NON-DESTRUCTIVE DETECTION OF MOLD IN MAIZE USING NEAR-INFRARED SPECTRAL FINGERPRINTING
/
基于近红外光谱指纹技术的霉变玉米籽粒无损检测

Longbao LIU[1)], Qixing TANG[1*)], Juan LIAO[1)], Lu LIU[1)], Yujun ZHANG[2)], Leizi JIAO[3)]
[1)]Anhui Agricultural University, Hefei 230061, China;
[2)] Key Laboratory of Environmental Optics & Technology, Anhui Institute of Optics and Fine Mechanics, Chinese Academy of Sciences, Hefei 230031, China;
[3)]Beijing Res Ctr Intelligent Equipment Agr, Beijing Academy of Agriculture and Forestry Sciences, Beijing 100097, China
Tel: +86-511-13637080689; E-mail: qxtang@ahau.edu.cn
DOI: https://doi.org/10.35633/inmateh-75-24

Keywords: *Mold infection, Feature wavelength, Machine learning, Precision classification*

ABSTRACT

Mold contamination of stored maize can cause significant economic losses, and it is crucial to effectively classify maize kernels without destroying their original structure. But existing studies have found it difficult to distinguish moldy maize. In this paper, a method for non-destructive detection of mold in maize using near-infrared spectral fingerprinting is proposed. The spectral raw data are initially acquired using a handheld near-infrared spectrometer. To enhance the signal quality, preprocessing is conducted, and a classification model is developed for full-band spectral data. In order to further optimize the model and enhance the classification accuracy, the feature wavelengths were extracted from the spectral data with effective preprocessing techniques in the full-band model. Finally, the maize kernel mold classification model is constructed. The classification accuracy of SG+SNV-SVM-ISFLA model can reach up to 97.22%, and the accuracy for the identification of asymptomatic moldy maize is 96.30%, which can realize the accurate grading of moldy maize and can well distinguish asymptomatic moldy maize. This work may significantly control the spread of molds in the food industry while improving storage economics and safety.

摘要

仓储玉米受到霉菌污染会造成重大经济损失，因此在不破坏玉米原有结构的情况下对玉米进行有效分类至关重要。但现有研究发现，轻微霉变玉米难以区分。本文提出了一种利用近红外光谱指纹的玉米霉变无损检测的方法。最初使用手持式近红外光谱仪获取光谱原始数据。为了提高信号质量，进行了预处理，并为全波段光谱数据建立分类模型。为了进一步优化模型并提高分类准确率，对预处理后的数据进行特征提取。最后，构建了玉米霉变分类模型。结果表明 SG+SNV-ISFLA-SVM 模型的分类准确率高达 97.22%，对无症状霉变玉米的识别准确率为 96.30%，可实现对玉米霉变的准确分级，并能很好地区分无症状霉变玉米。这项工作可大大控制霉菌在食品工业中的传播，同时提高贮藏的经济性和安全性。

INTRODUCTION

Maize, as an important crop, causes huge economic losses every year due to maize mold problems *(Long et al., 2022a).* Maize itself contains starch, proteins, sugar, amino acids and other nutrients for the reproduction of mold which provides the necessary nutritional conditions, while the temperature and humidity and other external factors in the storage room will also accelerate the reproduction of mold. The process of maize mold mainly produces aflatoxin, Deoxynivalenol (DON), fumonisins and other harmful substances. According to the estimation of the Food and Agriculture Organization (FAO) of the United Nations and a recent validation report, more than 25% of food crops are contaminated with mycotoxins due to the occurrence of molds *(Long et al., 2022b; Zhang et al., 2020)*. In order to prevent and control the occurrence of molds in food, there is a need for the efficient identification of molds in maize kernels *(Kang et al., 2022).* Existing mold identification methods are difficult to distinguish asymptomatic moldy maize. Therefore, researching non-destructive mold identification methods while being able to accurately detect asymptomatic moldy maize is a key technical challenge currently being faced.

Traditional methods of identifying mold in crops are divided into physical and chemical methods. The physical methods mainly identify the degree of mold by observing their color, odor and touch, which are simple to operate, but are affected by subjective factors and cannot be accurately quantified *(Milićević et al., 2010; Oyebanji et al., 1999)*.

The chemical methods are used to quantitatively analyze mold by preparing sample solution, which are highly sensitive and quantitative, but this method have high complexity and high demands for operators *(Paraginski et al., 2019)*. Visible light computer vision has been used by scholars for mold identification, which has the advantages of non-contact, high efficiency, rapidity and automation, but this technology cannot effectively identify asymptomatic moldy maize *(Fei et al., 2018; Qiang et al., 2014; Sun et al., 2022).* The rapid development of spectroscopic technology provides a new method for quality detection of crops, and the identification of moldy maize by spectroscopy offers the possibility of effective identification of healthy maize and asymptomatic moldy maize *(Bai et al., 2020).*

Near-infrared spectroscopy is often used for crop quality detection because of the fast analysis and high efficiency of the measurement process. Currently, the application of NIR spectroscopy to the detection of crop composition and diseases has achieved great results. The combined and multiplied frequencies of hydroxyl groups, such as O-H, N-H, C-H, and S-H, and C-O, are also in the NIR band *(Cui et al., 2019).* Hydroxyl and single bond hydroxyl groups are the main groups that make up organic compounds, including oils, proteins, and sugars. Therefore, the light absorption of organic compounds in the near-infrared region is mainly the absorption of light with the multiplicative or combined frequency of hydroxyl groups. *Jiahui Dai et al. (2024)* used near infrared spectroscopy to quantitatively analyze the main active components of betel nut, and the results showed that R2, values were close to 1, and the corresponding RPD values were all less than 3. *Jiangming Jia et al. (2022)* developed quantitative prediction models for sensory quality scores, total catechins and caffeine of different quality grades of Yuezhou Longjing tea using near-infrared spectroscopy. The results showed that the best prediction models for sensory scores, total catechins and caffeine were VCPA-IRIV + SVR, VCPA-IRIV + RF, and CARS + SVR. The relative percentage deviation (RPD) values were 2.485, 2.584 and 2.873, respectively. *Jiahui Zhang et al*. *(2023)* used a handheld miniature near-infrared spectrometer and applied 24 preprocessing methods in combination with a support vector machine (SVM) and a boosting algorithm. The results showed that the model exhibited high accuracy and stability during parameter tuning, with precision and F1 scores greater than 0.8 and a Kappa coefficient around 0.7. These findings confirm the feasibility of using near-infrared spectroscopy (NIRS) for the rapid identification of 'Dangshan' pear Mianhua reaction diseas. Pauline Ong et al. used visible and near infrared spectroscopy combined with a novel wavelength selection method called Modified Flower Pollination Algorithm (MFPA) to identify sugarcane diseases. The simplified SVM model developed utilized the MFPA wavelength selection method to obtain the best performance with an accuracy value of 0.9753, a sensitivity value of 0.9259, a specificity value of 0.9524 and an accuracy of 0.9487 *(Ong et al., 2023). Minhui An et al. (2023)* used NIR to establish a prediction model for walnut mold and developed Support Vector Machine (SVM) and Extreme Learning Machine (ELM) with an accuracy of 100% for the identification of walnut mol. *Hui Jiang et al. (2023)* quantified aflatoxin B1 in moldy peanuts using two-dimensional convolutional neural network in NIR and the results showed that R2=0.99, RPD=8.3, RPIQ=9.3. All of the above studies have validated the feasibility of NIR spectroscopy for crop composition analysis and disease detection as well as mold detection, but there is little research on early mold detection in maize, especially in healthy maize and asymptomatic mold.

Therefore, to address the challenge of distinguishing healthy maize from asymptomatic moldy maize, this paper proposes a method that utilizes SVM, PLS-DA, RF, KNN, and BP neural networks to analyze and determine the optimal classification model. By comparing the accuracy and F1 scores of these models, the most effective model-building approach is identified to achieve accurate classification of moldy maize. Additionally, this method enables the efficient detection of asymptomatic moldy maize, meeting the requirements of maize warehousing while enhancing its economic benefits.

MATERIALS AND METHODS

Samples

The samples for this experiment were taken from the same variety of maize that had been inoculated with molds in the breeding laboratory of Anhui Agricultural University and had been classified into four grades of moldy maize (grade 1, 2, 3, and 4), i.e., healthy maize, asymptomatic moldy maize, moderately moldy maize, and severely moldy maize, according to the method of LS/T 6132-2018, "Inspection of grain and oils—Storage fungal examination—Enumeration spores of fungi". A total of 683 maize kernel samples meeting the criteria were collected. This included 200 samples of grade 1, 153 samples of grade 2, 261 samples of grade 3, and 68 samples of grade 4. In accordance with an 8:2 ratio, all samples were randomly divided into a train set and a test set for the classification modeling study.

Near-infrared spectrum instrument

The instrument utilized to collect spectral data in this experiment is the handheld portable miniature near-infrared spectrometer "NIR-R210" created by Shenzhen Pynect Technology Co., Ltd. (Shenzhen, China). The spectral wavelength range spans from 900 nm to 1700 nm, comprising 228 bands for spectral detection, with a resolution of 3.89 nm and a signal-to-noise ratio (SNR) of 6000:1. The spectral data are collected by connecting the data line of the spectrometer to the computer. Prior to each measurement, the NIR spectrometer must undergo calibration using a black and white plate. Spectral data were acquired using the ISC-NIRScan-GUI (Windows v3.9.8-win64), supplied with the device.

Spectral data acquisition

For maximum measurement accuracy, the spectrometer underwent a 30-minute warm-up before measurement. The endosperm of each maize kernel was scanned with the endosperm side down, as illustrated in Fig. 1. Each sample underwent 5 scans, and the average of these 5 samples was considered as the final data. Each data file was named based on the sample number. Following the completion of spectral data acquisition, all files were imported. The data in each file included intensity, wavelength, absorbance, and reflectance. The absorption spectral data from the average of 5 scans were used as raw modeling spectral data.

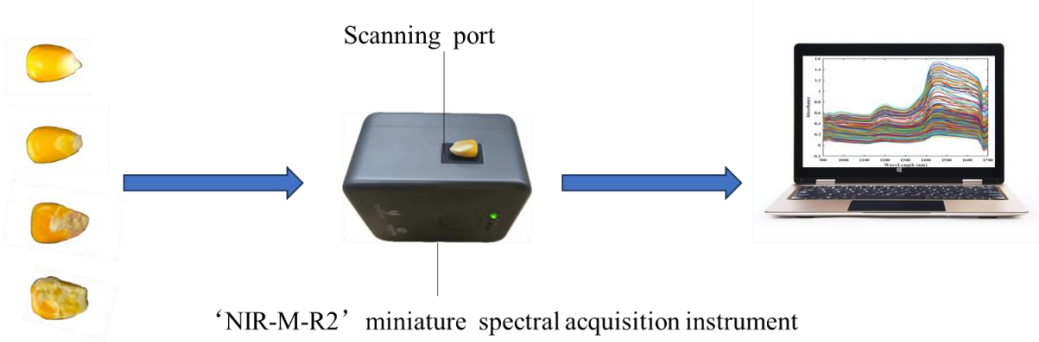


Fig. 1 - Spectral Data Scanning

RGB data acquisition

Due to the limitations of visible light cameras and the naked eye, detecting mold on asymptomatic moldy maize is challenging. Therefore, this paper adopts an electron microscope modeled as "HY-500BL" produced by Shenzhen HAYEAR Electronics Co., Ltd. to take pictures of moldy maize. It can find molds on the surface of asymptomatic moldy maize, as shown in Fig. 2. The collected microscopic photographs provide evidence that asymptomatic moldy corn is observable and allow for a secondary check of the classification results for asymptomatic mold.

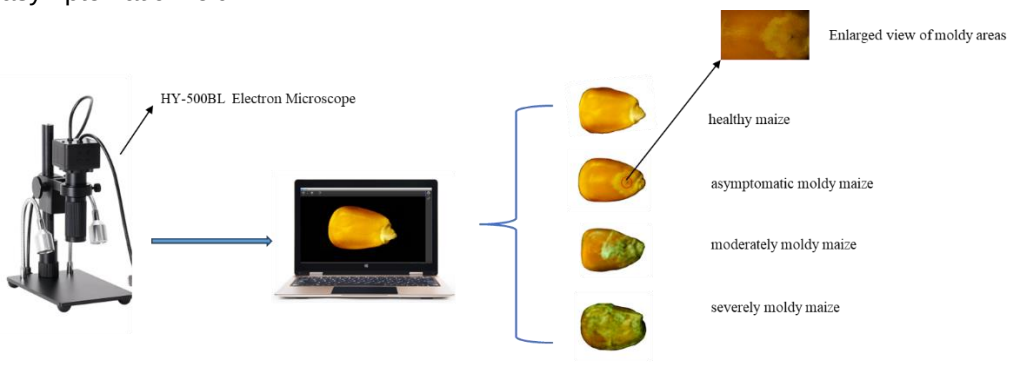


Fig. 2 - Plot of the extent of mold on maize under an electron microscope

Spectral processing method and feature wavelength extraction

During spectral data acquisition, certain disturbing factors unrelated to spectral information acquisition inevitably arise, such as the background noise of the experimental platform *(Liu et al., 2023).* This leads to some irrelevant information in the raw spectral data. Studies have demonstrated that data preprocessing is effective in mitigating invalid information in spectral curves, enhancing the utilization rate of spectrally valid information, improving SNR of NIR, and subsequently enhancing the accuracy and stability of the regression

model established at a later stage *(Xu et al., 2008).* In this paper, Baseline Correction (BC) *(Li et al., 2020)*, Multiplicative Scatter Correction (MSC) *(Sun et al., 2019)*, Standard Normal Variate Transform (SNV) *(Malavi et al., 2024)*, and Savitzky-Golay convolution smoothing (SG) *(Lanjewar et al., 2024)* were employed. Additionally, a combination of SG + SNV, SG + MSC, and SG + BC, constituting two single preprocessing algorithms, was explored to determine the optimal preprocessing method.

Specifically, BC eliminates the influence of the baseline by finding the flatter and more stable parts of the spectral data, and then subtracting or cancelling the resulting baseline from the original spectral signal by fitting the shape and position of the baseline, thus making the spectral signal features more prominent and clearer. As shown in Equation (1).

$$\begin{aligned} Y(\lambda) &= X(\lambda) - B(\lambda) \\ B(\lambda) &= a_0 + a_1\lambda + a_2\lambda^2 + \cdots + a_n\lambda^n \end{aligned} \tag{1}$$

where:

$X(\lambda)$ represents the original spectral data, $B(\lambda)$ represents the fitted baseline, and $Y(\lambda)$ represents the corrected spectral data.

MSC selects a set of representative spectral data as a benchmark, averages all the spectral data of the benchmark to obtain the average spectrum, then compares each spectral data with the average spectral data, calculates the scaling factor, i.e., the scaling factor at each wavelength point, and applies the obtained scaling factor to the original sample spectral data, to obtain the scaling-corrected spectral data. As shown in Equation (2).

$$\begin{aligned} \bar{X} &= \frac{1}{N}\sum_{i=1}^{N} X_i \\ X_{ij} &\approx a_i \bullet \bar{X}_j + b_i \\ Y_{ij} &= \frac{X_{ij} - b_i}{a_i} \end{aligned} \tag{2}$$

where: $\bar{X}$ represents the mean spectrum, N is the number of samples, Xi is the spectrum of the i-th sample, X_{ij} denotes the absorbance value of sample Xi at the j-th wavelength point, $\bar{X}_j$ denotes the absorbance value of the reference spectrum at the j-th wavelength point, a_i is the slope, b_i is the intercept, Y_{ij} represents the corrected spectrum.

SNV is centered by calculating the average of the spectral intensities at each wavelength for each sample, and then subtracting the average of the corresponding wavelengths for each sample, which ensures that the average spectral intensity at each wavelength is zero. For each wavelength point, the standard deviation of the spectral intensities of all samples at that wavelength point was calculated, and for each wavelength point, the spectral intensities of all samples were divided by the standard deviation of the corresponding wavelength point. This ensured that the variance of the spectral data was relatively consistent for the purpose of standardization, as shown in Equation (3).

$$\begin{aligned} \bar{X}_i &= \frac{1}{n}\sum_{j=1}^{n} X_{ij} \\ \sigma_i &= \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(X_{ij} - \bar{X}_i)^2} \\ Y_{ij} &= \frac{X_{ij} - \bar{X}_i}{\sigma}\infty \end{aligned} \tag{3}$$

where: $\bar{X}_i$ represents the mean of the samples, n is the total number of wavelength points in the spectrum, X_{ij} denotes the absorbance value of the i-th sample at the j-th wavelength point, σ_i represents the standard deviation, and Y_{ij} represents the standardized data.

Finally, SG determines the order of the polynomial by choosing the size of the smoothing window, i.e. the number of neighboring data points to be processed at a time, and then determines the order of the polynomial, which determines the complexity of the polynomial to be fitted. The selected window is slid over the data from the first data point to the last data point and within each window polynomial fitting and smoothing of the data is done. Polynomial fitting is done by finding a polynomial by least squares that is as close as possible to all the data points within the window. Smoothing of the data is done by using the fitted polynomial to estimate the value of the window centroid i.e., the value of the smoothed data point, and the above operation is repeated until all the data points are processed, as shown in Equation (4).

$$P(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_k x^k$$
$$Y_i = P(x) \bullet X_i \quad (4)$$

where: $P(x)$ represents the fitted polynomial, $a_{0,} a_{1,\ldots} a_{\mathrm{k}}$ denotes the coefficients of the fitted polynomial, and Y_i represents the smoothed data.

Following the preprocessing of spectral data, various feature extraction algorithms, including the Successive Projection Algorithm (SPA), Competitive Adaptive Reweighted Sampling (CARS), and Improved Shuffled Frog Leaping Algorithm (ISFLA), are utilized to identify characteristic wavelengths meeting the prediction criteria in the full-wavelength model. Subsequently, a model based on these extracted characteristic wavelengths is constructed to identify characteristic wavelengths. The model is further optimized to enhance accuracy.

The SPA algorithm first randomly selects a feature from the dataset as the initial feature. It then projects the remaining features onto the orthogonal complementary space of the already selected features, removing the influence of the selected features on the remaining ones. Afterward, the algorithm selects the feature that contributes the most to the model from the projected features. Finally, the newly selected feature is added to the set of selected features. Repeat the projection and selection steps until a predetermined number of features are selected, and the final set of selected features is representative and uncorrelated *(Pang et al., 2022; Vallese et al., 2023).*

CARS algorithm first randomly generates multiple feature subsets, each of which contains a random subset of the original features. For each feature subset, a PLS regression model is built and the regression coefficients are calculated for each feature. Based on the absolute value of the regression coefficients, weights are assigned to each feature. The larger the weight, the higher the importance of the feature to the model. Subsequently, a weighted sampling method was used to sample the features based on their weights. Features with higher weights have a higher probability of being retained. The relationship between the number of retained features and the prediction error in each iteration is calculated and a 'U' curve is plotted. Select the subset of features with the smallest prediction error, repeat the above process, continuously reduce the number of features, and recalculate the weights and feature selection in each round of iteration, and stop iterating when the predetermined number of iterations is reached or the selected subset of features is stable, and finally output the subset of features that contributes the most to the predictive ability of the model *(Tang et al., 2021; Xing et al., 2021)*.

ISFLA algorithm starts by randomly generating a certain number of frogs, each frog representing a possible solution, which form the initial population. The frogs are sorted by fitness value and divided into subpopulations, each containing a number of frogs. A local search is then performed to determine the optimal frog and the worst frog in each subpopulation. The solution of the optimal frog is the best solution in the current subpopulation, and the solution of the worst frog is the worst solution in the subpopulation. The worst frog updates its position based on the solution of the optimal frog. Specifically, the worst frog moves closer to the optimal frog to increase its fitness value. If the updated position is better than the previous position, the position is accepted; otherwise, further attempts are made to update based on the global optimal solution. After all subpopulations complete the local search, all frogs are remixed to break the local optimum among subpopulations. After mixing, the subgroups are reclassified to continue the next round of local search. In order to avoid the local search from falling into local optima, ISFLA introduces an adaptive step-size adjustment strategy, which dynamically adjusts the step-size of the frogs' jumps based on the feedback during the search process. In some iterations, a more exploratory global search strategy is used to increase the ability of the algorithm to jump out of the local optimum. The local and global search processes are repeated until the stopping condition is satisfied, and the solution of the optimal frog in the whole frog population, i.e., the selected optimal wavelength, is finally output *(Hsu and Wang, 2021; Kongsorot et al., 2022).*

Modeling method

The modeling methods utilized in this study comprise Support Vector Machines (SVM), Partial Least Squares Discriminant Analysis (PLS-DA), K Nearest Neighbor (KNN), Random Forest (RF), and Back Propagation Neural Network (BP).

The goal of SVM is to find a hyperplane that maximizes the interval between classes in order to classify the data. Specifically, a straight line (hyperplane) is computationally found to separate several classes of data while maximizing the distance between two classes of data points and the hyperplane. The closest data points to the hyperplane, known as support vectors, determine the position and direction of the hyperplane. The hyperplane is then adjusted to maximize the distance from the support vectors to the hyperplane. Finally, for new data points, their class is determined based on their position relative to the decision boundary (hyperplane) *(Khairunniza-Bejo et al., 2021).*

PLS-DA is a supervised learning method that combines Partial Least Squares Regression (PLS) and Linear Discriminant Analysis (LDA), and is commonly used for classification tasks, especially in high-dimensional, small-sample datasets. PLS-DA maximizes the variance of the samples on these variables by searching for latent variables, and minimizes the confounding between different classes, thereby achieving the classification purpose. Specifically, the input dataset X and the corresponding category labels Y are searched for a set of latent variables, T, so that these variables explain the maximum covariance between X and Y. Through an iterative approach, the weight vector W is computed to derive the latent variables $T = XW$, and a linear regression model of the latent variables T to the category labels Y is built, thus transforming the classification problem into a regression problem. Solve for the regression coefficient B so that $Y \approx TB$. Finally, use the results of the regression model to classify the new sample. The new sample is projected into the latent variable space and its category is determined based on the output of the regression model *(Daniels et al., 2021)*.

KNN A simple and intuitive supervised learning algorithm mainly used for classification and regression tasks. KNN selects the nearest K neighbors by calculating the distances between the samples to be classified and the training samples, and then determines the class of the samples to be classified by the class information of these neighbors. Specifically, the input data consists of a feature matrix X and corresponding labels Y, where X is a sample dataset containing multiple features, and Y is the category label corresponding to each sample, and the Euclidean distance metric Manhattan distance, etc. are chosen to compute the distances between the samples. For the to-be-classified sample, calculate its distance from all the samples in the training set and select the nearest K samples as its neighbors. Finally, based on the categories of these K neighbors, the categories of the samples to be classified are decided by voting, and the category with the highest frequency of occurrence is selected as the final classification result *(Cunningham and Delany, 2021)*. In this study, Euclidean distance is utilized to calculate the distances between points in the feature space. Cross-validation is employed to determine the optimal score of K *(Uddin et al., 2022).*

RF does this by generating multiple decision trees, each trained independently on a subset of the data, and then combining the predictions of these trees by majority voting. Specifically, the input dataset X and corresponding labels Y are input, multiple subsample sets are randomly selected from the original dataset with putback (i.e., Bootstrap sampling), each of which is used to train a decision tree, and during the construction of each tree, for each node split, a portion of the features are randomly selected to determine the optimal split thus increasing the diversity of the tree and reducing overfitting. Each decision tree is trained on the corresponding Bootstrap sample set to generate multiple different decision tree models, and finally, for new data points, the classification results of each decision tree are voted on and the category with the most votes is selected as the final prediction *(Wang et al., 2021).*

BP is a training algorithm for multilayer perceptron (MLP) which adjusts the network weights by calculating the errors and back propagating them so that the network is able to learn and approximate complex nonlinear functions. BP neural network consists of an input layer, a hidden layer, and an output layer, and each layer consists of a number of neurons. The core idea is to adjust the weights and bias of the network through the back propagation algorithm to minimize the error between the predicted output and the actual output, in which the input layer receives the input data, the hidden layer performs a linear transformation of the input data through the weights and bias, and then performs a nonlinear transformation through the activation function (e.g., sigmoid, ReLU, etc.), and the output layer performs a transformation of the output again through the weights and bias and through the activation function to get the final predicted output. The predicted output of the output layer is compared with the actual output (labels) and the error is calculated.

Starting from the output layer, the gradient of the error with respect to each weight and bias is calculated. By chain rule, forward propagation is performed layer by layer, calculating the gradient of each layer, and using gradient descent, the weights and biases of each layer are adjusted according to the calculated gradient to reduce the error. The process of forward propagation, error calculation and back propagation is performed repeatedly until the error converges to a preset threshold or the maximum number of iterations is reached *(Wang et al., 2023)*.

In this paper, spectral data were processed and modelled and analyzed using MATLAB 2023b.

Evaluation criteria

For the multi-class classification task of corn kernel quality detection, average evaluation metrics are more suitable. Therefore, this study uses accuracy, average recall, average precision, and average F1 score as evaluation metrics.

Acc is the ratio of the number of correctly classified samples to the total number of samples.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

where:

TP and TN represent the true prediction of the positive sample and the true prediction of the negative sample, respectively. FP and FN represent the false prediction of the positive sample and the false prediction of the negative sample, respectively.

P is the arithmetic mean of the precision (P) of each category, which is defined in the following:

$$P = \frac{1}{k}\sum_{i}^{k} P_i = \frac{1}{k}\sum_{i}^{k} \frac{TP_i}{TP_i + FP_i} \tag{6}$$

where:

i represents the target classes, k is set to 4, as four types of corn kernels are classified, P is the proportion of the predicted positive samples that are actually positive.

R is the arithmetic mean of the recall (R) of each category. It is defined as follows:

$$R = \frac{1}{k}\sum_{i}^{k} R_i = \frac{1}{k}\sum_{i}^{k} \frac{TP_i}{TP_i + FN_i} \tag{7}$$

$F1$ is the weighted harmonic average of P and R scores, which is defined as follows:

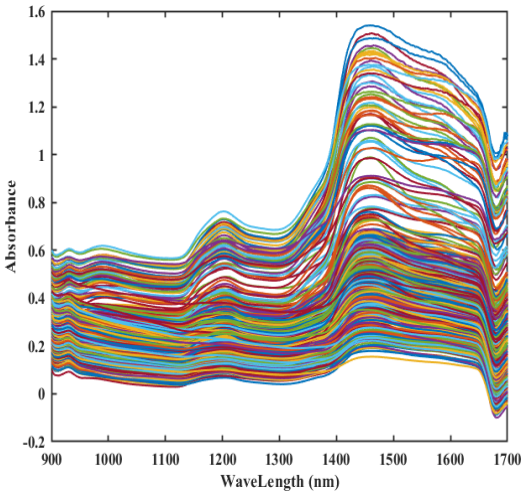
$$F1 = \frac{1}{k}\sum_{i}^{k} \frac{2 \bullet P_i \bullet R_i}{P_i + R_i} \tag{8}$$

RESULTS

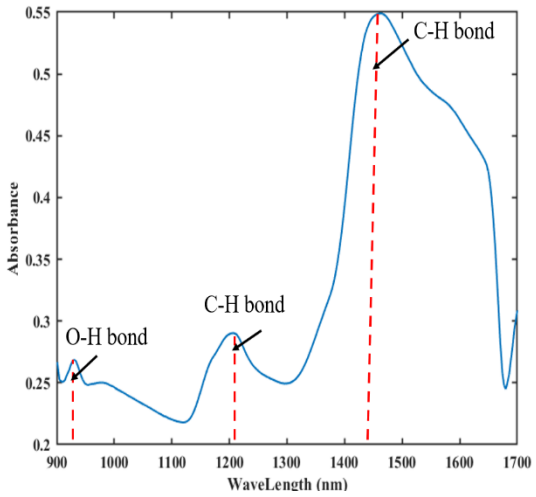
Analysis of original spectral curve

The spectral raw data were obtained using the experimental setup described above, as shown in Fig. 3(a). The overall shapes of the spectral curves of all the maize seed samples were similar. From Fig. 3(b), it can be clearly seen that there are obvious peaks at 930 nm, 1210 nm and 1460 nm, and clear troughs at 910 nm, 1220 nm and 1300 nm. The 930 nm peak corresponds to water absorption, associated with the first multiplicative frequency of the O-H bond stretching vibration in relevant carbohydrates. The 1210 nm peak corresponds to the first multiplicative frequency of the C-H bond stretching vibration in fats and oils within maize kernels, particularly fatty acids. The 1460 nm peak corresponds to C-H bond changes in fibers and polysaccharides inside the maize kernel.

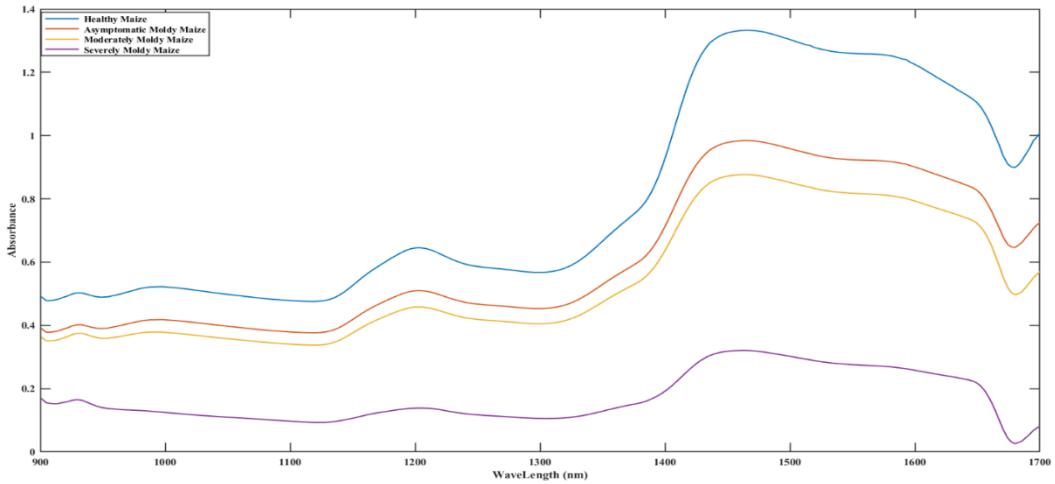
Fig 3(c) shows that the absorbance of maize kernels with different degrees of mold is different, which is mainly due to aflatoxin, erythromycin, vomitoxin and other molds, which will change the hydroxyl (OH), hydrocarbon (CH) and carbonyl (C=O) groups of maize kernels. Because of the complex compounds contained in the maize kernels, the spectral profiles are more complicated and correspond to the octave and frequency information of different chemical bond vibrations, which is the main reason for the different spectral profiles of different molded maize. It was also verified that it is feasible to use microscopy to observe asymptomatic moldy maize.



(a) raw spectral absorbance plot



(b) average spectral absorbance plot

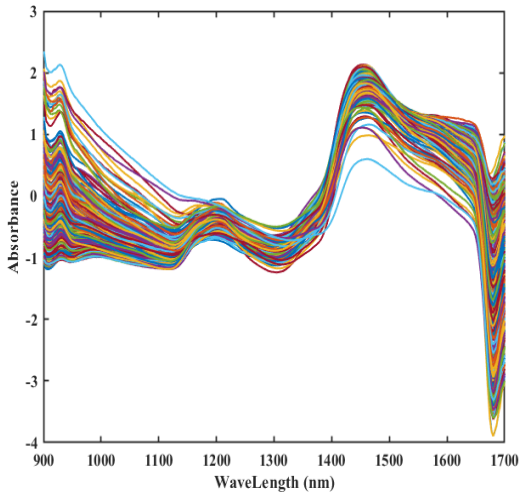


(c) Mean spectra of different maize kernel mold infection levels

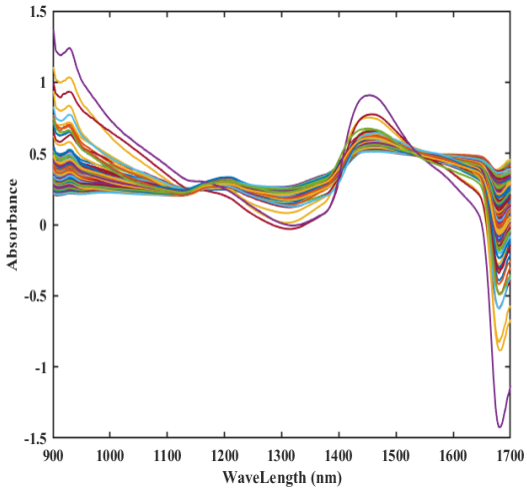
Fig. 3 -(a) raw spectral absorbance plot, (b) average spectral absorbance plot for all maize kernel samples and (c) Mean spectra of different maize kernel mold infection levels

Analysis of spectral data preprocessing results

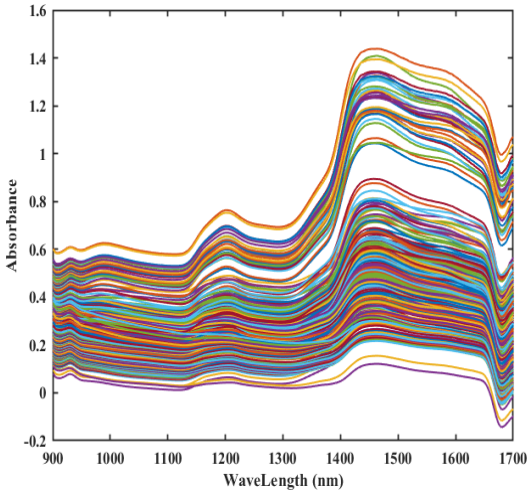
Preprocessing methods and combinations play a crucial role in reducing the interference of irrelevant information. The spectral data underwent preprocessing using SNV, MSC, SG, BC, SG+SNV, SG+MSC, and SG+BC. To ensure consistency in the observed sample data after applying different preprocessing methods, the resulting spectra were compared. The spectral curves resulting from the seven preprocessing methods are presented in Fig. 4, clearly demonstrating their distinct effects. From Fig. 4(a) and Fig. 4(b), it is evident that the spectral curves resulting from the SNV and MSC processing exhibit no significant or fluctuating trends. These methods, both belonging to the scattering processing category, effectively mitigate the scattering effect during spectral data acquisition. They enhance the signal-to-noise ratio of the spectral data, correct the spectral baseline shift, and do not impact the corresponding absorption data. The primary objective of SNV is to mitigate the scattering effect and eliminate optical range changes in the spectra. In contrast, MSC intends to correct the scattering effects due to uneven particle distribution and varied particle sizes on the sample surface. Fig. 4(c) illustrates the spectral curves obtained by applying the SG convolution smoothing method to the original spectral data. This method removes irrelevant noise from the spectra and results in a smoother curve than the original image. Fig. 4(d) displays the outcomes of BC baseline correction, revealing that the baseline drift of the spectral image improves significantly after correction. Additionally, the peaks and shapes of the spectral image are better aligned than those in the original image. This section analyses the relevance of the data preprocessing results. To evaluate the advantages and disadvantages of various preprocessing methods in this experiment, the next step is to model the preprocessed data.



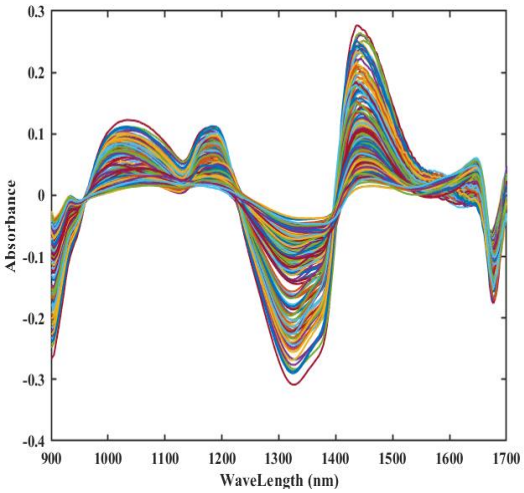
(a) spectrum with the SNV method



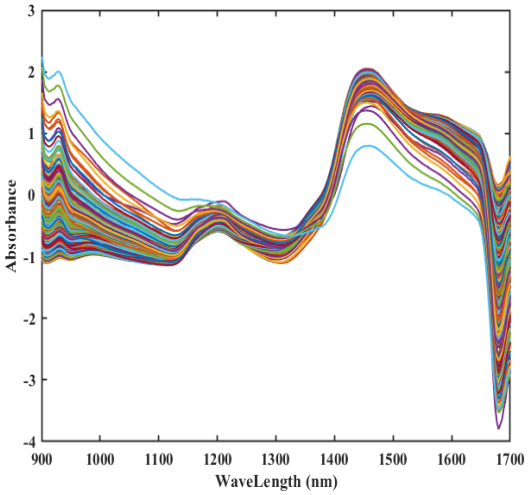
(b) spectrum with the MSC method



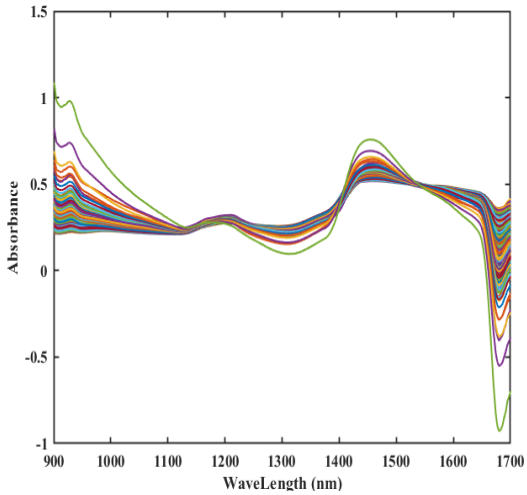
(c) spectrum with the SG method



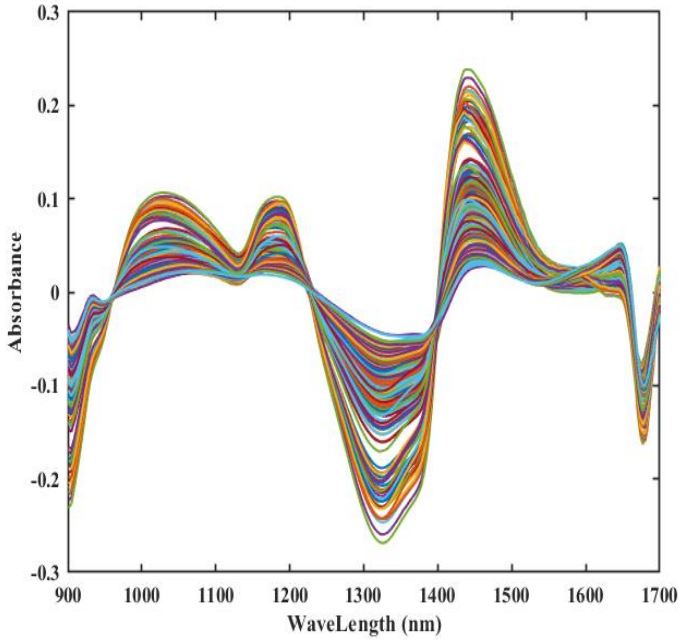
(d) spectrum with the BC method



(e) spectrum with the SG+SNV method



(f) spectrum with the SG+MSC method

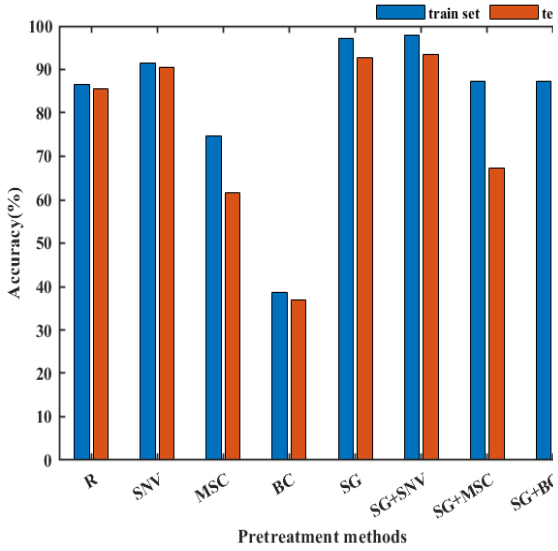


(g) spectrum with the SG+BC method

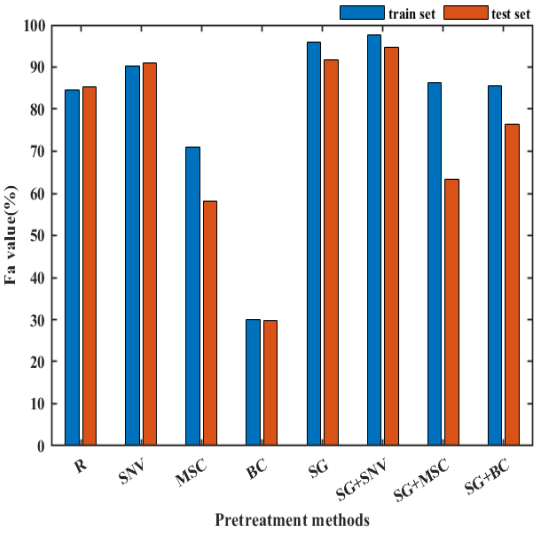
Fig. 4 -Spectral graphs after transformation based on different preprocessing methods

Full-band based modeling

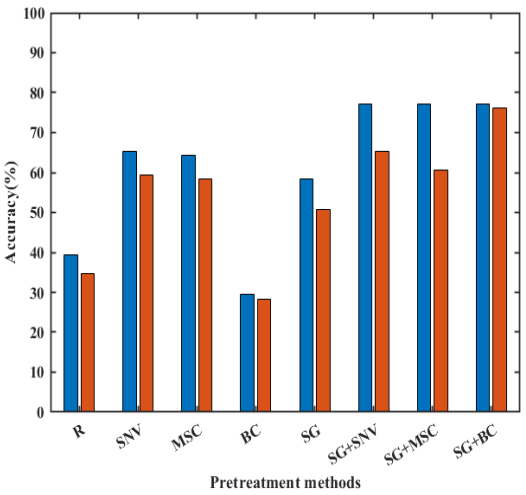
In this paper, five methods, SVM, PLS-DA, RF, KNN and BP, are used to model and analyze the spectral data after different preprocessing. Fig. 5 shows the comparison of accuracy and F1 value of each model with different preprocessing methods. From Fig. 5, it can be clearly seen that, except for the BP neural network and PLS-DA, the preprocessing effect of BC on the other models is poorer. This may be because the baseline and the signal sometimes overlap, leading to the removal of some real information during the BC process. On the other hand, the overall effect of the model after SG processing is better, mainly because SG removes high-frequency noise from the signal, making the spectral curves smoother and highlighting their key characteristics. In addition, for different models, SG+SNV, SG+MSC and SG+BC preprocessing methods all achieve good results, which indicates that the superposition of preprocessing methods is feasible.



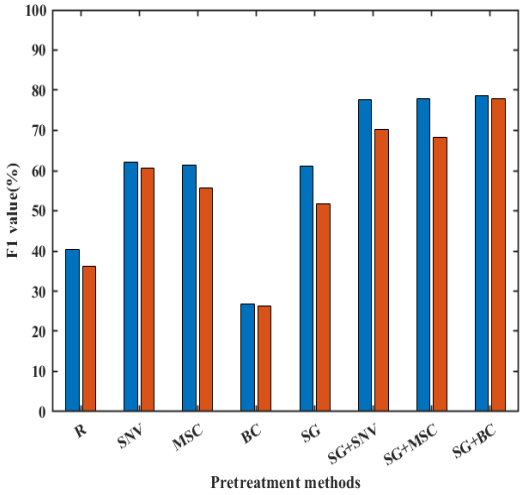
(a) Accuracy of SVM



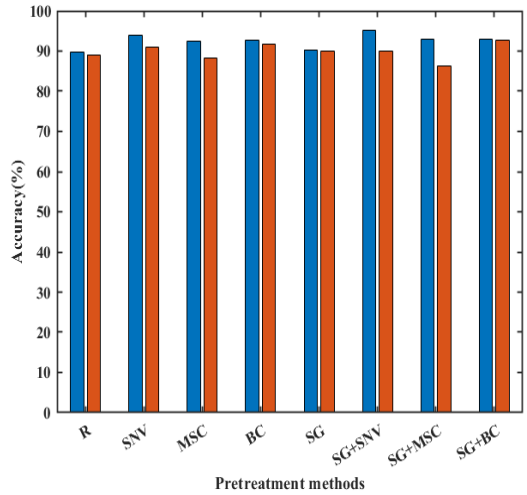
(b) F1 of SVM



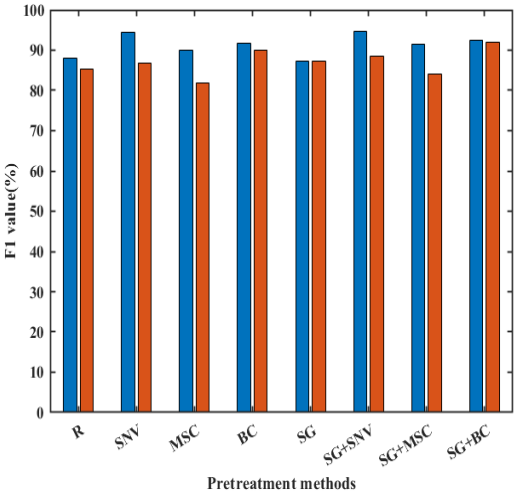
(c) Accuracy of KNN



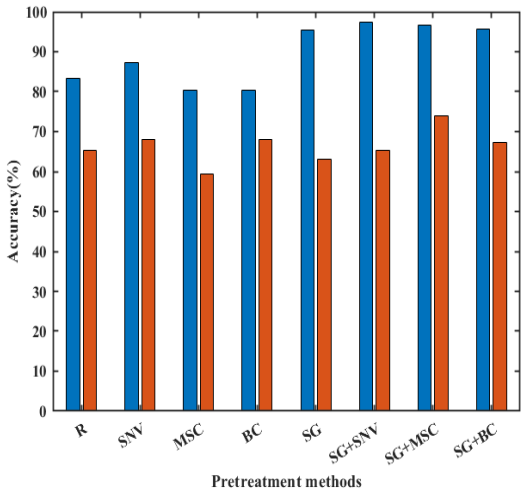
(d) F1 of KNN



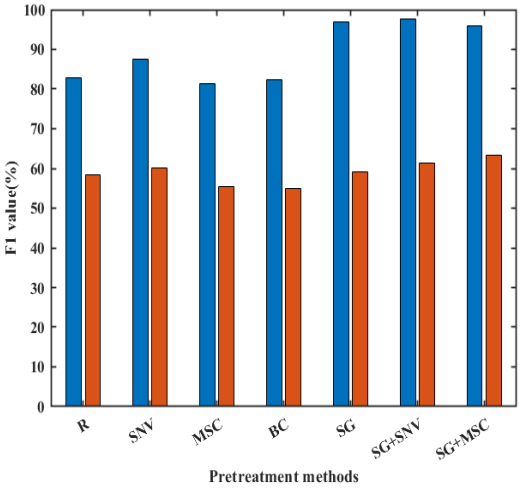
(e) Accuracy of BP



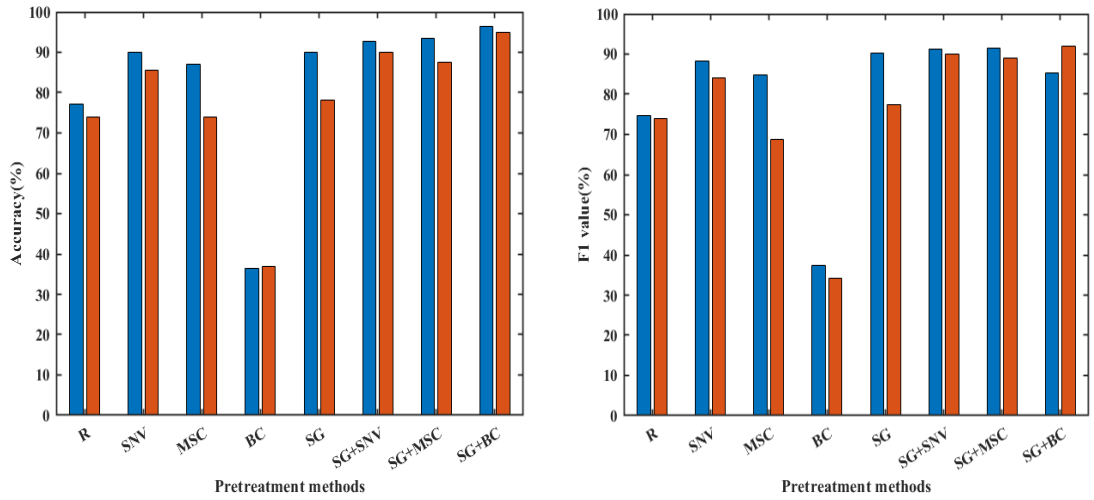
(f) F1 of BP



(g) Accuracy of PLS-DA



(h) F1 of PLS-DA



(i) Accuracy of RF (j) F1 of RF

Fig. 5 - Accuracy and F1 with the testing dataset

As shown in Table 1, regarding the accuracy and F1 of each model, SVM and RF consistently outperformed PLS-DA, KNN, and BP in classifying the extent of mold in maize kernels. The best results obtained by SVM were based on spectral data after SG+SNV preprocessing with 97.98% accuracy and 97.55% F1 value for the training set, 93.48% accuracy and 94.63% F1 value for the validation set, and 88.57% identification accuracy for asymptomatic moldy maize, and the best results obtained by RF were based on spectral data after SG+BC preprocessing that the training set accuracy is 96.33% with F1 value of 95.43% and the validation set accuracy is 94.50% with F1 value of 94.98% and the recognition accuracy for asymptomatic moldy maize is 91.43%, which is slightly higher than the SVM modeling, and both models have high discriminative ability. For the identification of asymptomatic moldy maize, the SG+SNV-SVM model incorrectly identified four asymptomatic moldy maize as three healthy maize and one moderately moldy maize. The SG+BC-RF model incorrectly identified three asymptomatic moldy maize as one healthy maize and two moderately moldy maize, which due to the fact that the surface of the asymptomatic moldy maize is similar to that of the healthy maize, but its internal chemical composition is similar to that of moderate moldy maize, which can easily lead to model misidentification.

Despite PLS-DA achieving high accuracy and F1 values in the training set (97.44% for SG+SNV, 96.71% for SG+MSC, and 95.61% for SG+BC), PLS-DA's performance in the validation set was subpar (65.22% for SG+SNV, 73.83% for SG+MSC, and 67.25% for SG+BC). This disparity arises from the limitations of PLS-DA in handling highly correlated independent variables and category imbalances. On the other hand, KNN demonstrated overall lower effectiveness in modeling both raw and preprocessed data, reaching a maximum accuracy of only 77.25% for SG+BC in the training set and 76.48% in the validation set, significantly below the anticipated performance. This inefficiency is attributed to KNN's inapplicability to data with high dimensionality and unbalanced feature weights. For maize mold samples, SVM, RF and BP gave better classification results due to their robustness and good handling of non-linear and large-scale datasets.

Table 1

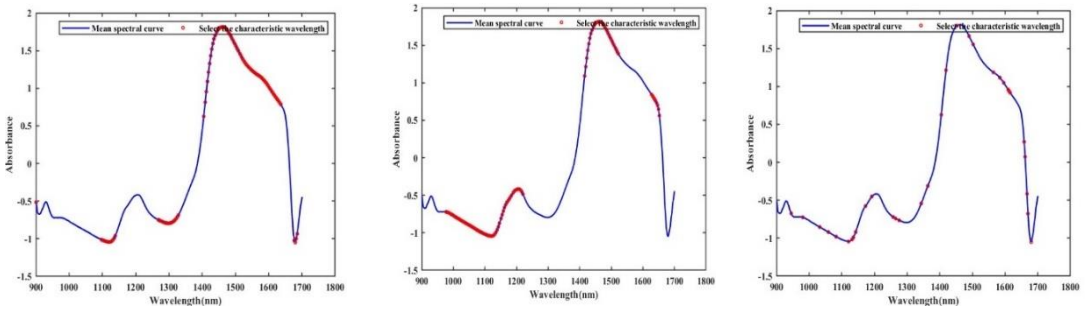
Classification results based on the SVM, PLS-DA, RF, KNN and BP models

Models	Preprocessing Methods	Training Set				Testing Set			
		Accuracy	P	R	F1	Accuracy	P	R	F1
SVM	RAW	86.47%	85.23%	84.08%	84.50%	85.51%	85.06%	86.63%	85.23%
	SNV	91.59%	91.51%	89.36%	90.31%	90.58%	92.87%	89.59%	91.02%
	MSC	74.59%	74.48%	67.92%	71.05%	61.59%	68.51%	66.24%	58.25%
	BC	38.57%	37.62%	25.00%	30.04%	36.96%	36.45%	25.00%	29.66%
	SG	97.06%	96.45%	95.34%	95.86%	92.75%	90.44%	94.24%	91.78%
	SG+SNV	**97.98%**	**98.35%**	**96.87%**	**97.55%**	**93.48%**	**95.01%**	**94.73%**	**94.63%**
	SG+MSC	87.34%	91.42%	83.82%	86.32%	67.39%	70.94%	71.46%	63.34%
	SG+BC	87.34%	91.72%	83.64%	85.47%	78.99%	86.33%	78.70%	76.35%

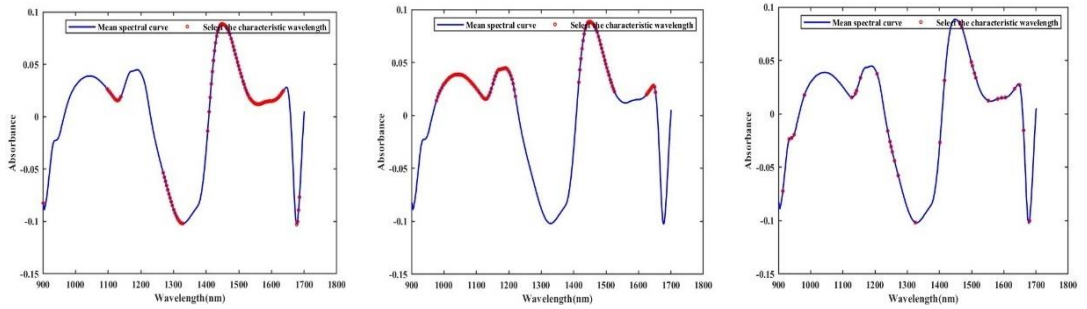
Models	Preprocessing Methods	Training Set				Testing Set			
		Accuracy	P	R	F1	Accuracy	P	R	F1
PLSDA	RAW	83.36%	81.54%	88.51%	82.79%	65.22%	63.65%	50.00%	58.52%
	SNV	87.20%	87.31%	88.24%	87.62%	68.12%	68.23%	53.68%	60.09%
	MSC	80.45%	80.37%	84.16%	81.34%	59.42%	60.35%	51.23%	55.42%
	BC	80.27%	81.35%	85.98%	82.30%	68.12%	67.25%	53.26%	59.45%
	SG	95.42%	96.70%	97.22%	96.85%	63.04%	64.59%	48.08%	54.97%
	SG+SNV	97.44%	98.41%	96.96%	97.61%	65.22%	64.89%	54.33%	59.14%
	SG+MSC	96.71%	96.54%	94.83%	95.68%	73.83%	72.56%	53.21%	61.40%
	SG+BC	95.61%	94.58%	97.67%	95.90%	67.25%	66.57%	60.54%	63.41%
KNN	RAW	39.31%	46.14%	49.53%	40.38%	34.78%	41.52%	44.65%	36.20%
	SNV	65.25%	62.24%	67.74%	62.15%	59.42%	60.41%	65.61%	60.54%
	MSC	64.35%	62.62%	62.20%	61.38%	58.41%	52.93%	58.76%	55.69%
	BC	29.62%	28.76%	25.00%	26.75%	28.26%	27.56%	25.00%	26.22%
	SG	58.32%	62.79%	66.26%	61.07%	50.72%	50.72%	58.53%	51.85%
	SG+SNV	77.15%	77.19%	83.38%	77.75%	65.22%	66.92%	74.04%	70.30%
	SG+MSC	77.15%	77.49%	82.40%	77.92%	60.58%	65.00%	71.75%	68.21%
	SG+BC	77.25%	81.00%	84.44%	78.76%	76.23%	75.34%	80.58%	77.87%
RF	RAW	77.06%	76.01%	73.73%	74.57%	73.91%	74.52%	74.57%	73.89%
	SNV	89.91%	92.72%	85.40%	88.23%	85.51%	87.64%	82.07%	84.17%
	MSC	87.16%	87.88%	82.71%	84.80%	73.92%	72.36%	65.39%	68.70%
	BC	39.45%	38.36%	36.57%	37.44%	36.96%	34.86%	33.58%	34.21%
	SG	89.91%	89.82%	90.67%	90.16%	78.26%	79.77%	80.10%	77.43%
	SG+SNV	92.66%	92.39%	90.41%	91.33%	89.86%	90.90%	90.93%	89.82%
	SG+MSC	93.58%	93.36%	89.85%	91.37%	87.56%	89.73%	88.25%	88.98%
	SG+BC	**96.33%**	**96.02%**	**94.96%**	**95.43%**	**94.50%**	**94.05%**	**96.24%**	**94.98%**
BP	RAW	89.68%	87.77%	88.30%	87.95%	89.09%	84.88%	86.82%	85.37%
	SNV	94.04%	94.59%	94.51%	94.36%	90.91%	86.69%	87.56%	86.87%
	MSC	92.43%	88.85%	91.28%	89.94%	88.18%	81.48%	84.27%	81.97%
	BC	92.66%	93.59%	90.20%	91.66%	91.82%	89.68%	90.73%	90.12%
	SG	90.37%	91.20%	85.09%	87.25%	90.09%	87.13%	87.36%	87.17%
	SG+SNV	95.18%	95.44%	94.18%	94.72%	90.09%	88.26%	89.26%	88.48%
	SG+MSC	92.89%	91.44%	92.01%	91.57%	86.36%	83.37%	85.74%	84.17%
	SG+BC	**92.89%**	**94.46%**	**90.66%**	**92.52%**	**92.73%**	**93.48%**	**90.88%**	**91.88%**

Characteristic wavelength extraction

Based on the results of the established full-band model it can be seen that the model classification under SG+SNV and SG+BC preprocessing is better, so SPA, CARS, and ISFLA were used to extract the feature wavelengths from the spectral data after SG+SNV and SG+BC preprocessing. In the case of SG+SNV, three feature wavelength extractions were performed, yielding 107, 113, and 32 feature wavelengths, respectively. For SG+BC, 107, 109, and 33 wavelengths were extracted. As shown in Fig 6, the feature extraction wavelengths for SG+BC-SPA were concentrated in the range of 1150 nm-1250 nm and 1300 nm to 1700 nm, while those for SG+BC-CARS were mainly concentrated in the range of 980 nm-1220 nm and 1410 nm-1650 nm. For the SG+BC-ISFLA method, a reduced number of feature wavelengths were identified, yet they predominantly clustered within the spectral ranges of 1100nm-1300nm and 1300nm-1650nm. These findings suggest that the maize kernel mold-related compounds associated with these feature wavelengths are consistently concentrated within their respective mentioned ranges. Although the specific values of the extracted feature wavelengths may vary across spectral curves, the observed similarity in their distribution ranges underscores the consistent concentration of these compounds in the aforementioned spectral intervals.



(a) feature selection result of the SG+SNV-SAP (b) feature selection result of the SG+SNV-CARS (c) feature selection result of the SG+SNV-ISFLA



(d) feature selection result of the SG+BC-SPA (e) feature selection result of the SG+BC-CARS (f) feature selection result of the SG+BC-ISFLA

Fig. 6 -Feature extraction results

Modeling of characteristic wavelengths

According to Table 2, the models constructed after feature extraction demonstrate improvements in prediction accuracy and F1 value *(Zhang et al., 2022)*. Specifically, for the SG+SNV-SVM model, ISFLA feature extraction yields the best results with an F1 value of 97.79%, indicating a 3.16% improvement over the pre-feature extraction performance. Conversely, SPA feature extraction shows the least favorable outcome for the SG+SNV-SVM model, resulting in a 0.63% increase compared to pre-extraction. In the case of the SG+BC-RF model, ISFLA feature extraction is the most effective, achieving an F1 value of 96.91% and a 1.93% improvement over the pre-extraction F1 value. However, SPA feature extraction for the SG+BC-RF model shows a minimal improvement of 0.09%, representing the least effective method. Finally, for the SG+BC-BP model, ISFLA feature extraction stands out with the highest F1 value of 95.20%, demonstrating a 3.32% improvement over pre-feature extraction. SPA feature extraction for SG+BC-BP achieves a 2.31% improvement compared to the other two methods. SG+SNV-SVM-ISFLA had the highest classification ability for moldy maize, with an F1 value of 97.22%. The recognition accuracy for asymptomatic moldy maize reached 96.30%, and this sample is a biological sample of high complexity, and the classification index of SG+SNV-SVM-ISFLA model is acceptable in this study.

In summary, while different feature extraction methods exhibit varying effects on the models, there is an overall enhancement in accuracy. This underscores the capability of the feature extraction process to eliminate irrelevant wavelengths and preserve information-rich features, thereby improving the predictive performance of the model.

Table 2

Modeling of maize grain mildew degree under different feature extraction methods.

Models	Characteristic Wavelength Extraction Method	Train set				Test set			
		Accuracy	P	R	F1	Accuracy	P	R	F1
SG+SNV-SVM	SPA	97.25%	97.49%	96.48%	96.91%	97.10%	97.12%	98.08%	97.54%
	CARS	97.61%	98.11%	97.81%	97.96%	97.08%	97.49%	97.87%	97.68%
	ISFLA	**98.16%**	**98.86%**	**97.04%**	**97.94%**	**97.22%**	**98.03%**	**97.56%**	**97.79%**
SG+BC-RF	SPA	98.17%	97.15%	97.15%	97.15%	95.63%	95.03%	95.12%	95.07%
	CARS	98.17%	98.89%	95.45%	96.93%	96.35%	95.90%	96.65%	96.13%
	ISFLA	96.33%	97.83%	96.37%	97.09%	96.30%	97.77%	96.43%	96.91%
	SPA	96.78%	97.20%	95.59%	96.25%	95.41%	95.47%	93.61%	94.19%

Models	Characteristic Wavelength Extraction Method	Train set				Test set			
		Accuracy	P	R	F1	Accuracy	P	R	F1
SG+BC-BP	CARS	97.47%	97.41%	96.09%	96.67%	95.41%	96.05%	93.92%	94.97%
	ISFLA	97.93%	98.23%	96.84%	97.46%	96.33%	96.71%	94.31%	95.20%

CONCLUSIONS

The aim of this paper is to investigate the use of near-infrared spectral fingerprints to classify moldy maize while at the same time accurately identifying asymptomatic moldy maize thereby reducing the probability of early mass infection of stored maize. Firstly, different preprocessing methods are used for comparative analysis, and the results show that the modeling effect of the spectral data after a series of preprocessing methods, such as SNV, MSC, SG, etc., is better than that of the original data, and better results can be obtained by using a combination of multiple preprocessing methods to deal with the original spectra, among which the SG+SNV and SG+BC are the most effective. Secondly, the full-band spectral data classification model was constructed, and the experimental results showed that the accuracy of the SG+BC-RF model for moldy maize classification could reach 94.50%. Then, feature extraction is performed on the different preprocessed data. Finally, the data after feature extraction were modelled, and the experimental results showed that the SG+SNV-SVM-ISFLA model could classify moldy maize with an accuracy of up to 97.22%, and identify asymptomatic moldy maize with an accuracy of up to 96.30%, which meets the identification requirements. This work provides new perspectives and ideas for the classification of moldy maize and the identification of asymptomatic moldy maize, as well as new ideas for early mold safety monitoring of stored maize.

ACKNOWLEDGEMENT

Project of National Natural Science Foundation of China (52105539), Anhui Natural Science Foundation (2108085QD179), National Engineering Technology Research Center (2005DP173065- 2022 - 01);

REFERENCES

[1] An M., Cao C., Wang S., Zhang X., (2023), Analysis, Non-destructive identification of moldy walnut based on NIR. *Journal of Food Composition and Analysis*, Vol. 121, pp. 105407, China.

[2] Bai X., Zhang C., Xiao Q., He Y., (2020), Application of near-infrared hyperspectral imaging to identify a variety of silage maize seeds and common maize seeds. *RSC advances*, Vol. 10, pp. 11707-11715, China.

[3] Cui Y., Ge W., Li J., Zhang J., An D., (2019), Screening of maize haploid kernels based on near infrared spectroscopy quantitative analysis. *Computers and electronics in agriculture*, Vol. 158, pp. 358-368, China.

[4] Cunningham P., Delany S.J., (2021), Delany, k-Nearest neighbour classifiers-A Tutorial. *ACM computing surveys (CSUR)*, Vol. 54, Issue 6, pp. 1-25, Ireland.

[5] Dai J., Tang W., Zhan, J., Kang X., Dai W., Ji J., (2024), Determination and quality evaluation of active ingredients in areca nut using near-infrared rapid detection technology. *Microchemical Journal*, Vol. 196, pp. 109586, China.

[6] Daniels A. J., Poblete-Echeverría C., Nieuwoudt H. H., Botha N., (2021), Classification of browning on intact table grape bunches using near-infrared spectroscopy coupled with partial least squares-discriminant analysis and artificial neural networks. *Frontiers in Plant Science*, Vol. 12, pp. 768046, South Africa.

[7] Hsu H.P., Wang C.N., (2021), A Hybrid Approach Combining Improved Shuffled Frog-Leaping Algorithm With Dynamic Programming for Disassembly Process Planning. *Ieee Access*, Vol. 9, pp. 57743-87756, China.

[8] Jia J., Zhou X., Li Y., Wang M., Liu Z., (2022), Establishment of a rapid detection model for the sensory quality and components of Yuezhou Longjing tea using near-infrared spectroscopy. *LWT*, Vol. 164, pp. 113625, China.

[9] Jiang H., Deng J., Zhu C., (2023), Quantitative analysis of aflatoxin B1 in moldy peanuts based on near-infrared spectra with two-dimensional convolutional neural network. *Infrared physics & technology*, Vol. 131, pp. 104672, China.

[10] Kang Z., Huang T., Zeng S., Li H., Dong L., (2022), A Method for Detection of Corn Kernel Mildew Based on Co-Clustering Algorithm with Hyperspectral Image Technology. *Sensors*, Vol. 22, Issue 14, pp. 5533.

[11] Khairunniza-Bejo S., Shahibullah M.S., Azmi A.N.N., (2021), Non-destructive detection of asymptomatic Ganoderma boninense infection of oil palm seedlings using NIR-hyperspectral data and support vector machine. *Applied Sciences*, Vol. 11, Issue 22, pp. 10878, Malaysia.
[12] Kongsorot Y., Musikawan P., Muneesawang P., (2022), An enhanced fuzzy-based clustering protocol with an improved shuffled frog leaping algorithm for WSNs. *Expert Systems with Applications*, Vol. 198, pp. 116767, Thailand.
[13] Lanjewar M. G., Morajkar P. P., Parab J. S., (2024), Portable system to detect starch adulteration in turmeric using NIR spectroscopy. *Food Control*, Vol. 155, pp. 110095, India.
[14] Li Y., Pan T., Li H., Chen S., (2020), Non-invasive quality analysis of thawed tuna using near infrared spectroscopy with baseline correction. *Journal of Food Process Engineering*, Vol. 43, Issue 8, pp. 13445,China.
[15] Liu L., Zhang H., Wu L., Gu S., Xu J., Jia B., (2023), An early asymptomatic diagnosis method for cork spot disorder in 'Akizuki'pear (Pyrus pyrifolia Nakai) using micro near infrared spectroscopy. *Food Chemistry: X*, Vol. 19, pp. 100851, China.
[16] Long Y., Huang W., Wang Q., (2022), Integration of textural and spectral features of Raman hyperspectral imaging for quantitative determination of a single maize kernel mildew coupled with chemometrics. *Food Chemistry*,Vol. 372, pp. 131246, China.
[17] Long Y., Wang Q., Tian X., (2022), Screening naturally mildewed maize kernels based on Raman hyperspectral imaging coupled with machine learning classifiers. *Journal of Food Process Engineering*, Vol. 45, Issue 11, pp. 14148, China.
[18] Malavi D., Nikkhah A., Alighaleh P., Einafshar S., Raes K., (2024), Detection of saffron adulteration with Crocus sativus style using NIR-hyperspectral imaging and chemometrics. *Food Control*, Vol. 157, pp. 110189, Belgium.
[19] Milićević D., Nikšić M., Baltić T., (2010), Isolation, characterization and evaluation of significant mycoflora and mycotoxins in pig feed from Serbian farms. *World Journal of Microbiology and Biotechnology*, Vol. 26, pp. 1715-1720, Serbia.
[20] Ong P., Jian J., Li X., Zou C., Yin J., (2023), New approach for sugarcane disease recognition through visible and near-infrared spectroscopy and a modified wavelength selection method using machine learning models. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, Vol. 302, pp. 123037, China.
[21] Oyebanji A O., Efiuvwevwere B.J.O., (1999), Growth of spoilage mould and aflatoxin B1 production in naturally contaminated or artificially inoculated maize as influenced by moisture content under ambient tropical condition. *International biodeterioration & biodegradation*, Vol.44, Issue 4, pp. 209-217, Nigeria.
[22] Pang L., Wang L., Yuan P., Yan L., (2022), Rapid seed viability prediction of Sophora japonica by improved successive projection algorithm and hyperspectral imaging. *Infrared Physics & Technology*, Vol. 123, pp. 104143, China.
[23] Paraginski R.T., Colussi R., Dias A.R.G., da Rosa Zavareze E., Elias M. C., (2019), Physicochemical, pasting, crystallinity, and morphological properties of starches isolated from maize kernels exhibiting different types of defects. *Food Chemistry*, Vol. 274, pp. 330-336, Brazil.
[24] Shen F., Wei Y., Zhang B., Shao X., (2018), Rapid detection of harmful mold infection in rice by near infrared spectroscopy. *Spectroscopy and Spectral Analysis*, Vol. 38, pp. 3748-3752, China.
[25] Sun J., Zhou X., Hu Y., Wu X., Zhang X., (2019), Visualizing distribution of moisture content in tea leaves using optimization algorithms and NIR hyperspectral imaging. *Computers and Electronics in Agriculture*, Vol. 160, pp. 153-159, China.
[26] Sun K., Zhang Y.J., Tong S.Y., Tang M.D., (2022), Study on rice grain mildewed region recognition based on microscopic computer vision and YOLO-v5 model. *Foods*, Vol. 11, Issue 24, pp. 4031, China.
[27] Tang N., Sun J., Yao K., Zhou X., Tian Y., Cao Y., (2021), Identification of Lycium barbarum varieties based on hyperspectral imaging technique and competitive adaptive reweighted sampling-whale optimization algorithm-support vector machine. *Journal of Food Process Engineering*, Vol. 44, Issue 1, pp. 13603, China.
[28] Uddin S., Haque I., Lu H., Moni M. A., (2022), Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. *Sci Rep 12*, Vol. 12, pp. 6256, Australia.
[29] Vallese F. D., Paoloni S. G., Springer V., de Sousa Fernandes D. D., (2023), Analysis, Exploiting the successive projections algorithm to improve the quantification of chemical constituents and

discrimination of botanical origin of Argentinean bee-pollen. *Journal of Food Composition and Analysis*, Vol. 126, pp. 105925, Argentina.

[30] Wang L., Huang Z., Wang R., (2021), Discrimination of cracked soybean seeds by near-infrared spectroscopy and random forest variable selection. *Infrared Physics & Technology*, Vol. 115, pp. 103731, China.

[31] Wang X., Jiang Z., Ji R., Han Y., Bian H., Yang, Y., (2003), Detection of Ningnanmycin Using Fluorescence Spectroscopy Combined with BP Neural Network. *Combinatorial Chemistry & High Throughput Screening*, Vol. 26, pp. 1414-1423, China,

[32] Xing Z., Du C., Shen Y., Ma F., (2021), A method combining FTIR-ATR and Raman spectroscopy to determine soil organic matter: Improvement of prediction accuracy using competitive adaptive reweighted sampling (CARS). *Computers and Electronics in Agriculture*, Vol. 191, pp. 106549, China.

[33] Xu L., Zhou Y. P., Tang L.J., Wu H.L., Jiang J.H., (2008), Ensemble preprocessing of near-infrared (NIR) spectra for multivariate calibration. *Analytica chimica acta*, Vol. 616, Issue 2, pp. 138-143, China.

[34] Zhang G., Tuo X., Zhai S., Zhu X., Luo L., (2022), Near-infrared spectral characteristic extraction and qualitative analysis method for complex multi-component mixtures based on TRPCA-SVM. *Sensors*, Vol. 22, Issue 4, pp. 1654, China.

[35] Zhang J., Liu L., Chen Y., Rao Y., Zhang X., (2023), The Nondestructive Model of Near-Infrared Spectroscopy with Different Pretreatment Transformation for Predicting “Dangshan” Pear Woolliness Disease. *Agronomy*, Vol. 13, Issue 5, pp. 1420, China.

[36] Zhang Q., Liu C., Sun J., Cui Y., Li Q., Rapid non-destructive detection for molds colony of paddy rice based on near infrared spectroscopy. *Journal of Northeast Agricultural University* (English Edition), Vol. 21, Issue 4, pp. 54-60, China.

[37] Zhang Y., Gao W., Cui C., Zhang Z., He L., Zheng J., (2020), Development of a method to evaluate the tenderness of fresh tea leaves based on rapid, in-situ Raman spectroscopy scanning for carotenoids. *Food Chemistry*, Vol. 308, pp. 125648, China.