Chemometric pre-processing can negatively affect the performance of near-infrared spectroscopy models for fruit quality prediction

Puneet Mishra, Douglas Rutledge, Jean-Michel Roger, Khan Wali, Haris Ahmad Khan

Talanta 229 (2021) 122303

Contents lists available at ScienceDirect

Talanta

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

Chemometric pre-processing can negatively affect the performance of near-infrared spectroscopy models for fruit quality prediction

Puneet Mishra [a,*], Douglas N. Rutledge [b,c], Jean-Michel Roger [d,e], Khan Wali [f], Haris Ahmad Khan [f]

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

ARTICLE INFO

Keywords:
Artificial intelligence
Neural network
Fruit quality

ABSTRACT

Chemometrics pre-processing of spectral data is widely performed to enhance the predictive performance of near-infrared (NIR) models related to fresh fruit quality. Pre-processing approaches in the domain of NIR data analysis are used to remove the scattering effects, thus, enhancing the absorption components related to the chemical properties. However, in the case of fresh fruit, both the scattering and absorption properties are of key interest as they jointly explain the physicochemical state of a fruit. Therefore, pre-processing data that reduces the scattering information in the spectra may lead to poorly performing models. The objectives of this study are to test two hypotheses to explore the effect of pre-processing on NIR spectra of fresh fruit. The first hypothesis is that the pre-processing of NIR spectra with scatter correction techniques can reduce the predictive performance of models as the scatter correction can reduce the useful scattering information correlated to the property of interest. The second hypothesis is that the Deep Learning (DL) can model the raw absorbance data (mix of scattering and absorption) much more efficiently than the Partial Least Squares (PLS) regression analysis. To test the hypotheses, a real NIR data set related to dry matter (DM) prediction in mango fruit was used. The dataset consisted of a total of 11,420 NIR spectra and reference DM measurements for model training and independent testing. The chemometric pre-processing methods explored were standard normal variate (SNV), variable sorting for normalization (VSN), Savitzky-Golay based 2nd derivative and their combinations. Further two modelling approaches i.e., PLS regression and DL were used to evaluate the effect of pre-processing. The results showed that the best root mean squared error of prediction (RMSEP) for both the PLS and DL models were obtained with the raw absorbance data. The spectral pre-processing in general decreased the performance of both the PLS and DL models. Further, the DL model attained the lowest RMSEP of 0.76%, which was 13% lower compared to the PLS regression on the raw absorbance data. Pre-processing approaches should be carefully used while analysing the NIR data related to fresh fruit.

1. Introduction

Spectral data pre-processing is widely performed in chemometrics to remove or reduce the undesired artefacts from the spectra so that model predictive performance can be improved [1]. Several methods for pre-processing the spectra are available such as smoothing, baseline correction, normalizations and scatter correction [2,3]. Different pre-processing methods and their combinations are usually explored in combination with modelling approaches such as partial-least squares (PLS) regression to find the one with the best performance in terms of lowest prediction error [4]. Recently, to reduce the need to explore each pre-processing combination, ensemble approaches for combining the information from different pre-processing methods have been proposed [5,6]. Hence, whether alone or in combinations, various pre-processing techniques can be very effective in improving the predictive performance of chemometric models [2,7,8].

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

https://doi.org/10.1016/j.talanta.2021.122303
Received 4 February 2021; Received in revised form 5 March 2021; Accepted 8 March 2021
Available online 11 March 2021

In the case of near-infrared (NIR) spectroscopy of fresh fruit, the interaction of light results in two major phenomena i.e., absorbance and scattering [9]. The absorbance is related to the chemical components present in the fresh fruit, whereas the scattering is caused by the physical microstructure of the fruit skin and flesh [10]. The spectra recorded therefore contain a mixture of absorbance and scattering information [11]. The absorbance is highlighted as the broad peaks and valleys in the spectra while the scattering results in the global differences in intensities leading to additive and multiplicative effects [4]. Hence, to predict chemical properties such as dry matter (DM) and total soluble solids in fresh fruit, it seems practical to reduce/remove the scattering contribution from the spectra so that the remaining absorbance information can be linearly correlated to reference properties using linear models such as PLS [11,12].

Some commonly used pre-processing techniques in the domain of fresh fruit analysis are the estimation of 2nd derivative using Savitzky-Golay filtering to reveal underlying peaks [13], standard normal variate (SNV) [14] and its variants such as variable sorting for normalization (VSN) [15] to normalize the global intensity differences, as well as combinations of normalization and 2nd derivatives such as SNV or VSN followed by 2nd derivative [16]. Although with chemometric pre-processing methods the scattering information is reduced/removed from the spectral profiles, it is clear from previous research works that scattering information may provide added value related to the physical structure of materials which could enhance the model performance. Therefore, in some cases it may be interesting to retain the scattering information in the data while in others the additive and multiplicative effects may add to the difficulty of the modelling. To NIR spectroscopy of fresh fruit, the scattering may contain important information related to the property of interest [10,17]. This is because the fruit structure is highly correlated with the fruit ripeness stage, hence a correlation can also be expected with the key ripeness parameters such as DM. The fruit structure is mainly responsible for the scattering information in the NIR spectra of fresh fruit, hence, performing a pre-processing task which removes/reduces the scattering may not be an ideal solution to attain optimal models. Further, modelling NIR data containing both scattering and absorption information with linear PLS techniques may also be not a practical solution as the mixture of scattering and absorption can be highly non-linear to model. For this reason, there is a need for advanced pattern learning approaches that can learn the patterns from both the scattering and absorption information present in the data and correlate them with the property of interest.

Although pre-processing plays a significant role in chemometrics, a recent study found that the effect of pre-processing gets less significant as the number of samples increases [18]. This is because the latent space-based models such as PLS can learn added information as extra latent variables which can compensate for the effects which would need to be removed by the pre-processing. However, in the case of huge data sets, standard PLS may not be an efficient solution to learn complex mixed patterns such as the absorption and scattering characteristics in data [19]. In such a case, non-linear methods that can automatically learn non-linear patterns in data could be of interest. Recently, interest is growing for the use of advanced machine learning algorithms, such as deep learning (DL), to model spectral data [20–22]. The interest in DL is increasing because the DL model keeps on learning as the data increases whereas the traditional machine learning algorithms become saturated in performance at a certain point. Unlike the conventional DL analysis performed for computer vision tasks using convolutional 2D neural networks (2D-CNN), spectral data requires 1-dimensional convolutional neural networks (1D-CNN) as the single spectrum is $1 \times n$ size, where n is the number of wavelengths [22]. Applications of 1D CNN models to spectral data has already shown better performance than PLS regression approaches [22], however, the effect of chemometric pre-processing methods on the performance of 1D CNN models on real-life big spectral data sets is still unexplored.

The objectives of this study are to test two hypotheses to explore the effect of pre-processing on NIR spectra of fresh fruit. The first hypothesis is that the pre-processing of NIR spectra with scatter correction techniques can reduce the predictive performance of models as the scatter correction can reduce the useful scattering information correlated to the property of interest. The second hypothesis is that the DL can model the raw absorbance data (mix of scattering and absorption) much more efficiently than the PLS regression analysis. To test the hypotheses, a real NIR data set related to dry matter (DM) prediction in mango fruit was used. The dataset consisted of a total of 11,420 NIR spectra and reference DM measurements for model training and independent testing. The chemometric pre-processing methods explored were SNV, VSN, 2nd derivative by Savitzky-Golay filtering, as well as their combinations. Further two modelling approaches i.e., partial least-squares (PLS) regression and 1-D CNN based DL were used to evaluate the effect of pre-processing and identify the best technique to model the raw absorbance spectra. The criteria to test both the hypotheses was the lowest RMSEP.

2. Material and methods

2.1. Data set

The data set used in this study is comprised of a total of 11,691 NIR spectra (742–990 nm) and reference dry matter measurements performed on 4,675 mango fruit collected across 4 harvest seasons 2015, 2016, 2017 and 2018. The NIR spectra were acquired using F750 Produce Quality Meter (Felix Instruments, Camas, USA). The DM was measured with hot air oven drying (UltraFD1000, Ezidri, Beverley, Australia). The data set was sourced from Prof. Kerry Walsh, Central Queensland University, Australia [23]. 10,243 spectra, corresponding to season 2015, 2016 and 2017, were used for training and tuning, and the remaining 1,448 spectra from season 2018 were used as an independent test set. Due to the presence of outliers in both the training and test sets, hoteling T^2 and Q statistics (with PLS) was used to remove the abnormal samples. After outlier removal, the final data consisted of 10,135 samples in the training set and 1,285 samples in the test set. The data used can be found in the supplementary file.

2.2. Pre-processing method implemented

Based on the popularity of use and practicality in NIR data analysis of fresh fruit, three main pre-processing techniques were chosen [12,16]. The three techniques were 2nd derivative with Savitzky-Golay polynomial fitting [13], SNV [14] and VSN [15]. Furthermore, the two normalization techniques (SNV and VSN) were combined with the 2nd derivative to unravel the underling peaks. A combination of normalization technique with the 2nd derivative has already proven to be more powerful than other techniques [16].

In this study, the 2nd derivative in the spectral domain was implemented with a fixed window size of 13 and a 2nd order polynomial. The pre-treatments were implemented using in-house codes in MATLAB 2018b, MathWorks, Natick, USA.

2.3. Partial least squares regression

PLS regression analysis was performed as the baseline method to compare the effect of different pre-processing on the DL model performance. The PLS was implemented using the non-linear iterative partial least squares (NIPALS) algorithm which starts by using the response variable (in the case of a single response variable) to estimate the weights $\mathbf{w}$ for the $\mathbf{X}$ matrix such that the covariance between $\mathbf{Xw}$ and $\mathbf{y}$ is maximized. The weight vector is further normalized to unit norm, i.e., $||\mathbf{w}|| = 1$. The $\mathbf{X}$-scores are then estimated as $\mathbf{t} = \mathbf{Xw}$ and $\mathbf{y}$ subsequently regressed against $\mathbf{t}$. Finally, $\mathbf{X}$ and $\mathbf{y}$ are deflated to remove the variation extracted by the current LV. In this work a 10-fold cross-validation was used to determine the optimal number of LVs for the final PLS model. The PLS analysis was carried out using the 'plsregress' function in

MATALB's 'machine learning and statistics' toolbox.

2.4. Deep learning architecture and modelling

A 1-dimensional convolutional neural network (1D-CNN) deep learning (DL) architecture inspired from Ref. [22] was used for training and testing. A summary of the architecture is presented in Fig. 1, where 6 layered networks were created with one input layer, one 1D- CNN layer with a fixed kernel of width = 5 and stride = 1, three fully connected layers with 36, 18 and 12 neurons, respectively, and the final output layer with one neuron. To capture the non-linearity in the data, rectified linear units were used as the activation function between the layers. 10,135 training samples were further split into calibration (66.6%) and tuning set (33.3%) using the 'test_train_split' function from SciKit learn. The model weights were optimized with adaptive moment (Adam) optimizer and the mean absolute error was used as the loss function to train the network. A batch size of 256 was used and each model was trained up to 500 epochs. To have a fair comparison for different pre-processing techniques, the same architecture settings were used.

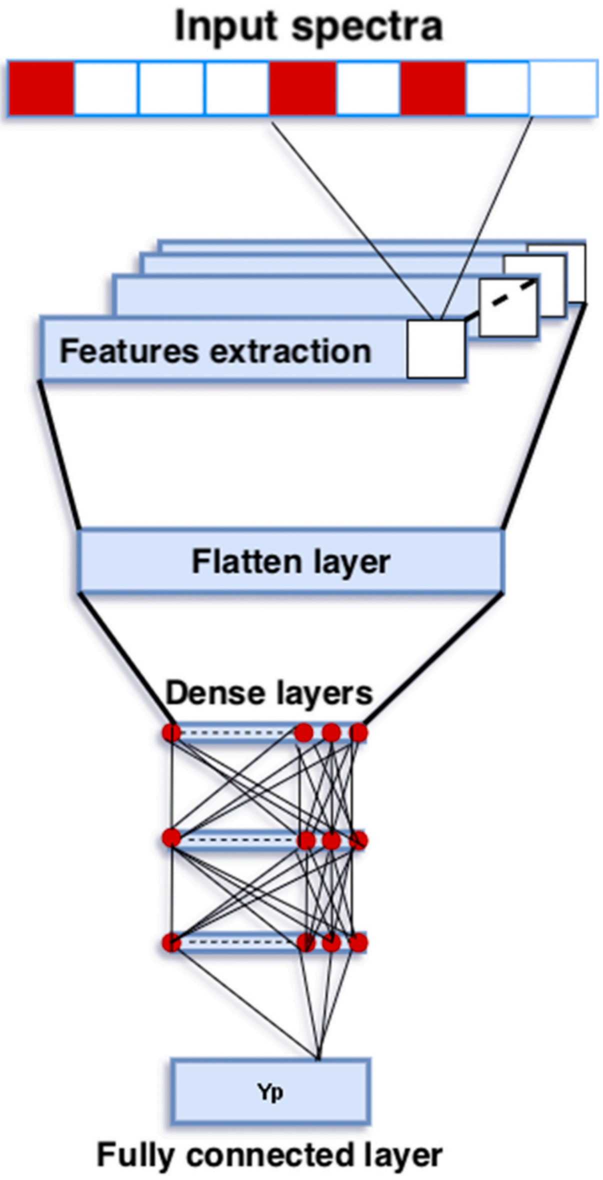

Fig. 1. A summary of the deep learning (DL) architecture implemented. The architecture consisted of one convolutional later and 3 dense fully connected layers.

All analyses were carried out using Tensorflow GPU 2.1.0 using the GeForce RTX 2080 Ti, Nvidia, Santa Clara, California, USA, using a desktop computer equipped with a 3.60 GHz Intel® Xeon® W-2133 processor (Intel Corporation, Santa Clara, CA) and 64 GB RAM, running Microsoft Windows 10 operating system (Microsoft Inc., Redmond, WA) and 64-bit MATLAB 2018b (The Mathworks, Natick, MA).

3. Results and discussion

3.1. Spectral profiles and pre-processings

Mean spectral profiles (742–990 nm) of the training and test sets as raw absorbance spectra and after several pre-processings are shown in Fig. 2. In the raw absorbance data (Fig. 2A), a global intensity difference can be seen in the mean spectral profiles of the training and test sets. Such global differences in intensity could be related to the presence of additive and multiplicative effects in the spectra due to changes in the light scattering caused by the interaction of light with the mango fruit at various levels of ripening [2] being influenced by changes in the cell structure of the fruit. For example, a hard fruit has a stiffer cell structure and exhibits refractive index steps that cause light scattering. On the contrary, a soft fruit has a looser cell structure [9] and the space between the cells is filled with liquid, which reduces refractive index steps, thus reducing light scattering. The raw spectra (Fig. 2A) also present a high absorbance near the spectral band at 960 nm which could be related to higher moisture content in fresh produce such a mango fruit [10,24]. The presence of high moisture usually masks peaks related to other chemical constituents such as sugars, fats and proteins, and methods such a 2nd derivative are usually required to reveal the underlying peaks [25]. Following the application of the 2nd derivative (Fig. 2B), several underlying peaks were in fact revealed, e.g., at 820, 870 and 920 nm which can be assigned to the 3rd overtones of NH and CH bonds, related to the protein and fats in the mango fruit [25]. The 2nd derivative also reduced the global difference in the spectra, but on the other hand, the shape of the spectral profile is lost. The SNV transformation (Fig. 2C) also allowed to reduce the global intensity differences while keeping the same shape of the spectra but did not reveal the underling peaks. However, the 2nd derivative of the SNV pre-processed data revealed the underlying peaks (Fig. 2E). The VSN pre-processing also reduced the overall intensity differences but only in the zones where the scattering dominates the absorption, while retaining the variability near the highly absorbing moisture bands i.e., 960 nm. Combining VSN with the 2nd derivative revealed the underlying peaks. A point to be noted is that the peak revealed after the 2nd derivative of the SNV or VSN pre-processed data were like the peaks revealed with 2nd derivative on raw absorbance data. A summary of reference dry matter (DM) values is presented in Fig. 3. The DM of the test set was well represented in the DM of the training set except for high DM samples which were less present in the training set compared to the test set. The DM of the training and test sets were 16.2 ± 2.4% and 16.9 ± 2.6%, respectively.

3.2. Partial least-squares vs deep learning vs pre-processings

A summary of PLS regression and DL model performances on raw absorbance and differently pre-processed data is shown in Table 1. For both PLS and DL, the model based on raw absorbance data outperformed the models based on pre-processed data, thus proving the first hypothesis in this study (scatter removal can deteriorate NIR models of fresh fruit) to be true. This claim is also supported by a recent study on pear fruit, where the best models for moisture content prediction (100% - DM) were attained when the raw absorbance information was incorporated in the PLS model [5]. The outcomes of PLS and DL models on raw absorbance data are shown in Fig. 4. The RMSEP for DL on the raw absorbance data was 13% lower than for the PLS regression, thus demonstrating the superiority of DL in modelling the raw absorbance

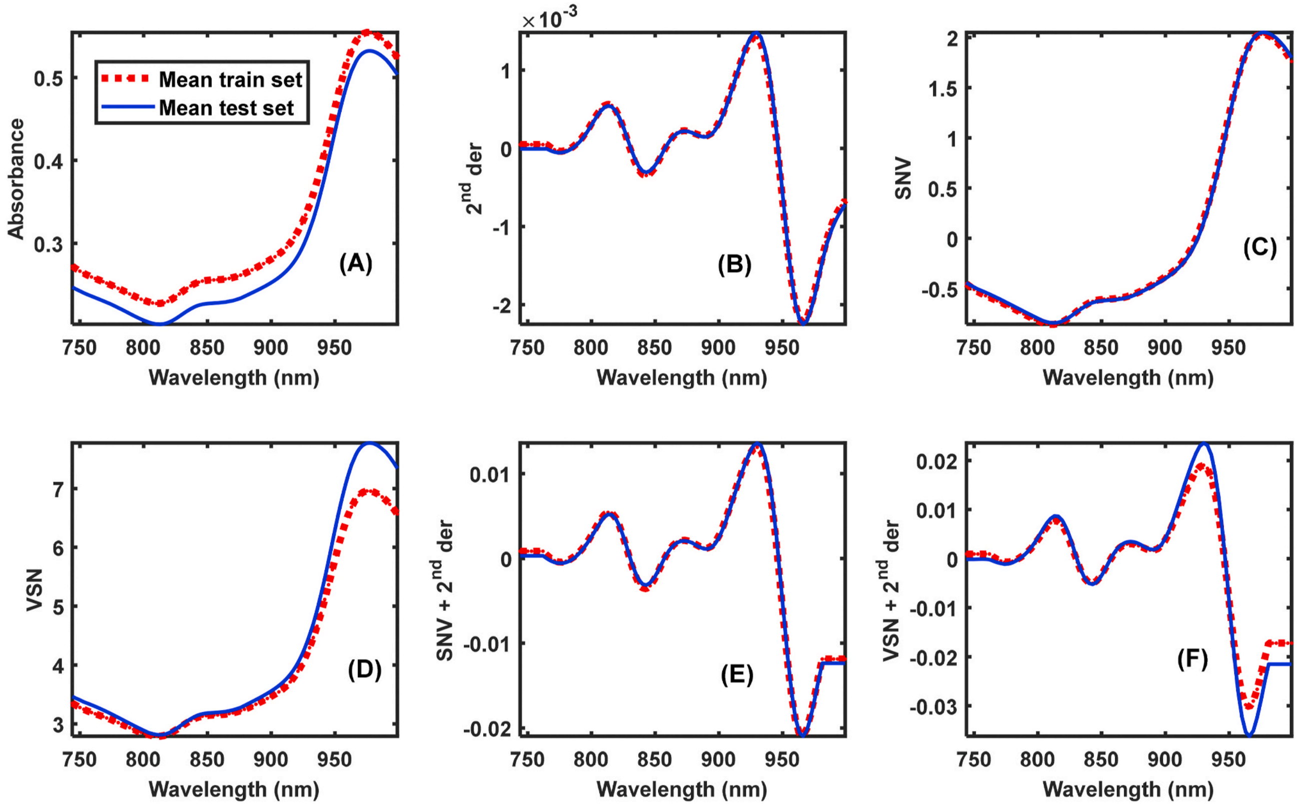

Fig. 2. Mean spectra of the training and test sets for absorbance spectra and after several pre-processings. (A) Absorbance, (B) 2nd derivative (window 13, order 2), (C) SNV, (D) VSN, (E) SNV followed by 2 nd derivative, and (F) VSN followed by 2 nd derivative.

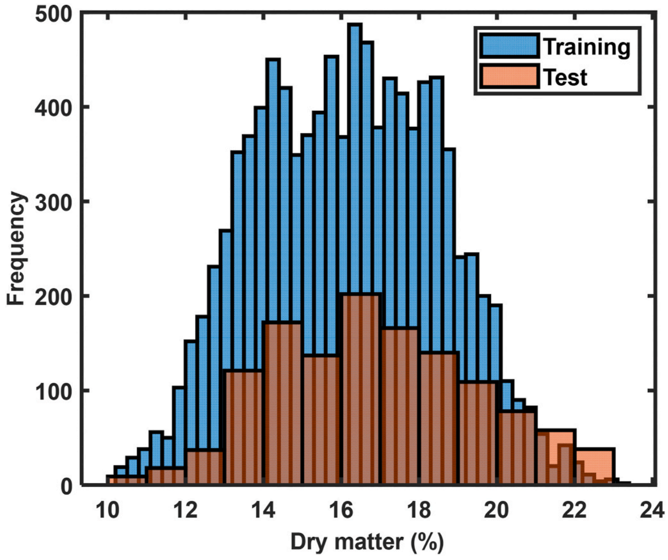

Fig. 3. Distribution of dry matter in mangoes for the training and test sets.

Table 1

A summary of performances of partial least squares (PLS) and deep learning (DL) models for predicting dry matter in mangoes using different pre-processing.

Pre-processing methods	RMSEP of PLS model/Latent Variables	RMSEP of DL
Raw absorbance	0.87/8 LVs	0.76
SNV	0.97/7 LVs	0.81
VSN	1/7 LVs	0.91
2nd derivative	0.88/8 LVs	0.92
SNV+2nd derivative	0.98/7 LVs	0.90
VSN+2nd derivative	0.95/9 LVs	0.98

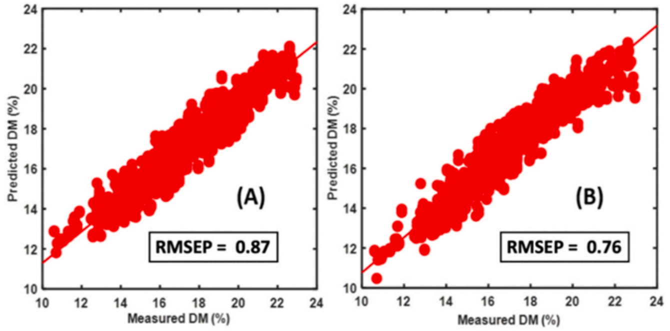

Fig. 4. Prediction of the test set by partial least squares regression (A) and deep learning (B), with a model based on raw absorbance data for dry matter prediction in mango fruit.

data. A 13% lower RMSEP of the DL model compared to the PLS regression model proves that the second hypothesis in this study is also true. The second-best DL model was obtained with the SNV pre-processed data and was also better than any PLS model. In the case of PLS modelling, the 2nd best performance was obtained on the 2nd derivative pre-processed data with same number of LVs as obtained for the raw absorbance data modelling. The 2nd derivative did not drastically reduce the performance of models compared to the scatter correction techniques, SNV and VSN. The better performance of 2nd derivative compared to the scatter correction technique is because the 2nd derivative is not only a scatter reduction technique, but also a peak revealing technique, thus completing the NIR data. And in fact, using scatter correction prior to 2nd derivative led to poorer PLS models compared to the PLS model on the data pre-processed only by the 2nd derivative.

The best performance of the PLS model for the raw absorbance data agrees with the results of Schoot et al. [18], who conclude that with big spectral data sets the effect of pre-processing is reduced and raw data is sufficient to attain high accuracy models. This study finds the same thing for DL, where the raw absorbance data give better models than the pre-processed data. The results were also in agreement with Ciu et al. [22], where the author concluded that pre-processing of spectral data may not be required as the DL model may be able to learn the required transformations automatically.

To get an enhanced understanding of the learning performed by PLS and DL on the raw data, the PLS regression weights and the mean activation of the convolutional layer of the DL model are shown in Fig. 5. At a first glance the PLS regression weights are dominated by the peak at 920 nm which can be related to the 3rd overtones of C–H bonds [25], however, the DL model shows near zero weights in that region. The DL model showed higher weights in the spectral range at 960 nm which is related to the 3rd overtones of the O–H bonds [25] related to the moisture in the fresh mangoes. A reason for the better performance of the DL model on the test set (which is a new harvest season) can be that moisture in fruit is a more generalized indicator of DM than the fatty acids which may vary between cultivars and be season specific.

3.3. Effect of pre-processing on PLS regression with different sample sizes

In this study, pre-processing of data in general lead to poor accuracy of models. The same holds when the data set size was reduced (Fig. 6) by multiple folds ((100%, 50%, 25%, 12.75%, and 3.37%). This analysis was only performed for the PLS regression models and not for DL, as DL requires big data. A recent study concluded that when the NIR data set is small, pre-processing significantly improved the model accuracies [18]. The present study shows that the conclusion proposed in [18] does not always hold true as here, in the case of fresh fruit analysis using NIR spectroscopy, the pre-processed data never performed better than the raw absorbance. In this study, exploring the individual pre-processings for different sample sizes showed that most pre-processings either maintained the same predictive performance or showed improved performance of the PLS models (Fig. 6) as the sample size increases (3.37%, 12.75%, 25%, 50% and 100%). This indicates that the use of more data in this study (irrespective of pre-processing) in general improved the performance of PLS models.

3.4. A posteriori analysis to determine whether any pre-processed data carry complementary information to raw absorbance

It has been shown that the PLS models based on raw absorbance performed the best compared to any other pre-processed data. However, differently pre-processed data may carry complementary information which may benefit the modelling performed with raw data alone [2,5,6,12]. To find any existing complementary information in differently pre-processed data, *a posterior* analysis was performed. At first the RMSEP for PLS models based on raw absorbance were explored in the LVs range from 1 to 30. In Fig. 7A, it can be noted that 7 LVs were sufficient to reach a RMSEP of 0.85%. After that, using the scores from the 7 LVs extracted from the raw absorbance data and the already explained part of the response variable (DM), the data matrices corresponding to each differently pre-processed data block and the response variable (DM) were orthogonalized. After orthogonalization the unique information present in the differently pre-processed data was used to model the unexplained part of the response variable. The PLS models for each differently pre-processed data block were explored in the range of 1–30 LVs. It can be noted in Fig. 7B that out of all pre-processing approaches, combining 4 LVs from the VSN pre-processed data (black solid line) with the 7 LVs of raw absorbance decreased the RMSEP from 0.85% to 0.82%, thus, indicating that VSN carries information complementary to the raw absorbance data. Although combining the information from raw absorbance and VSN pre-processed data decreased the RMSEP, it never achieved the RMSEP of 0.76% attained by the DL analysis. To understand the complementary information present in the VSN pre-processed data block, the regression vectors for raw absorbance and VSN pre-processed data are shown in Fig. 8. For raw absorbance: The left part of the coefficients has the same shape as the raw spectra. The regression is therefore based on baseline variations. The sine profile, positive for 920 nm and negative for 950 nm indicates that the regression is sensitive to a shift of the main peak flank. The peak of the coefficients at 975 nm is directly related to the water peak. These two characteristics are both related to the inflation of the main peak of the spectra. It may seem inquisitive to find positive coefficients at the peak related to water content, knowing that DM is negatively correlated with water content. However, one can hypothesize that the regression is sensitive to scattering, which has the effect of increasing the apparent absorbance. For VSN pre-processed spectra: The regression coefficients show peaks which correspond to chemical absorptions. Thus, the two negative peaks at around 760 nm and 975 nm can be directly attributed to water absorption [25].

4. Conclusions

This study tested two hypotheses concerning the effect of spectral pre-processing on the quality of predictive models. The first hypothesis was that the pre-processing of NIR spectra of mango fruit with scatter correction techniques can reduce the predictive performance of models. The second hypothesis was that the DL can model the raw absorbance data (mix of scattering and absorption) much more efficiently than the PLS regression analysis. The criteria to test both the hypotheses was the lowest RMSEP. The results from the study showed that both hypotheses are true as the lowest RMSEP models (both PLS and DL) for predicting DM in mango fruit were achieved with the raw absorbance data. The scatter correction methods, particularly SNV, VSN and their combination with 2nd derivative drastically degraded the PLS models for predicting DM in mango fruit. A reason put forward to explain this is that the scatter correction removes the useful scattering information correlated to the property of interest. Furthermore, for the raw data, DL

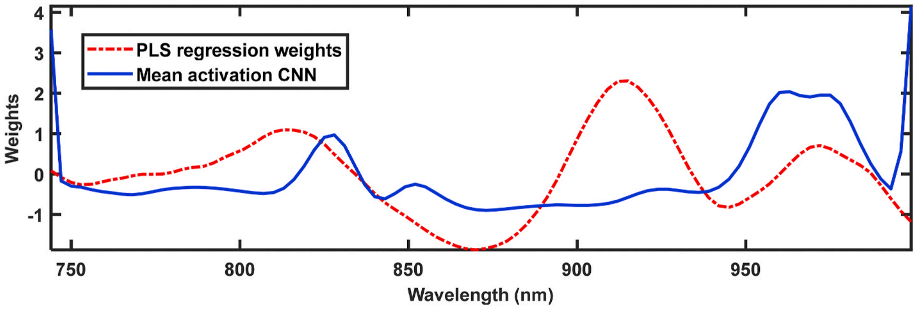

Fig. 5. Comparison of the spectral features detected by PLS regression and the mean activation of the convolutional layer, to predict dry matter in mango fruit.

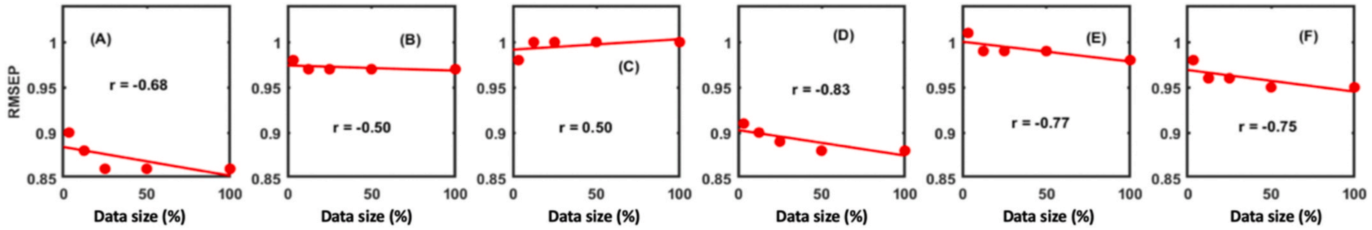

Fig. 6. A summary of correlation between the data size (3.37%, 12.75%, 25%, 50% and 100%) and the RMSEP of respective models (1st row). (A) Raw absorbance, (B) SNV, (C) VSN, (D) 2nd derivative, (E) SNV+2nd derivative, and (F) VSN+2nd derivative.

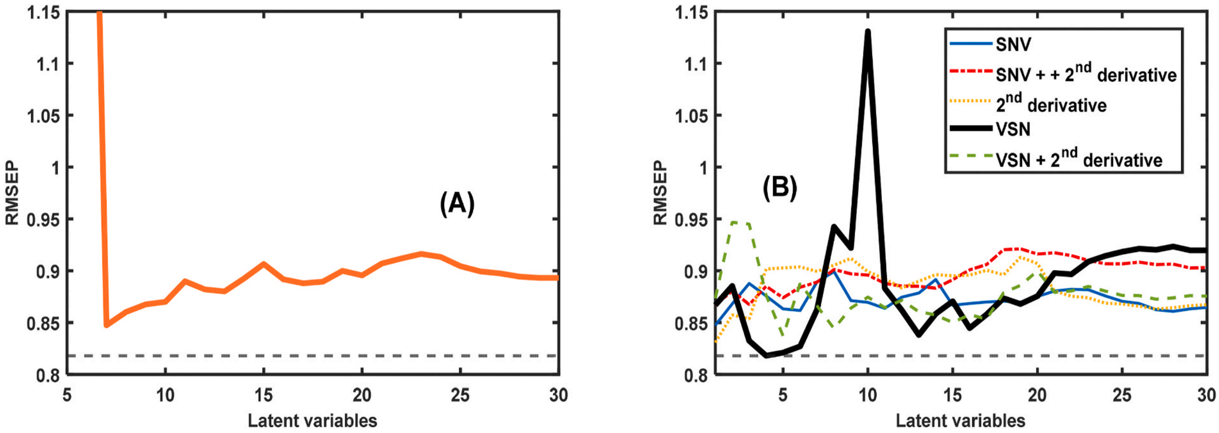

Fig. 7. (A) Evolution of RMSEP for PLS models for between 5 and 30 latent variables for raw absorbance data, and (B) evolution of RMSEP for differently pre-processed data, over the range of 1–30 latent variables, to find complementary information to the raw absorbance.

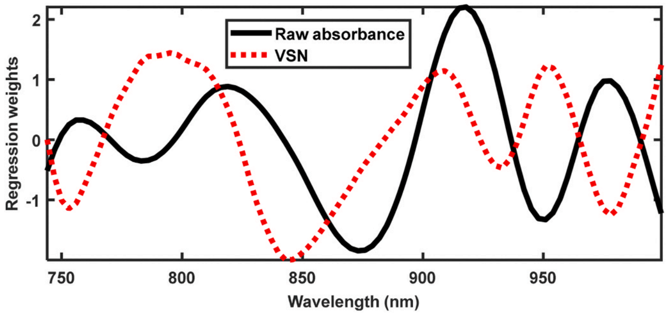

Fig. 8. The regression vector of the raw absorbance and the VSN pre-processed data.

models achieved a 13% lower RMSEP compared to the PLS models, indicating when big data is available, DL should be the preferred approach to attain high accuracy predictive models. This study also explored the effect of sample size and pre-processing on the performance of PLS models and found that increasing data in general improved the PLS modelling irrespective of the pre-processing. There was also complementary information present in the VSN pre-processed data which improved the performance of PLS models based on the raw absorbance, but the performance of DL models on the raw data outperformed all.

Author statement

Puneet Mishra: Conceptualization; Methodology; Software; Writing – original draft; Data curation, Douglas N. Rutledge: Conceptualization; Methodology; Writing – review & editing. Jean Michel Roger: Conceptualization; Methodology; Writing – review & editing. Khan Wali: Software; Formal analysis; Writing – review & editing. Haris Ahmad Khan: Software; Formal analysis; 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] J. Engel, J. Gerretzen, E. Szymańska, J.J. Jansen, G. Downey, L. Blanchet, L.M. C. Buydens, Breaking with trends in pre-processing? Trac. Trends Anal. Chem. 50 (2013) 96–106.
[2] 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.
[3] J.-M. Roger, J.-C. Boulet, M. Zeaiter, D.N. Rutledge, Pre-processing Methods, Reference Module in Chemistry, Molecular Sciences and Chemical Engineering, Elsevier, 2020, pp. 1–75.
[4] Å. Rinnan, F.v.d. Berg, S.B. Engelsen, Review of the most common pre-processing techniques for near-infrared spectra, Trac. Trends Anal. Chem. 28 (2009) 1201–1222.
[5] P. Mishra, J.M. Roger, F. Marini, A. Biancolillo, D.N. Rutledge, Parallel Pre-processing through Orthogonalization (PORTO) and its Application to Near-Infrared Spectroscopy, Chemometrics and Intelligent Laboratory Systems, 2020, p. 104190.
[6] 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.
[7] P. Mishra, A. Nordon, J.-M. Roger, Improved prediction of tablet properties with near-infrared spectroscopy by a fusion of scatter correction techniques, J. Pharmaceut. Biomed. Anal. (2020) 113684.
[8] P. Mishra, F. Marini, A. Biancolillo, J.-M. Roger, Improved prediction of fuel properties with near-infrared spectroscopy using a complementary sequential fusion of scatter correction techniques, Talanta (2020) 121693.
[9] R.F. Lu, R. Van Beers, W. Saeys, C.Y. Li, H.Y. Cen, Measurement of optical properties of fruits and vegetables: a review, Postharvest Biol. Technol. 159 (2020).
[10] K.B. Walsh, J. Blasco, M. Zude-Sasse, X. Sun, Visible-NIR 'point' spectroscopy in postharvest fruit and vegetable assessment: the science behind three decades of commercial use, Postharvest Biol. Technol. 168 (2020) 111246.
[11] W. Saeys, N.N. Do Trong, R. Van Beers, B.M. Nicolai, Multivariate calibration of spectroscopic sensors for postharvest quality evaluation: a review, Postharvest Biol. Technol. (2019) 158.

[12] P. Mishra, J.M. Roger, D.N. Rutledge, E. Woltering, SPORT pre-processing can improve near-infrared quality prediction models for fresh fruits and agro-materials, Postharvest Biol. Technol. 168 (2020) 111271.

[13] A. Savitzky, M.J.E. Golay, Smoothing and differentiation of data by simplified least squares procedures, Anal. Chem. 36 (1964) 1627–1639.

[14] 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.

[15] G. Rabatel, F. Marini, B. Walczak, J.-M. Roger, VSN: variable sorting for normalization, J. Chemometr. 34 (2020) e3164.

[16] X. Sun, P. Subedi, R. Walker, K.B. Walsh, NIRS prediction of dry matter content of single olive fruit with consideration of variable sorting for normalisation pre-treatment, Postharvest Biol. Technol. 163 (2020) 111140.

[17] R. Lu, R. Van Beers, W. Saeys, C. Li, H. Cen, Measurement of optical properties of fruits and vegetables: a review, Postharvest Biol. Technol. 159 (2020) 111003.

[18] M. Schoot, C. Kapper, G.H. van Kollenburg, G.J. Postma, G. van Kessel, L.M. C. Buydens, J.J. Jansen, Investigating the need for preprocessing of near-infrared spectroscopic data as a function of sample size, Chemometr. Intell. Lab. Syst. 204 (2020) 104105.

[19] R.D. Cramer, Partial least squares (PLS): its strengths and limitations, Perspect. Drug Discov. Des. 1 (1993) 269–278.

[20] M. Chatzidakis, G.A. Botton, Towards calibration-invariant spectroscopy using deep learning, Sci. Rep. 9 (2019) 2126.

[21] J.S. Larsen, L. Clemmensen, Deep Learning for Chemometric and Non-translational Data, 2019 arXiv preprint arXiv:1910.00391.

[22] C. Cui, T. Fearn, Modern practical convolutional neural networks for multivariate regression: applications to NIR calibration, Chemometr. Intell. Lab. Syst. 182 (2018) 9–20.

[23] N.T. Anderson, K.B. Walsh, P.P. Subedi, C.H. Hayes, Achieving robustness across season, location and cultivar for a NIRS model for intact mango fruit dry matter content, Postharvest Biol. Technol. 168 (2020) 111202.

[24] K.B. Walsh, V.A. McGlone, D.H. Han, The uses of near infra-red spectroscopy in postharvest decision support: a review, Postharvest Biol. Technol. 163 (2020) 111139.

[25] B.G. Osborne, Near-Infrared Spectroscopy in Food Analysis, Encyclopedia of Analytical Chemistry, 2006, https://doi.org/10.1002/9780470027318.a1018.