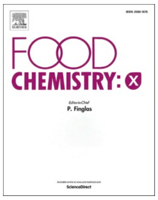
Contents lists available at ScienceDirect

Food Chemistry: X

journal homepage: www.sciencedirect.com/journal/food-chemistry-x



Rapid evaluation of Ziziphi Spinosae Semen and its adulterants based on the combination of FT-NIR and multivariate algorithms

Ming-xuan Li [b,1], Ya-bo Shi [b,1], Jiu-ba Zhang [b], Xin Wan [b], Jun Fang [b], Yi Wu [b], Rao Fu [b], Yu Li [b], Lin Li [b], Lian-lin Su [b,*], De Ji [b], Tu-lin Lu [b,*], Zhen-hua Bian [a,c,*]

[a] *Department of Pharmacy, Wuxi TCM Hospital Affiliated to Nanjing University of Chinese Medicine, Wuxi, 214071, China*
[b] *College of Pharmacy, Nanjing University of Chinese Medicine, Nanjing, 210023, China*
[c] *Jiangsu CM Clinical Innovation Center of Degenerative Bone & Joint Disease, Wuxi TCM Hospital Affiliated to Nanjing University of Chinese Medicine, Wuxi, 214071, China*

ARTICLE INFO

Keywords:
Adulteration
FT-NIR
Chromaticity
Ziziphi Spinosae Semen
Machine learning algorithm
Spectra preprocessing
Wavelength selection
Partial least squares regression model

ABSTRACT

Ziziphi Spinosae Semen (ZSS) is a valued seed renowned for its sedative and sleep-enhancing properties. However, the price increase has been accompanied by adulteration. In this study, chromaticity analysis and Fourier transform near-infrared (FT-NIR) combined with multivariate algorithms were employed to identify the adulteration and quantitatively predict the adulteration ratio. The findings suggested that the utilization of chromaticity extractor was insufficient for identification of adulteration ratio. The raw spectrum of ZMS and HAS adulterants extracted by FT-NIR was processed by SNV + CARS and 1d + SG + ICO respectively, the average accuracy of machine learning classification model was improved from 77.06 % to 97.58 %. Furthermore, the R^2 values of the calibration and prediction set of the two quantitative prediction regression models of adulteration ratio are greater than 0.99, demonstrating excellent linearity and predictive accuracy. Overall, this study demonstrated that FT-NIR combined with multivariate algorithms provided a significant approach to addressing the growing issue of ZSS adulteration.

1. Introduction

Food fraud, as a global issue, has attracted extensive attention and concern in various countries and regions (Cebi, Bekiroglu, Erarslan, & Rodriguez-Saona, 2023). Whether for profiteering or deceiving consumers, food adulteration seriously infringes upon the public's rights to food safety and poses an undeniable threat to social stability and health (Saadat, Pandya, Dey, & Rawtani, 2022). Despite the measures taken by governments and relevant institutions worldwide to combat food adulteration, this problem persists and is characterized by a certain level of complexity (Gizaw, 2019). Effectively addressing the issue of food adulteration and safeguarding the public's dietary safety has become an important and urgent issue.

Ziziphi Spinosae Semen (ZSS), derived from the dried mature seeds of *Ziziphus jujuba* Mill. Var. *Spinosa* (Bunge) Hu ex H. F. Chou, is primarily grown in the Asian, European, and Australian continents. Recognized as the primary origination of ZSS in the market of China, Shandong province commands a main market share and premium pricing, which has become a frequent target for unscrupulous merchants seeking to engage in adulteration. Modern research has indicated that ZSS contains over 150 active ingredients, which possess various beneficial effects such as sedation and hypnotic properties (Yang et al., 2023; Wang, Ho, & Bai, 2022). This has consequently resulted in a continuous increase in the price of ZSS, which has prompted unscrupulous

Abbreviations: 1d, 1st-derivative; 2d, 2nd-derivative; ANN, Artificial Neural Network; CARS, Competitive Adaptive Reweighted Sampling; FDR, False Discovery Rate; FNR, False Negative Rate; HAS, Hovenia Acerba Semen; ICO, Interval Combination Optimization; MSC, Multiplicative Scattering Correction; KNN, K-Nearest Neighbors; PCA, Principal Component Analysis; PLSR, Partial Least Squares Regression; PPV, Positive Predictive Value; RF, Random Frog; R_C^2, Calibration Determination Coefficient; R_P^2, Prediction Determination Coefficient; RMSEC, Root Mean Square Error of Calibration; RMSEP, Root Mean Square Error of Prediction; RPD, Relative Percent Deviation; SNV, Standard Normal Variate; SG, Savitzky Golay; SVM, Support Vector Machine; TPR, True Positive Rate; ZMS, Ziziphi Mauritianae Semen.

[*] Corresponding authors.

E-mail addresses: sll2020@njucm.edu.cn (L.-l. Su), ltl2021@njucm.edu.cn (T.-l. Lu), 20193096@njucm.edu.cn (Z.-h. Bian).

[1] They contributed equally to the present paper.

https://doi.org/10.1016/j.fochx.2023.101022
Received 7 September 2023; Received in revised form 9 November 2023; Accepted 19 November 2023
Available online 20 November 2023

merchants to use Ziziphi Mauritianae Semen (ZMS) and Hovenia Acerba Semen (HAS) in place of ZSS to deceive consumers and seek profits. Adulteration below 10 % is unprofitable for unscrupulous merchants, while above 50 % can be easily discerned by consumers, so the adulteration ratios in food markets typically range from 10 % to 50 %. ZSS, ZMS, and HAS exhibit similar shapes and colors, making it difficult to achieve precise distinctions solely with the naked eye. Nevertheless, the chemical composition and efficacy of ZMS and HAS differ significantly from those of ZSS, also with a price difference of nearly tenfold (Zhang et al., 2023). The aforementioned differences not only disrupts the competitive landscape of the market but also seriously harms the physical health of consumers who inadvertently consume adulterated products. Therefore, establishing a rapid and effective method to differentiate ZSS from its adulterants is of utmost importance.

Currently, numerous techniques such as untargeted ^{1}H NMR metabolomics (Yong et al., 2022), KASP, GC–MS/MS (Wang, Bai, Chen, Ren, Pang, & Han, 2022), and UHPLC-ELSD/UV (Sun, Lu, & Gao, 2021) have been employed for the identification of food adulterants. However, these methods suffer from limitations, including long analysis time, high cost, complex operation, and intricate sample preparation, which restrict their large-scale application in practical production settings. Fourier transform near-infrared spectroscopy (FT-NIR) is a highly efficient and rapid detection technology that integrates advancements in computer science, spectral analysis, and chemometrics across various disciplines (Beć, Grabska, & Huck, 2022b). FT-NIR has found widespread application in fast quality identification and control in the food industry (Cozzolino, 2021; Qu et al., 2015). It offers notable advantages such as rapid analysis speed, user-friendly operation, non-destructive testing, and environmentally friendly characteristics (Beć & Huck, 2019). Various studies have also indicated the tremendous potential of FT-NIR spectroscopy in rapidly detecting food adulteration. For instance, it has been successfully applied to differentiate natural honey adulterated with syrup (Huang et al., 2020), palm oil adulterated with lard (Basri, Hussain, Bakar, Sharif, Khir, & Zoolfakar, 2017), and butter adulterated with vegetable oil (Medeiros, Freitas, Correia, Teixeira, & Fernandes, 2023).

Although a recent study demonstrated that Flash GC e-nose and HS-GC–MS can successfully identify whether ZSS is adulterated (Zhang et al., 2023), but it did not involve the discrimination and quantitative prediction of ZSS adulteration ratio, and the high cost of the instruments used made them unsuitable for practical large-scale regulatory applications. The application of FT-NIR can not only fill the gap in the quantitative study of ZSS adulteration ratio, but also provide an economical, rapid and accurate approach of adulteration identification. Moreover, there were few reports that FT-NIR analysis was combined with multivariate algorithms such as wavelength selection and machine learning to solve food fraud.

Therefore, this experiment aimed to investigate the feasibility of integrating FT-NIR spectroscopy with multivariate algorithms such as characteristic wavelength selection for rapid identification and quantitative prediction of ZSS adulteration ratio, and the machine learning classification model and partial least squares regression (PLSR) quantitative correction model were established to verify the feasibility. The research findings are expected to contribute to a new quality control approach for ZSS and other seed foods.

2. Materials and methods

2.1. Materials

All samples of ZSS (15 batches) were collected from the Shandong province in China. Besides, ZMS and HAS for adulteration were purchased from the wholesale food market. All samples were identified by Prof. Jianwei Chen from the College of Pharmacy, Nanjing University of Chinese Medicine, and passed the food quality detection to ensure their authenticity and reliability.

2.2. Samples preparation

Before analysis, all samples were crushed into powder by a high-speed crusher (QE-300, Zhejiang Yili Industry and Trade Co., Ltd., Zhejiang, China), passed through a 65-mesh sieve (250 ± 9.9um), and stored in dry, sealed, and dark environment. Therefore, in this experiment, used as adulterants, each batch of ZMS and HAS were randomly added at gradients of 10 %, 20 %, 30 %, 40 %, and 50 % to 15 batches of pure ZSS. The samples were sufficiently mixed using a vortex oscillator (LC-Vortex-*P*2, Shanghai Lichen instrument Technology Co., Ltd., Shanghai, China). Two parallel samples were prepared for each gradient, resulting in a total of 30 samples for each adulteration gradient and 300 adulterated samples in total (150 each for ZMS and HAS adulterants). In addition, pure samples of ZSS, ZMS, and HAS were prepared as control groups for comparison.

2.3. Chromaticity measurement

After calibrating the instrument, the chromaticity (L^*, a^*, and b^*) of pure samples and different proportion adulterated powders were determined by a Chroma extractor (CM-5, KONICA MINOLTA, Tokyo, Japan) under transmission mode using CIE D65/10°. At room temperature, the sample powders were put into a quartz cuvette for detection, triplicate parallel measurements of each sample were undertaken to mitigate systematic errors and provide L^*, a^*, and b^* values, which were ultimately averaged to yield the measurement outcome. Moreover, due to potential disturbances caused by external light sources, tests on precision, stability, and repeatability were implemented. A higher L^* value signifies a brighter color. The larger the a^* value is, the redder the color is, and the larger the b^* value is, the more yellow the color is (Kulapichitr, Borompichaichartkul, Fang, Suppavorasatit, & Cadwallader, 2022). Total chromatic value (E^*ab), Chroma (C^*), hue angle (H^*), and color index (*CI*) were calculated using Eqs. (1), (2), (3), and (4):

$$E^{*}ab = \sqrt{L^{*2} + a^{*2} + b^{*2}} \tag{1}$$

$$C^{*} = \sqrt{a^{*2} + b^{*2}} \tag{2}$$

$$H^{*} = arctangent(b^{*}/a^{*}) \tag{3}$$

$$CI = (180^{\circ} - H^{*})/(L^{*} + C^{*}) \tag{4}$$

2.4. FT-NIR spectra acquisition

All tested samples were dried to constant weight before analysis to eliminate moisture interference. The Antaris II FT-NIR spectrometer (Thermo Fisher Scientific, USA) was utilized to collect spectral data of the test powders through diffusive reflection mode. Specially designed quartz vials were employed to hold the powders, with minor compression applied to ensure uniform filling densities. Spectra were measured in the 12,000 ~ 4000 cm^{-1} range, at a resolution of 16 cm^{-1} and with 32 scans per spectrum. Triplicate parallel analyses were conducted for each sample to obtain an average spectrum, which was subsequently employed for subsequent analyses.

2.5. FT-NIR spectra preprocessing

The spectral information acquired through FT-NIR is often contaminated by extraneous signals and noise, namely stray light, strong electrical noise, and human-induced noise during transmission, besides the desired sample information (Xiao et al., 2022). These interferences can significantly compromise the accuracy of the built models. Thus, to enhance the performance of quantitative models, it is indispensable to implement appropriate preprocessing steps on the raw spectral signals. A diverse range of preprocessing techniques, such as multiplicative scattering correction (MSC), standard normal variate (SNV), Savitzky

Golay (SG), 1st-derivative (1d), 2nd-derivative (2d), and their combinations, are employed through The Unscrambler X 10.4 software (CAMO, Inc., TEXAS, USA) in this study to optimize the initial spectra with the aim of establishing an optimal predictive model.

2.6. Characteristic wavelength selection

The collinearity problem carried by too many variables in FT-NIR raw spectra will cause varying degrees of influence on the accuracy and stability of the subsequent model predictions (Beć, Grabska, & Huck, 2022a). Therefore, the selection of characteristic wavelengths for preprocessed spectra is critical. The extraction of characteristic wavelengths can typically be divided into wavelength interval and wavelength point selection algorithms, which include competitive adaptive reweighted sampling (CARS), interval combination optimization (ICO), random frog (RF), and other characteristic wavelength selection algorithms.

The CARS algorithm is based on the absolute value of regression coefficients and serves as a wavelength point selection technique. In each screening step of the Partial Least Squares Regression (PLSR) model, the smallest regression coefficient's absolute value is removed, thereby preserving wavelengths with higher weights. This process culminates in identifying a subset of critical wavelength points for prediction objects through cross-validation, resulting in the lowest root-mean-square error of prediction (Li, Liang, Xu, & Cao, 2009). The ICO algorithm achieves optimal selection of wavelength segments for NIR raw spectra by applying a soft shrinkage method to optimize the wavelength interval combination, followed by automatic optimization of the width of the final selected interval through local search (Song, Huang, Yan, Xiong, & Min, 2016). RF algorithm operates iteratively, calculating the probability of each variable being selected in each iteration. The higher the probability, the more influential the variable is, and variables with higher selection probabilities are ultimately chosen as feature variables (Li, Xu, & Liang, 2012). The application of the above three wavelength selection algorithms can be achieved through MATLAB 2021a software (MathWorks Inc., Natick, MA, USA).

2.7. Application of classification model

As a multivariate statistical method, Principal Component Analysis (PCA) is used to transform and reduce the dimensionality of collected data, and linear classification is further performed on the reduced data (Yan et al., 2021), which can be achieved by Simca-p 14.1 software (SIMCA Imola s.c., Imola, Bologna, Italy). Besides, the algorithms of Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Artificial Neural Networks (ANN) were developed through MATLAB 2021a software (MathWorks Inc., Natick, MA, USA) to perform classification and discrimination on different levels of adulterants as well as pure samples. The 10-fold cross-validation was used to test the accuracy of the classification model. The clustering results were visualized by a confusion matrix diagram, and the prediction ability of the model was evaluated by true positive rate (TPR), false negative rate (FNR), positive predictive value (PPV), and false discovery rate (FDR), they were calculated using Eqs. (5), (6), (7), and (8). TP represents that the true class is positive, and the predicted class is also positive; FN represents that the true class is positive, but the predicted class is negative; and FP represents that the predicted class is positive, and the true class is negative (Zhang et al., 2023). Generally, higher values of TPR and PPV, as well as lower values of FNR and FDR, reflect the better predictive capability of the model, and the all-around performance of each classification model was judged by receiver operating characteristic (ROC) curve, a larger area under the ROC curve indicates the improved performance of the model.

$$TRP = \frac{TP}{TP + FN} \tag{5}$$

$$FNR = \frac{FN}{FN + TP} \tag{6}$$

$$PPV = \frac{TP}{TP + FP} \tag{7}$$

$$FDR = \frac{FP}{TP + FP} \tag{8}$$

2.8. Establishment and evaluation of PLSR model

The Partial Least Squares Regression (PLSR) algorithm was executed through MATLAB 2021a software (MathWorks Inc., Natick, MA, USA) to achieve quantitative prediction of adulteration levels. The Kennard and Stone (KS) algorithm was employed to divide all pure and adulterated samples into calibration and prediction sets at a ratio of 7:3, respectively, for the establishment and performance prediction of partial least squares regression (PLSR) models.

The accuracy of the regression model was evaluated by calibration determination coefficient ($R^2{}_C$), prediction determination coefficient ($R^2{}_p$), root mean square error of calibration (RMSEC), root mean square error of prediction (RMSEP), RMSEP/RMSEC ratio, and relative percent deviation (RPD). Generally, a model is considered to have good predictive performance when $R^2 > 0.8$, an appropriate fit is achieved when RMSEP/RMSEC ratio falls within 0.8–1.2, and a model is deemed reliable with an RPD greater than 2.0, which indicates its practical applicability for predicting and analyzing results (Guan, Ye, Yi, Hua, & Chen, 2022). Additionally, an excessive number of latent variables (LVs) can lead to overfitting of the PLSR model, while too few LVs can result in underfitting, significantly impacting the predictive performance of the model. Through cross-validation, the optimal value of LVs is determined by selecting the LVs value corresponding to the minimum RMSECV.

3. Results and discussion

3.1. Extraction and analysis of chromaticity value

3.1.1. The chromaticity characteristics analysis of adulterated samples

The chromaticity characteristics (L^*, a^*, b^*, E_{ab}, C, H^*, and CI) of ZSS and adulterated samples were obtained from the chroma extractor, precision, stability, and repeatability of the method were tested to observe the potential interference caused by external light sources. As shown in Table 1S, the relative standard deviations (RSD) values obtained were all <2.0 %, indicating that the measurement method is reliable and can be applied to experimental analysis.

As the adulteration ratio increased, an upward trend in both L^* value and b^* value was observed in ZMS adulterants compared to pure ZSS, while a declining trend occurred in the HAS adulterants, and except for the 10 % ZMS adulterants, where slight change was presented compared to pure ZSS in terms of L^* value, significant changes were observed in all other groups (Fig. 1A&C). As for a^* value (Fig. 1B), no significant variations occurred in both ZMS and HAS adulteration samples compared to pure ZSS as the adulteration ratio increased.

It was indicated that, compared to naked eye observation (Fig. 1S), the chromaticity extractor can objectively quantify the chromatic differences and changing trends of ZSS and its adulterants for further analysis.

3.1.2. Qualitative identification of the ZSS and its adulterants

After analyzing the changing trend of chromaticity parameters with the adulteration ratio, the unsupervised PCA model was applied to perform dimensionality reduction and linear classification, aiming to achieve an initial qualitative classification between ZSS and adulterated samples. Based on Fig. 1D, it was evident that for the ZSS, ZMS, and HAS samples, the cumulative contribution rate of the principal components reached 98.92 %, which proved that ZSS could be well distinguished

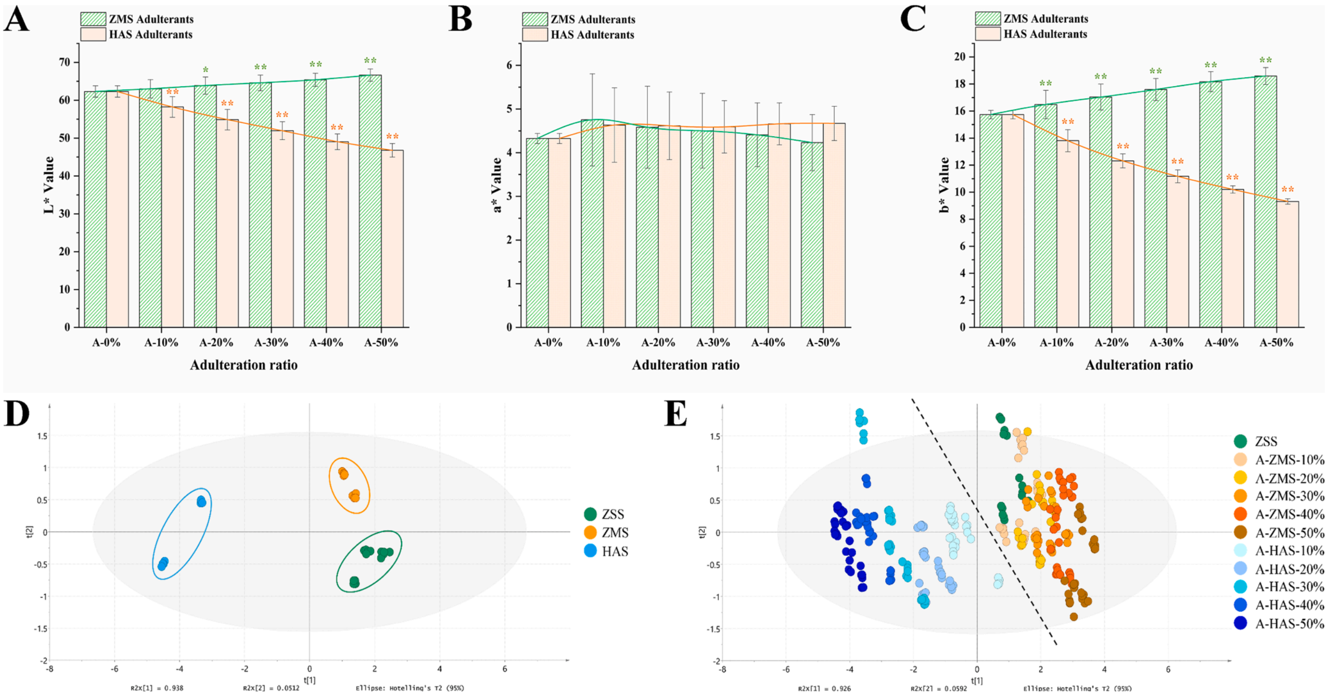


Fig. 1. The variations in L^* (A), a^* (B), and b^* (C) values of ZMS and HAS with different adulteration ratios. The PCA score plots of ZSS compared with ZMS & HAS (D) and adulterants with different ratios (E) based on chromaticity characteristics. $P^* < 0.05$, $P^{**} < 0.01$ vs. A-0 % group (Abbreviations: ZSS, Ziziphi Spinosae Semen; ZMS, Ziziphi Mauritianae Semen; HAS, Hovenia Acerba Semen).

from both ZMS and HAS. However, when comparing ZSS with adulterants at different proportions (Fig. 1E), it could be observed that two distinct regions can be formed between ZMS and HAS adulterants samples, and pure ZSS can be clearly distinguished from HAS adulterated samples, but not completely differentiated from ZMS adulterated samples. From the classification results of PCA, it can be observed that relying solely on chromatic analysis cannot achieve a good distinction between samples with different adulteration ratios.

Based on the preliminary classification, rapid discrimination formula, utilizing chromaticity parameters and Bayesian linear discriminant analysis method, was established to distinguish ZSS and its adulterants (Table 2S). The reliability of the mathematical discrimination formula was verified through cross-validation (Table 3S). The results indicated that the cross-validation rate for different ratios of HAS adulterants was 99.4 %. The method established could quickly identify HAS adulteration samples, while the classification accuracy for different proportions of ZMS adulterants was only 60 %, with a cross-validation rate of only 56.4 %.

In general, the chromatic features can be used to preliminarily determine whether ZSS was adulterated, but there were certain limitations when it came to accurately discerning the adulteration ratios of ZSS.

3.2. Analysis of FT-NIR raw spectral information

3.2.1. FT-NIR spectral features

In the FT-NIR spectrum, significant absorption peaks are primarily raised from O—H, C—H, C—C, C═C, and C═O functional groups. The spectra of ZSS, ZMS, HAS, and various adulterated samples all contained these functional groups. As shown in Fig.2A, the spectrum exhibited 6 characteristic absorption peaks. The broad absorption peak around 8300 cm^{-1} was caused by the second overtone absorption of C—H stretching vibrations (Laouni, El, Elhamdaoui, Karrouchi, El, & Bouatia, 2023). The absorption peak near 6920 cm^{-1} was attributed to the first overtone absorption of O—H stretching vibrations (Zhang et al., 2021). The broad absorption peak around 5650 cm^{-1} was induced by the first overtone absorption of C—H stretching vibrations (Zhan et al., 2017). The absorption peak near 5180 cm^{-1} may be due to the second overtone absorption of C═O stretching vibrations or a combination of O—H stretching and bending vibrations (Liu et al., 2019). The absorption peaks between 4000 and 4400 cm^{-1} were mainly caused by the combination frequencies of C—H, C-H_2, and C-H_3 (Wu et al., 2018).

3.2.2. Qualitative discrimination by PCA and machine learning algorithms

It can be observed that the spectral differences between ZSS and adulterated samples were minimal, making it difficult to distinguish them based solely on raw spectral information. Therefore, the utilization of classification models was required for further implementation of categorization. The results of PCA exhibited similarity to the results of chromatic discrimination. Three distinct regions were observed in the separation of pure ZSS, ZMS, and HAS samples, with a cumulative contribution rate of 97.4 % for principal components 1 and 2 (Fig. 2D). This allowed for the classification between pure and adulterated samples, but successful discrimination among samples with different adulteration ratios cannot be achieved (Fig. 2E).

Based on the preliminary classification, to obtain more accurate classification results for ZSS and its adulterants, three pattern recognition algorithms, including SVM, KNN, and ANN, were developed for in-depth analysis and data quantification, allowing for a more intuitive presentation of the classification results. Compared to the PCA model, SVM demonstrates strong generalization ability, can avoid the occurrence of overfitting effectively. KNN can exhibit high discriminative accuracy and is insensitive to outliers. ANN can possess strong distribution processing capability. Specifically, the classification accuracy of the three models for different ratios of ZMS adulterated samples is only 79.4 %, 57.0 % and 75.8 % (Fig. 2F), while for HAS adulterated samples, the classification accuracy is 83.0 %, 83.0 % and 84.2 % (Fig. 2G), respectively. Overall, it can be indicated from the classification results of machine learning algorithm that ZSS and its adulterants with different ratios cannot be accurately classified based on the information of raw spectra. However, the large number of variables in FT-NIR spectroscopy often results in the inclusion of instrument noise and irrelevant

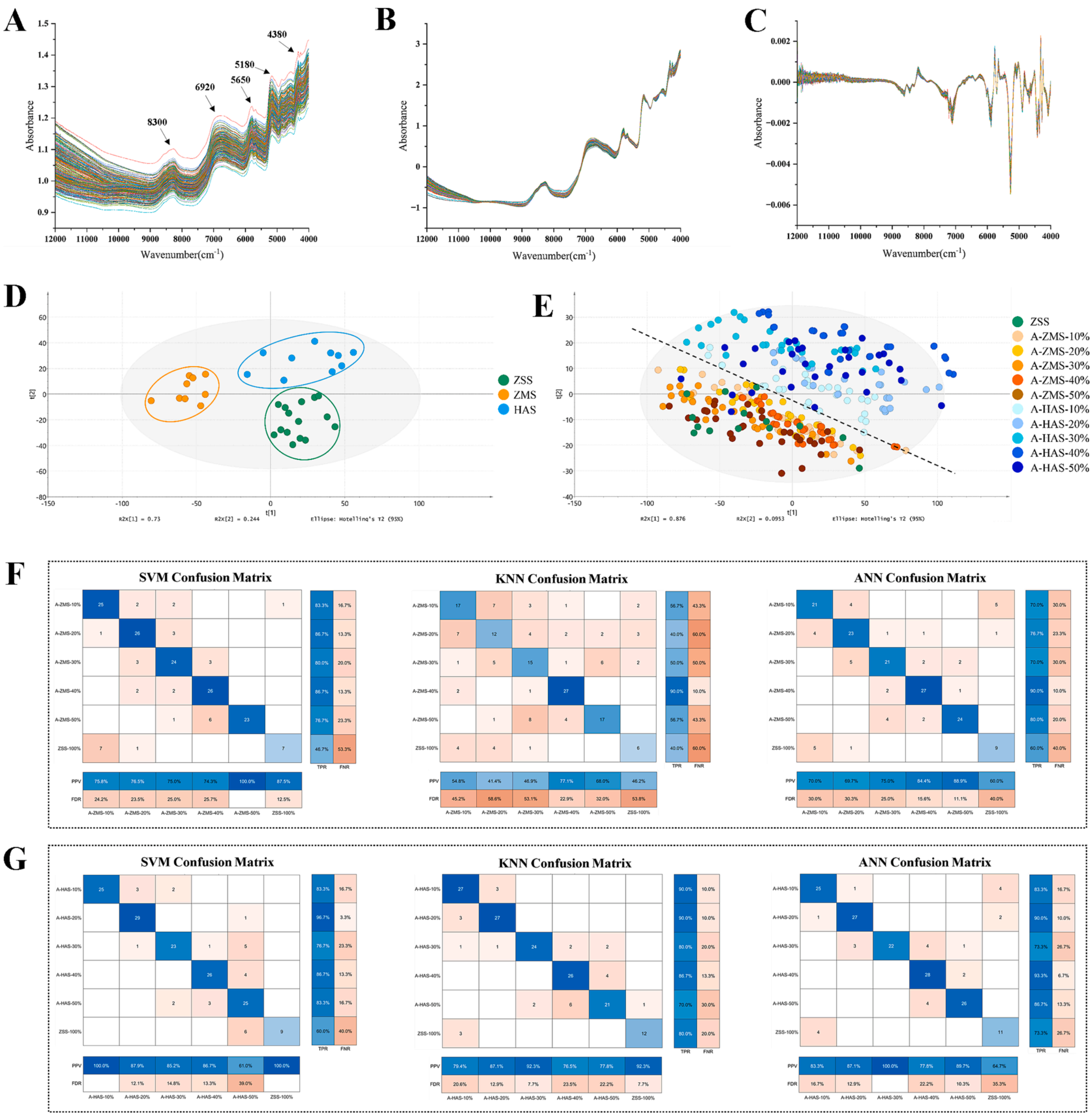


Fig. 2. Raw FT-NIR spectra of all pure and adulterated samples (A). FT-NIR spectra of ZMS adulterants samples preprocessed by SNV (B) and HAS adulterants preprocessed samples by 1d + SG (C). The PCA score plots of pure ZSS compared with pure ZMS & HAS (D) and ZMS & HAS with different adulteration ratios based on raw spectrum of FT-NIR(E). The confusion matrices and model score of SVM, KNN and ANN for ZMS (F) and HAS (G) adulterants based on raw spectrum of FT-NIR (Abbreviations: ZSS, Ziziphi Spinosae Semen; ZMS, Ziziphi Mauritianae Semen; HAS, Hovenia Acerba Semen; SVM, Support Vector Machine; KNN, K-Nearest Neighbors; ANN, Artificial Neural Network.).

variables, which significantly disrupts the accuracy of classification and prediction model establishment. Therefore, further preprocessing of the raw spectra and selection of characteristic wavelengths are required to optimize the spectral information and extract Key variables to enhance the overall quality of model construction.

3.3. *Processing and optimization of FT-NIR raw spectral information*

3.3.1. *Spectra preprocessing*

The raw spectra of ZMS and HAS samples with various adulteration ratios were subjected to optimization using 12 preprocessing methods, including SNV, MSC, SG, 1st-derivative, 2nd-derivative, and their combinations. The prediction accuracy and performance parameters were employed to evaluate the quality of the models, and the results are presented in Table 1. The model established based on the raw spectral data of ZMS adulterants achieved $R^2_C = 0.9587$, $R^2_p = 0.9614$, RPD = 5.23. After the SNV preprocessing method was applied, the R^2_C and R^2_p of the model improved to 0.9777 and 0.9766, respectively. The RMSEP/RMSEC ratio was 0.99, and RPD increased to 7.04, outperforming other preprocessing methods. These results indicated that the SNV method is most suitable for preprocessing the raw spectra of ZMS samples with different adulteration proportions. In the case of adulterated samples

Table 1
The results of different spectra preprocessing methods.

Adulteration category	Method	*LVs*	Calibration		Prediction		*RMSEP/RMSEC*	RPD
			$R^2{}_C$	*RMSEC*	$R^2{}_p$	*RMSEP*		
Adulterated with ZMS	RAW	5	0.9587	0.2782	0.9614	0.2917	1.05	5.23
	SG	6	0.9693	0.2443	0.9697	0.2446	1.00	5.95
	1d	4	0.9940	0.1082	0.9381	0.3277	3.03	4.38
	1d + SG	2	0.9577	0.28901	0.9152	0.3549	1.23	3.71
	2d	1	0.7179	0.6722	0.0623	0.8855	1.32	1.53
	2d + SG	1	0.7411	0.6442	0.4886	0.7547	1.17	1.86
	SNV	**5**	**0.9777**	**0.2090**	**0.9766**	**0.2068**	**0.99**	**7.04**
	SNV + 1d	1	0.9023	0.4290	0.8727	0.4261	0.99	3.20
	SNV + 2d	1	0.7366	0.6396	0.2218	0.8281	1.29	1.78
	MSC	5	0.9763	0.2162	0.9756	0.2164	1.00	6.65
	MSC + 1d	1	0.9021	0.4291	0.8288	0.4714	1.10	2.86
	MSC + 2d	1	0.7366	0.6396	0.2218	0.8281	1.37	1.78
Adulterated with HAS	RAW	6	0.9755	0.2070	0.9669	0.3008	1.45	5.37
	SG	7	0.9748	0.2242	0.9716	0.2348	1.05	5.98
	1d	2	0.9553	0.2966	0.8646	0.4566	1.54	3.05
	1d + SG	**5**	**0.9793**	**0.1964**	**0.9709**	**0.2005**	**1.12**	**6.26**
	2d	3	0.9065	0.1964	0.1321	0.9918	2.38	1.42
	2d + SG	3	0.7376	0.6603	0.4651	0.7613	1.15	1.77
	SNV	2	0.7294	0.6573	0.6175	0.7601	1.16	1.86
	SNV + 1d	1	0.8039	0.5901	0.7326	0.5947	1.01	2.27
	SNV + 2d	1	0.4960	0.8432	0.1195	1.0443	1.24	1.25
	MSC	5	0.9710	0.2328	0.97141	0.2414	1.04	6.47
	MSC + 1d	2	0.9567	0.2850	0.9366	0.3360	1.18	4.29
	MSC + 2d	1	0.6217	1.1760	0.2683	0.9293	0.79	1.54

The bold font: Optimal data processing combination.

with HAS, the preprocessing method combining 1d + SG demonstrated the best performance with $R^2{}_C = 0.9793$, $R^2{}_p = 0.9709$, RMSEP/RMSEC = 1.12, and RPD = 6.26, outperforming other processing methods.

Therefore, for subsequent analysis of the raw spectral data of ZMS and HAS adulterated samples, the SNV and 1d + SG methods were employed for preprocessing respectively. The optimized FT-NIR spectra were shown in Fig. 2B&C. It can be found that after spectral pretreatment, the originally complex spectrum became clearer and more concise, which is convenient for follow-up analysis.

Abbreviations: 1d, 1st-derivative; 2d, 2nd-derivative; SG, Savitzky Golay; MSC, Multiplicative Scattering Correction; SNV, Standard Normal Variate; R^2_C, Calibration Determination Coefficient; R^2_P, Prediction Determination Coefficient; RMSEC, Root Mean Square Error of Calibration; RMSEP, Root Mean Square Error of Prediction; RPD, Relative Percent Deviation;

3.3.2. *Characteristic wavelength selection*

After preprocessing, the characteristic wavelength selection algorithms were further used to extract the key spectral information. The prediction performance of samples with different adulteration ratios of ZMS and HAS was evaluated based on Full-PLS, ICO-PLS, CARS-PLS, and RF-PLS models.

As shown by Table 4S, compared with Full-PLS, both ICO-PLS and CARS-PLS improved the prediction performance of the model with the different adulteration ratio of ZMS and HAS samples, while RF-PLS had no obvious improvement, and even the prediction parameters of its calibration set decrease, which showed that the key wavelength extracted by ICO and CARS wavelength selection algorithms was effective. However, it was worth noting that when CARS-PLS and RF-PLS models were used to predict and evaluate the adulteration ratio of HAS, the RMSEP/RMSEC values were 1.89 and 0.69 respectively, which was not in the appropriate range of 0.8–1.2, indicating that the above two models did not have a reasonable degree of fitting in this case, and the corresponding wavelength selection algorithms were not suitable for extracting key information from the spectra of HAS adulterated samples. For ZMS samples with different adulteration ratios, after processed with the CARS algorithm, the model achieved $R^2{}_C = 0.9945$, $R^2{}_p = 0.9951$, RMSEP/RMSEC = 1.02, and RPD = 13.9989. These results indicated that the model was reliable and exhibited good predictive performance as well as an appropriate level of fit. After processed with the ICO algorithm, the $R^2{}_C$, $R^2{}_p$, and RPD parameters for adulterated HAS samples were superior to the results of the other three models.

Therefore, it can be concluded that using the CARS and ICO algorithms to extract characteristic wavelengths from ZMS and HAS samples with different adulteration ratios respectively produces the most desirable outcomes.

3.3.2.1. The application of CARS algorithm for ZMS adulterants. CARS algorithm is a classical wavelength selection algorithm, which has been widely used in the field of food adulteration identification (Li et al., 2017; Weng et al., 2020). Fig.3A represented the changing trend of number of sampled variables, RMSECV and regression coefficients path with the number of samplings runs respectively. The number of selected wavelengths decreased rapidly when the number of samplings runs increase from 0 to 10, and then tended to be smooth, which reflected that CARS algorithm can not only extract the spectral data quickly, but also refine the spectral information on this basis. During the early stages of operation, the mistaken characteristic spectral information was eliminated by the CARS algorithm. A decreasing trend followed by an increase can be observed in the value of RMSECV. When the iteration count reached 53 times (corresponding to the blue line in Fig. 3A (3)), the RMSECV value reached its minimum. At this point, a total of 72 characteristic wavelengths were extracted, which represented a significant reduction compared to the raw spectrum's complex wavelength information (Fig. 3B), and the red squares indicated the characteristic wavelengths that had been extracted by the CARS algorithm.

3.3.2.2. The application of ICO algorithm for HAS adulterants. In the iterative process of ICO algorithm, the weight coefficient of each wavelength interval changes with the increase of the number of iterations. The more yellow the color is, the closer the weight coefficient value is to 1, the bluer the color is, the closer the weight coefficient value is to 0. If the color is between blue and yellow, the weight coefficient value is between 0 and 1 (Song, Huang, Yan, Xiong, & Min, 2016). As

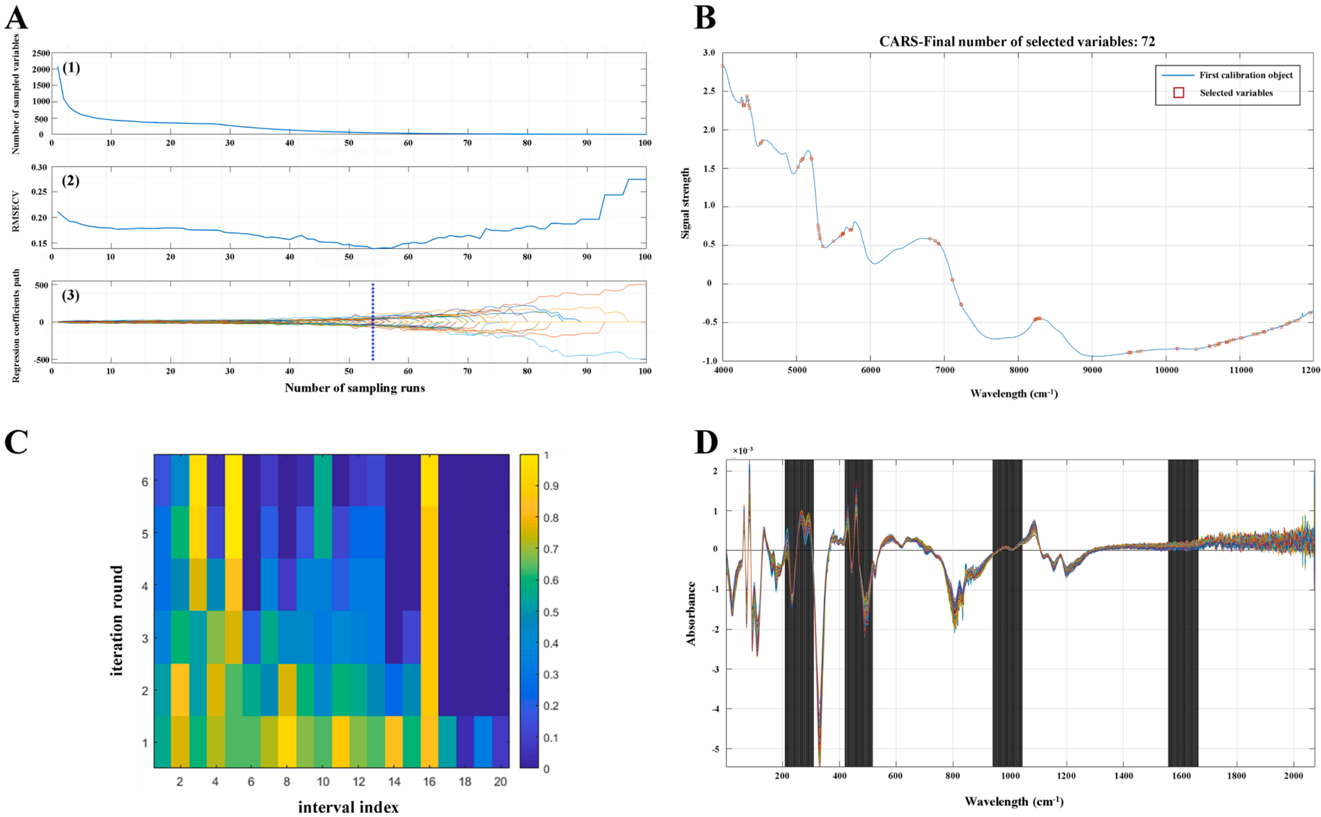


Fig. 3. The results of wavelength selection with CARS algorithm for ZMS adulterants: (A) (1) changes in the number of selected variables, (2) variation of RMSECV, (3) path of variable regression coefficients. (B) Characteristic wavelength selection results. The results of wavelength selection with ICO algorithm for HAS adulterants: (C) Sampling weights for each feature interval in the optimization process. (D) Characteristic intervals selected by ICO algorithm (Abbreviations: CARS, Competitive Adaptive Reweighted Sampling; ICO, Interval Combination Optimization.).

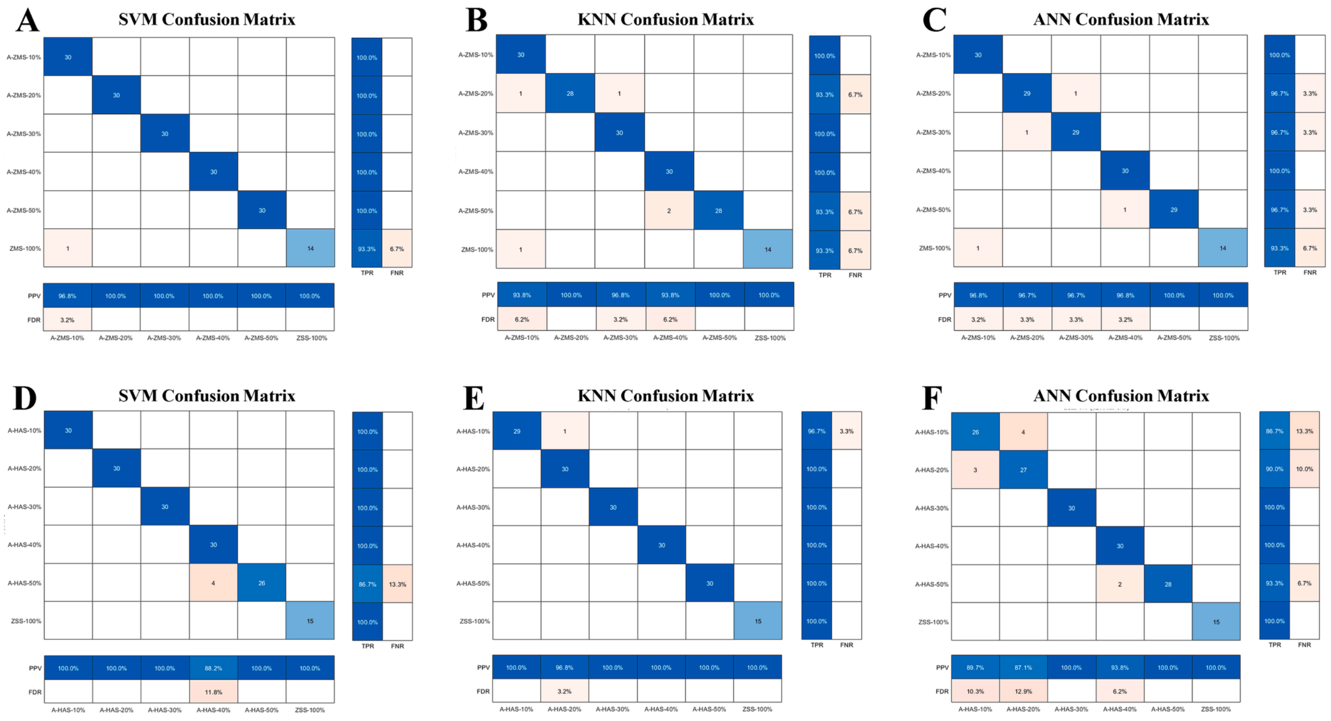


Fig. 4. The results of classification model after spectrum information processing: SVM (A), KNN (B) and ANN (C) model for pure ZSS and ZMS with different adulteration ratio; SVM (D), KNN (E) and ANN (F) model for pure ZSS and HAS with different adulteration ratio (Abbreviations: ZSS, Ziziphi Spinosae Semen; ZMS, Ziziphi Mauritianae Semen; HAS, Hovenia Acerba Semen; SVM, Support Vector Machine; KNN, K-Nearest Neighbors; ANN, Artificial Neural Network.).

shown by Fig.3C, took the fifth wavelength interval as an example, with the increase of the number of iterations, the weight coefficient was increased accordingly, and the weight coefficient reached 1 in the fifth iteration, so this wavelength interval was finally selected. At the same time, for the sixth wavelength range, with the increase of the number of iterations, the weight coefficient showed a downward trend, and most of the weight coefficients were between 0 and 0.5, so this interval was abandoned. On the contrary, the weight coefficient of the 16th wavelength interval was improved with the increase of the iteration's rounds, which was between 0.8 and 1, so it was selected. The final selected wavelength interval was shown in Fig. 3D. In the joint interval selected by ICO algorithm, the local search strategy was introduced to optimize the width automatically, and the total number of characteristic wavelengths selected was 630.

3.4. Analysis of FT-NIR spectral after preprocessing and wavelength selection

3.4.1. Qualitative discrimination based on machine learning algorithms

After preprocessing and characteristic wavelength selection, the background noise and unrelated interference of FT-NIR raw spectra were eliminated and the key spectral information was extracted. Based on this, classification models were established again using machine learning algorithms for both the pure ZSS and the samples with various adulteration ratios of ZMS and HAS to visually demonstrate the effectiveness and necessity of spectral preprocessing and wavelength selection.

From Fig. 4 A–C, it can be observed that the SVM, KNN, and ANN models can accurately classify the pure ZSS and ZMS adulterants, which raw spectra was processed by the SNV + CARS method. The classification accuracy of these three models reached 99.4 %, 97.0 %, and 97.6 % respectively, representing an improvement of 20.0 %, 40.0 %, and 21.8 % compared to the classification results of raw spectra, indicating their strong discriminative capability. Moreover, the area under the ROC curves for all three models was 1.00 (Table 5S), further demonstrating the reliability and effectiveness of the models. Similarly, for both the pure ZSS and HAS adulterants processed with 1d + SG + ICO method, the aforementioned three models also demonstrated excellent discriminative ability, with classification accuracies improved to 97.6 %, 99.4 %, and 94.5 %, respectively (Fig. 4 D–F). The areas under the ROC curves were all 1.00 (Table 5S), indicating that the established classification models were suitable and exhibit high discriminative accuracy.

The above results indicated that preprocessing and wavelength selection of FT-NIR raw spectra contribute to extracting valuable bands from complex spectral information, thereby improving the classification accuracy of ZSS samples with different adulteration ratios. This not only provided data processing technical support for the rapid and accurate differentiation of adulterated samples but also validated the necessity of spectral preprocessing and characteristic wavelength selection.

3.4.2. Quantitative prediction for adulteration ratio based on PLSR model

After pretreatment and characteristic wavelength selection, the optimized FT-NIR spectra of ZMS and HAS adulterated samples were obtained. On this basis, the samples with different adulteration ratios were quantitatively predicted and analyzed. Generally, if the sample is closely distributed near the regression line, it means that the establishment of the regression model is successful and has good prediction ability. In the quantitative prediction model based on FT-NIR spectroscopy (Fig. 5 A & B), the sample points were closely clustered around the regression line, with $R^2{}_C = 0.9924$, $R^2{}_p = 0.9920$ (for ZMS adulteration) and $R^2{}_C = 0.9965$, $R^2{}_p = 0.9952$ (for HAS adulteration). This further confirmed the reliability of the PLSR content prediction model based on FT-NIR spectra, indicating its strong performance in quantitatively predicting adulteration ratios, it also provides a positive technical for the quantitative prediction of ZSS adulteration ratio.

4. Conclusion

In this study, FT-NIR was first combined with multivariate algorithms to explore a new approach to address the increasingly severe issue of adulteration in ZSS. By preprocessing the NIR raw spectra and selecting wavelengths, the qualitative classification accuracy of adulterated samples and the quantitative prediction ability of adulteration ratios were improved, ultimately enabling a green and rapid evaluation of ZSS adulterants.

Compared to the observation of naked eyes, although chroma extractor can objectively quantify the chroma of adulterated samples and determine the changing trend, but relying solely on chromaticity features can only preliminarily determine whether ZSS is adulterated, with certain limitations in accurately identifying different adulteration ratios. FT-NIR spectra contains numerous physicochemical information and can reflect the absorption and vibrations of hydrogen-containing groups such as X-H (X = C, N, O), which includes the composition and molecular structure of most types of organic compounds. Typically, FT-NIR spectroscopy can provide specific spectral information of a sample within a few seconds, indirectly reflecting its chemical composition. Based on this, the potential of FT-NIR in identifying and quantitatively predicting the adulteration ratios were explored. After obtaining the raw spectral information, 12 preprocessing methods and 3 wavelength selection methods were used for optimization and comparison. Finally, the SNV + CARS method was applied to optimize the spectral information of ZMS adulterants samples, with the RMSEP/RMSEC ratio of 0.99 and an

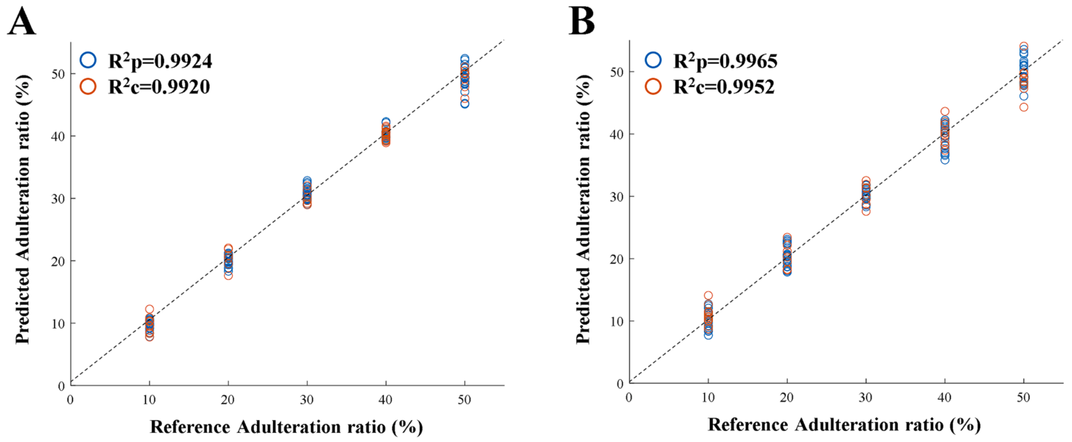


Fig. 5. Scatter plots between the predicted and reference values: Regression model Prediction for ZMS (A) and HAS (B) adulteration ratio based on FT-NIR (Abbreviations: R^2_C, Calibration Determination Coefficient; R^2_P, Prediction Determination Coefficient).

RPD value of 7.04. The 1d + SG + ICO method was applied to optimize the spectral data of HAS adulterants samples, with the RMSEP/RMSEC ratio of 1.12 and an RPD value of 6.26. Both two optimization methods exhibit good reliability and applicability. Compared to the raw spectra, the average classification accuracy of machine learning models for different adulteration ratios samples increased from 77.06 % to 97.58 % after spectral processing. Both two PLSR models achieved R^2 values exceeding 0.99 for the calibration and prediction sets, indicating good linearity and precision of the regression curves, based on the latest research advancements (Zhang et al., 2023), this study further achieves precise classification and discrimination of different counterfeit ratios. Additionally, quantitative prediction of the counterfeit ratio is accomplished. Compared to previous studies, this research achieves a balance between low cost and high accuracy.

In summary, compared to traditional FT-NIR detection methods, incorporating multiple algorithms such as spectral preprocessing and wavelength selection allowed for more accurate and comprehensive monitoring of adulteration in ZSS, enabling rapid tracking of adulteration and ensuring the authenticity of ZSS, which provided a beneficial solution to maintain stability in food consumption market and safeguard the interests and health of consumers. With the development of modern bionic technology, technologies such as e-nose and e-eye have been preliminarily applied in the food industry, but currently limited to the use of conventional chemometrics for the analysis of raw data (Fei et al., 2022; Fei et al., 2021). In the future, further development of characteristic data processing algorithms can enhance the application accuracy of bionic technology. In future practical applications, further emphasis should be placed on the lightweight design of instruments and the universality research of counterfeit detection, to address the increasingly severe issue of counterfeiting with non-destructive, convenient, and accurate detection.

CRediT authorship contribution statement

Ming-xuan Li: Conceptualization, Methodology, Software, Writing – original draft. **Ya-bo Shi:** Visualization, Investigation, Methodology, Formal analysis. **Jiu-ba Zhang:** Project administration, Writing – review & editing. **Xin Wan:** Project administration, Writing – review & editing. **Jun Fang:** Project administration, Writing – review & editing. **Yi Wu:** Methodology, Resources. **Rao Fu:** Methodology, Resources. **Yu Li:** Methodology, Resources. **Lin Li:** Investigation, Software. **Lian-lin Su:** Data curation, Formal analysis. **De Ji:** Data curation, Formal analysis. **Tu-lin Lu:** Data curation, Funding acquisition, Writing – review & editing. **Zhen-hua Bian:** Data curation, Funding acquisition, 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

The data that has been used is confidential.

Acknowledgments

This research study was supported by the Natural Science Foundation of Nanjing University of Chinese Medicine (Grant No. XZR2021067) and the “Double Thousand Plan” Project of Jiangxi Province’s in China (Grant No. S2019CXTD2300).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.fochx.2023.101022.

References

Basri, K. N., Hussain, M. N., Bakar, J., Sharif, Z., Khir, M., & Zoolfakar, A. S. (2017). Classification and quantification of palm oil adulteration via portable NIR spectroscopy. *Spectrochimica Acta. Part A, Molecular and Biomolecular Spectroscopy, 173*, 335–342.

Beć, K. B., Grabska, J., & Huck, C. W. (2022a). In silico NIR spectroscopy - A review. Molecular fingerprint, interpretation of calibration models, understanding of matrix effects and instrumental difference. *Spectrochimica Acta. Part A, Molecular and Biomolecular Spectroscopy, 279*, Article 121438.

Beć, K. B., Grabska, J., & Huck, C. W. (2022b). Miniaturized NIR spectroscopy in food analysis and quality control: Promises, challenges, and perspectives. *Foods, 11*(10).

Beć, K. B., & Huck, C. W. (2019). Breakthrough potential in near-infrared spectroscopy: Spectra simulation. A review of recent developments. *Frontiers Chemistry, 7*, 48.

Cebi, N., Bekiroglu, H., Erarslan, A., & Rodriguez-Saona, L. (2023). Rapid sensing: Hand-held and portable FTIR applications for on-site food quality control from farm to fork. *Molecules, 28*(9).

Cozzolino, D. (2021). The ability of near infrared (NIR) spectroscopy to predict functional properties in foods: Challenges and opportunities. *Molecules, 26*(22).

Fei, C., Ren, C., Wang, Y., Li, L., Li, W., Yin, F., … Yin, W. (2021). Identification of the raw and processed Crataegi Fructus based on the electronic nose coupled with chemometric methods. *Scientific Reports, 11*(1), 1849.

Fei, C., Xue, Q., Li, W., Xu, Y., Mou, L., Li, W., … Yin, F. (2022). Variations in volatile flavour compounds in Crataegi fructus roasting revealed by E-nose and HS-GC-MS. *Frontiers in Nutrition, 9*, 1035623.

Gizaw, Z. (2019). Public health risks related to food safety issues in the food market: A systematic literature review. *Environmental Health and Preventive Medicine, 24*(1), 68.

Guan, Y., Ye, T., Yi, Y., Hua, H., & Chen, C. (2022). Rapid quality evaluation of Plantaginis Semen by near infrared spectroscopy combined with chemometrics. *Journal of Pharmaceutical and Biomedical Analysis, 207*, Article 114435.

Huang, F., Song, H., Guo, L., Guang, P., Yang, X., Li, L., … Yang, M. (2020). Detection of adulteration in Chinese honey using NIR and ATR-FTIR spectral data fusion. *Spectrochimica Acta. Part A, Molecular and Biomolecular Spectroscopy, 235*, Article 118297.

Kulapichitr, F., Borompichaichartkul, C., Fang, M., Suppavorasatit, I., & Cadwallader, K. R. (2022). Effect of post-harvest drying process on chlorogenic acids, antioxidant activities and CIE-Lab color of Thai Arabica green coffee beans. *Food Chemistry, 366*, Article 130504.

Laouni, A., El, O. A., Elhamdaoui, O., Karrouchi, K., El, K. M., & Bouatia, M. (2023). A preliminary study on the potential of FT-IR spectroscopy and chemometrics for tracing the geographical origin of moroccan virgin olive oils. *Journal of AOAC International, 106*(3), 804–812.

Li, H. D., Xu, Q. S., & Liang, Y. Z. (2012). Random frog: An efficient reversible jump Markov Chain Monte Carlo-like approach for variable selection with applications to gene selection and disease classification. *Analytica Chimica Acta, 740*, 20–26.

Li, H., Liang, Y., Xu, Q., & Cao, D. (2009). Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration. *Analytica Chimica Acta, 648*(1), 77–84.

Li, S., Zhang, X., Shan, Y., Su, D., Ma, Q., Wen, R., & Li, J. (2017). Qualitative and quantitative detection of honey adulterated with high-fructose corn syrup and maltose syrup by using near-infrared spectroscopy. *Food Chemistry, 218*, 231–236.

Liu, X., Zhang, S., Ni, H., Xiao, W., Wang, J., Li, Y., & Wu, Y. (2019). Near infrared system coupled chemometric algorithms for the variable selection and prediction of baicalin in three different processes. *Spectrochimica Acta. Part A, Molecular and Biomolecular Spectroscopy, 218*, 33–39.

Medeiros, M., Freitas, L. A., Correia, G. M., Teixeira, G. H., & Fernandes, B. D. (2023). Portable near-infrared (NIR) spectrometer and chemometrics for rapid identification of butter cheese adulteration. *Food Chemistry, 425*, Article 136461.

Qu, J. H., Liu, D., Cheng, J. H., Sun, D. W., Ma, J., Pu, H., & Zeng, X. A. (2015). Applications of near-infrared spectroscopy in food safety evaluation and control: A review of recent research advances. *Critical Reviews in Food Science and Nutrition, 55*(13), 1939–1954.

Saadat, S., Pandya, H., Dey, A., & Rawtani, D. (2022). Food forensics: Techniques for authenticity determination of food products. *Forensic Science International, 333*, Article 111243.

Song, X., Huang, Y., Yan, H., Xiong, Y., & Min, S. (2016a). A novel algorithm for spectral interval combination optimization. *Analytica Chimica Acta, 948*, 19–29.

Sun, H., Lu, W., & Gao, B. (2021). Non-targeted detection of butter adulteration using pointwise UHPLC-ELSD and UHPLC-UV fingerprints with chemometrics. *Food Chemistry, 356*, Article 129604.

Wang, D., Ho, C. T., & Bai, N. (2022). Ziziphi Spinosae Semen: An updated review on pharmacological activity, quality control, and application. *Journal of Food Biochemistry, 46*(7), e14153.

Wang, G., Bai, X., Chen, X., Ren, Y., Pang, X., & Han, J. (2022). Detection of adulteration and pesticide residues in Chinese patent medicine Qipi Pill using KASP technology and GC-MS/MS. *Frontiers in Nutrition, 9*, Article 837268.

Weng, S., Guo, B., Tang, P., Yin, X., Pan, F., Zhao, J., … Zhang, D. (2020). Rapid detection of adulteration of minced beef using Vis/NIR reflectance spectroscopy with multivariate methods. *Spectrochimica Acta Part A, Molecular and Biomolecular Spectroscopy, 230*, Article 118005.

Wu, L., Su, Y., Yu, H., Qian, X., Zhang, X., Wang, Q., … Cheng, G. (2018). Rapid determination of saponins in the honey-fried processing of Rhizoma Cimicifugae by near infrared diffuse reflectance spectroscopy. *Molecules, 23*(7).

Xiao, Y., Wang, H., Xie, Z., Shen, M., Huang, R., Miao, Y., … Huang, W. (2022). NIR TADF emitters and OLEDs: Challenges, progress, and perspectives. *Chemical Science, 13*(31), 8906–8923.

Yan, H., Li, P. H., Zhou, G. S., Wang, Y. J., Bao, B. H., Wu, Q. N., & Huang, S. L. (2021). Rapid and practical qualitative and quantitative evaluation of non-fumigated ginger and sulfur-fumigated ginger via Fourier-transform infrared spectroscopy and chemometric methods. *Food Chemistry, 341*(Pt 1), Article 128241.

Yang, M., Wang, H., Zhang, Y. L., Zhang, F., Li, X., Kim, S. D., … Mao, J. J. (2023). The Herbal Medicine Suanzaoren (Ziziphi Spinosae Semen) for sleep quality improvements: A systematic review and meta-analysis. *Integrative Cancer Therapies, 22*, 1563477216.

Yong, C. H., Muhammad, S. A., Aziz, F. A., Nasir, F. I., Mustafa, M. Z., Ibrahim, B., … Seow, E. K. (2022). Detecting adulteration of stingless bee honey using untargeted ^{1}H NMR metabolomics with chemometrics. *Food Chemistry, 368*, Article 130808.

Zhan, H., Fang, J., Tang, L., Yang, H., Li, H., Wang, Z., … Fu, M. (2017). Application of near-infrared spectroscopy for the rapid quality assessment of Radix Paeoniae Rubra. *Spectrochimica Acta Part A, Molecular and Biomolecular Spectroscopy, 183*, 75–83.

Zhang, J. B., Li, M. X., Zhang, Y. F., Qin, Y. W., Li, Y., Su, L. L., … Lu, T. L. (2023). E-eye, flash GC E-nose and HS-GC-MS combined with chemometrics to identify the adulterants and geographical origins of Ziziphi Spinosae Semen. *Food Chemistry, 424*, Article 136270.

Zhang, J., Li, B., Hu, Y., Zhou, L., Wang, G., Guo, G., … Zhang, A. (2021). A parameter-free framework for calibration enhancement of near-infrared spectroscopy based on correlation constraint. *Analytica Chimica Acta, 1142*, 169–178.

Zhang, J., Li, Y., Wang, B., Song, J., Li, M., Chen, P., … Lu, T. (2023). Rapid evaluation of Radix Paeoniae Alba and its processed products by near-infrared spectroscopy combined with multivariate algorithms. *Analytical and Bioanalytical Chemistry, 415*(9), 1719–1732.