Analytica Chimica Acta 1163 (2021) 338520

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

Deep multiblock predictive modelling using parallel input convolutional neural networks

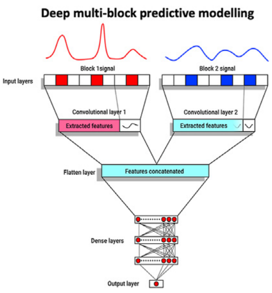
Puneet Mishra [a, *], Dário Passos [b]

[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, 8005-189, Faro, Portugal

HIGHLIGHTS

- A deep multi-block data modelling method is presented.
- An example of fusing two blocks, i.e., visible, and near-infrared data, is demonstrated.
- Proposed method outperformed classical multi-block chemometric modelling.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 1 February 2021
Received in revised form
27 March 2021
Accepted 13 April 2021
Available online 16 April 2021

Keywords:
Data fusion
Artificial intelligence
Spectroscopy
Chemistry

ABSTRACT

In the domain of chemometrics, multiblock data analysis is widely performed for exploring or fusing data from multiple sources. Commonly used methods for multiblock predictive analysis are the extensions of latent space modelling approaches. However, recently, deep learning (DL) approaches such as convolutional neural networks (CNNs) have outperformed the single block traditional latent space modelling chemometric approaches such as partial least-square (PLS) regression. The CNNs based DL modelling can also be performed to simultaneously deal with the multiblock data but was never explored until this study. Hence, this study for the first time presents the concept of parallel input CNNs based DL modelling for multiblock predictive chemometric analysis. The parallel input CNNs based DL modelling utilizes individual convolutional layers for each data block to extract key features that are later combined and passed to a regression module composed of fully connected layers. The method was tested on a real visible and near-infrared (Vis-NIR) large data set related to dry matter prediction in mango fruit. To have the multiblock data, the visible (Vis) and near-infrared (NIR) parts were treated as two separate blocks. The performance of the parallel input CNN was compared with the traditional single block CNNs based DL modelling, as well as with a commonly used multiblock chemometric approach called sequentially orthogonalized partial least-square (SO-PLS) regression. The results showed that the proposed parallel input CNNs based deep multiblock analysis outperformed the single block CNNs based DL modelling and the SO-PLS regression analysis. The root means squared errors of prediction obtained with deep multiblock analysis was 0.818%, relatively lower by 4 and 20% than single block CNNs and SO-PLS regression, respectively. Furthermore, the deep multiblock approach attained ~3% lower RMSE compared to the best known on the mango data set used for this study. The deep multiblock analysis approach based on parallel input CNNs could be considered as a useful tool for fusing data from multiple sources.

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

https://doi.org/10.1016/j.aca.2021.338520

1. Introduction

Data from multiple sources is widely encountered in the chemometrics domain [1,2]. For example, measurements performed on a single sample with multiple spectroscopic sensors [3,4], data measured on multiple batches [5], and same data pre-processed with several pre-processing techniques [6–8]. The traditional single block latent variables based chemometric approaches such as principal component analysis (PCA) [9] and partial least-square regression (PLS) [10,11] analysis are widely used but they are not the optimal analysis solution when comes to multiblock data [2]. Single-block chemometric techniques are inefficient in jointly modelling data from multiple sources and particularly when the data is in different scales [12–14]. Hence to deal with multi-source data, especial techniques called multiblock data analysis techniques exist in the domain of chemometrics [1,2,5,12–18].

Multiblock data analysis techniques exist for both data exploration [2,12,13] and predictive modelling [2,15–17]. Furthermore, novel applications of multiblock analysis even allow tasks such as calibration transfer [19] and optimal pre-processing selection and fusion [7,8]. The main aim of multiblock analysis for data exploration is to allow enhanced visualization of hidden patterns which is otherwise unachievable with the analysis made on data from a single source [5,12,13,20], while for predictive modelling the main aim of multiblock methods is to achieve a precise prediction of the property of interest by combining information from multiple sources [16,17,21]. Commonly used methods for multiblock predictive analysis are the extensions of latent space modelling approaches such as PLS regression [10,11] and principal covariates regression [22]. For example, some akin methods are multiblock partial least-square regression [23], response optimized sequential alteration [24], sequential orthogonalized partial least-square regression and parallel orthogonalized partial least-square regression [16,17]. Additionally, several feature selection methods are available such as the sparse covariate regression [25] and sequential orthogonalized covariate selection [21] that, while maintaining the predictive accuracy of models, allows extracting key hidden features from the multi-source data.

Recently, in the domain of chemometrics, a huge interest is emerging related to the use of deep learning frameworks for modelling multivariate signals such as spectral data [26–29]. Approaches such as convolutional neural networks (CNNs) [27,28] and autoencoders [30,31] are being increasingly applied to the field and have, in several occasions, shown to outperformed classical approaches such as PLS regression in terms of achieving high accuracy models. However, most of the CNNs DL models that currently exist in the chemometric literature are limited to deal with a single block of data. The CNNs based DL modelling can also be engineered to simultaneously deal with multiblock data. Multiblock modelling with CNNs can be performed by implementing a neural network architecture with several parallel convolutional layers blocks whose receptive fields are data coming from different sources. Each block then extracts complementary information from the different input data sources separately. The parallel conv. layers blocks are required as the data from each source may have different features, for example, if the data comes from near-infrared (NIR) spectroscopy and mass-spectroscopy (MS) domains, where the NIR data has broad spectral peaks while the MS data has sharp peaks, then, using the same type of convolutional filters on two such data may not be an ideal solution. In such a case, the size of the convolutional filters should be optimized individually to benefit the most from the complementary information present in the multi-source data. After the input parallel conv. layers block, extracted features can be combined using pooling layers or simply concatenated and passed through to a dense (a.k.a fully connected, FC) layers block to attain a suitable mapping with the property of interest. To the best of our chemometrics literature search [1,2,18], an implementation of parallel input CNNs for multiblock predictive modelling has never been done and this work is the first to implement and demonstrate its potential on a real-life spectroscopy data set.

The objective of this study is to implement the concept of parallel input CNNs based DL modelling for multiblock predictive analysis. The method was tested on a real, large Vis-NIR data set related to dry matter prediction in mango fruit. The multiblock data was crafted by treating the visible and near-infrared spectral bands of the original data set as two separate blocks. The performance of the parallel CNNs is compared with the traditional single block CNNs based DL modelling [27] and with the commonly used multiblock chemometric approach called sequentially orthogonalized partial least-squares (SO-PLS) regression [17]. Finally, the results from the deep multiblock modelling approach were compared with the best reported results on the mango data set [32] using only the NIR part of the spectra [33,34].

2. Materials and method

2.1. Data set

The data set used in this study was a visible and near-infrared spectroscopy data related to dry matter (DM) prediction in mango fruit (publicly available at [32]). The data in total have 11,691 Vis-NIR spectra (350–1200 nm in 3 nm sampling) and reference DM measurements performed on mango fruit across 4 harvest seasons 2015–2018. According to the description of the data [33,34], the spectral measurements were performed with F750 Produce Quality Meter (Felix Instruments, Camas, USA), while DM (%) was measured with oven drying (UltraFD1000, Ezidri, Beverley, Australia). The spectra, at the source repository, comes pre-partitioned into training and test sets, in order to be able to make a fair comparison with the results previously reported on the data set [33,34]. Out of 11,691 spectra, 10,243 training spectra are from the first three harvest seasons (2015–2017), while the remaining 1448 spectra (from 2018) are the independent test set. The original spectral range 350–1200 nm presented extreme noise at the beginning and end of the spectra. These noisy sections were removed, and the spectral range was reduced to (450–1030 nm). The original training set was pre-filtered using the Hotelling's T^2 or Q statistics obtained with PLS data decomposition that removed several outliers. A key point to note is that the outliers from the test set were not removed to enable a fair comparison of results with previous studies performed on the same data set [33,34]. The final training set, after outlier removal, comprised of 9914 samples. The training set was further partition into calibration (66.66% of training set) and validation set (33.33% of training set) for model training. To support the development of this multiblock analysis technique, the original spectra were partitioned into two blocks, visible (450–697 nm) and near-infrared (700–1030 nm) bands. Such a partition was performed as the fruit outer colour (450–697 nm) and the chemical absorption (700–1030 nm) related to OH and CH bond overtones are highly correlated to fruit properties such as DM. In previous works though, authors have only used mostly the NIR part of the spectra [33,34], but in this study the aim was to demonstrate how combining Vis and NIR information with deep multiblock analysis can provide a better accuracy model compared to the model developed on only NIR data.

2.2. Parallel CNNs based deep learning

The DL model architecture used in this study was an extension of the 1-dimensional convolutional neural network (1D-CNN) architecture presented in [27] and implemented in [35–37]. The 1D-CNN architecture used in [27] was introduced to deal with a single block of data. For multiblock analysis, the 1D-CNN architecture must be modified to accept multiple sources of data. An intuitive solution to solve this problem is to implement an architecture with input parallel layers that can simultaneously extract the features from multiple sources of data. Hence, the solution proposed in this study is to use parallel convolution layers to process different data types that were later concatenated and flatten before being fed to a dense layers block. A summary of the architecture proposed in this study for deep multiblock analysis is shown in Fig. 1. For an easier explanation of the concept, we opted to implement a simple CNN architecture with only two receptive fields composed by 2 convolution layers with 1 filter each and stride = 1 followed by a flatten layer that is connected to 3 fully FC layers with 36, 18 and 12 neurons, respectively and a final output layer with one neuron. The number of units (or neurons) in the FC layers follows the prescription of the original architecture in [25]. After each layer, the data flows through an exponential linear unit (eLU) activation function, except for the last output layer where a linear activation function was used. 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). We rely on the Adaptive moment estimation (Adam [32]) optimizer with the back-propagation algorithm to train the model weights. Adam was initiated with an initial learning rate (LR) given by 0.01 × (batch size)/256 and to increase the chances of convergence toward a global minimum, the LR was iteratively decreased by a factor of 2 when the validation loss wasn't improved by 10^5 after 25 epochs (using the tf.keras.ReduceLROnPlateau() in function). The maximum number of epochs allocated for the training was 700 but that value was almost never reached due to the use of the Early Stopping technique (tf.keras.EarlyStopping() function). This technique helps avoid overfitting by stopping the training process if the validation metrics don't improve after a certain consecutive number of epochs.

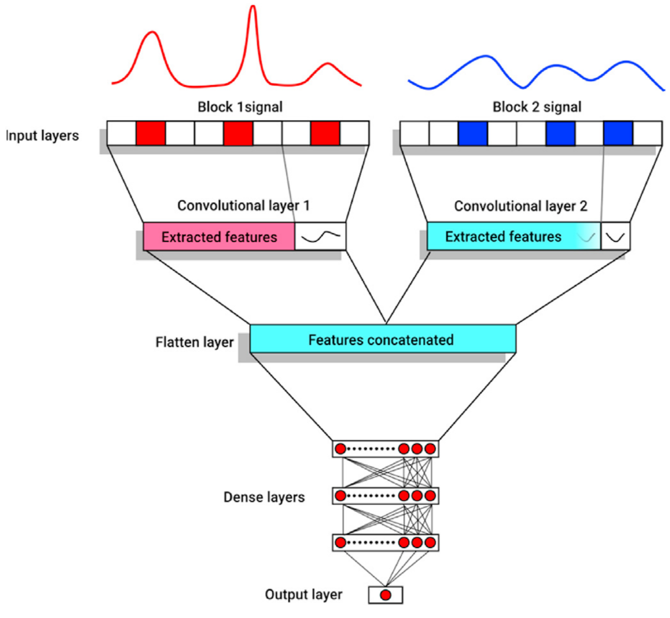


Fig. 1. A summary of the parallel CNNs architecture used for deep multiblock predictive modelling.

Since the main objective of this work is the introduction of a proof-of-concept architecture, for the sake of simplicity we chose to optimize only a limited number of models hyperparameters, i.e., the filter sizes for both conv. blocks 1 and 2, the size of the training mini-batch and the strength of the L2 regularization, β. A grid search was implemented that probed 1260 models with "filter sizes" 1 and 2 in the interval [5,10,15,20,25,30], batch size in [32, 64, 128, 256, 512] and β in [0.001, 0.003, 0.008, 0.01, 0.015, 0.02, 0.03]. The training set was further partitioned into calibration (66.66%) and validation (33.33%) using the 'test_train_split' function from sklearn (https://scikit-learn.org/stable/).

The optimisation strategy used first checks the effect of different β on the minima root mean squared errors (RMSE) of calibration and validation sets and identifies an "optimal" β as the one achieving lowest difference between the calibration and validation sets RMSE. A low difference in these RMSEs was a signal of less overfitting of the model on the training set. The optimal batch size was chosen based on the same criteria. Once the optimal β and batch size were set, the kernel sizes for each block were identified by searching for common minima in the calibration and the validation in filter-1-size vs filter-2-size RMSE contour plots. The models with optimal hyperparameters were used for predicting the test set. The parallel input CNN was implemented using the Python (3.6) language and the deep learning framework Tensorflow (2.4) with the tf.Keras API 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. The chemometric analysis related to outlier removal was performed in MATLAB 2018b, MathWorks, Natick, USA using the freely available MBA-GUI [1].

2.3. Benchmark analysis

The performance of the deep multiblock analysis based on parallel input CNN modelling was compared with two benchmark models. The first was the single block CNNs presented in Ref. [27], where the two data blocks were concatenated (to their original form) in the variable domain to make it a single block data [27]. This single block CNN was also optimized using a grid search approach for filter size in the convolutional layer, batch size and β over the same hyperparameter intervals previously defined in section 3.2. Optimal values were chosen based on the same criteria as presented in 3.2 with the difference that the final step involves a search for minima in "batch size" vs "filter size" contour plot.

The multiblock CNN analysis method was also compared in terms of accuracy with a popular multiblock predictive modelling technique called sequential orthogonalized partial least-square (SO-PLS) regression [17]. The SO-PLS at first builds a PLS regression with the first block of data to extract the scores related to the property of interest. Later, the scores were used to orthogonalize the data matrix from the second block and the response variable to remove the already explained part of the property of interest. The orthogonalized second block data was later used to build a new PLS model. At last, all the scores from the two different blocks were concatenated and used to build the final model. The SO-PLS regression was implemented with the freely available codes from MBA-GUI [1]. A key parameter to optimize in the SO-PLS was the number of latent variables (LVs) for each data block. The used approach was to try all possible combinations of LVs from all blocks and later choose the one carrying the lowest error [3,8]. However, in this work, to achieve a faster optimisation of the number of LVs, a sequential optimisation was performed. In sequential optimisation, at first, the total number of LVs for the first block were identified by increasing the LVs from 1 to 40 and monitoring the performance of the model on the validation set. The optimal number of LVs for the first block was selected as the elbow point in the error plot. The scores of the first block were then used to orthogonalize the second block and the property on interest. Later, the optimal number of LVs for the 2nd block was found by varying the LVs from 1 to 40 and monitoring the performance of the model on the orthogonalized validation set. Once again, the optimal LVs for the 2nd block was selected as the elbow point of the error plot. Finally, the multiblock SO-PLS model with optimal LVs was built and tested on the independent test set. In all cases, the performance of the models was judged based on the RMSE.

3. Results and discussion

3.1. Spectra and reference data

The mean spectra for two blocks i.e., Vis and NIR for mango fruit are shown in Fig. 2. Further, the reference dried matter distributions for calibration (red), validation (blue) and test (green) sets are shown in Fig. 2C. In the Vis spectra (Fig. 2A), some key peaks at 500 nm and 670 nm can be noted. These peaks are related to the colour of the outer peel which can range from green to yellow to red depending on the fruit cultivar and the maturity stage. During ripening the green colour of the outer peel changes toward red tones due to chlorophyll degradation. Hence, indirectly, the colour of the outer peel correlates with the maturity stage of the fruit, and thus, also to the DM (%) in the fruit. In the NIR spectra (Fig. 2B), the main peak at 960 nm related to the 3rd overtone of the OH bond related to H_2O can be noted. The overtone related to the OH is due to the high moisture in the fruit and is inversely related to the dried mater in the fruit (dried matter = 1 - moisture). With the distribution of reference DM (Fig. 2C), it can be noted that DM range for the test set was higher compared to the training and validation set. The data in the Vis range has different width peak such as the peak near 670 nm (Fig. 2A) is much thinner compared to the broad peak at 960 nm (Fig. 2B). Hence, in practical term for this study case, utilising the same convolutional filter size for Vis and NIR data may not be an optimal solution and exploration toward optimal convolutional filter size was required for different data blocks.

3.2. Benchmark single block deep learning and sequential orthogonalized partial least-square regression analysis

At first, the results of the benchmark analysis are presented. Fig. 3 shows the evolution of the lowest RMSE for both calibration and validation set for different β. As the L2 regularization strength increases, both RMSEs increase, but at the same time, they get closer to each other. Around $\beta = 0.01$ the difference between the RMSE of the calibration and validation was minimal and it was considered a good compromise point between model performance (low validation RMSE) and low overfitting (the smaller difference between RMSEs). For $\beta < 0.01$, the validation RMSE was lower but the higher difference to calibration RMSE indicates that the model was overfitting more, hence, losing its capacity of generalizing well when applied to the test set.

After choosing $\beta = 0.01$, the optimal filter and batch sizes were identified by identifying common minima for the RMSE on calibration (Fig. 4A) and validation (Fig. 4B) set. Kernel filter size = 25 and batch size = 64 were identified as optimal, as highlighted in Fig. 4B. The model based on these optimal parameters was tested on the independent test set and RMSE = 0.855% was obtained (Fig. 5).

The results of the benchmark SO-PLS modelling are shown in Fig. 6. It can be noted that the SO-PLS identified 15 (Figs. 6A) and 8 (Fig. 6B) LVs in the Vis and NIR data blocks, respectively. Finally, the model based on the optimal LVs was tested on the independent test set and RMSE = 1.03% was reached. The performance of the SO-PLS was poorer compared to the single block CNN modelling performed on the concatenated data. However, the performance of the SO-PLS was better compared to the single block PLS analysis (on NIR data) presented on the same data set in earlier studies [33,34,36]. Hence, the SO-PLS analysis demonstrates that combining the Vis information with the NIR could improve the model performance.

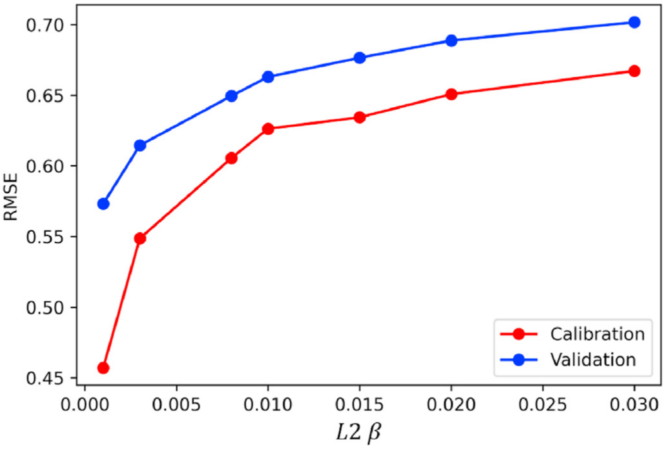


Fig. 3. An evolution of root mean squared error for calibration and validation set with increasing *L*2 *β*.

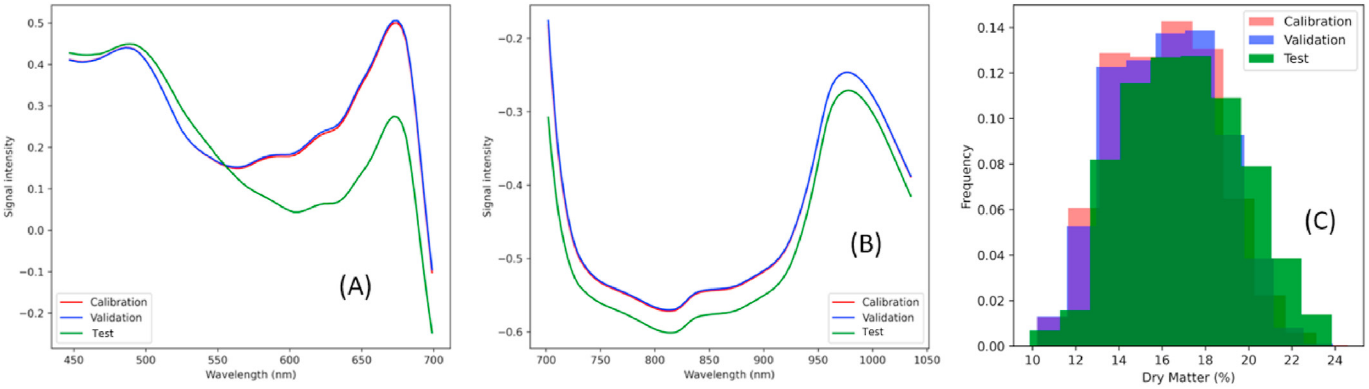


Fig. 2. Mean spectra mango for calibration, validation and test set. (A) Visible, and (B) near-infrared. (C) The histogram of reference dry matter (DM %) for calibration, validation and test set.

3.3. Deep multiblock analysis

A comparison between the models tested so far shows that the single block CNN (on concatenated data) performed better than the SO-PLS analysis. The better performance of the single block CNN model could be due to the non-linearity captured by CNN compared to the linear character of the SO-PLS model. The non-linearity could be an effect of both scattering and absorption features present in the spectra of the fruit [38,39]. Scattering effects occur mainly due to the changes in refractive indexes between cell walls and organelles and depend on the physical structure of fruit flesh and peel [38,40]. On the other hand, absorption is due to the interaction between light and the chemical components that constitute the fruit [40]. One limitation in the single block CNN model was that it forced the same convolutional filter size for the Vis and NIR data. In Fig. 2, the peaks for the NIR data were broader compared to the Vis spectra, hence, an optimal solution may only be found if we allow for different filter sizes for different blocks. The alternative scenario of using multiple filters in the conv. layer does not guaranty that each filter will be optimized to specific parts of the full block of data. The solution then is to use different filters sizes in parallel conv. layers, hence creating a deep multiblock CNN. This ensures that each filter size is optimized for the type of data input to that block.

In Fig. 7, the evolution of RMSE for two values of the L2 layer regularization strengths β, on the calibration and the validation set for varying batch size is shown. In this study, β, kernel width, and batch size were explored jointly, and it was found that β had the most effect on the overfitting of models and in lowering the validation set RMSE, hence, we at first selected the optimal β and later kernel width and batch size. For figure clarity, only two β are presented as other β showed higher RMSE differences. The difference between the RMSE of calibration and validation sets was high for $\beta = 0.001$, but smaller for $\beta = 0.003$. Such a higher difference for $\beta = 0.001$ indicates that the model was overfitting more, hence, the optimal β was chosen as $\beta = 0.003$. Furthermore, the increased batch size showed a slight increase in the difference between the RMSE of the calibration and validation set, hence, batch size = 32 was chosen as the optimal in this case. With $\beta = 0.003$and batch size = 32, the filters widths for Vis and NIR data blocks were explored and the summary is shown in Fig. 8. Based on the criteria of common minima between the calibration and validation RMSE, three filter widths combinations for Vis and NIR i.e. (5, 30), (15, 25), (25, 15) were identified as marked in Fig. 8B. The three models based on $\beta = 0.003$ and batch size = 32, and the three filter width combinations were further tested on the independent test set and the results are shown in Fig. 9. The lowest RMSE i.e. 0.818% was attained with the model made with β = 0.003, batch size = 32 and the filter width for visible = 5 and near-infrared = 30. The RMSE attained with the deep multiblock modelling was the lowest compared to both the benchmark approaches presented in earlier sections. The RMSE = 0.818% was also lower compared to the best known RMSE = 0.84% obtained in previous works using only the NIR data [33,34]. However, a key point to note is that, in total, three models were identified according to the criteria used in this study i.e., common minima in RMSE for calibration and validation set. Out of these three, only one model performed better than the RMSE = 0.84% obtained in previous works [33,34]. The other two models performed slightly poorer. Hence, the question remains related to automatically identifying a single model that can be used in practice. The other solution could be the use of an ensemble of three models and average out the output from three models for increased robustness.

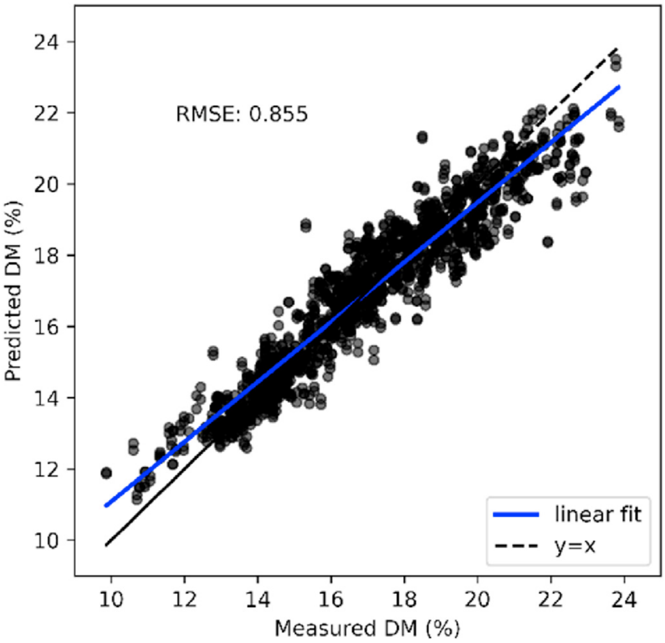


Fig. 5. Prediction for test set based on the optimal deep learning model for the single block analysis case.

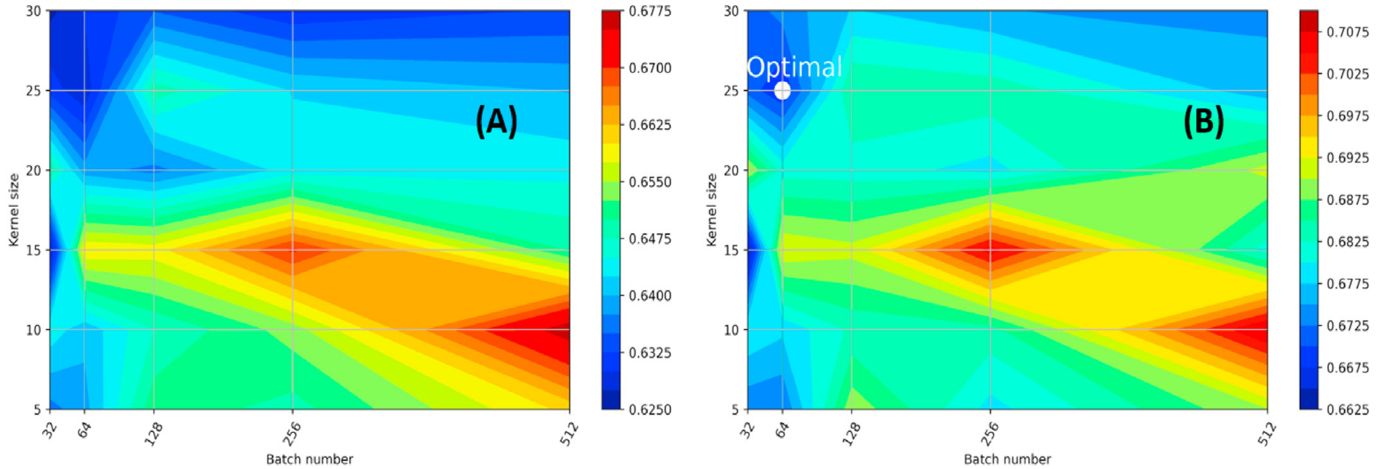


Fig. 4. A summary of root mean squared error (RMSE) obtained from the grid search for optimal filter (or kernel) and batch sizes. (A) Calibration set, and (B) validation set. The optimal hyperparameters correspond to common minima in both maps and are marked as “Optimal” in (B).

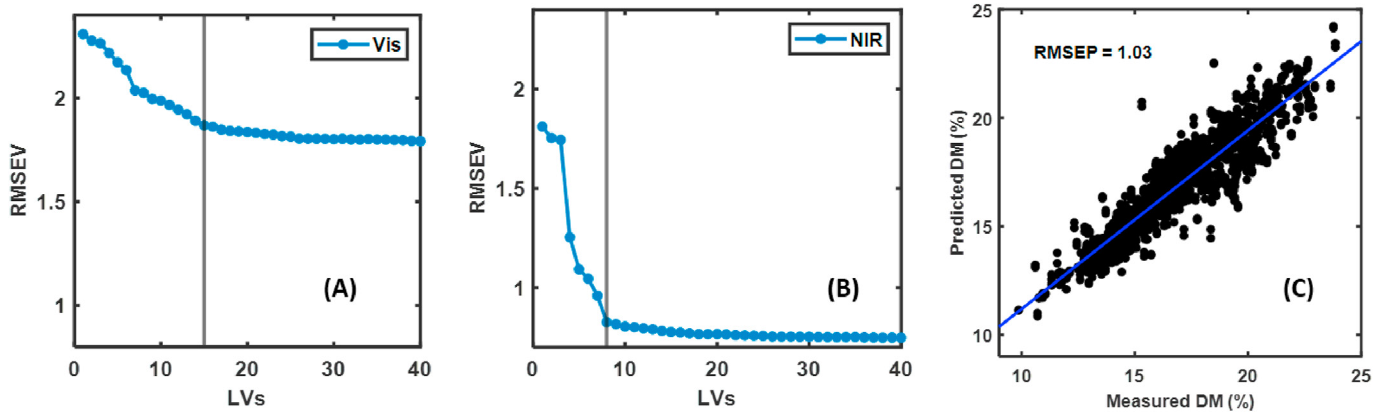


Fig. 6. A summary of SO-PLS model. (A) Latent variable modelled form the visible data block, (B) complementary latent variables extracted from near-infrared data block, and (C) performance of model on test set. The vertical lines in (A, B) shows the optimal latent variables extracted from visible and near-infrared spectral data.

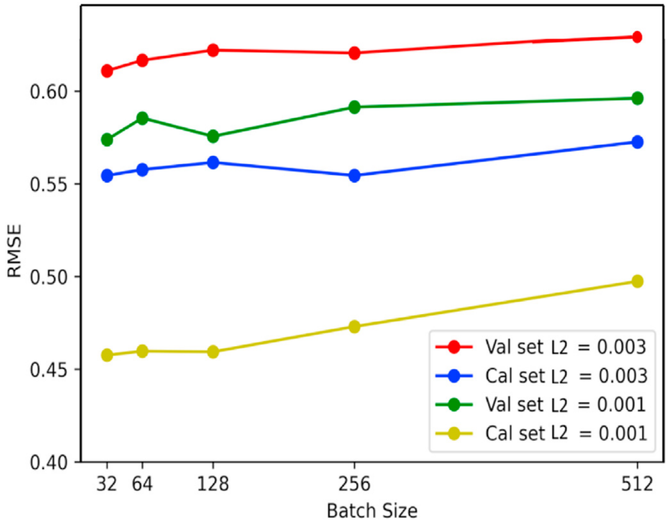


Fig. 7. An evolution of root mean squared error for calibration and validation set with increasing batch size for two different β'si.e. 0.001 and 0.003.

3.4. Activation weights for single block and multiblock CNN

In this study, the parallel CNNs based approach to multiblock modelling performed better than the single block CNNs modelling (performed on the concatenated data). The main difference between the parallel and the single block analysis was the block specific CNNs having different filter width for the convolutional layers, while the single block used a single filter width for both data blocks. For parallel input CNNs the optimal convolutional filters widths were 5 and 30 for Vis and NIR data blocks, respectively, while for single block CNNs the optimal width was 25. The effect of implementing a global filter (Fig. 10A) and block specific (Fig. 10B) filter on the mean activation response of the conv. layer is shown in Fig. 10. The main key feature to note is that the block specific convolution allowed to extract local features more accurately, for example, the peak at 960 nm related to the OH bond overtone [41] received higher mean activation weights for the block specific convolutional (Fig. 10B) compared to global convolution (Fig. 10A). Such enhanced features captured by the block specific convolution could be the cause of the better performance of parallel input CNNs based multiblock analysis compared to single-block based CNNs. In Fig. 10, it can also be noted that the mean activation weights near the edges of spectra also received higher activation weights, however, it is assumed that these were just numerical artefacts and were later compensated by the dense layers of the models. In the visible part of the CNN weights (Fig. 10B), a peak at 500 nm can be identified. The peak at 500 nm can be directly related to the green colour and which is often dominant in peel of mango fruit and changes as the fruit ripeness. Although, the change in fruit peel is dependent on cultivar and also get influenced by the biological variability of fruits.

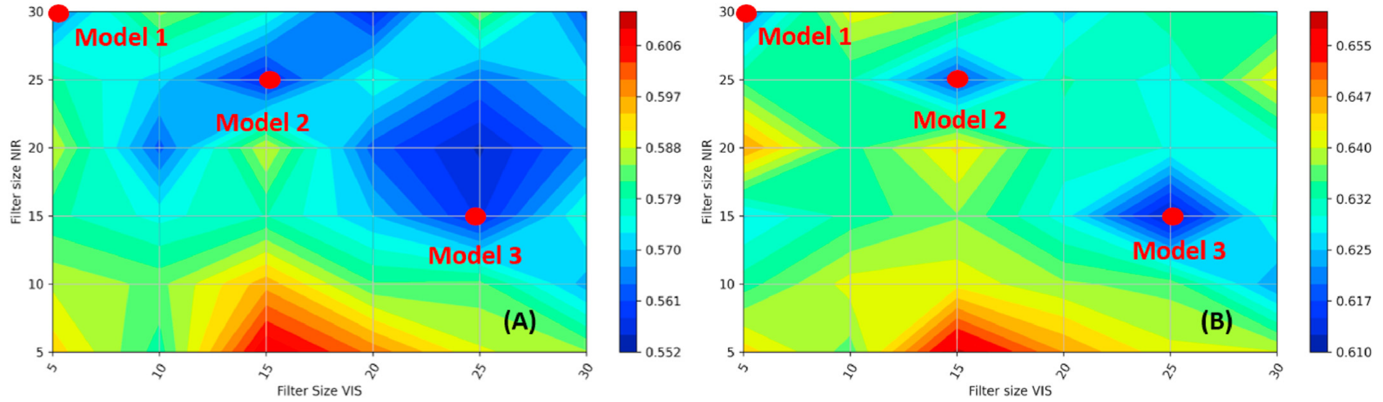


Fig. 8. A summary of root mean squared error (RMSE) obtained from the grid search for optimal kernel size for visible and near-infrared data (A) calibration set, and (B) validation set. Three sample models (Model 1, 2 and 3) were selected showing minima in both calibration and validation sets.

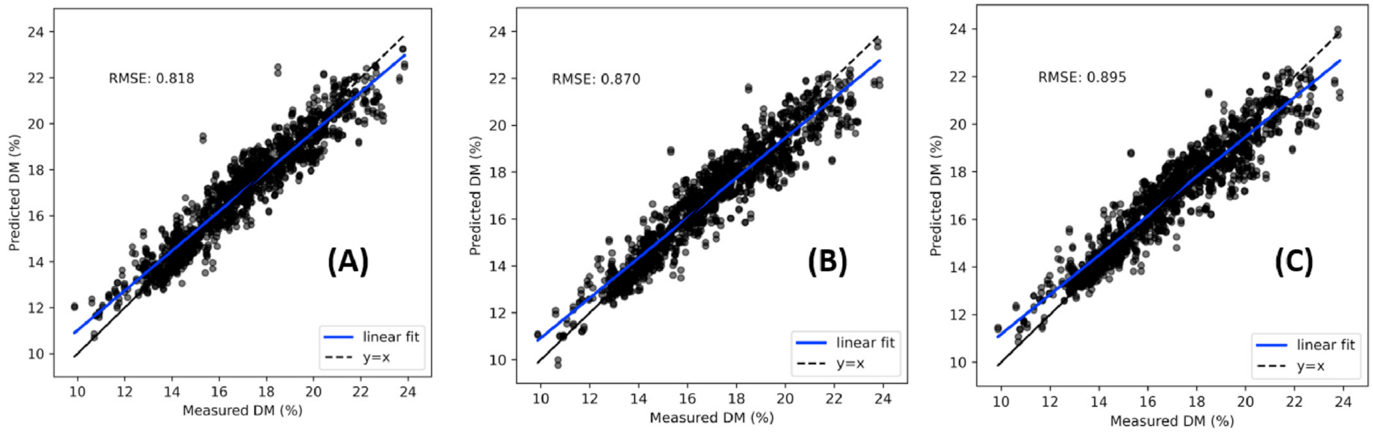


Fig. 9. A summary of performance of three multiblock models (A) Model 1, (B) Model 2, and (C) Model 3.

4. Conclusions

This study for the first time presented a new multiblock predictive modelling approach based on parallel input convolutional neural networks (CNNs). The method was also compared with a single block CNN and a popular chemometric technique called SO-PLS. The proposed parallel input CNNs based approach outperformed both the comparable approaches by attaining lower RMSE = 0.818%. Furthermore, the RMSE attained with parallel CNNs was even lower than the best-reported RMSE = 0.84% on the same data set by other authors using ensemble techniques [33]. On the optimisation front, there are also some challenges that need to be overcome and the process still requires further refinement in order to increase the confidence that optimal models are produced. The main advantage of the proposed technique is its capability to perform block specific convolutions to extract block specific features independently. Such, block specific feature extraction efficiently extracts relevant features which subsequently improves the predictive performance of models. In this study, a two-block case was presented as a proof-of-concept but the parallel input CNNs approach can be implemented for any number of desired blocks. However, with the increase of the number of blocks, the computational time and resources needed to perform model hyper-parameters optimizations will also increase. This parallel input CNNs architecture is not limited to the fusion of spectra but can also be extended to combine images (2D matrices) with spectra or even with 3D data cubes (e.g. video or hyperspectral images). At the present moment, due to the unavailability of relevant large data sets, such a demonstration was not included in this study. More complex convolutions blocks, composed of several conv. layers with multiples filters combined with pooling layers can be engineered to specific data types. The main benefits of deep multiblock analysis are expected to appear in the near future when large-sized multiblock data sets are available for deep learning tasks. The deep multiblock method can also be combined with recent spectral pre-processing augmentation techniques to further enrich the models [36].

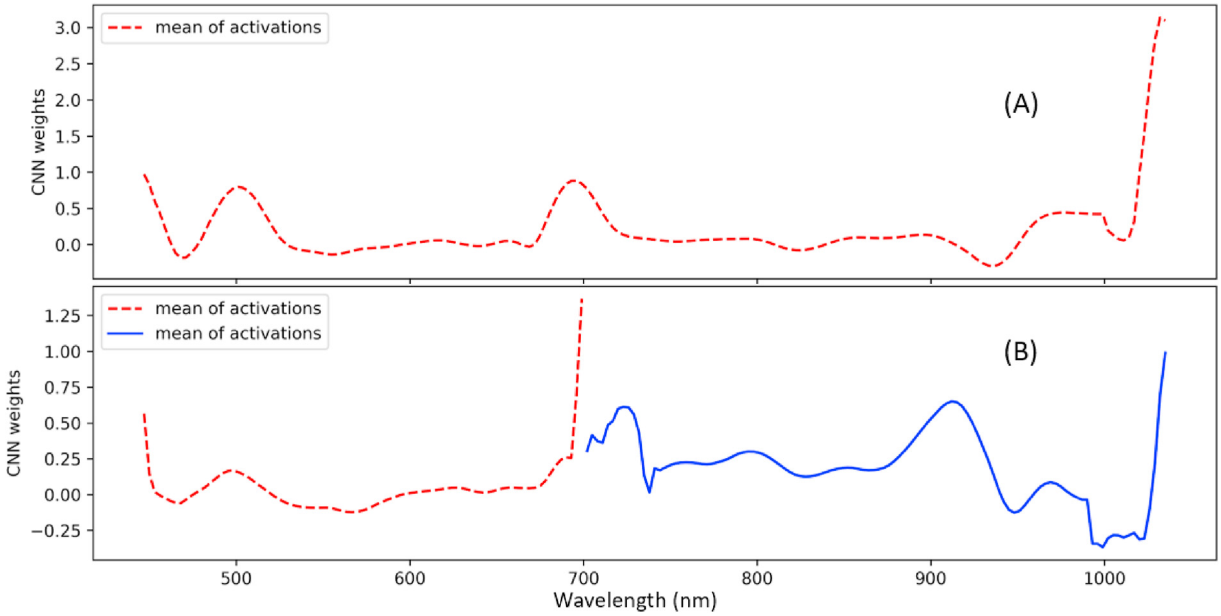


Fig. 10. A summary of mean activations of the CNN for the single block (A) and multiblock CNN case (B).

CRediT authorship contribution statement

Puneet Mishra: Conceptualization, Methodology, Software, Writing – original draft, Data curation. **Dário Passos:** Conceptualization, Software, 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] 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, 104139.

[2] 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.

[3] 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.

[4] A. Biancolillo, R. Bucci, A.L. Magrì, A.D. Magrì, F. Marini, Data-fusion for multiplatform characterization of an Italian craft beer aimed at its authentication, Anal. Chim. Acta 820 (2014) 23–31.

[5] R. Vitale, O.E. de Noord, J.A. Westerhuis, A.K. Smilde, A. Ferrer, Divide, et al., How disentangling common and distinctive variability in multiset data analysis can aid industrial process troubleshooting and understanding, J. Chemometr. (2020), e3266 n/a.

[6] 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.

[7] 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, 104190.

[8] 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.

[9] R. Bro, A.K. Smilde, Principal component analysis, Analytical Methods 6 (2014) 2812–2831.

[10] S. Wold, M. Sjostrom, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemometr. Intell. Lab. Syst. 58 (2001) 109–130.

[11] P. Geladi, B.R. Kowalski, Partial least-squares regression: a tutorial, Anal. Chim. Acta 185 (1986) 1–17.

[12] A.K. Smilde, I. Måge, T. Næs, T. Hankemeier, M.A. Lips, H.A.L. Kiers, E. Acar, R. Bro, Common and distinct components in data fusion, J. Chemometr. 31 (2017), e2900.

[13] Y. Song, J.A. Westerhuis, A.K. Smilde, Separating common (global and local) and distinct variation in multiple mixed types data sets, J. Chemometr. 34 (2020), e3197.

[14] I. Måge, A.K. Smilde, F.M. van der Kloet, Performance of methods that separate common and distinct variation in multiple data blocks, J. Chemometr. 33 (2019), e3085.

[15] A. Biancolillo, T. Næs, R. Bro, I. Måge, Extension of SO-PLS to multi-way arrays: SO-N-PLS, Chemometr. Intell. Lab. Syst. 164 (2017) 113–126.

[16] T. Næs, O. Tomic, N.K. Afseth, V. Segtnan, I. Måge, Multi-block regression based on combinations of orthogonalisation, PLS-regression and canonical correlation analysis, Chemometr. Intell. Lab. Syst. 124 (2013) 32–42.

[17] A. Biancolillo, T. Næs, M. Cocchi, Chapter 6 - the Sequential and Orthogonalized PLS Regression for Multiblock Regression: Theory, Examples, and Extensions, Data Handling in Science and Technology, Elsevier2019, pp. 157-177.

[18] 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.

[19] T. Skotare, D. Nilsson, S. Xiong, P. Geladi, J. Trygg, Joint and unique multiblock Analysis for integration and calibration transfer of NIR Instruments, Anal. Chem. 91 (2019) 3516–3524.

[20] M. Alinaghi, H.C. Bertram, A. Brunse, A.K. Smilde, J.A. Westerhuis, Common and distinct variation in data fusion of designed experimental data, Metabolomics 16 (2019) 2.

[21] A. Biancolillo, F. Marini, J.-M. Roger, SO-CovSel, A novel method for variable selection in a multiblock framework, J. Chemometr. 34 (2020), e3120.

[22] A.K. Smilde, J.A. Westerhuis, R. Boqué, Multiway multiblock component and covariates regression models, J. Chemometr. 14 (2000) 301–331.

[23] J.A. Westerhuis, T. Kourti, J.F. MacGregor, Analysis of multiblock and hierarchical PCA and PLS models, J. Chemometr. 12 (1998) 301–321.

[24] K.H. Liland, T. Næs, U.G. Indahl, ROSA—a fast extension of partial least squares regression for multiblock data analysis, J. Chemometr. 30 (2016) 651–662.

[25] S. Park, E. Ceulemans, K. Van Deun, Sparse common and distinctive covariates regression, J. Chemometr. (2020), e3270 n/a.

[26] 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.

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

[28] S. Malek, F. Melgani, Y. Bazi, One-dimensional convolutional neural networks for spectroscopic signal regression, J. Chemometr. 32 (2018), e2977.

[29] E.J. Bjerrum, M. Glahder, T. Skov, Data Augmentation of Spectral Data for Convolutional Neural Network (CNN) Based Deep Chemometrics, 2017 arXiv preprint arXiv:1710.01927.

[30] 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.

[31] 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.

[32] N. Anderson, K. Walsh, P. Subedi, Mango DMC and spectra Anderson et al, Mendley, Mendley data, 2020, 2020.

[33] 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.

[34] 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.

[35] P. Mishra, D. Passos, Realizing Transfer Learning for Updating Deep Learning Models of Spectral Data to Be Used in a New Scenario, Chemometrics and Intelligent Laboratory Systems, 2021, 104283.

[36] 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, 104287.

[37] P. Mishra, D.N. Rutledge, J.-M. Roger, K. Wali, H.A. Khan, Chemometric pre-processing can negatively affect the performance of near-infrared spectroscopy models for fruit quality prediction, Talanta (2021), 122303.

[38] 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.

[39] 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.

[40] 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.

[41] B.G. Osborne, Near-Infrared Spectroscopy in Food Analysis, Encyclopedia of Analytical Chemistry, 2006.