, respectively. The construction of
 294 new combinations in group B was consistent with those in group A as the results of the VIMs for
 295 a-2 and b-2 were similar.

296

297 4.2.2. Effect of variable combination on model prediction accuracy

298

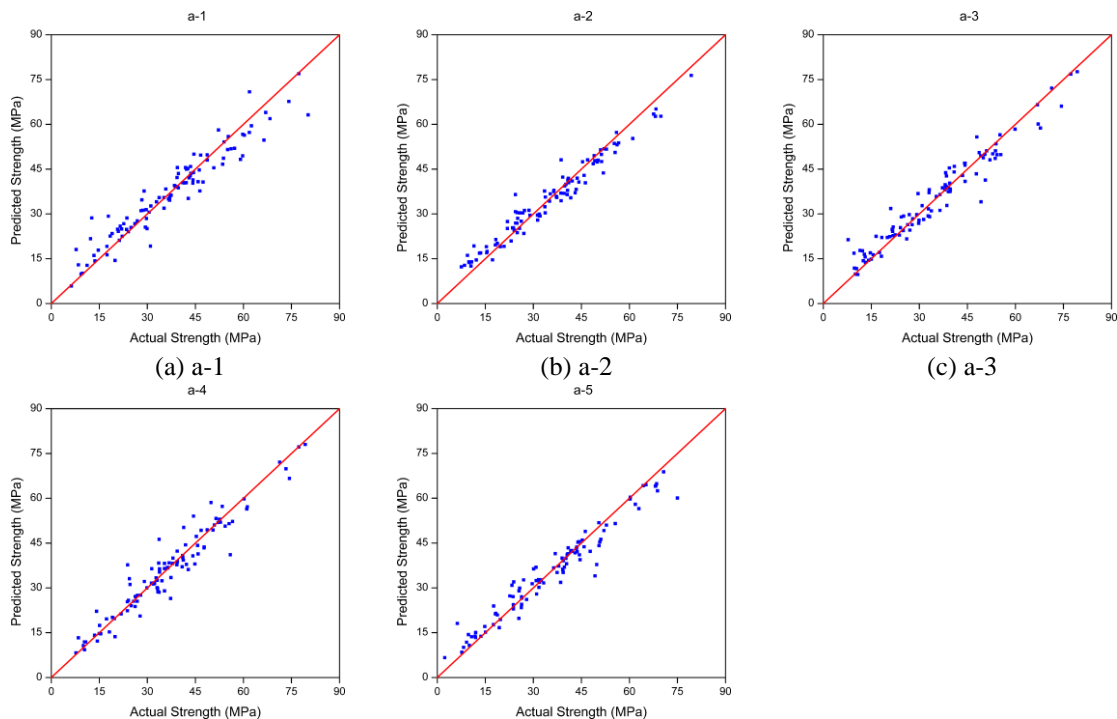
299 **Table 2** summarizes the performance measurements for models built with different variable
300 combinations. The best result for each performance measure is given in bold type. In group A, A-2
301 was the most complex combination, which corresponded to the lowest model prediction accuracy.
302 The simplest combination, A-1, corresponded to the second lowest model prediction accuracy.
303 This indicates that overabundant variables lead to poor prediction accuracy, as does a lack of key
304 variables.

305 The results for group B are similar to those for group A, implying that model prediction
306 accuracy is insensitive to the variable forms. The prediction accuracy of the model can be
307 improved by selecting appropriate variables. **Fig. 4** shows a comparison of the predicted and actual
308 values for each of the 10 models. The predicted values are very close to the corresponding actual
309 values.

310 Of the models, a-4 was the best model for minimizing MAE (3.1055), RMSE (4.4339), and SI
311 (0.0595), and the second-best model for minimizing MAPE (11.7850) and maximizing R (0.9655).
312 As a result, a-4 and A-4 were selected as the optimal model and the optimal variable combination,
313 respectively. **Table 3** summarizes a comparison of the performance measurements of a-1 and a-4.
314 All of the performance measurements for a-4 are better than that of a-2 indicating that the
315 proposed method was effective for improving model prediction accuracy.

316

317



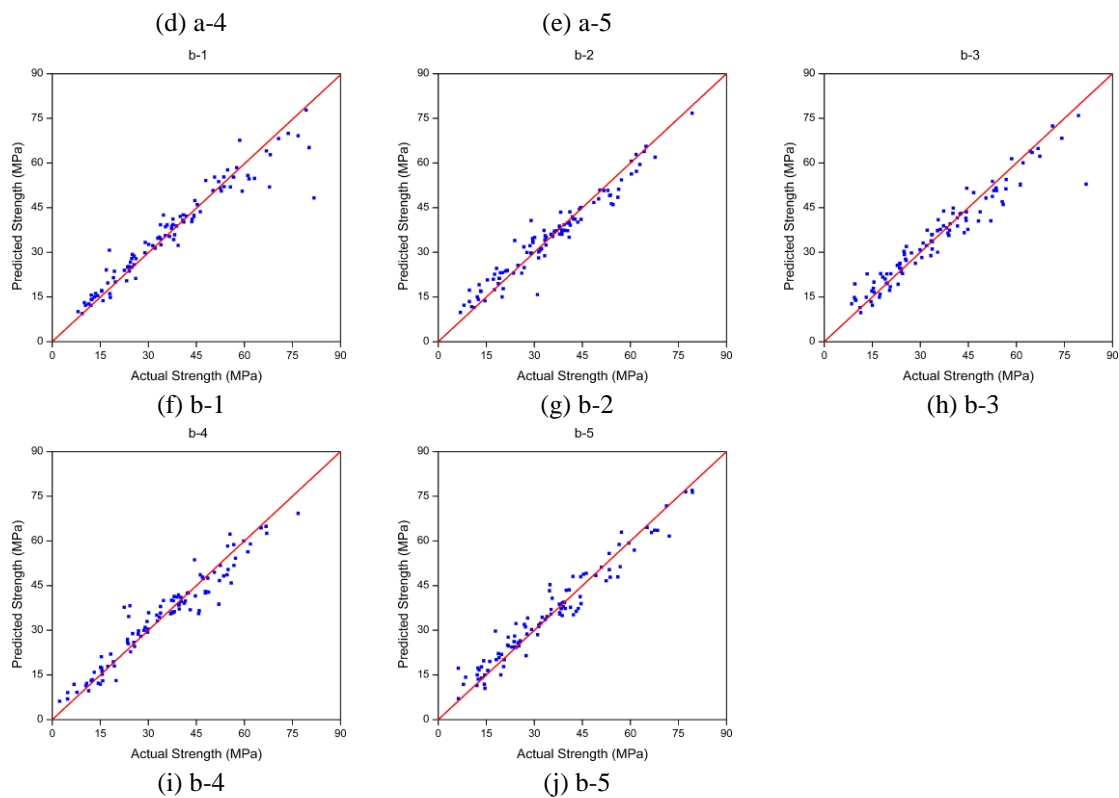


Fig. 4. Scatter plots of the predicted strength vs. the actual strength for the 10 models

318

Table 2 Prediction performance of the 10 models

Model	R	MAE (MPa)	RMSE (MPa)	MAPE (%)	SI
a-1	0.9623	3.3350	4.6650	12.0640	0.6591
a-2	0.9625	3.4065	4.7203	13.2777	0.9367
a-3	0.9637	3.2577	4.5281	12.2165	0.4289
a-4	0.9655	3.1055	4.4339	11.7850	0.0595
a-5	0.9662	3.1703	4.4455	11.8262	0.0974
b-1	0.9613	3.2228	4.6267	12.1511	0.5976
b-2	0.9655	3.3147	4.5432	12.5760	0.4464
b-3	0.9644	3.2078	4.4967	12.0204	0.2925
b-4	0.9622	3.1975	4.5481	11.7889	0.4073
b-5	0.9627	3.2150	4.6044	11.6018	0.4173

319

Table 3 Comparison of performance measurements for models a-1 and a-4

Performance measure	Model	Improvement (%)
	a-1	a-4

R	0.9623	0.9655	0.33
MAE (MPa)	3.3350	3.1055	6.88
RMSE (MPa)	4.6650	4.4339	4.95
MAPE (%)	12.064	11.785	3.31

320

321

4.2.3. Predictions of the optimal model

322

323

Fig.5 illustrates a set of residuals and the percentage error distribution for model a-4; this set selected randomly from the 50 sets of results. With increasing actual strength, the residuals and percentage error gradually fluctuated within a narrow range, except for a few outlying points. The residuals and percentage error of samples with an actual strength of less than 30 MPa were usually positive and larger than those for samples with actual strengths of greater than 30 MPa, which can be seen clearly in the upper left corner of **Fig. 5**. This indicates that the model tended to overestimate the compressive strength of samples with responses of less than 30 MPa.

329

330

The 50 sets of results were analyzed to obtain a more convincing conclusion. Each set was divided into two groups according to whether the response of sample was greater than 30 MPa, and each group was further divided into three subgroups based on the percentage error: greater than 10%, less than -10%, and between -10% and 10%. The percentage error distribution for each subgroup was counted, and the statistical results are summarized in **Table 4**. The results confirmed that this model has a tendency to overestimate the strength of samples with a strength less than 30 MPa. Surprisingly, this model could accurately predict the compressive strength of samples with strengths of greater than 30 MPa.

337

338

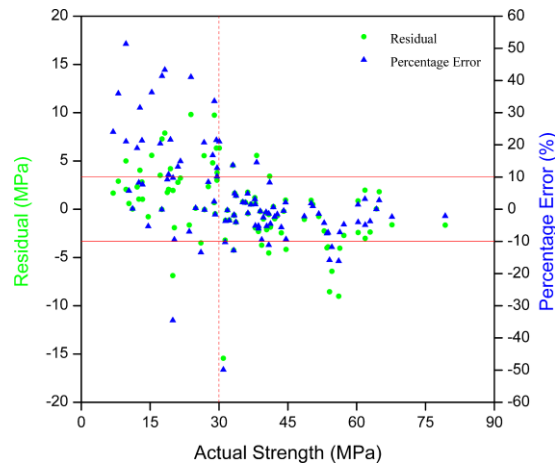


Fig. 5. Distribution of residuals and percentage error vs. actual strength for model a-4

339

Table 4 Statistical results for the number of samples in each subgroup, group, and the corresponding proportion.

Actual strength (MPa)	Percentage (%)	error	Number of samples in subgroup	Number of samples in group	Proportion (%)
[0,30]	$(-\infty,-10)$		174	2018	8.62
	$[-10,10]$		902		44.70
	$(10,+\infty)$		942		46.68
[30,82.6]	$(-\infty,-10)$		467	3132	14.91
	$[-10,10]$		2416		77.14
	$(10,+\infty)$		249		7.95
[0,82.6]	$(-\infty,-10)$		641	5150	12.45
	$[-10,10]$		3318		64.43
	$(10,+\infty)$		1191		23.12

340

4.3. Effect of parameter settings on predictions

341

342

343

By verifying that the HPCCS prediction obtained with the random forest model is insensitive to parameter settings, the goal of simplifying the parameter determination can be achieved.

344

345

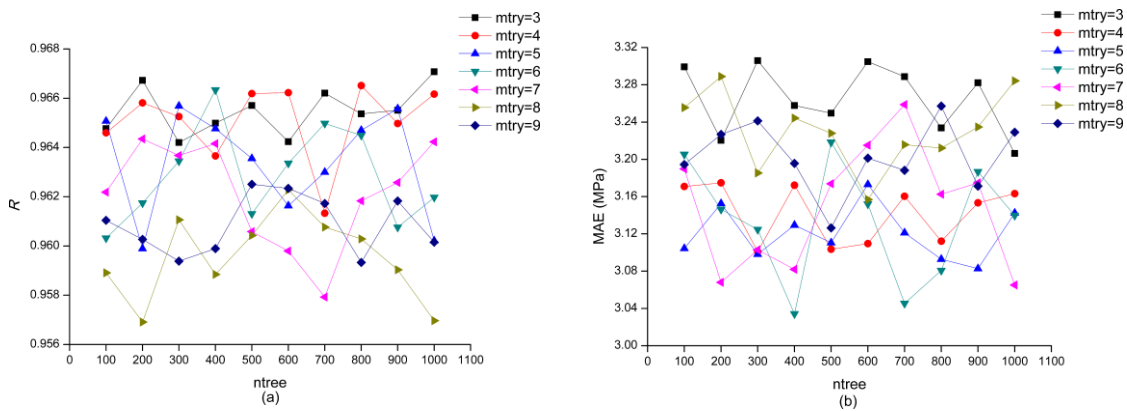
In this section, the effects of parameter settings on HPCCS predictions are compared. Fig. 6 shows the performance measurements for the models with different parameter settings. The maximum, minimum, and average values and the standard deviation for each performance measure are listed in Table 5. The results shown in Fig. 6 and Table 5 reflect that the prediction accuracy of each model is similar, and the parameter settings have little effect on model prediction accuracy as long as they are set within a reasonable range. In addition, these results confirm the robustness of the random forest algorithm for HPCCS prediction.

349

350

351

352



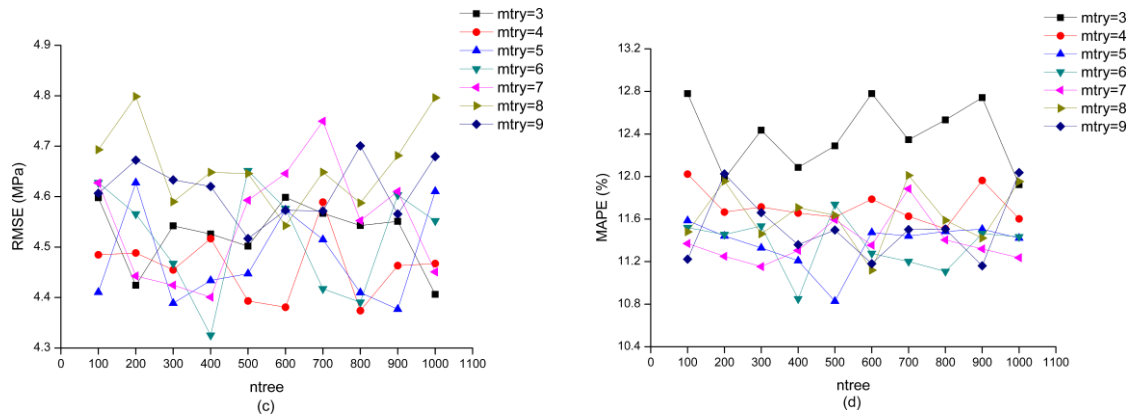


Fig. 6. Performance measurements for the model with different parameter settings: a) R , b) MAE, c) RMSE, and d) MAPE

353

Table 5 Statistical results of the performance measures

Performance measure	Max	Min	Avg	Standard deviation
R	0.9671	0.9569	0.9628	0.0026
MAE (MPa)	3.3060	3.0342	3.1777	0.0676
RMSE (MPa)	4.7985	4.3251	4.5444	0.1073
MAPE (%)	12.7796	10.8282	11.6245	0.4198

354

355

356

4.4. Comparison with previous work

357

358

359

360

361

362

363

Table 6 Comparison with previously proposed models

First author	Year	Ref.	Model	R	MAE (MPa)	RMSE (MPa)	MAPE (%)	Parameter tuning method
Yeh	1998	(Yeh 1998)	ANNs	0.9602	N/A	N/A	N/A	Hand tuning for ANNs
			LR	0.8826	N/A	N/A	N/A	
Chou	2011	(Chou et	ANNs	0.9534	N/A	5.0302	10.90	Hand tuning

		al. 2011)	SVM	0.9412	N/A	5.6192	12.77	
			MART	0.9544	N/A	4.9489	13.89	
			BRT	0.9436	N/A	5.5720	14.18	
Chou	2013	(Chou and Pham 2013)	ANNs	0.930	4.421	6.329	15.3	No tuning
			CART	0.840	6.815	9.703	24.1	
			CHAID	0.861	6.088	8.983	20.7	
			GENLIN	0.779	7.867	11.375	29.9	
			SVMs	0.923	4.764	6.911	17.3	
Chou	2014	(Chou et al. 2014)	MLP	N/A	6.19	7.95	20.84	No tuning
			CART	N/A	5.86	7.84	20.66	
			SVM	N/A	3.75	5.59	12.03	
Erdal	2013	(Erdal et al. 2013)	ANNs	0.9533	4.18	5.57	N/A	Hand tuning
			BANNs	0.9632	3.60	4.87	N/A	
			GBANNs	0.9628	4.09	5.24	N/A	
			WBANN	0.9694	3.30	4.54	N/A	
			s					
			WGBAN	0.9711	4.83	5.75	N/A	
			Ns					
This study	N/A	N/A	RF (a-4)	0.9655	3.1055	4.4339	11.79	No tuning
			RF (a-2)	0.9623	3.3350	4.6650	12.06	

364

365 Before input variable optimization, model a-1 in this study was the sixth best model for
366 maximizing R , the third best model for minimizing MAPE, and the second-best model for
367 minimizing MAE and RMSE. The optimized model (a-4) was the best model for minimizing MAE
368 and RMSE, the second-best model for minimizing MAPE, and the third best model for
369 maximizing R among these models. It is clear that the generalization capacity of the RF model was
370 greatly improved after input variable optimization.

371

372

373 5. Conclusions

374

375 In this study, a method was proposed to optimize input variables, simplify parameter
376 determination, and predict HPCCS. Unlike other studies aiming to develop advanced models to
377 predict HPCCS or compare the generalization ability of different models for HPCCS prediction,
378 this study attempts to improve the model generation efficiency and prediction accuracy.

- 379 • Measuring the importance of input variables revealed that Age and W/B are the two
380 variables that have the strongest influence on the HPCCS. The effect of variable forms on

381 HPCCS prediction was compared, and it was found that input variables in the form of either
382 relative mass or absolute mass have little effect on prediction. We suggested the use of the
383 absolute mass of HPC components as input variables to predict HPCCS.

384 • The quantity of input variables influences the prediction of HPCCS. The number of input
385 variables in the two models with the lowest prediction accuracy are the largest and the smallest
386 in groups A and B, respectively. The proposed method is effective for optimizing input
387 variables. The model built by the proposed method shows a stronger generalization capacity
388 than that built without input variable optimization.

389 • Random forest exhibits excellent performance for HPCCS prediction even with default
390 parameter settings, which was confirmed by a comparison with previously published models.
391 Moreover, we confirmed that the prediction of HPCCS is insensitive to parameter settings as
392 long as they are set within a reasonable range. In terms of computing expense, we recommend
393 using fewer trees and candidate variables for the predictions.

394 • In addition, the model built by the proposed method was inclined to overestimate the
395 compressive strength of samples with actual strengths of less than 30 MPa, but it could
396 accurately predict the compressive strength of samples with actual strengths greater than 30
397 MPa.

400 Acknowledgments

401
402 The authors acknowledge the support of National Natural Science Foundation of China (No.
403 51525803). UC Irvine Machine Learning Repository (<http://archive.ics.uci.edu/ml/>) and Professor
404 I-Cheng Yeh are also gratefully acknowledged for sharing the experimental data set.

406 **Declaration of interest:** None.

408 References

- 409
410 Archer, K. J. and Kimes, R. V. (2008). "Empirical characterization of random forest variable importance
411 measures." *COMPUT STAT DATA AN* **52**(4): 2249-2260.
- 412 Auret, L. and Aldrich, C. (2012). "Interpretation of nonlinear relationships between process variables by use
413 of random forests." *MINER ENG* **35**: 27-42.
- 414 Ayaz, Y., Kocamaz, A. F. and Karakoç, M. B. (2015). "Modeling of compressive strength and UPV of
415 high-volume mineral-admixtured concrete using rule-based M5 rule and tree model M5P classifiers."
416 *CONSTR BUILD MATER* **94**: 235-240.
- 417 Bui, DK., Nguyen, T., Chou, J.-S., Nguyen-Xuan, H., and Ngo, TD. (2018). "A modified firefly
418 algorithm-artificial neural network expert system for predicting compressive and tensile strength of
419 high-performance concrete." *CONSTR BUILD MATER* **180**: 320-333.
- 420 Behnood, A., Behnood, V., Modiri Gharehveran, M. and Alyamac, K. E. (2017). "Prediction of the
421 compressive strength of normal and high-performance concretes using M5P model tree algorithm."
422 *CONSTR BUILD MATER* **142**: 199-207.
- 423 Bharatkumar, B. H., Narayanan, R., Raghuprasad, B. K. and Ramachandramurthy, D. S. (2001). "Mix
424 proportioning of high performance concrete." *CEMENT CONCRETE COMP* **23**(1): 71-80.

425 Boulesteix, A.-L., Janitza, S., Kruppa, J. and König, I. R. (2012). "Overview of random forest methodology
426 and practical guidance with emphasis on computational biology and bioinformatics." *DATA MIN KNOWL*
427 *DISC* **2**(6): 493-507.

428 Breiman, L. (1996). "Bagging predictors." *MACH LEARN* **24**(2): 123-140.

429 Breiman, L. (2001). "Random forests." *MACH LEARN* **45**(1): 5-32.

430 Breiman, L., H.Friedman, J., Olshen, R. A. and Stone, C. J. (1984), *Classification and Regression Trees*,
431 Wadsworth statistics/probability, New York, NY, American.

432 Chang, T. P., Chuang, F. C. and Lin, H. C. (1996). "A mix proportioning methodology for high-performance
433 concrete." *J CHIN INST ENG* **19**(6): 645-655.

434 Cheng, M.-Y., Firdausi, P. M. and Prayogo, D. (2014). "High-performance concrete compressive strength
435 prediction using Genetic Weighted Pyramid Operation Tree (GW POT)." *ENG APPL ARTIF INTEL* **29**:
436 104-113.

437 Chiew, F. H., Ng, C. K., Chai, K. C. and Tay, K. M. (2017). "A Fuzzy Adaptive Resonance Theory-Based
438 Model for Mix Proportion Estimation of High-Performance Concrete." *COMPUT-AIDED CIV INF* **32**(9):
439 772-786.

440 Chithra, S., Kumar, S. R. R. S., Chinnaraju, K. and Alfin Ashmita, F. (2016). "A comparative study on the
441 compressive strength prediction models for High Performance Concrete containing nano silica and copper
442 slag using regression analysis and Artificial Neural Networks." *CONSTR BUILD MATER* **114**: 528-535.

443 Chou, J.-S., Chiu, C.-K., Farfoura, M. and Al-Taharwa, I. (2011). "Optimizing the Prediction Accuracy of
444 Concrete Compressive Strength Based on a Comparison of Data-Mining Techniques." *J COMPUT CIVIL*
445 *ENG* **25**(3): 242-253.

446 Chou, J.-S. and Pham, A.-D. (2013). "Enhanced artificial intelligence for ensemble approach to predicting
447 high performance concrete compressive strength." *CONSTR BUILD MATER* **49**: 554-563.

448 (2015). "Smart Artificial Firefly Colony Algorithm-Based Support Vector Regression for Enhanced
449 Forecasting in Civil Engineering." *COMPUT-AIDED CIV INF* **30**(9): 715-732.

450 Chou, J.-S. and Tsai, C.-F. (2012). "Concrete compressive strength analysis using a combined classification
451 and regression technique." *AUTOMAT CONSTR* **24**: 52-60.

452 Chou, J.-S., Tsai, C.-F., Pham, A.-D. and Lu, Y.-H. (2014). "Machine learning in concrete strength
453 simulations: Multi-nation data analytics." *CONSTR BUILD MATER* **73**: 771-780.

454 Dubeau, P., King, D., Unbushe, D. and Rebelo, L.-M. (2017). "Mapping the Dabus Wetlands, Ethiopia,
455 Using Random Forest Classification of Landsat, PALSAR and Topographic Data." *REMOTE*
456 *SENS-BASEL* **9**(10): 1056.

457 Erdal, H. I. (2013). "Two-level and hybrid ensembles of decision trees for high performance concrete
458 compressive strength prediction." *ENG APPL ARTIF INTEL* **26**(7): 1689-1697.

459 Erdal, H. I., Karakurt, O. and Namli, E. (2013). "High performance concrete compressive strength
460 forecasting using ensemble models based on discrete wavelet transform." *ENG APPL ARTIF INTEL* **26**(4):
461 1246-1254.

462 Fu, B., Wang, Y., Campbell, A., Li, Y., Zhang, B., Yin, S., Xing, Z. and Jin, X. (2017). "Comparison of
463 object-based and pixel-based Random Forest algorithm for wetland vegetation mapping using high spatial
464 resolution GF-1 and SAR data." *ECOL INDIC* **73**: 105-117.

465 Grömping, U. (2009). "Variable Importance Assessment in Regression: Linear Regression versus Random
466 Forest." *AMER.STATIST.ASSOC* **63**(4): 308-319.

467 Hanselmann, M., Köthe, U., Kirchner, M., Renard, B. Y., Amstalden, E. R., Glunde, K., Heeren, R. M. A.
468 and Hamprecht, F. A. (2009). "Toward Digital Staining using Imaging Mass Spectrometry and Random
469 Forests." *J PROTEOME RES* **8**(7): 3558-3567.

470 Kalman Šipoš, T., Miličević, I. and Siddique, R. (2017). "Model for mix design of brick aggregate concrete
471 based on neural network modelling." *CONSTR BUILD MATER* **148**: 757-769.

472 Krkač, M., Špoljarić, D., Bernat, S. and Arbanas, S. M. (2016). "Method for prediction of landslide
473 movements based on random forests." *LANDSLIDES* **14**(3): 947-960.

474 Lim, C.-H., Yoon, Y.-S. and Kim, J.-H. (2004). "Genetic algorithm in mix proportioning of

475 high-performance concrete." *CEMENT CONCRETE RES* **34**(3): 409-420.

476 Loh, W.-Y. (2011). "Classification and regression trees." Wiley Interdisciplinary Reviews: *DATA MIN*

477 *KNOWL DISC* **1**(1): 14-23.

478 Maghrebi, M., Waller, T. and Sammut, C. (2016). "Matching experts' decisions in concrete delivery

479 dispatching centers by ensemble learning algorithms: Tactical level." *AUTOMAT CONSTR* **68**: 146-155.

480 Matin, S. S., Farahzadi, L., Makaremi, S., Chelgani, S. C. and Sattari, G. (2017). "Variable selection and

481 prediction of uniaxial compressive strength and modulus of elasticity by random forest." *APPL SOFT*

482 *COMPUT*.

483 Mohamed, O. A., Ati, M. and Najm, O. F. (2017). "Predicting Compressive Strength of Sustainable

484 Self-Consolidating Concrete Using Random Forest." *KEY ENG MAT* **744**: 141-145.

485 Omran, B. A., Chen, Q. and Jin, R. (2016). "Comparison of Data Mining Techniques for Predicting

486 Compressive Strength of Environmentally Friendly Concrete." *J COMPUT CIVIL ENG* **30**(6): 04016029.

487 Ozcan, G., Kocak, Y. and Gulbandilar, E. (2017). "Estimation of compressive strength of BFS and WTRP

488 blended cement mortars with machine learning models." *COMPUT CONCRETE* **19**(3): 275-282.

489 Qi, C., Fourie, A. and Chen, Q. (2018). "Neural network and particle swarm optimization for predicting the

490 unconfined compressive strength of cemented paste backfill." *CONSTR BUILD MATER* **159**: 473-478.

491 Rao, W. (2017). "Application of Machine Learning in the Prediction of Compressive Strength of Concrete." *STAT APPL* **06**(01): 1-6.

492

493 Safarzadegan Gilan, S., Bahrami Jovein, H. and Ramezani pour, A. A. (2012). "Hybrid support vector

494 regression – Particle swarm optimization for prediction of compressive strength and RCPT of concretes

495 containing metakaolin." *CONSTR BUILD MATER* **34**: 321-329.

496 Schwarz, D. F., Knig, I. R. and Ziegler, A. (2011). "On safari to Random Jungle: a fast implementation of

497 Random Forests for high-dimensional data." *Bioinformatics* **27**(3): 439-439.

498 Sebastiá, M., Fernández Olmo, I. and Irabien, A. (2003). "Neural network prediction of unconfined

499 compressive strength of coal fly ash–cement mixtures." *CEMENT CONCRETE RES* **33**(8): 1137-1146.

500 Sonebi, M., Cevik, A., Grünwald, S. and Walraven, J. (2016). "Modelling the fresh properties of

501 self-compacting concrete using support vector machine approach." *CONSTR BUILD MATER* **106**: 55-64.

502 Svetnik V., Liaw A., Tong C., Wang T. (2004) *Application of Breiman's Random Forest to Modeling*

503 *Structure-Activity Relationships of Pharmaceutical Molecules*. Multiple Classifier Systems. MCS 2004.

504 Lecture Notes in Computer Science, vol **3077**. Springer, Berlin, Heidelberg

505 Wetzel, A and Middendorf, B. (2019). "Influence of silica fume on properties of fresh and hardened

506 ultra-high performance concrete based on alkali-activated slag. " *CEMENT CONCRETE COMP* **100**:

507 53-59.

508 Yeh, I. C. (1998). "Modeling of strength of high-performance concrete using artificial neural networks." *CEMENT CONCRETE RES* **28**(12): 1797-1808.

509

510 Zhu, H., Wang, Z. J., Xu, J. and Han, Q. H. (2019). "Microporous structures and compressive strength of

511 high-performance rubber concrete with internal curing agent." *CONSTR BUILD MATER* **215**: 128-134.

512

513