Analytica Chimica Acta 1058 (2019) 48–57

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

DeepSpectra: An end-to-end deep learning approach for quantitative spectral analysis

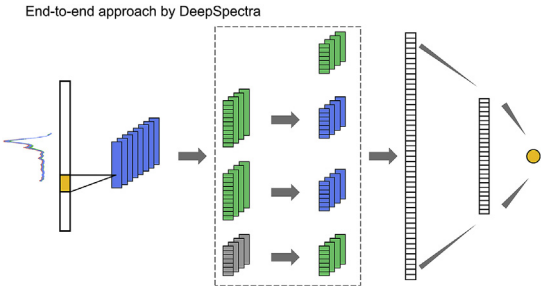
Xiaolei Zhang [a, b, 1], Tao Lin [a, b, 1], Jinfan Xu [a, b], Xuan Luo [a, b], Yibin Ying [a, b, c, *]

[a] College of Biosystems Engineering and Food Science, Zhejiang University, Hangzhou, Zhejiang, 310058, China
[b] Key Laboratory of On Site Processing Equipment for Agricultural Products, Ministry of Agriculture and Rural Affairs, China
[c] Faculty of Agricultural and Food Science, Zhejiang A&F University, Hangzhou, Zhejiang, 311300, China

HIGHLIGHTS

- DeepSpectra with the Inception module is developed for quantitative spectral analysis.
- DeepSpectra outperforms other CNN approaches on raw spectra analysis.
- Preprocessing strategies have little positive impact on DeepSpectra model performance.
- DeepSpectra on raw data comparable to the best calibration approach on preprocessing.
- The model repeatability and accuracy improved with increased sample sizes.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 25 September 2018
Received in revised form
24 November 2018
Accepted 2 January 2019
Available online 8 January 2019

Keywords:
Chemometrics
Inception
Convolutional neural network
Model accuracy
Repeatability

ABSTRACT

Learning patterns from spectra is critical for the development of chemometric analysis of spectroscopic data. Conventional two-stage calibration approaches consist of data preprocessing and modeling analysis. Misuse of preprocessing may introduce artifacts or remove useful patterns and result in worse model performance. An end-to-end deep learning approach incorporated Inception module, named DeepSpectra, is presented to learn patterns from raw data to improve the model performance. DeepSpectra model is compared to three CNN models on the raw data, and 16 preprocessing approaches are included to evaluate the preprocessing impact by testing four open accessed visible and near infrared spectroscopic datasets (corn, tablets, wheat, and soil). DeepSpectra model outperforms the other three convolutional neural network models on four datasets and obtains better results on raw data than in preprocessed data for most scenarios. The model is compared with linear partial least square (PLS) and nonlinear artificial neural network (ANN) methods and support vector machine (SVR) on raw and preprocessed data. The results show that DeepSpectra approach provides improved results than conventional linear and nonlinear calibration approaches in most scenarios. The increased training samples can improve the model repeatability and accuracy.

1. Introduction

The development of chemometrics boosts the spectroscopic based analysis in many different fields, including agricultural

* Corresponding author. College of Biosystems Engineering and Food Science, Zhejiang University, Hangzhou, Zhejiang, 310058, China.
E-mail address: yingyb@zju.edu.cn (Y. Ying).
[1] X.Z. and T.L. contributed equally to this work.

https://doi.org/10.1016/j.aca.2019.01.002
0003-2670/

products [1], pharmaceuticals [2], petrochemical [3], and soil [4]. Linear and nonlinear calibration approaches are developed to establish the relationships between the spectra to the concentration. Given the experimental changes and variability of sample origins, there exist various signals and artifacts from samples and environment. The uncertainty and variations from spectral signal and artifacts, however, lead to a critical challenge for model generalization and robustness. Practical limitations to the deployment of such techniques appear when existing calibration models are to be applied to measurements recorded by new samples and in different environmental conditions.

Preprocessing is usually required to enhance calibration model accuracy by reducing un-modeled variations such as instrumental and experimental artifacts [5,6]. A typical preprocessing approach for near infrared spectral analysis includes four steps: baseline correction, multiplicative scatter correction, noise removal, and scaling [7]. The choice of a suitable preprocessing or a combination of preprocessing methods may strongly affect the analysis performance and require significant computational resources by the trial-and-error approach. Misuse of preprocessing techniques, however, can decrease the model performance. Different datasets may require various optimal preprocessing approaches, possibly due to the differences of artifacts among various datasets [7,8].

Increasing the complexity of the model is another approach to improve model accuracy. Nonlinear calibration methods (e.g. Artificial Neural Network (ANN)) can capture both linear and nonlinear features of the spectra. It is, however, often criticized for the complex structure of the model, the risk of overfitting, and a requirement for large sample sizes [9]. Most existing nonlinear calibration analyses do not work on raw spectral data, often requiring principal component analysis (PCA) to reduce the dimension of spectra data matrix [9–11]. The reduction of data dimension, however, would possibly change the original patterns of spectra signal and lose useful information, not necessary for learning patterns from the data. Improving the model accuracy and robustness, therefore, requires an integrated approach to remove various artifacts and distill the information related to product characteristics from the spectroscopic data.

Recent developments in machine learning have demonstrated that data-driven deep learning approaches can discover intricate structures in high-dimensional data, reducing the need from prior knowledge and human effort in feature engineering [12]. A deep learning architecture is a multilayer stack composed of simple but non-linear modules which transform the representation at one level (starting with the raw input) into a representation at a higher, slightly more abstract level. With the composition of enough such transformations, very complex functions can be learned. In particular, Convolutional Neural Networks (CNN) are a popular deep learning approach that has been well applied in two- and three-dimensional data analysis, including image recognition [13,14], video analysis [15], drug discovery [16], and playing Go [17,18]. The CNN approach takes advantage of local sparse connections to study local patterns from the raw data and reduce the risk of overfitting by adopting the weight sharing. The Inception module developed in GooLeNet network achieves new state of art in deep learning by increasing both the width and depth of a neural network [13]. The details of DeepSpectra model are described in the section of Materials and Methods. To the best of our knowledge, some studies have developed deep learning based approaches for vibrational spectral analysis [19–22]. There, however, has few studies considering the increase of model width to capture multiple local spectra features hidden in the raw spectra for a better understanding of the relationship between spectra and sample concentration.

In this paper, we develop a spectral analytical approach without the requirement of data preprocessing for quantitative spectral analysis. The approach is named DeepSpectra for deep convolutional neural network based analysis. The DeepSpectra approach is tested by four datasets and its model accuracy is compared with three CNN models on the raw data. To evaluate the preprocessing impact on DeepSpectra, 16 preprocessing strategies are quantified in this study. We compare DeepSpectra model with three popular conventional calibration methods, including partial least square (PLS), artificial neural network (ANN), and support vector regression (SVR), on the raw and preprocessed data. Furthermore, the impact of sample size on the repeatability and accuracy of DeepSpectra approach is evaluated. The results demonstrate that DeepSpectra approach improves quantitative analysis of spectroscopic data without the need for data preprocessing.

2. Materials and methods

2.1. DeepSpectra model architecture

DeepSpectra model developed in this study is based on a convolutional neural network (CNN). CNN has the characteristics of local connections, weights sharing, pooling, and deeper layers [12]. Different neurons in each layer are locally connected to the neurons in the preceding layer by the filter with the shared weights. The 'deep' here means on the level on both the depth and width of the network (Fig. 1). The input of the model is one-dimensional raw spectroscopic data and the output is the object character to be estimated. The model has three convolutional layers (labeled as Conv1, Conv2, and Conv3), a flatten layer (labeled as Flatten), one fully connected layer (labeled as F1), and one output layer. Because of the local connection and shared weights, CNN has fewer trainable parameters than fully connected neural networks.

One unique feature of DeepSpectra model is its adoption of Inception structure [13] in layer Conv2 and Conv3 using four parallel different sizes of convolution modules as well as the pooling and 1×1 convolutions. To the best of our knowledge, the use of Inception module for one-dimensional spectral analysis has not been investigated yet. The increase of the depth of the network can possibly extract both low and high levels of features from raw spectra, without the need of principal component analysis for dimension reduction. The increase of the width of the network by stacking of parallel 1×1, $m \times 1$, and $n \times 1$ convolutions in layer Conv3 improves the adaptability of the network to the different scales of the local spectra features. DeepSpectra learns patterns from limited variables through concatenated outputs by combinations of these filter sizes. The Inception module allows DeepSpectra to have a large width and depth while keeping the computational complexity constant [13]. The 1×1 convolutions (green module) and pooling (gray module) in layer Conv2 reduce the number of weights of the network by decreasing the number and length of feature maps, respectively. A 1×1 convolution in Conv3 after the pooling operation is also used for feature map decreasing.

An example of three filter sizes of '1×1', '3×1', and '5×1' in the second layer is shown in Fig. 2. In the first layer, all the filters have a size of three and a stride of three. Consequently, one neuron in the Conv2 layer covers a receptive field of three, nine, and 15 original spectral variables, respectively. Once the neuron is activated, the pattern from the associated receptive field is learned as a feature.

Layer Conv1 is a convolutional layer with eight filters of the same filter size. After the first convolutional layer, one-dimensional spectra are transformed to eight feature maps with a smaller size. The filter size and stride vary for different datasets, which are optimized according to input spectra features and the training sample size (Table 1).

Layer Flatten is to concatenate the parallel outputs of layer

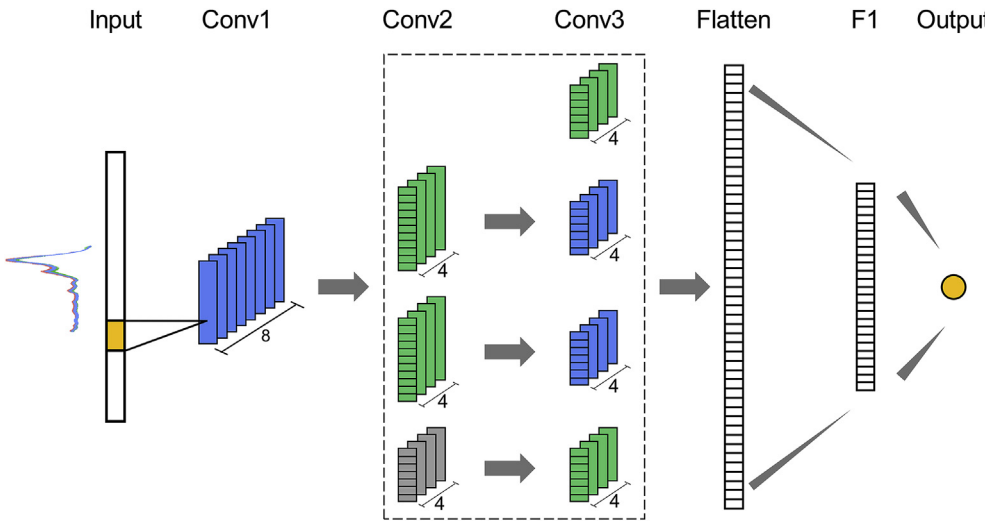


Fig. 1. The structure of DeepSpectra model. It consists of three convolutional layers (Conv1, Conv2, and Conv3), one flatten layer, one fully connected layer (F1), and one output layer. The blue module represents general convolution, the green module represents 1×1 convolution, and the gray module represents max pooling. The first 1×1 convolution module in layer Conv3 is directly connected to layer Conv1, which is called short connection. The last 1×1 convolution module in layer Conv3 and the two 1×1 convolution modules in layer Conv2 are used for channel number reduction to reduce the computation complexity. Small rectangles in each layer represent neurons, and large rectangles represent feature maps in Conv1, Conv2 and Conv3. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

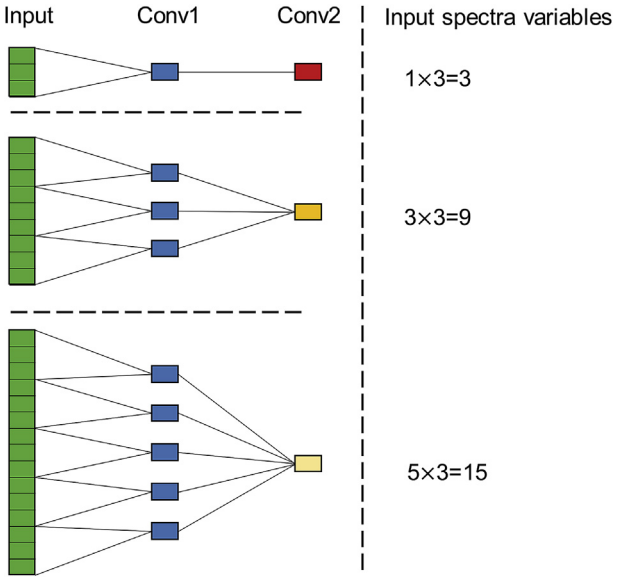


Fig. 2. The changes of the receptive spectra region by two convolutional layers with different filter sizes. An example of two convolutional layers with the first layer has a filter size of three and stride of three, and the second layer has three filter sizes of one, three, and five. A green rectangle in the input layer represents the input spectral variables, a rectangle in the Conv1 and Conv2 layers represents a neuron. One neuron in the Conv2 layer covers a receptive field of three, nine, and 15 original spectral variables, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Conv3 and convert them into a one-dimensional single vector. It is then fed into the fully connected layer F1. The Flatten layer has no parameters to be trained. The dropout function is adopted to prevent model overfitting and improve computing performance by randomly killing off a bunch of neurons [23].

Layer F1 is a dense layer fully connected to layer Flatten. The number of neurons in layer F1 is smaller than that in layer Flatten. The output layer is fully connected to layer F1 with the dropout function incorporated. Layer F2 has one neuron to output the objective to be estimated.

Table 1
Hyperparameters used in DeepSpectra model of four spectra datasets.

Hyperparameter	Corn	Tablet	Wheat	Soil
Kernel size 1	7	5	7	9
Kernel size 2	3	2	3	5
Kernel size 3	5	3	5	7
Stride 1	3	3	3	5
Stride 2	2	2	2	3
Hidden number	16	32	32	64
Batch size	32	128	128	512
Dropout rate	0.2	0.4	0.1	0.5
Regularization coefficient	0.001	0.001	0.01	0.01
Learning rate	0.01	0.01	0.01	0.01
Learning rate decay	0.001	0.001	0.001	0.001

For the objective function, we use mean squared error (MSE) and L2 norm regularization to minimize the sum of squares loss and prevent from overfitting (Eq. (1)).

$$\text{Loss} = \frac{1}{N}\sum_{n=1}^{N}[(y_n - \hat{y}_n)^2] + \lambda\|w\|^2 \quad (1)$$

Where y_n and $\hat{y}_n$ are measured values and predicted values, N is the number of samples in the training set, w is the weight matrix, and λ is the regularization coefficient.

We use the leaky rectified linear function [24] as the activation function for the convolutional layers and fully connected layer in the model. A batch normalization (BN) is used after layer Flatten and F1 to speed up training process and improve accuracy [25]. We use backpropagation together with Adam [26] optimizer to train the model and find a local minimum of the objective function. An early stopping strategy is used to avoid overfitting. To avoid gradient vanishing or exploding, we use variance scaling for weight initialization [27].

2.2. Design of experiments

2.2.1. Comparison of DeepSpectra with other CNN models

To evaluate the model performance of DeepSpectra, we compare

it with other three convolutional neural networks (CNN) based models (Fig. 3). The major difference of these three models lies in model depth. The first CNN model (Model 1) is the shallowest one with only one convolutional layer (Conv1), one fully connected layer (F1), and one output layer. Layer Conv1 has eight filters of the same filter size.

The second CNN model (Model 2) including two convolutional layers (Conv1 and Conv2), one fully connected layer (F1), and one output layer. Layer Conv1 includes eight filters of the same filter size. The second convolutional layer (Conv2) includes 12 filters of three different sizes to learn different features of data by using various filter sizes within one convolutional layer. Multiple filter sizes can learn multiple features with different receptive fields.

The third CNN model (Model 3) is deeper than other models, which consists of three convolutional and two pooling layers (Conv1, Pooling1, Conv2, Pooling2, Conv3), followed by a fully connected layer (F1) and an output layer.

The hyperparameters that include the filter sizes and strides in the convolutional layers and the hidden numbers in the flatten layer are listed in Table A1. The activation and objective functions of these three CNN models are the same as that for DeepSpectra.

2.2.2. Comparison of DeepSpectra with conventional calibration approaches

To compare DeepSpectra model with conventional calibration methods, we consider three popular linear and nonlinear

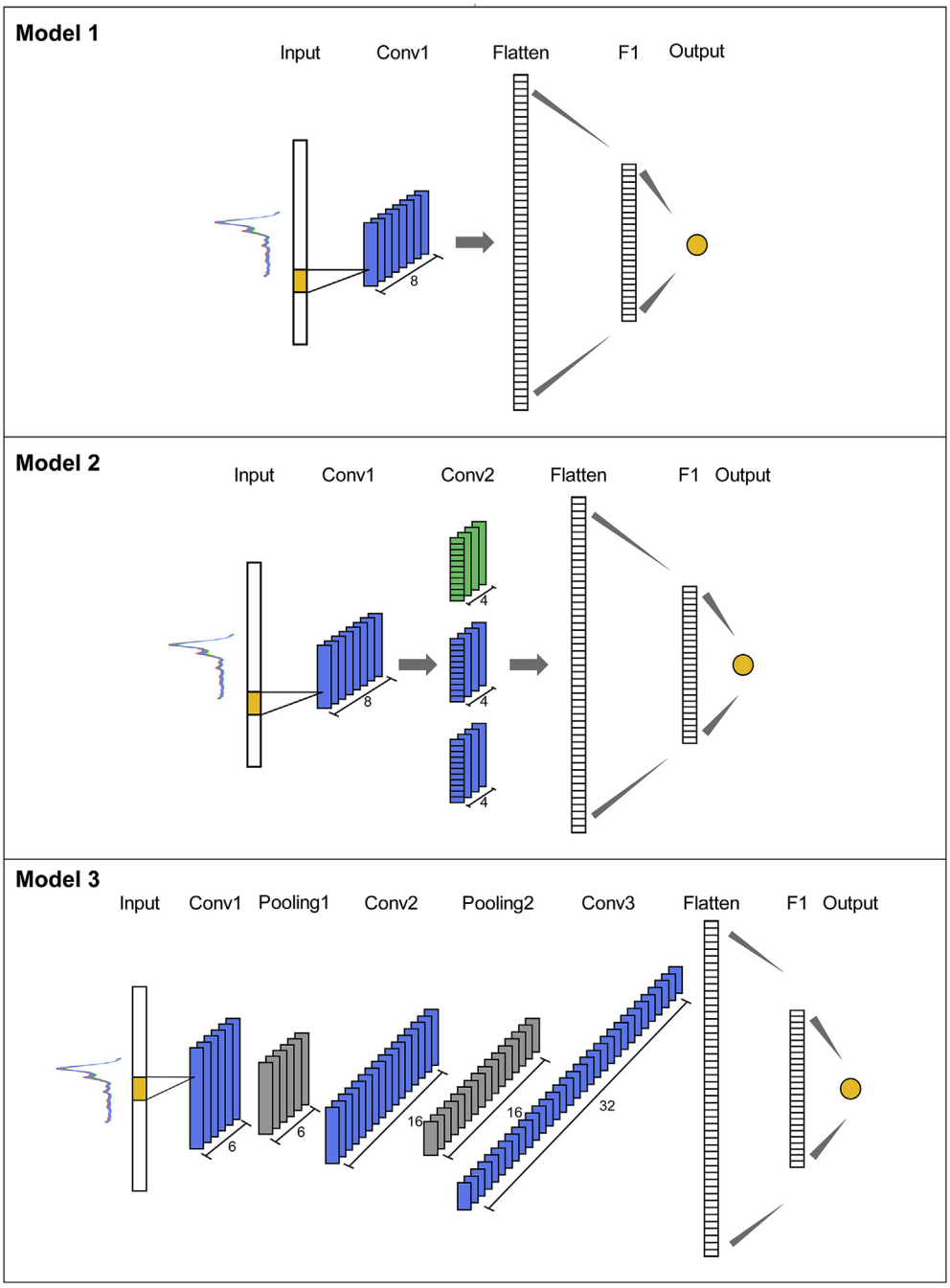


Fig. 3. Architectures of three CNN models. The blue module represents general convolution, the green module represents 1 × 1 convolution, and the gray module represents max pooling. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

calibration approaches, partial least square (PLS), principal component analysis - artificial neural network (PCA-ANN), and support vector regression (SVR) in this study. The flowchart of the experimental design is shown in Fig. 4. To understand the preprocessing effect on the DeepSpectra and other calibration approaches, 16 preprocessing strategies (Table A2) are considered based on a previous study [28], which includes a combination of baseline correction (asymmetric least squares, AsLS), multiplicative scatter correction (standard normal variate, SNV), smoothing (S-G filter), and scaling (Mean centering and Pareto scaling).

Parameters of all the models have been optimized toward an improved model performance. For the PLS approach, 5-fold cross validation is used to optimize the number of latent variables (LVs) from 2 to 40. For the PCA-ANN approach, the input spectral variables are reduced to lower dimensions through PCA before feeding into a typical artificial neural network with one hidden layer. We optimize the number of principal components (PCs) from a range of 2–40 by 5-fold cross validation. For the SVR approach, we use radial basis function (RBF) as a kernel function. We optimize the penalty parameter from a list of 0.01, 0.1, 1, 10, 100, 1000, and the kernel coefficient of 0.001 and 0.0001.

All the models are implemented on the Python platform using Keras and Scikit-learn library. The process is performed on a Linux workstation (Ubuntu 14.04 LTS) with 128 GB of RAM, and an Nvidia Geforce GTX1080Ti graphics card with 11 GB of RAM.

2.3. Model evaluation

The model performance is evaluated by root mean squared error of prediction (RMSEP) and R^2, which are shown in Eqs. (2) and (3).

$$RMSEP = \sqrt{\frac{\sum_{n=1}^{N}[(y_n - \hat{y}_n)^2]}{N}} \tag{2}$$

$$R^2 = \frac{\sum_{n=1}^{N}(\hat{y}_n - \bar{y})^2}{\sum_{n=1}^{N}(y_n - \bar{y})^2} \tag{3}$$

Where y_n and $\hat{y}_n$ are measured and predicted values, respectively. $\bar{y}$ is the average of y values, and N is the number of samples in the training set.

The coefficient of variation (CV) is used to evaluate the repeatability of model performance among different datasets. (Eq. (4)).

$$\mathrm{CV} = \frac{\sqrt{\frac{\sum_{j=1}^{N}(RMSEP_j - RMSEP_{avg})^2}{N}}}{RMSEP_{avg}} \tag{4}$$

Where the N is the repeated times, which is 25 in this study. The $RSMEP_j$ is the j-th root mean squared error of prediction. The $RMSEP_{avg}$ is the average root mean squared error of prediction for

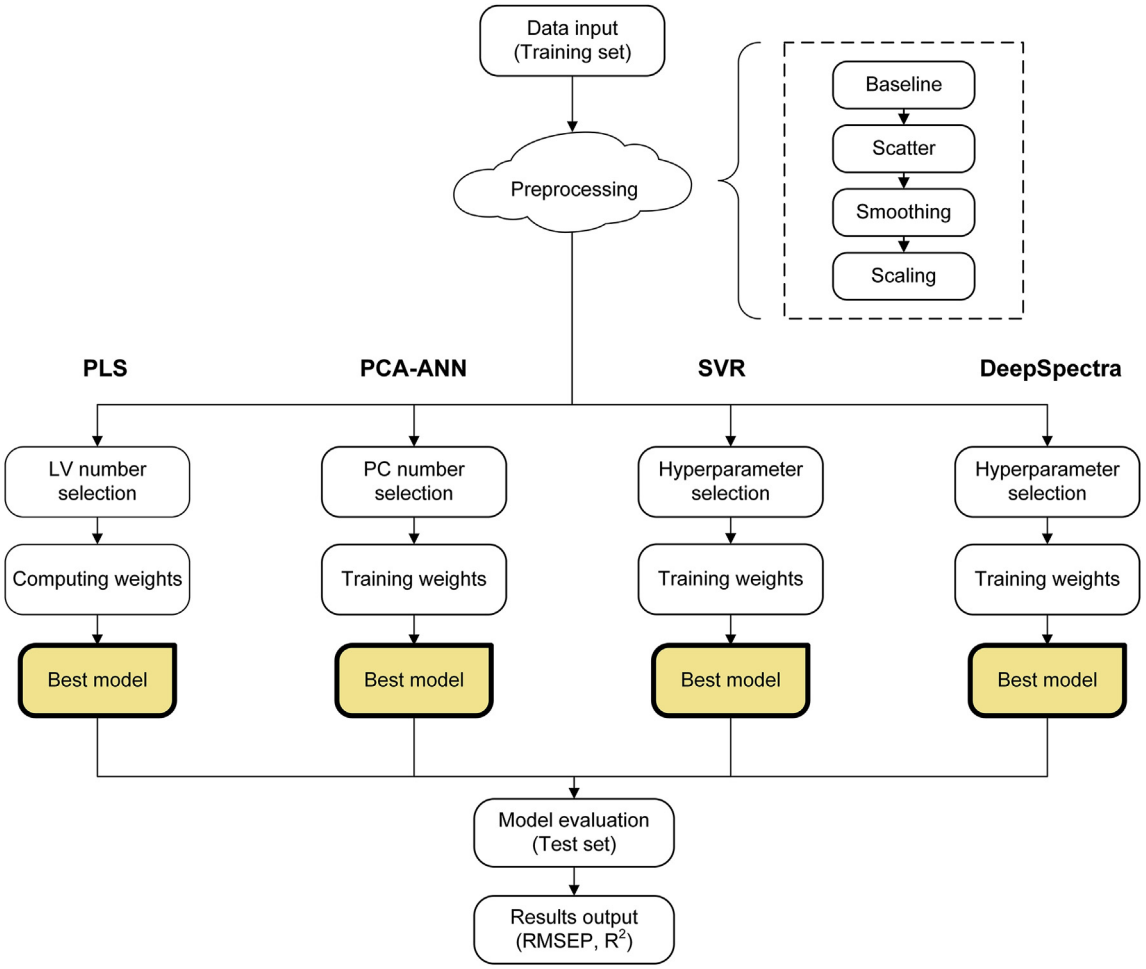


Fig. 4. Flowchart for the experimental design. DeepSpectra model as well as the PLS, PCA-ANN, and SVR models are implemented on raw and preprocessed data. The PC and LV numbers are optimized through 2 to 40. Abbreviations: AsLS: asymmetric least squares; SNV: standard normal variate; SG: Savitzky-Golay; MC: mean centering; PC: principal component; LV: latent variable.

25 repetitions.

2.4. Dataset description

Four datasets have been investigated in this study (Table 2). All of samples in each dataset are collected from either wide geographical origins, long time horizons, various production conditions, or different spectrometers. The raw spectra of four datasets are displayed in Figure A1.

The first dataset consists of 80 samples of corn measured on 3 different NIR spectrometers (M5, MP5, and MP6). We randomly select 27, 27, and 26 spectra samples from M5, MP5, and MP6 respectively, and divide them in training (60 samples) and test set (20 samples). The spectra cover a range from 1100 to 2498 nm at 2 nm intervals (700 channels). The protein content of the samples is to be predicted. The dataset is available at: www.eigenvector.com/Data/Corn/.

The second dataset consists of 238 samples of pharmaceutical tablets, which produced from a laboratory (89 samples), an intermediate scale tablet press (72), and an industrial scale press (67) [29]. The tablets produced in a laboratory and intermediate scale press make up training set, while those produced in industrial scale press are used for testing. The spectra are measured in transmittance mode with an ABB Bomem FT-NIR model MB-160. All spectra range from 952 to 1310 nm with 372 variables. The parameter to be predicted is active ingredient content (AC). The dataset is from 2012 International Diffuse Reflectance Conference (IDRC) Shootout competition: https://cnirs.clubexpress.com/content.aspx?page_id = 22&club_id = 409746&module_id = 148861.

The third dataset is wheat spectroscopic data with different classes of samples collected from eight crop years (1998–2005) [30]. 775 wheat samples grown from seven crop years (1998 and 2000–2005) are used as the training set while 107 samples grown in year 1999 is for the test set. The wheat spectra have a range from 400 to 2498 nm with a resolution of 2 nm, scanned on a NIRSystems 6500 instrument. The parameter to be predicted is protein content (PC). The dataset is from 2008 IDRC Shootout competition: https://cnirs.clubexpress.com/content.aspx?page_id= 22&club_id=409746&module_id=148863.

The fourth dataset is soil spectroscopic data collected from the U.S., Africa, Asia, South America and Europe [31]. It includes 2502 soil samples in the training set and 1291 in the test set, recorded in diffuse reflectance mode with a FieldSpec Pro-FR spectrometer. The spectra cover a range from 350 to 2500 nm with a resolution of 1 nm. The parameter to be predicted is soil organic carbon (SOC). An extreme value of SOC = 536.8 g kg^{-1} is removed according to a previous study [31]. The dataset is from 2006 IDRC Shootout competition: https://cnirs.clubexpress.com/content.aspx? page_id=22&club_id=409746&module_ id=148866.

In the base-case analysis, we compare the performance of DeepSpectra with three CNN models on the raw data. Statistical evaluation is provided by *t*-test to compare the performance (R^2) between DeepSpectra and other three models [32]. F-test is used for checking the equal variances assumption before conducting *t*-test. To evaluate the preprocessing impact on DeepSpectra, we apply 16 preprocessing strategies for four datasets. The PLS, PCA-ANN, and SVR models on the raw and preprocessed datasets are used as the benchmark. The soil dataset is used for sensitivity analysis of sample size by randomly selecting 10%–100% of the samples with a resolution of 10% from the original training set, respectively. The same test set is used for performance comparison.

3. Results and discussion

3.1. The comparisons between DeepSpectra and three CNN models

The end-to-end learning approach by DeepSpectra outperforms the other three convolutional neural network models on all four datasets in terms of a lower RMSEP and higher R^2 (Table 3). For corn protein analysis, DeepSpectra model obtains a mean RMSEP of 0.12 and a mean R^2 of 0.91, which is much better than other CNN models and a previous study [20]. The second best performance is from Model 2 which has a mean RMSEP of 0.18, or a 50% of RMSEP increase from the DeepSpectra model. For the dataset of pharmaceutical tablets, DeepSpectra model provides a mean RMSEP of active ingredient content (AC) of 0.35 and a mean R^2 of 0.88, which outperforms the other CNN models and the best result from the Shootout competition (RMSEP = 0.41) [29]. The Models 1 and 2 provide a low RMSEP of 0.38, or 10% increase from the result of DeepSpectra. For wheat protein analysis, DeepSpectra model generates a lowest mean RMSEP of 0.20, followed by the Models 2 and 1 with a mean RMSEP of 0.25 and 0.27, respectively. A previous study only achieves a best RMSEP of 0.37 [33]. For soil organic carbon analysis, the DeepSpectra model provides a lowest RMSEP of 8.88 among four calibration approaches, which is slightly better than Model 2 (RMSEP = 8.91) and a previous study using boosted regression trees model on preprocessed data of first derivative (RMSEP = 9.0) [31].

DeepSpectra model provides a significantly higher R^2 than all other three CNN models with 95% confidence interval for most datasets, except one scenario where it is not significantly better than Model 2 on the soil dataset (Table A3). It also provides a consistent improved performance compared to previous studies [22,29,31,32]. The advantages of DeepSpectra model is due to its unique neural network structure in layers Conv2 and Conv3. The parallel convolutions with different filter sizes in the third convolutional layer (Conv3) allow the model to capture both local and high abstracted features in one layer. The 1×1 convolutions and pooling in the second convolutional layer (Conv2) reduce the number of variables to alleviate computational challenges. These unique characteristics of network structure help to capture important features from the inputs and prevent possible overfitting. Model 2 with three filter sizes also performs relatively well, only second to DeepSpectra, which demonstrates that parallel combinations of convolutional operations with varied filter sizes have a promising structure to learn patterns from raw spectra with the increase of model width.

Model 3 with a deep neural network of three convolutional and two pooling layers obtains relatively poor performance on both training and test set, especially on the corn dataset where an RMSEC of 0.44 and RMSEP of 0.35 are provided. We also notice that Model 3 obtains closer performance to Models 1 and 2 on the datasets with larger sample size. It shows that increased sample

Table 2
Description of the datasets.

Datasets	# total samples	# training samples	# test samples	# features	Min	Max	Mean	SD
Corn protein content (%)	80	60	20	700	7.65	9.71	7.65	0.50
Tablet AC (%)	228	161	67	372	4.61	9.79	7.49	1.26
Wheat protein content (%)	882	775	107	1050	9.70	17.00	13.54	1.56
Soil organic content (g kg^{-1})	3793	2502	1291	2151	0	241.60	11.97	20.87

Table 3
Predictive results obtained by DeepSpectra as well as three CNN models on raw data for four datasets. The mean and standard deviation of RMSEP from 25 runs are presented. The last column presents the RMSEPs from previous studies.

Datasets	Model 1(Mean ± Std)	Model 2(Mean ± Std)	Model 3(Mean ± Std)	DeepSpectra(Mean ± Std)	Previous study(Range)
Corn protein content (%)					
RMSEP	0.22 ± 0.04	0.18 ± 0.05	0.35 ± 0.01	**0.12 ± 0.02**	–
R^2	0.72 ± 0.09	0.80 ± 0.10	0.26 ± 0.06	**0.91 ± 0.04**	0.74–0.82 [22]
Tablet AC (%)					
RMSEP	0.38 ± 0.05	0.38 ± 0.05	0.48 ± 0.06	**0.35 ± 0.05**	0.41–0.48 [29]
R^2	0.85 ± 0.05	0.85 ± 0.04	0.77 ± 0.06	**0.88 ± 0.03**	0.84–0.86 [29]
Wheat protein content (%)					
RMSEP	0.27 ± 0.07	0.25 ± 0.06	0.49 ± 0.15	**0.20 ± 0.01**	0.37–0.70 [33]
R^2	0.97 ± 0.02	0.97 ± 0.01	0.89 ± 0.07	**0.98 ± 0.002**	0.85–0.95 [33]
Soil organic carbon (g kg^{-1})					
RMSEP	9.27 ± 0.18	8.91 ± 0.17	9.34 ± 0.19	**8.88 ± 0.22**	9.00 [31]
R^2	0.82 ± 0.01	0.83 ± 0.01	0.82 ± 0.01	**0.84 ± 0.01**	0.55–0.82 [31]

size can improve the modeling fitting performance by a deep neural network. The advantage of deep neural network, however, is not reflected on the small dataset. With the increase of sample size, Model 3 demonstrates improved performance for better representing the structure of spectral data.

The stack of pooling layers could lead to information loss, which results in the poor performance by Model 3, especially on the corn dataset with small number of spectral features. Model 3 cannot learn enough effective patterns from the little information to fit data well.

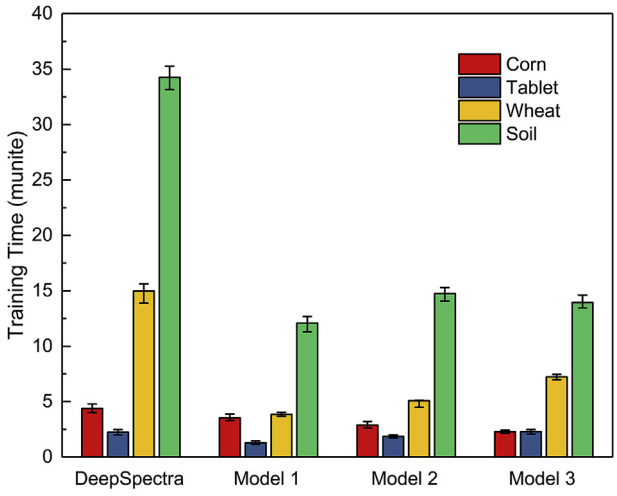


Fig. 5. Training time of four deep learning models on four datasets.

It is a trade-off between overfitting and underfitting for the design of the model structure. Compared to Model 3, DeepSpectra applies 1 × 1 convolutions and pooling within one layer in the Inception module, which reduces the information loss by the pooling operation only and thus retains the information in the model analysis. Models 1 and 2 do not have pooling layers so they can capture more features while bearing the risk of overfitting. The impact of information loss from pooling operations could be alleviated with the increase in the number of spectral features. The corn dataset has only 700 features with a resolution of 2 nm, while soil dataset has 2151 features with a resolution of 1 nm. The soil dataset, therefore, contains more information from the spectra. Model 3 is less sensitive to information loss resulted from pooling on soil dataset and provides higher accuracy. The tablet dataset has finer spectra resolution and larger sample size compared to the corn dataset, which result in a relatively good modeling performance for Model 3.

DeepSpectra, with similar depth and pooling operations as Model 3, has an Inception structure to extract features from the different scales simultaneously [13]. The Inception structure reduces information loss from pooling and prevents overfitting by weight reduction, which leads to a better matching to the four spectral datasets. The results reflect the advantage of Inception structure with increased width and depth over conventional deep neural networks in one-dimensional spectral analysis, especially on the dataset with few sample number and few spectral features such as the corn dataset.

DeepSpectra model, however, requires a longer computing time than other three CNN models for each dataset (Fig. 5). The training process of the DeepSpectra model costs up to 34.26 min for soil datasets in this study, followed by Model 2, Model 3, and Model 1 with 14.74, 13.94, and 12.09 min, respectively. The training time is

Table 4
RMSEPs obtained by DeepSpectra on the 16 kinds of preprocessed data and raw data.

Experiment	Baseline	Scatter	Smoothing	Scaling	Corn (%)	Tablet (%)	Wheat (%)	Soil (g kg^{-1})
1	AsLS	SNV	yes	Pareto	0.21 ± 0.01	0.42 ± 0.02	0.90 ± 0.02	13.83 ± 0.25
2	AsLS	SNV	yes	MC	0.22 ± 0.01	0.41 ± 0.03	0.90 ± 0.02	13.92 ± 0.21
3	AsLS	SNV	none	Pareto	0.23 ± 0.02	0.42 ± 0.03	0.90 ± 0.02	13.95 ± 0.26
4	AsLS	SNV	none	MC	0.22 ± 0.02	0.41 ± 0.02	0.90 ± 0.02	13.85 ± 0.25
5	AsLS	none	yes	Pareto	0.26 ± 0.01	0.93 ± 0.07	0.76 ± 0.02	12.95 ± 0.20
6	AsLS	none	yes	MC	0.27 ± 0.02	0.90 ± 0.08	0.77 ± 0.03	12.86 ± 0.23
7	AsLS	none	none	Pareto	0.27 ± 0.01	0.94 ± 0.06	0.77 ± 0.02	12.92 ± 0.23
8	AsLS	none	none	MC	0.27 ± 0.01	0.91 ± 0.11	0.76 ± 0.02	12.92 ± 0.29
9	none	SNV	yes	Pareto	0.16 ± 0.01	0.36 ± 0.03	**0.20 ± 0.01**	9.39 ± 0.19
10	none	SNV	yes	MC	0.15 ± 0.02	0.36 ± 0.03	**0.20 ± 0.01**	9.36 ± 0.21
11	none	SNV	none	Pareto	0.15 ± 0.01	0.37 ± 0.02	**0.20 ± 0.01**	9.36 ± 0.26
12	none	SNV	none	MC	0.15 ± 0.01	0.37 ± 0.02	**0.20 ± 0.01**	9.42 ± 0.23
13	none	none	yes	Pareto	0.14 ± 0.04	0.36 ± 0.08	**0.20 ± 0.01**	8.94 ± 0.22
14	none	none	yes	MC	**0.12 ± 0.01**	0.38 ± .311	0.23 ± 0.13	9.05 ± 0.22
15	none	none	none	Pareto	**0.12 ± 0.02**	0.38 ± 0.06	**0.20 ± 0.01**	8.94 ± 0.16
16	none	none	none	MC	0.13 ± 0.02	0.36 ± 0.04	**0.20 ± 0.02**	9.01 ± 0.24
17	none	none	none	none	**0.12 ± 0.02**	**0.35 ± 0.05**	**0.20 ± 0.01**	**8.88 ± 0.22**

related to sample size, input features, batch size, and others. Soil dataset requires most computing time to calibrate for all four models, given its large sample size (2502) and input features (2151). It is interesting to find that the tablet dataset needs least amount of training time though its sample size is larger than corn dataset. This is because the tablet dataset has only 372 input features while the corn dataset has 700 features. DeepSpectra model requires a very short time on testing with a few seconds for all four datasets.

3.2. Impact of preprocessing on DeepSpectra model

DeepSpectra model on raw data obtains better results than in preprocessed data for most scenarios (Table 4). For tablet and soil datasets, DeepSpectra model based on raw spectra outperforms the model using all 16 preprocessing strategies. For the tablet dataset, DeepSpectra model obtains a lowest mean RMSEP of 0.35 on the raw data, compared to a mean RMSEP in a range of 0.36–0.94 on the preprocessed data. For the soil dataset, the performance on the raw spectra is at a mean RMSEP of 8.88, improved from a range of RMSEP from 8.94 to 13.95 using the preprocessed data. For the corn dataset, DeepSpectra model provides a mean RMSEP of 0.12 with a standard deviation of 0.20 by raw data, which is comparable to the best results by two preprocessing approaches (smoothing + MC scaling and Pareto scaling only). For the wheat dataset, DeepSpectra model on seven preprocessing strategies achieve a mean RMSEP of 0.20, which is the same as the accuracy achieved using the raw data.

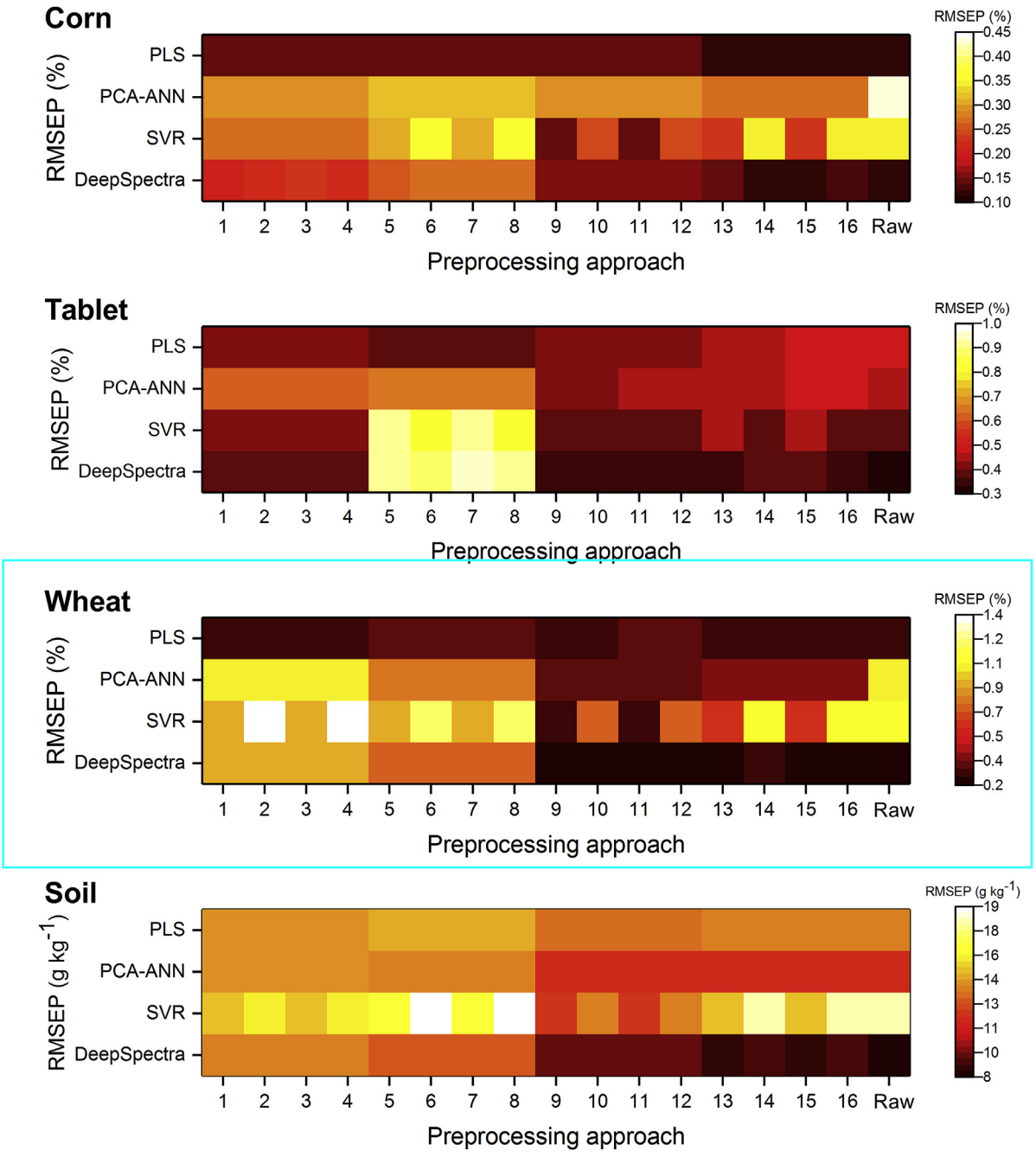


Fig. 6. Results of DeepSpectra, PLS, PCA-ANN, and SVR on the preprocessed data and raw data. RMSEP: root mean square error of prediction. The RMSEP values are presented in Table A4.

The results demonstrate that the DeepSpectra model can learn patterns from raw data and achieve the best quantitative analysis results without any preprocessing.

3.3. Comparison of DeepSpectra with conventional calibration approaches

DeepSpectra model on raw spectra data outperforms the conventional linear (PLS) and nonlinear (PCA-ANN and SVR) approaches using raw and preprocessed data in most scenarios on four datasets (Fig. 6, Table A4). Different datasets, however, would have their preferred conventional calibration methods (Fig. 6). The PLS method provides a lowest RMSEP for wheat and corn datasets, whereas the PCA-ANN and SVR approaches are best suited for soil organic carbon and tablet analysis, respectively.

DeepSpectra model provides an improved model performance, especially for the dataset with a large sample size such as soil samples. The mean RMSEP can be reduced to 8.88, which is much lower than the second best performance of 11.67 based on the PCA-ANN approach using the preprocessed data. DeepSpectra also improves the active ingredient content analysis of tablets to a mean RMSEP of 0.35, comparing to the best result by traditional approaches at 0.42 based on the SVR method after data preprocessing (strategies 9 and 11).

For protein content analysis of corn and wheat samples, DeepSpectra model on raw data provides a comparable performance to the best preprocessing strategy by the PLS model. This corresponds to a previous study suggesting a linear relation of the spectra to the concentration for protein analysis [9]. DeepSpectra model achieves the lowest REMSEP at 0.12 and 0.20 for corn and wheat, respectively, which compares to the best performance of PLS approach at 0.12 and 0.21. DeepSpectra model has no obvious advantage over PLS model for corn dataset with small sample size. The nonlinear PCA-ANN and SVR approaches, however, provide much poor results on both raw and preprocessed data for both datasets. Considering all four datasets, the results demonstrate that DeepSpectra model not only performs improved results on capturing nonlinear patterns but also provides comparable performance for samples with linear patterns.

The appropriate selection of preprocessing strategy does improve the performance by the traditional calibration approaches (Fig. 6). Different calibration models, however, have their preferred preprocessing methods for each dataset. For example, the preprocessing approaches 5–8 boost the PLS model for tablets analysis, but they unfortunately worsen the performance for PCA-ANN and SVR analysis.

3.4. Impact of sample size change on the model performance

DeepSpectra model provides a varied result on the same dataset with different runs as a result of stochastic optimization approach. The mean and standard deviation of RMSEP from 25 runs for each dataset are provided in Table 4. Although the hyperparameters are fixed, the trained weights would vary by different model runs due to some random factors such as randomly dropout some weights and stochastic optimization. The repeatability of DeepSpectra approach is improved with increased training samples, in terms of a lower coefficient of variation (Fig. 7). The coefficient of variation of RMSEP for soil dataset with 2502 training samples is only 2.37% while corn dataset with 60 training samples generates a coefficient of variation of 16.67%, followed by 11.43% and 5% with tablet and wheat dataset, respectively (Fig. 7).

The accuracy of DeepSpectra is improved with an increased sample size of a training dataset. The RMSEP by DeepSpectra approach reduces from 11.43 to 8.88 with the increasing number of samples in the training set of the soil dataset (Fig. 8). This indicates

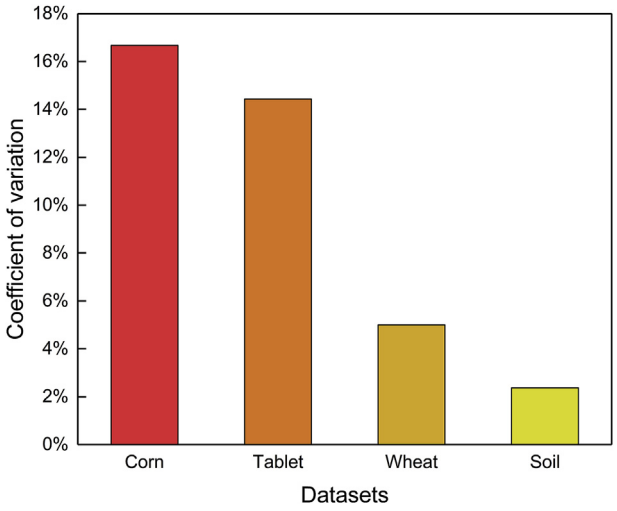


Fig. 7. Coefficient of variation by DeepSpectra model on the raw data for four datasets.

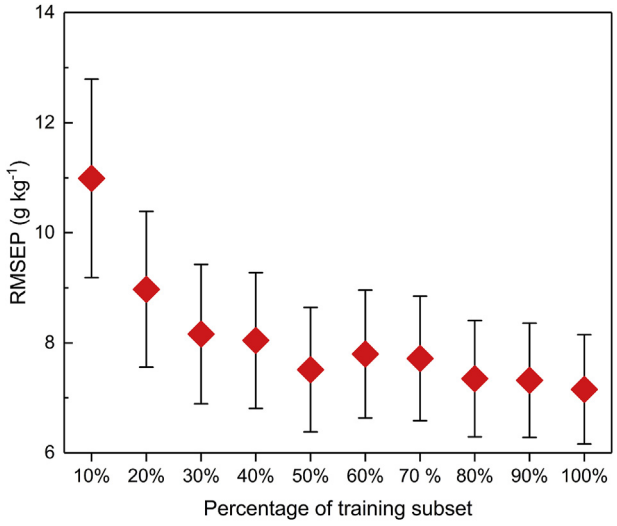


Fig. 8. DeepSpectra model performance of 25 repeated runs with a sample size from 10% to 100% randomly selected from original training set. A mean RMSEP and 90% confidence interval are presented for each scenario.

that the increased sample number would improve DeepSpectra model to learn patterns from various data sources. The magnitude of improvement, however, is not significant after using 80% of the original training dataset. The coefficient of variation of RMSEP is 11.18% with 10% samples (250 samples) used in the training set for soil data, which is much higher than 2.54% with 20% (500) samples used. DeepSpectra model shows a relatively high and stable repeatability with a coefficient of variation is around 2%, with more than 750 samples used in the training set.

4. Conclusion and future work

DeepSpectra model is developed to learn patterns from raw spectra without the need for data preprocessing and dimensional reduction. This study provides a detailed introduction of a deep learning approach for quantitative spectral analysis. The model includes three convolutional layers incorporated with Inception module to learn patterns from various spectra variables. Using four public spectral datasets, the results show that DeepSpectra

approach outperforms other CNN models on the raw data. The including of preprocessing methods does not improve DeepSpectra model performance for most scenarios. DeepSpectra model on raw spectra data outperforms the conventional linear (PLS) and nonlinear (PCA-ANN and SVR) approaches using raw and pre-processed data in most scenarios. DeepSpectra model provides a low RMSEP on small sample size dataset, but it shows poor stability with high coefficient of variation. The increased sample numbers can improve the DeepSpectra model repeatability and accuracy.

Model repeatability and interpretation capability are critical for the further development of deep learning based spectral analysis. It will be critical to understand the minimum and recommended requirement of sample size for deep learning based spectral analysis from model accuracy and repeatability perspective in future studies. It will also be interesting to understand and visualize the identification of critical spectral features related to component analysis. The improved understanding of mechanism of deep learning might be helpful for global analysis of component concentration among various products, such as protein content of grains and soluble solids content of fruits.

Declaration of interests

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.

Acknowledgements

This work was partially funded by the Thousand Young Talents Program of China and Zhejiang University. The authors would also like to acknowledge Chiyu Wu and anonymous reviewers for their helpful comments and suggestions, which substantially improved the quality of the paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.aca.2019.01.002.

References

[1] G. Binetti, L. Del Coco, R. Ragone, S. Zelasco, E. Perri, C. Montemurro, R. Valentini, D. Naso, F.P. Fanizzi, F.P. Schena, Cultivar classification of Apulian olive oils: use of artificial neural networks for comparing NMR, NIR and merceological data, Food Chem. 219 (2017) 131–138, https://doi.org/10.1016/j.foodchem.2016.09.041.

[2] E.M. Hetrick, Z. Shi, L.E. Barnes, A.W. Garrett, R.G. Rupard, T.T. Kramer, T.M. Cooper, D.P. Myers, B.C. Castle, Development of near infrared spectroscopy-based process monitoring methodology for pharmaceutical continuous manufacturing using an offline calibration approach, Anal. Chem. 89 (2017) 9175–9183, https://doi.org/10.1021/acs.analchem.7b01907.

[3] A. Palou, A. Miró, M. Blanco, R. Larraz, J.F. Gómez, T. Martínez, J.M. González, M. Alcalà, Calibration sets selection strategy for the construction of robust PLS models for prediction of biodiesel/diesel blends physico-chemical properties using NIR spectroscopy, Spectrochim. Acta Part A Mol. Biomol. Spectrosc. 180 (2017) 119–126, https://doi.org/10.1016/j.saa.2017.03.008.

[4] A. Gredilla, S. Fdez-Ortiz de Vallejuelo, N. Elejoste, A. de Diego, J.M. Madariaga, Non-destructive Spectroscopy combined with chemometrics as a tool for Green Chemical Analysis of environmental samples: a review, TrAC Trends Anal. Chem. 76 (2016) 30–39, https://doi.org/10.1016/j.trac.2015.11.011.

[5] Y. Bi, K. Yuan, W. Xiao, J. Wu, C. Shi, J. Xia, G. Chu, G. Zhang, G. Zhou, A local pre-processing method for near-infrared spectra, combined with spectral segmentation and standard normal variate transformation, Anal. Chim. Acta 909 (2016) 30–40, https://doi.org/10.1016/j.aca.2016.01.010.

[6] J. Gerretzen, E. Szymańska, J. Bart, A.N. Davies, H.J. van Manen, E.R. van den Heuvel, J.J. Jansen, L.M.C. Buydens, Boosting model performance and interpretation by entangling preprocessing selection and variable selection, Anal. Chim. Acta 938 (2016) 44–52, https://doi.org/10.1016/j.aca.2016.08.022.

[7] 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, https://doi.org/10.1016/j.trac.2013.04.015.

[8] Å. Rinnan, F. van den Berg, S.B. Engelsen, Review of the most common pre-processing techniques for near-infrared spectra, TrAC Trends Anal. Chem. 28 (2009) 1201–1222, https://doi.org/10.1016/j.trac.2009.07.007.

[9] W. Ni, L. Nørgaard, M. Mørup, Non-linear calibration models for near infrared spectroscopy, Anal. Chim. Acta 813 (2014) 1–14, https://doi.org/10.1016/j.aca.2013.12.002.

[10] M. Sun, D. Zhang, L. Liu, Z. Wang, How to predict the sugariness and hardness of melons: a near-infrared hyperspectral imaging method, Food Chem. 218 (2017) 413–421, https://doi.org/10.1016/j.foodchem.2016.09.023.

[11] F. Allegrini, A.C. Olivieri, Sensitivity, prediction uncertainty, and detection limit for artificial neural network calibrations, Anal. Chem. 88 (2016) 7807–7812, https://doi.org/10.1021/acs.analchem.6b01857.

[12] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444, https://doi.org/10.1038/nature14539.

[13] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: 2015 IEEE Conf. Comput. Vis. Pattern Recognit, IEEE, 2015, pp. 1–9, https://doi.org/10.1109/CVPR.2015.7298594.

[14] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conf. Comput. Vis. Pattern Recognit, IEEE, 2016, pp. 770–778, https://doi.org/10.1109/CVPR.2016.90.

[15] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei, Large-scale video classification with convolutional neural networks, in: 2014 IEEE Conf. Comput. Vis. Pattern Recognit, IEEE, 2014, pp. 1725–1732, https://doi.org/10.1109/CVPR.2014.223.

[16] M.H.S. Segler, M. Preuss, M.P. Waller, Planning chemical syntheses with deep neural networks and symbolic AI, Nature 555 (2018) 604–610, https://doi.org/10.1038/nature25978.

[17] D. Silver, A. Huang, C.J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, D. Hassabis, Mastering the game of Go with deep neural networks and tree search, Nature 529 (2016) 484–489, https://doi.org/10.1038/nature16961.

[18] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, D. Hassabis, Mastering the game of Go without human knowledge, Nature 550 (2017) 354–359, https://doi.org/10.1038/nature24270.

[19] J. Acquarelli, T. van Laarhoven, J. Gerretzen, T.N. Tran, L.M.C. Buydens, E. Marchiori, Convolutional neural networks for vibrational spectroscopic data analysis, Anal. Chim. Acta 954 (2017) 22–31, https://doi.org/10.1016/j.aca.2016.12.010.

[20] E.J. Bjerrum, M. Glahder, T. Skov, Data Augmentation of Spectral Data for Convolutional Neural Network (CNN) Based Deep Chemometrics, 2017, pp. 1–10. http://arxiv.org/abs/1710.01927.

[21] S. Malek, F. Melgani, Y. Bazi, One-dimensional convolutional neural networks for spectroscopic signal regression, J. Chemom. 32 (2018), https://doi.org/10.1002/cem.2977 e2977.

[22] C. Yuanyuan, W. Zhibin, Quantitative analysis modeling of infrared spectroscopy based on ensemble convolutional neural networks, Chemometr. Intell. Lab. Syst. 181 (2018) 1–10, https://doi.org/10.1016/j.chemolab.2018.08.001.

[23] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (2014) 1929–1958, https://doi.org/10.1214/12-AOS1000.

[24] A.L. Maas, A.Y. Hannun, A.Y. Ng, Rectifier nonlinearities improve neural network acoustic models, Proc. 30 Th Int. Conf. Mach. Learn. 28 (2013) 6.

[25] S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift, in: Int. Conf. Int. Conf. Mach. Learn, 2015, pp. 448–456, https://doi.org/10.1007/s13398-014-0173-7.2.

[26] D.P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, vols. 1–15, 2014. http://doi.acm.org.ezproxy.lib.ucf.edu/10.1145/1830483.1830503.

[27] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: surpassing human-level performance on ImageNet classification, in: 2015 IEEE Int. Conf. Comput. Vis, IEEE, 2015, pp. 1026–1034, https://doi.org/10.1109/ICCV.2015.123.

[28] J. Gerretzen, E. Szymańska, J.J. Jansen, J. Bart, H.-J. van Manen, E.R. van den Heuvel, L.M.C. Buydens, Simple and effective way for data preprocessing selection based on design of experiments, Anal. Chem. 87 (2015) 12096–12103, https://doi.org/10.1021/acs.analchem.5b02832.

[29] B. Igne, P. Berzaghi, D. Bu, P. Dardenne, P. Tillmann, M. Westerhaus, Summary of the 2012 IDRC software shoot-out, NIR News 23 (2012) 13–15, https://doi.org/10.1255/nirn.1331.

[30] C. Brenner, R. Pierce, D. Funk, "The good, the bad, and the ugly": 2008 IDRC software Shootout, NIR News 20 (2009) 12–15, https://doi.org/10.1255/nirn.1137.

[31] D.J. Brown, K.D. Shepherd, M.G. Walsh, M. Dewayne Mays, T.G. Reinsch, Global soil characterization with VNIR diffuse reflectance spectroscopy, Geoderma 132 (2006) 273–290, https://doi.org/10.1016/j.geoderma.2005.04.025.

[32] H. Motulsky, Intuitive Biostatistics: a Nonmathematical Guide to Statistical Thinking, fourth ed., Oxford University Press, 2014.

[33] X. Bian, P. Diwu, C. Zhang, L. Lin, G. Chen, X. Tan, Y. Guo, B. Cheng, Robust boosting neural networks with random weights for multivariate calibration of complex samples, Anal. Chim. Acta 1009 (2018) 20–26, https://doi.org/10.1016/j.aca.2018.01.013.