Avrupa Bilim ve Teknoloji Dergisi
Sayı 25, S. 475-484, Ağustos 2021
© Telif hakkı EJOSAT'a aittir
Derleme Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
No. 25, pp. 475-484, August 2021
Copyright © 2021 EJOSAT
Review Article

A Review of Data Analysis Techniques Used in Near-Infrared Spectroscopy

Özcan Çataltaş[1*], Kemal Tütüncü[2]

[1*] Selçuk University, Faculty of Technology, Department of Electrical and Electronic Eng., Konya, Turkey, (ORCID: 0000-0002-7136-6574), ozcancataltas@selcuk.edu.tr

[2] Selçuk University, Faculty of Technology, Department of Electrical and Electronic Eng., Konya, Turkey, (ORCID: 0000-0002-3005-374X), ktutuncu@selcuk.edu.tr

(First received 19 February 2021 and in final form 8 July 2021)

(**DOI:** 10.31590/ejosat.882749)

ATIF/REFERENCE: Çataltaş, Ö. & Tütüncü, K. (2021). A Review of Data Analysis Techniques Used in Near-Infrared Spectroscopy. *European Journal of Science and Technology*, (25), 475-484.

Abstract

Although the analysis of the structure of objects and the components that makeup them has been done for decades, it is one of today's research topics to do this analysis quickly and without damaging the sample. Near-infrared spectroscopy is used in many areas due to its non-contact measurement, fast analysis, and high accuracy features. Near-infrared spectroscopy is used in the classification or quality analysis of products, especially in the agriculture and food sector, due to the chemical bonds interacting in this region. The most critical part of achieving a successful result in near-infrared spectroscopy is pre-processing and analyzing the spectral data using the correct method. In this review, we perform a survey of recent studies that use near-infrared spectroscopy in food production and agriculture. Since there are many studies in this field in the literature, the survey is limited to cover works in the last five years. The review's main question is the pre-processing and data analysis methods used in these studies and the main features of these methods. Among the examined studies, the most frequently used pre-processing method was standard normal variate, and the most frequently used analysis method was partial least squares regression. In addition, the software tools and the spectrum range were also examined within the scope of the study.

Keywords: Near-infrared Spectroscopy, Chemometric, Pre-processing Methods, Regression, Classification.

Yakın Kızılötesi Spektroskopisinde Kullanılan Veri Analizi Tekniklerinin Bir Derlemesi

Öz

Nesnelerin yapılarının ve onları oluşturan bileşenlerin analizi onlarca yıldır yapılsa da bu analizi hızlı ve örneğe zarar vermeden yapmak günümüzün araştırma konularından biridir. Yakın kızılötesi spektroskopisi, temassız ölçüm, hızlı analiz ve yüksek doğruluk özellikleri nedeniyle birçok alanda kullanılmaktadır. Yakın kızılötesi spektroskopisi, bu bölgede etkileşen kimyasal bağlar nedeniyle özellikle tarım ve gıda sektöründe ürünlerin sınıflandırılmasında veya kalite analizinde sıkça kullanılmaktadır. Yakın kızılötesi spektroskopisinde başarılı bir sonuç elde etmenin en kritik kısmı, doğru yöntemi kullanarak spektral verileri ön işleme ve analiz etmektir. Bu çalışmada, gıda üretimi ve tarımda yakın kızılötesi spektroskopisi kullanan son çalışmaların bir derlemesini gerçekleştirdik. Literatürde bu alanda çok sayıda çalışma olduğu için çalışma son beş yıldaki çalışmaları kapsayacak şekilde sınırlandırılmıştır. Derlemenin ana sorusu, bu çalışmalarda kullanılan ön işleme ve veri analizi yöntemleri ve bu yöntemlerin temel özellikleridir. İncelenen çalışmalarda en sık kullanılan ön işleme yöntemi standart normal dağılım, en sık kullanılan analiz yöntemi ise kısmi en küçük kareler regresyon olarak bulunmuştur. Ayrıca, kullanılan yazılım araçları ve spektrum aralığı da çalışma kapsamında incelenmiştir.

Anahtar Kelimeler: Yakın Kızılötesi Spektroskopi, Kemometri, Ön İşleme Yöntemleri, Regresyon, Sınıflandırma.

[*] Corresponding Author: ozcancataltas@selcuk.edu.tr

1. Introduction

The awareness of using quality and healthy products in the food sector is increasing [1]. The quality analysis made by traditional methods generally includes physical and chemical processes. It is not a practical method because it takes longer, and the sample cannot be used again in most cases [2]. With the increasing need for quality analysis of products, the desire to develop more practical analysis methods has gradually increased [3].

Near-infrared spectroscopy (NIRS) is a contactless, relatively inexpensive, and rapid analysis tool for detecting the specifications or quality of subjects [4]. Because of these advantages, it is used in almost every field, especially in food analysis.

The infrared region lies between the visible region and the microwave region in the electromagnetic spectrum, divided into three subgroups: near-infrared, mid-infrared, and far-infrared. The near-infrared region covers wavelengths between 780 nm and 2500 nm (12821–4000 cm^{-1}) [5].

Near-infrared spectroscopy is the process of analyzing samples using near-infrared spectrum waves. Waves generated by an electromagnetic wave generator (e.g., light source) are directed to the sample. According to the chemical bonds in the sample structure, it reflects some waves and passes some waves. Reflected or transmitted waves from the sample are collected by a sensor [6]. Thus, the spectrum data of the sample is obtained. If the reflected waves from the sample are collected, it is called reflectance spectroscopy; if the waves passing through the sample are collected, it is called transmittance spectroscopy [7]. Also, there is a third category, namely absorbance spectroscopy. In this type, the spectrum is calculated using transmittance spectrum data with the notion that “waves that do not pass through the object are absorbed”. There are fundamental systemic design differences between the three structures [8].

In spectroscopy studies, every atomic bond does not affect every spectrum region. In the Near-infrared region, mostly overtones or combinations of -CH, -OH, -NH bands are observed [9]. These atomic bonds are common in vegetable and animal food products, containing water components, fat components, and protein structures [10].

Another benefit of NIRS in the food industry is that it helps prevent food fraud [11]. Especially high-priced products are subjected to fraud in order to gain unfair profit. Pereira et al., Mabood et al., Du et al., and Rodionova et al. are examples of such studies [12-15].

A review of chemometrics used in near-infrared spectroscopy has already been done by Roggo et al. [16]. However, this study was conducted 13 years ago, and this study focused on the methods used in pharmaceutical technologies. This review aims to handle near-infrared spectroscopy applications and data analysis methods, especially in the last five years. This review will provide readers with general information about data analysis methods used in NIRS.

2. Material and Method

2.1. Methodology

The literature search was carried out using keywords in academic databases such as IEEE Xplore, ScienceDirect, and Google Scholar. Also, the results were filtered to cover 2016 and later years to concentrate only on current studies. The keywords used are as follows:

“Near-infrared Spectroscopy”, “NIRS in Agriculture”, “Application of NIRS”

The studies obtained from the literature research were pre-examined, and those not related to the agricultural products or food sector were eliminated. After the pre-selection process, 35 studies remained. These studies have been analyzed in detail one by one according to the following questions:

- Which device was used to obtain spectrum data?
- What range of spectral data had been used?
- Which food product was targeted for analysis?
- Which pre-processing techniques were used?
- Which data analysis methods were used?
- Which pre-processing and analysis method combination had shown the most successful performance?
- What was the main conclusion of the study?

2.2. Pre-processing Methods

As mentioned above, NIRS data are acquired through photodiode-based sensors. These sensors generally output as analog signals and are transferred to the computer with the help of analog-digital converter circuits [17]. For this reason, deviations in the signal, called noise, which negatively affects the data analysis, occur during both the measurement phase and the conversion phase. Therefore, the pre-processing of spectral data is the most crucial step before analyzing it.

Pre-processing spectral data before using it, does not always mean that it will positively affect model success. Zhu and Tian [18] applied Savitzky-Golay (SG), standard normal variate (SNV), multiplicative scatter correction (MSC), normalization, 1st derivative and 2nd derivative pre-processing methods to determine the sugar content of Fuji apples, and it was observed that these methods did not give better results than the non-pre-processed condition. Sampaio et al. [19] used different pre-processing methods with PLS, iPLS, siPLS, and mwPLS regression models to determine amylose content in rice. For siPLS, the highest determination coefficient (Rp) was obtained with non-pre-processed spectral data.

In 2009, Rinnan et al. [20] conducted a literature review on pre-processing techniques in NIRS. In this section, the most used pre-processing methods among 35 studies are discussed in more detail.

2.2.1. Multiplicative Scatter Correction

The attenuation of a sample is ideally linearly related to the total absorption coefficient. However, in the presence of the scattering effect, the relationship between attenuation and absorption is nonlinear [21]. Multiplicative scatter correction (MSC) is one of the methods that reduce the scattering effect in spectral data. MSC is a row-oriented method, and the new value of one data is affected by its horizontal neighbors [22]. First, the mean value is calculated for each data point. The coefficients of the best-fitting curve for each sample's spectrum to this mean value are calculated using the least-squares method [23]. The MSC-treated spectrum data is obtained with the following equation:

$$x_i^{MSC} = \frac{(x_i - a_i)}{b_i} \quad (1)$$

where x_i is ith input row vector, x_i^{MSC} is ith MSE-treated output row vector, a_i and b_i are ith coefficients.

Rebellato et al. [24] applied the MSC and 1st derivative pre-processing method together with the partial least squares regression (PLSR) to assess the mineral content in hamburgers and obtained R^2_{pred} values of 0.72 for potassium, 0.93 for sodium, and 0.96 for calcium. Zhang et al. [25] used the NIRS to rapidly analyze the amount of free anthraquinone and total anthraquinone in rhubarb. Before PLSR and particle swarm optimization based least square support vector machines (PSO-LSSVM) methods, they tried different pre-processing methods that combinations of MSC, standard normal variate (SNV), and Savitsky-Golay (SG) methods. They obtained the most successful result with the combination of MSC and SG. Maestresalas et al. [26] used MSC to classify Lidia breed and foal meat mixed into beef samples and obtained the correct classification rate of 95.24% for Lidia meat and 100% for foal meat.

Although MSC reduces the scattering effect of light on spectral data, the failure to separate the physical light effect from the chemical light effect has led to emerging of the extended multiplicative signal correction (EMSC) method [27]. Quelal-Vásconez et al. [28] applied different methods and combinations as a pre-processing before creating the PLSR model to identify the cocoa shell in cocoa powder and obtained the best result with the EMSC - orthogonal signal correction (OSC) combination. In another study in which EMSC and MSC were used as pre-treatment, EMSC gave more successful results than MSC in protein estimation of weathered sorghum grain samples [29]. Monago-Maraña et al. was used EMSC to classify paprika powder [30].

2.2.2. Standard Normal Variate

Standard normal variate (SNV) is another pre-processing method that reduces the scattering effect in spectral data [31]. In SNV, the data is first centered by subtracting its average from the spectrum data. SNV-treated data is then obtained by dividing by its standard deviation [32]. SNV can be calculated with the following equation:

$$x_{i,j}^{SNV} = \frac{x_{i,j} - \bar{x}_i}{\sigma_i} \quad (2)$$

where $x_{i,j}$ is input data, $x_{i,j}^{SNV}$ is SNV-treated output data, σ_i is the standard deviation of ith row, $\bar{x}_i$ is mean of the ith row.

Sampaio et al. used the SNV method with other pre-processing methods to determine the amount of rice amylose and obtained the highest Rpred value with the combination of SNV – SG [19]. Mabood et al. combined SNV and unit vector normalization (UVN) methods to detect pork meat mixture in different meats and found the R^2_{cal} value as 0.977 [12]. Udompetaikul et al. [33] used the SNV method to analyze the soluble solids content of sugarcane billets, while Genisheva et al. [34] used the SNV method to analyze volatile compounds in wine. In the study done by Firmani et al. [35], Darjeeling tea was mixed with other teas. The highest correct classification rate was obtained with the combination of SNV, 1st derivative, and mean centering (MC) methods.

2.2.3. Savitzky-Golay Filter

Savitzky and Golay [36] have developed a type of digital smoothing filter known by their name in 1964. In this type of filter, the data set is fitted to a polynomial degree using the least-squares method with the help of convolution. Convolution coefficients to be used in the Savitsky-Golay (SG) filter are predetermined according to the differentiation order and polynomial degree. Luo et al. [37] improved the SG filter to be used for even-numbered data.

SG filter has been used as a pre-processing method in many studies on NIRS. Krepper et al. [38] applied different pre-processing techniques before PLS, iPLS, and iSPA-PLS models to determine fat content in chicken hamburgers. The most successful results were obtained with SG filters for all three methods. Puertas and Vázquez [39] analyzed egg yolk with a UV-VIS-NIR spectrometer to determine cholesterol content. In this study, where different pre-processing methods were used, they obtained the lowest Root Mean Squared Error of Calibration (RMSEC) value by applying SNV and SG pre-processing methods together with the Principle Component Regression (PCR) method. Lu et al. [40] applied five different pre-processing methods and five different modeling methods and their combination for analyzing the moisture content of coconut. They obtained the most successful result for each model with SG pre-processing method. Mishra et al. [41] used SNV and SG methods and PLSR method to improve moisture and SSC prediction in pear.

Apart from the methods mentioned above, other pre-processing methods are as follows: Linear Baseline Correction (LBC or BC) is used if the data graph has a constant base value. Puertas and Vázquez, Mabood et al., Krepper et al., and Femenias et al. used the LBC method in their studies [12, 38, 39, 42]. The orthogonal signal correction (OSC) introduced by Wold et al. [43] aims to delete unrelated orthogonal variations from the dataset, and Quelal-Vásconez et al. [28] have included this method in their work. Ning et al. [44] have used continuous wavelet transform (CWT) as a pre-treatment method. Mabood et al. [12] have applied unit vector normalization (UNV) with SNV to minimize the scattering effect. Mean centering (MC) provides that the dataset has zero mean, which is done by subtracting the mean value of the dataset from the dataset. MC was used in López-Maestresalas et al. and Yuan et al. [26, 45]. De-trending (DT) is a method used to correct baseline changes and curvilinear spectrum patterns. DT was used as a pre-processing method in López-Maestresalas et al. and Samadi et al. [26, 46]. Du et al. [14] were used Norris derivative as a pre-processing method to detect adulteration of different oils in camellia oil. Also, meaning average filter was used as a pre-processing method in Bahrami et al. and Udompetaikul et al. [33, 47].

Figure 1 shows the pre-treatment methods and usage frequencies used in 35 studies. MSC, SNV, and SG have been the most commonly used pre-processing methods in the studies. When we compare the pre-processing methods used in the latest studies, it is clear that there is no successful method in all spectral data. For this purpose, different pre-processing methods and their different combinations have been applied in many studies, and the most successful method has been preferred.

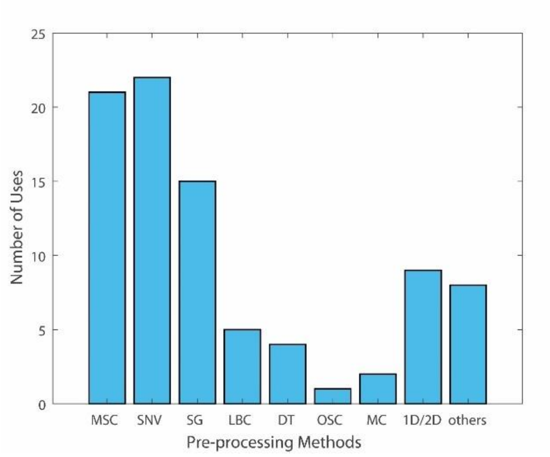

Figure 1. Pre-processing methods and frequency of use

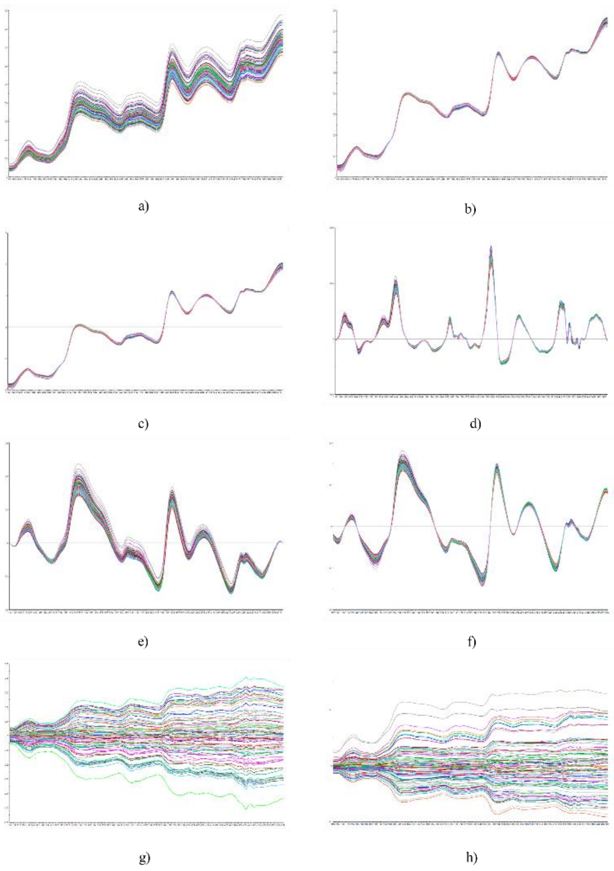

Figure 2. Data set of 80 corn samples measured between 1100-2498 nm wavelengths are taken from (https://eigenvector.com/resources/data-sets); unprocessed spectrum graph (a), pre-processed with MSC (b), SNV (c), SG (d), LBC (e), DT (f), OSC (g), MC (h).

In Figure 2, MSC, SNV, SG, LBC, DT, OSC, and MC pre-processing methods were applied on the corn data set taken from (https://eigenvector.com/resources/data-sets) [48], and the obtained spectrum graphics were given in Figure 1 (b), (c), (d), (e), (f), and (g), respectively. When Figure 1 is examined, it is very clearly seen that the uncertain details in the original spectrum graph appear after pre-processing.

2.3. Data Analysis Methods

In NIRS systems, it is almost impossible to extract information directly from the sensor's data. For this reason, data analysis methods are used to classify and analyze using spectral data [49]. These methods are also called chemometric methods. In this section, the most used chemometric methods among 35 studies are discussed in more detail.

2.3.1. Multiple Linear Regression

Multi Linear Regression (MLR) is a statistical method that tries to predict a variable using other variables. It is an extended version of the classical linear regression for multiple variables [50]. MLR aims to establish a linear connection between input variables and output variables, as in linear regression. MLR can be calculated with the following equation:

$$y_i = \alpha + \beta_1 * x_{i,1} + \beta_2 * x_{i,2} + \cdots + \beta_n * x_{i,n} \qquad (3)$$

where y_i is predicted value (output variable), x_i is input variables, α is the intercept of the fitted line, β_n is slopes. One of the success criteria of the model created with MLR is the coefficient of determination (R^2). R^2 value is a parameter that shows how much of the variation in the output variable is explained by the input parameters [51]. While R^2 value can take a value between 0 and 1, 0 indicates that the output value is entirely independent of the input variables, and 1 indicates that the actual value of the output variable can be calculated precisely with the input variables. However, R^2 increases as the number of input variables increase even if it is not related to the output variable. This situation reduces the reliability of the R^2. For this purpose, together with R^2, the p-value is used as a success criterion [52].

Wang et al. used MLR method after Successive Projection Algorithm (SPA) and Competitive Adaptive Reweighted Sampling (CARS) methods to evaluate black tea's taste attributes [53]. They found the highest Rp value for the bitterness of tea as 0.9437 with CARS-MLR. Huang et al. developed models with PLSR and MLR to determine the four main catechins and caffeine amounts, which are the most significant factors determining tea quality [54]. With the MLR model, better results were obtained for five parameters. Yuan et al. applied PLS and MLR regression models and their deviation fusion to analyze “Yunhe” pears attribute with NIRS, and as a result, they obtained a successful result with Rp value of 0.9026 [45]. Berhow et al. applied SG pre-processing method to the spectral data created to determine the isoflavone and saponin content of ground soybeans and modeled them with MLR. They have obtained promising results, especially in determining the number of isoflavones [55].

Another version of MLR modeling is Step-Wise Multiple Linear Regression (SMLR). In SMLR, MLR is applied to the data set multiple times. The least correlated input variable is removed each time. In this way, it continues until the highest R^2 value is achieved [51]. In Shen et al., SMLR method was employed to establish aflatoxin quantification models [56].

2.3.2. Partial Least Squares Regression

The partial least squares regression (PLSR) method is a technique that combines the basic features of the principal component analysis (PCA) and MLR methods [57]. It is beneficial in datasets where the number of observations is well below the number of the features [58]. It is a method frequently used in different branches of science.

Huang et al. analyzed the acoustic firmness, impact firmness, compression area, and puncture parameters of tomatoes using VIS / SWNIR and NIR spectrometers using PLS regression [59]. It was observed that the model created in the VIS / SWNIR region gave more successful results than the NIR region. Deng et al. analyzed the protein, carbohydrate, energy, and fat content of medical foods using the model they developed with PLSR [60]. In Shen et al. [56], successful results were obtained by obtaining Rp values between 0.922 and 0.973 thanks to the model created to determine the amount of aflatoxins in brown rice that causes cancer. Mabood et al. used the PLSR model to determine adulteration in camel milk with goat milk and obtained the R^2 value as 0.94 [61]. Das et al. combined PCA and PLSR methods with different machine learning methods for salinity stress phenotyping of rice [62]. As a result of their study, they obtained the performance of the methods as PLSR-combined > PCA-combined. Udompetaikul et al. used NIRS to determine the soluble solids contents of sugar cane with an on-line system on the conveyor and obtained the R^2 value as 0.805 in the test set with the PLSR model [33]. Maraphum et al. used the PLSR method to measure the starch content of cassava tubers [63]. Measurements can be made in the field with the system they have created. Genis et al. used the PLSR method to detect spinach and green pea adulteration in pistachio, a type of nuts common in Turkey, and obtained R^2 value of over 99% [64]. Yang et al. were designed a portable system to detect the main compositions of milk samples [65]. They used SG, MSC, SNV, and 2D pre-processing methods and PLSR method to establish a reliable method. Yi et al. used modified partial least squares (MPLS), a variant of PLS, to predict the chemical composition of the walnut kernel [66]. Bahrami et al. used PLSR method with SNV, meaning average (MA), and area normalization (AN) pre-processing methods to measure quality parameters of sugar beet juices [47].

2.3.3. Principal Component Regression

Although PLSR is a suitable modeling method for many spectral data, especially if the independent variables have a high correlation with each other, this negatively affects the regression coefficients [67]. This situation is called the collinearity problem. One method that gives more successful results in such datasets is Principal Component Regression (PCR) [68]. The PCR method is formed by the application of PCA and LSR methods together. First of all, the principal components of the spectral data are extracted with the PCA method. The ones that show the most variance in data are selected among these principal components, and the regression model is created using these principal components.

Puertas and Vázquez applied the PCR method and the PLS method to determine the amount of cholesterol in the egg yolk and generally obtained more successful results with the PLS method [39]. Samadi et al. analyzed nutritive parameters in feed using PCR [46]. They obtained R^2 value 0.83 for the Neutral Detergent Fibre (NDF) parameter and 0.867 for Acid Detergent Fibre (ADF) parameter with pre-processed spectral data.

2.3.4. Support Vector Machine

The support vector machine (SVM) can be defined as a vector space-based machine learning method that finds a decision boundary between the two furthest classes from any point in the training data. SVM is a technique developed by Vapnik et al., and is frequently used, especially in classification problems [69].

Dankowska and Kowalewski classified different tea types using QDA, SVM, RDA, and LDA in synchronous fluorescence, UV-Vis, and NIR spectra [70]. In the classification made using the fusion of all spectra, all methods correctly classified all samples. The correct classification rate was obtained in individual tests with the SVM method as 93.3%, but the most successful result was obtained with QDA.

SVM is used not only for classification problems but also for regression. This method, called Support Vector Machine Regression (SVMR), contains all the main features of SVM. SVMR was proposed in 1996 by Drucker et al. [71]. In SVMR, unlike classical regression, instead of the best fit line, an interval containing an acceptable error is obtained. Ning et al. used NIRS to identify pit mud properties that determine the quality of Chinese liquor [44]. First of all, they classified new and aged soil using four different methods, including SVM. Prediction accuracy was found as 100%. Then, they determined total carbon, total nitrogen, and total phosphorus in pit mud with high accuracy with SVR and PLSR. Das et al. analyzed the salt tolerance of rice phenotypes with VIS / NIRS to increase the yield of salty soils [62]. They applied different methods for data analysis, such as indices-based, PLSR-combined, and PCA-combined. It has been observed that the most successful result of the SVMR method, which is one of the methods they use, is given when combined with PLSR.

Suykens and Vandewalle proposed another version of SVM; the least squares support vector machines (LS-SVM) [72]. With this method, unlike the classical SVM method, the solution is found using a linear equation set instead of the quadratic programming problem for two-class applications. Yang et al. combined synchronous, asynchronous NIR-MIR spectroscopy and their fusion with LS-SVM to discriminate sesame oil adulterated with corn oil [73]. They obtained correct classification rates as 98.1% and 100% for the calibration and prediction set, respectively.

2.3.5. Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is a method used for dimensionality reduction [74]. Although its primary goal is similar to PCA, LDA's purpose is to determine the optimum new axis that maximizes the distance between different categories and minimizes the variance of the categories. Shen et al. used NIR and MIR regions to detect aflatoxin, a carcinogenic substance in brown rice [56]. In the study in which LDA was used as the classification technique, an average accuracy of 96.9% was obtained with NIR spectral data. One of the areas where NIRS benefits is food adulteration. De Girolamo et al. did one of the case studies for this purpose [11]. They analyzed durum wheat pasta adulteration with common wheat using LDA and PLS-DA. They classified the pasta samples. They divided into three groups with 95% accuracy using LDA.

2.3.6. Partial Least Squares Discriminant Analysis

Partial least squares discriminant analysis (PLS-DA) is a variant of PLSR dedicated to classification problems [75].

López-Maestresalas et al. were used PLS-DA to detect adulteration of minced lamb and beef with different meat types [26]. Moscetti et al. were utilized NIRS and image analysis to discriminate pine nuts from different geographic origins [75]. In the NIRS part of the study, they obtained an accuracy value of over 96% with PLS-DA. Firmani et al. used PLS-DA after 1D, 2D, MC, and SNV pre-processing methods, and the least average classification error was obtained as 4.55% [35].

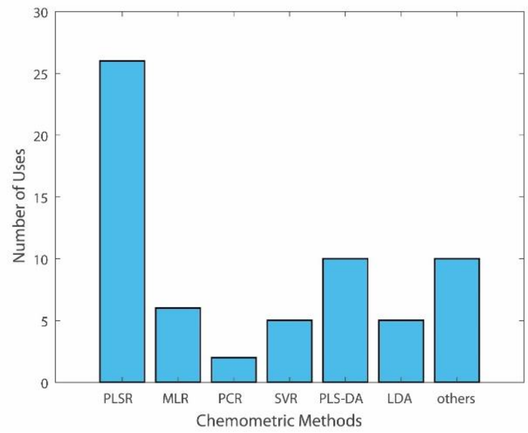

Figure 3. Chemometrics and frequency of use

The classification and analysis methods used in 35 studies are given in figure 3. When figure 3 is examined, it is seen that at least one of the PLSR and PLS-DA methods was used in 88% of the studies.

3. Results and Discussion

3.1. Performance Metrics

The authors used different metrics to determine the success rates of the created models. A list of these metrics with their definitions is given in table 1. RPD has been the most popular metric among 35 studies.

3.2. Spectrum Range Analysis

The spectrum ranges of the 35 studies are combined and given in figure 4. Only wavelengths between 300nm and 2500nm are included when forming the figure. When figure 4 is examined, it is seen that approximately 85% of the studies included 1000-1650 nm wavelengths, and approximately 65% of them included 1650-2500 nm wavelengths. Although analyzing in a wider spectrum increases the success rate as it contains more information, instrumentation costs are also increasing. In the SWNIR region (750-1100nm), Si photodiodes are used as a detector, much more affordable than InGaAs photodiodes. However, as shown in figure 4, only 45% of the studies were carried out in the SWNIR region. The main reason for this can be said that the atomic bonds that makeup food products generally have meaningful data around 1400, 1600, 1750, and 2300 nm [10].

3.3. Software

Data obtained from NIRS systems have to be analyzed using different methods. Nowadays, with the development of the software industry, these tasks are performed using popular software. The software used in the pre-processing and chemometric processes of the 35 studies is shown in figure 5.

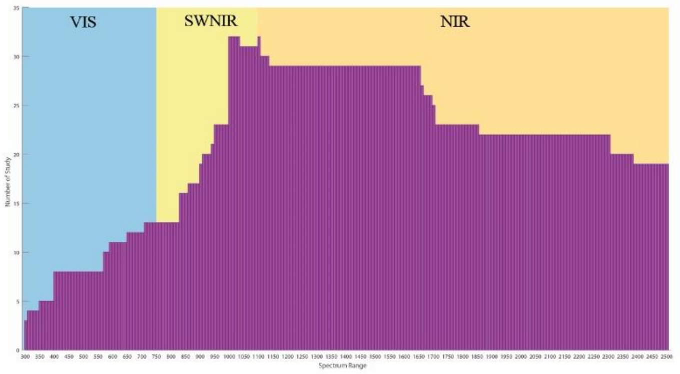

Figure 4. Spectral ranges of 35 studies

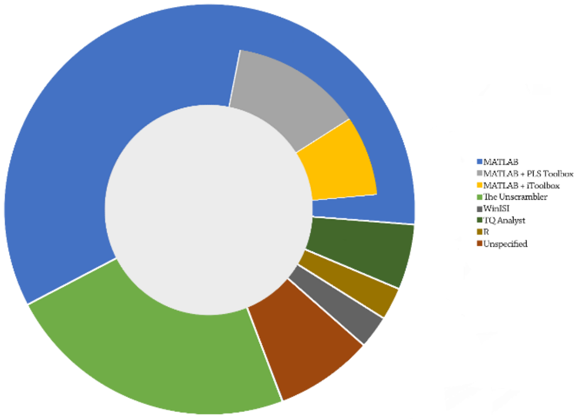

Figure 5. Software tools used for data analysis

When figure 5 is examined, it is seen that the most popular software used is MATLAB (Mathworks Inc., Natick, MA, USA) with a ratio of 60%. Also, PLS_Toolbox (Eigenvector Research Inc., WA, USA), and iToolbox (http://www.models.life.ku.dk/itoolbox) toolboxes provided by different companies or researchers with MATLAB were used. The second most used software was Unscrambler (Camo, Process, AS, Oslo, Norway) with a 24% usage rate. Other software used are WinISI (Infrasoft International, Port Matilda, PA, USA), TQ Analyst (Thermo Electron Corp., Madison, WI, USA), and R (cran.r-project.org/bin/windows/base/old/3.4.1).

4. Conclusions and Recommendations

In this review, data pre-processing and chemometric methods used in NIRS systems in recent years were examined. A summary of the reviewed 35 studies is presented in Table 2. It has been seen that although different pre-processing techniques have been used to eliminate the distortions in the spectrum data, no method can be called the best. In many studies, different pre-processing methods and their combinations have been applied. As the chemometric method, it has been seen that the PLS-DA method was the most commonly used in classification problems and the PLSR method in regression problems. The main reasons for this can be said that both are easy to perform and have high accuracy.

It has also been seen that artificial intelligence and deep learning, which are popular research areas today, have not been used much in NIRS systems.

Our aim is that this review would motivate more researchers to experiment with near-infrared spectroscopy, applying it for solving various food problems involving adulteration, food quality. With this review, researchers will be able to examine the pre-processing and chemometric methods used in recent years and shape their studies.

Table 1. Summary of performance metrics

Metric	Symbol	Formula	Description
Root Mean Square Error of Calibration	RMSEC	$\sqrt{\frac{1}{N_{cal}}\sum_{i=1}^{N_{cal}}(y_{cal,i}-\hat{y}_{cal,i})^2}$	RMSEC equals the root of the mean square of the difference between the actual and obtained values in the calibration set. The closer the RMSEC value to zero, the better.
Root Mean Square Error of Prediction	RMSEP	$\sqrt{\frac{1}{N_{pred}}\sum_{i=1}^{N_{pred}}(y_{pred,i}-\hat{y}_{pred,i})^2}$	RMSEP equals the root of the mean square of the difference between the actual and obtained values in the prediction set. The closer the RMSEC value to zero, the better.
The correlation coefficient of calibration	Rc	$\frac{\sum_{i=1}^{N_{cal}}[(y_{cal,i}-\bar{y}_{cal,i}) * (\hat{y}_{cal,i}-\bar{\hat{y}}_{cal,i})]}{\sqrt{\sum_{i=1}^{N_{cal}}(y_{cal,i}-\bar{y}_{cal,i})^2 * \sum_{i=1}^{N_{cal}}(\hat{y}_{cal,i}-\bar{\hat{y}}_{cal,i})^2}}$	Rc corresponds to how much obtained values correlated to actual values in the calibration set. It may have values between -1 to 1; higher is better.
The correlation coefficient of prediction	Rp	$\frac{\sum_{i=1}^{N_{pred}}[(y_{pred,i}-\bar{y}_{pred,i}) * (\hat{y}_{pred,i}-\bar{\hat{y}}_{pred,i})]}{\sqrt{\sum_{i=1}^{N_{pred}}(y_{pred,i}-\bar{y}_{pred,i})^2 * \sum_{i=1}^{N_{pred}}(\hat{y}_{pred,i}-\bar{\hat{y}}_{pred,i})^2}}$	Rp corresponds to how much obtained values correlated to actual values in the prediction set. It may have values between -1 to 1; higher is better.
The ratio of performance to deviation	RPD	$\frac{\sigma}{SEP}$	RPD is calculated by dividing the standard deviation to the standard error of prediction. RPD can be interpreted this way: it is suitable for screening if RPD>3, suitable for quality control if RPD>5, suitable for analytical tasks if RPD>8 [76].
The coefficient of determination	R^2	$\frac{Explained\ variation}{Total\ variation}$	R^2 corresponds to how the independent variables explain much variation of a dependent variable. R^2 can take a value between 0 and 1, 1 means that all the output variable variation is fully explained using input variables.
Correct classification ratio	CCR	$\%\frac{Correctly\ classified\ samples}{Total\ samples}$	CCR shows as a percentage of how many samples were correctly classified.
Sensitivity	Sens	$\frac{TP}{TP+FN}$	Sensitivity is the proportion of true positives among all positives.

N_{cal}, the number of samples for calibration; N_{cal}, the number of samples for prediction; TP, true positive; FN, false negative; $y_{cal,i}$, the real value of ith sample for calibration set; $y_{pred,i}$, the real value of ith sample for prediction set; $\hat{y}_{cal,i}$, the calculated value of ith sample for calibration set; $\hat{y}_{pred,i}$, the calculated value of ith sample for prediction set; σ, standard deviation; SEP, standard error of prediction.

References

[1] Grunert KG. Food quality and safety: Consumer perception and demand. European Review of Agricultural Economics 2005:32, 369-391, doi: https://doi.org/10.1093/eurrag/jbi011.

[2] Rajput H, Rehal J, Goswami D Mandge HM. Methods for food analysis and quality control. In State-of-the-art technologies in food science; ed.; Eds; 2017; 396.

[3] Porep JU, Kammerer DR Carle R. On-line application of near infrared (nir) spectroscopy in food production. Trends in Food Science & Technology 2015:46, 211-230, doi: https://doi.org/10.1016/j.tifs.2015.10.002.

[4] Johnson JB Naiker M. Seeing red: A review of the use of near-infrared spectroscopy (nirs) in entomology. Applied Spectroscopy Reviews 2019:55, 810-829, doi: https://doi.org/10.1080/05704928.2019.1685532.

[5] Salzer R. Practical guide to interpretive near-infrared spectroscopy. By jerry workman, jr. And lois weyer; 2008.

[6] Dix LML, van Bel F, Baerts W Lemmers PMA. Comparing near-infrared spectroscopy devices and their sensors for monitoring regional cerebral oxygen saturation in the neonate. Pediatric Research 2013:74, 557-563, doi: https://doi.org/10.1038/pr.2013.133.

[7] Woolley JT. Reflectance and transmittance of light by leaves. Plant Physiology 1971:47, 656-662, doi: https://doi.org/10.1104/pp.47.5.656.

[8] Siesler HW, Ozaki, Y.,Kawata, S., Heise, H.M. Near-infrared spectroscopy: Principles, instruments, applications; WILEY-VCH Verlag GmbH: 2001.

[9] *Handbook of near-infrared analysis*; Burns DA Ciurczak EW. Boca Raton: CRC Press, 2007.

[10] Petisco C, García-Criado B, Vázquez-de-Aldana BR, de Haro A García-Ciudad A. Measurement of quality parameters in intact seeds of brassica species using visible and near-infrared spectroscopy. Industrial Crops and Products 2010:32, 139-146, doi: https://doi.org/10.1016/j.indcrop.2010.04.003.

[11] De Girolamo A, Arroyo MC, Cervellieri S, Cortese M, Pascale M, Logrieco AF Lippolis V. Detection of durum wheat pasta adulteration with common wheat by infrared spectroscopy and chemometrics: A case study. LWT 2020:127, 109368, doi: https://doi.org/10.1016/j.lwt.2020.109368.

[12] Mabood F,Boqué R,Alkindi AY,Al-Harrasi A,Al Amri IS,Boukra S,Jabeen F,Hussain J,Abbas G,Naureen Z et al. Fast detection and quantification of pork meat in other meats by reflectance ft-nir spectroscopy and multivariate analysis. Meat Science 2020:163, 108084, doi: https://doi.org/10.1016/j.meatsci.2020.108084.

[13] Pereira EVdS, Fernandes DDdS, de Araújo MCU, Diniz PHGD Maciel MIS. Simultaneous determination of goat milk adulteration with cow milk and their fat and protein contents using nir spectroscopy and pls algorithms. LWT 2020:127, 109427, doi: https://doi.org/10.1016/j.lwt.2020.109427.

[14] Du Q, Zhu M, Shi T, Luo X, Gan B, Tang L Chen Y. Adulteration detection of corn oil, rapeseed oil and sunflower oil in camellia oil by in situ diffuse reflectance near-infrared spectroscopy and chemometrics. Food Control 2021:121, 107577, doi: https://doi.org/10.1016/j.foodcont.2020.107577.

[15] Rodionova OY, Fernández Pierna JA, Baeten V Pomerantsev AL. Chemometric non-targeted analysis for detection of soybean meal adulteration by near infrared spectroscopy. Food Control 2021:119, 107459, doi: https://doi.org/10.1016/j.foodcont.2020.107459.

[16] Roggo Y, Chalus P, Maurer L, Lema-Martinez C, Edmond A Jent N. A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies. Journal of Pharmaceutical and Biomedical Analysis 2007:44, 683-700, doi: https://doi.org/10.1016/j.jpba.2007.03.023.

[17] Pinheiro PP, Santos JCFD França MBDM. Development, testing, and validation of a prototype for qualification of substances based on near-infrared spectroscopy. IEEE Access 2019:7, 25650-25659, doi: https://doi.org/10.1109/ACCESS.2019.2900800.

[18] Zhu G Tian C. Determining sugar content and firmness of 'fuji' apples by using portable near-infrared spectrometer and diffuse transmittance spectroscopy. Journal of Food Process Engineering 2018:41, e12810, doi: https://doi.org/10.1111/jfpe.12810.

[19] Sampaio PS, Soares A, Castanho A, Almeida AS, Oliveira J Brites C. Optimization of rice amylose determination by nir-spectroscopy using pls chemometrics algorithms. Food Chemistry 2018:242, 196-204, doi: https://doi.org/10.1016/j.foodchem.2017.09.058.

[20] Rinnan Å, Berg Fvd Engelsen SB. Review of the most common pre-processing techniques for near-infrared spectra. TrAC Trends in Analytical Chemistry 2009:28, 1201-1222, doi: https://doi.org/10.1016/j.trac.2009.07.007.

[21] Lu B, Morgan SP, Crowe JA Stockford IM. Comparison of methods for reducing the effects of scattering in spectrophotometry. Applied Spectroscopy 2006:60, 1157-1166, doi: https://doi.org/10.1366/000370206778664725.

[22] Maleki MR, Mouazen AM, Ramon H De Baerdemaeker J. Multiplicative scatter correction during on-line measurement with near infrared spectroscopy. Biosystems Engineering 2007:96, 427-433, doi: https://doi.org/10.1016/j.biosystemseng.2006.11.014.

[23] Chen JY, Iyo C, Terada F Kawano S. Effect of multiplicative scatter correction on wavelength selection for near infrared calibration to determine fat content in raw milk. Journal of Near Infrared Spectroscopy 2002:10, 301-307, doi: https://doi.org/10.1255/jnirs.346.

[24] Rebellato AP, Caramês ETdS, Moraes PPd Pallone JAL. Minerals assessment and sodium control in hamburger by fast and green method and chemometric tools. LWT 2020:128, 109438, doi: https://doi.org/10.1016/j.lwt.2020.109438.

[25] Zhang S, Ma H, Pan H, Shao Q, Liu X Wu Y. Quantitative real-time release testing of rhubarb based on near-infrared spectroscopy and method validation. Vibrational Spectroscopy 2019:104, 102964, doi: https://doi.org/10.1016/j.vibspec.2019.102964.

[26] López-Maestresalas A, Insausti K, Jarén C, Pérez-Roncal C, Urrutia O, Beriain MJ Arazuri S. Detection of minced lamb and beef fraud using nir spectroscopy. Food Control 2019:98, 465-473, doi: https://doi.org/10.1016/j.foodcont.2018.12.003.

[27] Martens H, Nielsen JP Engelsen SB. Light scattering and light absorbance separated by extended multiplicative signal correction. Application to near-infrared transmission analysis of powder mixtures. Analytical Chemistry 2003:75, 394-404, doi: https://doi.org/10.1021/ac020194w.

[28] Quelal-Vásconez MA, Lerma-García MJ, Pérez-Esteve É, Arnau-Bonachera A, Barat JM Talens P. Fast detection of cocoa shell in cocoa powders by near infrared spectroscopy and multivariate analysis. Food Control 2019:99, 68-72, doi: https://doi.org/10.1016/j.foodcont.2018.12.028.

[29] Peiris KHS, Bean SR Jagadish SVK. Extended multiplicative signal correction to improve prediction accuracy of protein content in weathered sorghum grain samples. Cereal Chemistry n/a, doi: https://doi.org/10.1002/cche.10329.

[30] Monago-Maraña O, Eskildsen CE, Galeano-Díaz T, Muñoz de la Peña A Wold JP. Untargeted classification for paprika powder authentication using visible – near infrared spectroscopy (vis-nirs). Food Control 2021:121, 107564, doi: https://doi.org/10.1016/j.foodcont.2020.107564.

[31] Barnes RJ, Dhanoa MS Lister SJ. Correction to the description of standard normal variate (snv) and de-trend (dt) transformations in practical spectroscopy with applications in food and beverage analysis—2nd edition. NIR news 1994:5, 6-6, doi: https://doi.org/10.1255/nirn.248.

[32] Zeaiter M Rutledge D. Preprocessing methods. In Comprehensive chemometrics*;* 1st ed.; Brown, S. D., Tauler, R. ,Walczak, B., Eds; Elsevier, 2009; 121-231.

[33] Udompetaikul V, Phetpan K Sirisomboon P. Development of the partial least-squares model to determine the soluble

solids content of sugarcane billets on an elevator conveyor. Measurement 2021:167, 107898, doi: https://doi.org/10.1016/j.measurement.2020.107898.

[34] Genisheva Z, Quintelas C, Mesquita DP, Ferreira EC, Oliveira JM Amaral AL. New pls analysis approach to wine volatile compounds characterization by near infrared spectroscopy (nir). Food Chemistry 2018:246, 172-178, doi: https://doi.org/10.1016/j.foodchem.2017.11.015.

[35] Firmani P, De Luca S, Bucci R, Marini F Biancolillo A. Near infrared (nir) spectroscopy-based classification for the authentication of darjeeling black tea. Food Control 2019:100, 292-299, doi: https://doi.org/10.1016/j.foodcont.2019.02.006.

[36] Savitzky A Golay MJE. Smoothing and differentiation of data by simplified least squares procedures. Analytical Chemistry 1964:36, 1627-1639, doi: https://doi.org/10.1021/ac60214a047.

[37] Luo J, Ying K Bai J. Savitzky–golay smoothing and differentiation filter for even number data. Signal Processing 2005:85, 1429-1434, doi: https://doi.org/10.1016/j.sigpro.2005.02.002.

[38] Krepper G, Romeo F, Fernandes DDdS, Diniz PHGD, de Araújo MCU, Di Nezio MS, Pistonesi MF Centurión ME. Determination of fat content in chicken hamburgers using nir spectroscopy and the successive projections algorithm for interval selection in pls regression (ispa-pls). Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 2018:189, 300-306, doi: https://doi.org/10.1016/j.saa.2017.08.046.

[39] Puertas G Vázquez M. Cholesterol determination in egg yolk by uv-vis-nir spectroscopy. Food Control 2019:100, 262-268, doi: https://doi.org/10.1016/j.foodcont.2019.01.031.

[40] Lu B, Wang X, Liu N, He K, Wu K, Li H Tang X. Feasibility of nir spectroscopy detection of moisture content in coco-peat substrate based on the optimization characteristic variables. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 2020:239, 118455, doi: https://doi.org/10.1016/j.saa.2020.118455.

[41] Mishra P, Woltering E, Brouwer B Hogeveen-van Echtelt E. Improving moisture and soluble solids content prediction in pear fruit using near-infrared spectroscopy with variable selection and model updating approach. Postharvest Biology and Technology 2021:171, 111348, doi: https://doi.org/10.1016/j.postharvbio.2020.111348.

[42] Femenias A, Gatius F, Ramos AJ, Sanchis V Marín S. Near-infrared hyperspectral imaging for deoxynivalenol and ergosterol estimation in wheat samples. Food Chemistry 2021:341, 128206, doi: https://doi.org/10.1016/j.foodchem.2020.128206.

[43] Wold S, Antti H, Lindgren F Öhman J. Orthogonal signal correction of near-infrared spectra. Chemometrics and Intelligent Laboratory Systems 1998:44, 175-185, doi: https://doi.org/10.1016/S0169-7439(98)00109-9.

[44] Ning Y, Zhang H, Zhang Q Zhang X. Rapid identification and quantitative pit mud by near infrared spectroscopy with chemometrics. Vibrational Spectroscopy 2020:110, 103116, doi: https://doi.org/10.1016/j.vibspec.2020.103116.

[45] Yuan L-M, Mao F, Chen X, Li L Huang G. Non-invasive measurements of ‘yunhe’ pears by vis-nirs technology coupled with deviation fusion modeling approach. Postharvest Biology and Technology 2020:160, 111067, doi: https://doi.org/10.1016/j.postharvbio.2019.111067.

[46] Samadi, Wajizah S Munawar AA. Near infrared spectroscopy (nirs) data analysis for a rapid and simultaneous prediction of feed nutritive parameters. Data in Brief 2020:29, 105211, doi: https://doi.org/10.1016/j.dib.2020.105211.

[47] Bahrami ME, Honarvar M, Ansari K Jamshidi B. Measurement of quality parameters of sugar beet juices using near-infrared spectroscopy and chemometrics. Journal of Food Engineering 2020:271, 109775, doi: https://doi.org/10.1016/j.jfoodeng.2019.109775.

[48] https://eigenvector.com/resources/data-sets/

[49] Wold S. Chemometrics; what do we mean with it, and what do we want from it? Chemometrics and Intelligent Laboratory Systems 1995:30, 109-115, doi: https://doi.org/10.1016/0169-7439(95)00042-9.

[50] Mark H Workman J. Chapter 4 - matrix algebra and multiple linear regression: Part 1. In Chemometrics in spectroscopy (second edition)*;* ed.; Mark, H. ,Workman, J., Eds; Academic Press, 2018; 27-35.

[51] Riffenburgh RH Gillen DL. 16 - multiple linear and curvilinear regression and multifactor analysis of variance. In Statistics in medicine (fourth edition)*;* ed.; Riffenburgh, R. H. ,Gillen, D. L., Eds; Academic Press, 2020; 391-435.

[52] Fritz M Berger PD. Chapter 10 - can you relate in multiple ways? Multiple linear regression and stepwise regression. In Improving the user experience through practical data analytics*;* ed.; Fritz, M. ,Berger, P. D., Eds; Morgan Kaufmann, 2015; 239-269.

[53] Wang Y-J, Li T-H, Li L-Q, Ning J-M Zhang Z-Z. Evaluating taste-related attributes of black tea by micro-nirs. Journal of Food Engineering 2021:290, 110181, doi: https://doi.org/10.1016/j.jfoodeng.2020.110181.

[54] Huang Y, Dong W, Sanaeifar A, Wang X, Luo W, Zhan B, Liu X, Li R, Zhang H Li X. Development of simple identification models for four main catechins and caffeine in fresh green tea leaf based on visible and near-infrared spectroscopy. Computers and Electronics in Agriculture 2020:173, 105388, doi: https://doi.org/10.1016/j.compag.2020.105388.

[55] Berhow MA, Singh M, Bowman MJ, Price NPJ, Vaughn SF Liu SX. Quantitative nir determination of isoflavone and saponin content of ground soybeans. Food Chemistry 2020:317, 126373, doi: https://doi.org/10.1016/j.foodchem.2020.126373.

[56] Shen F, Wu Q, Shao X Zhang Q. Non-destructive and rapid evaluation of aflatoxins in brown rice by using near-infrared and mid-infrared spectroscopic techniques. Journal of Food Science and Technology 2018:55, 1175-1184, doi: https://doi.org/10.1007/s13197-018-3033-1.

[57] Abdi H Williams LJ. Partial least squares methods: Partial least squares correlation and partial least square regression. In Computational toxicology: Volume ii*;* ed.; Reisfeld, B. ,Mayeno, A. N., Eds; Humana Press, 2013; 549-579.

[58] Guebel DV Torres NV. Partial least-squares regression (plsr). In Encyclopedia of systems biology*;* ed.; Dubitzky, W., Wolkenhauer, O., Cho, K.-H. ,Yokota, H., Eds; Springer New York, 2013; 1646-1648.

[59] Huang Y, Lu R Chen K. Prediction of firmness parameters of tomatoes by portable visible and near-infrared spectroscopy. Journal of Food Engineering 2018:222, 185-198, doi: https://doi.org/10.1016/j.jfoodeng.2017.11.030.

[60] Deng Y, Wang Y, Zhong G Yu X. Simultaneous quantitative analysis of protein, carbohydrate and fat in

nutritionally complete formulas of medical foods by near-infrared spectroscopy. Infrared Physics & Technology 2018:93, 124-129, doi: https://doi.org/10.1016/j.infrared.2018.07.027.

[61] Mabood F,Jabeen F,Ahmed M,Hussain J,Al Mashaykhi SAA,Al Rubaiey ZMA,Farooq S,Boqué R,Ali L,Hussain Z et al. Development of new nir-spectroscopy method combined with multivariate analysis for detection of adulteration in camel milk with goat milk. Food Chemistry 2017:221, 746-750, doi: https://doi.org/10.1016/j.foodchem.2016.11.109.

[62] Das B, Manohara KK, Mahajan GR Sahoo RN. Spectroscopy based novel spectral indices, pca- and plsr-coupled machine learning models for salinity stress phenotyping of rice. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 2020:229, 117983, doi: https://doi.org/10.1016/j.saa.2019.117983.

[63] Maraphum K, Saengprachatanarug K, Wongpichet S, Phuphaphud A Posom J. In-field measurement of starch content of cassava tubers using handheld vis-near infrared spectroscopy implemented for breeding programmes. Computers and Electronics in Agriculture 2020:175, 105607, doi: https://doi.org/10.1016/j.compag.2020.105607.

[64] Genis HE, Durna S Boyaci IH. Determination of green pea and spinach adulteration in pistachio nuts using nir spectroscopy. LWT 2021:136, 110008, doi: https://doi.org/10.1016/j.lwt.2020.110008.

[65] Yang B, Zhu Z, Gao M, Yan X, Zhu X Guo W. A portable detector on main compositions of raw and homogenized milk. Computers and Electronics in Agriculture 2020:177, 105668, doi: https://doi.org/10.1016/j.compag.2020.105668.

[66] Yi J, Sun Y, Zhu Z, Liu N Lu J. Near-infrared reflectance spectroscopy for the prediction of chemical composition in walnut kernel. International Journal of Food Properties 2017:20, 1633-1642, doi: https://doi.org/10.1080/10942912.2016.1217006.

[67] Næs T Martens H. Principal component regression in nir analysis: Viewpoints, background details and selection of components. Journal of Chemometrics 1988:2, 155-167, doi: https://doi.org/10.1002/cem.1180020207.

[68] Mandel J. Use of the singular value decomposition in regression analysis. The American Statistician 1982:36, 15-24, doi: https://doi.org/10.2307/2684086.

[69] Smola AJ Schölkopf B. A tutorial on support vector regression. Statistics and Computing 2004:14, 199-222, doi: https://doi.org/10.1023/B:STCO.0000035301.49549.88.

[70] Dankowska A Kowalewski W. Tea types classification with data fusion of uv–vis, synchronous fluorescence and nir spectroscopies and chemometric analysis. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 2019:211, 195-202, doi: https://doi.org/10.1016/j.saa.2018.11.063.

[71] Drucker H, Burges CJC, Kaufman L, Smola AJ Vapnik V. Support vector regression machines. Advances in Neural Information Processing Systems 1997:9, 155-161, doi:

[72] Suykens JAK Vandewalle J. Least squares support vector machine classifiers. Neural Processing Letters 1999:9, 293-300, doi: https://doi.org/10.1023/A:1018628609742.

[73] Yang R, Dong G, Sun X, Yang Y, Liu H, Du Y, Jin H Zhang W. Discrimination of sesame oil adulterated with corn oil using information fusion of synchronous and asynchronous two-dimensional near-mid infrared spectroscopy. European Journal of Lipid Science and Technology 2017:119, 1600459, doi: https://doi.org/10.1002/ejlt.201600459.

[74] Hastie T, Tibshirani R Friedman J. The elements of statistical learning; Springer: California, 2009.

[75] Moscetti R, Berhe DH, Agrimi M, Haff RP, Liang P, Ferri S, Monarca D Massantini R. Pine nut species recognition using nir spectroscopy and image analysis. Journal of Food Engineering 2021:292, 110357, doi: https://doi.org/10.1016/j.jfoodeng.2020.110357.

[76] Ritthiruangdej P, Ritthiron R, Shinzawa H Ozaki Y. Non-destructive and rapid analysis of chemical compositions in thai steamed pork sausages by near-infrared spectroscopy. Food Chemistry 2011:129, 684-692, doi: https://doi.org/10.1016/j.foodchem.2011.04.110.

Table 2: Summary of the reviewed studies

Ref	Device/ Spectral Range	Sample	Pre-treatment method	Method	Goal	Best combination	Results							
							Rc	Rp	RMSEC	RMSEP	RPD	R^2	CCR Acc	Sens.
[53]	NIR-S-R2 950–1650 nm	Black Tea	SPA, CARS	PLSR, MLR	Bitterness	CARS-MLR	0.9330	0.9437	0.5012	0.5058	3.07			
					Astringency	CARS-PLSR	0.8779	0.9119	0.5650	0.5410	2.28			
					Caffeine	CARS-PLSR	0.9517	0.9509	2.7200	3.1300	3.29			
					EGCG	CARS-PLSR	0.9117	0.9387	2.0500	1.7400	2.91			
[54]	XDS 400-2498 nm	Green Tea	SG, SNV, MSC, CARS, SPA	PLSR, MLR	Epigallocatechin gallate	MLR	0.964	0.949	1.777	2.146				
					Epicatechin gallate	MLR	0.923	0.893	1.168	1.389				
					Epigallocatechin	MLR	0.967	0.968	1.143	1.137				
					Epicatechin	MLR	0.943	0.931	0.379	0.421				
					Caffeine	PLSR	0.922	0.918	0.810	0.828				
[45]	VIS-NIR scan Nano 590-1091nm	Pears	de-MC, MSC, SNV, 1st Derivative	PLSR, MLR	Soluble solids content	0.9077*PLS + 0.0923*MLR	0.9067	0.9026	0.52	0.59				
[55]	FOSS XDS 400-2498 nm	Ground Soybean	SG, SPA, siPLS, UVE	MLR	Total daidzein forms	SG+MLR			0.206	0.288	2.01	0.78		
					Total glycitein forms				0.083	0.108	0.93	0.34		
					Total genistein forms				0.280	0.350	1.80	0.75		
					Total isoflavones				0.407	0.616	2.06	0.80		
					Total saponins				1.342	1.750	0.63	0.58		
[40]	Flame NIR 940-1660 nm	Cocopeat substrate	SG, SNV, MSC, PCA	PLSR, MLR	Moisture	SG (2D) + MLR + SPA	0.9976	0.9963	1.0989	1.4029	11.28			
[39]	V-670 190-2500 nm	Egg yolk	SNV, MSC, SG, BC, DT	PLSR, PCR	Cholesterol	SG + BC + PLSR	0.93	0.88	0.6	0.82	2.85			
[46]	Thermo Nicolet Antaris II 1000-2500 nm	Feed	Smoothing, normalization, MSC, SNV, BC, DT	PLSR, PCR	NDF	PCR					1.925	0.830		
					ADF	PCR					1.972	0.867		
					IVOMD	SNV + PLSR					2.347	0.81		
					IVDMD	SNV + PLSR					3.911	0.86		
[70]	MPA, Genesis 6, Thermo 190-2500 nm	Tea	PCA	RDA, QDA, SVM, LDA	Classification	QDA, RDA, SVM							100	

[44]	Nicolet 6700 1000-2500 nm	Pit mud	SG, SNV, MSC, CWT	HCA, PLS-DA, ANN, SVM	Classification (new/aged)	PLS-DA, SVM						100	
				PLSR, ANN, SVR, ELM	Total Carbon	CWT+SVM		0.9814			5.2407		
					Total Nitrogen			0.9771			4.7096		
					Total Phosphorus			0.9430			3.0168		
[62]	GER 1500 282-1097 nm	Rice	PCA-Combined, PLSR-Combined	ELNET, SVMR, KNN, GPR, MARS, RF, XGB, GAM	K	PLSR + GAM	0.93	0.87					
					Ca	PLSR + GAM	0.88	0.88					
					Mg	PLSR+ELNET	0.90	0.82					
					Na	PLSR + GPR	0.96	0.87					
					Zn	PLSR+ELNET	0.92	0.93					
					Cu	PLSR + GAM	0.89	0.87					
[73]	Perkin 1000-2500 nm 2500-15000 nm	Sesame oil	PLS-DA	LS-SVM	Adulteration of corn oil	Synchronous						96.3	
						Asynchronous						96.3	
						Fusion						100	
[56]	MB3600 833-2500 nm	Brown rice	MSC	LDA	Classification	NIR Spectra						96.9	
				PLSR, SMLR	AFB1		0.981	0.973	105	70.4	4.0		
					AFB2		0.958	0.969	6.4	4.7	3.3		
					AFG1		0.941	0.936	180	144	2.8		
					AFG2		0.939	0.945	13.1	9.8	2.5		
					AFs		0.950	0.951	299	231	3.1		
[11]	Nicolet iS50 1000-25000 nm	Durum wheat pasta	PCA	LDA, PLS-DA	Classification	PLS-DA						97	
[24]	Perkin 1000-2500 nm	Hamburger	MSC, SG, PCA	PLS-DA, PLSR	Classification of Na intensity	PLS-DA							100
					Iron	SG + MSC + PLSR	0.94	0.85					
					Potassium		0.92	0.85					
					Sodium		0.98	0.96					
					Calcium		0.96	0.98					
[25]	Matrix-F 833-2500 nm	Rhubarb	CARS, MSC, SNV, SG	PLSR, PSO-LSSVM	Free anthraquinone	PSO-LSSVM	0.9891	0.9589	0.04865	0.07961	3.587		
					Total anthraquinone		0.9944	0.9735	0.03932	0.07810	4.370		

[26]	Luminar 5030 1100-2300 nm	Meet fraud	MC, MSC, SNV, 1D, 2D, DT	PLS-DA	Lamb-Pork	Raw							90
					Lamb-Chicken	MC							79.16
					Lamb- Lidia breed	1D + MC							86.36
					Lamb-Foal	2D + MC							85
					Beef-Pork	1D + MC							80
					Beef-Chicken	SNV+DT+MC							78.95
					Beef-Lidia breed	MSC+MC							95.24
					Beef-Foal	1D+MC							100
[28]	FOSS 5000 1100-2500 nm	Cocoa powder	EMSC, OSC, SNV, SG, PCA	PLS-DA, PLSR	Detection of Cocoa shell	EMSC + OSC+ PLSR			1.91	2.43	5.03	0.967	
[29]	Perten DA7250 950-1650 nm	Weathered sorghum grain	MSC, EMSC	PLSR	Protein	MSC+PLSR			0.69	0.47		0.92	
[19]	MPA 833-2500 nm	Rice	PCA, MSC, SNV, SG	PLSR, iPLS, siPLS, mwPLS	Amylose determination	SNV + SG + PLSR	0.92	0.90	2032	2435			
[12]	Perkin 1000-2500 nm	Pork Meat	SNV, BC, UNV, 1D	PLS-DA, PLSR	Discrimination of adulterated meat	SNV + UNV + PLSR			0.077	1.183		0.992	
[33]	AvaSpec-2048 350-1100 nm	Sugarcane	MA, SNV	PLSR	Prediction of soluble solids content	MA+ SNV + PLSR				0.31		0.805	
[34]	FTLA 2000 714-16666 nm	Wine	PCA	PLSR	Ethyl acete	PCA+PLSR					3.9	0.95	
					Methanol						4.4	0.96	
					2-Methyl-1-butanol						4.7	0.96	
					3-Methyl-1-butanol						4.7	0.96	
					2-Phenylethanol						4.8	0.97	
					3-Methylbutyl acetate						4.4	0.96	
					Ethyl lactate						4.5	0.96	
					Ethyl octanoate						3.7	0.94	
					Diethyl succinate						4.9	0.97	
					Diethyl malate						4.1	0.95	
[35]	Nicolet 6700 1000-2500 nm	Darjeeling black tea	SNV, 1D, 2D	PLS-DA, SIMCA	Classification	SNV + 2D + PLS-DA							95.45
[38]	Nicolet IS50 1000-2500 nm	Chicken hamburgers	SNV, MSC, SG, BC	PLS, iPLS, iSPA-PLS	Determination of fat content	SG (1D) + PLS	0.74	0.90	3.73	2.33	2.31		

[66]	InfraXact 570-1850 nm	Walnut kernel	SNV, DT, MSC, WMSC, IMSC	MPLS		Moisture	DT						0.965		
						Protein	SNV +DT						0.967		
						Fat	MSC						0.979		
[30]	XDS 400-2500 nm	Paprika powder	EMSC, PCA	PLS-DA, LDA, QDA		Classification	PLS-DA							0.95	
[41]	Felix F-750 310-1135 nm	Pear	SG, SNV	PLSR, iPLS		Moisture	iPLS				0.52		0.87		
						Soluble solids content					0.57		0.76		
[14]	Nicolet 460 1000-2380 nm	Camellia oil	MSC, SNV, SG, Norris	PLSR, DA		Classification	1D+MSC+Norris							96.7	
						Rapeseed oil	1D+MSC+Norris			1.45	2.50		0.9968		
						Corn oil	1D+MSC+Norris			1.01	3.30		0.9992		
						Sunflower oil	SNV+SG			6.79	4.98		0.9956		
[65]	Ocean Optics 650-1100nm	Milk	SG, SNV, MSC, 2D	PLSR	Raw milk	Fat	SG	0.97		0.18		3.6			
						Protein	SG + MSC	0.85		0.16		1.9			
						Lactose	SG + MSC	0.78		0.11		1.6			
						Total solids	SG	0.96		0.28		3.5			
					Homogenized milk	Fat	SG + SNV	0.99		0.11		6.9			
						Protein	SG + SNV	0.90		0.13		2.3			
						Lactose	SG + MSC	0.78		0.11		1.6			
						Total solids	SG	0.96		0.28		3.5			
[63]	P-TF1 570-1031 nm	Cassava tuber	SNV, BO, MSC, 1D, 2D, Normalization	PLSR		Starch content	2D+PLSR			1.93	2.21	1.6	0.73		
[18]	NIRQuest 900-1700 nm	‘Fuji’ apple	SG, SNV, MSC, 1D, 2D	MLR, PLSR, GRNN, ELM		Sugar content	SG+MSC+PLSR	0.952	0.956	0.653	0.57	3.366			
[47]	NIRQuest 860-2500 nm	Sugar beet juice	SNV, MA, AN	PLSR		Pol	AN			2.40	2.96		0.969		
						Brix	AN			2.40	2.29		0.984		
						Sucrose	SNV			3.41	4.75		0.921		
						pH	AN			0.35	0.67		0.671		
[64]	TanirPro 908-1695 nm	Pistachio	PCA	PLSR		Adult. of Green pea	PCA+ PLSR			6.37	7.87		0.9961		
						Adult. of Spinach				5.43	4.69		0.9968		
[75]	Luminar 5030 1100-2300 nm	Pine nut	1D, SNV, SG, MSC	PLS-DA, iPLS-DA		Classification	iPLS-DA							96	98

[42]	RESONON 900-1700 nm	Wheat	1D, BC	LDA, PLSR	Ergosterol	1D + PLSR	0.89	
					DON	1d + PLSR	0.73	
					Classification	BC + LDA		83.3

RMSEC, root mean squared error of calibration; RMSEP, root mean squared error of prediction; R_c, correlation coefficient of calibration; R_p, correlation coefficient of prediction; RPD, the ratio of prediction to deviation; R^2, the coefficient of determination; SPA, successive projection algorithm; CARS, competitive adaptive reweighted sampling; MLR, multiple linear regression; PCA, principle component analysis; CWT, continuous wavelet transform; BC, baseline correction; SNV, standard normal variate; MSC, multiplicative scatter correction; SG, Savitzky-Golay; PLSR, partial least squared regression; PCR, principle component regression; SVMR, support vector machine regression; LDA, linear discriminant analysis; PLS-DA, partial least squared-discriminant analysis; KNN, k-nearest neighbors; ELNET, elastic net; GPR, gaussian process regression; MARS, multivariate adaptive regression spline; XGB, extreme gradient boosting; GAM, generalized additive model; RF, random forest; *i*PLS, interval PLS; *i*SPA-PLS, interval selection PLS; 1D, 1st derivative; 2D, 2nd derivative; MC, mean centering.