Chapter

Near-Infrared Spectroscopy and Machine Learning: Analysis and Classification Methods of Rice

Pedro S. Sampaio and Carla M. Brites

Abstract

Nowadays, the conventional biochemical methods used to differentiate and characterize rice types, biochemical properties, authentication, and contamination issues are difficult to implement due to the high cost of reagents, time requirement and environmental issues. Actually, the success of agri-food technology is directly related to the quality of analysis of experimental data acquired by sensors or techniques such as the infrared-spectroscopy. To overcome these technical limitations, a rapid and non-destructive methodology for discrimination and classification of rice has been investigated. Near-infrared spectroscopy is considered as fast, clean, and non-destructive analytical tools and its spectra present significant biomolecular information that must be analysed by sophisticated methodologies. Machine learning plays an important role in the analysis of the spectral data being used several methods such as Partial Least Squares, Principal Component Analysis, Partial Least Squares-Discriminant Analysis, Support Vector Machine, Artificial Neuronal Network, among others which can successfully be applied for food classification and discrimination as well as in terms of authentication and contamination issues. The quality control of rice is extremely important at every stage of production, beginning with estimation of raw agricultural materials and monitoring their quality during storage, estimating food quality during the production process and of the final products as well as the determination of their authenticity and the detection of adulterants.

Keywords: Authentication, Classification, Machine Learning, Near-Infrared, Rice, Spectroscopy

1. Introduction

1.1 Rice (*Oryza sativa* L.): biochemical and physical characteristics

Rice (*Oryza sativa* L.), considered as the principal staple food for half of the world's population, is consumed from ancient times being considered one of the most important sources of dietary proteins, carbohydrates, vitamins, minerals and fiber [1]. Rice belongs to the family of cereal grasses, along with wheat, corn, millet, oats, barley, rye, and numerous others. Rice is a plant that normally grows for only one year, consisting of rounded, hollow, and articulated stalks (stems), has flat-looking leaves and a terminal panicle. Rice is considered the only cereal adapted to grow in either flooded or non-flooded soil. Rice is cultivated in different

climatic and geographic conditions and is the basis of food for a significant part of the world population. The diversity of rice grains and their quality are important factors for producers and consumers and depend on genetic characteristics and growing conditions. The grain is the seed of rice which, when the egg is fertilized, contains an embryo that has an ability to germinate and give rise to a new plant. It consists of the mature ovary, the lemma and palea (shell), the rachilla, the sterile lemmas and the wing (not always present). The embryo, present on the ventral side of the spikelet, close to the spikelet, has an embryonic root. The rest of the grain structure consists mainly of the endosperm (the edible portion), which contains starch, proteins, carbohydrates, fat, crude fiber and inorganic substance. The rough rice kernel includes the husks or hulls and pedicel, as well as the caryopsis (**Figure 1**). The weight distribution of rice caryopsis throughout the maturation phase is defined as follows: pericarp (1–2%), tegument and aleurone (5%), starchy endosperm (89–91%) and embryo (2–3%) [3]. A rice caryopsis (rice seed or whole rice grain) tends to accumulate rapidly during the developmental phase, over 5 to 15 days after fertilization under ideal conditions for development. Starch is accumulated in higher concentration in the starchy endosperm. Small amounts of starch are found in the subaleurone layer and very small amounts are present in the embryo and aleurone layer [4]. Functional proteins are present in different tissues of the embryo during development; the proteins considered storage are found accumulated in these tissues [5]. Storage proteins are found in high amounts in the starchy endosperm, however, the protein concentration is higher in the aleurone layer compared to the subaleurone layer and in the starchy endosperm [3]. Lipids, in the

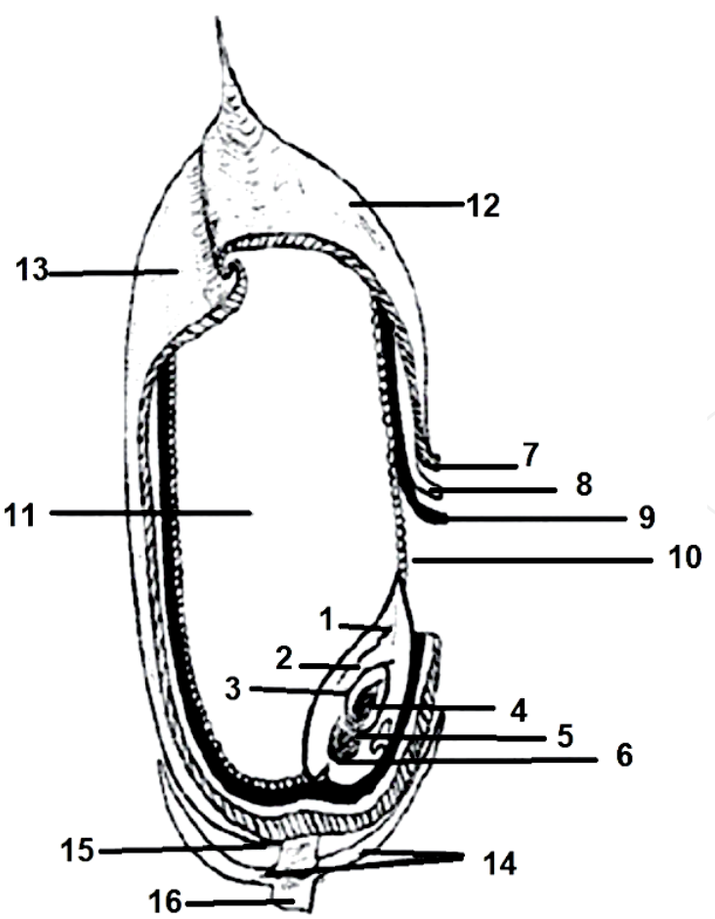

Figure 1.
Parts of rough rice grain. 1-Scutelium (Cotyledon); 2-Coleoptile; 3-Epicotyl (Plumule); 4-Apical meristem; 5-Radicle; 6-Coleorhiza; 7-Pericarp; 8-Tegmen (Seed coat); 9-Aleurone layer; 10-Subaleurone layer; 11-Starchy endosperm; 12-Lemma; 13-Palea; 14-Sterile lemmas; 15-Rachilla; 16-Part of pedicel. Adapted from: [2].

form of lipid bodies, begin to accumulate about five days after anthesis and increase in content in conjunction with starch and protein it can be accumulated for a longer period [6]. The biological activity of the pericarp and seed coat during development is important for cereals, including rice, but the synthetic activity of the seed covering the maternal tissue begins to decline before the endosperm and embryo maturity [7].

Many characteristics of grain quality, such as milling behaviour, appearance, nutritional properties, and cooking qualities, have been routinely evaluated [8]. The evaluation methods of rice varieties are based on their chemical composition, namely (protein, moisture, fat, and ash), apparent amylose concentration, gelatinization temperature, gel consistency and dough viscosity. These procedures are based on standardized methods, which are often considered to be slow and expensive [8]. The classification and characterization of different types of rice depends on several physicochemical parameters, namely, biometric data and protein, fat, ash, moisture, starch, amylose, among other.

Starch is one of main components in rice grain, being the essential carbohydrate reserve in the grain, and so its impact in the evaluated physico-chemical parameters. Starch is a complex polysaccharide of α-D-glucose units exclusively, which are joined by a sequence of α-D-(1,4)-glucosidic linkages thus giving rise to a linear or helical chain, being composed by two classes of glucose polymers: amylopectin and amylose. Amylose is a linear polymer of D-glucose units, and amylopectin is a highly branched polymer of glucose. These are referred to as amylose (20–30%). The much less frequent α-(1,6)-glucosidic linkages form the branch points between the chains thereby creating highly branched domains, denominated amylopectin (70–80%) [9]. Amylose is considered the most important determinant of the eating quality of rice and based on their contents, rice varieties can be classified as: waxy (0–2%); very low (3–12%); low (13–20%); intermediate (21–25%) and high (>26%) [10]. The classical and still commonly used method for the amylose and amylopectin determination is the iodine reaction coupled with potentiometric or amperometric titration. There are also other methods such as: differential scanning calorimetry [11], potentiometric [12], spectrophotometric [13], and chromatographic [14, 15] that can be used for classification and a detailed analysis. The fine structure of amylose, both molecular size and chain-length distribution, are also significant factors of the hardness of cooked rice [16]. Amylose content is correlated with the retrogradation behavior, influencing the textural properties of cooked rice and the viscoelasticity dynamic of rice starch gel [17]. The elongation of grains, volume expansion as well as water absorption characteristics are accounted for cooked rice quality [18].

Proteins and lipid content are also characteristics currently accepted to define rice quality [19]. After starch, the protein is the second main component of rice, being found by four fractions: albumin (soluble in water), globulin (soluble in salt), glutelin (soluble in alkali), which represents the dominant protein in brown rice and white rice, and prolamine (soluble alcohol), a secondary protein in all rice mill fractions [20, 21]. Lipids are the third major component of brown rice, next to carbohydrates and protein, playing a major role in the quality of rice during processing and storage. Fats or lipids are mainly concentrated in the outer bran layer of brown rice, up to 20% by mass; therefore, the lipids content of brown rice is greater than that of milled rice [19, 22].

Appearance quality is how the rice appears after milling and it is associated with grain length, width, length-width ratio (shape) and translucency/chalkiness of the endosperm. Generally, most markets prefer translucent rice as opposed to chalky ones. Appearance quality has a direct influence on marketability and success of commercial varieties. The physical properties of rice grain include all of its external

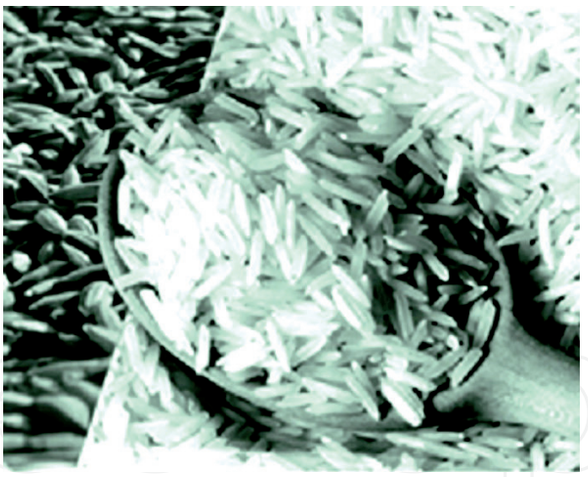

Figure 2.
Rice grains aspects.

or integral characteristics, such as its appearance (size, shape, smoothness, colour), weight, hardness, volume, flow properties and so on (**Figure 2**).

Rice classification and consequent analysis is a comprehensive quality indicator not only in terms of the appearance but also for its cooking and processing qualities. Physical properties of rice are fundamental in all activities related to the production, preservation and utilisation of rice [23]. The parameters such as dimensions, density, hardness, friction and mechanical properties are affected by the moisture content of the grain and its degree of milling, and also to a small extent by temperature. Cereal research, as well as grading and evaluation of food products, have encouraged the development of non-destructive, rapid and accurate analytical techniques to evaluate grain quality and safety being characterized by a huge amount of experimental data that must be accurately analysed [24]. Different types of rice vary in terms of size, shape, color and constitution, which cannot be accurately identified by human visualization. Often, rice seed cultivars, characterized by high quality, can be faked using low quality cultivars or confused with other cultivars, which complicates rice quality, yield and value. For this reason, the identification of rice seed cultivars is extremely important.

Grain appearance is characterized by biometric parameters (length, width, length/width ratio), total whiteness, vitreous whiteness, and chalkiness, being considered as crucial factor that affects its market acceptability. Grain shape can be described by biometric parameters, which are closely associated with grain weight [25, 26]. The ratio of the length and the width is used internationally to describe the shape and class of the variety. Grain weight provides information about the size and density of the grain. Grains of different density mill differently, and are likely to retain moisture differently and cook differently. Uniform grain weight is important for consistent grain quality [27]. Chalkiness, an opaque white discoloration of the endosperm, reduces the value of head rice kernels and decreases the ratio of head to broken rice produced during the milling process [28]. Viscosity is a characteristic that indicates some of the cooking properties of rice, being evaluated by Rapid Visco Analysis (RVA), which mimics the process of cooking and monitors the changes to a slurry of rice flour and water, during the test. Starch viscosity curves are useful for breeding because the shape of the curve is unique to each class of rice [29]. The primary RVA parameters include peak viscosity, PV (first peak viscosity after gelatinization); trough or hot paste viscosity, HPV (paste viscosity at the end of the 95 °C holding period) and final or cool paste viscosity, CPV (paste viscosity at the end of the test) [30]. The breakdown (BD = PV – HPV); setback (SB = CPV – PV); consistency (CS = CPV – HPV); set back ratio (SBR = CPV/HPV) and stability

(ST = HPV/PV) are considered as secondary parameters, once are derived from primary ones [30–32]. Other factors include peak time (time required to reach peak viscosity), and pasting temperature (temperature of initial viscosity increase) [33].

Industrial processing parameters such as the milling yield husked, milling yield milled, and milling industrial can influence positive and negatively the acceptability of rice by the industrials, can also affect the commercial value of rice. Rice yield and milling quality determine the economic value of rice from the field to the mill and in the industrial market. The rice commercial quality depends on several parameters that are evaluated separately or are involved several time-consuming experimental procedures. The evaluation of some parameters are related to biochemical or biological properties that allow more esasily its determination or prediction. Milling quality aspects affected by temperature during rice ripening include chalkiness, immature kernels, kernel dimensions, fissuring, protein content, amylose content, and amylopectin chain length [10]. Rice milling process can be subjected to dehusking of paddy which results in brown rice, and removing the bran from the kernel by polishing the brown rice to yield white rice. The milling quality of rice determines the yield and appearance of the rice after the milling process.

1.2 Near-infrared spectroscopy

Beer's law is generally applied in analytical spectroscopy to correlate the concentrations of standard samples with corresponding analyte absorbances to develop the calibration curve that is later used to evaluate the concentration of analyte of unknown samples, typically at lambda (λ_{max}). Variation in other wavelengths/wavenumber regions is often not considered but contains significant information that may be selected to represent analyte absorption fingerprint signatures and spectral profiles for ultimate pattern recognition and/or quantification of analytes in unknown samples.

Analytical infrared spectra are focus on the absorption or reflection of the electromagnetic radiation can be divided in three regions of IR: near IR (NIR) in the 12.000–4000 cm^{-1} region, mid IR (MIR) in the 4000–400 cm^{-1} region, and far IR (FIR) beyond 400 cm^{-1} (**Figure 3**). The MIR region (4000–400 cm^{-1}) is a well-recognized and reliable method through which different compounds can be identified and quantified, being used for biological applications, which includes the so-called fingerprint regions representative for lipids, proteins, amide I/II, carbohydrates, and nucleic acids (**Figure 3**). FIR spectroscopy (400–20 cm^{-1}) provides information on the highly ordered structures such as fibrillar formation and protein dynamics [35] since it is more sensitive to the vibrations from the peptide skeletons and hydrogen bonds than MIR [36]. NIR, known also "far-visible spectroscopy" or "overtone vibrational spectroscopy", can measure the chemical composition of biological materials using the diffuse reflectance or transmittance of the sample at several wavelengths [37]. The NIR spectrum, from 12.000 to 4000 cm^{-1} lies between the visible and mid-infrared regions of the electromagnetic spectrum, is characterized by a number of absorption bands that vary in intensity due to energy absorption by specific functional groups in a sample [38].

NIR is a spectroscopic technique used to study of hydrogen bonding because it evaluates the overtones and combinations of the molecule's vibrational modes, principally those involving hydrogen. NIR spectroscopy can measure the concentration of components, characterized by different molecular composition such as protein, water, or starch [39]. The chemical bonds present in food and crop components such as fats, water, and carbohydrates are easily detected by NIR spectroscopy due to the specificity of the radiation, in terms of the groups of interest such as N-H, C-H, and O-H bonds. Due to the macromolecular complexity of the rice sample, it is normal for these bands to overlap one another.

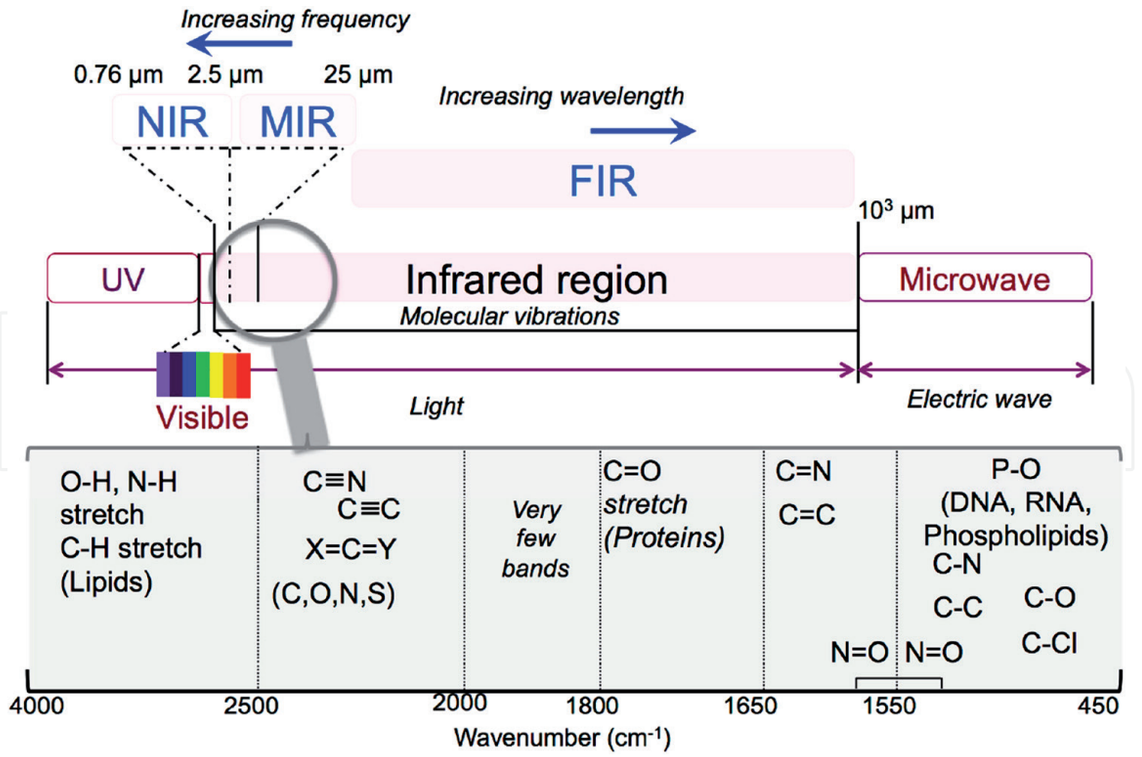

Figure 3.
Infrared spectral region (adapted by Balan et al. [34]).

The transmission and reflection are defined as the two major modes of NIR spectroscopy, that are used based on physical state of the sample. Transmission modes are more suitable for liquids, thin solids, and thick solids when inspecting a food item for its ripeness, or whether it contains pests or defects. In another side, reflectance mode is applied for measuring content in whole grains such as lipids, starch, amylose, protein, moisture, and oil content. Low reflectivity indicates that energy diffuses readily beneath the surface of most samples, including visually opaque samples. Low absorptivity represents that NIR light energy easily penetrates the samples without fast attenuation [40]. This technique is extensively used in breeding procedures for quality improvement of any cereals, and crop management, receivable testing, and on-line process control [41, 42].

The NIR methodology presents some advantages such as no sample preparation or pre-treatment process, no need for dangerous reagents or solvents, and no disposal problem, either. These advantages can eliminate sampling errors caused by manual sample handling and reagent contamination. The samples also can be used in additional studies, being carried out by technically untrained personnel. On the other hand, through NIR analysis, it is possible to obtain a set of spectra, simultaneously, in a certain range of wavelengths, which may serve as a basis for the development of specific calibration curves for each analyte. In the calibration process are transformed during modelling using, for this purpose, chemometric techniques that use a representative set of training to use the program to discriminate slight differences that exist in the specific spectra of the sample [43]. A single spectrum can be subjected to many different calibration models, to measure any number of constituents.

Different techniques such as machine vision and Visible/Near-Infrared spectroscopy have been developed and applied to determine and characterize rice varieties and evaluate the biochemical characteristics. Traditional techniques used for rice variety evaluation such as High-pressure Liquid Chromatography (HPLC) or Gas chromatography-mass spectrometry (GC-MS)

are time-consuming and hard to apply [44]. NIR spectroscopy, compared to the traditional analysis methods, is characterized by many advantages, such as is easy-to-use, real-time analysis, fast and accurate, highly reproducible results, non-destructive sampling, no sample preparation, multiple components analysis with a single measurement, high precision and non-destructive detection, being widely used in the measurement of agricultural and food products [45, 46].

1.3 Spectral pre-processing techniques

Over the years, several multivariate regression analysis methods have been developed in order to provide significant information from spectral data, due in part to the limitations of univariate spectral analysis. The processing of spectral data for chemical analysis usually uses the field of statistics and advanced mathematics for an analysis in terms of multivariate regression of spectral data. Simultaneous investigation of several wavenumbers or wavenumbers for biochemical analysis can be carried out through multivariate regression techniques, as these allow the analysis of different sample components without the need for spectral resolution and spectral deconvolutions. Pre-processing methods allowed eliminating noise caused by spectral data, which allow to remove the non-informative variability present in the spectra. Data pre-processing techniques such as normal variable transformation (SNV), multiplicative dispersion correction (MSC) and smoothing derivative are required for raw NIR spectra for proper qualitative classification and development of quantitative calibration models. MSC is used to compensate for particle size effects as it rotates the spectra to remove part of that effect, adjusting as close to the average spectrum as possible [47]. The first and second derivatives are calculated according to the Savitzky–Golay approach using a 19 point window and a 2nd or 3rd order polynomial, which allows to remove noise such as baseline drift, large, reverse and so on [48–50] (**Figure 4**).

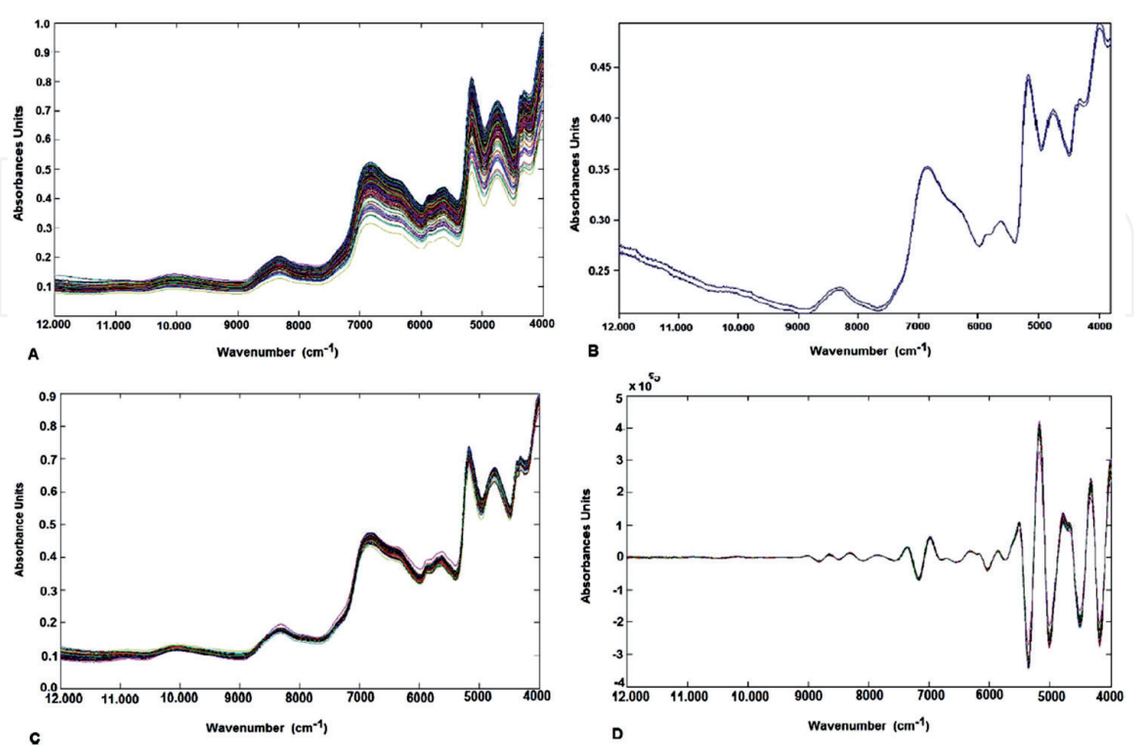

Figure 4.
Rice NIR spectra data without treatment (a); and after pre-processing procedure: baseline correction; (b, c) and first derivative process. (Adapted from Sampaio et al. [51]).

1.4 Machine learning methods

Machine learning is one of the most promising technologies in the field of artificial intelligence, that involve the use of algorithms that allow machines to learn by imitating the way humans learn step. Machine learning based on experimental data allows to optimize grouping or classification, developing models that allow to predict the behavior or properties of systems. There are two main types of machine learning: the supervised and the unsupervised process. Supervised machine learning uses algorithms that "learn" from the labeled data entered by a person without an algorithm. The algorithm generates expected output data as long as the input has been labelled and prior primary. There are two types of data that can be used in the development of the algorithm: (a) classification, which classifies an object into different classes, for example, it allows determining the type of rice according to its physical characteristics; (b) Regression, predicts a numerical value such as the concentration of any biochemical parameters such as the protein, lipids, or carbohydrates, etc. Supervised learning consists of learning a function from training examples, based on their attributes (inputs) and labels (outputs). In the unsupervised machine learning, unlike the previous case, there is no human intervention, and the algorithms learn process is based on the data with unlabeled elements, looking for patterns between them without human intervention. In this case two types of algorithms have been developed: (a) clustering, classifies the output data into groups according to its similarity; (b) association, the algorithm discovers rules within the data set. In semi-supervised learning, both labeled and unlabeled data is used for training, with usually only a small amount of labeled data, but a large amount of unlabeled data. Instead, the learning system receives some sort of a reward after each action, and the goal is to maximize the cumulative reward for the whole process. The much recognized machine learning methods are: Principal Component Analysis (PCA), the most basic feature extraction unsupervised techniques, based on the analysis of the variance of features within the full spectrum; the clustering unsupervised methods, used to identify biological subtypes within a sample, such as Hierarchical Cluster Analysis (HCA), k-Nearest Neighbors (KNN), Artificial Neural Networks (ANN), discriminant analysis (DA), Partial Least-Squares-Discriminant Analysis (PLS-DA), Partial Least-Squares (PLS), and Support Vector Machines (SVM).

1.4.1 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 [52]. The huge number of data are transformed into a reduced number of uncorrelated variables called principal components (PC) where each component represents a linear combination of the original data and the number of PCs is equal to the original variables. Early PCs explain most of the sample data, which allows for the reduction of data size. A PCA can reveal as variables that determine some inherent structure of the data, which can be interpreted in chemical or physicochemical terms. The scatter plot of PC1 and PC2 scores represent the most expressive variability among themselves, which account for most of the variability between samples and contain information from the entire spectrum. The PCA has been coupled with Mahalanobis distances to reduce dimensionality before carrying out the discriminant analysis [53]. Plots of PCs *versus* each other represents how the variables that they account for are related. To monitor the cluster together is important to determine a set of scaling coefficients, the scores. The scores for each

factor can be evaluate for every spectrum in the training set. The original spectra are constructed when the scores are multiplied by the load vectors and the results summed. In this way, knowing the set of charge vectors, how scores represent the spectra with the precision of the original responses at all wavelengths. PCA avoids the problem of overfitting by selecting too many wavelengths. This pattern recognition method was used to determine the Mahalanobis distances that are determined in units of standard deviations from the center (mean) of the training set cluster. Cross-validation is one method that is employed for evaluating the suitable number of factors. For performing this evaluation, each sample present in the calibration set is eliminated one by one and the remaining samples are used to build a Mahalanobis matrix for one, two, three factors, and so on. Then, the excluded sample is predicted, using the models developed for Mahalanobis grouping. The excluded sample is then put back to the calibration set, and a new sample is removed. The process continues until all changes have been removed from the calibration and prediction set. This represents an advantage of cross-validation compared to other methods, since the favors are not the same in relation to those used to define the model.

1.4.2 Discriminant Analysis

A Discriminant Analysis is a strategy that has been used successfully for a qualitative analysis, being called pattern recognition. This methodology aims to classify groups as groups into well-defined groups according to the similarities of a "training set" despite limited knowledge of the composition of those belonging to the group. Johnson and Wichern [54] concluded that the use of discriminant analysis uses several variables and analyzed how to solve the grouping together. The development of calibration models in discriminant analysis is based on two methods: Mahalanobis distances, considered the unit distance vector in multidimensional space, and PCA coupled with Mahalanobis distances [54, 55]. The Mahalanobis distance can be defined by an ellipsoid in a multidimensional space that circumscribes the data. This method is based on a matrix that represents the inverse of the matrix formed by combining the covariance matrices within the group of all groups, which is generated by combining information from all different materials of interest in a single matrix. Studies developed by and Williams considered the Mahalanobis distance as the mathematical number that defines the position, size and shape of the ellipsoid for all clusters [38]. According to of statistical perspective, the Mahalanobis distance considers the sample variability to be valid, while the Euclidean distance method does not consider the variability of values in all dimensions to be valid. The Mahalanobis distances look at not only variation between the responses at the same wavelengths, but also at the inter-wavelength variations. Instead of treating all values equally when calculating the distance from the mean point, it weights the differences by the range of variability in the direction of the sample point. The place of each cluster in multidimensional space is defined by the mean value of the absorbances (the group mean) at each wavelength. Dunmire and Williams indicated that the sample can be classified clearly if it falls within three times the Mahalanobis distance from the respective centroid and at least six times the Mahalanobis distance from the ellipses of other groups [38]. Meanwhile, the Mahalanobis distance represents a multidimensional distance D defined by the matrix equation as follows (Eq. (1)) [55]:

$$D^2 = (x - x')M(x - x') \tag{1}$$

where x represents a vector related to optical readings at several wavelengths which describes the position in multidimensional space corresponding to the spectrum of a given sample, x' is a vector that represents the position of a reference point in space, while M is the pooled inverse covariance matrix describing distance measures in the multidimensional space.

1.4.3 Partial Least Squares-Discriminant Analysis

Partial Least Squares-Discriminant Analysis (PLS-DA) is defined as a linear classification method that permits to estimate the predictive models based on partial least squares regression algorithm that follows for latent variables with maximum covariance, representing the significative sources of data variability with linear combinations of the original variables is considered an example of machine learning tool applied to conduct a global cellular analysis of bioprocess as an exploratory technique, gaining increasing attention as a useful feature selector and classifier [56–60]. 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. PLS-DA have shown promising results in the detection of food adulteration without identifying specific compounds [61]. PLS-DA is a discriminant classifier, being particularly suitable for handling correlated features (e.g., spectroscopic variables). The predicted value is a number, but not a dummy integer. Thus, a cut off value needs to be set to determine which class the sample belongs to. PLS-DA is computed based to full cross validation methods. More specifically, a predictor block is used to estimate (by PLS) a binary response called dummy Y (a binary response matrix encoding the class-belonging). Mathematically, the regression relation between the data matrix X and the dummy vector y for a two-class case is represented by the model represented in Eq. (2)

$$y = \hat{y} + e = X_b + e \tag{2}$$

where $\hat{y}$, b, and e represents, respectively, the vectors of predicted responses, regression coefficients, and residuals. When new samples (test set) need to be classified, their predicted responses, $ynew$, are calculated based on the measurements, X_{new}, and the regression coefficients, b, estimated on the training set, and the classification rule is then applied to assign each individual to one of the categories under study.

1.4.4 Support Vector Machine

Support Vector Machine (SVM) is a widely used supervised statistical learning algorithm, considered as a nonlinear classification technique, which works with supervised learning models that analyze data used for classification and regression analysis, producing linear boundaries between objects groups in a transformed space of the x-variables [62–64]. SVM was previously used to detect and quantify milk adulteration by mid-infrared spectrometry [64] and to identify rice seed cultivars [65]. SVM reveals advantages in dealing with small sample, non-linear and high dimensional data. The model performance depends of the selection of kernel function in SVM models, and the commonly used Radial Bias Function (RBF) is used as kernel function. The regularization parameter c, controls trade-off between the minimum training error and minimum model complexity, along with

the kernel parameter g of the kernel function. The parameter c reflects the degree of generalization, represents the width of the kernel function and reflects the degree of generalization are determined by a grid-search procedure in SVM.

1.4.5 Partial Least Squares

Partial Least Squares (PLS) regression and principal component regression (PCR) are examples of quantitative regression algorithms that are currently used for linear data, being considered as factor-based models. PLS and PCR use information from all wavelengths in the entire NIR spectrum to predict sample composition, instead of using a few selected wavelengths. PLS is similar to PCR but more sensitive in terms of variations in sample concentration. Studies performed by Wehling described that PLS and PCR, based on data reduction approaches, allowed to decrease a huge number of variables to a much smaller number of new variables that account for most of the variability in the samples [66]. The amount of a constituent in samples can then be predicted by these new variables. PLS is the most widely used supervised multivariate data analysis method that estimates and quantify components in a specific sample. Each training example is defined as a pair $(x, f(x))$, where x represents the input, and $f(x)$ is the output of the underlying unknown function. The objective of supervised learning is given a set of examples of f, return a function h that best approximates f. Osborne et al. indicated that PLS tends to generate solutions that need fewer factors than calibrations of comparable performance produced by PCR [53]. PLS is defined as a regression algorithm that uses concentration data during the decomposition process and involves information as much as possible into the first few loading vectors [67]. It performs, simultaneously, a decomposition on the spectral and concentration data. A small number of factors are developed as specific data linear and regression on the scores of the factors used to derive a prediction equation. To remove irrelevant spectral variables and to improve model performance, several methods have been studied to select the optimal variables for multivariate calibration. The multivariate calibration allowed builds a predictive model, relating variables (wavenumbers) to properties of interest (concentration data). To address this common problem, a variety of linear regression methods based on latent variables (LVs) have been developed, such as partial least squares (PLS), but due to several drawbacks such as the noise in spectral data, the calibration and prediction errors are high, and the model can be affected [68]. Regardless of the regression method, the initial stage of this process is related to a typical development, optimization and refinement. The main objective of any multivariate regression is to predict unknown the samples' with a degree of certainty and great accuracy using a process known in multivariate analysis as "validation". The established regression models must be sufficiently validated, usually with independent validation samples of known concentrations. Root-mean-square-error-of-prediction (RMSEP) and root-mean-square-percent-relative error (RMSRE) are utilized to calculate the reliability and performance of the regression model for accurate determination of analyte concentrations of validation or future samples.

The matrices containing the data provided by the NIR spectra, denominated by X and the vector Y containing the parameters that it will be determine are employed to build the regression model. The performance of the final PLS model is evaluated according to the RMSEP and the correlation coefficient (R). RMSEP was defined as:

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n}} \tag{3}$$

$$R = \sqrt{1 - \frac{\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2}} \tag{4}$$

where n represents the number of samples in test set validation, y_i is the reference measurement result for the test set sample i and $\hat{y}_i$ is the estimated result of the model for the test sample i. (Eq. (3)). Correlation coefficient (R) relatively to the predicted and the quantified value are determined for both the calibration and the test set which is determined based on the (Eq. (4)), where ȳ represents the mean of the reference measurement for all samples in the calibration and test set. The best combination of spectral regions and the pre-processing techniques were selected by picking the PLS model with a small RMSEP, a high R and a low number of latent variables covering enough data variance. The model construction was based on test set validation composed by randomly chosen samples from the entire dataset, not used for model calibration. Based on PLS models, there are some procedures that depends on specific algorithm, spectral region selection, can considerably improve the performance of the full-spectrum calibration techniques, avoiding non-modeled interferences and building a well-fitted model [69–71]. Studies then performed showed that it is fundamental to conduct a spectra region selection responsible for the property of interest to increase the prediction performance [72, 73]. These procedures can be categorized into two classes: single wavelength selection and wavelength interval selection. Different strategies have been suggested for selection of optimal set of spectral regions such an interval PLS (iPLS), synergy PLS (siPLS), and moving window PLS (mwPLS) [69, 74, 75]. The principle of iPLS involves of splitting the spectra into equal-width intervals, and developing sub-PLS models for each one. The sub-intervals with the lowest value of the root mean squared error of prediction (RMSEP) must be chosen as the best. Several methods based on iPLS were developed to optimize the combination of the selected intervals, such as synergy iPLS (siPLS) [74]. These methods present a significant advantage because it uses a graphical presentation to focus on a selection of better sub-intervals and perform comparison among the prediction execution of local models and the full-spectrum model. Instead of just testing a series of adjacent but non overlapping intervals, which would miss some more informative ones, mwPLS was proposed to overcome this drawback. This strategy develops a series in a window that moves through the complete spectra and then selects the informative intervals with low model complexity and low value of the sum of residuals. Because it considers all the possible continuous intervals, it can select all the possible informative intervals but not the optimized ones [76].

1.4.6 Soft Independent Modeling of Class Analogy

Soft Independent Modeling of Class Analogy (SIMCA) is a supervised discriminant analysis method based on PCA [77]. This methodology is a class-modeling approach, meaning that, in defining the class boundaries, the method focuses on the similarities among samples from the same category [61, 78]. For each class, a PCA model is created and consequently the residual variance of the modeled class with the residual variance of the unknown sample is compared to determine which category the sample belongs to. The number of PCs used in each class should be selected to achieve the best classification results. SIMCA results are presented in terms of "sensitivity" and "specificity", where the former specifies the percentage of samples truly belonging to the category correctly accepted by the class model, while

the latter expresses the percentage of the objects from other classes which have been correctly rejected. SIMCA starts from a principal component analysis (PCA) of only the training objects belonging to the category to be modeled, to "capture" the regular variability due to the similarities among samples of the same class [79, 80]. Once the PCA is calculated, objects are accepted or rejected by the class-model based to their reduced distance from the class space, referred as d. For a generic i^{th} sample, the d value is calculated by Eq. (5),

$$d_i = \sqrt{\left(\frac{T_i^2}{T_{0.95}^2}\right)^2 + \left(\frac{Q_i}{Q_{0.95}}\right)^2} = \sqrt{\left(T_{i,red}^2\right)^2 + Q_{i,red}^2} \quad (5)$$

where T^2 is the Mahalanobis distance of the sample from the center of the class space and Q is its orthogonal distance from the PC subspace. These values are divided by $T^2_{0.95}$ and $Q_{0.95}$, which are the 95th percentiles of the T^2 and $Q_{0.95}$ distributions, obtaining the reduced T^2 (T^2_{red}) and the reduced Q (Q_{red}), respectively [79]. Due to the normalization, T^2 and Q limit values are equal to 1; a sample will then be accepted by the class model if $d <$ sqrt(2), otherwise it is rejected.

1.4.7 k-Nearest Neighbor

k-Nearest Neighbor (k-NN) is methodology used for a classification step based on the closest training examples in the feature space. If most an unknown sample's *k*-Nearest Neighbors in training set belong to a specific class, then this unidentified sample is classified as this class. The parameter *k* affects the performance of k-NN model. The Euclidean distance is the most common algorithm used in k-NN [81].

1.4.8 Random Forest

Random Forest (RF) is a novel machine learning algorithm that presents many decision trees, and each tree is grown from a bootstrap sample of the response variable. The optimal split is chosen from a random subset of variables at each node of the tree, and then extends the tree to the maximum extent without cutting. Prediction procedure can be performed from new data by combining the outputs of all trees. RF is suitable and fast to deal with a large amount of data, showing the advantages to reduce variance and achieve comparable classification accuracy [82, 83].

1.4.9 Artificial Neural Networks

Artificial Neural Networks (ANNs) is defined a non-parametric regression models that capture any phenomena, to any degree of accuracy (depending on the adequacy of the data and the power of the predictors), without prior knowledge of the phenomena. ANNs are applied for classification and function mapping difficulties which are tolerant of some inaccuracy and have lots of training data available, but to which hard and fast rules cannot easily be applied [84]. In the ANN the input layer is linked to an output layer, either directly or through one or numerous hidden layers of interconnected neurons. The amount of hidden layers defines the depth of a ANN, and the width depends on the amount of neurons of each layer. Rapid optimization algorithms are used to iteratively develop forward and backward passes for minimization of a loss function and to learn the weights and biases of the layer. The activation functions are applied to the present values of the weights at

each layer in the forward pass. The final result of a forward pass is new predicted outputs. The backward pass computes the error derivatives among the expected outputs and the real outputs. These errors are then disseminated backwards updating the weights and calculating new error terms for each layer. Iterative repetitions of this process is designated as back-propagation [85]. A neural network is an adaptable system that learns relationships from the input and output data sets and then can predict a previously unseen data set of similar characteristics to the input set [86, 87]. Multilayer perceptron (MLP) and radial basis function (RBF) are widely used neural network architecture in literature for regression problems [88–90]. MLPs are usually used for prediction and classification using suitable training algorithms for the network weights. The MLP trained with the use of back propagation learning algorithm. **Figure 5a** represents a three-layer structure (MLP) the most basic ANN and its minimum configuration that consists of three layers of nodes (1) input layer, (2) hidden layer, and (3) output layer. The input layer accepts the data and the hidden layer processes them and finally the output layer displays the resultant outputs of the model [91, 92]. Each node, with the exception of the input, is a neuron that is based on a non-linear activation function. The MLP can be regarded as a hierarchical mathematical function planning some set of input values to output values via many simpler functions. Normally, the nodes are fully linked between layers and therefore the quantity of parameters quickly increases to huge numbers with a considerable risk of overfitting [93]. The RBF is considered the most broadly used structural design in ANN and simpler than MLP neural network (**Figure 5b**). The RBF has also an input, hidden and output layer. There are different types of radial basis functions, but the most widely used type is the Gaussian function.

1.4.10 Multiple Linear Regression

Multiple Linear Regression (MLR) is a commonly used machine learning algorithm that allows to determine a mathematical relationship among a number of random variables, analyzing how multiple independent variables are related to one dependent variable. Since each of the independent factors has been determined to predict the dependent variable, information about the multiple variables is used to develop an accurate prediction about the level of effect they have on the outcome variable. The model generates a relationship in the form of a straight line (linear) that best approximates all the individual data points. The most important advantage of MLR is it helps us to understand the relationships among variables present in the dataset. This will further help in understanding the correlation between dependent

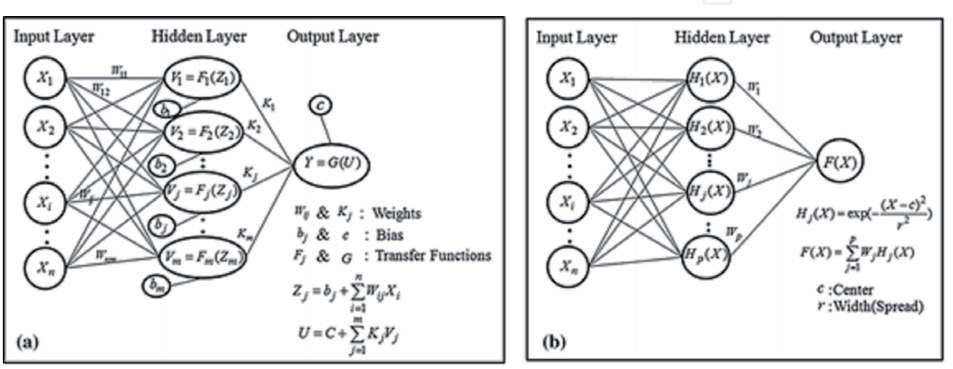

Figure 5.
A comparative study of artificial neural network (MLP, RBF) models for rice biochemical parameters prediction. Simple configuration of (a) MLP and; (b) RBF neural networks [86].

and independent variables. MLR is one of the oldest regression methods, being used to establish linear relationships between several independent variables (X_i) and the dependent variable (sample property) (Y) that depends by them. The developed model can be represented in the following the Eq. (6):

$$y_i = b_0 + \sum_{i=1}^{N} b_i x_i + e_{i,j} \tag{6}$$

where y; represents the sample property, b_i represents the computed coefficient for each variable x_i, while $e_{i,j}$ is the error. Each independent variable is analyzed and correlated with the specific property y_j. Regression coefficients b_i represent the effects of each determined term. After the MLR model has been developed the accuracy in prediction of the dependent variable is evaluated by computation of the correlation coefficient, which is calculated when true values are compared to predicted ones. Coefficient of determination R^2 is not reserved for MLR, as it is one of the most frequently used statistic parameters for assessment of validity of the developed model regardless of the model type (Eq. (7)).

$$R^2 = 1 - \frac{\Sigma\left(y_i - \hat{y}_i\right)}{\Sigma\left(y_i - \bar{y}\right)} \tag{7}$$

2. Practical applications of NIR spectroscopy and chemometrics

2.1 NIR spectroscopy in rice analysis: identification and classification

There are several studies that discribe the quantitative analysis by NIR spectroscopy in different types of food, providing an exceptional method for the evaluation of chemical composition (*i.e.* protein, starch, lipid, amylose, and moisture contents) in raw pork and beef [94], in cheese or other dairy products [95]. However, it is most widely used in the field of grains and cereal products. In some cases, such measurements are important to achieve the end-used objectives of a plant breeding program. The use of NIR spectroscopy for the quality assessment of processed foods has generated a lot of interest during the review period. Access to food with high quality is essential to human health. Thus, the accurate collection of agricultural food quality data in real-time is utmost importance, such as grains and flours. NIR spectroscopy has proven beneficial for the analysis of various cereals, grains, flours, and baked goods, including specific quality parameters, which influence classification, safety, grading, and price. By analyzing numerous factors and properties of crops during different steps in their development, crop quality can be expected early on. To maximize efficiency and lessen waste of produce, it is important that these data collection methods be non-invasive, non-destructive, and economical. Gas chromatography (GC), high-pressure liquid chromatography (HPLC), or mass spectrometry (MS) represent some quantitative instrumental techniques used for quality assessment of foods. However, these techniques are not applicable for real-time measurements. Spectroscopic instrumentation have recently utilized in agricultural industries for quality analysis. NIR spectroscopy allows a detailed food analysts to examine the quality, composition, and the authenticity of agricultural and food products quickly and accurately, based on physicochemical properties of

crops. Machine learning methodologies have been coupled with NIR spectroscopy for the prediction of rice quality factors [96] and the quantitative determination of amylose values [51]. There have been numerous applications of portable NIR instruments in recent years for specific analyses such as determining adulteration in rice and other food quality parameters. NIR spectroscopy is highly useful in analyzing shelf life and maturity of agricultural products like rice. However, the data collection and modeling are still time consuming for portable spectrometers to be efficient in some applications. This can be potentially overcome by combining NIR spectroscopy with other analytical methods. Studies developed by [97] allowed to develop a tandem approach of monitoring rice germ shelf life during storage using NIR and a portable *e*-nose. Le et al. proposed a study that combines deep learning with NIR to provide a much faster method of cereal analysis compratively to traditional NIR models [98]. The deep learning algorithm removes interference of spectral signal developing modeling significantly efficient. Jiang et al. developed a portable NIR spectrometer system to dynamically evaluate the fatty acid content of rice during storage [99]. Another challenge in NIR spectroscopy is determining authenticity and the geographical location of certain agricultural products like grains. Studies carried out by Sampaio et al. developed a strong and accurate classification model based on machine learning methods and NIR spectroscopy, allowing to sorting two genotypes of rice with high accuracy based on these characteristics [100]. Barnaby et al. correlated the grain chalk of rice to the genomic regions of NIR spectra. These spectral regions can be applied in the automation of grain chalk quantification or for other grain products as well [101].

There are several studies based on NIR to predict viscosity properties of rice. Delwiche et al. developed calibration models on whole-grain milled rice using PLS regression to predict viscosity properties of a flour-water paste as recorded by the RVA, that determine the cooking and processing characteristics of rice [102]. Meadows and Barton later used NIR to predict RVA data in rice flour [103]. A PLS regression of NIR spectra *vs*. RVA viscosity showed a highest correlation (R = 0.961–0.903) to NIR was at 212–228 sec, which is between the initial pasting time and peak viscosity. Furthermore, the pasting parameters of setback and break down, and gelatinization peak temperature of rice flour were predicted successfully using NIR [104]. Texture of cooked rice was also predicted by NIR analysis of whole grain rice [105]. Five of seven sensory texture attributes were predicted by NIR using PLS analysis, whose calibration models were developed based onf second derivative spectra. RVA peak viscosity and breakdown were also successfuly predicted based on NIR spectra and PLS regression models. Calibrations were developed using PLS and ANN analyses. The results showed limited precision of this method. However, it can be used as a rough screening method for starch amylose content. Xie et al. later reported that NIR spectra correlated strongly with differential scanning calorimetry (DSC) for measuring amylopectin retrogradation in bread staling [106]. Nowadays, requirements of quality control in grain milling and food processing increasingly call for on-line analyses [41]. Studies developed by Sampaio et al. based on NIR spectroscopy associated to PCA, PLS-DA, and SVM for discrimination and classification of rice varieties (Indica and Japonica) were explored after different spectra processing steps such as MSC, first derivative and second derivative [100]. The PCA allowed revealing the pattern and relationship of each variety and chemical similarities that were effectively distinguished by PLS-DA and SVM, according to their specific properties. The SVM model, showed a significant fitting accuracy (97%), cross-validation (93%), and prediction (91%). These data support the strength of the model for efficient rice types classification. The principal differences between both rice types were present at range 7476–7095 cm^{-1}, 7046 cm^{-1} and 4264–4153 cm^{-1}, which can be used for its discrimination, being possible to develop

a robust classification model for rice samples based on their specific physicochemical properties. The classification models developed using SVM tools were very robust compared to PLS-DA models, allowing to classify with high confidence both rice varieties. The machine learning tools can facilitate the process of classification and identification of different types of grains being possible, in the next future, to discriminate their origin, harvest season, state of conservation as well as the presence of contaminants and adulteration issues based on robust classification method, allowing to create a rice database and making *in situ*, real-time in classifying the types and origins of rice.

Studies developed by Osborne et al. using near infrared transmission spectroscopy allowed to discriminate between Basmati and other long-grain rice samples. A discriminant rule was derived using the Fisher linear discriminant function calculated from the first few principal component scores of the NIR spectra [107]. The discriminant rule was assessed by cross-validation. Based on this study, nine Basmati varieties and 53 other rice samples were classified correctly from NIR spectra, but 8% of the Basmatis and 14% of the others were misclassified on the basis of spectra of individual grains. NIR spectroscopy technique also offers effective quantitative capability for moisture, fat, protein and gluten content in rice cookies [108].

According to studies performed by Chen et al., the NIR diffuse reflectance spectroscopy of multi-grain seeds, a spectral discriminant analysis method for the variety identification of multi-grain rice seed was developed using the PLS-DA [109]. Due to the slight differences of seeds spectra in various varieties, it's necessary to propose the novel and valid methods. In this study, the SNV pretreatment combined with wavelength-screening methods improved the accuracy of the discriminant models. The selected optimal wavelength model was the combination of 54 discrete wavelengths within NIR region. NIR spectral discrimination total recognition accuracy rates reached 94.3% for a study that involves the identification of one type of differentiation (negative and excellent hybrid variety) and several interference groups (positive, four pure groups and four mixed groups).

The Hyperspectral Imaging (HSI) technique coupled with visible (vis) and/or NIR spectroscopy is generally used to identify or inspect different substances of seed by recognizing the molecular bonds in the sample, being considered the most feasible methods for rapidly and non-destructively detecting the substances of agricultural products, combining the technologies of spectroscopy and digital imaging. Studies developed by He et al. used the system NIR-HSI combined with multiple data preprocessing methods [110]. This approach allowed simultaneously to obtain spectral and spatial information from testing samples in the form of a hypercube constituted by two spatial dimensions and one spectral dimension. The HSI technique has the ability to collect hyperspectral information from samples of different sizes and shapes based on the spatial data. The detection speed of HSI is faster than that of point-based techniques, as many samples can be scanned and analyzed at the same time by using an HSI camera [111]. The classification models was developed to identify the vitality of rice seeds, presenting a great potential for identifying vitality and vigor of rice seeds. When detecting the seed vitality of the three different years, the extreme learning machine model with Savitzky–Golay preprocessing reached a significant classification accuracy of 93.67% by spectral data. In terms of the non-viable seeds identification from viable seeds of different years, the least squares support vector machine model coupled with raw data and selected wavelengths achieved a significant classification achievement (94.38% accuracy), and can be adopted as an optimal combination to identify non-viable seeds from viable seeds. In another study, carried out by Barnaby et al., NIR hyperspectral image consists of numerous bands with small spectrum gaps (every 4 nm in our study) and can assess

grain traits such as fat, starch, protein, moisture, color, and many other physico-chemical compounds at once [101]. Genome wide association study allowed to confirm known genes and to identify new genes that can affect grain quality traits based on hyperspectral imaging technique. The PLS-DA models of hyperspectral data identify spectral ranges that distinguished genetic and production environment differences, and this data can support to resolve the genetics of complex traits such as rice grain quality.

The nitrogen content is an important chemical indicator used for monitoring and management of plant due to its role in photosynthesis, productivity as well as its effect on carbon and oxygen cycle. The nitrogen content can be measured by laboratory analysis, meanwhile, its spectral reflectance of NIR (700–1075 nm) in the field was measured using hand held spectroradiometer. Studies performed by Afandia et al. evaluated nitrogen content in rice crop based on NIR reflectance using ANN [111]. The reported study allowed to conclude that the organic molecules (nitrogen, water, etc) present a specific absorption pattern in the NIR region and the comparison between measured and model estimation of nitrogen content presented a RMSE of 0.32.

A study developed by Lin et al., based on the imaging method, a system constituted by a NIR camera, filters, an automatically exchange filters device, and the imaging processing techniques allowed to detect the rice protein content based on the spectrum absorption. The NIR data allowed to establish the calibration model based on MLR, PLS, and ANN analysis models. In the MLR model, the NIR imaging system used the calibration model that take in account 5 wavelengths (880 nm, 910 nm, 920 nm, 1000 nm, and 1014 nm) to predict the rice protein content, and had R^2 validation (0.782) and standard error of predicition (SEP) 0.274%, and respectively. The NIR imaging system used 15 filters ranging from 870 to 1014 nm in the PLS model, the predictive results expressed a significant performance (R^2_{val} = 0.782, and SEP = 0.274%) comparatively tothe MLR model. The ANN model, the net input using the 5 spectrum wavelengths selected by the MLR, simplified the model, and the predicting results (R^2_{val} = 0.806, and SEP = 0.266%) were similar to those of the PLS. The prediction results indicated that the developed NIR imaging system has the advantages of simple, convenient operation, and high detection accuracy as well as it presents commercial potential in non-destructive high accurate predicting capability detection of rice protein content [112].

NIR spectroscopy was used to develop a new discrimination method of varieties of rice. The several variables compressed by PCA were used as inputs of multiple discriminant analysis (MDA). The study showed that the combinantion of spectroscopy and computer data processing technology based on PCA and MDA for the identification of rice from different areas allowed to identify correctly about 98% for the calibration process, and 100% for the prediction process. These results showed that the proposed alternative method is a feasible way for the identification of the specific production areas of rice [113].

2.2 NIR spectroscopy in rice authentication

NIR spectroscopy has been widely used in the evaluation of agricultural products due to its many advantages, such as being easy-to-use, non-destructive, fast and accurate, providing highly reproducible results, requiring minimum or, often, no sample preparation, and allowing the analysis of several constituents based on a single measurement. As consequence of the importance of rice at global level, in the literature it is possible to find several studies aimed at their analysis and characterization. Due to environmental reasons and the rice the market, non-destructive approaches are generally preferred. NIR spectroscopy has emerged as an important

tool to determine fraud, adulteration, contamination in grains and flours. A substantial instrumental improvements (e.g., hyperspectral imaging, FT-NIR) and advances in data analysis (e.g., deep learning) have allowed for the development of screening methods for detecting the presence of pests (e.g., rice weevil) across a range of stored grains [114–116].

Direct spectroscopic measurements have been widely applied for several foods and commodities, especially in the grain, cereal products, such for classification of rice [117–121]. Furthermore, in the structure of the evaluation of rice quality, NIR spectroscopy has been used for the discrimination of rice [122, 123]; varieties classificationand transgenic rice detection [124]; the physico-chemical properties quantification (such as moisture content, sound whole kernel, whiteness, translucency, color, and amylogram characteristics) [125]; cultivars classification [126], protein and amylose content prediction [127, 128]; wax rice detection [129]; and eating quality prediction [130]. Barnaby et al. correlated the grain chalk of rice to the genomic regions of NIR spectra [101]. These spectral regions can be applied in the automation of grain chalk quantification and potentially for other grain products as well [131].

Rapid and nondestructive detection of rice authenticity and quality were performed based on hand-held NIR spectrometer coupled with the appropriate chemometrics. The selection of different preprocessing methods with PCA and modeling with KNN and SVM multivariate calibration model showed that MSC + PCA plus KNN showed superiority in this study with more than 90% classification rate for all categories of rice samples studied. Based on these results, the hand-held spectrometer associated to an appropriate multivariate calibration model could be used for quick and non-destructive detection of rice quality and authenticity [132].

Food fraud remains a significant problem for food regulators, importers, merchants, law enforcement personnel, and the consumer. A key feature of food fraud is the use of a lower value ingredient to imitate an authentic product. NIR analysis technology, PLS-DA, and SVM have been used to detect whether high-quality rice was mixed with other varieties of rice. NIR spectral data analyzed using PLS-DA and a SVM algorithm, was shown to be a feasible method (5% detection limit) for the rapid identification of fraudulent rice varieties blended with authentic Wuchang rice samples [133].

Studies performed by Liu et al. showed that those techniques represent a significant support to qualitative discrimination [133]. PLS was used to establish the quantitative analysis model to support in the recognition of the degree of fraud. As consequence of the direct correlation between the results of NIR analysis and the homogeneity of the samples, four groups of samples with different physical forms (full granules, 40 mesh, 70 mesh, and 100 mesh) were prepared. Regarding qualitative analysis, the performance of the model has no obvious relationship with the physical state of the sample, the qualitative model of PLS-DA and SVM can detect the fraudulent rice with a 5% detection limit. The determination coefficient and root mean square errors of the optimal prediction result were 0.96 and 2.93, respectively. Based on this study, NIR analysis technology can be considered as a reliable and fast strategy to determine if the premium high-quality rice is adultered with inferior categories of rice.

Different preprocessing approache were used for NIR signals pretreatment. Besides considering raw data, the first derivative (Savitzky–Golay approach, 15 points window, 2nd order polynomial), second derivative (Savitzky–Golay approach, 15 points window, 3rd order polynomial), and standard normal variate (SNV) were also evaluated (**Figure 6**). NIR data were further mean-centered prior to the creation of any calibration model. The most suitable preprocessing approach, together with the optimal complexity (number of LVs or PCs to be extracted) of any classification model, were defined based on a cross-validation procedure. PLS-DA

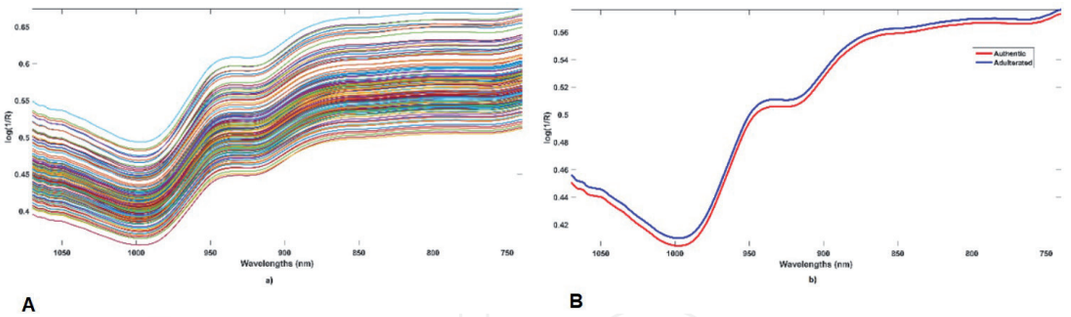

Figure 6.
NIR spectra (a) raw spectra of samples, (b) mean spectra of authentic (red line) and adulterated samples (blue line). (Adapted from [134]).

selection, specifically, was based on the combination of pre-processing and model complexity leading to the lowest mean classification error, whereas for SIMCA the maximum efficiency was sought. A study developed by Duy Le Nguyen Doan investigate the possibility of combination NIR spectroscopy and chemometric classifiers with the aim of detecting adulterated rice samples [134]. Two different strategies were exploited: discriminant classifier (PLS-DA), and class-modelling technique (SIMCA). Both strategies provided different results; in particular, SIMCA appeared unable to solve the investigated problem. On the other hand, PLS-DA analysis showed to be a suitable approach. These results indicate that the high within-class variability can have an impact on the possibility of detecting low levels of adulteration; simultaneously, was also suggested that the proposed approach could be useful for detecting samples adulterated. Then, this study demonstrates that the combination of NIR spectroscopy and PLS-DA can represent an effective, rapid and non-destructive tool for the determination of adulteration in jasmine rice [134].

2.3 NIR Spectroscopy in Rice Contamination

Fast determination of heavy metals is necessary and important to ensure the safety of crops. The potential of NIR spectroscopy coupled with chemometric technology for quantitative analysis of cadmium in rice was investigated. The spectrum was pre-processed using first derivation to reduce the baseline shift and several chemometric techniques, such as iPLS, mwPLS, siPLS, and biPLS were proposed to extract and optimize spectral interval from full-spectrum data. The PLS models based on four chemometric algorithms outperformed the full-spectrum PLS model then developed. Among the techniques, biPLS performed better with the optimal subinterval selection [135].

Heavy metals are spectrally featureless so that spectral responses could not be directly used for the assessment of heavy metals in rice. With a close combination of protein, crude fiber, and other ingredients, heavy metals present significant correlation with protein in rice [136]. The detection of heavy metal concentration in grain is mostly realized by physical and chemical direct methods that can exactly obtain the residual levels of heavy metal; however, it is time consuming, cumbersome, and inefficient. On the basis of the hypothesis that heavy metal concentration could be spectrally estimated through the correlation between heavy metal concentration and protein contents, the objectives of this study are to: (1) build quantitative model for the quick prediction of both heavy metal and protein content, and (2) to evaluate the feasibility of near-infrared spectroscopy in assessing heavy metal concentration in coarse rice.

Protecting people from heavy metal contamination is an important public-health concern and a major national environmental issue. The NIR spectral

technique is used to identify heavy metal concentration such as lead (Pb) and copper (Cu) in rice. The NIR spectral data were treated by some methods, including, logarithm, baseline correction, standard normal variate, multiple scatter correction, first derivates, and continuum removal. The lead (Pb) was accumulated in rice at a high level (17.05) compared with the others heavy metals. MSC-PLSR models were developed, respectively, for Pb (R^2 = 0.49, RMSE = 2.01 mg/kg) and Cu (R^2 = 0.29, RMSE = 0.75 mg/kg). It is achievable to identify Pb and Cu content in rice by using NIR spectral technique. However, further studies should be performed on the application of spectral technique in discriminating the other heavy metals in rice due to the limitations of few samples and particles size interference.

3. Conclusions

Based on the reported studies, it was possible to develop a robust classification, authentication or fraud detection model for rice samples considering their specific physicochemical properties and using machine learning tools such as PLS-DA, KNN, ANN, and SVM among other methodologies applied to NIR spectroscopy data, revealing the pattern and relationship of each variety and chemical similarities, according to their specific properties. The classification models developed using several models allow to classify with high confidence rice varieties using the spectral data. The results show that the use of these chemometric tools, combined with spectroscopy capabilities, can facilitate the process of classification and identification of different rice types. The rice discrimination by their origin, harvest season, state of conservation as well as the presence of contaminants and adulteration issues based on robust classification methods can facilitate the creation of a data base, a useful tool for rice authenticity that can increase the confidence and producer-consumer engagement in rice-based foods.

Acknowledgements

Acknowledge of funding: The study was supported by project TRACE-RICE -Tracing rice and valorising side streams along Mediterranean blockchain, grant n° 1934, (call 2019, Section 1 Agrofood) of the PRIMA Programme supported under Horizon 2020, the European Union's Framework Programme for Research and Innovation, and Research Unit, UIDB/04551/2020 (GREEN-IT, Bioresources for Sustainability).

Conflict of interest

The authors declare no conflict of interest.

Author details

Pedro S. Sampaio[1,2,3*] and Carla M. Brites[1,3]

1 Instituto Nacional de Investigação Agrária e Veterinária (INIAV), Quinta do Marquês, Oeiras, Portugal

2 DREAMS-Centre for Interdisciplinary Development and Research on Environment, Applied Management and Space, Faculty of Engineering, Lusófona University (ULHT), Lisbon, Portugal

3 GREEN-IT Bioresources for Sustainability, ITQB NOVA, Oeiras, Portugal

*Address all correspondence to: pedro.sampaio@ulusofona.pt

References

[1] "FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS Statistics.," 2016. [Online]. Available: http://faostat.fao.org/.

[2] Morphology of Rice Seed Development and Its Influence on Grain Quality. Paul A. Counce and Karen A. K. Moldenhauer; Nese Sreenivasulu (Ed.), Rice Grain Quality: Methods and Protocols, Methods in Molecular Biology, vol. 1892)

[3] L. Shin. B. S. Luh, " Properties of the rice caryopsis," in *Luh BS (ed) Rice.* , New York, Springer, 1991, p. 389-419.

[4] E.T. Champagne, D.F. Wood, B.O. Juliano, D.B. Bechtel, "The rice grain and its gross composition," in *In: Champagne ET (ed) Rice chemistry and technology.* , St. Paul, MN., AACC International, , 2004, p. 77-100.

[5] H. Yamagata, T. Sugimoto, K. Tanaka, Z. Kasai, "Biosynthesis of storage proteins in developing rice seeds.," *Plant Physiol,* vol. 70, no. 4, p. 1094-1100, 1982.

[6] K. Ichihara, N. Kobayashi, K. Saito, "Lipid synthesis and acyl-CoA synthetase in developing rice seeds," *Lipids,* vol. 38, no. 8, p. 881-884, 2003.

[7] N. Sreenivasulu, V. Radchuk, M. Strickert, O. Miersch, W. Weschke, U. Wobus, "Gene expression patterns reveal tissue-specific signaling networks controlling programmed cell death and ABA-regulated maturation in developing barley seeds," *Plant J,* vol. 47, no. 2, p. 310, 2006.

[8] X. Kong, P. Zhu, Z. Sui, J. Bao, "Physicochemical properties of starches from diverse rice cultivars varying in apparent amylose content and gelatinization temperature combinations," *Food Chem ,* vol. 172, pp. 433-440, 2015.

[9] M.K. Pandey, N S. Rani, M. S. Madhav, R. M. Sundaram, G. S. Varaprasad, A.K.P. Sivaranjani, A. Bohra, G. R. Kumar, A. Kumar, "Different isoforms of starch-synthesizing enzymes controlling amylose and amylopectin content in rice (*Oryza sativa* L.)," *Biotechnology Advances,* vol. 30, p. 1697-1706, 2012.

[10] B. O. Juliano, "A simplified assay for milled-rice amylose," *Cereal Science Today,* vol. 60, no. 16, pp. 334-340, 1971.

[11] D. Sievert, J.H. Holm, "Determination of Amylose by Differential Scanning Calorimetry," *Starch/Stärke,* vol. 45, pp. 136-139, 1993.

[12] W. Banks, C.T. Greenwood, D.D. Muir, "Studies on Starches of High Amylose-Content. Part 14. The Fractionation of Amylomaize Starch by Aqueous Leaching.," *Starch*, vol. 23, 1971.

[13] W.R. Morrison, B. Laignelet., " An improved colorimetric procedure for determining apparent and total amylose content in cereals and other starches.," *Journal Cereal Science,* vol. 1, pp. 9-20, 1983.

[14] N. K. Matheson, L. A. Welsh, "Estimation and fractionation of the essentially unbranched amylose and branched amylopectin component of starches with concanavalin A.," *Carbohydrate Research,* vol. 180, pp. 301-313, 1988.

[15] Y.H. Yun, H.D. Li, L.R. Wood, W. Fan, J.J. Wang, D.S. Cao, Q.S. Xu, Y.Z. Liang, "An efficient method of wavelength interval selection based on random frog for multivariate spectral calibration.," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy,* vol. 111, pp. 31-36, 2013.

[16] H. Li, S. Prakash, T.M. Nicholson, M.A. Fitzgerald, R.G. Gilbert, "The importance of amylose and amylopectin fine structure for textural properties of cooked rice grains.," *Food Chemistry,* vol. 196, pp. 702-711, 2016.

[17] Z.-H. Lu, T. Sasaki, Y.-Y. Li, T. Yoshihashi, L-T. Li, K. Kohyama, "Effect of Amylose Content and Rice Type on Dynamic Viscoelasticity of a Composite Rice Starch Gel.," *Food Hydrocolloids,* vol. 23, no. 7, pp. 1712-1719, 2009.

[18] X. J. Ge, Y.Z. Xing, C.G. Xu, Y.Q. He, "QTL analysis of rice grain elongation, volume expansion and water absorption using are combinant inbred population.," *Plant Breeding,* vol. 124, pp. 121-126, 2005.

[19] Z. Zhou, K. Robards, S. Helliwell, C. Blanchard, "Composition and functional properties of rice," *International Journal of Food Science and Technology,* vol. 37, p. 849-868, 2002.

[20] L. Amagliani, J. O'Regan, A. L. Kelly, J. A. O'Mahony, "The composition, extraction, functionality and applications of rice proteins: A review.," *Trends in Food Science & Technology,* vol. 64, pp. 1-12, 2017.

[21] R.J. Bryant, A. K. Jackson, K.M. Yeater, W. G. Yan, A. M. McClung, R. G. Fjellstrom, "Genetic variation and association mapping of protein concentration in brown rice using a diverse rice germplasm collection," *Cereal Chemistry,* vol. 90, no. 5, p. 445-452, 2013.

[22] Y. T. Thomas, R. Bath R. Y. Kuang, "Composition of amino acids, fatty acids, minerals and dietary fiber in some of the local and import rice varieties of Malaysia.," *International Food Research Journal,* vol. 22, no. 3, p. 1148-1155, 2015.

[23] K. R. Bhattacharya, P. V. Subba Rao, "Processing Conditions and Milling Yield in Parboiling of Rice," *Journal of Agricultural and Food Chemistry,* vol. 14, no. 5, pp. 473-475, 1966 .

[24] K. Liakos, P. Busato, "Sensors. Machine Learning in Agriculture: A Review.," *Sensors,* vol. 18, p. 2674, 2018.

[25] T. Q. Zheng J. L. Xu Z. K. Li H. Q. Zhai J.M. Wan "Genomic regions associated with milling quality and grain shape identified in a set of random introgression lines of rice (*Oryza sativa* L.)," *Plant Breed,* vol. 126, p. 158-163, 2007.

[26] X. Qiu, Y. Pang, Z. Yuan, D. Xing, J. Xu, M. Dingkuhn "Genome-wide association study of grain appearance and milling quality in a worldwide collection of Indica rice germplasm.," *PLoS ONE,* vol. 10, p. e0145577, 2015.

[27] X. Wang, Y. Pang, C. Wang, K. Chen, Y. Zhu, C. Shen, J. Ali, J. Xu, Z. Li, "New Candidate Genes Affecting Rice Grain Appearance and Milling Quality Detected by Genome-Wide and Gene-Based Association Analyses.," *Front. Plant Sci.* , vol. 7, p. 1998, 2017.

[28] A. J. Lisle, M. Martin, M. A. Fitzgerald "Chalky and Translucent Rice Grains Differ in Starch Composition and Structure and Cooking Properties," *Cereal Chemistry,* vol. 77, no. 5, 2000.

[29] B. Juliano, "Rice quality screening with the rapid visco analyser," in *Applications of the rapid visco analyser.*, H. J. In Walker CE, Ed., Sydney, Newport Scientific:, 1996, p. 19-24.

[30] J. Bao, Y. Xia, "Genetic control of paste viscosity characteristics in indica rice (*Oryza sativa* L.).," *Theoretical and Applied Genetics,* vol. 98, no. 6, p. 1120-1124, 1999.

[31] L.S. Collado, H. Corke, "Properties of starch noodles as affected by sweet potato genotype.," *Cereal Chemistry Journal,* vol. 74, no. 2, p. 182-187, 1997.

[32] J. Bao, S. Shen, M. Sun, H. Corke, "Analysis of genotypic diversity in the starch physicochemical properties of nonwaxy rice: apparent amylose content, pasting viscosity and gel texture.," *Starch—Starke,* vol. 58, no. 6, p. 259-67., 2006.

[33] L.Q. Wang, W.J.Liu, Y. Xu, Y.Q. He, L.J. Luo, Y.Z. Xing, C.G. Xu, Q. Zhang, "Genetic basis of 17 traits and viscosity parameters characterizing the eating and cooking quality of rice grain.," *Theoretical and Applied Genetics,* vol. 115, no. 4, p. 463-76., 2007.

[34] V. Balan, C.T. Mihai, F.D. Cojocaru, C.M. Uritu, G. Dodi, D. Botezat, I. Gardikiotis, "Vibrational Spectroscopy Fingerprinting in Medicine: from Molecular to Clinical Practice.," *Materials,* vol. 12, p. 2884, 2019.

[35] J. Durig, "Far-IR Spectroscopy, Applications.," in *In Encyclopedia of Spectroscopy and Spectrometry*, Amsterdam, NY, USA, Elsevier, 1999, p. pp. 498-504.

[36] Y. Han, S. Ling, Z. Qi, Z. Shao, X. Chen, "Application of far-infrared spectroscopy to the structural identification of protein materials.," *Phys. Chem. Chem. Phys. ,* vol. 20, p. 11643-11648, 2018.

[37] J. Workman, J. Shenk, "In: Near-Infrared Spectroscopy in Agriculture. Roberts C.A., Workman, J., Jr., and Reeves III, J.B., American Society of Agronomy, Inc., Crop Science society of America, Inc., and Soil Science Society of America, Inc.,," 2004, pp. 3-10.

[38] D. W. R. Dunmire, "Automated qualitative and quantitative NIR reflectance analyses," *Cereal Foods World,* vol. 35, pp. 913-918, 1990.

[39] W. Murray, "Chemical principle of near-infrared technology.," in *In: Near-infrared technology in the agricultural and food industries.* , P. W. a. K. N. (Eds.), Ed., American Association of Cereal Chemists. St. Paul, MN, 1990, pp. 17-34.

[40] B. M. Nicolai, T. Defraeye, B. De Ketelaere, E. Herremans, M. Hertog, W. Saeys, A. Torricelli, T. Vandendriessche, P. Verboven, "Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review.," *Journal of Postharvest Biology and Technology,* vol. 46, pp. 99-118, 2007.

[41] B. Osborne, "Review: Applications of near infrared spectroscopy in quality 36 sceening of early-generation material in cereal breeding programmes.," *J. Near Infrared Spectrosc. ,* vol. 14, pp. 93-101, 2006.

[42] B. Osborne, "Flours and breads. Ch. 8.1.," in *In: Near-Infrared Spectroscopy in Food Science and Technology.*, M. W. a. C. A. e. Ozaki Y., Ed., John Wiley & Sons, Inc., NJ,, 2007, pp. 281-296 pp.

[43] J.K.Drennen, B.D.Gebhart, E.G. Kraemer, R.A. Lodder, "Near-infrared spectrometric determination of hydrogen ion, glucose, and human serum albumin in a simulated biological matrix," *Spectroscopy,* vol. 6, pp. 28-32, 1990.

[44] J. Qu, "Applications of NIR-Spectroscopy in Food Safety Evaluation and Control: A Review of Recent Research Advances.," *Crit Rev Food Science &Nut,* vol. 55, p. 1939-54, 2015.

[45] J. Majumdar, "Analysis of agriculture data using data mining techniques: application of big data.," *J Big Data,* vol. 4, p. 20, 2017.

[46] M. Calingacion, "Diversity of Global Rice Markets and the Science Required for Consumer-Targeted Rice Breeding," *PLoSONE,* vol. 9, p. 85106, 2014.

[47] S. R. Delwiche, M. M. Bean, R. E. Miller, B. D. Webb, P. C. Williams

"Apparent amylose content of milled rice by nearinfrared reflectance spectrophotometry," Cereal Chemistry, vol. 72, p. 182-187, 1995.

[48] A. Savitzky, M. Golay, "Smoothing and Differentiation of Data by Simplified Least Squares Procedures," *Anal. Chem. ,* vol. 36, p. 1627-1639, 1964.

[49] A. G. M. Savitzky, " Smoothing and differentiation of data by simplified least squares procedures," *Analytical Chemistry,* vol. 36, p. 1627-1639, 1964.

[50] S. Xie, B. Xiang, L. Yu, H. Deng, "Tailoring noise frequency spectrum to improve NIR determinations," *Talanta,* vol. 80, p. 895-902, 2009.

[51] P. Sampaio, A. Soares, A. Castanho, A. Almeida, J. Oliveira, C. Brites, "Optimization of rice amylose determination by NIR-spectroscopy using PLS chemometrics algorithms,," *Food Chemistry,* vol. 242, pp. 196-204, 2018.

[52] I.T.Jolliffe, J. Cadima, "Principal component analysis: a review and recent developments," *Phil. Trans. R. Soc. ,* vol. A.3742015020220150202, 2016.

[53] B.G. Osborne, T. Fearn, P.H. Hindle "Near infrared calibration II.," in *Ch.7 In: Practical NIR Spectroscopy with Applications in Food and Beverage Analysis, 2nd ed.* , Longman Scientific & Technical, UK, 1993, pp. 121-144.

[54] R. A. Johnson, D. W. Wichern "Discrimination and classification. Ch.11," in *In: Applied Multivariate Statistical Analysis*, 4th ed. ed., New Jersey, A Simon and Schuster Company, Eaglewood Cliffs, 1998, pp. 629-725.

[55] H. L. Mark, D. Tunnel, "Qualitative near-infrared reflectance analysis using Mahalanobis distances.," *Anal. Chem.,* vol. 57, pp. 1449-1456, 1985.

[56] M. Barker, W. Rayens, "Partial least squares for discrimination.," *J. Chemom,* vol. 17, p. 166-173, 2003.

[57] H. Nocairi, E. Mostafa Qannari, E. Vigneau, D. Bertrand, "Discrimination on latent components with respect to patterns. Application to multicollinear data.," *Comput. Stat. Data Anal.,* vol. 48, pp. 139-147, 2005.

[58] U. Indahl, H. Martens, T. Næs, " From dummy regression to prior probabilities in PLS-DA.," *J. Chemom.,* vol. 21, pp. 529-536, 2007.

[59] M. Sjöström, S. Wold, B. Söderström, "PLS discriminant plots," in *In Pattern Recognition in Practice*, E. K. L. E. Gelsema, Ed., Amsterdam, The Netherlands, Elsevier , 1986, p. 461-470.

[60] L. Ståhle and S. Wold, "Partial least squares analysis with cross-validation for the two-class problem: A Monte Carlo study," *J. Chemom. ,* vol. 1, p. 85-196, 1987.

[61] S. Wold, "Pattern recognition by means of disjoint principal components models," *Pattern Recognit,* vol. 8, pp. 127-139, 1976.

[62] S. ML, Support vector machines, vol 1., New York: Springer, 2008.

[63] R. Balabin, R. Safieva and E. Lomakina, "Near-infrared (NIR) spectroscopy for motor oil classification: from discriminant analysis to support vector machines," *Microchem J,* vol. 98, pp. 121-128, 2011.

[64] P. Santos, E. Pereira-Filho, L. Rodriguez-Saona, "Rapid detection and quantification of milk adulteration using infrared microspectroscopy and chemometrics analysis," *Food Chem,* vol. 138, pp. 19-24, 2013.

[65] W. Kong, C. Zhang, F. Liu, P. Nie, Y. He, "Rice seed cultivar identification using near-infrared hyperspectral

imaging and multivariate data analysis.," *Sensors,* vol. 13, pp. 8916-8927, 2013.

[66] R. Wehling, "Infrared Spectroscopy. Ch.27," in *Food Analysis*, 2nd ed ed., S. (. Nielson, Ed., Gaithersburg, MD, Aspen Publishers, Inc., 1998, pp. 413-424.

[67] F. E. Dowell, J. Throne, J. E. Baker, "Automated Nondestructive Detection of Internal Insect Infestation of Wheat Kernels by Using Near-Infrared Reflectance Spectroscopy.," *J. Econ. Entomol,* vol. 91, pp. 899-904, 1998.

[68] S. Wold, M. Sjostrom, "PLS-regression: a basic tool of chemometrics," *Chemometrics and Intelligent Laboratory Systems,* vol. 58, pp. 109-130, 2001.

[69] M. Friedel, C.-D. Patz, H. Dietrich, "Comparison of different measurement techniques and variable selection methods for FT-MIR in wine analysis.," *Food Chemistry,* vol. 141, pp. 4200-4207, 2013.

[70] H.W. Lee, A. Bawn, S. Yoon, "Reproducibility, complementary measure of predictability for robustness improvement of multivariate calibration models via variable selections.," *Analytica Chimica Acta,* vol. 757, pp. 11-18, 2012.

[71] L. Nørgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck L, S.B. Engelsen, "Interval Partial least-squares regression (iPLS): A comparative chemometric study with an example from Near-Infrared spectroscopy.," Applied Spectroscopy, vol. 54, pp. 413-419, 2000.

[72] J. H. Kalivas, "Two data set for near infrared spectra," *Chemometrics Intelligent Laboratories Systems,* vol. 37, pp. 255-259, 1997.

[73] C. H. Spiegelman, M. J. McShane, M. J. Goetz, M. Motamedi, Q. L. Yue, G. L. Coté, "Theoretical justification of wavelength selection in PLS calibration: development of a new algorithm.," *Analytical Chemistry,* vol. 70, pp. 35-44, 1998.

[74] L. Leardi, J. Nørgaard, "Sequential application of backward interval partial least squares and genetic algorithms for the selection of relevant spectral regions.," *Chemometrics,* vol. 18, pp. 486-497, 2004.

[75] H.L. Ma, J.W. Wang, Y.J. Chen, J.L. Cheng, Z.T. Lai, "Rapid authentication of starch adulteration in ultrafine granular powder of Shanyao by near-infrared spectroscopy coupled with chemometric methods.," *Food Chemistry,* vol. 215, pp. 108-115, 2017.

[76] Y.-H. Yun, H.-D. Li, L. R. E. Wood, W.Fan, J.-J. Wang, D.-S. Cao, Q.-S. Xu, Y.-Z. Liang, "An efficient method of wavelength interval selection based on random frog for multivariate spectral calibration.," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy,* vol. 111, pp. 31-36, 2013.

[77] M. Daszykowski, J. Orzel, M. Wrobel, H. Czarnik-Matusewicz, B. Walczak, "Improvement of classification using robust soft classification rules for near-infrared reflectance spectral data Improvement of classification using robust soft classification rules for near-infrared reflectance spectral data.," *Chemometr. Intell. Lab.,* vol. 109, p. 86-93, 2011.

[78] S. Wold, M. Sjöström, "SIMCA: A Method for Analyzing Chemical Data in Terms of Similarity and Analogy.," in *In Chemometrics: Theory and Application*, vol. 52, Washington, DC, American Chemical Society, 1977, pp. 243-282.

[79] H. Yue, S. Qin, "Reconstruction-Based Fault Identification Using a Combined Index.," *Ind. Eng. Chem. Res.,* vol. 40, p. 4403-4414, 2001.

[80] M. Cocchi, A. Biancolillo, F. Marini, "Chapter Ten-Chemometric Methods for Classification and Feature Selection.," in *In Comprehensive Analytical Chemistry;*, J. B. C. T. R. E. Jaumot, Ed., Amsterdam, The Netherlands, Elsevier, 2018.

[81] J. Ruiz, T. Parello, R. Gomez, "Comparative study of multivariate methods to identify paper finishes using infrared spectroscopy.," *IEEE Trans. Instrum. Meas.,* vol. 61, pp. 1029-1036, 2012.

[82] L. Breiman, "Random forests.," *Mach. Learn.,* vol. 45, pp. 5-32, 2001.

[83] L. Fraiwan, K. Lweesy, N. Khasawneh, H. Wenz, H. Dickhaus, "Automated sleep stage identification system based on time-frequency analysis of a single EEG channel and random forest classifier.," *Comput. Meth. Programs Biomed.,* vol. 108, pp. 10-19, 2012.

[84] M.N. Vrahatis, G. D. Magoulas, K. Parsopoulos, V. P. Plagianakos, "Introduction to artificial neural network training and applications Conference Paper," 2000.

[85] D.E. Rumelhart, G. E. Hinton, R. J. Williams, "Learning representations by back-propagating errors.," *Nature,* vol. 323, pp. 533-536, 1986.

[86] S. Haykin, Neural Networks: a comprehensive foundation., New York.: Macmillan, 1999.

[87] ASCE Task Committee on Application of Artificial Neural Networks in Hydrology, "Artificial neural networks in hydrology I: preliminary concepts.," *J Hydrol Eng 5:115-123,* 2000.

[88] S. Cohen, N. Intrator, "Automatic model selection in a hybrid perceptron/radial network.," *Inf Fusion Special Issue Mult Experts,* vol. 3, no. 4, pp. 259-266, 2002.

[89] J. Kenneth, S. Wernter, J. MacInyre, "Knowledge extraction from radial basis function networks and multilayer perceptrons.," *Int J Comput Intell Appl,* vol. 1, no. 3, pp. 369-382, 2001.

[90] W. Loh, L. Tim, "A comparison of prediction accuracy, complexity, and training time of thirty three old and new classification algorithm.," *Mach Learn,* vol. 40, no. 3, pp. 203-238, 2000.

[91] Mohammad Ali Ghorbani, Rahman Khatibi, Behrouz Hosseini, Mehmet Bilgili, "Relative importance of parameters affecting wind speed prediction using artificial neural networks.," *Theor Appl Climatol,* vol. 114, no. 1, pp. 107-114, 2013.

[92] F. Rosenblatt, Principles of Neurodynamics, New York: Spartan, 1962.

[93] S. Theodoridis, Machine Learning: A Bayesian and Optimization Perspective, London, UK: Academic Press, 2015.

[94] E. Lanza, "Determination of Moisture, Protein, Fat, and Calories in Raw Pork and Beef By Near Infrared Spectroscopy.," *Journal of Food Science,* vol. 48, no. 2, pp. 471-474, 1983.

[95] J. Rodriguez-Otero, M. Hermida, A. Cepeda, "Determination of fat, protein, and total solids in cheese by near-infrared reflectance spectroscopy.," *J AOAC Int.,* vol. 78, no. 3, pp. 802-806, 1995.

[96] P. Williams, "Application of chemometrics to prediction of some wheat quality factors by near-infrared spectroscopy," *Cereal Chem.,* vol. 97, pp. 958-966, 2020.

[97] C. Malegori, S. Buratti, S. Benedetti, *P. oliveri*, S. Ratti, C. Cappa, M. Lucisano, "A modified mid-level data fusion approach on electronic nose and FT-NIR data for evaluating the effect of

different storage conditions on rice germ shelf life.," *Talanta,* vol. 206, p. 120208, 2020.

[98] B. Le, "Application of deep learning and near infrared spectroscopy in cereal analysis.," *Vib. Spectrosc.,* vol. 106, p. 103009, 2020.

[99] H. Jiang, T. Liu, Q. Chen, "Dynamic monitoring of fatty acid value in rice storage based on a portable near-infrared spectroscopy system.," *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.,* vol. 240, p. 118620., 2020.

[100] P. Sampaio, A. Castanho, A. Almeida, J. Oliveira, C. Brites, "Identification of rice flour types with near-infrared spectroscopy associated with PLS-DA and SVM methods.," *Eur. Food Res. Technol.,* vol. 246, pp. 527-537, 2019.

[101] J. Barnaby, T. Huggins, H. Lee, A. McClung, S. Pinson, M. Oh, G. Bauchan, L. Tarpley, K. Lee, M. Kim, "Vis/NIR hyperspectral imaging distinguishes sub-population, production environment, and physicochemical grain properties in rice.," *Sci. Rep.,* vol. 10, pp. 1-13, 2020.

[102] S. Delwiche, K. McKenzie, B.D. Webb, "Quality characteristics in rice by near-infrared reflectance analysis of whole-grain milled samples.," *Cereal Chem,* vol. 73, pp. 257-263, 1996.

[103] F. Meadows, F. Barton, "Determination of Rapid Visco Analyser Parameters in Rice by Near-Infrared Spectroscopy," *Cereal Chemistry,* vol. 79, pp. 563-566, 2002.

[104] J. Bao, Y. Cai, H. Corke, "Prediction of Rice Starch Quality Parameters by Near-Infrared Reflectance Spectroscopy," *Journal of Food Science,* vol. 66, pp. 936-939, 2001.

[105] J.-F. Meullenet, A. Mauromoustakos, T. Horner and B. Marks, "Prediction of Texture of Cooked White Rice by Near-Infrared Reflectance Analysis of Whole-Grain Milled Samples," *Cereal Chemistry,* vol. 79, pp. 52-57, 2002.

[106] F. Xie, F. Dowell and X. Sun, "Using visible and near-infrared reflectance spectroscopy and differential scanning calorimetry to study starch, protein and temperature effects on bread staling.," *Cereal Chem,* vol. 81, pp. 249-254, 2004.

[107] B. Osborne, B. Mertens, M. Thompson, T. Fearn, "The Authentication of Basmati Rice Using near Infrared Spectroscopy," *J. Near Infrared Spectrosc,* vol. 1, pp. 77-83, 1993.

[108] L. Wimonsiri, P. Ritthiruangdej, S. Kasemsumran, N. Therdthai, W. Chanput, Y. Ozaki, "Rapid analysis of chemical composition in intact and milled rice cookies using near infrared spectroscopy," *J. Near Infrared Spectrosc.,* vol. 25, pp. 330-337 , 2017.

[109] j. Chen, M. Li, T. Pan, L. Pang, L. Yao, J. Zhang, "Rapid and non-destructive analysis for the identification of multi-grain rice seeds with near-infrared spectroscopy," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy,* vol. 219, pp. 179-185, 2019.

[110] X. He, X. Feng, D. Sun, F. Liu, Y. Bao, Y. He, "Rapid and Nondestructive Measurement of Rice Seed Vitality of Different Years Using Near-Infrared Hyperspectral Imaging," *Molecules,* vol. 24, no. 12, p. 2227, 2019.

[111] S.D. Afandi, Y.Herdiyeni, L. B. Prasetyo, W. Hasbi, K. Arai, H. Okumura, "Nitrogen Content Estimation of Rice Crop Based on Near Infrared (NIR) Reflectance Using Artificial Neural Network (ANN)," *Procedia Environmental Sciences,* vol. 33, pp. 63-69, 2016.

[112] L.-H. Lin, F.-M. Lu, Y.-C. Chang, "Prediction of protein content in rice using a near-infrared imaging system as diagnostic technique.," *Int J Agric & Biol Eng,* vol. 12, no. 2, pp. 195-200, 2019.

[113] Z. Zi-li, W. Chun-Feng, J. Di, H. Yong, L. Xiao-li and S. Yong-ni, "Discrimination of varieties of rice using near infrared spectral by PCA and MDA model," in *6th International Conference on Computer Science & Education (ICCSE)*, 2011.

[114] J. Johnson, "An overview of near-infrared spectroscopy (NIRS) for the detection of insect pests in stored grains.," *J. Stored Prod. Res.,* vol. 86, p. 101558, 2020.

[115] F. Kosmowski, T. Worku, "Evaluation of a miniaturized NIR spectrometer for cultivar identification: The case of barley, chickpea and sorghum in Ethiopia," *PLoS ONE,* vol. 13, p. e0193620/1–e0193620/17., 2018.

[116] C. Blanch-Perez-del-Notario, W. Saeys, A. Lambrechts, "Fast ingredient quantification in multigrain flour mixes using hyperspectral imaging.," *Food Control ,* vol. 118, p. 107366, 2020.

[117] C. Viejo, D. Torrico, F. Dunshea, S. Fuentes, "Emerging Technologies Based on Artificial Intelligence to Assess the Quality and Consumer Preference of Beverages.," *Beverages,* vol. 5, p. 62, 2019.

[118] A. Kaya, A. Keçeli, C. Catal and B. Tekinerdogan, "Sensor Failure Tolerable Machine Learning-Based Food Quality Prediction Model.," *Sensors,* vol. 20, p. 3173, 2020.

[119] K. Böhme, P. Calo-Mata, J. Barros-Velázquez, I. Ortea, "Review of Recent DNA-Based Methods for Main Food-Authentication Topics," *J. Agric. Food Chem,* vol. 67, pp. 3854-3864, 2019.

[120] W. Liu, X. Wang, J. Tao, B. Xi, M. Xue, W. Sun, "A Multiplex PCR Assay Mediated by Universal Primers for the Detection of Adulterated Meat in Mutton.," *J. Food Prot. ,* vol. 82, p. 325-330, 2019.

[121] R. Yin, Y. Sun, K. Wang, N. Feng, H. Zhang, M. Xiao, "Development of a PCR-based lateral flow strip assay for the simple, rapid, and accurate detection of pork in meat and meat products.," *Food Chem,* vol. 318, p. 126541, 2020.

[122] J. Spink, D. Moyer, "Understanding and combating food fraud.," *Food Technol,* vol. 67, pp. 30-35, 2013.

[123] M. Xiao, Y. Chen, H. Chu, R. Yin, "Development of a polymerase chain reaction—Nucleic acid sensor assay for the rapid detection of chicken adulteration.," *LWT,* vol. 131, p. 109679, 2020.

[124] A. Vinayaka, T. Ngo, K. Kant, P. Engelsmann, V. Dave, M.-A. Shahbazi, A. Wol, D. Bang, "Rapid detection of *Salmonella enterica* in food samples by a novel approach with combination of sample concentration and direct PCR.," *Biosens. Bioelectron,* vol. 129, pp. 224-230, 2019.

[125] B. Pakbin, A. Basti, A. Khanjari, L. Azimi, A. Karimi, "Differentiation of stx1A gene for detection of *Escherichia coli* serotype O157: H7 and Shigella dysenteriae type 1 in food samples using high resolution melting curve analysis.," *Food Sci. Nutr.,* vol. 8, p. 3665-3672, 2020.

[126] D.-D. Li, C.-B. Hao, Z.-M. Liu, S.-J. Wang, Y. Wang, Z. Chao, S.-Y. Gao, S. Chen, "Development of a novel dual priming oligonucleotide system-based PCR assay for specific detection of Salmonella from food samples.," *J. Food Saf ,* vol. 40, p. 12789, 2020.

[127] Y. Geng, G. Liu, L. Liu, Q. Deng, L. Zhao, X. Sun, J. Wang, B. Zhao, J. Wang, "Real-time recombinase polymerase

amplification assay for the rapid and sensitive detection of Campylobacter jejuni in food samples.," *J. Microbiol. Methods,* vol. 157, p. 31-36, 2019.

[128] N. Salihah, M. Hossain, M. Abdul Hamid, M. Ahmed, "A novel, rapid, and sensitive real-time PCR assay for cost-effective detection and quantification of *Staphylococcus aureus* in food samples with the ZEN double quenched probe chemistry," *Int. Food Res. J. ,* vol. 26, p. 193-201, 2019.

[129] S. Rani, A. Pradhan, "Evaluation and meta-analysis of test accuracy of direct PCR and bioassay methods for detecting Toxoplasma gondii in meat samples.," *LWT,* vol. 131, p. 109666, 2020.

[130] R. Köppel, A. Sendic, H.-U. Waiblinger, "Two quantitative multiplex real-time PCR systems for the efficient GMO screening of food products.," *Eur. Food Res. Technol. ,* vol. 239, p. 653-659, 2014.

[131] D. Bwambok, N. Siraj, S. Macchi, N. Larm, G. Baker, R-L. Pérez, "QCM Sensor Arrays, Electroanalytical Techniques and NIR Spectroscopy Coupled to Multivariate Analysis for Quality Assessment of Food Products, Raw Materials, Ingredients and Foodborne Pathogen Detection: Challenges and Breakthroughs," *Sensors,* vol. 20, no. 23, p. 6982, 2020.

[132] E. Teye, C.L.Y. Amuah, T. McGrath, C. Elliott, "Innovative and rapid analysis for rice authenticity using hand-held NIR spectrometry and chemometrics," *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy,* vol. 217, pp. 147-154, 2019.

[133] Y. Liu, Y. Li, Y. Peng, Y. Yang, Q. Wang, "Detection of fraud in high-quality rice by near-infrared spectroscopy.," *J. Food Sci.,* vol. 85, p. 2773-2782, 2020.

[134] D. Le Nguyen Doan, Q. C. Nguyen, F. Marini, A. Biancolillo, "Authentication of Rice (*Oryza sativa* L.) Using Near Infrared Spectroscopy Combined with Different Chemometric Classification Strategies," *Appl. Sci.,* vol. 11, no. 1, p. 362, 2021.

[135] R. Hongyan, Z. Dafang, Y. Junxing, Y. Xinfang, "A Feasibility Study of NIR Spectra in Identifying Heavy Metal Contamination in Rice Around Abandoned Tailing Ponds: A Case Study in Guiyang County in South China Ren Hongyan.," *J Geophys Remote Sens,* vol. 2, p. 1, 2013.

[136] J. Yang, Y. Zha and H. Liu, "The distribution and chemical forms of Cd, Cu and Pb in polluted seeds.," *China Environmental Science,* vol. 19, pp. 500-504, 1999.