Elsevier Editorial System(tm) for
Construction & Building Materials
Manuscript Draft

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 performers 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	*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

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]	(-∞,-10)	174	2018	8.62
	[-10,10]	902		44.70
	(10,+∞)	942		46.68
[30,82.6]	(-∞,-10)	467	3132	14.91
	[-10,10]	2416		77.14
	(10,+∞)	249		7.95
[0,82.6]	(-∞,-10)	641	5150	12.45
	[-10,10]	3318		64.43
	(10,+∞)	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.

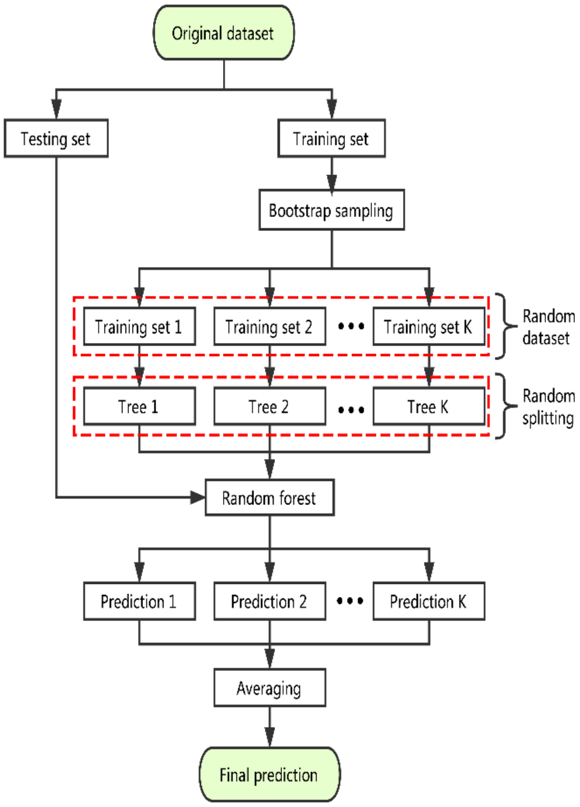


Fig. 1. Schematic of random forest generation and prediction

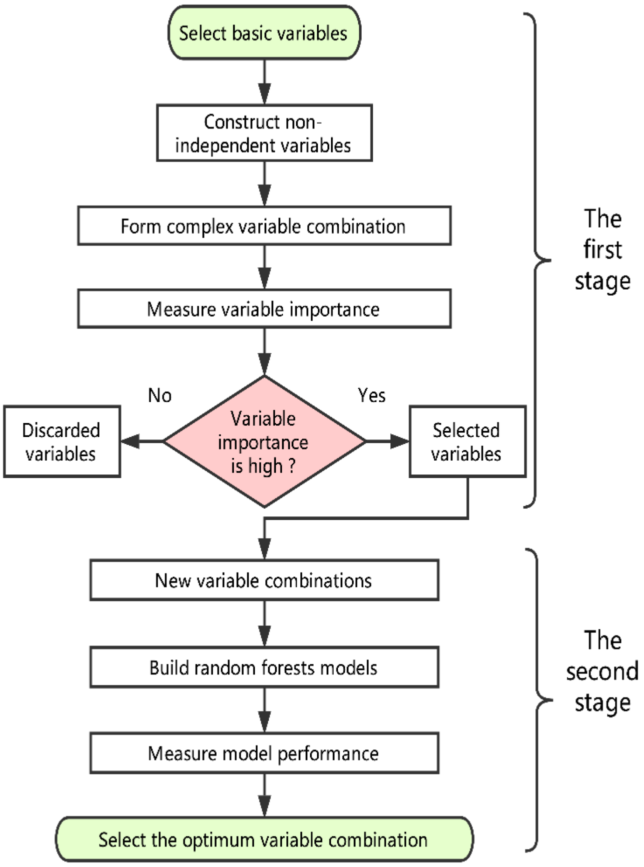

Select basic variables
Construct non-independent variables
Form complex variable combination
Measure variable importance
No
Yes
Discarded variables
Variable importance is high ?
Selected variables
The first stage
New variable combinations
Build random forests models
Measure model performance
Select the optimum variable combination
The second stage

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 performers 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 thrid 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-admixtured 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.

Safarzadegan Gilan, S., Bahrami Jovein, H. and Ramezanianpour, 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., Grünewald, 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

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

Qinghua Han[1, 2], Changqing Gui[2], Jie Xu[1, 2*], Giuseppe Lacidogna[3]

[1]*Key Laboratory of Earthquake Engineering Simulation and Seismic Resilience of China Earthquake Administration (Tianjin University), Tianjin 300350, China;*

[2]*School of Civil Engineering, Tianjin University/Key Laboratory of Coast Civil Structure Safety of China Ministry of Education, Tianjin University, Tianjin, 300072, China*

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

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

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

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 performers well when the actual compressive strength of HPC is high.

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

1. Introduction

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

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

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

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

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

On the other hand, it is tedious work to determine the proper parameter settings, which have great influence on the model prediction accuracy. Manual tuning requires a great deal of time and attention. Some scholars have summed up some empirical formulas; however, the results of these empirical formulas are often different (Kalman Šipoš et al. 2017). The application of an optimization algorithm can assist in determining the appropriate parameter settings, which increases the complexity of the model (Safarzadegan Gilan et al. 2012, Chou and Pham 2015, Qi et al. 2018).

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

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

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

2. Methods

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

2.1. Machine learning techniques

2.1.1 Bagging method

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

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

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

(1)

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

2.1.2 Classification and regression tree

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

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

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

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

where $|\boldsymbol{s}|$ is the number of samples in dataset $\boldsymbol{S}$ that reach the node; $\boldsymbol{S_i}$ is the dataset resulting from splitting at the node, which falls into a subspace according to the given variable $\boldsymbol{x_j}$ (j=1, 2, … , M) and segmentation $\boldsymbol{\alpha}$; and $\boldsymbol{y_i}$ is the response value of the ***ith*** sample in dataset $\boldsymbol{S}$.

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

2.1.3 Random forest

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

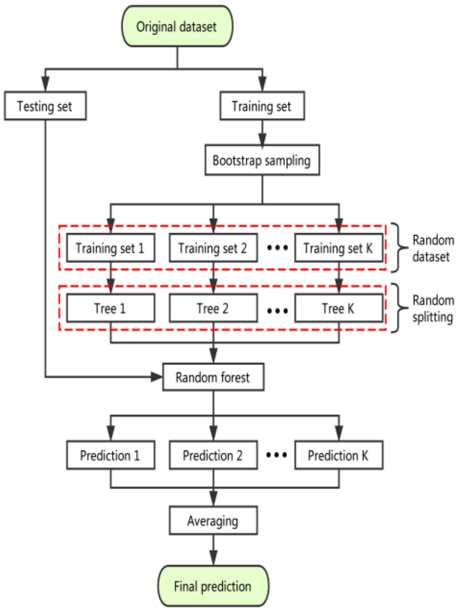


Fig. 1. Schematic of random forest generation and prediction

Variable importance measures (VIMs) are an inherent product of RF. The basic concept of VIMs is that if an input variable, $\boldsymbol{x_j}$, has an impact on the response, the prediction accuracy of the model will decrease with permutation of the values of variable $\boldsymbol{x_j}$. As a result, the values of variables are permuted one at a time and the resulting reduction in prediction accuracy of the new model is evaluated; the greater the decrease in prediction accuracy, the stronger the association between the permuted variable and the response. Generally, reduction in **MSE** has been used as the evaluation index. In RF, the out-of-bag (OOB) samples are permuted to measure variable importance in order to avoid training new forests (Archer and Kimes 2008, Auret and Aldrich 2012). For variable $\boldsymbol{x_j}$ in tree $\boldsymbol{i}$, the reduction in MSE can be calculated as follows:

$$Imp_j^i = \mathrm{MSE}\big(\mathrm{T}_{\mathrm{D}(\theta_i)}\big) - \mathrm{MSE}(\mathrm{T}_{\mathrm{D}^j_{\mathrm{OOB}}(\theta_i)}) \tag{5}$$

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

$$Imp_j = \frac{1}{K}\sum_{i=1}^{K} Imp_j^i \tag{6}$$

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

$$RMR_j = \frac{Imp_j}{\sum_{i=1}^{\mathrm{M}} Imp_i} \tag{7}$$

2.2 Proposed method

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

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

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

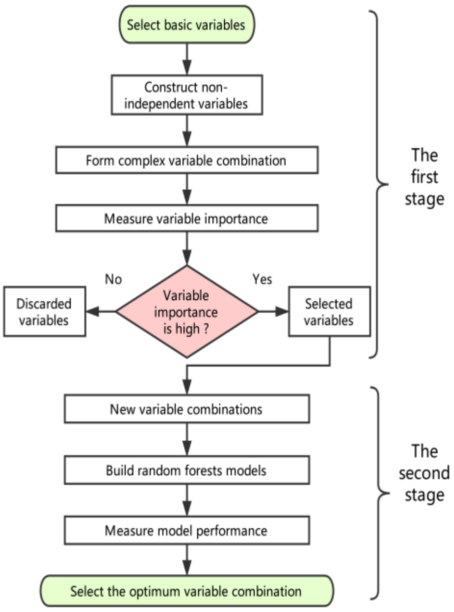

Select basic variables
Construct non-independent variables
Form complex variable combination
Measure variable importance
No
Yes
Discarded variables
Variable importance is high ?
Selected variables
The first stage
New variable combinations
Build random forests models
Measure model performance
Select the optimum variable combination
The second stage

Fig. 2. Flowchart of the proposed method

Note that the optimal input variables are for the RF model. However, the optimal input variables for other models can be obtained by replacing the RF model used in the second stage with the desired target model. The parameters of these other models may need to be tuned, which will increase the workload for input variable optimization.

3. Materials and modeling setting

3.1. Dataset

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

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

3.2. Input variable combinations

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

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

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

3.3. Performance evaluation

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

3.3.1. Linear correlation coefficient

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

where ***y*** is the actual response, $\boldsymbol{y'}$ is the predicted response, and ***n*** is the number of samples in the testing set.

3.3.2. Mean absolute error

$$\mathrm{MAE} = \frac{1}{n}\sum|y - y'| \qquad (9)$$

3.3.3. Root mean squared error

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

3.3.4. Mean absolute percentage error

$$\mathrm{MAPE} = \frac{1}{n}\sum\left|\frac{y - y\prime}{y}\right| \qquad (11)$$

3.3.5. Synthesis index

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

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

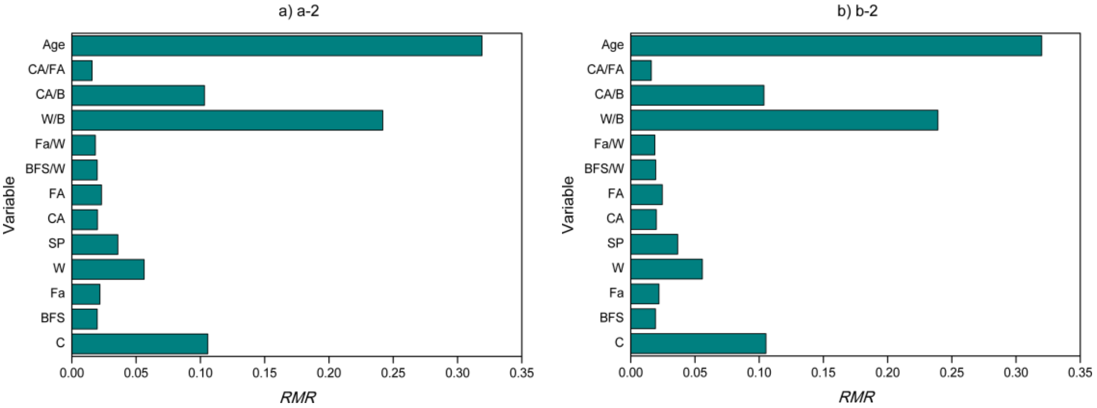
3.4. Modelling setting

The number of trees in the forest, ***ntree***, and the number of selected candidate variables when the node is splitting, ***mtry***, are two essential parameters which need to be set in the random forest model. The default parameter settings of ***ntree*** and ***mtry*** were 500 and the minimum integer that is greater than $\mathbf{log}_{\mathbf{2}}^{\boldsymbol{M}}$, respectively. To obtain precise measurements of variable importance, a large number of trees are needed (Grömping 2009). As a result, ***ntree*** was set to 1000 for the variable importance measurement. A total of 70 sets of parameter settings were constructed to verify the robustness of the RF model. The values of ***ntree*** range from 100 to 1000 in increments of 100, while ***mtry*** ranges from 3 to 9 with an increment of 1. Each dataset is divided into two subsets (i.e., the training set and the testing set). The training set consists of 927 samples selected randomly from the dataset, and the testing set contains the remaining 103 samples. Each RF model was run 50 times, and the average was taken as the final result to limit bias due to the random sampling.

4. Results and discussion

4.1. Results of the first stage

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



(a) a-2 (b) b-2

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

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

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

4.2. Results of the second stage

4.2.1. New variable combinations

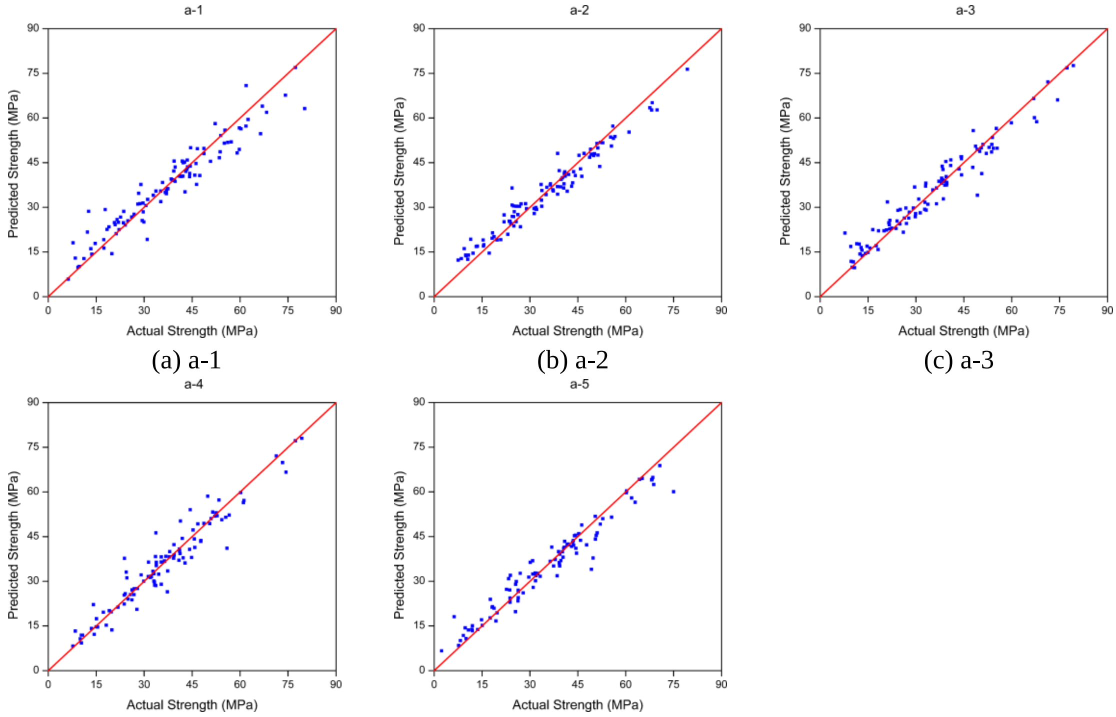
In group A, three new variable combinations were constructed by adding W/B, CA/B, and W/B and CA/B to A-1 in turn, which are named A-3, A-4, and A-5, respectively. The construction of new combinations in group B was consistent with those in group A as the results of the VIMs for a-2 and b-2 were similar.

4.2.2. Effect of variable combination on model prediction accuracy

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

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

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



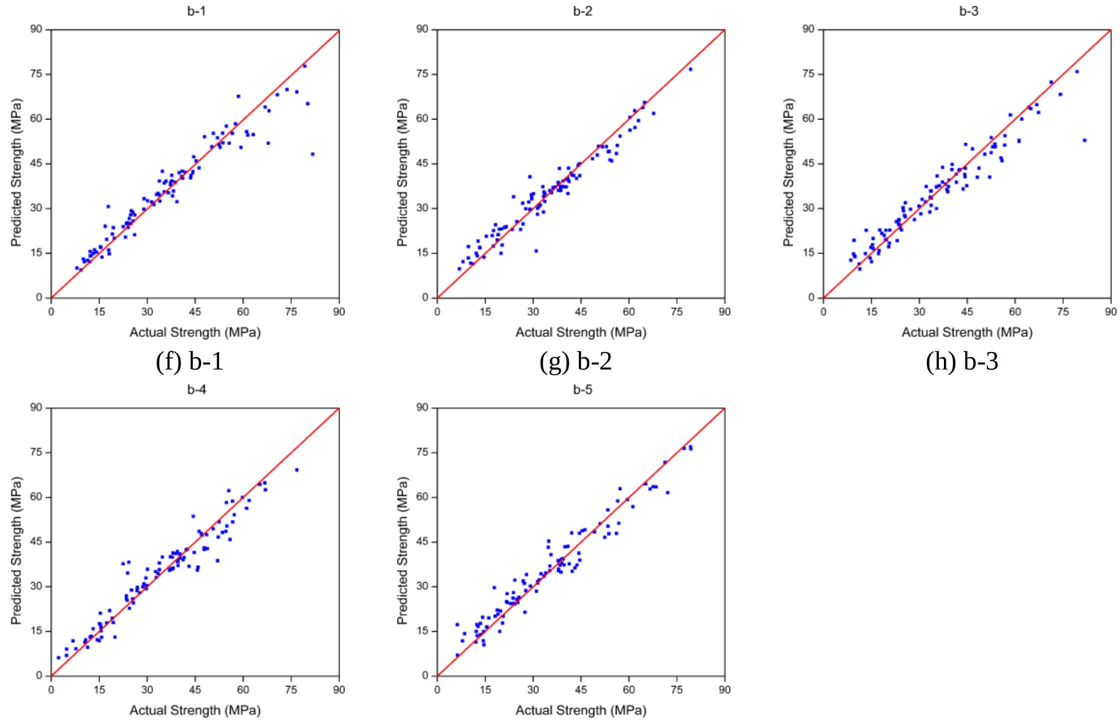
(a) a-1

(b) a-2

(c) a-3

(d) a-4

(e) a-5



(f) b-1 (g) b-2 (h) b-3

(i) b-4 (j) b-5

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

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

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

4.2.3. Predictions of the optimal model

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.

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.

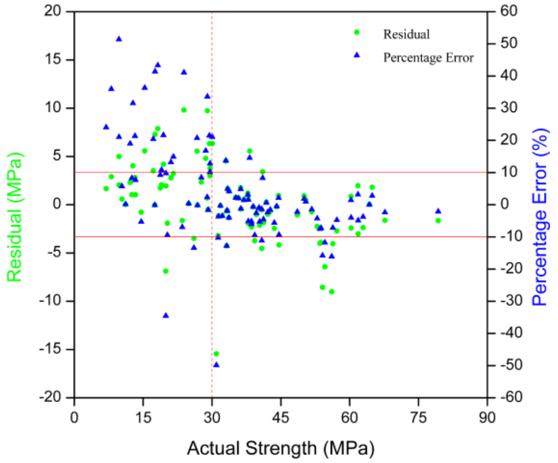


Fig. 5. Distribution of residuals and percentage error vs. actual strength for model a-4

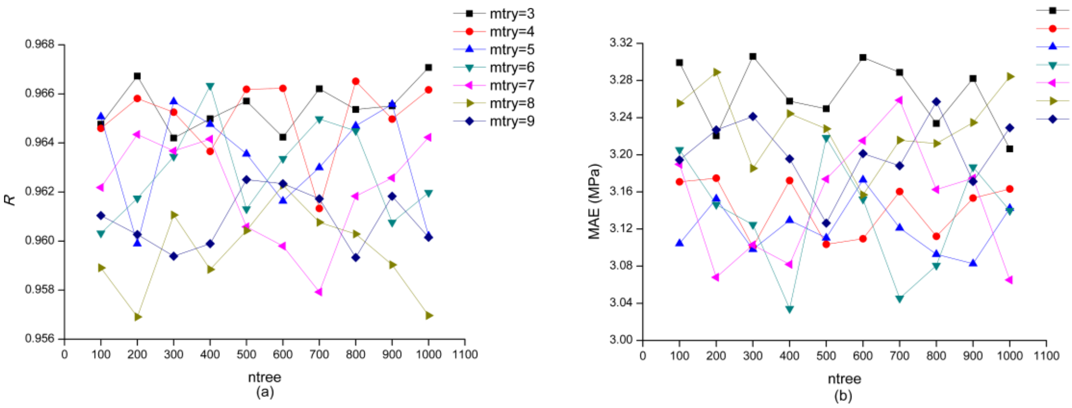
Table 4 Statistical results for the number of samples in each subgroup, group, and the corresponding proportion.

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

4.3. Effect of parameter settings on predictions

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.

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.



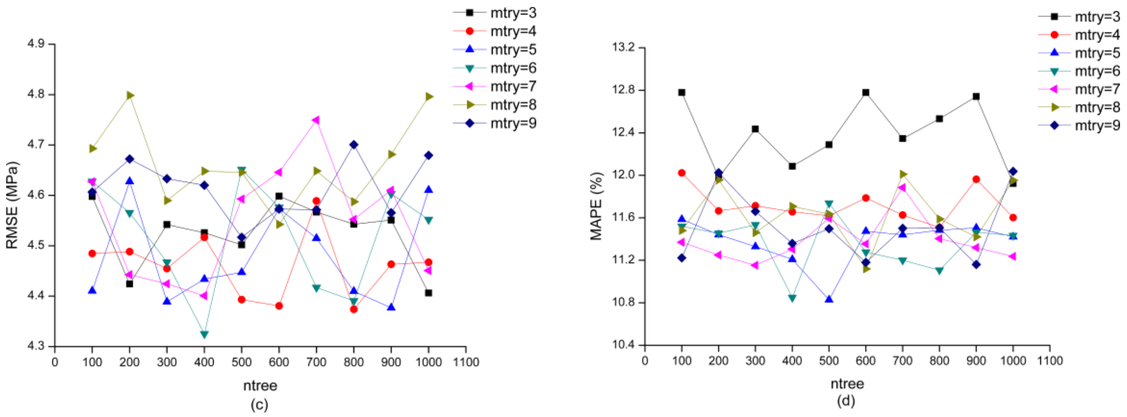


Fig. 6. Performance measurements for the model with different parameter settings: a) *R*, b) MAE, c) RMSE, and d) MAPE

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

4.4. Comparison with previous work

Many scholars have proposed different models to predict HPCCS in recent years. **Table 6** summarizes some previously published models for HPCCS prediction. The best performance measurements from these models are given in bold type. The datasets used in these studies are derived from the same dataset collected by Yeh used in this study (Yeh 1998). All $\boldsymbol{R}^2$ reported in the previous studies were converted to ***R*** for convenient comparison.

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 LR	0.9602 0.8826	N/A N/A	N/A N/A	N/A N/A	Hand tuning for ANNs
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	
			WBANNs	0.9694	3.30	4.54	N/A	
			WGBANNs	**0.9711**	4.83	5.75	N/A	
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	

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

5. Conclusions

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

- Measuring the importance of input variables revealed that Age and W/B are the two variables that have the strongest influence on the HPCCS. 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.

- The quantity of input variables influences the prediction of HPCCS. The number of input variables in the two models with the lowest prediction accuracy are the largest and the smallest in groups A and B, respectively. 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.
- 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. In terms of computing expense, we recommend using fewer trees and candidate variables for the predictions.
- In addition, 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.

Acknowledgments

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

Declaration of interest: None.

References

Archer, K. J. and Kimes, R. V. (2008). "Empirical characterization of random forest variable importance measures." *COMPUT STAT DATA AN* **52**(4): 2249-2260.

Auret, L. and Aldrich, C. (2012). "Interpretation of nonlinear relationships between process variables by use of random forests." *MINER ENG* **35**: 27-42.

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

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.

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.

Bharatkumar, B. H., Narayanan, R., Raghuprasad, B. K. and Ramachandramurthy, D. S. (2001). "Mix proportioning of high performance concrete." *CEMENT CONCRETE COMP* **23**(1): 71-80.

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

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

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

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

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

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

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

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

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.

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

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

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.

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

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

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

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

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

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

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.

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

Lim, C.-H., Yoon, Y.-S. and Kim, J.-H. (2004). "Genetic algorithm in mix proportioning of high-performance concrete." *CEMENT CONCRETE RES* **34**(3): 409-420.

Loh, W.-Y. (2011). "Classification and regression trees." Wiley Interdisciplinary Reviews: *DATA MIN KNOWL DISC* **1**(1): 14-23.

Maghrebi, M., Waller, T. and Sammut, C. (2016). "Matching experts' decisions in concrete delivery dispatching centers by ensemble learning algorithms: Tactical level." *AUTOMAT CONSTR* **68**: 146-155.

Matin, S. S., Farahzadi, L., Makaremi, S., Chelgani, S. C. and Sattari, G. (2017). "Variable selection and prediction of uniaxial compressive strength and modulus of elasticity by random forest." *APPL SOFT COMPUT*.

Mohamed, O. A., Ati, M. and Najm, O. F. (2017). "Predicting Compressive Strength of Sustainable Self-Consolidating Concrete Using Random Forest." *KEY ENG MAT* **744**: 141-145.

Omran, B. A., Chen, Q. and Jin, R. (2016). "Comparison of Data Mining Techniques for Predicting Compressive Strength of Environmentally Friendly Concrete." *J COMPUT CIVIL ENG* **30**(6): 04016029.

Ozcan, G., Kocak, Y. and Gulbandilar, E. (2017). "Estimation of compressive strength of BFS and WTRP blended cement mortars with machine learning models." *COMPUT CONCRETE* **19**(3): 275-282.

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.

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

Safarzadegan Gilan, S., Bahrami Jovein, H. and Ramezanianpour, 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.

Schwarz, D. F., Knig, I. R. and Ziegler, A. (2011). "On safari to Random Jungle: a fast implementation of Random Forests for high-dimensional data." *Bioinformatics* **27**(3): 439-439.

Sebastiá, M., Fernández Olmo, I. and Irabien, A. (2003). "Neural network prediction of unconfined compressive strength of coal fly ash–cement mixtures." *CEMENT CONCRETE RES* **33**(8): 1137-1146.

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

Svetnik V., Liaw A., Tong C., Wang T. (2004) *Application of Breiman's Random Forest to Modeling Structure-Activity Relationships of Pharmaceutical Molecules*. Multiple Classifier Systems. MCS 2004. Lecture Notes in Computer Science, vol **3077**. Springer, Berlin, Heidelberg

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.

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

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.