Journal of Soft Computing in Civil Engineering 5-4 (2021) 01-20

Contents lists available at SCCE

Journal of Soft Computing in Civil Engineering

Journal homepage: www.jsoftcivil.com

Machine Learning on Microstructural Chemical Maps to Classify Component Phases in Cement Pastes

Emily Ford[1], Kailasnath Maneparambil[2,3], Narayanan Neithalath[4*]

1. Graduate student, School of Sustainable Engineering and the Built Environment, Arizona State University, Tempe AZ 85287, USA
2. Intel Corporation, Chandler, AZ 85224, USA
3. Adjunct Faculty, Computer Science and Engineering, Arizona State University, Tempe AZ 85287, USA
4. Professor, School of Sustainable Engineering and the Built Environment, Arizona State University, Tempe AZ 85287, USA

Corresponding author: *narayanan.neithalath@asu.edu*

https://doi.org/10.22115/SCCE.2021.302400.1357

ARTICLE INFO

Article history:
Received: 30 August 2021
Revised: 13 September 2021
Accepted: 13 September 2021

Keywords:
Machine learning;
Nanoindentation;
Chemical mapping;
Microstructure;
Cement pastes;
Ultra-high performance concrete.

ABSTRACT

This paper implements machine learning (ML) classification algorithms on microstructural chemical maps to predict the constituent phases. Intensities of chemical species (Ca, Al, Si, etc.), and in some cases the nanomechanical properties measured at the corresponding points, form the input to the ML model, which predicts the phase label (LD or HD C-S-H, clinker etc.) belonging to that location. Artificial neural networks (ANN) and forest ensemble methods are used for classification. Confusion matrices and receiver-operator characteristic (ROC) curves are used to analyze the classification efficiency. It is shown that, for complex microstructures such as those of ultra-high performance (UHP) pastes, the classifier performs well when nanomechanical information augments the chemical intensity data. For simpler systems such as well-hydrated plain cement pastes, the classifier accurately predicts the phase label from the intensities of Ca, Al, and Si alone. The work enables fast-and-efficient phase identification and property forecasting from microstructural chemical maps.

How to cite this article: Ford E, Maneparambil K, Neithalath N. Machine learning on microstructural chemical maps to classify component phases in cement pastes. J Soft Comput Civ Eng 2021;5(4):01–20. https://doi.org/10.22115/scce.2021.302400.1357

1. Introduction

It is well-known that the microstructure of cementitious materials dictates the properties and performance of the material. The microstructure in turn is a function of time, processing techniques, as well as the constituent material properties and their proportions. Cement paste microstructures are generally constituted of solid and pore phases; the influence of porosity on the mechanical properties and durability of concrete has been well-elucidated for many decades. The solid phase generally consists of cement hydration products and unhydrated/unreacted materials, which depend on the water-to-binder ratio (w/b) and the reactivity of the starting materials [1–4]. While in well-hydrated plain ordinary Portland cement (OPC) pastes, C-S-H gel, calcium hydroxide (CH), and unhydrated clinker are invariably the only solid phases present, multi-component blends like ultra-high performance (UHP) cementitious pastes contain different types of C-S-H based on their density (e.g., low-density or LD, high-density or HD), ultra-high stiffness phases, mixed reaction products, and unreacted particles of cement, fly ash, and limestone [5–8]. Thus, the microstructural complexity increases with the use of multiple-blend binders, requiring more sophisticated and refined methods for microstructural characterization and analysis.

Typically, scanning electron microscopy (SEM) coupled with energy-dispersive X-ray spectroscopy (EDS) is used to extract the chemical information of the microstructure in cement-based materials [9–12]. Grid nanoindentation on these microstructures provide the nanomechanical properties (or more accurately, micromechanical properties since the region of influence of the indents is of the order of 1-3 μm) [13–15] Coupling nanoindentation data (i.e., modulus and/or hardness of the indented locations) with SEM-EDS-based microstructural chemical mapping (i.e., intensity of species such as Ca, Si, and Al) has been shown to provide much needed microscale chemistry-property relationships for cement-based materials [5,16,17]. Clustering algorithms such as k-means clustering or those based on Bayesian methods have been used in conjunction with nanoindentation and chemical maps of cement pastes [18–20]. The microscale properties thus obtained are upscaled using analytical or numerical tools to predict the bulk properties of the material such as elastic modulus, which are important in design [19,21,22].

Grid nanoindentation and chemical mapping produce large datasets, which when judiciously combined with machine learning (ML), enable the development of unbiased structure-property estimators. The use of ML to relate the properties of cement-based materials to the mixture proportions [23–27], or to a limited extent, to their constitutive phases [20,28] has been reported. A recent work by the authors demonstrated the use of ML to predict the nanoindentation modulus of different phases in UHP cementitious pastes using the intensity of chemical species at indentation locations as inputs [29]. It was shown that the efficiency of predicting the modulus suffers when the microstructure becomes more complex. In addition, acquisition of nanoindentation data can be time-and-cost-prohibitive. Thus, a ML-based classification approach is adopted in this work. If ML models can be trained on elemental maps from SEM-EDS and corresponding nanoindentation data, to classify locations in a SEM image as belonging to the appropriate microstructural phase (e.g., LD or HD C-S-H, unhydrated clinker, etc.), it facilitates

real-time characterization. In this paper, the focus is on using SEM-EDS information (with or without nanoindentation data) to identify the constitutive phases as labeled from clustering analysis of nanoindentation data and chemical intensities. This allows for very quick first-order determinations of the effective material properties. Artificial Neural Networks (ANN) and hierarchical decision trees are the ML approaches adopted in this study. The classification models are implemented on two UHP cement pastes, whose properties have been extensively reported [30,31], and validated on two other cement pastes whose characteristics are adopted from the literature [32,33].

2. Data and organization

2.1. UHP cement pastes

Nanoindentation and SEM-EDS chemical data utilized in this study belong to two UHP cementitious pastes (referred to as UHP-1 and UHP-2 in Table 1) which have been studied in detail by the authors [15,16,30,31]. As mentioned earlier, this dataset has been used in predicting the indentation modulus from chemical species intensities using ML [29]. Both UHPs contain multiple cement replacement materials (Class F fly ash, silica fume, fine limestone powder) of varying sizes and reactivity, and a low water-to-binder ratio (w/b), as shown in Table 1. Further details on chemical characteristics of the raw materials, mixture proportions, and mixing and curing conditions can be found in [15,16]. The paste mixtures were cured in moist conditions until their testing duration.

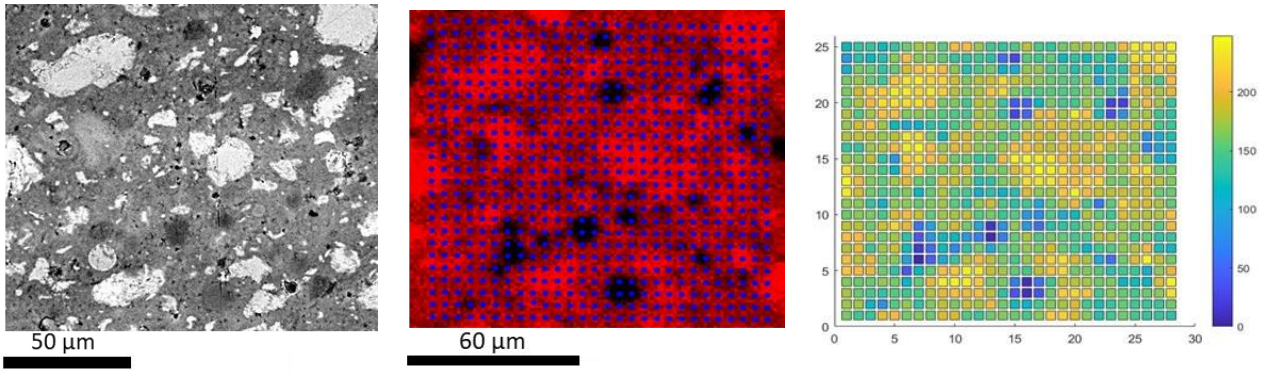
Table 1
Proportions (mass-based) of the UHP cementitious pastes employed in this study.

Mixture	Constituent mass fraction in the binder				w/b	Curing regime
	OPC	Fly ash	Silica fume	Limestone		
UHP-1	0.70	0.175	0.075	0.05	0.20	Moist curing, 30d, 90d
UHP-2	0.50	--	0.20	0.30	0.20	Moist curing, 30d

2.2. Nanoindentation and chemical mapping

A brief description of the procedure for nanoindentation and chemical mapping of UHP-1 and UHP-2 pastes is described here. The sample preparation included specimen cutting, ultrasonication in isopropyl alcohol (IPA) [34], and polishing [35–37]. Nanoindentation was carried out using an Ultra Nanoindentation Tester (UNHT^3; Anton Paar). Each sample had at least 1250 indents split among several grids in different locations to capture the heterogeneity in the microstructure of multi-component UHP paste systems. Indentations were performed in force control mode with a maximum displacement cutoff of 250 nm (0.25 μm) with loading profile detailed in [15,16]. This depth corresponded to an interaction volume idealized as a hemisphere with a radius 3 to 5 times the maximum cutoff [5,38,39]. The hardness (H) and the effective Young's Modulus (M) were determined following the Oliver and Pharr method [40,41].

The specimen surfaces were imaged after the nanoindentation tests using a SEM (SNE-4500M Plus) coupled with EDS (Bruker EDS with ESPRIT software). The application of SEM-EDS for compositional identification of cement hydration phases is discussed in [12,42]. Back-scattered electron (BSE) mode imaging (Fig. 1(a)) was performed with a beam current of 110 μA, a working distance of ~10 mm, and an accelerating voltage of 15 keV[16]. A BSE image was taken over the grid before EDS was performed at 50 kcps. It has been shown that, in cementitious materials, most of the characteristic X-rays escaping the material are generated within a depth of 2 μm [5,32], which is in line with the interaction depth for nanoindentation. To relate the elemental EDS information to the nanomechanical data, a MATLAB localization algorithm was implemented to align the optical image of the nanoindentation grid to the EDS chemical maps, as detailed in [16,18,29]. Brightness of the EDS chemical maps was auto-scaled by the data-collection software. Fig. 1(b) shows the Ca EDS map. Al, Si, and Fe maps were similarly obtained. Across different maps, the number of X-ray counts associated with the same brightness value varies, and hence EDS maps are qualitative measures of the concentration of elements in each indentation grid. For statistical analysis, the RGB intensities from the Al, Ca, Fe, and Si EDS maps (denoted as I_{Al}, I_{Ca}, I_{Fe}, and I_{Si} respectively) were matched with the corresponding nanomechanical data. Fig. 1(c) illustrates the translation of EDS map color intensity of Ca to the 0-255 scale. In BSE imaging, the cube of the brightness (γ^3) can be related to the density of the phase [9]. This local density information is also used as an input parameter in the ML models described later.



(a) (b) (c)

Fig. 1. (a) BSE image of the 30-day UHP-1 paste, (b) Ca EDS map with blue dots added to show the location of one of the indentation grids after the alignment procedure, and (c) MATLAB graphic translating EDS map color intensity into 0-255 scale for Ca.

2.3. Statistical cluster analysis from SEM-EDS and nanoindentation data

To generate the labels of the constitutive microstructural phases to train the ML classification models, a Bayesian Information Criterion (BIC) with negative log likelihood method was implemented for statistical deconvolution (clustering) of the chemical intensities and the micromechanical properties [18]. If there exists *n* phases in the microstructure with each phase occupying a volume fraction of ϕ_i (i = 1…n) such that $\sum_{i=1}^{n} \phi_i = 1$, the properties of each phase can be approximated by a Gaussian distribution with a probability density function (PDF) given as:

$$PDF = \sum_{i=1}^{n} \phi_i \psi_i \quad (1)$$

Here, ψ_i is the vector of classification variables of the phase. The classification variables utilized in cluster analysis were indentation modulus M, indentation hardness H, and the normalized intensities of aluminum I_{Al}, calcium I_{Ca}, iron I_{Fe}, and silicon I_{Si}. While the same statistical nanoindentation results can be fit using different number of phases and volume fractions [43], a maximum negative log likelihood estimation was used to find the PDFs that best represented the experimental data:

$$\mathrm{NlogL} = -\max(\log(\prod_{n} \mathrm{PDF}(n_i))) \quad (2)$$

Here, n_i represents the distribution parameters, in the case of a Gaussian distribution the mean and standard deviation, that are iterated to maximize the likelihood function. Then, the BIC was minimized such that:

$$\mathrm{BIC} = 2\,\mathrm{NlogL} + p\log(m) \quad (3)$$

In the above equation, m is the number of indentation points and p is the number of identifying parameters available at each indentation point (in this case six; four chemical intensities and two mechanical properties M and H) [18]. A summary of the constitutive phases identified from this clustering analysis is given in Table 2. They include low density (LD) C-S-H, high density (HD) C-S-H, an ultra-high stiffness (UHS) phase unique to the very low w/b cement pastes such as UHP mixtures, a mixed phase comprised of partially reacted starting materials such as fly ash or limestone and products such as carboaluminates, and residual clinker. The salient features of these phases have been elucidated in detail elsewhere [5,16,17,44,45]. As an example for the UHP-1 paste cured for 90 days, the clustering of M and H is shown in Fig. 2(a), while Fig. 2(b) depicts the normalized intensities of Ca at every indentation point and the corresponding M, and Fig. 2(c) showcases the normalized intensities of Ca vs. Si. Detailed analysis of the UHP paste clusters identified, and justification for their corresponding constitutive phase labels are described in [16].

Table 2
Constitutive phases identified and their volume fractions (ϕ) in the UHP pastes. FA, MS, L, and CA denotes fly ash, microsilica, limestone, and carboaluminates, respectively. The phase labels from 0-4 are the inputs to the ML classification algorithm.

Mixture	Phase	Phase Label	Volume fraction (ϕ)	
			30 d	90 d
UHP-1	LD CSH/Residual MS	0	0.18	-
	HD CSH	1	0.38	0.40
	UHS Phase	2	0.19	0.23
	Mixed (FA, L, MS, CA)	3	0.12	0.17
	Clinker	4	0.13	0.20
UHP-2	UHS Phase/CSH	2	0.42	-
	Mixed (L, MS)	3	0.41	-
	Clinker/Unreacted	4	0.17	-

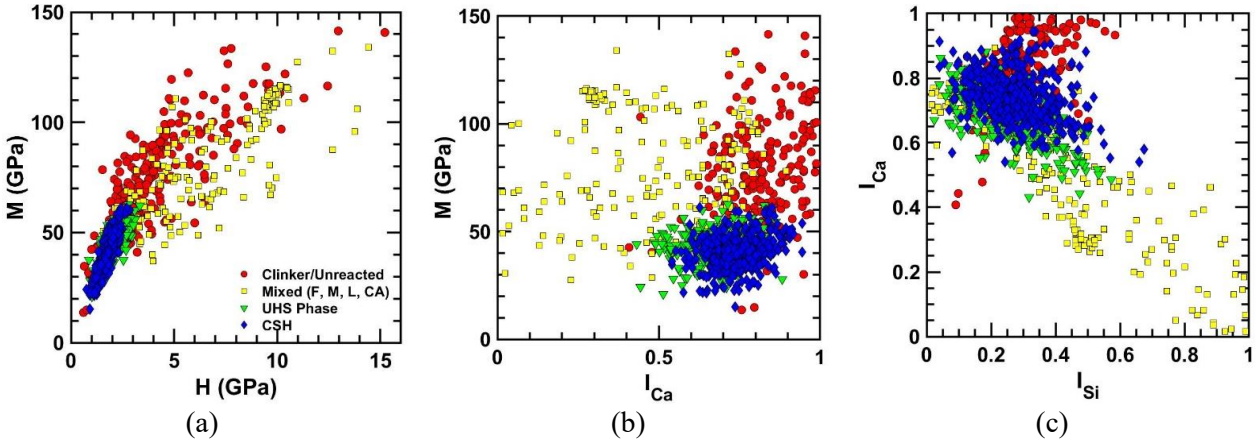


Fig. 2. Clustering analysis of the 90-day cured UHP-1 paste: (a) M vs. H, (b) M vs. I_{Ca}, and (c) I_{Ca} vs. I_{Si}.

2.4. Inputs to the machine learning classification model and the rationale

The ML classification models described in the forthcoming section uses the intensities at different indentation points to determine which of the phases (shown in Table 2), the point belongs to. The datasets for both mixtures and ages shown in Table 1 were combined to create the most generalizable ML classifier possible. The details of this large dataset are shown in Table 3. The ability of ML algorithms to accurately classify the constitutive phases in complex microstructures belonging to multiple mixtures at different ages is explored. For the first set of ML models, 7 inputs were used (i.e., the 4 chemical intensities, γ^3, and M and H values from nanoindentation). In the second set, the ML models were trained only using 5 inputs (i.e., the 4 chemical intensities and γ^3). M and H are used as inputs in one set of ML models since the actual nanomechanical information is expected to facilitate better learning of the ML models to identify the phases during the training stage. This is shown to be true later in this paper, especially for more complex microstructures such as the UHP pastes. To test the correlation between the predicted phase labels and the 7 inputs, Pearson correlation coefficients (or linear correlation coefficients) [20,25] were determined as shown in Fig. 3. It can be noticed that all the inputs are reasonably correlated to the phase label output. M and H have the greatest correlation with the phase labels, and all the chemical intensities are quite similarly related to the phase label output. The high correlation between the phase label output and M and H means that the efficiency of ML classification models that uses only chemical intensities from SEM-EDS (which is the preferred approach, since this data is easier to obtain than M and H) could suffer, which is evaluated in this paper. Generating ML models with and without nanoindentation data provides quantification of the tradeoff of only including SEM-EDS data as inputs.

Table 3
Details of the input dataset for the ML models, including γ^3, H, M, and RGB intensities of Al, Ca, Fe, and Si.

Dataset (see Table 1 for mixture details)	No. of data points	Statistic	I_{Al}	I_{Ca}	I_{Fe}	I_{Si}	γ^3	H (GPa)	M (GPa)
Combined dataset belonging to UHP-1 @ 30d, 90d and UHP-2 @ 30d	3476	Max	252	252	252	252	1.47×10^7	23.20	235.54
		Mean	49	159	81	57	2.18×10^6	3.45	56.92
		Min	4	4	4	4	6.85×10^3	0.43	12.87

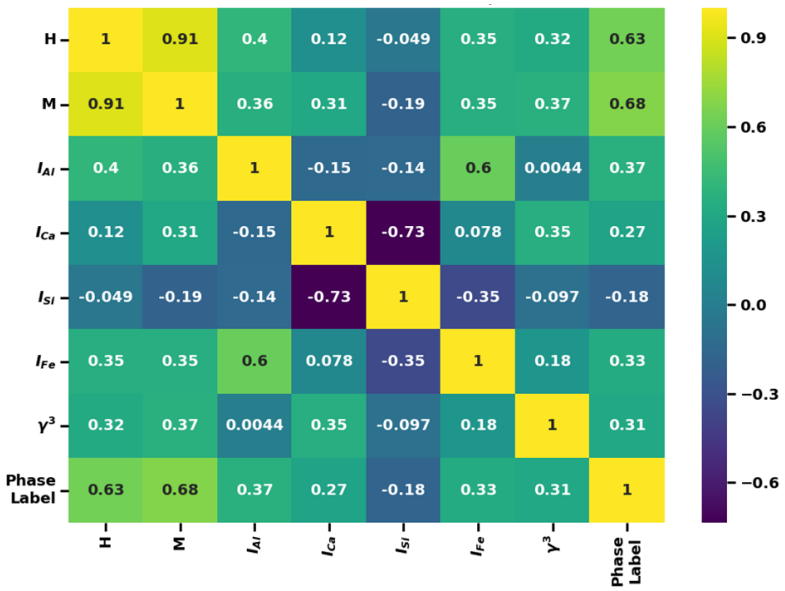


Fig. 3. Pearson coefficient heat map for the correlation between the 7 inputs and the phase label output.

3. Machine learning and data processing

The different machine learning (ML) techniques used for classification, along with the data pre-processing and parameter optimization methods, are summarized here.

3.1. Machine learning techniques

Artificial neural network (ANN) and forest ensemble methods are the ML algorithms used for the multi-classification (i.e., more than 2 classes or phase labels) reported in this paper. ANNs can learn very complex patterns of data, and thus is a preferred ML algorithm for many materials-related problems [23,26,46–48]. The ANNs used in this study utilize 2 to 3 hidden layers, which are appropriate for the number of unique data records used. The chosen activation function to relate neurons [46] is the rectified linear unit (ReLu) with optimization performed using RMSprop, which features an adaptive learning rate formula [49]. Backpropagation, using the gradient of the previous iteration to train the weights of the ANN, was performed automatically by the Keras neural network framework written in Python to build and train the ANNs [50]. To minimize over-fitting, a dropout rate, i.e., the probability that any neuron and its connections will be temporarily excluded from the network, was incorporated into the ANN [51].

Machine learning forest ensemble methods are based on the structure of a decision tree that finds logical splits in the data leading from one branch to the next until ending at the leaf node [23,52] To reduce prediction inaccuracy and over-fitting, the predictions from a collection of decision trees are bagged [23,53], termed ensembles. A basic form of forest ensemble is the Random Forest (RF) method in which the best split of the data is determined by considering all of the input features and checking a criterion, such as Gini impurity, to select the most discriminative threshold [52,54]. Each individual decision tree in the RF ensemble does not use the entire set of training data, but a bootstrap sample made from subsets of the training data with replacement [52,53]. Another forest ensemble is the Extra Trees (ET) regressor in which the splits are drawn at random for each feature and the best split, as measured by the chosen criteria, is selected as the splitting rule [52,54]. In the ET regression model, the entire dataset is incorporated into each individual tree [54]. The prediction results of the individual trees are averaged to produce the output prediction in the RF and ET regressions. In a Gradient Boosted Tree (GBT) ensemble, an initial tree is trained with the entire dataset. All subsequent trees in the forest are trained to minimize the residual between the predicted and actual values of the previous tree [23,54,55]. The final prediction is calculated as the weighted sum of the predictions of each tree. For each tree beyond the first, the prediction is multiplied by the learning rate, with typical values between 0.01 and 0.1 [23,54]. A specialized form of the GBT is Extreme Gradient Boosted (XGB) tree [55]. XGB performs shrinkage and column subsampling techniques to prevent overfitting between boosted trees and additionally offers scalability through parallel tree boosting (efficient computing regardless of data size) [55].

3.2. Preprocessing and evaluation

The input data points were pre-processed before separation into the testing and training sets to ensure that all the inputs and outputs lie in the range [0, 1] such that:

$$z_{new} = \frac{z - z_{min}}{z_{max} - z_{min}} \tag{4}$$

Here, z_{new} is the value of the variable after transformation, z is the current value of the variable, and z_{min} and z_{max} are the minimum and maximum values respectively, of that variable.

The dataset mentioned in Table 3 was shuffled along the rows of indentation points (Fig. 1(b)) such that adjacent points were separated, providing a greater chance of equal distribution of the various microstructural entities within the testing and training datasets. Training was performed by fitting the ML algorithm to the training dataset and allowing the algorithm to adjust its internal features to minimize the error. Model performance was evaluated using the testing dataset, which the ML algorithm has not yet seen, and measuring the resulting errors. To evaluate the accuracy of the ML predictions, a stratified n-fold cross-validation technique was employed [23,25,52] Stratified splitting refers to preserving the percentage of samples in each class within each fold [54]. A 3-fold cross-validation, deemed sufficient for the size of the datasets, was performed using the following steps: (i) randomizing the dataset and performing a 3-fold stratified split, (ii) training the model using 2 of the folds, (iii) testing the model using the remaining fold, (iv) repeating steps (ii) and (iii) until each fold has been used for testing once,

acquiring 3 independent performance measures, and (v) averaging the individual metrics measured to obtain the cross-validation value.

Among the several assessment methods for ML-based classification [56], the area under the Receiver Operator Characteristic curve (ROC-AUC) is chosen here since it is an important metric for checking any classification model's performance [56–58]. A ROC-AUC of 1.0 indicates most accurate classification. The ROC curve is created by plotting the true positive rate (also called sensitivity or recall) against the false positive rate (also called false alarm rate or fallout) at various threshold settings [56,57]. The ROC-AUC is a measure of how well a model can discriminate between two classes (or microstructural phase labels, in this case), and is insensitive to the changes in the class distribution [56,57]. In the case of multi-class labeling, however, ROC-AUC can be calculated using two different methods. The One-versus-Rest (OvR) strategy calculates the model's ability to discriminate between one class vs. the rest of the classes, while the One-versus-One (OvO) strategy pairs each class against another such that, for n phase labels, $\frac{n*(n-1)}{2}$ calculations are made [58]. The former is sensitive to class distribution changes [57] while the latter is insensitive to class distribution, but computationally more expensive when the class number increases. In this study with 5 phase labels and data that is not significantly imbalanced, which requires special class distribution considerations [59], the more general OvR method was employed. The multi-class dataset was one-hot encoded (i.e., represented as binary vectors), and a ML classifier trained to predict the probability that a data point belonged to each phase label. The phase label with the highest probability is taken as the prediction for each point. In training, the goal of the ML models was to maximize the objective function, which was the ROC-AUC. Other metrics tracked, but not used to train the models, were the accuracy and the F1 score, given as [56]:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{5}$$

$$F1 = \frac{2TP}{2TP+FP+FN} \tag{6}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives, predicted by the model for each class (or phase label). True positives indicate the success in identification of the correct phase label.

3.3. Hyperparameter optimization

For all the models, the parameters which maximized the 3-fold cross-validation ROC-AUC were used as the basis for the final models, with some additional fine-tuning. The parameters to optimize in the ANN models were the number of hidden layers, the number of neurons in each hidden layer, and the dropout rate. ReLu activation function with a learning rate of 0.001 and an RMSprop optimization scheme was used. For the RF, ET, and GBT models, the number of trees in the forest, the maximum depth of the trees, the minimum number of samples before splitting, and the minimum number of samples per leaf were tuned. Coarse optimization of the hyperparameters for ANN and the forest ensembles followed a random search pattern, found to be the most efficient method to optimize parameters [60], by randomly generating 20 different

combinations of hyperparameters. The hyperparameters for random testing were chosen from the uniform distributions shown in Table 4.

For the XGB models, there are many hyperparameters available to tune, nine of which were chosen for this study. The hyperparameters range from structure-based, such as the depth of the trees or the number of GBTs, to how splits are made via the subsample and colsample_bytree parameters, or even how big the leaf groups can be via min_child_weight. Additional parameters tuned included the learning rate, the minimum objective function loss required to split a leaf node called gamma, as well as the L1 and L2 regularization terms on the weights called alpha and lambda, respectively. Each hyperparameter was tested one at a time over a grid within the range of values indicated in Table 4, where the best value was used when searching for the next parameter. The order of hyperparameter selection is given by the order of parameters in Table 4 for XGB. This process was continued until the end when several different learning rates and number of trees were tested as a final tuning effort. Detailed breakdown of the allowed ranges and significance of each of these hyperparameters are given in the XGB code documentation [55].

Table 4
Hyperparameters tuned based on a uniform distribution range of potential values.

Model	Hyperparameter	Uniform Distribution Range
ANN	# hidden layers	[1, 4]
	# starting neurons	[20, 75]
	Drop rate	[0, 0.3]
Random Forest (RF), Extra Trees (ET) Forest, and Gradient Boosted Trees (GBT)	# of trees	[50, 400]
	Maximum depth	[3, 21]
	Minimum# of samples before split	[2, 25]
	Minimum # of samples on leaf	[1, 10]
XGB	# of trees	[0, 500]
	Maximum depth	[1, 9]
	min_child_weight	[1, 6]
	Gamma	[0, 0.8]
	Subsample	[0.5, 1.0]
	Colsample_bytree	[0.2, 1.0]
	Alpha	[1E-5, 1]
	Lambda	[1E-5, 1.05]
	Learning Rate	[0.05, 0.3]

4. ML-based classification of cementitious phases

4.1. UHP pastes

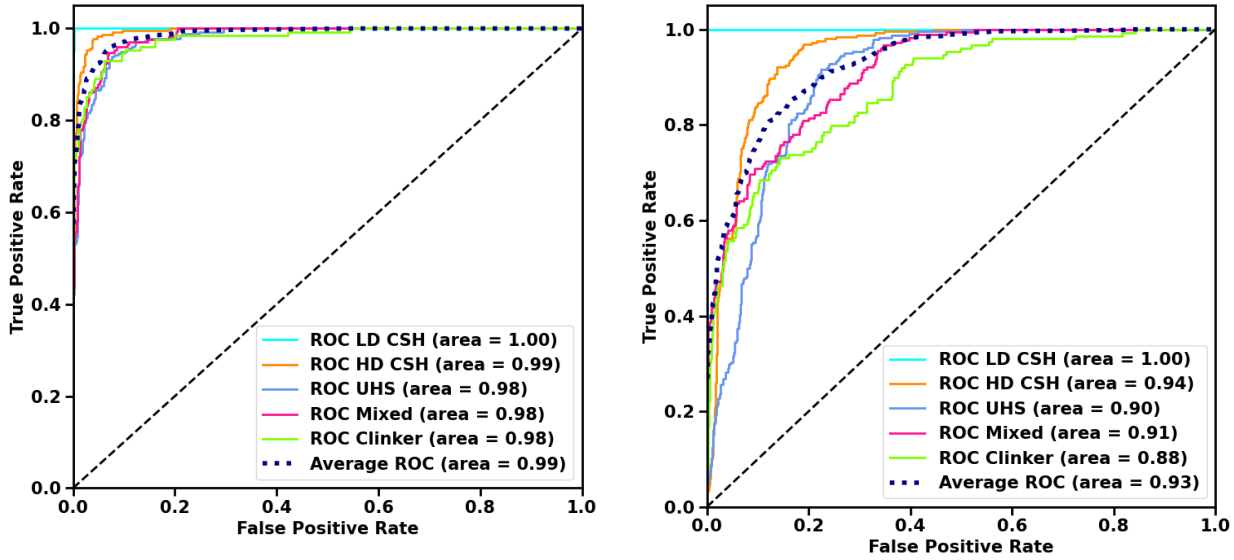
The predictive efficiency of the different ML models using SEM-EDS data with and without nanoindentation hardness (H) and stiffness (M) as inputs, to classify the UHP phase at each desired location is reported in this section. Each of the five ML algorithms (ANN, RF, ET, GBT, XGB) discussed above were implemented on the data to examine the applicability of the ML classification methodology to identify the phase labels in complex and heterogeneous UHP pastes. Table 5 lists the ROC-AUC, accuracy, and F1 values for the final ML classification models for the 7-input and 5-input cases. The bolded entries indicate the ML models where the OvR ROC-AUC results from 3-fold cross-validation were the highest. Note that the 3-fold cross-validation trials could not be plotted directly, instead, Fig. 4 and Fig. 5 were generated from a 75%/25% data split such that 75% of the data points were used for training and 25% were used for testing and displaying the plots, where the results were almost identical to the 3-fold cross-validation results reported in Table 5.

In the case of the datasets with 7 inputs (both SEM-EDS and nanomechanical data), all the ML models performed very well in terms of all three metrics (ROC-AUC, Accuracy, and F1), with the GBT model showing a slightly better performance. The ROC-AUC value was around 0.99 (1.0 being the absolute best) [56,57], indicating the efficiency of the classification algorithms in being able to determine the phase labels based on the given input data. Even when the nanomechanical data was removed from the datasets and the input matrix reduced to 5 SEM-EDS input parameters, the ML classification algorithms worked quite well with a ROC-AUC value of around 0.92. In this case, the ANN model provided the best ROC-AUC value, while the forest models also showed very similar performance. The high ROC-AUC values show that, in an OvR setting, the classification ML algorithms used are successful in distinguishing one class compared to all the others. However, it can be also seen from Table 5 that there is a sharp reduction in the accuracy and F1 scores, which both depend on the number of correctly identified data points as described using Equations 5 and 6 [56], when the nanomechanical information is absent. This is to be expected, since M and H had the highest correlation with the output phase label, as indicated in **Fig. 3**. It is observed that high accuracy and F1 values, along with high ROC-AUC, can be achieved when additional, relevant input data such as M and H are available.

Fig. 4(a) and (b) show the ROC curves obtained from these best-performing models for the 7-input and 5-input cases, respectively. As expected, and shown in Table 5, the ROC curves shift downward when the nanomechanical inputs are excluded from the ML classification analysis. However, it is important to note that not including M and H, which correlated the most with the phase label output (see Fig. 3), still produces reasonable identification of the microstructural phases just based on SEM-EDS information. This is significant in that, the use of SEM-EDS chemical maps along with a ML classification scheme allows for: (i) identification of potential phases present at those locations, which provides detailed insights into the influence of material composition on microstructure, and (ii) prediction of important paste properties (such as modulus) based on the known properties of the phases and their volume fractions.

Table 5
Efficiency metrics of the ML classification algorithms for phases in UHP pastes from SEM-EDS (5 inputs), and with two additional inputs, M and H, from nanoindentation (7 inputs). Average and standard deviation from 3-fold cross-validation is reported. The ML model with the greatest ROC-AUC for each number of inputs is shown in **bold**.

# of inputs	Model Type	ROC-AUC	Accuracy	F1
7	ANN	0.988 ± 0.003	0.906 ± 0.009	0.912 ± 0.009
	Random Forest	0.988 ± 0.003	0.903 ± 0.010	0.911 ± 0.010
	Extra Trees Forest	0.986 ± 0.003	0.889 ± 0.015	0.898 ± 0.014
	Gradient Boosted Trees	**0.989 ± 0.002**	**0.908 ± 0.011**	**0.914 ± 0.012**
	XGB	0.988 ± 0.003	0.907 ± 0.017	0.914 ± 0.016
5	**ANN**	**0.926 ± 0.002**	**0.726 ± 0.013**	**0.745 ± 0.014**
	Random Forest	0.924 ± 0.002	0.728 ± 0.010	0.749 ± 0.012
	Extra Trees Forest	0.924 ± 0.003	0.715 ± 0.012	0.729 ± 0.013
	Gradient Boosted Trees	0.919 ± 0.004	0.719 ± 0.015	0.746 ± 0.016
	XGB	0.921 ± 0.003	0.721 ± 0.008	0.743 ± 0.010



(a) (b)

Fig. 4. Receiver-Operator Curves (ROC) showing One-versus-Rest results for ML classification using 25% of data for testing: (a) GBT ML model with 7 inputs, (b) ANN ML model with 5 inputs. The dashed diagonal line represents the random guess of a class. The chosen models are the best performing ones based on Table 5.

Further information on the predictive performance of the classification models can be gleaned from confusion matrices presented in Fig. 5(a) and (b) for the 7-input GBT ML and 5-input ANN ML classification models, respectively. For both the input types the LD C-S-H phase is accurately identified in 94-100% of the points by the ML models as shown in Fig. 5. Similarly, the HD C-S-H phase is correctly classified in 83-94% of the points, depending on whether the 5-input or 7-input models are used. Since LD C-S-H and HD C-S-H have differences in their packing densities, which result in different mechanical properties [61,62], it is only natural that a ML model that is trained using nanomechanical data also shows near-perfect capability in accurately identifying these phases. However, cluster analysis in several past work [14,17,63]

have shown dissimilarities in chemical intensities between these phases, which enables the 5-input model also to perform satisfactorily in classifying these phases. As indicated in the authors' recent work [15,16], the remaining three hard-stiff phases, viz., UHS, mixed phase containing limestone, carboaluminates and fly ash, and clinker, with indentation moduli of ~43 GPa [61], ~75 GPa [21,44,64,65], and ~100 GPa [47] respectively, overlap in terms of chemical intensities and stiffnesses. This is clearly noticed in the scatter of points corresponding to these phases in Fig. 2(c). Reducing the number of inputs from 7 to 5 clearly has a significant adverse effect on the classification of these phases as noted from Fig. 5. In the 7-input model, the mixed phase is correctly identified in ~92% of the cases, while the classification accuracy drops down to ~63% in the 5-input model, where the mixed phase is confused with the UHS phase in many instances. In both the models, clinker classification has the lowest accuracy. In the 5-input model, the clinker classification accuracy is around 50%, with a significant number of clinker locations mis-identified as HD C-S-H phases due to the absence of corroborating nanoindentation data. It is also notable that the EDS chemical maps were obtained based on qualitative measurements and not on quantitative spot chemical analysis [18,32], and therefore only provide relative atomic ratios and not the exact ratios. As such, it is likely that cementitious phases with similar Ca/Si ratios, but different stiffnesses may be confused for one another in the 5-input ML model. Another explanation for the confusion between the clinker and HD C-S-H phases is that, in the UHP-2 mix there was no HD C-S-H cluster identified, and the reaction product belonged to the UHS phase [16] however, when the Ca and Si intensities were plotted for the clinker and UHS phases, they almost perfectly overlapped [16]. The high unreacted limestone content in this mixture could have resulted in excess Ca in the chemical map that contributed to a higher Ca/Si ratio, which is typical of clinker. This may have led to the confusion of the ML to differentiate between the clinker and UHS/HD C-S-H phases for the UHP-2 mixture. It is once again shown that, in complex microstructures where chemical intensities overlap between phases (as shown in Fig. 2(c)), the use of additional inputs in the form of nanomechanical properties help classification significantly.

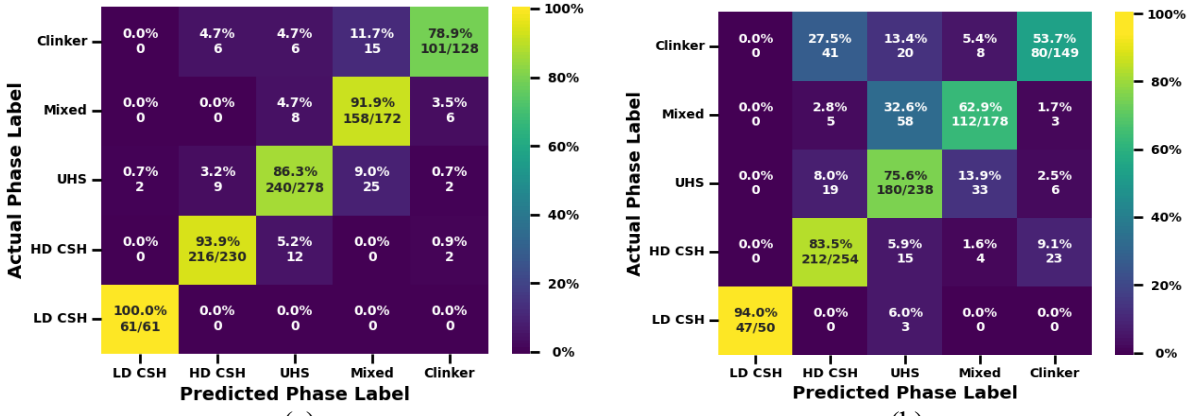


Fig. 5. Confusion matrices showing results for ML classification using 25% of data for testing: (a) GBT ML model with 7 inputs, and (b) ANN ML model with 5 inputs. Percentage accuracy in each row is given based on the total number of data points in each phase label, as shown along the diagonal. In an ideal case, it is desirable to have a classification accuracy near 100% on the boxes along the diagonal, which would result in little to no misidentification, and thus, close to 0% on all the other boxes.

4.2. Validation of the classification approach using other cement paste data

To validate the ML classification of microstructural phases through chemical intensities from SEM-EDS, two new datasets were curated from literature [32,33] and similar ML models developed to classify their phases. These datasets are referred to as OPC (plain cement paste) [32] and NP (20% of cement by mass replaced with a natural pozzolan) [33]. Nanoindentation and chemical mapping data reported in [32,33] identified several clusters of microstructural phases in these mixtures. The OPC data identified 5 clusters by BIC and negative log likelihood method in [32], however two clusters with the highest stiffness and hardness could be grouped together as part of the clinker phase to ensure that the same ML algorithms as described above can be used here. The remaining three clusters were labeled as LD C-S-H, HD C-S-H, and a mixed phase. For the NP data, 6 clusters were identified in [33], but clusters 5 and 6 were grouped as together as they were both identified as clinker phases [33], with LD C-S-H, HD C-S-H, UHS, and mixed phase labels given to the remaining clusters. The only available inputs were three elemental intensities, I_{Ca}, I_{Si}, and I_{Al}, along with M and H. Thus, ML models using all the 5 inputs, or just the 3 chemical signature inputs, were implemented. To keep the discussions succinct, only three forest ensemble models (RF, ET, and GBT; which generally are the faster ML models) are used here for the validation tests. Table 6 lists the resulting ROC-AUC, accuracy, and F1 values for these datasets. Similar to the UHP pastes, there is a decrease across all metrics of classification going from 5 inputs (which included the micromechanical M and H) to 3 inputs. However, this decrease is much lower, owing to the greatly reduced complexity in these microstructures that were well hydrated. As compared to the UHP pastes evaluated in the previous section, these pastes demonstrate reduced heterogeneity with fewer starting ingredients, proportioned using a higher w/b, and having undergone higher degrees of reaction, which influences the predictive accuracy as detailed in [29]. Based on the results in Table 6, the chosen ML classification methods can be considered to be successful in identifying the constituent phases, given only the chemical intensities, for less complex microstructures.

Table 6
Efficiency metrics of the ML classification algorithm for phases in OPC and NP pastes from SEM-EDS (3 inputs), and with two additional inputs, M and H, from nanoindentation (5 inputs). Average and standard deviation from 3-fold cross-validation is reported. The most accurate ML model for each number of inputs is shown in **bold**.

Dataset	# of inputs	Model Type	ROC-AUC	Accuracy	F1
OPC	5	Random Forest	0.975 ± 0.010	0.888 ± 0.018	0.893 ± 0.018
		Extra Trees Forest	**0.981 ± 0.005**	**0.897 ± 0.011**	**0.899 ± 0.011**
		Gradient Boosted Forest	0.968 ± 0.011	0.858 ± 0.043	0.867 ± 0.037
	3	Random Forest	0.951 ± 0.011	0.808 ± 0.062	0.801 ± 0.064
		Extra Trees Forest	**0.958 ± 0.012**	**0.829 ± 0.040**	**0.837 ± 0.037**
		Gradient Boosted Forest	0.945 ± 0.017	0.817 ± 0.040	0.827 ± 0.031
NP	5	Random Forest	0.988 ± 0.006	0.891 ± 0.023	0.891 ± 0.019
		Extra Trees Forest	0.989 ± 0.007	0.902 ± 0.038	0.897 ± 0.039
		Gradient Boosted Forest	**0.991 ± 0.006**	**0.925 ± 0.026**	**0.925 ± 0.022**
	3	**Random Forest**	**0.973 ± 0.008**	**0.860 ± 0.019**	**0.859 ± 0.022**
		Extra Trees Forest	0.973 ± 0.006	0.832 ± 0.041	0.826 ± 0.039
		Gradient Boosted Forest	0.965 ± 0.006	0.822 ± 0.028	0.819 ± 0.032

Fig. 6 shows the confusion matrices and ROC curves for the OPC and NP mixtures, for the 3-input cases. The classification accuracy is very high as noted from the confusion matrices for both the pastes, attributable to the relative simplicity of their microstructures as compared to the UHP pastes. There are very few mis-labeled indentation points even when the nanomechanical data is not provided. The results show the application of ML-based classification algorithms in labeling the microstructural phases in cementitious systems.

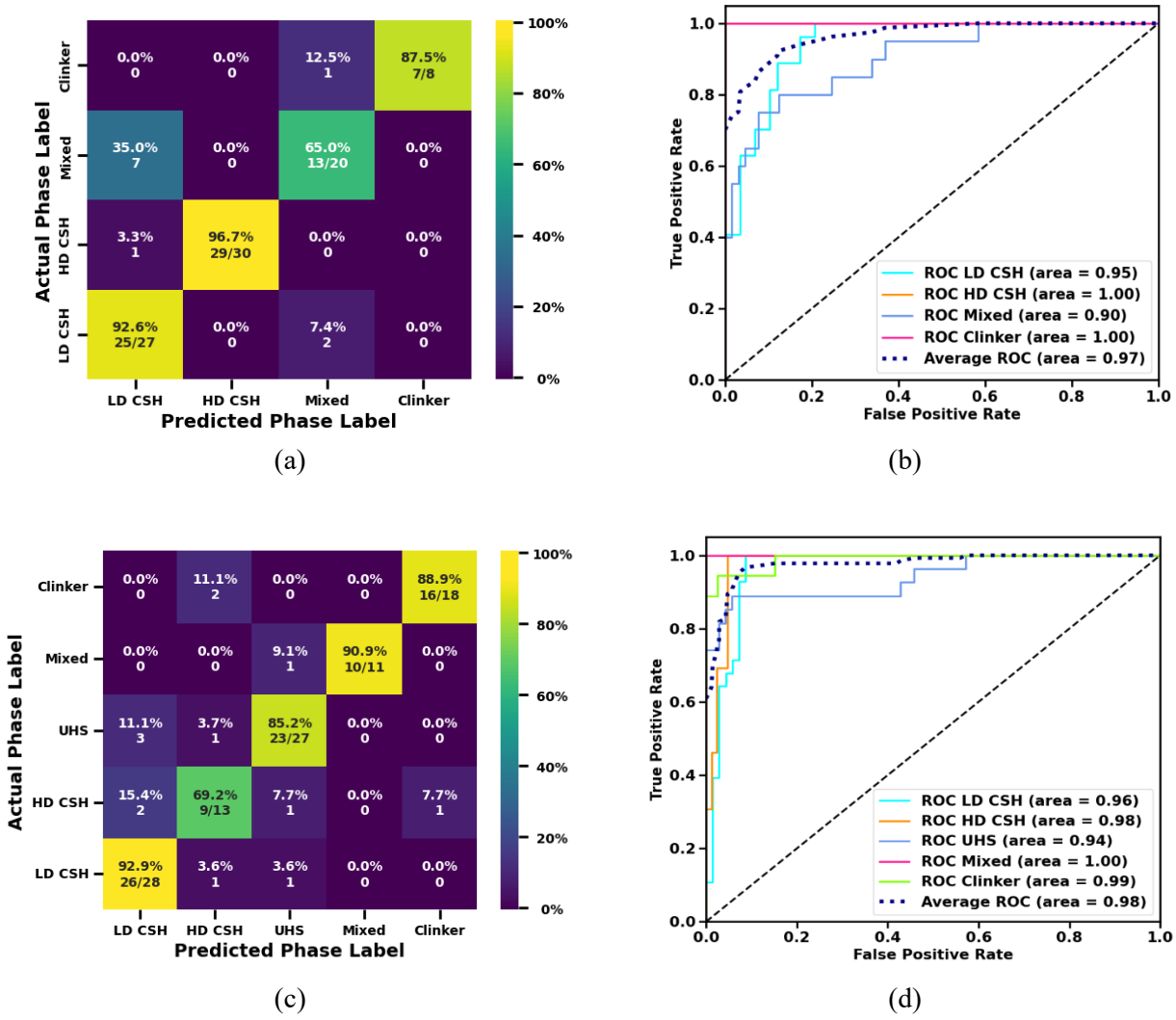


Fig. 6. ML classification using 25% of data for testing: (a) and (b) confusion matrix and ROC curves for the 3-input ET model for the OPC paste; (c) and (d) confusion matrix and ROC curves for the 3-input RF model for the NP paste.

5. Summary and Conclusions

This study has presented a novel approach to accurately predict cement hydration phases from chemical intensity maps, using ML methods. Chemical intensity data from SEM-EDS for different UHP cement paste datasets representing multiple cementing materials and hydration

ages were combined. Micromechanical information from nanoindentation as well the elemental intensities from qualitative EDS maps were then coupled with Bayesian statistical clustering. With the phase labels (e.g., LD or HD C-S-H, clinker etc.) thus identified, different ML classification techniques based on Artificial Neural Networks (ANN) and forest ensemble methods were implemented on the dataset. The area under the Receiver Operator Characteristic curve (ROC-AUC) was chosen as the indicator of model performance.

It was observed that, for the combined dataset of the UHP pastes, the removal of nanoindentation information from the datasets did impact the efficiency of classification. Confusion matrices demonstrated that the removal of nanoindentation information resulted in misidentification of some of the microstructural labels, especially where the chemical intensity data overlapped between multiple phases due to the unique composition of the UHP pastes. It was shown that, in such complex systems, the use of additional inputs in the form of nanomechanical properties help classification significantly. The same approach was also used on two less complex microstructures (i.e., fewer starting materials and more complete hydration), one of a plain OPC paste and the other a paste with 20% OPC replaced using a highly reactive natural pozzolan. Here, normalized intensities of just the three chemical species (Ca, Si, and Al) were deemed sufficient (without nanoindentation data) to generate a highly accurate classifier. It is shown that chemical intensity mapping of microstructures, coupled with machine learning, can be used to accurately (in the case of common cementitious microstructures) classify the microstructural phases, which can lead to *apriori* property (e.g., stiffness) predictions. ML models can thus classify the cementitious component phase at locations in a microstructure to facilitate real-time characterization and first-order estimation of bulk properties.

Acknowledgements

This study was partly supported by U.S. National Science Foundation (CMMI: 1727445). EF acknowledges the Dean's Fellowship from the Ira A. Fulton Schools of Engineering at Arizona State University.

References

[1] Suda Y, Saeki T, Saito T. Relation between chemical composition and physical properties of CSH generated from cementitious materials. J Adv Concr Technol 2015;13:275–90.

[2] Manzano H, Dolado JS, Ayuela A. Elastic properties of the main species present in Portland cement pastes. Acta Mater 2009;57:1666–74. https://doi.org/10.1016/j.actamat.2008.12.007.

[3] Wang D, Shi C, Farzadnia N, Shi Z, Jia H. A review on effects of limestone powder on the properties of concrete. Constr Build Mater 2018;192:153–66. https://doi.org/10.1016/j.conbuildmat.2018.10.119.

[4] Mondal P, Shah SP, Marks LD, Gaitero JJ. Comparative Study of the Effects of Microsilica and Nanosilica in Concrete. Transp Res Rec J Transp Res Board 2010;2141:6–9. https://doi.org/10.3141/2141-02.

[5] Chen JJ, Sorelli L, Vandamme M, Ulm F-J, Chanvillard G. A Coupled Nanoindentation/SEM-EDS Study on Low Water/Cement Ratio Portland Cement Paste: Evidence for C-S-H/Ca(OH) 2 Nanocomposites. J Am Ceram Soc 2010. https://doi.org/10.1111/j.1551-2916.2009.03599.x.

[6] Vance K, Aguayo M, Oey T, Sant G, Neithalath N. Hydration and strength development in ternary portland cement blends containing limestone and fly ash or metakaolin. Cem Concr Compos 2013;39:93–103. https://doi.org/10.1016/j.cemconcomp.2013.03.028.

[7] Yu R, Spiesz P, Brouwers HJH. Effect of nano-silica on the hydration and microstructure development of Ultra-High Performance Concrete (UHPC) with a low binder amount. Constr Build Mater 2014;65:140–50. https://doi.org/10.1016/j.conbuildmat.2014.04.063.

[8] Huang W, Kazemi-Kamyab H, Sun W, Scrivener K. Effect of cement substitution by limestone on the hydration and microstructural development of ultra-high performance concrete (UHPC). Cem Concr Compos 2017;77:86–101. https://doi.org/10.1016/j.cemconcomp.2016.12.009.

[9] Stutzman P. Scanning electron microscopy imaging of hydraulic cement microstructure. Cem Concr Compos 2004;26:957–66. https://doi.org/10.1016/j.cemconcomp.2004.02.043.

[10] Durdziński PT, Dunant CF, Haha M Ben, Scrivener KL. A new quantification method based on SEM-EDS to assess fly ash composition and study the reaction of its individual components in hydrating cement paste. Cem Concr Res 2015;73:111–22. https://doi.org/10.1016/j.cemconres.2015.02.008.

[11] Rößler C, Steiniger F, Ludwig H-M. Characterization of C-S-H and C-A-S-H phases by electron microscopy imaging, diffraction, and energy dispersive X-ray spectroscopy. J Am Ceram Soc 2017;100:1733–42. https://doi.org/10.1111/jace.14729.

[12] Rossen JE, Scrivener KL. Optimization of SEM-EDS to determine the C–A–S–H composition in matured cement paste samples. Mater Charact 2017;123:294–306. https://doi.org/10.1016/j.matchar.2016.11.041.

[13] Němeček J, Šmilauer V, Kopecký L. Nanoindentation characteristics of alkali-activated aluminosilicate materials. Cem Concr Compos 2011;33:163–70. https://doi.org/10.1016/j.cemconcomp.2010.10.005.

[14] Hu C, Li Z. Property investigation of individual phases in cementitious composites containing silica fume and fly ash. Cem Concr Compos 2015;57:17–26. https://doi.org/10.1016/j.cemconcomp.2014.11.011.

[15] Ford E, Arora A, Mobasher B, Hoover CG, Neithalath N. Elucidating the nano-mechanical behavior of multi-component binders for ultra-high performance concrete. Constr Build Mater 2020;243:118214. https://doi.org/10.1016/j.conbuildmat.2020.118214.

[16] Ford EL, Hoover CG, Mobasher B, Neithalath N. Relating the nano-mechanical response and qualitative chemical maps of multi-component ultra-high performance cementitious binders. Constr Build Mater 2020;260:119959. https://doi.org/10.1016/j.conbuildmat.2020.119959.

[17] Wilson W, Sorelli L, Tagnit-Hamou A. Unveiling micro-chemo-mechanical properties of C–(A)–S–H and other phases in blended-cement pastes. Cem Concr Res 2018;107:317–36. https://doi.org/10.1016/j.cemconres.2018.02.010.

[18] J. Krakowiak K, Wilson W, James S, Musso S, Ulm F-J. Inference of the phase-to-mechanical property link via coupled X-ray spectrometry and indentation analysis: Application to cement-based materials. Cem Concr Res 2015;67:271–85. https://doi.org/10.1016/j.cemconres.2014.09.001.

[19] Hu C, Li Z. A review on the mechanical properties of cement-based materials measured by nanoindentation. Constr Build Mater 2015;90:80–90. https://doi.org/10.1016/j.conbuildmat.2015.05.008.

[20] Konstantopoulos G, Koumoulos EP, Charitidis CA. Testing Novel Portland Cement Formulations with Carbon Nanotubes and Intrinsic Properties Revelation: Nanoindentation Analysis with Machine Learning on Microstructure Identification. Nanomaterials 2020;10:645. https://doi.org/10.3390/nano10040645.

[21] Gao X, Wei Y, Huang W. Effect of individual phases on multiscale modeling mechanical properties of hardened cement paste. Constr Build Mater 2017;153:25–35. https://doi.org/10.1016/j.conbuildmat.2017.07.074.

[22] Brown L, Allison PG, Sanchez F. Use of nanoindentation phase characterization and homogenization to estimate the elastic modulus of heterogeneously decalcified cement pastes. Mater Des 2018;142:308–18. https://doi.org/10.1016/j.matdes.2018.01.030.

[23] Young BA, Hall A, Pilon L, Gupta P, Sant G. Can the compressive strength of concrete be estimated from knowledge of the mixture proportions?: New insights from statistical analysis and machine learning methods. Cem Concr Res 2019;115:379–88. https://doi.org/10.1016/j.cemconres.2018.09.006.

[24] DeRousseau MA, Laftchiev E, Kasprzyk JR, Rajagopalan B, Srubar WV. A comparison of machine learning methods for predicting the compressive strength of field-placed concrete. Constr Build Mater 2019;228:116661. https://doi.org/10.1016/j.conbuildmat.2019.08.042.

[25] Chou J-S, Chiu C-K, Farfoura M, Al-Taharwa I. Optimizing the Prediction Accuracy of Concrete Compressive Strength Based on a Comparison of Data-Mining Techniques. J Comput Civ Eng 2011;25:242–53. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000088.

[26] Dao D Van, Adeli H, Ly H-B, Le LM, Le VM, Le T-T, et al. A Sensitivity and Robustness Analysis of GPR and ANN for High-Performance Concrete Compressive Strength Prediction Using a Monte Carlo Simulation. Sustainability 2020;12:830. https://doi.org/10.3390/su12030830.

[27] Nguyen KT, Nguyen QD, Le TA, Shin J, Lee K. Analyzing the compressive strength of green fly ash based geopolymer concrete using experiment and machine learning approaches. Constr Build Mater 2020;247:118581. https://doi.org/10.1016/j.conbuildmat.2020.118581.

[28] Koumoulos EP, Paraskevoudis K, Charitidis CA. Constituents Phase Reconstruction through Applied Machine Learning in Nanoindentation Mapping Data of Mortar Surface. J Compos Sci 2019;3:63. https://doi.org/10.3390/jcs3030063.

[29] Ford E, Kailas S, Maneparambil K, Neithalath N. Machine learning approaches to predict the micromechanical properties of cementitious hydration phases from microstructural chemical maps. Constr Build Mater 2020;265:120647. https://doi.org/10.1016/j.conbuildmat.2020.120647.

[30] Arora A, Aguayo M, Hansen H, Castro C, Federspiel E, Mobasher B, et al. Microstructural packing- and rheology-based binder selection and characterization for Ultra-high Performance Concrete (UHPC). Cem Concr Res 2018;103:179–90. https://doi.org/10.1016/j.cemconres.2017.10.013.

[31] Arora A, Almujaddidi A, Kianmofrad F, Mobasher B, Neithalath N. Material design of economical ultra-high performance concrete (UHPC) and evaluation of their properties. Cem Concr Compos 2019;104:103346. https://doi.org/10.1016/j.cemconcomp.2019.103346.

[32] Wilson W, Sorelli L, Tagnit-Hamou A. Automated coupling of NanoIndentation and Quantitative Energy-Dispersive Spectroscopy (NI-QEDS): A comprehensive method to disclose the micro-

chemo-mechanical properties of cement pastes. Cem Concr Res 2018;103:49–65. https://doi.org/10.1016/j.cemconres.2017.08.016.

[33] Wilson W, Rivera-Torres JM, Sorelli L, Durán-Herrera A, Tagnit-Hamou A. The micromechanical signature of high-volume natural pozzolan concrete by combined statistical nanoindentation and SEM-EDS analyses. Cem Concr Res 2017;91:1–12. https://doi.org/10.1016/j.cemconres.2016.10.004.

[34] Hoover CG, Ulm F-J. Experimental chemo-mechanics of early-age fracture properties of cement paste. Cem Concr Res 2015;75:42–52. https://doi.org/10.1016/j.cemconres.2015.04.004.

[35] Sorelli L, Constantinides G, Ulm F-J, Toutlemonde F. The nano-mechanical signature of Ultra High Performance Concrete by statistical nanoindentation techniques. Cem Concr Res 2008;38:1447–56. https://doi.org/10.1016/j.cemconres.2008.09.002.

[36] Stutzman PE, Clifton JR. Specimen preparation for scanning electron microscopy. Proc. Int. Conf. Cem. Microsc., vol. 21, International Cement Microscopy Association; 1999, p. 10–22.

[37] Stutzman PE. Microscopy of Clinker and Hydraulic Cements. Rev Mineral Geochemistry 2012;74:101–46. https://doi.org/10.2138/rmg.2012.74.3.

[38] Ulm F-J, Vandamme M, Jennings HM, Vanzo J, Bentivegna M, Krakowiak KJ, et al. Does microstructure matter for statistical nanoindentation techniques? Cem Concr Compos 2010;32:92–9. https://doi.org/10.1016/j.cemconcomp.2009.08.007.

[39] da Silva WRL, Němeček J, Štemberk P. Methodology for nanoindentation-assisted prediction of macroscale elastic properties of high performance cementitious composites. Cem Concr Compos 2014;45:57–68. https://doi.org/10.1016/j.cemconcomp.2013.09.013.

[40] Oliver WC, Pharr GM. An improved technique for determining hardness and elastic modulus using load and displacement sensing indentation experiments. J Mater Res 1992;7:1564–83. https://doi.org/10.1557/JMR.1992.1564.

[41] Oliver WC, Pharr GM. Measurement of hardness and elastic modulus by instrumented indentation: Advances in understanding and refinements to methodology. J Mater Res 2004;19:3–20. https://doi.org/10.1557/jmr.2004.19.1.3.

[42] Stutzman PE, Bullard JF, Feng P, Stutzman PE. Quantitative imaging of clinker and cement microstructure. US Department of Commerce, National Institute of Standards and Technology; 2016.

[43] Lura P, Trtik P, Münch B. Validity of recent approaches for statistical nanoindentation of cement pastes. Cem Concr Compos 2011;33:457–65. https://doi.org/10.1016/j.cemconcomp.2011.01.006.

[44] Haecker C-J, Garboczi EJ, Bullard JW, Bohn RB, Sun Z, Shah SP, et al. Modeling the linear elastic properties of Portland cement paste. Cem Concr Res 2005;35:1948–60. https://doi.org/10.1016/j.cemconres.2005.05.001.

[45] Puerta-Falla G, Balonis M, Le Saout G, Falzone G, Zhang C, Neithalath N, et al. Elucidating the Role of the Aluminous Source on Limestone Reactivity in Cementitious Materials. J Am Ceram Soc 2015;98:4076–89. https://doi.org/10.1111/jace.13806.

[46] Li X, Liu Z, Cui S, Luo C, Li C, Zhuang Z. Predicting the effective mechanical property of heterogeneous materials by image based modeling and deep learning. Comput Methods Appl Mech Eng 2019;347:735–53. https://doi.org/10.1016/j.cma.2019.01.005.

[47] Zhang J, Huang Y, Wang Y, Ma G. Multi-objective optimization of concrete mixture proportions using machine learning and metaheuristic algorithms. Constr Build Mater 2020;253:119208. https://doi.org/10.1016/j.conbuildmat.2020.119208.

[48] Konstantopoulos G, Koumoulos EP, Charitidis CA. Classification of mechanism of reinforcement in the fiber-matrix interface: Application of Machine Learning on nanoindentation data. Mater Des 2020;192:108705. https://doi.org/10.1016/j.matdes.2020.108705.

[49] Tieleman T, Hinton G. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA Neural Networks Mach Learn 2012;4:26–31.

[50] Chollet F. Keras, GitHub repository, 2015. [Online]. Available: https://keras.io/api/. n.d.

[51] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 2014;15:1929–58.

[52] Oey T, Jones S, Bullard JW, Sant G. Machine learning can predict setting behavior and strength evolution of hydrating cement systems. J Am Ceram Soc 2020;103:480–90. https://doi.org/10.1111/jace.16706.

[53] Huang BFF, Boutros PC. The parameter sensitivity of random forests. BMC Bioinformatics 2016;17:331. https://doi.org/10.1186/s12859-016-1228-x.

[54] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res 2011;12:2825–30.

[55] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., New York, NY, USA: ACM; 2016, p. 785–94. https://doi.org/10.1145/2939672.2939785.

[56] Tharwat A. Classification assessment methods. Appl Comput Informatics 2021;17:168–92. https://doi.org/10.1016/j.aci.2018.08.003.

[57] Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett 2006;27:861–74. https://doi.org/10.1016/j.patrec.2005.10.010.

[58] Hand DJ, Till RJ. A simple generalisation of the area under the ROC curve for multiple class classification problems. Mach Learn 2001;45:171–86.

[59] Su C, Ju S, Liu Y, Yu Z. Improving Random Forest and Rotation Forest for highly imbalanced datasets. Intell Data Anal 2015;19:1409–32. https://doi.org/10.3233/IDA-150789.

[60] Bergstra J, Bengio Y. Random search for hyper-parameter optimization. J Mach Learn Res 2012;13.

[61] Vandamme M, Ulm F-J, Fonollosa P. Nanogranular packing of C–S–H at substochiometric conditions. Cem Concr Res 2010;40:14–26. https://doi.org/10.1016/j.cemconres.2009.09.017.

[62] Yu Z, Zhou A, Lau D. Mesoscopic packing of disk-like building blocks in calcium silicate hydrate. Sci Rep 2016;6:36967. https://doi.org/10.1038/srep36967.

[63] Wei Y, Gao X, Liang S. A combined SPM/NI/EDS method to quantify properties of inner and outer C-S-H in OPC and slag-blended cement pastes. Cem Concr Compos 2018;85:56–66. https://doi.org/10.1016/j.cemconcomp.2017.09.017.

[64] Kim H, Ahn E, Cho S, Shin M, Sim S-H. Comparative analysis of image binarization methods for crack identification in concrete structures. Cem Concr Res 2017;99:53–61. https://doi.org/10.1016/j.cemconres.2017.04.018.

[65] Moon J, Yoon S, Wentzcovitch RM, Monteiro PJM. First-principles elasticity of monocarboaluminate hydrates. Am Mineral 2014;99:1360–8. https://doi.org/10.2138/am.2014.4597.