Current Research in Food Science 9 (2024) 100799

Contents lists available at ScienceDirect

Current Research in Food Science

journal homepage: www.sciencedirect.com/journal/current-research-in-food-science

Review Article

Hyperspectral imaging as a non-destructive technique for estimating the nutritional value of food

Juan-Jesús Marín-Méndez [a,*], Paula Luri Esplandiú [a], Miriam Alonso-Santamaría [a], Berta Remirez-Moreno [a], Leyre Urtasun Del Castillo [a], Jaione Echavarri Dublán [a], Eva Almiron-Roig [b,c], María-José Sáiz-Abajo [a]

[a] National Centre for Food Technology and Safety (CNTA), Crta.NA 134-km 53, 31570, San Adrian, Navarra, Spain
[b] Centre for Nutrition Research, University of Navarra. Pamplona, Spain
[c] Centro de Investigación Biomédica en Red de Fisiopatología de la Obesidad y Nutrición (CIBEROBN), Instituto de Salud Carlos III (ISCIII), Madrid, Spain

ARTICLE INFO

Handling Editor: Dr. Maria Corradini

Keywords:
Hyperspectral imaging
Ridge regression
Nutritional value
Chemometrics
Machine learning

ABSTRACT

Knowledge of the energy and macronutrient content of complex foods is essential for the food industry and to implement population-based dietary guidelines. However, conventional methodologies are time-consuming, require the use of chemical products and the sample cannot be recovered. We hypothesize that the nutritional value of heterogeneous food products can be readily measured instead by using hyperspectral imaging systems (NIR and VIS-NIR) combined with mathematical models previously fitted with spectral profiles.118 samples from different food products were collected for building the predictive models using their hyperspectral imaging data as predictors and their nutritional values as dependent variables. Ten different models were screened (Multivariate Linear regression, Lasso regression, Rigde regression, Elastic Net regression, K-Neighbors regression, Decision trees regression, Partial Least Square, Support Vector Machines, Gradient Boosting regression and Random Forest regression). The best results were obtained with Ridge regression for all parameters. The best performance was for estimating the protein content with a RMSE of 1.02 and a R^2 equal to 0.88 in a test set, following by moisture (RMSE of 2.21 and R^2 equal to 0.85), energy value (RMSE of 21.84 and R^2 equal to 0.76) and total fat (RMSE of 2.17 and R^2 equal to 0.72). The performance with carbohydrates (RMSE of 2.12 and R^2 equal to 0.61) and ashes (RMSE of 0.25 and R^2 equal to 0.38) was worse. This study shows that it is possible to predict the energy and nutrient values of processed complex foods, using hyperspectral imaging systems combined with supervised machine learning methods.

1. Introduction

Determining the nutritional value of foods is important for the food industry to allow for accurate package labelling and appropriate packaging conditions, as well as the development of new food products and innovation of existing ones. In addition, knowing the composition of foods in terms of nutrients and energy content is essential for the construction of food composition databases, which are essential for designing therapeutic diets and formulating population dietary guidelines, amongst other uses (Sociedad Española de Nutrición (SEN), 2017). Despite recent progress in the field of food technology, there is still a growing interest in finding faster, cheaper and nondestructive techniques for determining the nutritional composition of foods. With the incorporation of artificial intelligence (AI) and machine learning (ML) models in the equation, this problem is being progressively solved. The food and nutrition field is not an exception. There are a lot of applications of ML, from food safety (Deng, X. et al., 2021; Wang, X et al., 2022) to sales prediction (Tsoumakas, G., 2019).

ML is a subfield of AI that mimics the way in which humans learn. ML basically learns hidden patters from data. There are several ways of learning from data, and depending on this the algorithms used are very different. When the aim is to carry out regression or classification tasks, the type of ML learning is called supervised. These supervised ML models try to find the best way to predict a value (regression if is a continuous variable or classification when the variable is categorical)

* Corresponding author.
E-mail addresses: jmarin@cnta.es (J.-J. Marín-Méndez), pluri@cnta.es (P. Luri Esplandiú), malonso@cnta.es (M. Alonso-Santamaría), bremirez@cnta.es (B. Remirez-Moreno), lurtasun@cnta.es (L. Urtasun Del Castillo), jaioneechavarri@gmail.com (J. Echavarri Dublán), ealmiron@unav.es (E. Almiron-Roig), mjsaiz@cnta.es (M.-J. Sáiz-Abajo).

https://doi.org/10.1016/j.crfs.2024.100799
Received 8 March 2024; Received in revised form 7 May 2024; Accepted 23 June 2024
Available online 25 June 2024

Abbreviations

AI	artificial intelligence
ML	machine learning
HSI	hyperspectral imaging
SNV	standard normal variate
SG1	Savitzky–Golay +1st derivate
SG2	Savitzky–Golay +2nd derivate
VIS-NIR	visible-near infrared
NIR	near infrared
IR	infrared
nm	nanometers
MC	mean centering
RMSE	root mean squared error
RMSEC	root mean squared error in calibration
RMSECV	root mean squared error in cross-validation
RMSEP	root mean squared error in prediction in the test set
R^2	coefficient of determination
ROI	region of interest
PCA	principal component analysis
LR	Linear regression
KNN	K-Neighbors
PLS	Partial Least Square
SVM	Support Vector Machines
GBR	Gradient Boosting regression
RF	Random Forest regression

using features from the data. The more information is provided, the more accurate the model is making its predictions.

Nutrient levels in foods are very important for human health given their proven impact on the development of several diseases (Afshin, A. et al., 2019). Specifically, the amount and type of carbohydrate, protein and fat (particularly refined starch, non-lean protein and saturated fat), as well as the energy load present in food have been directly linked with the prevalence of obesity, type 2 diabetes, stroke, cancer and other chronic diseases, for which a controlled intake of these nutrients is recommended (World Health Organization, 2019; U.S. Department of Health and Human Services, 2024). Current methods to quantitatively characterize food composition and determine its energy value, are based on analytical chemistry. However, these conventional methodologies are slow processes that require chemical products and trained technicians, moreover, they are destructive processes in which the analyzed food cannot be recovered, and sometimes expensive materials are needed (Ingle, P.D. et al., 2016).

Hyperspectral imaging (HSI) techniques use different wavelength ranges, such as ultraviolet (200–400 nm), visible (380–800 nm), VIS-NIR (400–1000 nm), NIR (900–1700 nm), and short-wave infrared (970–2500 nm). These wavelengths have been developed for optical sensing of different types of samples (Lohumi, S.et al., 2015). One can add a third dimension to the data combining several images from the object and create a hypercube, where each pixel has spectral data, this is known as a hyperspectral image. Overtones of the fundamental vibrations occurring in the infrared (IR) region are the origin of the different absorptions, the predominantly features including: the methyl C–H stretching vibrations, methylene C–H stretching vibrations, aromatic C–H stretching vibrations, and O–H stretching vibrations. There are other less predominant features: methoxy C–H stretching, carbonyl associated C–H stretching; N–H from primary amides, secondary amides (both alkyl, and aryl group associations), N–H from primary, secondary, and tertiary amines, and N–H from amine salts (Wiley Analytical Science, 2014). The NIR methods relies on the correlation between harmonic vibrations and quantity of absorber and type of absorbing molecules that are present in a sample (Wiley Analytical Science, 2014).

The combination of HSI systems together with AI is cornerstone to finding low-cost, rapid and nondestructive methods that can be used alongside conventional techniques. It is called as “low-cost” because, although the initial outlay is high and must be calibrated, maintenance is simple afterwards, and its use only requires taking photos to extract the spectral values. HSI is already being widely applied to determine the nutritional composition of animal feed as well as food products for human consumption, thanks especially to its ability to accurately predict nutritive values for protein and fat (Givens, D. et al., 1997). The absorption spectrum in the NIR region can provide information about the nutritional components of the sample. The NIR spectral region can be useful for predicting the content of nutritional values because food components have absorptions peaks in this region (Ingle, P.D.et al., 2016) that represent the vibration of atoms bonding. On the other hand, VIS-NIR spectroscopy is a molecular/vibrational technique used to study the interactions of electromagnetic waves within a sample. Vibrations also provide information about the general molecular conformation, structure and intermolecular interactions within a sample (Ghidini, S. et al., 2019). The main two advantages of hyperspectral imaging are on the one hand its speed, and second, that it does not require sample homogenization because the entire sample can be scanned while accounting for sample heterogeneity (Kämper, W. et al., 2020), but the penetrance of NIR radiation is low, this means that stacked food will be an issue that has to be solved by expanding the feed well on the surface. Based on this principle, HSI could be useful to measure the nutritional composition of complex food products that consist of a combination of different ingredients like broad beans and green beans or potato, carrot, peas (Fig. 1). The ingredient combined in the products could be either vegetables that come from different plants, different vegetable parts (roots, seeds, tubers) or vegetables combined with meat elements like ham or minced meat. meat.

Currently there is a lack of tools to measure the nutritional composition of complex meals in a time-efficient way, beyond using manufacturers’ packaging information (food labels) and calculating proportions (which is only an estimative method). The use of HSI systems together with ML models offers a great opportunity to solve this problem, by obtaining objective results in real time from intact food products that do not need to be discarded and can be directly consumed or displayed in supermarket shelves.

Knowing the energy and macronutrient content of foods such as

Fig. 1. RGB images obtained from different dishes broad beans and green beans (left) and potato, carrot, and peas (right). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

protein, fat and carbohydrate levels is essential for food development and for human health. We hypothesize that we can estimate the energy and macronutrient content of heterogeneous food products using HSI systems combined with mathematical models previously fitted with spectral profiles. These techniques could replace time-consuming and destructive chemical analyses when fast results are needed.

2. Material and methods

2.1. Samples

One hundred and eighteen samples from different food products were collected between 2021 and 2022. The samples were purchased in local supermarkets in Navarra, Spain (Eroski, Alcampo and Mercadona). All the samples included more than one ingredient in their formulation and consisted of mixtures of cooked vegetables, legumes, meat, rice, pasta, sauces and similar combinations. As can be shown in the Figs. 1 and 2, the samples were very heterogeneous, them could be products ready to eat or mixes ready to cock. The ingredients of the samples can be of the same nature (vegetables) (Fig. 1) or mixes (meat and vegetables) (Fig. 2), and the section of the vegetable could be variable (roots, tubers, seeds …) (Fig. 1). All samples were stored according to the manufacturer's instructions (including at room temperature, 4 °C or −20 °C) until analysis.

2.2. Nutritional analysis

Conventional nutritional analyses were performed at the National Centre for Food Technology and Safety (CNTA) laboratory according to the following methods. Moisture/dry matter was obtained by the gravimetric method according to standard methods (Métodos Oficiales de Análisis de Alimentos. AMV Ediciones Mundi Prensa (1994)). The previously homogenized sample was dried at 102 °C until constant weight. Fat content was determined by gravimetry after Soxhlet extraction according to standard methods (Order of 17 September 1981; Order of 31 July 1979). Protein content in samples was determined by volumetric assay through Kjeldhal digestion according to standard methods (Métodos Oficiales de Análisis de Alimentos. AMV Ediciones Mundi Prensa. (1994)). The previously homogenized sample was first digested with sulfuric acid, then distilled to solubilize the ammonium cations and finally titrated with hydrochloric acid. Ash was determined by gravimetry after calcination of the sample according to standard methods (Métodos Oficiales de Análisis de Alimentos. AMV Ediciones Mundi Prensa. (1994)). The previously homogenized sample was pre-dried at 98 °C for 30 min and then calcinated in muffle at 550 °C for 8 h. The tempered sample was finally weighted. Finally, macronutrient and energy value were estimated with the Atwater conversion factors as described previously (Food energy – methods of analysis and conversion factors:2002Food energy – methods of analysis and conversion factors:2002) using the formulas (1 and 2):

$$Carbohydrates\left(\frac{g}{100g}\right) = 100 - Moisture\left(\frac{g}{100g}\right) - Ash\left(\frac{g}{100g}\right) - Fat\left(\frac{g}{100g}\right) - Protein\left(\frac{g}{100g}\right) \quad (1)$$

$$\frac{Kcal}{100g} = [4 * (Carbohydrates + Protein) + (9 * Fat)] \quad (2)$$

For whole nutritional analysis 400 g of sample was used. After homogenization the sample was split into aliquots of 1–5 g each to measure each nutritional parameter. Homogenization was performed using a GRINDOMIX or Ultra-Turrax equipment depending on the degree of sample heterogeneity.

Fig. 2. RGB picture form different dishes on the left tripes with chickpeas and on the right meatballs with peas and tomato sauce.

2.3. Hyperspectral imaging systems

The imaging system employed acquires the hyperspectral images in the reflectance mode. The HSI system consists of two components: a line-scanning with two spectrographs (Specim FX10 and Specim FX17, Specim, Spectral Imaging Ltd., Oulu, Finland) covering the spectral range of the VIS-NIR region (400–1000 nm) and the NIR region (900–1700 nm); and an illumination source including a group of six light bulbs of stabilized halogen (each bulb has a voltage of 12 V and a power of 20 W), where three light bulbs are located at each side and with a perpendicular angle to illuminate the mobile platform where the tray with the sample is placed. Additionally, the HIS equipment requires a computer system equipped with an imaging acquisition software (Lumo-Scanner, Specim, Spectral Imaging Ltd., Oulu, Finland) to allow adjusting the most important parameters to obtain high quality images. Moreover, the spectral resolution is 3.5 nm and the spatial sampling of the camera is 640 pixels. The Fig. 3 shows a picture of imagen acquisition system employee and in the Table 1 the attributes and parameters for both HIS systems is summarized.

2.4. Hyperspectral imaging sampling

The process to acquire the hyperspectral image was the following. Frozen samples were thawed and samples that were stored in the fridge were tempered at room temperature, before measuring. Samples were then placed on a tray before placing them in the conveyor. The tray dimensions are 17,2 × 12,7 cm, and it can fit around 350–400 g. A single photo was taken of each sample after configuring the parameters. The collected images are named "hypercubes" with three dimensions (x,y,λ). The dimension of the original RGB image were of 1614 × 640 x 3 pixels, after reducing the background and the white plastic tray the image is reduced obtaining a hypercube dimension of 524 × 320 x 3 pixels. After the image was obtained, a segmentation process was carried out to retain only the values of the spectrum that match the product, called the region of interest (ROI), and to avoid capturing information about the background. This segmentation process is based on the selection of three wavelengths (1056.9 nm, 1213.87 nm, 1347.3 nm for NIR range and 797.59 nm, 542.91 nm, 484.47 nm for VNIR range) and by developing code in Matlab. The size of ROI was adapted to each sample, because the quantity of each recipe was different. The Fig. 4 shows the different steps followed from RGB image acquisition to the final ROI image. After selecting the ROI, the extraction of the reflectance spectrum was carried out. This was obtained as the average of all the spectra associated with each pixel in the region of interest so finally, it was obtained one spectrum for each sample. Subsequently, the correction was made with the corresponding black and white according to the following equation (3), being reflectance the calibrated true reflectance, R_0 the measurement,

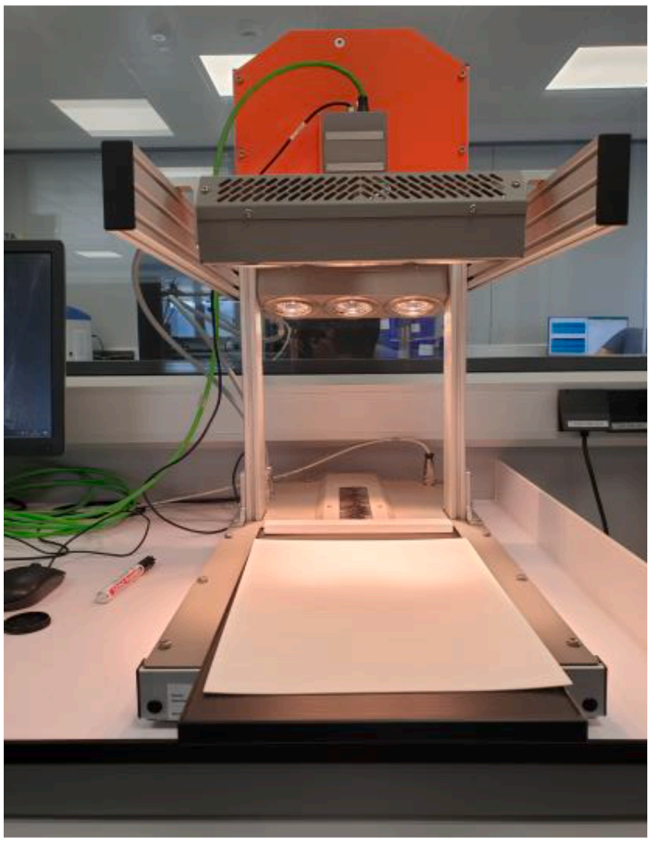

Fig. 3. Imagen acquisition system employee.

Table 1
Summary of attributes and parameters for HSI systems (VIS-NIR and NIR).

	Specim FX10 (VIS-NIR)	Specim FX17 (NIR)
Wavelength range	400–1000 nm	900–1700 nm
Variables	224	224
Spectral resolution	2,7 mm	3,5 mm
Image resolution (pixels)	1024 × 1024	640 × 640
Motor speed	10 mm/s	8 mm/s
Exposure time	19 ms	2 ms
Frame rate	40 Hz	20 Hz
Detector	CMOS	InGaAs

and R_w and R_d the white and dark reference, respectively.

$$Reflectance = \frac{R_0 - R_d}{R_w - R_d} \quad (3)$$

For this reason, a black and white reference image was captured beforehand. To capture reference white was used a uniform Teflon white tile with about 100% reflectance, and the dark reference image of 0% reflectance was captured automatically by closing the lens of the camera.

The same method was repeated for all ROI images resulting in a spectral matrix of approximately 118 samples × 224 bands.

After the hyperspectral imaging was obtained, the samples were grinded to measure the nutritional parameters.

2.5. Data preprocessing, analysis and modelling

Data analysis and modelling were performed with Python (Python Software Foundation). From the spectra obtained by the hyperspectral camera, an analysis pipeline was carried out including different preprocessing methods and models in order to find the combination that fits better the hyperspectral imaging with the actual nutritional values.

To evaluate the predictive ability and to determine the adjustment of the model two regression metrics were used: RMSE (root mean squared error) and R^2 coefficient. The first one measures the difference between the model's predicted values and the actual values (the smaller the RMSE value, the better). The second one is a metric that assesses the fit of the model to an adjusted regression line. It shows us how much of the target variable's variation can be explained by the model.

The normalized spectra obtained by the camera were preprocessed with different methods (standard normal variate (SNV) method, the Savitzky–Golay + 1st derivate (SG1) and the Savitzky–Golay + 2nd derivate (SG2), all followed by mean centering (MC). Savitzky–Golay methods include a smoothing step before applying the derivate.

The objective of each method is different. SNV belongs to scatter-corrective methods that try to remove the scatter and optical path variations. This method eliminates the multiplicative interferences produced by the diffraction and difference in the particle size and its effect is independent of the original absorption values (Barnes, R.J. et al., 1993; Echávarri-Dublán, J. et al., 2022). SNV centers and scales each spectra obtaining a mean value and variance equal to 0 and 1 respectively. On the other hand, Savitzky–Golay belongs to a class of smoothing methods and is used to reduce the random noise appearing in the raw signal. First and second derivatives are techniques used to remove a baseline shift from the signal. It is based on adjusting an appropriate polynomial degree for a small wavelength interval. This changes the original values but removes some of the noise that affects

Fig. 4. Original RGB image (on the left), original RGB image cropped (on the middle) and ROI image (on the right).

the spectra (Todeschini, R. 1998; Echávarri-Dublán, J.et al., 2022). The second derivative removes linear and constant background noises. Mean centering consists in changing the origin of the new variable scale to the mean of the variable before centering. The fundamental property of centered data is that the mean value of each of the variables is equal to zero. This pretreatment does not modify the variance of the data (Todeschini, R. 1998; Echávarri-Dublán, J. et al., 2022). Finally, before modelling (except for PLS), a dimension reduction method was applied to the data, Principal Component Analysis (PCA). This technique is useful for both exploring high dimensional data and reducing the complexity of the data. PCA is an unsupervised technique for pattern recognition based on the capture of the major source of explained variance. This technique transforms the predictive features into orthogonal components, which solves the problem of multicollinearity and reduces dimensionality (Vega-Vilca, JC, and Guzmán, JJ. 2011). PCA decomposes the original raw data (spectra) into loading and scoring matrices obtaining a lower number of principal components compared to the original features but retaining the source of variation present in the original data. Loadings are valuable for assessing the significance of each feature in explaining the variability observed in the principal components. After the transformation, the early first components are able to explain most of the variation in the original data (Zhu, F. et al., 2013; Jiang, H. et al., 2021; Cheng, J. H. et al., 2015; Li, P. et al., 2023). This allows exploring the data distribution using two dimensional or three-dimensional scatter plots with the first, second and third principal components, and thus identify possible outliers or new trends in the data.

To find which model better fitted our data, up to 10 algorithms were evaluated including linear and nonlinear models and ensemble methods using the Scikit learn library (Pedregosa, F. et al., 2011) in Python. This process is referred to as the "screening stage" (see details below). Multivariate Linear regression (MLR), Lasso regression (Lasso), Rigde regression (Ridge), Elastic Net regression (ElasticNet), K-Neighbors (KNN), Decision trees (Trees), Partial Least Square (PLS), Support Vector Machines (SVM), Gradient Boosting regression (GBR) and Random Forest regression (RF) were applied. The peculiarities of each method are out of the scope of this work but can be consulted on-line (Scikit-learn: machine learning in Python).

Previous modeling, to determine the presence of outliers in the sample, a combination of techniques and metrics was used (PCA and Leverge). PCA was used to visually evaluate the distribution of the data in two or three dimensions. Leverage is a PCA-based measure of outlyingness. Leverage can be used to assess the potential influence of each observation on the regression fit, and it is bounded between 0 and 1. This measure is related to the Mahalanobis distance (Mejia, A.F. et al., 2017).

To evaluate the performance of the models the dataset was randomly split into a training set (80% from original data) and a test set. The training set was used to select the best combination of preprocess method and model using 5 repeated 5-fold cross validation and to fit the selected combination. Cross validation is used to estimate the performance of algorithms with less variance than a single split into train and validation. The 5-fold cross validation splits the data in 5 parts, four of them are used to train the algorithm holding one to assess the performance of the model. This process is repeated as many times as each fold of the dataset is given a chance to be used to evaluate the model. So, when the cross validation is finished, you end up with 5 different performance scores. In our study, we ended up with 25 as a repeated 5-fold cross validation algorithm was used. Finally, you can summarize the performance of the model using mean and standard deviation.

The screening stage mentioned above tries to find the model that best fits the data and encompasses several steps. The first step consists in establishing a battery of different models to be evaluated (in our study we selected 10 algorithms) and select the metric that will be applied to evaluate the performance of the models. As this is a regression problem RMSE was used to compare which model fits better. After that, each algorithm in its raw configuration (without modifying any hyperparameter) is trained and its performance is evaluated using cross validation which results in an RMSE aggregated score for each model. After this step, each algorithm becomes ranked according to RMSE which allows to select the model (or models) that fits better to the data (lower values of RMSE).

Once the model with the best performance has been identified in its raw configuration, hyperparameter tuning is needed. The hyperparameter tuning is an iterative process in which every iteration changes an hyperparameter value following the evaluation of the performance of the model configuration to achieve the best configuration of hyperparameters that produces the best results according to RMSE. To develop this process there are several approaches, a systematic search which evaluates each possible combination of hyperparameters, a random search which evaluates a random selection of combinations of hyperparameters and a Bayesian optimization that chooses the combinations of hyperparameters based on previous results. To perform this process, a training set and cross validation is used again. The approach that we chose in our study was the systematic search. The hyperparameters that our group used to tune the algorithms are summarized in the Table 2.

After that process, the test set (test set denomination is used to avoid confusion with cross validation) was used to validate the performance of the fitted model. The use of training and test sets give us information about the performance of the model in terms of bias, variance and fitting. Bias is the error that occurs when there are differences between the predicted value and the actual value. These errors are systematic and occur when wrong assumptions are made in the process. On the other hand, the variance shows how the model depends on the data chosen to fit the predictive model, in other words, how the performance of the model changes when it is fitted with different subsets of data. The balance between these two concepts leads to the third one, fitting. When there is a low variance with a high bias, the model is underfitted; a high variance and a low bias leads to overfitting; a high variance and a high bias means the model is inconsistent and inaccurate. Finally, with a low variance and a low bias, the model is able to generalize well and the predictions will be consistent and accurate (although this is impossible in practice).

To perform all these process Python 3.11 was used. The libraries employed were pandas, numpy, sklearn, matplotlib, seaborn, statsmodels and scipy.

3. Results

3.1. Description of samples and nutritional analysis

After the identification of outliers, the original pictures were checked in order to assess if there were any artefact in it. This process concluded that the deviation of the 7 NIR samples were produced by technical sources so they were removed. The final set used to develop the different predictive models consisted of 111 NIR samples.

In the case of the Vis-NIR samples, no outliers were detected, so the process was carried out with 118 samples.

A basic description of different nutritional parameters measured by conventional analysis is summarized in Table 3.

As observed in Table 3, each objective variable was continuous, with a wide range of values, probably reflecting the wide range of food products included in the study. As expected, the attribute with the greatest range was energy content (range from 23.5 kcal/100 g to 204 kcal/100 g).

3.2. NIRs, Vis-NIRs analysis and modeling

Despite the raw wavelength range was from 935.61 to 1720 nm, the spectra were trimmed at the end (from 1670 to 1720 nm) because these wavelengths show some random noise. For the Vis-NIR spectra, there was no random noise so the whole spectra were used.

Table 2
Specific hyperparameters selected to be tuned in case that the algorithm is selected to hyperparameter tuning.

Model	Hyperparameters specific to each model
Multivariate Linear regression	There is not hyperparameter to be tunned
Lasso regression	- alpha (constant that multiplies L1 term) - random_state (Controls the randomness of the estimator)
Rigde regression	- alpha (constant that multiplies L2 term) - random_state (Controls the randomness of the estimator)
Elastic Net regression	- alpha (constant that multiplies penalization terms) - l1_ratio (elastic net mixing parameter) - random_state (Controls the randomness of the estimator)
K-Neighbors	- n_neighbors (Number of neighbors) - algorithm (algorithm used to compute the nearest neighbors) - metric (metric to use the distance computation) - weights (Weight function used in prediction) - leaf_size (Leaf size passed to BallTree or KDTree)
Decision trees	- criterion (to measure the quality of a split) - max_depth (Maximum depth of the tree) - min_samples_split (Minimum number of samples required to split an internal node) - min_samples_leaf (Minimum of samples required to be at a leaf) - random_state (Controls the randomness of the estimator)
Partial Least Square	- n_components (Number of components to keep, in other languages are known as latent variables) - tolerance (Is a convergence criteria)
Support Vector Machines	- kernel (Kernel type to be use in the algorithm) - gamma (kernel coefficient) - C (Regularization parameter) - tol (Criterion for stopping)
Gradient Boosting regression	- loss (Loss fuction to be optimized) - learning_rate (Learning rate shrinks the contribution of each tree) - n _estimators (number of boosting stages to perform) - criterion (Function to measure the quality of a split) - min_samples_leaf (Minimum of samples required to be at a leaf) - min_samples_split (Minimum number of samples required to split an internal node) - max_depth (Maximum depth of the individual regression estimator) - max_feature (The number of features to consider when looking for the best split) - random_state (Controls the randomness of the estimator)
Random Forest regression	- n _estimators (number of boosting stages to perform) - criterion (Function to measure the quality of a split) - min_samples_leaf (Minimum of samples required to be at a leaf) - min_samples_split (Minimum number of samples required to split an internal node) - max_depth (Maximum depth of the individual regression estimator) - max_feature (The number of features to consider when looking for the best split) - boostrap (Whether bootstrap samples are used when building trees) - random_state (Controls the randomness of the estimator)

After the random splitting, 80% of the spectra were assigned to the calibration (training) set, while the remaining 20% were assigned to the test (validation) set, for both the NIR and Vis-NIR samples sets.

In the process of screening, we evaluated the performance of combine the different preprocessing techniques (SNV, SG1, SG2 with and without mean centering and with and without dimensional reduction) and each algorithm evaluated (algorithms in raw configuration). It means that we evaluated 120 combinations for each nutritional parameter and energetic value. Using cross validation techniques (repeated 5-fold cross validation), we obtained an aggregated RMSE value for each combination of preprocessing technique and algorithm, this allows us to select the better (one or more) combination for each nutritional parameter. These were the algorithms selected to be subjected to hyperparameter tuning through cross validation. This process was done twice, once using the NIR spectra and once using VIS-NIR spectra.

Once the process of screening of different pipelines was done, the better performance was obtained using NIR spectra starting with SNV, followed by MC (preprocessing) and PCA (dimensional reduction), and using the mathematical algorithm of Ridge regression (linear regression penalized with Ridge) as predictive model for all the parameters evaluated.

Table 4 summarizes the results obtained for each nutritional parameter in both the training (cross validation and whole training set) and the test sets.

As it can be observed, there are no major differences between the results obtained in the training set vs. the test sets (except for ashes). This is indicative of a low variance and means that the model is able to generalize when new data are supplied.

According to the R^2 value, the nutritional parameter with the best fit, was protein content with a value of 0.88 in the test set, plus a RMSEP of 1.02. The next best parameter was moisture with a R^2 value of 0.85, and RMSEP of 2.21 (test set). The remaining parameters have values smaller than 0.80 but greater than 0.60, except for the model fitted with ashes which has a poor performance (R^2 of 0.38).

The Fig. 5 shows the different scatterplots showing the predicted values with Ridge regression models using NIR spectra versus the actual values for each parameter.

Table 5 summarizes the results obtained with the VIS-NIR spectra.

Table 4
Summary of regression metrics obtained for the different nutritional parameters in the pipeline SNV-MC-PCA-Ridge regression applied in data obtained in NIR wavelength range. RMSEC: root mean squared error in calibration set; RMSECV: root mean squared error in cross validation; RMSEP: root mean squared error in test set.

Nutritional parameter	Training set			Test set	
	RMSEC	R^2	RMSECV	RMSEP	R^2
Energy	16.20	0.86	20.01	21.84	0.76
Protein	0.79	0.95	1.56	1.02	0.88
Total fat	1.70	0.80	2.06	2.17	0.72
Carbohydrates	1.09	0.86	1.87	2.12	0.61
Ashes	0.15	0.85	0.35	0.25	0.38
Moisture	2.34	0.86	2.96	2.21	0.85

Table 3
Statistical descriptive analysis for each conventional nutritional parameter measured in the different food products.

Nutritional parameter	Whole dataset		Training set		Test set	
	Range	Mean ± SD	Range	Mean ± SD	Range	Mean ± SD
Energy (kcal/100 g)	23.5–204	113.96 ± 44.15	23.5–204	106.32 ± 41.23	38.00–190	116.96 ± 45.95
Protein (g/100 g)	1.12–17.36	6.54 ± 3.77	1.12–17.36	6.23 ± 3.61	1.30–16.19	6.46 ± 3.93
Total fat (g/100 g)	0.3–15	6.14 ± 3.92	0.3–15	5.32 ± 3.72	0.30–15.00	6.34 ± 4.29
Carbohydrates (g/100 g)	0.5–16	7.07 ± 3.12	0.5–16	7.33 ± 3.03	0.5–11.70	7.37 ± 3.12
Ashes (g/100 g)	0.3–2.20	1.34 ± 0.37	0.3–2.10	1.34 ± 0.34	0.3–2.20	1.43 ± 0.41
Moisture (g/100 g)	66.0–92.40	77.17 ± 6.40	66.0–92.40	77.83 ± 6.01	68.5–88.00	76.90 ± 5.97

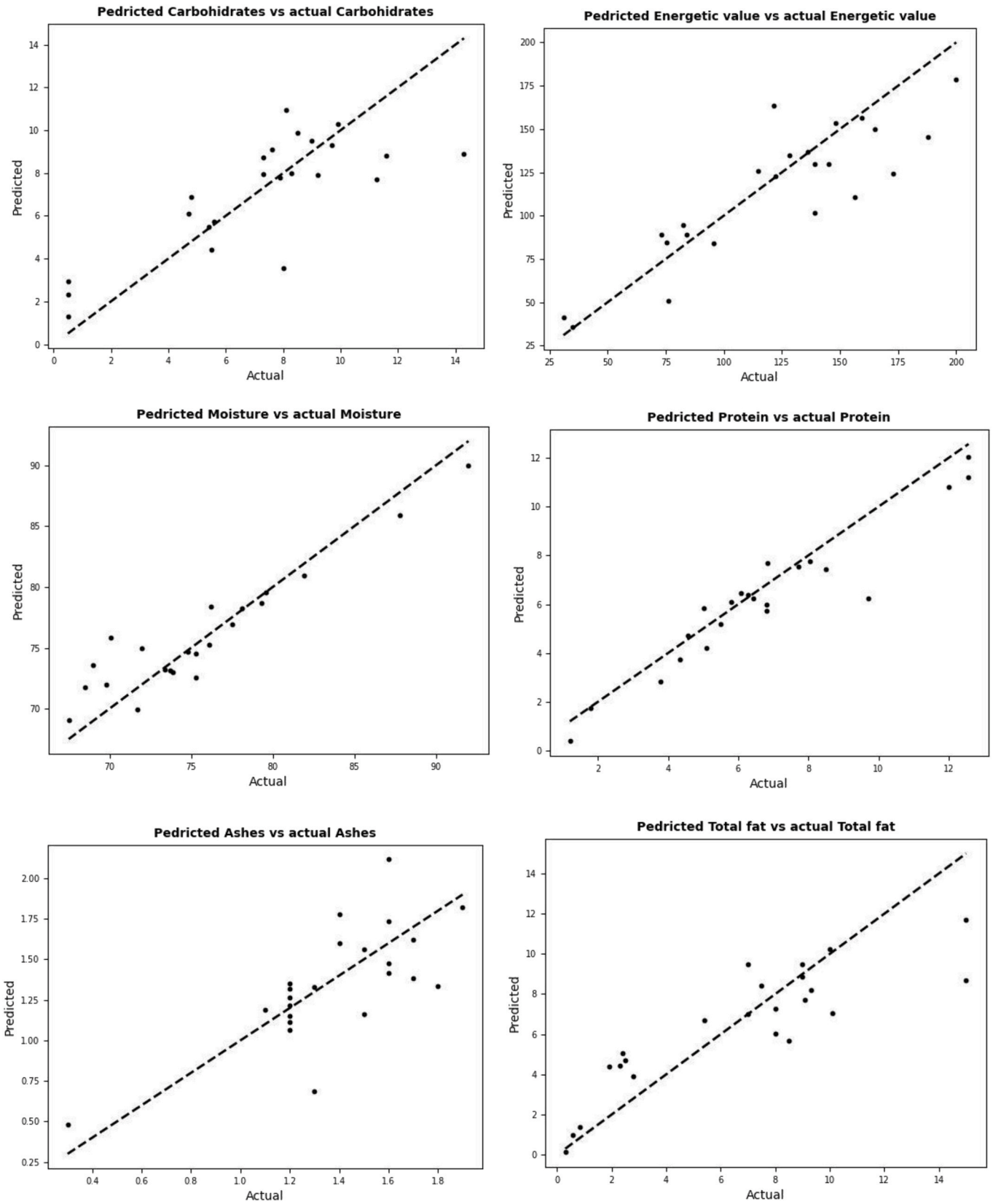

Fig. 5. Scatter plots of predicted values versus actual values for each parameter: carbohydrates (up-left), energetic value (up-right), moisture (middle-left), protein (middle-right), ashes (bottom-left) and total fat (bottom-right) using Ridge regression model trained with NIR spectra and processed with SNV + MC + PAC.

All the parameters showed worst results compared with the NIR spectra (Table 4). The best results with the Vis-NIR spectra were for energy and protein content with a R^2 of 0.63 plus RMSEP of 29.73 in the test set for energy; and R^2 of 0.63 with RMSEP of 2.38 for protein. The remaining parameters had values of R^2 under 0.60 and two parameters (carbohydrates and ashes) had R^2 values close to 0, indicating unsuitability of this method for these measurements. The Fig. 6 shows the different scatterplots showing the predicted values versus the actual values for each parameter using with Ridge regression models with VIS-NIR spectra.

The results obtained with both spectral range in the parameters measured, could indicate that most of the information related with protein, total fat and energy values are in the range of NIR spectra starting from 900 nm as models using NIR range performs better but models using VIS-NIR don't do it so bad. However, the results indicates that the information for ashes and carbohydrates could be in other spectral range, maybe in UV as neither using NIR and VIS-NIR spectra together ML models showed a good performance.

4. Discussion

Up to now the use of spectral imaging for food composition analysis has focused on the determination of components and/or classification in single ingredient food products. To our knowledge, the present work represents the first study where HSI is used for determining the energetic value provided by a wide range of complex food products and also their nutritional components (protein, carbohydrates and total fat).

Spectral imaging techniques measure the variance in wavelengths emitted by an object, and these differences provide information about its composition. HSI is a widely used technique in chemometrics that can provide useful information about the composition of a sample manly related with its nutritional components. Under the umbrella of HSI, a wide range of spectral wavelengths can be used depending on the objective of the experiment. Wavelengths range from ultraviolet (200–400 nm), visible (380–800 nm), VIS-NIR (400–1000 nm), NIR (900–1700 nm), to short-wave infrared (970–2500 nm). For the purpose of this study, both NIR and Vis-NIR were chosen. Both ranges are widely used in the field of food, for issues related to classification and regression, with good performance (Wang, L. et al., 2017; Yu, H. et al., 2021; Liu, F., and He, Y. 2008; You, H. et al., 2019; Kays, S.E. et al., 2000).

Over the last years, HIS has emerged as a non-destructive and a less time-consuming method to be applied with a diverse range of food products to determine their sensory properties (Özdoğan G. et al., 2021), internal food injuries (Guo X et al., 2023) or to predict micronutrients (Hu N et al., 2021). The use of predictive, trained AI models combined with HSI offers a much better alternative to conventional methods as can produce results in real time or close to real time and are easy to use with a little training. But the combination of AI and HIS include disadvantages like these techniques are still predictive, so they are not as accurate as the conventional biochemical techniques, and an initial investment in the equipment is necessary. However, on the whole, the use of HSI coupled with IA could be a cost-effective alternative, especially when there is a need to produce results in the very short time. Moreover, given that HSI + AI preserves the sample, it avoids measurement batch effects as it shows results for each individual sample, as opposed to a subsample, which no need to extrapolate to the rest of the sample.

Table 5
Summary of regression metrics obtained for the different nutritional parameters in the pipeline SNV-MC-PCA-Ridge regression applied in data obtained in Vis-NIR wavelength range. RMSEC: root mean squared error in calibration set; RMSECV: root mean squared error in cross validation; RMSEP: root mean squared error in test set.

Nutritional parameter	Training set			Test set	
	RMSEC	R^2	RMSECV	RMSEP	R^2
Energy	17.93	0.83	28.26	29.37	0.63
Protein	1.77	0.77	2.71	2.38	0.63
Total fat	1.92	0.75	2.71	2.58	0.58
Carbohydrates	1.54	0.74	2.70	3.35	0.08
Ashes	0.15	0.81	0.22	0.50	0.0
Moisture	3.62	0.66	4.40	4.91	0.45

As shown in the results, the best outcomes in this study were obtained using the NIR range for all the parameters, with Ridge regression. Ridge regression consists of a linear regression algorithm but with a penalization term, which changes the cost function, which means less overfitting in the model. Ridge regression adds to the cost function a penalty with the objective of avoiding very large parameters in the model, so the model ends up with two terms to minimize. Eventually this means that the model usually has a slightly higher bias but a lower variance. The lower variance makes the model better at generalizing and at making better predictions when new data (samples) are provided. The main difference between NIR and Vis-NIR is the wavelength range. That is, the optical requirements of both are different, so the cost is different being the technique of Vis-NIR cheaper than NIR.

Finally, and answering to our initial hypothesis (we can estimate the energy and macronutrient content of heterogeneous food products using HSI systems combined with mathematical models), this study has shown that it is possible to use HSI systems and machine learning models to estimate both energy and macronutrient content. However, the accuracy of the models will depend on which macronutrient we are interested in.

5. Conclusions

ML algorithms are powerful tools that can help to understand different types of complex data problems and to discover hidden patterns in the data which can be difficult to extract using other techniques. These methods have a huge dependency on the volume of the data available. The models were trained on a moderately sized dataset, and it is possible that their performance could be enhanced with a larger dataset. Another limitation is that only solid, processed food products and recipes were used, as opposed to freshly prepared meals. Therefore, for the application of this technique in food catering contexts for example, further research is needed to establish if the NIR method is still a valid option.

Based on the initial exploration of a sample of 118 commercially available foods, it is possible to predict the energy and nutritional values in processed complex foods, using hyperspectral imaging systems (both NIR and Vis-NIR), combined with artificial intelligence machine learning methods. The best predictive performance was achieved for protein content, with spectral data obtained in the NIR wavelength range following this pipeline: first, processing the spectra with SNV and apply MC; second, dimensional reduction with PCA algorithm; and finally, using the mathematical algorithm of Ridge regression as predictive model. These methods achieve good performance in a training set as well as in a test set, showing a nice generalization for new data, but the process should be validated in new data with foods of different textures and composition similar to the samples used to train the model. Also, the model's performance could be enhanced retraining the model with new data with similar composition, it means growing the number of samples using to achieve a greater knowledge from the data or using different compositions growing the variability of the sample and achieve a model that can be used to more types of dishes.

CRediT authorship contribution statement

Juan-Jesús Marín-Méndez: Conceptualization, Methodology, Software, Validation, Formal analysis, Data curation, Writing – original draft, Visualization. **Paula Luri Esplandiú:** Methodology, Investigation, Data curation, Writing – review & editing, Resources, Investigation, Validation, Software. **Miriam Alonso-Santamaría:** Methodology, Investigation, Writing – review & editing, Validation, Software. **Berta Remirez-Moreno:** Investigation, Writing – review & editing, Validation. **Leyre Urtasun Del Castillo:** Writing – review & editing,

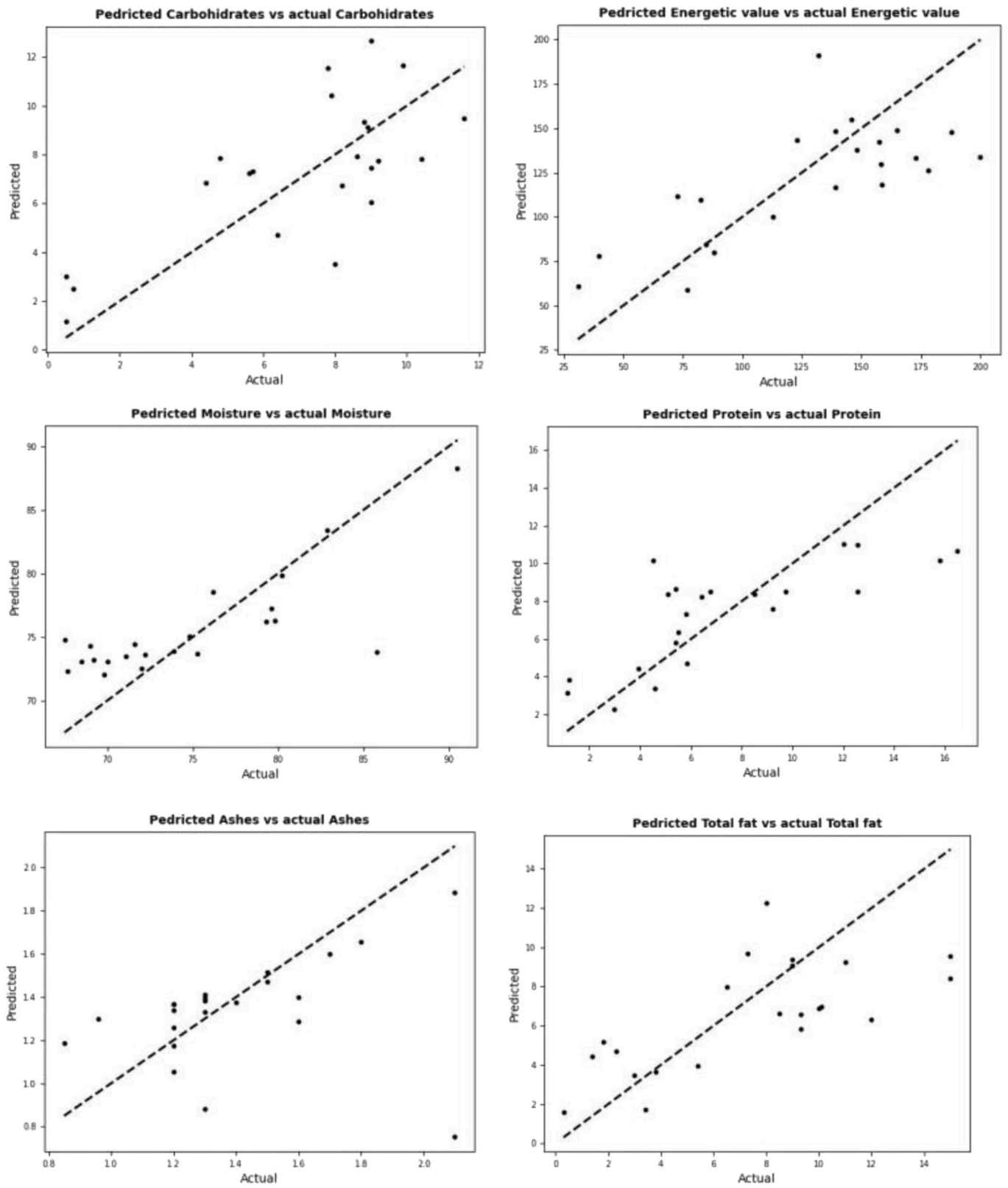

Fig. 6. Scatter plots of predicted values versus actual values for each parameter: carbohydrates (up-left), energetic value (up-right), moisture (middle-left), protein (middle-right), ashes (bottom-left) and total fat (bottom-right) using Ridge regression model trained with VIS-NIR spectra and processed with SNV + MC + PAC.

Supervision, Funding acquisition, Project administration, Methodology. **Jaione Echavarri Dublán:** Conceptualization, Methodology, Software, Writing – review & editing, Resources. **Eva Almiron-Roig:** Writing – review & editing, Funding acquisition, Project administration. **María-José Sáiz-Abajo:** Writing – review & editing, Supervision, Project administration, Methodology.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgement

This work was funded by the Government of Navarra, Departamento de Universidad, Innovación y Transformación Digital, grant number PC171-172 (PORTIONS-3), and Government of Navarra "Ayudas a Centros Tecnológicos para actividades de capacitación" (EvolTECH), with additional support from the National Centre for Food Technology and Safety (CNTA), and the Centre for Nutrition Research, University of Navarra (EA-R).

References

Afshin, A., Sur, P.J., Fay, K.A., Cornaby, L., Ferrara, G., Salama, J.S., Mullany, E.C., Abate, K.H., Abbafati, C., Abebe, Z., Afarideh, M., Aggarwal, A., Agrawal, S., Akinyemiju, T., Alahdab, F., Bacha, U., Bachman, V.F., Badali, H., Badawi, A., et al., 2019. Health effects of dietary risks in 195 countries, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet 393 (10184), 1958–1972. https://doi.org/10.1016/S0140-6736(19)30041-8.

Barnes, R.J., Dhanoa, M.S., Lister, S.J., 1993. Correction to the description of standard normal variate (SNV) and De-trend transformations in practical spectroscopy with applications in food and beverage analysis - 2nd edition. J. Near Infrared Spectrosc. 1 (3), 185. https://doi.org/10.1255/jnirs.21.

Cheng, J.H., Sun, D.W., Pu, H.B., Chen, X., Liu, Y., Zhang, H., Li, J.L., 2015. Integration of classifiers analysis and hyperspectral imaging for rapid discrimination of fresh from cold-stored and frozen-thawed fish fillets. J. Food Eng. 161, 33–39. https://doi.org/10.1016/j.jfoodeng.2015.03.011.

Deng, X., Cao, S., Horn, A.L., 2021. Emerging applications of machine learning in food safety. Annu. Rev. Food Sci. Technol. 12, 513–538.

Echávarri-Dublán, J., Alonso-Santamaría, M., P. Luri-Esplandiú, P., M.-J. Sáiz-Abajo, M. J., 2022. Comparison of different illumination systems for moisture prediction in cereal bars using hyperspectral imaging technology. J. Spectr. Imaging 11, a10. https://doi.org/10.1255/jsi.2022.a10.

Food energy – methods of analysis and conversion factors, 2002. FAO FOOD and NUTRITION PAPER 77. Report of a Technical Workshop, Rome, 3-6 December 2002. Food and Agriculture Organization of The United Nations, Rome, 2003 (ISBN 92-5-105014-7).

Ghidini, S., Varrà, M.O., Zanardi, E., 2019. Approaching authenticity issues in fish and seafood products by qualitative spectroscopy and chemometrics. In: Molecules, vol 24. MDPI AG.

Givens, D., De Boever, J., Deaville, E., 1997. The principles, practices and some future applications of near infrared spectroscopy for predicting the nutritive value of foods for animals and humans. Nutr. Res. Rev. 10 (1), 83–114. https://doi.org/10.1079/NRR19970006.

Guo, X., Tseung, C., Zare, A., Liu, T., 2023. Hyperspectral image analysis for the evaluation of chilling injury in avocado fruit during cold storage. Postharvest Biol. Technol. 206, 112548 https://doi.org/10.1016/j.postharvbio.2023.112548. ISSN 0925-5214.

Hu, N., Li, W., Du, C., Zhang, Z., Gao, Y., Sun, Z., Yang, L., Yu, K., Zhang, Y., Wang, Z., 2021. Predicting micronutrients of wheat using hyperspectral imaging. Food Chem. 343, 128473 https://doi.org/10.1016/j.foodchem.2020.128473. ISSN 0308-8146.

Ingle, P.D., Christian, R., Purohit, P., Zarraga, V., Handley, E., Freel, K., Abdo, S., 2016. Determination of protein content by NIR spectroscopy in protein powder mix products. J. AOAC Int. 99 (2), 360–363. https://doi.org/10.5740/jaoacint.15-0115.

Jiang, H., Ru, Y., Chen, Q., Wang, J., Xu, L., 2021. Near-infrared hyperspectral imaging for detection and visualization of offal adulteration in ground pork. Spectrochim. Acta Mol. Biomol. Spectrosc. 249 https://doi.org/10.1016/j.saa.2020.119307.

Kämper, W., Trueman, S.J., Tahmasbian, I., Bai, S.H., 2020. Rapid determination of nutrient concentrations in hass avocado fruit by Vis/NIR hyperspectral imaging of flesh or skin. Rem. Sens. 12 (20), 3409. https://doi.org/10.3390/rs12203409.

Kays, S.E., Barton, F.E., Windham, W.R., 2000. Predicting protein content by near infrared reflectance spectroscopy in diverse cereal food products. J. Near Infrared Spectrosc. 8 (1), 35–43. https://doi.org/10.1255/jnirs.262.

Li, P., Tang, S., Chen, S., Tian, X., Zhong, N., 2023. Hyperspectral imaging combined with convolutional neural network for accurately detecting adulteration in Atlantic salmon. Food Control 147. https://doi.org/10.1016/j.foodcont.2022.109573.

Liu, F., He, Y., 2008. Classification of brands of instant noodles using Vis/NIR spectroscopy and chemometrics. Food Res. Int. 41 (5), 562–567. https://doi.org/10.1016/j.foodres.2008.03.011. ISSN 0963-9969.

Lohumi, S., Lee, S., Lee, H., Cho, B.K., 2015. A review of vibrational spectroscopic techniques for the detection of food authenticity and adulteration. In: Trends in Food Science and Technology, vol 46. Elsevier Ltd, pp. 85–98.

Mejia, A.F., Nebel, M.B., Eloyan, A., Caffo, B., Lindquist, M.A., 2017. PCA leverage: outlier detection for high-dimensional functional magnetic resonance imaging data. Biostatistics 18 (3), 521–536. https://doi.org/10.1093/biostatistics/kxw050.

Métodos Oficiales de Análisis de Alimentos, 1994. AMV Ediciones Mundi Prensa.

Order of 17 September 1981 Establishing Official Methods of Analysis of Oils and Fats, Water, Meat and Meat Products (BOE 14-10-1981).

Order of 31 July 1979 Establishing the Official Methods of Analysis of Oils and Fats, Meat Products, Cereals and Derivatives, Fertilisers, Phytosanitary Products, Dairy Products, Animal Feed, Water and Grape Products (BOE of 2.9 August 1979).

Özdoğan, G., Lin, X., Sun, D.W., 2021. Rapid and noninvasive sensory analyses of food products by hyperspectral imaging: recent application developments. Trends Food Sci. Technol. 111, 151–165. https://doi.org/10.1016/j.tifs.2021.02.044. ISSN 0924-2244.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. 12, 2825–2830.

Python Software Foundation. Python language reference. from, version 3.9.13. http://www.python.org. (Accessed 27 July 2023).

Scikit-learn. Machine learning in Python — scikit-learn 1.2.1 documentation from. https://scikit-learn.org/stable/index.html. (Accessed 27 July 2023).

Sociedad Española de Nutricion (SEN), 2017. EuroFIR » food composition databases. from. https://www.sennutricion.org/es/2013/05/07/eurofir-food-composition-databases.

Todeschini, R., 1998. Introduzione Alla Chimiometria. EdiSES, Napoli.

Tsoumakas, G., 2019. A survey of machine learning techniques for food sales prediction. Artif. Intell. Rev. 52 (1), 441–447.

U.S. Department of Health and Human Services, 2024. Home | Dietary Guidelines for Americans. https://www.dietaryguidelines.gov/. (Accessed 3 May 2024).

Vega-Vilca, J.C., Guzmán, J.J., 2011. Regresión PLS Y PCA como solución al problema de multicolinealidad en regresión múltiple. Revista de Matemática: Teoría y Aplicaciones 18 (1), 9–20. CIMPA – URC ISSN: 1409-2433.

Wang, L., Sun, D.W., Pu, H., Cheng, J.H., 2017. Quality analysis, classification, and authentication of liquid foods by near-infrared spectroscopy: a review of recent research developments. Crit. Rev. Food Sci. Nutr. 57 (7), 1524–1538. https://doi.org/10.1080/10408398.2015.1115954.

Wang, X., Bouzembrak, Y., Lansink, A.O., van der Fels-Klerx, H.J., 2022. Application of machine learning to the monitoring and prediction of food safety: a review. Compr. Rev. Food Sci. Food Saf. 21 (1), 416–434.

Wiley Analytical Science, 2014. An introduction to near infrared spectroscopy -30 june 2014 - wiley analytical science written by jerry workman of argose, inc. from. https://analyticalscience.wiley.com/content/article-do/introduction-near-infrared-spectroscopy. (Accessed 30 April 2024).

World Health Organization, 2019. Healthy diet. 20. https://www.who.int/health-topics/healthy-diet#tab=tab_2. (Accessed 3 May 2024).

You, H., Kim, H., Joo, D.-K., Lee, S.M., Kim, J., Choi, S., 2019. Classification of food powders with open set using portable VIS-NIR spectrometer. In: International Conference on Artificial Intelligence in Information and Communication. ICAIIC, Okinawa, Japan, pp. 423–426. https://doi.org/10.1109/ICAIIC.2019.8668992.

Yu, H., Guo, L., Kharbach, M., Han, W., 2021. Multi-way analysis coupled with near-infrared spectroscopy in food industry: models and applications. Foods 10 (4), 802. https://doi.org/10.3390/foods10040802.

Zhu, F., Zhang, D., He, Y., Liu, F., Sun, D.W., 2013. Application of visible and near infrared hyperspectral imaging to differentiate between fresh and frozen-thawed fish fillets. Food Bioprocess Technol. 6, 2931–2937. https://doi.org/10.1007/s11947-012-0825-6.