Chemometrics and Intelligent Laboratory Systems 212 (2021) 104287

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

# A synergistic use of chemometrics and deep learning improved the predictive performance of near-infrared spectroscopy models for dry matter prediction in mango fruit

Puneet Mishra [a,*], Dário Passos [b]

[a] Wageningen Food and Biobased Research, Bornse Weilanden 9, P.O. Box 17, 6700AA, Wageningen, the Netherlands
[b] CEOT, Universidade do Algarve, Campus de Gambelas, FCT Ed.2, 8005-189 Faro, Portugal

ARTICLE INFO

*Keywords:*
1D-CNN
Neural networks
Fruit quality
Artificial intelligence
Ensemble pre-processing

ABSTRACT

This study provides an innovative approach to improve deep learning (DL) models for spectral data processing with the use of chemometrics knowledge. The technique proposes pre-filtering the outliers using the Hotelling's $T^2$ and Q statistics obtained with partial least-square (PLS) analysis and spectral data augmentation in the variable domain to improve the predictive performance of DL models made on spectral data. The data augmentation is carried out by stacking the same data pre-processed with several pre-processing techniques such as standard normal variate, 1st derivatives, 2nd derivatives and their combinations. The performance of the approach is demonstrated on a real near-infrared (NIR) data set related to dry matter (DM) prediction in mango fruit. The data set consisted of a total 11,961 spectra and reference DM measurements. The results showed that removing the outliers and augmenting spectral data improved the predictive performance of DL models. Furthermore, this innovative approach not only improved DL models but attained the lowest root mean squared error of prediction (RMSEP) on the mango data set i.e., 0.79% compared to the best known RMSEP of 0.84%. Further, by removing outliers from the test set the RMSEP decreased to 0.75%. Several chemometrics approaches can complement DL models and should be widely explored in conjunction.

## 1. Introduction

Deep learning (DL) after successfully solving several challenges in the domain of computer vision is now expanding in the chemometrics domain. Several primary applications of DL can be found to deal with chemometrics challenges such as spectral modelling [1], hyperspectral image processing [2–4], data clustering [5], molecular generative modelling [6], soft sensor modelling [7] and peaks identification in chromatographic data [8]. The basic foundations of DL are artificial neural network (NN) algorithms that enable the learning process of complex hidden non-linear patterns in the data, otherwise unachievable with classical machine learning and chemometrics techniques. The major role of DL comes into play in the availability of huge data sets, while in the case of small data sets, classical machine learning and chemometrics techniques can perform equally well. Further, based on the type of data, different approaches to creating DL models are available such as convolutional neural networks (CNNs) in the case of image processing and 1-D spectral data, whereas long-short term memory (LSTM) and gated rectified units (GRU) are mostly used for sequence data such as time series and natural language processing (NLP).

DL in spectral data modelling is currently emerging and a few works can be found related to regression modelling [1,3,4]. Basically, two main approaches to spectral data processing are available, first is the use of methods that are completely supervised and jointly perform feature extraction and learning, such as, 1-Dimensional (1D) CNNs with fully connected layers (FC) [1], and second is the combination of unsupervised feature extraction with the supervised regression modelling such as the use of auto-encoders for feature extraction as a first step and later training using support vector machines (SVM) to map the extracted features with the property of interest [4]. The first approach involving joint feature extraction and learning has the advantage that the features extracted are related to the property of interest whereas the in the second approach the feature extracted in an unsupervised way may not relate to the property of interest.

Apart from advancement in DL modelling approaches, in recent years, major developments have taken place in the development of miniature

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

https://doi.org/10.1016/j.chemolab.2021.104287
Received 14 December 2020; Received in revised form 23 February 2021; Accepted 7 March 2021
Available online 11 March 2021

spectrometers and several low-cost spectrometers are now available in the market [9,10]. These low-cost spectrometers are easy to use; hence, their usage range from high-end research facilities to an individual non-expert user [9]. A side outcome of such a wide distributed usage of low-cost spectrometers is the generation of huge data sets, which for DL modelling is of high-interest. A practical example is the portable spectroscopy of fresh fruit, where near-infrared is widely expanding with the availability of portable handheld spectrometers [11–18]. Further, a recent experiment using one of these handheld units was recently made available as an open access 'mango data set' comprising a total of 11,961 near-infrared (NIR) spectra and reference dry matter (DM) [17,19]. DM is a key quality trait in fresh fruit and allows to judge its maturity [18]. NIR spectroscopy is of huge interest as it can replace the laboratory-based DM analysis with a rapid non-destructive on-site measurement [20]. In previous works related to the analysis of this mango data set, several linear, non-linear, local, global and ensemble methods have been tried [17,19]. The lowest reported root mean squared error of prediction (RMSEP) on the mango data set is 0.839% and was achieved using an ensemble of the artificial neural network, Gaussian process regression and local PLS [19]. For individual models, the lowest RMSEP achieved was 0.881% using a Local Optimized by Variance Regression (LOVR) model [19]. However, no study yet reported the use of DL on this mango data set, hence, for the first time a DL analysis on mango data set is performed in this study.

Most of the previous works on DL analysis of spectral data were focussed on optimising the model architecture and showing that this kind of algorithms could achieve higher prediction accuracy than standard chemometrics approaches [1,3,4]. However, none of the spectral DL modelling approaches focussed on finding a cooperative strategy between chemometrics and DL. Initial works on DL spectral analysis suggest that the first layers of certain NN architectures can perform data transformations akin to classical pre-processing methods in an automated way [1,30]. However, it is not clear how deep these NN need to be to achieve this feature [31] and if this is indeed the best strategy for extracting information from spectral data. Chemometrics has several approaches that can directly benefit the DL such as pre-filtering outliers with the use of Hotelling's $T^2$ and Q statistics and spectral data pre-processing to remove the artefacts from spectral data. Furthermore, the recent concept of ensemble data pre-processing can play a key role in spectral data augmentation where benefits of different pre-processing techniques such as spectral normalisation and derivative can be combined to learn complementary information [21–23]. Augmentation of spectral data with several pre-processing methods can pre-enhance the features for the DL modelling, thus, making it more efficient and accurate.

The objective of this study is to show how the performance of DL models for spectral data analysis can be improved using chemometrics knowledge. In the field of machine learning, this is sometimes referred to as the introduction of expert knowledge into the models. The technique proposes pre-filtering the outliers using the classical Hotelling's $T^2$ and Q statistics approach using PLS analysis and spectral data augmentation in the variable domain to improve the predictive performance of DL models applied to spectral data. The data augmentation is carried out by stacking the same data pre-processed with several pre-processing techniques such as standard normal variate (SNV), 1st derivatives, 2nd derivatives and their combinations. The SNV and derivative pre-processing were chosen as they do not require any external reference spectra or weight estimation to execute and hence, can be easily translated to any external independent test set. The performance of the approach is demonstrated on a real NIR data set related to DM prediction in mango fruit. The data set consisted of a total 11,961 spectra and reference DM measurements. The performance of the DL model was compared with the best-known RMSEP obtained with individual and ensemble models for the same data.

**Table 1**
Summary of dry matter for training and test set before and after outlier's removal.

| Data set | Dry matter (%) with outliers | Dry matter (%) without outliers |
|---|---|---|
| **Train** | 16.17 ± 2.41 | 16.13 ± 2.38 |
| **Test** | 17.01 ± 2.67 | 17.01 ± 2.65 |

## 2. Materials and method

### 2.1. Data set

The data set used in this study comprises a total of 11,691 NIR spectra (684–990 nm in 3 nm sampling with a total 103 variables) and DM measurements performed on 4675 mango fruit across 4 harvest seasons 2015, 2016, 2017 and 2018 [24]. Portable F750 Produce Quality Meter (Felix Instruments, Camas, USA) was used for the non-destructive NIR measurements. DM was measured with oven drying (UltraFD1000, Ezidri, Beverley, Australia) on the samples extracted from the sample location on fruit where the NIR measurements were done. The data set is associated with the publications [17,19] and can also be accessed in the supplementary file section of the manuscripts. Out of 11,691 spectra, 10,243 spectra corresponding to harvest seasons 2015, 2016 and 2017 were used for training and tuning, whereas the remaining 1448 spectra from season 2018 were used as an independent test set. This data partition was used as it was previously used in Refs. [17,19] to compare the performance of the DL models to previously reported results. Even though the authors of the mango dataset report that they eliminated a few outliers in the data, in this study a more stringent outlier analysis was performed. The presence of outliers in both the training and test sets was detected by using Hotelling's $T^2$ and Q statistics from PLS decomposition with NIPALS algorithm [25] leading to the removal of the abnormal samples pertaining either high $T^2$ or Q statistics. For outlier removal, at first a PLS model was built (on the training set with outliers) with 10-fold cross-validation using the MBA-GUI [26], and later, the scores from the PLS decomposition were used to estimate the $T^2$ and Q statistics. The computed $T^2$ and Q statistics were plotted in a 2D plot, and later, manually by user, the outlying samples were eliminated by region of interest selection using MATLAB's 'roipoly' function. After, the outlier removal, a new PLS decomposition was performed and new $T^2$ and Q statistics were computed and plotted in a 2D plot. Once again based on the visual inspection of the samples, the user selects a region of interest using MATLAB's 'roipoly' function and remove the outlying samples. The process was repeated until samples with high $T^2$ and Q statistics were removed. Such a procedure is commonly used and is available in commercial chemometric toolboxes. However, in this study, in house MATLAB codes were used for performing the outlier removal. The outlier removal in this study was performed separately for training and test set. The separate outlier removal was performed such that the training set do not have any influence over the test set. However, the DL analysis was performed for data with and without outlier removal to have a fair comparison how outlier removal affects the DL performance as well to have a fair comparison with the work which were reported on data without removing these outliers [17,19]. The final training and test sets, after outlier removal, comprised of 9914 samples in the training set and 1413 samples in the independent test set. The final training and test sets are also provided as a supplementary file. A summary of reference DM before and after removal of outliers in the training and test set are shown in Table 1. It can be noted that before and after the outlier removal the means of training and test set were similar ~16% and ~17% respectively.

### 2.2. Data augmentation

Absorbance spectral data and differently pre-processed data in chemometrics carry complementary information [21,22,27–32]. For example, since the absorbance data carries both the absorption and scattering characteristics, the normalisation of spectral data with techniques such as standard normal variate (SNV) [33] enhances the absorption features by reducing the additive and multiplicative effects

related to light scattering, and derivatives can help in revealing underlying peaks which can enhance the data modelling [21]. Motivated from the complementary nature of spectral pre-processing, this study uses stacking of differently pre-processed data in the variable domain to perform the data augmentation to facilitate DL. The augmentation do not increase the samples size but extended the total number of variables from 103 to 618 (i.e. 103 × 6) in 3 nm sampling. A summary of total spectra before and after outlier removal and after data augmentation is shown in Table 2. Only SNV and smooth derivatives (Savitzky-Golay [34], window size = 13 and 2nd order polynomial) were used for spectral data augmentation as the task of data normalisation and extraction of underlying peaks can be carried out with SNV and derivatives, respectively. Further, there was no rationale behind the choice of order of pre--processing's as the DL architecture was not a sequential model; hence, no effect of pre-processing order is expected. SNV and derivative estimation were implemented using MBA-GUI [26] in MATLAB 2018b, MathWorks, Natick, USA.

The training samples were further shuffled and randomly partitioned into calibration (66.6%) and tuning (33.3%) sets using the 'test-train split' function (with random_state = 42) from SciKit-Learn (v.0.24.1) (https://scikit-learn.org/stable/). The training data were partitioned only once and all models were evaluated on the same partition. Single partition was used to lower the computation time for the hyperparameter optimization and grid search. Furthermore, the data already has many data points, therefore the chance of samples misrepresentation is low compared to the case when the chemometric modelling is performed on a few hundred spectra as in such a case, single partition with such a small number of samples may misrepresent some samples. These two sets are used for the DL model optimization and the test set is used for computing final metrics. Finally, all sets were standardized column-wise (using the mean and standard deviation of the train set to scale the test set) before feeding them to the NN algorithm. For this work, the root mean square error of prediction (RMSE) was used to assess the quality of the final models.

### 2.3. Deep leaning

#### 2.3.1. 1D-CNN architecture

The DL model architecture used was a 1-Dimensional convolutional neural network (1D-CNN) architecture developed in Ref. [1]. In Ref. [1] the details about the development of this CNN architecture including a simplified discussion about its hyper-parameters and how they affect the model is provided. A visual summary of the architecture is presented in Fig. 1, where a 6 layers network was created with 1 input layer, 1 convolution layer with 1 fixed kernel (a.k.a filter) and stride = 1 followed by 3 fully connected layers (FC) with 36, 18 and 12 units, respectively, and a final output layer with one unit.

To capture the non-linearity in the data, exponential linear unit (eLU) was used as the activation function between the layers except for the last layer that uses a linear activation function. The weights on the multiple layers were initialized using the 'HeNormal' initialization (tf.keras.he_normal (seed = 42)) and were trained using an adaptive moment optimizer algorithm (Adam) with an initial learning rate given by 0.01 × (batch size)/256. The mean squared error (MSE) was used as the loss function and layer regularization was implemented by adding an L2 penalty (β) on the model weighs (and added to the loss function). During training, the learning rate (LR) was iterative, decreasing by a factor of 2 when the validation loss was not improved by $10^{-6}$ after 25 epochs (using tf. keras.ReduceLROnPlateau () function). The maximum number of epochs allocated for the training process was 750, but during the optimization tests, the automatic Early Stopping algorithm (tf.keras.EarlyStopping () function) was used. This algorithm monitors the validation loss and terminates the training process if the validation metrics do not improve in $10^{-5}$ after 50 consecutive epochs. The calibration set was used for model training and the tuning set was used for validation. The test set was not used during the optimization phase.

To provide a first-hand comparison with a standard chemometric technique, PLS models [35] were also developed with outlier removed and augmented using the 'plsregress' function from MATLAB's statistics and machine learning toolbox. The latent variables for PLS models were optimized using as 10-fold Venetian blind cross-validation procedure [36].

The 1D-CNN was implemented using the Python (3.6) language and Keras/TensorFlow (2.5.0-dev20201204) running on a workstation equipped with a NVidia GPU (GeForce RTX 2080 Ti), an Intel® Core™ i7-4770k @3.5 GHz and 64 GB RAM, running Microsoft Windows 10 OS. Given the fact that this algorithm deals with a high degree of randomness during the NN training in the GPU, different versions of TensorFlow might produce slightly different results. The chemometric analysis was performed in MATLAB 2018b, MathWorks, Natick, USA using the freely available MBA-GUI [26].

**Table 2**
A summary of total near-infrared spectra in training and test before and after outlier's removal. The data augmentation was performed after outlier removal.

| Data set | Spectra (Samples × Variables) with outliers | Spectra (Samples × Variables) without outliers and after data augmentation |
|---|---|---|
| **Train** | 10,243 × 103 | 9914 × 618 |
| **Test** | 1448 × 103 | 1413 × 618 |

## 3. Results and discussion

### 3.1. Spectral profiles and reference properties

The absorbance mean spectra of mango fruit and the same spectra pre-processed with several pre-processing techniques are shown in Fig. 2. It can be noted that the absorbance signal was masked by other spectra with high-intensity. In the SNV spectra (red dotted line), the key chemical features appear near the ~960 nm which is correlated to the 3rd overtones of O–H bonds in $H_2O$ [37]. The 1st derivative of the spectra did not reveal any extra distinct peaks but small mounds near ~850 nm and ~950 nm, which can be related to the 3rd overtones of OH, CH and NH bonds. 1st derivative estimation on the SNV data has similar peaks compared to the lone estimation of 1st derivative. The 2nd derivative on absorbance as well as SNV pre-processed data showed similar peaks at ~820, 870 and 960 nm related to OH, CH and NH bonds [37]. Such extra peaks (previously hidden) revealed by different pre-processing are expected to facilitate the DL models.

To have a model that generalizes well, the rule of thumb is that the test set should come from the same distribution as the tuning set. This optimal modelling scenario is not possible for the data used in this study as the test set is based on fruits from a different harvest season (the year 2018 harvest). It is a truly external validation set and the data distributions were not similar (see Fig. 3). For this reason, a gap is expected in the metrics (RMSEP) between tuning and test sets.

### 3.2. PLS analysis

The benefit of outlier removal and data augmentation is first demonstrated and benchmarked with PLS analysis. A summary of the results from the PLS analysis is shown in Fig. 4. The PLS analysis on the training data (absorbance spectra) with outliers and using it on test data with (Fig. 4A) and without outliers (Fig. 4B) attained the RMSEP of 1.06% and 1.01%, respectively. The PLS model made on augmented data with outliers and tested on test data with (Fig. 4E) and without outliers (Fig. 4F) attained the RMSEP of 1.03% and 0.99%. The removal of outliers in the train set prior to PLS analysis reduced the RMSEP for test data with (Fig. 4C) and without outliers (Fig. 4D), but with a small fraction. However, PLS analysis on the outlier removed augmented data reduced the RMSEP to 0.99% and 0.95% for the test set with (Fig. 4G) and without outliers (Fig. 4H), respectively. Hence, the outlier removal and data augmentation improved PLS modelling.

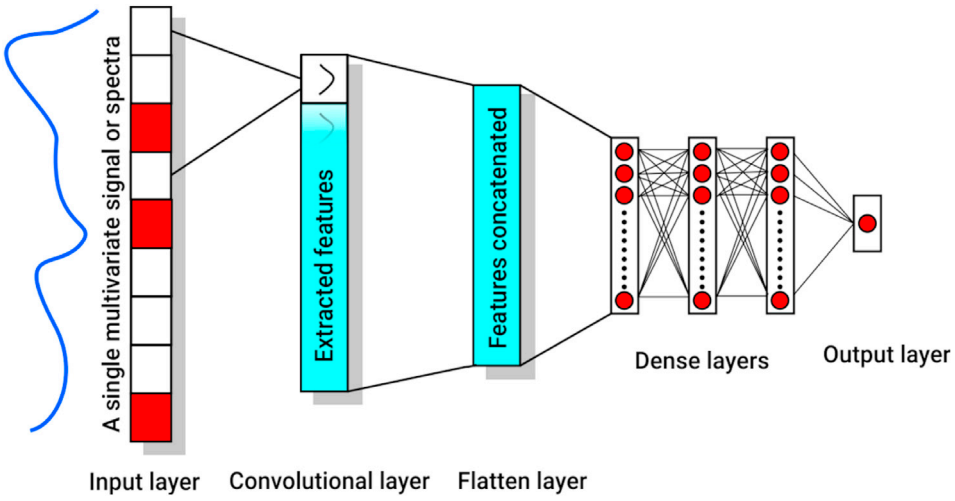


**Fig. 1.** The 1-D convolutional neural network architecture used to model the spectral (1 × n) data.

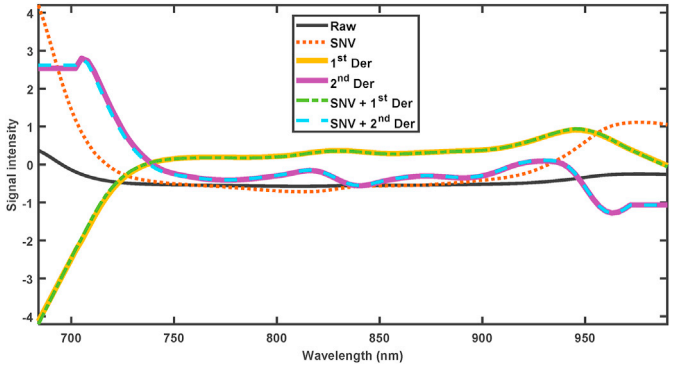


**Fig. 2.** Mean near-infrared spectral profile of mango fruit processed with different pre-processing.

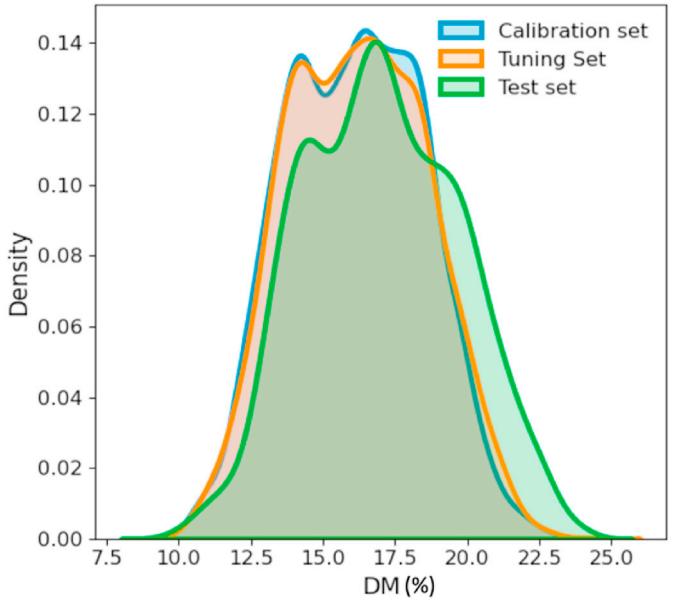


**Fig. 3.** Distributions for the dry matter (%) in the calibration, tuning and test sets. There is a clear excess of high DM content in the test set when compared to the sets used for calibration and tuning.

### 3.3. Deep learning analysis

To produce a fine-tuned model, the kernel width on the convolutional layer, the number of samples for the training batch (mini-batch) and the strength of the L2 regularization β were iteratively optimized. An initial kernel width was chosen based on previous experience with this CNN on other data sets. It was empirically found that for this specific CNN architecture, a kernel size of the same order as the window used to compute smooth Savitzky-Golay derivatives in pre-processing, is a good starting point. Given that a window of 13 points was used for the derivatives and a recent research works [17] uses 17 points, in this study an initial exploratory kernel width was set to 15 points. With this first guess fixed, an initial grid search on the two other hyper-parameters that control the model, $\beta \in [0.005, 0.007, 0.009, 0.011, 0.013, 0.015]$ and training batch size $\in$ [96, 128, 160, 192, 224], was performed. The results of the exploration are shown in Fig. 5, where the evolution of RMSE on the calibration and tuning sets as a function of batch size and L2 regularization strength (β) with respect to the convolutional filter size 15 are presented. In Fig. 5, it can be noted that the CNN overfitting decreases (red and black points come closer for each batch size) with increasing L2 regularization as expected. Considering that the optimal RMSE on the tuning set is not always the most robust solution, the value of $\beta = 0.011$ was chosen as a good compromise between decreased overfitting and tuning set RMSE stabilization.

With β set to 0.011, the model was further optimized by fine-tuning the kernel width on the interval [15,17,19,21,23,25] and batch size again. In Fig. 6, a 2D contour plot of the RMSE on the calibration and tuning sets in the hyper-parameter space is shown. With Fig. 6, two points in the hyperparameters space, namely (A) at kernel size = 19, and batch size = 160 and (B) at kernel size = 21 and batch size = 128, were chosen. The choice rational for the first point is the common RMSE minima for both calibration and tuning sets, while the second point was chosen by observing that area of the local RMSE of tuning set in this area expands when compared to the calibration set. This indicates that the model is generalising well in this corner of the hyper-parameter space.

The RMSEP's of these prototype models (corresponding to point A and B in Fig. 6A) are shown in Fig. 7. It can be noted that both models (A and B) attained higher RMSEPs compared to the calibration and tuning sets. This is because of the different distribution of the external test set where several samples had higher DM values compared to calibration and tuning sets. However, the RMSEP obtained with model A and B i.e., 0.80% and 0.79% were far lower compared to best achieved (0.84%) in the same dataset with an ensemble of several non-linear methods [19].

The two model prototypes (A and B) presented in Fig. 6A were further tested on the outlier removed external test data. The results were also

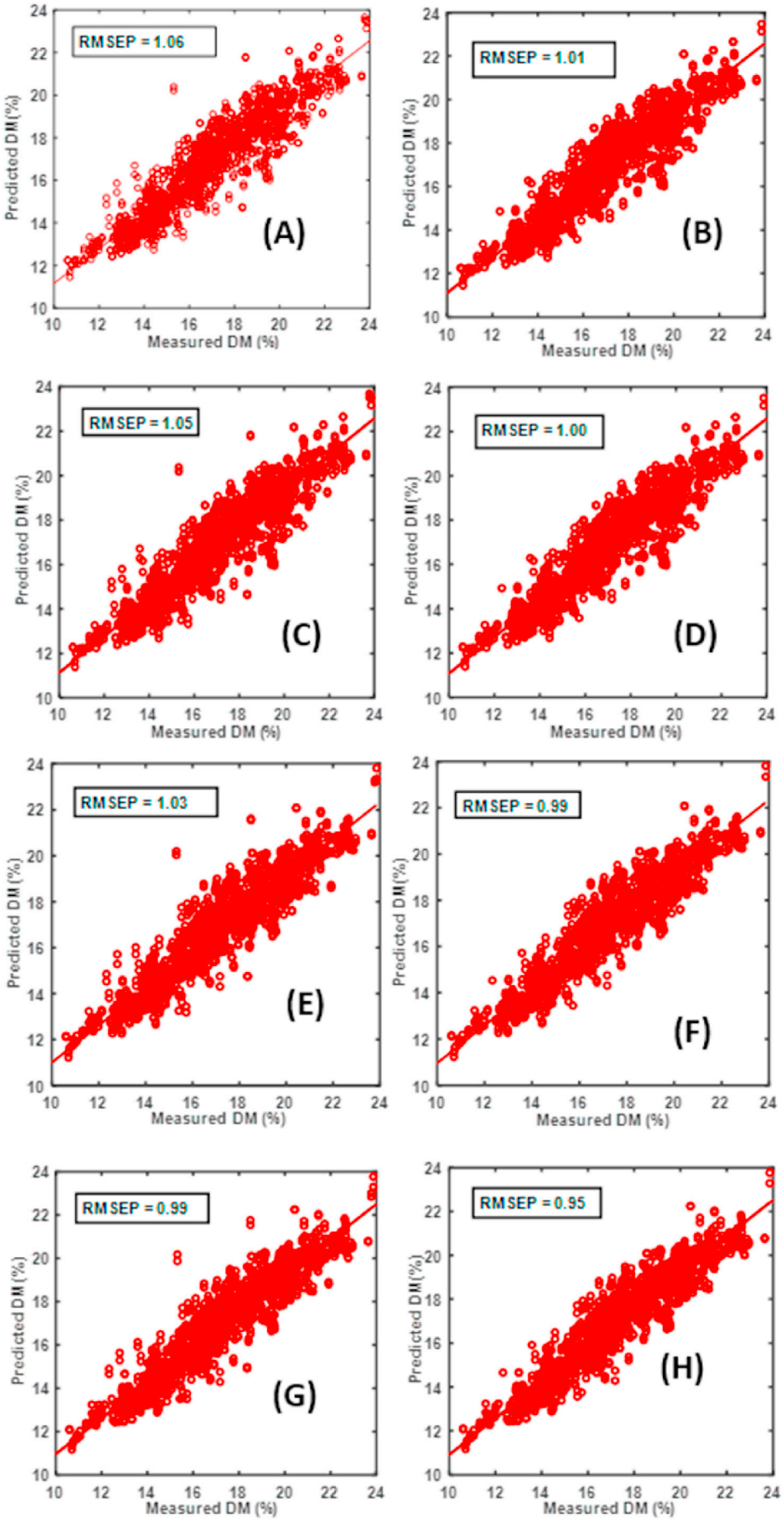


**Fig. 4.** A summary of partial least-square (PLS) models developed with absorbance and augmented data with and without outlier's removal. (A) PLS model (11 latent variables) made on absorbance data with outlier and tested on test data with outliers, (B) PLS model made on absorbance data with outlier and tested on test data without outliers, (C) PLS model (11 latent variables) made on absorbance data without outlier and tested on test data with outliers, (D) PLS model made on absorbance data without outlier and tested on test data without outliers, (E) PLS model (12 latent variables) made on augmented data with outlier and tested on test data with outliers, (F) PLS model (12 latent variables) made on augmented data with outlier and tested on test data without outliers, (G) PLS model (11 latent variables) made on augmented data without outlier and tested on test data with outliers, and (H) PLS model (11 latent variables) made on augmented data without outlier and tested on test data without outliers.

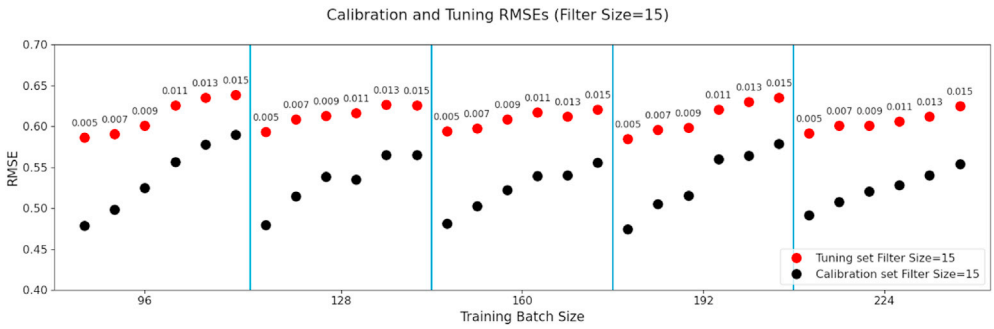


**Fig. 5.** RMSE of calibration (black) and tuning (red) sets as a function of batch size and β *for a kernel width of 15 points.* (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

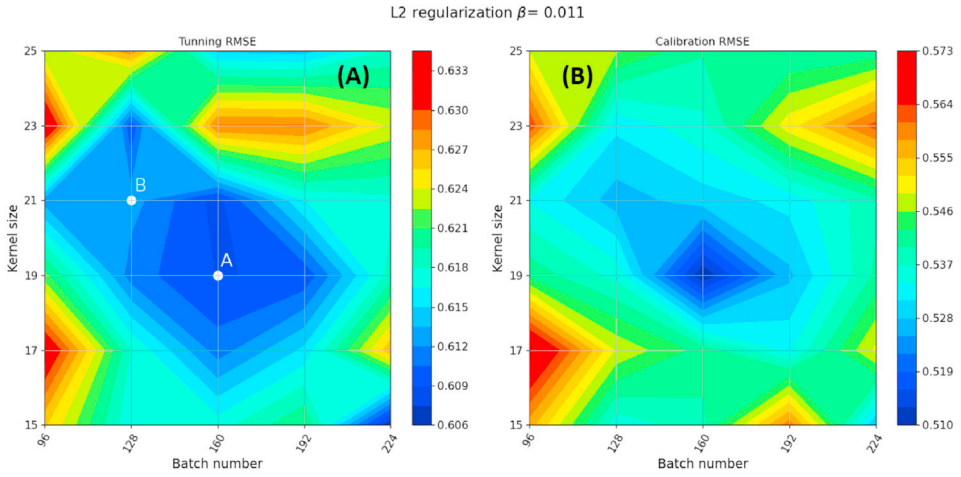


**Fig. 6.** Contour plot of RMSE of calibration (right) and tuning (left) sets as a function of hyper-parameters kernel size and batch size for β=*0.011. The colour scale represents the RMSE.* (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

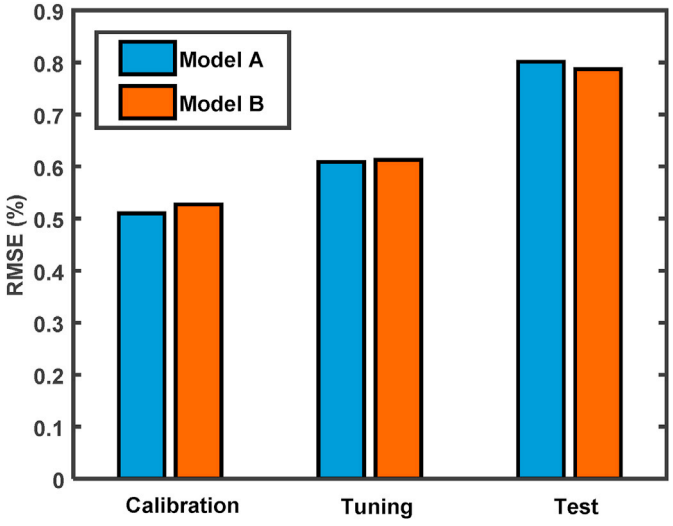


**Fig. 7.** Summary of prediction metrics for optimized CNN models A (blue) and B (orange). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

compared with outlier removed but not augmented data to demonstrate the benefit of the proposed data augmentation. Both models A (Fig. 8C) and B (Fig. 8D) based on augmented and outlier removed data showed the RMSEP of 0.75% compared to the non-augmented data.

The predictions made on the external validation test set showed the immense potential of DL modelling. The RMSE's obtained from the two prototype models, were lower than those obtained by other models [17, 19] on the same test set. However, the model optimization process at this moment involves decision making based on user experience, i.e., there is still not a quantifiable method to choose the best hyper-parameters. This is also a challenge in classical chemometric models (e.g. choice of a robust number of latent variable in PLS) that needs to be addressed in future works to avoid serendipitous hyper-parameter choices and ensure optimal solutions. With this concern in mind, a posterior analysis of the solutions space by probing the model over a larger hyper-parameter space is shown in Fig. 9. In Fig. 9, the RMSE maps for the calibration, tuning and test sets over a hyper-parameter space much broader than the one used in the model optimization phase are presented. For convenience, the values used for prototype models (A) and (B) are displayed in these 2D maps. On comparing the calibration (Fig. 9A) and tuning (Fig. 9B) RMSE maps, it can be noted that the kernel sizes ranging from 11 to 20 tend to generalize better with increased batch number (lower RMSE values). However, that same type of improvement did not extend to the test RMSE map (Fig. 9C). Furthermore, the local minima in the tuning RMSE map (Fig. 9B) do not map directly to local minima in the test RMSE map (Fig. 9C). This is a confirmation that the best solution for the tuning set does not always translate as the best solution for the test set. Nonetheless, even with this mismatch between tuning and test RMSE maps, it can be observed that most of the solutions obtained with the CNN in this hyper-parameter space were below 0.84% (green and blue shadings), with a small area covered by solutions between 0.84% and 0.88% and an even smaller area with solutions above 0.88%. This gives some assurance that even without the best optimization possible, the CNN model provides solutions better than the many linear, non-linear

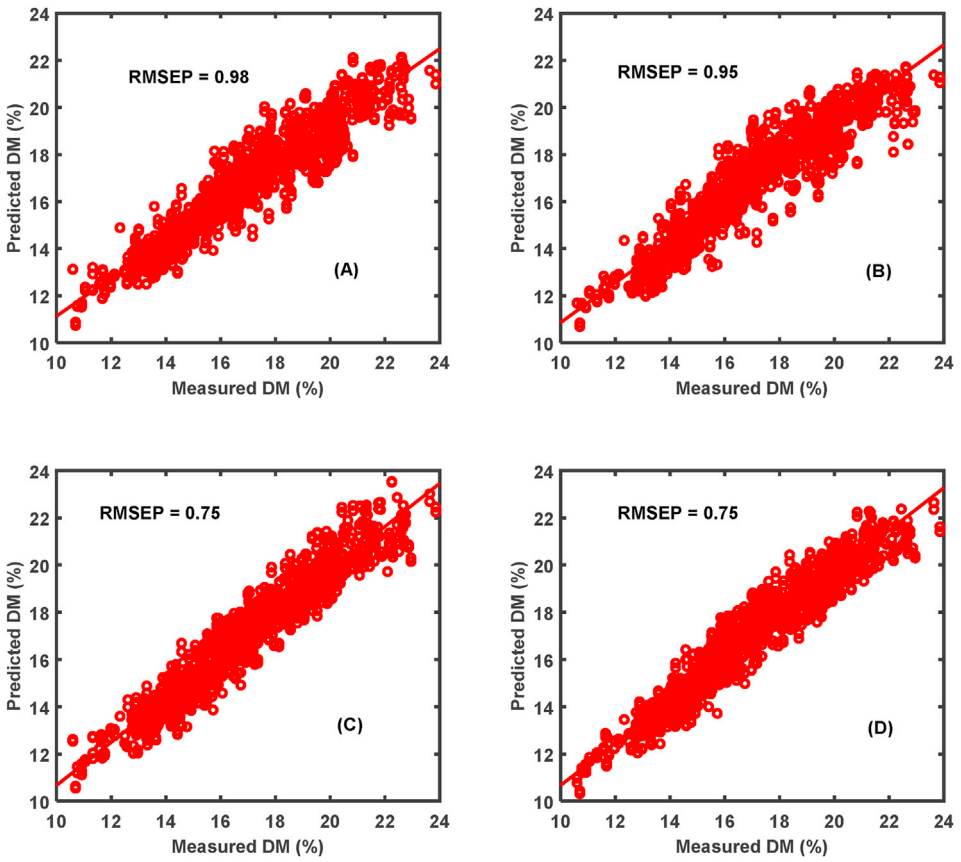


**Fig. 8.** A summary of deep learning models for absorbance and augmented data without outliers in test set. (A) DL model A made on absorbance data tested, (B) DL model B made on absorbance data test on data, (C) DL model A made on augmented data test, and (D) DL model B made on augmented data test.

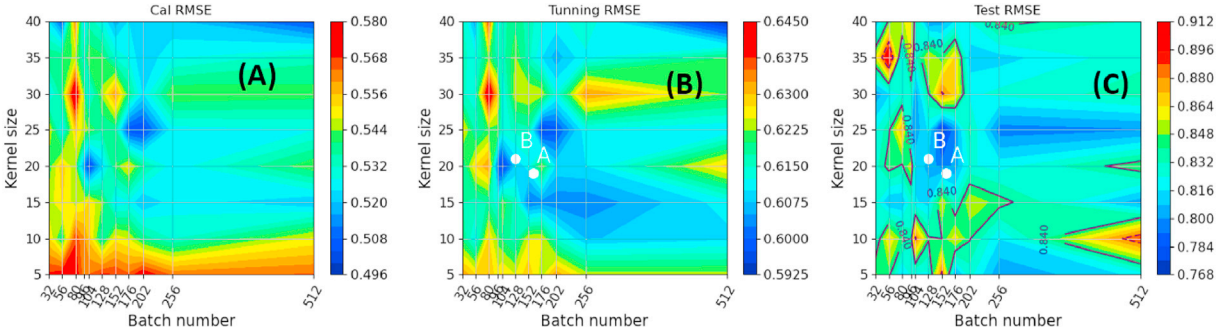


**Fig. 9.** RMSE maps for the calibration, tuning and test sets in a broader 'batch number' vs. 'kernel size' hyper-parameter space. The hyper-parameters for prototype models (A) and (B) are signalled as white dots on the tuning and test RMSE maps. In the test RMSE map, solid contour lines correspond to RMSE = 0.84% and dashed contour lines correspond to RMSE = 0.88%.

and ensemble models reported in Refs. [17,19].

Like the standard chemometrics analysis like PLS, the important spectral features in the data can also be visualized for CNN prediction. In Ref. [1], the authors a methodology, based on the numerical method (perturbation theory), to visualize something akin to a CNN regression coefficients. This is done by treating the CNN as a black-box function that maps input variable into target variables, and by comparing the behavior of the solutions based on the original data with solutions based on slightly perturbed data. This is displayed as CNN regression coefficients, $W_i$. In Fig. 10, the mean $W_i$ computed using 200 random samples and model (A) is shown. It can be noted that the larger coefficients related to the features extracted from the 1st derivative and from the SNV+2nd derivative sections. This hints that this model can lead to satisfactory (but not optimal) results using a more streamlined input vector (1st derivative and SNV+2nd derivative stacking) using the proposed augmentation approach.

## 4. Conclusions

This study showed that the combination of chemometrics approaches i.e., the removal of outliers with $T^2$ and Q statistics and augmenting data with spectral pre-processing approaches improved the accuracy of the DL models. Data augmentation also benefited the PLS regression analysis, hence, can be considered as a general tool to improve the predictive performance of calibration models. The results in this study set a new benchmark of RMSEP 0.79% for predicting DM in mango fruit with the use of mango data set. Readers are encouraged to use this big data set and produce innovative ideas and algorithms to achieve RMSEP better than 0.79%.

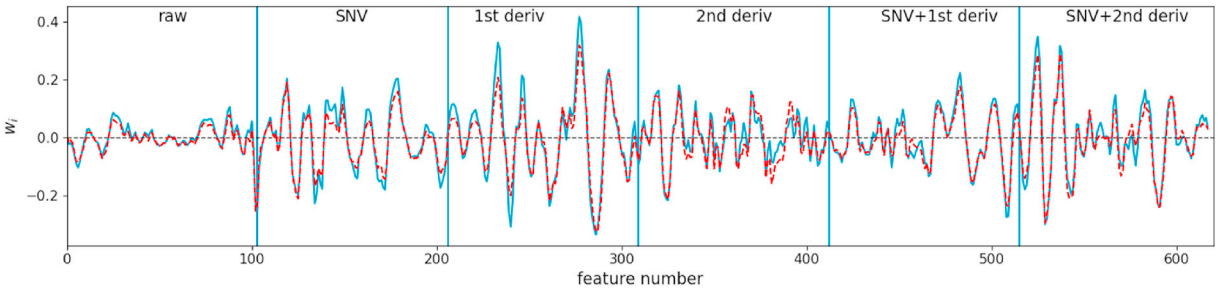


**Fig. 10.** Model (A) mean weight $w_i$ computed from 200 random samples from the tuning set (blue) and test set (red dashed). The input features are divided into sections with the corresponding pre-processing type identified. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

## Link to data

The original data related to this publication can be found at: https://data.mendeley.com/datasets/46htwnp833/1.

## Author statement

Puneet Mishra, Methodology, Conceptualization, Formal analysis, Writing – original draft, Dário Passos, Formal analysis, Investigation, 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] C. Cui, T. Fearn, Modern practical convolutional neural networks for multivariate regression: applications to NIR calibration, Chemometr. Intell. Lab. Syst. 182 (2018) 9–20.
[2] J.-E. Dong, Y. Wang, Z.-T. Zuo, Y.-Z. Wang, Deep learning for geographical discrimination of Panax notoginseng with directly near-infrared spectra image, Chemometr. Intell. Lab. Syst. 197 (2020) 103913.
[3] 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.
[4] 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.
[5] K. Balaji, K. Lavanya, A.G. Mary, Clustering of mixed datasets using deep learning algorithm, Chemometr. Intell. Lab. Syst. 204 (2020) 104123.
[6] T. Shi, S. Huang, L. Chen, Y. Heng, Z. Kuang, L. Xu, H. Mei, A molecular generative model of ADAM10 inhibitors by using GRU-based deep neural network and transfer learning, Chemometr. Intell. Lab. Syst. 205 (2020) 104122.
[7] L. Yi, J. Lu, J. Ding, C. Liu, T. Chai, Soft sensor modeling for fraction yield of crude oil based on ensemble deep learning, Chemometr. Intell. Lab. Syst. 204 (2020) 104087.
[8] A.B. Risum, R. Bro, Using deep learning to evaluate peaks in chromatographic data, Talanta 204 (2019) 255–260.
[9] R.A. Crocombe, Portable spectroscopy, Appl. Spectrosc. 72 (2018) 1701–1751.
[10] C.A.T. dos Santos, M. Lopo, R.N.M.J. Pascoa, J.A. Lopes, A review on the applications of portable near-infrared spectrometers in the agro-food industry, Appl. Spectrosc. 67 (2013) 1215–1233.
[11] P. Mishra, F. Marini, B. Brouwer, J.M. Roger, A. Biancolillo, E. Woltering, E.H.-v. Echtelt, Sequential fusion of information from two portable spectrometers for improved prediction of moisture and soluble solids content in pear fruit, Talanta 223 (2021) 121733.
[12] P. Mishra, E. Woltering, N. El Harchioui, Improved prediction of 'Kent' mango firmness during ripening by near-infrared spectroscopy supported by interval partial least square regression, Infrared Phys. Technol. 110 (2020) 103459.
[13] P.P. Subedi, K.B. Walsh, Assessment of avocado fruit dry matter content using portable near infrared spectroscopy: method and instrumentation optimisation, Postharvest Biol. Technol. (2020) 161.
[14] Y. Huang, R. Lu, K. Chen, Prediction of firmness parameters of tomatoes by portable visible and near-infrared spectroscopy, J. Food Eng. 222 (2018) 185–198.
[15] M. Li, Z.Q. Qian, B.W. Shi, J. Medlicott, A. East, Evaluating the performance of a consumer scale SCiO (TM) molecular sensor to predict quality of horticultural products, Postharvest Biol. Technol. 145 (2018) 183–192.
[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] 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.
[18] 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.
[19] N.T. Anderson, K.B. Walsh, J.R. Flynn, J.P. Walsh, Achieving robustness across season, location and cultivar for a NIRS model for intact mango fruit dry matter content. II. Local PLS and nonlinear models, Postharvest Biol. Technol. 171 (2021) 111358.
[20] 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.
[21] 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.
[22] 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.
[23] 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.
[24] N. Anderson, K. Walsh, P. Subedi, Mango DMC and spectra Anderson et al, Mendley, Mendley data, 2020, 2020.
[25] S. Wold, M. Sjostrom, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemometr. Intell. Lab. Syst. 58 (2001) 109–130.
[26] 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.
[27] 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.
[28] 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, p. 121693.
[29] 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.
[30] P. Mishra, J.M. Roger, D. Jouan-Rimbaud-Bouveresse, A. Biancolillo, F. Marini, A. Nordon, D.N. Rutledge, Recent trends in multi-block data analysis in chemometrics for multi-source data integration, Trac. Trends Anal. Chem. (2021) 116206.
[31] P. Mishra, S. Lohumi, Improved prediction of protein content in wheat kernels with a fusion of scatter correction methods in NIR data modelling, Biosyst. Eng. 203 (2021) 93–97.
[32] P. Mishra, T. Verkleij, R. Klont, Improved prediction of minced pork meat chemical properties with near-infrared spectroscopy by a fusion of scatter-correction techniques, Infrared Phys. Technol. 113 (2021) 103643.
[33] 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.
[34] A. Savitzky, M.J.E. Golay, Smoothing and differentiation of data by simplified least squares procedures, Anal. Chem. 36 (1964) 1627–1639.
[35] S. Wold, PLS Modeling with Latent Variables in Two or More Dimensions, 1987.
[36] F. Westad, F. Marini, Validation of chemometric models – a tutorial, Anal. Chim. Acta 893 (2015) 14–24.
[37] B.G. Osborne, Near-Infrared Spectroscopy in Food Analysis, Encyclopedia of Analytical Chemistry, 2006.