Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Between-study variance inflates apparent accuracy in literature-compiled machine learning models of sulphate resistance in SCM-blended cement

Abdulwahhab Abdullah
Kwara State University
Ibrahim Adedeji
adedejiibrahim121@gmail.com

Kwara State University

Research Article

Keywords: Leave-one-source-out validation, Data aggregation bias, Supplementary cementitious materials, External sulphate attack, Ridge regression, Compressive strength

Posted Date: November 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11132514/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Between-study variance inflates apparent accuracy in literature-compiled machine learning models of sulphate resistance in SCM-blended cement

Abdulwahhab Abdullah[1], Ibrahim Adedeji[2]

[1,2]Department of Civil and Environmental Engineering, Faculty of Engineering, Kwara State University, Malete, Nigeria

abdulwaha.abdullah@kwasu.edu.ng; adedejiibrahim121@gmail.com

Corresponding Author: adedejiibrahim121@gmail.com

Abstract

Machine learning models trained on compressive strength data aggregated from published experimental studies are increasingly used to predict the sulphate resistance of cement blended with supplementary cementitious materials (SCMs), routinely reporting coefficients of determination above 0.90. This study examines whether that accuracy reflects genuine predictive capability. A dataset of 94 observations was compiled from three peer-reviewed studies covering silica fume, nano-silica, fly ash, slag, rice husk ash and cassava peel ash at replacement levels of 0–30%, exposure durations of 0–500 days and sulphate concentrations of 0.352–0.704 mol/L. A ridge regression model ($\alpha = 0.078$) achieved $R^2 = 0.968$ on a randomly held-out test set, comparable to ordinary least squares (0.968), Lasso (0.966) and random forest (0.982). Under leave-one-study-out validation, however, every fold returned a negative R^2 (−12.28 to −162.35), with mean absolute errors up to 42.5 MPa on mortars whose measured strengths did not exceed 21.5 MPa. Variance decomposition showed that 96.1% of total strength variance lay between source studies rather than within them. Ten of eleven SCM categories occurred in only one study, and each study used its own sulphate cation and concentration, rendering material effects and study effects mathematically inseparable. For the single SCM type present in two studies, mean strengths differed by 7.3 MPa across studies, approximately half the 14.0 MPa range observed across all nine SCM types within a single study. Random train–test splitting cannot detect this confounding. Leave-one-study-out validation and between-study variance reporting should be treated as minimum requirements for literature-compiled durability models.

Keywords: Leave-one-source-out validation; Data aggregation bias; Supplementary cementitious materials; External sulphate attack; Ridge regression; Compressive strength.

1. Introduction

Cement manufacture accounts for close to 8% of anthropogenic carbon dioxide output worldwide, making it a persistent obstacle to lowering the carbon intensity of construction [1]. Substituting part of the clinker with supplementary cementitious materials (SCMs) remains one of the few mitigation routes available at scale. Fly ash, ground granulated blast-furnace slag and silica fume have served this purpose for decades. Attention has since extended to nano-silica and to agricultural residues

including rice husk ash and cassava peel ash, which matter particularly in economies where industrial by-products are limited but crop waste is plentiful.

Structures founded in sulphate-bearing soil, groundwater or seawater undergo external sulphate attack, among the more consequential chemical mechanisms limiting service life. Sulphate ions migrating through the pore network encounter hydrated cement phases, principally calcium hydroxide and monosulphate, and the reaction yields ettringite and gypsum [2]. Because these products occupy more volume than the phases they replace, the resulting crystallisation pressure fractures the matrix, and strength falls as cracking, spalling and mass loss progress [3].

Pozzolanic reaction consumes calcium hydroxide and refines pore structure, so SCM-blended systems are expected to resist this attack better than plain Portland cement. Verifying that expectation is the difficulty. Immersion testing under ASTM C1012 runs for months before any trend can be read, and the search for faster alternatives reflects how binding that constraint is: one recent protocol reports usable assessment at 21 days for mortar and 56 days for concrete, against twelve and six months respectively under the standard method, with outcomes that agree with C1012 interpretations [4]. Compounding the problem, replacement level, SCM reactivity, sulphate species and concentration, and exposure duration all interact, and their combined effect on strength is not additive. Testing the full factorial is not practical.

Statistical and machine learning models have accordingly been recruited to estimate durability outcomes without waiting for immersion tests to conclude, and the accuracies reported are consistently high. One stacking ensemble, fitted to 330 observations gathered from the published literature, returned a coefficient of determination of 0.992 for compressive strength under sulphate exposure [5]. Similar figures appear for sulphate resistance of blended cements across several algorithm families [6], and for expansion under external sulphate attack modelled by interpretable ensembles [7].

What these studies share, besides their accuracy, is where their data come from. Producing durability measurements is slow and costly, so training sets are assembled by extracting values from experiments other groups have already published and pooling them, rather than by generating observations under one controlled protocol. Surveying machine learning applied to concrete durability between 2013 and 2024, including sulphate attack, RILEM Technical Committee 315-DCS found the field drawing overwhelmingly on laboratory measurements, with little use of field data and few purpose-built datasets [8].

Accuracy in this literature is nearly always estimated the same way: the pooled dataset is partitioned at random into training and test subsets, or subjected to k-fold cross-validation over the same pool. Both procedures assume the observations are exchangeable, meaning that swapping any row for any other would leave the estimation problem unchanged. Pooled multi-study data violate that assumption by construction. Every contributing study fixes its own cement chemistry, water-to-binder ratio, aggregate, curing regime, specimen geometry and testing standard, and passes those attributes to each row it supplies. Randomly assigning rows to test and training subsets therefore guarantees that both contain material from the same studies. A model can then reach low test error by identifying the study a row originated from, without having learned anything about how composition and exposure govern strength.

The problem is recognised elsewhere in materials science. Meredig et al. [9] argued that models scored by conventional validation look stronger than they are whenever the intended application is prediction on unfamiliar material families, and proposed withholding entire chemical clusters as a corrective; errors rose sharply once folds were constructed this way. Durdy et al. [10] developed the approach further, observing that research effort is distributed unevenly across any design space, leaving some regions densely populated and others empty, which flatters models assessed by random splitting and exposes them when extrapolation is required. The same reasoning motivated redundancy control methods for materials property prediction [11]. Grouped validation of this kind has not passed into common use in cementitious durability modelling, where random partitioning remains standard. Whether the distinction matters in practice for sulphate durability prediction is an empirical question, and it is the question addressed here.

This study compares validation designs for a machine learning model of compressive strength in SCM-blended mortar under sulphate exposure, fitted to data compiled from three independent experimental studies. Three objectives are pursued: to quantify how estimated model performance differs between random train–test partitioning and leave-one-source-out validation, in which entire contributing studies are withheld; to decompose total variance in compressive strength into between-study and within-study components, establishing what proportion of the modelled signal originates in differences between source studies; and to determine whether SCM effects can be statistically separated from study effects in a dataset of the composition typical of this literature, together with what such datasets would need in order to support cross-study prediction. Ridge regression is used throughout. Regularisation was adopted for coefficient stability under correlated predictors rather than for any expected gain in accuracy, and the model is intentionally simple, since the subject of the study is how performance is measured rather than which algorithm measures best.

2. Materials and Methods

2.1 Data sources and extraction

No new experimental work was undertaken. Observations were compiled from three previously published experimental programmes [12], [13], [14], selected because each reports compressive strength of mortar or concrete containing supplementary cementitious materials after immersion in sulphate solution, with exposure duration and sulphate concentration stated. For each specimen the following were recorded: source study, cement type, SCM type, SCM replacement level by mass, exposure duration in days, sulphate concentration in mol/L, and measured compressive strength in MPa.

De Souza et al. [14] report results for both sodium and magnesium sulphate exposure. Only the sodium sulphate series was extracted, so that the study contributes a single exposure condition and its observations remain comparable with one another. The magnesium sulphate series was therefore excluded by design rather than by availability, and it is the series in which those authors record loss of compressive strength; the implications are considered in Sections 4.2 and 4.5.

Values reported in tables were transcribed directly. Values presented only in figures were recovered using WebPlotDigitizer. Digitisation introduces reading error that cannot be quantified without access to the underlying measurements; the resulting

values should be treated as approximations of the originals rather than exact reproductions, and this is noted among the limitations in Section 4.5. All quantities were converted to a common set of units before pooling.

2.2 Composition of the compiled dataset

The three studies differ from one another in nearly every respect other than the property measured. Table 1 and Fig. 1 are therefore presented before any modelling, because the structure they show governs the results reported in Section 3.

Table 1 Composition of the compiled dataset by source study

	Olonade & Abdullah [12]	**Wang et al.[13]**	**De Souza et al.[14]**
Observations	40	18	36
SCM categories	1	2	9
Categories present	CPA	RHA, SF10	Control, FA10, FA30, GBS10, GBS30, NS1, NS3, SF5, SF10
Replacement levels (%)	20	10	0, 1, 3, 5, 10, 30
Cement	Four commercial Portland cements	CPV-ARI	Portland cement
Sulphate concentration (mol/L)	0.352, 0.367	0.704	0.415
Exposure durations (days)	7, 14, 28, 56, 90	0, 14, 42, 70, 140	7, 28, 180, 500
Compressive strength (MPa)	6.12–21.50	43.00–64.00	47.00–66.00

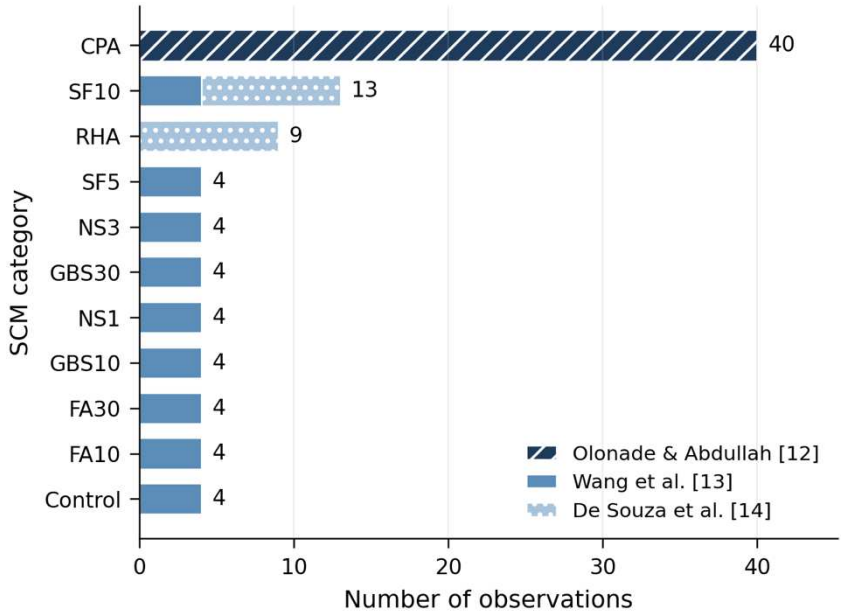


Fig. 1 Distribution of observations across source studies and SCM categories in the compiled dataset.

Four features are consequential. Sulphate concentration takes a distinct value in each study, so that concentration and study identity function as near-interchangeable predictors. The sulphate cation likewise varies between studies and is confounded with them. Compressive strength ranges do not overlap between Olonade & Abdullah [12] and the other two, leaving an unpopulated interval between 21.50 and 43.00 MPa across which any fitted model must interpolate without support. And of the eleven SCM categories in the pooled data, ten occur in one study only; the single exception, SF10, appears in two studies solely because of the category merge described in Section 2.3. Descriptive statistics for the pooled dataset are given in Table 2.

Table 2 Descriptive statistics for the pooled dataset (n = 94)

Variable	**Mean**	**SD**	**Minimum**	**Maximum**
Compressive strength (MPa)	36.37	22.67	6.12	66.00
SCM replacement (%)	14.64	8.16	0.00	30.00
Exposure duration (days)	96.37	142.18	0.00	500.00
Sulphate concentration (mol/L)	0.45	0.13	0.352	0.704

2.3 Data preparation

Four preparation steps were applied. Each is recorded in full, since three of them materially affect the coefficients reported in Section 3.

Two exact duplicate rows were identified and removed, reducing the compiled set from 96 to 94 observations. No missing values were present in any modelled variable,

so no imputation or case deletion was required. Two further rows, recording Sokoto-cement specimens at 56 and 90 days, had been assigned the calcium sulphate concentration when the source reports them under sodium sulphate; both were corrected against the original publication [12], as was one 90-day value that had been misread from the source figure.

Replacement levels recorded in the compiled spreadsheet disagreed with the dosage stated in the corresponding SCM label for five categories: SF5, NS1, NS3, FA30 and GBS30 had all been entered at 10%. Twenty rows were corrected so that replacement level matched the label, the label being the more specific record of what was tested.

Nine observations from De Souza et al. [14], labelled only as silica fume, were assigned to the SF10 category on the grounds of nominally equivalent material and replacement level. This merge is deliberate rather than incidental. It produces the only SCM category present in more than one study, and so permits the cross-study comparison reported in Section 3.4. The two sets of specimens were nevertheless prepared with different cements and exposed to different sulphate salts, and the comparison should be read with that in mind.

The sulphate cation could not be included as a predictor. Olonade & Abdullah [12] used calcium and sodium sulphate, Wang et al. [13] magnesium sulphate and De Souza et al. [14] sodium sulphate, so cation is fully determined by study for two of the three sources and cannot be separated from study identity by any estimation procedure. Cation is recorded in Table 1 and in the supplementary dataset, and its consequences for interpreting the concentration coefficient are taken up in Section 4.2.

Replacement level was excluded as a predictor for two reasons. Each SCM label encodes its own dosage, making replacement level a deterministic function of the SCM type indicators; its variance inflation factor was unbounded when both were retained, and the fitted coefficients took opposing signs that cancelled, neither of which could be interpreted. Independently, replacement level is itself close to a study marker, since Olonade & Abdullah [12] tested only at 20% and De Souza et al. [14] only at 10%, with Wang et al. [13] supplying the whole remaining range. Dosage information is therefore carried by the SCM type indicators alone.

SCM type was encoded as binary indicators, one per category. The unblended Portland cement mixture from Wang et al. [13] was designated the reference category and its indicator omitted, so that every SCM coefficient expresses a difference in predicted strength relative to that control. This treatment imposes no ordering among SCM types, which is appropriate given that the categories differ in material as well as dosage.

2.4 Ridge regression modelling

idge regression was fitted with compressive strength as the response and exposure duration, sulphate concentration and the SCM type indicators as predictors. Ridge applies an L2 penalty to the coefficient vector, shrinking estimates toward zero and stabilising them when predictors are correlated. It was selected for that stability rather than for any anticipated gain in accuracy; the comparison reported in Section 3.1 confirms that it confers none.

The two continuous predictors were standardised to zero mean and unit variance before fitting. The L2 penalty acts on raw coefficient magnitude, so variables measured on different scales are penalised unequally unless they are placed on a common scale first. Exposure duration spans 0 to 500 days and sulphate concentration 0.352 to 0.704 mol/L, a difference of roughly three orders of magnitude, and without standardisation the penalty would fall almost entirely on the concentration coefficient. Scaling was fitted inside the cross-validation pipeline so that no information from a held-out fold entered the transformation. The binary SCM indicators were left unscaled, which preserves their coefficients as differences in MPa relative to the reference mixture.

The regularisation parameter was selected by five-fold cross-validation over a logarithmic grid of 200 values spanning 10^{-4} to 10^{4}, returning $\alpha = 0.0784282$. The grid was made deliberately wide because a narrow search cannot distinguish a genuine preference for weak regularisation from a grid that excludes the stronger alternatives. The value returned lies near the lower end of the range and represents mild shrinkage, a point taken up in Section 3.1.

Because the fitted coefficients apply to standardised inputs, they are reported in Section 3.2 in two forms. Standardised coefficients express the change in predicted strength associated with a one standard deviation change in the predictor and are comparable across variables. Raw-unit coefficients, obtained by dividing each standardised coefficient by the standard deviation of its predictor and adjusting the intercept accordingly, apply directly to untransformed inputs and reproduce the model's predictions when substituted into the model equation.

Three further estimators were fitted on identical training folds for comparison: ordinary least squares, Lasso with its penalty selected by the same cross-validation procedure, and a random forest of 500 trees. These establish whether the accuracy reported in Section 3.1 is a property of the ridge model specifically or of the dataset generally.

All computation was carried out in Python 3.13.15 using scikit-learn 1.6.1, NumPy 2.1.3, pandas 2.2.3 and matplotlib 3.10.0. The pseudo-random seed was fixed at 42 for every operation involving random assignment. All results reported in Section 3 derive from a single execution of one analysis script, so that every table, figure and coefficient in the manuscript is mutually consistent.

2.5 Model validation

Two validation schemes were applied to the same model specification, and the comparison between them is the object of this study.

The first follows prevailing practice in the literature reviewed in Section 1. The pooled dataset was partitioned at random into training and test subsets in an 80:20 ratio, with the regularisation parameter tuned by five-fold cross-validation within the training subset alone. Under this scheme both subsets necessarily contain observations from all three source studies, since assignment is made at the level of the individual observation.

The second scheme withholds whole studies. The model was refitted three times, on each occasion using two studies for training and reserving the third entirely for testing. This asks whether the model transfers to an experimental programme it has not encountered, which is the task any predictive tool built from compiled data would face

in application. Because the withheld source contributes no observations to training, the scheme also removes the possibility of a model achieving low test error by recognising study-specific characteristics rather than material behaviour.

The design follows the grouped-validation principle developed for materials property prediction [9,10], with source study substituted for chemical cluster as the grouping variable. Study is the natural grouping here because the attributes shared within a group, namely cement chemistry, mixing and curing procedure, specimen geometry, sulphate salt and concentration, and testing apparatus, are precisely those that a compiled dataset leaves unrecorded and that a model trained on such data cannot observe.

Under leave-one-study-out validation the coefficient of determination admits negative values, indicating prediction worse than would be obtained by returning the training mean for every specimen. Negative values are reported as computed rather than truncated at zero, since their magnitude conveys how far the predictions fall from the withheld observations.

Reporting both schemes for one model isolates the effect of validation design. The model specification, predictors, regularisation parameter and software are identical throughout, so any difference in measured performance is attributable to how observations were assigned between training and testing.

2.6 Variance decomposition

Total variance in compressive strength across the pooled dataset was partitioned into within-source and between-source components. The within-source component was computed as the mean of the variances observed inside each source study, weighted by the number of observations each contributed, and the between-source component as the difference between total variance and that within-source figure. The between-source share is reported as the ratio of the between-source component to total variance. Variances were computed with denominator N rather than N − 1, so that the two components sum exactly to the total.

The decomposition establishes how much of the variation available to any model fitted on the pooled data arises from differences between the contributing studies rather than from variation in composition and exposure within them. A model can only be rewarded for explaining variation that exists. Where the between-source share is large, a high coefficient of determination is compatible with the model having learned to distinguish source studies while capturing little of the material behaviour the predictors are intended to represent.

Two further checks address whether SCM effects can be separated from study effects. First, the number of source studies contributing to each SCM category was tabulated. Second, for any category appearing in more than one source, mean strengths were compared across sources and the difference set against the range of mean strengths observed across all SCM categories within a single source. Multicollinearity among the retained predictors was assessed by variance inflation factor.

2.7 Evaluation metrics

Performance was reported as the coefficient of determination, mean absolute error and root mean square error. The coefficient of determination is reported for both

validation schemes. Mean absolute error is reported in MPa alongside the strength range of the withheld source, since an error of fixed magnitude carries different weight against specimens of 10 MPa than against specimens of 60 MPa.

3. Results

3.1 Model performance under random partitioning

Table 3 reports performance for all four estimators under the random 80:20 partition described in Section 2.5, and Fig. 2 plots predicted against measured strength for the ridge model on the held-out test set.

Table 3 Model performance under random train–test partitioning

Model	R^2 (train)	R^2 (test)	MAE (MPa)	RMSE (MPa)	CV R^2 (mean ± SD)
Ordinary least squares	0.9758	0.9678	3.03	3.85	0.9642 ± 0.0126
Ridge (α = 0.0784282)	0.9757	0.9674	3.04	3.88	0.9647 ± 0.0120
Lasso	0.9755	0.9661	3.13	3.96	0.9638 ± 0.0115
Random forest	0.9920	0.9821	2.29	2.88	0.9734 ± 0.0118

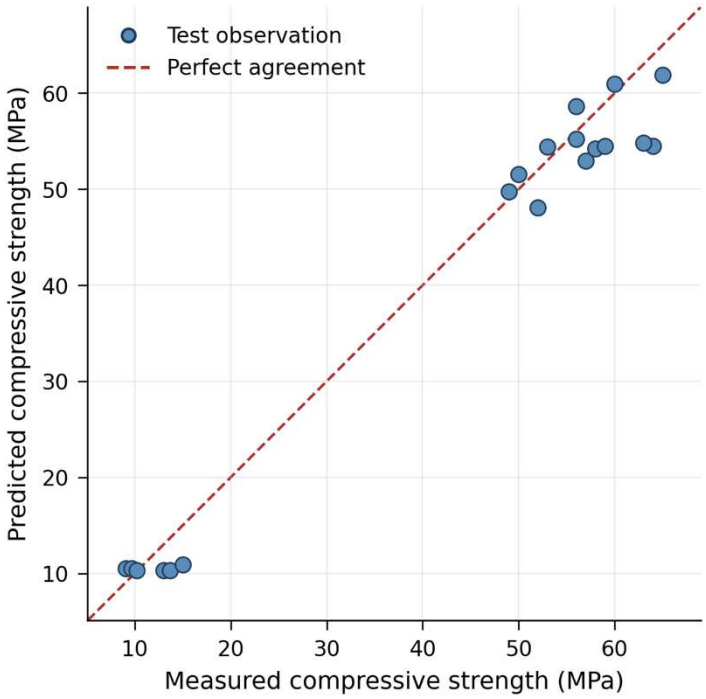


Fig. 2 Predicted against measured compressive strength for the ridge model on the randomly held-out test set. Dashed line indicates perfect agreement

All four estimators returned test-set coefficients of determination between 0.966 and 0.982. These values are consistent with those reported in the literature surveyed in Section 1 and would, on conventional reading, indicate a model capable of predicting compressive strength under sulphate exposure to within approximately 3 MPa.

Two observations qualify that reading before Section 3.3 addresses it directly. Ridge did not outperform ordinary least squares. Its test coefficient of determination was 0.9674 against 0.9678 for the unregularised model, and its mean absolute error was marginally higher at 3.04 MPa against 3.03 MPa. The selected regularisation parameter of 0.0784282 sits near the lower end of a grid spanning eight orders of magnitude and represents mild shrinkage, so the ridge and least-squares solutions are close to identical by construction. Ridge is retained in what follows for coefficient stability under correlated predictors, not because it confers any accuracy advantage. No such advantage was found.

The similarity of all four estimators is itself informative. A random forest, which imposes no linear structure and can represent arbitrary interactions, scored 0.9821 against 0.9674 for ridge, a difference of 1.5 percentage points. Where a flexible non-parametric model and a near-unregularised linear model perform comparably, the variation being explained is unlikely to be complex. Section 3.4 identifies what it is.

Variance inflation factors for the retained predictors ranged from 1.27 (exposure duration) to 7.03 (the CPA indicator), all below the conventional threshold of 10, confirming that the exclusion of replacement level described in Section 2.3 resolved the collinearity present in the initial specification.

3.2 Fitted coefficients

Table 4 reports the ridge coefficients in both standardised and raw-unit form, and Fig. 3 shows them graphically.

Table 4 Ridge regression coefficients. Reference category: unblended Portland cement control

Predictor	Standardised	Raw units	Unit
Exposure duration	−0.7567	−0.0051	MPa per day
Sulphate concentration	−3.0280	−23.7770	MPa per mol/L
CPA	−42.1519	−42.1519	MPa vs control
FA10	2.2427	2.2427	MPa vs control
FA30	−2.6300	−2.6300	MPa vs control
GBS10	7.0787	7.0787	MPa vs control

Predictor	Standardised	Raw units	Unit
GBS30	0.7349	0.7349	MPa vs control
NS1	2.8558	2.8558	MPa vs control
NS3	3.5420	3.5420	MPa vs control
RHA	10.4437	10.4437	MPa vs control
SF5	9.3526	9.3526	MPa vs control
SF10	11.1910	11.1910	MPa vs control
Intercept	50.4527	61.4811	MPa

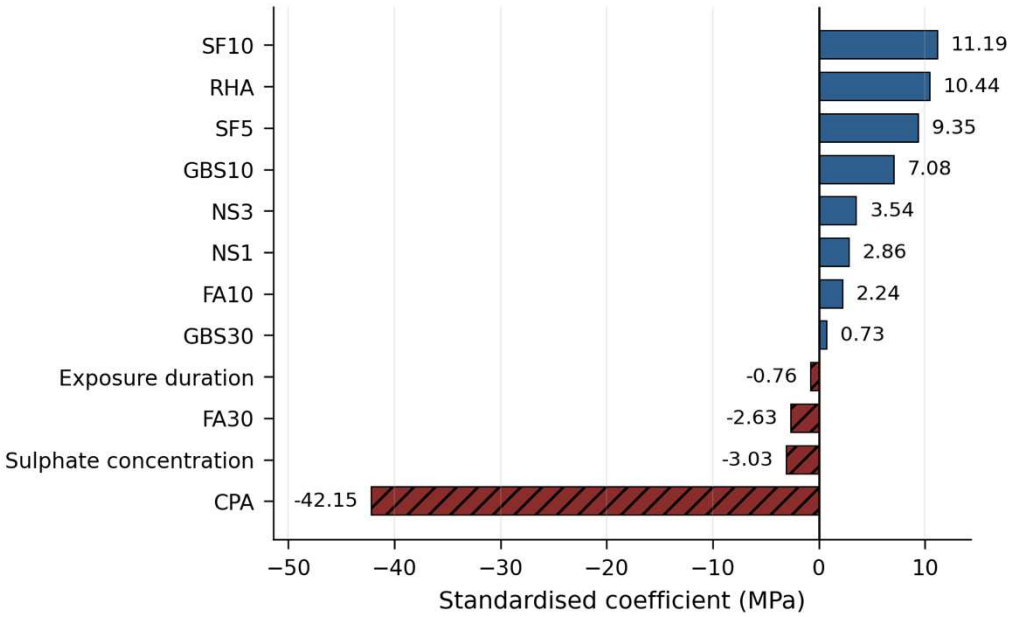


Fig. 3 Standardised ridge regression coefficients for all retained predictors. Positive values indicate higher predicted strength relative to the unblended control

In raw units the fitted model is

$$CS = 61.134 - 0.0050\,t - 23.0379\,c + \Sigma_j\,\beta_j\,x_j \qquad (1)$$

where CS is predicted compressive strength in MPa, t is exposure duration in days, c is sulphate concentration in mol/L, and xj are the binary SCM indicators with coefficients βj from Table 4. Substituting the values tabulated above reproduces the predictions reported in Table 7 exactly.

Read conventionally, these coefficients are unremarkable and largely consistent with expectation. Silica fume carries the largest positive coefficients, at 11.19 MPa for 10%

replacement and 9.35 MPa for 5%, which agrees with the established pozzolanic behaviour of the material. Sulphate concentration carries a substantial negative coefficient. Exposure duration is close to zero. Cassava peel ash carries a large negative coefficient of −42.15 MPa.

This section presents the coefficients because they are the output a reader of comparable published work would expect, and because their apparent reasonableness is part of what the remainder of Section 3 examines. They should not be read as estimates of material effects. Section 3.4 establishes that SCM type and source study are not statistically separable in this dataset, and Section 4.2 addresses the sulphate concentration coefficient specifically.

3.3 Leave-one-source-out validation

Table 5 and Fig. 4 report performance when whole source studies are withheld.

Table 5 Leave-one-source-out validation of the ridge model

Withheld study	n	R^2	MAE (MPa)	RMSE (MPa)	Strength range of withheld study (MPa)
Olonade & Abdullah [12]	40	−162.35	42.45	42.57	6.12–21.50
Wang et al. [13]	36	−12.28	15.90	17.91	47.00–66.00
De Souza et al. [14]	18	−15.80	20.79	22.71	43.00–64.00

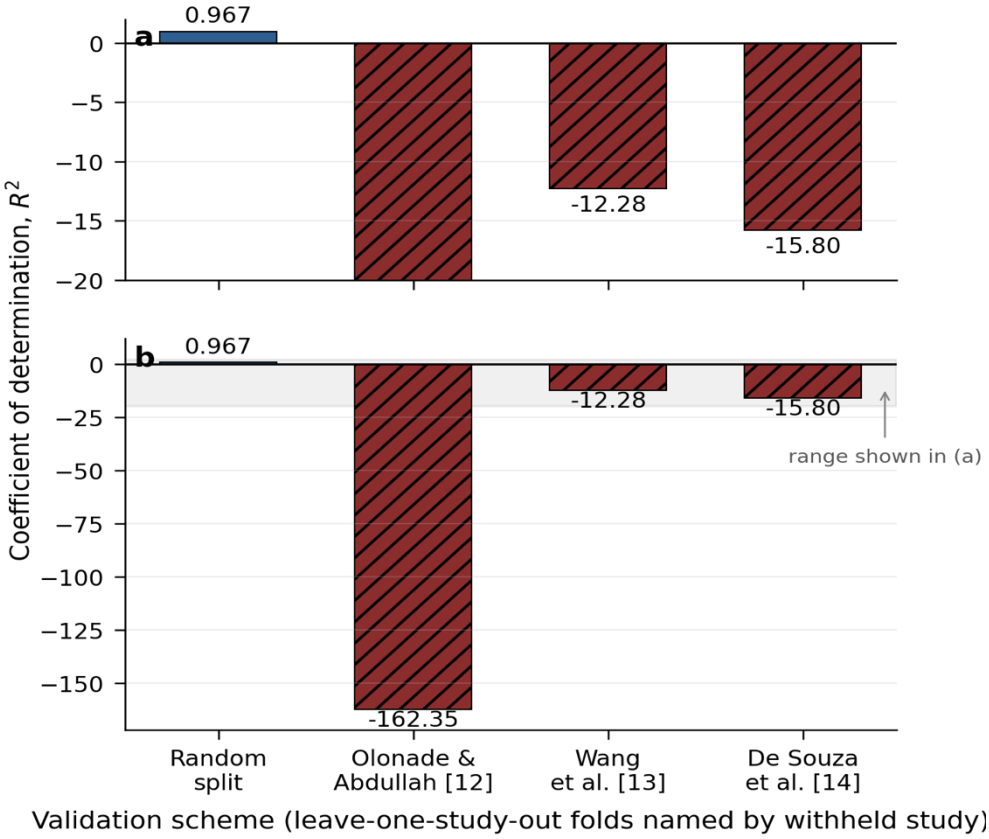


Fig. 4 Coefficient of determination under random partitioning compared with leave-one-source-out validation for each withheld source

Every fold returned a negative coefficient of determination. A negative value indicates that predicting the training mean for every specimen would have produced smaller squared error than the fitted model achieved. On the withheld source, in other words, the model carries no useful information at all.

The Olonade & Abdullah fold is the most severe. Trained on the other two studies, the model predicted specimens whose measured strengths lay between 6.12 and 21.50 MPa with a mean absolute error of 42.45 MPa, an error roughly three times the largest value it was attempting to predict. Both training sources consist of high-strength mixtures, and nothing in the training data indicates that a mortar of approximately 10 MPa is possible.

The contrast with Section 3.1 is the central empirical result of this study. The same model specification, fitted with the same predictors and the same regularisation parameter, returns a coefficient of determination of 0.968 under random partitioning and between −12.28 and −162.35 when whole studies are withheld. Nothing about the model changed. Only the assignment of observations between training and testing changed.

3.4 Source confounding

Section 3.3 establishes that the model does not transfer between studies. This section identifies why, through four findings.

Variance is almost entirely between-source. The decomposition described in Section 2.6 gives a total variance in compressive strength of 508.6 MPa², of which

19.9 MPa² is within-study and 488.7 MPa² between-study. The between-study share is 96.1%. Fewer than four percent of the squared deviations available to any model fitted on this dataset arise from variation inside the contributing studies; the remainder is variation between them. Mean strengths were 10.69 MPa for Olonade & Abdullah [12], 54.81 MPa for Wang et al. [13] and 56.56 MPa for De Souza et al. [14], and the ranges of the first and the other two do not overlap (Fig. 5). An interval between 21.50 and 43.00 MPa contains no observations at all.

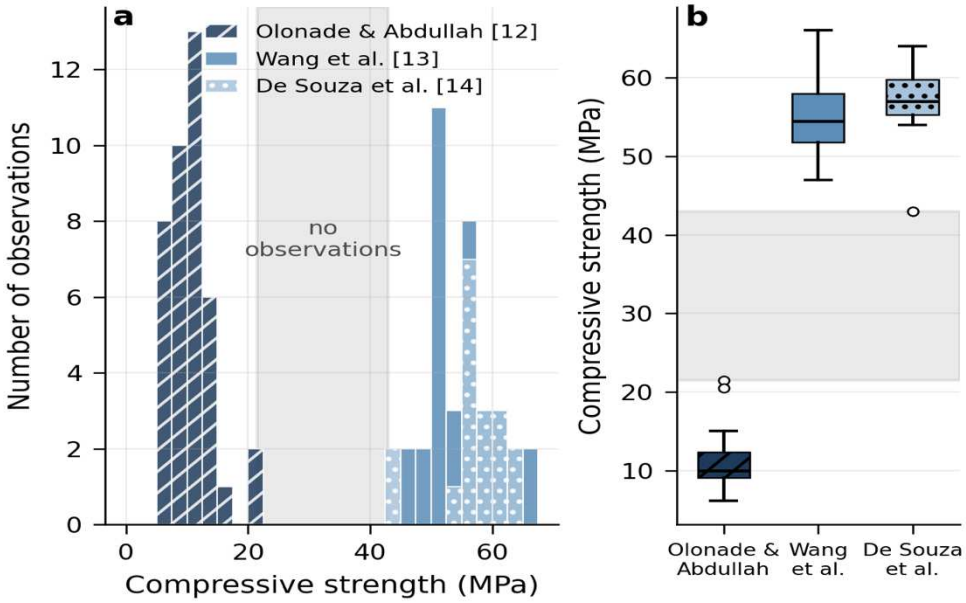


Fig. 5 Distribution of measured compressive strength within each source study. No observations fall between 21.50 and 43.00 MPa.

SCM categories are confined to single sources. Of the eleven SCM categories present after preparation, ten occur in one source study only. The single exception, SF10, appears in two studies solely as a result of the category merge described in Section 2.3. Where a categorical predictor takes a distinct set of values in each group, its indicators function as group indicators. The coefficients in Table 4 therefore cannot distinguish the effect of an SCM from the effect of the study that tested it, and no estimation procedure applied to this dataset could separate them. The −42.15 MPa coefficient on CPA is numerically indistinguishable from a coefficient identifying Olonade & Abdullah [12].

Sulphate exposure also identifies study. Olonade & Abdullah [12] used 0.352 and 0.367 mol/L, Wang et al. [13] 0.415 mol/L and De Souza et al. [14] 0.704 mol/L, with no concentration shared between studies. The sulphate cation varies in the same way: calcium and sodium sulphate in the first study, magnesium sulphate in the second and sodium sulphate in the third. Concentration and cation are therefore both study markers, and are addressed further in Section 4.2.

Study identity rivals SCM identity in magnitude. The SF10 category permits a direct comparison. Mean compressive strength for SF10 specimens was 63.00 MPa in Wang et al. [13] and 55.67 MPa in De Souza et al. [14], a difference of 7.33 MPa for material described identically in both. Within Wang et al. [13] alone, mean strengths across its nine SCM categories ranged from 49.00 MPa (FA30) to 63.00 MPa (SF10), a spread of 14.00 MPa. Changing the source study while holding the nominal material

fixed thus shifts mean strength by approximately half as much as changing the material entirely within one study.

This comparison should be read precisely. The two sets of SF10 specimens differ not only in provenance but in base cement (CPV-ARI against Portland P·I 42.5), sulphate cation (sodium against magnesium), sulphate concentration (0.704 against 0.415 mol/L) and exposure schedule. The 7.33 MPa difference is not attributable to any single one of these. That is the point: source study is a composite of every experimental choice a research group makes, and in a compiled dataset those choices are mostly unrecorded. The comparison measures the aggregate influence of that composite, which is what a model fitted on pooled data is exposed to.

3.5 Relationships within individual studies

If between-source variation dominates, the relationships the model is intended to capture must be sought inside each study. Table 6 reports the univariate regression of compressive strength on exposure duration within each source.

Table 6 Within-source relationship between exposure duration and compressive strength

Study	n	Slope (MPa per day)	R^2
Olonade & Abdullah [12]	40	+0.0118	0.012
Wang et al. [13]	36	−0.0066	0.070
De Souza et al. [14]	18	+0.0694	0.374

Exposure duration accounts for between 1.8% and 37.4% of within-source variance, and the slopes disagree in sign: strength increases with exposure in [12], [13] and decreases in [14]. The pooled coefficient of −0.0050 MPa per day reported in Table 4 is an average across studies that do not agree with one another, and describes none of them.

[14] is the only study with observations beyond 140 days, extending to 500, and is the only one returning a negative slope. Whether the disagreement reflects genuine differences in material behaviour, differences in exposure regime, or the limited duration of the other two programmes cannot be established from these data.

3.6 Consequences for prediction

The failure is visible in the model's predictions for individual mixtures. Table 7 and Fig. 6 compare measured and predicted compressive strength for CPA specimens at 0.367 mol/L.

Table 7 Measured and predicted compressive strength, CPA at 0.367 mol/L

Exposure (days)	Measured mean (MPa)	n	Predicted (MPa)

Exposure (days)	Measured mean (MPa)	n	Predicted (MPa)
7	6.86	4	10.57
14	8.61	4	10.53
28	11.69	4	10.46
56	11.48	4	10.32
90	8.77	4	10.14
180[a]	—	0	9.68

[a] Beyond the exposure range represented for this mixture; reported to illustrate model behaviour under extrapolation

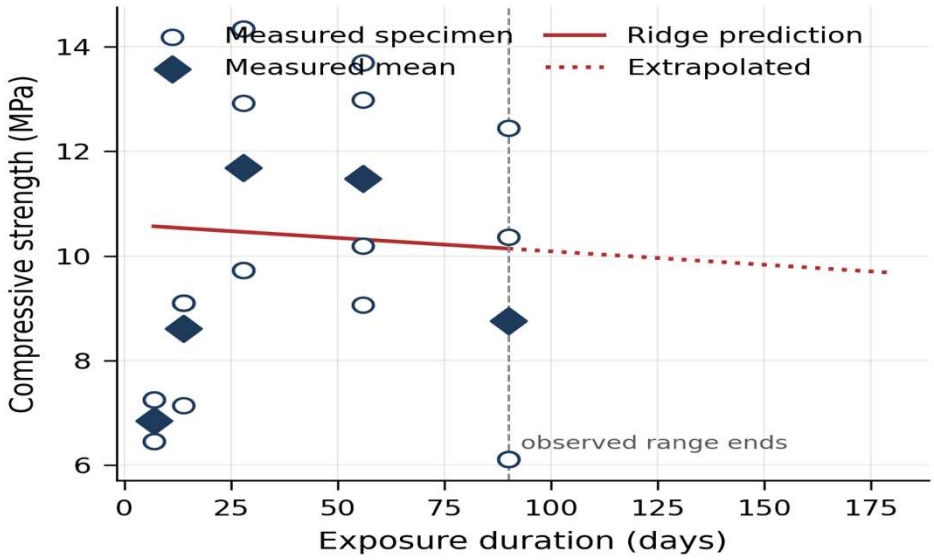


Fig. 6 Measured and predicted compressive strength of CPA mortar at 0.367 mol/L sulphate against exposure duration.

Measured group means vary between 6.86 and 11.69 MPa across the exposure series, a spread of 4.83 MPa. Predicted values fall from 10.57 to 10.14 MPa across the same series, a change of 0.43 MPa. The model returns approximately the mean strength of the category regardless of exposure duration, which is the behaviour expected of a model that has learned a level rather than a relationship.

Fig. 2 shows the same pattern for the random test partition. Predictions separate cleanly into two groups corresponding to the two strength regimes present in the dataset. Within the lower group, predictions cluster near 10.4 MPa irrespective of measured value, including for a specimen measured at 20.50 MPa. The high coefficient of determination reported in Section 3.1 arises from correctly assigning specimens to one of two levels approximately 45 MPa apart, not from resolving variation within either.

4. Discussion

4.1 Why random partitioning conceals the failure

Under random assignment, every test fold contains observations from all three source studies, and each source is represented in training by observations sharing its cement, mixing procedure, sulphate solution and testing protocol. A test observation from Olonade & Abdullah [12] is therefore predicted by a model that has already seen forty rows carrying that study's characteristics. Prediction reduces to recognising which of three well-separated strength regimes the observation belongs to, a task the SCM indicators and sulphate concentration solve directly, since each identifies its study uniquely.

Leave-one-study-out validation removes that support. The model must place a withheld study on a strength scale it has never observed, using indicators that took no value in training. Performance collapses because the quantity being estimated changes: interpolation between known study means becomes extrapolation to an unknown one.

This is not a defect of the ridge model, the predictors, or the size of the dataset. Any estimator fitted to this data would show the same behaviour, as the similarity between ridge, ordinary least squares, Lasso and random forest in Table 3 indicates. It is a property of how the data were assembled and how performance was measured.

4.2 Sulphate exposure as a proxy for study identity

The coefficient of −23.78 MPa per mol/L in Table 4 is the largest continuous effect in the model and would, taken at face value, quantify the strength penalty imposed by increasing sulphate severity. It cannot be read that way, for two reasons that compound one another.

The first is the correspondence between concentration and study already noted. Each study used a single concentration or a narrow pair: 0.352 and 0.367 mol/L in Olonade & Abdullah [12], 0.415 in Wang et al. [13] and 0.704 in De Souza et al. [14]. The coefficient can be traced to a single contrast. SF10 is the only category present in two studies, and its mean strength falls by 7.33 MPa across a concentration difference of 0.289 mol/L, an implied gradient of −25.4 MPa per mol/L. The fitted coefficient of −23.78 is close enough to that figure to be regarded as an estimate of it. In other words, the largest continuous effect in a model fitted to 94 observations rests almost entirely on the comparison of thirteen silica fume specimens drawn from two laboratories.

The second reason is that concentration is not the only thing varying along that contrast. The sulphate cation varies with it, and the cation is not a minor detail. Sodium and magnesium sulphate attack cement by different mechanisms: both form gypsum and ettringite, but magnesium sulphate additionally converts calcium hydroxide to brucite and decalcifies the C-S-H to magnesium silicate hydrate, which has negligible binding capacity. The consequence is measurable. De Souza et al. [14], whose data supply the higher-concentration end of the contrast, report that their silica fume and rice husk ash mortars resisted sodium sulphate almost completely over 140 days while losing 9.8% and 8.6% of compressive strength respectively under magnesium sulphate at the same sulphate ion concentration.

A coefficient estimated across studies that used different cations is therefore not a concentration effect at all. It is a composite of concentration, cation, base cement and laboratory, and the source studies themselves provide direct evidence that at least one of those components acts independently of concentration. Isolating a concentration effect would require observations at several concentrations of the same salt within one experimental programme. No such contrast exists in this dataset, and it is unlikely to exist in most compiled datasets, since individual studies typically fix both salt and concentration by reference to a testing standard.

4.3 What compiled datasets would require

A model able to predict across studies would need access to the attributes that differ between them. These include water-to-binder ratio, cement chemical composition and in particular tricalcium aluminate content, curing regime prior to exposure, aggregate type and grading, specimen geometry, the sulphate salt used, and the testing standard followed. Databases assembled for related purposes have incorporated several of these: the expansion database of Hilloulin et al. [7] combines cement composition, mixture proportions, sulphate solution characteristics and specimen geometry as inputs.

The obstacle is that these attributes are reported inconsistently. A study concerned with SCM replacement level may report water-to-binder ratio without cement oxide composition; another may report the reverse. Compiling across studies therefore forces a choice between retaining observations and retaining covariates, and the datasets that result are typically wide in observations and narrow in the variables that would explain differences between them. The present dataset records SCM type, replacement level, exposure duration, sulphate concentration and sulphate cation, of which the first three are largely determined by study and the last two entirely so.

Until reporting practice converges, the practical implication is that models compiled from heterogeneous studies should be presented as descriptions of the pooled data rather than as predictive tools, unless cross-study transfer has been demonstrated explicitly.

4.4 Implications for reporting

Two measures would materially improve the interpretability of published models in this area, and neither requires additional data.

The first is grouped validation. Where a dataset is compiled from multiple sources, holding out whole sources costs nothing beyond a modification to the validation loop and provides a direct estimate of transferability. In the present case it changed the reported coefficient of determination from 0.968 to values between −12.28 and −162.35. A study reporting only the former would give a materially misleading account of what its model can do.

The second is reporting the between-source share of variance in the response. This is a descriptive statistic requiring no modelling, and it establishes at the outset how much of the available signal originates in differences between contributing studies. Where that share approaches the 96.1% observed here, a high coefficient of determination should be interpreted as evidence that sources are distinguishable, not that the underlying material behaviour has been captured.

It follows that a coefficient of determination above 0.90 reported for a model trained on compiled durability data, absent either check, is uninformative about predictive validity. It is consistent with a model that transfers well and equally consistent with one that does not transfer at all.

4.5 Limitations

The dataset comprises three studies. Whether the confounding documented here is typical of compiled cementitious durability datasets, or particular to this combination, cannot be established from three cases. The mechanism, between-study heterogeneity in unrecorded attributes, is general, but its magnitude in other datasets is an empirical question.

Eight of the eleven SCM categories are represented by four observations each. Coefficients estimated on such categories are unstable, and the RHA coefficient in particular derives from nine observations in a single source. These values are reported for completeness and should not be read as material effects, a caution that Section 3.4 extends to the coefficient table as a whole.

The merge of the silica fume specimens of De Souza et al. [14] into the SF10 category, described in Section 2.3, treats materials prepared with different base cements and exposed to different sulphate salts as equivalent. The cross-study comparison in Section 3.4 depends on this merge and should be weighed accordingly. Without it, no SCM category would appear in more than one study and the comparison could not be made at all, which is itself indicative of the dataset's structure.

Compressive strength was the only durability indicator available across all three studies. Expansion, mass change and microstructural measurements were not uniformly reported and could not be included. Findings therefore concern strength prediction specifically.

Only the sodium sulphate series of De Souza et al. [14] was extracted, as recorded in Section 2.1. Their magnesium sulphate series, which shows losses of compressive strength that the sodium sulphate series does not, is absent from the compiled dataset. Including it would have introduced a second exposure condition within one study while leaving the other two studies with one each, and would not have resolved the confounding between cation and study documented in Section 3.4. Its exclusion nevertheless means that the dataset under-represents the more aggressive of the two exposures reported by that study.

Values extracted from figures by digitisation carry reading error that could not be quantified against the original measurements.

5. Conclusions

1. A ridge regression model of compressive strength in SCM-blended mortar under sulphate exposure achieved a coefficient of determination of 0.968 on a randomly held-out test set, comparable to ordinary least squares (0.968), Lasso (0.966) and random forest (0.982). Under leave-one-study-out validation, every fold returned a negative coefficient of determination, between −12.28 and −162.35, with mean absolute errors up to 42.45 MPa on specimens whose measured strengths did not exceed 21.50 MPa.

2. Variance decomposition attributed 96.1% of total variance in compressive strength to differences between source studies and 3.9% to variation within them.
3. Ten of the eleven SCM categories in the compiled dataset occurred in a single study, and each sulphate concentration and cation was used by one study only. Material effects and study effects are therefore not statistically separable in this dataset, and the fitted coefficients cannot be interpreted as material effects.
4. For the one SCM category present in two studies, mean compressive strength differed by 7.33 MPa between studies, approximately half the 14.00 MPa range observed across all nine SCM categories within a single study. The sulphate concentration coefficient is traceable to this single contrast, along which cation and base cement also vary.
5. Leave-one-study-out validation and reporting of the between-study variance share should be treated as minimum requirements for machine learning models trained on literature-compiled cementitious durability data. Absent these, a high reported coefficient of determination is uninformative about predictive validity.

Declarations

- **Funding Declaration:** No external funding was received to assist with the preparation of this manuscript.
- **Author Contribution:** All authors contributed equally to the conception, execution and preparation of this research.
- **Data Declarations:** The compiled datasets are provided as supplementary information files.
- **Ethics declarations:** Not applicable
- **Consent to Participate declarations:** Not applicable
- **Consent to Publish declaration:** Not applicable
- **Clinical Trial:** Not applicable.

Data availability

The dataset compiled and analysed in this study, together with the analysis code used to generate all tables and figures, is provided as supplementary information files accompanying this article. The underlying experimental measurements were originally reported in the published studies cited as references [12], [13] and [14].

References

[1] Andrew, R.M., 2019. Global CO2 emissions from cement production, 1928–2018. Earth System Science Data 11, 1675–1710. https://doi.org/10.5194/essd-11-1675-2019

[2] Rozière, E., Loukili, A., El Hachem, R., & Grondin, F. (2009). Durability of concrete exposed to leaching and external sulphate attacks. Cement and Concrete Research, 39(12), 1188-1198. https://doi.org/10.1016/j.cemconres.2009.07.021

[3] El-Hachem, R., Rozière, E., Grondin, F., & Loukili, A. (2012). Multi-criteria analysis of the mechanism of degradation of Portland cement based mortars exposed to external sulphate attack. Cement and Concrete Research, 42(10), 1327-1335. https://doi.org/10.1016/j.cemconres.2012.06.005

[4] Mousavinezhad S, Toledo WK, Newtson CM, Aguayo F. Rapid assessment of sulfate resistance in mortar and concrete. Materials. 2024;17(19):4678. https://doi.org/10.3390/ma17194678

[5] Zhang, M., Gu, Z., Zhao, Y., Fu, Y., & Kong, X. (2024). Compressive strength prediction of cement base under sulfate attack by machine learning approach. Case Studies in Construction Materials, 21, e03652. https://doi.org/10.1016/j.cscm.2024.e03652

[6] Hafidi A, Hilloulin B, Boudache S, Rejoice U, Loukili A. Comparison of machine learning algorithms for the prediction of the external sulphate attack resistance of blended cements. In: Jędrzejewska A, Kanavaris F, Azenha M, Benboudjema F, Schlicke D, editors. SynerCrete 2023. RILEM Bookseries, vol 44. Cham: Springer; 2023. https://doi.org/10.1007/978-3-031-33187-9_67

[7] Hilloulin, B., Hafidi, A., Boudache, S., & Loukili, A. (2023). Interpretable ensemble machine learning for the prediction of the expansion of cementitious materials under external sulfate attack. Journal of Building Engineering, 80, 107951. https://doi.org/10.1016/j.jobe.2023.107951

[8] Taffese WZ, Hilloulin B, Villagran Zaccardi Y, Marani A, Nehdi ML, Hanif MU, et al. Machine learning in concrete durability: challenges and pathways identified by RILEM TC 315-DCS towards enhanced predictive models. Mater Struct. 2025;58(4):145. https://doi.org/10.1617/s11527-025-02664-3

[9] Meredig B, Antono E, Church C, Hutchinson M, Ling J, Paradiso S, et al. Can machine learning identify the next high-temperature superconductor? Examining extrapolation performance for materials discovery. Mol Syst Des Eng. 2018. https://doi.org/10.1039/C8ME00012C

[10] Samantha Durdy, Michael W. Gaultois, Vladimir V. Gusev, Danushka Bollegala, Matthew J. Rosseinsky, Random projections and kernelised leave one cluster out cross-validation: universal baselines and evaluation tools for supervised machine learning of material properties. Digital Discovery. 2022;1(6):763-778, https://doi.org/10.1039/d2dd00039c

[11] Li, Q., Fu, N., Omee, S.S. et al. MD-HIT: Machine learning for material property prediction with dataset redundancy control. npj Comput Mater 10, 245 (2024). https://doi.org/10.1038/s41524-024-01426-zu

[12] Olonade KA, Abdullah AG. Effect of sulphate attack on the strength of cement brands blended with cassava peel ash. ASRIC J Eng Sci. 2021;1:40–45

[13] Wang Y, Ge Y, Wang X, Chen X, Li Q, Zhang Y. Magnesium sulfate attack resistance of SCM-modified cement-based materials. Ceram Silik. 2021;65(3):207–214. https://doi.org/10.13168/cs.2021.0020

[14] Souza DJ, Medeiros MHF, Hoppe Filho J. Evaluation of external sulfate attack (Na2SO4 and MgSO4): Portland cement mortars containing siliceous supplementary cementitious materials. Rev IBRACON Estrut Mater. 2020;13(4):e13403. https://doi.org/10.1590/S1983-41952020000400003

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- finalcleanedsulfatedatav4.csv
- Manuscript1.ipynb