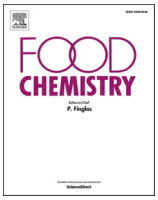
Food Chemistry 492 (2025) 145569

Contents lists available at ScienceDirect

Food Chemistry

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



# Spectral markers and machine learning: Revolutionizing Rice evaluation with near infrared spectroscopy

Pedro Sousa Sampaio [a,b,c,*], Bruna Carbas [a,d], Andreia Soares [a], Inês Sousa [a,b], Carla Brites [a,b]

[a] Instituto Nacional de Investigação Agrária e Veterinária (INIAV), Av. da República, Quinta do Marquês, 2780-157, Oeiras, Portugal
[b] GREEN-IT Bioresources for Sustainability, ITQB NOVA, Av. da República, 2780-157 Oeiras, Portugal
[c] COPELABS–Computação e Cognição Centrada nas Pessoas, Faculty of Engineering, Lusófona University, Campo Grande, 376, 1749-024 Lisbon, Portugal
[d] Centro de Investigação de Montanha (CIMO), Instituto Politécnico de Bragança, Campus de Santa Apolónia, 5300-253, Bragança, Portugal

ARTICLE INFO

*Keywords:*
Classification models
Machine learning techniques
NIR spectroscopy
PCA
PLS-DA
Rice
Spectral markers

ABSTRACT

The evaluation of rice varieties is a complex, time-consuming process requiring advanced equipment. This study aimed to discriminate 22 commercial rice varieties from six types by analyzing biochemical, physicochemical, and cooking properties. Near-infrared (NIR) spectroscopy, combined with machine learning, linked molecular properties with quality traits, offering a high-throughput solution. Partial Least Squares (PLS) models accurately predicted parameters such as whiteness ($R^2 = 0.94$), width ($R^2 = 0.94$), resilience ($R^2 = 0.96$), and springiness ($R^2 = 0.98$), highlighting key wavelength regions. Principal Component Analysis (PCA) revealed distinct clustering patterns, while Partial Least Squares Discriminant Analysis (PLS-DA) achieved a 17 % error rate in external predictions. Spectral markers at A6032/4457 $cm^{-1}$, A7004/5241 $cm^{-1}$, and A7004/4749 $cm^{-1}$ reflected biomolecular differences among varieties. This innovative approach enables precise quantification, classification, and differentiation of rice types, enhancing quality control, improving consumer satisfaction, and optimizing breeding selection processes efficiently.

## 1. Introduction

Rice (*Oryza sativa* L.) is unique among major cereals mostly consumed as a whole grain after cooking. Evaluating rice quality is essential to establish high standards and meet consumer expectations. This evaluation considers multiple parameters, including appearance, texture, aroma, taste, nutritional composition, and safety. These factors collectively determine the overall quality and market value of rice products. Rice quality is commonly determined through visual inspections and manual measurements being time-consuming, subjective, and prone to human error which is considered an important factor affecting buying decisions (Butardo & Sreenivasulu, 2019; Cuevas et al., 2016; Keith et al., 2007). Traditional approaches to rice variety evaluation focus on chemical composition (protein, moisture, fat, and ash), as well as apparent amylose content, gelatinization temperature, gel consistency, and pasting viscosity, are grounded in well-established international standard methods, although they are often expensive and time-consuming (Kong et al., 2015). Additionally, consumer perception of rice quality is influenced by traits such as grain length, uniformity of size and shape, colour, chalkiness, and the percentage of broken rice (Aznan et al., 2021; Custodio et al., 2019). Rice quality assessment also encompasses grain physical attributes, milling performance, biochemical composition, cooking properties, and basic chemical composition such as protein, moisture, fat, ash, and amylose content (Bhattacharya, 2011). Beyond sensory evaluation conducted by human panels, the textural properties of cooked rice are frequently analysed using texture profile analysis (TPA) with a textural analyser (Cameron & Wang, 2005; Li et al., 2016). TPA is a technique that has been extensively employed to mechanically and geometrically characterize food materials, involves measuring the mechanical response during a double compression, which attempts to mimic the first and second bites of food. For cooked rice, the two most meaningful parameters derived from TPA are hardness (the force required to attain a given deformation) and adhesiveness (a quantity that simulates the work required to overcome the attractive forces between the surface of the sample and the surface of the probe with which the same comes into contact) (Friedman et al., 1963). These factors can significantly impact buying decisions and consumer satisfaction. To address these challenges, advancements in technology and automation have led to the development of innovative approaches for rice quality evaluation, offering greater accuracy, efficiency, and

* Corresponding author at: Faculty of Engineering, Lusófona University, Campo Grande, 376, 1749-024 Lisbon, Portugal.
*E-mail address:* pedro.sampaio@ulusofona.pt (P.S. Sampaio).

https://doi.org/10.1016/j.foodchem.2025.145569
Received 2 March 2025; Received in revised form 13 July 2025; Accepted 14 July 2025
Available online 15 July 2025

consistency. Some of these technological solutions included digital imaging systems that provide objective measurements and eliminate the subjectivity associated with manual inspections.

Near-infrared spectroscopy (NIRS) is a fast and non-destructive technique that evaluates the chemical composition of rice grains based on their molecular interactions with light. This technology can rapidly determine parameters such as moisture content, amylose content, protein content, and starch properties, as well as various physiochemical properties such as whiteness, milling degree, and cooking quality, providing valuable insights into rice quality (Nicolai et al., 2007; Zhang et al., 2011). The NIRS spectral information may be analysed slightly using chemometrics methods such as the principal components analysis (PCA), which is the bilinear modelling method that gives an overview of the main information in a single multidimensional data (Jolliffe, 2002). Advanced data analytics techniques, including machine learning algorithms, are increasingly being used to analyse large datasets of rice quality attributes. These methods enable the evaluation of extensive datasets generated by various analytical techniques, providing deeper insights and greater precision. Their growing acceptance food science and technology fields highlights their potential to revolutionize data-driven research and applications (Ma, 2017). Classification techniques, also referred to as supervised pattern recognition, are used to analyse qualitative responses. These methods establish a mathematical relationship between descriptive variables (e.g., chemical measurements) and a qualitative variable, such as membership in a specific category. The classification identifies objects by grouping them into one of the finite sets of classes, which involves comparing measured features of a new object with those of a known object or other known criteria and determining if the new object belongs to a particular category of objects. A wide variety of approaches of pattern recognition techniques have been taken toward this task in the food quality evaluation such as Partial least squares discriminant analysis (PLS-DA), the k-Nearest Neighbours (kNN), Support vector machine (SVM), Artificial neural networks (ANN) (Dębska & Guzowska-Świder, 2011). PCA and PLS-DA are examples of machine learning tools applied to conduct a specific analysis of bioprocess as an exploratory technique (Scott et al., 2006). PLS-DA is gaining increasing attention as a useful feature selector and classifier (Brereton, 2009). Multivariate classification methods aimed at finding mathematical models able to recognize the membership of each sample to its appropriate class, by a set of measurements.

The current study aims to discriminate several rice types toward their biochemical, physicochemical properties, and cooking parameters such as the (*i*) appearance; (*ii*) chemical composition; (*iii*) water absorption; (*iv*) pasting parameters; and (*v*) texture. PLS models, based on NIR spectra, were developed for different parameters such as grain appearance, pasting, and TPA parameters, as well as the physicochemical, cooking, and texture characteristics. Based on NIR spectral data and supported by statistical analysis, spectral markers were identified, highlighting specific spectral regions and absorbance peaks associated with molecular differences among rice varieties. A classification method utilizing supervised techniques, specifically PLS-DA, was developed to enhance the accurate characterization and classification of rice.

## 2. Materials and methods

### 2.1. Rice samples

Twenty-two rice varieties were selected, being 20 cultivated in the Mediterranean Region (including Indica and Japonica subspecies), from Europe and Egypt (Giza 177, Giza 181), and the last two (Basmati varieties) were imported from outside of Europe. Based on commercial type, the rice varieties were classified: Long A (LA), Long B (LB), Medium Grain (MG), Short Grain (SG), European Aromatic (EA), and Basmati.

### 2.2. Milling yields and grain appearance

The rice samples were dehusked using a Satake mill (THU, Satake, Taito, Japan) and, consequently, polished (Suzuki MT98, Santa Cruz do Rio Pardo, São Paulo, Brazil) for assessing milling yields and obtain milled rice. About 20 g of rice was ground to flour on a Cyclone Sample Mill (Falling number 3100, Perten, Sweden), equipped with a 0.8 mm screen. Biometric parameters of polished rice grains, such as length, width, total whiteness, and vitreous whiteness, were evaluated in 50 g samples by image processing (S21 model and software, Suzuki, Brazil). The potential yields of husked, milled, and head rice were determined according to *ISO 6646:2011Rice — Determination of the potential milling yield from paddy and from husked rice*, 2025. The whiteness degree (KETT) was measured using a Milling Meter (Satake, Japan).

### 2.3. Pasting parameters

The pasting properties of rice were assessed using a viscosity analyser (RVA-4, Newport Scientific, Warriewood, Australia). Peak viscosity and setback were determined according to the AACC International Approved Method 61–02.01.

### 2.4. Cooking parameters

Each rice variety (12 g) was mixed with 120 g of distilled water and cooked for 20 min. Cooked rice was analysed according to the method described by (Ferreira et al., 2017) with some modifications concerning the amount of sample and using an electric pan, to obtain the water uptake ratio (WUp) and volumetric expansion ratio (VER). The cooking water was analysed for solids leached (SL), calculated according to (Altheide et al., 2012). The WUp was evaluated immediately after cooking, the rice sample was drained for 5 min and weighed. The WUp was calculated based on weight (g) gained during cooking (cooked rice weight) using the equation [Eq. 1]:

$$WUp\ (\%) = \frac{Cooked\ Rice\ (g) - Uncooked\ rice\ (g)}{Uncooked\ rice\ (g)} \times 100 \quad (1)$$

The volumetric expansion ratio (VER) was determined through ratio of the volume of cooked rice to the volume of uncooked rice (raw rice), according to the equation (Eq. 2):

$$\mathrm{VER} = \frac{Cooked\ rice\ volume\ (ml)}{Uncooked\ rice\ volume\ (ml)} \quad (2)$$

Solids leached were evaluated by drying an aliquot (50 ml) of cooking water, at a temperature of 102 °C, which was evaporated, for 24 h, in a glass container. SL was determined by calculating the difference between the weight of the glass container with the dried aliquot ($W_1$) and the weight of the empty glass container ($W_2$), using the equation (Eq. 3):

$$SL\ (\%) = \frac{W_1(g) - W_2(g)}{Uncooked\ rice\ (g)} \times 100 \quad (3)$$

### 2.5. Texture profile analysis (TPA)

The texture profile was evaluated using rice (5 g), cooked in excess water, being analysed through a Texture Analyser TA-XT2 (Stable Micro Systems, London, UK). A two-cycle compression, the force-versus-time program was used with a test speed of 2 mm/s and a rate of 80 % strain using a cylinder plunger with a 20 mm diameter. Several parameters from each test were taken: cohesiveness, chewiness, adhesiveness, hardness, gumminess, springiness, mean extrusion (Ferreira et al., 2017).

### 2.6. Physicochemical parameters

Starch, protein, fat, fibre, and ash content were assessed by NIR analyser (MPA, Bruker), using the cereals B-FING package calibration (Bruker Company, Massachusetts, USA). Amylose content was quantified using a standard curve developed from absorbance values of 4 calibrated samples from standard rice varieties (IR8, IR24, IR64, and IR65) obtained from the International Rice Research Institute. The amylose content was determined using a colorimetric technique with a spectrophotometer (Hitachi, Tokyo, Japan), at 720 nm, according to the *ISO 6647-2:2020Rice — Determination of amylose content. Part 2: Spectrophotometric routine method without defatting procedure and with calibration from rice standards,* 2025 method (Rice—Determination of Amylose Content, 2020).

### 2.7. Band ratios evaluation

The band ratios, represented by (A - absorbance), were evaluated based on the specific spectral peaks for a specific wavenumber. The statistical evaluation was performed using Student's *t*-test, at significance ($p = 0.05$) for each experimental condition (*EXCEL Microsoft* software).

### 2.8. Instrumentation and measurements

#### 2.8.1. Spectral acquisition

A total of 66 spectral observations were collected from 22 rice varieties, representing diverse geographic origins. A total of 66 spectral observations from 22 diverse rice varieties were selected to ensure a representative sampling of genetic and geographic variability, which is critical for developing robust and generalizable classification models. Each variety was measured in triplicate, resulting in three spectra per sample. This replication helps account for instrumental and sample preparation variability, improving measurement reliability and statistical confidence. The rice samples used in this study were constituted by different commercial types: Long A (11 varieties, 33 spectra); Long B (3 samples; 9 spectra); Short grain (3 samples, 9 spectra); Medium grain (2 samples, 6 spectra); Basmati (2 varieties; 6 spectra); and European Aromatic (1 sample; 3 spectra). The samples containing approximately 25 $cm^3$ of rice grain or flour were loaded in a circular sample cup and pressed slightly to obtain a similar packing density. In this study, rice samples were analysed in both grain and flour forms, depending on the specific conditions and parameters evaluated. Table 1 indicates the type of material used for each model.

Sample spectra were collected using NIR transflection MPA equipment (Bruker Optics, Germany). For each rice sample, 16 successive scans were performed, over a wavenumber range (12,000–4000 $cm^{-1}$), at 16 $cm^{-1}$ of resolution. For each rice sample, three spectra were registered.

#### 2.8.2. Spectral data pre-treatment

To enhance the spectral features and reduce systematic noise, such as baseline variation, light scattering, and path length differences, a mathematical pre-processing of the original spectra was needed. The spectral pre-treatment was optimized in a previous study (Borraz-Martínez et al., 2019) and consisted of a combination of standard normal variate (SNV) with Savitzky–Golay (SG) first (1st) derivative. SNV is a normalization procedure for spectral light scattering correction, being used to correct additive and multiplicative effects in spectra caused by particle size variation. SNV determines the standard deviation of all the variables in a given sample spectrum. SG first derivative was applied to remove the baseline drift and enhance small spectral differences. The SG

**Table 1**
Analysis of various physicochemical parameters of different rice types based on NIR spectra, utilizing PLS models with recorded spectral regions and applied spectral preprocessing methods.

| Parameter | $R^2_{cal}$ | RMSEC | Slope | LV | $R^2_{val}$ | RMSECV | Spectral region ($cm^{-1}$) | Spectral pre-processing method |
|---|---|---|---|---|---|---|---|---|
| Whiteness degree KETT (Grain) (%) | 0.97 | 1.564 | 0.97 | 8 | 0.94 | 1.730 | 9827–8038<br>6263–4481 | Straight line subtraction |
| Width average (Grain) (mm) | 0.96 | 0.109 | 0.96 | 8 | 0.94 | 0.111 | 10,715–9820<br>8933–7151<br>6263–4481 | Multiplicative scattering correction |
| Total whiteness (Grain) (%) | 0.98 | 2.376 | 0.98 | 9 | 0.95 | 2.490 | 10,715–8925<br>6263–3594 | Multiplicative scattering correction plus 1st Derivative |
| Vitreous whiteness (Grain) (%) | 0.93 | 1.14 | 0.93 | 8 | 0.91 | 1.18 | 9403–4597 | Vector normalization |
| Setback (Grain) (cP) | 0.79 | 357 | 0.79 | 4 | 0.77 | 365 | 9403–8447<br>4427–4242 | Vector normalization plus 1st Derivative |
| Gelatinization temperature (°C) | 0.99 | 0.51 | 0.99 | 10 | 0.99 | 0.52 | 9827–6256<br>5376–4481 | No spectral data processing |
| Peak Viscosity (cP) | 0.96 | 179 | 0.96 | 10 | 0.94 | 183 | 9403–4242 | Vector normalization plus 1st Derivative |
| Hardness (%) | 0.92 | 219 | 0.92 | 8 | 0.89 | 223 | 8933–4481<br>10,715–9820 | Mulpitplicative scattering correction |
| Adhesiveness (%) | 0.84 | 32.4 | 0.84 | 8 | 0.81 | 33.6 | 7158–6256<br>5376–4481<br>9403–7498 | Constant offset elimination |
| Resilience (%) | 0.97 | 0.532 | 0.97 | 7 | 0.96 | 0.545 | 5777–5446<br>4605–4420 | 2nd Derivative |
| Springiness (%) | 0.98 | 1.64 | 0.98 | 7 | 0.98 | 1.65 | 9827–8038 | Straight line subtraction |
| Gumminess (%) | 0.93 | 75.4 | 0.93 | 8 | 0.91 | 76.5 | 9403–4242 | Multiplicative scattering correction |
| Chewinness (%) | 0.84 | 91.4 | 0.84 | 7 | 0.81 | 92.2 | 10,715–9820<br>7158–4481<br>9404–7498 | Constant offset elimination |
| Cohesiviness (%) | 0.91 | 0.011 | 0.91 | 5 | 0.89 | 0.013 | 6102–5446<br>4605–4420 | 2nd Derivative |
| Solid leach (%) | 0.98 | 0.087 | 0.98 | 9 | 0.99 | 0.091 | 12,498 - 11,594<br>8933–8038 | Min-max normalization |
| Water up ratio (%) | 0.90 | 0.097 | 0.90 | 9 | 0.87 | 0.099 | 8933–7151<br>6264–4482 | Straight line subtraction |

RMSECV – Root mean square error cross-validation; RMSEC – Root mean square error of calibration; $R^2_{cal}$ – Determination coefficient of calibration; $R^2_{val}$ - Determination coefficient of validation; LV – Latent variables.

algorithm requires the selection of the order of the polynomial, the order of the derivative, and the filter width, which corresponds to the size of the window (Borraz-Martínez et al., 2019; Savitzky & Golay, 1964).

### 2.9. Machine learning tools

#### 2.9.1. Partial least squares (PLS) regression

After the outliers' identification, the PLS regression was performed. The matrices containing the NIR spectra, represented by *X*, and the vector *Y,* containing the response parameter, were employed to build the regression model. The performance of the PLS model was evaluated according to the root mean square error prediction (RMSECV) and the determination coefficient ($R^2$), where *n* is the number of samples in the test set validation, $y_i$ corresponds to the reference measurement for the test set of sample *i,* and $\widehat{yi}$ represents the estimated values for test sample *i* which is one of the most used statistical parameters for the assessment of the developed model, defined by the following [Eq. 4,5]:

$$RMSECV = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{n-1}} \tag{4}$$

$$R^2 = \frac{\sum_{i=1}^{n}(\widehat{yi} - yi)^2}{\sum_{i=1}^{n}(yi - \bar{yi})^2} \tag{5}$$

The variables tested were defined by whiteness degree (KETT), width average, total whiteness, vitreous whiteness, setback, gelatinization temperature, peak viscosity, hardness, adhesiveness, resilience, springiness, gumminess, chewiness, cohesiveness, solid leach, and water up ratio. The correlation coefficient (R) between the predicted and the measured values are evaluated for the calibration and test set using the [Eq. 6], where ȳ represents the average of the reference measurement results for all samples in the calibration and test set. The statistical analysis was performed using the specific toolbox in Microsoft-*Excel* software for ANOVA processing.

$$R = \sqrt{1 - \frac{\sum_{i=1}^{n}(\hat{y}\boldsymbol{i} - \boldsymbol{yi})^2}{\sum_{i=1}^{n}(\boldsymbol{yi} - \bar{y})^2}} \tag{6}$$

#### 2.9.2. Principal component analysis

Principal component analysis (PCA) is an unsupervised technique that allows the dimensionality reduction of the multivariate data to n principal components that preserves the variance of initial data as possible in the lower dimensionality output data (Jolliffe & Cadima, 2016). The large datasets are transformed into a small number of uncorrelated variables (called Principal Components, PCs). The scores scatter plot shows the most significant variability among samples, and each PC is composed of scores and loadings. The scores represent the variance on sample direction, thus being used to identify patterns of similarity among the samples. The loadings represent the variance on wavelength direction, thus being used to identify possible spectral markers responsible for the score pattern. PCA analysis was performed using MATLAB® (2023a) software. PCA toolbox for MATLAB, version 1.3 (May 2017); Milano Chemometrics and QSAR Research Group. http://michem.disat.unimib.it/chm/

#### 2.9.3. Partial least squares-discriminant analysis (PLS-DA)

Partial Least Squares-Discriminant Analysis (PLS-DA) is a linear classification tool used to build predictive models. It is based on the partial least squares regression algorithm, which identifies latent variables that maximize covariance. These latent variables represent key sources of data variability through linear combinations of the original variables (Ballabio & Consonni, 2013; Barker & Rayens, 2003). A PLS-DA model creates a set of prediction values for different classes, with values close to one for the target class and values near zero for the other classes. An optimized threshold is then calculated to distinguish between these classes. Once validated with a test set of samples, the PLS-DA model can be used to predict the class of an unknown sample. Six rice varieties were analysed, so the Y-block contained six columns, one for each class. Class 1-Long A (LA); Class 2-Short grain (SG); Class 3-Long B (LB); Class 4-European Aromatic (EA); Class 5-Medium grain (MG); Class 6-Basmati rice type. A Venetian blinds cross-validation, with a data split of 10 and one sample per blind (thickness), was used to find the optimal number of factors (LVs) for the PLS-DA model (Spring, 2008). For classification models (PLS-DA), spectral data related to the rice samples were divided into calibration samples (43) and for the external validation procedure (23), which corresponds to 65 % and 35 % of the samples, respectively. The accuracy, sensitivity, and specificity parameters were calculated for the test set, being calculated individually per class; herein, the average for all classes is reported. The accuracy (AC) represents the total number of samples correctly classified, considering true and false negatives [Eq. 7]; the sensitivity (SENS) represents the proportion of positives that are correctly classified, and the specificity (SPEC) represents the proportion of negatives that are correctly identified [Eq. 8 and 9]. These metrics are calculated as follows: TP stands for true positives, TN for true negatives, FP for false positives, and FN for false negatives (Emsley et al., 2022). Spectral data and pre-processing procedure were carried out using the classification toolbox for MATLAB® (version 2.0), developed by Milano Chemometrics and QSAR Research Group (http://michem.disat.unimib.it/chm).

$$AC(\%) = \left(\frac{TP + TN}{TP + FP + TN + FN}\right) \times 100 \tag{7}$$

$$SENS(\%) = \left(\frac{TP}{TP + FN}\right) \times 100 \tag{8}$$

$$SPEC(\%) = \left(\frac{TN}{TN + FP}\right) \times 100 \tag{9}$$

## 3. Results and discussion

### 3.1. PCA analysis of rice types based on traits and biochemical properties

After PCA analysis, characterized by PC1 (29.99 %) and PC2 (17.81 %), six clusters were defined, highlighting the differences among the rice varieties (Fig. 1A). The loadings were defined according to the specific parameters such as biochemical and physical traits, texture profile analysis (TPA), and biometrics parameters (Fig. 1B). The largest cluster, defined fundamentally by the LA variety, characterized by its long and slim shape, being defined, in terms of the loadings, by the starch, vitreous whiteness, peak viscosity, and gumminess parameters. The clusters defined by the varieties LA (Giza 181) and the SG variety (Gageron) due to the biochemical similarity are overlap, being defined by the specific loadings such as resilience, cohesiveness, solids leached, and fibre. Grain elongation, volume expansion, and water absorption properties are essential traits for determining the quality of cooked rice (Ge et al., 2005). Long-grain rice, which contains higher amylose levels, produces drier kernels that hold their shape better, resulting in a less sticky texture compared to short-grain rice. Medium-grain rice falls between these two in terms of size and texture characteristics (Amanullah & Fahad, 2017). Physical parameters of rice grains, such as their appearance (size, shape, smoothness, and colour), weight, hardness, volume, and flow properties, are of paramount importance for the entire rice production and utilization chain, from harvesting, drying, handling, and storage to milling, packaging, marketing, cooking and product development (Bhattacharya, 2011). The cluster defined by LB variety comprises the CL-28, Maçarico, and Puntal rice types, which are placed near the Basmati, EA, and LA clusters. This cluster is characterized by the length of the grain, hardness, and gumminess properties. The LB and EA varieties are closely aligned as indicated by loadings, which

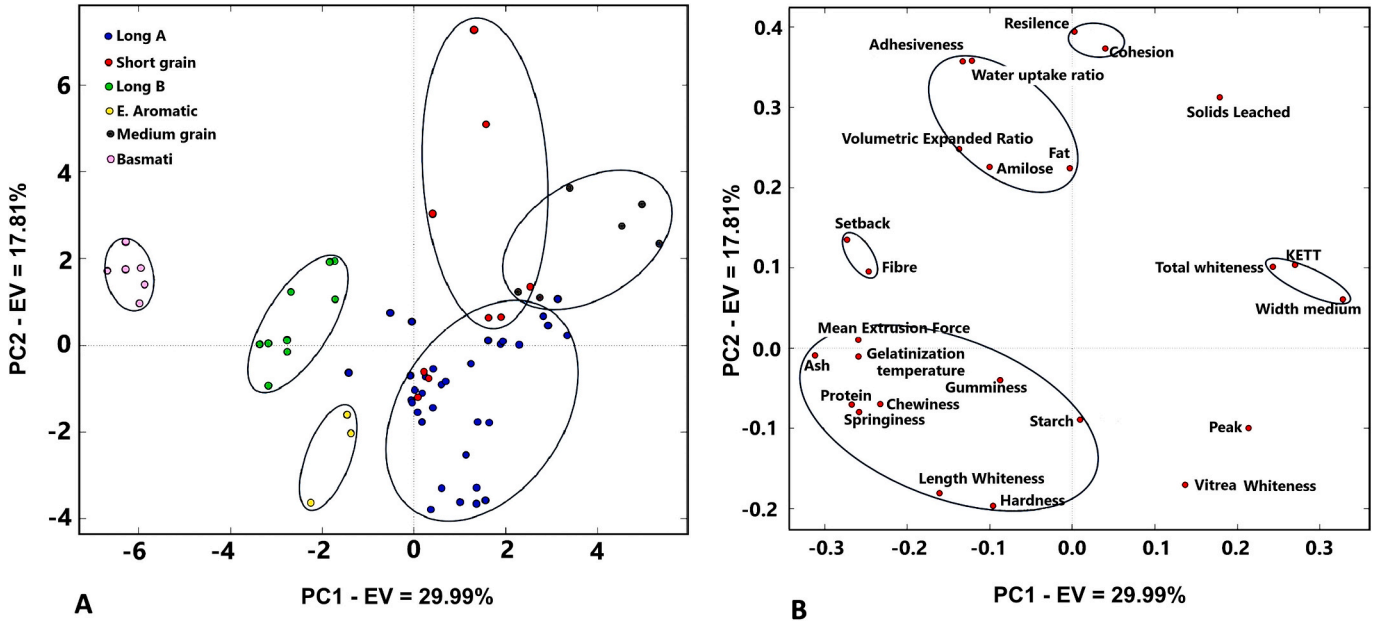



**Fig. 1.** PCA analysis of rice samples from different commercial types, explaining 29.99 % of the variance for PC1 and 17.81 % for PC2, with six distinct clusters identified based on biochemical properties, physical traits, and cooking parameters. (A); B-PC loadings highlighting traits and biochemical properties of rice types: Whiteness degree; Whiteness grain length; Grain width; Total whiteness; Vitreous whiteness; Amylose; Ash; Fat; Fibre; Protein; Starch; Gelatinization temperature; Peak viscosity; Setback; Hardness; Adhesiveness; Resilience; Cohesiveness; Springiness; Gumminess; Chewiness; Mean extrusion force; Water uptake ratio; Volumetric expanded ratio; Solids leached (B).

reveal the relationships to hardness, grain length, whiteness, gumminess, chewiness, springiness, resilience, ash, and protein levels, underscoring the similarities in their physicochemical and textural properties. Based on these results, there is a significant correlation between resilience and cohesiveness ($R = 0.95$); gumminess and hardness ($R = 0.91$); chewiness and hardness ($R = 0.75$), as well as the chewiness and the gumminess (0.74) (Fig. 1B). In terms of the physicochemical composition, such as the amylose content, significant differences were registered between the LA (19.66 %) and LB (27.46 %) types, factor that can justify the differences that exist between both rice varieties. The difference in amylose content was clear, with Indica varieties LB showing a higher concentration compared to the intermediate levels observed in Japonica varieties LA (Pereira et al., 2024). Peak of viscosity and setback properties reported a strong negative correlation ($R = -0.91$) which depends on the physical and biochemical properties of the rice grain. Between TPA parameters and amylose content low correlation was observed (Fig. 1B). This aligns with findings by Lyon et al. (2000), who reported low correlations between instrumental TPA assessments, such as hardness, adhesiveness, cohesiveness, and sensory evaluations of hardness across different rice varieties. The cluster defined by the medium-grain varieties (JSendra and Manobi types) was characterized by whiteness degree (KETT), mean width, showing a relative correlation ($R = 0.78$), as well as the total whiteness, and peak of viscosity parameter (Fig. 1B). Arborio types stand out for their significant starch content, resulting in a chewier and stickier consistency. Biometric analysis describes Arborio grains as LA variety, chubby, and starchy, with the capacity to swell and clump together, giving them a glue-like or sticky texture. In another side, starch is one of the primary characteristics of rice grain, and its physicochemical properties such as apparent amylose content, gelatinization temperature, gel consistency, and pasting viscosity are widely used as indicators of cooking and eating quality (Kong et al., 2015). Short- and medium-grain rice varieties present higher levels of amylopectin, which results in sticky, cohesive rice when cooked, often prone to breaking down and turning mushy. The aromatic rice cluster, including the Basmati variety (Super Basmati and Tipo III types) is characterized by ash, fibre, protein, gelatinization temperature, setback, springiness, gumminess, and mean extrusion force. The Basmati type presents high fibre contains and a low glycaemic index (GI) compared with other rice varieties (Pereira et al., 2024). The EA cluster (Ellectra variety) is characterized by whiteness degree (KETT), starch, vitreous whiteness, resilience, cohesiveness, solids leached, fibre, and gumminess as shown by the loadings. The Basmati (1.94 % and 9.09 %) and MG types (1.12 % and 6.99 %) were also significantly differentiated in terms of fibre and protein, while for fat, no significant differences among the different rice types were observed (Pereira et al., 2024). Significant differences were detected in terms of peak of viscosity across the various rice types. Cooking under identical conditions enables us to comprehend the characteristics of distinct grain types, including the water absorption capacity, expansion volume, and resulting grain softening. According to studies performed by Pereira et al. (2024), significant differences were observed between the Basmati (246.9 %) and LA (187.3 %) rice varieties for this cooking parameter (Pereira et al., 2024). While cooking rice, starch and other soluble solids may leach out. Hence, weight variation is not solely indicative of water absorption. The solids leached parameter can quantify the loss of these substances accurately in particular. Significant differences in SL content were observed between the Basmati (4.95 %) and MG (7.50 %) types of rice (Pereira et al., 2024). In terms of the fibre and protein, a relative positive correlation was registered ($R = 0.70$), showing also a strong correlation with cooking parameters. Notably, fibre present a relative positive correlation with the cooking parameter WUp ($R = 0.60$), while protein showed a negative correlation with the cooking parameter SL ($R = -0.60$). This means that varieties with high fibre content show higher WUp, whereas varieties with higher protein content exhibit lower SL for cooking water. Basmati varieties stand out with the highest fibre and protein contents (1.94 % and 9.09 %, respectively), while MG varieties exhibit the lowest values (1.11 % and 6.99 %) (Pereira et al., 2024). Chemical properties, such as amylose content, are key indicators of cooking quality, correlating positively or negatively with cooking characteristics. Rice grains with high amylose content typically exhibit a less crystalline structure and a lower gelatinization temperature (Bhat & Riar, 2017). Gelatinization temperature, as demonstrated by Cuevas and Fitzgerald (2012), provides a reliable indicator of cooking time. A negative correlation between amylose content and gel consistency indicates that these traits are unlikely to

respond in the same direction during selective breeding efforts.

### 3.2. Partial least squares models

Machine learning algorithms, such as partial least squares (PLS), were used to develop the prediction models based on the relationship that exists between the specific components and the spectral regions. NIR spectroscopy exploits changes in molecular vibrations caused by the absorption of infrared light to gather information about the chemical composition of a sample. PLS models for physicochemical parameters such as texture profile analysis (TPA), cooking parameters, pasting parameters, and grain physical chemicals were developed (Table 1). To reduce the influence of noise and select an appropriate analytical method, the spectral data were pre-processed using standard normal variate (SNV), multiplicative scatter correction (MSC), as well as first and second derivatives. SNV, a widely used mathematical transformation technique, was employed to eliminate the effects of solid particles, variations in light intensity, and surface scattering on the spectral data (Zhang et al., 2019). This normalization method is particularly effective when there are variations in the effective path length among samples - an issue commonly encountered when analyzing powdery materials, as in this study. Such variations can arise due to differences in particle size and sample colour (Sampaio et al., 2018).

These methods also hold significant potential for applications in screening varieties for breeding programs and the milling industry. The whiteness degree (KETT) model is characterized by $R^2 = 0.94$, root mean square error cross-validation (RMSECV) = 1.730, for a spectral region 9827–8038; 6263–4481 $cm^{-1}$. The PLS model for rice grain width average is characterized by $R^2$ = 94, RMSECV = 0.111, for three different spectral regions (10,715–9820 $cm^{-1}$; 8933–7151 $cm^{-1}$; 6263–4481 $cm^{-1}$). Meanwhile, the total whiteness model is defined by $R^2$ = 0.95; RMSECV = 2.490, for the spectral region (10,715–8925 $cm^{-1}$; 6263–3594 $cm^{-1}$), while the vitreous whiteness is characterized by $R^2 = 0.91$, RMSECV = 1.18, for a spectral region 9403–4597 $cm^{-1}$. For the setback parameter, the evaluation models based on the spectral regions (9403–8447 $cm^{-1}$; 4427–4242 $cm^{-1}$) showed an $R^2 = 0.77$ and an RMSECV = 365, after vector normalization and first derivative processing. The gelatinization temperature model, without spectral processing, was defined by $R^2$ = 0.99; RMSECV = 0.52, for the spectral regions 9827–6256 $cm^{-1}$ and 5376–4481 $cm^{-1}$. The peak viscosity model exhibited an $R^2 = 0.94$, RMSECV = 183, for the spectral region 9403–4242 $cm^{-1}$. The PLS model for hardness parameter, after multiplicative scattering correction preprocessing, is defined by $R^2$ = 0.89, RMSECV = 223, for the spectral range 8933–4481 $cm^{-1}$.

Hardness models were previously determined for rough, brown, and milled rice forms at 5952 $cm^{-1}$, while a specific PLS model for brown rice was developed based on the wavelength range 7140–4170 $cm^{-1}$ (Webb et al., 1986). Hardness is influenced by the composition of molecules such as amylose, amylopectin, compact arrangement of starch granules, protein, and lipids (Zhou et al., 2002). The adhesiveness parameter is defined as the work required to overcome the forces of attraction between the material and the probe surface, represented by the negative force. It is measured as the ability of the product to recover its original height after the first penetration, before the start of the waiting period. The PLS model for the adhesiveness is characterized by $R^2$ = 0.81, and the RMSECV = 33.6, for a spectral region defined by 10,715–9820 $cm^{-1}$, 7158–6256 $cm^{-1}$, and 5376–4481 $cm^{-1}$. The PLS model for resilience, which represents the ratio between the area after the peak force and the area before the peak force, is characterized by $R^2$ = 0.96, RMSECV = 0.545 for the specific spectral regions defined by 9403–7498 $cm^{-1}$, 5777–5446 $cm^{-1}$, 4605–4420 $cm^{-1}$ after the spectral preprocessing, using the 2nd derivative (Table 1). The PLS model for the springiness, after the spectral preprocessing (straight line subtraction), is defined by $R^2$ = 0.98, for RMSECV = 1.65, for the spectral range 9827–8038 $cm^{-1}$ (Table 1). Springiness is defined as the physical recovery of the product height between the first compression and the start of the second compression, determined by the ratio of the two compression distances. Elasticity refers to how effectively the product physically recovers its shape during this interval. It is given by the ratio between the two deformations. The gumminess, related to the energy required to chew the product, combining both its hardness and cohesiveness, the PLS model is defined by $R^2 = 0.91$, RMSECV = 76.5, using the spectral range 9403–4242 $cm^{-1}$. The chewiness is related to gumminess, which represents the energy required to chew a portion of food. The PLS model for chewiness is defined by $R^2 = 0.81$; RMSECV = 92.2, defined by the spectral ranges 10,715–9820 $cm^{-1}$ and 7158–4481 $cm^{-1}$. However, while both parameters are related to swallowing, they represent mechanical parameters derived from hardness and, therefore, were considered to evaluate the samples to verify the existing relationships (Zangirolami et al., 2023). Meanwhile, cohesion is defined as the cohesive energy, calculated similarly to adhesion, but divided by the contact area between two rice grains (Yu et al., 2019). The cohesion PLS model achieved an $R^2$ = 0.89 and RMSECV = 0.013 for the spectral regions 9404–7498 $cm^{-1}$, 6102–5446 $cm^{-1}$, and 4605–4420 $cm^{-1}$. Cohesion represents the product's resistance to a second deformation relative to its resistance to the first deformation, reflecting the strength of the internal bonds that define the food's structure. The PLS model for solid leach (SL) is defined by $R^2$ = 0.99 and RMSECV = 0.091 for the spectral regions 12,498–11,594 $cm^{-1}$, and 8933–8038 $cm^{-1}$. The SL during cooking involves the removal of substances such as minerals or undesirable flavours through soaking, altering the taste, texture, or nutritional profile of the food. In rice grains, the fine structure and molecular size of starch, along with starch surface protein influences, significantly, starch leaching and stickiness. Rice varieties with higher amylose content are prone to leaching out solids into the cooking water as starch granules during cooking (Juliano, 1971). The composition of the leached solids is mainly a mixture of starch, lipids, proteins, and other minor components. Starch represents, approximately, 90 % of the mass of the dried kernel, and it is the main component that leaches out during the cooking process. The PLS model for Wup is characterized by $R^2$ = 0.87, RMSECV = 0.099, for the spectral regions 8933–7151 $cm^{-1}$ and 6264–4482 $cm^{-1}$, after straight line subtraction processing.

### 3.3. PLS-DA classification model

NIR spectra from various rice types facilitated the classification model based on biochemical and physicochemical properties using machine learning methodologies. The spectra of rice samples were preprocessed using the multiplicative scattering correction and 1st derivative process, once the quality of spectral data represents a key characteristic for the suitable performance of the classification model (Fig. 2-A and B). PCA was applied to the NIR spectra allowing to cluster the rice types according to their physical and biochemical properties, being registered some overlaps (Fig. 3). Although being a robust exploratory technique, PCA was not able to classify samples objectively; therefore, a supervised classifier was added, and the samples were analysed by PLS-DA technique. The PLS-DA model was performed using the NIR spectra pre-processed (1st derivative), which highlights the differences. The optimal number of latent variables (LV) was selected using the criterion of lowest prediction error (highest accuracy) in cross-validation (random subsets), i.e. of optimal prediction of *y*-values for the external validation samples not used in the calibration step.

The classification accuracy and error rate associated with fitting, validation, and prediction steps for each spectra processing were evaluated (Table 2). The PLS-DA model was developed with 6 LV, explaining 56 % of the total variance, and it was also successful in discriminating all spectra in the test subset. The error rate for the fitting and validation procedure (17 %) is a very common parameter and allows evaluation of the quality of the classification model. The accuracy for the training process was significant, registering a not-assigned value for the samples used for the calibration step (23 %). The cross-validation process was characterized by an accuracy of 68 %, an error rate of 21 %, and a not-

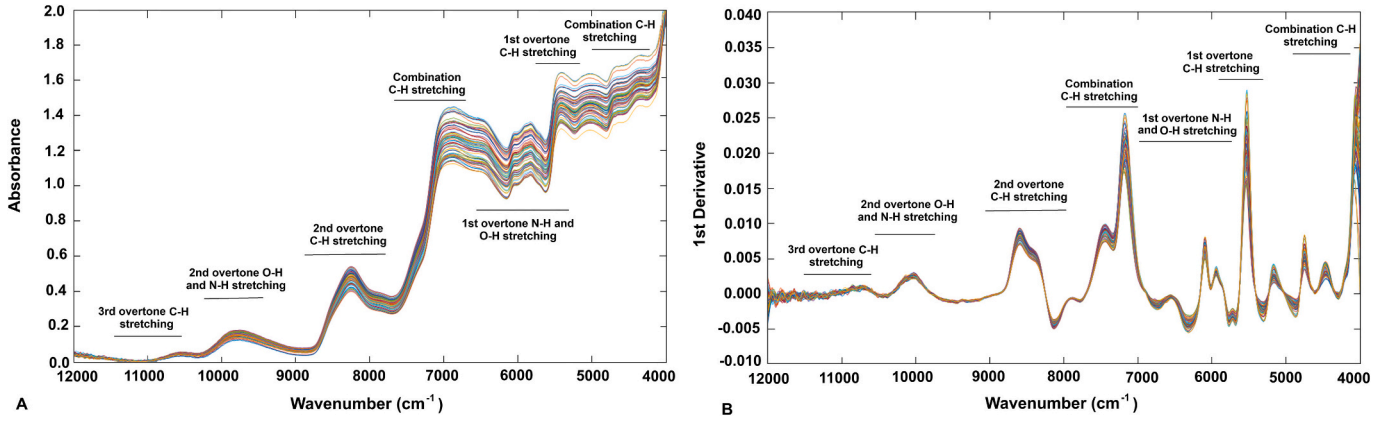



**Fig. 2.** NIR spectra of the different rice types after multiplicative scattering correction preprocessing (A); Spectral data after using the first derivative preprocessing (B).

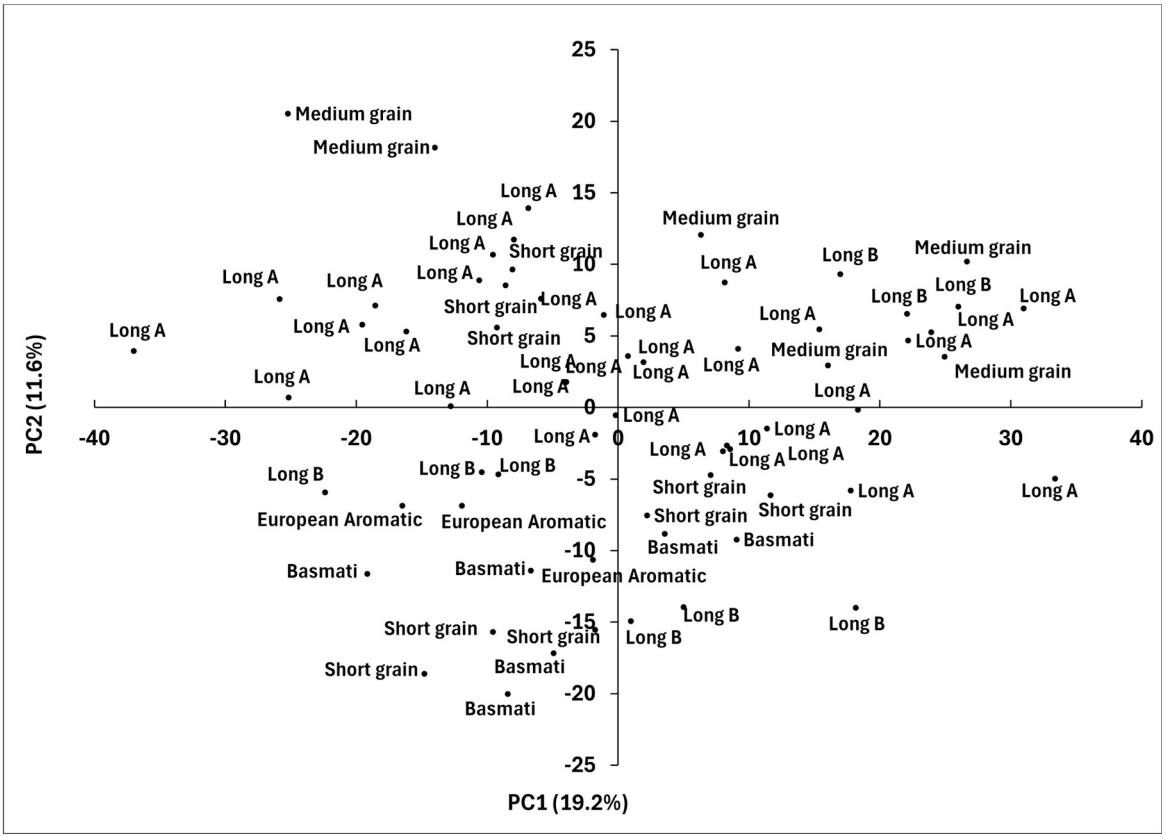



**Fig. 3.** PCA developed using spectral data related to the rice grain evaluation. The resulting clusters are categorized based on the distinct rice types (Long A; Long B; Medium grain; Short grain; European Aromatic; and Basmati type).

assigned value (28 %). The cross-validation process was characterized by an accuracy of 68 %, an error rate of 21 %, and a not-assigned value (28 %). For the prediction process, based on external samples, the PLS-DA model was characterized by an error rate of 38 %, and an accuracy of 65 %, a not-assignment for 26 %, due to the registered clusters' overlap.

The PLS-DA model demonstrates promising performance across a diverse classification range, highlighting its potential in handling complex sample sets such as different rice classes and varieties with subtle compositional differences. Despite the challenges posed by overlapping clusters, the model achieved an overall accuracy of 65 % and an error rate of 38 %, reflecting a reasonably robust predictive capability given the intricacy of the dataset. In scenarios where class boundaries are not sharply defined, such an accuracy indicates that the model was able to extract and leverage relevant discriminant features, allowing correct classification in the majority of assignable cases.

The presence of a 26 % non-assignment rate and a 38 % misclassification rate is largely attributed to spectral overlap among varieties, a common limitation in datasets with closely related classes. Nonetheless, the model's performance varied notably across individual rice classes, offering insights into its discriminative power. For the LA class, the model demonstrated good performance, with a sensitivity of 79 % and precision of 75 %, indicating it correctly identified most true LA varieties and maintained a satisfactory proportion of correct positive classifications. For the SG class, the model achieved outstanding results, with 100

**Table 2**
Classification parameters related to the calibration, validation, and testing process as well as the sensitivity, specificity, precision, and accuracy related to six rice types studied. (LA – Long A; LB – Long B; SG – Short grain; MG – Medium Grain; EA – European Aromatic; and Basmati).

| Parameter | Class | Sensitivity | Specificity | Precision |
|---|---|---|---|---|
| **CALIBRATION STEP** | LA | 100 | 100 | 100 |
| Non-Error rate: 83 % | SG | 100 | 100 | 100 |
| Accuracy: 100 % | LB | 0 | 100 | 0 |
| Not-assigned: 17 % | EA | 100 | 100 | 100 |
| | MG | 100 | 100 | 100 |
| | Basmati | 100 | 100 | 100 |
| **CROSS-VALIDATION** | LA | 79 | 58 | 75 |
| Non-error: 69 % | SG | 100 | 100 | 100 |
| Accuracy: 68 % | LB | 0 | 93 | 0 |
| Not-assigned: 28 % | EA | 100 | 97 | 67 |
| | MG | 33 | 93 | 33 |
| | Basmati | 100 | 100 | 100 |
| **PREDICTION ON EXTERNAL SAMPLES** | LA | 88 | 56 | 64 |
| Non-error rate: 62 % | SG | 33 | 100 | 100 |
| Accuracy: 65 % | LB | 0 | 100 | 0 |
| Not-assigned: 26 % | EA | 100 | 100 | 100 |
| | MG | 50 | 93 | 50 |
| | Basmati | 100 | 94 | 50 |

% sensitivity, specificity, and precision, confirming its ability to perfectly distinguish SG varieties from all others.

In contrast, the model performance for the LB class was less effective. Although specificity was high (95 %), the sensitivity and precision were considerably lower, suggesting that the model struggled to accurately detect LB variety, likely due to spectral similarities with other varieties that hinder clear differentiation.

The quality of spectral data represents a key characteristic for the suitable performance of the classification model. For the EA class, the model performed very well, achieving 100 % sensitivity and 97 % specificity, with a moderate precision of 67 %, reflecting a strong ability to detect true positives while limiting false positives. In the case of the MG class, the model exhibited high specificity (93 %) but low sensitivity and precision, indicating that while it could correctly exclude non-MG varieties, it faced challenges in positively identifying MG, likely due to overlapping characteristics with other types. The Basmati variety stood out, with the model reaching 100 % sensitivity, specificity, and precision, demonstrating its excellent capacity to differentiate this aromatic rice type with complete accuracy (Table 2).

These results collectively highlight the strengths and limitations of the PLS-DA model and support its applicability for varietal classification in rice, especially when supported by high-quality, well-resolved spectral data. During the validation step, the test samples from different clusters were placed close to their corresponding training samples, showing the high accuracy of the PLS-DA algorithm for the classification model, being a valuable feature for classification purposes. The proximity of test samples to the calibration clusters during the cross-validation process confirmed the reduced error in classification accuracy, as shown in Table 2. This spatial consistency during validation highlights the strength of PLS-DA in modelling underlying patterns, even when faced with closely related classes.

The PC loading plot represents the relative contribution of spectral region that defined the differences among several clusters according to the specific biochemical parameters (Fig. 4B). Concerning the PC loadings, the rice types were resolved according to the different spectral peaks, such as 9793 $cm^{-1}$, 9786 $cm^{-1}$, 8877 $cm^{-1}$, 7399 $cm^{-1}$, 7108 $cm^{-1}$, 5970 $cm^{-1}$, 5283 $cm^{-1}$, 5179 $cm^{-1}$, and 4451 $cm^{-1}$ (Fig. 4-B). The most intense vibration for positive values of the PC1 is located around 7142–7067 $cm^{-1}$, which is a region with vibration properties for aliphatic hydrocarbons, such as the C—H (2$\nu$ $CH_2$ and $\delta CH_2$) combination when C—H is associated with linear aliphatic $R(CH_2)NR$ or associated with branched aliphatic $RC(CH_3)_3$ or $RCH(CH_3)_2$ (Workman & Weyer, 2012). There is also a signal around 7462–7353 $cm^{-1}$, which is usually assigned to C—H from methyl ($CH_3$) structures, besides the signal in 8620 $cm^{-1}$, assigned to the carbonyl (C=O) vibration to aliphatic hydrocarbons. The strongest absorption bands observed at 5184 $cm^{-1}$ are related to the combination of stretching and bending of the O—H group of amylose, while the peak at 6835 $cm^{-1}$ is related to the combination of the first overtone of (O—H) anti-symmetric stretching and O—H symmetric stretching of amylose molecule, respectively. The fundamental vibrations for amylose are the number one carbon in the α(1–4) linked carbohydrate and normal C—H stretch. The selected NIR spectra region, characterized by the second overtone (anti-symmetric stretching, for the methyl group, ($-CH_3$) (8941–8194 $cm^{-1}$) is included, being close to the spectral region (8183 and 6850 $cm^{-1}$) mainly associated with the C—H second overtone and combination bands as described by (Bagchi et al., 2016). The spectral range (5592–5054 $cm^{-1}$) is close to the interval (5875–5495 $cm^{-1}$), as identified in studies by (Fertig et al., 2004) and (Vichasilp & Kawano, 2015), which can be associated with the vibration of amylose. The bands between 5149 and 5050 $cm^{-1}$ correspond to the O—H stretch and O—H band combination, as well as the H-O-H deformation combination, which is indicative of starch content (Aenugu et al., 2011), and N-H/C-H bending in the plane is at 4878–4830 $cm^{-1}$ (Burns & Curczak, 1992). The spectral range selected (4683–4335 $cm^{-1}$) can also be related to some starch bands (4760 $cm^{-1}$) and the protein band at 4587 $cm^{-1}$ according to (Vichasilp & Kawano, 2015).

### 3.4. NIR spectra markers

The NIR spectral ratio can be associated with pasting, cooking, biochemical, and other specific properties registered in NIR spectra, allowing to distinguish rice varieties. The clusters were defined according to the spectral regions and, consequently, the functional groups, specific to the biochemical and biophysical characteristics of rice. Based on the spectral information, five band ratios were studied (A7455/5436 $cm^{-1}$; A7004/5241 $cm^{-1}$; A8544/4749 $cm^{-1}$; A6032/4457 $cm^{-1}$, and A7004/4749 $cm^{-1}$), evaluating statistically, using the Student's *t*-test, for significance ($p = 0.05$). The NIR spectral markers should be evaluated as a reliable alternative to other biochemical methods for differentiating rice varieties, which can be considered as a suitable reference to develop a robust classification step. The A7455/5436 $cm^{-1}$ band ratios shows the differences between the six rice types studied (Table 3; Fig. 5-A). These spectral regions may be quite common, i.e. they are not specific to a particular rice variety. Similarly, the band ratios A8544/4749 $cm^{-1}$ also showed no significant differences between the varieties. This may be related to the similarities in the different spectral regions, which are quite similar between the rice varieties evaluated (Table 3; Fig. 5-C). On the other hand, the statistically validated band ratios A6032/4457 $cm^{-1}$, A7004/5241 $cm^{-1}$, and A7004/4749 $cm^{-1}$ exhibited significant differences among rice varieties (Fig. 5-D and E). The spectral regions used to evaluate the band ratios can be considered appropriate for developing models to classify rice varieties (Table 3). Based on these achievements, the LA variety presents some similarities with the SG and LB rice variety, just as the LB variety has similarities with the Basmati variety. Meanwhile, based on the statistical analysis, the EA variety is significantly different from LA, SG, and MG, which is also different from Basmati variety. The band ratios A6032/4457 $cm^{-1}$, A7004/5241 $cm^{-1}$, and A7004/4749 $cm^{-1}$ showed significant differences among the several rice varieties, based on the significance level ($p = 0.05$). According to the band ratios evaluation, the Basmati variety is different from MG, LB, and EA varieties. On the other hand, the difference between the LB varieties and the EA and MG varieties is evident. In contrast, the LA and LB varieties present a significant similarity, as well as between the LA and Basmati varieties and SG – MG. Based on these

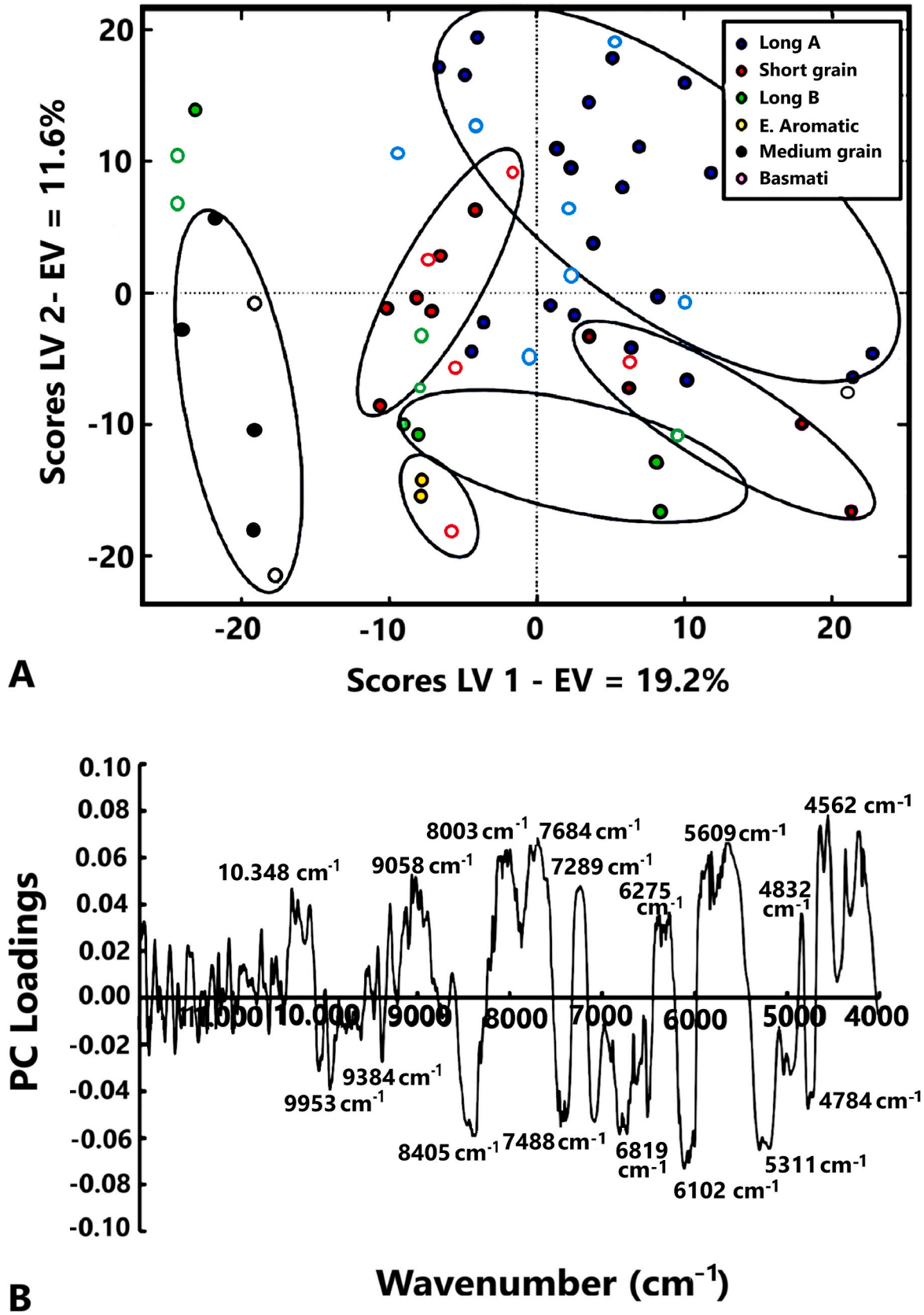



**Fig. 4.** Clusters created for rice type classification, with six clusters based on training samples and external validation samples (A); B-PC loadings related to the classification process using PLS-DA methods (B).

**Table 3**
Assessment of the classification of different rice types using NIR spectral markers selected based on their biochemical and biophysical characteristics. (LA – Long A; LB – Long B; SG – Short grain; MG – Medium Grain; EA – European Aromatic; and Basmati varieties).

| Rice Type | NIR band's spectral ratio | | | | |
|---|---|---|---|---|---|
| | A7455/5436 | A7004/5241 | A8544/4749 | A6032/4457 | A7004/4749 |
| LA/SG | 0.251 | 0.001 | 0.093 | 0.000 | 0.004 |
| LA/LB | 0.216 | 0.225 | 0.339 | 0.001 | 0.098 |
| LA/EA | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| LA/MG | 0.141 | 0,000 | 0.004 | 0.000 | 0.000 |
| LA/ Basmati | 0.002 | 0,181 | 0.020 | 0.000 | 0.415 |
| SG/LB | 0.091 | 0.000 | 0.026 | 0.000 | 0.000 |
| SG/EA | 0.003 | 0.000 | 0.000 | 0.000 | 0.000 |
| SG/MG | 0.065 | 0.331 | 0.091 | 0.009 | 0.045 |
| SG/ Basmati | 0.007 | 0.001 | 0.294 | 0.000 | 0.001 |
| LB/EA | 0.000 | 0.000 | 0.000 | 0.117 | 0.000 |
| LB/MG | 0.383 | 0.000 | 0.001 | 0.000 | 0.000 |
| LB/ Basmati | 0.003 | 0.010 | 0.003 | 0.163 | 0.025 |
| EA/MG | 0.003 | 0.000 | 0.000 | 0.000 | 0.000 |
| EA/ Basmati | 0.001 | 0.000 | 0.000 | 0.449 | 0.000 |
| MG/ Basmati | 0.000 | 0.000 | 0.030 | 0.000 | 0.000 |

findings, band ratios contain significant biochemical information and can serve as an alternative tool for the detailed study of rice types. This evaluation enables a precise quantification, classification, and differentiation of various rice varieties with enhanced sensitivity, reducing, significantly, the time and resources required for processing and evaluation. These characteristics are linked to fundamental biological data captured in NIR spectra, allowing the identification of spectral markers that facilitate rice discrimination. These potential biomarkers contribute to the identification and classification of rice varieties, providing a rapid, high-throughput alternative for breeding selection. Band ratios offer a valuable strategy for evaluating biochemical and biophysical differences among rice varieties. Derived from specific NIR spectral wavelengths, they provide detailed insights into rice type variations, enabling the identification of unique molecular markers. These markers highlight specific wavelengths and biomolecular changes, reflecting the distinctive characteristics of each rice variety.

## 4. Conclusion

The findings of this study demonstrate the successful development of a classification model for rice samples based on their distinct physicochemical properties, using machine learning techniques such as PLS-DA applied to NIR spectroscopy data. The PLS-DA classification models achieved high accuracy in differentiating rice varieties based on spectral data, underscoring the potential of integrating chemometric tools with spectroscopy to enhance grain classification and identification. These results suggest that such methods could be effectively employed to distinguish rice origin, harvest season, and storage conditions, as well as to detect contaminants or adulteration, providing significant value to both the industry and producers. NIR spectral markers, which capture the biomolecular traits of each rice variety, serve as powerful tools for precise quantification, classification, and differentiation. Their use enhances sensitivity and efficiency in rice quality evaluation while reducing time and resource consumption.

## CRediT authorship contribution statement

**Pedro Sousa Sampaio:** Writing – original draft, Supervision, Formal

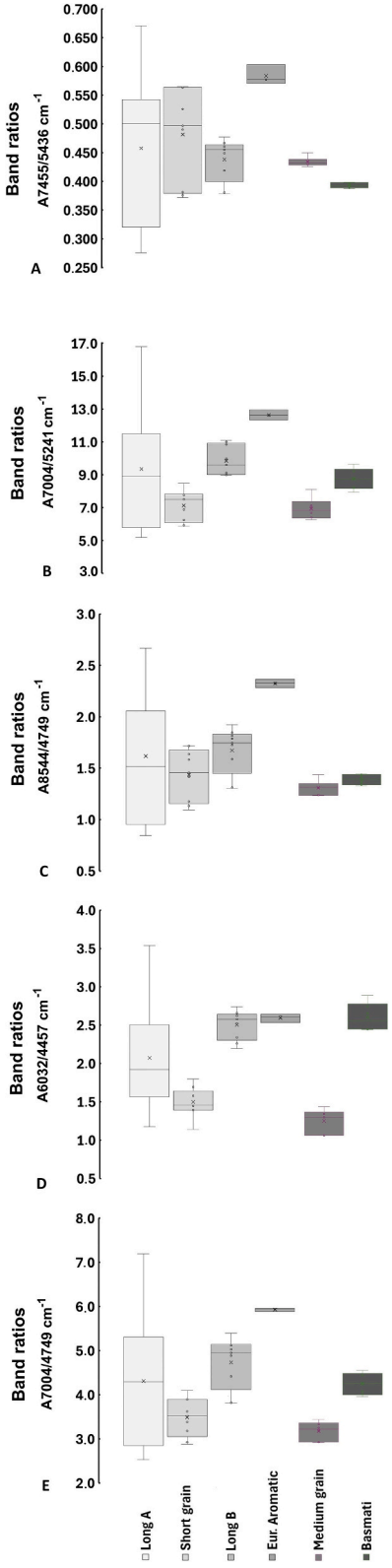



**Fig. 5.** Distribution plot of each five spectral ratios with average referencing. A box-and-whisker plot is applied to present a statistical summary of each variable across the band ratios for each rice type. The grey tons boxes represent the groups of rice varieties studied. The central box spans from the lower to the upper quartile (25th to 75th percentile), with the middle line indicating the median. The horizontal line represents the range from the minimum to the maximum non-outlier values, with outliers plotted as individual data points. Band ratios: (A) A7455/5436 $cm^{-1}$; (B) A7004/5241 $cm^{-1}$; (C) A8544/4749 $cm^{-1}$; (D) A6032/4457 $cm^{-1}$; and (E) A7004/4749 $cm^{-1}$.

analysis, Data curation, Conceptualization. **Bruna Carbas:** Writing – review & editing, Formal analysis. **Andreia Soares:** Resources. **Inês Sousa:** Formal analysis. **Carla Brites:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization.

## 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.

## Acknowledgments

Funding for this research has been received from TRACE-RICE—-Tracing rice and valorizing side streams along with Mediterranean blockchain, grant no. 1934 (call 2019, Section 1 Agrofood) of the PRIMA Program supported under Horizon 2020, the European Union's Framework Program for Research and Innovation, and Research Unit, GREEN-IT Bioresources for Sustainability Base Funding https://doi.org/10.54499/UIDB/04551/2020.

## Data availability

Data will be made available on request.

## References

Aenugu, H. P., Kumar, D. S., Srisudharson, N. P., Ghosh, S. S., & Banji, D. (2011). Near infrared spectroscopy—An overview. *International Journal of ChemTech Research, 3*(2), 825–836.

Altheide, M., Morawicki, R., & Hager, T. (2012). Impact of milling and water-to-rice ratio on cooked rice and wastewater properties. *Food Science and Technology International, 18*(3), 291–298.

Amanullah, K., & Fahad, S. (Eds.). (2017). *Rice: Technology and Production*. BoD–Books on Demand.

Aznan, A., Gonzalez Viejo, C., Pang, A., & Fuentes, S. (2021). Computer Vision and Machine Learning Analysis of Commercial Rice Grains: A Potential Digital Approach for Consumer Perception Studies. *Sensors, 21*, 6354.

Bagchi, T. B., Sharma, S., & Chattopadhyay, K. (2016). Development of NIRS models to predict protein and amylose content of brown rice and proximate compositions of rice bran. *Food Chemistry, 191*, 21–27.

Ballabio, D., & Consonni, V. (2013). Classification tools in chemistry. Part 1: linear models PLS-DA. *Analytical Methods, 5*, 3790–3798.

Barker, M., & Rayens, W. (2003). Partial least squares for discrimination. *Journal of Chemometrics, 17*, 166–173.

Bhat, F. M., & Riar, C. S. (2017). Physicochemical, cooking, and textural characteristics of grains of different rice (i *Oryza sativa* L.) cultivars of temperature region of India and their interrelationships. *Journal of Texture Studies, 48*(2), 160–170.

Bhattacharya, K. (2011). *Rice Quality: A Guide to Rice Properties and Analysis (USA; New Delhi, India, 2011. ed.)*. Oxford, UK; Cambridge, UK: Woodhead Publishing Limited.

Borraz-Martínez, S., Boqué, R., Simó, J., Mestre, M., & Gras, A. (2019). Development of a methodology to analyze leaves from *Prunus dulcis* varieties using near infrared spectroscopy. *Talanta, 204*, 320–328.

Brereton, R. G. (2009). *Chemometrics for pattern recognition*. Chichester: Wiley.

Burns, D. A., & Curczak, E. W. (1992). *Handbook of near-infrared analysis* (pp. 393–395). Practical spectroscopy series New York: Marcel Dekker Inc.

Butardo, V. M., & Sreenivasulu, N. (2019). Improving head rice yield and milling quality: State-of-the-art and future prospects. In *Rice Grain Quality: Methods and Protocols; Sreenivasulu, N.* New York, USA: Springer.

Cameron, D. K., & Wang, Y. J. (2005). A better understanding of factors that affect the hardness and stickiness of long-grain rice. *Cereal Chemistry, 82*, 113–119.

Cuevas, R. P., & Fitzgerald, M. A. (2012). Genetic diversity of rice grain quality. *Genetic diversity in plants, 10*, 35119.

Cuevas, R. P., Pede, V. O., McKinley, J., Velarde, O., & Demont, M. (2016). Rice Grain Quality and Consumer Preferences: A Case Study of Two Rural Towns in the Philippines. *PLoS One, 11*, Article e0150345.

Custodio, M. C., Cuevas, R. P., Ynion, J., Laborte, A. G., & Velasco, M. L. (2019). Rice quality: how is it defined by consumers, industry, food scientists, and geneticists? *Trends in Food Science and Technology, 92*, 122–137.

Dębska, B., & Guzowska-Świder, B. (2011). Application of artificial neuronal network in food classification. *Analytica Chimica Acta, 705*, 283–291.

Emsley, N. E., Holden, C. A., Guo, S., Bevan, R. S., Rees, C., McAinsh, M. C., & Morais, C. L. (2022). Machine Learning Approach Using a Handheld Near-Infrared (NIR) Device to Predict the Effect of Storage Conditions on Tomato Biomarkers. *ACS Food Science & Technology, 2*(1), 187–194.

Ferreira, A. R., Oliveira, J., Pathania, S., Almeida, A. S., & Brites, C. (2017). Rice quality profiling to classify germplasm in breeding programs. *Journal of Cereal Science, 76*, 17–27.

Fertig, C. C., Podczeck, F., Jee, R. D., & Smith, M. R. (2004). Feasibility study for the rapid determination of the amylose content in starch by near-infrared spectroscopy. *European Journal of Pharmaceutical Sciences, 21*(2–3), 155–159.

Friedman, H. H., Whitney, J. E., & Szczesniak, A. S. (1963). The texturometer—a new instrument for objective texture measurement. *Journal of Food Science, 28*, 390–396.

Ge, X. J., Xing, Y. Z., Xu, C. G., & He, Y. Q. (2005). QTL analysis of rice grain elongation, volume expansion and water absorption using a recombinant inbred population. *Plant Breeding, 124*, 121–126.

*ISO 6646:2011Rice — Determination of the potential milling yield from paddy and from husked rice*.(2025).

*ISO 6647-2:2020Rice — Determination of amylose content. Part 2: Spectrophotometric routine method without defatting procedure and with calibration from rice standards*. (2025).

Jolliffe, I. T. (2002). *Principal component analysis for special types of data* (pp. 338–372). New York: Springer.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences, 374*(2065), Article 20150202.

Juliano, B. O. (1971). A simplified assay for milled rice amylose. *Cereal Science Today, 16*, 334–338.

Keith, T., John, M., Joseph, G., Bernice, K., & Tamakloe, I. (2007). Study of sensory evaluation, consumer acceptability, affordability and market price of rice. *Journal of the Science of Food and Agriculture, 87*, 1564–1575.

Kong, X., Zhu, P., Sui, Z., & Bao, J. (2015). Physicochemical properties of starches from diverse rice cultivars varying in apparent amylose content and gelatinization temperature combinations. *Food Chemistry, 172*, 433–440.

Li, H., Prakash, S., Nicholson, T. M., Fitzgerald, M. A., & Gilbert, R. G. (2016). The importance of amylose and amylopectin fine structure for textural properties of cooked rice grains. *Food Chemistry, 196*, 702–711.

Lyon, B. G., Champagne, E. T., Vinyard, B. T., & Windham, W. R. (2000). Sensory and instrumental relationships of texture of cooked rice from selected cultivars and postharvest handling practices. *Cereal Chemistry, 77*(1), 64–69.

Ma, H. W. (2017). Rapid authentication of starch adulteration in ultrafine granular powder of Shanyao by near-infrared spectroscopy coupled with chemometric methods. *Food Chemistry, 215*, 108–115.

Nicolai, B. M., Beullens, K., Bobelyn, E., Peirs, A., Saeys, W., Theron, K. I., & Lammertyn, J. (2007). Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review. *Postharvest Biology and Technology, 46*(2), 99–118.

Pereira, C. L., Sousa, I., Lourenço, V. M., Sampaio, P., Gárzon, R., Rosell, C. M., & Brites, C. (2024). Relationship between Physicochemical and Cooking Quality Parameters with Estimated Glycaemic Index of Rice Varieties. *Foods, 13*(135).

Sampaio, P., Soares, A., Castanho, A., Almeida, A., Oliveira, J., & Brites, C. (2018). Optimization of rice amylose determination by NIR-spectroscopy using PLS chemometrics algorithms. *Food Chemistry, 242*, 196–204. https://doi.org/10.1016/j.foodchem.2017.09.058

Savitzky, A., & Golay, M. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry, 36*, 1627–1639.

Scott, S. M., James, D., & Ali, Z. (2006). Data analysis for electronic nose systems. *Microchimica Acta, 156*, 183–207.

Spring, M. L. (2008). (Springer, Ed.) New York: Support vector machines.

Vichasilp, C., & Kawano, S. (2015). Prediction of starch content in meatballs using near-infrared spectroscopy (NIRS). *International Food Research Journal, 22*, 1501–1506.

Webb, B. D., Pomeranz, Y., Afework, S., & Lai, F. S. (1986). Rice grain hardness and its relation to some milling, cooking and processing characteristics. *American Association Cereal Chemistry, 63*(1), 27–30.

Workman, J. J., & Weyer, L. (2012). *Practical guide and spectral atlas for interpretive near-infrared spectroscopy* ((2nd edition ed.).). CRC Press.

Yu, L., Witt, T., Rincon Bonilla, M., Turner, M. S., Fitzgerald, M., & Stokes, J. R. (2019). New insights into cooked rice quality by measuring modulus, adhesion and cohesion at the level of an individual rice grain. *Journal of Food Engineering, 240*, 21–28.

Zangirolami, M. S., Moreira, T. F., Leimann, F. V., Valderrama, P., & Março, P. H. (2023). Texture profile and short-NIR spectral vibrations relationship evaluated through Comdim: The case study for animal and vegetable proteins. *Food Control, 143*(109290).

Zhang, G., Cheng, Z., Zhang, X., & Wan, J. (2011). Double repression of soluble starch synthase genes SSIIa and SSIIIa in rice (*Oryza sativa* L.) uncovers interactive effects on the physicochemical properties of starch. *Genome, 54*(6), 448–459.

Zhang, J., Li, M., Pan, T., Yao, L., & Chen, J. (2019). Purity analysis of multi-grain rice seeds with non-destructive visible and near-infrared spectroscopy. *Computers and Electronics in Agriculture, 164*, Article 104882. https://doi.org/10.1016/j.compag.2019.104882

Zhou, Z., Robards, K., Helliwell, S., & Blanchard, C. (2002). Ageing of stored rice changes in chemical and physical attributes. *Journal of Cereal Science, 35*, 65–78.