Machine learning can predict setting behavior and strength evolution of hydrating cement systems

Tandré Oey ([a]), Scott Jones ([b]), Jeffrey W. Bullard ([b]), and Gaurav Sant ([a,c,d,e])*

Abstract

Setting and strength development of ordinary portland cement (OPC) binders is a complex process that involves multiple interacting chemical reactions, which result in the formation of a solid microstructure. A long-standing yet elusive goal of the cementitious materials community has been to establish a basis for prediction of the properties and performance of concrete using knowledge of the chemical and physical attributes of its components – OPC, sand, stone, water, and chemical admixtures – together with the environmental conditions under which they react. Machine learning provides a *data-driven* basis for the estimation of properties, and has recently been applied to estimate the 28 d (compressive) strength of concrete simply from knowledge of its mixture proportions [1]. Building on this success, the current work uses a diverse dataset of different ASTM C150 cements, the chemical composition and other attributes of which have been measured. Machine learning (ML) estimators were trained with this dataset to estimate both paste setting time and mortar strength development as a function of the OPC composition and fineness. The ML estimation errors are typically similar to or lower than the measurement repeatability of the relevant ASTM test methods. ML therefore can be used to estimate the influence of binder composition and fineness on the engineering properties of cementitious systems. This creates new opportunities to apply data intensive methods to optimize concrete formulations under multiple constraints of cost, CO_2 impact, and performance attributes.

Keywords: cement composition, fineness, strength, setting, machine learning

[a] Laboratory for the Chemistry of Construction Materials (LC^2), Department of Civil and Environmental Engineering, University of California, Los Angeles, CA, USA
[b] Engineering Laboratory, National Institute of Standards and Technology, Gaithersburg, MD, USA
[c] Institute for Carbon Management, University of California, Los Angeles, CA, USA
[d] Department of Materials Science and Engineering, University of California, Los Angeles, CA, USA
[e] California Nanosystems Institute, University of California, Los Angeles, CA, USA

Corresponding author: Tandre Oey (Email: tandreoey@gmail.com)

1.0 Introduction

The hydration of ordinary portland cement (PC) entails multiple concurrent chemical reactions [2]. These reactions cause extensive changes in phase assemblage and microstructure, which in turn determine the time-dependent evolution of concrete properties and performance such as setting time and compressive strength. Mature (28 d) compressive strength is the metric most commonly used to specify and qualify a concrete for structural design [3]. Multi-scale simulations suggest the need to couple microstructural and mechanical models as a means to predict time-dependent mechanical properties [4]. However, these approaches are still severely limited by gaps in knowledge of OPC's hydration process and its constituent mechanisms, and are generally unable to forecast the evolution of properties and performance unless they are experimentally, and narrowly, calibrated to the specific system of interest [5].

In the absence of knowledge needed to predict cement hydration rates and associated changes in properties, data-driven machine learning (ML) methods offer an attractive, mechanism-agnostic approach for estimating engineering properties such as the 28 d compressive strength of concrete [6–15]. Young *et al*. [1] have recently demonstrated that ML can, when trained on enough data, make reasonable estimations of the 28 d compressive strength of field-produced concretes as a function of its attributes such as water-to-cement mass ratio (w/c), aggregate content, and the content and type of mineral and chemical admixtures. Such results demonstrate the potential of ML approaches for predicting concrete performance because the data that were used therein were obtained for concrete produced under the *relatively uncontrolled* conditions of diverse construction sites. The predictions could likely be made even more accurate by including site-specific variables such as temperature and humidity changes with time. However, the study was limited to concretes produced with Type I/II PC.

To supplement and extend existing models, the current study takes another step toward truly predictive models of concrete properties by applying ML methods to estimate the effects of OPC characteristics, such as chemical composition and fineness, on target performance characteristics such as paste setting time and mortar compressive strength. In addition, a tentative lower bound on the number of data records that are required for future estimation of other concrete properties is established. Special focus is paid to identify potential technical barriers faced by ML methods to identifying *general trends* among thousands of data points and, more importantly, to accurately predict the properties of any *one* material of interest.

2.0 Background and Methods

2.1 Machine learning algorithms

Young *et al*. showed that bootstrap-aggregated (or bagged) decision tree ensembles can accurately estimate the 28 d compressive strength of concrete when trained on large datasets with potentially high inherent variability [1]. These rule-based estimators identify logical splits in data, partitioning the input space into a tree of decision nodes that are traversed until arriving at a final prediction of the target, called a leaf node. A simple operation, such as the multiplication of the input by a constant, produces the output estimation from each leaf node. A collection, or ensemble of trees are constructed, each tree being trained on different data sets and attributes, and their results are then averaged to produce the final prediction of the target [16]. This study focuses on three different decision tree ensembles because of their ability to estimate field concrete compressive strength [1]. The first method is a bagged[*] tree ensemble, which bootstrap samples *n* different subsets of the training data with replacement to train *n* trees. Other than the random sampling from the training data, the method is deterministic in the sense that the decision nodes are chosen from among all attributes using a deterministic function such as information gain or Gini index [17]. In addition, the threshold value for splitting at a decision node is chosen to be that which optimizes that deterministic function. The second method, a random forest ensemble, differs from the bagged tree ensemble in that it selects the attribute chosen for each decision node from among a randomly chosen small subset of the attributes. The third method, called extra[†] trees, is the same as a random forest except that the threshold value for splitting a decision node is also chosen at random instead of being prescribed by optimization of a thresholding function [18]. Other ML estimators besides these three tree ensembles were also examined, including basic linear regression and K-nearest neighbor (K-NN) regression [19–21]. The tree ensembles provided the highest prediction accuracy for every attribute, although the results of the other regression methods are also shown for comparison. All the algorithms used for estimator construction are regressors from the scikit-learn library, and can be accessed and downloaded, along with their documentation, at http://scikit-learn.org/stable/ [19].

2.2 Data collection and preprocessing

Two datasets were utilized. The first dataset was provided by the Cement and Concrete Reference Laboratory (CCRL) Proficiency Sample Program, which issues four OPCs each year for comprehensive physical and chemical testing by nearly 200 different laboratories. This dataset consists of measurements of 48 attributes of a given OPC sample (see Table 1), as established by ASTM test methods [22]. The second dataset is a compilation of different industry survey data supplied by the Portland Cement Association (PCA) and the National Institute of Standards and Technology (NIST), formerly the National Bureau of Standards (NBS). This dataset comprises 2211 different PCs characterized by an unknown number of testing institutions using standard test methods. It also includes the averages[‡] of 19 of the 48 attributes for each of the CCRL cements (marked in bold in Table 1). Two other attributes, normal consistency and final setting time, were also reported in the majority of records available, and so were also considered in this study (italicized in Table 1). The bolded entries in the “Chemical Tests” column of Table 1 were used as inputs to the final ML estimators, along with Blaine fineness, while the bolded and italicized entries in the “Physical Tests” column of Table 1, with the exception of Blaine fineness, were used as targets for ML prediction using these estimators.

Prior to use as inputs and targets in the machine learning estimators, the data were preprocessed to remove obvious errors and to ensure they would be compatible with all the ML algorithms used. First, on an attribute-by-attribute basis, unphysical or meaningless values were deleted. Among these were percentages outside the range of 0 % to 100 %[§] and unphysical values such as negative setting time or compressive strength. Second, a filter was applied to each attribute to delete any outliers, which we defined according Chauvenet’s criterion [23] as more than four

[*] The term “bagged” is a portmanteau of the terms “bootstrap” and “aggregated”.

[†] The term “extra” is a portmanteau of the terms “extremely” and “randomized”.

[‡] Use of averages was necessary to ensure that no cement was over-represented in the input data to ML models, as this is known to negatively impact ML estimator performance.

[§] Any percentage values in excess of 100 % or below 0 % were retained only if they were physically meaningful. For example, negative percentages in autoclave expansion measurements correspond to shrinkage.

standard deviations from the mean of that attribute across all cements. The mean(s) were recalculated after those outliers were removed and the filter was reapplied, the process being repeated until no more outliers were identified. Less than 0.05 % of the data were discarded by this filtering for any given attribute, and the process of omitting outliers required only three iterations. Afterward, duplicate records (that is, identical cements) were deleted and any missing attributes were replaced by mean imputation, setting each missing value to the mean for the appropriate attribute as determined using data from the other cements. This is the simplest of all methods of data imputation, used in situations when data are missing completely at random, i.e., when the absence of a value is unrelated to the state of the system or values of other variables.

Of the two datasets, that from the CCRL contains the greater number of records, nearly 31 000, and has a more comprehensive list of potential attributes to be used as inputs or targets for ML estimators. However, that dataset is also missing more data, contains many more duplicates (consisting of only about 200 unique cements), and consequently was unable to train any ML estimators as accurately as the composite survey dataset. The CCRL data were incorporated in the composite survey dataset, however, by using the mean value of each attribute for each cement instead of the individual records. Randomly shuffling the order of data records proved essential for effectively training the ML estimators regardless of the algorithm used. This indicates that the ranges of attribute values are not homogeneously distributed across the different surveys in the compilation, and that leaving the data grouped by survey alone introduces an inadvertent bias in the sampling of input attributes toward one particular study. Therefore, random shuffling as implemented herein is an effective and necessary way to ameliorate that artifact.

Table 1: The cement attributes provided in the datasets, and the ASTM standards [22] (in square brackets) used to measure them. The boldfaced entries are reported consistently for nearly all cements in the full dataset, and italicized entries are reported in at least 50 % of the records in the dataset. Other entries were not consistently reported and were excluded from inputs to ML estimators. All boldfaced and italicized entries listed under "Physical Tests," with the exception of Blaine Fineness, were utilized as target attributes in this study, and as such were also excluded from inputs to ML estimators. All other entries that were excluded from inputs to ML estimators were verified to be of minimal importance to estimator performance, as outlined in Section 3.2.

Chemical Tests	**Physical Tests**
SiO_2 (mass %) [C114]	*Paste Normal Consistency (%), [C187]*
Al_2O_3 (mass %) [C114]	**Vicat Paste Initial Set (minutes), [C191]**
Fe_2O_3 (mass %) [C114]	*Vicat Paste Final Set (minutes) [C191]*
CaO (mass %) [C114]	Gillmore Initial Set (minutes) [C266]
C_3S (mass %) [C150]	Gillmore Final Set (minutes) [C266]
C_2S (mass %) [C150]	False Set (%) [C451]
C_3A (mass %) [C150]	Autoclave Expansion (%) [C151]
C_4AF (mass %) [C150]	Air Content (%) [C185]
Free CaO (mass %) [C114]	Air Content Mixing Water (%) [C185]
MgO (mass %) [C114]	Air Content Mixture Flow (%) [C185]
SO_3 (mass %) [C114]	**3 Day Mortar Compressive Strength (MPa) [C109]**
Na_2O (mass %) [C114]	**7 Day Mortar Compressive Strength (MPa) [C109]**
K_2O (mass %) [C114]	**28 Day Mortar Compressive Strength (MPa) [C109]**
Loss on Ignition (mass %) [C114]	Mortar Compressive Strength Mixture Flow (%) [C109]
Insoluble Residue (mass %) [C114]	**Blaine Fineness (m^2/kg) [C204]**

Carbon Dioxide (mass %) [C114]	Wagner Fineness (m^2/kg) [C115]
Limestone (mass %) [C114]	Sieve Fineness (% passing) [C430]
ZnO (mass %) [C114]	0 Day Heat of Solution (cal/g) [C186]
Mn_2O (mass %) [C114]	7 Day Heat of Solution (cal/g) [C186]
P_2O_5 (mass %) [C114]	28 Day Heat of Solution (cal/g) [C186]
TiO_2 (mass %) [C114]	7 Day Heat of Hydration (cal/g) [C186]
Cl (mass %) [C114]	28 Day Heat of Hydration (cal/g) [C186]
	Mortar Bar Expansion (%) [C1038]
	Mortar Bar Mixing Water (%) [C1038]
	Mortar Bar Flow [C1038]

2.3 Estimator optimization

Numerous machine learning estimators were constructed and applied to predict initial set (minutes), 3 d compressive strength (MPa), 7 d compressive strength (MPa), and 28 d compressive strength (MPa). These targets were chosen because the first three affect the scheduling of construction operations, and the 28 d strength is both an input for structural design and a specification criterion. Each estimator was trained and tested on the combined datasets with the performance of each estimator being evaluated using several error metrics. Both training and testing were conducted on different portions of data using a standard low-bias resampling procedure called k-Fold Cross-Validation* [24,25]. The data records were randomly split into k = 10 "folds," nine of which were used to train the estimator, and one of which was used to evaluate the estimator after training. The process was then repeated nine additional times, each time using a different fold as the test set, and the remaining nine folds as the training set.

The estimators used in this study are sensitive to the magnitude of the attributes in the sense that they will be biased to assign more importance to attributes with inherently greater values. For example, merely changing the units of Blaine fineness of the powder from m^2/kg to cm^2/g increases the numerical value by a factor of ten and can influence the accuracy of the estimators even though the physical data are the same. To address this kind of artifact, after the training and testing sets were identified and separated, the data for each attribute were rescaled to a standard normal distribution (mean = 0, variance = 1). This step was taken after the separation of the training and testing sets to avoid data leakage (i.e., the unintentional passing of information about the test set to the training set) which could potentially happen if the combined testing and training data were rescaled together.

Estimator optimization was performed by determining extremal values of one of three objective functions that characterize the overall fidelity of the predictions to the actual values in the testing set. The objective functions are the root mean square error (RMSE), the mean absolute percentage error (MAPE), and the coefficient of determination (R^2):

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(P_i - A_i)^2}{n}} \tag{1}$$

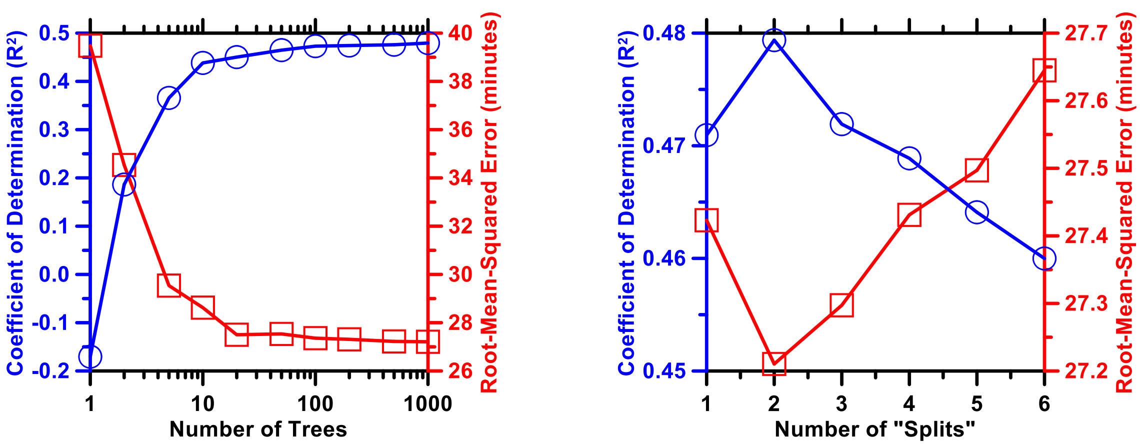
* Cross-validation is necessary to evaluate how machine-learning estimators are likely to perform when making predictions on previously unseen data: a portion of the data are taken as a training set and used to train and optimize the model, and the remainder of the data are withheld as a testing set to evaluate the model's performance.

$$\text{MAPE} = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{P_i - A_i}{A_i}\right| \tag{2}$$

$$\text{R}^2 = 1 - \frac{\sum_{i=1}^{n}(P_i - A_i)^2}{\sum_{i=1}^{n}(A_i - \bar{A})^2} \tag{3}$$

where n is the number of records in the testing set, P_i and A_i are the predicted and actual target value of the i^{th} record in the testing set, respectively, and $\bar{A}$ is the arithmetic average of the actual target values. RMSE and MAPE indicate the average departure of estimated values from actual values, whereas R^2 is the fraction of the variance of the target values that is predictable from the attributes using the model. As described in Section 3.1, low RMSE and MAPE values may still be achieved even when the data are relatively scattered and the R^2 value is low. This has also been observed previously [1] and suggests that an error-based metric such as MAPE is a better test of estimator performance than R^2 because it can be compared more directly with the acceptable range of physical test values for attributes such as setting time or strength [22].

Each of the machine learning estimators were finalized by optimizing their estimation performance via hyperparameter tuning. This procedure varied both the number of trees used in random forest estimators and the number of attributes considered per tree split when partitioning the input space. The results of this hyperparameter tuning, shown in Figure 1, indicate that estimator performance improves only marginally beyond a certain number of trees. Consequently, the final estimators reported here employ only 100 trees to avoid over-fitting of the training data, and the extremely-random forest estimators employ only two attributes per "split," for similar reasons. These fully-optimized ML estimators are a substantial improvement (roughly a two-fold reduction in MAPE) over prior work [1], by merit of their consideration of cement composition.



(a) (b)

Figure 1: The results of a representative parameter tuning exercise for the extremely random forest estimators constructed to estimate initial setting time, showing: **(a)** A plateau in estimator performance with increasing number of trees (i.e., in each case using two attributes to determine each partitioning of the input space), and **(b)** A modest optimum of two splits is observed when using 1000 trees.

3.0 Results and discussion

3.1 Estimation accuracy for a given target is comparable to ASTM repeatability limits

Among the algorithms examined, ensembles of decision trees consistently produced the lowest errors, as shown in Table 2. Of the tree ensembles, the extra trees estimator most accurately estimated every primary attribute as measured by MAPE. The error metrics are not much greater in magnitude than the reported repeatability of the corresponding ASTM test methods, reported as a coefficient of variation, though there is no standard ML error parameter that would enable more direct comparisons [22]. For example, the MAPE for 7 d compressive strength predictions by extra trees estimators is 6.58 %, less than twice the single-operator coefficient of variation of the measurement using ASTM C109 (3.8%). Similarly, the variability in initial set time for that estimator, 25.5 minutes, is considerably less than the acceptable range of two successive measurements using ASTM C191 (34 minutes). This suggests that, for cement compositions covered by ASTM C150, ensemble machine learning approaches may reliably estimate the *average properties* and performance of paste / mortar formulations nearly as well or better than they can be repeatably measured in the lab.

Table 2: The results of 10-fold cross-validation using the following error metrics: root mean square error (RMSE), coefficient of determination (R^2), and mean absolute percentage error (MAPE). The input attributes were SiO_2 (mass %), Al_2O_3 (mass %), Fe_2O_3 (mass %), CaO (mass %), SO_3 (mass %), and Blaine fineness (m^2/kg), as determined by attribute importance in Section 3.2.

Target Attributes for Ten-Fold Cross-Validation												
	Initial Set Time[b]			**3 Day Strength[c]**			**7 Day Strength[c]**			**28 Day Strength[c]**		
Estimator	***RMSE (min)***	***R^2***	***MAPE (%)***	***RMSE (MPa)***	***R^2***	***MAPE (%)***	***RMSE (MPa)***	***R^2***	***MAPE (%)***	***RMSE (MPa)***	***R^2***	***MAPE (%)***
Linear	29.6	0.392	17.7	3.26	0.676	9.01	3.59	0.573	7.90	4.01	0.305	7.06
K-NN[a]	27.8	0.437	15.9	3.25	0.691	8.32	3.48	0.614	7.27	3.77	0.394	6.35
Decision Tree Ensemble Estimators:												
Bagged	26.2	0.524	15.0	2.79	0.766	7.29	3.18	0.668	6.67	3.50	0.489	5.91
Random	25.6	0.541	14.9	2.82	0.763	7.35	3.15	0.674	6.68	3.48	0.497	5.87
Extra	**25.5**	**0.547**	**14.7**	**2.82**	**0.762**	**7.29**	**3.14**	**0.675**	**6.58**	**3.44**	**0.506**	**5.79**
Boosted[a]	29.0	0.417	17.7	3.44	0.646	10.0	3.63	0.567	8.19	3.89	0.368	6.81
Gradient[a]	26.9	0.495	15.7	2.89	0.749	7.74	3.30	0.642	6.93	3.59	0.460	6.13

[a] K-nearest neighbors, boosted decision trees, and gradient boosted decision trees were also used, among other estimators (not shown), as they are likely to perform similarly to bagged decision trees. None performed better for these target attributes. For details regarding the implementation, see http://scikit-learn.org/stable/.

[b] ASTM C191.

[c] ASTM C109.

3.2 Higher errors for late-age strength suggest missing data attributes

Table 2 shows that estimator performance for predicting compressive strength is progressively poorer at later ages, regardless of the estimator used. For example, the RMSE of the extremely randomized forest estimator increases from 2.82 MPa at 3 d to 3.14 MPa and 3.44 MPa at 7 d and 28 d, respectively. Despite the somewhat poorer estimator performance for 28 d strength compared to earlier times, both the MAPE and RMSE for 28 d strength estimates are modestly better than those determined by Young *et al*. [1] for industrially produced concretes using similar estimators, likely due to more detailed knowledge of mixture and material characteristics in the current study (cement composition, fineness). In any case, the greater errors at later ages may indicate that the available datasets are missing some important attributes that influence compressive strength at later ages.

One possible reason for this decrease in accuracy at later ages may be inconsistent or poorly controlled curing conditions in practice, the effects of which would become progressively more important with time. It is impossible to assess the likelihood of that possibility based on the data alone, however, because there are no requirements in ASTM C109 to report the imposed degree of control over curing temperature or moisture conditions. A second possible reason for increased error is air entrainment in some subset of the measurements, given that ASTM C109 allows the user to decide whether or not the sample will contain entrained air – macroscopic air voids stabilized by chemical admixtures to improve freeze-thaw resistance – requiring a lower water-cement mass ratio (w/c) of 0.460 than the value of 0.485 required for samples without air entrainment. Finally, differences in water content may play a significant role in poor estimator performance for initial setting time measured using ASTM C191, wherein the mixture must be prepared with "normal consistency" as measured by ASTM C187, which is the empirically determined water content required to achieve a prescribed paste stiffness after 30 s of mixing with 0.65 kg of cement powder (varying from about 22 % to 30 % of the powder mass among different PCs). Therefore, ML estimation of normal consistency has also been investigated, as it may serve as a proxy for w/c and is available in some of the compiled survey data.

Table 3: Results of 10-fold cross-validation for the final machine learning estimators of secondary targets with partial data records, evaluated using the same error metrics given in Table 1. The best-performing estimator (lowest MAPE) is marked in bold. The number of available data points used in each estimator is also reported.

Target Attribute for Ten-Fold Cross-Validation								
	Final Set				**Normal Consistency**			
Estimator	***RMSE (min)***	***R^2***	***MAPE (%)***	***Data Points***	***RMSE (%)***	***R^2***	***MAPE (%)***	***Data Points***
Linear	60.1	0.422	18.4	1144	1.04	0.292	2.96	1447
K-NN	59.6	0.432	17.6	1144	0.935	0.427	2.29	1447
Trees:								
Bagged	55.5	0.505	16.6	1144	0.920	0.446	2.28	1447
Random	55.5	0.505	16.6	1144	0.935	0.427	2.29	1447
Extra	**54.7**	**0.513**	**16.4**	1144	**0.894**	**0.471**	**2.23**	1447
Boosted	57.6	0.461	17.8	1144	1.11	0.193	3.31	1447
Gradient	57.9	0.461	17.5	1144	0.999	0.358	2.49	1447

[a]ASTM C191.
[b]ASTM C187.

3.3 Estimation of secondary targets suggests a limited ability to account for missing attributes

Among the other attributes in the dataset besides initial set and compressive strength, both normal consistency and final setting time were reported frequently enough to construct viable estimators. Estimators for these secondary targets, results of which are given in Table 3, were indeed about as accurate as those for primary targets in Table 2. However, in contrast to the primary targets, the errors in estimating normal consistency are significantly higher than the tolerances listed in its associated ASTM C187 test method. Nevertheless, the normal consistency estimators have the lowest MAPE of any estimator used in this study. ASTM C187 uses OPC pastes prepared with normal consistency, so the estimator's ability to capture the dependence of normal consistency on composition and fineness may explain why ML estimators are able to predict initial and final setting times from those same attributes despite the fact that the w/c used can be different for each cement. In other words, cement details such as fineness are able to at least *somewhat* capture this indicator of "water demand" of a cement, but there likely are other powder characteristics – perhaps microscale texture or grinding aid type or dose – that affect normal consistency but are currently not being measured by standard test methods. This example highlights both a limitation of, and an opportunity for, ML methods: they can estimate certain aspects of concrete performance from routinely collected data, but they can also identify other performance attributes, the systematic estimation of which requires additional or perhaps qualitatively different material characterization. Similarly, as taken up in the next section, it is helpful for understanding to identify which currently-measured attributes contribute most strongly to the quality of ML estimations of different targets.

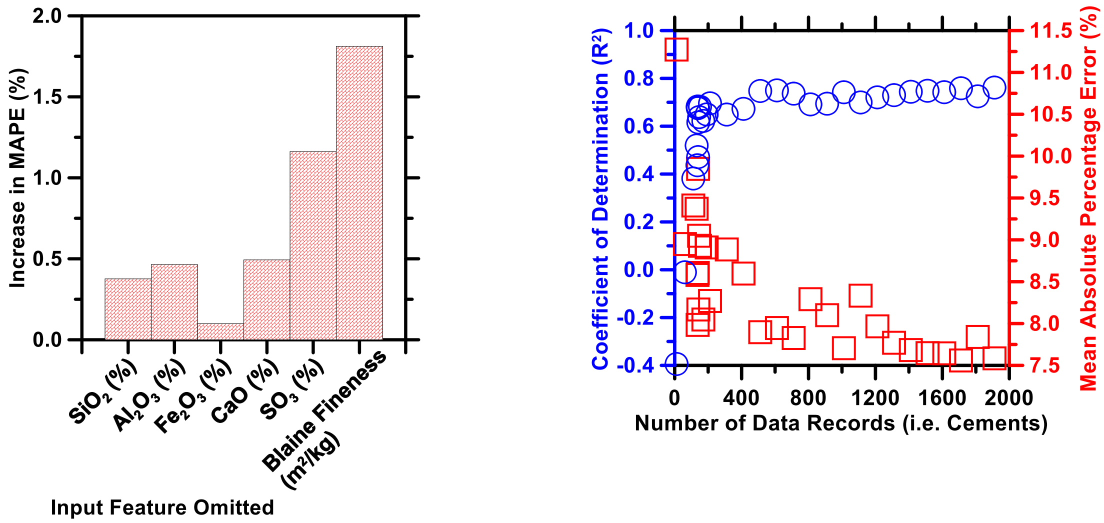
3.4 Selective omission identifies six attributes necessary for estimation of set and strength

One can evaluate the relative importance of the different attributes in determining estimator performance in predicting the primary targets (initial set, compressive strength) by eliminating them one at a time from the training set. The corresponding increase in MAPE was used as a quantitative measure of attribute importance, as shown in Figure 2(a). Unsurprisingly, cement fineness is by far the most influential input attribute, followed by the oxides of sulfur, calcium, aluminum, silicon, and iron. Similar attribute rankings were obtained for all targets estimated. This is reassuring because (i) available surface area is well known to be a key factor that affects cement reaction rates and water demand, (ii) calcium and aluminum bearing cement phases such as tricalcium silicate (C_3S[*]) and tricalcium aluminate (C_3A), are known to be the most reactive cement phases, and (iii) proper sulfation of a cement is empirically known to influence setting and early-age strength gain. Predictions showed only marginal improvement upon inclusion of any other other attributes from Table 1 besides these six, such as minor oxides (Mg, Na, K), loss on ignition, or free lime content. Whether added alone or in combination with other such attributes, none affected the MAPE by more than 0.1 %. Replacing the four major oxides with the Bogue estimates of the four major clinker phases also did not improve estimator performance, which is understandable because the Bogue estimates are merely linear functions of the oxide proportions.

3.5 Random omission identifies a tentative lower bound on data needed to train estimators

Having now established the minimum attributes necessary for predicting the primary targets, we now turn attention to determining the minimum number of data records needed to make accurate target estimates. This measure of robustness of the different ML algorithms, when applied to these datasets, can be evaluated by retraining them with a sparse subset of the data. Specifically, learning curves were constructed by randomly omitting data records from the input, as illustrated in Figure 2(b). For convenience in terminology, we define "data-sufficiency" as the minimum number of data records at which the learning curves plateau. Figure 2(b) shows that the estimators approach peak performance, at least with respect to R^2, with less than 10 % of the available dataset; those trained with a random selection of at least 200 of the 2211 total available data records performed within about 1 % of the MAPE of the same estimators that were given access to the full training set. This suggests the viability of applying such estimators even for relatively smaller datasets and is an encouraging sign that these methods can also be used reliably even with limited field data. However, the error metrics frequently used to evaluate the quality of ML estimators, such as MAPE, are not necessarily suitable for the direct comparison between estimator accuracy *on average* and the ability of the estimator to make consistently accurate predictions of engineering properties of particular cement systems.

[*] Conventional cement chemistry notation is used: C = CaO, S = SiO_2, A = Al_2O_3.



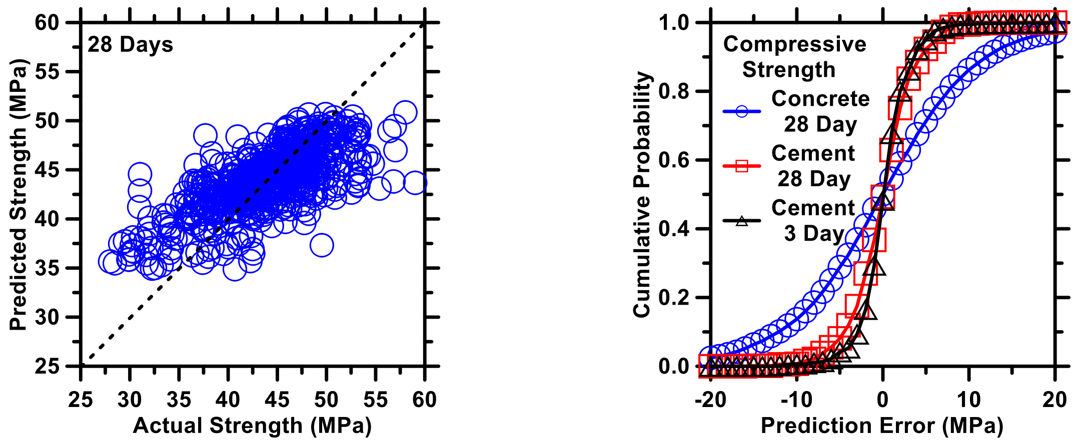
(a) **(b)**

Figure 2: Representative evaluations of estimator performance shown for the extremely random forest estimators constructed to estimate 3 d compressive strength which highlight **(a)** Attribute importance as determined by an increase in MAPE upon omission of a given input attribute, and **(b)** so-called "learning curves" for the estimator showing the minimum number of input records required to construct an adequate estimator.

3.6 New evaluative metrics are needed to properly reflect estimator prediction accuracy

The three objective functions used to score the estimator performance in this study, which are among the most commonly used scoring metrics in other machine learning efforts, reflect the estimator's performance on average for the entire dataset, which comprises many cements. However, indicators of average error such as RMSE and MAPE do *not* indicate the estimator's accuracy in predicting the target value of any *particular* cement in the testing set. Just as a significant fraction of a normally distributed population lies outside one standard deviation of its mean, so does a given estimator produce individual errors much greater than the RMSE for a significant fraction of the cements. As an example, Figure 3(a) shows the individual predictions of 28 d compressive strength made by an extremely random forest regressor with 500 trees applied to a testing set after training. The predicted value for each data record is plotted against the actual target value for that record. The RMSE for this estimator is less than 5 MPa, but the maximum error for any particular cement could be as high as 20 MPa and corresponds to a relative error of about 50 %.

To view the situation in a different way, the absolute prediction errors for 28 d strength of individual cements were collected in a histogram with 1 MPa bin widths. The histogram was converted into a normalized probability density plot, the positive portion of which is shown in Figure 3(b). For comparison, the same figure shows the corresponding histograms for 3-d compressive strength obtained in this study and for 28-d concrete strength obtained by Young *et al*. [1]. The errors have an approximately normal distribution with a peak near 2 MPa and a standard deviation of approximately 3.6 MPa. A tolerance interval for an ML estimator can then be established in a similar manner to the ASTM standard test methods. For example, given that the 28 d strength errors in Figure 3(b) are approximately normally distributed with a mean of 2 MPa and a standard deviation of 3.6 MPa, there is a 95 % probability that 90 % of the predictions will be no less than 6.2 MPa below the actual value and no more than 10.2 MPa above it. A tolerance interval this large is far from ideal. However, for comparison the interval for similar estimations from concrete mixture proportions by Young *et al.* [1] comes in at about ±15.5 MPa. As illustrated in Figure 3(b) by comparison to predictions on concrete, as well as 3 d strength, the current results clarify both the substantial improvement achieved by inclusion of attributes such as cement composition, as well as the potential future improvements that may arise from inclusion of additional attributes such as curing conditions.



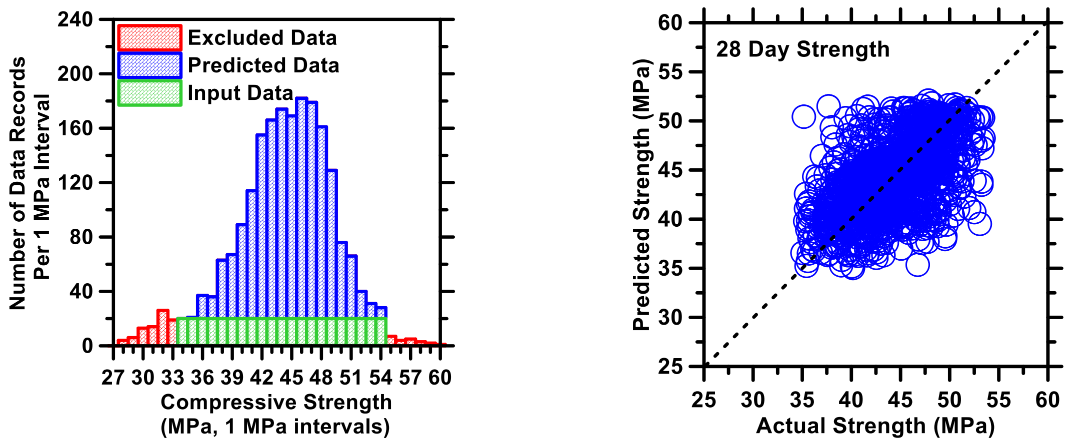
(a) **(b)**

Figure 3. The prediction results of an optimized 500-tree extremely random forest regression estimator, shown as **(a)** predicted vs actual strength values with a dashed line of identity provided to guide the eye, and **(b)** the normalized cumulative probability distribution of a prediction by the estimator having a given error. Also shown for comparison are distributions for a similar estimator applied to prediction of 3 d compressive strength of mortars (this study) and 28 d compressive strength of concretes (Young *et al.* [1]).

If ML estimators are to be used confidently for concrete mixture design and optimization, they will need to achieve much lower tolerance intervals in their predictions than are indicated herein. In statistical treatments such as those discussed above, the only way to reduce the probability that a particular estimate is outside a tolerable limit is to significantly reduce the average error values such as RMSE and MAPE, or to effectively tighten the distribution of errors about these average values. The ways to decrease average error are to provide the estimator with data that more uniformly span the range of possible values, to acquire better curated data, or to identify and collect data on other attributes that may relate more meaningfully to the target being estimated. Within the narrowly prescribed range of cement compositions and characteristics considered herein, namely ASTM C150 PCs, the dataset would appear to be easily large enough to train the estimators according to the plateau in learning curves demonstrated in Figure 2(b). Consequently, the only feasible way to reduce the unexplained variance is to develop a means for identifying relatively more inconsistent data within the currently applied dataset, or to supplement the data with measurements of other material or processing characteristics that are currently not being routinely captured including, but not restricted to, the types and dosages of chemical admixtures, the particle size distribution of the OPC, clinker grinding parameters, curing conditions, and data on the mineralogy, texture, and impurities in the individual cement components.

3.7 Under-sampling intermediate strength values reduces estimator bias

The correlation between predicted and actual 28 d compressive strength values, as illustrated in Figure 3(a), exhibits a distinct bias: low actual compressive strength values are consistently over-predicted, while high values are consistently under-predicted. This suggests that such regression estimators, including ensemble models such as extremely random forests, suffer from an imbalance in the input data used to train them, specifically in that a scarcity of very low and very high compressive strength values leads to less accurate predictions in these ranges. This issue has been frequently addressed in the field of ML classification [26] by resampling, that is, omitting or adding data records in the ML training set. Development of this practice for regression estimators is only in its early stages [26], with primary interest so far in its ability to allow for prediction of rare extremal values [27]. In the current case, where more accurate predictions within a narrowly prescribed range are the goal, resampling methods provide a ready means to reduce estimator bias by simply omitting a selection of the input data.



(a) (b)

Figure 4. **(a)** The distribution of measured compressive strength values from the full dataset, with data that was used as input to train ML estimators, predictions to test ML estimators, and excluded data marked in green, blue, and red, respectively. **(b)** Prediction results of an optimized 500-tree extremely random forest regressor trained on an input set subject to under-sampling (as illustrated in part (a)), shown as predicted vs actual strength values with a dashed line of identity provided to guide the eye.

A tentative under-sampling procedure, developed specifically for the dataset under consideration, demonstrates that the input of *fewer data* is preferable when predicting the compressive strength of cement mortars (Figure 4b). The under-sampling in this case was conducted by analyzing the distribution in actual compressive strength values (Figure 4(a)), divided arbitrarily into 1 MPa intervals. About 90 % of the data records have compressive strengths between 34 MPa and 54 MPa. At least 20 data records were available within each 1 MPa interval in that range, but not outside that range. Therefore, 20 data records were randomly selected from each 1 MPa interval within the range of 34 MPa to 54 MPa, and the remainder of the records in that range were used to test prediction accuracy. The input set constructed in this manner consisted of 420 data records, more than enough to optimally train estimators according to Figure 2(b). Moreover, the new restricted training set corrected the bias in 28 d strength predictions, as can be seen by comparing Figure 4(b) with Figure 3(a).

The under-sampling procedure described above provides marginal improvements in previously discussed average error metrics; for example, R^2 for 28 d strength correlations increased from 0.506 to 0.582. However, the error in any specific prediction, as before, is still considerably larger than that achieved by repeated experimental measurements. Nonetheless, this result highlights an important guideline that should be taken into account, both when using existing datasets and when acquiring new data with a broader array of attributes: prediction bias can be reduced when the training set contains data that are more evenly spread over the entire range of possible target values. The same principle might apply to imbalances also in specific attributes, which then reduce estimator performance but are not easily identifiable. This is likely most applicable to cases for which some of the input attributes are known to vary widely, like those of concrete mixture proportions, as opposed to the relatively well-bounded cement compositions considered currently. The potential applications of under-sampling and/or over-sampling across many attributes to improve the performance of ML regression estimators represents a significant area for future research, with particular relevance to cement and concrete-type materials.

4. Summary and conclusions

This study takes another important step toward predictive ML models of concrete properties by including the effects of OPC characteristics on the properties and performance of cement pastes and mortars. ML methods are applied to estimate 3 d, 7 d, and 28 d compressive strength and the time of initial set across numerous ASTM C150 compliant PCs – attributes that are typically measured in a laborious and time-intensive manner using standard test methods. At a minimum, accurate estimation of these properties by ML requires knowledge of the cement fineness and the mass fractions of the oxides of silicon, aluminum, iron, calcium, and sulfur. Additionally, a lower bound of approximately 200 data records for different cements is required to enable this nature of estimations, with estimator performance improving only marginally with provision of more data records, likely due to the relatively narrow range of cement compositions and finenesses that are included. This implies that suitably-trained ML approaches may be used even when limited data are available.

A distinction of the dataset used in this study is that all the attributes and targets were measured following standard test methods that are intended to minimize the variability of measurement conditions. One advantage of this is that it enables the ML estimators to isolate and discover the influences of OPC powder characteristics on engineering performance without the complications of variability among other important parameters such as mixture proportions and curing temperature. In the field, these latter variables are not held constant and can have a decisive influence on concrete performance. However, prior applications of ensemble ML estimators to field concrete performance have demonstrated that realistic mixture proportioning, and production procedures and curing conditions can be accommodated and still yield reasonably accurate estimations of 28 d compressive strength [1]. Therefore, in a limited sense, this effort confirms the ability of ML methods to estimate how OPC powder characteristics affect binder properties, while outlining the limitations, such as the difference between an estimator's average accuracy and its accuracy in making single predictions. Tight tolerance intervals are a major goal in the ongoing effort to develop more comprehensive ML approaches to predicting the field performance of concrete with multicomponent binders. ML approaches are all the more desirable in this context, however, because they can, if provided with suitable and sufficient data, capture the effects of variable environmental conditions and curing practices on concrete properties and performance.

Acknowledgements

The authors acknowledge financial support for this research provided by the NIST Engineering Laboratory's Exploratory Research Program, the National Science Foundation (CAREER:

1253269, CMMI: 1562066), and the Henry Samueli Fellowship. The contents of this paper reflect the views and opinions of the authors, who are responsible for the accuracy of data presented herein. This research was conducted in the Laboratory for the Chemistry of Construction Materials (LC^2) at the University of California, Los Angeles (UCLA) and the Inorganic Materials Group of the Materials and Structural Systems Division of the Engineering Laboratory at NIST. The authors gratefully acknowledge the support that has made these laboratories and their operations possible.

References

[1] B.A. Young, A. Hall, L. Pilon, P. Gupta, G. Sant, Can the compressive strength of concrete be estimated from knowledge of the mixture proportions?: New insights from statistical analysis and machine learning methods, Cement and Concrete Research. 115 (2019) 379–388.

[2] J.W. Bullard, H.M. Jennings, R.A. Livingston, A. Nonat, G.W. Scherer, J.S. Schweitzer, K.L. Scrivener, J.J. Thomas, Mechanisms of cement hydration, Cement and Concrete Research. 41 (2011) 1208–1223.

[3] S. Mindess, J.F. Young, D. Darwin, Concrete, Prentice Hall, 2003. http://www.bcin.ca/Interface/openbcin.cgi?submit=submit&Chinkey=302193.

[4] H.M. Jennings, J.W. Bullard, J.J. Thomas, J.E. Andrade, J.J. Chen, G.W. Scherer, Characterization and modeling of pores and surfaces in cement paste, Journal of Advanced Concrete Technology. 6 (2008) 5–29.

[5] J.J. Thomas, J.J. Biernacki, J.W. Bullard, S. Bishnoi, J.S. Dolado, G.W. Scherer, A. Luttge, Modeling and simulation of cement hydration kinetics and microstructure development, Cement and Concrete Research. 41 (2011) 1257–1278.

[6] I.-C. Yeh, Modeling of strength of high-performance concrete using artificial neural networks, Cement and Concrete Research. 28 (1998) 1797–1808.

[7] J.-S. Chou, C.-K. Chiu, M. Farfoura, I. Al-Taharwa, Optimizing the prediction accuracy of concrete compressive strength based on a comparison of data-mining techniques, Journal of Computing in Civil Engineering. 25 (2010) 242–253.

[8] K.O. Akande, T.O. Owolabi, S. Twaha, S.O. Olatunji, Performance comparison of SVM and ANN in predicting compressive strength of concrete, IOSR Journal of Computer Engineering. 16 (2014) 88–94.

[9] M.F. Zarandi, I.B. Türksen, J. Sobhani, A.A. Ramezanianpour, Fuzzy polynomial neural networks for approximation of the compressive strength of concrete, Applied Soft Computing. 8 (2008) 488–498.

[10] U. Atici, Prediction of the strength of mineral admixture concrete using multivariable regression analysis and an artificial neural network, Expert Systems with Applications. 38 (2011) 9609–9618.

[11] J. Kasperkiewicz, J. Racz, A. Dubrawski, HPC strength prediction using artificial neural network, Journal of Computing in Civil Engineering. 9 (1995) 279–284.

[12] H.-G. Ni, J.-Z. Wang, Prediction of compressive strength of concrete by neural networks, Cement and Concrete Research. 30 (2000) 1245–1250.

[13] A. Öztaş, M. Pala, E. Özbay, E. Kanca, N. Caglar, M.A. Bhatti, Predicting the compressive strength and slump of high strength concrete using neural network, Construction and Building Materials. 20 (2006) 769–775.

[14] M.H. Rafiei, W.H. Khushefati, R. Demirboga, H. Adeli, Supervised Deep Restricted Boltzmann Machine for Estimation of Concrete., ACI Materials Journal. 114 (2017).
[15] I.B. Topcu, M. Sarıdemir, Prediction of compressive strength of concrete containing fly ash using artificial neural networks and fuzzy logic, Computational Materials Science. 41 (2008) 305–311.
[16] Z.Q. John Lu, The elements of statistical learning: data mining, inference, and prediction, Journal of the Royal Statistical Society: Series A (Statistics in Society). 173 (2010) 693–694.
[17] L. Breiman, Bagging predictors, Machine Learning. 24 (1996) 123–140.
[18] P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees, Machine Learning. 63 (2006) 3–42.
[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research. 12 (2011) 2825–2830.
[20] E. Fix, J.L. Hodges Jr, Discriminatory analysis-nonparametric discrimination: consistency properties, California Univ Berkeley, 1951.
[21] T. Cover, Estimation by the nearest neighbor rule, IEEE Transactions on Information Theory. 14 (1968) 50–55.
[22] ASTM International, Annual Book of ASTM Standards, (2012).
[23] J.O. Irwin, On a criterion for the rejection of outlying observations, Biometrika. (1925) 238–250.
[24] G. McLachlan, K.-A. Do, C. Ambroise, Analyzing microarray gene expression data, John Wiley & Sons, 2005.
[25] J. Brownlee, Machine Learning Mastery with Python, Machine Learning Mastery Pty Ltd. (2016) 100–120.
[26] B. Krawczyk, Learning from imbalanced data: open challenges and future directions, Progress in Artificial Intelligence. 5 (2016) 221–232.
[27] L. Torgo, P. Branco, R.P. Ribeiro, B. Pfahringer, Resampling strategies for regression, Expert Systems. 32 (2015) 465–476.