Contents lists available at ScienceDirect

Current Research in Food Science

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

High-accuracy classification and origin traceability of peanut kernels based on near-infrared (NIR) spectroscopy using Adaboost - Maximum uncertainty linear discriminant analysis

Rui Zhu [a], Xiaohong Wu [b,c,*], Bin Wu [d,**], Jiaxing Gao [e]

[a] Mengxi Honors College, Jiangsu University, Zhenjiang, China
[b] School of Electrical and Information Engineering, Jiangsu University, Zhenjiang, China
[c] High-tech Key Laboratory of Agricultural Equipment and Intelligence of Jiangsu Province, Jiangsu University, Zhenjiang, China
[d] Department of Information Engineering, Chuzhou Polytechnic, Chuzhou, China
[e] School of Automotive and Traffic Engineering, Jiangsu University, Zhenjiang, China

ARTICLE INFO

Handling Editor: Dr. Maria Corradini

Keywords:
Peanut kernel
Near-infrared spectroscopy
Classification
Maximum uncertainty linear discriminant analysis
Adaboost

ABSTRACT

Peanut kernels, known for their high nutritional value and palatability, are classified as nut food. In this study, peanut kernel samples from six distinct cities in Shandong Province, China, were examined to categorize and trace their origins. Near-infrared (NIR) spectra of samples were captured using a portable NIR-M-R2 spectrometer. After the application of Savitzky-Golay (SG) filtering, the classification was attempted using principal component analysis (PCA) plus linear discrimination analysis (LDA). Additionally, maximum uncertainty linear discriminant analysis (MLDA) was applied for comparison. A specific number of eigenvectors could respectively maximize the classification accuracies, 81.48% for PCA + LDA and 76.54% for MLDA. In order to further improve the classification accuracies, Adaboost-MLDA was proposed to develop a stronger classifier. This method, after 18 iterations, achieved remarkable effects, achieving a high accuracy of 95.06%. In a similar vein, the enhancement with preprocessing techniques multiplicative scatter correction (MSC) + SG and standard normal variate (SNV) + SG raised accuracies to 98.77% and 97.53%, respectively. The results of classifying first-order and second-order derivative spectra using Adaboost-MLDA were also described, achieving accuracies near 100%. The experiment demonstrates that integrating Adaboost with NIR spectroscopy offers a highly accurate method for peanut kernel classification, promising for practical applications in food quality control.

1. Introduction

As a globally cultivated economic crop, peanuts are among the principal oilseed crops worldwide (Wadood et al., 2022). In China, peanuts are beloved for their rich nutritional value, comprising 44–56% high-quality fats, 22–30% protein, 9.5%–19% carbohydrates, as well as other substances like dietary fiber, minerals, essential amino acids, unsaturated fatty acids, vitamin E and resveratrol (Asibuo et al., 2018; Norlia et al., 2019). Peanut kernels deliver both macronutrients and micronutrients to the human body (Shokunbi et al., 2012). In light of this, peanuts have been shown to offer protective benefits against various health conditions, such as complications of diabetes (Liu et al., 2019), cancer (Amba et al., 2019), cognitive impairments (de Camargo et al., 2017), and cardiovascular diseases (Jafari Azad et al., 2020). Additionally, they possess anti-aging properties, not only reducing cholesterol but also maintaining smooth skin, hence peanuts are often referred to as “long-life nuts” (de Oliveira Sousa et al., 2011; Yao, 2004). Introduced to China during the late Tang dynasty, peanuts have been cultivated for over 400 years, and this made China one of the largest peanut producers in the world (Yang et al., 2020) as well as an indispensable exporter (Wang et al., 2021), accounting for over 40% of the global peanut trade (Wu et al., 2016). In 2018, the peanut cultivation area in China reached 4.62 million hectares, with a production of 17.33 million tons (National Bureau of Statistics of China, 2019), a significant share compared to the global cultivation area of 22.67 million hectares and an annual production of 35 million tons in 2020 (Yang et al., 2020). China has four main peanut-producing regions: the Southeast Coast, the Yangtze River region, the Yellow River region, and the Northeast region

* Corresponding author. School of Electrical and Information Engineering, Jiangsu University, Xuefu Road No. 301, Zhenjiang, 212013, China.
** Corresponding author. Department of Information Engineering, Chuzhou Polytechnic, Fengle Street No. 2188, Chuzhou, 239000, China.
E-mail addresses: wxh419@ujs.edu.cn (X. Wu), wubin@chzc.edu.cn (B. Wu).

https://doi.org/10.1016/j.crfs.2024.100766
Received 8 January 2024; Received in revised form 18 April 2024; Accepted 7 May 2024
Available online 8 May 2024

(Zhang et al., 2017), with Shandong Province in the Yellow River region being particularly prominent. Shandong, a major producer of peanuts and peanut oil in China, contributes about one-third of the country's total peanut oil output (Dong et al., 2023). The high value of peanuts hinges on food safety. Studies by Lu et al. indicate that aflatoxins and heavy metal content are primary limiting factors for peanut safety (Lu et al., 2013). Peanut kernels from different regions exhibit significant variations in quality and nutritional value, leading to their suitability for diverse application fields. However, the sale of counterfeit and inferior peanut kernels by certain merchants, who resort to illicit practices to deceive consumers for substantial profits, is profoundly unethical (Pan et al., 2024). Therefore, establishing effective traceability techniques for peanuts is crucial for recalling non-compliant products and minimizing economic losses, also safeguarding the interests of consumers from a commercial standpoint (Deniz et al., 2018).

Near-infrared (NIR) spectroscopy is an electromagnetic spectrum located between the mid-infrared and infrared regions of the visible spectrum. This powerful analytical technique is widely used in traceability studies to assess and analyze the molecular composition of materials through their interaction with NIR light (Jiang et al., 2021; Cheng et al., 2022). It offers a small absorption coefficient, fast measurement, and a large dynamic range of sample thicknesses (Zareef et al., 2021; Sun et al., 2024). NIR spectroscopy has the potential to replace or supplement traditional methods (Guo et al., 2020; Cheng et al., 2023). In NIR analysis, the interaction of a sample with incident light causes absorption at specific wavelengths corresponding to molecular vibrations, such as overtone and combination vibrations of fundamental modes. These absorption bands provide a unique fingerprint-like spectrum for each material, allowing for both qualitative and quantitative analyses without extensive sample preparation. In peanuts, the internal structure affects the band positions on the spectrum, with regions carrying information about specific molecular concentrations (Deniz et al., 2018). On this basis, Li et al. were able to quantify the content of peanut aflatoxin B_1 using NIR spectroscopy (Li et al., 2023). Variations in the spectral characteristics of different peanut classes arise from their unique internal structures, formed by varying geological and climatic conditions (temperature, rainfall, sunlight, etc.) of their geographical origins. These differences result in variations in the concentrations of chemical constituents like fats and proteins, providing a reliable basis for tracing the origins of peanuts (Holaday and Pearson, 1974; Zhao et al., 2013).

Spectroscopy analysis is widely used in food classification and traceability (Wu et al., 2022; Chen et al., 2023). Wang et al. applied Fourier transform infrared (FT-IR) spectroscopy for peanut identification in Shandong, China. They proved stepwise linear discriminant analysis (SLDA) methods are feasible for this purpose (Wang et al., 2021). Zhang et al. used a portable NIR spectrometer for milk origin identification, achieving high accuracy with fuzzy unrelated discriminant transformation (FUDT) (Zhang et al., 2022). Chen et al. identified the origins of roasted green tea using transform near-infrared (FT-NIR) spectroscopy and supervised techniques (Chen et al., 2009). Long et al. combined NIR spectroscopy with nanocomposites for the origin identification of lilies, showing promise for food and medicine authentication (Long et al., 2022). Chen et al. used NIR spectroscopy for ginseng origin identification, demonstrating the effectiveness of the random subspace ensemble (RSE) algorithm (Chen et al., 2024). Uríčková and Sádecká, 2015. determined the origins of alcoholic beverages using ultraviolet, visible, and infrared spectroscopy, highlighting the importance of spectral range and recognition methods (Uríčková and Sádecká, 2015).

This paper focuses on NIR spectroscopy analysis, combining Adaboost with maximum uncertainty linear discriminant analysis (MLDA) for classifying peanut kernels from different cities in Shandong, China, to trace their origins. Adaboost-maximum uncertainty linear discriminant analysis (Adaboost-MLDA) fundamentally entails ensemble learning of data, generating stronger classifiers through continuous training and voting, thereby improving classification accuracy. This study demonstrates the significance of Adaboost-MLDA in the classification of peanut kernels incorporation with NIR spectroscopy.

2. Data acquisition and preprocessing

2.1. Sample collection

486 peanut kernel samples were selected, containing six different classes for experimentation. These samples, all sourced from Shandong Province, a key peanut-producing area in China, were evenly distributed among the six classes. Each class consisted of 81 samples from each region in Shandong Province, specifically Qingdao (QD), Yantai (YT), Heze (HZ), Jinan (JN), Linyi (LY), and Weihai (WH). The samples were obtained fresh and raw through local distributors. For each class, 81 samples were meticulously chosen based on uniform size, good color, and smooth surface, ensuring their freshness was preserved for the experiments.

2.2. Apparatus

In this study, a portable near-infrared spectrometer NIR-M-R2 (Pynect, Shenzhen, China) was applied to obtain spectral data of peanut kernel samples. The device has a wavelength range of 900–1700 nm (11,100 to 5880 cm^{-1}), a signal-to-noise ratio of 6000:1, a slit size of 1.8 × 0.025 mm, and a detector material of 1 mm uncooled InGaAs. Its optical resolution is typically 10 nm, with a maximum of 12 nm. Wavelength accuracy is typically ±1 nm and ranges up to ±2 nm.

2.3. Spectral acquisition

Before the experiment, all peanut kernel samples and the spectrometer were placed in the laboratory for over 24 h to ensure that the environmental conditions of the samples were consistent with those of the instrument, thereby minimizing the impact of temperature and humidity on the spectral measurements of the peanut seeds. Each sample was carefully cleaned to remove dust, and samples that were shriveled or had obvious defects were discarded. The samples selected for measurement were further required to be similar in size and regular in shape.

Spectral measurements for each class were conducted using the NIR-M-R2 spectrometer. Before spectral scanning, the NIR spectrometer was preheated for 30 min. The spectrometer was set at a distance of 1–2 mm from the samples. Once fixed, spectral data were acquired using the Column scan configuration (spectral wavelength range of 900–1700 nm, digital resolution of 228, and an average of 6 scans per sample) to determine the absorption rates of the peanut kernels. Then NIR spectra were measured at equidistant intervals along the equatorial region of the peanut kernels. Therefore, for each sample, three NIR spectra were acquired, and their average value was taken as the spectrum of that sample. For each class of peanut kernels, 81 samples were measured, resulting in a total of 486 NIR spectra.

2.4. Denoising

To enhance the smoothness of the acquired peanut kernel spectra, the Savitzky-Golay (SG) filter was employed to diminish noise in the spectral data (Savitzky and Golay, 1964). Primarily used for filtering noise from chemical spectrometry data, the SG filter operated on local polynomial regression, a principle grounded in Weierstrass's Theorem (Sury, 2011). The SG filter, a finite impulse response kernel, convolved with the data to approximate the polynomial for the selected filter parameter set (Menon and Seelamantula, 2014). In simpler terms, the SG filter employed the least squares method to fit local data segments and used the fitted function to estimate the value of each data point, thereby achieving smoothing.

The SG filtering algorithm has rapid computational speed and effectively eliminates high-frequency noise, demonstrating strong adaptability to nonlinear signals. However, a notable limitation of this algorithm lies in the significant impact of the chosen parameters for the fitting polynomial on the filtering outcomes.

Beyond the solitary application of SG filtering, this study investigated a hybrid technique that combined multiplicative scatter correction (MSC) with SG, referred to as MSC + SG. MSC adjusted the spectrum of each sample via linear regression, reducing dataset variability and enhancing the signal-to-noise ratio to some extent (Shen et al., 2021). Similarly, standard normal variate (SNV) was also employed in conjunction with SG, known as SNV + SG. SNV standardized sample data by first subtracting the mean and then dividing by the standard deviation. Mapping data to the origin, followed by comparing the spectra of samples on a uniform scale, could enhance the accuracy and reliability of further analyses significantly.

3. Methods of analysis

3.1. PCA + LDA

Principal component analysis (PCA), an unsupervised technique, is typically customary to favor the principal components that exhibit a significant contribution to the spectrum. As the magnitude of contribution increases, a greater amount of component information can be retained (Wu et al., 2020). Linear discriminant analysis (LDA) is a supervised learning approach. It efficiently classifies data by projecting it into a lower-dimensional space when the number of spectral variables meets or exceeds the number of samples (Lasalvia et al., 2022).

Overall, LDA demonstrates superior classification performance on data compared to PCA. To address the limitations of LDA, especially the challenge of small sample size problem encountered with high-dimensional data (Chen et al., 2000), the solution is a combination of PCA and LDA: applying LDA on the data pre-processed by PCA. This approach applies LDA to the PCA scores, thereby enhancing the classification accuracy (Lasalvia et al., 2022).

Generally, in the PCA stage, spectral data undergo preliminary dimension reduction; during the LDA stage, test data can be mapped to corresponding discriminant vectors using discriminant information extracted from the training data, achieving a second dimension reduction (Wu et al., 2017).

3.2. Maximum uncertainty linear discriminant analysis

Due to the significantly greater estimation errors of non-dominant or small eigenvalues compared to dominant or large eigenvalues (Pudil et al., 1990) and the impact of PCA principal component selection on classification accuracy as well as the quantity of useful discriminative information, Thomaz et al. proposed the MLDA method to address these issues (Thomaz et al., 2004). The distinguished feature of MLDA is its approach, which enlarges the less reliable smaller eigenvalues of the pooled covariance matrix in LDA, while it maintains the majority of the larger eigenvalues unchanged (Thomaz et al., 2006). MLDA mitigates the instability inherent in LDA effectively, leveraging the maximum entropy selection method to stabilize the scaling of the identity matrix within the within-class covariance matrices (Wu et al., 2023).

Given a dataset containing n samples, which are divided into c classes, let $X = \{x_1, x_2, ..., x_n\} \in R^d$ denote the mean of all samples, $\overline{x}$ represent the mean of the samples in class i, and signify the number of samples in class i. The implementation process of the MLDA algorithm is as follows (Thomaz et al., 2006; Wu et al., 2023):

(1)Calculate S_w and S_b according to the aforementioned formula; (2) Compute $S_p = S_w/(n- c)$, its eigenvalue diagonal matrix Λ, and eigenvector matrix φ; (3) Determine the eigenvalues λ of S_p and the average eigenvalue $\overline{\lambda}$; (4) For eigenvalues λ in Λ, which are also smaller than $\overline{\lambda}$, replace them with $\overline{\lambda}$ to generate a new eigenvalue diagonal matrix Λ^*; (5) Calculate the within-class scatter matrix $S_w{}^* = S_p{}^*(n- c) = (\varphi\Lambda^*\varphi^T)(n- c)$, and finally acquire the discriminant transformation vectors by $S_w^*{}^{-1}S_b$.

MLDA is a validated method. Initially, it computes $c-1$ discriminant transformation vectors from the training samples, onto which the test samples are projected in the next stage. Based on this, classification is performed using a K-nearest neighbors (KNN) algorithm. The accuracy of this classification serves to evaluate the effectiveness of the validation.

3.3. Adaboost–MLDA

Adaboost, an acronym for Adaptive Boosting, is an ensemble method that transforms multiple weak classifiers into a singular strong classifier. Its fundamental principle involves iterative weight adjustments for samples, particularly focusing on those misclassified samples in previous rounds. This iterative process trains new weak classifiers, each time prioritizing previously misclassified samples, and eventually combines them into a stronger, more predictive classifier. AdaBoost excels in handling the high-dimensional data, minimizing overfitting, and showcasing robust performance in diverse classification tasks.

In this context, the integration of Adaboost.M1 with MLDA, forming Adaboost-MLDA, is proposed to develop an effective strong classifier. This combination initiates with equal weight distribution across all samples. The MLDA classifier then uses this data to compute classification errors and adjust weights accordingly. Misclassified samples receive increased weights, while those correctly classified samples get reduced weights. This iterative weight adjustment process, focusing on error-prone samples, leads to the creation of several MLDA classifiers. The effectiveness of each classifier is enhanced by its ability to concentrate on samples that were previously misclassified. In the final combination of the Adaboost-MLDA classifier, the influence of each MLDA classifier within the ensemble is determined by its classification accuracy. This approach ensures that classifiers with the higher accuracy have a more significant impact on the overall result.

3.4. KNN algorithm

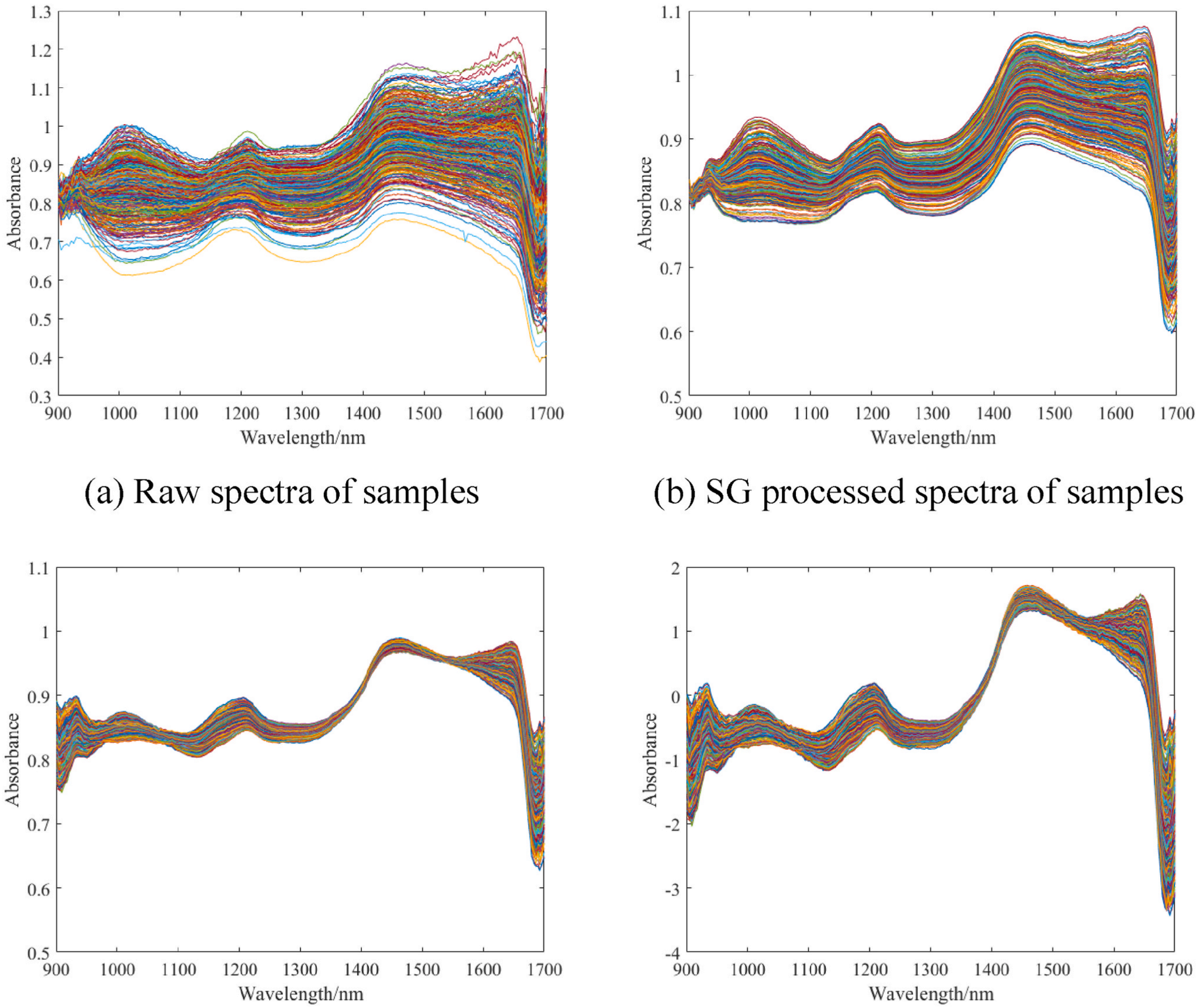
The KNN algorithm is a commonly used supervised learning method for regression and classification tasks. In essence, its principle involves first calculating the spatial distances between the data to be classified and the known data, then identifying the k data points closest in this space, referred to as neighbors. Classification of the data in question is based on the majority class among these neighbors. It is noteworthy that the outcome of the classification is influenced by the value of k, and selecting an appropriate k is a crucial prerequisite for promoting classification accuracy.

4. Results and discussion

4.1. Data preprocessing

In this study, NIR spectra were acquired using the NIR-M-R2 spectrometer, which operates in conjunction with the computer software DLP NIRscan Nano. The scanning configuration employed was the inherent Column method in the system, capturing the absorption rates of peanut kernel samples within the wavelength range of 900–1700 nm, with each sample having 228 sampling points. The absorption spectra of all 486 samples are depicted in a single graph, as shown in Fig. 1(a).

The processing and evaluation of all raw data in this study were performed using MATLAB 2021A (The Mathworks). Fig. 1 shows the raw and preprocessed NIR spectra of peanut kernels. The raw peanut kernel spectrogram contains some noise, rendering the spectral curves less smooth and unfavorable for subsequent classification processes. To address this, the SG filter is utilized for noise cancellation, with a polynomial order of 2 and a frame length of 25. The filtered spectrogram



(a) Raw spectra of samples (b) SG processed spectra of samples

(c) MSC+SG processed spectra of samples (d) SNV+SG processed spectra of samples

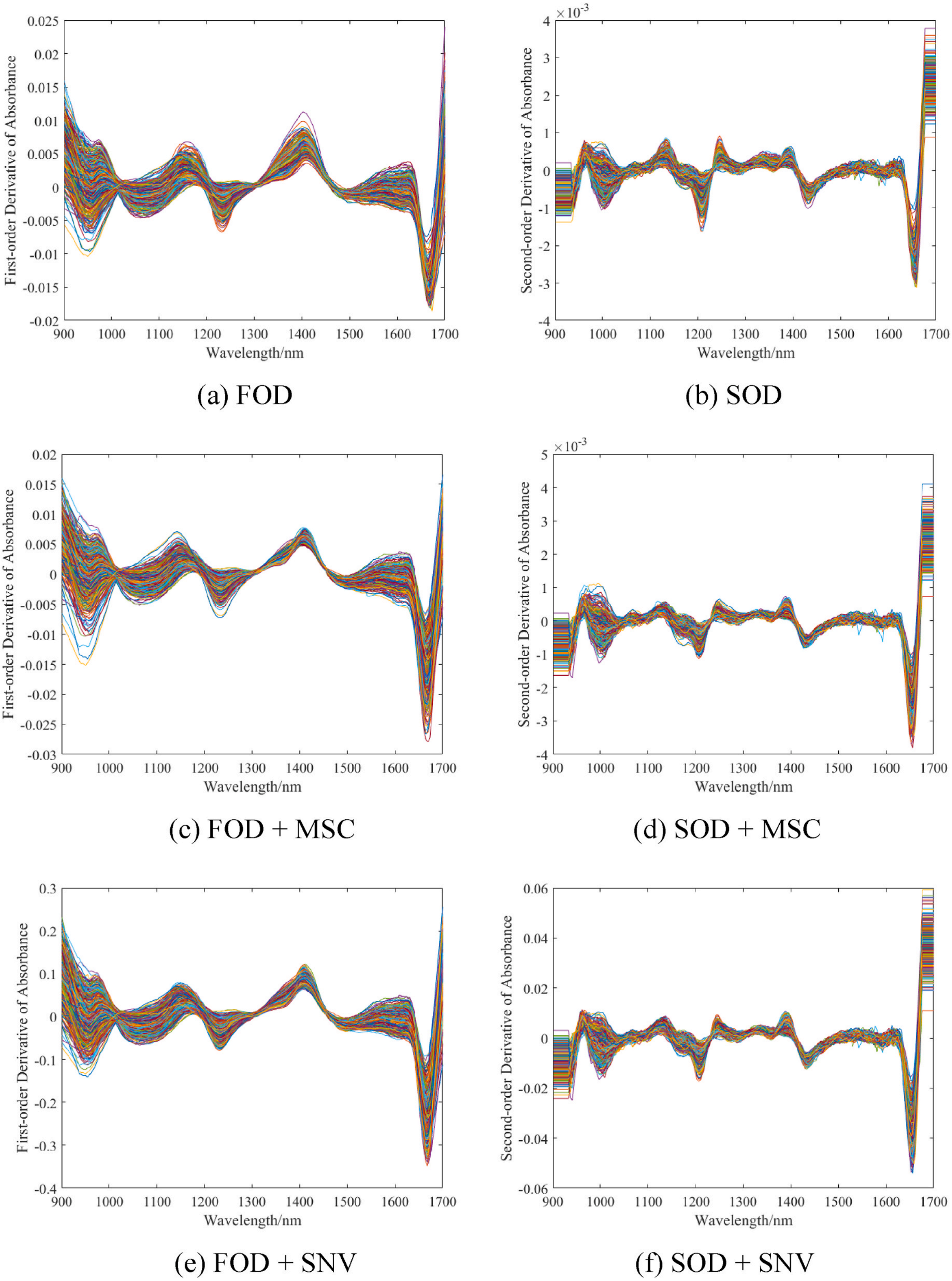
Fig. 1. The raw and preprocessed NIR spectra of peanut kernels.

is presented in Fig. 1(b). It is noticeable that compared to Fig. 1(a), the peanut kernel spectra have become smoother, and the overall arrangement of the spectra demonstrates greater consistency. The spectra on both the upper and lower ends have converged towards the center, becoming more compact, and the few previously extreme data points have disappeared. Fig. 1(c) illustrates the results of MSC + SG preprocessing, where the data appears more concentrated compared to SG alone, especially around the wavelength of 1400 nm. Conversely, spectra at shorter and longer wavelengths are more dispersed. In Fig. 1 (d), the spectra processed with SNV + SG are displayed, exhibiting an overall shape similar to MSC + SG. The primary distinction lies in the vertical axis values being shifted closer to the origin. The merits and drawbacks of these three preprocessing methods remain to be evaluated by subsequent analyses.

To enhance the characteristic peaks and improve the resolution of the spectra, the first-order derivative (FOD) and second-order derivative (SOD) of the raw spectra were calculated, followed by filtering using MSC and SNV. The processed NIR spectra are displayed in Fig. 2. Differentiating the spectra helps to mitigate the effects of interference, and this technique will subsequently be applied to the classification of peanut kernels.

4.2. *Analysis of NIR spectra*

Differences in the spectral behavior of samples may be attributed to their chemical compositions (Ghosh et al., 2016). The NIR spectroscopy measured in experiments reflects the overtones and combinations of molecular chemical bond vibrations, which makes some parts of the spectrum interesting for the spectroscopy of organic materials (Hakkel et al., 2022). Variations in information stemming from differences in hydrogen-containing functional groups (X–H, where X represents C, O, N, S, etc.). A related study indicates that the wavelength of 900 nm is associated with the absorbance of proteins, while 1450 nm corresponds to that of starch. The protein in peanut kernels is noted to have a high amino group content, with its 1st overtone band spanning 1450–1550 nm as well as the 2nd and 3rd overtones distributed within 970–1000 nm. Starch, abundant in hydroxyl groups, corresponds to two spectral ranges: 1410–1480 nm and 920–945 nm, aligning with the mentioned overtone bands (Sundaram et al., 2009). The peanut kernel samples were collected in the 900–1700 nm wavelength range, as illustrated in Fig. 1. Notable visual differences in the spectra near 1000 nm and 1550 nm are identified as arising from overtones. Both amino and hydroxyl groups are likely significant factors in determining spectral distinctions,



(a) FOD

(b) SOD

(c) FOD + MSC

(d) SOD + MSC

(e) FOD + SNV

(f) SOD + SNV

Fig. 2. NIR spectra preprocessed by FOD, SOD, MSC and SNC.

owing to the close alignment of their overtone wavelength ranges with the characteristic bands. These functional groups, along with chemical bond stretching vibrations, are integral to the analytical results produced by spectrometers, where the number of vibrational absorptions correlates with these results. The variations in the quantity of hydrogen-containing functional groups in different peanut kernel classes lead to distinct absorption peaks. Regardless of the extent of differences among samples, NIR spectroscopy remains a viable method for the qualitative identification of peanut kernels (Li et al., 2022).

4.3. PCA + LDA

In this study, PCA was employed to compress the data, selecting ten eigenvectors, thereby reducing the original 228-dimensional space data to a 10-dimensional space. This compression method effectively preserved most of the information of original NIR data, clustering similar or identical data together after projection in a specific direction. Within these ten dimensions, the first three play a predominant role. This is due to the relatively minor values of the subsequent seven dimensions, which hold less significance. Take data preprocessed using SG as an example, the contribution rates of the first three dimensions are 95.90%, 2.55%, and 0.93%, respectively, cumulating a total contribution of 99.38%. In other words, these first three dimensions are the principal factors influencing the spectral data after PCA. The PCA plot, constructed using these three dimensions as axes, is shown in Fig. 3. NIR spectra from peanut kernels originating from the same region are almost exclusively grouped. As illustrated in Fig. 3, peanut kernels pretreated by SG exhibit a more compact aggregation compared to MSC + SG, and SNV + SG pretreatment results in scattered clustering. However, there are still noticeable central positional differences in the spectra of the six peanut classes, likely attributable to inherent varietal differences in the peanuts themselves. We will discuss the merits of each method based on their accuracy later.

In the subsequent phase, LDA is carried out for feature extraction. As a supervised method, it is essential to partition the data into training and testing sets before extraction. The data, after PCA processing, consists of 486 instances, each represented in a 10-dimensional space, with each class comprising 81 samples in total. Within each class, two-thirds, equating to 54 samples, are allocated for training, while the remaining 27 were set aside for testing. This generates a total of 324 training samples and 162 testing samples. Choosing 5 discriminant vectors, the samples are trained to yield the 5-dimensional LDA data.

At this point, the PCA + LDA preprocessing was completed. Finally, the KNN algorithm was applied for data classification with the chosen value of k being 1. After voting, all the results were computed and stored in Table 1, where the classification accuracy under SG is 62.96%. As a matter of fact, these are still not particularly impressive results, prompting the exploration of an alternative method to enhance classification accuracy.

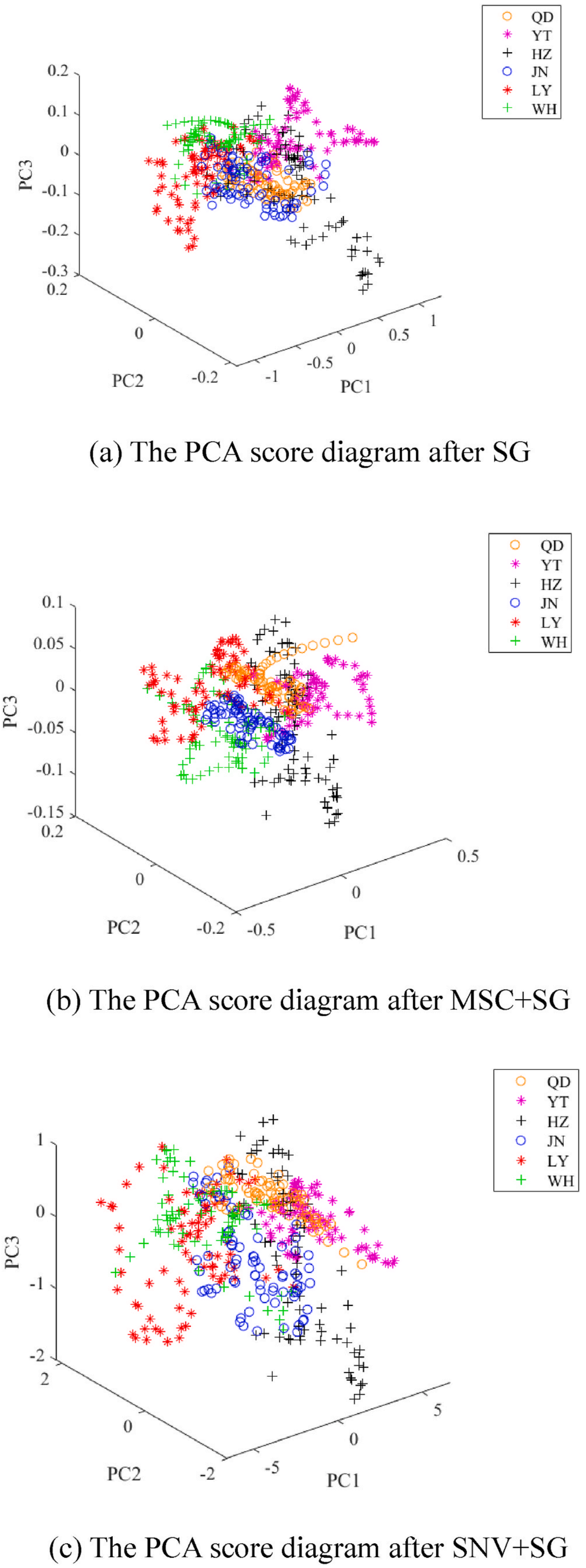
4.4. MLDA

To address the small sample size problem encountered by classical LDA, this study utilizes MLDA, an improved version of LDA. MLDA, based on the Fisher criterion, modifies the within-class scatter matrix with scalar processing. This modification aims to eliminate the need for matrix inversion and guarantees the consistent existence of the matrix in various conditions. Furthermore, the algorithm treats scalars as the basic units for processing various objects, effectively reducing computational load and enhancing operational efficiency.

The MLDA algorithm first projects the training samples to determine the optimal projection direction, which is also where the test samples will be projected. The projected test samples are then classified in the KNN classifier along with the training samples to obtain the classification results. The workflow of MLDA is illustrated in Fig. 4.

The data processing methodology initially employed the PCA approach, followed by scalar processing of the within-class scatter matrix S_w based on LDA, and finally applied KNN classification. After SG preprocessing, the final calculation yielded a classification accuracy of 70.99% for MLDA. Compared to the accuracy of 62.96% utilizing PCA + LDA, MLDA represents a significant improvement of 8.03%, demonstrating that MLDA is capable of raising the classification accuracy of LDA.

This result is only for the case when the number of eigenvectors is set to 10. To analyze the impact of the number of eigenvectors on the classification accuracy of PCA + LDA and MLDA, each number within



(a) The PCA score diagram after SG

(b) The PCA score diagram after MSC+SG

(c) The PCA score diagram after SNV+SG

Fig. 3. PCA score plot of six classes of peanut kernels based on SG, MSC + SG and SNV + SG preprocessing.

Table 1
The final classification accuracies of PCA + LDA, MLDA and Adaboost-MLDA based on SG, MSC + SG and SNV + SG (%).

	PCA + LDA	MLDA	Adaboost-MLDA
SG	62.96	70.99	95.06
MSC + SG	64.81	48.15	98.77
SNV + SG	74.69	59.26	97.53

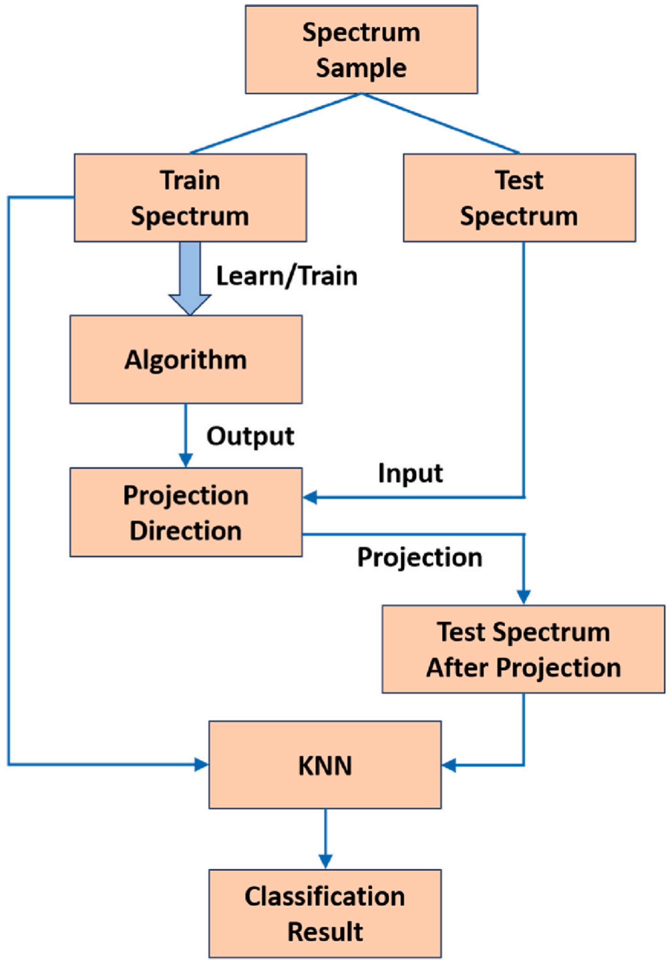


Fig. 4. Algorithm flowchart of MLDA used to classify the NIR spectra.

the range [10, 228] was selected as the number of eigenvectors to calculate the accuracies of both methods. Additionally, the average accuracies of the two methods were calculated for comparison. As shown in Fig. 5, PCA + LDA exhibits higher classification accuracy within fewer eigenvectors. However, its accuracy decreases with more eigenvectors, fluctuating around an average of 63.85%. The maximum accuracy of PCA + LDA, 81.48%, occurs at eigenvector numbers 32, 33 and 40, a remarkable difference of 17.63% from the average. On the contrary, the accuracy of the MLDA method generally increases with the number of eigenvectors, though fluctuations still occur. The highest accuracy for MLDA, approximately 76.54%, occurs in the interval [139,167], with only a 2.68% difference from its average accuracy of 73.86%. Given the minor impact of the number of eigenvectors on the classification accuracy of MLDA, it is reasonable to slightly sacrifice MLDA accuracy to enhance that of PCA + LDA. At this point, selecting 40 eigenvectors is deemed optimal in this context, as it allows the classification accuracy of MLDA to peak at 81.48%. In contrast, PCA + LDA reaches only 70.37% accuracy, which is 6.17% below its maximum and 3.49% below its average, yet still within an acceptable range. This selection reflects a strategic balance between accuracy and computational efficiency in the classification process. Furthermore, the chart indicates that MLDA improves classification accuracy primarily when the number of eigenvectors is either low or high.

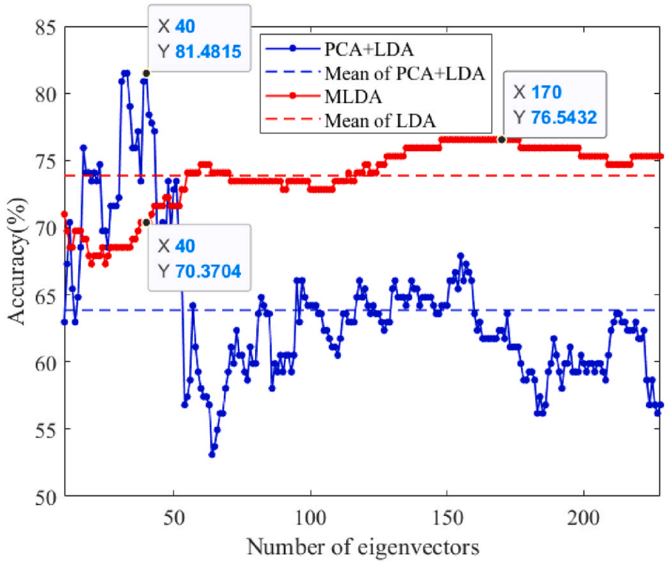


Fig. 5. Comparison of classification accuracy of PCA + LDA and MLDA utilizing SG preprocessing with changes in the number of eigenvectors.

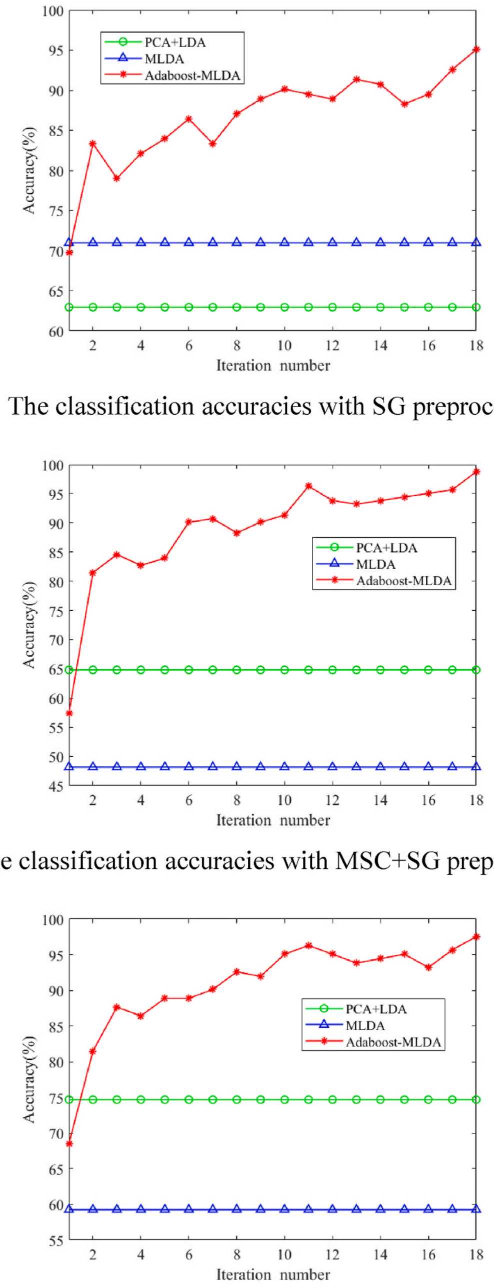
Despite these considerations, in this experiment, the classification accuracies of MLDA for the peanut kernel spectrum are still not particularly high, which might be attributed to certain limitations of MLDA. For instance, MLDA tends to have a slightly lower classification accuracy compared to PCA + LDA, and its algorithmic performance is more easily influenced by the dimensionality of features (Liu and Wang, 2010). On account of this, a more accurate classification approach, Adaboost-MLDA, has been developed.

4.5. *Adaboost–MLDA*

Adaboost can integrate weak classifiers into a more robust classifier, with each iteration potentially improving classification accuracy. In this experiment, each class was assigned 54 training data and 27 test data. To draw a comparison with Adaboost-MLDA, PCA + LDA and MLDA were executed to evaluate their classification accuracies. For the Adaboost-MLDA, the number of iterations was set to 18. Following data input and iteration, the computed results are presented in Table 1 and Fig. 6.

The classification accuracy of the Adaboost-MLDA method generally exhibits an upward trend with an increase in the number of iterations. Focusing on SG, although the accuracy of Adaboost-MLDA initially stands at only 69.75%, positioning it between PCA + LDA and MLDA, it gradually increases and then stabilizes. After 18 iterations, the classification accuracy of Adaboost-MLDA reaches 95.06%, significantly surpassing PCA + LDA and MLDA. Similar conclusions can be obtained for data preprocessed by MSC + SG and SNV + SG. This is attributable to the iterative training function of Adaboost, which is an adaptive feature extraction process. The classification errors of each round of weak classifiers contribute to generating the optimal feature selection for the training model. Therefore, Adaboost can improve classification accuracy substantially, providing valuable insights for research into more accurate discriminant analysis methods.

As illustrated in Fig. 6, For MLDA, The classification accuracy of SNV + SG was higher than that of MSC + SG, and SG achieved the highest accuracy. Different from Fig. 6(b) and (c)–. (a) lies in its higher accuracy of MLDA compared to PCA + LDA. This could be attributed to the selection of an optimal feature dimension specifically for MLDA. While the relative merits of SG, MSC + SG, and SNV + SG were not readily



(a) The classification accuracies with SG preprocessing

(b) The classification accuracies with MSC+SG preprocessing

(c) The classification accuracies with SNV+SG preprocessing

Fig. 6. The classification accuracies of Adaboost-MLDA, PCA + LDA and MLDA with SG, MSC + SG and SNV + SG.

definable, it was important to note that under the enhancements of Adaboost-MLDA, all three preprocessing methods achieved higher accuracies.

Similarly, for the spectra in Fig. 2 that underwent derivation and filtering, we classified them using the same method described above, with the accuracy presented in Table 2. As indicated in Table 2, although the classification capabilities of PCA + LDA and MLDA are not significantly prominent, Adaboost-MLDA is notably exceptional. After 18 iterations, its accuracy approaches 100%, making it the most effective classification method. This finding is consistent with the conclusions of the previous experiment.

To sum up, in this experiment, MLDA demonstrates relatively high classification accuracy when the number of eigenvectors is either low or high, surpassing the accuracy of PCA + LDA under similar conditions. However, the classification accuracy of MLDA alone is still not optimal. Combining Adaboost with MLDA significantly improves the classification accuracy of MLDA, making Adaboost-MLDA the most effective classification method in this study.

Table 2
The final classification accuracies of PCA + LDA, MLDA and Adaboost-MLDA based on NIR spectra via FOD and SOD (%).

	FOD	FOD + MSC	FOD + SNV	SOD	SOD + MSC	SOD + SNV
PCA + LDA	30.25	35.19	29.01	27.78	29.01	25.31
MLDA	32.10	30.86	32.72	29.63	28.40	37.65
Adaboost-MLDA	95.06	96.30	95.68	92.59	99.38	96.91

Adaboost-MLDA significantly enhances classification accuracy, achieving an accuracy of 100%, thereby establishing itself as an effective method for peanut kernel classification. For any given test sample, its spectrum can be measured and compared with known peanut spectra, ultimately determining the most probable classification or origin of the sample. This straightforward approach is advantageous in combating the sale of counterfeit and substandard peanuts in the market, thus safeguarding consumer rights.

5. Conclusion

Adaboost-MLDA, integrating Adaboost with MLDA, creates a stronger classifier, markedly improving the classification accuracy of MLDA through adaptive feature extraction. While PCA + LDA sometimes shows low accuracy, MLDA enhances it, yet not optimally. Therefore, Adaboost-MLDA is crucial, initially boosting the performance of the classifier before classification. Experimental results support that Adaboost significantly elevates classification accuracy, addressing the need for a more efficient classification method in scenarios where standard techniques like PCA + LDA or MLDA show poor performance.

The exceptional classification accuracy of Adaboost-MLDA in identifying peanut spectra is a key aspect of its utility. For new peanut kernel samples, measuring and comparing their NIR spectra with known NIR spectra is feasible. Adaboost-MLDA, used in classifying these NIR spectra, identifies the most likely class of the peanut kernels effectively. This approach plays a pivotal role in tracing their origins, guaranteeing the quality of peanuts, safeguarding consumer rights and eliminating market disorder.

CRediT authorship contribution statement

Rui Zhu: Investigation, Methodology, Software, Writing – original draft. **Xiaohong Wu:** Conceptualization, Methodology, Project administration, Supervision, Writing – review & editing. **Bin Wu:** Funding acquisition, Writing – review & editing, Validation, Software. **Jiaxing Gao:** Investigation, Resources, Formal analysis, Visualization, All authors have read and agreed to the published version of the manuscript.

Declaration of competing interest

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

Data availability

Data will be made available on request.

Acknowledgment

This research was supported by the Major Natural Science Research

Projects of Colleges and Universities in Anhui Province (2022AH040333), the Youth and Middle-aged Teachers Cultivation Action Project in Anhui Province (JNFX2023136) and the Undergraduate Innovation and Entrepreneurship Training Program of Jiangsu Province (202310299096Y).

References

Amba, V., Murphy, G., Etemadi, A., Wang, S., Abnet, C.C., Hashemian, M., 2019. Nut and peanut butter consumption and mortality in the National institutes of health-AARP diet and health study. Nutrition 11, 1508.

Asibuo, J.Y., Forpoh, A.S., Akromah, R., 2018. Genotype X envionment interactions of groundnut (Arachis hypogaea L.) for pod yield. Ecol. Genet. Genom. 7, 27–32.

Chen, H., Tan, C., Lin, Z., 2024. Geographical origin identification of ginseng using near-infrared spectroscopy coupled with subspace-based ensemble classifiers. Spectrochim. Acta Part A Mol. Biomol. Spectrosc. 304, 123315.

Chen, L.F., Liao, H.Y.M., Ko, M.T., Lin, J.C., Yu, G.J., 2000. A new LDA-based face recognition system which can solve the small sample size problem. Pattern Recogn. 33, 1713–1726.

Chen, Q., Zhao, J., Lin, H., 2009. Study on discrimination of Roast green tea (Camellia sinensis L.) according to geographical origin by FT-NIR spectroscopy and supervised pattern recognition. Spectrochim. Acta Part A Mol. Biomol. Spectrosc. 72, 845–850.

Chen, Y., Guo, Y.Z., Wang, W., Wu, X.H., Jia, H.W., Wu, B., 2023. Clustering analysis of FTIR spectra using fuzzy K-Harmonic-Kohonen clustering network. Spectrosc. Spectr. Anal. 43, 268–272.

Cheng, J., Sun, J., Yao, K., Dai, C., 2023. Generalized and hetero two-dimensional correlation analysis of hyperspectral imaging combined with three-dimensional convolutional neural network for evaluating lipid oxidation in pork. Food Control 153, 109940.

Cheng, J., Sun, J., Yao, K., Xu, M., Tian, Y., Dai, C., 2022. A decision fusion method based on hyperspectral imaging and electronic nose techniques for moisture content prediction in frozen-thawed pork. Lebensm. Wiss. Technol. 165, 113778.

de Camargo, A.C., Regitano-d'Arce, M.A.B., Rasera, G.B., Canniatti-Brazaca, S.G., do Prado-Silva, L., Alvarenga, V.O., Sant'Ana, A.S., Shahidi, F., 2017. Phenolic acids and flavonoids of peanut by-products: antioxidant capacity and antimicrobial effects. Food Chem. 237, 538–544.

de Oliveira Sousa, A.G., Fernandes, D.C., Alves, A.M., De Freitas, J.B., Naves, M.M.V., 2011. Nutritional quality and protein value of exotic almonds and nut from the Brazilian Savanna compared to peanut. Food Res. Int. 44, 2319–2325.

Deniz, E., Altuntaş, E.G., Ayhan, B., İğci, N., Demiralp, D.Ö., Candoğan, K., 2018. Differentiation of beef mixtures adulterated with chicken or Turkey meat using FTIR spectroscopy. J. Food Process. Preserv. 42, e13767.

Dong, Y., Wang, L., Cai, D., Zhang, C., Zhao, S., 2023. Risk assessment on dietary exposure to aflatoxin B1, heavy metals and phthalates in peanuts, a case study of Shandong province, China. J. Food Compos. Anal. 120, 105359.

Ghosh, S., Mishra, P., Mohamad, S.N.H., de Santos, R.M., Iglesias, B.D., Elorza, P.B., 2016. Discrimination of peanuts from bulk cereals and nuts by near infrared reflectance spectroscopy. Biosyst. Eng. 151, 178–186.

Guo, Z., Barimah, A.O., Shujat, A., Zhang, Z., Ouyang, Q., Shi, J., El-Seedi, H.R., Zou, X., Chen, Q., 2020. Simultaneous quantification of active constituents and antioxidant capability of green tea using NIR spectroscopy coupled with swarm intelligence algorithm. Lebensm. Wiss. Technol. 129, 109510.

Hakkel, K.D., Petruzzella, M., Ou, F., van Klinken, A., Pagliano, F., Liu, T., van Veldhoven, R.P.J., Fiore, A., 2022. Integrated near-infrared spectral sensing. Nat. Commun. 13, 103.

Holaday, C.E., Pearson, J.L., 1974. Effects of genotype and production area on fatty-acid composition, total oil and total protein in peanuts. J. Food Sci. 39, 1206–1209.

Jafari Azad, B., Daneshzad, E., Azadbakht, L., 2020. Peanut and cardiovascular disease risk factors: a systematic review and meta-analysis. Crit. Rev. Food Sci. Nutr. 60, 1123–1140.

Jiang, H., He, Y., Chen, Q., 2021. Determination of acid value during edible oil storage using a portable NIR spectroscopy system combined with variable selection algorithms based on an MPA-based strategy. J. Sci. Food Agric. 101 (8), 3328–3335.

Lasalvia, M., Capozzi, V., Perna, G., 2022. A comparison of PCA-LDA and PLS-DA techniques for classification of vibrational spectra. Appl. Sci. 12, 5345.

Li, J., Deng, J., Bai, X.Monteiro, Monteiro, D.D.N., Jiang, H., 2023. Quantitative analysis of aflatoxin B1 of peanut by optimized support vector machine models based on near-infrared spectral features. Spectrochim. Acta 303, 123208.

Li, Q., Wu, X., Zheng, J., Wu, B., Jian, H., Sun, C., Tang, Y., 2022. Determination of pork meat storage time using near-infrared spectroscopy combined with fuzzy clustering algorithms. Foods 11, 2101.

Liu, Z.B., Wang, S.T., 2010. Modified linear discriminant analysis method MLDA. Comput. Sci. 37, 239–242.

Liu, G., Guasch-Ferré, M., Hu, Y., Li, Y., Hu, F.B., Rimm, E.B., Manson, J.E., Rexrode, K. M., Sun, Q., 2019. Nut consumption in relation to cardiovascular disease incidence and mortality among patients with diabetes mellitus. Circ. Res. 124, 920–929.

Long, W., Hu, Z., Wei, L., Chen, H., Liu, T., Wang, S., Guan, Y., Yang, X., Yang, J., Fu, H., 2022. Accurate identification of the geographical origins of lily using near-infrared spectroscopy combined with carbon dot-tetramethoxyporphyrin nanocomposite and chemometrics. Spectrochim. Acta Part A Mol. Biomol. Spectrosc. 271, 120932.

Lu, Z., Zhang, Z., Su, Y., Liu, C., Shi, G., 2013. Cultivar variation in morphological response of peanut roots to cadmium stress and its relation to cadmium accumulation. Ecotoxicol. Environ. Saf. 91, 147–155.

Menon, S.V., Seelamantula, C.S., 2014. Robust savitzky-golay filters. In: 2014 International Conference on Digital Signal Processing (DSP), pp. 688–693.

National Bureau of Statistics of China, 2019. China Statistical Yearbook. China Statistics Press, Beijing.

Norlia, M., Jinap, S., Nor-Khaizura, M.A.R., Radu, S., Samsudin, N.I.P., Azri, F.A., 2019. Aspergillus section Flavi and aflatoxins: occurrence, detection, and identification in raw peanuts and peanut-based products along the supply chain. Front. Microbiol. 10, 2602.

Pan, W., Liu, W., Huang, X., 2024. Rapid identification of the geographical origin of Baimudan tea using a Multi-AdaBoost model integrated with Raman Spectroscopy. Curr. Res. Food Sci. 8, 100654.

Pudil, P., Somol, P., Haindl, M., 1990. Introduction to Statistical Pattern Recognition. Academic, San Diego, pp. 303–304.

Savitzky, A., Golay, M.J., 1964. Smoothing and differentiation of data by simplified least squares procedures. Anal. Chem. 36, 1627–1639.

Shen, Y., Wu, X., Wu, B., Tan, Y., Liu, J., 2021. Qualitative analysis of lambda-cyhalothrin on Chinese cabbage using mid-infrared spectroscopy combined with fuzzy feature extraction algorithms. Agric. For. 11, 275.

Shokunbi, O.S., Fayomi, E.T., Sonuga, O.S., Tayo, G.O., 2012. Nutrient composition of five varieties of commonly consumed Nigerian groundnut (Arachis hypogaea L.). Grasas Aceites 63, 14–18.

Sun, J., Yang, F., Cheng, J., Wang, S., Fu, L., 2024. Nondestructive identification of soybean protein in minced chicken meat based on hyperspectral imaging and VGG16-SVM. J. Food Compos. Anal. 125, 105713.

Sundaram, J., Kandala, C.V., Butts, C.L., 2009. Application of near infrared spectroscopy to peanut grading and quality analysis: overview. Sens. Instrum. Food Qual. Saf. 3, 156–164.

Sury, B., 2011. Weierstrass's theorem—leaving no 'Stone'unturned. Reson 16, 341–355.

Thomaz, C.E., Gillies, D.F., Feitosa, R.Q., 2004. A new covariance estimate for Bayesian classifiers in biometric recognition. IEEE Trans. Circ. Syst. Video Technol. 14, 214–223.

Thomaz, C.E., Kitani, E.C., Gillies, D.F., 2006. A maximum uncertainty LDA-based approach for limited sample size problems—with application to face recognition. J. Braz. Comput. Soc. 12, 7–18.

Uríčková, V., Sádecká, J., 2015. Determination of geographical origin of alcoholic beverages using ultraviolet, visible and infrared spectroscopy: a review. Spectrochim. Acta Part A Mol. Biomol. Spectrosc. 148, 131–137.

Wadood, S.A., Nie, J., Li, C., Rogers, K.M., Zhang, Y., Yuan, Y., 2022. Geographical origin classification of peanuts and processed fractions using stable isotopes. Food Chem. X, 16.

Wang, L., Yang, Q., Zhao, H., 2021. Sub-regional identification of peanuts from Shandong Province of China based on Fourier transform infrared (FT-IR) spectroscopy. Food Control 124, 107879.

Wu, B., Shen, J., Wang, X., Wu, X., Hou, X., 2022. NIR spectral classification of lettuce using principal component analysis sort and fuzzy linear discriminant analysis. Spectrosc. Spectr. Anal. 42, 3079–3083.

Wu, L., Ding, X., Li, P., Du, X., Zhou, H., Bai, Y., Zhang, L., 2016. Aflatoxin contamination of peanuts at harvest in China from 2010 to 2013 and its relationship with climatic conditions. Food Control 60, 117–123.

Wu, X., Fu, H., Tian, X., Wu, B., Sun, J., 2017. Prediction of pork storage time using Fourier transform near infrared spectroscopy and Adaboost-ULDA. J. Food Process. Eng. 40.

Wu, X., He, F., Wu, B., Zeng, S., He, C., 2023. Accurate classification of chunmee tea grade using NIR spectroscopy and fuzzy maximum uncertainty linear discriminant analysis. Foods 12, 541.

Wu, X., Zhu, J., Wu, B., Huang, D., Sun, J., Dai, C., 2020. Classification of Chinese vinegar varieties using electronic nose and fuzzy Foley–Sammon transformation. J. Food Sci. Technol. 57, 1310–1319.

Yang, B., Zhang, C., Zhang, X., Wang, G., Li, L., Geng, H., Liu, Y., Nie, C., 2020. Survey of aflatoxin B1 and heavy metal contamination in peanut and peanut soil in China during 2017–2018. Food Control 118, 107372.

Yao, G., 2004. Peanut production and utilization in the People's Republic of China. Peanut Local Glob. Food Syst. Ser. Rep. 4.

Zareef, M., Arslan, M., Hassan, M.M., Ahmad, W., Ali, S., Li, H., Ouyang, Q., Wu, X., Hashim, M.M., Chen, Q., 2021. Recent advances in assessing qualitative and quantitative aspects of cereals using nondestructive techniques: a review. Trends Food Sci. Technol. 116, 815–828.

Zhang, C., Selvaraj, J.N., Yang, Q., Liu, Y., 2017. A survey of aflatoxin-producing Aspergillus sp. from peanut field soils in four agroecological zones of China. Toxins 9, 40.

Zhang, T., Wu, X., Wu, B., Dai, C., Fu, H., 2022. Rapid authentication of the geographical origin of milk using portable near-infrared spectrometer and fuzzy uncorrelated discriminant transformation. J. Food Process. Eng. 45.

Zhao, H., Guo, B., Wei, Y., Zhang, B., 2013. Near infrared reflectance spectroscopy for determination of the geographical origin of wheat. Food Chem. 138, 1902–1907.