Rapid classification of peanut varieties for their processing into peanut butters based on near-infrared spectroscopy combined with machine learning

Journal of Food Composition and Analysis

Yu, Hongwei; Erasmus, Sara W.; Wang, Qiang; Liu, Hongzhi; Ruth, Saskia M.

https://doi.org/10.1016/j.jfca.2023.105348

Journal of Food Composition and Analysis 120 (2023) 105348

Contents lists available at ScienceDirect

Journal of Food Composition and Analysis

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

Rapid classification of peanut varieties for their processing into peanut butters based on near-infrared spectroscopy combined with machine learning

Hongwei Yu [a,b], Sara W. Erasmus [b], Qiang Wang [a,*], Hongzhi Liu [a,*], Saskia M. van Ruth [b]

[a] Institute of Food Science and Technology, Chinese Academy of Agricultural Sciences/Key Laboratory of Agro-Products Processing, Ministry of Agriculture and Rural Affairs, P.O. Box 5109, Beijing 100193, China
[b] Food Quality and Design, Wageningen University and Research, P.O. Box 17, 6700 AA Wageningen, The Netherlands

ARTICLE INFO

Keywords:
Cluster analysis
Efficient processing
Near-infrared spectroscopy
Peanut butters
Random forest
Support vector machine

ABSTRACT

Peanut classification based on processing purposes is becoming mainstream. In order to speed up the classification procedure, near-infrared (NIR) spectroscopy for classifying peanut varieties for their processing into peanut butters was assessed for the first time. Peanut varieties were primarily classified by principal component analysis (PCA) combined with cluster analysis based on the structural characteristics (texture and rheology) and roast characteristics (colour and volatile compounds) of the resulting peanut butters. After the completion of spectral collection and subsequent spectral pre-treatments, the performances of classification models built by partial least squares discriminant analysis, support vector machine, and random forest were compared. PCA, variable importance, and random forest selection by filter were investigated as feature extraction methods. The sensitivity, specificity, and accuracy of the filtered cross validation and external validation models were all over 90%, while the kernel density estimation presented the acceptable distribution results of categories probabilities in the selected models. These results showed that NIR spectroscopy combined with machine learning methods is a promising approach to provide a reliable evaluation of peanuts for efficient processing.

1. Introduction

Peanuts are one of the paramount nuts in the world with the total production in 2019 equalling 48.8 million tons (Faostat, 2020). Peanuts can be consumed as either raw materials or as processed products, such as roasted peanut, peanut oil, and peanut butter, to satisfy consumer preferences and nutritional requirements (Wang, 2018). Generally, peanuts could be divided based on variety or the size and appearance of peanut kernels. In contrast, product-oriented classification can better insure the stable characteristics of final products such as peanut butter and enhance their market values (Gong et al., 2018). The industrial scale manufacture of peanut butter is fully matured through roasting and grinding of peanuts (Wang, 2016). Peanuts impact the structural and roast characteristics of their resulting peanut butter given that different peanut varieties have different chemical compositions (Dhamsaniya et al., 2012). Peanut butter with improved L* and rheological qualities can be produced by using the varieties with higher levels of tyrosine and threonine content (Yu et al., 2021). Some researchers classified peanut varieties based on the processing suitability linked with the quality traits of peanut products (Wang et al., 2017). Typically, the characteristics of peanut butter are evaluated based on its structural characteristics (texture and rheology) and roast characteristics (colour and volatile compounds) through time-consuming and costly laboratory analyses (Yu et al., 2021). Hence, the development of speedy, precise, and stable methods is vital to select suitable peanuts for processing to match the growing need for high-quality peanut butter production.

Currently, near-infrared (NIR) spectroscopy based on the vibration of hydrogen-containing molecules has been systematically applied for the quality evaluation of agricultural products. On one hand, it has obvious advantages as it can rapidly collect and analyse spectral data without laborious preparations and it is also an environmentally and economically friendly method without chemical waste and high expenditure compared with the chemical methods. On the other hand, it initially needs stable calibration models built through representative sample collection, spectral and reference data analysis, and model establishment. Previously, abundant studies have been carried out to analyse the

* Corresponding authors.
E-mail addresses: wangqiang06@caas.cn (Q. Wang), lhz0416@126.com (H. Liu).

https://doi.org/10.1016/j.jfca.2023.105348
Received 24 November 2022; Received in revised form 19 March 2023; Accepted 16 April 2023
Available online 17 April 2023

qualities of peanuts, and satisfactory results were obtained for fat (Yu et al., 2016), protein and protein subunits (Zhao et al., 2021), amino acids (Wang et al., 2013), sucrose (Yu et al., 2020a), and fatty acids (Yu et al., 2020b). Meanwhile, NIR spectroscopy was also used to grade peanuts (Sundaram et al., 2009) and determine their maturity (Windham et al., 2010). Evidently, NIR spectroscopy can be used to evaluate the properties of peanuts, and the quality traits of peanut butter are related to the peanuts used. Therefore, it can be hypothesised that the spectral data of peanuts could be applied to reflect the characteristics of peanut butters at least under the decided preparation process. However, the challenge remains in accurately predicting the characteristics of peanut butters using only the spectra data of corresponding raw materials. In recent years machine learning algorithms such as random forest (RF) and support vector machine (SVM) have been developed to obtain better classification and regression models because of their ability to deal with complex systems and multi-variables (Monforte et al., 2021; Phan and Tomasino, 2021). Therefore, machine learning has the huge potential to analyse the spectral data of peanuts and their relationship with peanut butters.

The aim of this study was to develop a rapid and robust method for sorting peanut varieties for efficient process of peanut butter using NIR spectroscopy combined with machine learning. A low-cost, reliable, and efficient method is optimal for the assessment of raw materials to produce products with high quality. Hence, a suitable approach for this study was to firstly conduct a cluster analysis of peanut butters based on the characteristics analysis results to achieve scientific grouping of the corresponding peanut varieties. Secondly, classification models of peanuts were built combined with spectral data through chemometrics and machine learning. To build better performance classification models, partial least squares discriminant analysis (PLS-DA), SVM, and RF as modelling algorithms were compared in this study based on the different pre-treatment spectral data. Principal component analysis (PCA), variable importance (VarImp) (Kuhn, 2008), and random forest selection by filter (RFSBF) (Kuhn and Johnson, 2013) were also conducted and compared for feature extraction to establish simplified and stable models. The sensitivity, specificity, and accuracy were used to assess the performances of all models, while the kernel density estimation (KDE) assessed the distributions of class probabilities. The flowchart of the analysis procedure is presented in Fig. 1.

2. Materials and methods

2.1. Peanut and peanut butter

A total of 40 peanut varieties are listed in Table S1, which includes the main-planting varieties and high oleic acid varieties. For each variety, about 5 kg of unshelled peanut samples were expressed from different provinces to our lab and stored at commercial 4 ℃ cold storage (Yuandong Co., Ltd., Tianjin, China). The peanut butters were prepared according to a general process (Yu et al., 2021). Concisely, approximately 0.5 kg plump-shelled peanut kernels per variety were placed on a steel bakeware and roasted at 160 ℃ for 30 min in an electric oven with the top and bottom heating mode. The peeled roasted peanuts were then ground in a colloid grinder (Langfang Tongyong Machinery Co., Ltd., Hebei, China). The initial 10 min grinding was done to roughly grind all peanut kernels, and the subsequent 30 s fine grinding was done to achieve 0.4 kg 100% pure peanut butter per variety. Peanut butter per variety was separately stored in a glass bottle at room temperature, out of direct sunlight.

2.2. Spectral data collection

A benchtop spectrometer with a rotating sphere (Bruker Scientific Instruments, Karlsruhe, Germany) was used to acquire the spectral data. About 100 g kernels were placed in a ring cup (9.7 cm diameter and 4.5 cm depth) for each sample measurement of which each spectrum was the average of 32 scans. Each variety was subsampled five times, producing a total of 200 (40 × 5) spectra for the subsequent analysis. The reflectance spectra were recorded by an indium gallium arsenide (InGaAs) detector where the wavelength ranged from 12489.49 cm^{-1} to 3996.02 cm^{-1}. Spectral acquisition and conversion were conducted using the OPUS 7.5 software (Bruker Scientific Instruments, Karlsruhe,

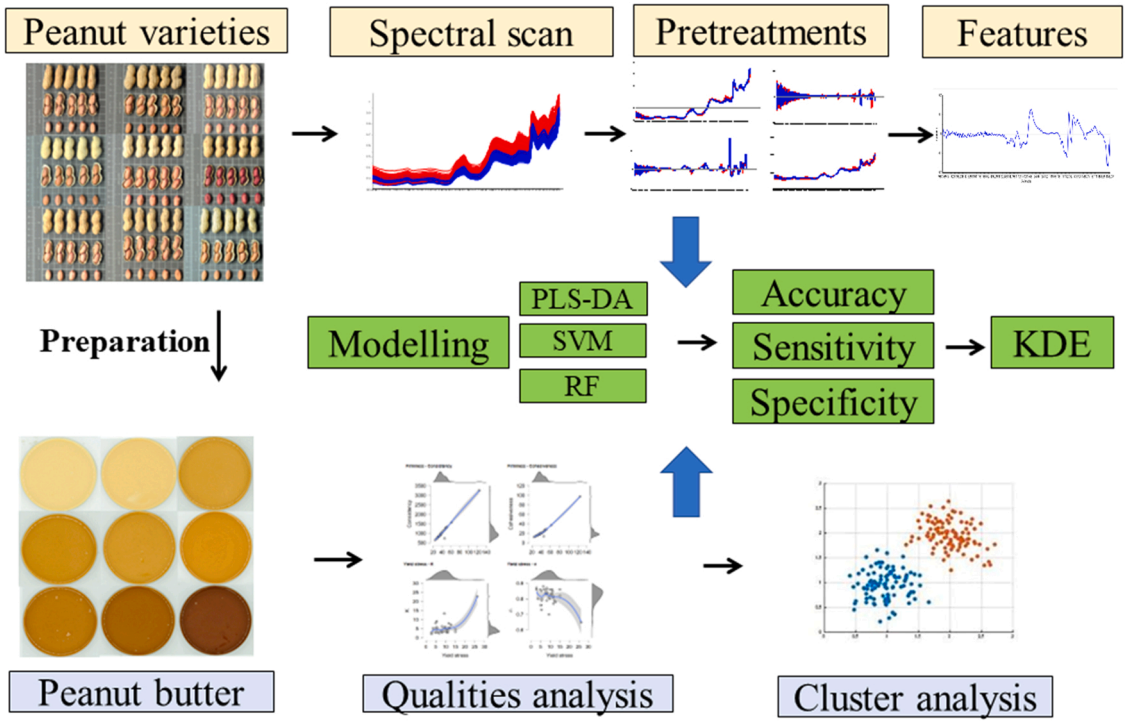

Fig. 1. The flowchart of this research. (KDE) Kernel density estimation; (PLS-DA) Partial least squares discriminant analysis; (RF) Random forest; (SVM) Support vector machine.

Germany).

2.3. Physicochemical characteristics of peanut butters

2.3.1. Texture

Peanut butter texture was described using firmness (g). A TA-XTplus texture analyser (Micro Stable System Co., UK) with the back-extrusion penetration model was used to measure the texture characteristics. The analyser was coupled with a long roller and a 35 mm diameter compression disc. Each sample was put into the cylindrical jar (50 mm internal diameter) with the same volume (75%). The trigger pressure of penetration was 5 g, and the penetration depth was 30 mm with a speed of 1 mm s^{-1}. Samples were analysed in duplicate to measure the force as firmness and then averaged to obtain the mean value per sample.

2.3.2. Rheology

An HR-2 rheometer (TA Instruments, New Castle, USA) was applied to evaluate the rheological characteristics of the peanut butters. Approximately 2 g samples were positioned on the crosshatched plate (40 mm diameter) with 1 mm gap geometry and subsequently assessed. After 2 min of equilibration, steady-state detection was conducted at a shear rate range (1–300 s^{-1}) for the shear rate sweep test. The Herschel-Bulkley's model was applied to model the flow behaviour of the peanut butters (Ahmed and Ramaswamy, 2006). The yield stress as an important parameter of the model means the lowest shear stress required to trigger peanut butter to flow, which was confirmed for the next analysis. The storage modulus (G') and loss modulus (G'') were confirmed by dynamic oscillatory experiments. All experiments were performed inside the linear viscoelastic range determined at the 1 Hz frequency. The results were collected in the frequency range of 0.1–100 Hz. As indicated in Eq. (1) and Eq. (2), the G' and G'' were fitted by power function law equations of oscillatory frequency (ω), expressing the viscoelastic characteristics of peanut butter.

$$G' = a \times \omega^{b} \tag{1}$$

$$G'' = c \times \omega^{d} \tag{2}$$

Where a (kPa Hz^{-b}) and c (kPa Hz^{-d}) demonstrate the quantity of G′ and G″ correspondingly at a specific frequency, and b and d (×100) indicate the slope of the connections between the modulus and frequency (Liu et al., 2019; Resch and Daubert, 2002). Samples were analysed in duplicate and then averaged to obtain the mean value per sample.

2.3.3. Colour

Peanut butter colour was confirmed using a CS-600 colour spectrophotometer (CHNSpec Technology Co., Ltd, Hangzhou, China). For each colour assessment, 2 g samples of each peanut butter were put in a circular quartz cell and subsequently the L*(darkness at 0 to lightness at 100), a*(greenness at −128 to redness at 127), and b*(blueness at −128 to yellowness at 127) colour values were determined. Triplicate measurements were performed, and the mean value was calculated for each peanut butter sample.

2.3.4. Volatile compounds

The volatile compounds measurement of the peanut butter samples was based on headspace solid-phase microextraction gas chromatography mass spectrometry (HS-SPME-GS-MS). The whole analysis protocol could be found in our previous study (Yu et al., 2021). Briefly, samples were prepared by weighing 5 g of each peanut butter into a 20 mL glass vial, and 50 μL internal standard 1,2,3-trichloropropane (0.5 mg mL^{-1} in methanol, Sinopharm Chemical Reagent Co., Ltd., Beijing, China) was added to each sample vial for the concentration calculation. Each vial was sealed with a Teflon diaphragm and an aluminium lid. The samples were put in shaking incubators at 55 °C for 10 min pre-equilibrium after which the SPME fibre was introduced to the headspace for 40 min. The absorbed volatiles were transferred in the hot injection section (260 °C) for 150 s desorption. The splitless mode was applied for the GS-MS analysis with helium as the carrier gas and a flow rate of 1 mL min^{-1}. The analyser's temperature was set at 250 °C, while the oven temperature programme was initiated at 40 °C for 5 min and subsequently raise at a rate of 5 °C min^{-1} to 250 °C with a holding time of 5 min. Mass spectra were obtained using the electron impact ionization mode (70 eV) in the mass range of 35–500 m z^{-1}. By comparing the data to the mass spectral library and calculating the retention indices (RI) (Yu et al., 2021), volatile compounds were recognized. The calculation of RI, as shown in the Eq. (3), relies on the n-alkane standard (C7-C40) (0.5 mg mL^{-1} in n-hexane, Smart Solutions, North Charleston, American) as the reference.

$$RI_x = 100n + 100 \times (tR_x - tR_n)/(tR_{n+1} - tR_n) \tag{3}$$

Where tR indicates the retention time, n indicates the number of atom carbon, and x indicates one of volatile compounds. After that, the effective volatile compounds were confirmed by calculating the ratio of the concentrations / the odour threshold (>1) (Yu et al., 2021). Duplicate measurements were performed, and the mean value was calculated for each peanut butter sample.

2.4. Chemometrics and machine learning

After all the physicochemical characteristics analyses were performed on the peanut butter samples, K-means as unsupervised clustering was applied to acquire the distinct groups as the reference values. Kruskal-Wallis tests were used to assess whether the structural and roast characteristics of the various groups differed significantly. The strength of the connections between the characteristics of peanut butters and the absorbances of the original spectral data was assessed using correlation analysis. Principal component analysis (PCA) was conducted to reduce data dimension to explore the relationships of the peanut butter characteristics. The outliers of peanut varieties were estimated by the Hotelling's T^2 based on the PCA results (Liu et al., 2018).

All spectral datasets were prior pre-processed in various methods to improve the signal-to-noise ratio and uncover more relevant data, including standard normal variable (SNV), the first derivative (FD), the second derivative (SD), normalization, and multiple scatter correlation (MSC). Partial least squares discriminant analysis (PLS-DA), support vector machine (SVM), and random forest (RF) were used for mathematical modelling. The spectral datasets were split using a 4:1 ratio into the training dataset and external validation dataset after deleting outliers. The repetitive spectra of the same varieties were assigned to one of the datasets. The three repeats of 10-fold cross validation as internal validation based on the training dataset were used to prevent training model overfitting. Since there are 1102 spectral variables, some features were extracted by PCA, variable importance (VarImp), and random forest selection by filter (RFSBF) to simplify the models. PCA converted the original spectral data into new orthogonal and non-overlapping principal components (PCs). VarImp scores were generated to determine the feature importance by using the caret package (Kuhn, 2008). The variables (score value >1) were selected as the important spectra. RFSBF in the caret package with 10-fold cross validation tests were applied to select the best feature spectra.

The performances of all classification models were assessed based on the following various parameters: accuracy (ACCU); sensitivity (SENS); and specificity (SPEC). The formulas for calculating these parameters are indicated in Eq. (4), Eq. (5), and Eq. (6).

$$\text{Accuracy} = (\text{Number of true assessments}) / (\text{Number of all assessments}) \times 100\% \tag{4}$$

$$\text{Sensitivity} = (\text{Number of true group 1 assessments}) / (\text{Number of all group 1 assessments}) \times 100\% \tag{5}$$

Specificity = (Number of true group 2 assessments) / (Number of all group 2 assessments) × 100% (6)

A kernel density estimation function (KDE) was used to generate the genuinely positive and negative rate distribution. KDE is a kind of non-parametric distribution assessment that is similar to histograms, but allows for distribution interpolation and modification (Alewijn et al., 2016). All calculations were performed by R 4.0.3 software (R Foundation for Statistical Computing, Austria).

3. Results and discussion

3.1. The cluster results of peanut varieties

Fig. 2a shows the cluster analysis results of peanut butters except for Jihuatian1 based on the textural and rheological characteristics. The first two principal components (PCs) explained near 90% of the total variance and therefore contained most of the relevant information. Group 1 (blue) was in the positive direction of PC1, indicating that it had positive relationships with firmness, field stress, G'-a, and G''-c, while group 2 (red) was in the negative direction of PC1. Correspondingly, group 1 had significantly higher ($P < 0.05$) values for firmness (43.4 ± 8.0 g), yield stress (11.13 ± 1.81 Pa), G'-a (60.94 ± 14.01 kPa Hz^{-b}), and G''-c (48.59 ± 6.99 kPa Hz^{-d}) than group 2 (34.6 ± 5.6 g, 6.12 ± 1.83 Pa, 32.61 ± 9.94 kPa Hz^{-b}, and 24.68 ± 6.34 kPa Hz^{-d}, respectively). Firmness is related to the hardness of peanut butters, while field stress, G'-a, and G''-c present the flow and deformation of peanut butters under stress and strain (Sun and Gunasekaran, 2009). The textural and rheological qualities reflect the structural state of peanut butter which is connected to the sensory attributes evaluated (Shakerardekani et al., 2013). The clustering results clearly show that there were two groups of peanut butters based on the texture and rheological characteristics. Based on the cluster results, the spectral data of raw materials are shown in Fig. 2b. Similarly, the blue lines stood for group 1, while the red lines stood for group 2. Overall, the spectral absorbance values of group 2 were higher than that of group 1. The spectral absorbance values derived from the C-H, O-H group of fat, protein, and sucrose in peanuts (Hourant et al., 2000) which is known to have a great influence on the texture and rheological characteristics of peanut butters (Dhamsaniya et al., 2012; Mohd Rozalli et al., 2015). For example, the first overtone of O-H stretching in fat caused 7067 cm^{-1} absorption and the first overtone of C-H stretching in protein may lead to the absorption from 6173 to 5882 cm^{-1} (Workman and Weyer, 2012). PCA was also used to explore the spectral data of peanut varieties to reduce multicollinearity. The sum of PC1 and PC2 was 98%. Group 1 (blue) was mostly grouped in the right direction of PC1, while group 2 (red) was in the left direction in Fig. 2c, which had the same trend as PCA results in Fig. 2a. Therefore, it shows the potential interrelation between peanut butters and the spectra data of peanut varieties.

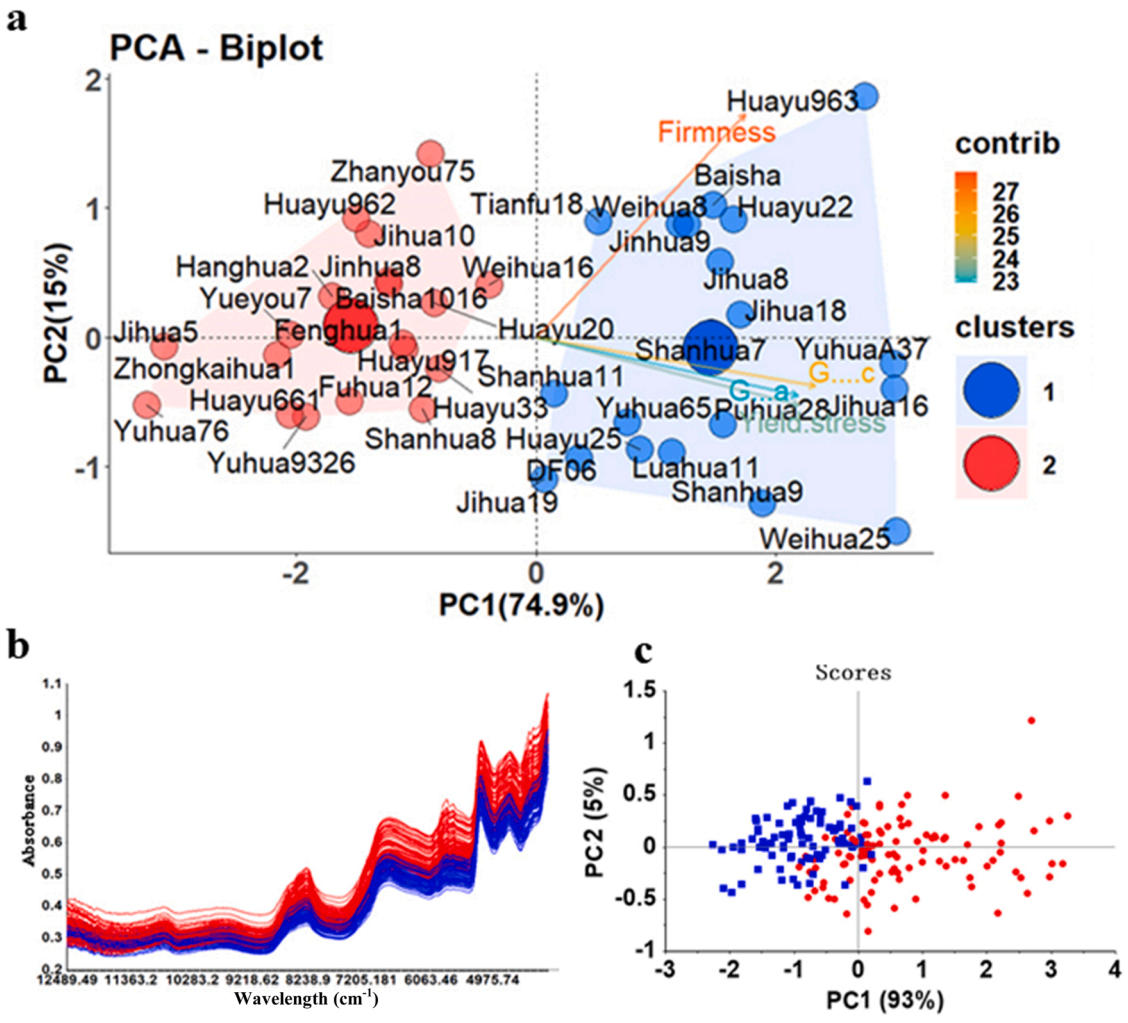

Fig. 2. The principal component analysis (PCA) biplot (a), the raw spectral data of peanut varieties (b), and the PCA score plot of the spectral data (c) for the structure characteristics of peanut butters. Different coloured polygons indicate the different sample clusters and arrows indicate the contributions (contrib) of parameters in a. Vertical axis is absorbance and horizontal axis is the wavelength range from 12489.49 cm^{-1} to 3996.02 cm^{-1} in b. (PC1) Principal component 1; (PC2) Principal component 2.

Fig. 3 presents the strength of the relationship between the characteristics of peanut butter (vertical) and the full wavelength (horizontal). Compared with the results in Fig. 2a, yield stress, G'-a, and G'-c had a similar relationship with the wavelength. Specifically, these three indicators had higher negative correlation coefficient values with longer wavelength ranges (8786.62 cm^{-1} - 3996.02 cm^{-1}) and positive correlation coefficient values with short wavelength ranges (12489.49 cm^{-1} - 11100.92 cm^{-1}). Firmness had positive correlation values with short wavelength ranges (12489.49 cm^{-1} - 11100.92 cm^{-1}). Group 1 had higher these indicators, resulting in lower absorbance in longer wavelength ranges shown in Fig. 2b. Wavelengths in these ranges are likely derived from the 1st overtone region, the 3rd overtone region, and the combination bands of the main chemical compositions (fat, protein, and sucrose) in peanuts (Barbin et al., 2014).

The cluster results of peanut varieties except for zhanyou75, Jinhua8, and tianfu18 based on the roast characteristics including colour and volatile compounds are presented in Fig. 4a. PC1 and PC2 accounted for near 72% of the total variance. Group 1 (green) was located in the negative direction of PC1, where L* made the great contributions, while group 2 (yellow) was located in the positive direction along PC1 where the major contributions were derived from most of the roast characteristics such as pyrazines, 2-acetylpyrrole, and a*. There were also significant differences ($P < 0.05$) in the roast characteristics between group 1 and group 2. Specifically, group 2 had higher 2,5-dimethylpyrazine (6.19 ± 1.97 mg kg^{-1}), 2,3,5-trimethylpyrazine (2.65 ± 1.01 mg kg^{-1}), 3-ethyl-2–5-dimethylpyrazine (1.79 ± 0.68 mg kg^{-1}), furaneol (1.68 ± 0.72 mg kg^{-1}), 2-acetylpyrrole (1.15 ± 0.30 mg kg^{-1}), 2-methoxy-4-vinylphenol (2.54 ± 0.82 mg kg^{-1}) than group 1 (3.90 ± 1.30 mg kg^{-1}, 1.66 ± 0.71 mg kg^{-1}, 1.00 ± 0.36 mg kg^{-1}, 0.86 ± 0.49 mg kg^{-1}, 0.68 ± 0.31 mg kg^{-1}, and 1.53 ± 0.45 mg kg^{-1}, respectively). These volatile compounds, especially pyrazines with nutty and roast aromas, are known as some of the primary volatiles of peanut products (Baker et al., 2003; Li and Hou, 2018). The L* value (51.40 ± 4.68) of group 1 was lower than that of group 2 (57.04 ± 3.39 and 8.16 ± 2.06), while the a* value (10.94 ± 2.25) of group 1 was higher than that of group 2 (8.16 ± 2.06). Therefore, the manufacturers can select peanut varieties from group 2 to manufacture peanut butters with rich flavour and bright colour.

The spectra of different groups based on the roast characteristics cluster analysis can be seen in Fig. 4b. The green lines (group 1) and the yellow lines (group 2) overlapped on the full wavelength range. The PCA results (Fig. 4c) show more distinct relationships between the two groups. The total of PC1 and PC2 accounted for 98% of the total variance in spectral data. Group 1 and group 2 were intertwined and there was no clear separation between them. The main reason for the incomplete separation is because Fig. 4a presents that some varieties of group 1 (e.g. Huayu917, Shanhua11, and Yuhua9326) and group 2 (e.g. Jihua8, Jihua19, Baisha, and Jihuatian1) are close, which has a negative effect on the classification. Meanwhile, the roast characteristics of peanut butter are determined by raw materials and the roasting process, while the spectral data of raw materials only show partial information. Therefore, it cannot fully reflect the roast characteristics of peanut butters. The correlation heatmap offered more detailed information about the relationship between the spectral data of peanut varieties and the roast characteristics of the resulting peanut butters. It was found that a* value and 2-methoxy-4-vinylphenol had significantly positive correlations (correlation coefficient values > 0.15, $P < 0.05$) with most of the whole wavelength. This is because a* value stands for the redness mechanically linked with the infra-red band where the NIR wavelength exists. There were also significant negative correlations (correlation coefficient values < -0.15, $P < 0.05$) between 2,5-dimethylpyrazine and 3-ethyl-2–5-dimethylpyrazine and most of longer wavelength ranges (11332.34 cm^{-1} – 5315.17 cm^{-1}). This could imply that the spectral data of raw materials can reflect the characteristics of peanut butters.

3.2. The classification models built based on the full wavelength range

The PLS-DA, SVM, and RF were used to establish classification models for the discrimination of peanut varieties except for Jihuatian 1 with different structural characteristics based on different pre-treatment spectral datasets. The results are shown in Table 1. Overall, all training and cross validation models for structural characteristics had good results. The SENS, SPEC, and ACCU of training models were over 97%,

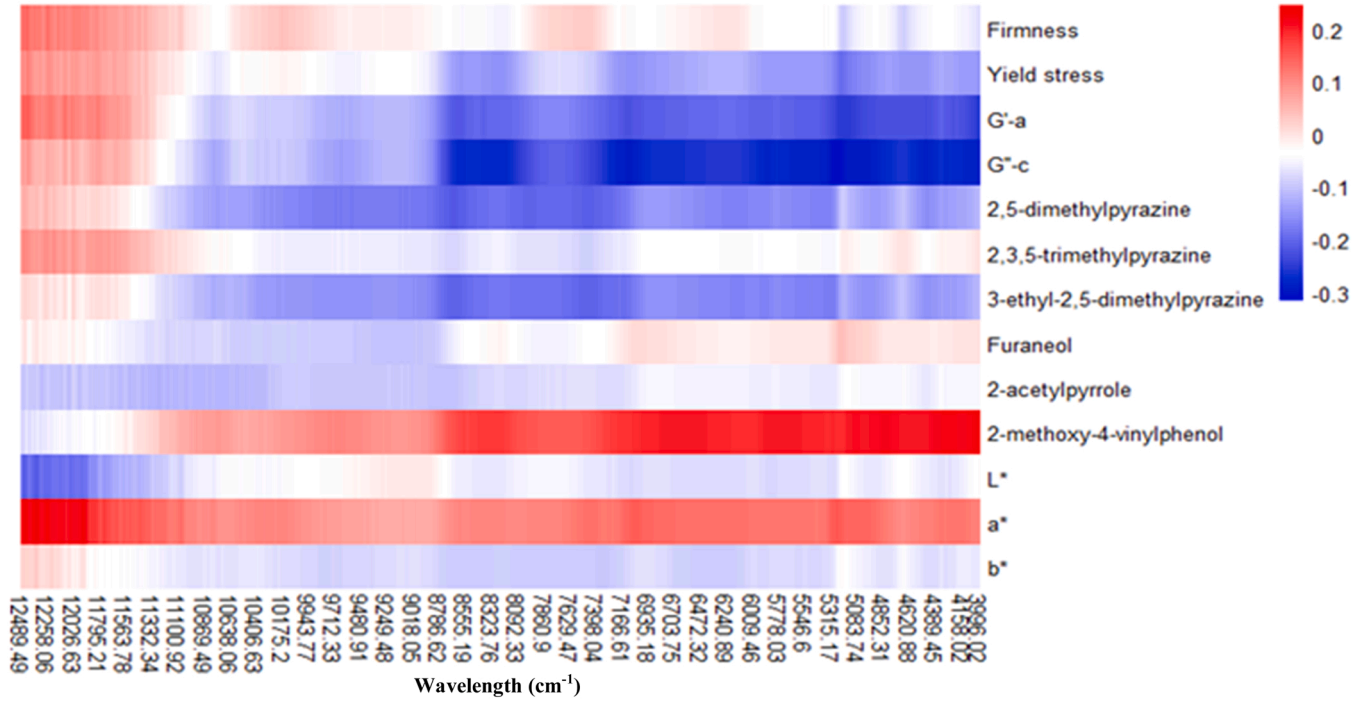

Fig. 3. The correlation analysis between raw spectral data of peanut varieties (vertical axis) and characteristics of peanut butters (horizontal axis). Positive and negative coefficients are coloured red and blue, respectively. Correlation coefficients < -0.15 and > 0.15 indicate significant correlations ($P < 0.05$).

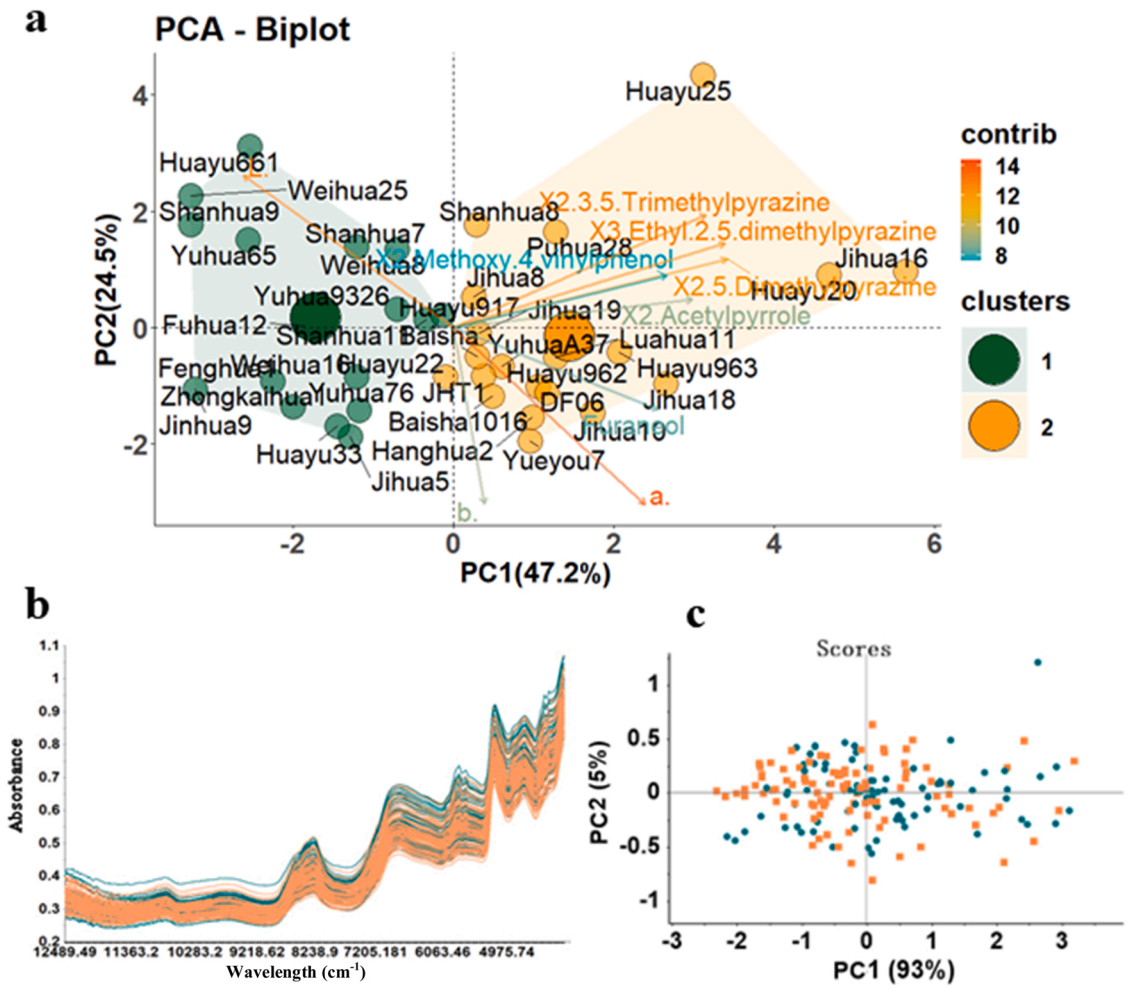

Fig. 4. The principal component analysis (PCA) biplot (a), the raw spectral data of peanut varieties (b), and the PCA score plot of the spectral data of peanut varieties (c) for the roast characteristics of peanut butters. Different coloured polygons indicate the different sample clusters and arrows indicate the contributions (contrib) of parameters in a. Vertical axis is absorbance and horizontal axis is the wavelength range from 12489.49 cm^{-1} to 3996.02 cm^{-1} in b. (PC1) Principal component 1; (PC2) Principal component 2.

Table 1
The sensitivity, specificity, and accuracy results for the discrimination of peanut varieties based on the structure characteristics of peanut butters combined with the different spectral algorithm and pre-treatment methods.

Algorithms	Pre-treatments	Training			Cross validation			External validation		
		SENS (%)	SPEC (%)	ACCU (%)	SENS (%)	SPEC (%)	ACCU (%)	SENS (%)	SPEC (%)	ACCU (%)
PLS-DA	Original	100	98	99	87	89	88	70	50	60
	SNV	100	100	100	92	93	93	75	65	70
	FD	100	100	100	89	94	91	80	90	85
	SD	100	100	100	83	86	84	70	75	73
	MSC	98	100	99	95	89	92	70	95	83
	Normalization	99	100	99	95	93	94	70	85	78
SVM	Original	98	97	97	93	91	92	80	100	90
	SNV	100	100	100	99	99	99	65	100	83
	FD	100	100	100	96	95	96	70	95	83
	SD	100	100	100	94	97	95	65	95	80
	MSC	100	100	100	98	99	98	65	100	83
	Normalization	98	97	97	95	92	94	75	100	88
RF	Original	100	100	100	89	90	90	80	100	90
	SNV	100	100	100	95	97	96	65	100	83
	FD	100	100	100	93	93	93	80	100	90
	SD	100	100	100	89	94	91	65	100	83
	MSC	100	100	100	94	96	95	85	100	93
	Normalization	100	100	100	91	91	91	85	100	93

(ACCU) Accuracy; (FD) The first derivative; (MSC) Multiple scatter correlation; (n) Number of samples; (PLS-DA) Partial least square discrimination analysis; (RF) Random forest; (SD) Second derivative; (SENS) Sensitivity; (SPEC) Specificity; (SNV) Standard normalization variable; (SVM) Support vector machine.

while the results of cross validation models were over 90% except for a few models. But in contrast, the performances of external validation models were quite unequal that the range of SENS, SPEC, and ACCU was 65–85%, 50–100%, and 60–93%, respectively. Therefore, the results of the external validation models were critical to determine which algorithms and pre-treatment methods had the best performances.

In terms of the modelling algorithms, PLS-DA is an effective, multivariate regression-based algorithm for peanut classification. Although compelling, PLS-DA incurs performance degeneration under complex situations such as class imbalance and multiclass, which are common in peanut varieties (Song et al., 2018). The full wavelength including 1102 variables, has nonlinearity effects on the PLS-DA models. Therefore, SVM and RF as non-parametric machine learning algorithms have more advantages. Although the cross validation results showed that the SVM models had better results than the RF models, the external validation results of the SVM models had lower performances than the RF models, which meant that the SVM models existed overfitting effects. This is because when using SVM with high dimensional inputs, there is a risk of overfitting which could result in misleading outcomes (Zhang et al., 2018). Hence, RF had the best prediction capacity compared to other algorithms, while the average SENS, SPEC, and ACCU of the RF external models were 77%, 100%, and 89%, respectively. Concerning the pre-treatment methods, all methods increased the performances of the models. Among them, FD pre-treatment had the best improvement in the performances of the models. FD pre-treatment can remove undesired physical scatter caused by the shape of peanut kernels (Rinnan et al., 2009). The average performances (SENC, SPEC, and ACCU) of the models based on FD pre-treatment were 93%, 94%, and 93% (respectively) for cross validation, and 77%, 95%, and 86% (respectively) for external validation. The results showed that the models established based on the full wavelength have great performances to classify peanut varieties to produce different structural characteristics of peanut butters.

The classification models for classifying peanuts varieties except for zhanyou75, Jinhua8, and tianfu18 based on the roast characteristics (colour and volatile compounds) are compared in Table 2. The performances of the classification models for roast characteristics were overall not good when compared with the classification models for structural characteristics. This is because the roast characteristics of some samples of the two groups were similar. Meanwhile, the roast characteristics are not only derived from the raw materials, but also are generated by brown reactions and caramelisation during the roasting process. Therefore, the spectral data of the raw materials lacked information about roasting, leading to less control over controlling roasting variation. Among all pre-treatment methods, SNV was the best pre-treatment and the average performances (SENC, SPEC, and ACCU) were 88%, 85%, and 86% (respectively) for cross validation, and 87%, 92%, and 89% (respectively) for external validation. In respect of the modelling algorithms, SVM and RF showed better modelling capability when compared to PLS-DA algorithm. Among all models, the SNV-SVM models had the best performances (SENC, SPEC, and ACCU) with all parameters 100% for the training models, 98%, 97%, and 97% (respectively) for the cross validation models, and 100%, 94%, and 98% (respectively) for the external validation models.

3.3. The classification models built based on the extracted features

The results of the PLS-DA, SVM, and RF models based on the features extracted by PCA, VarImp, and RFSBF are presented in Table 3 and Table 4. PCA is the conventional method to reduce the dimension of data through building new non-linear variables. The retrieved features in this research were the sum of the top five PCs, accounting for 99% of total variances. The classification models for the structural characteristics are shown in Table 3. The top five PCs of the FD spectral data were selected as the new variables to build models. The PCA-RF models had the best prediction performances compared to the other methods. The ACCU of the training, cross validation, and external validation models was 100%, 95%, and 88%, respectively, which was similar to the results obtained for the full wavelength models. For VarImp and RFSBF, 19 wavelengths and 719 wavelengths were selected as features, respectively. The VarImp method estimated model performances by using the minimum number of wavelengths. The ACCU of the external validation models by VarImp was 90%, 80%, and 83% for PLS-DA, SVM, and RF, respectively. The RFSBF method have already effectively selected the discriminating features from images (Longlong et al., 2020). The ACCU of the training, cross validation, and external validation models was 100%, 96%, and 90% for FD-RFSBF-SVM and 100%, 93%, and 88% for FD-RFSBF-RF. There are slight differences among the above models. Overall, the FD-RFSBF-SVM model had the best performance. RFSBF method maintained or improved the performances of the corresponding full wavelength models. These results proved that features extracted by PCA, VarImp, and RFSBF could be effectively used to build high accuracy and belief models for classifying peanut varieties to process different structural characteristics of peanut butters.

The classification models of the roast characteristics based on the

Table 2
The sensitivity, specificity, and accuracy results for the discrimination of peanut varieties based on the roast characteristics of peanut butters combined with the different spectral algorithm and pre-treatment methods.

Algorithms	Pre-treatments	Training			Cross validation			External validation		
		SENS (%)	SPEC (%)	ACCU (%)	SENS (%)	SPEC (%)	ACCU (%)	SENS (%)	SPEC (%)	ACCU (%)
PLS-DA	Original	91	96	94	75	70	71	80	88	84
	SNV	99	94	97	76	75	75	85	94	89
	FD	81	79	80	66	63	65	50	76	62
	SD	100	100	100	66	64	65	55	82	68
	MSC	94	88	91	79	72	75	75	88	81
	Normalization	94	91	93	72	69	71	85	88	86
SVM	Original	100	100	100	84	81	82	65	76	70
	SNV	100	100	100	98	97	97	100	94	98
	FD	100	100	100	80	76	78	100	71	86
	SD	100	100	100	75	73	74	65	59	62
	MSC	100	100	100	96	99	98	85	65	76
	Normalization	100	100	100	82	76	80	65	77	70
RF	Original	100	100	100	66	57	62	70	41	57
	SNV	100	100	100	90	84	87	75	88	81
	FD	100	100	100	76	75	75	95	82	89
	SD	100	100	100	76	53	65	65	47	57
	MSC	100	100	100	90	85	88	60	88	73
	Normalization	100	100	100	66	63	64	65	35	51

(ACCU) Accuracy; (FD) The first derivative; (MSC) Multiple scatter correlation; (n) Number of samples; (PLS-DA) Partial least square discrimination analysis; (RF) Random forest; (SD) second derivative; (SENS) Sensitivity; (SPEC) Specificity; (SNV) Standard normalization variable; (SVM) Support vector machine.

Table 3
The sensitivity, specificity, and accuracy results for the discrimination of peanut varieties based on the structure characteristics of peanut butters combined with the different algorithm methods using features extracted by PCA, RFSBF, and VarImp.

Algorithms	Data type	Na	Training			Cross validation			External validation		
			SENS (%)	SPEC (%)	ACCU (%)	SENS (%)	SPEC (%)	ACCU (%)	SENS (%)	SPEC (%)	ACCU (%)
PLS-DA	Full	1102	100	100	100	89	94	91	80	90	85
	PCA	5	88	97	93	87	95	91	80	100	90
	RFSBF	719	94	100	97	91	92	91	80	95	88
	VarImp	19	89	96	93	89	96	92	80	100	90
SVM	Full	1102	100	100	100	96	95	96	70	95	83
	PCA	5	99	99	99	97	98	97	75	100	88
	RFSBF	719	100	100	100	96	96	96	80	100	90
	VarImp	19	98	100	99	94	94	94	65	95	80
RF	Full	1102	100	100	100	93	93	93	80	100	90
	PCA	5	100	100	100	93	96	95	80	95	88
	RFSBF	719	100	100	100	93	93	93	75	100	88
	VarImp	19	100	100	100	93	93	93	65	100	83

(ACCU) Accuracy; (n) Number of samples; (PCA) Principal component analysis; (PLS-DA) Partial least square discrimination analysis; (RF) Random forest; (RFSBF) Random forest selection by filter; (SENS) Sensitivity; (SPEC) Specificity; (SNV) Standard normalization variable; (SVM) Support vector machine; (VarImp) Variable importance.
[a] N is the number of variables.

Table 4
The sensitivity, specificity, and accuracy results for the discrimination of peanut varieties based on the roast characteristics of peanut butters combined with the different algorithm methods using features extracted by PCA, RFSBF, and VarImp.

Algorithms	Data type	Na	Training			Cross validation			External validation		
			SENS (%)	SPEC (%)	ACCU (%)	SENS (%)	SPEC (%)	ACCU (%)	SENS (%)	SPEC (%)	ACCU (%)
PLS-DA	Full	1102	99	94	97	76	75	75	85	94	89
	PCA	5	69	56	63	69	53	60	60	53	57
	RFSBF	728	91	94	88	80	74	77	85	94	89
	VarImp	5	80	66	74	84	62	70	55	59	57
SVM	Full	1102	100	100	100	98	97	97	100	94	98
	PCA	5	99	99	99	95	97	96	90	88	89
	RFSBF	728	100	100	100	98	97	97	95	94	95
	VarImp	5	96	99	97	89	85	87	75	88	81
RF	Full	1102	100	100	100	90	84	87	75	88	81
	PCA	5	100	100	100	89	85	86	85	77	81
	RFSBF	728	100	100	100	90	82	86	75	88	81
	VarImp	5	100	100	100	86	84	84	70	77	73

(ACCU) Accuracy; (n) Number of samples; (PCA) Principal component analysis; (PLS-DA) Partial least square discrimination analysis; (RF) Random forest; (RFSBF) Random forest selection by filter; (SENS) Sensitivity; (SPEC) Specificity; (SNV) Standard normalization variable; (SVM) Support vector machine; (VarImp) Variable importance.
[a] N is the number of variables.

features are presented in Table 4. The five new variables extracted from the SNV spectral data by PCA methods were used to establish the models. The SNV-PCA-SVM models had the best prediction performance compared to the other methods. The ACCU of the training, cross validation, and external validation models was 99%, 96%, and 89%, respectively, which was in line with the full wavelength models. The performances of the models built by the 728 feature wavelengths based on RFSBF were consistent with the results based on the corresponding whole wavelength models. The SNV-RFSBF-SVM models had the best performance with 100%, 97%, and 95% for the ACCU of the training, cross validation, and external validation models, respectively. Only five wavelengths (4528.31 cm^{-1}, 4536.02 cm^{-1}, 10044.05 cm^{-1}, 10051.77 cm^{-1}, and 10537.77 cm^{-1}) were selected to build models by VarImp. VarImp was very helpful and efficient to select the important features in the previous research (Dewi and Chen, 2019). Although the performances of the VarImp models were relatively poor when compared to the other feature extraction methods in this study, VarImp greatly simplified the complexity of the models. In summary, it can thus be suggested that different feature extraction methods could be used to simplify models and SNV-RFSBF-SVM was the best method for peanut varieties classification for processing different roast characteristics of peanut butters.

3.4. Kernel density estimation distribution of the selected models

Fig. 5 presents the kernel density estimation (KDE) distribution of the selected models including FD-RF and FD-RFSBF-SVM for structural characteristics, and SNV-SVM and SNV- RFSBF-SVM for roast characteristics. In contrast with binary analysis, KDE distribution analysis offers more content than a single value. The conventional binary analysis classifies samples based on a threshold value. Peanuts with probability values below the threshold are grouped into one class, while peanuts with probability values above the threshold are grouped into the other group. Generally, the number of peanuts classified accurately will be demonstrated. However, KDE distribution plots further show the distances between probability values to the threshold. The smaller the distance, the higher the risk of misclassification (Yan et al., 2019).

In this study, the default threshold was 0.5. Fig. 5a shows two subgroups in all FD-RF models. Specifically, the main body of the two sub-groups was separately located on both sides and the location of the peak were far from the threshold value of 0.5 in the training model. KDE distribution for cross validation and external validation models were similar. Although there was some superposition between the two groups, the main body of different groups were separate. The KDE results of the FD-RFSBF-SVM models shown in Fig. 5b had the better distribution. Two sub-groups in the training model were completely

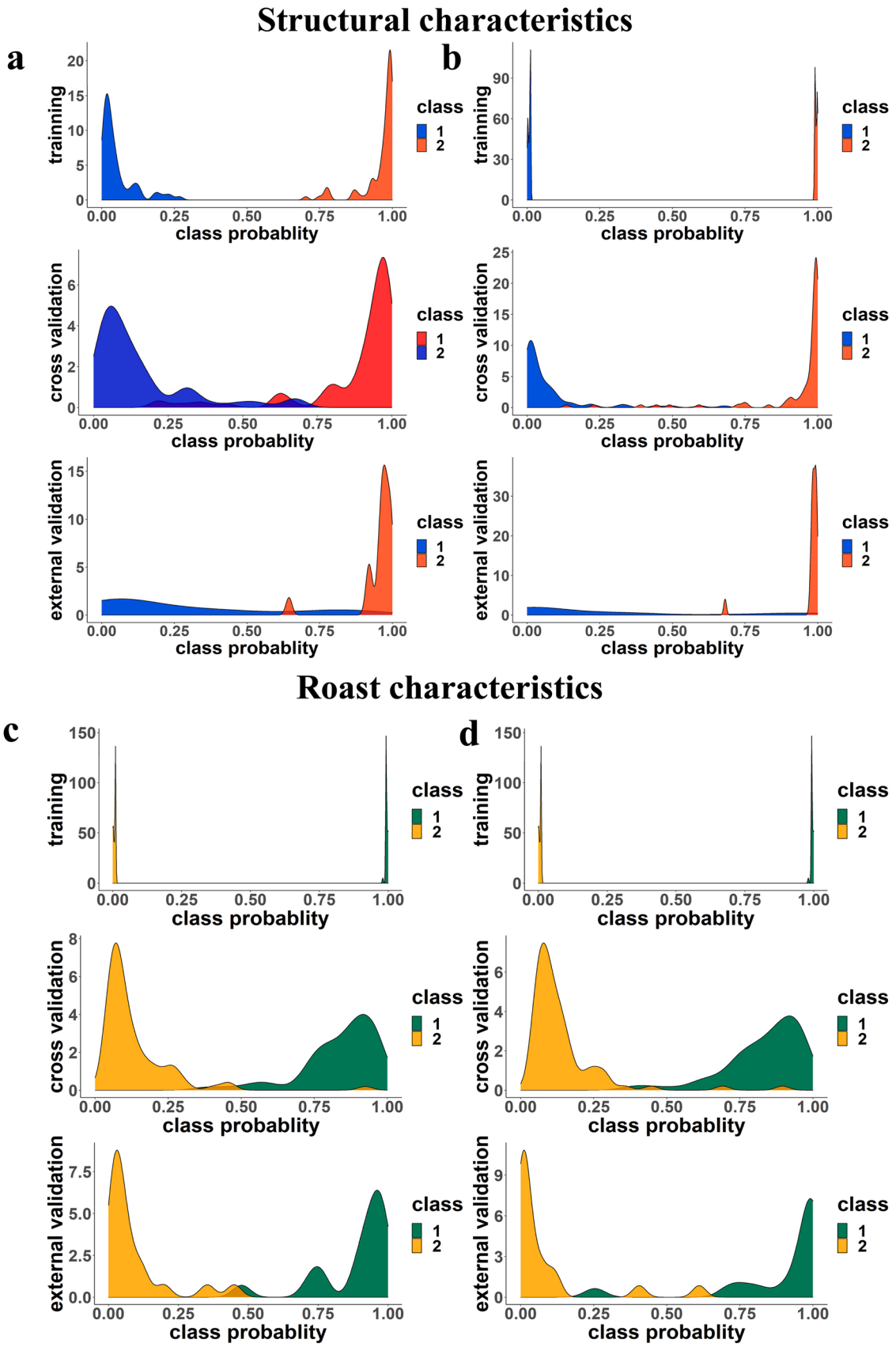

Fig. 5. The kernel density estimation of probability distributions of different models. (a) FD-RF, (b) FD-RFSBF-SVM, (c) SNV-SVM, and (d) SNV-RFSBF-SVM. (FD) The first derivative; (RF) Random forest; (RFSBF) Random forest selection by filter; (SNV) Standard normalization variable; (SVM) Support vector machine.

distributed at both ends. Regarding the cross validation and external validation models, both tails at the sides were slight and the locations of the peaks were far from the threshold value of 0.5. These results corroborated the previous results shown in Table 3.

In respect of the classification models for the roast characteristics, Fig. 5c and Fig. 5d show that the training models of SNV-SVM and SNV-RFSBF-SVM had the greatest distribution of class probability as the different groups located on both ends of the X-axis. For the cross validation models, group 1 and group 2 were mainly separated with a limit of the threshold value (0.5) and the principal part of class probability were 0.05–0.25 for group 1 and 0.75–1.00 for group 2, respectively, which matched the model performances shown in Table 2 and Table 4. One could also see that the external validation model of SNV-SVM had a great segregation between different groups with very little overlap, while the same model of SNV-RFSBF-SVM had a relatively poor distribution that each group had a small peak in the wrong position. Despite this, the SNV-RFSBF-SVM models only used parts of the full wavelength. These results provided further support for the hypothesis that the extracted features could be used to simplify models under the premise of ensuring the stability of model performances.

4. Conclusions

NIR spectroscopy combined with various machine learning approaches was explored in this study to classify peanut varieties for efficient processing based on the structural and roast characteristics of the resulting peanut butters. To date, manufacturers in the peanut processing industry have solely used their inherent knowledge and experience to produce blends of peanuts to obtain peanut butter with different characteristics. The overall results of the study showed the feasibility of using NIR spectroscopy to sort or select peanut varieties based on the expected peanut butter qualities. Systematically scanning all peanuts could provide some objective data to predict the final product characteristics and thus reduce the waste of materials along the processing chain. Technically, machine learning algorithms such as RF and SVM show their wide application space for improving the accuracy of the classification models, especially in the field of food analysis. The extracted feature wavelengths could be used to design low-cost classifier sensors as part of the Internet of things. Further work is needed to increase the sample size and investigate the interactions of the processing conditions to provide guidance for adapted processing procedures to attain stable peanut butters.

CRediT authorship contribution statement

Hongwei Yu: Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – original draught. **Sara W. Erasmus:** Supervision, Writing – review & editing. **Qiang Wang:** Conceptualization, Funding acquisition, Project administration, Supervision, Resources. **Hongzhi Liu**: Conceptualization, Data curation, Methodology, Supervision, Resources. **Saskia van Ruth:** Conceptualization, Project administration, Supervision, Writing – review & editing.

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.

Acknowledgements

This research was financially supported by the National peanut industry technology system "Research and application of key technologies for peanut processing suitability evaluation, quality improvement and efficiency increase" (CARS-13), Hebei oil innovation team of modern agro-industry technology research system (HBCT2019090203), Shandong province Taishan industrial leading talent project (Innovative Category) "Research and application of key Technologies for precision processing and quality control of stable peanut butter", Weihai Wendeng district industry-university-research distinguished expert support programme “Evaluation of the suitability of peanut butter and high-quality product development and industrialization demonstration”, CAAS-WUR joint PhD programme (MOE11NL1A20151701N) and the first author received a scholarship (201903250121) from China Scholarship Council.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jfca.2023.105348.

References

Ahmed, J., Ramaswamy, H.S., 2006. Viscoelastic properties of sweet potato puree infant food. J. Food Eng. 74 (3), 376–382.
Alewijn, M., van der Voet, H., van Ruth, S., 2016. Validation of multivariate classification methods using analytical fingerprints - concept and case study on organic feed for laying hens. J. Food Compos. Anal. 51, 15–23.
Baker, G., Cornell, J., Gorbet, D., O'keefe, S., Sims, C., Talcott, S., 2003. Determination of pyrazine and flavor variations in peanut genotypes during roasting. J. Food Sci. 68 (1), 394–400.
Barbin, D.F., Felicio, A.Ld.S.M., Sun, D.W., Nixdorf, S.L., Hirooka, E.Y., 2014. Application of infrared spectral techniques on quality and compositional attributes of coffee: an overview. Food Res. Int. 61, 23–32.
Dewi, C., Chen, R.-C., 2019. Human activity recognition based on evolution of features selection and random Forest, 2019 IEEE international conference on systems, man and cybernetics (SMC). IEEE, pp. 2496–2501.
Dhamsaniya, N.K., Patel, N.C., Dabhi, M.N., 2012. Selection of groundnut variety for making a good quality peanut butter. J. Food Sci. Technol. 49 (1), 115–118.
Faostat, F., 2020. Statistical databases. Food and Agriculture Organization of the United Nations,.
Gong, A., Shi, A., Liu, H., Yu, H., Liu, L., Lin, W., Wang, Q., 2018. Relationship of chemical properties of different peanut varieties to peanut butter storage stability. J. Integr. Agric 1003–1010.
Hourant, P., Baeten, V., Morales, M.T., Meurens, M., Aparicio, R., 2000. Oil and fat classification by selected bands of near-infrared spectroscopy. Appl. Spectrosc. 54 (8), 1168–1174.
Kuhn, M., 2008. Building predictive models in R using the caret package. J. Stat. Softw. 28 (5), 1–26.
Kuhn, M., Johnson, K., 2013. An introduction to feature selection. Applied predictive modeling. Springer,, New York, pp. 487–519.
Li, C., Hou, L., 2018. Review on volatile flavor components of roasted oilseeds and their products. Grain Oil Sci. Technol. 1 (4), 151–156.
Liu, N., Parra, H.A., Pustjens, A., Hettinga, K., Mongondry, P., van Ruth, S.M., 2018. Evaluation of portable near-infrared spectroscopy for organic milk authentication. Talanta 184, 128–135.
Liu, Y., Yu, Y., Liu, C., Regenstein, J.M., Liu, X., Zhou, P., 2019. Rheological and mechanical behavior of milk protein composite gel for extrusion-based 3d food printing. LWT Food Sci. Technol. 102, 338–346.
Longlong, Z., Xinxiang, L., Yaqiong, G., Wei, W., 2020. Predictive value of the texture analysis of enhanced computed tomographic images for preoperative pancreatic carcinoma differentiation. Front. Bioeng. Biotechnol. 8, 719.
Mohd Rozalli, N.H., Chin, N.L., Yusof, Y.A., 2015. Particle size distribution of natural peanut butter and its dynamic rheological properties. Int. J. Food Prop. 18 (9), 1888–1894.
Monforte, A.R., Martins, S., Silva Ferreira, A.C., 2021. Discrimination of white wine ageing based on untarget peak picking approach with multi-class target coupled with machine learning algorithms. Food Chem. 352, 129288.
Phan, Q., Tomasino, E., 2021. Untargeted lipidomic approach in studying pinot noir wine lipids and predicting wine origin. Food Chem. 355, 129409.
Resch, J.J., Daubert, C.R., 2002. Rheological and physicochemical properties of derivatized whey protein concentrate powders. Int. J. Food Prop. 5 (2), 419–434.
Rinnan, Å., Van Den Berg, F., Engelsen, S.B., 2009. Review of the most common pre-processing techniques for near-infrared spectra. TrAC Trends Anal. Chem. 28 (10), 1201–1222.
Shakerardekani, A., Karim, R., Ghazali, H.M., Chin, N.L., 2013. Textural rheological and sensory properties and oxidative stability of nut spreads-a review. Int. J. Mol. Sci. 14 (2), 4223–4241.
Song, W., Wang, H., Maguire, P., Nibouche, O., 2018. Collaborative representation based classifier with partial least squares regression for the classification of spectral data. chemom. Intell. Lab. Syst. 182, 79–86.

Sun, A., Gunasekaran, S., 2009. Yield stress in foods: measurements and applications. Int. J. Food Prop. 12 (1), 70–101.
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 (3), 156–164.
Wang, L., Wang, Q., Liu, H., Liu, L., Du, Y., 2013. Determining the contents of protein and amino acids in peanuts using near-infrared reflectance spectroscopy. J. Food Sci. Technol. 93 (1), 118–124.
Wang, Q., 2016. Peanuts: Processing technology and product development. Academic Press,, Cambridge.
Wang, Q., 2018. Peanut processing characteristics and quality evaluation. Springer,, New York.
Wang, Q., Liu, H., Shi, A., Hu, H., Liu, L., Wang, L., Yu, H., 2017. Review on the processing characteristics of cereals and oilseeds and their processing suitability evaluation technology. J. Integr. Agric. 16 (12), 2886–2897.
Windham, W., Kandala, C., Sundaram, J., Nuti, R., 2010. Determination of peanut pod maturity by near-infrared reflectance spectroscopy. Trans. ASABE 53 (2), 491–495.
Workman Jr, J., Weyer, L., 2012. Practical guide and spectral atlas for interpretive near-infrared spectroscopy. CRC press,, Boca Raton.
Yan, J., Stuijvenberg, L., Ruth, S.M., 2019. Handheld near-infrared spectroscopy for distinction of extra virgin olive oil from other olive oil grades substantiated by compositional data. Eur. J. Lipid Sci. Technol. 121, 12.
Yu, H., Liu, H., Wang, Q., van Ruth, S., 2020b. Evaluation of portable and benchtop NIR for classification of high oleic acid peanuts and fatty acid quantitation. LWT - Food Sci. Technol. 128.
Yu, H., Liu, H., Wang, N., Yang, Y., Shi, A., Liu, L., Hu, H., Mzimbiri, R.I., Wang, Q., 2016. Rapid and visual measurement of fat content in peanuts by using the hyperspectral imaging technique with chemometrics. Anal. Methods 8 (41), 7482–7492.
Yu, H., Liu, H., Erasmus, S.W., Zhao, S., Wang, Q., van Ruth, S.M., 2020a. Rapid high-throughput determination of major components and amino acids in a single peanut kernel based on portable near-infrared spectroscopy combined with chemometrics. Ind. Crop. Prod. 158.
Yu, H., Liu, H., Erasmus, S.W., Zhao, S., Wang, Q., van Ruth, S.M., 2021. An explorative study on the relationships between the quality traits of peanut varieties and their peanut butters. LWT Food Sci. Technol. 151.
Zhang, D., Qian, L., Mao, B., Huang, C., Huang, B., Si, Y., 2018. A data-driven design for fault detection of wind turbines using random forests and XGboost. IEEE Access 6, 21020–21031.
Zhao, S., Yu, H., Gao, G., Chen, N., Wang, B., Wang, Q., Liu, H., 2021. Rapid determination of protein components and their subunits in peanut based on near infrared technology. Spectrosc. Spect. Anal. 41 (3), 912–917.