A short note on achieving similar performance to deep learning with practical chemometrics

Puneet Mishra, Jean-Michel Roger, Douglas N Rutledge

Chemometrics and Intelligent Laboratory Systems 214 (2021) 104336

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics

Short communication

A short note on achieving similar performance to deep learning with practical chemometrics

Puneet Mishra [a,*], Jean-Michel Roger [b,c], Douglas N. Rutledge [d,e]

[a] Wageningen Food and Biobased Research, Bornse Weilanden 9, P.O. Box 17, 6700AA, Wageningen, the Netherlands
[b] ITAP, INRAE, Institut Agro, University Montpellier, Montpellier, France
[c] ChemHouse Research Group, Montpellier, France
[d] Université Paris-Saclay, INRAE, AgroParisTech, UMR SayFood, 75005, Paris, France
[e] National Wine and Grape Industry Centre, Charles Sturt University, Wagga Wagga, Australia

ARTICLE INFO

Keywords:
Spectroscopy
Chemometrics
Pre-processing fusion
Ensemble learning

ABSTRACT

Deep learning (DL) is appearing as a powerful tool for spectral data modelling. In many cases, the DL models have been shown to outperform the classical chemometric spectral data modelling approaches. However, a main challenge with the DL models is the limited model interpretability as there are no scores and loadings generated in DL models. Scores and loadings are key parts of chemometric modelling as they facilitate the interpretation of the models. Furthermore, most of the time in the reported literature, the performance of DL models is compared with basic chemometric approaches with less attention being paid to optimizing these chemometric models. This study aims to test the hypothesis that proper chemometric modelling of spectral data can lead to performance equivalent to that of DL models while having all the useful information such as scores, loadings, and regression coefficients to support model interpretation. To test this, a case study is presented for the prediction of nitrogen content in rapeseed (*Brassica napus* L.) by Vis-NIR spectroscopy. On the classical chemometric side, two recently developed pre-processing fusion approaches, i.e. sequential pre-processing through orthogonalization (SPORT) and parallel pre-processing through orthogonalization (PORTO) were used. On the DL side, the previously published with DL modelling results for the same data set were used as the benchmark. Such a comparison was valid as the chemometric analysis was performed on the same calibration and test sets as used previously for the DL modelling. Results showed that the sequential and parallel learning approaches attained the same accuracy as that of the previously reported DL procedure on the same data set. The information related to scores, loading and regression coefficients could be used for model interpretation.

1. Introduction

Deep learning (DL) is starting to appear as a powerful tool for spectral data modelling [1–3]. Currently, two DL approaches exist for spectral predictive modelling; the first is the supervised approach using convolutional neural networks (CNNs) with fully connected layers [2–5] to model the property of interest, while the second approach uses unsupervised feature extraction, such as autoencoders, which are then used to calibrate machine learning models [6–8]. In many cases, DL models have outperformed classical chemometric spectral data modelling approaches such as partial least-square (PLS) regression [6–8]. However, a main problem with DL models is the limited model interpretability since scores and loadings are not generated [2]. Scores and loadings are key parts of chemometric modelling as they facilitate the interpretation of the models [9,10]. Furthermore, most of the time in the reported literature, the performance of DL models has been compared with basic chemometric approaches with less attention being paid to optimizing the chemometric models [6–8].

Visible and near-infrared (Vis-NIR) spectroscopy is widely used in remote sensing of agricultural systems [11,12]. Applications can be found ranging from close-range greenhouse and field spectroscopy [13] to high-end satellite imaging [14,15]. The main reason for the use of Vis-NIR spectroscopy is the possibility to extract information on the physicochemical properties of plants in a non-destructive way [16]. The interaction of electromagnetic radiation with plants, and particularly with leaves, varies with the wavelength [17], resulting in two main phenomena, i.e. light scattering, and absorption [11]. Due to the scattering and absorption phenomena, the measured Vis-NIR signals of leaves

* Corresponding author.
E-mail address: puneet.mishra@wur.nl (P. Mishra).

https://doi.org/10.1016/j.chemolab.2021.104336
Received 4 February 2021; Received in revised form 26 March 2021; Accepted 4 May 2021
Available online 8 May 2021

often have a mixture of scattering as additive and multiplicative effects combined with the broad absorption peaks related to overtones of functional group vibrations [11,18]. To develop models to predict chemical properties, it is often recommended to reduce or remove the light scattering effects from the signals so that the models can be focussed on the absorption information [13,18,19]. In the field of chemometrics, several scatter correction techniques are available and often implemented prior to regression modelling [20]. However, focussing the modelling solely on the scatter-corrected data in the case of leaves may not be an efficient approach since the overall spectral signal of leaves is affected by the internal structure [11]. This means that there may exist extra correlations in the scattering information which may be useful to explain the property of interest. Hence, a synergistic modelling of the raw reflectance spectra (with scattering) and the scatter-corrected reflectance spectra can allow the full use of all the information present in the data, i.e. both the scattering and the absorption features.

A DL model does not make a clear distinction between absorption and scattering features, it will feed the data into a non-linear model to learn the patterns automatically. On the other hand, for the chemometric modelling, it is important to optimize the model with respect to the absorption and scatter information present in the spectra. This indicates that the comparison previously performed between DL modelling and raw reflectance-based PLS regression modelling for leaf properties prediction [6,8] may have been sub-optimal and a proper chemometric modelling approach could lead to the same performance as the DL models.

This study aims to test the hypothesis that proper chemometric modelling of spectral data can lead to performances equivalent to those of DL models while retaining all the useful information such as scores, loadings, and regression coefficients that facilitate model interpretation. To test this, a case study is presented for the prediction of nitrogen content in rapeseed (*Brassica napus* L.*)* by Vis-NIR spectroscopy. On the classical chemometric side, two recently developed pre-processing fusion approaches were used, i.e., sequential pre-processing through orthogonalization (SPORT) [21] and parallel pre-processing through orthogonalization (PORTO) [22]. On the DL side, the DL modelling results previously published on the same calibration and test data sets were used as a benchmark [8].

2. Materials and methods

2.1. Data set

The data set used in this study was first presented in Ref. [8] and consisted of 192 mean Vis-NIR (380–1030 nm) hyperspectral signals of rapeseeds (*Brassica napus, Zheyou51*). According to the original article, the plants were grown in an experimental field located in Ningbo, Zhejiang, China (30.32°N, 121.18°E) and were given four nitrogen level treatments including no nitrogen, or 100, 200 and 300 kg/ha, using urea. Eight experimental plots of 2.0 m × 1.5 m were used, with two plots for each nitrogen level. Leaf sampling was carried out before the ripeness stage on 12 February, 9 March, 22 March, 8 April, and 26 April 2016. The collected leaves were then used for hyperspectral data acquisition and reference chemical analysis of nitrogen concentration within 4 h. The hyperspectral data acquisition was performed with a Vis-NIR imaging system consisting of a spectrograph (380–1030 nm) and a CCD camera. Line scan data acquisition was performed at a constant speed of 4.5 mm s^{-1} and an exposure time of 16 ms. After hyperspectral data acquisition, the leaves were oven-dried at 105 °C for half an hour, and then ground into a powder which was used to measure nitrogen concentration by the Kjeldahl method.

In this study, the calculated mean spectra and the reference N values were used for the data modelling. The mean spectra were supplied as the supplementary material to the article [8]. Out of the 192 samples, 128 samples were used for model calibration and tuning, and 64 for the independent test set. The calibration and the testing sets were the same as those used in the original article so that the results of the present study will be directly comparable to those obtained previously with several machine learning and deep learning techniques [8]. A further description of the calibration and test set is provided in Table 1.

2.2. Pre-processing fusion modelling techniques

Ensemble approaches to spectral data modelling are gaining popularity in chemometrics [23]. A particular use of ensemble methods is to process the same data with different pre-treatments and learn complementary information [20]. Two approaches that are currently available are SPORT [21] and PORTO [22]. The SPORT approach is based on sequential orthogonalized PLS regression and PORTO is based on parallel orthogonalized PLS regression. For an intuitive understanding, a schematic explaining the concept behind SPORT and PORTO modelling is shown in Fig. 1. In the case of leaves, the SPORT approach can be used to sequentially learn unique complementary subspaces from the raw reflectance and the scatter-corrected reflectance spectra. The PORTO approach for leaf modelling can be implemented to extract the common and unique subspaces of the reflectance and scatter-corrected reflectance spectra that explain the properties of interest. For more information on the SPORT and PORTO approaches, readers are referred to Refs. [21,22]. In the present work, three commonly used scatter correction techniques were selected. One technique requires estimation of external parameters i.e., weights for each wavelengths (variable sorting for normalization [24]) and two are model-free (standard normal variate [25] and 2nd derivative [26]). After processing the data with the three pre-processing techniques, a total of four data blocks were available for SPORT and PORTO analysis, one raw reflectance and three pre-processed data blocks.

All the pre-processing methods and SPORT modelling were implemented in MATLAB 2018b (The Mathworks, Natick, MA, USA) using the MBA-GUI [27]. PORTO was implemented in MATLAB 2018b using the multi-block data analysis codes from Nofima (https://nofima.no/en/) for the implementation of the parallel orthogonalized partial least-squares. All the models were evaluated based on the coefficient of determination (R^2_P) and the root mean square error of prediction (RMSEP).

3. Results

3.1. Spectral characteristics

The spectral characteristics for rapeseed leaves are shown in Fig. 2. The calibration and test spectra are shown in solid blue lines and dotted red lines, respectively. The spectral range of 400–670 nm is the visible part of the spectrum and is related to pigments such as chlorophyll, carotenoids, and anthocyanins. The sharp rise in the reflectance between 670 nm and 750 nm is the red-edge part, after which the cellular structure of the leaves reflects most of the light, which avoids overheating of the plants. The region after 750 nm is related to the overtones of the fundamental vibrations of chemical bonds such as OH, NH and CH. In leaves, these bonds are present in molecules such as water, sugar, protein, and other minor biochemical components. Typically, chlorophyll and several other chemical components containing nitrogen, such as amino acids, are correlated with the nitrogen content in the leaves. Vis-NIR spectroscopy does not supply a direct quantification of the physicochemical components but requires a prior calibration modelling to be

Table 1
A summary of calibration and test sets used both here and in the original study [8].

Data	Spectral range (nm)	Spectra (Samples × wavelengths)	Nitrogen content (%) (mean ± std)
Calibration	380–1020	128 × 512	4.62 ± 1.03
Test	380–1020	64 × 512	4.63 ± 0.99

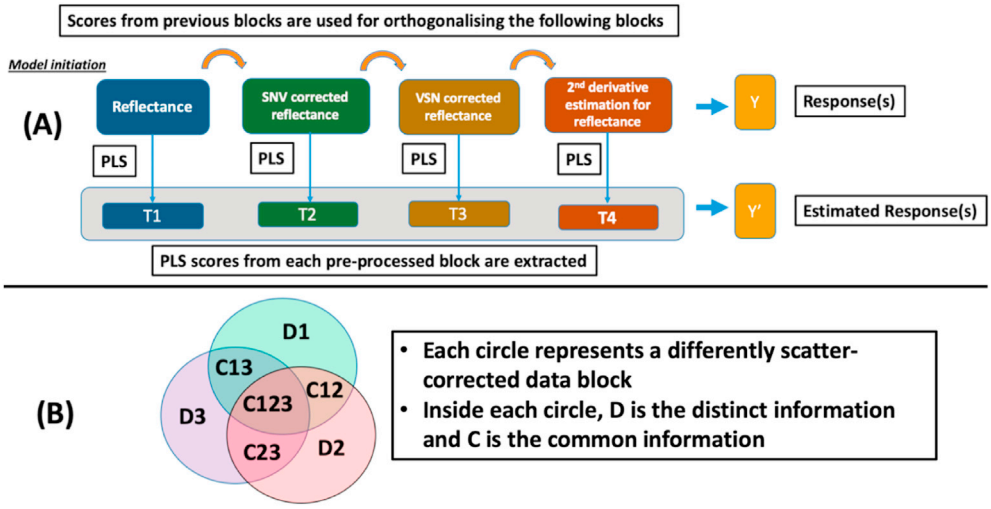

Fig. 1. Framework of sequential (A) and parallel (B) preprocessing fusion. The sequential approach aims to extract complementary information one by one from each differently scatter-corrected data block. The parallel approach aims at finding the common and the distinct information within differently scatter-corrected data blocks.

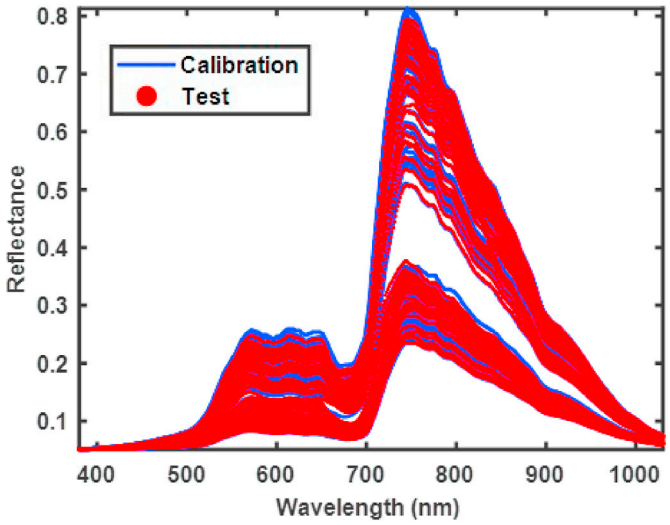

Fig. 2. Spectral characteristics of rapeseed for calibration (blue) and test set (red). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

able to predict the components in future samples. The spectra have differences in global intensities, caused by the light scattering. In chemometrics, such differences in global intensities are classed as additive and multiplicative effects [20,28,29] and dedicated pre-processing methods try to reduce/remove these global differences so as to focalise the data modelling on absorption peaks.

3.2. Sequential and parallel information learning for nitrogen prediction

The results of the sequential and parallel learning approaches are shown in Fig. 3. The sequential approach, i.e., SPORT, found complementary information in the raw reflectance and the 2nd derivative-corrected reflectance. The SPORT modelled 4 latent variables from the raw reflectance and 3 latent variables from the 2nd derivative-corrected reflectance (Fig. 3A). Both SNV and VSN found zero latent variables, hence did not contribute to the final model. This shows that the removal of information from the spectra by normalization is detrimental to the prediction of nitrogen content. In the case of PORTO, the optimal models were obtained by modelling 3 common latent variables (common between raw reflectance and the different scatter correction techniques) and 1 distinct component for raw reflectance. The choice of a distinct component for raw reflectance data shows that the raw reflectance carries unique information which is not present in the data after scatter correction. This unique information can be none other than the scattering

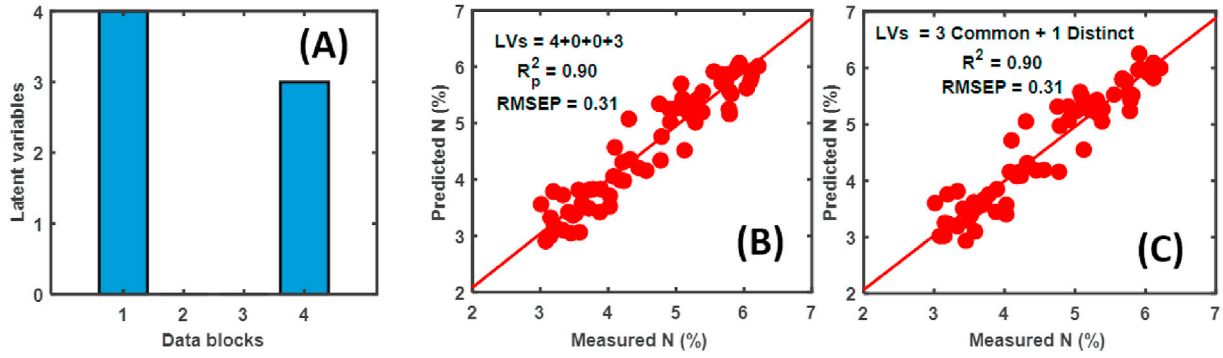

Fig. 3. Summary of the sequential pre-processing through orthogonalization (SPORT) and the parallel pre-processing through orthogonalization (PORTO) models. (A) Number of latent vectors extracted from raw reflectance (1st data block), standard normal variate pre-processed data (2nd data block), variable sorting for normalization pre-processed data (3rd data block) and 2nd derivative pre-processed data (4th data block). In total, SPORT modelled 4 latent variables from raw reflectance and 3 from the 2nd derivative. (B) SPORT predictions, and (C) PORTO predictions.

profiles of the leaves which would appear to be correlated in some way to the nitrogen content. With both the SPORT and PORTO approaches, the prediction R^2_P was 0.90 and the RMSEP was 0.31%. These results are better than those obtained in the original study on the same calibration and test sets by PLS regression (prediction R^2_P and RMSEP were 0.85 and 0.38%) and by LS-SVM regression (prediction R^2_P and RMSEP of 0.88 and 0.35%). As well, they are equivalent to those obtained in that article by the Deep Learning approach (R^2_P and RMSEP of 0.90 and 0.31%). This is of particular interest since the SPORT and PORTO approaches are simpler, using linear algebraic operations, thus saving time and resources for training the deep learning models. The other main benefit of the SPORT and PORTO approaches compared to the deep learning model (based on stacked autoencoders and a fully connected neural network) is that these techniques give all the usual multivariate data analysis outputs, such as loadings, scores and regression vectors which facilitate the spectrochemical interpretation of the models as shown in the following section.

3.3. An example of complementary learning from raw reflectance and scatter corrected reflectance

The optimal model with the SPORT approach was obtained by modelling 4 latent variables from the raw reflectance and 3 latent variables from the 2nd derivative corrected reflectance. The main point to be noted is that SPORT being a sequential approach, it first modelled the raw reflectance and afterwards explored the scatter-corrected data to search for any extra information correlated to the nitrogen content in the rapeseed leaves. Fig. 4A shows the regression coefficients of the raw reflectance block, superimposed on the mean spectrum of the block. The same peaks as in the PLS regression analysis performed in Ref. [8] have positive correlations (530, 700, 750 nm) and negative correlations (670, 720, 790 and 840 nm). This is as was expected since the first step of SPORT is just the PLS regression on the raw reflectance data. In addition to the peaks reported in this publication, it should be noted that the peak at 530 nm is characteristic of the green color, which in reflectance is closely related to the nitrogen content [30]. Note also that there is a trough at 661 nm, a peak at 695 nm and again a trough at 716 nm, which reflect two phenomena: the intensity of chlorophyll absorption at 680 nm and the shift of the red-edge, which is around 715 nm [31]. Fig. 4B shows the regression coefficients for the second derivative block, superimposed on the mean spectrum of the block. The raw spectra show waves (Fig. 2), certainly due to an acquisition problem. Therefore, the derivative spectra present small peaks that are not related to leaf chemistry. However, the largest peaks remain interpretable. Thus one can notice in the b-coefficients a succession of a positive peak near 680 nm, a negative peak near 705 nm and a positive peak near 720 nm. This profile shows that the regression is sensitive to the intensity of the positive peak of the second derivative at 700 nm. This peak is related to the curvature of the foot of the red-edge. This curvature is caused by the presence of the chlorophyll trough at 680 nm and the red-edge, whose position varies between 710 and 720 nm [31], which also varies with the chlorophyll content, and therefore nitrogen, as seen previously. It should be noted that here the SPORT method has allowed the extraction of complementary information from the two blocks: the chlorophyll absorption and the position of the red-edge in the raw reflectance block, and the intensity of the red edge in the second derivative block. There is no peak centered at 680 nm in the coefficients related to the 2nd derivative block, which would have been the case if the information extracted from this block was related to the chlorophyll absorption.

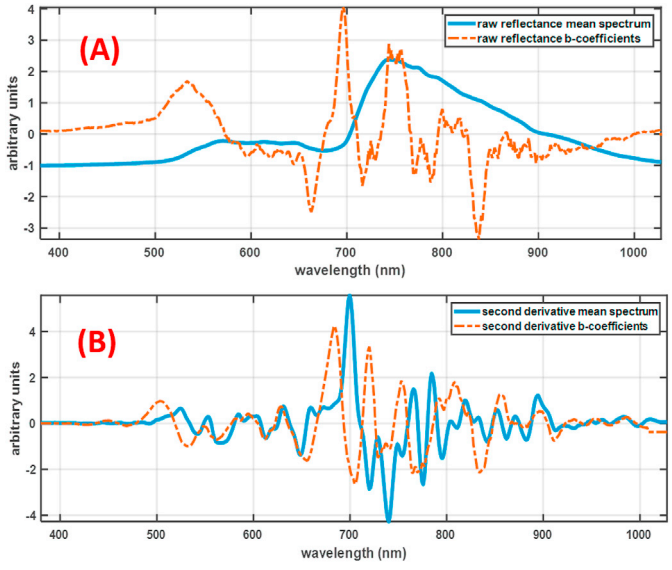

Fig. 4. The complementary B-coefficients vectors obtained by sequential preprocessing through orthogonalization, superimposed on the mean spectrum, from raw reflectance (top) and 2nd derivative pre-processed reflectance (bottom).

4. Discussion

In this study, linear methods such as SPORT and PORTO achieved predictive performance like that of the non-linear DL modelling performed in Ref. [8]. The underlying assumption behind spectral modelling is Beer's law [32] which considers that the amount of light absorbed is directly proportion to the concentration of the analyte [33]. Hence, according to Beer's law [32] a pure absorbance signal measured on the sample can be related to the concentration of the corresponding analyte. However, in a practical scenario this is unachievable as the signal measured by the spectrometer consists of a mixture of light scattering and absorption bringing non-linearities into the signal [20]. In the case of many classical chemometric procedures, the linearity can be recovered by reducing/removing the scattering information by spectral pre-processing and retaining only the absorption information which is linearly related to the concentration of the analyte [20]. However, in some cases both the scattering and absorption information are related to the property of interest, as for example in the demonstrated case of leaf nitrogen measurement, where both the scattering due to the internal structure of leaves and the absorption due to N-containing chemical components are related to the total N content in the leaves. It is of no surprise that the DL modelling carried out in Ref. [8] was able to deal with the non-linearity caused by mixed scattering and absorption information while a simple PLS (without any pre-processing) applied to the raw data performed poorly [8]. In more general terms, in the case presented here of leaf property modelling, the underlying mathematical relationship (linear or non-linear) between input and output information was unclear, hence, the DL was able to take advantage of its flexibility to adjust a model according to the data, resulting in a better performance than the PLS modelling on raw data alone. The case would have been different if the relation between input and output had been linear, since in that case, the DL modelling would have performed similarly to the PLS model. However, as demonstrated in this study, the same linear PLS when complemented by pretreatment procedures such as SPORT and PORTO can achieve the same performance as the DL presented in Ref. [8]. The main contribution of SPORT and the PORTO in this study was to allow the PLS to learn the specific information present in the raw as well as differently pre-processed data, which otherwise was not possible with a simple PLS.

In the chemometric modelling performed in this study as well the PLS modelling in the previously reported study [8], only 192 mean Vis-NIR spectra and reference properties were used. However, the DL modelling performed in Ref. [8], utilised at first 800 spectra per leaf i.e., $192 \times 800 = 153,600$ spectra, to train the primary autoencoder for feature extraction. Subsequently, the autoencoder was used to extract the features from the 192 mean spectra and to train the neural network model. It

cannot be said how the PLS reported in Ref. [8] would have performed if it had been trained on 153,600 spectra instead of 192 mean spectra; however, the PLS reported in Ref. [8] had comparable performance to the DL model while being trained on a data set 800 times smaller. Furthermore, in this study, the SPORT and PORTO model trained on only 192 mean spectra performed similarly to the DL model trained on 153,600 spectra. In DL modelling, the data size plays a key role in extracting complex mathematical relationships and several recent studies have shown that for spectral modelling can indeed outperform traditional chemometric approaches such as PLS [34] and even advanced approaches such as SPORT [35]. This is because, in the presence of a large amount of data, the DL model not only learns the relation between the input and output but also learns the hidden trends and patterns in the data which can make the model more robust and precise.

In this study, a key focus of the advantages of classical chemometric modelling over the DL modelling presented in Ref. [8] was the interpretability of the models. For example, with the SPORT modelling, all the scores, loading and regression vectors were available to facilitate the interpretability of the PLS model, while in Ref. [8] model interpretability was not discussed, and the main aim was to achieve a low RMSEP. To some extent, DL models can also be interpreted but the interpretation provided by the chemometric models such as PLS is incomparable with that of DL models. For DL model interpretation and in particular the visualisation of spectral CNN models, several authors have recently tried to visualise the activation weights of the CNN layers [2,34,35] which can give an insight into the important spectral regions as they have higher weights than the unimportant regions after the passing through the CNN layers [2,34,35]. Furthermore, with the recent advances in the DL approaches such as grad-CAM [36] and attention layers [37], it is expected that the interpretability of the DL models will further improve with time.

One key point to note is that both the SPORT and PORTO modelling, as used in this study, required optimisation of the total number of latent variables. For SPORT, the number of LVs was optimised for each block of data, while for PORTO, the total number of both common and shared LVs was optimised. Hence, based on the 4 data blocks used, the SPORT required optimisation of 4 parameters i.e., LVs corresponding to 4 data blocks, while for the PORTO only 2 main parameters i.e., common, and shared LVs were optimised. However, in the previous DL modelling [8], the total number of parameters was far greater compared to both SPORT and PORTO. Such a huge number of parameters in DL modelling carries an elevated risk of overfitting and requires extensive hyperparameter optimisation to achieve optimal models [34]. Hence, a key benefit of these classical chemometric approaches, apart from achieving similar predictive performance to DL models, is the low number of parameters to be optimised, which can usually be done in a fraction of the time and limits the risk of overfitting.

5. Conclusions

The calibration modelling of Vis-NIR spectra of leaves has for a long time been limited to either the use of traditional machine learning approaches or to PLS regression approaches applied to either the raw or scatter-corrected data. Recently deep learning approaches have been proposed with improved performances compared to the standard chemometric approaches. However, those DL models were compared with a basic chemometric approach such as PLS regression on raw data. In this study, for the first time two new chemometric approaches, SPORT and PORTO, for synergistic sequential and parallel learning from raw reflectance and scattering-corrected reflectance are presented. The application of these two approaches to rapeseed leaves for nitrogen prediction showed that they both reached the same accuracy as that obtained with the DL modelling on the same data set. Furthermore, the SPORT and PORTO scores, loadings and regression vectors could be interpreted to make a link with the plant biochemistry. As well, both the PORTO and SPORT approaches are based on simple linear algebraic operations, and hence are faster and require less computational resources. A key point to note is that the data set used in this study has only 192 mean spectra to calibrate and test the model, while the earlier study based on DL modelling utilised 153,600 spectra. Hence, a conclusion to be drawn from this is that both chemometric approaches i.e., SPORT and PORTO, led to similar performance as that of the DL model, despite being trained on a data set 800 times smaller.

Author statement

Puneet Mishra: Conceptualization; Methodology; Software; Writing – original draft; Data curation, Jean-Michel Roger: Conceptualization; Methodology; Writing – review & editing, Douglas N. Rutledge: Conceptualization; Methodology; Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] A.B. Risum, R. Bro, Using deep learning to evaluate peaks in chromatographic data, Talanta 204 (2019) 255–260.
[2] C. Cui, T. Fearn, Modern practical convolutional neural networks for multivariate regression: applications to NIR calibration, Chemometr. Intell. Lab. Syst. 182 (2018) 9–20.
[3] S. Malek, F. Melgani, Y. Bazi, One-dimensional convolutional neural networks for spectroscopic signal regression, J. Chemometr. 32 (2018) e2977.
[4] W. Ng, B. Minasny, M. Montazerolghaem, J. Padarian, R. Ferguson, S. Bailey, A.B. McBratney, Convolutional neural network for simultaneous prediction of several soil properties using visible/near-infrared, mid-infrared, and their combined spectra, Geoderma 352 (2019) 251–267.
[5] E.J. Bjerrum, M. Glahder, T. Skov, Data Augmentation of Spectral Data for Convolutional Neural Network (CNN) Based Deep Chemometrics, arXiv preprint arXiv:1710.01927, 2017.
[6] Z. Xin, S. Jun, T. Yan, C. Quansheng, W. Xiaohong, H. Yingying, A deep learning based regression method on hyperspectral data for rapid prediction of cadmium residue in lettuce leaves, Chemometr. Intell. Lab. Syst. 200 (2020) 103996.
[7] X.J. Yu, H.D. Lu, D. Wu, Development of deep learning method for predicting firmness and soluble solid content of postharvest Korla fragrant pear using Vis/NIR hyperspectral reflectance imaging, Postharvest Biol. Technol. 141 (2018) 39–49.
[8] X. Yu, H. Lu, Q. Liu, Deep-learning-based regression model and hyperspectral imaging for rapid detection of nitrogen concentration in oilseed rape (Brassica napus L.) leaf, Chemometr. Intell. Lab. Syst. 172 (2018) 188–193.
[9] P. Geladi, Chemometrics in spectroscopy. Part 1. Classical chemometrics, Spectrochim. Acta B Atom Spectrosc. 58 (2003) 767–782.
[10] S. Wold, C. Albano, W.J. Dunn, U. Edlund, K. Esbensen, P. Geladi, S. Hellberg, E. Johansson, W. Lindberg, M. Sjöström, Multivariate data analysis in chemistry, in: B.R. Kowalski (Ed.), Chemometrics: Mathematics and Statistics in Chemistry, Springer Netherlands, Dordrecht, 1984, pp. 17–95.
[11] S. Jacquemoud, F. Baret, PROSPECT: a model of leaf optical properties spectra, Rem. Sens. Environ. 34 (1990) 75–91.
[12] S. Jay, R. Bendoula, X. Hadoux, J.B. Feret, N. Gorretta, A physically-based model for retrieving foliar biochemistry and leaf orientation using close-range imaging spectroscopy, Rem. Sens. Environ. 177 (2016) 220–236.
[13] M.S.M. Asaari, P. Mishra, S. Mertens, S. Dhondt, D. Inze, N. Wuyts, P. Scheunders, Close-range hyperspectral image analysis for the early detection of stress responses in individual plants in a high-throughput phenotyping platform, ISPRS J. Photogrammetry Remote Sens. 138 (2018) 121–138.
[14] C. Jia, P.J. Minnett, High latitude sea surface temperatures derived from MODIS infrared measurements, Rem. Sens. Environ. 251 (2020) 112094.
[15] X. Chen, T.A. Warner, D.J. Campagna, Integrating visible, near-infrared and short-wave infrared hyperspectral and multispectral thermal imagery for geological mapping at Cuprite, Nevada, Rem. Sens. Environ. 110 (2007) 344–356.
[16] P. Mishra, S. Lohumi, H. Ahmad Khan, A. Nordon, Close-range hyperspectral imaging of whole plants for digital phenotyping: recent applications and illumination correction approaches, Comput. Electron. Agric. 178 (2020) 105780.
[17] P. Mishra, M.S.M. Asaari, A. Herrero-Langreo, S. Lohumi, B. Diezma, P. Scheunders, Close range hyperspectral imaging of plants: a review, Biosyst. Eng. 164 (2017) 49–67.
[18] N. Al Makdessi, M. Ecarnot, P. Roumet, G. Rabatel, A spectral correction method for multi-scattering effects in close range hyperspectral imagery of vegetation scenes: application to nitrogen content assessment in wheat, Precis. Agric. 20 (2019) 237–259.

[19] N. Vigneau, M. Ecarnot, G. Rabatel, P. Roumet, Potential of field hyperspectral imaging as a non destructive method to assess leaf nitrogen content in Wheat, Field Crop. Res. 122 (2011) 25–31.
[20] P. Mishra, A. Biancolillo, J.M. Roger, F. Marini, D.N. Rutledge, New data preprocessing trends based on ensemble of multiple preprocessing techniques, Trac. Trends Anal. Chem. (2020) 116045.
[21] J.-M. Roger, A. Biancolillo, F. Marini, Sequential preprocessing through ORThogonalization (SPORT) and its application to near infrared spectroscopy, Chemometr. Intell. Lab. Syst. 199 (2020) 103975.
[22] P. Mishra, J.M. Roger, F. Marini, A. Biancolillo, D.N. Rutledge, Parallel Preprocessing through Orthogonalization (PORTO) and its Application to Near-Infrared Spectroscopy, Chemometrics and Intelligent Laboratory Systems, 2020, p. 104190.
[23] X. Bian, P. Diwu, Y. Liu, P. Liu, Q. Li, X. Tan, Ensemble calibration for the spectral quantitative analysis of complex samples, J. Chemometr. 32 (2018) e2940.
[24] G. Rabatel, F. Marini, B. Walczak, J.-M. Roger, VSN: variable sorting for normalization, J. Chemometr. 34 (2020) e3164.
[25] R.J. Barnes, M.S. Dhanoa, S.J. Lister, Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra, Appl. Spectrosc. 43 (1989) 772–777.
[26] A. Savitzky, M.J.E. Golay, Smoothing and differentiation of data by simplified least squares procedures, Anal. Chem. 36 (1964) 1627–1639.
[27] P. Mishra, J.M. Roger, D.N. Rutledge, A. Biancolillo, F. Marini, A. Nordon, D. Jouan-Rimbaud-Bouveresse, MBA-GUI: A Chemometric Graphical User Interface for Multi-Block Data Visualisation, Regression, Classification, Variable Selection and Automated Pre-processing, Chemometrics and Intelligent Laboratory Systems, 2020, p. 104139.
[28] T. Isaksson, T. Næs, The effect of multiplicative scatter correction (MSC) and linearity improvement in NIR spectroscopy, Appl. Spectrosc. 42 (1988) 1273–1284.
[29] M.S. Dhanoa, S.J. Lister, R. Sanderson, R.J. Barnes, The link between multiplicative scatter correction (MSC) and standard normal variate (SNV) transformations of NIR spectra, J. Near Infrared Spectrosc. 2 (1994) 43–47.
[30] S. Khamis, T. Lamaze, Y. Lemoine, C. Foyer, Adaptation of the photosynthetic apparatus in maize leaves as a result of nitrogen limitation, Plant Physiol. 94 (1990) 1436.
[31] P.J. Curran, W.R. Windham, H.L. Gholz, Exploring the relationship between reflectance red edge and chlorophyll concentration in slash pine leaves, Tree Physiol. 15 (1995) 203–206.
[32] Beer, Bestimmung der Absorption des rothen Lichts in farbigen Flüssigkeiten, Ann. Phys. 162 (1852) 78–88.
[33] C. Pasquini, Near infrared spectroscopy: a mature analytical technique with new perspectives – a review, Anal. Chim. Acta 1026 (2018) 8–36.
[34] P. Mishra, D. Passos, A Synergistic Use of Chemometrics and Deep Learning Improved the Predictive Performance of Near-Infrared Spectroscopy Models for Dry Matter Prediction in Mango Fruit, Chemometrics and Intelligent Laboratory Systems, 2021, p. 104287.
[35] P. Mishra, D.N. Rutledge, J.-M. Roger, K. Wali, H.A. Khan, Chemometric preprocessing can negatively affect the performance of near-infrared spectroscopy models for fruit quality prediction, Talanta (2021) 122303.
[36] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-Cam, Visual explanations from deep networks via gradient-based localization, Int. J. Comput. Vis. 128 (2020) 336–359.
[37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention Is All You Need, arXiv preprint arXiv:1706.03762, 2017.