Towards accelerating the development of calcined clay cements: data-driven prediction of compressive strength exploiting machine learning algorithms

Yassine El Khessaimi, Youssef El Hafiane, Claire Peyratout, Karim Tamine, Samir Adly, Moulay Barkatou, Agnès Smith

Towards accelerating the development of calcined clay cements: data-driven prediction of compressive strength exploiting machine learning algorithms

Yassine El Khessaimi [(1) (2)]*, Youssef El Hafiane [(1)], Agnès Smith [(1)], Claire Peyratout [(1)], Karim Tamine [(2)], Samir Adly [(2)], Moulay Barkatou [(2)]

[(1)] Institute of Research on Ceramics, IRCER, UMR CNRS 7315, University of Limoges, 12 rue Atlantis, 87068 Limoges, France.

[(2)] MathIS, XLIM Laboratory, UMR CNRS 7252, University of Limoges, 123 Av. Albert Thomas, 87000 Limoges, France.

*Corresponding author: Yassine El Khessaimi (yassine.el-khessaimi@unilim.fr)

1. Introduction

The development of calcined clay cement has been receiving a particular attention from industrials and scientific researchers as an option for decreasing the CO_2 emissions coming from production of Portland cement (PC). The goal is to develop novel cements, in which the percentage of clinker is reduced as much as possible, or completely eliminated, while having materials consolidated with hopefully equivalent performances to Portland cements in terms of mechanical properties and durability. These cements consist of a mixture of clinker with gypsum, calcined clay and limestone [1–3]. The major interest of these new binders lies in the substitution of clinker by calcined clay [4]. The clays are calcined between 600 and 900 °C, while the manufacture of Portland clinker requires heat treatment at about 1450 °C [3,5]. As a result, energy consumption and CO_2 emissions are reduced by 35 to 40% [6,7].

It is highly difficult to model mechanical properties of calcined clay cements using empirical models, because of the complex and dynamic behavior of cement hydration, adding to that, the not complete comprehension of pozzolanic reactivity [8]. But, machine learning (ML) algorithms demonstrate high prediction performance for Portland cement. It should be the

case for calcined clay, since ML need data as feeding input without deep theoretical considerations [9,10].

In the last decades, data-driven methods such as ML and artificial intelligence were the solution to predict many materials properties knowing their composition and their synthesis parameters. ML approach consists of feeding an algorithm with input data to carry out models allowing the optimization of materials properties. ML was applied in many studies dedicated to predict properties of cementitious materials. This research topic can be considered as an emerging area of research. According to published documents indexed in Scopus concerning the two key words “machine learning” & “cement” [11], it can be noted that the number of articles published has increased significantly starting from 2018, to reach 123 published documents in 2021 (Figure 1). Then, to date, the peer-reviewed journal in which the largest number of papers are published is ‘Construction & Building Materials’ with a dozen of articles only in 2022 (Figure 2). Finally, at the international scale, China, followed by the United States and India are the largest providers of scientific articles in the field of exploiting ML to study cementitious materials (Figure 3). European countries also show significant activity.

The application of ML approach requires good comprehension of the problem and expert choice of data size and models. Several authors applied ML to predict properties of Portland cement or alkali-activated materials [12–36,36–43]. They handled different databases and parameters. These parameters are summarized in Appendix A. The size of the database is the first parameter to be considered. This size varies from tens to thousands of samples. In these studies, the most used ML algorithms were support vector regressor (SVR), random forest (RF), XGboost and artificial neural network (ANN). For example, to predict the plastic viscosity of cement, Sathyan et al. [12] used the XGboost algorithm applied to a dataset containing 252 experimental formulations of blended PC. The input parameters were cement amount, superplasticizer amount, and temperature. The ML predictions were very close to the observed results. The hydration heat of cement materials was as well the subject of ML

application. Cook et al. [25] used the random forest algorithm for predicting heat of hydration based on 10 input parameters. The database consisted of 7800 calorimetric measurements. The results showed high accuracy obtained by the RF algorithm expressed by a determination coefficient of 0.93.

From the data given in Appendix A, it is worth noting that input parameters cover a broad spectrum: (i) chemical and mineralogical composition; (ii) physical properties of cement (specific surface area, microstructure and grain size); (iii) rheology of cement paste; and (iv) hardening conditions. Target parameters to judge the performance of consolidated materials are essentially: (i) composition of the hydrated cement; (ii) microstructure; (iii) compressive strength. The parameter which has been the object of prediction by ML algorithms in most of articles is the compressive strength (Appendix A), which is a justified fact. For structural purposes, the compressive strength of cement material is usually the parameter to be evaluated in the first rang before rheology and durability [44,45].

The rapid development of calcined clay cements will help to achieve as quickly as possible carbon neutrality in the cement sector. The main interest of this work is to contribute to this development through the application of novel prediction paradigm. To meet this task, the present paper aims to apply ML approach, especially supervised regression algorithms, to predict compressive strength of calcined clay cements. Then, to investigate how raw materials composition, calcination conditions of clays and hardening environment, influence the compressive strength.

2. Method

2.1. Experimental database

The size and the quality of the dataset are significant for the accuracy of the ML model [46]. An experimental database of 323 mix design (10692 data values), containing partial replacement of Portland cement with calcined clay and limestone, was compiled from

previous studies that were reported in the literature [1,2,47–57]. Data splitting is a usually used method for model validation, where the dataset is split into two separate parts: the first for training, and the second for testing and validation [58]. The data was randomly partitioned into training, testing and validation sets: 80% of the data was used for training and testing (258 samples) and the remaining 20% was used for validation (65 samples). This data splitting ratio (80/20) is the most adopted according to previous works in literature (Appendix A) [12–36,36–43].

The collected database is destined to link the chemical composition of raw materials, calcination conditions of the clay and hardening conditions, with the resulting compressive strength of the mortar. The corresponding data presented in literature are mostly unstructured. For example, the compressive strength values are presented in the form of histogram graphs, so the extraction of these values was carried out manually. Henceforth, during structuring the database different selection criteria were respected. This selection approach was involved by Zhang et al. [18] for other type of cement materials:

- Only articles published in journals indexed in Scopus database have been retained for data collection.
- A detailed chemical composition of binders must have been clearly presented.
- The mix proportion of each component of the low carbon cement must have been provided.
- Calcination conditions and powder fineness of the involved clay must have been provided.
- A detailed description of hydration and hardening conditions must have been presented.
- Compressive strength of the mortar must have been established according to a normalized protocol (for example: EN 197 or ASTM C109). Mortars were prepared using a sand to cement ratio equal to 3:1.

Given the considerable number of oxides in each component of the calcined clay cement, the dataset has a large dimension, which affects the accuracy and robustness of the ML models. Hence, the chemical compositions of calcined clay, limestone and Portland cement are substituted with their reactivity ratios [59]. Chemical analysis of clinker, limestone and clays are commonly expressed in terms of weight percentages of oxides, but it is often useful to employ quantities derived from these percentages [18,59,60]. In the equations that follow, chemical formulae also denote weight percentages. The following parameters are widely used:

$$RR = \frac{CaO + MgO + Al_2O_3}{SiO_2} \quad (1)$$

$$AR = \frac{Al_2O_3}{Fe_2O_3} \quad (2)$$

$$SR = \frac{SiO_2}{Al_2O_3 + Fe_2O_3} \quad (3)$$

$$HR = \frac{CaO}{SiO_2 + Al_2O_3 + Fe_2O_3} \quad (4)$$

It is worth noting that RR (reactivity ratio) and HR (hydraulic ratio) are generally applied to evaluate hydraulic binding properties, while AR (alumina ratio) and SR (silica ratio) are applied to evaluate pozzolanic properties [18].

The reactivity ratio of each cement mix was defined through the weighted average as follows:

$$RM = \frac{\sum_{i=1}^{n}(RR_i.wr\%)}{100\%} \quad (5)$$

$$AM = \frac{\sum_{i=1}^{n}(AR_i.wr\%)}{100\%} \quad (6)$$

$$SM = \frac{\sum_{i=1}^{n}(SR_i.wr\%)}{100\%} \quad (7)$$

$$HM = \frac{\sum_{i=1}^{n}(HR_i.wr\%)}{100\%} \quad (8)$$

where $wr\%$ is the weight percentage of each constituent in the calcined clay cement. The constituents are calcined clay, limestone and Portland cement.

2.2. Description of the data features

From the collected data [1,2,47–57], 14 input features are considered. The inputs contain the weight percentage of clay, OPC, and limestone, chemical composition of each constituent expressed in terms of reactivity ratios, calcination conditions of the clay, and hardening conditions. For the output, there is one targeted feature which is the compressive strength of the low carbon cement. Table 1 shows units and statistical parameters of the features. The analysis shows that there is no outlier or aberrant values. Thereby the data can be used as it is for the next steps of machine learning application.

2.3. Machine learning models

- Linear Regression (LR)

Linear regression is a regression model in which the target value is expected to be a linear combination of the features noted x_1 to x_p [61,62]. In mathematical notation, $\hat{y}$ is the predicted value (equation (9)):

$$\hat{y}(w, x) = w_0 + w_1 x_1 + \cdots + w_p x_p \qquad (9)$$

The vector $w = (w_1, \ldots, w_p)$ is the model coefficients and w_0 as intercept value of the model.

Ordinary Least Squares method aims to fit a linear model with coefficients $w = (w_1, \ldots, w_p)$ to minimize the residual sum of squares between the observed targets in the dataset, and the targets predicted by the linear approximation. Mathematically it solves a problem of the form given by equation (10):

$$\min_w \|Xw - y\|_2^2 \qquad (10)$$

It is worth noting that the coefficient estimated for Ordinary Least Squares method relies on the independence of the features.

- Linear Regression with the Ridge Regularization (LR-RR)

Ridge regression addresses some of the problems of Ordinary Least Squares method by imposing a penalty on the size of the coefficients [61,63]. The ridge coefficients minimize a penalized residual sum of squares given by equation (11):

$$\min_w \|Xw - y\|_2^2 + \alpha \|w\|_2^2 \qquad (11)$$

The complexity parameter $\alpha \geq 0$ controls the amount of shrinkage. The higher the value of α, the larger is the amount of shrinkage.

- Support Vector Regressor (SVR)

The aim of SVR is to seek the hyperplane which will separate at best, then to try to estimate the target values y_i. It amounts to determine $w \in \mathbb{R}^p$ and $b \in \mathbb{R}$, such as [61,64]:

$$|\langle w, x_i \rangle + b - y_i| \leq \varepsilon \qquad (12)$$

where $\varepsilon > 0$ is a small positive value. Then (w, b) is determined to minimize:

$$\frac{1}{2}\|w\|^2 \qquad (13)$$

under the constraints

$$|y_i - \langle w, x_i \rangle - b| \leq \varepsilon, i = 1, \dots, n \qquad (14)$$

The constraints imply that all observations must be defined in a margin of 2ε. This hypothesis may lead the user to use large ε values, and consequently prevent the solution to adjust the scatter points. To overcome this, a spring variable is introduced to allow certain observations to be outside the margin. The problem to be solved then comes to finding (w, b, ξ, ξ^*) that minimizes:

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \qquad (15)$$

Here again the parameter C will have to be calibrated and the kernel method is used.

Under the constraints

$$\begin{cases} y_i - \langle w, x_i \rangle - b \leq \varepsilon + \xi_i, i = 1, \dots, n \\ \langle w, x_i \rangle + b - y_i \leq \varepsilon + \xi_i^*, i = 1, \dots, n \\ \xi_i \geq 0, \xi_i^* \geq 0, i = 1, \dots, n \end{cases} \qquad (16)$$

The solution vector w^* is written as a linear combination of support vectors:

$$w^* = \sum_{i=1}^{n} (\alpha_i^* - \alpha_i) x_i \qquad (17)$$

The supporting vectors are the observations verifying $\alpha_i^* - \alpha_i \neq 0$

- Decision Trees for Regression (DTR)

Decision Trees for Regression (DTR) are a supervised learning method used for regression [61,65]. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. The algorithm used to train a decision tree is called CART, for Classification And Regression Tree. It is an algorithm for partitioning space through a gluttonous, recursive and divisive approach [66,67]. At each node of a decision tree built by CART corresponds a splitting variable $j \in \{1, \ldots, p\}$ according to which the data will be partitioned [68]. This separator variable defines two regions, corresponding to the children of the node in question [69,70].

In the case where the separation variable is a real variable, it is then accompanied by a splitting point (s) which is the value of the attribute against which the decision will be made. The two regions are then:

$$R_l(j, s) = \{\vec{x} : x_j < s\}; R_r(j, s) = \{\vec{x} : x_j \geq s\} \qquad (18)$$

At each iteration of the CART algorithm, all possible values of j and, if applicable, all possible values of (s) are used to determine (j, s) that minimizes a predefined criterion.

This criterion is usually the mean square error. So, the variable and the splitting point were chosen as:

$$\arg\min_{j,s}\left(\sum_{i:\vec{x}^i\in R_l(j,s)}\left(y^i-y_l(j,s)\right)^2+\sum_{i:\vec{x}^i\in R_r(j,s)}\left(y^i-y_r(j,s)\right)^2\right) \quad (19)$$

were $y_l(j,s)$ (respectively $y_r(j,s)$) is the label value associated with the $R_l(j,s)$ region (respectively $R_r(j,s)$) at this stage, which is the average of the labels in this region.

- Random Forest (RF)

The “random forest” algorithm performs a parallel learning on multiple decision trees built randomly and driven on different subsets of data. The ideal number of trees can go up to several hundred or more [66,71]. Practically, each tree in the random forest is trained on a random part of data according to the principle of bagging, with a random part of features according to the principle of «random projections». Predictions are then averaged when the data are quantitative, as it is the case for regression [61,72].

Let us assume that n samples are randomly collected from S_n(the training set) with a probability of selection $1/n$ for each sample. These n samples are called bootstrap sample S_n^θ, where θ is an independently distributed vector. Assume that q bootstrap samples $(S_n^{\theta 1}, S_n^{\theta 2}, \dots, S_n^{\theta q})$ are chosen using the bagging algorithm and that q regression trees are trained on the subsets $\hat{h}\left(X, S_n^{\theta 1}\right), \hat{h}\left(X, S_n^{\theta 2}\right), \dots, \hat{h}\left(X, S_n^{\theta q}\right)$. The q outputs are: $\widehat{Y_1} = \hat{h}\left(X, S_n^{\theta 1}\right), \widehat{Y_2} = \hat{h}\left(X, S_n^{\theta 2}\right), \dots, \widehat{Y_q} = \hat{h}\left(X, S_n^{\theta q}\right)$. The final output is the average value of the q outputs. The illustration of this concept is shown in Figure 4.

- eXtreme gradient boosting (XGboost)

XGboost algorithm consists on growing sequentially decision trees; each tree is grown using information from previously grown trees to progress its performance [61,73].

In the case of XGboost, the learning objective consists of two parts: the loss function and the regularization term. Mathematically, XGboost's learning objective may be defined as follows:

$$obj(\theta) = l(\theta) + \Omega(\theta) \quad (20)$$

Here, $l(\theta)$ is the loss function, which is the Mean Squared Error (MSE) for regression, and $\Omega(\theta)$ is the regularization function, which is a penalty term to prevent over-fitting.

The loss function, defined as the MSE for regression, can be written in summation notation, as follows:

$$l(\theta) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (21)$$

Here, y_i is the target value for the i[th] row and $\hat{y}_i$ is the value predicted by the machine learning model for the i[th] row.

Let w be the vector space of leaves. Then, f, the function mapping the tree root to the leaves, can be recast in terms of w, as follows:

$$f_t(x) = w_{q(x)}, w \in R^T, q: R^d \rightarrow \{1,2,\dots,T\} \quad (22)$$

Here, q is the function assigning data points to leaves and T is the number of leaves. XGboost settles on the following as the regularization function, where γ and λ are penalty constants to reduce overfitting:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \quad (23)$$

- Multi-layered perceptron (MLP)

Artificial neural networks are basically systems inspired by the functioning of biological neurons [74]. The most famous of these is the multi-layered perceptron (MLP), an artificial system capable of learning by the experience. It was introduced in 1957 by Franck

Rosenblatt, it has only been used since 1982 after its improvement [75,76]. Thanks to the computational power of the 2000s, the perceptron is used in a set of neurons organized in layers (Figure 5). From one layer to another, the input signal propagates to the output, activating or not as the neurons progress.

The output formula of a hidden neuron will therefore always be of the form [61,77]:

$$y = f_{activation}\left(b + \sum_i w_i \,.\, x_i\right) \qquad (24)$$

where w_i are the weights of the system. In practice, they are randomly initialized when the neural network is created, and b refers to the bias.

The contribution of an input neuron will be $y = x$

that the contribution of an output neuron will be $y = \sum_i w_i \,.\, x_i$

The activation function decides whether a neuron should be activated or not. Thereafter some activation functions applied for the MLP algorithm:

- Sigmoid/Logistic function

$$f(x) = \frac{1}{1 + e^{-x}} \qquad (25)$$

- Tanh function

$$f(x) = \frac{(e^{x} - e^{-x})}{(e^{x} + e^{-x})} \qquad (26)$$

- Rectified Linear Unit (RELU) function

$$f(x) = \max(0, x) \qquad (27)$$

2.4. Hyperparameter tuning and K-fold cross-validation

Hyperparameters are tunable parameters that allow the control of the model training process. Hyperparameter tuning is the method of finding the hyperparameter configuration that produces the best performance of the ML algorithm. In the present work, Bayesian optimization method was used to tune hyperparameters [78]. The search intervals for hyperparameters tuning used in the present work are given in Table 2.

Bayesian optimization needs a prior knowledge to guess the hyperparameter [61]. It is based on a probabilistic model matching hyperparameters with a probability function of a score on the objective function. These probability functions are defined below:

$$P(score|hyperparameters) = \frac{P(hyperparameters|score)\ P(score)}{p(hyperparameters)} \qquad (28)$$

This function is also called “surrogate” of the objective function. It is much easier to optimize than the objective function. Below are the stages for applying Bayesian optimization for hyperparameter tuning:

1- Construct a substitution probability model of the objective function;
2- Determine the hyperparameters of the surrogate model;
3- Apply these hyperparameters to the original objective function;
4- Evaluate the surrogate model using new results;
5- Repeat stages 2,3 and 4 up to a defined iterations number.

To check whether the developed machine learning model is efficient enough to predict the outcome of a test data set, performance evaluation of the applied machine learning model was carried out using the K-fold cross-validation. This technique is basically a method of resampling the dataset in order to evaluate ML trained algorithm and to prevent overfitting and underfitting [79]. In this technique, the parameter K is the number of different subsets into which the given data set should be divided. Additionally, K-1 subsets are used to train

the model and the left-out subsets are used as the testing set. An illustration is given in Figure 6.

2.5. Model performance evaluation

The predictive performance of the applied machine learning algorithms can be evaluated using three different indicators [80], which are defined as:

- Coefficient of determination (R^2)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(P_i - T_i)^2}{\sum_{i=1}^{n}(T_i - \bar{T})^2} \quad (29)$$

- Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(P_i - T_i)^2}{m}} \quad (30)$$

- Mean Absolute Error (MAE)

$$MAE = \frac{\sum_{i=1}^{n}|P_i - T_i|}{n} \quad (31)$$

where P_i and T_i are the Predicted and Tested compressive strength values, respectively; $\bar{T}$ is the mean value of all the tested values; n is the total number of samples in the dataset.

The metric R^2 shows the extent of the linear correlation between the predicted and tested values. The closer R^2 is to 1, the better is the performance of the model. RMSE provides information on the deviation between the predicted values and the tested values. MAE indicates the prediction error. Among the three performance indicators, R^2 is mainly used as the representative quantity to discuss the performance of the trained models in this study.

2.6. Experimental scenarios

Different experimental scenarios were designed to simulate real cases. The goal is to predict compressive strength of different mixes of low carbon cements using the best performed ML algorithm. These scenarios have been organized in such a way to evaluate quantitatively the effect of inputs on the compressive strength. The evaluated input parameters were: (i) weight fraction of each constituent of the calcined clay cement, (ii) chemical composition of the clay, (iii) calcination conditions of the clay, (iv) and hardening conditions. It leads to the design of four scenarios. Each one contains variable inputs to be evaluated and fixed inputs. Table 3 represents this designation.

3. Results and discussion

3.1. Linear correlation between variables and feature importance analysis

The statistical correlations between the input features and the target are shown in Table 4. The analysis results show that the features named age_D (hardening age) and OPC% (weight percentage of OPC) have the strongest positive correlation with the compressive strength as expected, with correlation coefficients equal to 0.50 and 0.24, respectively. This suggests that the compressive strength can possibly increase as the hardening age and weight percentage of OPC increase. The input feature named CL% (weight percentage of calcined clay) has a negative correlation of -0.21 with the compressive strength, indicating that the increase of CL% has likely a negative effect on the compressive strength. Besides, there is some weak correlation between the other features with the compressive strength.

Even if statistical correlations are useful for initial data analysis, they do not show the non-linear dependencies between the different features and the compressive strength which is extremely important for prediction studies and that’s why ML algorithms were adopted.

Feature importance analysis denotes to techniques that allocate a score to input features based on how useful they are at predicting a target variable [81]. For the case of linear regression algorithms, the feature importance score is merely the coefficients of the model. These coefficients can be used directly as a basic type of feature importance score. Based on the linear regression with Ridge regularization algorithm, the features importance is calculated and shown in Figure 7-a. The results show that RM reactivity ratio and W/B are the most important features to predict the compressive strength, whereas the hardening age (age_D) was shown as the feature with the least importance (Figure 7-a). These results do not fit with our common knowledge on cementitious materials, nor with correlation results (Table 4). It is well known that hardening age is the most impacting parameter on the development of compressive strength. This unpredictable result shown in Figure 7-a, can be explained by the difference in the standard deviation of the features (Table 1). For example, the standard deviation value of the hardening age is 41.3 MPa and that of the reactivity is 0.5, indicating a large difference between the two values, which induce errors during feature importance analysis. To correct this mismatch, all the features were rescaled using the following formula:

$$Z = \frac{X - \mu}{\sigma} \qquad (32)$$

where Z is the rescaled variable, X is the original variable, μ is the average, and σ is the standard deviation.

The normalization assumes that the input variables have the same scale, and the results of feature importance calculation can then well be interpreted. The feature importance analysis after rescaling is given in Figure 7-b. From the figure, it can be observed that hardening age (age_D) is the most important variable affecting compressive strength. This is in alignment with the practical knowledge related to cementitious materials. The second important identified feature are alumina ratio (AM), reactivity ratio (RM) and calcination duration of the

clay (time_calcin(h)). The weight fraction of OPC (%OPC) is ranked as the third important input variable. From this figure, it can be observed that the hardening relative humidity (RH_cure), water to binder ratio (W/B), calcination temperature (T_calcin) and hardening temperature (T_cure) are the least important variables. Lastly, hydraulic ratio (HM), silica ratio (SM), weight fraction of calcined clay (CL%) and weight fraction of limestone (CC%) have a negative importance on the prediction of compressive strength. These results offer new perceptions on the parameters affecting the compressive strength of calcined clay cements and open the door to further studies to better understand the governing mechanisms.

3.2. Predictive performance of the ML algorithms

The hyperparameters of the seven selected ML algorithms are optimized using the Bayesian optimization method coupled with K-fold cross-validation. Since the iteration number is an interesting hyperparameter for MLP to optimize the calculation cost, Figure 8 shows the convergence of the prediction error using the MLP algorithm. As the iteration number increases from 1 to 5, the loss function of compressive strength decreases from 870 to 98. As the iteration number further increases from 5 to 100, the loss function decreases from 98 to 0. Such tendency shows that 100 is a reasonable number of iterations to achieve convergence for the MLP loss function.

Figure 9 shows the error metrics of the seven selected ML algorithms including their tuned hyperparameters. The linear regression (LR, LR-RR) and SVR models show lower R^2 with higher RMSE and MAE compared to DTR, RF, XGboost and MLP models. During the training stage, the DTR and MLP models show $R^2 = 1$ with RMSE = MAE = 0 indicating an overfitting problem of data training, thereby they are eliminated for further predictions. Overfitting problems usually occur when very complex models are used with a lack of generalization for simple problems. To detect a model that overfit, it offers very good performance on training data, but does not succeed when it is actually in production, as it

was the case for DTR and MLP (Figure 9). Higher R^2 values with lower RMSE and MAE values for training and testing were found for the RF and XGboost models. In particular, XGboost can be distinguished as the best performing model with high determination coefficients: $R^2_{training}$ = 0.99 and $R^2_{testing}$ = 0.95. XGboost shows a very close link between predicted and actual values with $RMSE_{training}$ = 2.21 MPa and $RMSE_{testing}$ = 11.4 MPa. And it exhibits a high accuracy with $MAE_{training}$ = 1.1 MPa and $MAE_{testing}$ = 2.5 MPa.

Figure 10 presents a comparison of predicted compressive strength and measured values during the testing stage. The red line refers to the perfect fit line, that is predicted compressive strength is equal to actual value. Compared to the other models, the graph corresponding to XGboost illustrates clearly a majority of points close to the perfect fit line (the red line), confirming visually the best prediction performance character of the XGboost model in comparison to the other models.

Since there is an absence of works dealing with the prediction of compressive strength of calcined clay cements, it can be difficult to compare the prediction performance results of our work with previous studies from literature. However, if the comparison is exclusively made with previous works dealing with alkali-activated materials or Portland cements, it can be worth nothing that our results agree with almost previous studies. For the prediction of compressive strength of alkali-activated cementitious materials, Zhang et al. [18] revealed that the XGboost algorithm (676 samples) performed better than SVR and DTR, with an $R^2_{training}$ = 0.98 and $R^2_{testing}$ = 0.94. Gomaa et al. [82] found that RF algorithm (202 samples) performed well for the prediction of compressive strength in the case of alkali-activated concretes. This means that although a large set of data samples might produce a better generalization of the model, the prediction accuracy can be different. In fact, the selection of adequate hyperparameter optimization method can be a crucial parameter in improving the accuracy of the algorithm. Besides, boosting algorithms confirms their high prediction accuracy for the prediction of compressive strength of OPC-MK cement mortars. Asteris et

al. [83] found that AdaBoost algorithm (424 samples) shows $R^2_{training}$ = 0.99 and $R^2_{testing}$ = 0.94, performing better than SVR, DTR and RF.

Among the seven algorithms selected in our case, XGboost algorithm presents the best prediction performance during training and testing, and thus it is selected for further prediction of compressive strength in different experimental scenarios.

3.3. Prediction of compressive strength of different experimental scenarios

- Effect of weight fraction of each constituent of the calcined clay cement (scenario 1)

Figure 11 shows the effect of weight fraction of each constituent of the calcined clay cement on the predicted compressive strength at 3, 7 and 28 days using the XGboost model. The different colors indicate the values of predicted compressive strength, with a mean absolute error of 2.5 MPa. It can be observed that among the input variables, the weight fraction of OPC was found generally to be the most impacting on the variation in compressive strength at 3, 7 and 28 days. By increasing the weight fraction of OPC from 30 to 55 wt.%, the compressive strength can increase from 21 to 26 MPa at 3 d (Figure 11-a), from 26 to 45 MPa at 7 d (Figure 11-b), and from 32 to 52 MPa at 28 d (Figure 11-c). This result is in good agreement with many previous studies [47,48,54] because the major part of compressive strength development is coming from the hydration of the clinker cementitious phases, namely alite, belite, alumino-ferrite and tricalcium aluminate [84]. It is worth noting that in the region of composition corresponding to 55–75 wt.% of OPC, 20–35 wt.% of calcined clay, and 0–15 wt.% of limestone (Figure 11), we can assign a significant increase of compressive strength. It reaches a value of 59 MPa at 28 d for 65 wt.% of OPC, 30 wt.% of calcined clay and 5 wt.% of limestone. This region of composition corresponds to the LC3 cement composition as expected. The increase of compressive strength can be explained by the presence of calcined kaolin, which through a synergetic reaction with limestone enables a denser microstructure, and consequently a higher compressive strength [5,53,84,85]. During hydration of LC3, additional reactions occur compared with OPC. Metakaolin, the product of

kaolinite calcination, reacts as a pozzolanic material, consuming Portlandite and forming mainly calcium aluminum silicate hydrate (C-A-S-H). Limestone also reacts with the calcium aluminate coming from clinker to form carboaluminate hydrates. The formation of these additional fine hydrate phases leads to a denser microstructure showing higher compressive strength [85].

In addition, in either hardening age at 3, 7 or 28 d, there is a drop of compressive strength when weight fraction of limestone ranges between 20 and 30 %, and that of calcined clay is between 20 and 40%. In comparison with experimental values of compressive strength reported in literature, the predicted compressive strength values of the present work have the same order of magnitude. For example, Antoni et al. [1] tested a mix containing 40 wt.% Metakaolin, 40 wt.% OPC and 20 wt.% limestone, and they found a compressive strength of 46 MPa at 28 days. Our predicted values for this mix is 47$\pm$ 2.5 MPa at 28 days, which confirms the prediction performance of the XGboost algorithm.

- Effect of calcination conditions of clay (scenario 2)

Given the other input features fixed, the effect of calcination conditions of clay on the compressive strength at 3, 7 and 28 d are given in Figure 12. In the calcination temperature range between 600 and 920 °C, the compressive strength shows a slight downward trend at 3 and 7 d. At 28 d, however, a drop of values is observed for a calcination temperature between 680 and 825 °C (Figure 12-c). This drop of values can be neglected considering the mean absolute error of 2.5 MPa. Above 920 °C, the compressive strength deceases drastically by 5 to 7 MPa at all hardening ages, and visually observing. This decrease corresponds to the changeover of bar color from green to blue at the three hardening ages (Figure 12). According to Rashad et al. [86], the best calcination temperature of kaolin in order to obtain amorphous metakaolin is likely in the range of 600 to 850 °C for heating period of 2–5 h. Extending calcination temperature above 920 °C of the metakaolin does not noticeably improve its pozzolanicity [87]. High calcination temperature can lead to the

formation of non-reactive phases, namely mullite, resulting from structural rearrangement and recrystallization of the amorphous phases [88]. Based on solid-state NMR Studies, the reactivity of metakaolinite is at a maximum at 750-800 °C when the population of hexacoordinate Al is at a minimum and tetra- and pentacoordinate populations at a maximum [89]. Concerning the calcination duration, there is no significant effect on compressive strength at 3 and 7 d. Whereas at 28 d, the effect becomes almost more important. For example, at 800 °C (Figure 12-c), the compressive strength decreases continuously from 58 to 50 MPa when the duration goes from 3 to 0.5 h.

- Effect of chemical composition of calcined clay (scenario 3)

To evaluate the influence of chemical composition of clay on the compressive strength, Figure 13 depicts predictive compressive strength according to cement reactivity ratios. To avoid message ambiguity, only results at 28d are presented. It is noted that in this simulated experiment, the variation in reactivity ratios values is induced exclusively by variation in the clay chemistry, as the chemical composition of OPC and limestone were fixed (Table 5). At 28 d, the higher compressive strength values (48 to 49.7 MPa) are attained for HM = 1.6–1.7 (Figure 13-a and c), SM = 1.8–2 (Figure 13-a), RM = 2.9–3.1 (Figure 13-b) and AM = 4–4.5 (Figure 13-b and c). By considering an analysis on overall values of compressive strength including low values, it can be obviously noted that predicted compressive strength at 28 d decreases by increasing SM and HM (Figure 13-a and c), and it shows an upward trend by increasing AM (Figure 13-b and c), without any flagrant trend for RM variation (Figure 13-b). The depletion of compressive strength by increasing HM ratio, can be explained by the negative effect of CaO on the development of compressive strength. In fact, calcareous clay is often regarded as not suitable for blended cements, explained by the decomposition of $CaCO_3$ to CaO after calcination. During service, CaO may react with water forming $Ca(OH)_2$ which can result in swelling and cracks [90]. However, substituting MK by 10–20% of calcined marl (containing 40 wt.% CaO) can be effective, because calcium hydroxide which is formed during hydration of this CaO reacts with dehydroxylated clay minerals through

pozzolanic reaction to form more binding hydrates [91,92]. This later case corresponds to HM between 1.6 and 1.7 (Figure 14-a and c).

Briki et al. [84] observe a slowdown of metakaolin reaction degree when it is replaced with silica fume, which can possibly explain the decrease of compressive strength by increasing SM moduli. On another hand, the enhancement of compressive strength by increasing AM ratio can possibly explained by the considerable contribution of alumina-rich calcined clays on the development of compressive strength. It was shown that the pozzolanic behavior of metakaolin is affected by its chemical and mineralogical composition, notably (SiO_2+Al_2O_3+Fe_2O_3) content, calcination temperature and duration which fix the degree of amorphousness/dehydroxylation, morphology and size of the clay particles [93–96].

- Effect of hardening conditions (scenario 4)

Figures 14 presents the effect of hardening conditions on the predicted compressive strength at (a) 3, (b) 7 and (c) 28 d. In the range of W/B between 0.4 and 0.6, the graphs indicate that the compressive strength is quite influenced by water content. In general, adding more water to the concrete mixture leads to an excess of free water when the mortar hardens. This free water will create a higher porosity, leading to a reduction in compressive strength. Briki et al. [84] found that the reaction of metakaolin in calcined clay slows down after 28 days with the water to binder ratio of 0.6. Similar effect is observed for hardening temperature, in the range of 14–30 °C (Figure 14), there is no significant variation in compressive strength.

4. Conclusion and perspectives

This study presented the application of machine learning models on the prediction of compressive strength of calcined clay cement. For this purpose, seven algorithms of supervised machine learning were exploited to predict the compressive strength. The models were trained and tested with an experimental database of 323 mix design.

The results show that rescaling of input features is indispensable for a good interpretability of the feature importance analysis. Among the evaluated models, XGboost model was identified as the most accurate for the prediction. K-fold cross validation and Bayesian optimization method were combined to find the optimal hyperparameters of the developed XGboost model. Further, these novel machine learning results reveal that once the model is correctly trained, prediction of compressive strengh of calcined clay cement with different mixture design is possible, with mean absolute error of 2.5 MPa.

The deployment of the XGboost model shows that alumina to silica ratio (alumina ratio) of clays is the most impacting input feature on compressive strength of calcined clay cements. Then, interesting composition domains of raw materials were identified for further optimization of calcined clay cements. Prediction of compressive strengh exploiting machine learning approach show a downtrend of strength above 920 °C, which confirms experimental results published in literature.

The original work within this paper denotes the successful feasibility of machine learning approach to predict the compressive strength of calcined clay cement which represents an interesting initial milestone for design of potential compositions. Additionally, our future studies will deal with the prediction of rheology and durability properties exploiting machine learning which constitutes a potential opportunity to improve the design optimization of this low carbon cement.

Appendix A.

Data available in the literature concerning the application of ML for cements. The meanings of the abbreviations are given below this table*.

Size of the dataset	Training data (%)	Testing data (%)	Cement type	ML algorithm	Input parameters	Target parameters	Réf.
252	85	15	OPC	XGboost	Amount of cement Amount of superplasticizer	Shear flow	[12]
93	80	20	OPC	SVR	Surface exposed to the acid pH Hydration age Acid concentration	Compressive strength	[13]

130	70	30	OPC + CPB	ANN	Chemical composition Particle size Input velocity Amount of solid Ratio cement/mine tailing Density Coefficient of curvature Coefficient of uniformity	Pressure drops	[14]
-	-	-	CEM II/A-LL 32.5R	ANN	Cracking Patterns	Image segmentation	[15]
-	90	10	OPC	ANN-SVM	Amount of cement Amount of water Amount of dispersant Hardening age	Compressive strength	[97]
50	-	-	Foam cement slurries	ANN	Amount of foam	Electrical resistivity	[17]
676 (shared dataset)	-	-	Alcali-activated material	SVR, RFR, ETR, GBR	Hardening age Hardening humidity Water-to-cement ratio Chemical composition	Compressive strength	[18]
215	70	30	Clinker	Multiple kernel learning	Sample weight Furnace temperature Gas temperature Bogue's modules	Amount of free lime	[98]
192	70	30	CEM I 42,5N + FA	ANN- PSO	Chemical composition Hardening age Amount of FA	Compressive strength	[20]
90	70	30	Class G Cement	LM	Nanoclay fraction Temperature Shear rate	Apparent viscosity Shear stress Plastic viscosity Yield point	[99]
2416	75	25	OPC OPC +	SVR	Chemical composition by EDS	Nano-indentation ratio	[22]

			Pozzolans				
31000	-	-	OPC	DTE	Chemical composition Physical tests (ASTM)	Setting time Compressive strength	[100]
5808 (shared dataset)	97	3	OPC + MK + $CaCO_3$	CART-RF	Chemical composition	Hydration heat	[24]
8112	96	4	OPC + MK + $CaCO_3$ + Quartz	RF	Hardening age Specific Surface area Amount of additive Type of additive Amount of cement	Hydration heat Cumulative hydration heat	[101]
713	70	30	Blended OPC	GBRT model	Density Chemical composition Formulation Compressive strength Exposure time	Carbonation depth	[26]
154	80	20	OPC + PCM	RFR ETR GBR XGBR	Melting temperature PCM content Latent heat Chemical composition Hydration age Temperature	Compressive strength	[27]
-	80	20	Blended OPC + Soil	ANN	Type of cement Type of soil Composition of the mixture cement-soil Hardening age Plasticity index	Compressive strength	[28]
1030	80	20	OPC	Ada-Boost predictive model	Chemical composition Formulation	Compressive strength	[29]
250	70	30	Blended OPC + fibers	ANN	Amount of cement Type and amount of fibers Water-to-cement ratio	Self-healing capacity	[30]

					Size of crack Hardening age		
45 544	98	2	OPC Clinker	ELMARE	Coal feeding The baffle opening of the high-temperature fan	NO_x concentration O_2 concentration	[31]
751	75	25	Composite material based on OPC	ANN SVR CART XGboost	Sand-to-binder ratio Water-to-binder ratio Superplasticizer content Fiber length Fiber elastic ratio	Compressive strength Tensile strength Tensile strain capacity	[32]
132	90	10	OPC	MGGP MARS	Hardening age Depth of measured position Diffusion dimension Presence of reinforcement Chloride ion concentration	Chloride diffusion rate	[102]
1030	-	-	OPC	ANN SVM RF	Formulation Hardening age	Compressive strength	[34]
215	56	44	OPC + Soil + Fibers	ANN SVM RF MR	Formulation Atterberg limits Water content Hardening age Fiber content Mechanical properties of the fiber	Compressive strength Tensile strength	[35]
40	65	35	OPC	ANN SVM MLP	Chemical composition	Blaine Fineness	[36]
1200	-	-	OPC	ANN	3D Microstructures generated by HydratiCA software	Microstructure of hydrating tricalcium silicate	[37]
638	86	14	OPC	ANN SVM	Formulation Water-to-cement ratio Sulfate ions concentration	Compressive strength	[38]

					Exposure conditions		
304	70	30	OPC	LMBP-ANN	Chemical composition Blaine Fineness Temperature Water content	Compressive strength	[39]
114 (shared dataset)	80	20	CEM I 42.5R CEM I 52.5R + Carbone nanotubes	ANN SVM	Type of cement Amount of nanotubes Size of nanotubes Functionalisation method Hardening age Temperature Method of dispersion	Compressive strength Tensile strength	[40]
52	70	30	Blended OPC + Soil	ANN RF GP M5P tree	Cement content Soil content Amount of fly ash Hardening age	Compressive strength	[41]
50	60	40	OPC	ANN SVR RVM GPR	Chemical composition Blaine Fineness	Compressive strength	[42]
512	80	20	OPC + Silica microparticles	SVM	Water-to-cement ratio Sand-to-cement Silica nanoparticles/cement Silica microparticles/cement Hardening age Porosity	Compressive strength Tensile strength	[43]

***Abbreviations**

AAPE	Average absolute percentage error
ANN	Artificial neural network
CART	Classification and regression trees
CPB	Cemented paste backfill
DTE	Decision tree ensembles
ELMARE	Extreme learning machine autoregressive exogenous model
ETR	Extra trees regressor

GBR	Gradient boosting regressor
GPR	Gaussian process regression
LM	Levenberg Marquardt
LMBP	Levenbarg-Marquardt back-propagation
MAE	Mean absolute error
MAPE	Mean absolute percentage error
MARS	Multivariate adaptive regression splines
MGGP	Multi-gene genetic programming
MLP	Multi-layer perception
MR	Multiple regression
MSE	Mean square error
OPC	Ordinary Portland Cement
PSO	Particle swarm optimization
R2	Coefficient of determination
RF	Random forest
RFR	Random forest regressor
RMSE	Root-mean-square error
RRSE	Root relative squared error
RVM	Relevance vector machine
SOS	Symbiotic organism search
SVM	Support vector machine
SVR	Support vector regression
XGBR	eXtreme gradient boosting regression

Acknowledgement

This work is supported by institutional grants from the National Research Agency under the Investments for the future program with the reference ANR 10 LABX 0074 01 Sigma LIM.

References

[1] M. Antoni, J. Rossen, F. Martirena, K. Scrivener, Cement substitution by a combination of metakaolin and limestone, Cem. Concr. Res. 42 (2012) 1579–1589.
[2] R. Fernandez, F. Martirena, K.L. Scrivener, The origin of the pozzolanic activity of calcined clay minerals: A comparison between kaolinite, illite and montmorillonite, Cem. Concr. Res. 41 (2011) 113–122.
[3] F. Avet, K. Scrivener, Investigation of the calcined kaolinite content on the hydration of Limestone Calcined Clay Cement (LC3), Cem. Concr. Res. 107 (2018) 124–135.
[4] G. Rojo-López, S. Nunes, B. González-Fonteboa, F. Martínez-Abella, Quaternary blends of portland cement, metakaolin, biomass ash and granite powder for production of self-compacting concrete, J. Clean. Prod. 266 (2020) 121666.

[5] F. Avet, E. Boehm-Courjault, K. Scrivener, Investigation of CASH composition, morphology and density in Limestone Calcined Clay Cement (LC3), Cem. Concr. Res. 115 (2019) 70–79.
[6] M. Sharma, S. Bishnoi, F. Martirena, K. Scrivener, Limestone calcined clay cement and concrete: A state-of-the-art review, Cem. Concr. Res. 149 (2021) 106564.
[7] D. Zhang, B. Jaworska, H. Zhu, K. Dahlquist, V.C. Li, Engineered Cementitious Composites (ECC) with limestone calcined clay cement (LC3), Cem. Concr. Compos. 114 (2020) 103766.
[8] M. Cyr, P. Lawrence, E. Ringot, Efficiency of mineral admixtures in mortars: Quantification of the physical and chemical effects of fine admixtures in relation with compressive strength, Cem. Concr. Res. 36 (2006) 264–277.
[9] H. Van Damme, Concrete material science: Past, present, and future innovations, Cem. Concr. Res. 112 (2018) 5–24.
[10] B. Luzu, R. Trauchessec, A. Lecomte, Packing density of limestone calcined clay binder, Powder Technol. 408 (2022) 117702.
[11] https://www.scopus.com.
[12] D. Sathyan, D. Govind, C.B. Rajesh, K. Gopikrishnan, G.A. Kannan, J. Mahadevan, Modelling the Shear Flow Behaviour of Cement Paste Using Machine Learning–XGBoost, in: J. Phys. Conf. Ser., IOP Publishing, 2020: p. 012026.
[13] L. Wu, C. Hu, W.V. Liu, Forecasting the deterioration of cement-based mixtures under sulfuric acid attack using support vector regression based on Bayesian optimization, SN Appl. Sci. 2 (2020) 1–16.
[14] C. Qi, L. Guo, H.-B. Ly, H. Van Le, B.T. Pham, Improving pressure drops estimation of fresh cemented paste backfill slurry using a hybrid machine learning method, Miner. Eng. 163 (2021) 106790.
[15] M. Szeląg, Application of an automated digital image-processing method for quantitative assessment of cracking patterns in a lime cement matrix, Sensors. 20 (2020) 3859.
[16] E.E. Nyakilla, G. Jun, N.A. Kasimu, E.F. Robert, N. Innocent, T. Mohamedy, M. Shaame, M.R. Ngata, P.E. Mabeyo, Application of machine learning in the prediction of compressive, and shear bond strengths from the experimental data in oil well cement at 80°C. Ensemble trees boosting approach, Constr. Build. Mater. 317 (2022) 125778.
[17] C. Vipulanandan, A. Maddi, Characterizing the thermal, piezoresistive, rheology and fluid loss of smart foam cement slurries using artificial neural network and Vipulanandan Models, J. Pet. Sci. Eng. 207 (2021) 109161.
[18] L.V. Zhang, A. Marani, M.L. Nehdi, Chemistry-informed machine learning prediction of compressive strength for alkali-activated materials, Constr. Build. Mater. 316 (2022) 126103.
[19] P. Zhao, Y. Chen, Z. Zhao, Cholesky Factorization Based Online Sequential Multiple Kernel Extreme Learning Machine Algorithm for a Cement Clinker Free Lime Content Prediction Model, Processes. 9 (2021) 1540.
[20] U.K. Sevim, H.H. Bilgic, O.F. Cansiz, M. Ozturk, C.D. Atis, Compressive strength prediction models for cementitious composites with fly ash using machine learning techniques, Constr. Build. Mater. 271 (2021) 121584.
[21] Z. Tariq, M. Murtaza, M. Mahmoud, Development of new rheological models for class G cement with nanoclay as an additive using machine learning techniques, ACS Omega. 5 (2020) 17646–17657.
[22] E. Ford, S. Kailas, K. Maneparambil, N. Neithalath, Machine learning approaches to predict the micromechanical properties of cementitious hydration phases from microstructural chemical maps, Constr. Build. Mater. 265 (2020) 120647.
[23] T. Oey, S. Jones, J.W. Bullard, G. Sant, Machine learning can predict setting behavior and strength evolution of hydrating cement systems, J. Am. Ceram. Soc. 103 (2020) 480–490.
[24] J. Lapeyre, T. Han, B. Wiles, H. Ma, J. Huang, G. Sant, A. Kumar, Machine learning enables prompt prediction of hydration kinetics of multicomponent cementitious systems, Sci. Rep. 11 (2021) 3922.

[25] R. Cook, T. Han, A. Childers, C. Ryckman, K. Khayat, H. Ma, J. Huang, A. Kumar, Machine learning for high-fidelity prediction of cement hydration kinetics in blended systems, Mater. Des. 208 (2021) 109920.
[26] I. Nunez, M.L. Nehdi, Machine learning prediction of carbonation depth in recycled aggregate concrete incorporating SCMs, Constr. Build. Mater. 287 (2021) 123027.
[27] A. Marani, M.L. Nehdi, Machine learning prediction of compressive strength for phase change materials integrated cementitious composites, Constr. Build. Mater. 265 (2020) 120286.
[28] E.U. Eyo, S.J. Abbey, Machine learning regression and classification algorithms utilised for strength prediction of OPC/by-product materials improved soils, Constr. Build. Mater. 284 (2021) 122817.
[29] D.-C. Feng, Z.-T. Liu, X.-D. Wang, Y. Chen, J.-Q. Chang, D.-F. Wei, Z.-M. Jiang, Machine learning-based compressive strength prediction for concrete: An adaptive boosting approach, Constr. Build. Mater. 230 (2020) 117000.
[30] S. Gupta, S. Al-Obaidi, L. Ferrara, Meta-Analysis and Machine Learning Models to Optimize the Efficiency of Self-Healing Capacity of Cementitious Material, Materials. 14 (2021) 4437.
[31] M. Wang, E. Chen, P. Liu, W. Guo, Multivariable nonlinear predictive control of a clinker sintering system at different working states by combining artificial neural network and autoregressive exogenous, Adv. Mech. Eng. 12 (2020) 1687814019896509.
[32] P. Guo, W. Meng, M. Xu, V.C. Li, Y. Bao, Predicting mechanical properties of high-performance fiber-reinforced cementitious composites by integrating micromechanics and machine learning, Materials. 14 (2021) 3143.
[33] N.-D. Hoang, C.-T. Chen, K.-W. Liao, Prediction of chloride diffusion in cement mortar using multi-gene genetic programming and multivariate adaptive regression splines, Measurement. 112 (2017) 141–149.
[34] V. Vakharia, R. Gujar, Prediction of compressive strength and portland cement composition using cross-validation and feature ranking techniques, Constr. Build. Mater. 225 (2019) 292–301.
[35] J. Tinoco, A.A.S. Correia, P.J. Venda Oliveira, Soil-cement mixtures reinforced with fibers: a data-driven approach for mechanical properties prediction, Appl. Sci. 11 (2021) 8099.
[36] B. Tutmez, A data-driven study for evaluating fineness of cement by various predictors, Int. J. Mach. Learn. Cybern. 6 (2015) 501–510.
[37] D. Cruz, D.A. Talbert, W. Eberle, J. Biernacki, A neural network approach for predicting microstructure development in cement, in: Proc. Int. Conf. Artif. Intell. ICAI, The Steering Committee of The World Congress in Computer Science, 2016: p. 328.
[38] H. Chen, C. Qian, C. Liang, W. Kang, An approach for predicting the compressive strength of cement-based materials exposed to sulfate attack, PLoS One. 13 (2018) e0191370.
[39] N. Kumar, V. Naranje, S. Salunkhe, Cement strength prediction using cloud-based machine learning techniques, J. Struct. Integr. Maint. 5 (2020) 244–251.
[40] J. Huang, J. Liew, K. Liew, Data-driven machine learning approach for exploring and assessing mechanical properties of carbon nanotube-reinforced cement composites, Compos. Struct. 267 (2021) 113917.
[41] S. Mohanty, N. Roy, S.P. Singh, P. Sihag, Estimating the strength of stabilized dispersive soil with cement clinker and fly ash, Geotech. Geol. Eng. 37 (2019) 2915–2926.
[42] M. Verma, A. Thirumalaiselvi, J. Rajasankar, Kernel-based models for prediction of cement compressive strength, Neural Comput. Appl. 28 (2017) 1083–1100.
[43] S. Jueyendah, M. Lezgy-Nazargah, H. Eskandari-Naddaf, S. Emamian, Predicting the mechanical properties of cement mortar using the support vector machine approach, Constr. Build. Mater. 291 (2021) 123396.

[44] Q.D. Nguyen, S. Afroz, A. Castel, Influence of calcined clay reactivity on the mechanical properties and chloride diffusion resistance of limestone calcined clay cement (LC3) concrete, J. Mar. Sci. Eng. 8 (2020) 301.
[45] C. Rodriguez, J.I. Tobon, Influence of calcined clay/limestone, sulfate and clinker proportions on cement performance, Constr. Build. Mater. 251 (2020) 119050.
[46] Y. Zhang, C. Ling, A strategy to apply machine learning to small datasets in materials science, Npj Comput. Mater. 4 (2018) 1–8.
[47] A.A. Akindahunsi, F. Avet, K. Scrivener, The Influence of some calcined clays from Nigeria as clinker substitute in cementitious systems, Case Stud. Constr. Mater. 13 (2020) e00443.
[48] R.-S. Lin, H.-S. Lee, Y. Han, X.-Y. Wang, Experimental studies on hydration–strength–durability of limestone-cement-calcined Hwangtoh clay ternary composite, Constr. Build. Mater. 269 (2021) 121290.
[49] A. Dixit, H. Du, S. Dai Pang, Performance of mortar incorporating calcined marine clays with varying kaolinite content, J. Clean. Prod. 282 (2021) 124513.
[50] S. Krishnan, S.K. Kanaujia, S. Mithia, S. Bishnoi, Hydration kinetics and mechanisms of carbonates from stone wastes in ternary blends with calcined clay, Constr. Build. Mater. 164 (2018) 265–274.
[51] G. Mishra, A.C. Emmanuel, S. Bishnoi, Influence of temperature on hydration and microstructure properties of limestone-calcined clay blended cement, Mater. Struct. 52 (2019) 1–13.
[52] Y. Dhandapani, M. Santhanam, Assessment of pore structure evolution in the limestone calcined clay cementitious system and its implications for performance, Cem. Concr. Compos. 84 (2017) 36–47.
[53] F. Avet, R. Snellings, A.A. Diaz, M.B. Haha, K. Scrivener, Development of a new rapid, relevant and reliable (R3) test method to evaluate the pozzolanic reactivity of calcined kaolinitic clays, Cem. Concr. Res. 85 (2016) 1–11.
[54] N.S. Msinjili, G.J. Gluth, P. Sturm, N. Vogler, H.-C. Kühne, Comparison of calcined illitic clays (brick clays) and low-grade kaolinitic clays as supplementary cementitious materials, Mater. Struct. 52 (2019) 1–14.
[55] A. Alujas, R. Fernández, R. Quintana, K.L. Scrivener, F. Martirena, Pozzolanic reactivity of low grade kaolinitic clays: Influence of calcination temperature and impact of calcination products on OPC hydration, Appl. Clay Sci. 108 (2015) 94–101.
[56] B. Lorentz, H. Zhu, Y. Stetsko, K.A. Riding, A. Zayed, Feasibility Study for Calcined Clay Use in the Southeast USA, in: Calcined Clays Sustain. Concr., Springer, 2020: pp. 27–36.
[57] A. Machner, M. Zajac, M.B. Haha, K.O. Kjellsen, M.R. Geiker, K. De Weerdt, Portland metakaolin cement containing dolomite or limestone–Similarities and differences in phase assemblage and compressive strength, Constr. Build. Mater. 157 (2017) 214–225.
[58] J. Larsen, C. Goutte, On optimal data split for generalization estimation and model selection, in: Neural Netw. Signal Process. IX Proc. 1999 IEEE Signal Process. Soc. Workshop Cat No 98TH8468, IEEE, 1999: pp. 225–234.
[59] H.F. Taylor, Cement chemistry, Thomas Telford, 1997.
[60] T. Xie, P. Visintin, A unified approach for mix design of concrete containing supplementary cementitious materials based on reactivity moduli, J. Clean. Prod. 203 (2018) 68–82.
[61] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[62] S. Ray, A quick review of machine learning algorithms, in: 2019 Int. Conf. Mach. Learn. Big Data Cloud Parallel Comput. Com., IEEE, 2019: pp. 35–39.
[63] C. Saunders, A. Gammerman, V. Vovk, Ridge regression learning algorithm in dual variables, (1998).

[64] A.J. Smola, B. Schölkopf, A tutorial on support vector regression, Stat. Comput. 14 (2004) 199–222.
[65] C. Apté, S. Weiss, Data mining with decision trees and decision rules, Future Gener. Comput. Syst. 13 (1997) 197–210.
[66] A. Cutler, D.R. Cutler, J.R. Stevens, Random forests, in: Ensemble Mach. Learn., Springer, 2012: pp. 157–175.
[67] L. Breiman, Bagging predictors, Mach. Learn. 24 (1996) 123–140.
[68] Y.-Y. Song, L.U. Ying, Decision tree methods: applications for classification and prediction, Shanghai Arch. Psychiatry. 27 (2015) 130.
[69] C.-A. Azencott, Machine learning and genomics: precision medicine versus patient privacy, Philos. Trans. R. Soc. Math. Phys. Eng. Sci. 376 (2018) 20170350.
[70] C.-A. Azencott, Statistical machine learning and data mining for chemoinformatics and drug discovery, University of California, Irvine, 2010.
[71] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[72] P.F. Smith, S. Ganesh, P. Liu, A comparison of random forest regression and multiple linear regression for prediction in neuroscience, J. Neurosci. Methods. 220 (2013) 85–91.
[73] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proc. 22nd Acm Sigkdd Int. Conf. Knowl. Discov. Data Min., 2016: pp. 785–794.
[74] J.J. Hopfield, Artificial neural networks, IEEE Circuits Devices Mag. 4 (1988) 3–10.
[75] F. Rosenblatt, The perceptron: a probabilistic model for information storage and organization in the brain., Psychol. Rev. 65 (1958) 386.
[76] A.G. Barto, R.S. Sutton, C.W. Anderson, Neuronlike adaptive elements that can solve difficult learning control problems, IEEE Trans. Syst. Man Cybern. (1983) 834–846.
[77] F. Murtagh, Multilayer perceptrons for classification and regression, Neurocomputing. 2 (1991) 183–197.
[78] K. Swersky, J. Snoek, R.P. Adams, Multi-task bayesian optimization, Adv. Neural Inf. Process. Syst. 26 (2013).
[79] P. Refaeilzadeh, L. Tang, H. Liu, Cross-validation., Encycl. Database Syst. 5 (2009) 532–538.
[80] A. Botchkarev, Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology, ArXiv Prepr. ArXiv180903006. (2018).
[81] A. Altmann, L. Toloşi, O. Sander, T. Lengauer, Permutation importance: a corrected feature importance measure, Bioinformatics. 26 (2010) 1340–1347.
[82] E. Gomaa, T. Han, M. ElGawady, J. Huang, A. Kumar, Machine learning to predict properties of fresh and hardened alkali-activated concrete, Cem. Concr. Compos. 115 (2021) 103863.
[83] P.G. Asteris, M. Koopialipoor, D.J. Armaghani, E.A. Kotsonis, P.B. Lourenço, Prediction of cement-based mortars compressive strength using machine learning techniques, Neural Comput. Appl. 33 (2021) 13089–13121.
[84] Y. Briki, F. Avet, M. Zajac, P. Bowen, M.B. Haha, K. Scrivener, Understanding of the factors slowing down metakaolin reaction in limestone calcined clay cement (LC3) at late ages, Cem. Concr. Res. 146 (2021) 106477.
[85] K. Scrivener, F. Avet, H. Maraghechi, F. Zunino, J. Ston, W. Hanpongpun, A. Favier, Impacting factors and properties of limestone calcined clay cements (LC3), Green Mater. 7 (2018) 3–14.
[86] A.M. Rashad, Metakaolin as cementitious material: History, scours, production and composition–A comprehensive overview, Constr. Build. Mater. 41 (2013) 303–318.
[87] S. Salvador, Pozzolanic properties of flash-calcined kaolinite: a comparative study with soak-calcined products, Cem. Concr. Res. 25 (1995) 102–112.
[88] G. Kakali, T.H. Perraki, S. Tsivilis, E. Badogiannis, Thermal treatment of kaolin: the effect of mineralogy on the pozzolanic activity, Appl. Clay Sci. 20 (2001) 73–80.
[89] J. Rocha, J. Klinowski, Solid-state NMR studies of the structure and reactivity of metakaolinite, Angew. Chem. Int. Ed. Engl. 29 (1990) 553–554.

[90] T. Danner, G. Norden, H. Justnes, The effect of calcite in the raw clay on the pozzolanic activity of calcined illite and smectite, in: Calcined Clays Sustain. Concr., Springer, 2020: pp. 131–138.
[91] R.Z. Rakhimov, N.R. Rakhimova, A.R. Gaifullin, V.P. Morozov, Properties of Portland cement pastes enriched with addition of calcined marl, J. Build. Eng. 11 (2017) 30–36.
[92] G. Cardinaud, E. Rozière, O. Martinage, A. Loukili, L. Barnes-Davin, M. Paris, D. Deneele, Calcined clay–Limestone cements: Hydration processes with high and low-grade kaolinite clays, Constr. Build. Mater. 277 (2021) 122271.
[93] C. Bich, J. Ambroise, J. Péra, Influence of degree of dehydroxylation on the pozzolanic activity of metakaolin, Appl. Clay Sci. 44 (2009) 194–200.
[94] M. Cyr, M. Trinh, B. Husson, G. Casaux-Ginestet, Effect of cement type on metakaolin efficiency, Cem. Concr. Res. 64 (2014) 63–72.
[95] A. Shvarzman, K. Kovler, G.S. Grader, G.E. Shter, The effect of dehydroxylation/amorphization degree on pozzolanic activity of kaolinite, Cem. Concr. Res. 33 (2003) 405–416.
[96] A. Tironi, M.A. Trezza, A.N. Scian, E.F. Irassar, Kaolinitic calcined clays: Factors affecting its performance as pozzolans, Constr. Build. Mater. 28 (2012) 276–281.
[97] E.E. Nyakilla, G. Jun, N.A. Kasimu, E.F. Robert, N. Innocent, T. Mohamedy, M. Shaame, M.R. Ngata, P.E. Mabeyo, Application of machine learning in the prediction of compressive, and shear bond strengths from the experimental data in oil well cement at 80° C. Ensemble trees boosting approach, Constr. Build. Mater. 317 (2022) 125778.
[98] P. Zhao, Y. Chen, Z. Zhao, Cholesky Factorization Based Online Sequential Multiple Kernel Extreme Learning Machine Algorithm for a Cement Clinker Free Lime Content Prediction Model, Processes. 9 (2021) 1540.
[99] Z. Tariq, M. Murtaza, M. Mahmoud, Development of new rheological models for class G cement with nanoclay as an additive using machine learning techniques, ACS Omega. 5 (2020) 17646–17657.
[100] T. Oey, S. Jones, J.W. Bullard, G. Sant, Machine learning can predict setting behavior and strength evolution of hydrating cement systems, J. Am. Ceram. Soc. 103 (2020) 480–490.
[101] R. Cook, T. Han, A. Childers, C. Ryckman, K. Khayat, H. Ma, J. Huang, A. Kumar, Machine learning for high-fidelity prediction of cement hydration kinetics in blended systems, Mater. Des. 208 (2021) 109920.
[102] N.-D. Hoang, C.-T. Chen, K.-W. Liao, Prediction of chloride diffusion in cement mortar using multi-gene genetic programming and multivariate adaptive regression splines, Measurement. 112 (2017) 141–149.

Table 1: Description and statistical parameters of the data features

	Data items	Features	Symbol	Units	Mean	STD	min	max
14 Inputs	Calcined clay	Proportion of calcined clay	CL%	wt.%	23.5	7.9	10	40
		BET surface area	CL_Ss	m^2/g	18.5	7.3	2.5	45.7
	Calcination conditions of the clay	Temperature	T_calcin	°C	760.7	84.06	600	925
		Duration	time_calcin	hours	1.27	0.75	0.2	3
	Portland Cement	Proportion of OPC	OPC%	wt.%	68.9	12.13	37.6	90
	Limestone	Proportion Limestone	CC%	wt.%	7.5	7.7	0	31.1
	Chemical composition of the binder	Reactivity ratio	RM	-	2.3	0.5	1.6	3.8
		Silica ratio	SM	-	2	0.6	1.2	4.3
		Alumina ratio	AM	-	3.9	2.5	1.6	17.8
		Hydraulic ratio	HM	-	1.2	0.3	0.7	2.1
	Hardening conditions	Water to binder ratio	W/B	-	0.5	0.09	0.1	0.9
		Hardening temperature	T_cure	°C	22.6	6.0	5	50
		Hardening relative humidity	RH_cure	%	92.3	5.3	80	100
		Hardening age	age_D	days	30.7	41.3	1	270
1 Output	Compressive	Compressive	R	MPa	39.2	16.6	5	75

strength	strength	

Table 2: Search intervals of hyperparameters for the seven selected ML algorithms.

Algorithms	Hyperparameters	Search interval
LR	-	-
LR-RR	-	-
SVR	kernel coefficient (γ)	[0.1, 5]
	epsilon (ε)	[0.1, 20]
	regularization (C)	[0.01, 500]
DTR	max_depth	[1, 30]
	max_features	[1, 14]
	min_samples_split	[2, 50]
	min_samples_leaf	[1, 50]
	random_state	[1, 80]
RF	n_estimators	[10, 600]
	max_depth	[1, 30]
	max_features	[1, 14]
	min_samples_split	[2, 50]
	min_samples_leaf	[1, 50]
	random_state	[1, 80]
XGboost	n_estimators	[10, 200]

	max_depth	[1, 30]
	subsample	[0.5, 10]
	seed	[5, 100]
MLP	hidden_layer_sizes	[(150, 100, 50)
		, (120, 80, 40)
		, (100, 50, 30)]
	max_iter	[50, 300]
	activation	[‘tanh’, ’relu’]
	solver	[‘sgd’, ’adam’]
	alpha	[0.0001, 0.05]
	learning_rate	[‘constant’, ’adaptive’]

Table 3: Designed experimental scenarios with their fixed and variable parameters.

	Scenario 1	Scenario 2	Scenario 3	Scenario 4
Input features	Effect of weight fraction of the cement constituents	Effect of the chemical composition of the clay	Effect of calcination conditions of the clay	Effect of the hardening conditions
CL%	Variable	30	30	30
CL_Ss	20	20	20	20
T_calcin	800	800	Variable	800
time_calcin	1	1	Variable	1
OPC%	Variable	55	55	55
CC%	Variable	15	15	15
RM	Calculated and fixed (*)	Variable (**)	Calculated and fixed (*)	Calculated and fixed (*)

SM	Calculated and fixed (*)	Variable (**)	Calculated and fixed (*)	Calculated and fixed (*)
AM	Calculated and fixed (*)	Variable (**)	Calculated and fixed (*)	Calculated and fixed (*)
HM	Calculated and fixed (*)	Variable (**)	Calculated and fixed (*)	Calculated and fixed (*)
W/B	0.5	0.5	0.5	Variable
T_cure	20	20	20	Variable
RH_cure	90	90	90	90
age_D	Variable	Variable	Variable	Variable

(*) calculated from the chemical compositions given in Table 5.
(**) different clay compositions were taken from literature ([1,2,47–57]) in order to evaluate their effect on the compressive strength.

Table 4: Multi-correlation matrix of the 14 inputs and the compressive strength.

	R	CL%	T_calcin	CL_Ss	time_calcin (h)	CC%	OPC%	W/B	T_cure	RH_cure	age_D	RM	SM	AM	HM
R	1	-0.21	0.04	-0.01	0.11	-0.17	0.25	0.06	-0.05	0.00	0.50	0.17	0.00	0.07	0.13
CL%	-0.21	1	0.22	-0.05	-0.29	0.20	-0.78	-0.04	0.16	0.08	-0.07	-0.81	-0.30	0.28	-0.94
T_calcin	0.04	0.22	1	0.05	-0.06	0.30	-0.33	0.07	-0.08	0.26	0.05	-0.30	-0.09	0.11	-0.35
CL_Ss	-0.01	-0.05	0.05	1	0.15	0.00	0.03	-0.01	-0.15	0.03	-0.04	0.01	0.02	-0.12	0.03
time_calcin (h)	0.11	-0.29	-0.06	0.15	1	-0.43	0.46	-0.21	-0.23	0.39	-0.17	-0.08	0.76	-0.07	0.25
CC%	-0.17	0.20	0.30	0.00	-0.43	1	-0.76	-0.07	-0.02	0.06	0.08	-0.02	-0.36	0.34	-0.22
OPC%	0.25	-0.78	-0.33	0.03	0.46	-0.76	1	0.07	-0.09	-0.09	0.00	0.55	0.43	-0.40	0.76
W/B	0.06	-0.04	0.07	-0.01	-0.21	-0.07	0.07	1	0.04	-0.31	0.10	0.08	-0.15	-0.32	0.01
T_cure	-0.05	0.16	-0.08	-0.15	-0.23	-0.02	-0.09	0.04	1	0.03	0.00	-0.11	-0.10	-0.16	-0.13
RH_cure	0.00	0.08	0.26	0.03	0.39	0.06	-0.09	-0.31	0.03	1	-0.08	-0.41	0.41	-0.05	-0.22
age_D	0.50	-0.07	0.05	-0.04	-0.17	0.08	0.00	0.10	0.00	-0.08	1	0.07	-0.11	-0.01	0.01
RM	0.17	-0.81	-0.30	0.01	-0.08	-0.02	0.55	0.08	-0.11	-0.41	0.07	1	-0.20	-0.04	0.88
SM	0.00	-0.30	-0.09	0.02	0.76	-0.36	0.43	-0.15	-0.10	0.41	-0.11	-0.20	1	-0.27	0.26
AM	0.07	0.28	0.11	-0.12	-0.07	0.34	-0.40	-0.32	-0.16	-0.05	-0.01	-0.04	-0.27	1	-0.25
HM	0.13	-0.94	-0.35	0.03	0.25	-0.22	0.76	0.01	-0.13	-0.22	0.01	0.88	0.26	-0.25	1

Table 5: Chemical composition of the cement constituents used for the calculation of reactivity ratios given in Table 3.

	CaO	SiO_2	Al_2O_3	Fe_2O_3	MgO	Na_2O	K_2O	CO_2	Other elements
Kaolin	0.03	56.79	35.63	2	0.34	0.2	3.49	0	1.52
OPC	65.21	21.27	4.57	3.25	1.62	0.08	0.76	0	3.24
Limestone	54.98	0.33	0.24	0.04	0.88	0	0.04	43.46	0.03