Contributed Papers from Materials Science and Technology 2019 (MS&T19)
September 29–October 3, 2019, Oregon Convention Center, Portland, Oregon, USA

DOI 10.7449/2019/MST_2019_604_611

PREDICTION OF COMPRESSIVE STRENGTH AND MODULUS OF ELASTICITY OF CONCRETE USING MACHINE LEARNING MODELS

Taihao Han[1], Kamal Khayat[2], Hongyan Ma[3], Jie Huang[4], and Aditya Kumar[5]

1. Graduate Student, Department of Materials Science and Engineering, Missouri University of Science and Technology, Rolla, MO, USA 65409, Email: thy3b@mst.edu
2. Professor, Department of Civil, Architectural, and Environmental Engineering, Missouri University of Science and Technology, Rolla, MO, USA 65409, Email: khayat@mst.edu
3. Assistant Professor, Department of Civil, Architectural, and Environmental Engineering, Missouri University of Science and Technology, Rolla, MO, USA 65409, Email: mahon@mst.edu
4. Assistant Professor, Department of Electrical and Computer Engineering, Missouri University of Science and Technology, Rolla, MO, USA 65409, Email: jieh@mst.edu
5. Assistant Professor, Department of Materials Science and Engineering, Missouri University of Science and Technology, Rolla, MO, USA 65409, Email: kumarad@mst.edu

Keywords: machine learning; random forest; support vector machine; artificial neural network; compressive strength; and modulus of elasticity

Abstract

This paper presents machine learning (ML) models for high fidelity prediction of compressive strength and modulus of elasticity (MOE) of concrete in relation to primary attributes of its mixture design. Two comprehensive databases, consisting of over 1000 and 500 data-records consolidated from technical publications, were used for training and testing the ML models that included random forests (RF), support vector machine (SVM) and multilayer perceptron artificial neural network (MLP-ANN). The metrics used for evaluation of prediction performance included five different statistical parameters and composite performance index (CPI). Results show that the RF model consistently outperforms the other two ML models in terms of prediction accuracy. Overall, machine learning is a very powerful and efficient tool for prediction of concrete properties as well as for the optimization of its mixture design to meet a set of desired performance criteria.

Introduction

Concrete is the most produced and used material in the world. There is burgeoning interest among researchers to develop numerical models that can reliably predict mechanical properties (e.g., compressive strength, and modulus of elasticity) of concrete in relation to the composition of its precursors and its mixture design [1]. Mechanical properties, such as compressive strength and modulus of elasticity can determine the workability, serviceability, durability, and quality control of concrete. However, the properties of concrete are complex – highly nonlinear, and, often, non-monotonous – and cannot be predicted using simple linear regression models because of the staggeringly large compositional degrees of freedom in concrete and the inherent nonlinear

relationships between mixture design variables and properties of concretes. Sophisticated approaches, such as Machine Learning (ML), have the capability to reveal the hidden, and complex, semi-empirical rules that govern the correlation between mixture design and properties of concrete. Previous studies [2], [3] have successfully employed ML models to predict concretes' properties using their mixture design variables and age as inputs.

This study presents two-fold contributions. The first one is to compare the performance of three common ML models – random forests (RF), support vector machine (SVM) and multilayer perceptron artificial neural network (MLP-ANN) – on the prediction of compressive strength and modulus of elasticity (MOE) of concretes. The second one is to identify the best prediction model of the three models, which can not only enable reliable predictions but also support future research, as a baseline prediction model for comparison against other ensemble models.

Machine Learning Model

This section presents an overview of three machine learning (ML) models implemented in this study. In each of the following sub-sections, one or more original references are provided that describe the formulation and implementation of the ML model in more detail.

Support Vector Machine (SVM)

SVM, a ML model for both classification and regression tasks, predicts new patterns based on the training data as the goal of learning a maximum-margin hyperplane in the feature space [4]. The maximum hyperplane happens while the decision boundary has the maximal distance from any training data. During the training step, input variables are mapped from a low-dimension to a high-dimension feature space via kernel functions function included polynomial function, sigmoidal function, and Gaussian radial basis kernel function [5], [6]. The SVM attempts to determine a linear objective function $f_{SVM}\ (x,\ \omega)$ (see **Eq. 1**) that has a maximum deviation of ε with respect to the actual value in the training dataset.

$$f_{SVM}\ (x, \omega) = \sum_{i=1}^{n} \omega_i K_i\ (x) + b \qquad \textbf{(1)}$$

In **Eq. 1**, K_i represents the set of n nonlinear kernel functions which are used to transform the original input data (x) into higher dimensional feature space, b represents a bias term, and ω is the weight vector consisting of n choice coefficients.

Introduce non-negative slack variables ξ_i and ξ_i^* to the function, and the above optimization problem is reduced to the following problem

$$\text{Minimize: } \frac{1}{2}\|\omega\|^2 + C \sum_{i=1}^{n} \xi_i + \xi_i^* \qquad \textbf{(2)}$$

$$\text{subject to} \begin{cases} y_i - f_{SVM}\,(x_i, \omega) - b \leq \varepsilon + \xi_i \\ f_{SVM}\,(x_i, \omega) + b - y_i \leq \varepsilon + \xi_i^* \\ \xi_i \; and \; \xi_i^* \geq 0, and \; i = 1, 2, 3 \ldots n \end{cases} \quad \textbf{(3)}$$

Here, C is constant which represents the degree of penalty of the sample with error exceeding the magnitude of the insensitive loss function (ε). In order to derive the optimum objective function, the parameters – ε and C – and any parameter associated with kernel function should be optimized. In this study, based on comparisons of prediction performances of each kernel, the 5th order polynomial kernel was chosen for SVM model.

Random Forest (RF)

Random Forest, a collection of tree predictors, is based on two machine learning techniques: bagging and random feature selection [7], [8]. During the training process, a series of decision trees are grown, resulting in an output that averages all decision trees. Prior to split at each node, each individual tree is independently constructed with a subset of the original training set. The general construction step of the RF model can be summarized as follows:

1) "n_t" bootstrap samples are generated randomly from the original training dataset.
2) Grow unpruned regression tree for each of the n_t bootstrap datasets. The number of leaves of each tree is held constant across the entire model.
3) Each tree predicts a data-point outside of the selected bootstrap space. The output of the prediction is designated as out of bag (OOB) prediction [9].
4) Predict the output of testing data by, for a given input vector (x), aggregating and averaging the overall OOB prediction $\hat{f}_{RF}^{n_t}(x)$ and OOB error rate (see **Eq. 4**).

$$\hat{f}_{RF}^{n_t}(x) = \frac{1}{n_t}\sum_{j=1}^{n_t} f_{RF_j}(x) \quad \textbf{(4)}$$

Multilayer Perceptron Artificial Neural Network (MLP-ANN)

Artificial neural network (ANN) consists of several computational elements (termed as neurons) arranged in layers, resembling the network of neurons in the human brain responsible for processing information in a hierarchical fashion [10]. Multilayer perceptron artificial neural network (MLP-ANN) is a subclass of ANN with strong self-learning capabilities [11]. The hierarchical structure of MLP-ANN is comprised of: (i) one input layer containing a set of input nodes (ii) one or more hierarchical hidden consisting of computation nodes (iii) one output layer containing one computation node. Each neuron in any given hidden layer is functionally related – as shown in **Eq. 5** – to all neurons in the previous layer.

$$N_j = \sum w_{ji} o_i \quad \textbf{(5)}$$

Here, N_j represents the activation of the j^{th} neuron, i indicates the set of all neurons in the previous layer, w_{ji} represents the weight of the connection between neurons j and i, and o_i is the

output of the neuron. Each neuron uses activation functions to calculate intermediate-output values, which are subsequently passed on as input values to the next neuron layer. This process proceeds throughout the network until reaching the final neuron layer that produces the final output. Activation functions are represented as sigmoidal or logistic-transfer functions (**Eq. 6**)[11].

$$y_j = f(N_j) = \frac{1}{1 + e^{w_{ji} \cdot N_j}} \quad \textbf{(6)}$$

where $y_j = f(N_j)$ is the activation function of the j^{th} neuron. During the training step, a back-propagation algorithm [12] is used to minimize deviation between actual and predicted values. This is accomplished by iteratively adjusting and finally determining the optimal connection weights (i.e., w_{ji}) – pertaining to each activation function – by using the gradient descent approach or the Levenberg-Marquardt algorithm [11].

Data Collection and Evaluation of Prediction of Machine Learning Models

Dataset-1, published by Yeh et al. [13], [14], consists of 1030 data-records, featuring 278 unique concrete mixture designs and compressive strengths. In each data-record, there are eight input variables – contents of cement (kg. m^{-3}), blast furnace slag (kg. m^{-3}), fly ash (kg. m^{-3}), superplasticizer (kg. m^{-3}), water (kg. m^{-3}), fine aggregate (kg. m^{-3}) and coarse aggregate (kg. m^{-3}), and age (days); and one output – compressive strength (MPa). Statistical parameters pertaining to Dataset-1 are shown in **Table I**.

Table I: A summary of statistical parameters pertaining to each of the 9 attributes (8 input and 1 output) of Dataset-1. The dataset consists of 1030 unique data-records.

Attribute	Unit	Min.	Max.	Mean	Std. Dev.
Cement	kg. m^{-3}	102.00	540.00	281.27	104.51
Blast Furnace Slag	kg. m^{-3}	0.0000	359.40	73.896	86.279
Fly Ash	kg. m^{-3}	0.0000	200.10	54.188	63.997
Water	kg. m^{-3}	121.80	247.00	181.57	21.354
Superplasticizer	kg. m^{-3}	0.0000	32.200	6.2050	5.9740
Coarse Aggregate	kg. m^{-3}	801.00	1145.0	972.92	77.754
Fine Aggregate	kg. m^{-3}	594.00	992.60	773.58	80.176
Age	Days	1.0000	365.00	45.662	63.170
Compressive Strength	MPa	2.3300	82.600	35.818	16.706

Dataset-2 was published by Sadati et al. [15]. This dataset comprised 526 unique data-records. This dataset contains 13 inputs and 1 output. The 13 inputs parameters included: type of binder ("0" for plain binder and "1" for binary/ternary binder); contents (in kg. m^{-3}) of cement, supplementary concrete material (SCM), natural coarse aggregate, recycled concrete aggregate (RCA), fine aggregate, and water; and density (in kg. m^{-3}), water absorption capacity (in %) and maximum aggregate size (in mm) of natural coarse aggregate and RCA. The output parameter included the 28-day MOE (in GPa) of all concretes. Statistical parameters pertaining to the database are summarized in **Table II**.

Table II: A summary of statistical parameters pertaining to each of the 14 attributes (13 input and 1 output) of the Dataset-2. The database consists of 526 unique data-records.					
Attribute	**Unit**	**Min.**	**Max.**	**Mean**	**Std. Dev.**
Binder type	Unitless	1	2		
Cement content	kg. m^{-3}	150.00	597.00	338.68	77.21
SCMs (fly ash and/or slag) content	kg. m^{-3}	0.00	225.09	32.32	57.82
Natural aggregate (coarse) content	kg. m^{-3}	0.00	1950.00	563.09	434.25
RCA (coarse) content	kg. m^{-3}	0.00	1800.00	495.38	423.50
Fine aggregate content	kg. m^{-3}	465.00	1301.10	730.69	121.87
Natural agg. water absorption capacity	%	0.20	6.10	1.22	0.77
RCA water absorption capacity	%	1.93	18.91	5.38	2.33
Natural aggregate density	kg. m^{-3}	2482.79	2880.00	2616.63	84.67
RCA density	kg. m^{-3}	1800.00	2602.00	2312.22	121.88
Natural aggregate max. particle size	mm	8.00	32.00	20.00	3.80
RCA max. particle size	mm	8.00	32.00	18.95	4.76
Water	kg. m^{-3}	108.30	234.00	170.69	31.55
28-day MOE (output)	GPa	11.30	54.80	30.41	7.81

For training and evaluation of prediction performances of above three ML models, each above database was randomly split into two subsets: a training set (75% data-records) and a testing set (remaining 25% data-records). The training set was used to finalize and optimize ML model parameters, and testing set used to determine the cumulative error between predicted and measured values. Five different statistical parameters were used to assess the prediction performances of three ML models. The five statistical parameters include Person correlation coefficient (R), coefficient of determination (R^2), mean absolute percentage error (MAPE), mean absolute error (MAE), and root mean squared error (RMSE). To gain a comprehensive evaluation of prediction performance of each model, the five statistical parameters were unified into a composite performance index (CPI) [16].

Results and Discussion

As described in the previous section, three machine learning models were firstly trained by 75% of the database and then prediction performances of each model were evaluated by against the rest 25% of the database. Predictions of mechanical properties of concretes from Database 1, and Database 2, as predicted by three ML models implemented in this study, are shown in **Figure 1, and 2**, respectively, and statistical parameters of each model are enumerated in **Table III, and IV**.

As shown in **Figure 1** and **Table 3**, all ML models presented in this study were able to predict the age-dependent the compressive strength of concrete with reasonable accuracy. This is evidenced by the relatively low and high values of RMSE (ranging between 4.51-and-6.33 MPa) and R^2 (ranging between 0.86-and-0.93). Based on the values of CPI, the prediction performances of ML models can be ranked as RF > SVM > MLP-ANN. As can be seen in **Figure 2** and **Table 4**, all ML models produced predictions with reasonable accuracy, with the values of R ranging

from 0.66 to 0.91, and RMSE ranging from 6.02 GPa to 3.34 GPa. Based on the values of CPI, the prediction performances of the ML models can be ranked as RF > MLP-ANN> SVM.

Usually, the prediction performance of SVM is good, especially when applied to predict the compressive strength of concrete [2], [17]. The inferior prediction performance of Database-2 can be explained as that SVM models, very much like ANN models, rely on local search and optimization algorithms; as such, they suffer from the drawback of converging to a local minimum rather than the global minimum, especially when the relationship between input variables and output in the training dataset contains several closely-placed local minima.

Predictions made by the MLP-ANN model were more accurate compared to the SVM model. Nevertheless, the prediction performance of the MLP-ANN model could also be compromised due to its inherent susceptibility to converge to a local – as opposed to the global – minimum [18], [19]. However, in this study, the hyper-parameters of the MLP-ANN model were rigorously optimized through the 10-fold CV method; on account of this optimization, it is expected that the aforementioned drawback of the model was, at least partially, overcome, thereby allowing the model to produce predictions with reasonable accuracy.

The RF model outperformed the two aforementioned models in terms of prediction accuracy. This is expected because, in the RF model, a large number of trees are grown without pruning or smoothening. On account of having a large number of unpruned trees, splits into data are more logical, and, therefore, errors resulting from generalization are minimized and overfitting of the training data is mitigated [9]. Furthermore, because of the two-stage randomization employed in the RF model correlation among unpruned trees is minimized (diversity among trees is high), the bias is kept low, and variance is significantly reduced.

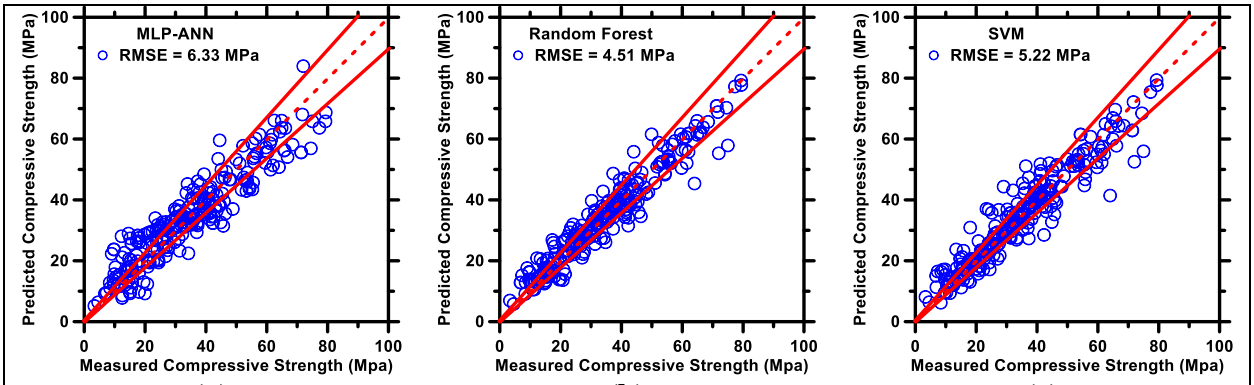
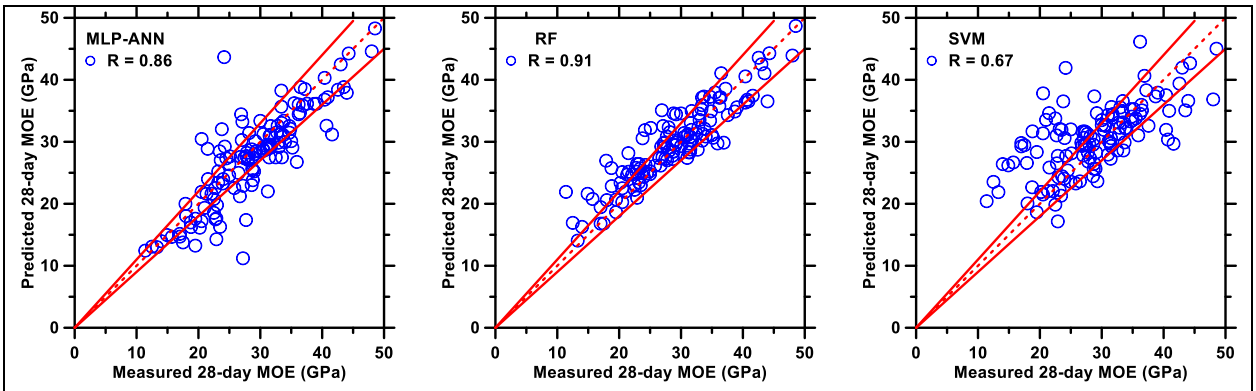


Figure 1: Predictions made by ML models: **(a)** MLP-ANN; **(b)** RF; **(c)** SVM compared against actual compressive strength of concretes (drawn from Dataset-1). The dashed line represents the line of ideality and the solid lines represent a ±10% bound.

Table III: Prediction performance of ML models, measured on the basis of the test set of Dataset-1. Five statistical parameters (i.e., R, R^2, MAE, MAPE, and RMSE) and the composite performance index (CPI) are shown.

ML Model	**R**	**R^2**	**MAE**	**MAPE**	**RMSE**	**CPI**
	Unitless	***Unitless***	***MPa***	***%***	***MPa***	***Unitless***
MLP-ANN	0.9308	0.8664	5.0421	36.143	6.3300	1.0000
SVM	0.9525	0.9073	3.5756	25.624	5.2234	0.4385

RF	0.9654	0.9320	3.2674	23.443	4.5103	0.1999



(a) **(b)** **(c)**

Figure 2: Predictions made by ML models: **(a)** MLP-ANN; **(b)** RF; **(c)** SVM method compared against actual MOE of concretes (drawn from Dataset-2). The dashed line represents the line of ideality and the solid lines represent a ±10% bound.

Table IV: Prediction performance of ML models, measured on the basis of the test set of Dataset-2. Five statistical parameters (i.e., R, R^2, MAE, MAPE, and RMSE) and the composite performance index (CPI) are shown.

ML Model	**R**	**R^2**	**MAE**	**MAPE**	**RMSE**	**CPI**
	Unitless	***Unitless***	***GPa***	***%***	***GPa***	***Unitless***
SVM	0.6672	0.4452	4.4568	69.999	6.0226	0.9880
MLP	0.8559	0.7326	3.0973	48.646	4.3859	0.3770
RF	0.9119	0.8316	2.5198	39.577	3.3448	0.1241

Conclusion

As shown in this work, in spite of the inherently nonlinear and complex nature of the relationship between input (concrete mixture design) and output (concrete's mechanical properties), machine learning (ML) models can reliably perform predictions. The modulus of elasticity (MOE) and compressive strength of concretes are predicted by random forests (RF) model, support vector machine (SVM), and multilayer perceptron artificial neural network (MLP-ANN) models. The prediction performance of the RF model was superior compared to the other two models implemented in this study. The large number of unpruned trees, which develop logical input-output correlation and mitigate overfitting and generalization errors, contributed to the accuracy of the RF model. The further studies are needed to explore how the ensemble ML models can predict and optimize the properties of concretes.

Acknowledgement

Computational tasks were conducted in the Materials Research Center of Missouri S&T. The first and fourth author would like to acknowledge funding provided by the Leonard Wood Institute (LWI). The second author would like to acknowledge funding provided by the RECAST

University Transportation Center (at Missouri S&T) and Missouri Department of Transportation (MoDOT). The third and last authors would like to acknowledge funding provided by the National Science Foundation (NSF; CMMI: 1661609).

Reference

[1] J. Sobhani, M. Najimi, A. R. Pourkhorshidi, and T. Parhizkar, "Prediction of the compressive strength of no-slump concrete: A comparative study of regression, neural network and ANFIS models," *Construction and Building Materials*, vol. 24, no. 5, pp. 709–718, May 2010.

[2] K. O. Akande, T. O. Owolabi, S. Twaha, and S. O. Olatunji, "Performance comparison of SVM and ANN in predicting compressive strength of concrete," *IOSR Journal of Computer Engineering*, vol. 16, no. 5, pp. 88–94, 2014.

[3] A. Behnood, V. Behnood, M. M. Gharehveran, and K. E. Alyamac, "Prediction of the compressive strength of normal and high-performance concretes using M5P model tree algorithm," *Construction and Building Materials*, vol. 142, pp. 199–207, 2017.

[4] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing*, vol. 14, no. 3, pp. 199–222, 2004.

[5] S. M. Clarke, J. H. Griebsch, and T. W. Simpson, "Analysis of Support Vector Regression for Approximation of Complex Engineering Analyses," *J. Mech. Des*, vol. 127, no. 6, pp. 1077–1087, Aug. 2004.

[6] P. Garg and J. Verma, "In silico prediction of blood brain barrier permeability: an artificial neural network model," *Journal of chemical information and modeling*, vol. 46, no. 1, pp. 289–297, 2006.

[7] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC Bioinformatics*, vol. 8, no. 1, p. 25, Jan. 2007.

[8] K. J. Archer and R. V. Kimes, "Empirical characterization of random forest variable importance measures," *Computational Statistics & Data Analysis*, vol. 52, no. 4, pp. 2249–2260, Jan. 2008.

[9] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, Aug. 1996.

[10] R. J. Schalkoff, *Artificial neural networks*, vol. 1. McGraw-Hill New York, 1997.

[11] M. W. Gardner and S. R. Dorling, "Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences," *Atmospheric Environment*, vol. 32, no. 14, pp. 2627–2636, Aug. 1998.

[12] T. Mueller, A. G. Kusne, and R. Ramprasad, "Machine learning in materials science: Recent progress and emerging applications," *Reviews in Computational Chemistry*, vol. 29, pp. 186–273, 2016.

[13] I.-C. Yeh, "Modeling of strength of high-performance concrete using artificial neural networks," *Cement and Concrete research*, vol. 28, no. 12, pp. 1797–1808, 1998.

[14] I.-C. Yeh, "Modeling concrete strength with augment-neuron networks," *Journal of Materials in Civil Engineering*, vol. 10, no. 4, pp. 263–268, 1998.

[15] S. Sadati, L. E. Brito da Silva, D. C. Wunsch, and K. H. Khayat, "Artificial Intelligence to Investigate Modulus of Elasticity of Recycled Aggregate Concrete," *ACI Materials Journal*, vol. 116, no. 1, Jan. 2019.

[16] V. Chandwani, V. Agrawal, and R. Nagar, "Modeling slump of ready mix concrete using genetic algorithms assisted training of Artificial Neural Networks," *Expert Systems with Applications*, vol. 42, no. 2, pp. 885–893, 2015.

[17] J.-S. Chou, C.-F. Tsai, A.-D. Pham, and Y.-H. Lu, "Machine learning in concrete strength simulations: Multi-nation data analytics," *Construction and Building Materials*, vol. 73, pp. 771–780, Dec. 2014.

[18] G. Zhang, B. E. Patuwo, and M. Y. Hu, "Forecasting with artificial neural networks:: The state of the art," *International journal of forecasting*, vol. 14, no. 1, pp. 35–62, 1998.

[19] R. Cook, J. Lapeyre, H. Ma, and A. Kumar, "Prediction of Compressive Strength of Concrete: A Critical Comparison of Performance of a Hybrid Machine Learning Model with Standalone Models," *ASCE Journal of Materials in Civil Engineering*, vol. Re-submitted after minor revision, p. 41, 2019.