Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 13, No. 2, 2025

Research on Apple Internal Quality Classification Based on Near-Infrared Spectroscopy

Zhipeng Li *

School of Electronic Information, Xijing University, Xi 'an, Shaanxi, China
*** Corresponding author Email:** 3473326046@qq.com

Abstract: In this study, Luochuan Red Fuji apple in Shaanxi Province was taken as the experimental object, and the sugar content and spectral data were measured and averaged at three locations at the upper distance of the equator. SPXY method was used to divide the data set, MAS, SNV, MC three data preprocessing methods were used, and CARS were used to select the characteristic wavelength, and three classification models SVC, DT and KNN were established. The results show that SPXY + MC + CARS + DT model has the best classification results, and the accuracy rate, accuracy rate, recall rate and F1 score reach 0.955 respectively. In summary, the use of near-infrared spectroscopy technology can be used in Apple's internal quality classification, which improves the basis and reference for the application of Apple's non-destructive testing technology.

Keywords: Apple; Near Infrared Spectroscopy; Nondestructive Testing; Internal Quality; Classification Model.

1. Introduction

With the improvement of people's living standards and the development and progress of society, more and more consumers are constantly raising their requirements for the quality of food. Apples are a kind of fruit rich in various vitamins and minerals. They are highly nutritious and beneficial to human health, and are deeply loved by consumers. All along, the external quality of apples has always been the focus of people's attention when purchasing, but the internal quality is often the decisive factor determining the taste and nutritional value of apples. Throughout the entire growth process of the fruit, the soluble solids in apples are the key to determining the internal quality and play a crucial role in the market value of apples. Soluble solids are one of the main factors reflecting the internal quality and maturity of apples, among which sugar content is the most crucial indicator of soluble solids. Therefore, determining the sugar content of apples and classifying them based on their internal quality is of great significance for the post-harvest grading processing of apples and improving commercial benefits.

2. Experimental Materials and Methods

2.1. Materials and Instruments

A total of 112 red Fuji apples from Luochuan, Shaanxi Province, with uniform color, consistent size and no external damage, were selected as experimental samples. All apple samples completed spectral data collection and sugar content determination at the same experimental site on the same day.

The near-infrared spectral data of apple samples were collected using a near-infrared spectrometer (model ATP8600, wavelength range 997-1708nm, Optiancheng Company), as shown in Figure 1. The sugar content values of apple samples were collected using a handheld refractometer (ATAGO PAL-BX|ACID8, with a detection range of soluble sugar from 0.0 to 90%, an accuracy of ±0.2%, and a suitable temperature range of 9.0 to 99.9℃), as shown in Figure 2.

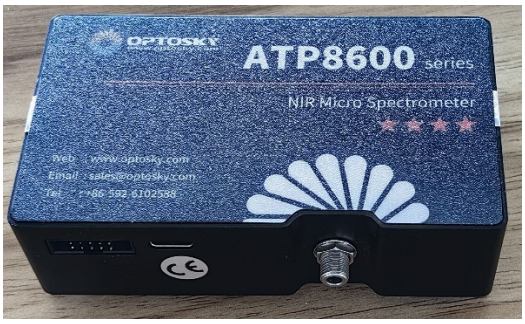


Fig 1. Near-infrared spectrometer

Fig 2. Handheld refractometer

2.2. Experimental Method

2.2.1. Experimental Data Collection

(1) Sample marking

At the circumequatorial position of the apple, select a collection point at equal intervals of 60 degrees and mark it. A total of three points is marked as the data collection positions.

(2) Collect spectral data

Set up the spectral equipment; Adjust the optical path and use the calibration whiteboard to collect the reference

spectrum; Start collecting the spectral data of the experimental sample. Collect the data three times for each apple sample and calculate the average as the original spectral data of the sample, and save it in an excel table; After the spectral data collection is completed, export the spectral data.

(3) Collect sugar content data

Before measurement, calibrate the refractometer with a sucrose solution. When collecting sugar content data, peel the apples, take small samples at the marked positions with a clean fruit knife, extract the juice, filter it, and titrate it into the measurement area of the refractometer. Take the average value of the three measurement positions of the apples as the sugar content value of the sample.

2.2.2. Spectral Data Preprocessing

To suppress the influence of noise on spectral data and improve the accuracy of the classification model. The experiment selected three data preprocessing methods, namely Moving average smoothing (MAS), Standard normal variate (SNV), and Mean centering (MC), to preprocess the original spectral data respectively.

(1) Moving average smoothing

The Moving average smoothing algorithm (MAS) works by using a smoothing window of a specific window width [1], where the wavelength points within a single window are all odd. The average of the measurement values at the center wavelength point k and the X points before and after in the window is used to replace the measurement values at the wavelength points. The algorithm is moved from left to right by k to smooth all the points.

(2) Standard normal variate

The main purpose of the Standard normal variate (SNV) is to eliminate the influence of the size of solid particles and the changes in the optical path of surface scattering of the sample on the near-infrared diffuse reflection spectrum [2]. The main principle of SNV is to eliminate the background noise of the sample by changing the scale range and intensity of the spectral signal.

(3) Mean centering

Mean centering (MC) is centered on calculating the average spectrum of the entire dataset and subtracting the average spectrum from each sample, thereby eliminating the overall offset and bias caused by the data [3]. This process adjusts the mean of the data to zero, causing the spectral data to fluctuate around the zero point. Mean centralization can simplify the data structure, enhance the interpretability of the model, and improve the performance of the analysis method simultaneously.

2.2.3. Dataset Partitioning

Before establishing the model, the Sample Set Partitioning based on Joint X-Y Distance (SPXY) dataset division method was adopted, and the sample set was divided into 89 training sets and 23 prediction sets at a ratio of 4:1. The divided training set is used to fit the data to establish a classification model. The prediction set does not participate in the model training and is used to evaluate the actual effect of the model established by the training set. The SPXY algorithm is a sample partitioning method based on statistical principles [4]. Due to its efficient coverage ability in the multi-dimensional vector space, it can significantly improve the prediction accuracy of the established model.

2.2.4. Characteristic Wavelength Selection

The Competitive Adaptive Reweighted Sampling Algorithm (CARS) is a feature selection method based on Monte Carlo sampling and partial least squares regression, which is used to screen out the feature bands that contribute the most to the model prediction from the high-order spectral data [5]. The core principle is to dynamically adjust the selection probability of each band by using the exponential attenuation function, and gradually propose the unimportant bands based on the absolute value weights of the regression coefficients.

2.2.5. Establishment of Classification Model

(1) Support vector classification

Support Vector Classification (SVC) is a supervised learning classification algorithm. Its core is to maximize the intervals between data points of different categories through an optimal hyperplane, and at the same time use support vectors (the points closest to the hyperplane) to determine the position and direction of the hyperplane, separating data points of different categories as much as possible.

(2) Decision tree

Decision Tree (DT) is a classification algorithm based on tree structure. It divides the data set into smaller subsets by recursively selecting the optimal features to construct the model. The partitioning effect of features is measured by indicators such as information gain. The features that maximize the improvement of data purity are selected for node splitting until the stopping condition is met.

(3) K-nearest neighbor algorithm

The K-nearest Neighbor Algorithm (KNN) is a simple and intuitive classification method. Its core idea is to calculate the distance between the sample to be classified and the samples of known categories, find the K samples with the closest distance, and then determine the category of the sample to be classified through voting or weighted voting based on the category information of such samples.

2.3. Evaluation Index

In the experiment, the training results of the three classification models were evaluated through four evaluation indicators: accuracy rate, precision rate, recall rate and F1 score.

2.3.1. Accuracy

Accuracy is an important metric to measure the performance of a model. It describes how much of the model's predictions agree with the true results. That is, the proportion of the total number of samples that the model correctly classified. Accuracy is intuitive and easy to interpret and understand. Calculate according to Equation (1)

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

TP (True Positive) is the number of samples predicted as positive by the model.

TN (True Negative) is the number of examples that the model predicts to be negative.

FP (False Positive) represents the number of samples that the model incorrectly predicted as positive.

FN (False Negative) represents the number of samples that the model incorrectly predicted as negative class.

2.3.2. Precision

Precision, also known as precision, is the fraction of examples that the model predicts to be positive that are also positive. Calculate according to Equation (2)

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

2.3.3. Recall

Recall, also known as recall, is the fraction of examples that are actually positive that are correctly predicted to be positive by the model. Calculate according to Equation (3)

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

2.3.4. F1 Score

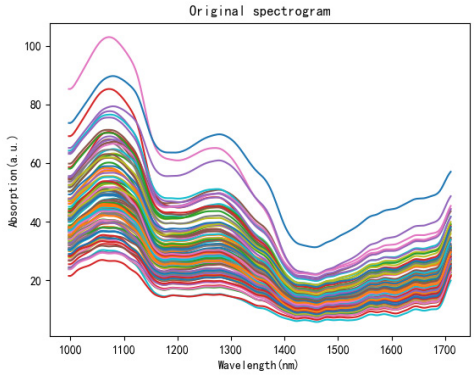
The F1 Score is the harmonic mean of precision and recall and is used to strike a balance between precision and recall. The F1 score is a useful performance metric when both precision and recall are important. The F1 score ranges from 0 to 1, with higher values indicating better model training results. Calculate according to Equation (4)

$$\text{F1 Score} = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (4)$$

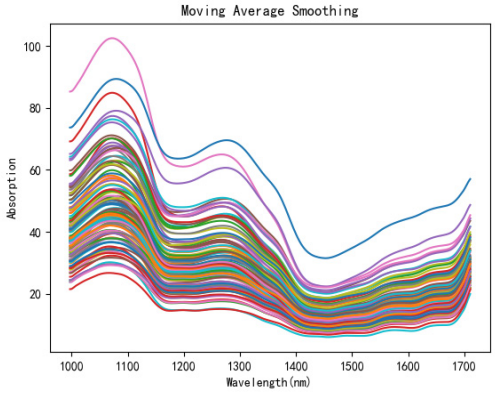
3. Results and Analysis

3.1. Comparison of Preprocessing Methods

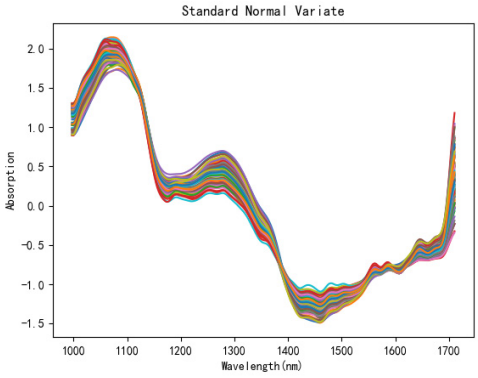
For the original near-infrared spectral data, there is a lot of redundant information in the spectral band that has nothing to do with the sample itself, which will affect the training and prediction effect of the established model. In order to improve the spectral signal-to-noise ratio and the prediction effect of the model, the data preprocessing method is used to reduce the error in the model training process. The spectrograms processed by different preprocessing methods are shown in Fig 3. (b) - (d), and from the original spectrogram (a), it can be seen that the overall trend of the spectra after the preprocessing methods is basically consistent with the original spectra.



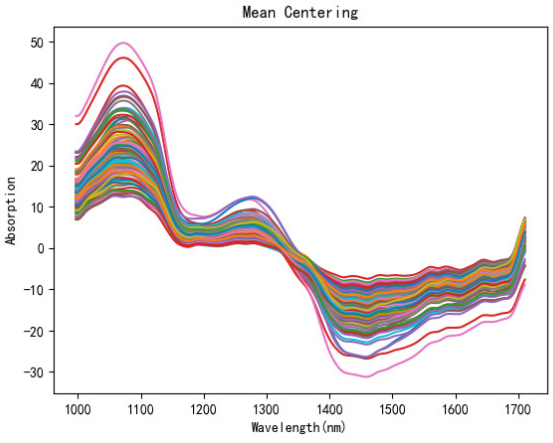
a Original spectrogram



b Spectra after moving average smoothing



c Spectra after standard normal variate



d Spectra after mean centering

Fig 3. The spectrograms processed by different preprocessing methods

According to the preprocessed result figure, MSC significantly reduces the noise level, and the processed spectral curve is clearer and smoother, while retaining important spectral features such as peaks and valleys, which improves the overall quality of the data while retaining all the information points of the original data. SVN scales the absorption value to a smaller and uniform range, reduces the intensity variation caused by the light scattering effect, improves the data signal-to-noise ratio, and makes the comparison between different samples more accurate. After MC preprocessing, the features in the spectra are more obvious, the main features and original distribution characteristics of the spectral data are preserved, the baseline of the spectra is more stable, and the baseline drift caused by the measurement conditions or sample differences is reduced.

3.2. Comparison of Modeling Methods

In the experiment, the performance of DT, SVC and KNN classification models under MAS, SNV and MC preprocessing methods are compared, and the experimental results are shown in Table 1. The results show that the preprocessing method has a significant impact on the model performance. DT performs best under MC preprocessing, with the accuracy, precision, recall and F1 score all reaching 0.955. The performance of SVC is improved to 0.864 after MAS and SNV preprocessing, and the performance of KNN is improved to 0.91 after MC preprocessing. The CARS feature selection method adaptively adjusts the selection

probability of each band through Monte Carlo sampling and exponential decay function, and finally selects the optimal band combination that contributes the most to the modeling performance. Under different models and preprocessing methods, CARS have a certain impact on the performance, but the effect depends on the specific model and preprocessing method. In general, DT under MC preprocessing combined with CARS feature selection performs best, and the classification effect is the best.

Table 1. Classification results of different models

Preprocessing	Model	Feature extraction	Accuracy	Precision	Recall	F1 Score
No	DT	CARS	0.73	0.73	0.73	0.73
MAS			0.73	0.73	0.73	0.73
SNV			0.64	0.64	0.64	0.64
MC			0.955	0.955	0.955	0.955
No	SVC	CARS	0.773	0.773	0.773	0.773
MAS			0.864	0.864	0.864	0.864
SNV			0.864	0.864	0.864	0.864
MC			0.773	0.773	0.773	0.773
No	KNN	CARS	0.73	0.73	0.73	0.73
MAS			0.682	0.682	0.682	0.682
SNV			0.82	0.82	0.82	0.82
MC			0.91	0.91	0.91	0.91

4. Summary

This paper discusses the internal quality classification research of Shaanxi Luochuan Red Fuji apple, using near infrared spectroscopy technology combined with SPXY data set division method, as well as MAS, SNV, MC three data preprocessing methods, using CARS algorithm for data feature wavelength selection, based on SVC, DT, KNN algorithm to establish the internal quality classification model of apple. In the constructed classification model, the best apple internal quality classification model based on near infrared spectroscopy is SPXY + MC + CARS + DT model, which has an accuracy of 0.955, a precision of 0.955, a recall of 0.955, and an F1 score of 0.955. The experimental results show that the NIR spectroscopy method can effectively classify the internal quality of apples, which can be used to guide the classification processing of apples after harvest, and can also provide theoretical basis for the classification processing of different varieties of fruits after harvest. With the research and development of the near-infrared spectroscopy model, the advantages of the model for fruit internal quality classification can be further optimized in the future, so as to improve the robustness and practicability of the detection technology.

References

[1] Guo Z, Chen X ,Zhang Y , et al.Dynamic Nondestructive Detection Models of Apple Quality in Critical Harvest Period Based on Near-Infrared Spectroscopy and Intelligent Algorithms [J].Foods,2024,13(11).

[2] Sanqing L, Shuxiang F ,Lin L , et al.An improved method for predicting soluble solids content in apples by heterogeneous transfer learning and near-infrared spectroscopy [J].Computers and Electronics in Agriculture,2022,203.

[3] Zhao C, Yin Z ,Zhang W , et al.Identification of apple watercore based on ConvNeXt and Vis/NIR spectra[J].Infrared Physics and Technology,2024.

[4] Tian H ,Zhang L ,Li M , et al.Weighted SPXY method for calibration set selection for composition analysis based on near-infrared spectroscopy[J].Infrared Physics and Technology, 2018.

[5] Run C. Determination of fatty acid of wheat by near-infrared spectroscopy with combined feature selection based on CARS and NSGA-III[J]. Infrared Physics and Technology,2023,129.