# Comparison of Machine Learning Algorithms for the Prediction of the External Sulphate Attack Resistance of Blended Cements

Abdelhamid Hafidi[1,2], Benoit Hilloulin[1](✉), Sonia Boudache[1], Umunnakwe Rejoice[1,3], and Ahmed Loukili[1]

[1] Nantes Université, Ecole Centrale Nantes, CNRS, GeM, UMR 6183, 44000 Nantes, France
benoit.hilloulin@ec-nantes.fr
[2] Edycem, Rue du Fléchet, 85600 Boufféré, France
[3] Covenant University, KM 10 Idiroko Rd, Ota 112104, Nigeria

**Abstract.** Using supplementary cementitious materials (SCM) can help increase the sulphate resistance of cement blends. However, formulating sulphate-resistant materials with increasing amounts of SCM is challenging, and the required standard tests last several months. Therefore, creating new tools that can be easily applied and understandable could help develop novel materials in the future. Machine learning techniques have been widely used recently to predict cementitious materials' properties such as strength, creep, or shrinkage. However, their usage is relatively limited regarding durability properties, maybe because of the large number of parameters involved in durability processes, some of them intrinsic to the material and others related to the environment. In this study, an extensive database has been built using more than 300 cementitious sample characteristics from different studies. A large collection of inputs related to cement composition, mix composition, sample geometry, and environmental conditions such as sulphate concentration has been gathered. Then several machine learning algorithms were applied to assess the resistance of blended cements to the external sulphate attack. Two groups of algorithms, e.g., classification and Regression algorithms, incorporating several models from linear to ensemble models, have then been compared. The results show that most classification models can very quickly assess the sulphate resistance of cementitious materials using the extensive database, and the best Regression models can efficiently predict the temporal evolution of the degradation. The most influential parameters can be identified, and recommendations can be drawn regarding future blended cement compositions.

**Keywords:** Blended Cements · SCM · Sulphate Attack · Machine Learning

## 1 Introduction

Cement is the most widely used materiel in the word. The chemical interactions between cement's components and the environment usually lead to degradation of the building materials. Among all chemical attacks affecting cementitious materials, sulphate attack


A. Jędrzejewska et al. (Eds.): SynerCrete 2023, RILEM Bookseries 44, pp. 725–735, 2023.
https://doi.org/10.1007/978-3-031-33187-9_67

is the most documented [1–3]. Several factors are influencing the resistance of concrete faced with external sulphate attack (ESA) [4–7]. The reactions between sulphate ions and concrete lead to the precipitation of expansive agents like gypsum and ettringite, which damages concrete structure [8]. However, the complexity of chemico-physical aspects behind these reactions requires a good understanding of ESA process [9, 10]. Nevertheless, the development of an observable expansion takes several months, and the sensitivity of phenomenon affects the precision of laboratory tests [11].

Some experiments are made to accelerate the deterioration process. It is hard to validate their compatibility with the real process of ESA, even though they might be effective [8]. Furthermore, there is ongoing discussion regarding the relative significance of the various influential factors of ESA. For instance, the geometrical characteristics of the samples and the composition of the mortar and concrete mix significantly impact the volumetric expansion that is recorded during laboratory tests, affecting the final categorization of cements regarding their sulphate resistance.

Machine Learning (ML) is a branch of artificial intelligence techniques increasingly employed in civil engineering studies, and several recent applications opened new research paths. For example, convolutional neural networks and decision trees approaches have been successfully employed in the assessment of concrete qualities at many scales, from structural integrity to the microscale identification of cracks [12–14]. Similarly, sophisticated ensemble ML models have recently shown promising results in interpreting the physical aspects in different cases, such as creep [15], shrinkage [16], and compressive strength [17].

In this study a new approach based on artificial intelligence is adopted to determine concrete durability faced with ESA and determine the most influential factors and their contribution to the global behaviour of concrete structure exposed to sulphate solutions.

## 2 Database Description

### 2.1 Data Collection

This study aims to assess the effectiveness of ML models in studying the process of ESA. To this end, a detailed database has been constructed from different laboratory tests. This database, introduced in another study [18], comprises several experimental characteristics, including detailed cement composition, different additions, mix properties, aggressive solution characteristics, and mold properties. More than 20 parameters and 483 formulations grouping cement paste, mortar, and concrete were used to create an extensive and solid database. The choice of the inputs of the ML models had gone through several stages, ended up with a group of parameters summarized in Table 1. These parameters are then normalized and standardized to have comparable variances.

**Table 1.** Database description.

| | Mean | Std | Min | Max |
|---|---|---|---|---|
| $C_3S$ % | 56.66 | 7.95 | 17.7 | 74.0 |
| $C_2S$ % | 14.57 | 6.39 | 2.29 | 39.0 |
| $C_3A$ % | 7.51 | 2.86 | 0.0 | 11.9 |
| $C_4AF$ % | 9.01 | 2.7 | 0.0 | 19.7 |
| Cement (kg/m$^3$) | 459.13 | 251.14 | 55.0 | 1238.27 |
| Gravel (kg/m$^3$) | 87.26 | 282.49 | 0.0 | 1258.92 |
| Sand (kg/m$^3$) | 1166.78 | 518.2 | 0.0 | 1665.42 |
| Fly Ash (kg/m$^3$) | 34.03 | 84.66 | 0.0 | 469.79 |
| Slag (kg/m$^3$) | 44.75 | 106.09 | 0.0 | 788.15 |
| Water (kg/m$^3$) | 287.01 | 125.0 | 105.0 | 630.31 |
| Limestone (kg/m$^3$) | 23.91 | 58.98 | 0.0 | 388.0 |
| Metakaolin (kg/m$^3$) | 14.02 | 36.59 | 0.0 | 174.43 |
| Silica Fume (kg/m$^3$) | 3.6 | 16.19 | 0.0 | 132.44 |
| Water/Binder | 0.52 | 0.16 | 0.29 | 1.83 |
| Aggregates/Cement | 3.78 | 2.72 | 0.0 | 17.75 |
| Concentration % | 5.94 | 5.03 | 0.3 | 20.0 |
| pH | 9.05 | 1.66 | 3.0 | 12.3 |
| Mold properties | 0.05 | 0.21 | 0.0 | 1.0 |
| Surface/perimeter (cm) | 0.89 | 0.58 | 0.25 | 2.5 |
| $f_{c28}$ (MPa) | 49.53 | 12.97 | 20.6 | 100.0 |

## 2.2 Data Pre-processing

The experimental dataset contains some missing inputs as well as categorical parameters that cannot be directly fed into the algorithms. An initial pre-processing has been performed to circumvent this problem, starting by inferring missing values. For clinker composition, missing values were calculated using Bogue equations based on the oxide composition, and missing values of uncontrolled pH were set to 10.

For categorical parameters, e.g., type of cation, type of cement and mold form, encoding has been used. For example, two types of cation were considered in this study: sodium (Na) and magnesium (Mg), which were represented by using a simple encoder with the value 1, resp. 0. The same operation has been performed concerning the mold form.

## 2.3 Data Splitting

Good data is the crucial key to accurate AI modeling [19]. In this study, specific attention was paid to the database collection due to the sensitivity of expansion and the likely correlation between parameters presented in Table 1. The final database has been randomly divided into training data which constitute 80% (386 different formulations) of the database, and the last 20% are testing data (97 different formulations).

# 3 Methodology

## 3.1 Machine Learning Models

In this work, four ML models were used. These models are some of the best models employed to solve and predict civil engineering phenomena. These models were validated using 5-fold cross-validation.

**Decision Tree (DT)**
Decision Trees are a type of non-parametric model used in Machine Learning. Decision trees can be used for both classification and regression, meaning that decision trees are flexible models that do not increase their number of parameters if more features are added. DT are composed of two types of components: nodes and branches. Each data feature is treated at each node to partition the observations during the training phase or to make a single data point follow a certain path while producing a prediction.

**Random Forest (RF)**
Breiman [20] was the first to introduce the random forest model. It is a Supervised Machine Learning Algorithm that is widely used in classification and regression problems, combining the concepts of bootstrap aggregation and random subspace. The model is made of many independent decision trees and each of the trees is produced using a random process. First, bootstrap sampling was used to do many rounds of sampling. A subset of inputs is randomly chosen in each sampling cycle, and several decision trees may be trained. Finally, the RF model's forecast may be determined by voting or averaging the decision tree outcomes as illustrated in Fig. 1, where $R(x)$ is the average output of a total number of $n$, $r_i(x)$ is the result of $i$-th base learner trained with a subset of the dataset and randomly selected features.

**Extreme Gradient Boosting (XGB)**
By sequentially training a series of weak learners, the gradient boosting machine XGBoost uses a training approach to produce a strong learner. A gradient descent optimization algorithm is used in each step to train a weak learner to minimize the loss function. Recognized as a more sophisticated gradient boosting machine implementation, XGBoost makes use of a more regularized model generation to control over-fitting more successfully. XGB is a gradient boosting (GRAB) method that has been improved. The GRAB uses the first-order derivative for optimization, but the XGB expands the loss function in a second-order Taylor manner and uses both the first-order and second-order derivatives. Through the learning model, XGB can handle missing data and automatically choose the optimum default splitting direction.

**Light Gradients Boosting Machine (LGBM)**
LGBM is a recent ML technique built on Gradient Boosting Decision Trees. Microsoft

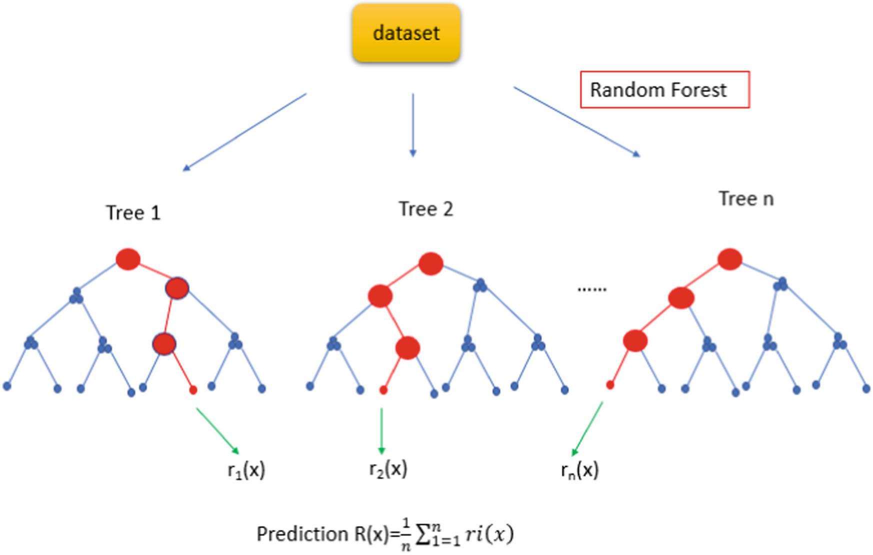


**Fig. 1.** Schematic view of Random Forest

introduced LGBM approach, which is similar to XGB. However, LGBM does not develop a tree level-wise (row by row/horizontally), unlike most Tree-based ML models. Instead, it uses the leaf-wise tree development strategy, which leads to vertical development. This indicates that it chooses and develops on the leaf with the greatest potential for loss reduction. The cost of making a poor prediction had been reduced by this construction strategy. By restricting its tree depth, LGBM also avoids overfitting. The large number of hyperparameters covered by LGBM, makes it harder to adjust, which is the only main disadvantage.

### 3.2 Statistical Indicators

To assess the precision of classification ML models, three metrics Precision, Recall and F1-score have been employed. All classification statistical indicators are based on the concept of True Positive (TP) when the predicted and measured value are 0, False Positive (FP) when predict value is class 0 and measured value is class 1, True Negative (TN) when predicted and measured values are 1 and False Negative (FN) when measured value is 0 and predicted value is 1. The metrics are expressed are expressed in Eqs. (1), (2) and (3):

$$Precision = \frac{TP}{TP+FP} \tag{1}$$

$$Recall = \frac{TP}{TP+FN} \tag{2}$$

$$F1-score = 2 \times \frac{recall \times precision}{recall + precision} \tag{3}$$

### 3.3 Classification Methodology Using Regression Models

Before comparing regression and classification models, it is necessary to build a classification methodology using the regression models. The methodology needs to respect the goals of the regression study, which are, in our case, the prediction of temporal expansion, and also comply with the classification objective. To this end, if the measured or predicted expansion exceeds 0.2%, the studied cementitious material is considered non-sulphate-resistant (NSR). Figure 2 illustrates the methodology adopted to classify

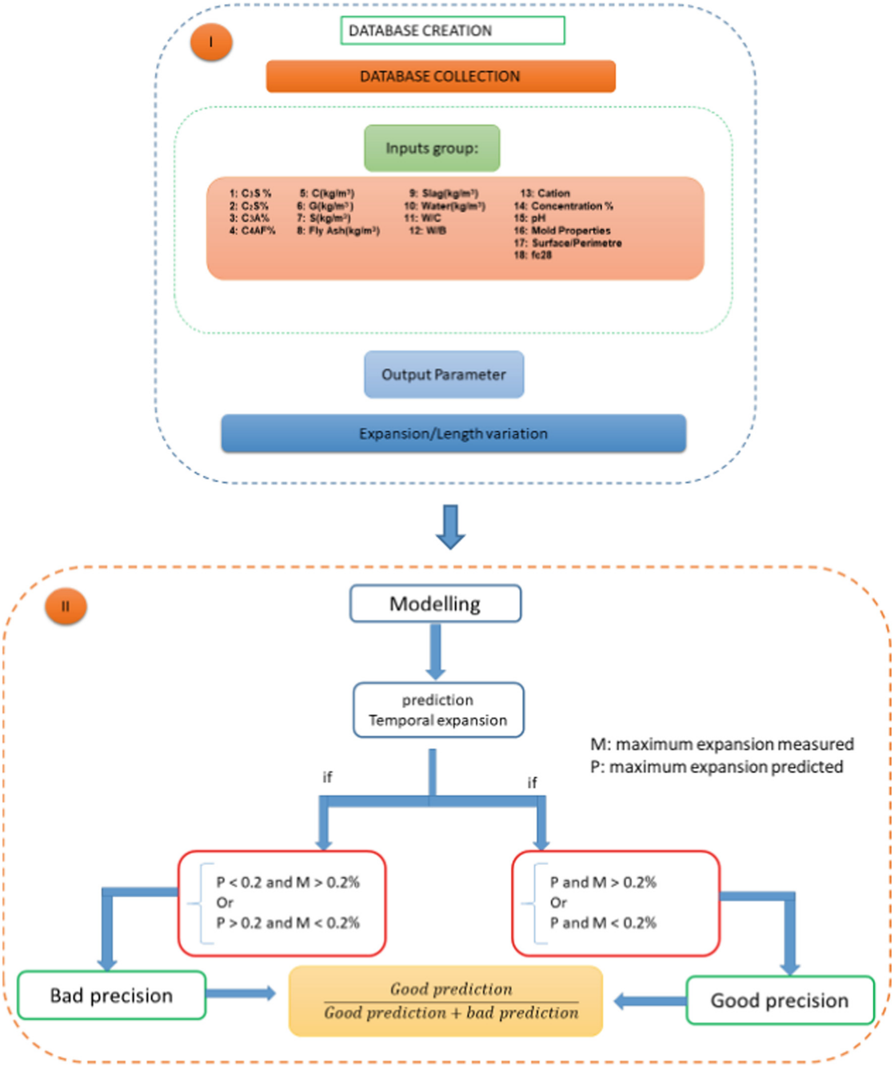


**Fig. 2.** Methodology flowchart for classification.

cements, starting by predicting the evolution of expansion over time before applying the classification rule and, finally, classifying cements.

The comparability between regression and classification models was ensured by using the same training and testing formulations.

# 4 Results

## 4.1 Classification Scores

The results of ML models were described using classification reports containing the evaluated aforementioned metrics and adding statistical treatment as micro average and accuracy.

Macro-average compute an average after calculating the metric regardless of the classes. The accuracy of a ML model indicates the proportion of times it will accurately predict a result out of all the predictions it has made. In our study, we presented only precision, recall and F1-score. These three metrics constitute the references to compare the robustness of the models. Table 2 shows the scores of each model. The difference between the scores is not very significant, for class 0 (SR) in which RF is the most precise model with F1-score equal to 86%, for class 1 which is more important since NSR is more problematic than SR. In this case LGBM had good results with F1-score around 78%. In both cases, all the models show close results, which confirm the high precision of chosen models. The comparability of results helps to determine the origin of prediction difficulties.

**Table 2.** Classification results.

| Models | | Precision | Recall | F1-score | Support |
|---|---|---|---|---|---|
| DT | 0 | 0.81 | 0.84 | 0.82 | 55 |
| | 1 | 0.78 | 0.74 | 0.76 | 42 |
| RF | 0 | 0.80 | 0.93 | 0.86 | 55 |
| | 1 | 0.88 | 0.69 | 0.77 | 42 |
| XGB | 0 | 0.80 | 0.87 | 0.83 | 55 |
| | 1 | 0.81 | 0.71 | 0.76 | 42 |
| LGBM | 0 | 0.82 | 0.89 | 0.85 | 55 |
| | 1 | 0.84 | 0.74 | 0.78 | 42 |

## 4.2 Confusion Matrix

The confusion matrix is a technique for evaluating the effectiveness of classification models with two or more classes representing the various combinations of actual and predicted values in the binary situation (i.e., with two classes, the simplest example). It will not only show which forecasts were right and wrong, but more importantly, it will show what kinds of mistakes were made. The predicted values of each class are reported in columns, and the actual values in rows.

Figure 3 shows the confusion matrix of four used ML models using the test set predictions. As expected, all used models had good results. The difference between the models is not significant. The best results were obtained using LGBM and DT, LGBM had correctly predicted 31/42 NSR cements and 49/55 SR cements. DT model had the same results as LGBM.

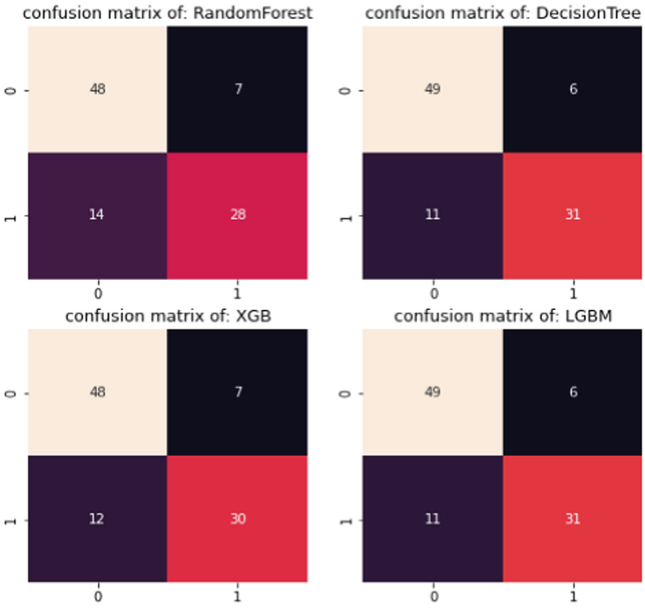


**Fig. 3.** Confusion matrix of classification models for test set.

### 4.3 Comparison Between Regression and Classification Models

As illustrated in Fig. 4, the best regression model found here was DT, with a score of 88.42%, and the best classification model was LGBM, with a score of 82%. The results of both techniques are satisfying. The comparison between these two types of techniques (Fig. 4) shows the effectiveness of regression models to classify cements regarding their resistance to ESA. The results of the regression models are thus more promising. However, the major drawback of this kind of model is the need for an extensive database with a precise description of the temporal evolutions of the expansions. In addition, it is worth noting that the reported performances could be further improved with hyperparameter optimization.

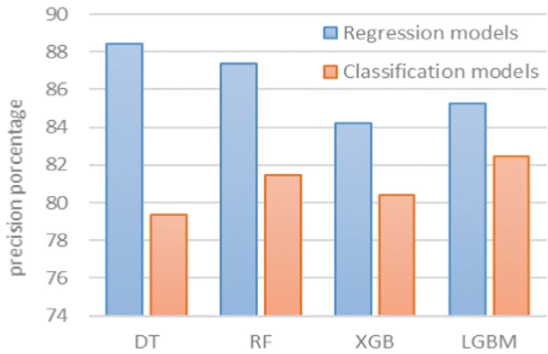


**Fig. 4.** Comparison between classification and regression ML models precision for ESA resistance classification problem.

### 4.4 Model Interpretability (LIME)

The interpretability of the ML models did not attract much attention, even if it is the origin of model predictions. To fully understand the forecast of concrete resistance, Local Interpretable Model-Agnostic Explanations (LIME) has been employed to highlight the underlying patterns governing predictions. LIME principle is based on randomly producing more features that are close to the pointed value. LIME will then calculate the prediction of these values after weighting each random value proportionally to its proximity to the target value. Finally, because LIME works at a microscopic level, it will provide a linear model that will help to understand feature impacts. Figure 5 shows an example of Lime table for randomly chosen cements from test data, 'good' correspond to SR and 'bad' correspond to class NSR.

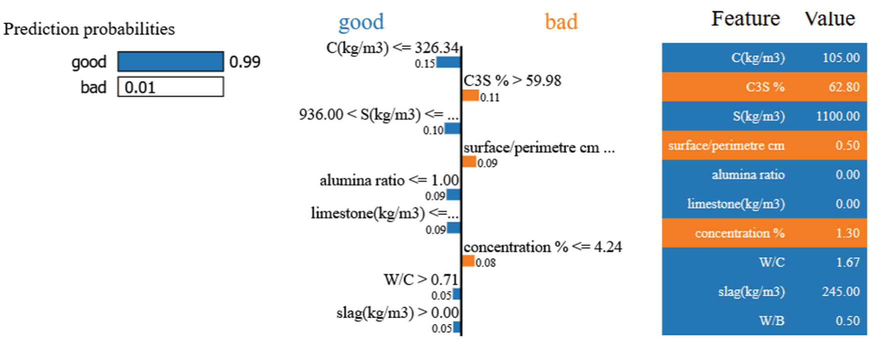


**Fig. 5.** Model interpretability and features impact.

As expected, the quantity of cement had been interpreted as the first indicator for ESA resistance. For the selected specimen, the small quantity of cement of 105 kg/m$^3$ due to the high slag replacement ratio justifies the resistance of this mortar. The proportion of $C_3S$ is the second most important indicator. The percentage of $C_3S$ in this specimen is relatively high 62.80%, which negatively affects sulphate resistance. On the other hand, the large quantity of slag improves the resistance since the durability of concrete is increased with large proportions of slag. Last but not least, geometrical properties and sulphate solution characteristics have been found to have a relatively important negative effect on this sample. Overall, LIME is an interesting tool to understand the mechanism of ESA and, more globally, interpret ML models in civil engineering.

## 5 Conclusions

This study compared classification and regression ensemble learning techniques for predicting cement sulphate resistance. Four ML algorithms DT, RF, XGB, and LGBM, were employed to analyze a database containing 483 specimens gathered from the literature. Cement composition, mixture proportions, mold properties, and sulphates solution characteristics were chosen as model inputs. Based on the obtained results, the following conclusions can be drawn:

(1) All ML models show good results, and the prediction of concrete resistance has been assessed with high precision.
(2) Comparable results can be obtained with the tested ML models, which confirms the consistency and the accuracy of these techniques.
(3) LGBM and DT had the best results and predicted 80/97 specimens from the test set with high precision.
(4) The introduced classification criterion helped obtain better results with the regression models than with the classification models.
(5) LIME technique was employed to analyse feature importance for each prediction, and, according to this analysis, cement proportion, aggregates and SCM are the most influential factors.

Even though concrete durability problems are intrinsically complicated because of the variety of phenomena at stake, this study opens up a novel research path for using Machine Learning techniques for cementitious materials classification regarding some of the most impactful environmental degradations.

## References

1. Thorvaldson, T.: Chemical aspects of the durability of cement products. In: Proceedings of the 3rd International Symposium on Chemistry of Cement, London, UK (1952)
2. Marchand, J., Skalny, J. (eds.): Sulphate Attack Mechanism. The American Ceramic Society, Westerville, USA (1999)
3. Skalny, J., Marchand, J., Odler, I. (eds.): Sulphate attack on concrete. Spon Press, London, UK (2002)
4. Zhang, G., Wu, C., Hou, D., Yang, J., Sun, D., Zhang, X.: Effect of environmental pH values on phase composition and microstructure of Portland cement paste under sulphate attack. Compos. Part B: Eng. **216**, 108862 (2021)
5. Wei, Y., et al.: Effect of fly ash on mechanical properties and microstructure of cellulose fiber-reinforced concrete under sulphate dry–wet cycle attack. Constr. Build. Mater. **302**, 124207 (2021)

6. González, M.A., Irassar, E.F.: Effect of limestone filler on the sulphate resistance of low C3A Portland cement. Cem. Concr. Res. **28**(11), 1655–1667 (1998)
7. Makhloufi, Z., Aggoun, S., Benabed, B., Kadri, E.H., Bederina, M.: Effect of magnesium sulphate on the durability of limestone mortars based on quaternary blended cements. Cem. Concr. Compos. **65**, 186–199 (2016)
8. Boudache, S., Rozière, E., Loukili, A., Izoret, L.: Towards common specifications for low- and high-expansion cement-based materials exposed to external sulphate attacks. Constr. Build. Mater. **294**, 123586 (2021)
9. Skalny J.: Sulphate attack issues: an overview. In: Jacques, M., Jan, S., (edn.): Materials science of concrete – sulphate attack mechanisms. The American Ceramic Society (1999)
10. Moon, H.Y., Lee, S.T., Kim, S.S.: Sulphate resistance of silica fume blended mortars exposed to various sulphate solutions. Can. J. Civ. Eng. **30**(4), 625–636 (2003)
11. Skalny, J., Pierce, J.: Sulphate attack issues, material science of concrete. In: Sulphate Attack Mechanism, American Ceramic Society, Westerville, OH (1999)
12. Cha, Y.-J., Choi, W., Büyüköztürk, O.: Deep learning-based crack damage detection using convolutional neural networks: deep learning-based crack damage detection using CNNs. Comput. Aided Civ. Infrastruct. Eng. **32**(5), 361–378 (2017)
13. Dorafshan, S., Thomas, R.J., Maguire, M.: Comparison of deep convolutional neural networks and edge detectors for image-based crack detection in concrete. Constr. Build. Mater. **186**, 1031–1045 (2018)
14. Liu, Z., Cao, Y., Wang, Y., Wang, W.: Computer vision-based concrete crack detection using U-net fully convolutional networks. Autom. Constr. **104**, 129–139 (2019)
15. Chen, P., Zheng, W., Wang, Y., Chang, W.: Creep model of high-strength concrete containing supplementary cementitious materials. Constr. Build. Mater. **202**, 494–506 (2019)
16. Hilloulin, B., Tran, V.Q.: Using machine learning techniques for predicting autogenous shrinkage of concrete incorporating superabsorbent polymers and supplementary cementitious materials. J. Build. Eng. **49**, 104086 (2022)
17. Wu, X., et al.: Prediction of the frost resistance of high-performance concrete based on RF-REF: a hybrid prediction approach. Constr. Build. Mater. **333**, 127132 (2022)
18. Hilloulin, B., Hafidi, A., Boudache, S., Loukili, A.: Interpretable Ensemble Machine Learning for the prediction of the expansion of cementitious materials under external sulphate attack, submitted (2022)
19. Li, Z., Yoon, J., Zhang, R., Rajabipour, F., W.V. Srubar III, I. Dabo, A. Radlińska.: Machine Learning in Concrete Science: Applications, Challenges, and Best Practices . Npj Computational Materials (2022)
20. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)