

Compositional Analysis of Alternative Protein Blends Using Near and Mid-Infrared Spectroscopy Coupled with Conventional and Machine Learning Algorithms

dos Santos, R.¹, Cruz, J.², Muñoz, I.¹, Gou, P.¹, Nordon, A.³, Fulladosa, E.^{1*}

¹ IRTA. Food Quality and Technology. Finca Camps i Armet, s/n, 17121 Monells, Girona, Spain.

² Escola Universitària Salesiana de Sarrià, Passeig Sant Joan Bosco, 74, 08017 Barcelona, Spain.

³ WestCHEM, Department of Pure and Applied Chemistry and CPACT, University of Strathclyde, 295 Cathedral Street, Glasgow, G1 1XL, UK

*Email: elena.fulladosa@irta.cat

Highlights:

- Alternative protein source blends are a mixture of pulses, cereals, and pseudocereals.
- NIRS and MIRS can determine the composition of alternative protein source blends.
- Model accuracy for prediction of composition is sufficient for monitoring purposes.
- PLSDA, CNN and SVMDA algorithms can identify gluten-free alternative protein blends.

Abstract

The non-invasive real-time analysis of the composition of alternative, plant-based protein sources is important to control high moisture extrusion processes and ensure the quality and texture of the final extrudates used in the elaboration of meat analogues. This study aims to analyse the composition and presence of gluten in blended plant-based alternative protein sources from pulse, cereal and pseudocereal origin by means of near infrared spectroscopy (NIRS) and mid infrared spectroscopy (MIRS) using conventional and machine learning algorithms. Blends were prepared using five alternative protein sources (barley, wheat, fava bean, lupin, and buckwheat) and spectra were acquired using a low-cost and a benchtop near-infrared spectrometer, and a mid-infrared spectrometer. Using the acquired spectra, partial least square regression (PLSR), support vector machine discriminant analysis (SVM-DA), partial least square discriminant analysis (PLS-DA), and convolutional neural networks (CNN) were used to develop predictive models to determine the composition and to identify samples containing gluten. The protein, moisture, carbohydrates and fat content in blends of alternative protein sources was determined with a RMSEP of 1.59, 0.18, 1.41, and 0.19 %, respectively, when using the benchtop NIR spectrometer and PLSR. Gluten-free samples were identified with high sensitivity (0.85) and accuracy (0.93) using PLS-DA. The study demonstrated that infrared

spectroscopy can be used to analyse the composition of blends of alternative protein sources including pulses, cereals, and pseudocereals, as well as to identify gluten-free samples.

Keywords: Proximate composition, process monitoring, partial least squares, chemometrics, machine learning, plant-based proteins

1.- Introduction

Meat analogues are usually plant-based products that mimic the texture and flavour of meat products (Boukid, 2021). The most common alternative protein sources used to develop these products are soy (*Glycine max*) and wheat (*Triticum aestivum*), although other alternative protein sources such as barley (*Hordeum vulgare*), fava bean (*Vicia faba*), lupin (*Lupinus angustifolius*), and buckwheat (*Fagopyrum esculentum*) are also used (Azzollini et al., 2019). However, these plant-based protein sources lack the fibrousness found in meat. One method to develop fibrous textures is high moisture extrusion processing (HMEP) (Dekkers et al., 2018; Kołodziejczak et al., 2022). In HMEP, protein powders are fed into an extruder and exposed to high temperature, pressure, and friction. The resulting products are high moisture extrudates (HMEs) with a certain degree of fibrousness, which can be used to elaborate meat analogues.

The composition of the protein powders (which are, in many cases, blends of different alternative protein sources), defines the final texture and the nutritional quality of the HMEs. Parameters such as the moisture, protein, fat, and carbohydrate content of these raw materials is key to achieving products with optimal texture. Since traditional methods for determining these parameters are time and resource-consuming, the use of non-invasive techniques like near-infrared spectroscopy (NIRS) and mid-infrared spectroscopy (MIRS) is an emerging alternative (Holroyd et al., 2023).

NIRS has been demonstrated to be an ideal method for moisture determination in different kinds of food products due to the intense bands in the spectra produced by the O-H bond vibration (Zhang et al., 2022; Büning-Pfaue, 2003). Protein, carbohydrate, and fat content can also be determined from peaks arising from N-H, C-O-H, and C-H bonds, respectively. The composition of plant-based protein sources has been studied previously using NIRS (Bertrand et al., 1985; Cozzolino, 2014, 2016). Recently, alternative protein sources have also been analyzed. In Pellacani et al. (2024), buckwheat and barley flours were analyzed using visible-near infrared (VIS-NIR) spectroscopy and the models developed were able to discriminate samples from different origins. Platov et al. (2021) studied different types of buckwheat with VIS-NIR spectroscopy and concluded that it was possible to determine certain quality features like grain

size, harvest time, and roasting method using the spectra and multivariate analysis. Sato et al. (2001) also studied the composition of different varieties of buckwheat harvested in different years using NIRS showing low prediction errors and biases for protein, moisture, and fat content. A similar study showed the capability of NIRS to discriminate amaranth and buckwheat germplasms based on their protein, oil, amino acids and fatty acids (Shruti et al., 2023). However, to our knowledge, the composition of alternative protein blends from different sources has not been reported before. Miralbés (2004) and Albanell et al. (2012) demonstrated that NIRS is also a valid method to detect and quantify gluten in the production line. Although some studies using low-cost handheld instruments (Yan et al., 2023) to determine salt content in fish (Fulladosa et al., 2023), the quality of apples (Grabska et al., 2023), and the staling of bread (De Géa Neves et al., 2023) have also been reported, studies related to the detection of fraud and adulteration of alternative protein powder blends are scarce. These non-invasive spectroscopic sensors are tools used in Process Analytical technologies (PAT) and, together with multivariate data analysis, have proven effective for real-time monitoring, control, and analysis of different processes within production environments (Pomerantsev & Rodionova, 2012).

In contrast to NIRS, MIRS allows for more specific identification of components within samples by providing detailed information on fundamental transitions, such as the absorption of light by proteins due to their specific bonds (Beć, Grabska, Bonn, et al., 2020; Türker-Kaya & Huck, 2017). Beck et al. (2017) studied the effect of low moisture extrusion on the structure of proteins and García-Gutiérrez et al. (2021) and Mellado-Carretero et al. (2020) classified edible insects using MIRS. MIRS seems more suitable for the authentication and classification of food products, and less suitable for the determination of composition than NIRS (Hell et al., 2016). Nonetheless, MIRS has proven potential for rapid analysis of quality and composition characteristics at line, although it cannot be used in-line and so is not as suitable as NIRS for PAT applications.

The modelling techniques used to analyse the data retrieved from NIRS and MIRS are also crucial. Thus, recently, many studies have reported the use of machine learning algorithms such as support vector machines (SVMs) and convolutional neural networks (CNNs) to give improved classification scores compared to PLS-DA. Examples include the determination of protein content in barley (Singh et al., 2023), the prediction of dry matter in fruit (Passos & Mishra, 2023), and the nutritional composition of food products (Ahn et al., 2019). According to Marini (2009), artificial neural networks (ANNs) are better suited when a linear function cannot approximate the relation between predictors and responses.

The main objective of this study was to determine the composition and presence of gluten in blended plant-based alternative protein sources from different types (pulse, cereal and pseudocereals) using NIRS and MIRS as part of a wider approach to control HMEP. The feasibility of employing different spectrometers and modelling approaches, based on conventional and machine learning algorithms was also evaluated.

2.- Materials and methods

2.1.- Near infrared and mid infrared spectrometers

A low-cost handheld near infrared (NIR) spectrometer called SCiO (henceforth referred to as low-cost NIR spectrometer) (Consumer Physics Inc., Tel Aviv, Israel) measured absorbance spectra over the 740–1070 nm (13,500–9,350 cm^{-1}) spectral region with a 1 nm resolution. A shade accessory was used to minimize the influence of external light and to measure all samples at a 10 mm distance. Spectra were collected in reflection mode using the SCiO smartphone app (The Lab, version 1.3.1.81), and a smartphone with a Bluetooth connection, and then uploaded to the Consumer Physics Cloud database.

A benchtop Fourier Transform NIR spectrometer model Matrix-F duplex (henceforth referred to as benchtop NIR spectrometer) (Bruker Optik GmbH, Germany) measured absorbance spectra using the OPUS™ software (Bruker Optik GmbH, Germany) over the 830–2500 nm (12,000–4,000 cm^{-1}) spectral region with a contact probe IN 268-2 (spot size 3 mm \varnothing) (Solvias AG, Switzerland). Each spectrum was obtained in absorption mode from 32 scans performed at 8 cm^{-1} resolution. The probe was manually pressed directly onto the sample during spectral acquisition.

A Thermo Scientific™ iD7 Nicolet™ iS™5 Attenuated Total Reflectance - Fourier Transform Infrared (ATR-FTIR) spectrometer (henceforth referred to as MIR spectrometer) measured ATR-FTIR spectra of the samples. Spectra (wavenumber range of 4000–500 cm^{-1} , 16 scans per spectrum, spectral resolution of 4 cm^{-1}) were recorded by attenuated total reflection (ATR) using a diamond crystal. For spectra acquisition, a small sample was placed onto the ATR crystal using a metal spatula, and the clamp was screwed down onto the sample to ensure good contact with the crystal.

2.2.- Alternative protein sources

Five commercial plant-based protein sources were used: barley (*Hordeum vulgare*) flour (Casa Ruiz Granel Selecto SL, Madrid, Spain); wheat (*Triticum aestivum*) vital gluten (*amygluten*)

isolate (Collelldevall SL, Girona, Spain); fava bean (*Vicia faba*) protein concentrate (AGT foods, USA); lupin (*Lupinus angustifolius*) protein isolate (Prolupin GmbH, Germany); buckwheat (*Fagopyrum esculentum*) flour (Hort del Silenci, Lleida, Spain).

To increase the range of moisture content, a separate batch of the five commercial protein sources mentioned above (dry protein sources) were humidified with a water sprayer (humidified protein sources) and later used to prepare the powder blends. In total, ten separate protein sources (five dry protein sources and five humidified protein sources) were used.

2.3.- Experimental design and sample preparation

An experimental design to develop predictive models for protein, moisture, carbohydrate, and fat content in different blends of alternative protein sources was defined using The Design-Expert® (Stat-Ease Inc., USA). The D-Optimal algorithm was used to iteratively search for and evaluate different combinations of alternative protein sources to identify the optimal set of experimental runs (de Aguiar et al., 1995). The ten alternative protein sources were used to prepare a total of 100 samples. The first 90 samples comprised different combinations of the ten different protein sources, with each sample containing no more than 50% of any single source, ensuring a mix of at least two protein sources in each sample. The remaining 10 samples consisted of pure protein sources. Table 1 (Supplementary Material) details the compositions of all 100 blends used in the experimental design. Samples without wheat or barley powders were designated as gluten-free (although gluten was not chemically determined) while the rest of samples were considered as containing gluten.

2.4.- Chemical analysis

Protein content was measured using the Kjeldahl method (AOAC, 2005) by digesting the powders in a KjeldDigester K-446 (Büchi Labortechnik AG, Flawil, Switzerland), measuring the nitrogen content, and multiplying by a conversion factor (6.25) to obtain protein content. The moisture was measured using the gravimetric method, drying the powders at 102 ± 2 °C until constant weight (AOAC, 1990). Fat content was measured using the Soxhlet method (AOAC, 2000), by digesting the powders and using an organic dissolver (diethyl ether) to separate lipidic content. Ash content was determined by direct calcination in porcelain crucibles (AOAC, 2000), and the carbohydrate content was estimated as the difference between the total weight and the sum of the protein, fat, moisture, and ash content.

2.5.- Qualitative analysis and development of predictive models for composition

Seven samples (A, B, C, D, E, F, and G) covering a representative range of protein, moisture, fat, and carbohydrates were chosen from the original 100 samples to perform a qualitative evaluation of spectral changes due to composition variations and protein type (legumes, cereals, and pseudocereals or seeds).

The PLS_Toolbox 9.2 software (Eigenvector Inc., Manson, WA, USA) running in the MATLAB 2022b (Mathworks Inc., Natick, MA, USA) environment was used to develop the predictive models for the composition of the 100 samples elaborated according to the experimental design. Each sample was analysed in triplicate with each spectrometer, the spectra were averaged, and a single spectrum was used for each sample.

Predictive models for composition (moisture, protein, carbohydrates, and fat content) were developed using PLSR (Martens & Næs, 1989). For each parameter, data were divided into a calibration data set (3/4 of samples) and a prediction data set (the remaining 1/4 of samples) using the Kennard-Stone algorithm and the Mahalanobis distance. The Kennard-Stone method of creating two subsets (calibration and prediction) uses the covariance matrix to choose for the calibration set those spectra that are the most distant from each other, capturing most of the variability and allowing the development of robust predictive models (Kennard & Stone, 1969). The prediction set was used after the development of the predictive models to study overfitting (Jung & Hu, 2015; Stone, 1978). Before constructing the predictive models, NIR spectra were subject to various pretreatments, including mean centring, first-derivative, second-derivative, and standard normal variate (SNV), which are commonly used to correct scattering effects and improve the signal-to-noise ratio, (Barnes et al., 1989), reducing the offset and highlighting differences and changes (Savitzky & Golay, 1964), and wavelengths were selected manually or using the interval PLS (iPLS) technique. MIR spectra were subjected to different pretreatments like vector normalization, constant offset elimination, orthogonal scattering correction, straight line subtraction, min-max normalization, multiplicative scatter correction, and first derivative to correct scattering effects caused by the physical properties of the samples (i.e. remove the effect of different particle sizes or their compaction degree on the spectra (Jin et al., 2012)). Cross-validation was carried out using Venetian blinds, with 10 data splits and 1 sample per blind. The best combination of pretreatments and wavelength selection methods was chosen based on the lowest root mean square error of cross-validation (RMSECV), and the lowest number of latent variables (LV).

The coefficient of determination (R^2) shows how well the predictions fit into the developed model. The bias is the difference between the average of the predictions and the average of the reference measurements.

The range error ratio (RER) was calculated to evaluate the performance of the predictive models. RER is the ratio between the range of the reference values (i.e. the difference between the maximum reference value, and the minimum reference value) and the RMSEP. The RER values were interpreted according to the thresholds given by Malley et al. (2004) as follows: $RER > 20$ indicates an excellent prediction model; $15 \leq RER \leq 20$ successful; $10 \leq RER < 15$ moderately successful and $8 \leq RER < 10$ indicates a moderately useful prediction model.

The models obtained using the three spectrometers were compared using a Fisher-Snedecor distribution (or F-distribution) in Microsoft Excel 2021 (Microsoft Corporation, USA). A p -value < 0.05 in this test shows a significant difference between the errors of each model (Mohr et al., 2021).

2.6.- Development of classification models for gluten detection

PLS-DA, SVM-DA, and CNNs were used to predict the presence of gluten in the samples. PLS-DA and SVM-DA models were developed using PLS_Toolbox 9.2 software (Eigenvector Inc., Manson, WA, USA) running on the MATLAB 2022b (Mathworks Inc., Natick, MA, USA). Python 3.12 (Grus, 2015; Matthes, 2023), the pandas library (v. 1.5.3), the numpy library (v. 1.23.5) (McKinney, 2022), the TensorFlow library (v. 2.17.0), the Keras library (v. 3.5.0) (Chollet, 2017), and the scikit-learn library (v. 1.5.2) (Géron, 2022) were used to create and train CNNs. Cross-validation was performed using a 5-fold procedure where samples were randomly selected for each fold. In this case, because of the low number of samples, only cross-validation was performed. The parameters of the models were fine-tuned depending on the data set (low-cost NIR, benchtop NIR, and MIR data sets) being analysed to achieve a better performance. For the SVM-DA, the optimal cost parameters (C), and gamma values were chosen by PLS Toolbox. The selected kernel type was the radial basis function. For the CNN, different combinations of architectures and learning rates were manually checked and an architecture of 4 Conv1D, 3 MaxPooling1D, 2 dense layers, and 1 dropout layer was finally used as it achieved the best results during cross-validation. In contrast to PLS-DA and SVM-DA, the CNN was not systematically optimised because of the high computational cost of running the experiments. As a result, the performance of this CNN configuration could be lower than that corresponding to the optimal configuration.

The optimal parameters and pre-processing techniques for the models were selected depending on the average accuracy and Matthew's correlation coefficient (MCC) during the cross-

validation. The performance of the classification models was evaluated using a confusion matrix in which true negatives (TN, i.e. correctly identified samples without gluten), true positives (TP, i.e. correctly identified samples with gluten), false negatives (FN, i.e. samples with gluten misidentified as gluten-free), and false positives (FP, i.e. samples without gluten misidentified as having gluten) were obtained. The sensitivity, accuracy, and the MCC were calculated. Sensitivity is the quotient of TP and TP plus FN. It is used to estimate the model's ability to minimize false negatives (Ballabio & Consonni, 2013). Accuracy is the ratio between correctly classified and total samples, showing the percentage of correctly classified samples in a prediction. However, accuracy can lead to misleading and overoptimistic results when the dataset is unbalanced (i.e. there are many more samples in one class than the other, as is the case in the present work where there were 75 samples with gluten and only 25 without gluten) (Chicco & Jurman, 2020). In this regard, the MCC might help to interpret the confusion matrix. MCC is a statistical tool used to measure the difference between the predicted values and the reference values. Values close to 1 mean that the correlation is very high, and the model correctly predicts most samples. Value close to 0 are expected for random predictions (Chicco & Jurman, 2023; Matthews, 1975). Finally, the variable importance in the projection (VIP) scores were retrieved from the PLS model to study how each wavelength contributed to the gluten predictions (Chong & Jun, 2005).

3.- Results and discussions

3.1.- Chemical and qualitative analysis of the spectra of the protein sources

Figure 1 shows the variation in the composition of the samples used in this study (Table 2 Supplementary Material). The experimental design provided a set of samples with enough variation to cover the variation of these parameters in commercial alternative protein sources used to produce high moisture extrudates.

Raw spectra acquired using the low-cost NIR, benchtop NIR, and MIR spectrometers for a set of selected samples (A, B, C, D, E, F, and G), representative of the protein, moisture, fat, and carbohydrate content (Table 1), are presented in Figure 2. For the low-cost NIR spectra (Figure 2a), two absorbance peaks are seen at around 880 and 940 nm. The absorbance increases with water content at around 940 nm, showing the maximum absorbance in sample D (14.50 % moisture) and the minimum absorbance in sample C (6.96 % moisture). This band is related to the O-H bond. Other elements of the composition have no specific bands in the spectral range covered by this spectrometer. In contrast, in the case of the benchtop NIR spectra, differences in absorbance attributed to moisture content variations were also observed around the 1250

and 1850 nm bands marked in blue, which are known bands of water absorption (Figure 2b) (Büning-Pfaue, 2003). Nonetheless, it seems that the spectra are dominated by other components of the samples, like proteins and carbohydrates, and it is difficult to see how absorption increases in relation to moisture in the peaks usually attributed to water. For these other elements of the composition (see protein, fat, and carbohydrate bands in red, yellow, and green, respectively), the relationship between differences in absorbance and composition is not straightforward. For the MIR spectra, there does not seem to exist a straightforward relation between water content and absorbance. The MIR region has been described as a complex region, with many overlapping bands specific to molecular structure. Relatively sharp absorbance peaks can be attributed to specific chemical functional groups such as carbohydrates and proteins (around 3300 cm^{-1}), and lipids (2900 cm^{-1} and 1700 cm^{-1}) (Türker-Kaya & Huck, 2017) (Figure 2c).

In Figure 3, the spectra of pure protein sources used to create the blended samples are shown. Differences in low-cost NIR spectra (Figure 3a) can be attributed to changing composition of the samples. In the benchtop NIR spectra (Figure 3b), the different absorption peaks occurring between 2000 and 2500 nm seen in Figure 3b can be attributed to the type of protein source (legume, cereal, or pseudocereal), showing that NIR can be used to identify the different types of protein source used. In MIR region (Figure 3c), it can also be seen that the peaks around 1700 cm^{-1} and 1000 cm^{-1} have varying shapes. Wheat (cereal) has the highest absorbance at 1700 cm^{-1} while buckwheat (pseudocereal) has the highest absorbance at 1000 cm^{-1} . In other parts of the spectra, like the 2900 cm^{-1} region, it seems lupin (a legume) has the highest absorbance. Cereals, pseudocereals and pulses have also different peak positions (e.g. lupin around 1750 cm^{-1} and 2900 cm^{-1} , and wheat at around 1450 cm^{-1}) related to different functional groups which could be related to different source type. This is in line with what is described by Roberts et al. (2018), who reported that peak position in the MIR range of the spectrum allows the unambiguous identification of chemical functional groups.

3.2.- Predictive models for composition

Table 2 shows the performance of the PLSR to predict the composition of the protein source blends using the low-cost NIR, benchtop NIR, and MIR spectrometers, respectively. The models shown in the tables are the best prediction models obtained after trying different combinations of pre-processing techniques (mean centring, SNV, and derivatives) and wavelength selection methods (full wavelength, iPLS, and manual selection).

Benchtop NIR showed the best overall prediction capacity for all the parameters, with low errors and high linearities for protein, moisture, carbohydrates, and fat content (Figure 4). The best model for moisture was obtained using only the bands in the 1350-1550 nm and 1840-2024 nm regions (Table 2). These bands were manually selected as the peaks of absorbance are related to water content. For carbohydrates and fat content, the best models were obtained using iPLS. The bands selected by the iPLS algorithm were 906-1098 nm and 1228-1393 nm for carbohydrates and 1098-1393 nm and 1609-2333 nm for fats (Table 2).

The predicted compositions using spectra acquired with the low-cost NIR and MIR spectrometers had higher errors than those obtained using the benchtop NIR spectrometer. This could be because the wavelength range of the benchtop NIR (833-2500 nm) is wider than that of the low-cost NIR (740-1070 nm), and NIR is better suited for the quantification of these parameters than MIR, which, as explained previously, has sharp peaks that allow the identification of functional groups but might be less suited for their quantification (Shi & Yu, 2017). It could also be because the low-cost NIR spectrometer had lower sensitivity and resolution than the benchtop spectrometer. In the case of the low-cost NIR spectrometer, the full range of the spectra was used to develop the models for all parameters except moisture, where performing iPLS (which selected the interval between 940 nm and 1019 nm) achieved higher accuracy. This is in line with the existence of a band sensitive to water around 970 nm, as previously described. In the case of MIR, the full spectral range was used to develop the models for all parameters. The RER values show that low-cost NIR and MIR are appropriate for screening purposes (with RER values between 8 and 20) while benchtop NIR is useful for quality control purposes (> 20). However, compositional variation might be narrower in certain industrial applications, making it necessary to calculate RER values for each case.

The RMSEP of the PLSR models and the results of the F-distribution (p -values) obtained for each parameter of the composition using the low-cost NIR, benchtop NIR, and MIR spectrometers are also shown in Table 2 using significance letters. For moisture, protein, and carbohydrates, there were significant differences between the NIR benchtop and the other two spectrometers (p -value < 0.05), but no statistical differences between the low-cost NIR and MIR spectrometers (p -value > 0.05) were found. In contrast, the RMSEPs of the benchtop NIR and MIR spectrometers models have no statistical differences for fat content, and they both have lower errors than the low-cost NIR spectrometer. No significant differences were found for fat, and this could be attributed to the small fat content range. The models developed using the data from the benchtop NIR spectrometer are more robust and, therefore, have lower RMSEPs than the others.

3.3.- Gluten discrimination models

Table 3 shows the results obtained using PLS-DA, SVM-DA, and CNN for the classification of samples depending on the presence or absence of gluten using the low-cost NIR, benchtop NIR, and MIR spectrometers, respectively. The overall best models were developed using the benchtop NIR spectra, with average accuracies of 0.93, 0.89, and 0.85 for PLS-DA, SVM-DA, and CNN, respectively. The sensitivities to false negatives were higher than 0.77 in all cases, and the highest overall MCC was obtained with PLS-DA (0.82). The models developed using the low-cost NIR and MIR spectrometers showed their potential to predict the presence of gluten in the samples but with lower performances than the benchtop NIR spectrometer. The low-cost NIR spectrometer achieved a sensitivity to false negatives of 0.79 using CNN, but the highest MCC, obtained with SVM-DA (0.52), can be regarded as low. The MIR spectrometer had similar accuracies (0.88, 0.88, and 0.86) and MCCs (0.67, 0.72, 0.60) irrespective of the classification method used. As previously described in the Materials and Methods section, the systematic optimisation of CNN algorithms could help further improve its prediction capability when compared to PLS-DA and SVM-DA.

The VIP scores of the PLS-DA for each spectrometer are shown in Figure 5. Overall, for the low-cost NIR spectrometer, the most important wavelengths seem to be the ones around 750 nm (visible part of the spectrum) and 950 nm. For the benchtop NIR spectrometer, the most important wavelengths seem to be after 1800 nm according to previously reported results (Beć, Grabska, Bonn, et al., 2020; Beć, Grabska, & Huck, 2020; Izutsu et al., 2006). In the MIR spectrometer, the most important peaks seem to be around 3000 cm^{-1} , 1600 cm^{-1} , and 1100 cm^{-1} .

Overall, using a non-linear method like CNN did not improve the results of the models when compared to PLS-DA or SVM-DA. All models that used the data from the benchtop NIR and the MIR spectrometers had similar results, and they both were better than the ones that used data acquired using the low-cost NIR spectrometer, as expected. This could be explained by the narrower spectral range of the low-cost NIR spectrometer, which does not include information related to gluten content. False negatives, which could be a problem for people with coeliac disease, were observed in all cases, although in smaller numbers for the models developed using the benchtop NIR spectrometer as shown by the higher sensitivity.

For the models developed using the benchtop NIR spectrometer, there is a higher MCC (> 0.8) than for the other two spectrometers. CNN had lower MCCs than the other algorithms, which indicates difficulty in modelling the presence of gluten using spectral data and deep-learning

algorithms. SVM-DA and PLS-DA gave overall better results, except for data acquired using the low-cost NIR, where the results had similarly low MCCs.

The accuracy of these categorical models could be improved by increasing the number of samples included in the model. Given that the European Commission ((Commission implementing regulation (EU) No 828/2014, 2014)) regulates 'gluten-free' and 'very low gluten' claims, after verification of the detection limits, these sensors could serve as valuable tools for on-site monitoring by regulatory agencies.

4.- Conclusions

Low-cost NIR, benchtop NIR, and MIR spectroscopy, combined with chemometrics, have demonstrated their ability to analyse the composition of alternative protein source blends, including pulses, cereals, and pseudocereals. The models developed had low predictive errors for moisture, protein, carbohydrates, and fat content and can be considered useful for the control and monitoring of high moisture extrusion processes to improve the quality and texture of the final product. The benchtop NIR spectrometer demonstrated superior accuracy in predicting protein, moisture, and carbohydrate contents compared to fat content. The narrower fat content range (3.03% to 6.75%) may have limited the ability to develop an accurate calibration model.

The performance of gluten-free detection models could also ensure food safety for the final consumer. However, the robustness of the models should be improved using data from samples with a wider range of gluten contents and the detection limit should be determined. The benefits of incorporating in-line NIR sensors into plant-based product elaboration processes have yet to be evaluated.

Acknowledgements

This research was supported by project Sensanalog [PID2021-122285OR-I00] funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU. The first author received the grant [PRE2022-103798] funded by MICIU/AEI/10.13039/501100011033 and ESF+. Acknowledgements are extended to the consolidated research group (2021 SGR 00461) and CERCA program from Generalitat de Catalunya. Elena Fulladosa was supported by a mobility grant within the Incentives for Research Program 2023 by the Institute of Agrifood Research and Technology (IRTA).

CRedit authorship contribution statement

dos Santos, R.: formal analysis, investigation, methodology, writing – original draft preparation, and writing – review & editing **Cruz, J.:** formal analysis, investigation, methodology, supervision, and writing – review & editing **Muñoz, I.:** formal analysis, software, and writing – review & editing **Gou, P.:** supervision, writing – review & editing, **Nordon, A. –** Resources, writing – review & editing **Fulladosa, E.:** conceptualization, supervision, funding acquisition, investigation, project administration, writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing interests that could have influenced the work reported in this article.

Data availability

Data will be made available on request.

5.- References

- Ahn, D., Choi, J.-Y., Kim, H.-C., Cho, J.-S., Moon, K.-D., & Park, T. (2019). Estimating the Composition of Food Nutrients from Hyperspectral Signals Based on Deep Neural Networks. *Sensors*, *19*(7), Article 7. <https://doi.org/10.3390/s19071560>
- Albanell, E., Miñarro, B., & Carrasco, N. (2012). Detection of low-level gluten content in flour and batter by near infrared reflectance spectroscopy (NIRS). *Journal of Cereal Science*, *56*(2), 490–495. <https://doi.org/10.1016/J.JCS.2012.06.011>
- AOAC. (1990). *Official Methods of Analysis* (Association of Official Analytical Chemist, Ed.; 15th Edition).
- AOAC. (2000). *AOAC (2000) Official Methods of Analysis*. (17th Edition).
- AOAC. (2005). *Official Methods of Analysis* (Association of Official Analytical Chemist, Ed.; 22nd Edition). <https://www.aoac.org/official-methods-of-analysis/>
- Azzollini, D., Wibisaphira, T., Lakemond, C. M. M., & Fogliano, V. (2019). Toward the design of insect-based meat analogue: The role of calcium and temperature in coagulation behavior of *Alphitobius diaperinus* proteins. *LWT*, *100*, 75–82. <https://doi.org/10.1016/J.LWT.2018.10.037>
- Ballabio, D., & Consonni, V. (2013). Classification tools in chemistry. Part 1: Linear models. PLS-DA. *Analytical Methods*, *5*(16), 3790–3798. <https://doi.org/10.1039/C3AY40582F>
- Barnes, R. J., Dhanoa, M. S., & Lister, S. J. (1989). Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Applied Spectroscopy*, *43*(5), 772–777. <https://doi.org/10.1366/0003702894202201>
- Beć, K. B., Grabska, J., Bonn, G. K., Popp, M., & Huck, C. W. (2020). Principles and Applications of Vibrational Spectroscopic Imaging in Plant Science: A Review. *Frontiers in Plant Science*, *11*, 548519–548519. <https://doi.org/10.3389/FPLS.2020.01226/BIBTEX>

- Beć, K. B., Grabska, J., & Huck, C. W. (2020). Near-Infrared Spectroscopy in Bio-Applications. *Molecules* 2020, Vol. 25, Page 2948, 25(12), 2948–2948. <https://doi.org/10.3390/MOLECULES25122948>
- Beć, K. B., Grabska, J., & Huck, C. W. (2021). Principles and Applications of Miniaturized Near-Infrared (NIR) Spectrometers. *Chemistry – A European Journal*, 27(5), 1514–1532. <https://doi.org/10.1002/CHEM.202002838>
- Beck, S. M., Knoerzer, K., & Arcot, J. (2017). Effect of low moisture extrusion on a pea protein isolate's expansion, solubility, molecular weight distribution and secondary structure as determined by Fourier Transform Infrared Spectroscopy (FTIR). *Journal of Food Engineering*, 214, 166–174. <https://doi.org/10.1016/j.jfoodeng.2017.06.037>
- Bertrand, D., Robert, P., & Loisel, W. (1985). Identification of some wheat varieties by near infrared reflectance spectroscopy. *Journal of the Science of Food and Agriculture*, 36(11), 1120–1124. <https://doi.org/10.1002/jsfa.2740361114>
- Boukid, F. (2021). Plant-based meat analogues: From niche to mainstream. *European Food Research and Technology*, 247, 297–308. <https://doi.org/10.1007/s00217-020-03630-9>
- Büning-Pfaue, H. (2003). Analysis of water in food by near infrared spectroscopy. *Food Chemistry*, 82(1), 107–115. [https://doi.org/10.1016/S0308-8146\(02\)00583-6](https://doi.org/10.1016/S0308-8146(02)00583-6)
- Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1), 6. <https://doi.org/10.1186/s12864-019-6413-7>
- Chicco, D., & Jurman, G. (2023). The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData Mining*, 16, 4. <https://doi.org/10.1186/s13040-023-00322-4>
- Chollet, F. (2017). *Deep Learning with Python* (First Edition). Manning.

- Chong, I.-G., & Jun, C.-H. (2005). Performance of some variable selection methods when multicollinearity is present. *Chemometrics and Intelligent Laboratory Systems*, 78(1), 103–112. <https://doi.org/10.1016/j.chemolab.2004.12.011>
- COMMISSION IMPLEMENTING REGULATION (EU) No 828/2014. (2014). *On the requirements for the provision of information to consumers on the absence or reduced presence of gluten in food*. https://food.ec.europa.eu/food-safety/labelling-and-nutrition/specific-groups/gluten-free-food_en
- Cozzolino, D. (2014). An overview of the use of infrared spectroscopy and chemometrics in authenticity and traceability of cereals. *Food Research International*, 60, 262–265. <https://doi.org/10.1016/j.foodres.2013.08.034>
- Cozzolino, D. (2016). 16—Authentication of Cereals and Cereal Products. In G. Downey (Ed.), *Advances in Food Authenticity Testing* (pp. 441–457). Woodhead Publishing. <https://doi.org/10.1016/B978-0-08-100220-9.00016-3>
- de Aguiar, P. F., Bourguignon, B., Khots, M. S., Massart, D. L., & Phan-Thau-Luu, R. (1995). D-optimal designs. *Chemometrics and Intelligent Laboratory Systems*, 30(2), 199–210. [https://doi.org/10.1016/0169-7439\(94\)00076-X](https://doi.org/10.1016/0169-7439(94)00076-X)
- De Géa Neves, M., Noda, I., & Siesler, H. W. (2023). Investigation of bread staling by handheld NIR spectroscopy in tandem with 2D-COS and MCR-ALS analysis. *Microchemical Journal*, 190, 108578. <https://doi.org/10.1016/j.microc.2023.108578>
- Dekkers, B. L., Boom, R. M., & van der Goot, A. J. (2018). Structuring processes for meat analogues. *Trends in Food Science & Technology*, 81, 25–36. <https://doi.org/10.1016/J.TIFS.2018.08.011>
- Fulladosa, E., Barnés-Calle, C., Cruz, J., Martínez, B., Giró-Candanedo, M., Comaposada, J., Font-i-Furnols, M., & Gou, P. (2023). Near infrared sensors for the precise characterization of

Compositional analysis of alternative protein blends using near and mid-infrared spectroscopy coupled with conventional and machine learning algorithms

salt content in canned tuna fish. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 303, 123217. <https://doi.org/10.1016/j.saa.2023.123217>

García-Gutiérrez, N., Mellado-Carretero, J., Bengoa, C., Salvador, A., Sanz, T., Wang, J., Ferrando, M., Güell, C., & Lamo-Castellví, S. de. (2021). ATR-FTIR Spectroscopy Combined with Multivariate Analysis Successfully Discriminates Raw Doughs and Baked 3D-Printed Snacks Enriched with Edible Insect Powder. *Foods*, 10(8), Article 8. <https://doi.org/10.3390/foods10081806>

Géron, A. (2022). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems* (3rd edition). O'Reilly Media.

Grabska, J., Beć, K. B., Ueno, N., & Huck, C. W. (2023). Analyzing the Quality Parameters of Apples by Spectroscopy from Vis/NIR to NIR Region: A Comprehensive Review. *Foods 2023, Vol. 12, Page 1946*, 12(10), 1946–1946. <https://doi.org/10.3390/FOODS12101946>

Grus, J. (2015). *Data science from Scratch: First principles with Python* (2nd ed.). O'Reilly.

Hell, J., Prückler, M., Danner, L., Henniges, U., Apprich, S., Rosenau, T., Kneifel, W., & Böhmendorfer, S. (2016). A comparison between near-infrared (NIR) and mid-infrared (ATR-FTIR) spectroscopy for the multivariate determination of compositional properties in wheat bran samples. *Food Control*, 60, 365–369. <https://doi.org/10.1016/j.foodcont.2015.08.003>

Holroyd, S. E., Nickless, E., & Watkinson, P. (2023). Raman and mid-infrared spectroscopy to assess changes in Cheddar cheese with maturation. *International Journal of Dairy Technology*, 76(2), 408–417. <https://doi.org/10.1111/1471-0307.12929>

Compositional analysis of alternative protein blends using near and mid-infrared spectroscopy coupled with conventional and machine learning algorithms

- Izutsu, K., Fujimaki, Y., Kuwabara, A., Hiyama, Y., Yomota, C., & Aoyagi, N. (2006). Near-infrared analysis of protein secondary structure in aqueous solutions and freeze-dried solids. *Journal of Pharmaceutical Sciences*, *95*(4), 781–789. <https://doi.org/10.1002/jps.20580>
- Jin, J.-W., Chen, Z.-P., Li, L.-M., Steponavicius, R., Thennadil, S. N., Yang, J., & Yu, R.-Q. (2012). Quantitative Spectroscopic Analysis of Heterogeneous Mixtures: The Correction of Multiplicative Effects Caused by Variations in Physical Properties of Samples. *Analytical Chemistry*, *84*(1), 320–326. <https://doi.org/10.1021/ac202598f>
- Jung, Y., & Hu, J. (2015). A K-fold Averaging Cross-validation Procedure. *Journal of Nonparametric Statistics*, *27*(2), 167–179. <https://doi.org/10.1080/10485252.2015.1010532>
- Kennard, R. W., & Stone, L. A. (1969). Computer Aided Design of Experiments. *Technometrics*, *11*(1), 137–148. <https://doi.org/10.1080/00401706.1969.10490666>
- Kołodziejczak, K., Onopiuk, A., Szpicer, A., & Poltorak, A. (2022). Meat Analogues in the Perspective of Recent Scientific Research: A Review. *Foods*, *11*(1). <https://doi.org/10.3390/FOODS11010105>
- Malley, D. F., Martin, P. D., & Ben-Dor, E. (2004). Application in Analysis of Soils. In *Near-Infrared Spectroscopy in Agriculture* (pp. 729–784). John Wiley & Sons, Ltd. <https://doi.org/10.2134/agronmonogr44.c26>
- Marini, F. (2009). Artificial neural networks in foodstuff analyses: Trends and perspectives A review. *Analytica Chimica Acta*, *635*(2), 121–131. <https://doi.org/10.1016/j.aca.2009.01.009>
- Martens, Harald., & Næs, Tormod. (1989). *Multivariate calibration*. Wiley.
- Matthes, E. (2023). *Python Crash Course, 3rd Edition: A Hands-On, Project-Based Introduction to Programming*.

- Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2), 442–451. [https://doi.org/10.1016/0005-2795\(75\)90109-9](https://doi.org/10.1016/0005-2795(75)90109-9)
- McKinney, W. (2022). *Python for Data Analysis: Data Wrangling with pandas, NumPy, and Jupyter* (3rd ed.).
- Mellado-Carretero, J., García-Gutiérrez, N., Ferrando, M., Güell, C., García-Gonzalo, D., & Lamo-Castellví, S. D. (2020). Rapid discrimination and classification of edible insect powders using ATR-FTIR spectroscopy combined with multivariate analysis. *Journal of Insects as Food and Feed*, 6(2), 141–148. <https://doi.org/10.3920/JIFF2019.0032>
- Miralbés, C. (2004). Quality control in the milling industry using near infrared transmittance spectroscopy. *Food Chemistry*, 88(4), 621–628. <https://doi.org/10.1016/j.foodchem.2004.05.004>
- Mohr, D. L., Wilson, W. J., & Freund, R. J. (2021). *Statistical Methods* (p. 754). Elsevier. <https://doi.org/10.1016/B978-0-12-823043-5.00015-1>
- Passos, D., & Mishra, P. (2023). Deep Tutti Frutti: Exploring CNN architectures for dry matter prediction in fruit from multi-fruit near-infrared spectra. *Chemometrics and Intelligent Laboratory Systems*, 243, 105023. <https://doi.org/10.1016/j.chemolab.2023.105023>
- Pellacani, S., Borsari, M., Cocchi, M., D’Alessandro, A., Durante, C., Farioli, G., & Strani, L. (2024). Near Infrared and UV-Visible Spectroscopy Coupled with Chemometrics for the Characterization of Flours from Different Starch Origins. *Chemosensors*, 12(1), Article 1. <https://doi.org/10.3390/chemosensors12010001>
- Platov, Yu. T., Metlenkin, D. A., Platova, R. A., Rassulov, V. A., Vereshchagin, A. I., & Marin, V. A. (2021). Buckwheat Identification by Combined UV-VIS-NIR Spectroscopy and

Compositional analysis of alternative protein blends using near and mid-infrared spectroscopy coupled with conventional and machine learning algorithms

Multivariate Analysis. *Journal of Applied Spectroscopy*, 88(4), 723–730.
<https://doi.org/10.1007/s10812-021-01231-2>

Pomerantsev, A. L., & Rodionova, O. Y. (2012). Process analytical technology: A critical view of the chemometricians. *Journal of Chemometrics*, 26(6), 299–310.
<https://doi.org/10.1002/CEM.2445>

Roberts, J. J., Power, A., Chapman, J., Chandra, S., & Cozzolino, D. (2018). Chapter Three—Vibrational Spectroscopy Methods for Agro-Food Product Analysis. In J. Lopes & C. Sousa (Eds.), *Comprehensive Analytical Chemistry* (Vol. 80, pp. 51–68). Elsevier.
<https://doi.org/10.1016/bs.coac.2018.03.002>

Sato, T., Morishita, T., Hara, T., Suda, I., & Tetsuka, T. (2001). Near-Infrared Reflectance Spectroscopic Analysis of Moisture, Fat, Protein, and Physiological Activity in Buckwheat Flour for Breeding Selection. *Plant Production Science*, 4(4), 270–277.
<https://doi.org/10.1626/pp.4.270>

Savitzky, A., & Golay, M. J. E. (1964). Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry*, 36(8), 1627–1639.
https://doi.org/10.1021/AC60214A047/ASSET/AC60214A047.FP.PNG_V03

Shi, H., & Yu, P. (2017). Comparison of grating-based near-infrared (NIR) and Fourier transform mid-infrared (ATR-FT/MIR) spectroscopy based on spectral preprocessing and wavelength selection for the determination of crude protein and moisture content in wheat. *Food Control*, 82, 57–65. <https://doi.org/10.1016/j.foodcont.2017.06.015>

Shruti, Shukla, A., Rahman, S. S., Suneja, P., Yadav, R., Hussain, Z., Singh, R., Yadav, S. K., Rana, J. C., Yadav, S., & Bhardwaj, R. (2023). Developing an NIRS Prediction Model for Oil, Protein, Amino Acids and Fatty Acids in Amaranth and Buckwheat. *Agriculture*, 13(2), Article 2. <https://doi.org/10.3390/agriculture13020469>

- Singh, T., Garg, N. M., Iyengar, S. R. S., & Singh, V. (2023). Near-infrared hyperspectral imaging for determination of protein content in barley samples using convolutional neural network. *JOURNAL OF FOOD MEASUREMENT AND CHARACTERIZATION*, 17(4), 3548–3560. <https://doi.org/10.1007/s11694-023-01892-x>
- Stone, M. (1978). Cross-validation: A review. *Statistics: A Journal of Theoretical and Applied Statistics*. <https://doi.org/10.1080/02331887808801414>
- Türker-Kaya, S., & Huck, C. W. (2017). A Review of Mid-Infrared and Near-Infrared Imaging: Principles, Concepts and Applications in Plant Tissue Analysis. *Molecules* 2017, Vol. 22, Page 168, 22(1), 168–168. <https://doi.org/10.3390/MOLECULES22010168>
- Yan, H., Neves, M. D. G., Wise, B. M., Moraes, I. A., Barbin, D. F., & Siesler, H. W. (2023). The Application of Handheld Near-Infrared Spectroscopy and Raman Spectroscopic Imaging for the Identification and Quality Control of Food Products. *Molecules*, 28(23), Article 23. <https://doi.org/10.3390/molecules28237891>

Compositional analysis of alternative protein blends using near and mid-infrared spectroscopy coupled with conventional and machine learning algorithms

Tables

Table 1: Composition of the samples chosen for the qualitative analysis.

Code	Protein (%)	Moisture (%)	Fat (%)	Carbohydrates (%)
A	34.05	10.45	6.61	45.83
B	63.32	9.87	4.97	18.41
C	44.72	6.96	5.45	37.79
D	11.17	14.50	3.06	69.65
E	10.26	12.89	2.98	72.48
F	32.18	10.05	3.54	49.98
G	73.74	10.92	5.85	8.99

Compositional analysis of alternative protein blends using near and mid-infrared spectroscopy coupled with conventional and machine learning algorithms

Table 2: Results of the PLSR models for protein, moisture, carbohydrate, and fat content using the low-cost NIR spectrometer, the benchtop NIR spectrometer, and the MIR spectrometer and comparison of the models developed for each parameter using F-distribution.

	Spectral range used	Pre-processing	LV	RMSEP*	R ²	Bias	RER
Protein content							
Low-cost NIR	Full spectra	2 nd Der. + MC	3	2.36 _b	0.97	1.37	18.18
Benchtop NIR	Full spectra	SNV + MC	3	1.59 _a	0.98	-0.10	25.94
MIR	Full spectra	Smooth. + MSC + 1 st Der.	4	3.27 _b	0.97	1.18	19.07
Moisture content							
Low-cost NIR	940-1019 nm	SNV + MC	2	0.49 _b	0.87	-0.13	11.25
Benchtop NIR	1350-1550 nm and 1840-2040 nm	1 st Der. + MC	4	0.18 _a	0.99	-0.07	40.84
MIR	Full spectra	Smooth. + MSC	7	0.56 _b	0.82	-0.18	9.22
Carbohydrates content							
Low-cost NIR	Full spectra	2 nd Der. + MC	4	2.92 _b	0.96	-0.91	17.76
Benchtop NIR	906-1098 nm and 1228-1393 nm	1 st Der. + MC	3	1.41 _a	0.99	-0.34	35.15
MIR	Full spectra	Smooth. + MSC + 1 st Der.	4	3.43 _b	0.96	-1.31	17.60
Fat content							
Low-cost NIR	Full spectra	1 st Der. + MC	5	0.44 _b	0.80	0.24	7.35
Benchtop NIR	1098-1393 nm and 1609-2333 nm	1 st Der. + MC	5	0.19 _a	0.96	-0.08	16.62
MIR	Full spectra	Smooth. + MSC	5	0.20 _a	0.94	-0.03	13.83

LV: Latent variables. RMSEP: Root mean square error of prediction. R²: Regression coefficient of determination of the prediction set. Bias: The difference between the average of predictions and references. RER: Range error ratio. MC: Mean centre, 1st Der.: first derivative, 2nd Der.: second derivative, SNV: Standard Normal Variate, Smooth.: Smoothing of the spectra, MSC: Multiplicative Scatter Correction (mean). *Different significance letters indicate significant differences between the errors according to the F-distribution (p-value < 0.05).

Compositional analysis of alternative protein blends using near and mid-infrared spectroscopy coupled with conventional and machine learning algorithms

Table 3: Results of the gluten discrimination models developed using SVM-DA, PLS-DA, and CNN obtained from the cross-validation sets using a 5-fold cross-validation for the data collected using the low-cost NIR, the benchtop NIR and the MIR spectrometer. MCC: Matthews Correlation Coefficient.

	PLS-DA		SVMDA*		CNN**	
Low-cost NIR	Mean center, SNV, 3 LV		Mean center, SNV		Mean center, SNV	
	Sensitivity	0.46 7	Sensitivity	0.70 0	Sensitivity	0.78 6
	Average Accuracy	0.72 0	Average Accuracy	0.83 0	Average Accuracy	0.83 0
	MCC	0.45 3	MCC	0.52 0	MCC	0.49 9
Benchtop NIR	Mean center, 1st der., 5 LV		Mean center, SNV		Mean center, SNV	
	Sensitivity	0.84 6	Sensitivity	0.85 0	Sensitivity	0.77 8
	Average Accuracy	0.93 0	Average Accuracy	0.89 0	Average Accuracy	0.85 0
	MCC	0.81 6	MCC	0.71 3	MCC	0.62 8
MIR	MSC (mean), 6 LV		No pretreatments		Mean center, SNV	
	Sensitivity	0.69 7	Sensitivity	0.78 3	Sensitivity	0.82 4
	Average Accuracy	0.87 8	Average Accuracy	0.87 8	Average Accuracy	0.85 7
	MCC	0.72 2	MCC	0.67 0	MCC	0.59 7

*The optimal C and gamma values of the SVM-DA found for the low-cost NIR were 100 and 0.001, respectively, for the benchtop NIR were 31.62 and 0.001, respectively, and for MIR were 31.62 and 1, respectively.

**For the CNN, 256 epochs were used, the layers had 16, 32, 64, and 128 filters respectively, with a kernel size of 4, 8, 8, and 8 respectively; the strides of each layer were 2, 4, 4, and 4 respectively. For the pooling of the first three layers, a pool size of 2, 3, and 3 respectively were used and the strides were set to 2. The dense layer had 16 nodes and a dropout of 0.25. The activation functions used for these layers was ReLU and the loss function used was categorical cross-entropy. The last layer, used for classification, had 2 neurons and the activation was softmax. The Adam optimizer's learning rate was 0.00001.

Compositional analysis of alternative protein blends using near and mid-infrared spectroscopy coupled with conventional and machine learning algorithms

Figures

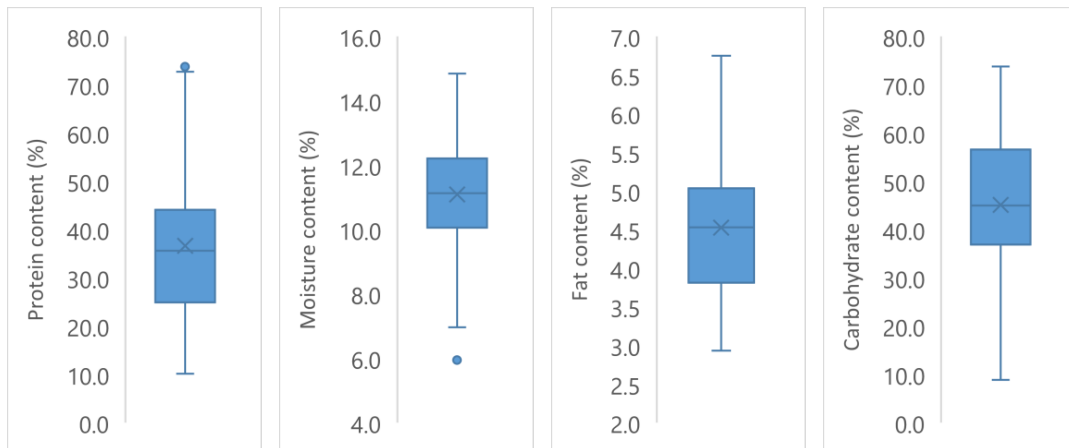


Figure 1: Distribution of the composition of the alternative protein source blends.

Compositional analysis of alternative protein blends using near and mid-infrared spectroscopy coupled with conventional and machine learning algorithms

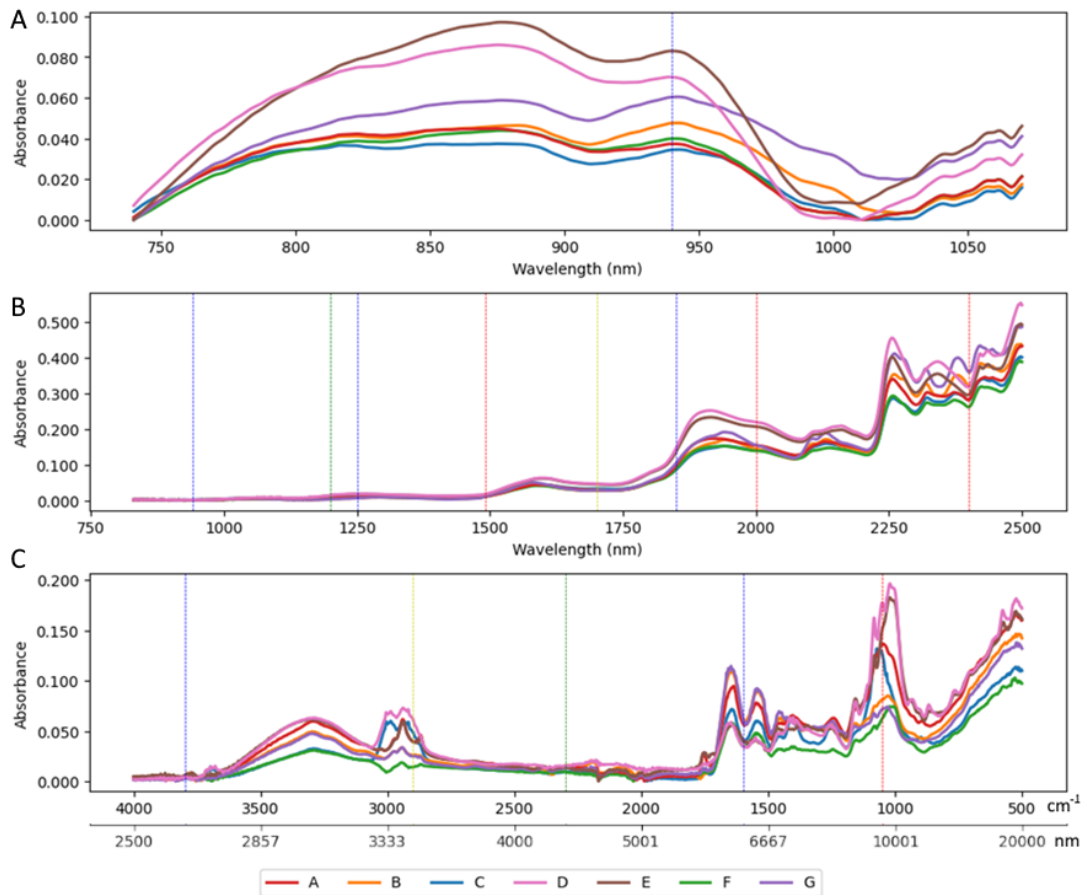


Figure 2: Raw spectra from low-cost NIR (a), benchtop NIR (b), and MIR (c) spectrometers of the selected alternative protein source blends. The dotted vertical lines in red (protein), blue (moisture), yellow (fat), and green (carbohydrates) are the specific bands, found in bibliography, for the different components of the samples.

Compositional analysis of alternative protein blends using near and mid-infrared spectroscopy coupled with conventional and machine learning algorithms

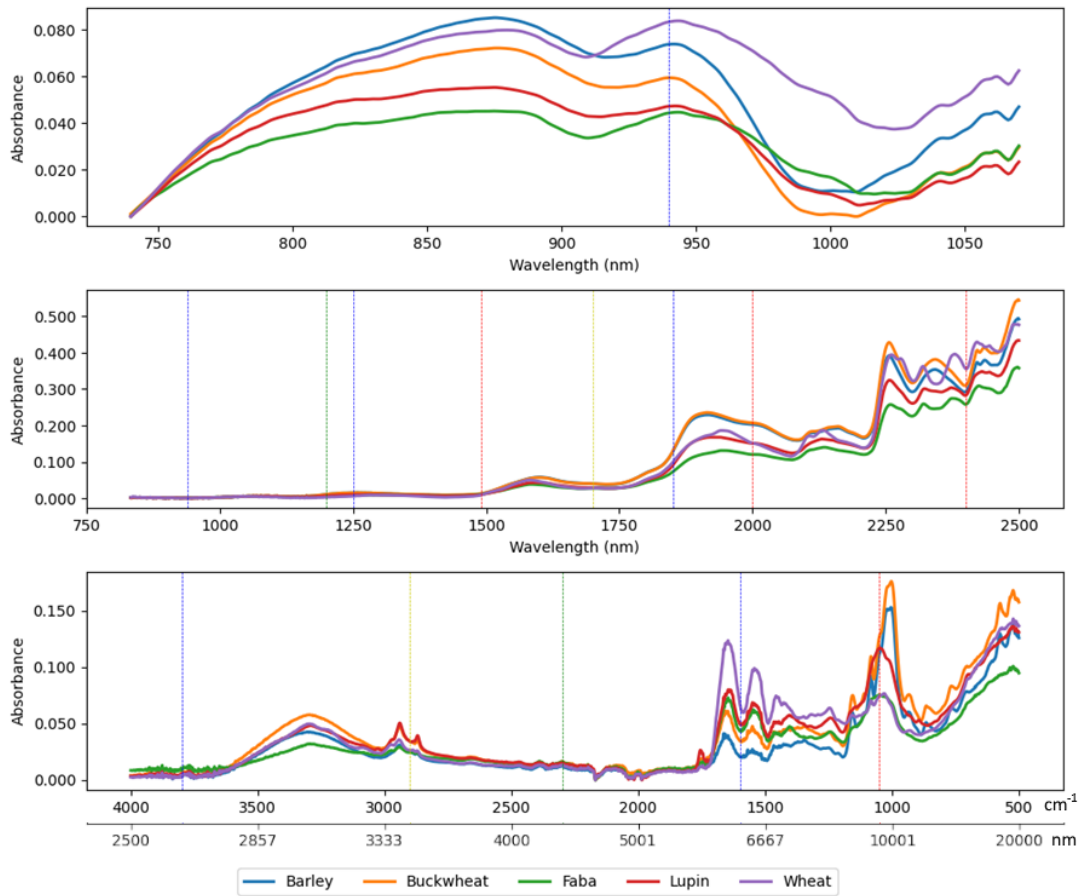


Figure 3: Raw spectra from low-cost NIR (a), benchtop NIR (b) and MIR (c) of the different, dry pure protein sources. The dotted vertical lines in red (protein), blue (moisture), yellow (fat), and green (carbohydrates) are the specific bands, found in the bibliography, for the different components of the samples.

Compositional analysis of alternative protein blends using near and mid-infrared spectroscopy coupled with conventional and machine learning algorithms

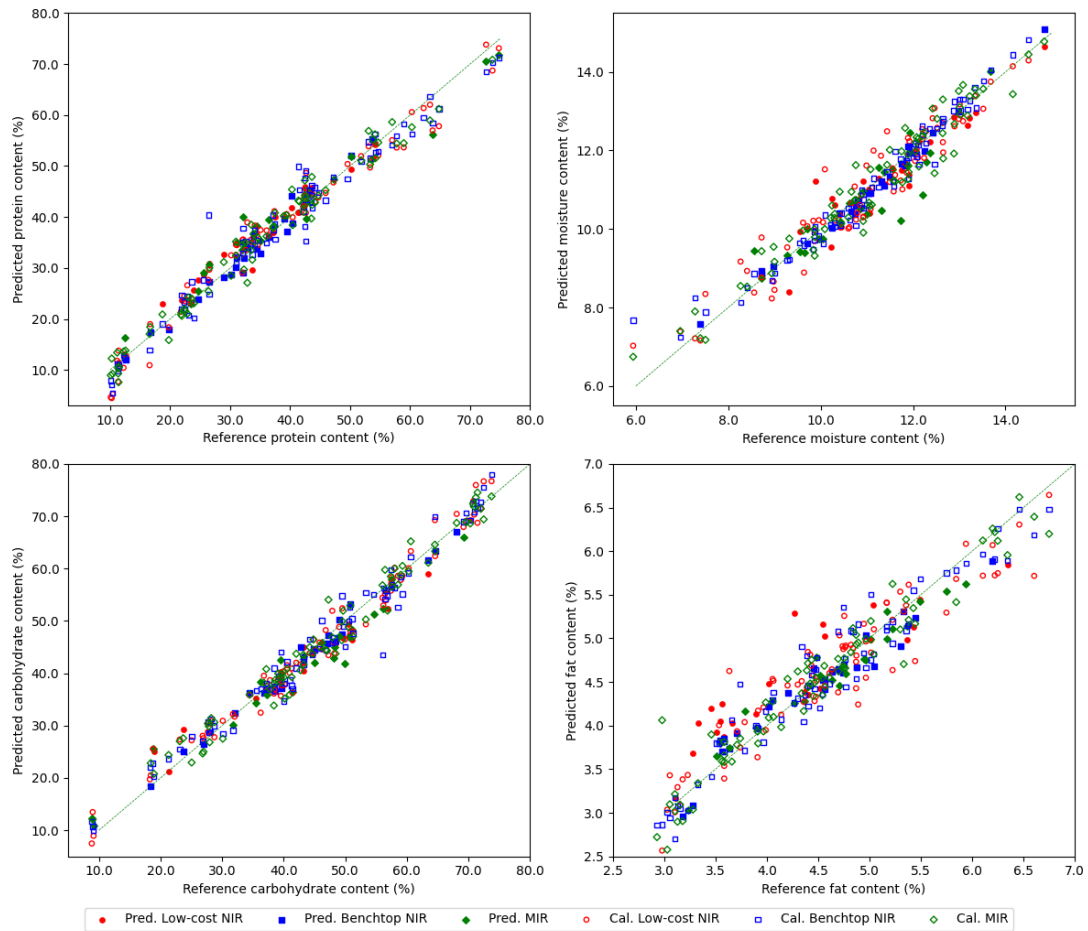


Figure 4: Relationship between reference and predicted contents using models developed with data from each spectrometer used. *Pred.*: Prediction data set on which the models were tested, *Cal.*: Calibration data set where the cross-validation was performed and on which the models were calibrated.

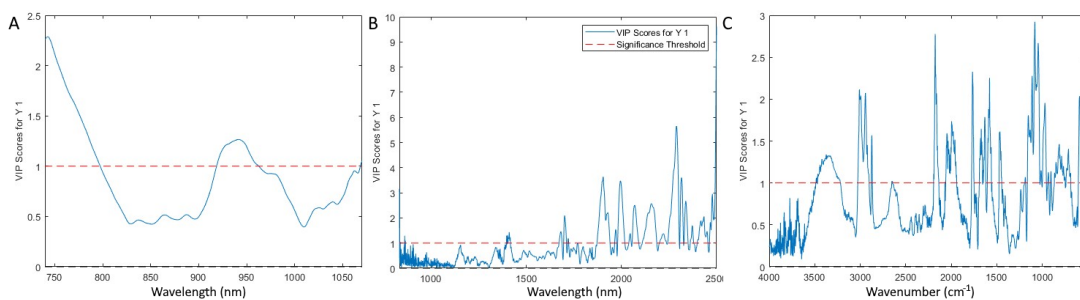


Figure 5: VIP scores of the PLS-DA performed for the low-cost NIR (A), benchtop NIR (B), and MIR (C) spectrometers.