Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Detection of dextran, maltodextrin and soluble starch in the adulterated *Lycium barbarum* polysaccharides (LBPs) using Fourier-transform infrared spectroscopy (FTIR) and machine learning models

Lulu Chen [a,1], Siyue Yang [b,**,1], Zhuan Nan [a], Yanping Li [c], Jianlong Ma [d,e], Jianbao Ding [a,c], Yi Lv [f], Jin Yang [a,d,*]

[a] *School of Chemistry and Chemical Engineering, North Minzu University, Yinchuan 750021, China*
[b] *Department of Statistical Sciences, University of Toronto, Toronto M5T 1P5, Canada*
[c] *Ningxia Wuxing Science and Technology Co. Ltd, Yinchuan 750021, China*
[d] *Ningxia Research Center for Natural Medicine Engineering and Technology, Yinchuan 750021, China*
[e] *College of Chemistry and Chemical Engineering, Ningxia University, Yinchuan 750021, China*
[f] *Ningxia Food Testing and Research Institute (Key Laboratory of Quality and Safety of Wolfberry and Wine for State Administration for Market Regulation), Yinchuan 750001, China*

ARTICLE INFO

Keywords:
LBPs
FTIR
Adulteration identification
PCA
Machine learning models
Quantitative and qualitative analysis

ABSTRACT

Due to the similar chemical structures and physicochemical properties, it is challenging to distinguish dextran, maltodextrin, and soluble starch from the polysaccharide products of plant origin, such as *Lycium barbarum* polysaccharides (LBPs). Using the first-order derivatives of Fourier-transformed infrared spectroscopy (FTIR, wave range 1800–400 cm^{-1}), this study proposed a two-step pipeline to identify dextran, maltodextrin, and soluble starch from adulterated LBPs samples qualitatively and quantitatively. We applied principal component analysis (PCA) to reduce the dimensionality of FTIR features. For the qualitative step, a set of machine learning models, including logistic regression, support vector machine (SVM), Naïve Bayes, and partial least squares (PLS), were used to classify the adulterants. For the quantitative step, linear regression, LASSO, random forest, and PLS were used to predict the concentration of LBPs adulterants. The results showed that logistic regression and SVM are suitable for classifying adulterants, and random forests is superior for predicting adulterant concentrations. This would be the first attempt to discriminate the adulterants from the polysaccharide's product of plant origin. The proposed two-step methods can be easily extended to other applications for the quantitative and qualitative detection of samples from adulterants with similar chemical structures.

* Corresponding author. School of Chemistry and Chemical Engineering, North Minzu University, Wenchang Road, Yinchuan 750021, China.
** Corresponding author.
E-mail addresses: syue.yang@mail.utoronto.ca (S. Yang), yang_jin@nun.edu.cn (J. Yang).

[1] The authors contributed equally.

https://doi.org/10.1016/j.heliyon.2023.e17115
Received 20 February 2023; Received in revised form 3 June 2023; Accepted 7 June 2023
Available online 9 June 2023

1. Introduction

The adulteration of medicinal herbs always influences the herbs' effectiveness and safety, thereby this issue has been paid growing attention [1,2]. A series of analytical techniques coupled with chemometrics methods had been developed for the quantitative and qualitative detection the adulterants in botanical ingredients. Among these analytical platform, Fourier-transform infrared spectroscopy (FTIR) is a simple, fast, and green technique and plays an important role to detect adulteration of the exogenous substances [3–6]. These well-documented studies focused on the fraud detection of raw herbal medicines, such as kudzu root, saffron, and Radix Astragali. Litter research has been conducted to distinguish adulterants from the extracts of medicinal plants.

The adulteration of herbal extracts is carried out by the addition of universal adulterants being represented by starches [1]. An investigation found the starch-like substances in *Ganoderma lucidum* dietary supplements in the U.S. market [7]. These universal additives might be safe for consumers, but must reduce the quality of the extract. This issue will be very complicated when polysaccharide is considered as the main bioactive constituent of herbal extract. Identifying adulterants, such as dextran, maltodextrin, and soluble starch, in plant-derived polysaccharide products like *Lycium barbarum* polysaccharides (LBPs), presents a challenge due to their similar chemical structures and physicochemical properties.

LBPs are a class of water-soluble macromolecular sugar complexes contained in the fruit of *Lycium barbarum* L. (Solanaceae) and considered as the most important active ingredients in this herbal medicine [8]. Due to the superior biological activities of LBPs and their widespread industrial application, LBPs extracts have grown fast in the herbal extract market [9]. However, a practical problem in the real world is:

Whether a sample of LBPs is adulterated and, if so, at what concentration?

The challenge stems from the limited availability of suitable tools for evaluating the quality of LBPs extract. The general approach for quality control of LBPs is to utilize the phenol-sulfate acid method according to *the Pharmacopoeia of Peoples Republic of China* (2020 Edition). The accuracy of this method is hard guarantee [9,10]. Moreover, it is challenging to distinguish LBPs from other substances with similar chemical structures, which has led to adulteration [11]. Therefore, there is an urgent need to develop a method to identify purity of polysaccharide extracts from plant origins and ensure compliance of commercially available products.

Infrared technology can be applied to characterize complex mixtures. Unfortunately, when two chemicals with similar chemical structures are mixed, it is difficult to distinguish them by IR spectrum. Machine learning approaches can handle this type of sample-informative, noise-laden data collection. The methods may be used to identify and characterize polysaccharide extracts and their adulterants, making it possible to detect adulterants. In this study, we proposed a machine learning-based method to identify the adulteration of LBPs using FTIR. LBPs were selected as the model plant polysaccharides and dextran, maltodextrin and soluble starch were chosen as adulterants due to the similar chemical structures and physicochemical properties to polysaccharide. We develop a two-step method that can effectively identify the adulterants of LBPs and the concentration of adulterants. In the first step, dextran, maltodextrin, soluble, and LBPs were classified using a set of common machine learning models including the multinomial logistic regression, the support vector machine (SVM), the Naïve Bayes, and the partial least square (PLS) model based on the first derivatives of FTIR values. In the second step, the quantitative model of adulterated LBPs was established for predicting the concentrations of adulterants in the samples *via* the standard regression methods. This would be the first attempt to discriminate the adulterants from the polysaccharide's product of plant origin.

2. Materials and methods

2.1. Materials

Wolfberry was purchased from Zhongning International Wolfberry Trade Center, Zhongning County, Ningxia. The samples were identified by Prof. J. Ding at Ningxia Wuxing Science and Technology Co., Ltd., and the voucher specimens were kept under −20 °C at North Minzu University.

KBr (spectroscopical pure), dextran, soluble starch, and maltodextrin (analytically pure) were obtained from Shaanxi National Pharmaceutical Reagent Co. (Xi'an, Shaanxi). Dextran, soluble starch, and maltodextrin were dried at 105 °C until the mass did not change anymore.

Table 1
Preparation of LBPs and adulterated LBPs.

Name	Saccharides (g/g)		Proportion
	LBPs	Adulterated substances	
Adulterated LBPs	100	0	0%
	95	5	5%
	90	10	10%
	70	30	30%
	50	50	50%
	30	70	70%
	0	100	100%

2.2. *Preparation of LBPs*

Following removal of the small molecule compounds with 70% ethanol, wolfberry was decocted twice with purified water for 1 h each time. The decoction was evaporated in a vacuum to a suitable volume, and then 95% alcohol (v/v) was added to bring the ethanol content to 50%. After centrifugation (5430R, Eppendorf Corporate, Germany) to remove the residue, the supernatant was successively mixed with 95% ethanol until the final ethanol content was 85%. After overnight storage, the precipitate was centrifuged and washed twice with 95% EtOH to obtain LBPs, which were then lyophilized (FD-1C-50, Beijing Biocool Co., Ltd., China). A trace moisture meter (AKD-K5, Yangzhou Accurate Instrument Co. Ltd., China) revealed that the water content of LBPs was less than 5%.

2.3. *Preparation of adulterated LBPs' samples*

LBPs were mixed with dextran, maltodextrin, and soluble starch at 0%, 5%, 10%, 30%, 50%, 70%, and 100% (% adulterated substances) with 50 replicates per group. A total of 1250 samples and the corresponding FTIR spectra were collected. Table 1 provides an overview of the LBPs and adulterated LBPs considered in this study.

2.4. *Generation of infrared spectra*

The FTIR spectra of LBPs and adulterated LBPs samples were obtained by the KBr pellet method. The powder of each sample was ground at a ratio of 1:100 with KBr powder. The mixture was accurately weighed to 0.1 g and pressed into pellet form to record the infrared (IR) spectra. The IR spectrum recorded on a Spectrum GX FTIR spectrometer (PerkinElmer Inc., USA) was in the 4000 - 400 cm^{-1} (2.5–25 μm) range with a resolution of 4 cm^{-1} and 32 scans. Each sample was measured 5 times, and the average value was taken

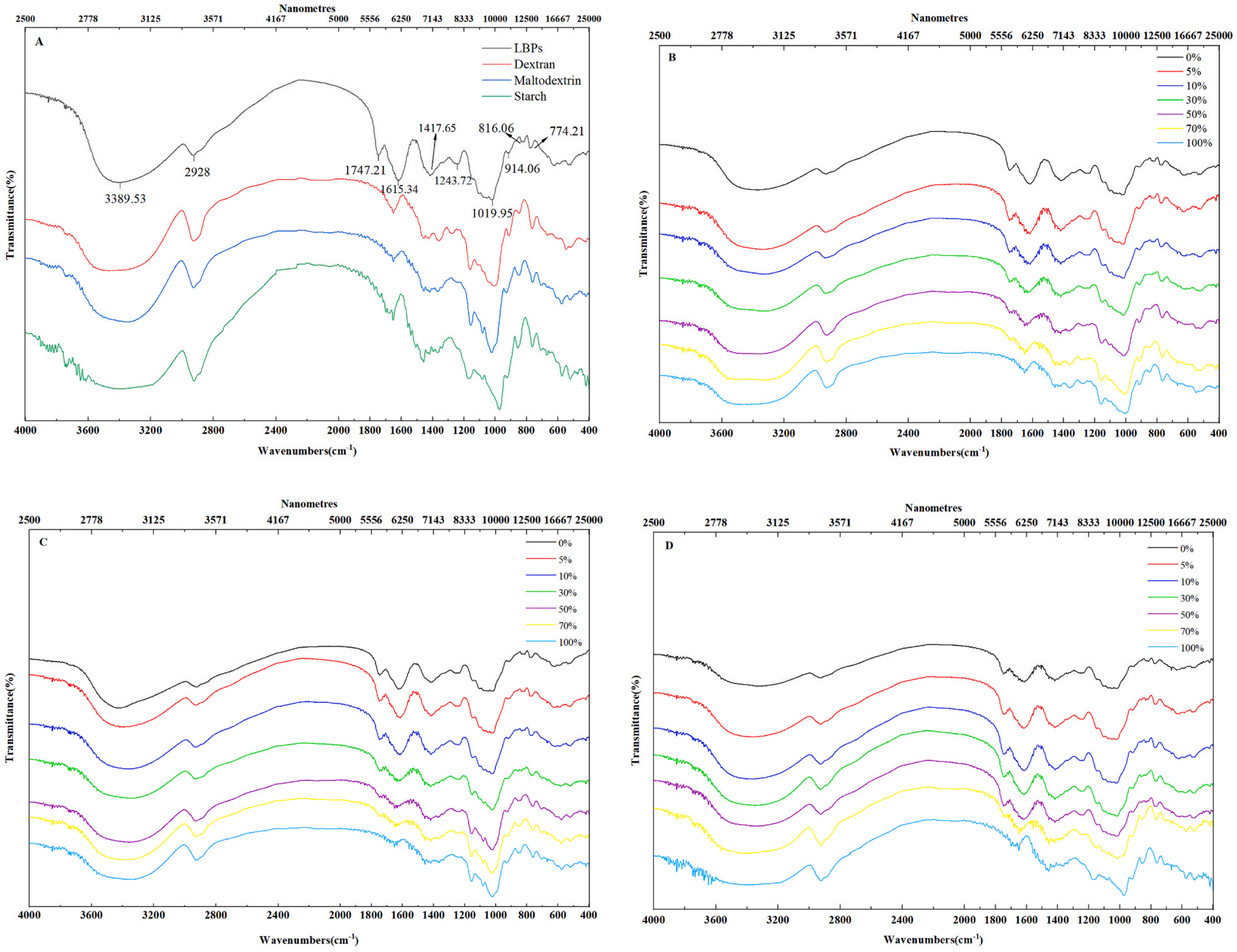

Fig. 1. Infrared spectra of LBPs (0%) and mixed with different concentrations of adulterants (5%, 10%, 30%, 50%, 70% & 100%). A: Spectra of LBPs, dextran, maltodextrin, and soluble starch; B: Spectra of LBPs mixed with dextran; C: Spectra of LBPs mixed with maltodextrin; D: Spectra of LBPs mixed with soluble starch.

as the final spectrum. The background interference of moisture and CO_2 was removed instantly during the scanning process. The IR spectra were processed by the Spectrum 2.0 software equipped on the spectrometer.

2.5. Data processing

Spectrum 2.0 and origin 2018 software were used to process the IR spectra. A series of preprocessing were applied, including curve smoothing, the standard normal variate methods (SNV), leveling baseline correction, and derivative processing to eliminate the instability in IR spectra due to measurement errors and inconsistent compression [12].

To reduce the features and avoid overfitting, principal component analysis (PCA) was performed. Then, the adulterated LBPs were identified in two steps. Firstly, the adulterated components were classified using all data from 5% adulterated substances. The classification models included logistic regression, support vector machine (SVM), Naïve Bayes, and partial least squares discriminant analysis (PLS-DA). Secondly, the concentration of the identified adulterated components was predicted using machine learning models, including linear regression, the least absolute shrinkage, and selection operator (LASSO), random forest, and partial least squares (PLS) model.

3. Results

3.1. Infrared spectral analysis

The IR spectrum provides a wealth of structural information that can be used to characterize materials. Chemical bonds or functional groups in molecules vibrate and absorb when exposed to infrared light. The frequencies of vibration and absorption vary, and the differences are reflected in the IR spectrum [13]. Therefore, the IR spectrum can provide the structural information of the material. Fig. 1A shows the IR spectra of LBPs, dextran, maltodextrin, and soluble starch, respectively. In the IR spectra, the board absorptive band at 3600–3200 cm^{-1} (2778–3125 nm) belonging to the carboxyl group can be observed, which suggested that LBPs and the adulterants possessed the hydroxyl groups [14]. The stretching vibrations of methylene at 2928 cm^{-1} (3415 nm) were the characteristic absorption peak of sugars [15]. In the IR spectrum of LBPs, the absorption bands at 1019, 816, and 771 cm^{-1} (9814, 12, 255, and 12,970 nm) were attributed to the stretching vibration of the pyran ring [16], respectively. The spectra of adulterants, the adsorption peaks at 1010 cm^{-1} (9901 nm) and 767 cm^{-1} (13,038 nm) of dextran, 1019 cm^{-1} (9814 nm) of maltodextrin, and 774 cm^{-1} (12,920 nm) of soluble starch were observed. The information above displayed that Dextran, maltodextrin, and soluble starch possessed similar chemical structures to LBPs.

Meanwhile, the IR spectrum LBPs exhibited the distinctive characteristics. In the spectrum of LBPs (Fig. 1A), the ketone carbonyl vibrations between 1747 and 1615 cm^{-1} (5724 and 6192 nm) indicated that LBPs contained carboxylic acid or an amide bond [17]. The inferences were further supported by the absorption peak at 1243 cm^{-1} (8045 nm), which was induced by O–H variable angle variation of –COOH [18]. The absorption peak at 1417 cm^{-1} (7057 nm) was potentially led by N–H groups from the amide bond

Table 2
Characteristic IR absorptions of LBPs, dextran, maltodextrin, and soluble starch.

Substances	Frequency (cm^{-1}/nm)	Bond	Functional group	References
LBPs	3600–3200/2778–3125	O–H stretch	–OH	[14]
	2928/3415	C–H stretch	$–CH_2$	[15]
	1747/5724	C]O stretch	–COOH	[17]
	1615/6192	C]O stretch	–CONH	[17]
	1417/7057	N–H stretch	–CONH	[17]
	1243/8045	O–H stretch	–COOH	[18]
	1019/9814	C–H stretch	Pyran ring	[16]
	914/10,941	O–H stretch	β-glycosidic bond	[19,20]
	816/12,255	C–H stretch	Pyran ring	[16]
	771/12,970	C–H stretch	Pyran ring	[16]
Dextran	3600–3200/2778–3125	O–H stretch	–OH	[14]
	2928/3415	C–H stretch	$–CH_2$	[15]
	1274/7849	C–O stretch	C–O–C	[16]
	1010/9901	C–H stretch	Pyran ring	[16]
	846/11,820	O–H stretch	α-Glycosidic bond	[19]
	767/13,038	C–H stretch	Pyran ring	[16]
Maltodextrin	3600–3200/2778–3125	–	–OH	[14]
	2928/3415	C–H stretch	$–CH_2$	[15]
	1274/7849	C–O stretch	C–O–C	[16]
	1019/9814	C–H stretch	Pyran ring	[16]
	848/11,792	O–H stretch	α-Glycosidic bond	[19]
Soluble starch	3600–3200/2778–3125	O–H stretch	–OH	[14]
	2928/3415	C–H stretch	$–CH_2$	[15]
	1274/7849	C–O stretch	C–O–C	[16]
	854/11,710	O–H stretch	α-Glycosidic bond	[19]
	774/12,920	C–H stretch	Pyran ring	[16]

(−CONH), demonstrating that LBPs contained a glycoprotein complex. The characteristic absorption peak at 914 cm^{-1} (10,941 nm) indicated that the structure of LBPs was characterized by the pyranose linked by a β-glycosidic bond [19,20].

Fig. 1B–D show the IR spectra of LBPs mixed with different amounts of dextran, maltodextrin, and soluble starch, respectively. Compared with the IR spectrum of LBPs (Fig. 1A), the IR spectra of adulterated LBPs have ketone carbonyl vibrations in 1747 and 1615 cm^{-1} (5724 and 6192 nm) and a stretching band of the amide group at 1417 cm^{-1} (7057 nm), which gradually flattened out with increasing adulteration. One potential reason was that none of the glucuronic acid or the amide bonds exist in dextran, maltodextrin, and starch. Furthermore, the characteristic absorption band of α-glycan at 846 cm^{-1} (11,820 nm, Fig. 1B), 848 cm^{-1} (11,792 nm, Fig. 1C), and 854 cm^{-1} (11,710 nm, Fig. 1D) indicated that the monosaccharide of each adulterated LBPs was linked by an α-glycosidic bond, respectively [19]. Moreover, both LBPs and the adulterated substances contained pyran rings, which coincided with the evidence of the increasing intensity of related vibration peaks in the IR spectrum. Notably, the absorption band at 1274 cm^{-1} (7849 nm) indicated a C–O group from C–O–C in LBPs mixed dextran, and the fluctuation of the band was more intense with the increasing dextran (Fig. 1B). The peculiar vibration peaks of the pyran ring at 1010 and 767 cm^{-1} (9901 and 13,038 nm) fluctuated similarly. The same pattern can be found in the spectra of LBPs mixed maltodextrin (the peak at 1019 cm^{-1}/9814 nm, Fig. 1C) and soluble starch (the peak at 774 cm^{-1}/12,920 nm, Fig. 1D) [21,22]. Table 2 listed the characteristic absorptions of LBPs and adulterants. As shown in Table 2, IR spectra in the range of 1800–400 cm^{-1} (5556–25,000 nm) best reflected the differences between LBPs and adulterated LBPs, which made it possible to distinguish adulterated LBPs from LBPs by suitable infrared spectral analysis.

To capture the characteristics of the IR spectra such as peaks and fluctuations, the numerical discrete gradients of the given 729 data points were calculated using the R package, pracma, which was based on the "central difference formula". Fig. 2 presents the first order derivatives of the IR spectra for LBPs mixed with adulterated substances.

3.2. Dimension reduction

Traditional substance identification using the FTIR spectra typically requires comparison with standard spectra. However, it is challenging to identify adulterated LBPs because there are few standard spectra for LBPs. Instead, we leveraged machine learning

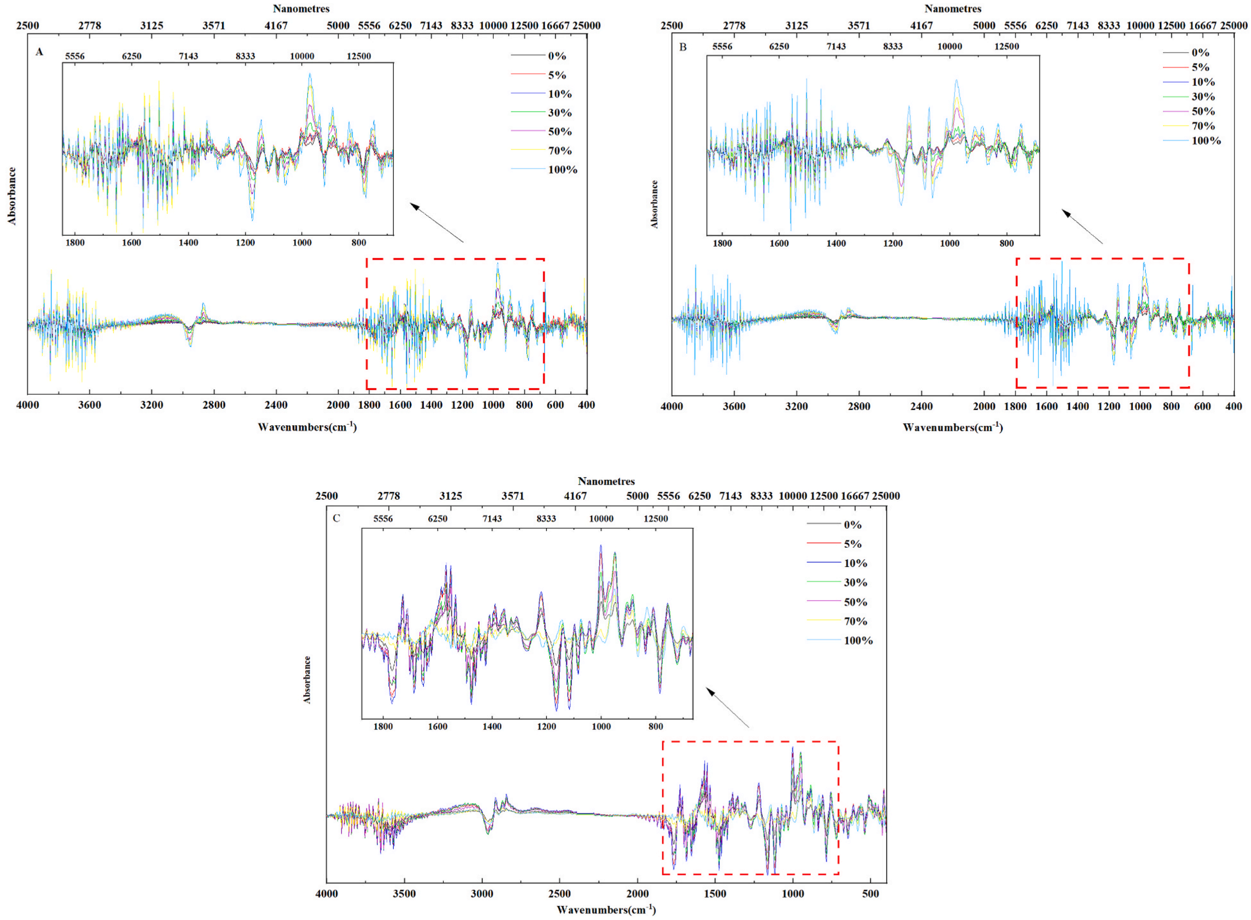

Fig. 2. The first derivative IR spectra of LBPs (0%) and LBPs mixed with different concentrations of adulterants (5%, 10%, 30%, 50%, 70% & 100%). A: Dextran; B: Maltodextrin; C: Soluble starch.

models to distinguish LBPs directly based on the IR spectra' characteristics.

Throughout the study, the 729 derivatives of the IR spectra were used as the features together with 1250 samples for adulterated LBPs classification and prediction. Note that in this dataset, the number of features was comparable to the number of samples in this dataset. Dealing with such a dataset is problematic due to the "curse of dimensionality" [23]. Standard classification models such as logistic regression can easily overfit, leading to low out-of-sample prediction performance [24]. PCA was applied to avoid overfitting.

PCA is a common statistical method for dimensionality reduction [25]. The key idea of PCA is to project the original data onto a smaller dimensional space such that the projections in the new space are separated as much as possible. In other words, PCA aims to maximize the variance in data while maintaining most of the information. The directions of the maximum variance in the new space are principal components and orthogonal. Typically, the principal components are ranked by the explained variance of the whole data. The explained variance is used as the criteria for PCA to select a few principal components with the most variations to achieve dimension reduction.

All samples at 5% concentration of adulterated substances were collected and PCA was applied on the 250 samples with scaled 729 features. Fig. 3 shows the projected data of derivative IR spectra in three-dimensional space. Fig. 3A shows that the LBPs and the adulterated LBPs at 5% concentration are separable in the new space. With the increasing concentrations, i.e., 10%, 30%, and 50%, it was easier to distinguish the LBPs from the adulterated LBPs (Supplementary Material Fig. S1).

With the first ten principal components, the cumulative explained variance reached 98%, indicating that they kept the most useful information of the original dataset (Fig. 3B). Therefore, 10 features processed by PCA were selected for the downstream classification and prediction. Our two-step framework, including code to reproduce results, is publicly available at https://github.com/Yangjcn/adulterated_LBPs_detection.

3.3. Step 1: classification of adulterant contents in LBPs

Machine learning models were used to identify the adulterated samples of LBPs based on the processed derivative IR spectra data. Several common classification models were considered: logistic regression, support vector machine (SVM), Naïve Bayes classifier, and partial least squares discriminant analysis (PLS-DA). The features transformed by the first ten principal components were used to fit logistic regression, SVM, and Naïve Baye. PLS-DA is designed to incorporate high-collinear features [26]; thereby, all scaled features were fitted to PLS-DA. The cross-validation procedure was utilized to compare the model performance and select the best model. Samples containing 5% concentration of the adulterants and samples containing pure LBPs were collected for classification modeling. The dataset was randomly divided into a training set with 160 samples and a testing set with 40 samples (Table 3).

For model training, a 5-fold cross-validation approach [27] with multiple iterations was applied to evaluate the models' ability to generalize into a new dataset. Specifically, the training dataset was randomly equally split into 5 complementary subsets, where four subsets were used for training and the rest was used for model evaluation. The procedure was repeated 5 times, with each of the samples used exactly once as the validation data. The model performance for one iteration was evaluated by the averaged estimation (i.e., mean square error or accuracy) across all the validation folds. 50 rounds of cross-validation on different partitions of the training dataset were performed to obtain a final model for the logistic regression, SVM, Naïve Bayes, and PLS-DA, respectively. The repeated partitions for the model training were designed to generate a stable model and to avoid bias due to experiment or measurement error when collecting and processing the IR spectra of LBPs.

For model testing, the metrics considered included accuracy, sensitivity, precision, and F1-score, which described the model performance from different perspectives. The accuracy, the ratio of correctly classified samples to the total samples, is often used to evaluate the overall model performance. Sensitivity and precision are common metrics used in binary classification. Sensitivity refers to how many LBPs samples are detected, while precision describes how many identified samples are indeed LBPs. In practice, an

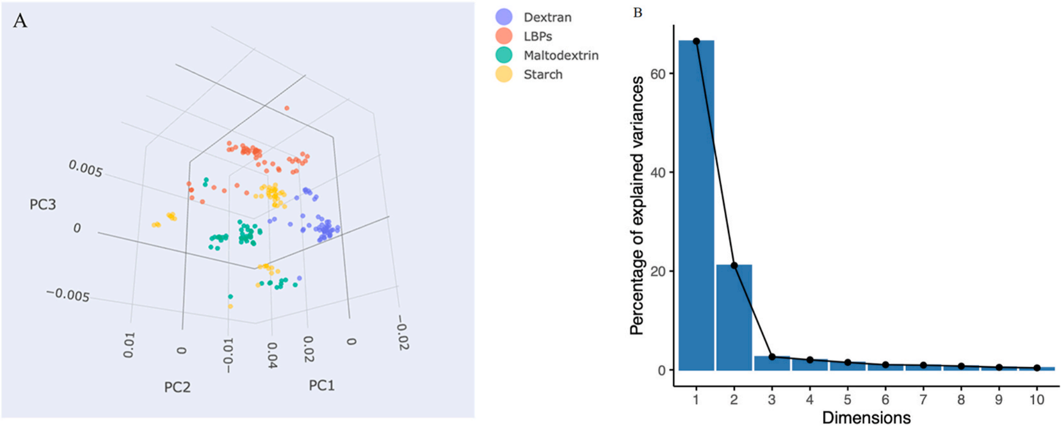

Fig. 3. Result of PCA analysis for LBPs and LBPs mixed with dextran, maltodextrin and soluble starch at 5% concentrations. A: Clustering of LBPs and LBPs adulterants at 5% concentrations; B: Scree plot for cumulative contribution rate.

Table 3
Number of samples in the training set and the testing set to classify dextran, maltodextrin and soluble starch from LBPs.

Sample	Total number of samples	Number of training set samples	Number of test set samples
LBPs	50	44	6
Dextran	50	40	10
Maltodextrin	50	38	12
Soluble starch	50	38	12
Total	200	160	40

increase in sensitivity often comes at the expense of a decrease in precision [28]. Therefore, we used the F1 score that combines the two metrics (Equation (1)).

$$\mathrm{F1-score} = 2 \times \frac{\mathrm{Sensitivity} \times \mathrm{Precision}}{\mathrm{Sensitivity} + \mathrm{Precision}} \tag{1}$$

The binary classification results are summarized in Table 4, while the macro values for sensitivity, precision, and F1-score were calculated by averaging the binary prediction results for each model.

As shown in Table 4, all models achieved high accuracy except for PLS-DA. The logistic regression and SVM reached the highest class-specific accuracy, sensitivity, and precision and the highest overall accuracy, macro-sensitivity, and macro-precision (i.e., 1.000). These observations suggested that the logistic regression and SVM could distinguish pure LBPs from adulterated LBPs. Although the Naïve Bayes is widely used in classification, it did not achieve the values of the precision and F1-score as high as the ones of the logistic regression and SVM when classifying LBPs. Surprisingly, PLS-DA achieved the lowest performance among the four machine learning models. One explanation was that we used 10 spectra derivative features transformed by PCA when fitting logistic regression, SVM, and Naïve Bayes, while for PLS-DA, the input features were all 729 spectra derivatives. The testing accuracy of PLS-DA was probably decreased due to the irrelevant and redundant information in the high-dimensional features. Another potential reason was that PCA and PLS-DA handled high-colinear features from different perspectives; PCA selected a subset of orthogonal projections that captured the most variance in the features, while PLS-DA considered a subset of orthogonal combinations that had the most variance of the features times the square of the correlation to the outcome variable.

The models achieved good classification performance when testing samples of adulterated LBPs with 5% concentration. Next, the rest of the data, containing adulterated LBPs samples at concentrations of 10%, 30%, 50%, 70%, and 100%, were used as an additional testing set to validate the model further. A summary of the additional testing set, and the classification results are available in Supplementary Material Table S1. Consistent with the previous results, the logistic regression and SVM achieved an overall accuracy of over 0.95, suggesting that the model trained with 5% adulterated LBPs samples could distinguish LBPs from dextran, maltodextrin, and starch samples with different concentrations of adulterants.

3.4. Step 2: prediction of adulterant contents in LBPs

Once identifying the adulterated samples by classification machine learning models, regression models were used to predict the concentrations of the samples. 4 regression models, linear regression, LASSO, random forest, and PLS, were considered to study the model performance comparatively and choose the optimal one. Each model was used 3 times to predict the concentration of dextran, maltodextrin, and starch in the LBPs samples. For each adulterated substance, the samples with different concentration levels were divided into 240 training samples and 60 testing samples (Table 5).

Table 4
Comparison of the performance of machine learning models to classify dextran, maltodextrin, soluble starch, and LBPs at the concentration of 5%.

Substances	Evaluation indexes	Machine learning classification methods			
		Logistic regression	SVM	Naïve Bayes	PLS
Overall	Accuracy	1.000	1.000	0.875	0.625
	Macro-Sensitivity	1.000	1.000	0.892	0.688
	Macro-Precision	1.000	1.000	0.875	0.614
	Macro-F1	1.000	1.000	0.868	0.623
LBPs	Sensitivity	1.000	1.000	1.000	1.000
	Precision	1.000	1.000	0.600	0.545
	F1-score	1.000	1.000	0.750	0.706
Dextran	Sensitivity	1.000	1.000	0.900	1.000
	Precision	1.000	1.000	1.000	0.909
	F1-score	1.000	1.000	0.947	0.952
Maltodextrin	Sensitivity	1.000	1.000	0.917	0.250
	Precision	1.000	1.000	1.000	0.500
	F1-score	1.000	1.000	0.957	0.333
Soluble starch	Sensitivity	1.000	1.000	0.750	0.500
	Precision	1.000	1.000	0.900	0.500
	F1-score	1.000	1.000	0.818	0.500

Table 5
Number of samples in the training set and the testing set to predict the concentration of adulterants.

Adulterants	Concentration (%)	Total number of samples	Number of training set	Number of test set
Dextran	5	50	42	8
	10	50	37	13
	30	50	42	8
	50	50	41	9
	70	50	39	11
	100	50	39	11
Total		300	240	60
Maltodextrin	5	50	42	8
	10	50	37	13
	30	50	42	8
	50	50	41	9
	70	50	39	11
	100	50	39	11
Total		300	240	60
Soluble starch	5	50	42	8
	10	50	37	13
	30	50	42	8
	50	50	41	9
	70	50	39	11
	100	50	39	11
Total		300	240	60

Table 6 reports the root mean square error (RMSE) and the mean absolute error (MAE) of prediction models for LBPs mixed with dextran, maltodextrin, and starch. Random forest achieved the best prediction for the concentration of maltodextrin and starch with the smallest MAE and RMSE. For dextran, the linear regression achieved the lowest MAE, while random forest achieved the lowest RMSE. As a result, the random forest was suitable for predicting adulterated concentration.

Fig. 4 compares the predicted values by the random forest and the actual values in the testing set, demonstrating that the predicted values of concentration were close to the actual values. This observation was further verified by the two-sample *t* test, indicating no significant differences between the two groups (p-value $>$0.05, Supplementary Material Table S2).

4. Discussion

Due to the similar chemical structures and physicochemical properties, it is challenging to distinguish dextran, maltodextrin, and starch from the polysaccharide products of plant origin. To identify the adulterants in LBPs, we propose a two-step approach based on the characteristics of LBPs adulterants stored in the IR spectra and chemometric. The two steps are identifying the LBPs adulterants by classification models and detecting the adulterant concentration levels by prediction models.

While our proposed two-step approach has demonstrated great potential to identify adulteration, it also comes with some limitations that are important to note. One potential limitation is the sensitivity of the method to data quality. To address this, we have made every effort to collect as many samples of each adulteration as possible to ensure that the dataset is comprehensive and generalizable. Another limitation is the interpretability of the models. While interpretability is an essential aspect of statistical models, our primary goal in this study was to achieve high predictive power. Therefore, we placed less emphasis on model interpretability and instead focused on the accuracy and reliability of the models. Finally, the computational and implementation complexity of the method is another consideration. While we have employed some non-linear algorithms in our analysis, the size of the dataset was relatively small, and the computational cost was manageable. To ensure transparency and reproducibility, we have made our code publicly available for implementation and replication.

In a preliminary analysis, we also tried a one-step approach that directly predicted the adulterants and concentration with one model. However, PCA cluster analysis with all concentration levels and all types of adulterants indicated that the clusters were indistinguishable in 3D visualization. This can be attributed to the fact that the adulterants' concentrations significantly impacted spectra, leading to a difference between dextran, maltodextrin, starch, and LBPs on IR spectra (Fig. 2). As a result, including samples with all levels of concentration and types of adulterants in one model was not interpretable nor practical. On the other hand, the two-

Table 6
Comparison of the performance of machine learning models to predict the concentration of dextran, maltodextrin and soluble starch.

Adulterated substance	Dextran		Maltodextrin		Starch	
Assessment of the accuracy of concentration predictions	RMSE	MAE	RMSE	MAE	RMSE	MAE
Linear regression	***0.077***	***0.049***	0.074	0.059	0.079	0.060
LASSO	0.082	0.052	0.079	0.064	0.076	0.057
Random forest	0.087	***0.068***	***0.062***	***0.047***	***0.010***	***0.061***
PLS	0.088	0.047	0.083	0.064	0.161	0.135

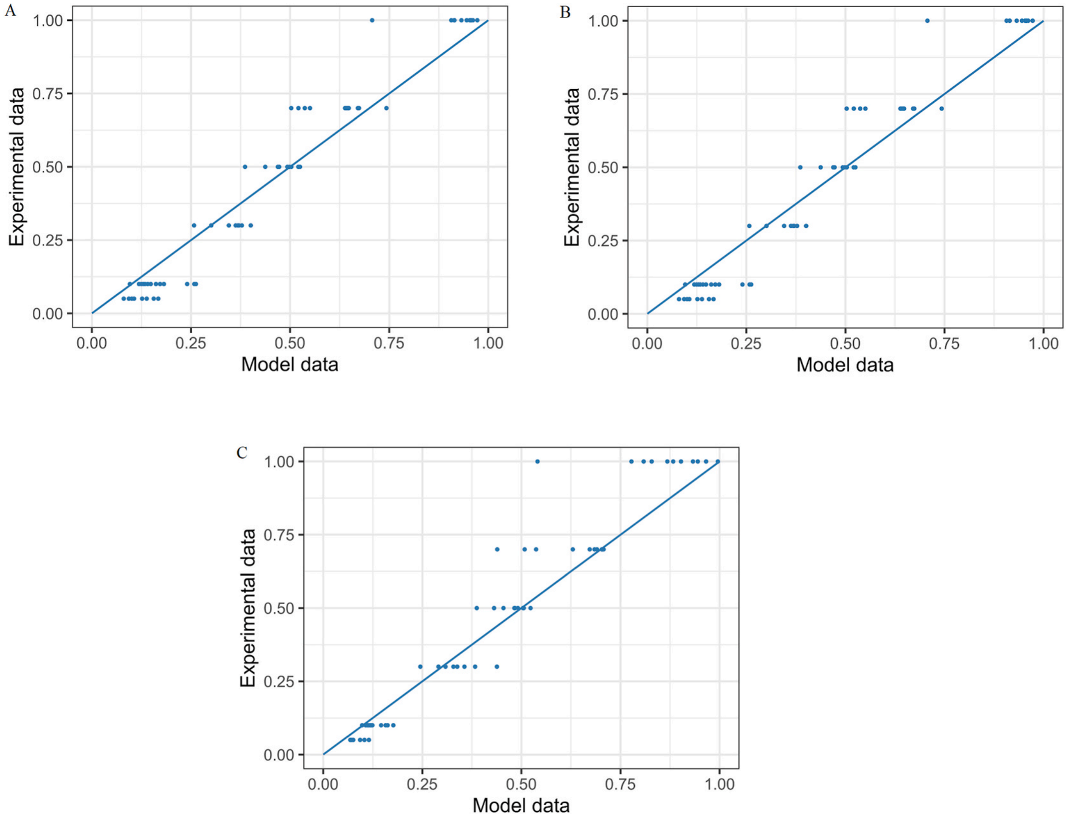

Fig. 4. Comparison of the predicted values of the random forest and the actual values for the concentration of dextran, maltodextrin and soluble starch mixed with LBPs. A: Dextran; B: Maltodextrin; C: Soluble starch.

step approach controlled the concentration level when performing classification by training with the data with 5% adulterants mixed with LBPs. Compared to the one-step approach, the two-step method required more steps and analysis, but remained safer for practical applications. Moreover, the procedure can be extended to distinguish adulterant samples with similar chemical structures.

5. Conclusion

This study proposed a two-step method to detect adulterants of LBPs using mid-infrared spectroscopy and machine learning models. 729 features extracted from the first-order derivative spectra at the range of 1800 cm^{-1} and 400 cm^{-1} were processed to reduce the dimensionality via PCA. In step 1, classification models, including logistic regression, SVM, Naïve Bayes, and PLS, were applied to classify adulterated LBPs at 5% concentration. Logistic regression and SVM classified LBPs adulterants at different concentration levels with the highest accuracy, sensitivity, precision, and F1-score, when tested using samples of concentration at 5%, 10%, 30%, 50%, 70%, and 100%. In step 2, the prediction models, including linear regression, LASSO, the random forest, and PLS, were used to predict the concentration of dextran, maltodextrin, and soluble starch in LBPs, respectively. Random forest achieved the lowest RMSE and MAE and was suitable for predicting adulterants' concentration. Combining mid-infrared spectroscopy and machine learning models to identify adulterants was non-destructive, non-polluting, convenient, and efficient. Moreover, the proposed two-step method can easily distinguish samples from adulterants with similar chemical structures.

Author contribution statement

Lulu Chen: Performed the experiments; Wrote the paper. Siyue Yang: Conceived and designed the experiments; Analyzed and interpreted data; Wrote the paper. Zhuan Nan: Performed the experiments. Yanping Li: Analyzed and interpreted data; Wrote the paper. Jianlong Ma, Yi Lv: Contributed reagents, materials, analysis tools or data. Jianbao Ding, Jin Yang: Conceived and designed the experiments.

Data availability statement

Data associated with this study has been deposited at https://github.com/Yangjcn/adulterated_LBPs_detection.

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.

Acknowledgements

This study was supported by the Key Research and Development Program of Ningxia Hui Autonomous Region [2020BFG02016], China.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2023.e17115.

References

[1] A.G. Osman, V. Raman, S. Haider, Z. Ali, A.G. Chittiboyina, I.A. Khan, Overview of analytical tools for the identification of adulterants in commonly traded herbs and spices, J. AOAC Int. 102 (2019) 376–385, https://doi.org/10.5740/jaoacint.18-0389.

[2] Z. Liu, M.Q. Yang, Y. Zuo, Y. Wang, J. Zhang, Fraud detection of herbal medicines based on modern analytical technologies combine with chemometrics approach: a review, Crit. Rev. Anal. Chem. 52 (2022) 1606–1623, https://doi.org/10.1080/10408347.2021.1905503.

[3] L. Hu, C. Yin, Fourier transform infrared spectroscopy coupled with chemometrics for determining the geographical origin of kudzu root and the detection and quantification of adulterants in kudzu root, Anal. Methods 9 (2017) 3643–3652, https://doi.org/10.1039/c7ay00876g.

[4] A. Amirvaresi, N. Nikounezhad, M. Amirahmadi, B. Daraei, H. Parastar, Comparison of near-infrared (NIR) and mid-infrared (MIR) spectroscopy based on chemometrics for saffron authentication and adulteration detection, Food Chem. 344 (2021), 128647, https://doi.org/10.1016/j.foodchem.2020.128647.

[5] J. Yang, C. Yin, X. Miao, X. Meng, Z. Liu, L. Hu, Rapid discrimination of adulteration in Radix Astragali combining diffuse reflectance mid-infrared Fourier transform spectroscopy with chemometrics, Spectrochim. Acta A: Mol. Biomol. Spectrosc. 248 (2021), 119251, https://doi.org/10.1016/j.saa.2020.119251.

[6] S.-W. Hou, W. Wei, Y. Wang, J.-H. Gan, Y. Lu, N.-P. Tao, X.-C. Wang, Y. Liu, C.-H. Xu, Integrated recognition and quantitative detection of starch in surimi by infrared spectroscopy and spectroscopic imaging, Spectrochim. Acta A: Mol. Biomol. Spectrosc. 215 (2019) 1–8, https://doi.org/10.1016/j.saa.2019.02.080.

[7] D.-T. Wu, Y. Deng, L.-X. Chen, J. Zhao, A. Bzhelyansky, S.-P. Li, Evaluation on quality consistency of Ganoderma lucidum dietary supplements collected in the United States, Sci. Rep. 7 (2017) 7792, https://doi.org/10.1038/s41598-017-06336-3.

[8] X. Tian, T. Liang, Y. Liu, G. Ding, F. Zhang, Z. Ma, Extraction, structural characterization, and biological functions of Lycium barbarum polysaccharides: a review, Biomolecules 9 (2019) 389, https://doi.org/10.3390/biom9090389.

[9] D.-T. Wu, H. Guo, S. Lin, S.-C. Lam, L. Zhao, D.-R. Lin, W. Qin, Review of the structural characterization, quality evaluation, and industrial application of Lycium barbarum polysaccharides, Trends Food Sci. Technol. 79 (2018) 171–183, https://doi.org/10.1016/j.tifs.2018.07.016.

[10] F. Yue, J. Zhang, J. Xu, T. Niu, X. Lü, M. Liu, Effects of monosaccharide composition on quantitative analysis of total sugar content by phenol-sulfuric acid method, Front. Nutr. 9 (2022), 963318, https://doi.org/10.3389/fnut.2022.963318.

[11] H.M. Salo, N. Nguyen, E. Alakärppä, L. Klavins, A.L. Hykkerud, K. Karppinen, L. Jaakola, M. Klavins, H. Häggman, Authentication of berries and berry-based food products, Compr. Rev. Food Sci. Food Saf. 20 (2021) 5197–5225, https://doi.org/10.1111/1541-4337.12811.

[12] Y. Jiao, Z. Li, X. Chen, S. Fei, Preprocessing methods for near-infrared spectrum calibration, J. Chemom. 34 (2020), https://doi.org/10.1002/cem.3306.

[13] Y.K. Ke, H.R. Dong, Handbook of Analytical Chemistry: Molecular Spectroscopy Analysis, Chemical Industry Press Co. Ltd, Beijing, 2016.

[14] D. Yuan, C. Li, Q. Huang, X. Fu, Ultrasonic degradation effects on the physicochemical, rheological and antioxidant properties of polysaccharide from Sargassum pallidum, Carbohydr. Polym. 239 (2020), 116230, https://doi.org/10.1016/j.carbpol.2020.116230.

[15] N.A. Nikonenko, D.K. Buslov, N.I. Sushko, R.G. Zhbankov, Investigation of stretching vibrations of glycosidic linkages in disaccharides and polysaccharides with use of IR spectra deconvolution, Biopolymers 57 (2000) 257–262, https://doi.org/10.1002/1097-0282(2000)57:4<257::AID-BIP7>3.0.CO;2-3.

[16] Y.-J. Liu, X.-L. Mo, X.-Z. Tang, J.-H. Li, M.-B. Hu, D. Yan, W. Peng, C.-J. Wu, Extraction optimization, characterization, and bioactivities of polysaccharides from Pinelliae Rhizoma Praeparatum cum alumine employing ultrasound-assisted extraction, Molecules 22 (2017) 965, https://doi.org/10.3390/molecules22060965.

[17] M. Zhang, H. Zu, X. Zhuang, Y. Yu, Y. Wang, Z. Zhao, Y. Zhou, Structural analyses of the HG-type pectin from notopterygium incisum and its effects on galectins, Int. J. Biol. Macromol. 162 (2020) 1035–1043, https://doi.org/10.1016/j.ijbiomac.2020.06.216.

[18] E.E. Santos, R.C. Amaro, C.C.C. Bustamante, M.H.A. Guerra, L.C. Soares, R.E.S. Froes, Extraction of pectin from agroindustrial residue with an ecofriendly solvent: use of FTIR and chemometrics to differentiate pectins according to degree of methyl esterification, Food Hydrocolloids 107 (2020), 105921, https://doi.org/10.1016/j.foodhyd.2020.105921.

[19] E. Wiercigroch, E. Szafraniec, K. Czamara, M.Z. Pacia, K. Majzner, K. Kochan, A. Kaczor, M. Baranska, K. Malek, Raman and infrared spectroscopy of carbohydrates: a review, Spectrochim. Acta A: Mol. Biomol. Spectrosc. 185 (2017) 317–335, https://doi.org/10.1016/j.saa.2017.05.045.

[20] M. Kacuráková, M. Mathlouthi, FTIR and laser-Raman spectra of oligosaccharides in water: characterization of the glycosidic bond, Carbohydr. Res. 284 (1996) 145–157, https://doi.org/10.1016/0008-6215(95)00412-2.

[21] T. Hong, J.-Y. Yin, S.-P. Nie, M.-Y. Xie, Applications of infrared spectroscopy in polysaccharide structural analysis: Progress, challenge and perspective, Food Chem. X 12 (2021), 100168, https://doi.org/10.1016/j.fochx.2021.100168.

[22] Yu-Xiao Wang, Xin Yue, Jun-Yi Yin, Xiao-Jun Huang, Jun-Qiao Wang, Jie-Lun Hu, Fang Geng, Shao-Ping Nie, Revealing the architecture and solution properties of polysaccharide fractions from Macrolepiota albuminosa (Berk.) Pegler, Food Chem. 368 (2022), 130772, https://doi.org/10.1016/j.foodchem.2021.130772.

[23] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, second ed., Springer, New York, NY, 2017 https://doi.org/10.1007/978-0-387-84858-7.

[24] D.M. Hawkins, The problem of overfitting, ChemInform 35 (2004), https://doi.org/10.1002/chin.200419274.

[25] I.T. Jolliffe, J. Cadima, Principal component analysis: a review and recent developments, Philos. Trans. A Math. Phys. Eng. Sci. 374 (2016), 20150202, https://doi.org/10.1098/rsta.2015.0202.

[26] V.E. Vinzi, W.W. Chin, J. Henseler, H. Wang (Eds.), Handbook of Partial Least Squares, Springer, Berlin, Germany, 2016.

[27] S. Arlot, A. Celisse, A survey of cross-validation procedures for model selection, Stat. Surv. 4 (2010) 40–79, https://doi.org/10.1214/09-ss054.

[28] H.B. Wong, G.H. Lim, Measures of diagnostic accuracy: sensitivity, specificity, PPV and NPV, Proc. Singapore Healthc. 20 (2011) 316–318, https://doi.org/10.1177/201010581102000411.