Systems & Control Transactions

Research Article - Peer Reviewed Conference Proceeding
ESCAPE 35 - European Symposium on Computer Aided Process Engineering
Ghent, Belgium. 6-9 July 2025
Jan F.M. Van Impe, Grégoire Léonard, Satyajeet S. Bhonsale,
Monika E. Polańska, Filip Logist (Eds.)

PSE PRESS

Linear and non-linear convolutional approaches and XAI for spectral data: classification of waste lubricant oils

Ruben Gariso[a], João P. L. Coutinho[a], Tiago J. Rato[a], Marco S. Reis[a,*]

[a] CERES, Department of Chemical Engineering, University of Coimbra, Portugal
* Corresponding Author: marco@eq.uc.pt.

ABSTRACT

Waste lubricant oil (WLO) is a hazardous residual that requires proper management, being WLO regeneration the preferred approach. However, regeneration is only viable if the WLO does not coagulate in the equipment. Otherwise, the process needs to be shut down for cleaning and maintenance. To mitigate this risk, a laboratory test is currently used to assess the WLO coagulation potential before it enters the process. This laboratory test is, however, time-consuming, presents several safety risks, and is subjective. To expedite decision-making, process analytics technology (PAT) and machine learning were used to develop a model to classify WLOs according to their coagulation potential. Three approaches were followed, spanning linear and non-linear models. The first approach (benchmark) uses partial least squares for discriminant analysis (PLS-DA) and interval PLS combined with standard chemometric preprocessing techniques (27 model variants). The second approach uses wavelet transforms to decompose the spectra and PLS-DA for feature selection (10 model variants). Finally, the third approach uses convolutional neural networks (CNN) to estimate the optimal filter for feature extraction (1 model variant). These models were trained on real industrial data. The results show that the three modelling approaches can attain high accuracy, with an average of 91%. The spectral filtering using wavelet transforms proved to be a viable option to reduce the number of models to explore compared to the benchmark approach. The CNN was also able to streamline the preprocessing selection by implicitly estimating the optimal filter. The models also provided physical insight into the WLO coagulation phenomenon.

Keywords: Waste lubricating oil, Multiblock analysis, PLS, CNN, Classification.

INTRODUCTION

Waste lubricant oil (WLO) is a hazardous material with the potential to cause severe environmental harm if not properly managed. Within the European Union's waste management hierarchy, regeneration is identified as the preferred method for addressing this type of residual. In Portugal, the entire WLO regeneration supply chain is managed by Sogilub, with the primary objective of recovering base oil – the main component of lubricant oil – from WLO. However, WLO regeneration is only viable if the WLO does not coagulate during processing. Otherwise, the process must be stopped for cleaning and maintenance, and subsequent disposal of the entire production batch. To mitigate this risk, the coagulation potential of WLO is currently assessed using a laboratory test involving an alkaline treatment with potassium hydroxide (KOH) [1]. While effective, this test is time-intensive, poses safety concerns, due to the formation of foams and use of high temperatures, and relies on a subjective visual interpretation by the analyst. As an alternative, the implementation of a process analytical technology (PAT)-based classification system offers a promising solution to accelerate WLO processing, improve safety, and enhance the analytical capacity of laboratories.

This study aims to evaluate and compare the performance of traditional linear models and modern non-linear models in leveraging mid-infrared (MIR) spectroscopy to classify WLO before regeneration. Three methodologies were considered: (i) partial least squares (PLS) [2] and interval partial least squares (iPLS) [3] combined with chemometric preprocessing techniques; (ii) iPLS

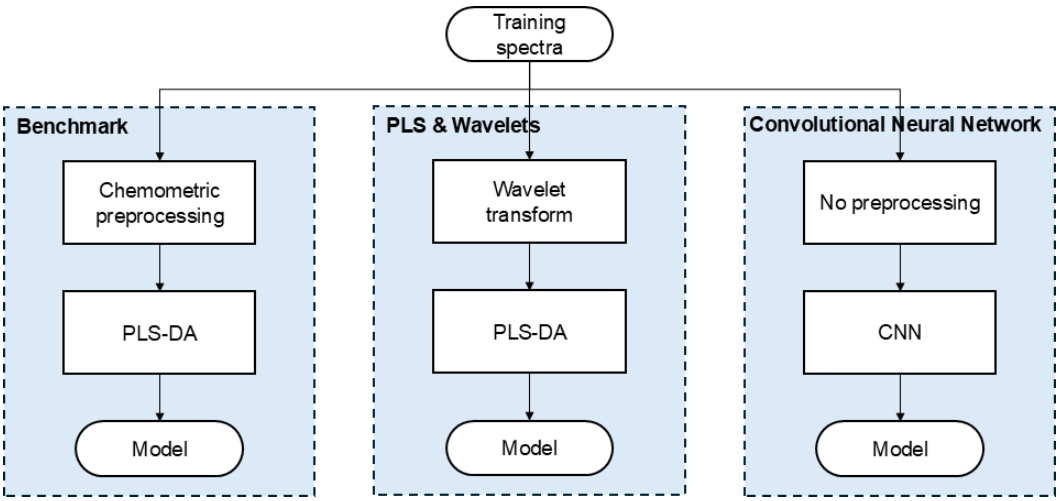

Figure 1 Overview of the models under comparison using combinations of preprocessing and modeling approaches.

combined with wavelet transforms [4]; and (iii) convolutional neural networks (CNNs) [5].

The remainder of this article is structured as follows. The Methodology section encompasses the preprocessing techniques and modeling methodologies employed. In the Results section, an analysis of the results obtained in this study is made. Finally, the Conclusion section offers a concise summary of the conclusions drawn from the findings.

METHODOLOGY

This study aims to develop a suitable data-driven model to predict the coagulation potential of WLO using PAT techniques. To achieve this end, three modeling approaches were considered and systematically compared using the SS-DAC framework [6].

In the following subsections, an overview of the preprocessing techniques and modelling methodologies used for model development are described. A simplified diagram of the models under comparison is presented in Figure 1.

Spectral Preprocessing

Preprocessing of the spectral data is often used to mitigate potential artifacts that may arise due to unintended interactions between light and the sample under examination. Two main classes of preprocessing techniques were considered: (i) classical chemometric approaches; and (ii) wavelet transforms.

Chemometric Preprocessing

From the class of classical preprocessing techniques typically found in the literature (e.g. [7]), the following were considered:

- Standard Normal Variate (SNV) [8];
- Multiplicative Scatter Correction (MSC) [9];
- Savitzky-Golay differentiation (SGD) [10].

Different combinations of these preprocessing techniques, as well as different parameterizations of SGD (identified as SGD-{derivative order}-{window size}) were examined, leading to nine distinct preprocessing variations (see Table 1).

Wavelet Transforms

As an alternative to the classical chemometric preprocessing, wavelet transforms were used to decompose the spectra into multiple frequency components by convolution with (fixed) linear filters. The obtained wavelet coefficients were then fed to the modeling methodologies for variable selection.

The discrete wavelet transform (DWT) can be efficiently performed using the recursive algorithm proposed in [4]. Several wavelet filter can be found in the literature [11–13], being the Haar, Daubechies and Symlets wavelets the most common and those used in this study (see Table 1).

Modeling Methodologies

Two main classes of modeling methodologies were considered in this study: (i) linear; and (ii) non-linear.

From the linear class, partial least squares (PLS) for discriminant analysis (PLS-DA) [14] using the full spectra and two interval-based extensions of PLS, namely forward interval PLS (FiPLS) and backward interval PLS (BiPLS) were used. These models were combined with classical chemometric preprocessing approaches and represent the benchmark. Additionally, wavelet

transforms were used to decompose the spectra by convolution with fixed filters and their wavelet coefficients were then used as features by FiPLS and BiPLS.

From the non-linear class, convolutional neural networks (CNNs) were used. CNNs have recently been considered as a potential alternative to classic chemometric methods for modeling of spectroscopic data [5,15]. The main benefit of CNN models lies in their ability to perform automatic feature extraction of the raw spectra, tailored for each specific dataset and target variables. For better interpretability, we consider only a single convolutional layer with a single filter. After the convolutional layer, the extracted features are passed through a non-linear activation function and multiple hidden layers. Since the problem is a binary classification problem, the output layer has a logistic activation function, and the Binary Cross Entropy loss function is used to estimate the CNN model parameters using the Adam optimizer. We further consider L_2 regularization of the loss function to prevent model overfitting.

Feature Importance

For MIR data, different spectral bands can be associated with specific functional groups. Thus, the analysis of feature importance can provide insights on the critical chemical compounds related with the coagulation phenomenon. For the PLS-based approaches, feature importance was assessed using the Variable Importance in Projection (VIP) method [16,17]. In turn for the CNN model, the feature importance was determined using the Integrated Gradients (IG) method [18]. Whereas the VIP values are calculated based on the entire training dataset, IG values are calculated for each individual spectral sample. In addition, due to the stochastic weight initialization, the IG feature importance was repeated 30 times and the results averaged.

RESULTS

WLO samples were provided by Sogilub, a non-profit organization responsible for the management of WLO in Portugal. A total of 109 WLO samples were collected for analysis. The coagulation potential of each sample was assessed using a coagulation test involving an alkaline treatment with potassium hydroxide (KOH). Based on the outcomes of this test, the laboratory analyst labeled the samples into two categories: "coagulates" (43 samples) or "does-not-coagulate" (66 samples). Additionally, MIR spectra comprising 1814 wavenumbers in the range of 4000 to 500 cm^{-1} with a resolution of 2 cm^{-1} were obtained for each sample in triplicate.

A total of 38 models, encompassing both linear and non-linear models, were compared using the SS-DAC framework. The first class of models, serving as the benchmark, employed PLS-DA, FiPLS and BiPLS, combined with standard chemometric preprocessing techniques, yielding 27 model variants. The second approach used wavelet transforms to decompose the spectral data into multiple frequency components through convolution with linear filters, followed by feature selection using FiPLS and BiPLS, resulting in 10 model variants. Finally, the third approach used CNN (without preprocessing) for non-linear feature extraction and modeling (1 model variant).

For the BiPLS and FiPLS methodologies using the chemometric preprocessing approaches, the MIR spectra were split into 30 equal intervals. For FiPLS and BiPLS using wavelet transforms the spectra were divided in intervals with a dyadic length of 64 cm^{-1} (leading to into 29 intervals).

In the first stage of the SS-DAC framework, the raw dataset was randomly divided into two datasets: a

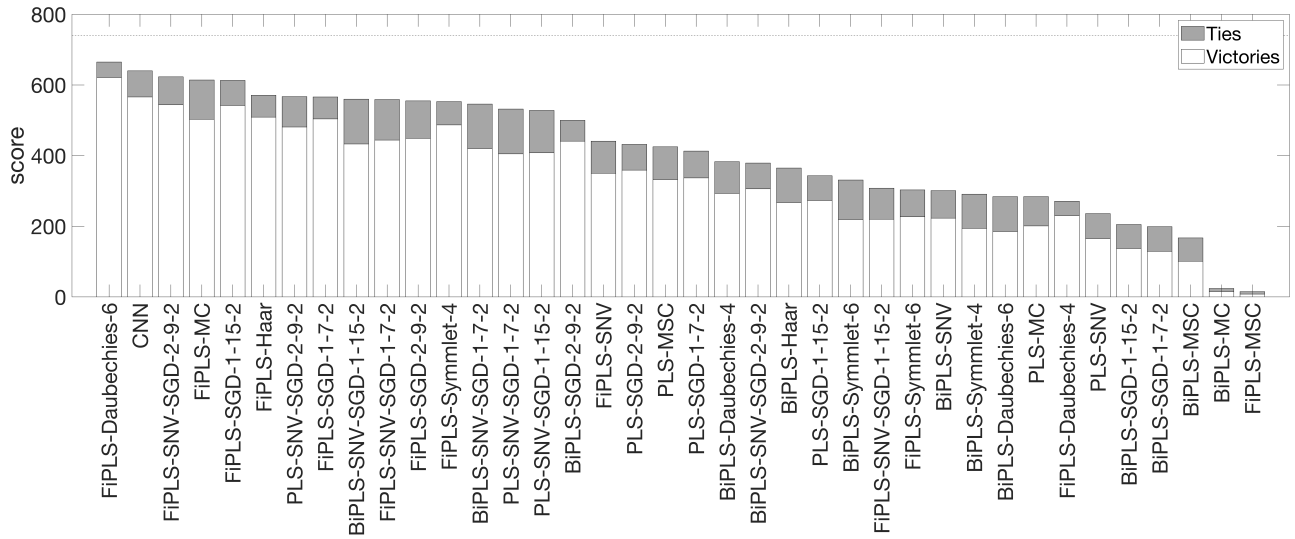

Figure 2 Score of the models' under comparison using the SS-DAC fremework.

Table 1 Mean accuracy for each model. The models with the highest rank in each class are highlighted in bold.

Class	Preprocessing Techniques	Modeling Methodology			
		PLS	FiPLS	BiPLS	CNN
PLS & Chemometric (benchmark)	Mean centering (MC)	0.82	0.88	0.69	-
	Standard normal variate (SNV)	0.84	0.64	0.77	-
	Multiplicative scatter correction (MSC)	0.78	0.85	0.79	-
	Savitzky-Golay differentiation (SGD-{1-7}) *	0.84	0.90	0.78	-
	Savitzky-Golay differentiation (SGD-{1-15}) *	0.84	0.89	0.78	-
	Savitzky-Golay differentiation (SGD-{2-9}) *	0.86	0.88	0.88	-
	SNV & SGD (SNV-SGD-{1-7}) *	0.87	0.88	0.88	-
	SNV & SGD (SNV-SGD-{1-15}) *	0.88	0.82	0.88	-
	SNV & SGD (SNV-SGD-{2-9}) *	0.89	**0.90**	0.85	-
PLS & Wavelet transform	Haar	-	0.90	0.84	-
	Daubechies 4	-	0.80	0.85	-
	Daubechies 6	-	**0.93**	0.83	-
	Symmlet 4	-	0.89	0.83	-
	Symmlet 6	-	0.84	0.83	-
CNN	None	-	-	-	**0.90**

* The short name of the Savitzky-Golay differentiation preprocessing is coded as SGD-{D,W}, where D is the devative order and W is the window size.

training dataset comprising 80% of the samples and a test dataset containing the remaining 20% of the samples, ensuring balanced distributions across both datasets. Additionally, the replicates of the same sample were assigned to the same dataset to maintain consistency. The hyperparameters of the PLS-based models (number of latent variables and intervals kept in the model) were selected using the training dataset and Monte Carlo Cross-Validation (MCCV). In turn, the hyperparameters of the CNN (learning rate, regularization weights, kernel size, number and size of hidden layers, activation function and type of pooling), were selected by Bayesian Optimization (BO) [19] based on stratified 10-fold cross validation. All models were trained by maximization of the accuracy (ACC) given by:

$$ACC = \frac{TP+TN}{n} \quad (1)$$

where, TP is the number of true positives, TN is the number of true negatives, and n is the total number of samples.

Subsequently, in the second stage of SS-DAC, the models' performance was evaluated on the test dataset using the accuracy as the key performance indicator (KPI).

In the third stage of the SS-DAC, the models were compared using bootstrap to obtain multiple realizations of the KPI in test conditions, and the Wilcoxon signed-rank test [20] to statistically assess the significance of the differences in accuracy of every pair of models. For each pairwise comparison, if a model had a statistically significant higher KPI, it was awarded a "victory", while the competing model received a "defeat". In cases where the accuracy of both models was deemed not statistically different, the comparison outcome was recorded as a "tie". A performance score was then calculated for each model by summing up the number of victories and ties. Models achieving higher scores, characterized by a larger number of victories and ties, were considered to exhibit consistently superior accuracy. The model's score, ranked from highest to lowest score are presented in Figure 2. Table 1 provides the mean accuracy obtained by each model.

The results obtained show that PLS with wavelet transforms, namely the FiPLS-Daubechies-6 model, achieved the best performance (highest score; first place), with a mean ACC of 0.93 and a standard deviation of 0.07. The CNN attained the second place, with a mean ACC of 0.92 with a standard deviation of 0.11. In the benchmark approach (PLS & Chemometric preprocessing), the best model was FiPLS-SNV-SGD-{2-9}, which ranked third place with a mean ACC of 0.90 and a standard deviation of 0.09. Thus, all model classes were able to attain high accuracy levels.

To achieve a high accuracy level using PLS and chemometric preprocessing it was however necessary to screen over several models, being that 26% of them have a mean accuracy lower then 0.80. The full-spectrum PLS models exhibited low performance since they rely on the entire spectral range, which makes them more susceptible to spectral noise. Similarly, the BiPLS-based models also showed low performance since they retained many intervals, leading to a less focused feature selection.

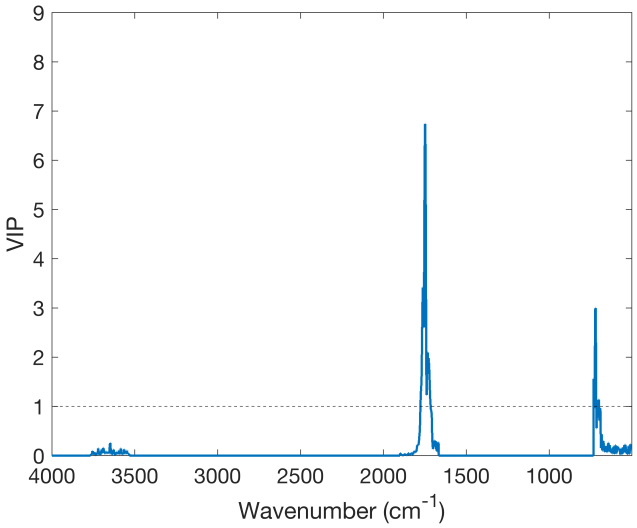

Figure 3 Feature importance for FiPLS-SNV-SGD-{2-9}.

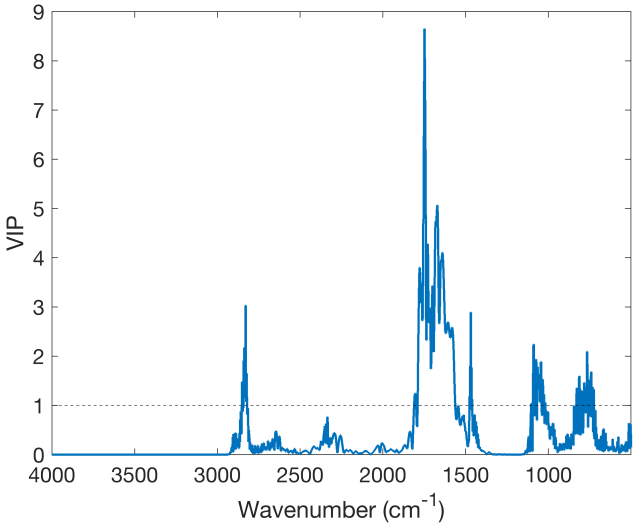

Figure 4 Feature importance for FiPLS-Daubechies-6.

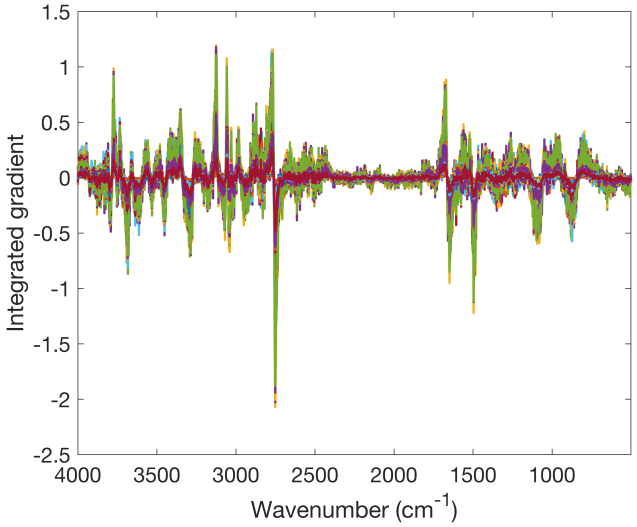

Figure 5 Feature importance for CNN, calculated for each spectral sample. Refer to the online version for color.

In turn, all models based on PLS with wavelet transforms and CNN have a mean accuracy greater than 0.80. Thus, they are deemed as better modelling alternatives. The CNN model shows good performance even without any preliminary preprocessing of the spectra, which highlights their ability to perform automatic feature extraction. The BO of the CNN hyperparameters resulted in a large kernel, with a size of 184 and a larger structure, with 3 hidden layers with 82 units each. This high number of parameters can justify the slightly larger standard deviation of the CNN model compared to the more parsimonious PLS models.

In terms of feature importance, the FiPLS-SNV-SGD-{2-9} model distinctly identified the spectral band between 1775 and 1700 cm^{-1} as the most critical for classification (Figure 3). Similarly, the FiPLS-Daubechies-6 model highlighted a broader region, identifying the 1790–1574 cm^{-1} band as relevant (Figure 4). In contrast, other FiPLS models, whether using chemometric or wavelet preprocessing, included additional bands beyond this range, which negatively impacted their performance. Unlike the interval PLS models, the identification of the relevant features is less clear for the CNN model, due to the lack of explicit feature selection. Contrary to the other models, the band around 2700 cm^{-1} is identified as the most important, which is related to the C-O stretch, possibly indicating a non-linear relationship with this band and the coagulation class.

Among the spectral bands consistently selected by most models, the 1750–1735 cm^{-1} band stands out. This region is associated with the ester functional group [21]. Based on this, it is conjectured that esters may have a significant effect on the WLO coagulation phenomenon.

As the reference laboratory test is inherently subjective, there is some uncertainty in the WLO labels. Nevertheless, as the large majority of the samples are well discriminated, we believe this label uncertainty does not have a significant impact on the models' overall performance.

CONCLUSION

The modeling methodologies employed in this study produced models with high predictive accuracy, with a mean ACC ranging from 0.90 to 0.93. The performance of PLS-based models was significantly affected by the choice of preprocessing techniques, emphasizing the importance of evaluating multiple preprocessing strategies during model development.

The use of wavelet transforms to preprocess the spectrum proved to be competitive against the traditional chemometric preprocessing approaches and can reduce the search space during preprocessing selection. The CNN exhibited good performance even without preprocessing. This advantage can reduce the effort typically required by exhaustive preprocessing selection by trial

and error. However, the CNNs requires the optimization of hyperparameters, for which BO is a compelling approach within the realm of automated machine learning.

The top-performing models allowed for the identification of esters as a potential factor in WLO coagulation, offering critical insights into the underlying mechanisms of the coagulation phenomenon.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the financial support of SOGILUB – Sociedade de Gestão Integrada de Óleos Lubrificantes Usados, Lda. The authors also acknowledge the support from the Chemical Engineering and Renewable Resources for Sustainability (CERES), which is financed by national funds from FCT/MCTES (with references DOI:10.54499/UIDB/00102/2020 and DOI:10.54499/UIDP/00102/2020).

REFERENCES

1. Pinheiro CT, Ascensão VR, Reis MS, Quina MJ, Gando-Ferreira LM. A data-driven approach for the study of coagulation phenomena in waste lubricant oils and its relevance in alkaline regeneration treatments. Sci Total Environ 599–600:2054–2064 (2017).
2. Wold S, Sjöström M, Eriksson L. PLS-Regression: A Basic Tool of Chemometrics. Chemom Intell Lab Syst 58:109–130 (2001).
3. Nørgaard L, Saudland A, Wagner J, Nielsen JP, Munck L, Engelsen SB. Interval Partial Least-Squares Regression (iPLS): A Comparative Chemometric Study with an Example from Near-Infrared Spectroscopy. Appl Spectrosc 54:413–419 (2000).
4. Mallat SG. A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans Pattern Anal Mach Intell 11:674–693 (1989).
5. Yang J, Xu J, Zhang X, Wu C, Lin T, Ying Y. Deep learning for vibrational spectral analysis: Recent progress and a practical guide. Anal Chim Acta 1081:6–17 (2019).
6. Rato TJ, Reis MS. SS-DAC: A systematic framework for selecting the best modeling approach and preprocessing for spectroscopic data. Comput Chem Eng 128:437–449 (2019).
7. Naes T, Isaksson T, Fearn T, Davies T. A User-Friendly Guide to Multivariate Calibration and Classification. NIR Publications (2002).
8. Barnes RJ, Dhanoa MS, Lister SJ. Standard Normal Variate Transformation and De-trending of Near-Infrared Diffuse Reflectance Spectra. Appl Spectrosc 43:772–777 (1989).
9. Geladi P, MacDougall D, Martens H. Linearization and Scatter-Correction for Near-Infrared Reflectance Spectra of Meat. Appl Spectrosc 39:491–500 (1985).
10. Savitzky A, Golay MJE. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. Anal Chem 36:1627–1639 (1964).
11. Mallat S. A Wavelet Tour of Signal Processing. Academic Press (2009).
12. Daubechies I. Orthonormal bases of compactly supported wavelets. Commun Pur Appl Math 41:909–996 (1988).
13. Press WH, Teukolsky SA, Vetterling WT, Flannery BP. Numerical Recipes: The Art of Scientific Computing. Cambridge University Press (2007).
14. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer (2009).
15. Cui C, Fearn T. Modern practical convolutional neural networks for multivariate regression: Applications to NIR calibration. Chemom Intell Lab Syst 182:9–20 (2018).
16. Wold S, Johansson E, Cocchi M. PLS: Partial Least Squares Projections to Latent Structures. In: 3D QSAR in Drug Design: Theory, Methods and Applications. Ed. Kubinyi K. 523–550 ESCOM Science Publisher (1993).
17. Eriksson L, Johansson E, Kettaneh-Wold N, Trygg J, Wikström C, Wold S. Multi- and Megavariate Data Analysis: Parte I - Basic Principles and Applications. Umetrics Academy (2006).
18. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. In: Proceedings of the 34th international conference on machine learning. Eds. Doina P, Yee Whye T. 70:3319–3328 PMLR (2017).
19. Shahriari B, Swersky K, Wang Z, Adams RP, de Freitas N. Taking the Human Out of the Loop: A Review of Bayesian Optimization. Proc IEEE 104:148–175 (2016).
20. Wilcoxon F. Individual Comparisons by Ranking Methods. Biometrics Bull 1:80–83 (1945).
21. Larkin P. Infrared and Raman Spectroscopy: Principles and Spectral Interpretation. Elsevier (2017).