Food Control

# USE OF CONVOLUTIONAL NEURAL NETWORK (CNN) COMBINED WITH FT-NIR SPECTROSCOPY TO PREDICT FOOD ADULTERATION: A CASE STUDY ON COFFEE

--Manuscript Draft--

| | |
|---|---|
| **Manuscript Number:** | |
| **Article Type:** | Research Paper |
| **Keywords:** | Deep learning; artificial intelligence; espresso grounds; admixtures; economically motivated frauds |
| **Corresponding Author:** | Roberto Moscetti<br>Università degli Studi della Tuscia<br>Viterbo, Lazio ITALY |
| **First Author:** | Swathi Sirisha Nallan Chakravartula |
| **Order of Authors:** | Swathi Sirisha Nallan Chakravartula |
| | Roberto Moscetti |
| | Giacomo Bedini |
| | Riccardo Massantini |
| **Abstract:** | Food systems are negatively affected by food frauds with food recalls challenging the system's sustainability and consumer confidence in food safety. Coffee, an economically important commodity is frequently adulterated for economic gains, thereby requiring fast and reliable detection techniques. Of the various tracing strategies, spectroscopic techniques have seen considerable commercial success but rely heavily on human-engineered features. Thus, this study aims to evaluate feasibility of deep chemometrics (i.e., convolutional neural network, CNN ) for coffee adulterant quantification in comparison to standard chemometrics approaches (i.e., partial least squares, PLS; and interval-PLS, iPLS ). Commercial 'espresso' coffee was admixed with chicory, barley, and maize (0 – 25 %, w/w) and subjected to Fourier Transformed-Near Infrared (FT-NIR) spectral analysis. The results confirmed the feasibility of CNN algorithm for adulterant quantification from FT-NIR spectra with excellent performances ($R^2 > 0.98$). Moreover, CNN with Autoscaling pre-treatment showed better prediction performances with lower BIAS values than PLS or iPLS models. The study showed that deep learning algorithms can be potential alternatives to standard methods with little to no human interference for feature extraction during real-time applications of spectroscopic tools targeted to overcome food fraud crisis. |
| **Suggested Reviewers:** | Ron Haff<br>ron.haff@usda.gov |
| | Peishih Liang<br>peishih.liang@usda.gov |
| | Pietro Ibba<br>Pietro.Ibba@natec.unibz.it |

**DIBAF**
*DEPARTMENT FOR INNOVATION IN BIOLOGICAL, AGRO-FOOD AND FOREST SYSTEM*

*Via S. Camillo De Lellis - 01100 Viterbo*

August 03, 2021

Dear Editor,

Kindly find enclosed the manuscript titled, 'USE OF CONVOLUTIONAL NEURAL NETWORK (CNN) COMBINED WITH FT-NIR SPECTROSCOPY TO PREDICT FOOD ADULTERATION: A CASE STUDY ON COFFEE', being submitted for consideration to be published in Food Control journal.

The said manuscript deals with the novel aspect of implementation of CNN, a deep chemometrics algorithm in coffee adulteration quantification in combination with FT-NIR spectral data. This deep learning algorithm even though gaining attention in image-based data elaboration, has not been implemented so far for 1D and 2D spectral data, specifically for coffee adulteration. In this regard, relevant existing articles on coffee adulteration and use of spectroscopy have been sufficiently cited to present the state-of -art and identify the knowledge gap. The authors believe strongly that this article serves to demonstrate the potential use of computational advances and the less experimented paradigms for researchers to probe, apart from demonstrating the advantages and feasibility of a CNN algorithm for potential simplification and automation for food fraud detection and quantification.

We hope that the manuscript is suitable for publication in your journal. We look forward for your response and would be glad to answer any further questions and comments that you may have.

On behalf of all authors,
Dott. Roberto Moscetti*
Prof. Riccardo Massantini**
DIBAF – Department for Innovation in Biological, Agro-food and Forest systems
Tuscia University, Viterbo, Italy.
*Tel.: +39 0761 357539
*E-mail: rmoscetti@unitus.it
**Tel.: +39 0761 357496
**E-mail: massanti@unitus.it

# USE OF CONVOLUTIONAL NEURAL NETWORK (CNN) COMBINED WITH FT-NIR SPECTROSCOPY TO PREDICT FOOD ADULTERATION: A CASE STUDY ON COFFEE

Swathi Sirisha Nallan Chakravartula, Roberto Moscetti*, Giacomo Bedini, Riccardo Massantini**

Department for Innovation in Biological, Agro-food and Forest systems, University of Tuscia, Via S. Camillo de Lellis snc, 01100, Viterbo, Italy

*, **Corresponding authors: Tuscia University, Department for Innovation in Biological, Agro-food and Forest systems, Via S. Camillo De Lellis snc, 01100, Viterbo, Italy. Tel.: +39 0761 357539; e-mail addresses: *rmoscetti@unitus.it (Moscetti, R.); **massanti@unitus.it (Massantini, R.).

## ABSTRACT

Food systems are negatively affected by food frauds with food recalls challenging the system's sustainability and consumer confidence in food safety. Coffee, an economically important commodity is frequently adulterated for economic gains, thereby requiring fast and reliable detection techniques. Of the various tracing strategies, spectroscopic techniques have seen considerable commercial success but rely heavily on human-engineered features. Thus, this study aims to evaluate feasibility of deep chemometrics (i.e., convolutional neural network, CNN) for coffee adulterant quantification in comparison to standard chemometrics approaches (i.e., partial least squares, PLS; and interval-PLS, iPLS). Commercial 'espresso' coffee was admixed with chicory, barley, and maize (0 – 25 %, w/w) and subjected to Fourier Transformed-Near Infrared (FT-NIR) spectral analysis. The results confirmed the feasibility of CNN algorithm for adulterant quantification from FT-NIR spectra with excellent performances ($R^2$ > 0.98). Moreover, CNN with Autoscaling pre-treatment showed better prediction performances with lower BIAS values than PLS or iPLS models. The study showed that deep learning algorithms can be potential alternatives to standard methods with little to no human interference for feature extraction during real-time applications of spectroscopic tools targeted to overcome food fraud crisis.

**KEYWORDS**

Deep learning; artificial intelligence; espresso grounds; admixtures; economically motivated frauds

**ABBREVIATIONS**

| | |
|---|---|
| Adj-$R^2$ | Adjusted Coefficient of Determination |
| ANOVA | Analysis of Variance |
| AS | Autoscaling |
| C | Calibration |
| CV | Cross-Validation |
| P | Prediction |
| CNN | Convolutional Neural Network |
| CSV | Comma-Separated-Values |
| EMA | Economically Motivated Adulteration |
| EU | European Union |
| FEA | Feature extraction algorithms |
| iPLS | Interval Partial Least Squares |
| ISTAT | Italian National Statistical Institute |
| MC | Mean Centering |
| MSC | Multiplicative Scatter Correction |
| PC | Principal Components |
| PCA | Principal Component Analysis |
| PLS | Partial Least Squares |
| RMSE | Root Mean Square Error |
| SNV | Standard Normal Variate |
| t-SNE | t-distributed Stochastic Neighbour Embedding |
| ReLU | Rectified Linear Unit |
| MSE | Mean Square Error |
| LV | Latent Variable |
| EMSC | Extended Multiplicative Scatter Correction |
| DA | Data Augmentation |

## 1. INTRODUCTION

Food fraud is an ever-growing challenge for the food industry and can be broadly explained as per the EU (European union) regulation 2017/625 as fraudulent or deceptive activities unintentionally or intentionally carried out by an individual or business in violation of food laws to have an undue advantage in the market (European Comission, 2020; European commission, 2009). These activities even though economically motivated have potential implications at food safety and public health. In fact, EU food fraud network in their recent publication considers intentional adulteration to be '*ideologically motivated to harm through bio-terrorism*' (European Comission, 2020).

These implications make it ever-so important to recognize and distinguish the type of fraudulent activities not only to restore consumer confidence but also to improve transparency in food safety control (Moore et al., 2012). Recently, combating food fraud has evolved as a key priority for the EU, particularly in foods of high commercial value such as olive oil, fish, meat, honey, coffee, tea, spices, and milk (European Comission, 2020; European commission, 2009). Of the listed products, coffee is a ubiquitous beverage with a fast market growth in the food trade registering a global consumption of 166 million bags (60 kg) for the year 2020-21 (ICO-January, 2021). Among different regions, Europe registered a consumption of 54 million 60 Kg bags for the year 2020-21 with Italy ranking second in terms of the total consumption. In practice, according to Italian National Statistical Institute (ISTAT) the Italian coffee market is occupied mainly by espresso coffee (~70 %) with home consumption accounting for 87 % of which roasted ground coffee occupies ~83 % of the market segment (ICO-January, 2021; Italiaimballaggio, 2020)

Roasted coffee is principally obtained from pure and/ or blends of the two economically important species: *Coffea arabica* (Arabica) and *Coffea canephora* (Robusta) (Barbin et al., 2014; Illy & Viani, 2005). The roasting and milling processes makes it a vulnerable target to undeclared mixtures (Barbin et al., 2014; Dankowska et al., 2017; Toci et al., 2016). In fact, only coffee substitutes declared on the label like chicory are permitted (Morin & Lees, 2018) and other vegetal materials i.e., barley, coffee husks, maize, wheat, coffee husks and stems are not allowed. Despite these regulations, vegetal materials have been frequently identified in commercial coffee as part of economically motivated adulteration (EMA) (European Comission, 2020; Souto et al., 2015; Tavares et al., 2016). EMAs apart from resulting in regulation non-compliance (both quality and safety) cause potential health risk to consumers (European commission, 2009; Moore et al., 2012) thereby

necessitating coffee authentication to ensure the product safety, integrity as well as true product price (Ebrahimi-Najafabadi et al., 2012).

Several methods and strategies have been investigated over the years to detect and verify the authenticity, adulteration and quality of coffee which can be broadly classified as subjective (sensorial) and objective (physical, chemical, physico-chemical, biological) methods (Combes et al., 2018; Esteban-Díez et al., 2004; Ferreira et al., 2016; Pauli et al., 2014; Tavares et al., 2016; Toci et al., 2016; Uncu & Uncu, 2018). Most of these techniques usually require expertise and involve wet chemistry that is environmentally harmful, labour-intensive and time consuming. Among the objective physical approaches, near and mid-infrared spectroscopies gained considerable interest, particularly for detection of coffee adulteration as a rapid, low-cost, and non-destructive method alone or in combination with other reference techniques (Toci et al., 2016). These approaches are being explored widely to detect quality of green and roasted beans (Caporaso et al., 2018; Oliveri et al., 2019), admixtures with different species (Dankowska et al., 2017; Esteban-Díez et al., 2007), adulteration with corn, barley, chicory, husk and spent coffee (Pauli et al., 2014; Reis et al., 2013a, 2013b; Winkler-Moser et al., 2015).

Within coffee adulteration, different researchers have used spectroscopic techniques in mid to near-infrared zones (Ebrahimi-Najafabadi et al., 2012; Reis et al., 2013a, 2013b) combined with classical chemometrics. Although advantageous, classical approaches are frequently criticized for their requirement of expertise and subjectivity in elaborating the complex spectral data (Acquarelli et al., 2017). To overcome some of these issues, artificial neural network (ANN) approaches have been used to improve model accuracy but were observed to be complicated structures with evident risks of over-fitting. Moreover, these strategies although robust and accurate, have drawbacks related to efforts and practical feasibility (Acquarelli et al., 2017; Engel et al., 2013). Consequently, researchers are probing for a shift in the paradigm towards the application of deep neural network (DNN) algorithms for resolving the issues related to classical approaches. One of the state-of-art approach of considerable interest is convolutional neural networks (CNN) implemented for recognition, classification, identification of images and spectroscopic data (Acquarelli et al., 2017; Malek et al., 2018). The CNN algorithms in fact can also be used for regression tasks wherein this supervised approach can perform both extraction and learning related to features of interest (Cui & Fearn, 2018; Mishra & Passos, 2021a). Furthermore, the use of CNN can simplify the task of pre-processing and model building through transfer learning thereby improving the speed of food fraud identification and quantification.

To the best of our knowledge, even though coffee adulteration was investigated, no study focused on the use of neural networks as advanced computational algorithms for quantification of adulterants by FT-NIR spectroscopy. Thus, considering food fraud as one of the important areas of food safety and control that requires rapid and automated methods for detection, identification and quantification of adulterants, the objective of this study was to (i) investigate the potential use of CNN models for quantifying adulterants in coffee and (ii) compare their performance against the standard chemometrics approach.

## 2. MATERIALS AND METHODS

### 2.1. Adulterated coffee preparation

Commercially available roasted, ground coffee (*Coffea arabica*), barley for espresso, roasted chicory nibs for espresso and whole corn kernels were acquired from the local market (Viterbo, Italy). The corn kernels were roasted at 200 °C for 10 min in a convection oven (Smeg, mod. CX60SVPZ9). The roasted corn kernels, barley and chicory nibs were ground in a laboratory scale mill (MF 10, IKA®-Werke, Germany) using a sieve of 0.25 mm at minimum speed to reduce heat production. The sampling scheme consisted of three classes of adulterated samples as follows: 'coffee + chicory', 'coffee + barley' and 'coffee + maize'. Each adulterant was added at 0, 2.5, 5, 10, 15, 20, 25 % (w/w) concentration levels. Each sample consisted of 5 biological replicates x 5 technical replicates with predetermined quantities of coffee and adulterants that were prepared by carefully weighing and mechanically mixing for 1 minute.

### 2.2. Spectral data acquisition

The spectral acquisition was carried out using a Fourier Transform-NIR spectrophotometer mod. Antaris II (Thermo Fisher Scientific, WI, USA) operating in the spectral range of 10000 to 4000 $cm^{-1}$ (1000 - 2500 nm) with a resolution of 4 $cm^{-1}$. Six grams of sample were weighed, spread evenly, and compacted into a 12-cm rotary-cup spinner accessory. The spectra were acquired at room temperature (20 ± 1 °C) in diffuse reflectance mode and directly converted to absorbance by means of RESULT 3 software (Thermo Fisher Scientific, WI, USA). Each spectrum had an average of 32 interferometer sub-scans, with the internal instrument standard as reference. A dataset of 550 samples composed of pure and adulterated coffee was acquired. Subsequently, spectra for each adulteration level were randomly split between calibration set (70 % spectra) and prediction set (30 % spectra).

### 2.3. Spectral pre-treatment and data augmentation

This study investigated two different approaches to quantify and predict the adulterant levels: (i) a (standard) chemometrics linear approach by both Partial Least Squares (PLS) and

interval-PLS (iPLS) algorithms and (ii) a (novel) deep chemometrics non-linear approach by convolutional neural network (CNN).

Prior to data exploration and standard chemometrics, the spectra were subjected to spectral pre-processing to reduce non-linearities, scattering effects and noise. For the intended purpose, the following pre-processing treatments were applied and tested for their effectiveness: (i) Standard Normal Variate (SNV), (ii) Multiplicative Scatter Correction (MSC), and (iii) Savitzky-Golay algorithm as smoothing and/or differentiation filter. Each smoothing/derivative pre-treatment was applied with a second or third order polynomial fitted over a window of 11, 13 or 15 features. As a final pre-treatment step, spectral data were normalized through mean centering (MC) or autoscaling (AS). Every possible combination of these spectral treatments was tested and only the one improving model performances were retained.

In the case of deep chemometrics, raw spectral dataset was augmented by adding random offset, multiplication, and slope effects using the method proposed by Bjerrum et al., (2017) with the following modifications: offset was ±0.60 times the standard deviation of the calibration set; multiplication was 1±0.60 times the standard deviation of the calibration set; and the slope was randomly adjusted between 0.95 and 1.05. Data augmentation is a technique commonly used to pursue an effective training of neural networks and then circumvent the risk of developing poorly generalized models. Before the CNN training, the augmented dataset was subjected to normalization (i.e., MC or AS) alone or after SNV.

### 2.4. Data exploration and visualization

Feature extraction algorithms (FEAs) were exclusively used as exploratory tools to identify patterns and evaluate pairwise similarities within adulterant types and adulteration levels. Specifically, both Principal Component Analysis (PCA) and t-distributed Stochastic Neighbour Embedding (t-SNE) algorithms were individually applied. Preliminary results from PCA showed its well-known limitations in keeping close the similar representations and scattered results among replications within a limit of three PCs. Thus, prior PCA, spectral data were averaged assuming adulterant type and adulteration levels as grouping elements. Consequently, averaged spectra were subjected to PCA, and the required principal components (PCs) were selected using the scree-plot criterion.

Differently from PCA, t-SNE was successfully computed on the whole dataset. The t-SNE is a non-linear FEA based on manifold learning with usually better performance than PCA for its capability to capture global and local structures of high-dimensional data, as well as to identify the presence of clusters at various scales (Luo et al., 2021). The algorithm initially

reduces high-dimensional Euclidean distances using similarities between two data points into conditional probabilities. Subsequently, the Student's t-distribution is used to further transform the probabilities into a low-dimensional space, using a single degree of freedom. Dimensionality reduction through t-SNE was performed by setting the 'perplexity' and the 'output dimensionality' parameters at 20 and 2, respectively. Specifically, the 'perplexity' parameter refers to the effective number of neighbours used in the t-SNE algorithm; it may affect data visualization if outside the limit of 5-50 (Luo et al., 2021). The 'output dimensionality' corresponds to the number of dimensions of the low-dimensional space in which spectral data are compressed.

### 2.5. Adulterant concentration prediction (model development)

*Chemometrics approach*

Within this approach two extensively used algorithms, PLS and iPLS, were applied. Prediction models were individually developed for each adulterant, using pre-processed spectra as X-matrix and adulterant concentrations (0-25 % w/w) as Y-vectors. PLS was used for prediction of the adulterant concentration based on linear combinations in independent variables. Subsequently, iPLS was tested for feature extraction to potentially improve the regression results. The iPLS algorithm was set to stepwise forward mode with a selection of a maximum of 5 waveband intervals of 10 features (wavelengths) each. Each PLS and iPLS model was cross validated for the correct number of latent variables (LVs) to find the optimal trade-off between under-fitting and over-fitting using a venetian blinds cross-validation with 10 data splits. Model performances were evaluated in terms of Root Mean Square Error, BIAS and coefficient of determination ($R^2$) for calibration (RMSEC, BIASC and $R^2C$, respectively), cross-validation (RMSECV, BIASCV and $R^2CV$, respectively), and prediction (RMSEP, BIASP and $R^2P$) (Moscetti et al., 2019). The Hotelling's t-squared and Q-residuals tests were used in combination to identify potential outliers of each given model. Chemometrics were performed using Matlab environment R2017b (Mathworks Inc., USA) coupled with PLS toolbox v8.6.2 (Eigenvector research Inc., USA). Finally, the best performing PLS and iPLS models per adulterant were selected based on the calibration and prediction performance metrics.

*Deep chemometrics approach*

The Convolutional Neural Networks (CNNs) were developed following the method proposed by Bjerrum et al., (2017). CNNs generally have several layers that enable the algorithm to carry out spectral pattern recognition such as input, convolutional, grouping, fully connected and dense layers. The architecture of neural network used in this study consisted of

two main parts (Table 1): (i) a convolution section deputed to separate and identify features from the FT-NIR spectra (from layer #1 to #6) and (ii) fully connected layers (layers #7 and #8) that use output from the convolutional process to predict adulterant concentration. Specifically, the input (i.e., augmented spectrum) was fed to a regularization layer (layer #1) to apply an additive zero-centered Gaussian noise with a standard deviation of 0.05, as a form of additional random data augmentation. Subsequently, data was reshaped (layer #2) and then convoluted towards two one-dimensional convolutional layers (layers #3 and #4) with kernel sizes of (8, 32) and (16, 32), respectively, and a Rectified Linear Unit (ReLU) activation. Before connection with the two fully connected layers, data was flattened (layer #5) to converge into a one-dimensional array and regularized through a dropout (layer #6), which reduced unnecessary feature dependencies and then improved the generalization abilities of the network. In other words, the dropout layer helped to reduce over-fitting by dropping out random nodes from the network. The last layers were two fully connected dense layers in which the regression process took place. The first dense layer (layer #7) was composed by 128 units (nodes) working with a ReLU activation function, while the second dense layer (layer #8) was a single output node with a linear activation function.

CNNs were built by using Python v3.7.3 coupled with the deep learning library Keras v2.3.1 using Tensorflow v1.14.0 as computation back end. The calibration was performed on a personal computer equipped with an Intel(R) core (TM) i7-8565U CPU at 1.80 GHz clock frequency, 16 GB RAM and GPU Nvidia Geforce MX250 (Pascal GP108). Parallel computing was done using CUDA v.10.1. Each network was implemented for 50 epochs using an early stopping criterion, with batch size of 16 and an initial learning rate of 0.01. The ReduceLROnPlateau () function in Keras, with a ‘patience’ of 5 and a ‘decrease factor’ of 0.5 (both empirically selected), was used to optimize the network during training. The Mean Square Error (MSE) was assumed as a calibration and prediction loss function, to circumvent under-fitting or over-fitting problems through selection of the optimal number of epochs. Finally, network performances were evaluated in terms of RMSE, BIAS and $R^2$ for calibration (C) and prediction (P). Finally, the best CNN per adulterant was selected.

## 3. RESULTS AND DISCUSSION

### 3.1. Spectral reference characteristics and overview

The average spectra obtained for the pure coffee and adulterants (chicory, barley, and maize) were characterized by assigning functional groups based on the published infrared molecular overtones and combinations data as reported in Table 2.

As can be observed from Fig. 1a, the spectral profiles of pure coffee and adulterants were similar and characterized by few substantial differences in peak positions and curve trends. In general, for all the samples considered i.e., pure coffee and adulterants, few characteristic overlapping peaks contributed by the presence of carbohydrates, water, lipids and proteins were observed in the following wavebands: 4200-4400 $cm^{-1}$ related predominantly to O-H and C-H bands; 4700-4800 $cm^{-1}$ to C=O 1$^{st}$ overtones and combination bands of O-H and N-H; 5167-5177 $cm^{-1}$ and 5670-5790 $cm^{-1}$ to resonance bands of O-H and 1$^{st}$ overtone of C-H bonds attributed to carbohydrates (starch), water and other hydrocarbons; 6890-6940 $cm^{-1}$ to the O-H stretch related to water and carbohydrates; and 8280-9340 $cm^{-1}$ to the resonance bands of O-H of water and starch as well as the 2$^{nd}$ overtone of C-H stretching (Barbin et al., 2014; Cozzolino et al., 2013; Ribeiro et al., 2011; Workman & Weyer, 2007) .

Specifically, in pure coffee two distinguishable sharp peaks were identified at 4256 $cm^{-1}$ and 4332 $cm^{-1}$, attributed to the degradation of chlorogenic acid during roasting and the subsequent release of quinic acid with its concentration increasing with higher degree of roast (Alessandrini et al., 2008; Milani et al., 2020). A similar small blunt peak at 4382 $cm^{-1}$ was observed in chicory which can be attributed to probable presence of chlorogenic acid and chicorin (Zawirska-Wojtasiak et al., 2014). Whereas these peaks in range of 4250 $cm^{-1}$ to 4390 $cm^{-1}$ caused by chemical changes during roasting, were not distinguishable in case of barley and corn. Continuing with specifics of coffee, the wavenumbers at 5689 and 5786 $cm^{-1}$ were related to the 1$^{st}$ overtone region; 5175 $cm^{-1}$ to the 2$^{nd}$ overtone of C=O stretching and 4734 $cm^{-1}$ to the 1$^{st}$ overtone of O-H combination bands to the presence of alkaloid caffeine (Ribeiro et al., 2011). In case of chicory, barley and corn, these wavebands are represented by lipids, proteins and carbohydrates (Cozzolino et al., 2012, 2013; Hódsági et al., 2010; Workman & Weyer, 2007) leading to overlapping spectral regions (Fig 1a).

Subsequently PCA as an unsupervised tool was used to visualize patterns on the pre-treated and averaged spectral datasets. Figure 1a represents the two-dimensional score plot with PC1 and PC2 accounting for 85.31 % and 11.40 %, respectively, of the 96.71 % explained total variance. As expected, pure coffee and adulterants were distinctly separated, indicating that FT-NIR effectively identified the differences in their chemical composition. This distinctive disposition of pure coffee might be attributed to the presence of compounds like chlorogenic and quinic acids, and caffeine (Reis et al., 2013a).

Furthermore, PCA was observed to distinguish effectively the samples based on the type of adulterant and adulterant concentration. Among the adulterants, three groups were observed for each type of adulterant with barley samples clearly distinguished along PC2

compared to the closely positioned chicory and maize samples. The close positioning of samples in the bidimensional plot was indicative of similar and overlapping functional groups and observed to be a common phenomenon in agro-food samples owing to their complex chemical interactions and structures by Luo et al. (2021). As for the different concentrations, as expected samples with lower adulterant concentration were spatially closer to pure coffee (black square) with samples moving towards the pure adulterant with increasing adulteration concentration. In fact, a linear behaviour of sample distribution for all types of adulterants was observed, say for example barley adulterated samples at 25 % w/w concentration were positioned closer to pure barley sample (green triangle). On the hindsight it is important to consider that only averaged spectra were used in this analysis as the use of whole dataset resulted in spread out variance with less than 80 % of total variance explained by the first 3 PCs (data not shown).

To better visualize the whole spectral dataset and circumvent the issue of sample overlaps, the t-SNE algorithm was utilized. Figure 1c represents the t-SNE plot wherein the adulterated coffees were effectively grouped based on the adulterant type and concentration. Unlike PCA, each adulterant type was distinctly positioned with barley and maize samples grouped along Y1-axis and chicory samples along both Y1 and Y2 axes. Further, the samples were positioned moving away from the pure coffee in quadrant I towards pure adulterants positioned in quadrants II and III with the increasing adulterant concentration. The use of t-SNE allowed identification of certain patterns in the adulterated samples as follows: (i) in all adulterant types, the clusters at low adulterant concentrations (< 10 %) were observed to be spatially closer with slight overlapping among the replicates which might be due to spectral similarities and can potentially pose challenges for prediction at low levels of admixing; and (ii) in case of chicory a distant group at 10 % w/w concentration in the plot was identified as potential outliers that are to be considered in further analysis. Overall, t-SNE was observed to categorize the samples more closely and consistently based on the extracted features thereby providing a better overview of the complete dataset and indications for further tasks of data elaboration.

### 3.2. Adulterant concentration prediction models

In this study two approaches using chemometrics (i.e., PLS and iPLS) and deep chemometrics (i.e., CNN) were tested and compared.

In chemometrics approach, PLS was used as a full-spectrum approach in which all variables were used for modelling; and iPLS was used as a feature selection approach wherein a sequential search was carried out by the algorithm to select the best variable or combination

thereof based on minimizing the RMSECV using forward selection. In deep chemometrics a one-dimensional CNN, an emerging approach with fully connected (FC) layers was used. This is a completely supervised algorithm which performs both feature extraction and learning related to characteristics of interest (Mishra & Passos, 2021b) . This approach is like PLS in that it utilizes the full spectrum whereas that of iPLS in selecting specific features. However, CNN is unique in its ability to identify and use an optimized filter for feature engineering with little to no human intervention (Cui & Fearn, 2018; Mishra & Passos, 2021b).

Among the chemometrics approaches, PLS was used as a benchmark to select the pre-treatments and identify the possible outliers. Subsequently, among the various evaluated pre-treatments (data not shown), only those models with simplest pre-treatment combinations, lower number of latent variables and better performances i.e., AS and SNV+AS were retained and further discussed. Furthermore, in case of chicory adulterated samples, two outliers as identified in t-SNE plot (Fig. 1c) were confirmed based on the Q-residuals and the Hotelling's $T^2$ values and excluded from the regression analysis.

Figure 2 shows an overview of the prediction models of the adulterant concentrations for the selected pre-treatments (i.e., AS and SNV+AS). The proposed sub-plots were built on average predictands to limit superimposed dots and, thus, obtain cleaner representations. It can be observed from the figure that both the pre-treatments performed well for the algorithms and adulterants chosen. However, the use of CNN on AS pre-treated spectra seemed to predict better for the tested adulterants with values closer to the ideal regression line (grey dotted line) compared to standard chemometrics. Whereas the use of SNV+AS even though improved the performances of PLS and iPLS models, negatively affected the CNN models, except in case of barley. It can be clearly seen in Fig. 2d and 2f that the predictands in chicory and maize for CNN compared to PLS and/or iPLS models were visibly away from the ideal line indicative of potential bias. However, it must be noted that in all cases of graphical representation the variability within each predictand was masked due to the use of averaged values.

The regression models can be evaluated in depth from Table 3 detailing the performance metrics for the selected approaches. The complete calibration and prediction performances show that all the tested algorithms (PLS, iPLS and CNN) resulted in similar models with good to excellent prediction performances for all the tested adulterants i.e., chicory, barley, and maize with $R^2P \geq 0.96$ and RMSE values comparable to those of calibration errors in range of $0.64 \leq RMSEP \leq 1.47$. In case of PLS and iPLS algorithms, the optimal model complexity ranged from 2-4 latent variables. No specific reduction of LVs in case of iPLS models was observed indicating feasibility of feature selection for modelling of

adulterant concentrations. As for the deep chemometrics model, the validation loss in terms of RMSE was used to select the optimal number of epochs to build the model. The optimal epochs were considered when ideally a low difference between calibration and validation loss in the training was observed to prevent under-and /or over-fitting.

Among the pre-treatments, the use of SNV+AS in general compared to AS pre-treatment seemed to improve the model performances for PLS and iPLS models whereas decreased the performances in CNN models, except in case of barley. To further elucidate, in PLS models of chicory adulterated samples the use of SNV+AS decreased the RMSEP to 1.00 and BIASP to $-1.70 \times 10^{-1}$ compared to AS pre-treatment with RMSEP of 1.39 and BIASP of $-3.06 \times 10^{-1}$, respectively. Similar performances were observed also in case of iPLS models, whereas for CNN models AS pre-treatment showed better performances (RMSEP: 0.76; BIASP: $1.00 \times 10^{-2}$) than the SNV+AS pre-treatment (RMSEP: 1.06; BIASP: $5.90 \times 10^{-1}$). Similar behaviour was observed also in case of maize adulterated samples with use of SNV improving the PLS (RMSEP: 0.64, BIASP: $-1.33 \times 10^{-1}$) and iPLS (RMSEP: 0.73, BIASP: $-7.40 \times 10^{-2}$) models whereas for CNN model the use of only AS registered comparable performances (RMSEP: 0.82; BIASP: $1.00 \times 10^{-1}$). As for the case of barley adulterated samples, it was interesting to observe that SNV+AS resulted in superior metrics, particularly lower BIASP (PLS: $3.36 \times 10^{-1}$, iPLS: $2.01 \times 10^{-1}$, CNN: $-4.00 \times 10^{-1}$) values compared to AS pre-treatment (PLS: $3.58 \times 10^{-1}$, iPLS: $7.11 \times 10^{-1}$, CNN: $1.10 \times 10^{-1}$). Furthermore, in terms of comparison among the models, the CNN models had comparable performances in terms RMSE, and $R^2$ values whereas considering BIASP values exhibited superior performances to that of classical chemometrics (i.e., PLS, iPLS).

These differential performances of CNN models with different pre-treatments as also observed in figures 2d-f can be explained by the effects of combining data augmentation with the spectral pre-treatments (Bjerrum et al., 2017). As observed in previous discussion the predictive capacity with the use of SNV+DA deteriorated in chicory and maize adulteration predictions, whereas was improved in barley adulteration prediction. This might be related to the fact that DA principally involves in addition of random offset to simulate scatter effects (as in section 2.4) which counteracts the removal of scatter effects by SNV as observed in case of chicory and maize. Whereas, in case of barley samples these opposing effects might have allowed the neural networks to focus on features represented by the residual variations not affected by pre-treatments in the spectrum as also explained by Bjerrum et al., (2017) in use of extended multiplicative scatter correction with data augmentation (EMSC+DA). This indicates

that pre-processing and data augmentation can have a synergistic or antagonistic effect on the models and should be evaluated on a case-to-case basis.

Further, the regression vector plots of the PLS models were evaluated to identify the most contributing wavebands and verify those corresponding to the wavebands qualitatively identified. As can be observed from Figure 3a-c, in general, the ß-coefficient was used as threshold to identify the positioning of peaks (i.e., positively, or negatively related to the prediction). The peaks with high ß-coefficients are generally the most contributing bands among which the important peaks corresponding to the functional groups (ref. Table 2) were found, indicative of the effective role of the identified components in the regression models.

In case of chicory adulteration among the most contributing bands the wavenumbers 6890, 6431, 4739 and 4332/ 4382 $cm^{-1}$ relative to water, carbohydrates, proteins, caffeine and chlorogenic acid seemed to positively contribute whereas wavenumbers 5774 and 5167 $cm^{-1}$ corresponding to hydrocarbons (lipids), carbohydrates and water seemed to negatively contribute to the PLS model. Of these wavenumbers, 10 bands within the region of 4870 - 4350 $cm^{-1}$ (Table 3) were selected by the iPLS algorithm confirming the major contribution of chlorogenic and caffeine components to the models. Similarly, in barley adulterated coffee all the identified peaks for pure barley except 6335 $cm^{-1}$ (Table 2) were observed to be among the most contributing bands for PLS regression. Whereas only bands falling within the range of 4715 – 4080 $cm^{-1}$ were selected as intervals (Table 3) and were identified to be related to C-H + C-H, N-H and O-H combination bands of starch in barley and chlorogenic acids, caffeine in coffee. Finally, in maize adulterated coffee only the wavenumbers 6939, 5782, 5175, 4764 $cm^{-1}$ related to water, starch, lipids, and protein of maize that overlap with water, lipid, and caffeine bands of coffee (Table 2) were the most contributing peaks to the model. Furthermore, the selection of bands within 5740 – 4000 $cm^{-1}$ by the iPLS model confirm the role of the said groups in addition to chlorogenic acids to the model. Overall, the regression vectors and the intervals selected allowed us to observe that regions related to caffeine, chlorogenic acids and carbohydrates were the major modelling factors in chemometrics. Further, these observations can allow to hypothesize that these identified wavebands were also the features extracted and selected by CNN (Cui & Fearn, 2018).

However further investigations on the loadings and CNN black box which was out of the scope of present work can provide better insights into the contributing features and their significance in the models.

## 4. CONCLUSIONS

The presented study evaluated the feasibility of applying emerging convolutional neural networks (CNNs) in comparison to classical chemometrics algorithms (PLS, iPLS) for predicting coffee adulteration. The FT-NIR spectral data were acquired in the range of 4000 – 10000 $cm^{-1}$ on coffee adulterated with chicory, barley, and maize. The qualitative evaluation of the pure spectra provided insights on the important peaks and functional groups that might overlap and present challenges for regression tasks. Subsequently, the use of t-SNE allowed for better visualization and identification of potential outliers in each adulterant group compared to the principal component analysis (PCA). Based on the qualitative analysis, partial least squares regression was used to screen spectral pre-processing methods and remove the outliers. CNN algorithm in general presented excellent performances comparable to those of classical chemometrics approaches (PLS, iPLS). Further, AS-based CNN models presented superior performances for all the adulterants tested, particularly in terms of BIAS with lower under-/over-fitting. Also, it was observed that the use of deep learning algorithm needs to be evaluated on case-to-case basis for exploring the advantages of counterintuitive approaches (i.e., use of data augmentation in combination with spectral pre-processing). Moreover, the PLS regression vectors confirmed the contribution of identified functional groups such as chlorogenic acid, caffeine, starch, and lipids to the PLS and iPLS models, allowing to hypothesize the features probably selected for CNN. It would be of interest to extend investigations on the CNN as a spectral pre-processing and structure of the neural network black box to improve the model robustness and transferability. Overall, CNN, a deep learning algorithm was observed to be a feasible and practical alternative to classical chemometrics for coffee adulterant quantification wherein its feature and transfer learning advantages can be exploited for real-time applications in tools targeted to overcome food fraud crisis.

## 5.0 ACKNOWLEDGMENTS

The authors gratefully acknowledge [1] the ‘Departments of excellence 2018’ program (i.e., ‘Dipartimenti di eccellenza’) of the Italian Ministry of Education, University and Research for the financial support through the ‘Landscape 4.0 food, wellbeing and environment’ (DIBAF department of University of Tuscia). Moreover, our sincere thanks to the master student Linda Monesi for the valuable help and support during the experimentation.

## DECLARATIONS

The authors declare no conflict of interest.

## DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author, upon reasonable request.

## REFERENCES


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

Alessandrini, L., Romani, S., Pinnavaia, G., & Rosa, M. D. (2008). Near infrared spectroscopy: An analytical tool to predict coffee roasting degree. *Analytica Chimica Acta*, *625*(1), 95–102. https://doi.org/10.1016/j.aca.2008.07.013

Barbin, D. F., Felicio, A. L. de S. M., Sun, D. W., Nixdorf, S. L., & Hirooka, E. Y. (2014). Application of infrared spectral techniques on quality and compositional attributes of coffee: An overview. In *Food Research International* (Vol. 61, pp. 23–32). Elsevier Ltd. https://doi.org/10.1016/j.foodres.2014.01.005

Bjerrum, E. J., Glahder, M., & Skov, T. (2017). *Data Augmentation of Spectral Data for Convolutional Neural Network (CNN) Based Deep Chemometrics*.

Caporaso, N., Whitworth, M. B., Grebby, S., & Fisk, I. D. (2018). Rapid prediction of single green coffee bean moisture and lipid content by hyperspectral imaging. *Journal of Food Engineering*, *227*, 18–29. https://doi.org/10.1016/j.jfoodeng.2018.01.009

Combes, M. C., Joët, T., & Lashermes, P. (2018). Development of a rapid and efficient DNA-based method to detect and quantify adulterations in coffee (Arabica versus Robusta). *Food Control*, *88*, 198–206. https://doi.org/10.1016/j.foodcont.2018.01.014

Cozzolino, D., Roumeliotis, S., & Eglinton, J. (2012). Relationships between Swelling Power, Water Solubility and Near-Infrared Spectra in Whole Grain Barley: A Feasibility Study. *Food and Bioprocess Technology 2012 6:10*, *6*(10), 2732–2738. https://doi.org/10.1007/S11947-012-0948-9

Cozzolino, D., Roumeliotis, S., & Eglinton, J. (2013). Exploring the Use of Near Infrared (NIR) Reflectance Spectroscopy to Predict Starch Pasting Properties in Whole Grain Barley. *Food Biophysics 2013 8:4*, *8*(4), 256–261. https://doi.org/10.1007/S11483-013-9298-Z

Cui, C., & Fearn, T. (2018). Modern practical convolutional neural networks for multivariate regression: Applications to NIR calibration. *Chemometrics and Intelligent Laboratory Systems*, *182*, 9–20. https://doi.org/10.1016/j.chemolab.2018.07.008

Dankowska, A., Domagała, A., & Kowalewski, W. (2017). Quantification of Coffea arabica and Coffea canephora var. robusta concentration in blends by means of synchronous fluorescence and UV-Vis spectroscopies. *Talanta*, *172*, 215–220. https://doi.org/10.1016/j.talanta.2017.05.036

Ebrahimi-Najafabadi, H., Leardi, R., Oliveri, P., Chiara Casolino, M., Jalali-Heravi, M., & Lanteri, S. (2012). Detection of addition of barley to coffee using near infrared spectroscopy and chemometric techniques. *Talanta*, *99*, 175–179. https://doi.org/10.1016/j.talanta.2012.05.036

Engel, J., Gerretzen, J., Szymańska, E., Jansen, J. J., Downey, G., Blanchet, L., & Buydens, L. M. C. (2013). Breaking with trends in pre-processing? In *TrAC - Trends in Analytical Chemistry* (Vol. 50, pp. 96–106). Elsevier B.V. https://doi.org/10.1016/j.trac.2013.04.015

Esteban-Díez, I., González-Sáiz, J. M., & Pizarro, C. (2004). Prediction of sensory properties of espresso from roasted coffee samples by near-infrared spectroscopy. *Analytica Chimica Acta*, *525*(2), 171–182. https://doi.org/10.1016/j.aca.2004.08.057

Esteban-Díez, I., González-Sáiz, J. M., Sáenz-González, C., & Pizarro, C. (2007). Coffee varietal differentiation based on near infrared spectroscopy. *Talanta*, *71*(1), 221–229. https://doi.org/10.1016/j.talanta.2006.03.052

European Comission. (2020). *The EU food fraud network and the administrative assistance and cooperation system health and food safety*. https://doi.org/10.2875/326318

European commission. (2009). *2013/2091(INI)*. http://ec.europa.eu/food/food/horsemeat/plan_en.htm.

Ferreira, T., Farah, A., Oliveira, T. C., Lima, I. S., Vitório, F., & Oliveira, E. M. M. (2016). Using Real-Time PCR as a tool for monitoring the authenticity of commercial coffees. *Food Chemistry*, *199*, 433–438. https://doi.org/10.1016/j.foodchem.2015.12.045

Hódsági, M., Gergely, S., Gelencsér, T., & Salgó, A. (2010). Investigations of Native and Resistant Starches and Their Mixtures Using Near-Infrared Spectroscopy. *Food and Bioprocess Technology 2010 5:1*, *5*(1), 401–407. https://doi.org/10.1007/S11947-010-0491-5

ICO-January. (2021). *Arabica prices continue to rise in January while Robusta falls*. *44*, 7612.

Illy, A., & Viani, R. (2005). *Espresso Coffee: The Science of Quality*. Elsevier Science. https://books.google.it/books?id=AJdlfSFCmVIC

Italiaimballaggio. (2020). *Coffee & tea: the market in Italy*.

https://www.italiaimballaggio.network/en/coffee-tea-market--consumption-italy-italian-packaging-institute?__cf_chl_jschl_tk__=040322461bc063c234a35719bbb4a33c1514d2ff-1618215374-0-AX1xP2wZNcftFZdHBJu8z_GdIWR3FVNFcDNmE5zY4SjiUjHVJf4WVUCTbjKvpUxkDfLLIbaCU8P

Luo, N., Yang, X., Sun, C., Xing, B., Han, J., & Zhao, C. (2021). Visualization of vibrational spectroscopy for agro-food samples using t-Distributed Stochastic Neighbor Embedding. *Food Control*, *126*, 107812. https://doi.org/10.1016/j.foodcont.2020.107812

Malek, S., Melgani, F., & Bazi, Y. (2018). One-dimensional convolutional neural networks for spectroscopic signal regression. *Journal of Chemometrics*, *32*(5), e2977. https://doi.org/10.1002/cem.2977

Milani, M. I., Rossini, E. L., Catelani, T. A., Pezza, L., Toci, A. T., & Pezza, H. R. (2020). Authentication of roasted and ground coffee samples containing multiple adulterants using NMR and a chemometric approach. *Food Control*, *112*, 107104. https://doi.org/10.1016/j.foodcont.2020.107104

Mishra, P., & Passos, D. (2021a). 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*, *212*, 104287.

Mishra, P., & Passos, D. (2021b). Realizing transfer learning for updating deep learning models of spectral data to be used in new scenarios. *Chemometrics and Intelligent Laboratory Systems*, *212*, 104283. https://doi.org/10.1016/j.chemolab.2021.104283

Moore, J. C., Spink, J., & Lipp, M. (2012). Development and Application of a Database of Food Ingredient Fraud and Economically Motivated Adulteration from 1980 to 2010. In *Journal of Food Science* (Vol. 77, Issue 4). https://doi.org/10.1111/j.1750-3841.2012.02657.x

Morin, J.-F., & Lees, M. (Eds.). (2018). *FoodIntegrity Handbook*. Eurofins Analytics France. https://doi.org/10.32741/fihb

Moscetti, R., Massantini, R., & Fidaleo, M. (2019). Application on-line NIR spectroscopy and other process analytical technology tools to the characterization of soy sauce desalting by electrodialysis. *Journal of Food Engineering*, *263*, 243–252. https://doi.org/10.1016/j.jfoodeng.2019.06.022

Oliveri, P., Malegori, C., Casale, M., Tartacca, E., & Salvatori, G. (2019). An innovative

multivariate strategy for HSI-NIR images to automatically detect defects in green coffee. *Talanta*, *199*, 270–276. https://doi.org/10.1016/j.talanta.2019.02.049

Pauli, E. D., Barbieri, F., Garcia, P. S., Madeira, T. B., Acquaro, V. R., Scarminio, I. S., da Camara, C. A. P., & Nixdorf, S. L. (2014). Detection of ground roasted coffee adulteration with roasted soybean and wheat. *Food Research International*, *61*, 112–119. https://doi.org/10.1016/j.foodres.2014.02.032

Reis, N., Franca, A. S., & Oliveira, L. S. (2013a). Discrimination between roasted coffee, roasted corn and coffee husks by Diffuse Reflectance Infrared Fourier Transform Spectroscopy. *LWT - Food Science and Technology*, *50*(2), 715–722. https://doi.org/10.1016/j.lwt.2012.07.016

Reis, N., Franca, A. S., & Oliveira, L. S. (2013b). Quantitative evaluation of multiple adulterants in roasted coffee by Diffuse Reflectance Infrared Fourier Transform Spectroscopy (DRIFTS) and chemometrics. *Talanta*, *115*, 563–568. https://doi.org/10.1016/j.talanta.2013.06.004

Ribeiro, J. S., Ferreira, M. M. C., & Salva, T. J. G. (2011). Chemometric models for the quantitative descriptive sensory analysis of Arabica coffee beverages using near infrared spectroscopy. *Talanta*, *83*(5), 1352–1358. https://doi.org/10.1016/j.talanta.2010.11.001

Souto, U. T. de C. P., Barbosa, M. F., Dantas, H. V., de Pontes, A. S., Lyra, W. da S., Diniz, P. H. G. D., de Araújo, M. C. U., & da Silva, E. C. (2015). Screening for Coffee Adulteration Using Digital Images and SPA-LDA. *Food Analytical Methods*, *8*(6), 1515–1521. https://doi.org/10.1007/s12161-014-0020-7

Tavares, K. M., Lima, A. R., Nunes, C. A., Silva, V. A., Mendes, E., Casal, S., & Pereira, R. G. F. A. (2016). Free tocopherols as chemical markers for Arabica coffee adulteration with maize and coffee by-products. *Food Control*, *70*, 318–324. https://doi.org/10.1016/j.foodcont.2016.06.011

Toci, A. T., Farah, A., Pezza, H. R., & Pezza, L. (2016). Coffee Adulteration: More than Two Decades of Research. In *Critical Reviews in Analytical Chemistry* (Vol. 46, Issue 2, pp. 83–92). Taylor and Francis Ltd. https://doi.org/10.1080/10408347.2014.966185

Uncu, A. T., & Uncu, A. O. (2018). Plastid trnH-psbA intergenic spacer serves as a PCR-based marker to detect common grain adulterants of coffee (Coffea arabica L.). *Food Control*, *91*, 32–39. https://doi.org/10.1016/j.foodcont.2018.03.029

Winkler-Moser, J. K., Singh, M., Rennick, K. A., Bakota, E. L., Jham, G., Liu, S. X., & Vaughn, S. F. (2015). Detection of Corn Adulteration in Brazilian Coffee (Coffea arabica) by Tocopherol Profiling and Near-Infrared (NIR) Spectroscopy. *Journal of*

*Agricultural and Food Chemistry*, *63*(49), 10662–10668.
https://doi.org/10.1021/acs.jafc.5b04777
Workman, J. . J., & Weyer, L. (2007). Practical Guide to Interpretive Near-Infrared Spectroscopy. *Practical Guide to Interpretive Near-Infrared Spectroscopy*.
https://doi.org/10.1201/9781420018318
Zawirska-Wojtasiak, R., Wojtowicz, E., Przygoński, K., & Olkowicz, M. (2014). Chlorogenic acid in raw materials for the production of chicory coffee. *Journal of the Science of Food and Agriculture*, *94*(10), 2118–2123.
https://doi.org/10.1002/JSFA.6532

**Highlights**

1. Deep chemometric algorithm tested for coffee adulteration prediction.
2. Convolutional neural network led to better performance than chemometrics.
3. Outlier removal and data augmentation benefitted the CNN algorithm.
4. CNN can be a feasible alternative for FT-NIR data analysis.

Figure 1

Click here to access/download;Figure;Figure_1.png

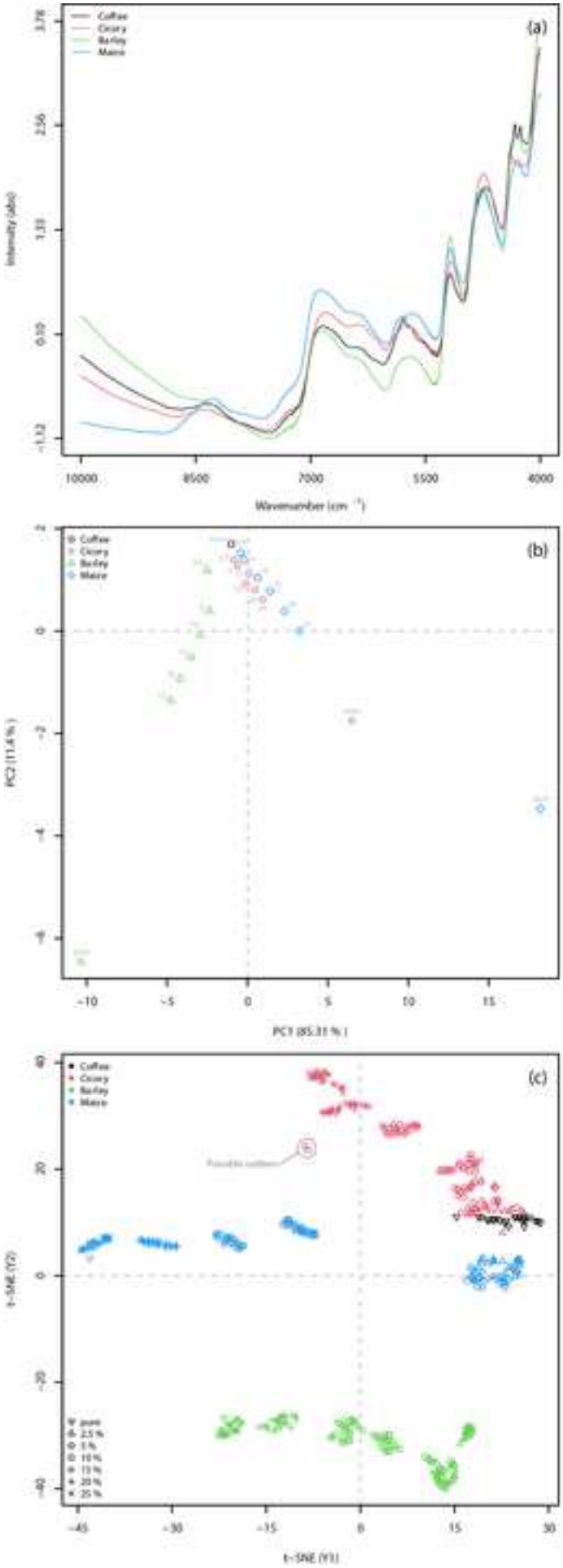


Figure 2



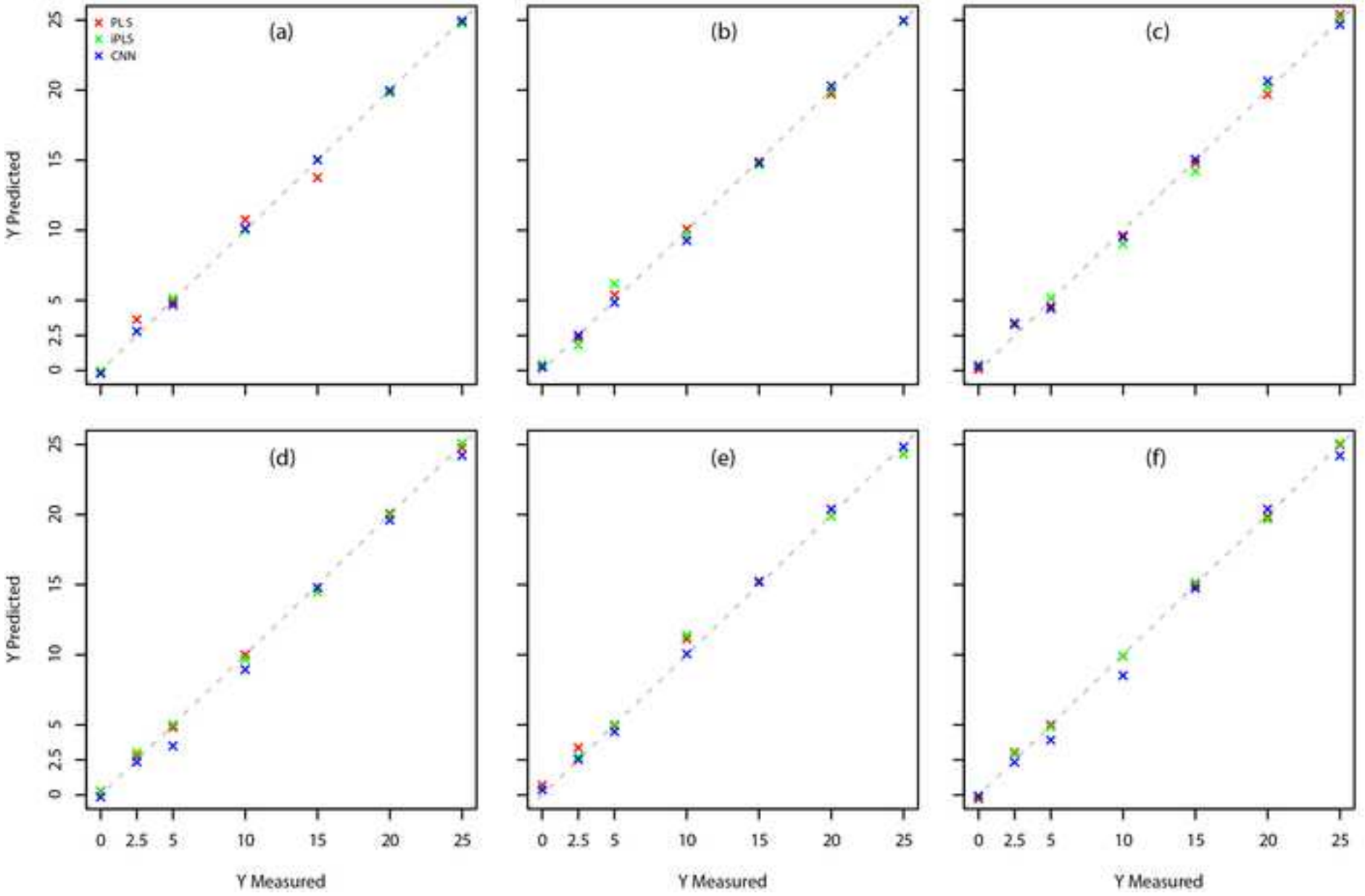
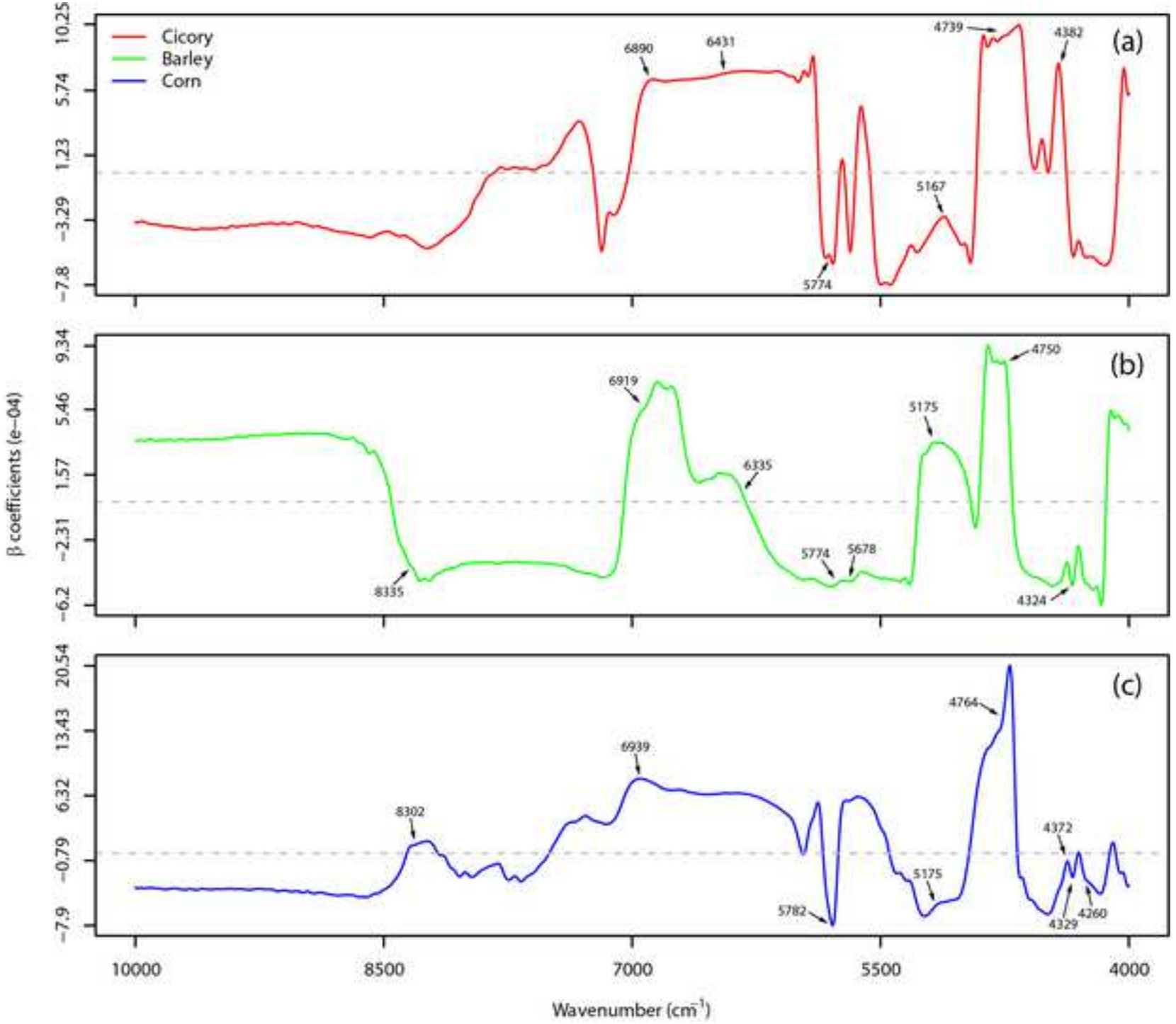


Figure 3

Click here to access/download;Figure;Figure_3.png



## Figure Captions

**Figure 1.** (a) The mean raw spectra of the pure coffee and adulterants; spatial distribution plots obtained for the adulterated coffee sets with (b) Principal Component Analysis on the average prediction datasets and (c) t-Distributed Stochastic Neighbour Embedding (t-SNE) analysis on the complete prediction dataset.

**Figure 2.** Regression plots of Y-measured (reference concentration) and Y-predicted (predicted concentration) values by PLS (red marks x), iPLS (green marks x) and CNN (blue marks x) algorithms for coffee adulterated with (a) chicory, (b) barley, and (c) maize. The rows correspond to pre-treatments with (1) AS and (2) SNV + AS. The dotted grey line (---) in each figure represents the ideal prediction line.

**Figure 3.** Regression plots based on ß-coefficients for the PLS models of the NIR spectral features of the coffee adulterated with (a) chicory, (b) barley and (c) maize.

**TABLE 1**

| Layer # | Layer (type) | Hyperparameters | Output Shape | Parameters |
|---|---|---|---|---|
| 1 | Gaussian Noise | *stdev = 0.05 (float)* | 3112 | 0 |
| 2 | Reshape | *shape = 1 (int)* | 3112, 1 | 0 |
| 3 | 1D convolution | *kernel_size = (8, 32) (int, int)*<br>*activation = "relu"*<br>*padding = "same"* | 3112, 8 | 264 |
| 4 | 1D convolution | *kernel_size = (16, 32) (int, int)*<br>*activation = "relu"*<br>*padding = "same"* | 3112, 16 | 4112 |
| 5 | Flatten | *none* | 49792 | 0 |
| 6 | Dropout | *rate = 0.5 (float)* | 49792 | 0 |
| 7 | Dense | *units = 128 (int)*<br>*activation = "relu"* | 128 | 6373504 |
| 8 | Dense | *units = 1 (int)*<br>*activation = "linear"* | 1 | 129 |

**Table 2.**

| Wave number ($cm^{-1}$) | Functional group | Note |
|---|---|---|
| **Coffee (A)** | | |
| 4256, 4332 | Combination bands of O-H vibration and N-H bands | Degradation of chlorogenic acids and release of quinic acid; characteristic peaks in roasted coffee |
| 4734 | 1st overtone of C=O and O-H combination bands with interference of N-H combination bonds | Caffeine, chlorogenic acids and polysaccharides |
| 5175 | 2nd overtone of C=O and O-H stretching | Caffeine, chlorogenic acid and polysaccharides |
| 5689, 5786 | 1st overtone C-H | Lipids and caffeine |
| 6333 | 1st overtone of O-H and N-H | Amino acids and chlorogenic acid |
| 6892 | 1st overtone O-H stretch | Water |
| 7317 | C-H stretching and deformation vibration | Caffeine |
| 8319 | 2nd overtone of C-H stretching | Caffeine and carbohydrates |
| **Chicory (C)** | | |
| 4382 | Combination bands of O-H and N-H | Chlorogenic acids and carbohydrates (starch) |
| 4739 | 1st overtone of C=O and O-H combination bands | Carbohydrates |
| 5167 | 1st O-H overtone region | Carbohydrates and water |
| 5774 | C-H bonds | Hydrocarbons |
| 6431 | N-H overtone | Proteins |
| 6890 | O-H bonds | Water and carbohydrates (starch) |
| **Barley (B)** | | |
| 4324 | C-H bending | Carbohydrates (starch) |
| 4750 | Combination bands of O-H and C=O stretch | Carbohydrates (starch) |
| 5175 | O-H bonds | Water and carbohydrates (starch) |
| 5678, 5774 | 1st overtones of C-H stretch | Hydrocarbons - Lipids |
| 6335 | 1st overtone region of O-H and N-H bonds | Water and protein |
| 6919 | 1st overtone O-H | Water and carbohydrates (starch) |
| 8335 | 2nd C-H overtone | Hydrocarbons - Lipids |
| **Maize (M)** | | |
| 4260, 4329 | O-H Combination bands | Hydrocarbons and Lipids |
| 4372 | C=O and N-H bonds | Proteins |
| 4764 | C=O stretch and O-H bend combination bands | Carbohydrates (Polysaccharides) |
| 5175 | O-H bonds | Carbohydrates and water |
| 5782 | C-H stretch 1st overtone | Lipids |
| 6939 | O-H and N-H of primary amines | Proteins and water |
| 8302 | C-H stretching | Lipids |

**References (Alessandrini et al., 2008; Barbin et al., 2014; Cozzolino et al., 2013, 2012, 2006; Hódsági et al., 2010; Ribeiro et al., 2011; Toci et al., 2016; Workman and Weyer, 2007)*

**TABLE 3**

| Adulterant | Spectral pre-treatment | Algorithm | Wavenumbers ($cm^{-1}$) | LV | RMSE | | BIAS | | $R^2$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | *Cal* | *Pred* | *Cal* | *Pred* | *Cal* | *Pred* |
| Chicory | AS | PLS | 4869-4852, 4811-4794, 4464-4447, 4367-4350 | 3 | 1.54 | 1.39 | -1.07E-14 | -3.06E-01 | 0.97 | 0.98 |
| | | iPLS | | 3 | 1.03 | 1.06 | 3.85E-13 | 1.68E-01 | 0.99 | 0.99 |
| | | CNN | | - | 0.59 | 0.76 | 4.00E-02 | -1.00E-02 | 1.00 | 0.99 |
| | SNV+AS | PLS | 4869-4852, 4753-4736, 4464-4447, 4387-4369 | 3 | 1.03 | 1.00 | 5.32E-15 | 1.70E-01 | 0.99 | 0.99 |
| | | iPLS | | 2 | 1.06 | 1.05 | 4.44E-14 | 1.54E-01 | 0.99 | 0.99 |
| | | CNN | | - | 0.68 | 1.06 | 5.20E-01 | 5.90E-01 | 0.99 | 0.98 |
| Barley | AS | PLS | 4753-4736, 4406-4389, 4290-4254 | 3 | 0.94 | 1.31 | -1.78E-15 | 3.58E-01 | 0.99 | 0.98 |
| | | iPLS | | 2 | 1.03 | 1.63 | -4.97E-14 | 7.11E-01 | 0.99 | 0.97 |
| | | CNN | | - | 0.61 | 0.80 | -6.00E-02 | 1.10E-01 | 1.00 | 0.99 |
| | SNV+AS | PLS | 4715-4697, 4445-4427, 4329-4312, 4097-4080 | 3 | 0.63 | 1.39 | -1.07E-14 | 3.36E-01 | 1.00 | 0.99 |
| | | iPLS | | 3 | 0.41 | 1.06 | 3.85E-13 | 2.01E-01 | 1.00 | 0.99 |
| | | CNN | | - | 0.33 | 0.76 | 4.00E-02 | -4.00E-02 | 1.00 | 1.00 |
| Maize | AS | PLS | 5216-5199, 4734-4717, 4290-4254 | 4 | 0.86 | 0.75 | 1.59E-14 | -3.00E-02 | 0.99 | 0.99 |
| | | iPLS | | 3 | 1.29 | 1.47 | -2.87E-13 | -1.61E-01 | 0.98 | 0.97 |
| | | CNN | | - | 0.46 | 0.82 | -1.00E-02 | -1.00E-01 | 1.00 | 0.99 |
| | SNV+AS | PLS | | 3 | 0.64 | 0.64 | -1.42E-14 | -1.33E-01 | 0.99 | 1.00 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | iPLS | 5737-5719, 4290-4254, 4020-4003 | 3 | 0.59 | 0.73 | -1.30E-13 | -7.40E-02 | 1.00 | 0.99 |
| | CNN | | - | 0.54 | 0.98 | 3.70E-01 | 5.10E-01 | 1.00 | 0.99 |

## Table Captions

**Table 1.** CNN architecture implemented in the experimentation.

**Table 2.** Spectra–structure correlation and absorption regions of major peaks found in pure coffee and selected pure adulterants (chicory, barley, and maize).

**Table 3.** Summary of performance metrics for the combinations of spectral pre-treatments and regression algorithms (PLS, iPLS and CNN) for the adulterated coffee samples with chicory, barley, and maize.

Conflict of Interest and Authorship Conformation Form

Please check the following as appropriate:

- ☑ All authors have participated in (a) conception and design, or analysis and interpretation of the data; (b) drafting the article or revising it critically for important intellectual content; and (c) approval of the final version.
- ☑ This manuscript has not been submitted to, nor is under review at, another journal or other publishing venue.
- ☑ The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript
- ☐ The following authors have affiliations with organizations with direct or indirect financial interest in the subject matter discussed in the manuscript:

| **Author's name** | **Affiliation** |
|---|---|
| | |