a

Predicting the incidence of alkali-aggregate reaction in Finnish bridges with machine learning

Previsão da incidência das reações álcalis-agregado em pontes finlandesas usando aprendizagem automática

Tandré Oey
Tapio Vehmas
Antti Torkki
Miguel Ferreira
Edgar Bohner

Abstract

Alkali-aggregate reaction (AAR) is a prominent degradation mechanism of concrete structures, which results from the dissolution of reactive silicate aggregates and the associated formation of damaging, expansive AAR gels. Due to the complex nature of AAR reactions, it has remained ambiguous which factors contribute the most to its occurrence on a global scale. Similarly, concrete monitoring in Finland has only recently begun to adapt to the reality that AAR often occurs concomitantly with other degradation mechanisms such as freeze-thaw damage, highlighting the need for critical evaluation of current methods to identify AAR occurrence and distinguish it from these other mechanisms. Machine learning (ML) provides a data-driven framework for both evaluating the relative "importance" of various data features, as well as predicting their consequent influence on AAR damage to concrete. Building on the success of previous such data-driven learning for the prediction of well-defined concrete properties such as compressive strength and setting time, the current work evaluates the feasibility of extending ML methods to AAR-relevant predictions. Results provide new insights into several of the most relevant concrete characteristics linked with AAR occurrence, and establish a basis for future work to extend and enhance such predictions to supplement monitoring and risk management of concrete structures.

Keywords: Alkali-aggregate reaction / Bridges / Feature importance / Machine learning / Prediction

Resumo

A reação álcalis-agregado (AAR) constitui um mecanismo de degradação proeminente das estruturas de betão, que resulta da dissolução de silicatos de agregados reativos e da formação associada de geles expansivos, que provocam danos. Devido à natureza complexa da AAR, são ambíguos os fatores que mais contribuem para a sua ocorrência, na escala global. Por outro lado, a monitorização das estruturas de betão, na Finlândia, só recentemente começou a adaptar-se à realidade de que a AAR ocorre muitas vezes em simultâneo com outros processos de deterioração, como degradação por ciclos de gelo-degelo, destacando a necessidade de avaliação crítica dos métodos atuais para identificar a ocorrência de AAR e distingui-la dos outros mecanismos. A aprendizagem automática (ML) constitui uma ferramenta, baseada em dados, para avaliar a importância relativa desses dados e prever a sua influência nos danos provocados pela AAR. Com base na aprendizagem anterior, baseada em dados bem definidos de previsão das propriedades do betão, como a resistência à compressão e tempo de cura, o trabalho pretende avaliar a viabilidade de estender os métodos de ML para previsões relevantes sobre os efeitos da AAR. Os resultados fornecem novos elementos sobre várias características do betão que são relevantes para a ocorrência de AAR e estabelecem uma base para trabalhos futuros, para estender e melhorar as previsões, em complemento da monitorização e da gestão das estruturas de betão.

Palavras-chave: Reação álcalis-agregado / Pontes / Importância do carácter dos dados / Aprendizagem automática / Previsão

Tandré Oey

VTT Technical Research Centre
Espoo, Finland
tandre.oey@vtt.fi

Tapio Vehmas

VTT Technical Research Centre
Espoo, Finland

Antti Torkki

VTT Technical Research Centre
Espoo, Finland

Miguel Ferreira

VTT Technical Research Centre
Espoo, Finland

Edgar Bohner

VTT Technical Research Centre
Espoo, Finland

OEY, T. [*et al.*] – Predicting the incidence of alkali-aggregate reaction in Finnish bridges with machine learning. **Revista Portuguesa de Engenharia de Estruturas**. Ed. LNEC. Série III. n.º 15. ISSN 2183-8488. (março 2020) 65-76.

1 Introduction

Alkali aggregate reaction (AAR), also commonly referred to as alkali silica reaction (ASR), is a prevalent durability problem in concrete infrastructure. The chemical and physical mechanisms leading to AAR damage of concrete remain poorly understood [1]. While broadly, AAR is acknowledged to be caused by reactive silica-containing aggregates, and exacerbated by high levels of alkalinity and moisture, more detailed understanding is still being developed, complicating efforts to develop general methods for AAR mitigation. In Finland in particular, the likely co-occurrence of AAR with other damage mechanisms (e.g., freeze-thaw damage) has even confounded efforts to diagnose this process, let alone mitigate it. In the absence of knowledge needed to predict AAR occurrence, data-driven machine learning (ML) methods offer an attractive, mechanism-agnostic approach that may achieve such predictions even while lacking detailed physical understanding. For example, ML methods have recently been demonstrated to make reasonable estimations of the 28 day compressive strength of concrete, from either mixture proportions [2], or cement composition and fineness [3]. Although limited application of ML methods to modelling of AAR has been attempted [4], these efforts have likely suffered somewhat from a reliance on largely synthetic, and possibly biased data, as well as relying on accelerated laboratory test methods. The current study focuses on a dataset acquired from inspection reports of several concrete bridge structures currently in service across Finland, with the aim of evaluating feasibility of such ML methods for the prediction and/or diagnosis of AAR under realistic operating conditions.

2 Background and machine learning procedures

2.1 Machine learning algorithms

Several recent studies have demonstrated the ability of bootstrap-aggregated (or bagged) decision tree ensembles to accurately estimate compressive strength of concrete [2, 3]. These rule-based models identify logical splits in the data and partition the input space into a "tree" of decision nodes, which are then followed to arrive at predictions of a target data feature (i.e., "leaf nodes") for each given set of input features. A collection, or ensemble of trees are constructed, with each tree then being trained on different subsets of the data and their results averaged to produce the final prediction [5]. Recent application of these models to concrete has largely been in the context of regression problems, i.e., prediction of a given property (e.g., compressive strength) as a continuous function of input variables. The success of such bagged decision tree ensembles for concrete datasets, which may have relatively high inherent variability, highlights their potential for successful extension to classification problems in this domain, i.e., as necessitated by the nature of the dataset currently under consideration for prediction of AAR incidence in Finnish bridges. As such, bootstrap-aggregated decision tree ensembles are the primary focus of the current study for classification of AAR. More specifically, two particular methods

for decision tree construction are given primary consideration due to their past success in the domain of concrete materials: (1) "random forest," wherein a random subset of the input features is considered when determining the split at each decision node, and (2) "extra trees," wherein the threshold for each split is also randomly determined ("extra" being a portmanteau of "extremely" and "randomized"). Other classification algorithms beyond these two were also examined, though it was confirmed that the "extra trees" type random forest model obtained the highest accuracy among these (see *Section 3.1*). All the algorithms and auxiliary functions used for classification were sourced from the scikit-learn library, and can be accessed and downloaded, along with their documentation, at http://scikit-learn.org/stable/.

2.2 Data collection and pre-processing

The dataset under consideration was provided by Väylävirasto, the Finnish Transport Infrastructure Agency. It originally consisted of 156 data records, selected from bridge inspection reports collected on concrete bridge structures across Finland between the years 2016 and 2018 to provide a representative sample of different regions and construction years (methods defined in [6, 7]). Data feature types represented in these original data records are listed in Table 1, along with several that were derived from these features for the purpose of this study. Several data records were noted to be missing some of these original features, and were omitted prior to data pre-processing, resulting in a new input set of 136 data records. Features which consisted of text were then converted to numerical values as follows:

1) Features descriptive of "types" of bridges (e.g., use, maintainer, etc.) were converted to integer values, with lower values arbitrarily assigned to those types that occurred first in a randomly shuffled list of the data records.
2) Features descriptive of "abundances" of materials (e.g., the occurrence of various rock types in concrete aggregates) were assigned integer values, with lower values assigned to relatively lower reported abundance (i.e., "none" = 0, "little" =1, "moderate" = 2, "abundant" = 3).

Table 1 Input features used to train machine learning models for classification of AAR gel (the target feature, shown in gray text). Features from the original bridge inspection reports are shown in plain text, while features determined from these reports for the purposes of the current study are shown in *italic*. The final set of nine features used as inputs for optimized classification models are underlined.

Type-Features
Use (1-7)
Maintainer (1-9)
Environment (0-3)
Marine Exposure (0-1)
Pre-stress (0-1)

Abundance-Features	
Thin Section	**Aggregate Rock**
AAR Gel (0-3)	Slate (0-3)
Cracking (0-2)	Granite (0-3)
Air Voids (0-3)	Gneiss (0-3)
Ettringite (0-3)	Amphibole (0-3)
Calcite (0-2)	Meta-Tuff (0-3)
	Diabase (0-3)
	Quartzite (0-2)
	Sandstone (0-1)
	Diorite (0-1)
	Phyllite (0-2)
	Limestone (0-2)
	Gabbro (0-3)
	Mylonite (0-1)

Numerical-Features	
Records	**Testing**
Year Built	Tensile Strength (Minimum)
Year Inspected	Tensile Strength (Maximum)
Location (Latitude)	Carbonation Depth (Minimum)
Location (Longitude)	Carbonation Depth (Maximum)
Condition Score	Chloride Content (Mass % Concrete, Minimum)
Condition	Chloride Content (Mass % Concrete, Maximum)

Visual Inspection
Map Cracking (0-2, RILEM)
Deformation (0-2, RILEM)
Discoloration (0-2, RILEM)
Exudations (0-2, RILEM)
Pop-Outs (0-1, RILEM)
Environment (2-3, RILEM)
Other Damages (0-3)

Numerical features that displayed low or zero variance (i.e., those that were nearly uniform across every data record) were omitted, following confirmation of negligible feature importance (as described in *Section 3.2*) to verify that they did not contribute significantly to classification accuracy. Remaining feature data was inspected to ensure that no unphysical or meaningless values were present (e.g., negative carbonation depths, etc.), of which none were found. Data records were also checked for duplicates, with none being found. Finally, the data records were randomly shuffled to avoid introducing inadvertent bias in the sampling of input attributes during model construction.

Several data features not typically included in inspection reports were also determined and included in the input dataset. These consisted of visual criteria for AAR diagnosis, determined according to RILEM guideline 6.1 [8], and assignment of numerical values to quantify observations of "other damage" (e.g., freeze-thaw damage) as provided from the bridge inspection reports. These additional features, which were assessed similarly to the original features descriptive of "abundances" as described in (2) above, are also included in Table 1, and distinguished from original inspection data using *italic* text.

2.3 Training of classification models

AAR gel abundance was chosen as the target feature because the available dataset contains no other features that would provide a simple (e.g., expansion), unambiguous indicator of AAR occurrence. Even if such features did exist, the available data are also lacking in important details relevant to prediction of concrete's mechanical strength (i.e., cement composition, fineness and mixture proportions [2, 3]), which are likely necessary for ML to extrapolate from the presence of AAR gel to the manifestation of AAR-induced expansion and damage. Lastly, and of particular relevance to AAR incidence in Finland, AAR damage often occurs concomitantly with other damage mechanisms (e.g., freeze-thaw damage), making it difficult to even unambiguously link characteristic cracking patterns or other indicators with AAR. For these reasons, classification by AAR gel occurrence represents a small yet critical first step in moving toward models that would someday be able to predict the extent and kinetics of AAR damage.

Numerous ML classification models were constructed and applied to predict the abundance or occurrence of AAR gel. Training and testing of each model took between one-to-ten seconds, depending on the model used (i.e., for a personal computer with an Intel Core i5-8350u 1,7 GHz processor, and 8GB RAM). Performance of each classification model was evaluated using the prediction accuracy metric, i.e., what fraction of predictions on the test data produced the correct classification for AAR gel abundance (with accuracy ranging from 0 to 1, wherein a value of 1 would indicate that every prediction was correct). Both training and testing were conducted on different portions of data using a standard low-bias resampling procedure called k-Fold Cross-Validation [5]. The data records were randomly split into k = 10 "folds," nine of which were used to train the model, and one of which was used to evaluate the model 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 classification models used in this study were not noted to be sensitive to the magnitude of the data attributes, in the sense that they may potentially have been biased to assign more importance to attributes with inherently greater values. Nonetheless, to address this potential artefact, 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, i.e., using sklearn.preprocessing.StandardScaler). 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. In any case, no difference in model accuracy was found between results obtained with and without this rescaling procedure.

2.4 Hyper-parameter optimization

Following model selection and feature selection (detailed in *Section 3.1* and *3.2*), the most accurate model was optimized for the selected feature set in terms of its hyper-parameters (i.e., those that are set before model training, which dictate the speed and quality of the learning process), to arrive at a current best-in-class predictor of AAR occurrence (i.e., the occurrence or abundance of gel). This classification model, an "extra trees" type random forest model, contains two embedded hyper-parameters: (1) the number of data features considered at each "split" when building decision trees, and (2) the total number of decision trees. Additionally, the k-fold cross validation procedure contains a third hyper-parameter: (3) the number of "folds" into which the data is divided prior to construction of the training and test sets.

In the number of features per "split," an optimum in accuracy for five features is observed (out of nine total features, selected as described *Section 3.2*). A similar optimum in accuracy occurs for six "folds" in cross-validation. This is likely due to the small size of the dataset being considered: whereas too few "folds" decreases the size of training sets to a degree sufficient to impair classification accuracy, too many decreases the likelihood that the test set will be representative of broader trends in the data. Lastly, accuracy plateaus with number of trees, with diminishing returns to adding decision trees beyond about 400 (for this specific dataset and ML model). Final models are limited to the use of 400 trees to avoid "overfitting," i.e., building a model that sacrifices its ability to generalize to predictions on unseen data by too closely fitting the available data. Final values for each of the three optimized hyper-parameters with the "extra trees" model are marked in red in Figure 1, corresponding to five features per "split," six "folds," and 400 trees.

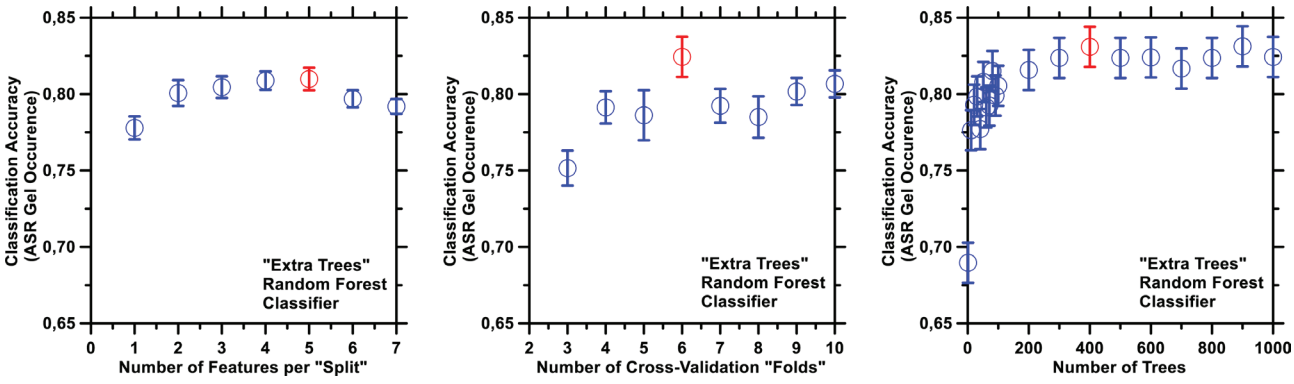

Figure 1 The results of a representative hyper-parameter tuning exercise using the "extra trees" algorithm to classify bridge concretes for ASR gel occurrence, showing (a) an optimum of five features per "split," (b) an optimum of six "folds" of data during cross-validation, and (c) an optimum of 400 trees. Points shown in red indicate parameter values used for final models. Error bars represent the standard deviation between 20 separate train-test repetitions of the models.

3 Results and discussion

3.1 Ensemble models perform best for classification of AAR gel amount

Unsurprisingly, ensemble models of decision trees outperform most other classification models in predicting the abundance of AAR gel in the Finnish bridge concrete dataset (Table 2). Specifically, the "extra trees" random forest model provides the highest classification accuracy, both before and after the feature selection described in *Section 3.2*. This again highlights the ability of such models to perform well, even for data obtained from real concretes that may be prone to high inherent variability [2], a finding now borne out in the context of classification and AAR. Further consideration during feature selection and later steps of analysis is given only to the "extra trees" model, in the interest of defining an upper bound on AAR prediction accuracy that can be obtained using existing bridge inspection data.

3.2 Recursive feature elimination improves classification accuracy

The best performing classification model using all features of the bridge inspection dataset, an "extra trees" type random forest model, was able to obtain only an accuracy score of 0,649, corresponding to about 65% of correct predictions of AAR gel abundance. For comparison, the same classification model trained and tested using inputs where reported values of AAR gel abundance were *randomly* assigned to the data records yielded an accuracy score of 0,450, i.e., the worst-possible performance that could likely be achieved using

Table 2 Accuracy of several common classification models when predicting AAR gel abundance. All models used are open-source, and their documentation can be found at http://scikit-learn.org/stable/.

	Accuracy (0-1)	
Classification Model	After Feature Selection	Before Feature Selection
Non-ensemble models		
K-nearest neighbours	0,600	0,571
Logistic regression	0,630	0,565
Linear discriminant analysis	0,632	0,584
Support vector machine	0,683	0,580
Decision tree	0,676	0,515
Ensemble models		
Bagged trees	0,670	0,612
Random forest	0,691	0,632
Gradient boosted trees	0,692	0,581
"Extra trees" random forest	**0,728**	**0,649**

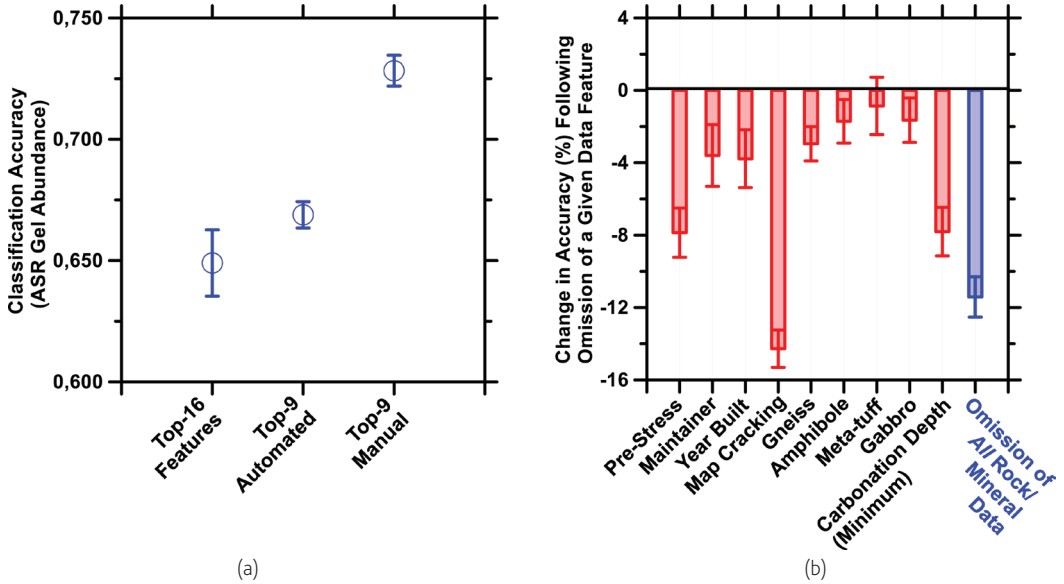

Figure 2 Results of feature selection, demonstrating (a) the better classification accuracy achieved during recursive feature elimination by preferentially removing features whose omission improved the model prior to those that had negligible significance, and (b) the final "feature importances" of the nine features remaining after recursive feature elimination (i.e., all of which resulted in significantly reduced classification accuracy when omitted from the input data). The cumulative "importance" of the combined rock/mineral data (i.e., decline in accuracy when omitting gneiss, amphibole, meta-tuff, and gabbro together) is also shown for comparison. Error bars represent the standard deviation between 20 separate train-test repetitions of the models.

this dataset. While an accuracy of 0,649 is appreciably higher than this value, it is nonetheless desirable to further refine the "extra trees" model by performing feature selection.

Feature selection typically consists of sequentially removing various data features from the inputs, to identify features that may be (1) of negligible importance to predicting the target feature, or (2) detrimental to such predictions due to being redundant or partially redundant with other features (thus introducing bias). Two methods for such feature selection were tested, both of which fall under the umbrella of recursive feature elimination: omitting features one-by-one and then determining which feature to eliminate from the model based on how much of a change in accuracy its omission produced. The default recursive feature elimination algorithm provided by the scikit-learn library sequentially eliminates the features whose omission produces the *lowest absolute change* in model accuracy. However, it was found that first eliminating those features whose omission produced increase in model accuracy (followed by those producing the lowest absolute decrease) led to a more accurate final model, as illustrated in Figure 2(a). As an example, omission of granite, quartzite, and sandstone abundance led to a roughly 5% increase in classification accuracy, likely due to co-occurrence with other rock types (and thus redundancy with data features corresponding to those rock types). The final feature set arrived at via recursive feature elimination, consisting of nine data features (Figure 2(b)), resulted in a classification accuracy improvement for the "extra trees" model to 0,728, or about 73%.

3.3 Feature importances highlight need for improved monitoring practices

Feature importance, defined herein by the magnitude of accuracy loss (%) upon omission of a given feature from ML model inputs (Figure 2(b)), provides a rough indicator of which available data features one can expect to be most relevant to AAR incidence. Among the nine features remaining after recursive feature elimination, the extent of "map-cracking," a feature often used to distinguish AAR occurrence by visual inspection, appears to be of primary importance. On the one hand, this lends credence to the RILEM method developed to evaluate such visible damage in concrete structures, by demonstrating a significant link with observed AAR gel formation. Additionally, model performance appears not to depend on the "other damage" feature, implying that such ML models represent a reliable means to distinguish AAR from other concurrent damage mechanisms (at least based on the current methods for evaluating such concurrent damage mechanisms). On the other hand the importance of "map-cracking," along with minimum carbonation depth, implies that the current model depends strongly on observation of damage that has already occurred (i.e., rather than being able to estimate AAR risk *a priori*). While this link should be further explored, e.g., by pursuing more quantitative metrics for AAR damage such as via computer-aided image analysis or other non-destructive testing techniques [9–11], it would ultimately be more desirable to train such models exclusively on data reflecting the

state of the concrete *before* AAR damage has occurred (e.g., mixture proportions, aggregate mineralogy, etc.).

Importance of unexpected data features, specifically minimum carbonation depth and maintainer, may suggest either (1) the importance of "embedded" data, which correlates with these features but is absent from inspection records (e.g., moisture-resistant coatings that may have been applied to the concrete by specific maintainers), or (2) artificial differences in data due to variable inspection methods (e.g., near-surface sampling that may have impacted AAR gel assessments due to carbonation). While it is currently impossible to distinguish between these two options, both point toward the need for more extensive record keeping and/or more standardized inspection and sampling procedures. This is likely all the more important for one feature that is conspicuously absent from the current dataset, yet widely acknowledged to play a critical role in AAR: moisture [1]. Although several features were present that may have been related to the bridge concrete's moisture state, i.e., environment and location, neither was found to be necessary in accurately classifying bridge concretes by AAR gel abundance. This may of course be due to a similar moisture state among bridges because of the geographically limited nature of the dataset (within Finland), and it is likely that the current classification models would require more detailed data on the moisture state of concrete in order to be applied under circumstances where environmental conditions vary more widely.

Another factor expected to vary between concretes undergoing AAR is their stress state, or confinement, which has been reported to result in preferential expansion along the direction of minimum applied load (i.e., resulting in cracking primarily transverse to this direction [12]). Though past efforts to incorporate the influence of stress into understanding of AAR progression have mainly focused on mechanical effects [13], the importance of pre-stress to AAR gel abundance for the bridge concretes currently under study (Figure 2(b)) suggests that even this rough indicator of mechanical stress is significantly linked to *chemical* driving forces for AAR, i.e., those necessary to influence gel formation. Though it is difficult to comment further without more detailed data on the stress state of these concretes, this would be in line with reported mechanisms for chemo-mechanical coupling in mineral-water systems [14, 15], which have been employed to propose a dissolution-precipitation mechanism at the origin of concrete creep [16]. Current results indicate that more detailed investigation along these lines is merited, specifically the possibility that applied loads may influence dissolution and precipitation in the context of AAR.

Among the lowest "importance" values were assigned to the four rock/mineral features retained by the recursive feature elimination process. However, their cumulative importance was substantial (i.e., when *all* were omitted from the model), second only to the importance of the "map cracking" feature (Figure 2(b)). Obtaining more detail as to the AAR reactivity of these rocks/minerals thus represents a very promising area for future improvements in the accuracy of such models, in particular because they represent a *pre-damage* descriptor of potential AAR risk (as opposed to features reliant on existing AAR damage like cracking or carbonation). Recent work has shown success in linking aqueous reactivity to chemical structure for silicate materials, i.e., those most relevant to AAR due to their release of silicon upon dissolution [17, 18]. To illustrate the potential of this concept, several simple assumptions were tested regarding the reactivity of each rock type or mineral provided in the input data:

1) Cumulative contribution to "reactivity," as relevant to AAR, was assumed to be proportional to the reported abundance of each rock type or mineral (i.e., with reactivity being calculated as a weighted average as outlined in Equation 1 below).
2) "Reactivity" of each rock type, constituted of several minerals, was assumed to be dictated by the most abundant of these minerals in a typical specimen of the given rock (e.g., orthoclase for granite, plagioclase for diorite, pyroxene for gabbro, etc.).
3) "Reactivity" of each silicate mineral (whose structures are defined by SiO_4 tetrahedral units, Figure 3) was assumed to be inversely proportional to the average number of inter-tetrahedral bonds within its chemical structure, i.e., 4 for orthosilicates, 2 for single-chain silicates, 1,5 for double-chain silicates, 1 for sheet silicates, and 0 for tectosilicates [19]. This is an extension of the idea that silicon dissolution is required for AAR to occur [1], and furthermore that silicon dissolution proceeds via sequential breaking (hydrolysis) of inter-tetrahedral Si-O-Si bonds [17]. Such a formulation is the equivalent to the average number of "non-bridging" oxygen atoms that make up the corners of the SiO_4 tetrahedral units in silicates [20], as illustrated in Figure 3.
4) "Reactivity" of each non-silicate mineral (i.e., calcite, from limestone) was assumed to be 0.

$$\text{"Si-Bonding Weighted Average"} = \frac{\sum_1^i A_i N_i}{\sum_1^i A_i} \quad (1)$$

where A_i is the reported abundance for mineral i, and N_i is the average number of "non-bridging" oxygen atoms per tetrahedral unit in the mineral's molecular structure (or 0, for non-silicates).

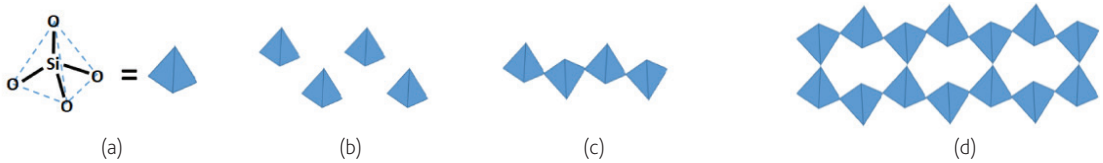

Figure 3 Illustrations of local molecular structure of (a) an SiO_4 tetrahedral unit, and such units in (b) an orthosilicate: 0 inter-tetrahedral bonds, 4 "non-bridging" oxygens per unit; (c) a single-chain silicate: 2 inter-tetrahedral bonds, 2 "non-bridging" oxygens per unit; and (d) a double-chain silicate: alternating between 2 and 3 inter-tetrahedral bonds, 2 and 1 "non-bridging" oxygens per unit, etc.

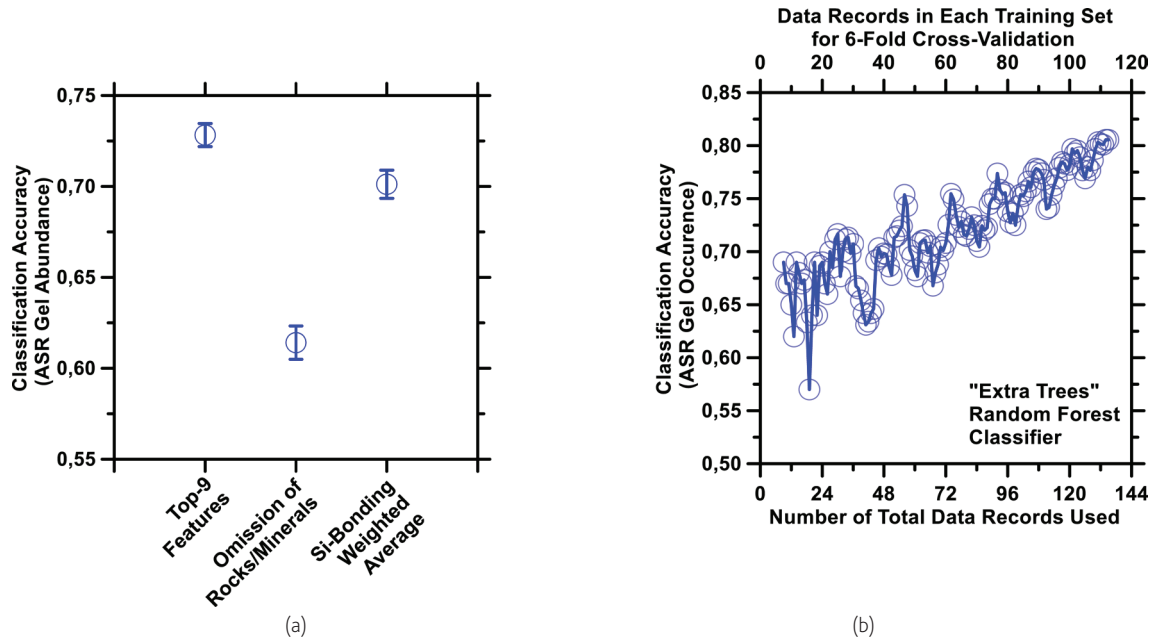

Figure 4 (a) ML classification accuracy following recursive feature elimination (i.e., "Top-9 Manual" from Figure 2(a)) as compared with other ML models wherein (1) all rock and mineral data features were omitted from the inputs, and (2) the "Si-bonding weighted average" was re-introduced to the inputs following such omissions. Error bars represent the standard deviation between 20 separate train-test repetitions of the models. (b) Classification accuracy as a function of the total number of data records used in ML inputs, i.e., an illustrative "learning curve" for the ML classification models.

The end result of the simple assumptions outlined above regarding the influence of silicate aggregates' chemical structure on their AAR reactivity is a weighted average (Equation 1), meant to reflect the cumulative influence of each rock type or mineral on the formation of AAR gel. As demonstrated by Figure 4(a), this "Si-bonding weighted average" produced models with comparable classification accuracy when substituted for the raw data on rock type and mineral abundance in ML inputs. The surprising success of even these very simple assumptions in distilling the data from rock and mineral abundance to a single descriptive parameter, without sacrificing classification accuracy, highlights that improved aggregate characterization methods, e.g., reporting mineralogy data rather than just rock types, hold great potential to improve such predictive modelling efforts.

Though not reflected among the importances of individual data features, the number of data records available should also be noted for its substantial influence on the quality of current classification models. The dataset under consideration, consisting of only 136 data records (i.e., data from each of 136 unique bridge inspection reports), is quite small relative to the size of datasets typically utilized for training ML models. Though recent work has highlighted the ability of such models to perform well even with only a few hundred data records [3], it is nonetheless expected that the AAR gel classification models would continue to improve as more data becomes available. To illustrate this point, "learning curves" were constructed by sequentially omitting full data records from the inputs (Figure 4(b)). It can be seen that, though current models are unstable (accuracy fluctuates with small perturbations to input data), there is a consistent improvement in accuracy with the inclusion of increasingly more data records. The lack of a "plateau" in such a learning curve for the current, limited dataset suggests the likelihood that provision of even more data to the model would, in fact, be able to produce additional accuracy improvements.

3.4 Addressing imbalance in AAR gel class distribution improves accuracy

Though accuracy improvements often necessitate the provision of greater amounts of more detailed data, it is important to recognize that this is not always the case [3]. As an example, skewed distributions between the classes in the target feature (AAR gel abundance) could be expected to somewhat bias the model toward predicting the occurrence of no AAR gel, as this is the majority class (Table 3). For comparison, prediction of only the absence of AAR gel would result in a classification accuracy of 58%. Two methods have been investigated to deal with this possible class bias in ML models: (1) binarization of the classes, i.e., shifting from consideration of *how much* AAR gel has formed to the simpler consideration of *whether* any has formed or not; and (2) stratification of the "folds" during k-fold cross validation, i.e., ensuring that each subdivision of the data used to train and test the models contains an equal number of members of each target feature class.

Table 3 Class distributions for the originally marked dataset according to AAR gel abundance, and the new class distribution after application of binarized class markings based on AAR gel occurrence.

AAR Gel *Abundance* Class Distribution	
Class	Population
"None" (0)	79
"Little" (1)	45
"Moderate" (2)	11
"Abundant" (3)	1

AAR Gel *Occurrence* Class Distribution	
Class	Population
"None" (0)	79
"Any" (1)	57

Binarization of target feature data from AAR gel abundance to AAR gel occurrence does somewhat reduce the class imbalance problem (Table 3). This results in a marked improvement in model accuracy to about 80% (Figure 5(a)), and may also benefit from the fact that such binarization helps simplify the evaluation of AAR by obscuring any bias accompanying inter-operator variability in quantification of gel abundance from thin section analyses (i.e., removing the degree of subjectivity associated with what any given bridge inspector regards as "little" relative to "moderate" AAR gel). Stratification during cross-validation produces a similar improvement in accuracy to about 78%, or 82% with the combination of both binarization and stratification (Figure 5(a)). This result highlights that current ML classification models are at present best-suited to use as binary AAR diagnostic tools, which identify whether observed damage can be attributed to AAR rather than some other mechanism. Furthermore, it highlights the importance of providing such models with balanced training data, i.e., indicating the need for preferential collection of data from concrete structures that experience AAR, and even more so from those that experience moderate-to-abundant formation of AAR gel. While over-sampling or under-sampling are also known to help reduce such bias [3], i.e., by preferentially including duplicate data records of underrepresented classes or excluding data records of overrepresented classes, neither of these methods were found to be feasible for the current dataset due to its relatively small size.

To verify that class imbalance had been suitably addressed by binarization, the specific nature of each correct or incorrect classification by the model was examined (Figure 5(b)). For example, if the model was biased to preferentially classify data records as having no AAR gel due to this being the majority class, many of its classifications would be false negatives, i.e., they would predict no AAR for concrete where in reality AAR had occurred. Grouping predictions as such yields four types: true negatives (correct predictions of no AAR), true positives (correct predictions of AAR), false positives (incorrect predictions of AAR), and false negatives (incorrect predictions of no AAR). Looking at the split between true positives and true negatives in Figure 5(b), these appear to be in

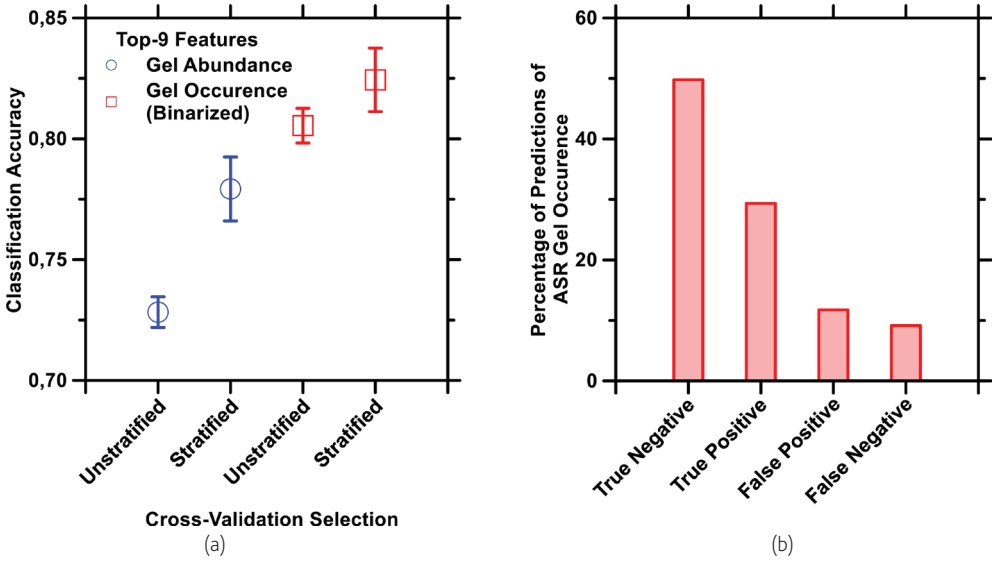

Figure 5 (a) A comparison in classification accuracy between models with stratified and un-stratified cross-validation for classifying AAR gel abundance (four classes) or occurrence (two classes). (b) The distribution among classification predictions by correctness as well as class, e.g., making the additional distinction between incorrect predictions for those which did or did not truly have AAR gel. Error bars represent the standard deviation between 20 separate train-test repetitions of the models.

rough proportion to the class distribution between concrete with and without AAR gel, respectively. This result confirms that class bias has been largely addressed by binarization and stratification for the current dataset and ML models.

It should also be noted, however, that the "cost" of the various types of predictions has not been factored into current ML models. For example, a model that falsely suggests AAR has occurred when in reality it has not may be deemed as less of a risk than a model that falsely suggests AAR has not occurred when in reality it has. This highlights a promising area for future research regarding use of ML for AAR risk assessment: rather than using algorithms that maximize the number of correct predictions, they could instead minimize the number of false negatives (or any other type of prediction). In the current case of AAR evaluation in concrete, this would mean the difference between the minor short-term cost of applying unnecessary AAR mitigation measures, and the potentially catastrophic long-term cost of structural failure. In this light, if one includes false positives in the calculation of accuracy for the current best-in-class model, it performs at 90% "reliability," only *failing to detect* AAR 10% of the time. Likewise, in the pursuit of a predictive rather than a diagnostic model for AAR (i.e., one not dependent on observation of damage), even previously "poor" models perform somewhat well. When the rate of false positive predictions is combined with classification accuracy, a model trained to make binary predictions for AAR incidence *without* the "map cracking" and carbonation depth features performs with a "reliability" of 85%. This is promising in that, even without improvements in the size or quality of input data, new ML algorithms for cost-sensitive optimization may also have the potential to improve reliability of AAR classification to the point required for their more widespread use in risk assessment or mix design.

4 Summary and conclusions

This study evaluated the feasibility of applying ML methods to the prediction and/or diagnosis of AAR in concrete under realistic service conditions. It was found that using an "extra trees" type random forest algorithm, an accuracy of 78% could be achieved when classifying concretes by AAR gel abundance. This accuracy was improved to 82% when applied to the simpler classification of the existence (or not) of AAR gel in the concretes, and to 90% when factoring in the relatively lower "cost" of falsely predicting the occurrence of AAR (relative to that of falsely predicting that AAR did not occur).

Several significant trends in the importance of various data features were also highlighted, particularly the promise shown by an exercise to incorporate a more detailed understanding of aggregate reactivity while maintaining model accuracy. These trends suggest the need for both:

(1) More thorough record keeping of concrete's raw materials, particularly with regard to details about the cement composition, mixture proportions, and aggregate mineralogy, and

(2) More in-depth monitoring activities, particularly of the moisture and stress states of the concrete.

While the AAR classification models currently seem best-suited to limited diagnostic use, i.e., to facilitate more confident and rapid deployment of AAR mitigation measures, their accuracy and applicability is likely to improve upon the development of either:

(1) Larger, more detailed datasets, and/or

(2) Improved "cost-sensitive" ML algorithms, potentially qualifying them for future application to risk assessment or even concrete mix design.

5 References

[1] Rajabipour F, Giannini E, Dunant C, Ideker JH, Thomas MD (2015) Alkali–silica reaction: Current understanding of the reaction mechanisms and the knowledge gaps. Cem Concr Res 76:130–146. https://doi.org/10.1016/j.cemconres.2015.05.024

[2] Young B, Hall A, Pilon L, Gupta P, Sant G (2019) Can the compressive strength of concrete be estimated from knowledge of the mixture proportions?: New insights from statistical analysis and machine learning. Cem Con Res 115:379-388. https://doi.org/10.1016/j.cemconres.2018.09.006

[3] Oey T, Jones S, Bullard JW, Sant G (2020) Machine learning can predict setting behavior and strength evolution of hydrating cement systems. J Am Ceram Soc 103:480–490. https://doi.org/10.1111/jace.16706

[4] Allahyari H, Heidarpour A, Shayan A, Nguyen VP (2020) A robust time-dependent model of alkali-silica reaction at different temperatures. Cem Concr Compos 106:103460. https://doi.org/10.1016/j.cemconcomp.2019.103460

[5] Brownlee J (2016) Machine Learning Mastery. Machine Learning Mastery Pty Ltd.

[6] Nummelin M, Äijälä M (2018) Taitorakenteiden erikoistarkastusten laatuvaatimukset - Sillat. https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=2ahUKEwiEwMKw4OTmAhVow4sKHX-8CsUQFjAAegQIBBAC&url=https%3A%2F%2Fjulkaisut.liikennevirasto.fi%2Fpdf8%2Flo_2018-28_taitorakenteiden_erikoistarkastusten_web.pdf&usg=AOvVaw163JCX1pvtux_55gggYQcw. Accessed 2 January 2020.

[7] Nummelin M, Äijälä M (2017) Sillantarkastuskäsikirjan sovellusohje. https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=2ahUKEwjZv-_n4OTmAhWow4sKHbmKAeUQFjAAegQIAhAC&url=https%3A%2F%2Fjulkaisut.liikennevirasto.fi%2Fpdf8%2Fohje_2017_sillantarkastuskasikirjan_sovellusohje_web.pdf&usg=AOvVaw2A6DbKCBHpF8aVKpqZuv6T. Accessed 2 January 2020.

[8] Godart B, Rooij M de, Wood J (2013) Guide to diagnosis and appraisal of AAR damage to concrete in structures. Springer Verlag, New York.

[9] Voltolini M, Marinoni N, Mancini L (2011) Synchrotron X-ray computed microtomography investigation of a mortar affected by alkali–silica reaction: a quantitative characterization of its microstructural features. J Mater Sci 46:6633–6641. https://doi.org/10.1007/s10853-011-5614-5

[10] Clayton DA, Ezell ND, Albright A, Santos-Villalobos HJ (2018) Detecting alkali-silica reaction in thick concrete structures using linear array ultrasound. In: Shull PJ (ed) Nondestructive Characterization and Monitoring of Advanced Materials, Aerospace, Civil Infrastructure, and Transportation XII. SPIE, p 4.

[11] Ezell NDB, Albright A, Floyd D, Clayton D, Khazanovich L (2019) A novel use of frequency-banded synthetic aperture focusing technique for reconstructions of alkali-silica reaction in thick-reinforced concrete structures. In: AIP Conf Proc 2102:110003. AIP Publishing LLC, Melville.

[12] Multon S, Toutlemonde F (2006) Effect of applied stresses on alkali–silica reaction-induced expansions. Cem Concr Res 36:912–920. https://doi.org/10.1016/j.cemconres.2005.11.012

[13] Pignatelli R, Comi C, Monteiro PJM (2013) A coupled mechanical and chemical damage model for concrete affected by alkali–silica reaction. Cem Concr Res 53:196–210. https://doi.org/10.1016/j.cemconres.2013.06.011

[14] Kristiansen K, Valtiner M, Greene GW, Boles JR, Israelachvili JN (2011) Pressure solution–The importance of the electrochemical surface potentials. Geochim Cosmochim Acta 75:6882-6892.

[15] Greene GW, Kristiansen K, Meyer EE, Boles JR, Israelachvili JN (2009) Role of electrochemical reactions in pressure solution. Geochim Cosmochim Acta 73:2862-2874.

[16] Pignatelli I, Kumar A, Alizadeh R, Le Pape Y, Bauchy M, Sant G (2016) A dissolution-precipitation mechanism is at the origin of concrete creep in moist environments. J Chem Phys 145:054701. https://doi.org/10.1063/1.4955429

[17] Pignatelli I, Kumar A, Bauchy M, Sant G (2016) Topological Control on Silicates' Dissolution Kinetics. Langmuir 32:4434–4439. https://doi.org/10.1021/acs.langmuir.6b00359

[18] Oey T, Kumar A, Pignatelli I, Yu Y, Neithalath N, Bullard JW, Bauchy M, Sant G (2017) Topological controls on the dissolution kinetics of glassy aluminosilicates. J Am Ceram Soc 100:5521–5527. https://doi.org/10.1111/jace.15122

[19] Nesse W (2012) Introduction to mineralogy. Oxford University Press, New York.

[20] Shelby J (2005) Introduction to glass science and technology. Royal Society of Chemistry, London.