Journal of Food Measurement and Characterization
https://doi.org/10.1007/s11694-024-02841-y

ORIGINAL PAPER

Improving discrimination accuracy of pest-infested crabapples using Vis/NIR spectral morphological features

Yuanhao Zheng[1,2] · Ying Zhou[3] · Penghui Liu[1,2] · Yingjie Zheng[1,2] · Zichao Wei[1,2] · Zetong Li[1,2] · Lijuan Xie[1,2]

Received: 14 June 2024 / Accepted: 22 August 2024

Abstract

The visible/near-infrared (Vis/NIR) spectroscopy technique is effective for fruit quality detection. The distinct spectral features can reflect the internal composition of fruits, while variations in external orientation may induce interference. Considering both external and internal factors, we improved the discrimination accuracy of pest-infested crabapples by compensating for variations in orientation and amplifying differences in spectral morphological features (SMFs). Firstly, spectral intensity variations caused by orientations and morphological differences caused by pest infestation were analyzed. Based on these differences, the global model was established to mitigate the external orientation influence. Subsequently, SMFs, derived from spectral peaks and troughs, were employed to amplify spectral features. Finally, with the supplementation using 1st deviation, SMFs improved the discrimination performance of the partial least square–linear discriminant analysis (PLS-LDA) model for pest infestation, yielding results of sensitivity, specificity, and accuracy as 95.14%, 96.32%, and 95.94%, respectively. Overall, compensating for external orientation variations and exploiting internal spectral features enhanced the detection accuracy of pest infestation, providing valuable insights for internal defect discrimination based on Vis/NIR spectroscopy.

Keywords Visible/Near-infrared spectra · Discriminant analysis · Spectral morphological feature · Global model compensation

Introduction

Pest infestation poses significant threats to the fruit industry, leading to substantial losses in agricultural and food production worldwide [1]. Devastating pests such as the codling moth (CM, Cydia pomonella) cause damage to apples, pears, and other fruits through shallow-feeding scars, direct damage to the pulp or seeds, and indirect contamination by larval feces [2]. These symptoms not only cause economic and ecological losses [3] but also potentially endanger consumers' health. Therefore, the accurate and rapid detection of internal pest damage in fruits is crucial before commercialization.

A multitude of non-destructive techniques have been developed to distinguish insect infestation under post-harvest conditions. Currently, hyperspectral imaging (HSI), magnetic resonance imaging (MRI), and X-ray imaging (XRI) are impractical for large-scale online detection due to their cost, redundant data, and time consumption [4]. Machine vision (MV) also encounters obstacles because there are no obvious surface defects when larvae bore their way into the fruit through the calyx [5]. Visible/near-infrared (Vis/NIR) spectroscopy, combined with chemometric methods, has been considered a feasible technique for in-line inspection of the entire production [6]. The range of Vis/NIR radiation has good penetration abilities through biological tissues. As photons traverse through the fruit, they interact with hydrogen-containing groups (C-H, O-H, N-H) [7], allowing for the acquisition of information about the internal components. Based on the spectral information, the

✉ Lijuan Xie
ljxie@zju.edu.cn

1 College of Biosystems Engineering and Food Science, Zhejiang University, 866 Yuhangtang Road, Hangzhou 310058, P.R. China

2 Key Laboratory of Intelligent Equipment and Robotics for Agriculture of Zhejiang Province, Hangzhou 310058, P.R. China

3 Hangzhou Customs Technology Center, 398 Jianshe San Road, Xiaoshan District, Hangzhou 310058, P.R. China

Published online: 13 September 2024

Vis/NIR spectroscopy can capture the differences between healthy and pest-infested fruits. However, in practical applications, it is worth exploring how to eliminate the influence of external physical properties [8] and amplify the spectral differences of internal biological features [9] to improve the Vis/NIR detection accuracy.

Regarding external physical properties, orientation is one of the factors influencing the Vis/NIR model performance [10]. The interaction between samples and photons may vary with random orientations, potentially causing spectral changes due to variations in optical path length and propagation characteristics [11, 12]. Some spectral correction methods may not apply to online practice because of the imponderable influence of random orientations [13]. In this scenario, the 'global model' incorporating spectral information from various orientations could serve as a solution to mitigate the impact of random orientations. For example, the global model outperforms the local model in predicting the soluble solid content (SSC) of apples [14] and identifying moldy apple cores [15]. Thus, a global model is adopted to mitigate the influence of fruit orientations on pest infestation detection to meet the requirements of in-line inspection.

Regarding internal biological features, the Vis/NIR discrimination of pest infestation primarily relies on spectral differences caused by variances in fruit tissue or components [16]. For example, codling moth invasion can result in the consumption of fruit pulp and seeds, leading to tissue browning and the accumulation of metabolic frass [17]. These changes manifest as corresponding features in Vis/NIR spectra. Spectral morphological features (SMFs) can intuitively reflect these differences using indices such as spectral difference (SD), spectral ratio (SR), and normalized spectral intensity difference (NSID) [18]. These SMFs indices have been successfully applied to enhance the prediction performance of SSC, maturity, and dry matter content (DMC) in apples [19, 20], moisture content in pork [21], internal flaws of hazelnuts [22], and ripening stages of peaches [23]. However, typically only a few spectral peaks/valleys are utilized to calculate SMFs indices, which may lead to underutilization of information due to the overlapping nature of NIR spectra [24]. Therefore, it is crucial to investigate methods to fully exploit wavelengths associated with spectral shape to improve the representativeness and accuracy of SMFs.

Chinese pear-leaved crabapple (Malus asiatica), which is gradually gaining popularity among consumers but facing the challenge of pest infestation [25], is chosen as the illustrative sample for pest damage detection. The purpose and novelty of this study are to enhance the discrimination accuracy of healthy/infested crabapples through raw data quality improvement, by mitigating the influence of external orientations and amplifying the spectral differences associated with internal features using SMFs. The specific contents are to (1) analyze spectral differences caused by orientations and internal properties, (2) mitigate the impact of crabapple orientations using the global model, (3) improve the discrimination accuracy of pest infestation for crabapples using SMFs, and (4) further enhance SMFs-based model performance through spectral information supplementation.

Materials and methods

Samples

The crabapple samples were obtained from a post-harvest processing factory in Hulunbuir City, Inner Mongolia Autonomous Region, China. These samples were selected through simple random sampling from the factory, including crabapples with varying degrees of pest damage, from very mild to severe, regardless of whether pest infestation was visible on the exocarp. This approach ensured that the distribution of healthy and pest-damaged crabapples reflected real-world scenarios. These samples were transported under refrigeration to our laboratory in Hangzhou, Zhejiang Province, where subsequent spectra collection and model construction were conducted. After being numbered, these crabapples were stored at ambient conditions (25°C) for 12 h to eliminate the effects of refrigerated transport temperature on spectral collection [26].

Experiments

Spectra collection

The transmittance spectra within 400–1100 nm were obtained using a self-developed inspection system, as shown in Fig. 1a. The spectral acquisition system is similar in design to Vis/NIR spectral acquisition modules on the commercial sorting line. Detailed information about the spectral acquisition system can be found in Zheng et al. [27]. The power of the halogen lamp was set to 70 W to provide adequate illumination. Each crabapple was placed on the fruit holder in four orientations alternately, as shown in Fig. 1b, to assess the influence of orientation on the spectra. After interacting with the samples, the light was collected by a spectrometer (QE65 Pro, Ocean Optics, USA) through an optical fiber. All spectra were collected in the dark box with an integration time of 50 ms and automatically smoothed in the SpectrumSuite software.

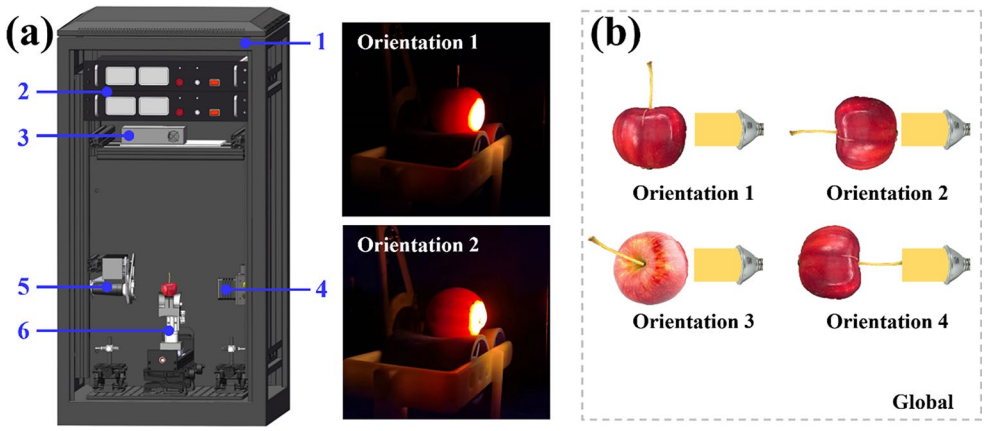

Fig. 1 Vis/NIR spectra acquisition system: **(a)** structure diagram of the acquisition system (1-dark box, 2-power system, 3-spectrometer, 4-light source, 5-receiver, 6-fruit holder) and spectrum acquisition cases, **(b)** schematic diagram of different detection orientations

Pest information acquisition

After the spectral acquisition, the size information (length and height) of each sample was measured using a vernier caliper and recorded as the optical path length of these four orientations. Then, given that pest damage is easily distinguishable by the naked eye, we opted for manual identification to ensure quick and efficient processing to meet production needs. The crabapples were classified as infested or non-infested through manual identification according to the following steps. These samples were cut into thin slices of approximately 5 mm along the direction parallel to the fruit equator to display their internal condition. The interior of pest-damaged samples exhibited noticeable characteristics, such as the presence of the larva, accumulation of frass (larval droppings), wormholes, internal tissue darkening, and others [28]. Two researchers with extensive experience in post-harvest detection carefully inspected the slices to accurately determine whether each sample was infested with pests. Those samples exhibiting pest infestation were marked as -1, and other crabapples were labeled as + 1.

Orientation compensation

The spectra from all samples under the same orientation were averaged to investigate the influence of different orientations on the spectra. Subsequently, to mitigate the impact of orientation-induced spectral variations on pest infestation discrimination, prediction models were constructed for each orientation separately (local models) and for all orientations collectively (global model). Specifically, all samples were randomly divided into calibration and prediction sets. The calibration set for local models was based on spectra under the same orientation, whereas the calibration set for the global model included spectra from various orientations [15]. The results of different models on the prediction sets were used to evaluate their model performance.

Effective wavelengths extraction

Generally, spectral information can be corrected and adjusted to improve model performance through preprocessing and effective wavelength extraction methods [29]. This is because the original spectra contain noise such as baseline drift and light scattering, as well as data redundancy phenomena such as multicollinearity. In this research, the spectra were automatically smoothed during the collection process, which can improve the signal-to-noise ratio of spectra. Standard normal variate (SNV) transform was selected as the preprocessing method for scattering correction.

Considering the computational cost of the algorithm, competitive adaptive reweighted sampling (CARS), random frog (RF), and Monte Carlo - uninformative variable elimination (MC-UVE) were utilized as representative feature wavelength extraction methods. The CARS algorithm, based on the 'survival of the fittest' principle, selects wavelengths with high absolute regression coefficients as effective variables in an iterative and competitive manner [30]. The RF method, drawing from the framework of the reversible jump Markov Chain Monte Carlo (MCMC) approach, determines the importance of each wavelength by selection probability [31]. The MC-UVE technique, a modified version of UVE, eliminates variables with poor stability and selects effective wavelengths by measuring the stability of corresponding variable coefficients according to a large number of PLS models [32]. For these three methods, the maximum number of latent variables (LVs) was set to 20 and the number of Monte Carlo simulations was set to 100. The variable number for RF and MC-UVE was set to 60.

Morphological features analysis

Spectral morphological features may exhibit slight differences between healthy and pest-infested crabapples, particularly in peak/trough intensity and position, attributed to varying internal compositions. To analyze the differences

in spectral shape, separate calculations were performed for the average spectral intensity and absorbance of healthy and infested crabapples. Subsequently, raw spectral bands with prominent peak/valley locations were selected as feature wavelengths. SMFs can characterize and amplify spectral differences by calculating metrics such as SD, SR, and NSIR [18]. For each single spectrum, these metrics were calculated as follows:

$$SD = I_i - I_j (1 \le i < j \le n) \tag{1}$$

$$SR = I_i / I_j (1 \le i < j \le n) \tag{2}$$

$$NSIR = \frac{I_i - I_j}{I_i + I_j} (1 \le i < j \le n) \tag{3}$$

where I_i and I_j represent the spectral intensity at ith and jth effective wavelengths out of n variables, respectively. In this case, for each SMFs index, n spectral values can generate $\frac{n(n-1)}{2}$ new variables. When I_i is greater than I_j, SD, and NSIR are positive, and SR is larger than 1, otherwise the opposite. In this way, the spectral shape features can be characterized using SD, SR, and NSIR.

In addition to peaks and valleys, other spectral regions also contain sample-associated information due to the overlap of NIR spectra. These regions typically exhibit subtle variations in the change rate caused by the differences in peaks and troughs at both ends. To utilize this information, the first derivative (1st D) was employed to obtain the characteristic wavelengths. The bands where extremums of the rate of change occur in the average spectra are selected as supplementary information for the feature wavelengths. The calculated SMFs values using the original peak/trough intensity and supplementary data are denoted as SD*, SR*, and NSIR*.

Chemometric methods

Modeling method

PLS-LDA, an effective machine learning algorithm for classification problems based on high dimensional data, was adopted to construct discrimination models for healthy/infested crabapples. Partial least squares (PLS) regression was used to extract LVs to reduce the dimensionality of the original data. Then these components were projected to a low-dimensional space to achieve the optimal separability of different sample classes based on linear discriminant analysis (LDA) [33]. To prevent model overfitting, the maximum number of LVs was set to 20 for the original spectra and 10 for the extracted effective wavelengths and SMFs indices, depending on the number of variables. This limitation could help control the model's complexity. The optimal number of LVs was determined through 10-fold cross-validation, where the prediction accuracy plateaued, with no further improvement. In this way, the optimal LVs were determined to build the discriminant model. Before modeling, all samples were randomly divided into calibration and prediction sets at a ratio of 3:1. To address the potential variability introduced by random partitioning, the modeling process was repeated 100 times with different random partitions. The final prediction result was obtained by averaging the outcomes from these 100 iterations. This approach could mitigate the randomness of a single partition and provide a more reliable and stable estimate of the data quality of SMFs and the model's performance.

To further compare the effectiveness of information from the raw spectra and SMFs, a probabilistic model, Gaussian Naive Bayes (GNB) classifier, was applied for the classification of crabapples. The classification principle of the GNB model assumes that each feature follows a Gaussian distribution and calculates the likelihood of the observed data under each class, combining it with the prior probability to determine the posterior probability for classification. The spectral preprocessing and calibration-prediction set partitioning methods used for the GNB model were consistent with the previously described procedures. Finally, PLS-LDA and GNB models were constructed using raw spectra and SMFs to compare the quality of their spectral information.

Model evaluation

Sensitivity, specificity, error, and accuracy were used to evaluate the performance of the discrimination model. These indicators were calculated as follows:

$$Sensitivity = \frac{TP}{TP + FN} \times 100\% \tag{4}$$

$$Specificity = \frac{TN}{TN + FP} \times 100\% \tag{5}$$

$$Error = (1 - \frac{Sensitivity + Specificity}{2}) \times 100\% \tag{6}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \tag{7}$$

where TP (true positive) and TN (true negative) are the number of samples correctly predicted as positive and negative, respectively. While FP (false positive) and FN (false negative) are the number of samples incorrectly predicted as positive and negative, respectively. Sensitivity represents the prediction ability for positive samples, while specificity

represents the prediction ability for negative samples. In this research, pest-damaged crabapples were labeled as negative (-1) and other crabapples were labeled as positive (+1). Accuracy was used as the main evaluation index because the number of positive and negative samples (60:68) was relatively balanced. The model performance was free of overfitting or underfitting by confirming that no significant difference existed between calibration and prediction results.

All spectra were obtained using SepctrumStuite (Ocean Optics, USA). The data processing and modeling procedure was conducted using MATLAB (R2021b) with libPLS toolbox [34]. Figures were created using Origin 2021. One-way analysis of variance (ANOVA) was conducted using SPSS (ver.25.0).

Results and discussion

Spectral data analysis

Figure 2a depicts the average transmittance spectra of healthy/pest-infested samples. The absorption spectra are shown in Fig. 2b to present the absorption information intuitively. The absorption peak around the 980 nm region corresponds to the combined absorption of the water molecule's symmetric and asymmetric stretching vibrations, while the absorption around 740 nm and 840 nm is associated with its second combination frequency [35]. Additionally, the absorption near 650 nm is attributed to chlorophyll [36], and β-carotene generates the absorption around 455 nm [37]. It is worth noting that peaks near 590 nm and 1110 nm are caused by spectrum intensity variations of the halogen tungsten lamp rather than the absorption feature of specific substances.

While all spectra share similar shapes and trends, spectral intensity differences still exist between intact and pest-infested samples (Fig. 2a). These differences can be explained by the comprehensive changes in chemical composition and biological variability due to infestation [38]. Specifically, the higher absorbance of infested crabapples within 600–800 nm is attributed to the vibration of N-H and C-H bonds in organic nitrogen and amines present in insect excrement [39], as well as the strong absorption of dark-colored substances generated by pest metabolism and fruit deterioration. On the contrary, the moisture loss caused by the incomplete structure of the infested fruits [40, 41] leads to lower absorbance in the 800–1100 nm range, which is also reflected in the slowing absorbance trend near 740 nm and 840 nm. Another advantage of using Vis/NIR spectroscopy for identifying internal pest infestations is its ability to detect infestations in fruits without visible surface damage. As Vis/NIR light can pass through the entire sample, allowing the capture of internal information. These spectra differences can be utilized to distinguish between healthy and pest-infested samples.

Orientation compensation

Global orientation compensation

To explore the influence of measure orientation on discrimination performance, the average spectra from different orientations are presented in Fig. 3a. The spectral intensity of O2 and O4 is higher than that of O1 and O3, attributed to the difference in optical path and propagation characteristics among orientations. As shown in Fig. 3a, the length of the crabapple is approximately 1.25 times its height, resulting in longer optical paths for O1 and O3. In addition, the propagation path of O2 and O4 is through the stem-calyx, passing through more fruit core cavities [15]. Spectral intensity differences caused by detection orientations may affect prediction accuracy in practical online detection.

The global model is an effective method to compensate for the influence of orientations. The prediction accuracy of local and global models for healthy/infested crabapples is shown in Fig. 3b. It can be observed that the local models have poor generalization ability. They perform well in

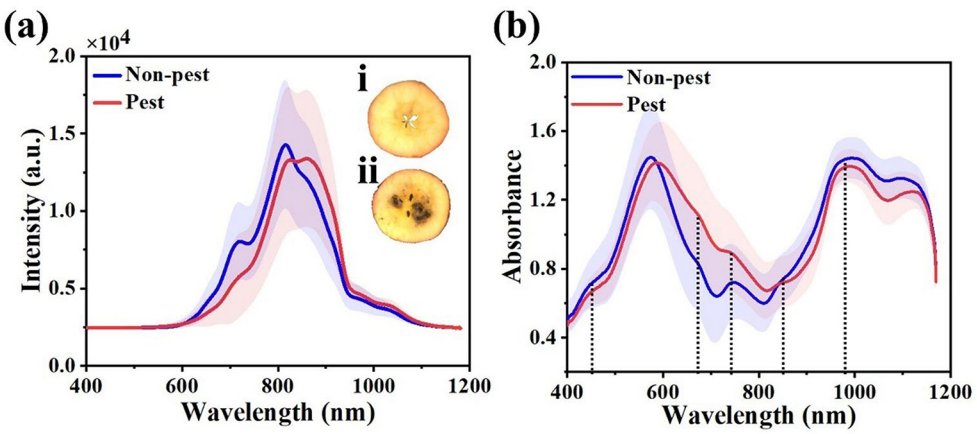

Fig. 2 Raw spectra information: **(a)** the average spectra of non-pest (i) and pest-infested (ii) crabapples, **(b)** the average absorption of two kinds of samples. The color range represents the standard deviation of all raw transmittance spectra

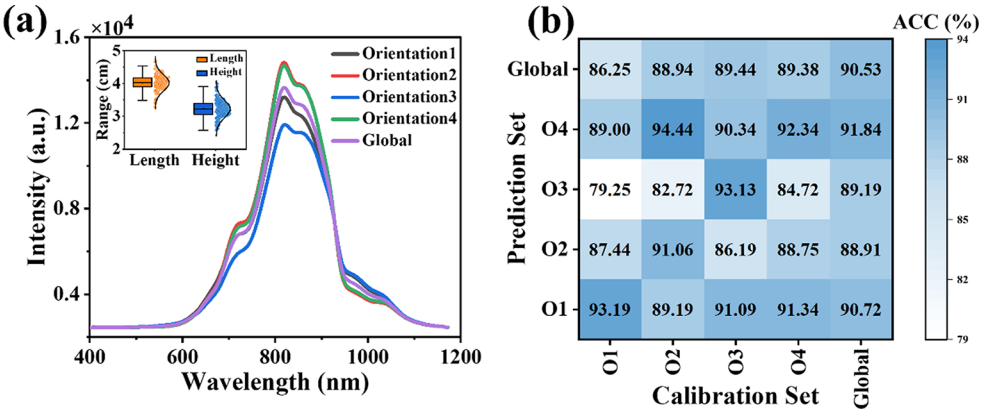

Fig. 3 The average spectra of crabapples under different orientations **(a)**, the prediction accuracy (ACC) of local and global models **(b)**

Table 1 Prediction results of the global model with different effective wavelength extraction methods

Extraction Method	Calibration Set				Prediction Set			
	Sensitivity (%)	Specificity (%)	Error (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)	Error (%)	Accuracy (%)
--	93.96	90.33	8.04	97.39	88.59	92.60	9.41	90.53[a]
CARS	**96.32**	**92.67**	**5.70**	**98.43**	**88.27**	**93.94**	**8.90**	**90.97[a]**
RF	95.13	91.77	6.74	98.00	86.40	93.22	10.19	86.69[a]
UVE	94.06	91.32	7.44	97.57	87.52	93.11	9.68	90.47[a]

Notes Prediction accuracy values with different lowercase letters indicate a significant difference ($p < 0.05$) among effective wavelength extraction methods

predicting samples under the same orientation as the calibration set (deep blue diagonal), but show slightly poorer performance in predicting samples under other orientations. By contrast, the global model demonstrates more consistent discrimination accuracy across crabapples under different orientations, because it incorporates more comprehensive spectral information from various poses. Moreover, the global model exhibits a higher prediction accuracy (90.53%) when the prediction dataset includes crabapples of all orientations, which indicates it may achieve more accurate and stable prediction results in practice. Therefore, compared to the local model, the global model can compensate for orientation differences and better meet the needs of practical sorting.

Effective wavelength extraction

Although the global model can compensate for the impact of spectral differences from various orientations, its prediction accuracy could be further improved. Here, we evaluate the effect of traditional feature wavelength extraction methods such as CARS, RF, and UVE on the global model performance (Table 1). The prediction accuracy results of CARS, RF, and UVE on the prediction dataset are 90.97%, 89.69%, and 90.47%, respectively. There is no significant difference compared to the prediction accuracy of the full-band model (90.53%). While a balance has been made between avoiding overfitting and achieving more accurate predictions for fewer feature variables, these results still tend to be overfitting. In addition, these classical effective wavelength extraction methods introduce instability and increase computation costs [42]. The embedded random sampling technique in these algorithms can cause the program to select different effective variables at different times, and the iterative process also results in a longer computation time. In this case, it is necessary to adopt other feasible methods to improve the prediction performance of the global model.

Spectral morphological features

The spectral shape feature is one of the possible solutions to improve the accuracy of infested fruit distinguishment. Each sample spectrum contains 1044 wavelengths, among which peaks and valleys are often taken as morphological feature points due to their potential association with characteristic absorption [18]. To highlight differences in healthy/infested crabapple spectra and use as few data points as possible, SMF values were calculated with eight selected prominent peaks/valleys (Figs. 4a and 581.50, 718.69, 739.18, 815.33, 840.03, 860.92, 956.27, and 1004.45 nm). Among them, the range around 718.69 nm is related to the lesions of the crabapple core [43]. Wavelengths at 739.18, 840.03, and 956.27 nm correspond to the absorption feature of water. The combined vibration and rotation frequency of C-H causes the absorption at 815 nm, while the spectrum near 1004.45 nm is attributed to the vibration of O-H [35].

The combined calculation of these bands yields a total of 84 SMFs variables (SD + SR + NSIR). The PCA results

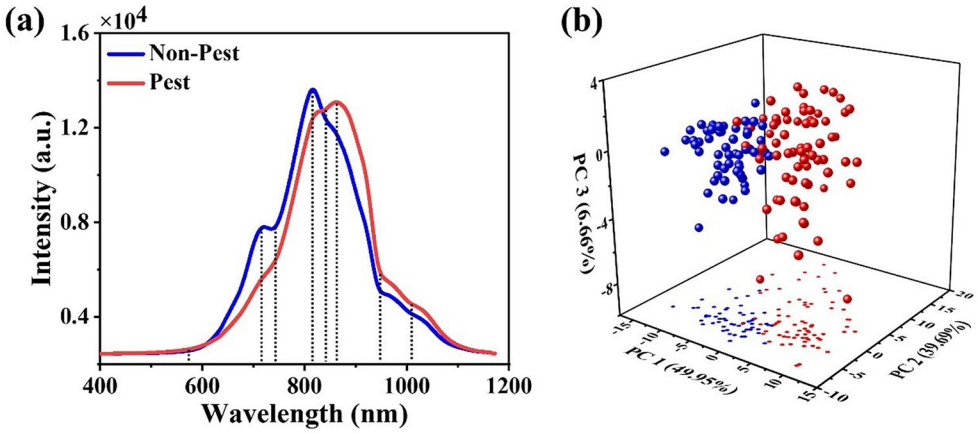

Fig. 4 The wavelengths of spectral morphological features **(a)** and clustering result of principal component analysis (PCA) based on SD + SR + NSIR **(b)**

Table 2 Prediction results of the global model with different spectral morphological feature strategies

Extraction Method	Calibration Set				Prediction Set			
	Sensitivity (%)	Specificity (%)	Error (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)	Error (%)	Accuracy (%)
--	93.96	90.33	8.04	97.39	88.59	92.60	9.41	90.53[a]
SD	96.45	94.41	4.70	98.80	90.52	94.36	7.56	92.53[b]
SR	98.05	92.31	5.18	99.06	87.24	97.40	7.68	92.31[b]
NSIR	98.41	91.93	5.22	99.19	86.89	97.50	7.80	92.31[b]
SD + SR	96.51	95.10	4.28	98.79	91.66	94.41	6.97	93.03[b]
SD + NSIR	97.11	95.27	3.92	98.97	91.27	94.67	7.03	92.91[b]
SR + NSIR	97.69	92.37	5.27	99.14	88.23	96.51	7.63	92.38[b]
SD + SR + NSIR	**96.97**	**95.84**	**3.66**	**98.99**	**91.35**	**95.04**	**6.81**	**93.25**[b]

Notes Prediction accuracy values with different lowercase letters indicate a significant difference ($p < 0.05$) among SMFs indices

based on SMFs values are present in Fig. 4b, revealing that the first three principal components account for a cumulative contribution rate of 96.3%. Although most of the healthy/pest-infested samples could be effectively distinguished, there are still some notable overlaps that cannot be correctly classified. This indicates that the SMFs matrix could be used to identify internal infestation in crabapples, but more precise chemometric methods need to be adopted to ensure discrimination accuracy.

Table 2 displays the prediction results for infested crabapples using the PLS-LDA modeling method based on spectral morphological features and their combinations. As expected, the introduction of SMFs leads to a significant improvement in prediction accuracy ($p < 0.05$). Among them, the combination of three SMFs achieves the optimal prediction results. For the calibration set, the average accuracy is 98.99%, and the average sensitivity and specificity are 96.97% and 95.84%, respectively. For the prediction set, the average accuracy increases from 90.53% for the full-spectrum model to 93.25% for the SMFs combination, and the average prediction error rate decreases from 9.41% to 6.81%. These results provide evidence that, for discriminating healthy/infested crabapples, the SMFs are superior to traditional full-spectrum models and classical effective wavelength extraction methods (i.e., CARS). It is also worth noting that although using 8 representative wavelengths can calculate spectral morphological features and improve prediction accuracy, the spectral data may still not be fully utilized with the 1044 original data points. Further exploration is needed on how to completely utilize the spectral data and improve prediction accuracy on this basis.

Improvement of SMFs

Due to the overlapping nature of the NIR spectrum [24], other spectral bands also contain information about the internal quality of fruits besides the peaks or valleys of the spectra. The spectra outside the peaks/valleys exhibit larger slopes, so we choose the original spectral points corresponding to extremum points of the 1st derivative as supplementary information for primary SMFs. The supplementary wavelengths are 640.02, 689.75, 727.8, 789.03, 907.67, 930.55, 982.6, and 1022.6 nm. As shown in Fig. 5a, the selected representative wavelengths cover almost the entire effective range of 550–1100 nm after the addition of 8 new bands. The combination of initial and supplementary wavelengths produces 360 SMF values (SD*+SR*+NSIR*), which ensures the representativeness and integrity of spectral morphological features.

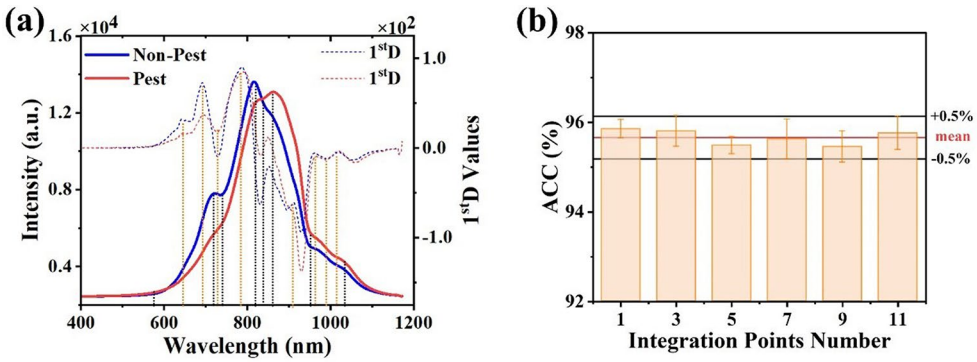

Fig. 5 Supplementary information on the spectral morphological feature **(a)** and the prediction accuracy based on different integration point numbers **(b)**

Table 3 Prediction results of global model of the original and supplementary spectral morphological features

Extraction Method	Calibration Set				Prediction Set			
	Sensitivity (%)	Specificity (%)	Error (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)	Error (%)	Accuracy (%)
--	93.96	90.33	8.04	97.39	88.59	92.60	9.41	90.53[a]
SD + SR + NSIR	96.97	95.84	3.66	98.99	91.35	95.04	6.81	93.25[b]
SD*+SR*+NSIR*	**97.02**	**98.11**	**2.36**	**99.46**	**95.14**	**96.32**	**4.27**	**95.94[c]**

Notes Prediction accuracy values with different lowercase letters indicate a significant difference ($p < 0.05$) among SMFs indices

The prediction results using the PLS-LDA model based on initial and supplementary SMF values are shown in Table 3. The addition of feature wavelengths on primary SMFs has led to a significant improvement in prediction accuracy ($p < 0.05$). In the calibration set, the sensitivity and specificity values are 97.02% and 98.11%, respectively. And the average accuracy is 99.46%. In the prediction set, the average accuracy increases to 95.94%, and the values of sensitivity and specificity are 95.14% and 96.32%, respectively. The prediction error is further reduced from 6.81% of the original SMFs to 4.27%. This improvement in prediction performance can be attributed to the supplementation and enhancement of spectral morphological information. The addition of spectral data at slope extremum points compensates for the deficiencies of initial SMFs, which only include spectral peaks/troughs. As a result, these 16 wavelengths could cover almost all effective wavelength bands, and the prediction accuracy based on the new SMFs matrix has been further improved. However, despite the high discrimination accuracy (95.94%), it has not reached 100%. The detailed reasons will be discussed in later Sect. 3.5 Misclassification analysis.

In the spectral acquisition process, the spectral shifts of different samples may affect spectral morphological features. To validate the representativeness and stability of the selected wavelengths, the average of the integrated spectral region is utilized to compare the prediction performance of different spectral bandwidths. The results of prediction accuracy for different numbers of spectral wavelengths are shown in Fig. 5b. There is no significant change in discrimination accuracy as the number of spectral points increases. As depicted in Fig. 5b, the average prediction accuracy fluctuates within a narrow range of ±0.5%. This indicates the representativeness of the selected SMF wavelengths and the stability of the prediction results based on these features. This stability can be attributed to the fact that SMFs are based on the overall trend of the spectrum, which is less susceptible to external interferences. The differences in spectral shape are mainly influenced by internal infestation, and even with slight shifts, the overall spectral morphology does not change drastically. Thus, the SMFs can be used to distinguish healthy/infested crabapples and also exhibit certain stability.

To further validate the effectiveness of SMFs in improving spectral data quality, a probabilistic model, GNB, was employed to detect pest-infested crabapples. Table 4 presents the discrimination results of the GNB model and its performance comparison with the non-probabilistic PLS-LDA model. When using the same spectral data (RawS, SMFs), the GNB model performs slightly worse than the PLS-LDA model, underscoring the superior discriminatory ability of PLS-LDA in identifying pest-infested crabapples. Additionally, the GNB model based on SMFs achieved an average discriminant accuracy of 93.84%, exceeding the accuracy of 84.15% obtained with the raw spectra. Under both probabilistic and non-probabilistic machine learning models, the classification models based on SMFs outperform those using raw spectra, confirming the enhancement in spectral quality due to SMFs.

Misclassification analysis

As shown in Table 3, the average prediction accuracy for healthy/infested crabapples based on SMFs can reach

Table 4 Comparison of classification performance between PLS-LDA and GNB models

Modeling method	Calibration Set				Prediction Set			
	Sensitivity (%)	Specificity (%)	Error (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)	Error (%)	Accuracy (%)
RawS[1] + PLS-LDA	93.96	90.33	8.04	97.39	88.59	92.60	9.41	90.53
SMFs[2] + PLS-LDA	**97.02**	**98.11**	**2.36**	**99.46**	**95.14**	**96.32**	**4.27**	**95.94**
RawS + GNB	81.89	91.63	13.23	86.86	78.50	90.27	15.61	84.15
SMFs + GNB	92.23	96.15	5.81	94.38	91.50	95.72	6.39	93.84

[1] RawS: Raw spectral data, [2] SMFs: SD*+SR*+NSIR*

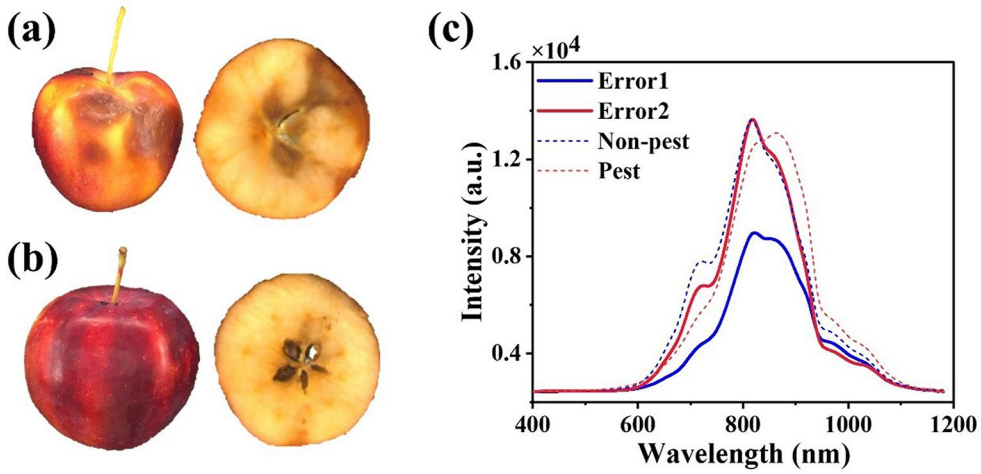

Fig. 6 Samples with low classification accuracy: **(a)** bruised crabapple, **(b)** pest-infested sample without obvious lesions, and **(c)** the corresponding spectra

95.94%, while there is still a small proportion of samples being misclassified. These misclassified samples can be categorized into two types: crabapples with severe bruises but no infestation (Fig. 6a), and crabapples with mild infestation but no obvious lesions (Fig. 6b). For the first misclassification, the bruise produces internal lesions with characteristics similar to infestation, such as tissue browning. Thus, its spectral shape resembles that of pest-infested crabapples (Fig. 6c, Error1), making it easily misclassified as the infested sample. Typically, online detection equipment includes external and internal quality inspection modules, with the former capable of identifying objects with surface bruising. On the contrary, for the second misclassification, only the seeds are damaged since it is in the early stage of infestation. There are no obvious accumulations of pest metabolites or tissue lesions inside the fruit at this stage, which presents a significant challenge for discrimination. Thus, it tends to be identified as intact crabapple (Fig. 6c, Error2) [18]. For such mild infestations, postharvest treatments such as phytosanitary irradiation can be employed in the post-sorting process to kill pests and ensure the quality of crabapples in trade [44].

To address misclassification, periodic assessment could also be an effective measure because pest-infestation or non-pest rotting may vary over time. Considering the short storage duration of crabapples, such periodic assessments were not conducted to meet the needs of actual production. However, for fruits with longer storage durations, such as apples and citrus, periodic assessment might be meaningful.

Discussion about SMFs

To date, no previous studies have explored the simultaneous mitigation of sample orientation impact and amplification of internal spectral features. For external orientations, Huang et al. [45] investigated the optimal orientation and spectral acquisition methods for detecting defects in apples. Tian et al. [15] analyzed the spectral propagation characteristics under different orientations and used a global model to compensate for the spectral differences from various orientations. These studies have improved the accuracy of defect detection in fruits, but they focus solely on the external orientation effects, without considering the internal spectral feature differences. For morphological features, Liu et al. [18] combined characteristic wavelengths (CWs) and SMFs to classify apples with moldy cores, achieving an accuracy of 97.3%. However, since only five SMF wavelengths (651, 687, 715, 773, and 808 nm) were used, the classification accuracy using SMFs alone was only 91.9%, likely due to the poor representativeness of these five points for the overall spectral information. Similarly, Zhang et al. [46] achieved an accuracy of 98.48% in identifying watercore apples using only two SMF data points, but this result

was based solely on the O3 orientation. The accuracy was only 89.39% in another O2 orientation, failing to overcome the impact of orientation. In this study, the classification accuracy for pest-infested apples was improved to 95.94% through enhancing the quality of modeling data, by eliminating the influence of external orientation and amplifying differences in spectral morphological features. Additionally, the SMFs used in this study were supplemented with first-order derivatives, making the data more comprehensive and representative. As a result, our study enhanced the quality of spectral data from both external and internal dimensions and improved the representativeness of SMFs.

According to the previous results, compared to classical variable selection methods, spectral morphological features can significantly enhance the discrimination accuracy of pest-infested crabapples. Detection of pest-infested fruits based on SMFs may present potential advantages such as better interpretability and relative stability [42]. Specifically, the differences in the content of pest metabolites, water, pigments, and browning, along with the corresponding vibration of functional groups such as C-H, O-H, and N-H at the molecular level [7], result in observable distinctions in spectral shape. These differences have been thoroughly analyzed in Sect. 3.1. In addition, no preprocessing methods are used for SMFs in this study, and the number of variables required has been reduced from the full spectrum of 1044 to 360, which can lower the model complexity and computation cost. However, some challenges of the SMF method, such as generalization ability, are still worth attention. The biological variability of fruits, such as cultivar, geographical origin, and season, may affect their composition and concentration, thereby influencing the spectral shape [8]. Therefore, how to balance the accuracy and generalizability of the prediction model based on presentative SMF wavelengths is a challenge to be addressed.

Conclusion

This study aims to enhance the prediction accuracy of pest-infested crabapples by mitigating the influence of external orientations and amplifying the spectral differences caused by internal biological properties. The spectral intensities of crabapples vary with orientations, while the spectral shapes of healthy and infested samples exhibit distinctions. Based on this, the global model, which is more stable than the local model, is employed to mitigate the external influence of crabapple orientation. Subsequently, compared to classical feature extraction methods, SMFs can characterize and amplify the spectral differences between effective wavelengths for pest infestation prediction. Finally, with the supplementation of spectral information using 1st D, SMFs improve the average prediction sensitivity, specificity, and accuracy from 88.59%, 92.60%, and 90.53% of full-spectra to 95.14%, 96.32%, and 95.94%, respectively. Additionally, SMFs exhibit more intuitive interpretability and greater stability, as the effective bands are directly derived from spectral shape differences.

In summary, SMFs combined with the global model can mitigate the impacts of external orientations and enhance spectral differences between healthy and infested crabapples, ultimately improving discrimination performance. Given that this study relies on laboratory spectral acquisition equipment, despite its structural similarity to that of production lines, robust model transfer methods are necessary for practical use in sorting lines. For future research, it is important to investigate the influence of sample colors (primarily varying shades of yellow and red for crabapples) on spectral morphology to further improve prediction accuracy. We anticipate the application of these findings in practical production settings.

Acknowledgements The authors gratefully acknowledge the financial support provided by the National Key R&D Program of China (2023YFD2201301).

Author contributions Yuanhao Zheng: Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing - original draft, Writing - review & editing. Ying Zhou: Conceptualization, Investigation, Methodology, Writing - review & editing. Penghui Liu: Conceptualization, Methodology, Software, Visualization, Writing - review & editing. Yingjie Zheng: Visualization, Writing - review & editing. Zichao Wei: Data curation, Investigation. Zetong Li: Data curation, Investigation. Lijuan Xie: Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing - review & editing.

Data availability Data will be made available on request.

Declarations

Competing interests 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.

References

1. F. Mas, L.-A. Manning, M. Alavi, T. Osborne, O. Reynolds, A. Kralicek, Early detection of fruit infested with Bactrocera tryoni. Postharvest Biol. Technol. **175**, 111496 (2021). https://doi.org/10.1016/j.postharvbio.2021.111496
2. S.C. Welter, Chapter 50, Codling Moth. In: V. H. Resh & R. T. Cardé (Eds.), Encyclopedia of Insects (Second Edition)Academic Press,. (2009). pp. 174–175. https://doi.org/10.1016/B978-0-12-374144-8.00059-X
3. B. Veltman, D. Harpaz, A. Sadeh, E. Eltzov, Whole-cell bacterial biosensor applied to identify the presence of *Thaumatotibia leucotreta* larva in citrus fruits by volatile sensing. Food Control. **160**, 110388 (2024). https://doi.org/10.1016/j.foodcont.2024.110388

4. M. Mei, J. Li, An overview on optical non-destructive detection of bruises in fruit: technology, method, application, challenge and trend. Comput. Electron. Agric. **213**, 108195 (2023). https://doi.org/10.1016/j.compag.2023.108195
5. N. Ekramirad, A.Y. Khaled, C.A. Parrish, K.D. Donohue, R.T. Villanueva, A.A. Adedeji, Development of pattern recognition and classification models for the detection of vibro-acoustic emissions from codling moth infested apples. Postharvest Biol. Technol. **181**, 111633 (2021). https://doi.org/10.1016/j.postharvbio.2021.111633
6. V. Cortés, J. Blasco, N. Aleixos, S. Cubero, P. Talens, Monitoring strategies for quality control of agricultural products using visible and near-infrared spectroscopy: a review. Trends Food Sci. Tech. **85**, 138–148 (2019). https://doi.org/10.1016/j.tifs.2019.01.015
7. X. Li, L. Zhang, Y. Zhang, D. Wang, X. Wang, L. Yu, W. Zhang, P. Li, Review of NIR spectroscopy methods for nondestructive quality analysis of oilseeds and edible oils. Trends Food Sci. Tech. **101**, 172–181 (2020). https://doi.org/10.1016/j.tifs.2020.05.002
8. B. Zhang, D. Dai, J. Huang, J. Zhou, Q. Gui, F. Dai, Influence of physical and biological variability and solution methods in fruit and vegetable quality nondestructive inspection by using imaging and near-infrared spectroscopy techniques: a review. Crit. Rev. Food Sci. Nutr. **58**(12), 2099–2118 (2018). https://doi.org/10.1080/10408398.2017.1300789
9. W. Long, Z. Hu, L. Wei, H. Chen, T. Liu, S. Wang, Y. Guan, X. Yang, J. Yang, H. Fu, Accurate identification of the geographical origins of lily using near-infrared spectroscopy combined with carbon dot-tetramethoxyporphyrin nanocomposite and chemometrics. Spectrochim Acta Mol. Biomol. Spectrosc. **271**, 120932 (2022). https://doi.org/10.1016/j.saa.2022.120932
10. G. Fan, J. Zha, R. Du, L. Gao, Determination of soluble solids and firmness of apples by Vis/NIR transmittance. J. Food Eng. **93**(4), 416–420 (2009). https://doi.org/10.1016/j.jfoodeng.2009.02.006
11. C.J. Clark, V.A. McGlone, R.B. Jordan, Detection of Brownheart in 'Braeburn' apple by transmission NIR spectroscopy. Postharvest Biol. Technol. **28**(1), 87–96 (2003). https://doi.org/10.1016/S0925-5214(02)00122-9
12. J. Sun, R. Künnemeyer, A. McGlone, N. Tomer, Investigations of optical geometry and sample positioning in NIRS transmittance for detecting vascular browning in apples. Comput. Electron. Agric. **155**, 32–40 (2018). https://doi.org/10.1016/j.compag.2018.09.041
13. Y. Hao, Q. Wang, S. Zhang, Online accurate detection of soluble solids content in navel orange assisted by automatic orientation correction device. INFRARED PHYS. TECHNOL. **118**, 103871 (2021). https://doi.org/10.1016/j.infrared.2021.103871
14. S. Fan, B. Zhang, J. Li, W. Huang, C. Wang, Effect of spectrum measurement position variation on the robustness of NIR spectroscopy models for soluble solids content of apple. Biosyst Eng. **143**, 9–19 (2016). https://doi.org/10.1016/j.biosystemseng.2015.12.012
15. S. Tian, M. Zhang, B. Li, Z. Zhang, J. Zhao, Z. Zhang, H. Zhang, J. Hu, Measurement orientation compensation and comparison of transmission spectroscopy for online detection of moldy apple core. INFRARED PHYS. TECHNOL. **111**, 103510 (2020). https://doi.org/10.1016/j.infrared.2020.103510
16. J. Wang, K. Nakano, S. Ohashi, K. Takizawa, J.G. He, Comparison of different modes of visible and near-infrared spectroscopy for detecting internal insect infestation in jujubes. J. Food Eng. **101**(1), 78–84 (2010). https://doi.org/10.1016/j.jfoodeng.2010.06.011
17. A. Rady, N. Ekramirad, A.A. Adedeji, M. Li, R. Alimardani, Hyperspectral imaging for detection of codling moth infestation in GoldRush apples. Postharvest Biol. Technol. **129**, 37–44 (2017). https://doi.org/10.1016/j.postharvbio.2017.03.007
18. H. Liu, Z. Wei, M. Lu, P. Gao, J. Li, J. Zhao, J. Hu, A Vis/NIR device for detecting moldy apple cores using spectral shape features. Comput. Electron. Agric. **220**, 108898 (2024). https://doi.org/10.1016/j.compag.2024.108898
19. L. Li, Y. Peng, C. Yang, Y. Li, Optical sensing system for detection of the internal and external quality attributes of apples. Postharvest Biol. Technol. **162**, 111101 (2020). https://doi.org/10.1016/j.postharvbio.2019.111101
20. M. Zhang, M. Shen, Y. Pu, H. Li, B. Zhang, Z. Zhang, X. Ren, J. Zhao, Rapid Identification of Apple Maturity based on Multispectral Sensor combined with spectral shape features. Horticulturae. **8**(5)), Article 5. ((2022). https://doi.org/10.3390/horticulturae8050361
21. J. Ma, D.-W. Sun, H. Pu, Spectral absorption index in hyperspectral image analysis for predicting moisture contents in pork *longissimus dorsi* muscles. Food Chem. **197**, 848–854 (2016). https://doi.org/10.1016/j.foodchem.2015.11.023
22. R. Moscetti, R.P. Haff, B. Aernouts, W. Saeys, D. Monarca, M. Cecchini, R. Massantini, Feasibility of Vis/NIR spectroscopy for detection of flaws in hazelnut kernels. J. Food Eng. **118**(1), 1–7 (2013). https://doi.org/10.1016/j.jfoodeng.2013.03.037
23. L. Lleó, J.M. Roger, A. Herrero-Langreo, B. Diezma-Iglesias, P. Barreiro, Comparison of multispectral indexes extracted from hyperspectral images for the assessment of fruit ripening. J. Food Eng. **104**(4), 612–620 (2011). https://doi.org/10.1016/j.jfoodeng.2011.01.028
24. T. Leng, F. Li, L. Xiong, Q. Xiong, M. Zhu, Y. Chen, Quantitative detection of binary and ternary adulteration of minced beef meat with pork and duck meat by NIR combined with chemometrics. Food Control. **113**, 107203 (2020). https://doi.org/10.1016/j.foodcont.2020.107203
25. X. Zhai, X. Wang, X. Wang, H. Zhang, Y. Ji, D. Ren, J. Lu, An efficient method using ultrasound to accelerate aging in crabapple (Malus Asiatica) vinegar produced from fresh fruit and its influencing mechanism investigation. Ultrason. Sonochem. **72**, 105464 (2021). https://doi.org/10.1016/j.ultsonch.2021.105464
26. Z. Zhang, H. Liu, D. Chen, J. Zhang, H. Li, M. Shen, Y. Pu, Z. Zhang, J. Zhao, J. Hu, SMOTE-based method for balanced spectral nondestructive detection of moldy apple core. Food Control. **141**, 109100 (2022). https://doi.org/10.1016/j.foodcont.2022.109100
27. Y. Zheng, Y. Cao, L. Xie, Design of a multi-function experimental system for online internal quality evaluation of fruits. Food Measure. **18**(1), 26–39 (2024). https://doi.org/10.1007/s11694-023-02143-9
28. M. Kadoić Balaško, R. Bažok, K.M. Mikac, D. Lemic, & Pajač Živković, I. Pest Management Challenges and Control Practices in Codling Moth: A Review. *Insects*, *11*(1), Article 1. (2020) https://doi.org/10.3390/insects11010038
29. Y. Zheng, S. Tian, L. Xie, Improving the identification accuracy of sugar orange suffering from granulation through diameter correction and stepwise variable selection. Postharvest Biol. Technol. **200**, 112313 (2023). https://doi.org/10.1016/j.postharvbio.2023.112313
30. H. Li, Y. Liang, Q. Xu, D. Cao, Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration. Anal. Chim. Acta. **648**(1), 77–84 (2009). https://doi.org/10.1016/j.aca.2009.06.046
31. H.-D. Li, Q.-S. Xu, Y.-Z. Liang, Random frog: an efficient reversible jump Markov Chain Monte Carlo-like approach for variable selection with applications to gene selection and disease classification. Anal. Chim. Acta. **740**, 20–26 (2012). https://doi.org/10.1016/j.aca.2012.06.031
32. W. Cai, Y. Li, X. Shao, A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra. Chemometr Intell. Lab. Syst. **90**(2), 188–194 (2008). https://doi.org/10.1016/j.chemolab.2007.10.001

33. S. Li, X. Zhang, Y. Shan, D. Su, Q. Ma, R. Wen, J. Li, Qualitative and quantitative detection of honey adulterated with high-fructose corn syrup and maltose syrup by using near-infrared spectroscopy. Food Chem. **218**, 231–236 (2017). https://doi.org/10.1016/j.foodchem.2016.08.105
34. H.-D. Li, Q.-S. Xu, Y.-Z. Liang, libPLS, An integrated library for partial least squares regression and linear discriminant analysis. Chemometr Intell. Lab. Syst. **176**, 34–43 (2018). https://doi.org/10.1016/j.chemolab.2018.03.003
35. J. Workman, L. Weyer, Spectra-structure correlations for Near-Infrared. *Practical Guide to Interpretive Near-Infrared Spectroscopy*. (CRC, 2007), 57–58, 219–220. https://doi.org/10.1201/9781420018318
36. E. Bertone, A. Venturello, R. Leardi, F. Geobaldo, Prediction of the optimum harvest time of 'Scarlet' apples using DR-UV–Vis and NIR spectroscopy. Postharvest Biol. Technol. **69**, 15–23 (2012). https://doi.org/10.1016/j.postharvbio.2012.02.009
37. S.B. Lohan, K. Vitt, P. Scholz, C.M. Keck, M.C. Meinke, ROS production and glutathione response in keratinocytes after application of β-carotene and VIS/NIR irradiation. Chem. Biol. Interact. **280**, 1–7 (2018). https://doi.org/10.1016/j.cbi.2017.12.002
38. R. Moscetti, R.P. Haff, S. Saranwong, D. Monarca, M. Cecchini, R. Massantini, Nondestructive detection of insect infested chestnuts based on NIR spectroscopy. Postharvest Biol. Technol. **87**, 88–94 (2014). https://doi.org/10.1016/j.postharvbio.2013.08.010
39. V.B. Wigglesworth Chapter 5, Excretion. In: Insect Physiology. (Springer, M.A. Boston, 1974), pp. 62–77. https://link.springer.com/book/10.1007/978-1-4899-3202-0
40. Omkar (ed.), Chapter 14, pests of Apple. *Pests and Their Management* (Springer, Singapore, 2018), 480–484. https://doi.org/10.1007/978-981-10-8687-8
41. S. Sarker, Y.H. Woo, U.T. Lim, Developmental stages of peach, plum, and apple fruit influence development and fecundity of Grapholita molesta (Lepidoptera: Tortricidae). Sci. Rep. **11**(1), 2105 (2021). https://doi.org/10.1038/s41598-021-81651-4
42. Y.-H. Yun, H.-D. Li, B.-C. Deng, D.-S. Cao, An overview of variable selection methods in multivariate analysis of near-infrared spectra. Trends Anal. Chem. **113**, 102–115 (2019). https://doi.org/10.1016/j.trac.2019.01.018
43. S. Tian, J. Zhang, Z. Zhang, J. Zhao, Z. Zhang, H. Zhang, Effective modification through transmission Vis/NIR spectra affected by fruit size to improve the prediction of moldy apple core. Infrared Phys. Technol. **100**, 117–124 (2019). https://doi.org/10.1016/j.infrared.2019.05.015
44. J.B. Golding, A. Uthairatanakij, J. de Ornelas-Paz J., A. Prakash, Phytosanitary irradiation effects on fresh produce quality – A review. Postharvest Biol. Technol. **211**, 112855 (2024). https://doi.org/10.1016/j.postharvbio.2024.112855
45. Y. Huang, R. Lu, K. Chen, Detection of internal defect of apples by a multichannel Vis/NIR spectroscopic system. Postharvest Biol. Technol. **161**, 111065 (2020). https://doi.org/10.1016/j.postharvbio.2019.111065
46. Y. Zhang, X. Yang, Z. Cai, S. Fan, H. Zhang, J. Li, Online detection of watercore apples by Vis/NIR full-transmittance spectroscopy coupled with ANOVA method. Foods. **10**(12), 2983 (2021). https://doi.org/10.3390/foods10122983