Interpretable ensemble machine learning for the prediction of the expansion of cementitious materials under external sulfate attack

Benoit Hilloulin, Abdelhamid Hafidi, Sonia Boudache, Ahmed Loukili

Interpretable ensemble machine learning for the prediction of the expansion of cementitious materials under external sulfate attack

Benoît Hilloulin, Abdelhamid Hafidi, Sonia Boudache, Ahmed Loukili

PII: S2352-7102(23)02131-9

DOI: https://doi.org/10.1016/j.jobe.2023.107951

Reference: JOBE 107951

To appear in: *Journal of Building Engineering*

Received Date: 21 August 2023

Revised Date: 7 October 2023

Accepted Date: 14 October 2023

Please cite this article as: Benoî. Hilloulin, A. Hafidi, S. Boudache, A. Loukili, Interpretable ensemble machine learning for the prediction of the expansion of cementitious materials under external sulfate attack, *Journal of Building Engineering* (2023), doi: https://doi.org/10.1016/j.jobe.2023.107951.

Interpretable Ensemble Machine Learning for the prediction of the expansion of cementitious materials under external sulfate attack

Benoît Hilloulin*, Abdelhamid Hafidi, Sonia Boudache, Ahmed Loukili

Nantes Université, Ecole Centrale Nantes, CNRS, GeM, UMR 6183, F-44000 Nantes,France – e-mail: benoit.hilloulin@ec-nantes.fr

*Corresponding author: Benoit.Hilloulin@ec-nantes.fr

Abstract

External sulfate attack (ESA) is a key degradation mechanisms of cementitious materials. Although the advantages of low-C_3A cement and supplementary cementitious materials (SCM) have been confirmed, there remains a need for a better understanding of the phenomenon and guidance on accelerated testing due to the numerous parameters affecting this degradation. This study introduces a machine learning framework for predicting the expansion of cementitious materials incorporating SCM because of ESA. A comprehensive database is constructed, and four optimized machine learning models are compared. Among them, extreme Gradient Boosting (XGBoost) showed the best performance with a R^2 accuracy of 0.933 and 0.788 on the training and the test set resp. Additionally, SHapley Additive exPlanations (SHAP) enabled the identification of the most influential inputs and their relative influence. It has been found that clinker composition, mix proportion, sample geometry, and sulfate solution characteristics play an important role, with their relative contribution being 34%, 36%, 3% and 27% resp. Furthermore, a thorough analysis of the model predictions on some expansive and non-expansive mortar and concrete samples demonstrated its reliability. Finally, the model was shown to be able to accurately predict the time required to reach a given expansion.

Keywords: Concrete; Sulfate attack; Expansion, Machine Learning; SHAP.

1. Introduction

Among all the degradations that can happen to cementitious materials, the external sulfate attack is one of the most studied and documented. The attack mechanism can be summarized as follows: sulfates in solution progress into the cement matrix by diffusion and react with ions in the

pore solution to form expansive products. The most common expansive product is the aluminosulfate phase ettringite [1]. According to the crystal pressure theory [2], secondary ettringite forming from a supersaturated pore solution into the cement matrix nanoporosity will lead to expansion and, after that, cracking [3,4].

Many lab tests have been performed to highlight the parameters influencing ESA. Those parameters can be related to the sulfate solution, the conditions of exposure, or the cementitious material properties. First, the type of cation and the sulfate concentration of the sulfate solution directly impact the products formed during the attack. The magnesium reacts with hydrates from the cement paste to form brucite and M-S-H that replace C-S-H during magnesium sulfate attack [5]. Gypsum will form over ettringite when the sulfate solution has a sulfate concentration of over 30 g/L [6,7]). A high sulfate concentration accelerates the diffusion of sulfates. The supersaturation regarding ettringite is reached faster: the higher the sulfate concentration, the shorter the response time [8]. Secondly, exposure conditions such as pH and temperature also affect the kinetic of the attack. A controlled pH of 7 spurs the $Ca(OH)_2$ leaching [9] and subsequently provides Ca^{2+} to the pore solution that can react with sulfate to form expansive products [10]. While high temperatures induce a shortening in response time, temperatures below 5 °C lead to the formation of calcium sulfate carbonate phase, e.g., thaumasite, as a product of ESA [11].

While sulfate solution and exposure conditions greatly influence the ESA mechanism, most of the chemical elements reacting with sulfates are brought by the cement matrix. The clinker C_3A content is considered the most important aluminates source. Thus CEM I cement with low C_3A content are considered sulfate-resistant cements (SR0, SR3, SR5) [12]. SCM such as slag, fly ash or pozzolans consume portlandite by pozzolanic reactions and form C-A-S-H besides the C-S-H formed by C_3S and C_2S hydration, lessening the gypsum formation and ettringite recrystallization [13]. Recently, the positive effect of calcined clay has been reported [14,15]. The fine porosity of C-A-S-H slows the sulfate diffusion through the cementitious matrix and has a positive effect on the resistance to ESA. Last, diffusion being controlled by porosity, the water/cement ratio greatly affects the sulfate resistance of a sample [16]. However, the behavior of the ternary or quaternary cement blends with high substitution ratios remains an open question.

The collected information helps set up more efficient and representative lab tests. One of the most debated points is the duration of the test because of the number of parameters influencing the time to reach a given expansion. As stated before, the formation of expansive products will depend

on the sulfate concentration in the pore solution; the sulfate diffusion is thus the limiting step. Tests may require several months before giving significant results. Thus, to address this issue, some lab tests are designed to accelerate the degradation process. While they may be effective, it is difficult to say if they are accurate to the ESA mechanism [17]. Moreover, the relative importance of the various parameters involved in ESA is still debated. For example, mortar and concrete mix compositions and samples' geometrical features play a decisive role in the volumetric expansion measured in laboratory tests, influencing the final classification of types of cement regarding their sulfate resistance.

In recent years, Machine Learning has been increasingly employed for predicting and analyzing cementitious materials' properties. Deep Learning techniques and Convolutional Neural networks help assess concrete properties at various scales: from crack and defect detection [18–20] to concrete microscopic image analysis [21–23] or mechanical properties [24]. Gaussian processes, Bayesian techniques, and exploration-exploitation techniques close to reinforcement learning have been successfully employed to infer mechanical characteristics from microindentation and nanoindentation [25] or quantitatively estimate uncertainties concerning concrete properties such as susceptibility to sulfate degradation [26]. However, these techniques are relatively limited in terms of interpretability. For this reason, supervised learning models have been further developed as they can be accompanied by mature interpretability tools in order to gain insights about the most influencing parameters governing a phenomenon. Numerous research works have been published to predict concrete compressive strength [27–30], fresh properties [31], creep [32,33], shrinkage[34–36], chloride [37], carbonation resistance [38], frost resistance [39]. Advances such as features analysis and dubbed SHapley Additives exPlanations (SHAP) [40] introduced novel ways to explore feature impact, and it has been shown that machine learning can provide similar or better predictions than analytical models [36]. However, to the authors' knowledge, no study has been reported to predict the expansion of mortar and concrete samples due to external sulfate attack using machine learning, even though such models would help identify the major influencing parameters, the role of SCM, and likely help provide guidelines relative to the laboratory tests definition towards a quick and representative expansion assessment.

This study provides insight into the potential of machine learning models based on conventional or ensemble techniques to predict the expansion of cementitious materials, eventually incorporating supplementary cementitious materials due to the external sulfate attack. A database

has been specifically built based on the available literature. The theory and procedures associated with the models are briefly presented in the manuscript. Then, the results of the models are discussed, and the best model candidate is further examined using SHapley Additive exPlanation (SHAP) theory to understand the most influential features and derive partial difference plots. Finally, a comparison is made between the experimental and predicted time to reach specific expansions.

2. Database description and initial processing

2.1 Database construction

Observations of linear expansions of samples completely immersed in sulfate solutions, also called length variations, were selected from different sources in the literature [8,13,14,17,41–64]. These studies reported expansions of various mortar and concrete specimens with sulfate-resistant and non-sulfate-resistant cement from standard to high-strength mixes. Only mixes based on CEM I cements, eventually with SCM (fly ash, slag, pozzolan, limestone, silica fume and calcined clay, abbreviated with MK in Table 1), were selected because of the lack of information about some standardized blended cement (classified as CEM II to CEM V according to Eurocodes). In total, 336 mortar and concrete expansion curves were obtained and used to interpolate expansion values at increasing ages relative to the square root of time due to the diffusive nature of the process, e.g., $1\sqrt{d}$, $2\sqrt{d}$, $3\sqrt{d}$, …, until the end of each corresponding measurement. Then, after the interpolation and the filtering steps, 5294 expansion data points were generated. Only positive expansions smaller than 0.4% were considered in order to limit the influence of external phenomena on the results, such as leaching or extensive cracking. Some curves were not included because of excessive or very rapid unexplained expansions. No further cleaning or filtering was applied.

Four types of inputs parameters were considered: clinker composition (C_3S, C_2S, C_3A and C_4AF), mix proportion and characteristics (cement mass, aggregate-to-cement ratio, water-to-binder ratio and SCMs proportions, 28-day compressive strength reflecting porosity which is rarely reported), sample geometry (shape and surface-to-perimeter ratio), and sulfate solution and environment characteristics (cation type, concentration, pH, and temperature). Categorical values such as cation type and mold properties were encoded to be used as inputs in the model. We attributed the value 1 to Na, the value 0 to Mg, and, concerning the mold shape, we attributed the value 0 for prismatic specimens and 1 for cylindrical specimens.

2.2 Imputation of missing values

2.2.1 Compressive strength inference using XGBoost

Since porosity and strength-related properties might have an influence on expansion during the external sulfate attack process and considering that strength is reported more often than porosity [63], a dedicated XGBoost model has been used to input missing 28-day compressive strength data, amounting to 44% of the database (148 formulations), based on the cement composition and the mix formulation as described in the previous section. The choice has been made to demonstrate the suitability of the present database without the need for an external database. The model has been trained using 75% of the 188 experimental compressive strengths available in the original database and tested on the remaining 25%. Between 20 and 80 MPa, the 28-day cubic compressive strength of the test samples was estimated with an R^2 value of 0.83, which is better than the classic mean inference or regression methods. All missing strength values were thus inferred using the dedicated XGBoost model.

2.2.2 Other missing values

Two different methods were used to infer other missing values [17-19]. First, the univariate imputation consists of filling in the missing values by statistical characteristics of the existing data (median, most frequent, mean). This simple approach is used for the missing oxides in the cement composition. The second method, relying on the physical aspects, is used for pH. Due to leaching phenomena, the missing values of uncontrolled pH are filled by 10. Last, missing C_3S, C_2S, C_3A, and C_4AF values were calculated using Bogue's equation based on oxides compositions.

2.3. Final database description

After inferring all missing values, the final database containing 5294 expansion values relative to 21 parameters was obtained. A description of the database is given in Table 1, and the cumulated distribution functions associated with most of the inputs are given in Fig. 1. Mean and median values have been reported, as well as minimum and maximum values. The database covers a wide range of cement compositions, mortar and concrete formulations with various W/B ratios, SCM types, and sulfate solution compositions and environmental conditions. No particular data imbalance can be observed. However, temperature values were almost all equal to 20 or 23°C and nanosilica content equal to 0%, their effects cannot thus be analyzed in-depth.

Correlations between the input variables were calculated before the machine learning algorithms' application in order to avoid excessive correlations between variables. The correlation matrix is given in Fig. 2. As expected, some formulation parameters (water-binder ratio, water-binder ratio, aggregate-cement ratio, and cement content) were particularly correlated. Apart from these correlations concerning formulations, no other significant correlation was observed.

Table 1. Description of the database used in this study

	mean	std	min	25%	50%	75%	max
C_3S (%)	57.7	5.9	34.0	54.95	56.66	61.0	73.5
C_2S (%)	14.9	6.0	2.29	12.0	14.57	18.1	39.0
C_3A (%)	8.0	2.6	0.9	5.8	7.6	10.1	11.9
C_4AF (%)	9.4	2.7	0.0	7.8	9.7	10.98	19.7
Cement (kg/m^3)	355.4	129.5	55.0	269	375	469	553
A/C	4.36	1.9	1.64	3.0	3.9	5.2	15.4
W/B	0.5	0.08	0.33	0.48	0.5	0.55	0.77
Fly Ash (%mb)	3.6	25.8	0.0	0.0	0.0	0.0	50.0
Slag (%mb)	14.3	23.6	0.0	0.0	0.0	20.0	85.0
Pozzolan (%mb)	0.25	3.5	0.0	0.0	0.0	0.0	50.0
Limestone (%mb)	2.9	8.6	0.0	0.0	0.0	0.0	50.0
MK (%mb)	2.37	7.0	0.0	0.0	0.0	0.0	34.0
Silica Fume (%mb)	0.29	1.6	0.0	0.0	0.0	0.0	13.0
Nanosilica (%mb)	0.30	1.2	0.0	0.0	0.0	0.0	8.0
Cation	0.5	0.5	0.0	1.0	1.0	1.0	1.0
Concentration %	5.7	5.1	0.3	3.0	5.0	5.0	20.0
pH	9.0	1.8	3.0	7.5	10.0	10.0	12.3
Temperature (°C)	19.5	4.7	1.0	20.0	20.0	20.	35.0
Mold properties	0.04	0.2	0.0	0.0	0.0	0.0	1.0
Surface/perimeter (cm)	1.0	0.6	0.25	0.625	0.625	1.25	2.5
fc28 (MPa)	49.4	13.4	20.6	41.5	47.9	54	100

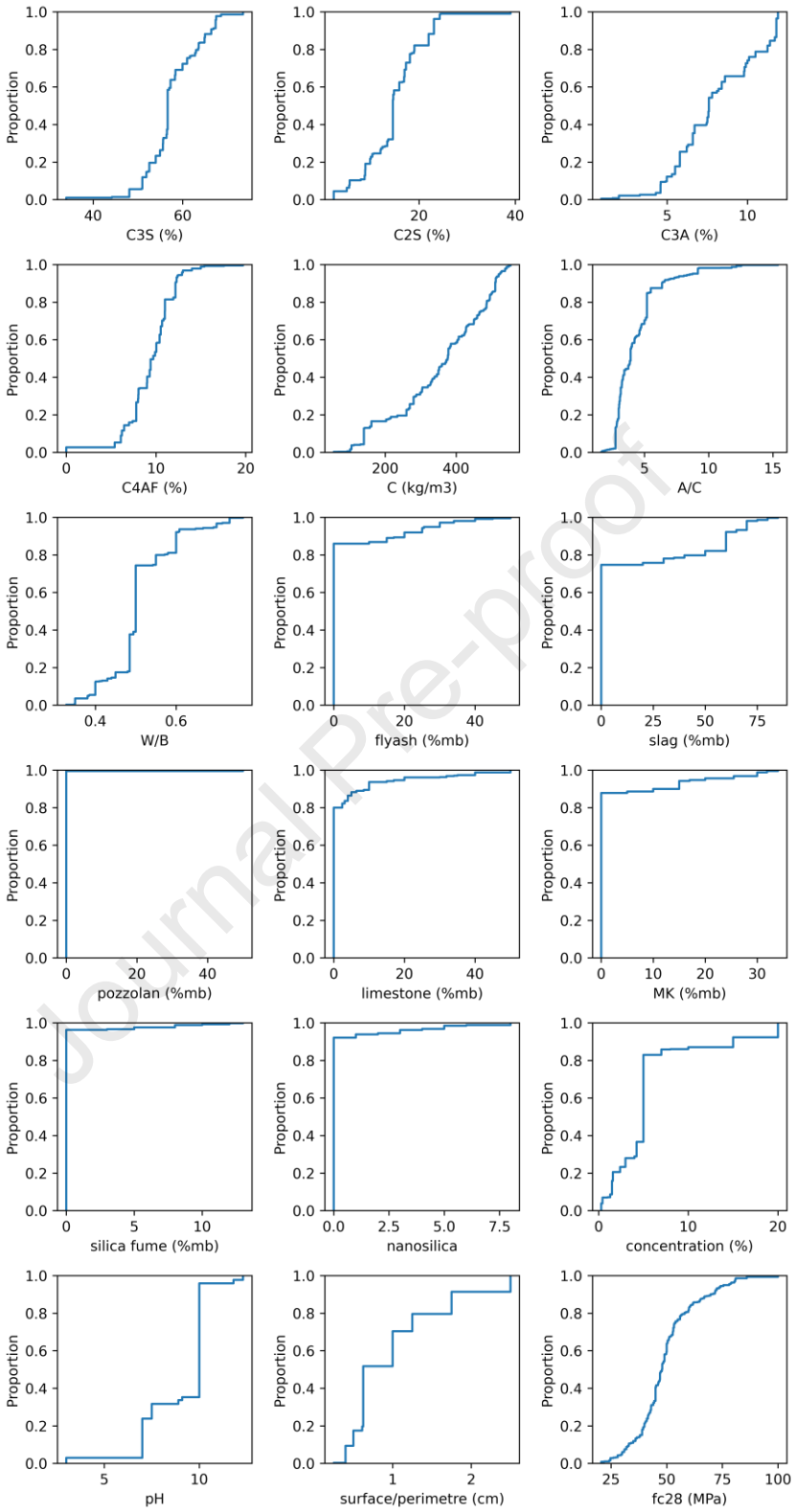


Fig 1. Experimental cumulative distribution functions of the inputs

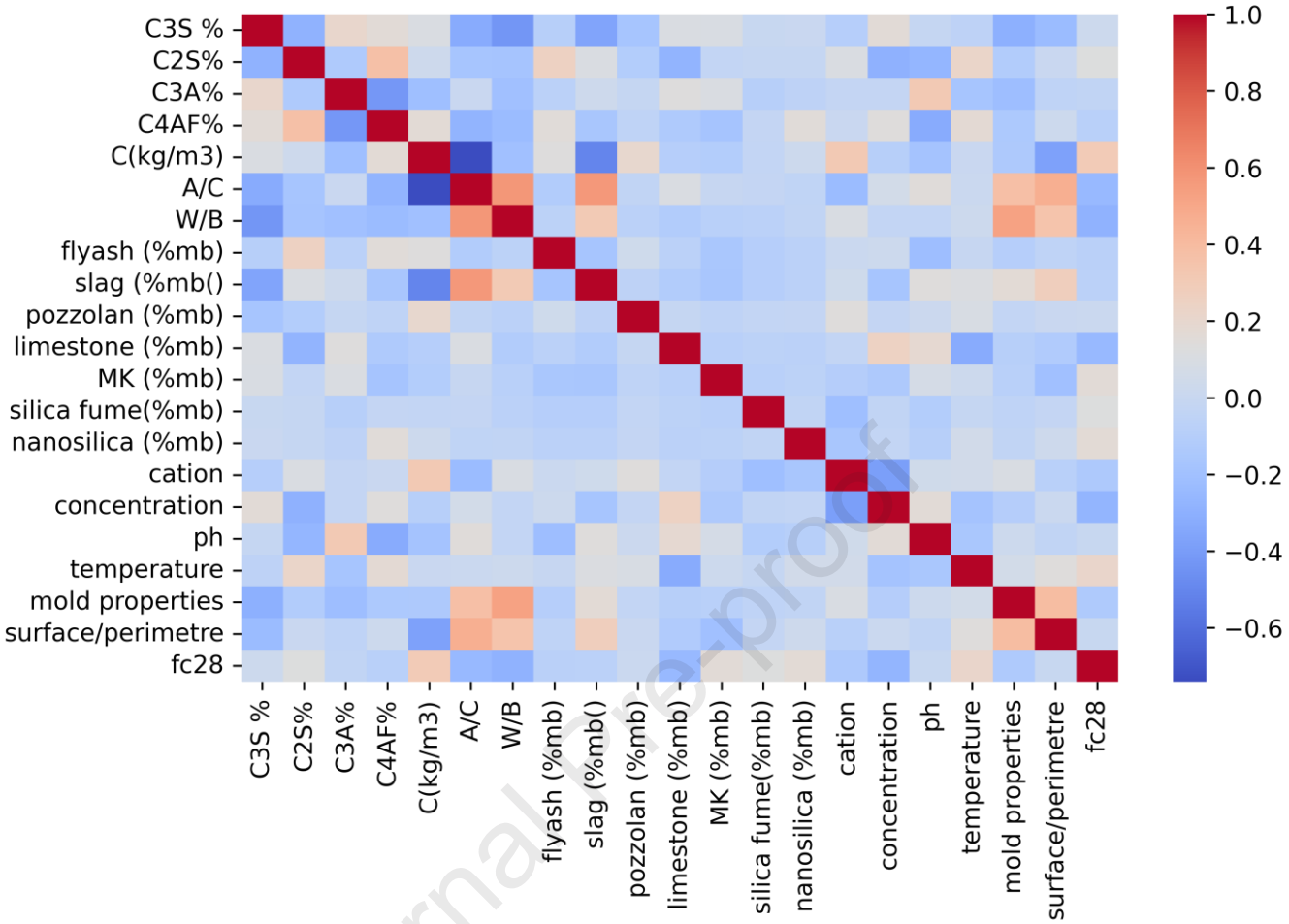


Fig. 2. Correlation matrix of the input variables

3. Methods

3.1. Machine learning models

3.1.1 Linear regression

Regression analysis was first provided by Francis Galton in the second half of the 19th century. Regression analysis is a statistical approach that uses the relation between quantitative variables, so that a simple linear curve can predict a result or response variable [65]. Regression models have only linear parameters; therefore, the predicted variables are linear. The linear curve is constructed to have a lesser error between the variables and the curve. Linear regression (LR) is used as a benchmark model in this study to highlight the benefits of using ensemble models.

3.1.2 Decision trees

Decision Trees (DT) are basic estimators and non-parametric models used in Machine Learning. DT comprise two types of components: nodes and branches [66]. Each one of the data's features is examined at each node. For this reason, DT are flexible models that do not increase their number of parameters when adding more features. The internal nodes indicate an attribute test, and each branch and leaf represent the test result and the class tag, respectively. These nodes come in three different categories, and each one has a distinct geometric shape, such as a circle, rectangle, or triangle. DT is a simple ML learning model in term of interpretability and constitute the element piece of more advanced ensemble models described hereafter.

3.1.3 XGBoost

XGB is known as an upgraded gradient boosting machine implementation, that uses a more regularized model generation to control over-fitting more successfully using Friedman's gradient boosting method [67]. The prediction is made by using several additive functions:

$$Y_{ik} = Y_{ik+1} + \mu f_k \tag{1}$$

Where Y_{ik} is the predicted value of ith iteration, f_k is an estimator corresponding to a tree structure, Y_{i0} is the mean of predictions of training dataset, μ is the learning rate, facilitating steady model improvement while including new trees and preventing overfitting. It is important to remember that overfitting is the main issue with all ML models. At the k step, kth estimator is added to the model, and the estimation of Y_k can be done according to equation (2),f_k can be determined by minimizing the following objective function:

$$objective = \lambda \mathrm{T} + \sum_{j=0}^{T} [E_j \omega + \frac{1}{2}(F_j + \gamma) w_j^2] \tag{2}$$

Where T represents the total number of leaves in the kth decision tree and w_j are the weights of each leaf. λ and γ are regularization parameters that control the simplicity of the tree structure to reduce overfitting. E_j and F_j are the sums of the samples associated with the jth leaf of the first and second gradients of the loss function, respectively.

3.1.4 Light Gradient Boosting (LGBM)

Light Gradient Boosting Machine (LGBM) is a new Gradient Boosted Decision Tree-based algorithm for machine learning [68]. It was originally introduced by Microsoft and is very similar

to XGB model. However, unlike most other implementations, LGBM does not grow a tree level-wise (row by row/horizontally). Instead, it implements the leaf-wise tree growth method (it grows vertically). This means that it selects the leaf that will have a maximum decrease in loss and grows on it. This building approach lowers the penalty for a wrong prediction. LGBM can also avoid over-fitting by limiting its tree depth. The main disadvantage of LGBM is that it covers many hyperparameters, making it harder to tune.

LGBM was created as a significant rival to XGBoost to increase training speed, use less memory, and retain excellent accuracy. The primary distinction between LGBM and XGBoost is how the trees are grown, as illustrated in Fig.3.

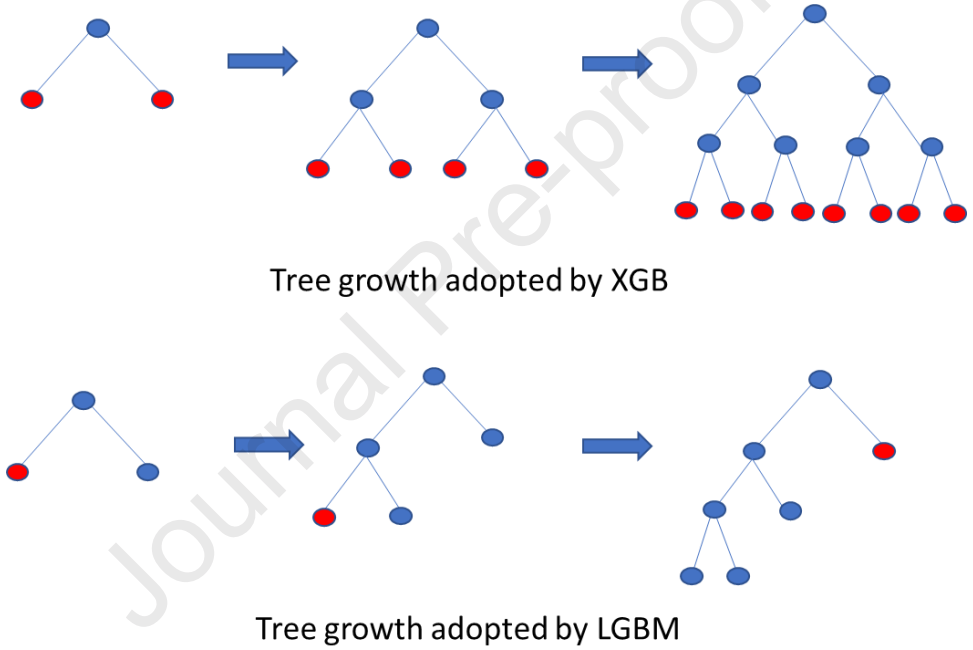


Fig. 3. Different ways for growing trees between LGBM and XGB [69].

3.2. Hyperparameters optimization

A Bayesian-based hyperparameter optimization algorithm has been employed, namely the Tree-structured Parzen Estimator (TPE), which is a Sequential Model-Based Method [70]. The interest of TPE algorithm lies in its reduced runtime and the better scores of the optimized models on the final test set as compared to other optimization algorithms, especially Random Search or manually-based optimization. TPE algorithm can be briefly described as an exploration-exploitation algorithm that looks to optimize the expected improvement (EI) function, defined in eq. 12, at each iteration using a surrogate loss function and selecting the best couple of hyperparameters within a given search space.

$$EI_{y^*}(x) = \int_{-\infty}^{\infty} \max(y^* - y, 0)\, p(y|x) dy \quad (3)$$

Where y* is a target performance, i.e. threshold value of the loss, e.g. objective, function, x is the proposed set of hyperparameters, y is the actual value of the loss using hyperparameters x, and p(y | x) is the surrogate probability model expressing the probability of y given x. Maximizing the Expected Improvement with respect to x means finding the best hyperparameters under the surrogate function p (y | x).

For TPE, p(y | x) is approximated using Bayes' rule:

$$p(y|x) = \frac{p(x|y) \times p(y)}{p(x)} \quad (4)$$

Where p(x|y), which is the probability of the hyperparameters given the score on the objective function, in turn, is expressed considering a split according to two different distributions for the hyperparameters: one where the value of the objective function is less than the threshold, l(x), and one where the value of the objective function is greater than the threshold, g(x):

$$p(x|y) = \begin{cases} l(x), if\ y < y^* \\ g(x), if\ y \geq y^* \end{cases} \quad (5)$$

The EI to maximize is then proportional to g(x)/l(x) that is to be minimized:

$$EI_{y^*}(x) \propto \left(\gamma + \frac{g(x)}{l(x)}(1-\gamma)\right)^{-1} \quad (6)$$

Where γ is the quantile of search result: $\gamma = p(y < y^*) = \int_{-\infty}^{y^*} p(y) dy$.

The hyperparameters search spaces of the four models used in this study are given in Table 2.

Table 2. Tuning ranges of hyperparameters.

LR	No hyperparameter				
DT	Max depth [2, 100]	Min samples split [1, 30]	Min samples leaf [1, 20]	Min weighted fraction at leaf node [0, 0.5]	
XGB	Nb of trees	Learning rate	Max depth	Min child weight	Subsample ratio

	[10, 1000]	[0.005, 0.30]	[2, 100]	[1, 30]	[0.8, 1]
LGBM	Nb of trees	Learning rate	Max depth	Min child weight	Subsample ratio
	[10, 1000]	[0.005, 0.30]	[2, 100]	[1, 30]	[0.8, 1]

3.3. Results analysis

3.3.1 Performance evaluation

To calculate the scores of the models, three metrics indexes are used: mean absolute error (MAE), coefficient of determination (R^2) and root mean square error (RMSE), which can be expressed as follows:

- Mean absolute error (MAE):

Mean absolute error is defined as the mean of differences between the predicted values and the experimental values in the data, MAE is expressed as $MAE = \frac{1}{N}\Sigma|ei - \hat{e}_i|$ where e_i is the value of expansion of i-th point in the database, $\hat{e}_i$ is the predicted value given by the models i-th sample point.

- Root mean square error (RMSE):

The error of a model in predicting quantitative data is often measured using the Root Mean Square Error (RMSE). It is officially defined as $RMSE = \sqrt{\frac{1}{N}\Sigma(e_i - \hat{e}i)^2}$

- Coefficient of determination (R^2)

The coefficient of determination is a statistical measurement that looks at how variations in one variable may be explained by changes in a second variable. R^2 is expressed as $R^2 = 1 - \frac{\Sigma(e-\hat{e}_i)^2}{\sum(ei-\bar{e})^2}$ where $\bar{e}$ is the averaged value of expansion. Both MAE and RMSE can provide an accurate assessment of model performance by clearly describing the residual error at each sample point. In contrast, R^2 creates a dimensionless score that ranges from 0 to 1 by normalizing the squared residual error with the database variance, in this study we worked with a logarithm of expansion, that is why the statistical indicators were adapted to this transformation.

3.3.2 Model interpretation and features importance using SHAP

Although several ML-based investigations in solid materials have successfully predicted their outputs with high accuracy, the interpretability of the ML models has received little attention. SHAP reveals the underlying pattern that the database's EML models show, which can give a thorough insight into the prediction of expansion behavior [71]. It is a means to determine how a

feature will affect the value of the target variable. The key idea is that the influence of features depends on the full set of characteristics in the database rather than just one particular feature. Therefore, SHAP retrains the model through all the combinations of features that contain the one we are investigating to determine the influence of each feature on the expansion, an indicator of a feature's significance is the average absolute magnitude of its effect on the output. The Shapley value identifies the relative importance of each trait. The approach developed by SHAP for comprehending model predictions may be used to understand even the most complex models.

3.3.3 Prediction of time to reach specific expansions

The time to reach 0.2% expansion was calculated with a $1\sqrt{d}$ precision for the test specimens based on the model outputs to evaluate the model prediction capability, similar to experimentally-measured time to failure [72]. For this purpose, the expansions of a given specimen were predicted from 1 day to 1000 days using the optimized model. The time to reach 0.2% expansion was taken as the first age at which the specimen's expansion reached this threshold value or a default value of 1000 days if the threshold was not reached. Two cases could thus be distinguished: i) non-sulfate-resistant cementitious materials that achieved an expansion higher than 0.2%. For these samples, the ML models predicted time to reach this expansion was compared with the experimental values; ii) sulfate-resistant cementitious materials. In this case, the default time of 1000 days was compared to the experimentally measured values, which were eventually set to 1000 days if no expansion had been observed.

3.4. Methodology flowchart

The investigation design is recapitulated in the methodology diagram presented in Fig. 4. Four crucial steps, which have been previously described, can be highlighted: step (I) database creation and description, step (II) predictions of fc28 and other missing values, step (III) selection of the optimal ML model with the highest performance, and step (IV) expansion prediction and sensitivity analysis using SHAP. Database creation was based on literature studies, 18 parameters were selected as the model inputs, and the expansion was the output. In step II, fc28 prediction was done to complete the database and prove the performance of chosen ML models. In step III, the database is randomly split into 20% for testing and 80% for training. LR, DT, XGB, and LGBM were trained, and their hyperparameters optimized. The performance of ML models was assessed using R^2, RMSE, and MAE. The best ML model was then used to predict concrete expansions,

and the influencing inputs on the output expansion were evaluated using SHAP.

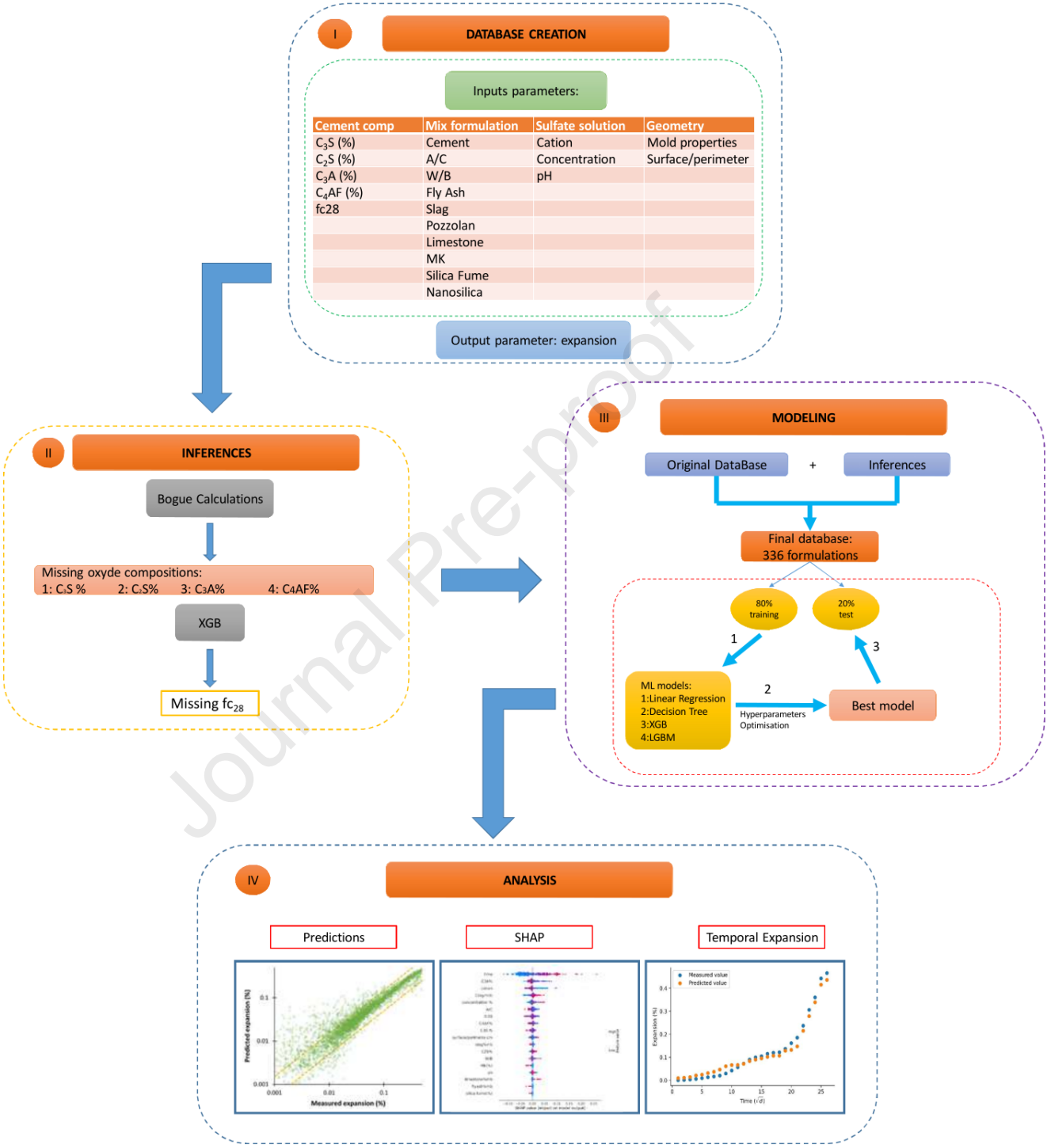


Fig. 4. Methodology flowchart of the study

4. Results and discussion

4.1. Hyperparameters optimization

Tree-structured Parzen Estimator leads to a fast optimization hyperparameters tuning, as illustrated in Fig. 5. It can be seen from this figure that the algorithm RMSE on the five validation sets quickly decreased in around 20 iterations. The decrease in RMSE between the first model evaluated and the optimized model is around 0.001 for LGBM, 0.004 for XGB, and more than 0.009 for DT model, highlighting the potential of this optimization procedure in this situation, in contrast with longer classical Bayesian optimizations and largely longer Random Search. The best mean result on the five validation sets after optimization has been obtained by the XGB model exhibiting an RMSE of 0.0515. The optimized hyperparameters of the DT, the XGB, and the LGBM models are reported in Table 3.

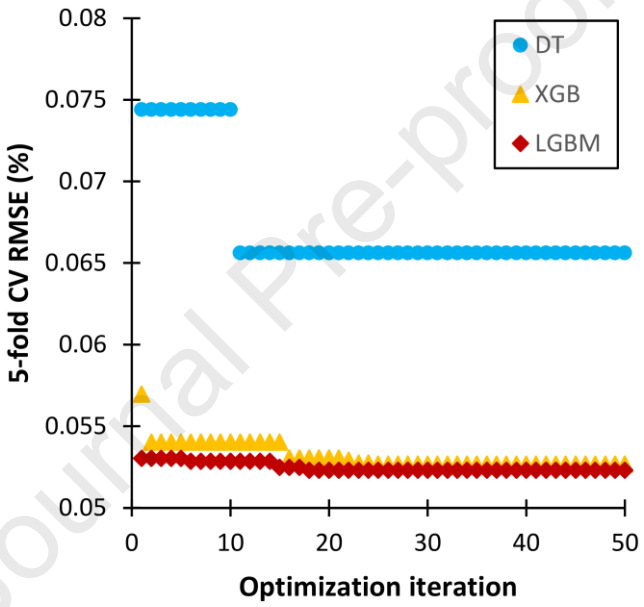


Fig 5. Optimization history using Tree-structure Parzen Estimator.

Table 3. Tunned hyperparameters of the ML models.

LR	No hyperparameter				
DT	Max depth	Min samples split	Min samples leaf	Min weighted fraction at leaf node	
	11	6	13	0.02	
XGB	Nb of trees	Learning rate	Max depth	Min child weight	Subsample ratio
	200	0.08	52	28	0.841
LGBM	Nb of trees	Learning rate	Max depth	Min child weight	Subsample ratio
	590	0.215	41	29	0.973

4.2. Sulfate attack expansion predictions

The prediction capacity of the optimized models on both the training and the test datasets are illustrated in Fig. 6 and Fig. 7 resp. It can be observed that, after optimization, the models performed well on the training set, especially for relatively large values of expansions higher than 0.05-0.1%. Though the LR model showed relatively poor performance in some cases, which can be attributed to its low complexity, the other three models could provide predicted expansions close to the measured values in most cases. Indeed, a high proportion of the predicted values lie within a 50% interval compared to the experimentally measured values, as illustrated by the dashed lines. This is a good result considering the relatively low precision of some reported results since most of the studies generally aim to compare different mortar formulations with one non-sulfate resistant formulation reducing the other samples' expansions resolution. Only a tiny number of expansions higher than 0.2% were incorrectly predicted. These values, which are often associated with formulations containing both SCM, represent a small proportion of the dataset due to the lack of data in the literature. It is worth noting that these values can be either underestimated or overestimated. The latter case is more favorable for building a secure model that might be looked for in the future. Overall, from the graphs, it can be concluded that the LGBM and XGB models performed the best on the training dataset, followed by DT model, while LR performed relatively poorly.

The comparisons between the predicted and measured expansion values from the test set are illustrated in Fig. 7. It can be observed that the predictions still match relatively well the experimentally measured expansions in the case of entirely unknown samples. Based on the visual spread of the predicted values, it can be concluded that XGB and LGBM models performed the best. However, these models tend to slightly overestimate some values, especially for low expansions.

Typical time evolutions of the expansions predicted by the optimized XGB are reported in Fig. 8 and the corresponding mortar and concrete formulations from the test set are given in Table 4. As illustrated in all subfigures, and in agreement with the general results detailed in the paragraph above, most predictions are close to the measured values, even in the case of irregular, probably noisy, time evolutions. In most cases, the model was able to reproduce the two-stage expansion with a limited expansion increase at the beginning, and then a sudden expansion

increase due to the formation of macro-cracks in the specimens. Moreover, the time to reach specific expansions such as 0.05% and 0.1% limit guidelines has been predicted with a relatively good precision of around $2\ to\ 3\sqrt{days}$ in most cases. Some rare samples were incorrectly predicted as mentioned above, which opens rooms for improvement.

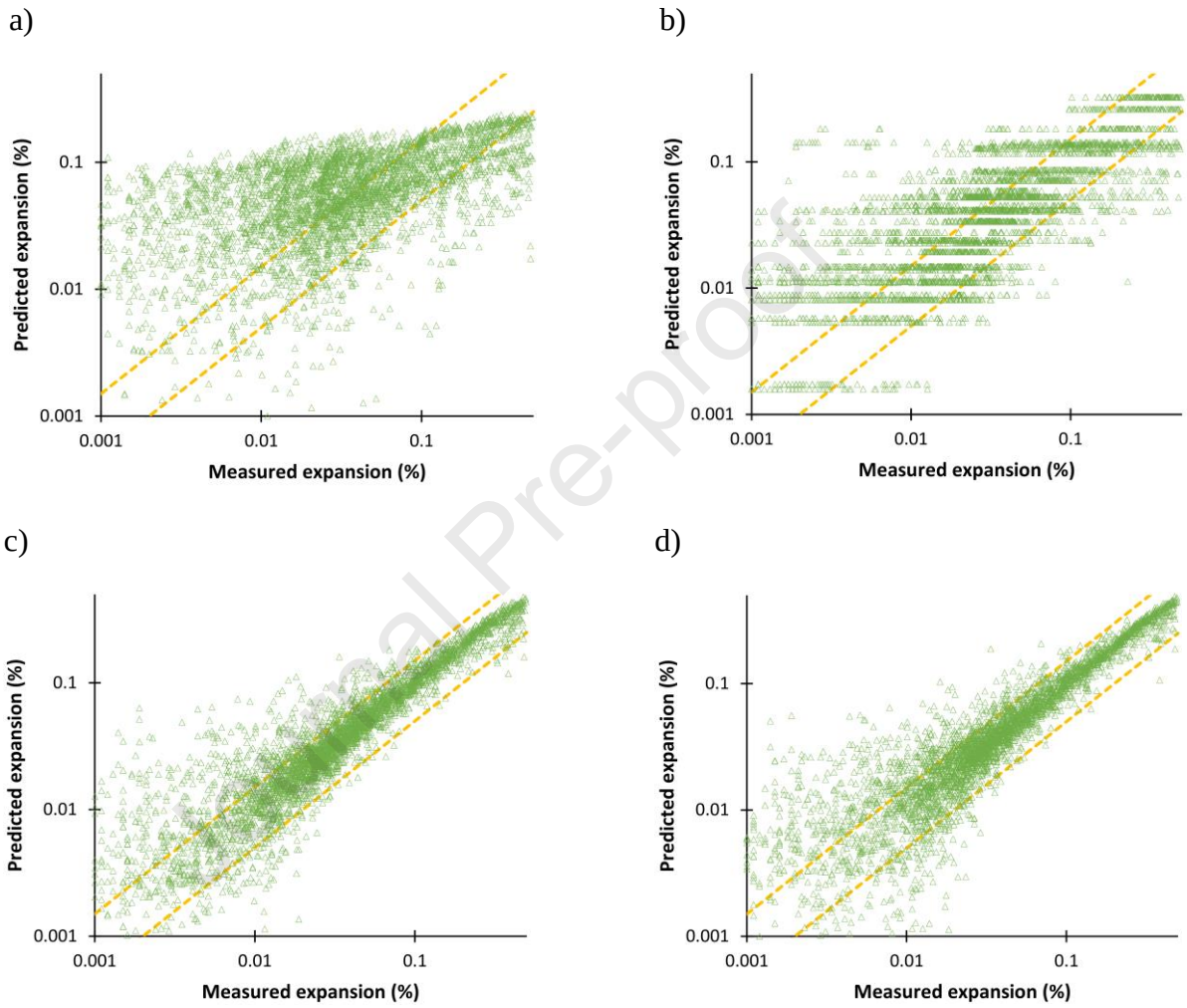


Fig. 6. Comparison between target and calculated outputs of the models on the training database: a) LR, b) DT, c) XGB, d) LGBM

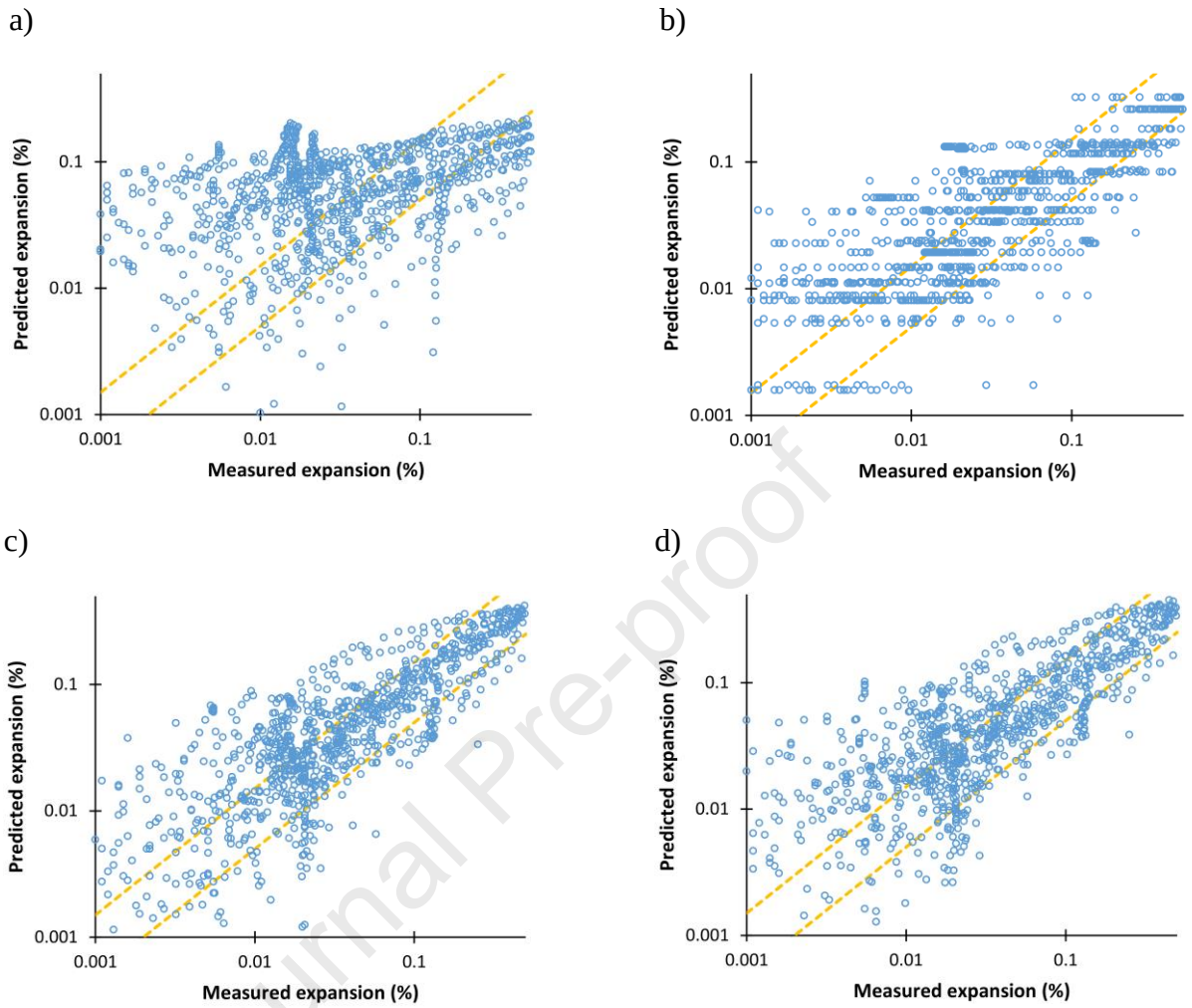


Fig. 7. Comparison between target and calculated outputs of the models on the testing database: a) LR, b) DT, c) XGB, d) LGBM

Table 4. Mortar and concrete compositions of some typical test samples.

Ref	NSR-M [63]	SR-M [73]	NSR-C [54]	SR-C [8]
Type	CEM I	CEM I	CEM I	CEM I
C_3S %	56.6*	52.57	58.4	73.5
C_2S%	14.5*	18.11	14.65	5.5
C_3A%	5.7*	7.59	6.22	2
C_4AF%	12.1*	10.04	8.99	12.9
Cement(kg/m^3)	499	320	350	352
A/C	3.0	4.58	4.79	5.19
W/B	0.55	0.5	0.55	0.49
Fly Ash%	0	0	0	0

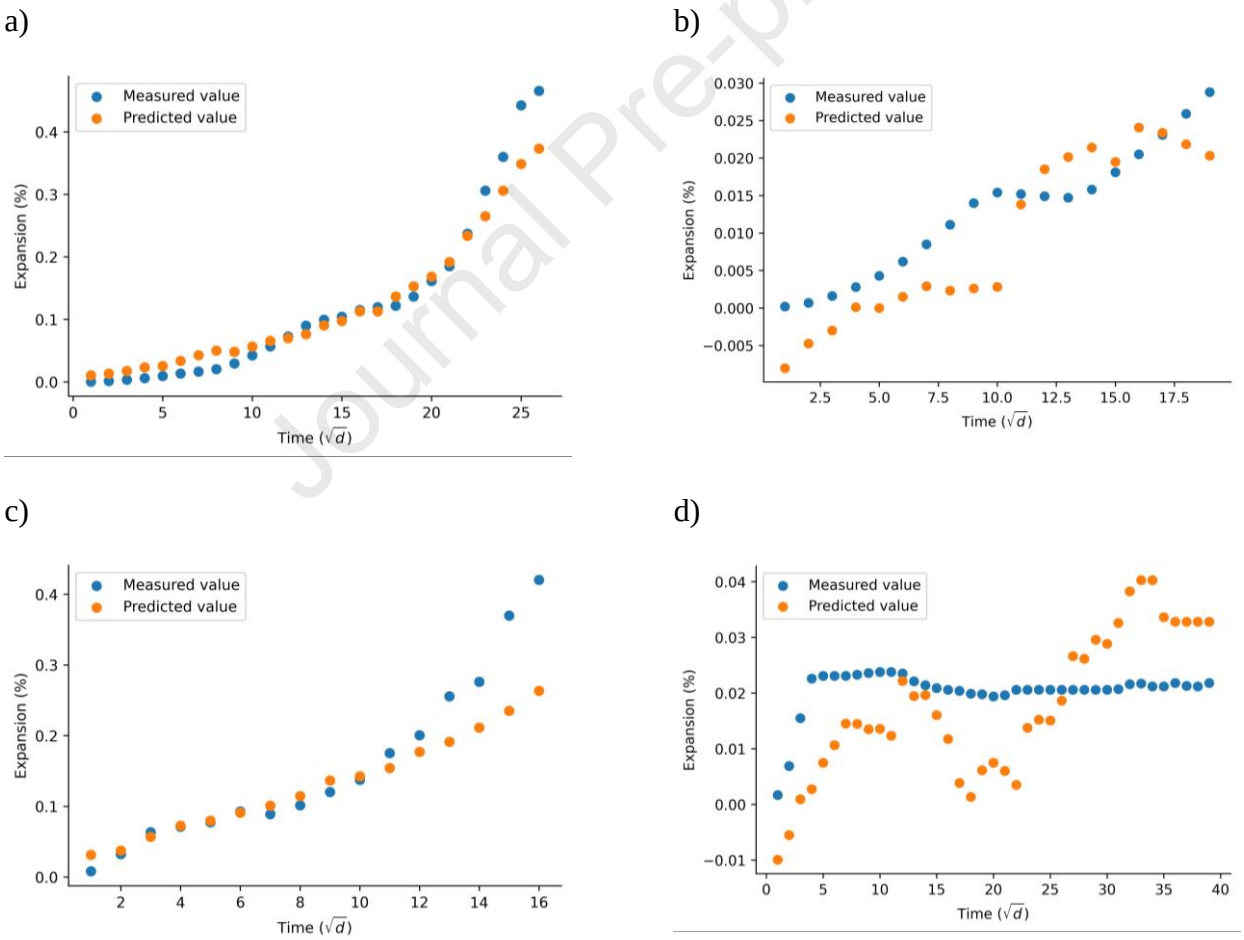
Slag%	0	40	0	0
Limestone%	0	0	0	0
MK (%)	0	0	0	0
Silica Fume (%)	0	0	0	0
Cation	Mg	Na	Na	Na
Concentration %	4.24	5	5	0.3
pH	7	10***	10***	7.5
Mold properties	prismatic	prismatic	cylindrical	prismatic
Surface/perimeter cm	1	1	2.5	1.75
fc28 (MPa)	46.1**	48.5**	43.1	52.7**

354 * Calculated using Bogue calculation

355 ** Inferred using the dedicated XGB model (cf section 2.2.1)

356 *** Imputed (cf section 2.2.2)

357



358 **Fig 8.** Typical predictions of expansion evolution of mortar and concrete samples due to external sulfate
359 attack: a) non sulfate resistant mortar (NSR-M), b) sulfate resistant mortar (SR-M), c) non sulfate
360 resistant concrete (NSR-C), d) sulfate resistant concrete (SR-C).

4.3. Model universality

The universality, i.e. generalization capacity of the model was further validated by predicting the expansion of samples from a study not included in the database [74]. The comparison of measured and predicted values shown in Figure 9 confirms the good generalization capacity of the model, as all four expansion kinetics were correctly reproduced. Additionally, predictions of sulfate-resistant mixtures from this study were also satisfactorily predicted.

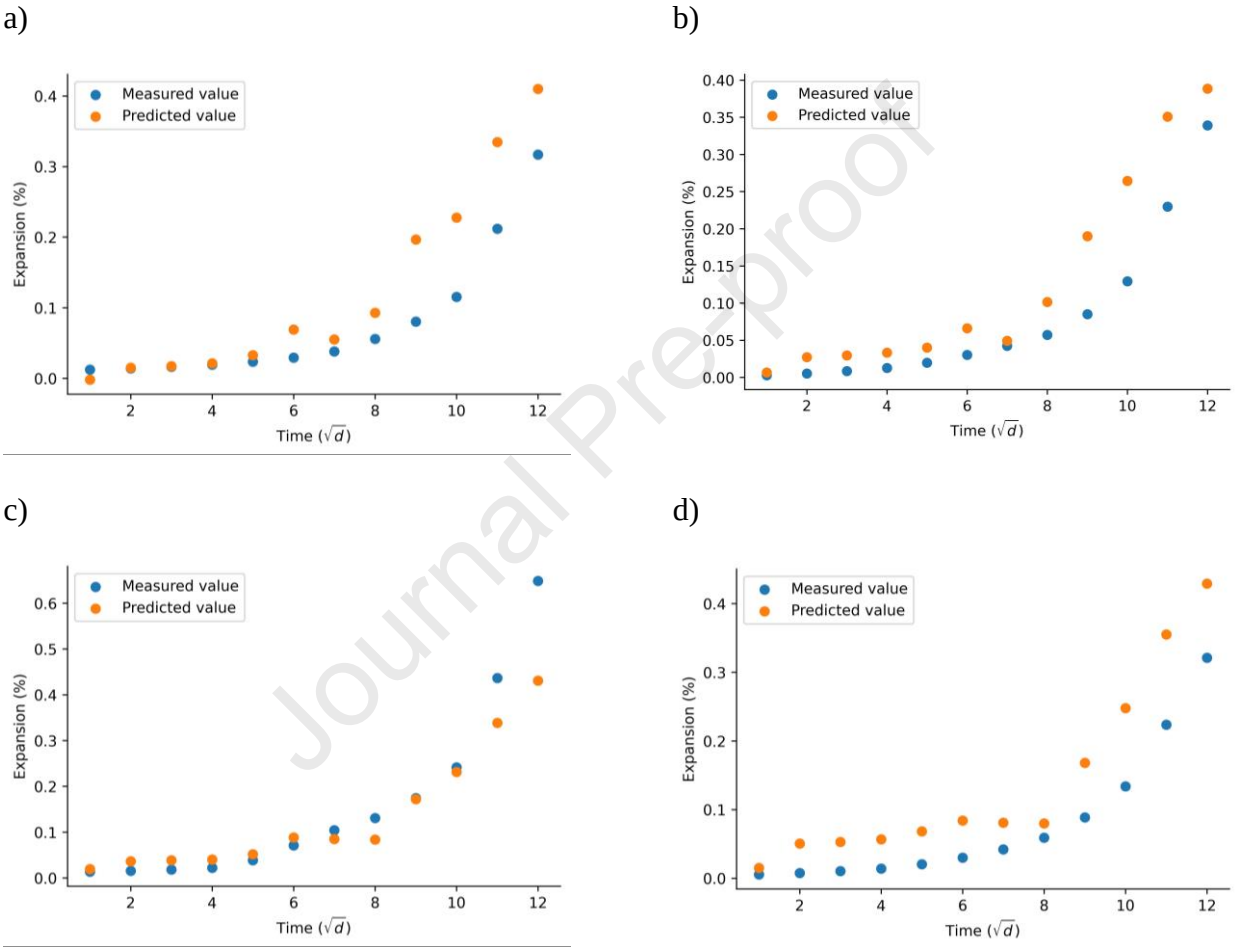


Fig 9. Model universality test: expansion predictions of four samples from a study not included in the database.

4.4. Optimized models' performances

The performance of the optimized models on the training and the test sets has then been evaluated by comparing the predicted to the interpolated measured expansions relatively to the square root of time. The mean values of the performance predictions of the models have been reported in Table 5. As illustrated by the table, the best results have been obtained by the ensemble

models, e.g., XGB and LGBM. Indeed, the best results on the training set have been obtained by the LGBM model with mean R^2 values of 0.947 and a corresponding RMSE value of 0.0205. This model can very well learn the mechanisms at the origin of expansion. XGB model also performed very well on the training set with R^2 and RMSE values of 0.933 and 0.0230 resp. The other models, LR and DT, were found to be less performant with R^2 values of 0.358 and 0.591 resp., which agrees with the literature relative to the prediction of concrete properties using Machine Learning. This observation can be attributed to the lower complexity of these models, which is insufficient to efficiently learn the hidden patterns underlying the external sulfate attack expansion.

The generalization capacity of the models and their performance on unknown data has been assessed by predicting the test data, that is, full expansion curves knowing only cement composition, mortar composition, specimens' geometry, and sulfate solution characteristics. As illustrated in Fig. 7 and Table 5, the XGB model obtained the best results on the test set. R^2, RMSE and MAE values of 0.788, 0.0466%, and 0.0273% were achieved resp. LGBM model performed slightly worse with R^2, RMSE and MAE values of 0.762, 0.0495%, and 0.0307% resp. Again, LR and DT models performed significantly worse, which agrees with the abovementioned observations made on the training set results. Though these values are slightly lower than the scores on the training set, which can be explained by the difficult and highly nonlinear problem and the relatively limited amount of data sources, these values can be acceptable to achieve good predictions and parameters interpretation, as we will see in the next sections.

Table 5. Mean machine learning algorithms performance for external sulfate attack expansion prediction.

Algorithm	Training set			Test set		
	R^2	RMSE (%)	MAE (%)	R^2	RMSE (%)	MAE (%)
LR	0.358	0.0714	0.0465	0.240	0.0882	0.0595
DT	0.591	0.0570	0.0306	0.596	0.0642	0.0375
XGB	0.933	0.0230	0.0112	0.788	0.0466	0.0273
LGBM	0.947	0.0205	0.0110	0.762	0.0493	0.0307

An overall evaluation of the models can be visualized using a Taylor diagram, as illustrated in Fig. 10, which summarizes three valuable characteristics of the models compared to the measured values: the standard deviation associated with the model's predictions, the correlation

between the predictions and the experimental values, and the centered root-mean-square difference (RMSD). Similarly to what has been reported previously concerning RMSE, the best RMSD has been obtained by the LGBM model, followed by XGB, DT, and LR models. Concerning the correlation of the models with the experimental values, the best correlation is obtained by the LGBM and XGB models with values around 0.86 and 0.88 resp. Though XGB performed better regarding RMSD and correlation, the standard deviation of the experimental data of around 0.1% is almost reproduced by the LGBM model, which is probably due to the highest number of trees composing the optimized model (see Table 3). However, the standard deviations associated with the two other models are significantly smaller than the experimental value, which confirms the difficulty of DT and LR models in grasping the experimental data diversity. For this reason, it can be concluded that XGB is the best model per se on the test data in this study, but the LGBM model is more complex due to a higher number of trees. Thus, the optimized XGB model results will be discussed in the following sections related to the model interpretation.

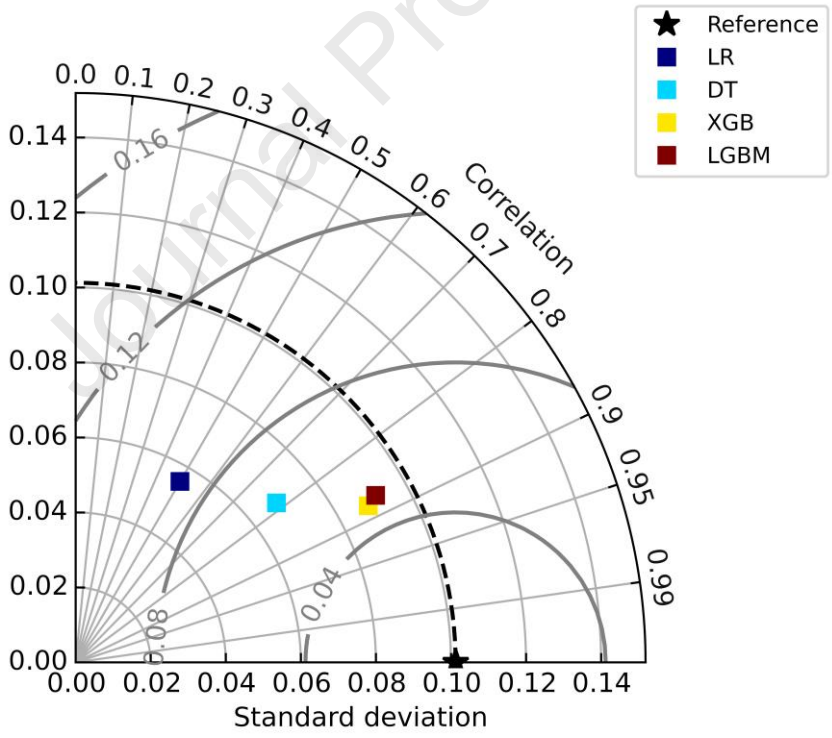


Fig. 10. Taylor diagram representing ML models performance on the test set.

4.5. Model interpretation and feature importance analysis using Shapley Additive Explanations (SHAP)

4.5.1. Global interpretation

The most influential features on the expansion predictions have been obtained using SHAP, and the global SHAP values are reported in Fig. 11. In this figure, the features are classified in descending order based on their influence from top to bottom and colored depending on their type. Notably, the top 18 parameters accounted for 98.5% of the cumulative mean absolute SHAP value. Excluding time, the respective contributions of clinker composition, mix proportion, sample geometry, and sulfate solution characteristics are 34%, 36%, 3%, and 27%, illustrating the intricate nature of the ESA mechanism and the need to construct intricate databases to effectively utilize ML models to anticipate ESA-induced expansion.

According to the model, the cation type, sodium or magnesium, is the most influential parameter, besides time. While additional studies exploring magnesium would have been advantageous, this finding aligns with existing literature. While Na^+ has no influence on the attack mechanism, Mg^{2+} ions from $MgSO_4$ react with cement paste hydrates and prevail over sulfate attack. The most significant consequence of $MgSO_4$ is the formation of M-S-H in place of C-S-H. As M-S-H do not provide the same cohesive characteristics as C-S-H, the cementitious sample will lose compressive strength over expansion [75].

Among the most influential parameters predicted by the model, the water and cement content, C_3A content, and sulfate solution concentration have already been extensively documented and are already taken into account in various standard recommendations. The model reveals 28-day compressive strength as another influential parameter. Strength is commonly used as a degradation indicator more than a parameter [76–78]. This result is related to porosity: a sample with high porosity is prone to sulfate ingress and, simultaneously, has less compressive strength. Finally, time of exposure is the most influential parameter. It is worth noting that among these most influential parameters, various types of parameters, e.g., cement composition, mixture proportions, and sulfate solution characteristics are all present, which highlights the complexity of the sulfate attack degradation.

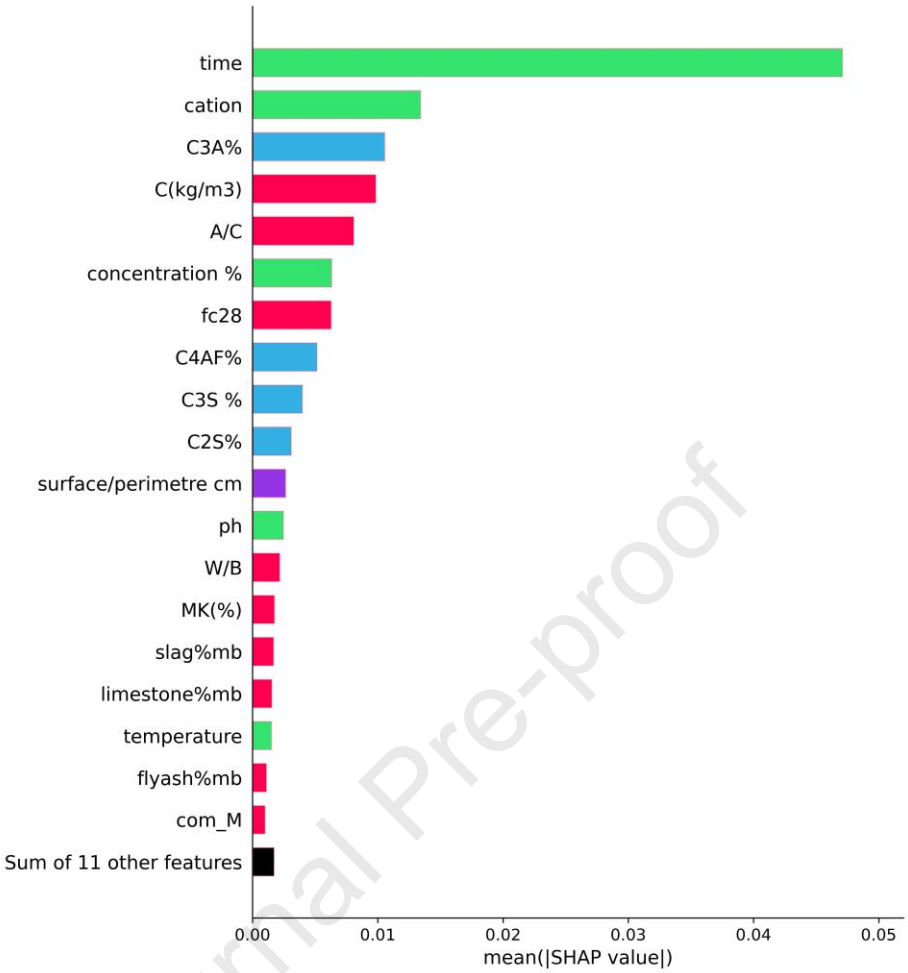


Fig 11. Feature importance plot of the optimized XGB model (clinker parameters, mix factors, sample geometry-related parameters and environmental factors are colored in blue, red, purple and green resp.)

From the SHAP summary plot illustrated in Fig. 12, the relative influence of the most influential parameters can be qualitatively determined using the spread of the dots and their color. It has been found that increasing water and cement contents generally increase the expansion. Increasing C_3A content also increases the expansion, especially when transitioning from very low C_3A content close to 0% to moderate C_3A content of around 7-8%. Regarding the other factors related to clinker composition, it was found that the C_4AF content is the second most important clinker factor, the higher its concentration the lower the expansion. This can be partly explained by its negative correlation with C_3A content, as illustrated in Fig. 2. C_3S and C_2S contents have a relatively small influence on the expansion. Fig. 12 also shows that an increase in 28-day compressive strength leads to a decrease in sulfate expansion, while higher aggregate-to-cement ratios are associated with smaller expansions, which can be explained by the fact that only the

cement paste reacts.

Though significantly less important according to the SHAP analysis, the influence of SCM can also be quantified and agreed with the literature [79–82]. Based on the model results, it has been found that limestone can harm sulfate resistance at high dosage, while calcined clay and slag addition have a powerfully positive impact on sulfate resistance, and fly ash has a relatively smaller yet positive influence on sulfate resistance. Conversely, silica fume addition might negatively influence sulfate attack resistance, especially for moderate to high dosages.

Among the specimen geometrical parameters, it has been found that the surface-to-perimeter ratio has the most substantial influence, which agrees with the experimental observations [83,84]. Indeed, the higher the surface-to-perimeter, the smaller the expansion. Other geometrical parameters such as the specimen shape (mold properties) and height have a smaller influence, though quantitatively close to some SCM additions. Thus, the model results support the development of accelerated tests involving specimens with small sections. Additionally, a careful analysis of the sulfate induced expansion results must be performed regarding these parameters.

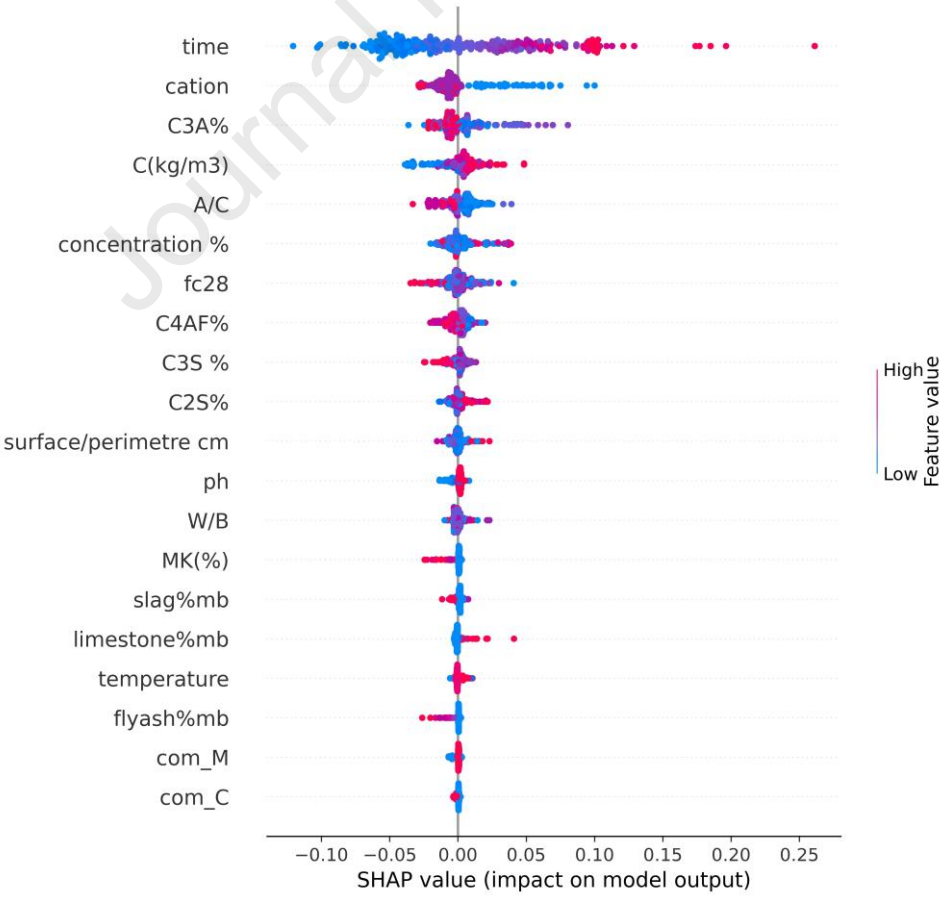


Fig 12. SHAP summary plot of the optimized XGB model.

4.5.2 Feature dependence plots

The feature dependency of the results has been analyzed in detail using the SHAP feature dependence plots. As illustrated in the dependence plots reported in Fig. 13, the relative influence of each parameter on the computed SHAP values and, therefore, their influence on the predicted expansion can be calculated. Apart from the cation type of the sulfate solution, which greatly influences the development of expansion, as explained in the previous section, the most influential parameters belong to the cement composition and the mixture proportion. Among these parameters, cement content and water content, which is not reported, have a similar effect and strongly and almost linearly influence the expansion as illustrated in Fig. 13 a). Indeed a difference in SHAP values of around 0.05% is almost systematically observed between mixes with cement contents around 100 kg/m^3 and mixes with more than 500 kg / m^3 of cement. Similarly, as illustrated in Fig. 13 b), SHAP values increase around 0.1% when C_3A content increases from around 0% to 8%, meaning that higher expansions are associated to C_3A contents of around 8% as compared to smaller to C_3A contents. However, higher C_3A contents (<8%) have a smaller influence on SHAP value (around 0%). Even though this result might be due to the relatively small number of high-C_3A content cements in the database, this could also be explained by the fact that other factors predominate in the expansion phenomenon of these specific concretes.

Mixtures characteristics such as the 28-day compressive strength and the aggregate-to-cement ratios clearly influence the expansion values. As illustrated in Fig. 13 c), the SHAP values associated with high 28-day compressive strength values decrease. In the same way, increasing the A/C ratio reduces the predicted expansion, especially when the A/C ratio increases from around 1-2 to 4-5.

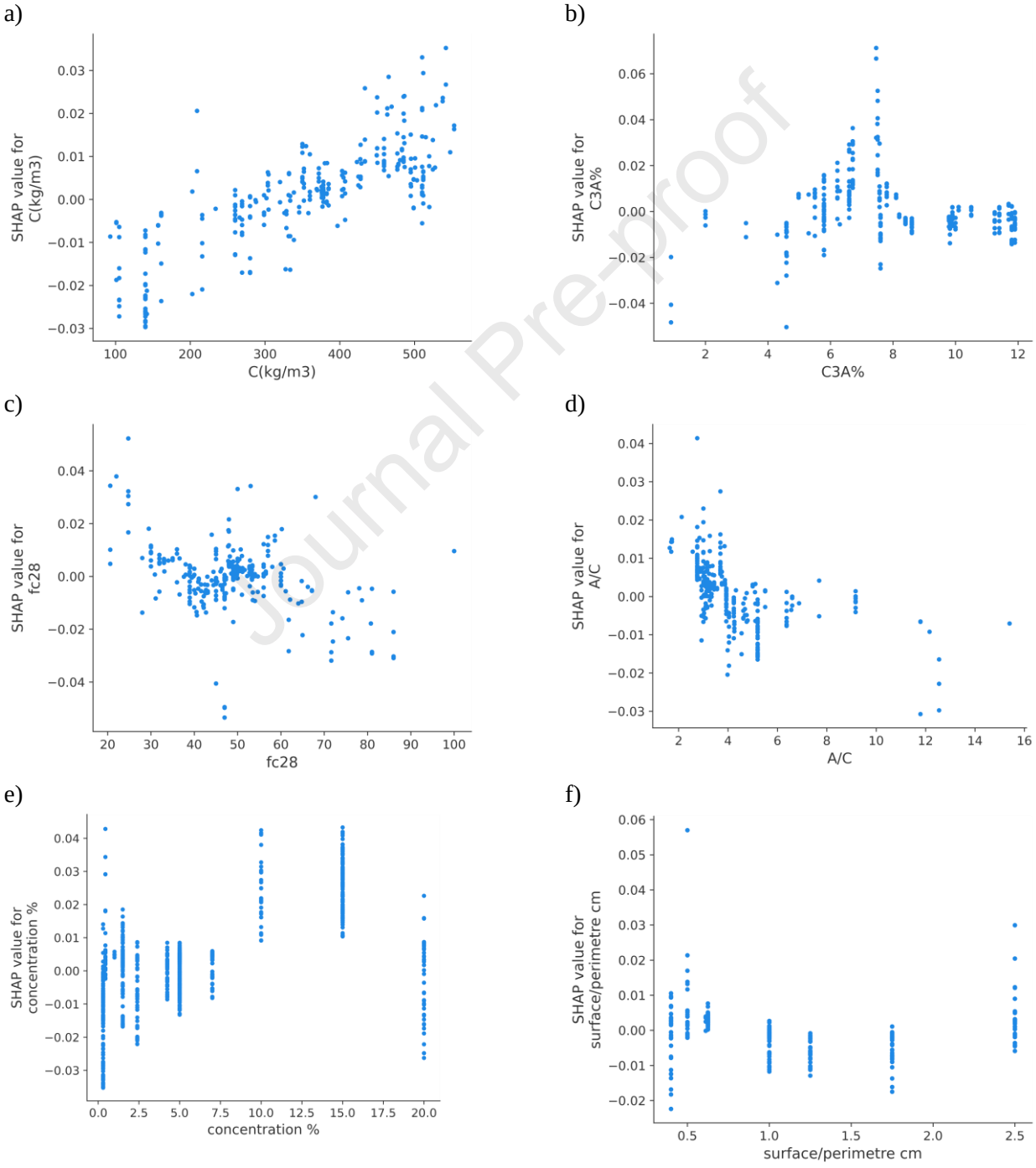
Accelerated test conditions such as sulfate solution concentration and geometrical properties of the samples can be analyzed thoroughly. As illustrated in Fig. 13 e), increasing sulfate concentration values, in the range of 0% to 15%, generate more expansion. However, sulfate concentrations around 20% might not accelerate the degradation. Similarly, when the surface-to-perimeter ratio of the specimens increases from around 0.5 to 1.7, SHAP values decrease, which means that the samples are less prone to sulfate expansion.

The partial substitution of cement by SCM can also be analyzed though more data would be

502 welcome. As illustrated in Fig. 13 g), cement replacement by slag generally leads to lower
503 expansions, especially for replacement rates higher than 30-40% [17,85]. In the same way, cement
504 replacement by calcined clay usually decreases expansions, although 30% replacement rates might
505 not always have positive effects.

506

507



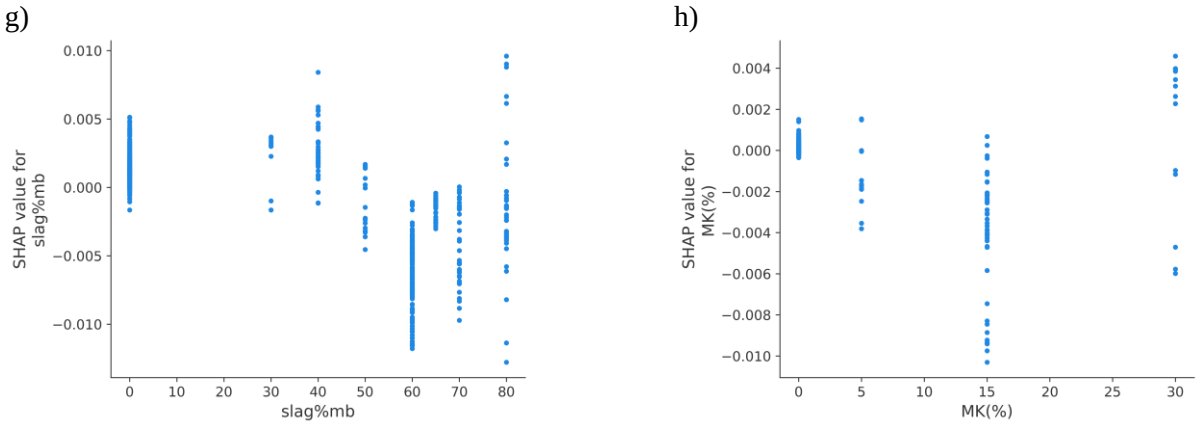


Fig 13. SHAP feature dependence plot: a) C_3A, b) cement content, c) fc28, d) A/C, e) sulfate concentration, f) surface over perimeter, g) slag content, and, h) calcined clay content.

4.5.3. Local interpretation

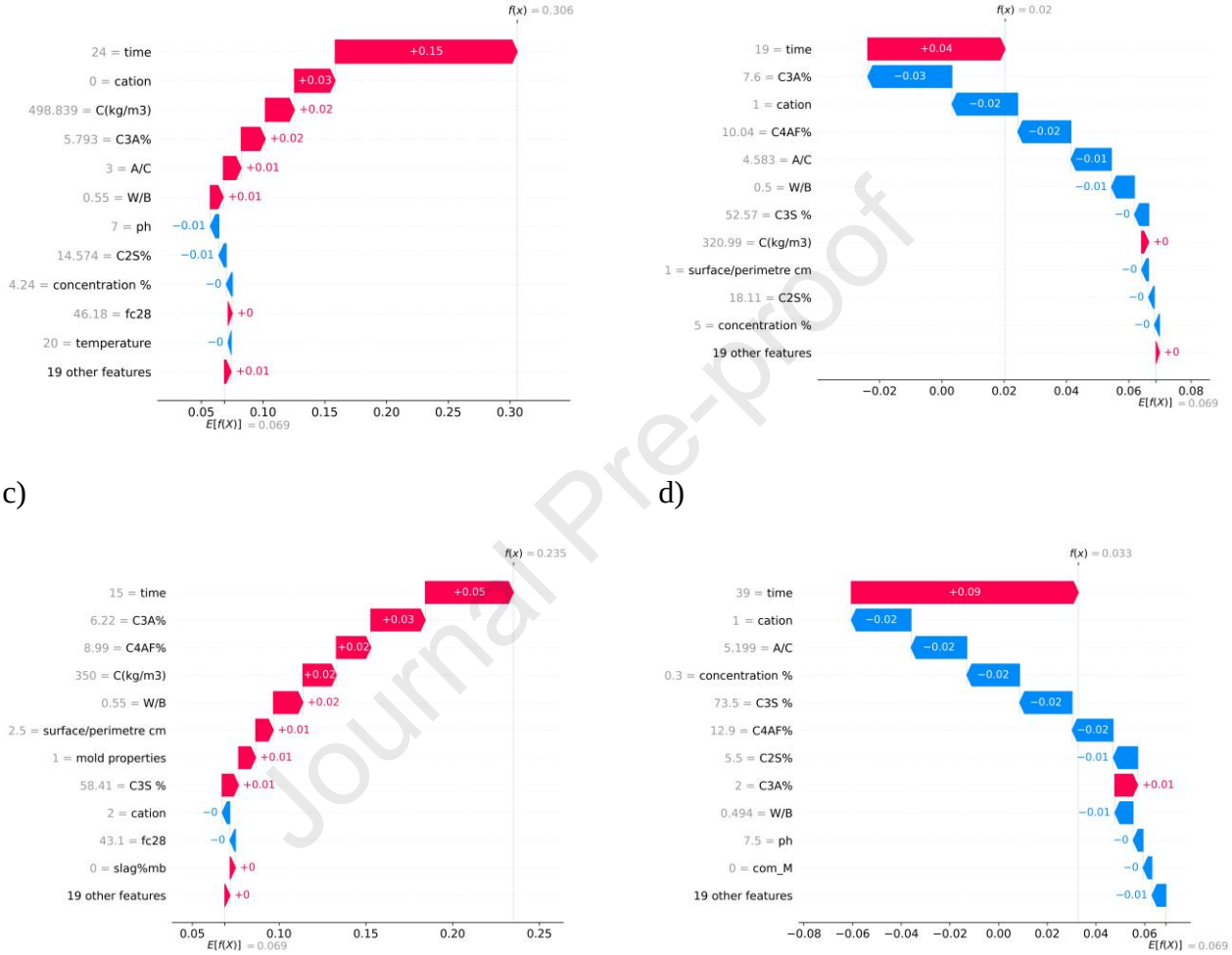
The local SHAP analysis of four typical mortar and concrete samples whose composition is given in Table 4 are reported in Fig. 14. With a fixed base value of 0.069%, the final prediction for expansion, displayed atop each graph, is the counteracting result of SHAP values of different features.

According to the SHAP analysis of the results concerning mortar samples (Fig. 14 a and b), the amount of cement has a negative impact on the expansion results. Thus, mortar samples made with cement containing 5.8 % of C_3A can have more expansion than mortar samples made with 7.6 % of C_3A. This indicates that the C_3A content can represent a limit between SR and NSR cement. However, above this limit there is no proportionality between expansion and C_3A content [86]. Moreover, other parameters can compensate the poor performance of a given cement such as the low cement content, a relatively high aggregate-to-cement ratio or a moderate water-to-binder ratio, as in sample SR-M (Fig 14 b).

The SHAP local analysis also gives consistent and valuable indications for NSR-C and SR-C concrete specimens qualification, as illustrated in Fig. 14 c) and d) resp. The C_3A content of the clinker part is the parameter that has the most negative impact on the expansion for NSR-C. The concrete samples made with cement containing an important proportion of C_3A must present assets to offset the great reactivity of the cement to sulfate ions. It can be a lower cement content or a low w/b ratio. Here it is not the case for the NSR-C concrete, as the high cement content and w/b ratio have, associated with a low aggregate-to-cement ratio, favor expansion. Conversely, in SR-C sample, the formulation of concrete, the cement composition and the conditions of exposure

decrease the expansion due to the reaction to sulfate. Those parameters compensate for the slightly negative effect of the 2% C_3A content.

a) b)



c) d)

Fig 14. Local interpretation of mortar and concrete samples expansion due to external sulfate attack: a) non sulfate resistant mortar (NSR-M), b) sulfate resistant mortar (SR-M), c) non sulfate resistant concrete (NSR-C), d) sulfate resistant concrete (SR-C).

Table 6. Comparison of measured and predicted expansions of typical test samples.

Ref	NSR-M [63]	SR-M [87]	NSR-C [54]	SR-C [8]
Time (days)	576	361	225	1521
Measured expansion (%)	0.360	0.029	0.369	0.022

Predicted expansion (%)	0.306	0.020	0.235	0.033

4.6. Prediction of time to reach specific expansions

The time to reach 0.2% expansion of the test specimens has been calculated following the methodology that has been aforementioned. The results are represented in Fig. 15. It can be observed in this figure that among the 68 samples, 43 were correctly predicted to not reach 0.2% expansion (points with x and y values equal to 1000 days on the top right), 18 samples showing an extension have been very well predicted (bottom-left part of the graph), while 7 samples, i.e. 10%, expansion time have been incorrectly predicted (6 samples have been underestimated and 1 samples have been overestimated). In total, in 84% of the cases, the prediction and measured time differed less than 100 days. The R^2 and correlation values between the predicted and experimentally measured expansion time reached 0.64 and 0.80 resp., and the regression line and 95% confidence interval reported on the graph can be used to practically infer the time ti reach a specific expansion. Therefore, it can be concluded that, even though some improvements may be needed to precisely predict the time to reach a given expansion, the optimized Ensemble Machine Learning models can be advantageously employed to estimate the degradation time of mortar and concrete specimens subjected to accelerated external sulfate attack degradation tests and can help select non-expansive formulations.

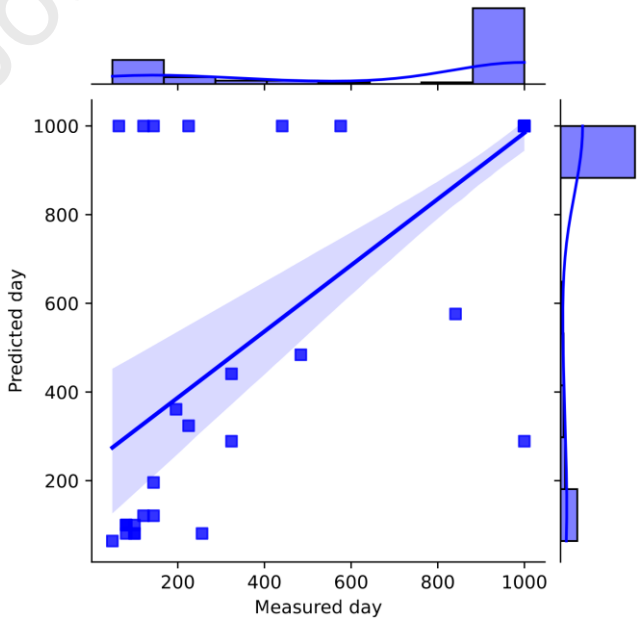


Fig 15. Comparison of XGB predicted and measured time needed to reach a 0.2% expansion with

559 regression line and 95% confidence interval (1000d values are attributed to specimens not reaching the
560 0.2% expansion threshold)

561 **5. Conclusions and perspectives**

562 This study investigated the potential of machine learning models to predict the expansion of
563 cementitious materials incorporating supplementary cementitious materials (SCM) subjected to
564 external sulfate attack (ESA) degradation. For the first time, a comprehensive database
565 documenting the expansion of mortar and concrete specimens has been built. Various types of
566 inputs have been considered: cement composition, mixture proportions, sulfate solution
567 characteristics, and geometrical features of the samples. Several techniques have been developed
568 to fill in the missing values, such as using an XGB model to infer the missing 28-day compressive
569 strengths. Then four machine learning (ML) models of ascending intricacy were used to predict
570 the temporal evolution of expansion. The machine learning models have been optimized using the
571 Tree-structured Parzen Estimator algorithm. Next, Shapley Additives Explanations (SHAP) were
572 then used to interpret the machine learning models' performances, and the time to reach specific
573 expansion were calculated using the test predictions. The results were discussed and compared to
574 the literature, demonstrating the ability of ensemble ML models to predict ESA-induced
575 expansion. The following main conclusions can be drawn:

576 - ML models can be applied to predict ESA-induced expansion thanks to a comprehensive
577 database gathering from the literature 336 conventional, binary and ternary mortar and
578 concrete sample expansions with 20 parameters. Expansion values have been interpolated,
579 and values higher than 0.5% have been filtered out to finally generate a database
580 containing 5294 expansion values at different times.

581 - Two ensemble machine learning models, Light Gradient Boosting (LGBM) and Extreme
582 Gradient Boosting (XGB), were trained on the completed database using a 5-fold CV
583 procedure after using a train-test split on samples, and outperformed the other two models
584 under investigation. The best optimized XGB model achieved a good R^2 accuracy of 0.933
585 and 0.788 on the training and the test set resp. Its universality has been validated on
586 unknown data.

587 - The most influential parameters were the cation type in the sulfate solution, the water and
588 cement contents in the mixture, the C_3A amount within the cement, the 28-day

compressive strength (which serves as a porosity indicator), the sulfate solution concentration, and the aggregate-to-cement ratio. The clinker composition, the mix proportion, the geometry of the sample, and the quantity of sulfate solution contributed 34%, 36%, 3%, and 27%, respectively, demonstrating the complex nature of ESA.

- An in-depth analysis of some predicted expansions from specimens in the test set ensured the consistency of the model and helped quantify the impact of the input parameters in these specific cases.
- Time to reach specific expansions due to ESA can be estimated using ML models. In 84% of the cases, the predicted and measured time for specimens from the test set to reach a 0.2% expansion differed less than 100 days.

The study might open up new research paths related to the design of cementitious materials with SCM that can resist the external sulfate attack. Moreover, this model could help develop more efficient accelerated tests to assess the sulfate resistance of cementitious materials quickly. Such advanced models might be of interest in the future regarding highly durable eco-friendly cementitious materials development.

Conflict of Interest: The authors declare that there is no conflict of interest.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Availability of data and material: Data will be provided upon request.

References

[1] B.A. Clark, P.W. Brown, Phases formed during hydration of tetracalcium aluminoferrite in 1.0M magnesium sulfate solutions, Cement and Concrete Composites. 24 (2002) 331–338. https://doi.org/10.1016/S0958-9465(01)00084-1.

[2] X. Ping, J.J. Beaudoin, MECHANISM OF SULPHATE EXPANSION, 22 (n.d.) 10.

[3] C. Yu, W. Sun, K. Scrivener, Mechanism of expansion of mortars immersed in sodium sulfate solutions, Cement and Concrete Research. 43 (2013) 105–111. https://doi.org/10.1016/j.cemconres.2012.10.001.

[4] W. Müllauer, R.E. Beddoe, D. Heinz, Sulfate attack expansion mechanisms, Cement and Concrete Research. 52 (2013) 208–215. https://doi.org/10.1016/j.cemconres.2013.07.005.

[5] Z. Makhloufi, S. Aggoun, B. Benabed, E.H. Kadri, M.Bederina, Effect of magnesium sulfate on the durability of limestone mortars based on quaternary blended cements, Cement and Concrete Composites. 65 (2016) 186–199. https://doi.org/10.1016/j.cemconcomp.2015.10.020.

[6] D. Damidot, S.J. Barnett, F.P. Glasser, D.E. Macphee, Investigation of the CaO–Al $_2$ O $_3$ –SiO $_2$ –CaSO $_4$ –CaCO $_3$ –H $_2$ O system at 25°C by thermodynamic calculation, Advances in Cement Research. 16 (2004) 69–76. https://doi.org/10.1680/adcr.2004.16.2.69.

[7] F. Bellmann, B. Möser, J. Stark, Influence of sulfate solution concentration on the formation of gypsum in sulfate resistance test specimen, Cement and Concrete Research. 36 (2006) 358–363. https://doi.org/10.1016/j.cemconres.2005.04.006.

[8] R. El-Hachem, E. Rozière, F. Grondin, A. Loukili, New procedure to investigate external sulphate attack on cementitious materials, Cement and Concrete Composites. 34 (2012) 357–364. https://doi.org/10/fggr6x.

[9] D. Planel, J. Sercombe, P. Le Bescop, F. Adenot, J.-M. Torrenti, Long-term performance of cement paste during combined calcium leaching–sulfate attack: kinetics and size effect, Cement and Concrete Research. 36 (2006) 137–143. https://doi.org/10.1016/j.cemconres.2004.07.039.

[10] J. Bizzozero, C. Gosselin, K.L. Scrivener, Expansion mechanisms in calcium aluminate and sulfoaluminate systems with calcium sulfate, Cement and Concrete Research. 56 (2014) 190–202. https://doi.org/10.1016/j.cemconres.2013.11.011.

[11] N.J. Crammond, The thaumasite form of sulfate attack in the UK, Cement and Concrete Composites. 25 (2003) 809–818. https://doi.org/10.1016/S0958-9465(03)00106-9.

[12] NF EN 197-1.pdf, (n.d.).

[13] H.T. Cao, L. Bucea, A. Ray, S. Yozghatlian, The effect of cement composition and pH of environment on sulfate resistance of Portland cements and blended cements, Cement and Concrete Composites. 19 (1997) 161–171. https://doi.org/10.1016/S0958-9465(97)00011-5.

[14] Z. Shi, S. Ferreiro, B. Lothenbach, M.R. Geiker, W. Kunther, J. Kaufmann, D. Herfort, J. Skibsted, Sulfate resistance of calcined clay – Limestone – Portland cements, Cement and Concrete Research. 116 (2019) 238–251. https://doi.org/10.1016/j.cemconres.2018.11.003.

[15] A. Rossetti, T. Ikumi, I. Segura, E.F. Irassar, Sulfate performance of blended cements (limestone and illite calcined clay) exposed to aggressive environment after casting, Cement and Concrete Research. 147 (2021) 106495. https://doi.org/10.1016/j.cemconres.2021.106495.

[16] M. Sahmaran, O. Kasap, K. Duru, I.O. Yaman, Effects of mix composition and water–cement ratio on the sulfate resistance of blended cements, Cement and Concrete Composites. 29 (2007) 159–167. https://doi.org/10.1016/j.cemconcomp.2006.11.007.

[17] S. Boudache, E. Rozière, A. Loukili, L. Izoret, Towards common specifications for low- and high-expansion cement-based materials exposed to external sulphate attacks, Construction and Building Materials. 294 (2021) 123586. https://doi.org/10.1016/j.conbuildmat.2021.123586.

[18] Y.-J. Cha, W. Choi, O. Büyüköztürk, Deep Learning-Based Crack Damage Detection Using Convolutional Neural Networks: Deep learning-based crack damage detection using CNNs, Computer-Aided Civil and Infrastructure Engineering. 32 (2017) 361–378. https://doi.org/10/f928ww.

[19] S. Dorafshan, R.J. Thomas, M. Maguire, Comparison of deep convolutional neural networks and edge detectors for image-based crack detection in concrete, Construction and Building Materials. 186 (2018) 1031–1045. https://doi.org/10.1016/j.conbuildmat.2018.08.011.

[20] Z. Liu, Y. Cao, Y. Wang, W. Wang, Computer vision-based concrete crack detection using U-net fully convolutional networks, Automation in Construction. 104 (2019) 129–139. https://doi.org/10.1016/j.autcon.2019.04.005.

[21] Y. Song, Z. Huang, C. Shen, H. Shi, D.A. Lange, Deep learning-based automated image segmentation for concrete petrographic analysis, Cement and Concrete Research. 135 (2020) 106118. https://doi.org/10.1016/j.cemconres.2020.106118.

[22] B. Hilloulin, I. Bekrine, E. Schmitt, A. Loukili, Modular Deep Learning Segmentation Algorithm for Concrete Microscopic Images, Construction and Building Materials. 349 (2022). https://doi.org/10.1016/j.conbuildmat.2022.128736.

[23] B. Hilloulin, I. Bekrine, E. Schmitt, A. Loukili, Open-source deep learning-based air-voids detection algorithm for concrete microscopic images, Journal of Microscopy. (2022) jmi.13098.

https://doi.org/10.1111/jmi.13098.
[24] M. Liang, Y. Gan, Z. Chang, Z. Wan, E. Schlangen, B. Šavija, Microstructure-informed deep convolutional neural network for predicting short-term creep modulus of cement paste, Cement and Concrete Research. 152 (2022) 106681. https://doi.org/10.1016/j.cemconres.2021.106681.
[25] B. Hilloulin, M. Lagrange, M. Duvillard, G. Garioud, ε–greedy automated indentation of cementitious materials for phase mechanical properties determination, Cement and Concrete Composites. 129 (2022) 104465. https://doi.org/10.1016/j.cemconcomp.2022.104465.
[26] Y. Yu, W. Gao, A. Castel, A. Liu, X. Chen, M. Liu, Assessing external sulfate attack on thin-shell artificial reef structures under uncertainty, Ocean Engineering. 207 (2020) 107397. https://doi.org/10.1016/j.oceaneng.2020.107397.
[27] J.-S. Chou, C.-F. Tsai, A.-D. Pham, Y.-H. Lu, Machine learning in concrete strength simulations: Multi-nation data analytics, Construction and Building Materials. 73 (2014) 771–780. https://doi.org/10.1016/j.conbuildmat.2014.09.054.
[28] D.K. Bui, T. Nguyen, J.S. Chou, H. Nguyen-Xuan, T.D. Ngo, A modified firefly algorithm-artificial neural network expert system for predicting compressive and tensile strength of high-performance concrete, Construction and Building Materials. 180 (2018) 320–333. https://doi.org/10.1016/j.conbuildmat.2018.05.201.
[29] M.J. Munir, S.M. Saleem Kazmi, Y.-F. Wu, X. Lin, M.R. Ahmad, Development of a novel compressive strength design equation for natural and recycled aggregate concrete through advanced computational modeling, Journal of Building Engineering. (2022) 104690. https://doi.org/10.1016/j.jobe.2022.104690.
[30] H.S. Ullah, R.A. Khushnood, J. Ahmad, F. Farooq, Predictive modelling of sustainable lightweight foamed concrete using machine learning novel approach, Journal of Building Engineering. 56 (2022) 104746. https://doi.org/10.1016/j.jobe.2022.104746.
[31] W.E. Elemam, A.H. Abdelraheem, M.G. Mahdy, A.M. Tahwia, Optimizing fresh properties and compressive strength of self-consolidating concrete, Construction and Building Materials. 249 (2020) 118781. https://doi.org/10.1016/j.conbuildmat.2020.118781.
[32] J. Karthikeyan, A. Upadhyay, N.M. Bhandari, Artificial Neural Network for Predicting Creep and Shrinkage of High Performance Concrete, ACT. 6 (2008) 135–142. https://doi.org/10.3151/jact.6.135.
[33] P. Chen, W. Zheng, Y. Wang, W. Chang, Creep model of high-strength concrete containing supplementary cementitious materials, Construction and Building Materials. 202 (2019) 494–506. https://doi.org/10.1016/j.conbuildmat.2019.01.005.
[34] L. Bal, F. Buyle-Bodin, Artificial neural network for predicting drying shrinkage of concrete, Construction and Building Materials. 38 (2013) 248–254. https://doi.org/10.1016/j.conbuildmat.2012.08.043.
[35] B. Hilloulin, V.Q. Tran, Using machine learning techniques for predicting autogenous shrinkage of concrete incorporating superabsorbent polymers and supplementary cementitious materials, Journal of Building Engineering. 49 (2022) 104086. https://doi.org/10.1016/j.jobe.2022.104086.
[36] B. Hilloulin, V.Q. Tran, Interpretable machine learning model for autogenous shrinkage prediction of low-carbon cementitious materials, Construction and Building Materials. 396 (2023) 132343. https://doi.org/10.1016/j.conbuildmat.2023.132343.
[37] R. Cai, T. Han, W. Liao, J. Huang, D. Li, A. Kumar, H. Ma, Prediction of surface chloride concentration of marine concrete using ensemble machine learning, Cement and Concrete Research. 136 (2020) 106164. https://doi.org/10.1016/j.cemconres.2020.106164.
[38] I. Nunez, M.L. Nehdi, Machine learning prediction of carbonation depth in recycled aggregate concrete incorporating SCMs, Construction and Building Materials. 287 (2021) 123027. https://doi.org/10.1016/j.conbuildmat.2021.123027.
[39] X. Wu, S. Zheng, Z. Feng, B. Chen, Y. Qin, W. Xu, Y. Liu, Prediction of the frost resistance of high-performance concrete based on RF-REF: A hybrid prediction approach, Construction and Building Materials. 333 (2022) 127132. https://doi.org/10.1016/j.conbuildmat.2022.127132.
[40] S.M. Lundberg, S.-I. Lee, A Unified Approach to Interpreting Model Predictions, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA, 2017: pp. 4768–4777.
[41] A.K.S. T. H. Wee S.F. Wong, and A.K.M. Anisur Rahman, Sulfate Resistance of Concrete Containing

Mineral Admixtures, ACI Materials Journal. 97 (2000). https://doi.org/10.14359/9286.
[42] D.D. Higgins, Increased sulfate resistance of ggbs concrete in the presence of carbonate, Cement and Concrete Composites. 25 (2003) 913–919. https://doi.org/10.1016/S0958-9465(03)00148-3.
[43] L. Courard, A. Darimont, M. Schouterden, F. Ferauche, X. Willem, R. Degeimbre, Durability of mortars modified with metakaolin, Cement and Concrete Research. 33 (2003) 1473–1479. https://doi.org/10.1016/S0008-8846(03)00090-5.
[44] M. Zeljkovic, Metakaolin effects on concrete durability, University of Toronto, 2009.
[45] A. Bonakdar, B. Mobasher, Multi-parameter study of external sulfate attack in blended cement materials, Construction and Building Materials. 24 (2010) 61–70. https://doi.org/10.1016/j.conbuildmat.2009.08.009.
[46] H.N. Atahan, D. Dikme, Use of mineral admixtures for enhanced resistance against sulfate attack, Construction and Building Materials. 25 (2011) 3450–3457. https://doi.org/10.1016/j.conbuildmat.2011.03.036.
[47] W. Kunther, B. Lothenbach, K.L. Scrivener, On the relevance of volume increase for the length changes of mortar bars in sulfate solutions, Cement and Concrete Research. 46 (2013) 23–29. https://doi.org/10.1016/j.cemconres.2013.01.002.
[48] H. Siad, S. Kamali-Bernard, H.A. Mesbah, G. Escadeillas, M. Mouli, H. Khelafi, Characterization of the degradation of self-compacting concretes in sodium sulfate environment: Influence of different mineral admixtures, Construction and Building Materials. 47 (2013) 1188–1200. https://doi.org/10.1016/j.conbuildmat.2013.05.086.
[49] Ş. Yazıcı, H.Ş. Arel, D. Anuk, Influences of Metakaolin on the Durability and Mechanical Properties of Mortars, Arab J Sci Eng. 39 (2014) 8585–8592. https://doi.org/10.1007/s13369-014-1413-z.
[50] A.M. Hossack, M.D.A. Thomas, The effect of temperature on the rate of sulfate attack of Portland cement blended mortars in Na2SO4 solution, Cement and Concrete Research. 73 (2015) 136–142. https://doi.org/10.1016/j.cemconres.2015.02.024.
[51] F. Mittermayr, M. Rezvani, A. Baldermann, S. Hainer, P. Breitenbücher, J. Juhart, C.-A. Graubner, T. Proske, Sulfate resistance of cement-reduced eco-friendly concretes, Cement and Concrete Composites. 55 (2015) 364–373. https://doi.org/10.1016/j.cemconcomp.2014.09.020.
[52] A. Trümer, H.-M. Ludwig, Sulphate and ASR Resistance of Concrete Made with Calcined Clay Blended Cements, in: K. Scrivener, A. Favier (Eds.), Calcined Clays for Sustainable Concrete, Springer Netherlands, Dordrecht, 2015: pp. 3–9. https://doi.org/10.1007/978-94-017-9939-3_1.
[53] H.Ş. Arel, B.S. Thomas, The effects of nano- and micro-particle additives on the durability and mechanical properties of mortars exposed to internal and external sulfate attacks, Results in Physics. 7 (2017) 843–851. https://doi.org/10.1016/j.rinp.2017.02.009.
[54] V. Bulatović, M. Melešev, M. Radeka, V. Radonjanin, I. Lukić, Evaluation of sulfate resistance of concrete with recycled and natural aggregates, Construction and Building Materials. 152 (2017) 614–631. https://doi.org/10.1016/j.conbuildmat.2017.06.161.
[55] A. Baldermann, M. Rezvani, T. Proske, C. Grengg, F. Steindl, M. Sakoparnig, C. Baldermann, I. Galan, F. Emmerich, F. Mittermayr, Effect of very high limestone content and quality on the sulfate resistance of blended cements, Construction and Building Materials. 188 (2018) 1065–1076. https://doi.org/10.1016/j.conbuildmat.2018.08.169.
[56] M. Sullivan, M. Chorzepa, H. Hamid, S. Durham, S. Kim, Sustainable Materials for Transportation Infrastructures: Comparison of Three Commercially-Available Metakaolin Products in Binary Cementitious Systems, Infrastructures. 3 (2018) 17. https://doi.org/10.3390/infrastructures3030017.
[57] Namik Kemal University, V. Akyuncu, M. Uysal, Istanbul University Cerrahpasa, H. Tanyildizi, Firat University, M. Sumer, Sakarya University, Modeling the weight and length changes of the concrete exposed to sulfate using artificial neural network, Rdlc. 17 (2019) 337–353. https://doi.org/10.7764/RDLC.17.3.337.
[58] F. Nosouhian, M. Fincan, N. Shanahan, Y.P. Stetsko, K.A. Riding, A. Zayed, Effects of slag characteristics on sulfate durability of Portland cement-slag blended systems, Construction and Building Materials. 229 (2019) 116882. https://doi.org/10.1016/j.conbuildmat.2019.116882.
[59] X. Lv, Y. Dong, R. Wang, C. Lu, X. Wang, Resistance improvement of cement mortar containing silica fume to external sulfate attacks at normal temperature, Construction and Building Materials. 258 (2020) 119630. https://doi.org/10.1016/j.conbuildmat.2020.119630.

[60] A. Rossetti, T. Ikumi, I. Segura, E. Irassar, Sulfate Resistance of Blended Cements (Limestone Illite Calcined Clay) Exposed Without Previous Curing, in: XV International Conference on Durability of Building Materials and Components. EBook of Proceedings, CIMNE, 2020. https://doi.org/10.23967/dbmc.2020.224.
[61] Y. Yang, B. Zhan, J. Wang, Y. Zhang, W. Duan, Damage evolution of cement mortar with high volume slag exposed to sulfate attack, Construction and Building Materials. 247 (2020) 118626. https://doi.org/10.1016/j.conbuildmat.2020.118626.
[62] G. Cordoba, E.F. Irassar, Sulfate performance of calcined illitic shales, Construction and Building Materials. 291 (2021) 123215. https://doi.org/10.1016/j.conbuildmat.2021.123215.
[63] Q. Huang, X. Zhu, G. Xiong, M. Zhang, J. Deng, M. Zhao, L. Zhao, Will the magnesium sulfate attack of cement mortars always be inhibited by incorporating nanosilica?, Construction and Building Materials. 305 (2021) 124695. https://doi.org/10.1016/j.conbuildmat.2021.124695.
[64] A. Rossetti, T. Ikumi, I. Segura, E.F. Irassar, Sulfate performance of blended cements (limestone and illite calcined clay) exposed to aggressive environment after casting, Cement and Concrete Research. 147 (2021) 106495. https://doi.org/10.1016/j.cemconres.2021.106495.
[65] A. Schneider, G. Hommel, M. Blettner, Linear Regression Analysis, Dtsch Arztebl Int. 107 (2010) 776–782. https://doi.org/10.3238/arztebl.2010.0776.
[66] Y. SONG, Y. LU, Decision tree methods: applications for classification and prediction, Shanghai Arch Psychiatry. 27 (2015) 130–135. https://doi.org/10.11919/j.issn.1002-0829.215044.
[67] T. Chen, C. Guestrin, XGBoost: A Scalable Tree Boosting System, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, New York, NY, USA, 2016: pp. 785–794. https://doi.org/10.1145/2939672.2939785.
[68] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, LightGBM: A Highly Efficient Gradient Boosting Decision Tree, in: 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 2017.
[69] M. Liang, Z. Chang, Z. Wan, Y. Gan, E. Schlangen, B. Šavija, Interpretable Ensemble-Machine-Learning models for predicting creep behavior of concrete, Cement and Concrete Composites. (2021) 104295. https://doi.org/10.1016/j.cemconcomp.2021.104295.
[70] J. Bergstra, R. Bardenet, Y. Bengio, B. Kég, Algorithms for hyper-parameter optimization, Advances in Neural Information Processing Systems. (2011).
[71] M.T. Ribeiro, S. Singh, C. Guestrin, "Why Should I Trust You?": Explaining the Predictions of Any Classifier, in: KDD, San Francisco, CA, USA, 2016. https://doi.org/10.48550/arXiv.1602.04938.
[72] P.J.M. Monteiro, K.E. Kurtis, Time to failure for concrete exposed to severe sulfate attack, Cement and Concrete Research. 33 (2003) 987–993. https://doi.org/10.1016/S0008-8846(02)01097-9.
[73] H.N. Atahan, D. Dikme, Use of mineral admixtures for enhanced resistance against sulfate attack, Construction and Building Materials. 25 (2011) 3450–3457. https://doi.org/10.1016/j.conbuildmat.2011.03.036.
[74] P.J. Tikalsky, D. Roy, B. Scheetz, T. Krize, Redefining cement characteristics for sulfate-resistant Portland cement, Cement and Concrete Research. 32 (2002) 1239–1246. https://doi.org/10.1016/S0008-8846(02)00767-6.
[75] E. Gruyaert, P. Van den Heede, M. Maes, N. De Belie, Investigation of the influence of blast-furnace slag on the resistance of concrete against organic acid or sulphate attack by means of accelerated degradation tests, Cement and Concrete Research. 42 (2012) 173–185. https://doi.org/10.1016/j.cemconres.2011.09.009.
[76] H. Binici, O. Aksoğan, Sulfate resistance of plain and blended cement, Cement and Concrete Composites. 28 (2006) 39–46. https://doi.org/10.1016/j.cemconcomp.2005.08.002.
[77] M.A. González, E.F. Irassar, Effect of limestone filler on the sulfate resistance of low C3A portland cement, Cement and Concrete Research. 28 (1998) 1655–1667. https://doi.org/10.1016/S0008-8846(98)00144-6.
[78] B. Persson, Sulphate resistance of self-compacting concrete, Cement and Concrete Research. 33 (2003) 1933–1938. https://doi.org/10.1016/S0008-8846(03)00184-4.
[79] X. Lv, L. Yang, J. Li, F. Wang, Roles of fly ash, granulated blast-furnace slag, and silica fume in long-term resistance to external sulfate attacks at atmospheric temperature, Cement and Concrete Composites. 133 (2022) 104696. https://doi.org/10.1016/j.cemconcomp.2022.104696.
[80] T. Aye, C.T. Oguchi, Resistance of plain and blended cement mortars exposed to severe sulfate attacks,

Construction and Building Materials. 25 (2011) 2988–2996. https://doi.org/10.1016/j.conbuildmat.2010.11.106.
[81] T. Schmidt, B. Lothenbach, M. Romer, J. Neuenschwander, K. Scrivener, Physical and microstructural aspects of sulfate attack on ordinary and limestone blended Portland cements, Cement and Concrete Research. 39 (2009) 1111–1121. https://doi.org/10.1016/j.cemconres.2009.08.005.
[82] E.F. Irassar, V.L. Bonavetti, M. González, Microstructural study of sulfate attack on ordinary and limestone Portland cements at ambient temperature, Cement and Concrete Research. 33 (2003) 31–41. https://doi.org/10.1016/S0008-8846(02)00914-6.
[83] X. Brunetaud, M.-R. Khelifa, M. Al-Mukhtar, Size effect of concrete samples on the kinetics of external sulfate attack, Cement and Concrete Composites. 34 (2012) 370–376. https://doi.org/10.1016/j.cemconcomp.2011.08.014.
[84] G. Massaad, E. Rozière, A. Loukili, L. Izoret, Do the geometry and aggregates size influence external sulfate attack mechanism?, Construction and Building Materials. 157 (2017) 778–789. https://doi.org/10.1016/j.conbuildmat.2017.09.117.
[85] A. Leemann, R. Loser, ACCELERATED SULFATE RESISTANCE TEST FOR CONCRETE - CHEMICAL AND MICROSTRUCTURAL ASPECTS, (2012) 8.
[86] S. Boudache, A. Loukili, L. Izoret, E. Rozière, Investigating the role played by portlandite and C-A-S-H in the degradation response of pozzolanic and slag cements to external sulphate attack, Journal of Building Engineering. 67 (2023) 106053. https://doi.org/10.1016/j.jobe.2023.106053.
[87] M. Zeljkovic, Metakaolin effects on concrete durability, University of Toronto, 2009.

Highlights

- A database of concrete expansion under external sulfate attack is built
- Ensemble Machine learning models predict the expansion due to the external sulfate attack
- XGBoost is the most precise ensemble model for sulfate attack expansion prediction
- Inputs importance and relative influence can be assessed by SHAP
- Time to reach a given expansion can be inferred by the models

CRediT authorship contribution statement

Benoit Hilloulin: Conceptualization, Methodology, Investigation, Software, Formal analysis, Validation, Writing -original draft, Writing - review & editing, Supervision.

Abdelhamid Hafidi: Investigation, Software**,** Formal analysis, Writing -original draft.

Sonia Boudache: Validation, Writing -original draft, Writing - review & editing, Supervision.

Ahmed Loukili: Supervision, Project administration, Funding acquisition