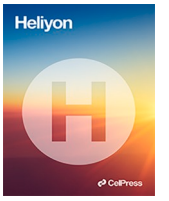
Heliyon 9 (2023) e15898

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

The synergic approach between machine learning, chemometrics, and NIR hyperspectral imagery for a real-time, reliable, and accurate prediction of mass loss in cement samples

Abderrahim Diane [a,b,*], Taoufiq Saffaj [a], Bouchaib Ihssane [a], Reda Rabie [b]

[a] University Sidi Mohamed Ben Abdellah, Faculty of Sciences and Techniques of Fez, Laboratory of Applied Organic Chemistry, Fez, Morocco
[b] Moroccan Foundation for Advanced Science, Innovation and Research, MAScIR Rabat, Morocco

ARTICLE INFO

Keywords:
Hyperspectral imaging
Near InfraRed
Machine learning
Chemometrics
Spectroscopy

ABSTRACT

Alternative and non-destructive analytical methods that predict analyte concentration accurately and immediately in a specific matrix are becoming vital in the analytical chemistry domain. Here, a new innovative and rapid method of predicting mass loss of cement samples based on a combination of Machine Learning (ML) and the emerging technique called Hyperspectral Imaging (HSI) is presented. The method has proved its reliability and accuracy by providing a predictive ML model, with satisfactory best validation scores recorded using partial least squared regression, with a reported ratio of performance to inter-quartile distance and root mean squared error of 12,89 and 0.337, respectively. Moreover, the possibility of optimizing and boosting the performance of the method by optimizing the predictive model performance has been suggested. Therefore, a features selection approach was conducted to disqualify non-relevant wavelengths and stress only relevant ones in order to make them the only contributors to a final optimized model. The best selected features subset was composed of 28 wavelengths out of 121, found by applying genetic algorithm combined to partial least squares regression as a feature selection method, on spectra preprocessed consecutively by the first-order savitzky-golay derivative calculated with 7-point quadratic SG filter, and multiplicative scatter correction method. The overall results show the possibility of combining HSI and ML for fast monitoring of water content in cement samples.

1. Introduction

Cement is widely regarded as one of the most important building materials on the global scale and is one of the main components involved in the formulation of high-strength and high-performance concrete. Magnesium Oxychloride Cement (MOC) known as Sorel is one of the attractive cement types used in civil engineering, it is generally formulated by blending magnesium chloride solution (MgCl2–H2O) with magnesium oxide powder (MgO), and is famous for its high physical properties expressed by relatively high early strength and a low coefficient of thermal expansion. However, in the case of prolonged water contact, those famous properties degrade rapidly due to its notorious weak water resistance property [1–3]. Recently, significant contributions from research and development

* Corresponding author. University Sidi Mohamed Ben Abdellah, Faculty of Sciences and Techniques of Fez, Laboratory of Applied Organic Chemistry, Fez, Morocco .
E-mail address: abderrahim.diane@usmba.ac.ma (A. Diane).

https://doi.org/10.1016/j.heliyon.2023.e15898
Received 14 June 2022; Received in revised form 21 April 2023; Accepted 25 April 2023
Available online 30 April 2023

scientists have been made to enhance the water resistance property by introducing new MOC formulations derived from the original one [2,3]. However, evaluating these formulations requires a reliable and feasible analytical method for water monitoring. Several rapid and non-invasive analytical methods that helps achieving this task, including the RGB imaging and the spectral imaging, have been introduced in the literature. However, most of which do not offer detailed information about the sample.

HyperSpectral Imaging (HSI) is an innovative technique that has drawn the attention of chemists as it provides not only the spectral signature of the sample as spectral methods do, but also the spatial information that helps the analyst obtain a detailed overview of the sample state, composition, homogeneity, etc. HSI technique consists of a 2D image acquisition at each Wavelength (Wl) over a specific range of Wls. And therefore, provides three-dimensional data, often named datacube, whose first two dimensions correspond to the spatial information of the sample, and the third one reflects its spectral information [4]. Making profit from this technique requires having a relatively important background in Machine Learning (ML) and chemometric methods. These methods are a cluster of statistical and mathematical tools that are devoted to developing either predictive or descriptive models (multivariate calibration performing) which performance would be the decider on whether the technique might be utilized for reliable prediction making or not. Multivariate calibration, here, could be defined as the art of manipulating high-dimensional data so as to estimate a statistical model or equation that could predict unknown continuous or discrete values or even make a decision, with a certain level of accuracy. This manipulation was a time-consuming and difficult decades ago, since the available hardware performance at that time was not sufficient to implement hard and complex computations. This was the case until the dawn of hardware acceleration which facilitated the task and attracted scientists to head researches in the field in order to innovate new techniques based on the light-matter interactions such as spectral and HSI techniques that have been successfully used through the past few years to achieve different objectives including matter detection, identification, and quantification, chemical process monitoring, quality monitoring, cancer diagnosis and so on [4–8].

The overall objective of this study is to show the result of coupling HSI, ML and chemometric methods to develop a new reliable analytical method devoted to predicting the Mass Loss (MaL) in cement samples. In the course of this study, raw spectra were extracted from samples HSI images, assessed to remove any kind of unwanted noise probably affecting them, then an exploratory data analysis was conducted by the mean of Principal Components Analysis (PCA) on these spectra to identify the number of factors able to explain the variability and also detect any outlier spectrum. Subsequently, multivariate calibration, and a features selection approach were performed where various regression methods and features selection methods available in the literature have been used to develop high-performance ML predictive models.

2. Materials and methods

2.1. Data collection and softwares

The data used in this study consists of three magnesium oxychloride cement sample replicates, each of which was scaled at nine different time points for MaL determination and immediately imaged using a NIR-HSI camera that works in the range that extends from 880 to 1720 nm. Data preprocessing, spectral preprocessing and features selection were performed with R language (v4.1.2), while full range ML models were developed with Python Language (v3.10.1).

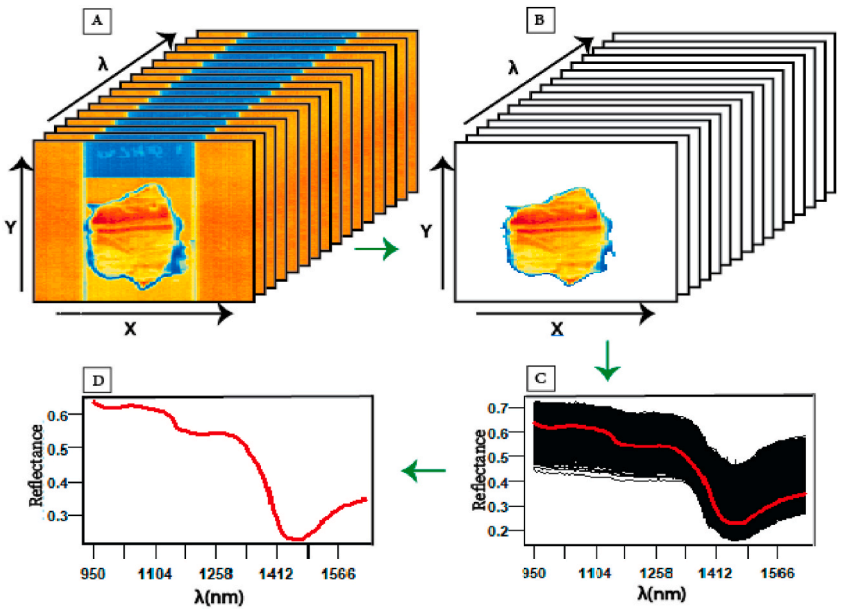


Fig. 1. HSI images processing steps: (A) image acquisition, (B) Background removal, (C) Spectra acquisition, (D) Mean spectrum extraction.

2.2. Chemometrics & machine learning

After data collection, NIR-HSI images (Fig. 1A) were preprocessed to remove the irrelevant background (Fig. 1B), subsequently, a mean spectrum was extracted from each image and considered as the sample spectrum (Fig. 1C and D). All extracted mean spectra were then gathered together and smoothed using the moving average filter to remove the low signal fluctuations [9–11,18].

2.3. Spectral preprocessing

Preprocessing of Near-InfraRed (NIR) spectra has been an integral part of our multivariate calibration or modeling process. It was performed to improve the quality of acquired spectra as it consists of biased spectra correction and irrelevant variation removal. Therefore, it helps to develop high-performance predictive models. For our raw spectra, the performed spectral preprocessing process consisted of using three signal preprocessing methods after bias identification and quantification, which are: Multiplicative Scatter Correction (MSC) and Standard Normal Variate (SNV) for Light scattering and particles size effects reduction [14–17], the first-order Savitzky-Golay (SG) derivative calculated with 7-point quadratic SG filter for smoothing and removing probable band overlapping and bands shift [12,13,18] and finally the combination of SG derivative and MSC.

2.4. Exploratory data analysis

Exploratory data analysis was conducted by means of Principal Components Analysis (PCA), the core of chemometrics. It is been used for overcoming the curse of high dimensionality, as it projects data from high-dimensional space onto a new subspace made of new independent synthetic variables named Principal Components (PCs). The PCs themselves are synthetic variables made of a linear combination of correlated variables that concentrate and compress the information contained in those correlated variables in the linear combination. The projection onto a simple subspace of two or three dimensions gives us an insight about outliers, the sufficient number of PCs able to compress and resume relevant information in the high-dimensional data, homogeneity, and the form of distribution of the data, as well as possible samples clusters [19].

2.5. Multivariate calibration

Exploratory data analysis was then followed by the multivariate calibration, in other terms, ML models development that predicts a reference value Y (slow, time-consuming measurement and sometimes hard or impossible to obtain) from the corresponding recorded spectrum X (fast and easy measurement). To succeed in this stage, multiple regressors available in the literature including Partial Least Squares (PLSR) [20–23], Principal Components (PCR) [24], Support Vector Machine (SVMR) [20,25,26], Decision Tree (DTR) [27], K-Nearest Neighbors (KNNR) [28,29], eXtreme Gradient Boosting (XGBR) [30], Light Gradient Boosting Machine (LGBMR) [31], Categorical Boosting (CBR) [32], and MultiLayer Perceptron (MLPR) [33] with two hidden layers, were used for developing our predictive models. Due to the low number of samples, leave-triplicate-out Cross Validation (CV) [34] scores were adopted to assess the developed predictive models, overcome underfitting and overfitting problems as well as to drive models benchmarking study.

2.6. Models evaluation

The evaluation of models performance is a critical step in the predictive ML development process. It enables a thorough understanding of how developed predictive models will perform in the future on unseen data. Different statistical metrics, such as the coefficient of determination (R^2) (1), the Root Mean Squared Error (RMSE) (2), the Ratio of Prediction to Deviation (RPD) (3), and the Ratio of Performance to Inter-Quartile distance (RPIQ) (4), are commonly used to evaluate models. In our study, all the aforementioned metrics were used to evaluate the developed predictive models. The computational formula for each metric is provided below [30].

$$\boldsymbol{R^2 = 1 - \frac{\sum_{i=1}^{N} (yi - \hat{y}i)^2}{\sum_{i=1}^{N} (yi - y)^2}} \tag{1}$$

$$\boldsymbol{RMSE = \sqrt{\frac{\sum_{i=1}^{N} (\hat{y}i - y)^2}{N}}} \tag{2}$$

$$\boldsymbol{RPD = \frac{S}{RMSE}} \tag{3}$$

$$\boldsymbol{RPIQ = \frac{IQ}{RMSE}} \tag{4}$$

Where N and $\hat{y}$ are the total number of samples and the predicted values, respectively. And yi, $\bar{y}$, S, IQ, are the measured values, the average, the standard deviation and the interquartile range of the observed values, respectively.

The best model is typically characterized by having the highest R^2, the lowest RMSE, and the highest RPD and RPIQ.

2.7. Hyperparameter tuning

Since all implicated regressors have at least two or three hyperparameter to be tuned, simulated annealing [35,36] has been embedded in each regression method to achieve this tuning. It is a Meta-heuristic algorithm that has been widely used for solving several optimization problems. The best tune was the one that minimizes the Root Mean Squared Error of Cross Validation (RMSECV) of the predictive model.

2.8. Features subset selection

Subsequently, a Features Subset Selection (FSS) approach has been conducted to identify and extract the influential regions of the spectra, remove probable interferences and noise. This approach generally helps enhancing the interpretation ability and the predictive quality of ML models. FSS approach in our case consisted of using four regression methods widely used for features selection purposes: Lasso (LR) [29,37], and ElasticNet (ENR) [37,38] regulizers, Backward Interval-Partial Least Squares (BI-PLSR) [30,39,40] which is used mostly for spectroscopic data, and lastly Genetic Algorithm-Partial Least Squares (GA-PLSR) that consists of embedding the evolutionary meta-heuristic algorithm named Genetic Algorithm (GA) [30,39] which is inspired by the process of biological natural selection theory, in PLSR algorithm.

3. Results and discussion

The evolution of the spatial distribution of one sample replicate at nine time points is shown in Fig. 2. The samples Mass Loss (MaL) vector is visualized in Fig. 3. It can be seen that the samples progressively lose their mass through the first 3 h due to water evaporation. Then, due to the effect of uncontrolled factors variation such as temperature range and moisture level in the air, all the three replicates start becoming heavier. This presumably can be explained by the increase in their moisture level. The corresponding reflectance spectra are visualized in Fig. 4. They were acquired by computing the mean spectrum of each of the 27 HSI images. Gathered mean spectra show a large band that extends from about 1350 to 1450 nm that is found to be the first overtone of the water molecule [41–43]. Irrelevant variability in the form of a difference in reflectance value between spectra through the entire Wl range is remarkable, which requires to be suppressed or reduced using spectral preprocessing methods.

3.1. Exploratory data analysis

Application of PCA algorithm on the spectra preprocessed differently has yielded the result shown in the Table 1. It shows, for each preprocessing method, that the first two PCs were able to explain over 96% of total variability with the first PC alone explaining over

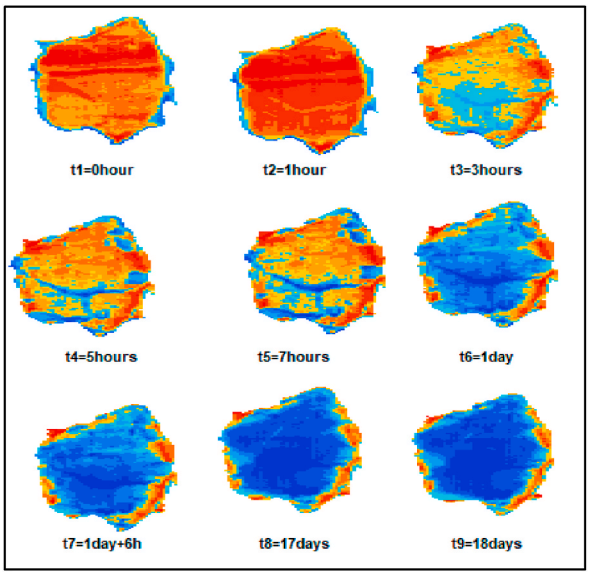


Fig. 2. HSI image of the first sample replicate at nine different time points at wavelength of 1223 nm indicating the evolution spatial distribution of the sample.

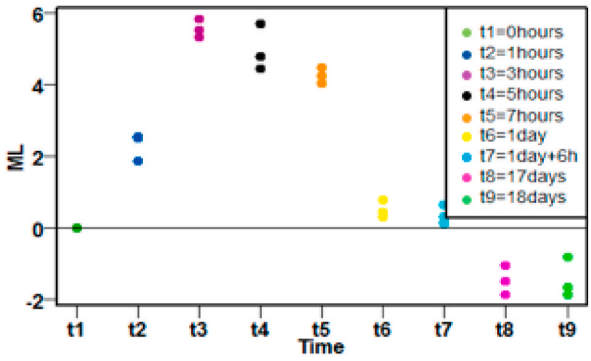


Fig. 3. The evolution of mass loss in analyzed samples over time.

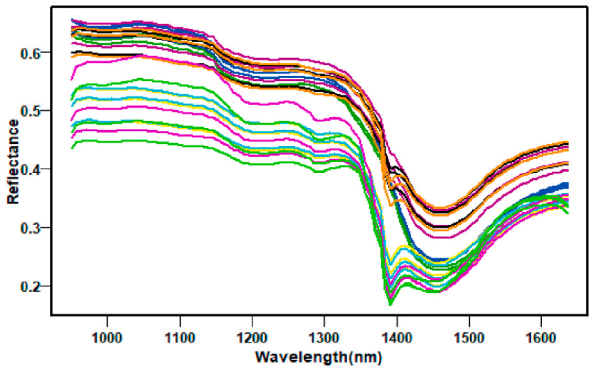


Fig. 4. Near-infrared spectra of cement samples extracted from the HSI images.

90% of the total variability in the data, this means that the information contained in the data in not complicated. The scores plot Fig. 5 shows, for each preprocessing method, that with only the first two PCs, spectra clusters could be created based on their recording time. This means that these two PCs hold the largest part of the relevant information contained in the spectra. Notice that when applying PCA to spectra preprocessed using SG_MSC, the first PC, which explained approximately 94% of the total variability allowed to make this clustering which pushed us to make the hypothesis that the SG_MSC preprocessing method might separate relevant overlapped bands.

3.2. Multivariate calibration

Multivariate Calibration performing on data preprocessed differently has given the recorded scores presented in the Table 2. The scores of the best developed predictive models are presented in Table 3, they were selected based on validation scores, the number of hyperparameters and the complexity level. It can be seen that when trying to model the relationship between MaL and the NIR spectra. PLSR and PCR methods in conjunction with either MSC and SG_MSC spectral preprocessing methods allowed obtaining the most performant and efficient predictive models characterized with the lowest RMSE both in the calibration phase RMSEC (0.05 and 0.11, respectively) and the validation phase RMSECV (0.333 and 0.489, respectively), and highest R^2CV(99 and 97, respectively). PLSR and PCR are two of the preferred linear regression algorithms used for spectral data modeling as they are efficient and possess only few

Table 1
The percentage of explained variability by each PC for raw spectra and spectra preprocessed differently.

Number of PCs	Preprocessing			
	Raw (1)	MSC (2)	SNV (3)	SG_MSC (4)
1	90.94	91.14	90.80	94.02
2	7.80	5.99	6.11	3.41
3	1.01	2.12	2.30	1.82
4	0.18	0.50	0.51	0.50
5	0.06	0.13	0.14	0.17
6	0.01	0.07	0.08	0.04
7	0.01	0.02	0.03	0.02

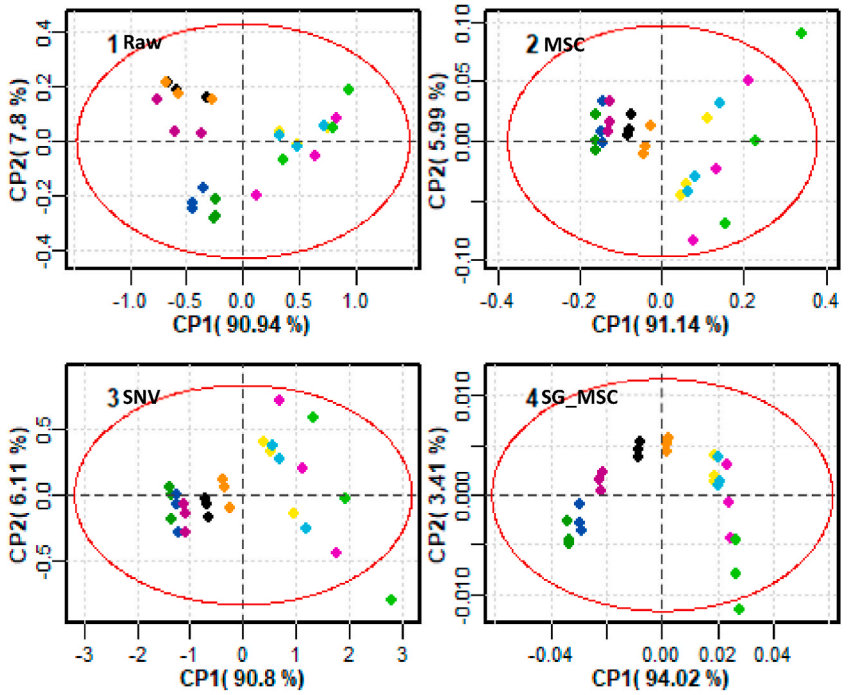


Fig. 5. PCA scores plot.

hyperparameters that need tuning. LGBM also allowed the development of high-performance and robust models, however, on the contrary to PLSR and PCR, it possesses numerous hyperparameters that need a performant computer to be tuned.

NIR-HSI has been innovated to provide both the spatial and the spectral information of samples over the NIR range. The spectral information consists of a large number of contiguous Wls which present a source of informative information that is useful for developing high-performance calibration models, noise that lower the predictive performance of the calibration models, and redundant information which contributes to the curse of dimensionality. ML regression algorithms do not always succeed in catching only informative information, a thing that results in complex and low-performance models [30]. Therefore, the identification of informative Wls by testing the significance of models parameters and conducting a feature selection approach to retain them were conducted.

3.3. Significance testing of the retained prediction models' coefficients using jack-knife based t-test

The significance test for regression coefficients, as the name indicates, is a diagnostic tool that provides a general overview of the coefficient estimates, allowing for the identification of passive features and the stressing of active ones. It is a tool that helps spectroscopy researchers to determine which Wls carry the relevant spectral signature or information of the sample. Various methods, such as the Variable Importance in Projection (VIP) score and the Jack-knife-based *t*-test, have been proposed in the literature for testing the significance of the estimated regression coefficients. The former consists of determining the statistical contribution of each Wl to the overall fitted PLSR model, whereas the latter consists of assessing the coefficients' uncertainty and stability using the jack-knifing method [44–48].

The jack-knife-based *t*-test was used in this study to identify useful Wls based on the inference on the regression coefficient. The selection probability for PLSR model coefficients at a 95% significance level was reported in Fig. 6. It clearly shows that only a few features are statistically stable and contribute significantly to model construction, indicating that not all the NIR range implicated in this study provides relevant information. This means that the three retained PLSR models carry uninformative and noisy Wls, and the corresponding coefficients add complexity and error to the models. These Wls could be suppressed using a different method or approaches, including simple filtering based on model coefficient selection probability and driving a feature subset selection approach.

3.4. Features subset selection (FSS)

Regardless of how satisfying the predictive quality of our best models is, high dimensionality, irrelevant information, and noise remain a curse that must be broken since they reduce predictive model performance and efficiency, reduce interpretability and comprehension levels, and frequently lead to over-fitting.

The application of FSS methods on the data (121 Wls) allowed for the development of models with performance metrics summarized in the Table 4. It shows that ENR and LR were effective at reducing the number of Wls (Nvar), but the developed models did not outperform the models developed without FSS, implying that both selection methods either eliminated some relevant information

Table 2
Scored metrics of the developed prediction models.

PCR	MSC	0.111	97	0.489	5.25	8.88	SNV	0.191	99	0.338	7.59	12.85	SG_MSC	0.194	97	0.496	5.17	8.76
PLSR		0.05	99	0.333	7.7	13.04		0.042	99	0.339	7.57	12.81		0.132	99	0.337	7.61	12.89
DTR		0.516	81	1.338	1.92	3.25		0.592	86	1.158	2.22	3.75		0.0	88	1.048	2.45	4.14
KNNR		0.0	94	0.726	3.53	5.98		0.0	94	0.746	3.44	5.82		0.0	96	0.633	4.05	6.86
SVMR		0.937	88	1.064	2.41	4.08		0.927	88	1.056	2.43	4.11		0.595	94	0.765	3.35	5.68
RFR		0.791	79	1.419	1.81	3.06		0.785	80	1.359	1.89	3.2		0.568	92	0.847	3.03	5.13
LGBMR		0.0	93	0.813	3.16	5.34		0.146	88	1.041	2.46	4.17		0.0	99	0.35	7.33	12.41
XGBR		0.0	92	0.854	3.0	5.09		0.0	92	0.876	2.93	4.96		0.0	96	0.647	3.96	6.71
CTR		0.0	90	0.973	2.64	4.46		0.0	91	0.9	2.85	4.83		0.026	95	0.698	3.68	6.22
MLPR		1.833	60	1.931	1.33	2.25		0.792	90	0.982	2.61	4.42		2.072	52	2.114	1.21	2.05
Model		RMSEC	R^2CV	RMSECV	RPD	RPIQ		RMSEC	R^2CV	RMSECV	RPD	RPIQ		RMSEC	R^2CV	RMSECV	RPD	RPIQ

Table 3
The scores of the best predictive models.

MODÈL	PREPROCESSING	R^2CV	RMSECV	RPDCV	RPIQCV
PLSR	MSC	99	0.333	7.7	13.04
PCR	SNV	99	0.338	7.59	12.85
PLSR	SNV	99	0.339	7.57	12.81
PLSR	SG_MSC	99	0.337	7.61	12.89
LGBMR	SG_MSC	99	0.35	7.33	12.41

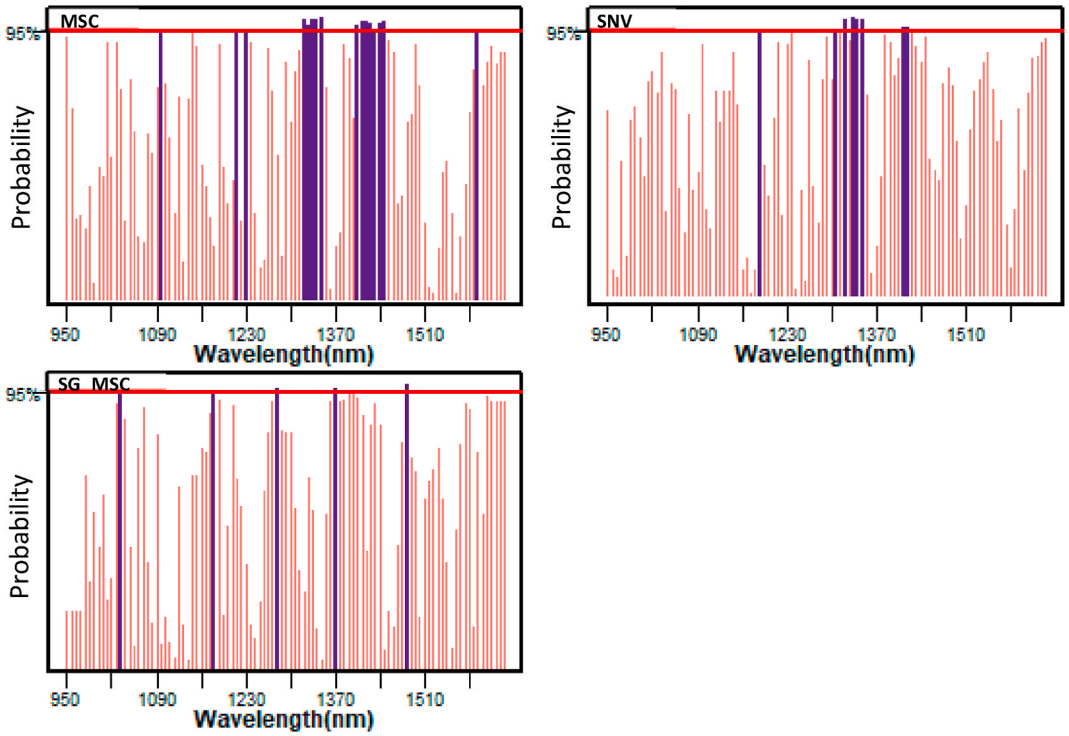


Fig. 6. Jack-knife based *t*-test results for PLSR models parameters significance indicating the selection probability of each wavelength.

Table 4
Summary of the Feature selection approach.

Regression	Preprocessing	RMSEC	R^2CV	RMSECV	Nvar
ENR	MSC	0.112	97	0.564	58
	SNV	0.172	93	0.821	52
	SG_SNV	0.129	98	0.445	63
LR	MSC	0.239	97	0.561	30
	SNV	0.193	93	0.816	46
	SG_SNV	0.126	98	0.44	64
GA_PLSR	MSC	0,017	100	0,113	25
	SNV	0,011	100	0,113	27
	SG_SNV	0,01	100	0,112	28
BI-PLSR	MSC	0,0373	100	0,1118	39
	SNV	0,0428	100	0,1468	45
	SG_SNV	0,072	100	0,1284	40

or retained noise. The GA-PLS and BI-PLS, on the other hand, have enabled the development of models that outperform models developed without FSS and that with fewer Wls, as expressed by lower RMSEC, RMSECV, Nvar, and higher R^2CV. The findings demonstrated that FSS played an important role in improving the performance of predictive models by removing irrelevant and redundant information partially or completely from the data.

4. Conclusion

Hyperspectral imaging is a complex, multidisciplinary field that can be defined as the simultaneous acquisition of spatial images at multiple spectral wavelengths. It permits the gathering of detailed information about the sample analyzed. In this study, a novel method for estimating mass loss in magnesium oxychloride cement samples using the synergy of hyperspectral imaging, machine learning, and chemometrics was introduced. A variety of techniques were involved to achieve the goal, beginning with hyperspectral image acquisition and ending with the development and optimization of machine learning models. The overall findings demonstrated the method's reliability by allowing for high-accuracy mass loss estimation. As a result, it could be concluded that machine learning algorithms and hyperspectral imaging will become indispensable tools for cement quality control in the future.

Author contribution statement

Abderrahim DIANE: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Wrote the paper. Taoufiq SAFFAJ, Bouchaib IHSSANE, Reda RABIE: Performed the experiments; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Data availability statement

Data included in article/supp. material/referenced in article.

References

[1] Y. Tan, Y. Liu, Z. Zhao, J.Z. Paxton, L.M. Grover, Synthesis and in vitro degradation of a novel magnesium oxychloride cement, J. Biomed. Mater. Res. 103 (1) (Jan. 2015) 194–202.
[2] K. Li, et al., Superhydrophobic magnesium oxychloride cement based composites with integral stability and recyclability, Cem. Concr. Compos. 118 (Apr. 2021) 103973.
[3] Y. Guo, Y.X. Zhang, K. Soe, R. Wuhrer, W.D. Hutchison, H. Timmers, Development of magnesium oxychloride cement with enhanced water resistance by adding silica fume and hybrid fly ash-silica fume, J. Clean. Prod. 313 (Sep. 2021) 127682.
[4] R. Dorrepaal, C. Malegori, A. Gowen, Tutorial: time series hyperspectral image analysis, J. Near Infrared Spectrosc. 24 (2) (2016) 89–107.
[5] U. Siripatrawan, Y. Makino, Y. Kawagoe, S. Oshita, Rapid detection of Escherichia coli contamination in packaged fresh spinach using hyperspectral imaging, Talanta 85 (1) (Jul. 2011) 276–281.
[6] M. Grosjean, B. Amann, C. Butz, B. Rein, W. Tylmann, Hyperspectral imaging: a novel, non-destructive method for investigating sub-annual sediment structures and composition, Past Global Changes Magazine 22 (1) (Apr. 2014) 10–11.
[7] C. Malegori, et al., Identification of invisible biological traces in forensic evidences by hyperspectral NIR imaging combined with chemometrics, Talanta 215 (Aug. 2020) 120911.
[8] C. Manis, et al., Non-destructive age estimation of biological fluid stains: an integrated analytical strategy based on near-infrared hyperspectral imaging and multivariate regression,, Talanta (Apr. 2022) 123472.
[9] R. Webster, Spectral analysis of gilgai soil, Soil Res. 15 (3) (1977) 191–204.
[10] C. Vaiphasa, Consideration of smoothing techniques for hyperspectral remote sensing, ISPRS J. Photogrammetry Remote Sens. 60 (2) (Apr. 2006) 91–99.
[11] R. Webster, Spectral analysis of gilgai soil, Soil Res. 15 (3) (1977) 191–204.
[12] C. Vaiphasa, Consideration of smoothing techniques for hyperspectral remote sensing, ISPRS J. Photogrammetry Remote Sens. 60 (2) (Apr. 2006) 91–99.
[13] J. Ji, et al., [Spectral smoothing with adaptive multiscale window average], Guang Pu Xue Yu Guang Pu Fen Xi 35 (5) (May 2015) 1445–1449.
[14] J. Luypaert, S. Heuerding, Y. Vander Heyden, D.L. Massart, The effect of preprocessing methods in reducing interfering variability from near-infrared measurements of creams, J. Pharm. Biomed. Anal. 36 (3) (Nov. 2004) 495–503.
[15] A. Kohler, M. Zimonja, V. Segtnan, H. Martens, Standard normal variate, multiplicative signal correction and extended multiplicative signal correction preprocessing in biospectroscopy, Compr. Chemom. 2 (2009) 139–162.
[16] R.J. Barnes, M.S. Dhanoa, S.J. Lister, Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra 43 (5) (Aug. 2016) 772–777.
[17] T. Isaksson, T. Naes, The effect of multiplicative scatter correction (MSC) and linearity improvement in NIR spectroscopy 42 (7) (Aug. 2016) 1273–1284.
[18] R.W. Schafer, What is a savitzky-golay filter? IEEE Signal Process. Mag. 28 (4) (2011) 111–117.
[19] I.T. Jolliffe, Principal Components in Regression Analysis, 1986, pp. 129–155.
[20] R. Reda, et al., A comparative study between a new method and other machine learning algorithms for soil organic carbon and total nitrogen prediction using near infrared spectroscopy, Chemometr. Intell. Lab. Syst. 195 (Dec. 2019) 103873.
[21] P. Geladi, B.R. Kowalski, Partial least-squares regression: a tutorial, Anal. Chim. Acta 185 (C) (Jan. 1986) 1–17.
[22] J.C.L. Alves, R.J. Poppi, Biodiesel content determination in diesel fuel blends using near infrared (NIR) spectroscopy and support vector machines (SVM), Talanta 104 (2013) 155–161.
[23] J.P.M. Andries, G.H. Tinnevelt, Y. Vander Heyden, Improved modelling for low-correlated multiple responses by common-subset-of-independent-variables partial-least-squares, Talanta 239 (Mar. 2022).
[24] T. Naes, H. Martens, Principal component regression in NIR analysis: viewpoints, background details and selection of components, J. Chemom. 2 (2) (Apr. 1988) 155–167.
[25] Y. Zhang, Q. Cong, Y. Xie, JingxiuYang, B. Zhao, Quantitative analysis of routine chemical constituents in tobacco by near-infrared spectroscopy and support vector machine, Spectrochim. Acta Mol. Biomol. Spectrosc. 71 (4) (Dec. 2008) 1408–1413.
[26] P.R. Filgueiras, L.A. Terra, E.V.R. Castro, L.M.S.L. Oliveira, J.C.M. Dias, R.J. Poppi, Prediction of the distillation temperatures of crude oils using ^{1}H NMR and support vector regression with estimated confidence intervals, Talanta 142 (Sep. 2015) 197–205.
[27] E. Pekel, Estimation of soil moisture using decision tree regression, Theor. Appl. Climatol. 139 (3) (Nov. 2019) 1111–1119.
[28] O. Kramer, K-nearest Neighbors, 2013, pp. 13–23.
[29] J. Ranstam, J.A. Cook, LASSO regression, Br. J. Surg. 105 (10) (Aug. 2018) 1348.
[30] R. Reda, et al., Predicting soil phosphorus and studying the effect of texture on the prediction accuracy using machine learning combined with near-infrared spectroscopy, Spectrochim. Acta Mol. Biomol. Spectrosc. 242 (Dec. 2020) 118736.
[31] N. Aziz, E.A.P. Akhir, I.A. Aziz, J. Jaafar, M.H. Hasan, A.N.C. Abas, A study on gradient boosting algorithms for development of AI monitoring and prediction systems, in: International Conference on Computational Intelligence, ICCI, 2020, pp. 11–16.
[32] M. Luo, et al., Combination of feature selection and CatBoost for prediction: the first application to the estimation of aboveground biomass, Forests 12 (2) (2021) 216.

[33] S. Chen, et al., Monitoring soil organic carbon in alpine soils using in situ vis-NIR spectroscopy and a multilayer perceptron, Land Degrad. Dev. 31 (8) (May 2020) 1026–1038.
[34] G.I. Webb, et al., Leave-one-out cross-validation, in: Encyclopedia of Machine Learning, 2011, pp. 600–601.
[35] M. Fischetti, M. Stringher, Embedded Hyper-Parameter Tuning by Simulated Annealing, Jun. 2019.
[36] N. Pathik, P. Shukla, Simulated annealing based algorithm for tuning LDA hyper parameters, Adv. Intell. Syst. Comput. 1154 (2020) 515–521.
[37] H. Zou, T. Hastie, Regression Shrinkage and Selection via the Elastic Net, with Applications to Microarrays Human Connectome Project for Disordered Emotional States (HCP-DES) View Project Cracking the Speech Code: A Cross-Linguistic Neurobehavioral Approach to Language Learning in Typical and Atypical Populations. View Project Regression Shrinkage and Selection via the Elastic Net, with Applications to Microarrays, 2004.
[38] C. De Mol, E. De Vito, L. Rosasco, Elastic-net regularization in learning theory, J. Complex 25 (2) (Apr. 2009) 201–230.
[39] R. Leardi, A. Lupianez Gonzalez, Genetic algorithms applied to feature selection in PLS regression: how and when to use them, Chemometr. Intell. Lab. Syst. 41 (2) (Jul. 1998) 195–207.
[40] J.P. Nielsen, A. Saudland, L. Norgaard, J. Wagner, S.B. Engelsen, L. Munck, Interval partial least-squares regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy, Appl. Spectrosc. 54 (3) (Mar. 2000) 413–419. Accessed: Dec. 05, 2022. [Online]. Available: https://opg.optica.org/abstract.cfm?uri=as-54-3-413.
[41] J. Muncan, Comparative study on structure and properties of water by infra red and opto-magnetic spectroscopy, Contemp. Mater. 1 (3) (Oct. 2012).
[42] E. Chatani, Y. Tsuchisaka, Y. Masuda, R. Tsenkova, Water molecular system dynamics associated with amyloidogenic nucleation as revealed by real time near infrared spectroscopy and aquaphotomics, PLoS One 9 (7) (Jul. 2014) e101997.
[43] D. Gastaldi, F. Canonico, S. Irico, D. Pellerej, M.C. Paganini, Near-infrared spectroscopy investigation on the hydration degree of a cement paste, J. Mater. Sci. 45 (12) (Jun. 2010) 3169–3174.
[44] H. Martens, M. Hoy, F. Westad, D. Folkenberg, M. Martens, Analysis of designed experiments by stabilised PLS Regression and jack-knifing, Chemometr. Intell. Lab. Syst. 58 (2) (Oct. 2001) 151–170.
[45] J.H. Matthes, S.H. Knox, C. Sturtevant, O. Sonnentag, J. Verfaillie, D. Baldocchi, Predicting landscape-scale CO2 flux at a pasture and rice paddy with long-term hyperspectral canopy reflectance measurements, Biogeosciences 12 (15) (Aug. 2015) 4577–4594.
[46] A. Oussama, F. Elabadi, S. Platikanov, F. Kzaiber, R. Tauler, Detection of olive oil adulteration using FT-IR spectroscopy and PLS with variable importance of projection (VIP) scores, JAOCS (J. Am. Oil Chem. Soc.) 89 (10) (Jun. 2012) 1807–1812.
[47] H. Martens, M. Martens, Modified Jack-knife estimation of parameter uncertainty in bilinear modelling by partial least squares regression (PLSR), Food Qual. Prefer. 11 (1–2) (Jan. 2000) 5–16.
[48] H. Martens, M. Hoy, F. Westad, D. Folkenberg, M. Martens, Analysis of designed experiments by stabilised PLS Regression and jack-knifing, Chemometr. Intell. Lab. Syst. 58 (2001) 151–170.