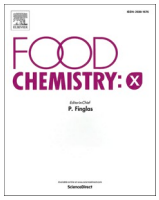
Food Chemistry: X 24 (2024) 101798

Contents lists available at ScienceDirect

Food Chemistry: X

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



# FT-NIR combined with machine learning was used to rapidly detect the adulteration of pericarpium citri reticulatae (*chenpi*) and predict the adulteration concentration

Ying Chen [a,1], Si Li [a,1], Jia Jia [a,1], Chuanduo Sun [b,1], Enzhong Cui [a], Yunyan Xu [a], Fangchao Shi [a,*], Anfu Tang [a,*]

[a] Department of Pharmacy, Jinling Hospital, Nanjing University School of Medicine, Nanjing, PR China
[b] Central Medical Branch of PLA General Hospital, PR China

ARTICLE INFO

*Keywords:*
Pericarpium citri reticulatae
Food adulteration
FT-NIR
Machine learning
Quantitative analysis

ABSTRACT

Pericarpium citri reticulatae (PCR) has been used as a food and spice for many years and is known for its rich nutritional content and unique aroma. However, price increases are often accompanied by adulteration. In this study, two kinds of adulterants (Orange peel-OP and Mandarin Rind-MR) were identified by chromaticity analysis, FT-NIR and machine learning algorithm, and the doping concentration was predicted quantitatively. The results show that colorimetric analysis cannot completely differentiate between PCR and adulterants. Using spectral preprocessing combined with machine learning algorithms, PCR and two adulterants were successfully distinguished, with classification accuracy reaching 99.30 % and 98.64 % respectively. After selecting characteristic wavelengths, the $R_P^2$ of the adulterated quantitative model is greater than 0.99. Generally, this study proposes to use FT-NIR to study the adulteration of PCR for the first time, which fills the technical gap in the adulteration research of PCR, and provides an important method to solve the increasingly serious adulteration problem of PCR.

## 1. Introduction

Pericarpium citri reticulatae (PCR) is the dried mature peel and its cultivated varieties of Citrus reticulate Blanco. It has been used as medicine and food for more than 2000 years (Pan et al., 2022). PCR mainly contains volatile oils, flavonoids and phenolic acids, and has good antiasthmatic, anti-inflammatory and antioxidant effects. In addition, because it is rich in nutrients such as vitamins and dietary fiber, it is not only widely used in food, beverage, condiment processing and fruit tea and fruit wine production processes (Yi et al., 2015), but also can be used as a popular dietary supplement (Ho & Kuo, 2014; Manthey et al., 2001).

Generally speaking, the longer the aging time of PCR, the stronger its nutritional content and medicinal effect, but the price is also more expensive (Bian et al., 2022; Luo et al., 2019). Therefore, unscrupulous merchants often use orange peels (OP) and mandarin rind (MR) of the same family to pass them off as PCR and sell them to make huge profits. This behavior not only seriously damages the economic interests of consumers, but also harms the health of consumers if consumed improperly, and is not conducive to the healthy development of the industry. Therefore, combating the adulteration of PCR has become an important and urgent issue.

When used as spices and food, PCR often needs to be processed into powder to make it easier to season and process. However, OP and MR are used for adulteration , their appearance and smell are very similar, making it difficult for consumers to accurately identify them through their senses. With the continuous advancement of science and technology, plant metabolomics methods such as UPLC-Q-TOF/MS and GC–MS can comprehensively analyze chemical components and identify metabolic differences as markers for identification (Duan et al., 2016; Shi et al., 2020; Wang et al., 2019). Unfortunately, these technologies all face challenges such as long detection cycles and complex sample preprocessing. Therefore, exploring rapid non-destructive testing methods is of great significance to combat adulteration of PCR.

Fourier transform near infrared (FT-NIR) spectroscopy has the advantages of being fast, convenient, accurate, and does not require any

* Corresponding authors.
*E-mail addresses:* 79382196@qq.com (F. Shi), ahf5499@sina.com (A. Tang).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.fochx.2024.101798
Received 24 July 2024; Received in revised form 30 August 2024; Accepted 30 August 2024
Available online 2 September 2024

pretreatment of samples, making it very popular in the fields of food and agriculture (Chen et al., 2023; Ditcharoen et al., 2023; Raghavendra et al., 2021). With the rapid development of artificial intelligence, the combination of NIR spectroscopy and machine learning algorithms has also shown great development potential in food adulteration detection. It has been used to predict adulteration of ginger powder (Yu et al., 2022), butter cheese (Medeiros et al., 2023), olive oil (Vieira et al., 2021) and other foods. However, to date, a large number of studies have focused on trace the origin and aging time of PCR, and there are almost no reports on PCR and its adulterants (Dai et al., 2023; Pu et al., 2023).

This study mainly uses 5 methods to preprocess NIR spectra, uses machine learning algorithms to distinguish PCR and two types of adulterants with different concentrations, and establishes a PLS regression model to quickly quantify and predict the concentrations of the two adulterants, and improve the accuracy of doping concentration quantification by selecting characteristic wavelengths. The purpose of this study is to explore a simple method to solve the adulteration of PCR, so as to fill the missing technical gap in the adulteration research of PCR.

## 2. Materials and methods

### 2.1. Sample collection

We purchased 20 batches of fresh PCR from Xinhui District of Guangdong Province (Authentic producing area of PCR). In addition, OP and MR used for adulteration are purchased from food markets. We use the same method to remove moisture from all samples and process them into dry samples for follow-up experiments. All the collected samples were identified by Food inspection agencies as the dried mature pericarp of *Citrus reticulata* Blanco of Rutaceae plant and its cultivated varieties, and the authenticity and reliability were ensured by food quality testing.

### 2.2. Sample preparation

First we processed all the samples into powder, sifted through 50 mesh and stored in a dry, sealed and dark environment. The adulteration rate of food and spices on the market is basically between 10 % and 50 %. This is because an adulteration rate of less than 10 % is unprofitable for businessmen, while adulteration of more than 50 % will be easily detected by consumers. Therefore, in this study, OP and MR of the same batch number were selected as adulterants and were randomly added to 20 batches of PCR in gradients of 10 %, 20 %, 30 %, 40 %, and 50 %, respectively, and the adulterated samples were mixed well with a vortex shaker to finally obtain 20 samples of each adulterant gradient.We simultaneously prepared 20 pure samples each of PCR, OP, and MR as controls, making the total number of samples in this study 260.

### 2.3. Chroma analysis

The chromaticity values of PCR and its adulterants were extracted by CM-5 spectrophotometer (KONICA MINOLTA, Tokyo, Japan). Firstly, the pulsed xenon lamp is set as the lighting source, the acquisition mode is set to SCE, the viewing angle is set to 10°, and the color data are collected within the wavelength range of 360–740 nm. The methodology was examined before the formal data collection, and the results are shown in Table S1. Among them, L* represents brightness; a* represents the red green value; b* represents yellow and blue values.

### 2.4. FT-NIR spectrum acquisition

The spectral information of the sample was collected by the Antaris II FT-NIR spectrometer (Thermo Fisher Scientific, USA), the sample was poured into a special quartz cup for determination, gently pressed to obtain a uniform filling density, the instrument was adjusted to diffuse mode, the 10,000–4000 $cm^{-1}$ spectral range was collected, the resolution was set to 16 $cm^{-1}$, each spectrum was scanned 32 times, and each sample was measured three times in parallel.

### 2.5. Spectral preprocessing

Near-infrared spectrum can reflect the overall characteristics, so it can not only solve the target problem, but also carry a lot of noise, baseline drift, spectral overlap and other problems, so we need spectral preprocessing to improve the accuracy and reliability of the model. In this study, multiplicative scattering correction (MSC) and standard normal variate (SNV) are used to eliminate the absorption shift and skew shift caused by light scattering (Y. Shi et al., 2023), and the first derivative (1d) and second derivative (2d) are used to eliminate vertical offset and linear tilt (X. Zhang et al., 2021). In addition, Savitzky-Golay (SG) algorithm is used to reduce the background noise of the instrument (Oliveira et al., 2020).Each preprocessing method has its own unique principle, and they can play different roles for the heterogeneity of different samples. In the case of MSC, for example, it first calculates the mean of all spectral data as a reference spectrum. A linear regression is then performed on each measured spectrum, fitting a linear model using the least squares method. Finally, the corrected spectra are calculated and the resulting regression coefficients a and b are used to correct the original spectra. Briefly, MSC calculates the covariance between the measured and reference spectra and the variance of the reference spectrum, and then calculates the regression coefficients. These coefficients are ultimately utilized to correct the raw spectra in order to reduce the influence of scattering effects in the spectral data and to improve the accuracy and reliability of the data.

### 2.6. Spectral characteristic wavelength selection

When the modeling effect after spectral preprocessing still does not achieve the desired results, we will select the characteristic wavelength. Its function is to eliminate the redundant information contained in the spectrum and reduce the data dimension. The wavelength selection algorithms of NIR spectral data mainly include wavelength interval selection algorithms such as interval partial least squares (iPLS) and interval combinatorial optimization (ICO)(Song et al., 2016), and wavelength point selection algorithms such as competitive adaptive reweighted sampling (CARS) and successive projections algorithm (SPA) (Araújo et al., 2001; Yuan et al., 2020). In addition, feature wavelengths can be extracted by random importance feature selection algorithms such as RF and VIP scores (VIPs). In this study, we use ICO, CARS, RF and VIPs to extract the feature wavelength.

### 2.7. Machine learning model

#### 2.7.1. Classification model

In recent years, with the rise of artificial intelligence, machine learning algorithms have shown great potential in food safety, raw material quality control and so on. In this study, we use five machine learning algorithms: Support vector machine (SVM), K-nearest-neighbor (KNN), random forest (RF), partial least squares discriminant analysis (PLS-DA) and gradient boosting machine (GBM) to quickly distinguish tangerine peel and its adulterants. All models are validated using ten fold cross validation. The performance of all models is evaluated by accuracy (Acc), precision (Pr), recall rate (Re), and F1 score (F1).

#### 2.7.2. Quantitative model

We choose PLS as the regression quantitative model, and use the correlation coefficient $R^2$, the root mean square error (RMSE), and the relative prediction deviation (RPD) as indicators to evaluate the accuracy and robustness of the model. Prior to this, we used the Kennard Stone algorithm to divide all samples into calibration and prediction sets in a 3:1 ratio. Generally speaking, the best model is the one that considers higher $R^2$ and lower RMSE (Ye et al., 2018). The RPD value reflects the overall predictive ability of the PLS regression model. When

the RPD value is greater than 3, it indicates excellent performance (Ndlovu et al., 2021).

## 3. Results and discussion

### 3.1. Analysis of appearance and color of PCR and its adulterants

As can be seen from Fig. 1A, the PCR powder is orange-yellow, the surface of the OP powder is yellow-white, and the surface of the MR powder is gray-yellow. It can be clearly seen with the naked eye that the surface colors of the three pure samples are significantly different. However, when PCR powder is mixed with different proportions (10 %, 20 %, 30 %, 40 %, 50 %) of OP and MR, the surface color of the sample will change to a certain extent. Specifically, when the adulteration ratio is low, the color is basically similar to that of PCR powder. As the adulteration ratio gradually increases, the color of the sample gradually changes to the adulterants, which makes it difficult for us to distinguish it with the naked eye.

In order to further study the color changes of PCR and its adulterants, we used colorimetric analysis technology to measure the colors of three pure samples and two adulterants with different proportions, and obtained the color values L*, a* and b*. In CIELAB chromaticity space, they represent brightness value, red-green value and yellow-blue value respectively. Among the three pure samples, PCR has the lowest L* value

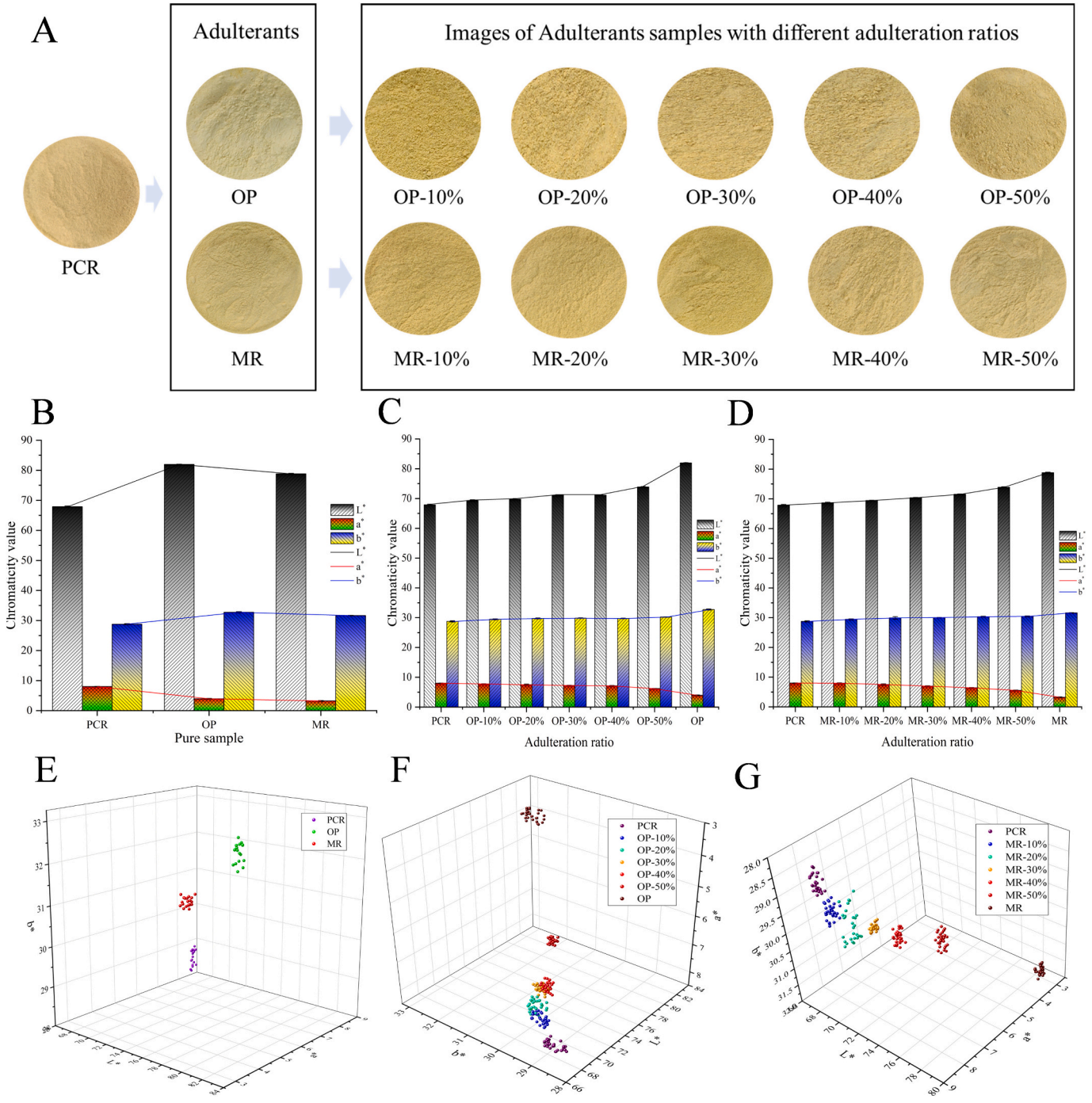



**Fig. 1.** Images of different proportions of adulterants (A), the color change trend of different proportions of adulterants (B-pure sample, C-adulterated with OP, D-adulterated with MR), color clustering of adulterants with different proportions (*E*-pure sample, F-adulterated with OP, G-adulterated with MR).

and b* value, and the highest a* value (Fig. 1B). As the adulteration ratio increases, whether it is adulterated with OP powder (Fig. 1C) or MR powder (Fig. 1D), the L* and b* values of the sample gradually increase, and the a* value gradually decreases. Interestingly, all color value changes were insignificant at different proportions of adulterants. This may be because PCR, OP and MR are all products of Rutaceae plants and have certain similarities in certain chemical components. Therefore, the color parameters of pure PCR and adulterants are not significantly different. This is also brings great difficulty to distinguish PCR and adulterants through color.

Furthermore, we established a three-dimensional scatter plot of PCR and its adulterants based on the L*, a* and b*. As can be seen from Fig. 1E, the three pure products can be clustered into one category separately in three-dimensional space, and the clustering distance of each sample is very far, which is consistent with the results of naked eye observation, that is, the color difference of the three pure samples huge. In Fig. 1F and Fig. 1G, PCR and adulterants are at different positions in the three-dimensional space. Interestingly, the colors of adulterants in adjacent proportions are very similar. Specifically, PCR and pure adulterants are each grouped into one category, but low-proportion adulterants are close together and basically indistinguishable, and high-proportion adulterants can be grouped into a separate category. Overall, although color digitization cannot completely differentiate between genuine PCR and adulterants, compared with naked eye observation, it can accurately and objectively reflect the information of adulterants and provide relatively better classification.

### 3.2. Qualitative analysis based on NIR spectroscopy

#### 3.2.1. NIR spectroscopy and exploratory analysis

Fig. 2A shows the raw near-infrared spectra of three pure samples. It can be seen that different samples have different absorption intensities at different positions. Research shows that the absorption peak near 4725 $cm^{-1}$ is caused by the stretching vibration of C—C and C═C (Zhang et al., 2023). The absorption peak close to 5180 $cm^{-1}$ may be caused by the stretching vibration of C═O or the second overtone absorption of the stretching and deformation vibration of O—H (Liu et al., 2019). The absorption peak at 6835 $cm^{-1}$ is caused by the first overtone absorption of the stretching vibration of O—H (Ma et al., 2020). The generation of these absorption peaks may be caused by flavonoid compounds. In addition, according to the knowledge of spectral analysis, the peaks near 5750 $cm^{-1}$ and 8350 $cm^{-1}$ may be caused by the stretching

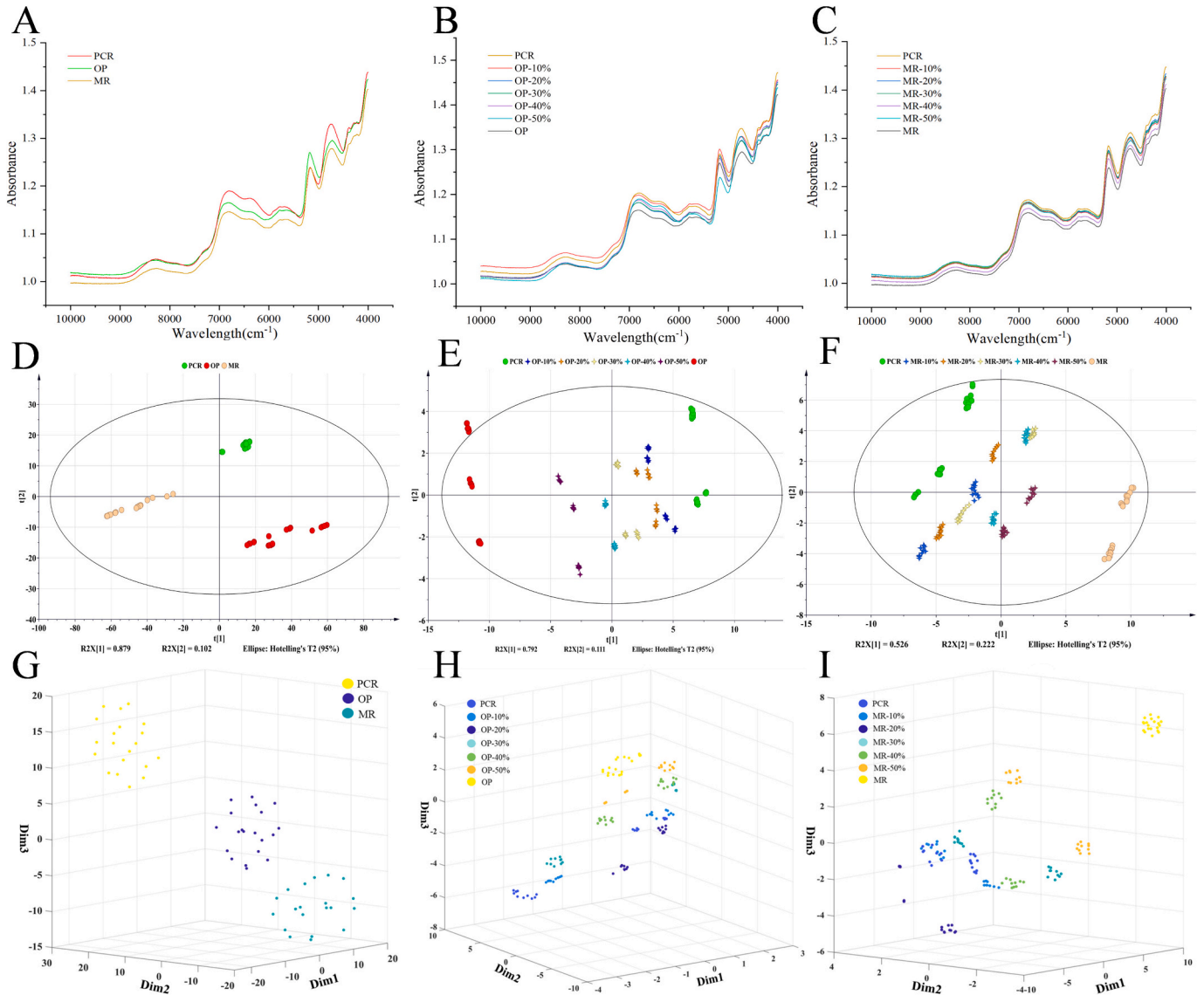



**Fig. 2.** NIR spectra (A-C), PCA clustering (D–F), and T-SNE clustering (G-I) of different proportions of adulterants (ADG-pure sample, BEH-adulterated with OP, CFI-adulterated with MR).

vibration of C—H in some aromatic compounds or the overtones produced by them (Barra et al., 2022; Zhan et al., 2017). Similarly, these absorption peaks also exist in the spectra of adulterated OP (Fig. 2B) and MR (Fig. 2C) with different proportions. Compared with pure PCR, they only show differences in absorption intensity, so it is speculated that these absorption peaks may be an important factor in distinguishing PCR from adulterants.

In order to further distinguish PCR and adulterants, we used PCA and t-SNE to perform visual analysis on the original NIR spectral information. Principal component analysis (PCA) is a widely used linear dimensionality reduction technique that transforms original variables into new variables that are linear combinations of the original variables. In contrast, t-distributed Stochastic Neighbor Embedding (t-SNE) is a nonlinear dimensionality reduction technique that can map high-dimensional data to a low-dimensional space while retaining the local structure of the data (Yan et al., 2022).

It can be seen from the PCA plots of the three pure samples (Fig. 2D) that principal components 1 and 2 contributed 87.9 % and 10.2 % to the variance respectively, and the first two principal components accounted for more than 98 % of the variance, which shows that the PCA model captures most of the information in the NIR spectrum. The visualization results show that the three pure samples can be clearly distinguished, and the samples adulterated with OP (Fig. 2E) and MR (Fig. 2F) can also be distinguished from pure PCR, but the adulterants in adjacent proportions are closer. This is basically consistent with the colorimetric analysis results. Similarly, it can be seen from Fig. G, H, and I that the results of cluster analysis using t-SNE technology are basically consistent with the PCA. However, it is gratifying that t-SNE has better results for different proportions of adulterants. All in all, through PCA and t-SNE visual analysis, it can be seen that PCR and adulterants with different proportions have obvious clustering trends, so it is necessary to combine spectral preprocessing and classification models to distinguish them.

#### 3.2.2. Machine learning discriminant analysis based on spectral preprocessing

In order to find out the difference between PCR and adulterants, we further preprocessed the spectra and developed five machine learning algorithms for differentiation. SVM takes the optimal hyperplane as the core to maximize the distance between different categories of points and the hyperplane (Amirvaresi & Parastar, 2021). RF make predictions by building multiple decision trees and combining them. GBM iteratively train a series of decision trees to gradually improve the performance of the overall model (Sun et al., 2021). KNN calculates the distance between the unknown sample and all samples in the training set, then selects the K closest samples to predict by voting or averaging. PLS-DA uses dimensionality reduction technology to find a new feature space that can distinguish different categories to the greatest extent, and then uses this new feature space for classification and prediction. Before building the machine learning model, we first adjust the hyperparameters of all models to ensure that the model achieves the best performance. The optimal hyperparameters of all models are listed in Table S2.

Fig. S1 shows the spectra produced by five different preprocessing methods. It can be seen that the shape and absorbance of the spectra have changed significantly after being processed by different methods, indicating that preprocessing is an effective method. Table S3 lists the performance indicators of machine learning classification models built using different preprocessing methods. By analyzing the performance indicators of the classification model of the raw NIR spectra of PCR and adulterants, it was found that GBM has the best classification effect. The classification accuracy of adulterated OP and MR reached 90.82 % and 90.21 % respectively. Relatively speaking, PLS-DA classification effect is the worst. We chose the GBM model to evaluate different preprocessing methods and found that the model established after SNV processing had the highest accuracy. The classification accuracy for adulterated OP and MR reached 99.30 % and 98.64 % respectively, fully realizing the classification of PCR and adulterants. In addition, SNV-GBM has the highest precision, recall and F1 scores. These indices are used to measure the accuracy of the model. Generally speaking, the best results are close to 1. Fig. 3 shows the confusion matrix and ROC curve of RAW-GBM and SNV-GBM of PCR and adulterants. It can be seen that in the RAW-GBM confusion matrix, the samples mixed with different proportions of OP and MR are always misidentified, and the misclassification basically does not occur after SNV treatment, and the area under the ROC curve of SNV-GBM is 1, which proves again that the model is very accurate. The results show that the spectral preprocessing combined with machine learning model can be used to identify PCR and adulterants, and the performance of SNV-GBM is the best.

### 3.3. Quantitative analysis based on FT-NIR spectroscopy

#### 3.3.1. Spectral preprocessing based on PLS regression

Although the established machine learning model can successfully identify PCR and adulterants, the concentration of adulterants can not be predicted by the classification model, so we use PLS to establish a quantitative calibration model of adulteration concentration. First of all, the spectrum is optimized by pretreatment to improve the accuracy of the PLS model. Generally speaking, too many principal components will lead to over-fitting of the model, and too few principal components will make the extraction information incomplete, so this study uses Monte Carlo cross-validation method to obtain the best principal components.

Table 1 shows the model performance index of the two adulterants, and Lvs is the potential best principal component of PLS. By comparing the performance of the PLS model established by different pretreatment methods, it is found that SNV and MSC give the best performance for the concentration prediction of adulterated OP and MR, respectively. The $R^2$ of the validation of the two models is 0.8966 and 0.9019 respectively, indicating that the true value and the predicted value are close. The RMSE is 0.0959 and 0.0957 respectively, indicating that the prediction error is small, and their ratio is close to 1, indicating that the division of the data set is more reasonable, and RPD reflects the overall performance of the model. When the RPD is greater than 3, the performance of the model is better, we can see that the performance of the model is better. Although the spectral pretreatment improves the accuracy of the model prediction, it is not enough to accurately predict the adulteration concentration.

#### 3.3.2. Feature wavelength selection based on PLS regression

In order to improve the performance of PLS model and achieve accurate prediction of adulteration concentration, we further use feature extraction algorithm to collect wavelengths which are closely related to modeling. CARS, ICO, RF and VIP algorithms are mainly used to extract feature wavelengths. The extracted feature wavelength is visualized in Fig. 4 (taking adulterated with OP as an example).

In the operation of CARS algorithm, with the increase of sampling times, the sampling variable decreases rapidly at first, and then tends to be stable (Fig. 4A1), and the number of RMSECV decreases rapidly before the sharp increase (Fig. 4A2). This is because a large number of uninformative variables were eliminated in the previous stage, and then some key variables were mistakenly eliminated, resulting in the loss of information. RMSECV reached the lowest value (marked by the blue line in Fig. 4A3) after 41 iterations, and 106 wavelengths were selected (Fig. 4B). The wavelength range of the NIR spectrum of the ICO algorithm is divided into 20 equal width intervals, and the weighted bootstrap sampling (WBS) is used to iteratively find the optimal interval combination by soft contraction. Finally, the local search strategy is used to optimize the selected wavelength interval width. The more yellow the color in Fig. 4C, the closer the sampling weight value is to 1, and the final selected wavelength is presented in the form of interval (Fig. 4D). Each tree in the RF algorithm carries out random feature sampling of the wavelength, and realizes feature selection by calculating the importance of each wavelength. Each point in Fig. 4E represents the wavelength,

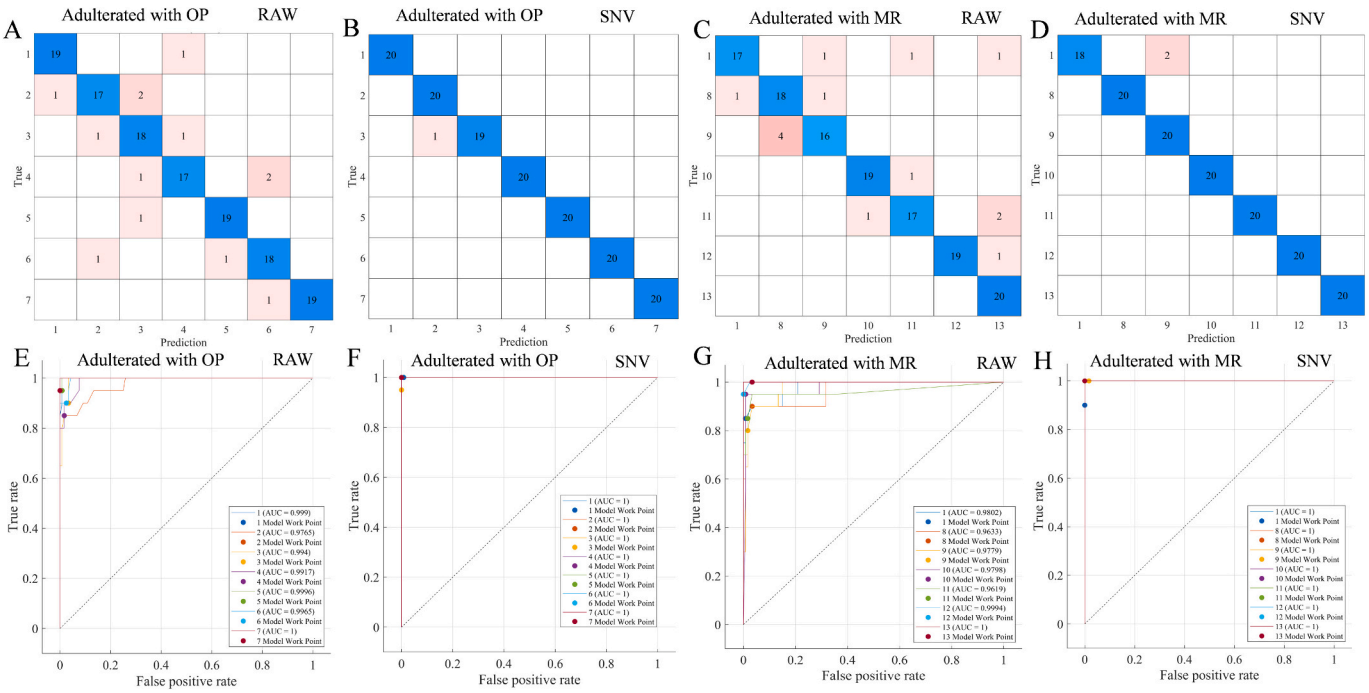



**Fig. 3.** Confusion matrix (A-D) and ROC curve (E-H) of RAW-GBM and SNV-GBM of tangerine peel and adulterants (A, B, E, F-Adulterated with OP, C, D, G, H-Adulterated with MR; 1-PCR, 2–7- Adulterated with OP10%–50 %, 8–13- Adulterated with MR10%–50 %).

**Table 1**
PLS regression prediction results of different spectra preprocessing methods.

| Adulteration category | Methods | Lvs | Calibration | | | Validation | | |
|---|---|---|---|---|---|---|---|---|
| | | | $R^2$ | RMSE | RPD | $R^2$ | RMSE | RPD |
| Adulterated with OP | RAW | 10 | 0.8971 | 0.1139 | 2.91 | 0.8621 | 0.1237 | 2.45 |
| | **SNV** | **8** | **0.9015** | **0.1017** | **3.11** | **0.8966** | **0.0959** | **3.10** |
| | MSC | 8 | 0.8845 | 0.1106 | 2.86 | 0.8831 | 0.1114 | 2.84 |
| | SG | 11 | 0.9047 | 0.1148 | 3.03 | 0.8785 | 0.1208 | 2.14 |
| | 1d | 3 | 0.6164 | 0.2808 | 1.44 | 0.6012 | 0.2620 | 1.55 |
| | 2d | 5 | 0.7098 | 0.4447 | 1.83 | 0.6219 | 0.3740 | 1.49 |
| Adulterated with MR | RAW | 11 | 0.7380 | 0.1601 | 1.92 | 0.7159 | 0.1772 | 1.82 |
| | SNV | 9 | 0.8923 | 0.1077 | 2.95 | 0.8860 | 0.1036 | 2.94 |
| | **MSC** | **9** | **0.9045** | **0.1013** | **3.14** | **0.9019** | **0.0957** | **3.18** |
| | SG | 11 | 0.8719 | 0.1126 | 2.77 | 0.8114 | 0.1410 | 2.21 |
| | 1d | 3 | 0.8451 | 0.1891 | 2.19 | 0.8496 | 0.1333 | 1.80 |
| | 2d | 6 | 0.8100 | 0.3527 | 2.27 | 0.7981 | 0.4032 | 1.57 |

and the size represents the importance. It can be seen that the characteristic wavelength is mainly concentrated between 4500 $cm^{-1}$ and 7000 $cm^{-1}$. In the VIP algorithm, the importance score of each wavelength to the model prediction is calculated, and then the most important wavelength is selected according to these scores. Usually VIP greater than 1 is the selection criterion (Fig. 4F).

Table S4 shows the performance of the PLS model constructed using characteristic wavelengths. For the adulterated with OP, the SNV-RF-PLS model has the best prediction performance, and the $R^2$, RMSE and RPD of the validation are 0.9930, 0.0247 and 11.91, respectively, and for the adulterated with MR, the MSC-ICO-PLS model gives the best prediction performance, and the $R^2$, RMSE and RPD of the validation are 0.9855, 0.0354 and 8.31.

### *3.3.3. Construction of quantitative regression curve*

Through spectral pretreatment and characteristic wavelength selection, the best quantitative calibration model for the concentration of two kinds of adulterants in PCR was obtained, and the quantitative regression curve was constructed. Fig. 5 shows the scatter diagram between the predicted and reference values of the concentrations of the two adulterants. As we all know, the closer the scatter is to the diagonal, the better the prediction performance of the model (Yang et al., 2017). After pretreatment and characteristic wavelength selection, the contours of all scattered points are in a straight line, distributed near the diagonal, indicating that the model has a good prediction ability and can realize the quantitative prediction of adulteration concentration of PCR.

## 4. Conclusion

As a popular dual-use medicine, PCR can be used not only as food, but also as daily spice. With the increasing price of PCR, unscrupulous businessmen often pretend to sell OP and MR of the same family, which greatly affects the fairness and stability of the market. Previous studies focused on the geographical origin and aging time of PCR, and almost ignored the adulteration of PCR. Although plant metabolomics methods such as UHPLC-Q-TOF-MS and GC–MS have been used to identify the chemical composition and quality of PCR, and can comprehensively analyze the chemical composition and identify metabolic differences as markers for identification (Duan et al., 2016; Sanches et al., 2022; L. Shi et al., 2020; Wang et al., 2019). Unfortunately, compared with non-

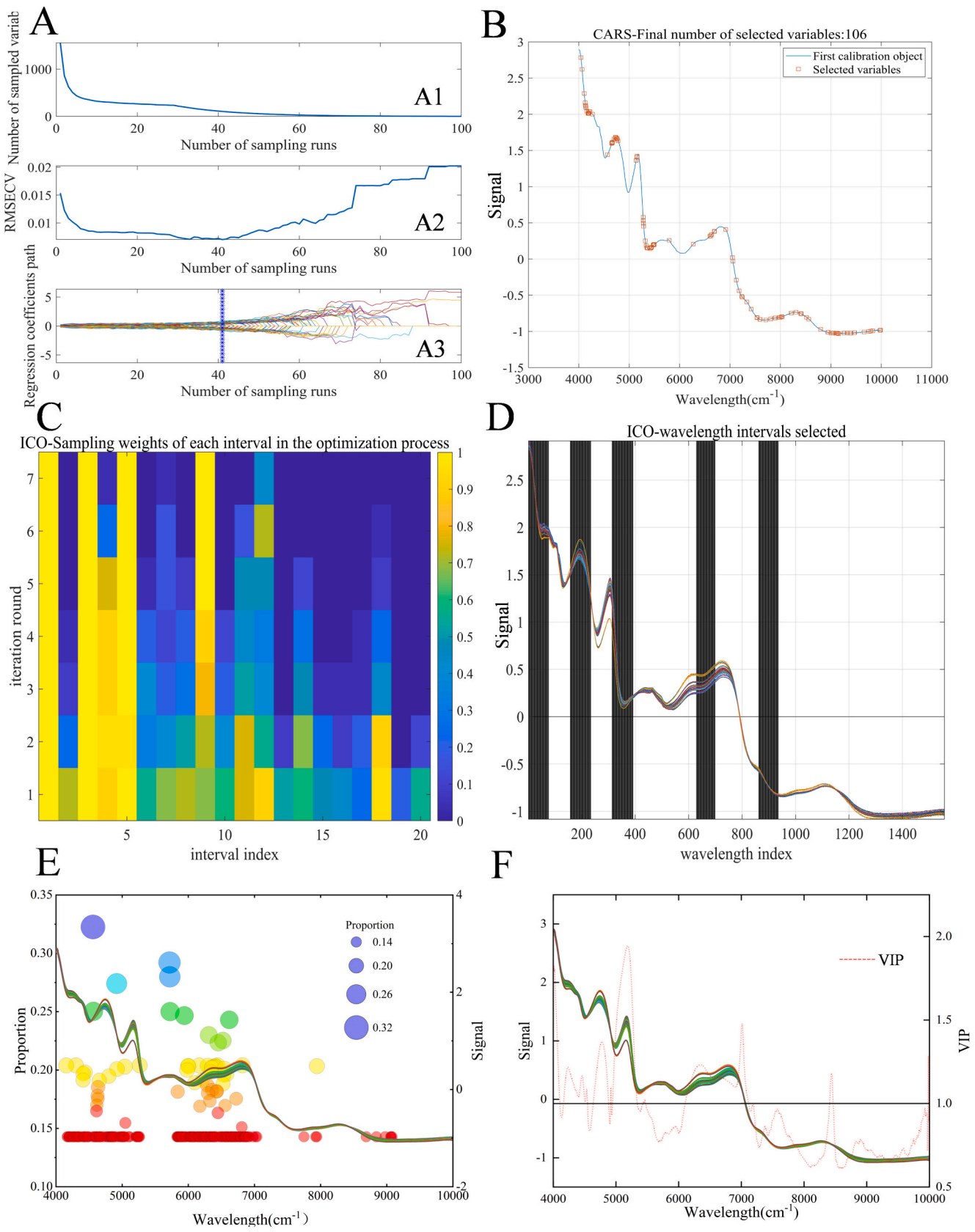



**Fig. 4.** The results of wavelength selection with CARS algorithm for OP adulterants:(A) (A1) changes in the number of selected variables, (A2) variation of RMSECV, (A3) path of variable regression coefficients. (B) Characteristic wavelength selection results. The results of wavelength selection with ICO algorithm for OP adulterants: (C) Sampling weights for each feature interval in the optimization process. (D) Characteristic intervals selected by ICO algorithm. Results of wavelength selection for OP adulterants using RF(E) and VIP(F) algorithms.

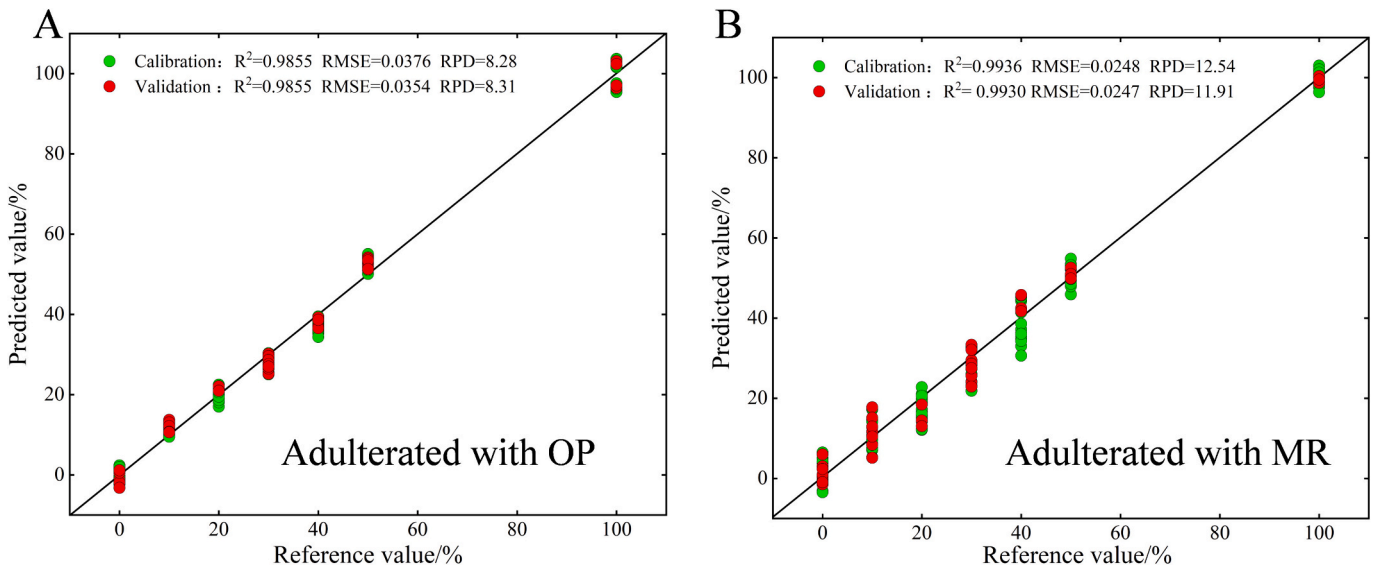



**Fig. 5.** Regression prediction model for OP (A) and MR (B) adulteration ratio based on FT-NIR.

destructive testing technologies such as FT-NIR, these technologies face challenges such as long detection cycles and complex sample pretreatment. Therefore, exploring rapid non-destructive testing methods is of great guiding significance for combating PCR adulteration.

The purpose of this study is to explore an economical and simple method to solve the adulteration problem of PCR. Through the study, it is found that the difference between PCR and adulterants can not be completely realized by color value, so it is further analyzed by NIR spectroscopy. Through the visual analysis of spectra by PCA and t-SNE, it is found that PCR and adulterants have the trend of clustering. Spectral preprocessing technology combined with machine learning algorithm is used to classify and evaluate them, and the SNV-GBM model is constructed. The classification accuracy of the two kinds of adulterants is 99.30 % and 98.24 %, respectively. Then the PLS regression quantitative models of two kinds of adulterants concentrations are established. The model established by the combination of SNV-RF (adulterated with OP) and MSC-ICO (adulterated with MR) has the best performance, and the $R^2$ of the validation is 0.9930 and 0.9855, respectively. It shows that the regression model has good linearity and accuracy.

This study pioneered the use of FT-NIR technology to study PCR adulteration, providing a novel solution in this field. Traditional PCR detection methods often face challenges such as time-consuming, complicated detection steps, and limited accuracy. In contrast, by innovatively combining the advantages of efficient and nondestructive detection of FT-NIR spectroscopy with a variety of spectral preprocessing methods, such as MSC, this study realizes a more accurate and comprehensive monitoring of PCR adulteration, which brings a new technological breakthrough for the solution of this problem. This method not only has obvious advantages in terms of time and cost, but also provides deeper spectral characterization to accurately target adulterated components and behaviors. This breakthrough brings new technical means and ideas for food safety testing, which not only effectively ensures the authenticity and reliability of PCR, but also provides strong support for maintaining the stability of the food consumption market and protecting the health and rights of consumers. In the future, this research direction is expected to continue to develop and inject new vitality into scientific and technological innovation in the field of food safety.

## CRediT authorship contribution statement

**Ying Chen:** Writing – original draft, Methodology, Data curation. **Si Li:** Writing – original draft, Formal analysis, Data curation. **Jia Jia:** Data curation, Formal analysis, Project administration, Supervision. **Chuanduo Sun:** Conceptualization, Data curation, Visualization, Writing – review & editing. **Enzhong Cui:** Conceptualization, Formal analysis, Supervision, Visualization. **Yunyan Xu:** Data curation, Investigation, Resources, Software. **Fangchao Shi:** Formal analysis, Funding acquisition, Supervision. **Anfu Tang:** Writing – review & editing, Supervision, Methodology, Funding acquisition.

## Declaration of competing interest

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

## Data availability

Data will be made available on request.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.fochx.2024.101798.

## References

Amirvaresi, A., & Parastar, H. (2021). External parameter orthogonalization-support vector machine for processing of attenuated total reflectance-mid-infrared spectra: A solution for saffron authenticity problem. *Analytica Chimica Acta, 1154*, Article 338308.

Araújo, M. C. U., Saldanha, T. C. B., Galvão, R. K. H., Yoneyama, T., Chame, H. C., & Visani, V. (2001). The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. *Chemometrics and Intelligent Laboratory Systems, 57*(2), 65–73.

Barra, I., Briak, H., & Kebede, F. (2022). The application of statistical preprocessing on spectral data does not always guarantee the improvement of the predictive quality of multivariate models: Case of soil spectroscopy applied to Moroccan soils. *Vibrational Spectroscopy, 121*, Article 103409.

Bian, X., Xie, X., Cai, J., Zhao, Y., Miao, W., Chen, X., … Wu, J.-L. (2022). Dynamic changes of phenolic acids and antioxidant activity of Citri Reticulatae Pericarpium during aging processes. *Food Chemistry, 373*, Article 131399.

Chen, R., Li, S., Cao, H., Xu, T., Bai, Y., Li, Z., … Huang, Y. (2023). Rapid quality evaluation and geographical origin recognition of ginger powder by portable NIRS in tandem with chemometrics. *Food Chemistry*, Article 137931.

Dai, G., Wu, L., Zhao, J., Guan, Q., Zeng, H., Zong, M., … Du, C. (2023). Classification of Pericarpium Citri Reticulatae (Chenpi) age using surface-enhanced Raman spectroscopy. *Food Chemistry, 408*, Article 135210.

Ditcharoen, S., Sirisomboon, P., Saengprachatanarug, K., Phuphaphud, A., Rittiron, R., Terdwongworakul, A., Malai, C., Saenphon, C., Panduangnate, L., & Posom, J.

(2023). Improving the non-destructive maturity classification model for durian fruit using near-infrared spectroscopy. *Artificial Intelligence in Agriculture, 7*, 35–43.
Duan, L., Guo, L., Dou, L.-L., Zhou, C.-L., Xu, F.-G., Zheng, G.-D., … Liu, E. H. (2016). Discrimination of *Citrus reticulata* Blanco and *Citrus reticulata* 'Chachi' by gas chromatograph-mass spectrometry based metabolomics approach. *Food Chemistry, 212*, 123–127.
Ho, S.-C., & Kuo, C.-T. (2014). Hesperidin, nobiletin, and tangeretin are collectively responsible for the anti-neuroinflammatory capacity of tangerine peel (Citri reticulatae pericarpium). *Food and Chemical Toxicology, 71*, 176–182.
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.
Luo, Y., Zeng, W., Huang, K.-E., Li, D.-X., Chen, W., Yu, X.-Q., & Ke, X.-H. (2019). Discrimination of *Citrus reticulata* Blanco and *Citrus reticulata* 'Chachi' as well as the *Citrus reticulata* 'Chachi' within different storage years using ultra high performance liquid chromatography quadrupole/time-of-flight mass spectrometry based metabolomics approach. *Journal of Pharmaceutical and Biomedical Analysis, 171*, 218–231.
Ma, H., Shao, Y., Chen, J., Pan, D., Si, L., Liu, X., … Wu, Y. (2020). Maintaining the predictive abilities of near-infrared spectroscopy models for the determination of multi-parameters in White Paeony Root. *Infrared Physics & Technology, 109*, Article 103419.
Manthey, J. A., Grohmann, K., & Guthrie, N. (2001). Biological properties of citrus flavonoids pertaining to cancer and inflammation. *Current Medicinal Chemistry, 8*(2), 135–153.
Medeiros, M. L.d. S., Freitas Lima, A., Correia Gonçalves, M., Teixeira Godoy, H., & Fernandes Barbin, D. (2023). Portable near-infrared (NIR) spectrometer and chemometrics for rapid identification of butter cheese adulteration. *Food Chemistry, 425*, Article 136461.
Ndlovu, P. F., Magwaza, L. S., Tesfay, S. Z., & Mphahlele, R. R. (2021). Rapid spectroscopic method for quantifying gluten concentration as a potential biomarker to test adulteration of green banana flour. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 262*, Article 120081.
Oliveira, M. M., Cruz-Tirado, J. P., Roque, J. V., Teófilo, R. F., & Barbin, D. F. (2020). Portable near-infrared spectroscopy for rapid authentication of adulterated paprika powder. *Journal of Food Composition and Analysis, 87*, Article 103403.
Pan, S., Zhang, X., Xu, W., Yin, J., Gu, H., & Yu, X. (2022). Rapid on-site identification of geographical origin and storage age of tangerine peel by near-infrared spectroscopy. *Spectrochimica Acta. Part A, Molecular and Biomolecular Spectroscopy, 271*, Article 120936.
Pu, H., Yu, J., Sun, D. W., Wei, Q., & Li, Q. (2023). Distinguishing pericarpium citri reticulatae of different origins using terahertz time-domain spectroscopy combined with convolutional neural networks. *Spectrochimica Acta. Part A, Molecular and Biomolecular Spectroscopy, 299*, Article 122771.
Raghavendra, A., Guru, D. S., & Rao, M. K. (2021). Mango internal defect detection based on optimal wavelength selection method using NIR spectroscopy. *Artificial Intelligence in Agriculture, 5*, 43–51.
Sanches, V. L., Cunha, T. A., Viganó, J., de Souza Mesquita, L. M., Faccioli, L. H., Breitkreitz, M. C., & Rostagno, M. A. (2022). Comprehensive analysis of phenolics compounds in citrus fruits peels by UPLC-PDA and UPLC-Q/TOF MS using a fused-core column. *Food Chemistry: X, 14*, Article 100262.
Shi, L., Wang, R., Liu, T., Wu, J., Zhang, H., Liu, Z., … Liu, Z. (2020). A rapid protocol to distinguish between Citri Exocarpium rubrum and Citri Reticulatae Pericarpium based on the characteristic fingerprint and UHPLC-Q-TOF MS methods. *Food & Function, 11*(4), 3719–3729.
Shi, Y., He, T., Zhong, J., Mei, X., Li, Y., Li, M., Zhang, W., Ji, D., Su, L., Lu, T., & Zhao, X. (2023). Classification and rapid non-destructive quality evaluation of different processed products of *Cyperus rotundus* based on near-infrared spectroscopy combined with deep learning. *Talanta, 268*(Pt 1), Article 125266.
Song, X., Huang, Y., Yan, H., Xiong, Y., & Min, S. (2016). A novel algorithm for spectral interval combination optimization. *Analytica Chimica Acta, 948*, 19–29.
Sun, Y., Liu, N., Kang, X., Zhao, Y., Cao, R., Ning, J., Ding, H., Sheng, X., & Zhou, D. (2021). Rapid identification of geographical origin of sea cucumbers *Apostichopus japonicus* using FT-NIR coupled with light gradient boosting machine. *Food Control, 124*, Article 107883.
Vieira, L. S., Assis, C., de Queiroz, M. E. L. R., Neves, A. A., & de Oliveira, A. F. (2021). Building robust models for identification of adulteration in olive oil using FT-NIR, PLS-DA and variable selection. *Food Chemistry, 345*, Article 128866.
Wang, P., Zhang, J., Zhang, Y., Su, H., Qiu, X., Gong, L., … Xu, W. (2019). Chemical and genetic discrimination of commercial Guangchenpi (*Citrus reticulata* "Chachi") by using UPLC-QTOF-MS/MS based metabolomics and DNA barcoding approaches. *RSC Advances, 9*(40), 23373–23381.
Yan, W., Zhao, M., Fu, Z., Pearlson, G. D., Sui, J., & Calhoun, V. D. (2022). Mapping relationships among schizophrenia, bipolar and schizoaffective disorders: A deep classification and clustering framework using fMRI time series. *Schizophrenia Research, 245*, 141–150.
Yang, Y., Liu, X., Li, W., Jin, Y., Wu, Y., Zheng, J., … Chen, Y. (2017). Rapid measurement of epimedin A, epimedin B, epimedin C, icariin, and moisture in Herba Epimedii using near infrared spectroscopy. *Spectrochimica Acta. Part A, Molecular and Biomolecular Spectroscopy, 171*, 351–360.
Ye, D., Sun, L., Zou, B., Zhang, Q., Tan, W., & Che, W. (2018). Non-destructive prediction of protein content in wheat using NIRS. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 189*, 463–472.
Yi, L., Dong, N., Liu, S., Yi, Z., & Zhang, Y. (2015). Chemical features of Pericarpium Citri Reticulatae and Pericarpium Citri Reticulatae Viride revealed by GC-MS metabolomics analysis. *Food Chemistry, 186*, 192–199.
Yu, D.-X., Guo, S., Zhang, X., Yan, H., Zhang, Z.-Y., Chen, X., … Duan, J.-A. (2022). Rapid detection of adulteration in powder of ginger (*Zingiber officinale* Roscoe) by FT-NIR spectroscopy combined with chemometrics. *Food Chemistry: X, 15*, Article 100450.
Yuan, L.-M., Mao, F., Huang, G., Chen, X., Wu, D., Li, S., Zhou, X., Jiang, Q., Lin, D., & He, R. (2020). Models fused with successive CARS-PLS for measurement of the soluble solids content of Chinese bayberry by vis-NIRS technology. *Postharvest Biology and Technology, 169*, Article 111308.
Zhan, H., Fang, J., Tang, L., Yang, H., Li, H., Wang, Z., Yang, B., Wu, H., & 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., 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.
Zhang, X., Sun, J., Li, P., Zeng, F., & Wang, H. (2021). Hyperspectral detection of salted sea cucumber adulteration using different spectral preprocessing techniques and SVM method. *LWT, 152*, Article 112295.