ISSN 2616-5767 Vol.7, Issue 3: 84-88, DOI: 10.25236/AJETS.2024.070312

Exploration of an Apple Grading Model Based on Near-Infrared Spectroscopy

Fengkang Qiao[1], Tingting Liu[1], Zhipeng Li[1], Yanan Xue[1]

[1]*School of Electronic Information, Xijing University, Xi'an, Shaanxi, China*

Abstract: *With food safety issues gaining widespread attention, monitoring the quality of fruits has become an essential part of ensuring consumer health. Apples, being a significant component of daily diets, have their safety and quality highly prioritized by consumers. This study employs Near-Infrared Spectroscopy (NIR) to develop two classification models: Partial Least Squares Discriminant Analysis (PLS-DA) and One-Dimensional Convolutional Neural Network (1D CNN), aimed at performing non-destructive detection of apple quality and precise grading of sweetness. Through the analysis of experimental data, especially after preprocessing with SG smoothing and partitioning using the SPXY algorithm, the 1D CNN model achieved an accuracy rate of 0.8856 in apple quality classification, demonstrating its significant potential for application in apple sweetness grading. This study validates the efficiency and reliability of Near-Infrared Spectroscopy in the assessment and grading of agricultural product quality, providing substantial support for improving agricultural product quality and ensuring consumer food safety and health.*

Keywords: *Food Safety, Non-Destructive Testing, Near-Infrared Spectroscopy, 1D-CNN*

1. Introduction

In recent years, as people's living standards improve, the demand for higher quality agricultural products has increased. Apples, being one of the most consumed fruits globally, have their quality control and grading becoming a crucial part of the supply chain. Traditional apple grading primarily relies on human vision, which is not only inefficient but also subject to individual subjective judgment, leading to inconsistencies in grading results. Thus, developing a rapid, objective, and accurate automatic grading technology is particularly important. Near-Infrared Spectroscopy (NIR), known for its non-destructive, rapid, and chemical composition revealing characteristics, shows great application potential in the field of agricultural product quality assessment. By analyzing the near-infrared spectrum data of apples, it's possible to accurately assess their internal quality, including key indicators such as moisture, sugar content, and acidity, without damaging the fruit.

2. Experimental Section

2.1. Experimental Equipment

The Near-Infrared Spectroscopy instrument used is OPTOSKY's ATP8600, with a wavelength range of 996-1710nm, taking a wavelength point every 3nm for detection. The light source comprises two 100W halogen lamps, with optical fibers connecting the detection probe, collimating mirror, etc. The refractometer used is ATAGO's (Japan ATAGO Company) PAL-BX/ACID8.

2.2. Apple Sample Spectral Data Collection

Following the study results by Liu Yande et al[1], employing a diffuse transmission light path for measuring the near-infrared spectral data values of apples can achieve better experimental outcomes. The light path setup, as shown in Figure 1, aims to minimize the influence of ambient stray light on experimental results, with operations conducted in a darkroom.

ISSN 2616-5767 Vol.7, Issue 3: 84-88, DOI: 10.25236/AJETS.2024.070312

Figure 1: Optical Path Construction

2.3. Collection of Apple Physicochemical Value Information

A total of 300 apple samples were collected in this experiment, with 24 samples excluded due to abnormal spectral data caused by improper operation.

3. Data Preprocessing and Sample Division

3.1. Data Preprocessing

Near-infrared spectral data often suffer from various interferences during the collection process, such as instrument noise, changes in ambient light, and the heterogeneity of the samples themselves. These interferences may mask or distort the true spectral features of the samples. Therefore, data preprocessing is a necessary step to improve the accuracy and stability of the model. Effective data preprocessing can remove or reduce these interferences and enhance the features related to sample properties in the spectral signal, providing reliable data support for precise grading. The raw spectral data collected are shown in Figure 2:

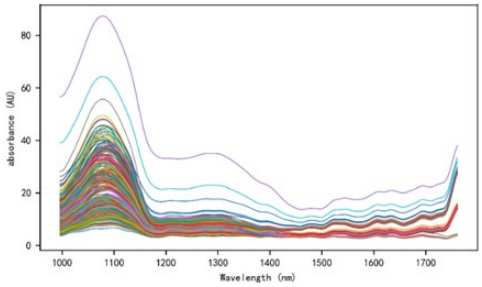

Figure 2: Raw spectral data of apples

The data preprocessing methods used in this article mainly include:

(1) Standard Normal Variate Transformation (SNV)

Standard Normal Variate Transformation is a common method used to eliminate the effects of particle size and path length differences in spectral data. It standardizes each spectrum so that the data have zero mean and unit variance [2], thereby improving the comparability between different samples. The spectral graph of the data processed by Standard Normal Variate Transformation is shown in Figure 3:

ISSN 2616-5767 Vol.7, Issue 3: 84-88, DOI: 10.25236/AJETS.2024.070312

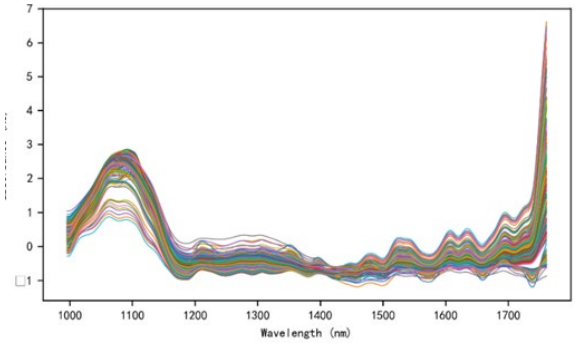

Figure 3: SNV spectral graph

(2) Savitzky-Golay Smoothing Filter (SG Filtering)

The basic principle of Savitzky-Golay filtering is to fit a polynomial to the data points in the neighborhood of each data point (including the point itself and the data on both sides) using the least squares method. For each data point, its value is recalculated based on the value of this polynomial. During this process, one can choose to adjust the order of the polynomial (such as first or second order) and the window size (i.e., the considered neighborhood range) to achieve better data preprocessing effects. The spectra after processing with SG smoothing filtering are shown in Figure 4:

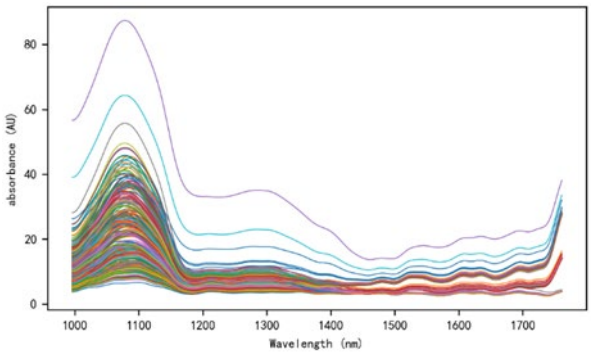

Figure 4: SG spectral graph

3.2. Dataset Division

In machine learning and statistical modeling, correctly dividing the dataset is a key step to ensure that the model has good generalization ability. Common sample division algorithms include the SPXY algorithm and the KS algorithm, both aimed at selecting a representative training set to ensure that the model can capture and learn the main variability and structure of the data.

(1) KS (Kennard-Stone) Algorithm

The KS algorithm aims to select highly representative samples from the original dataset to form the training set. It emphasizes covering all samples uniformly in the feature space to ensure that the training set represents the characteristics of the entire dataset. It uses the two samples with the largest Euclidean distance among the available samples as the starting samples. Among the remaining samples, it finds the sample with the largest Euclidean distance from all samples in the current training set and adds it to the training set. This process is repeated until the training set reaches a predetermined size.

(2) SPXY (Sample set Partitioning based on joint X–Y distances) Algorithm

The SPXY algorithm is a strategy for dataset division, particularly suitable for Partial Least Squares Regression (PLSR) and other modeling methods that depend on sample representativeness. The goal of the algorithm is to ensure that the training and test sets have good statistical consistency, allowing the model to generalize better to new, unseen data. The algorithm first calculates the similarity or distance between every pair of samples in the dataset, typically using Euclidean distance or other similarity metrics. Based on the similarity assessment, it selects the pair of samples with the highest similarity (i.e., the shortest distance). One sample from each pair is randomly assigned to the training set, and the other to the test set, proceeding in this manner until the predetermined size ratio of the training and test sets is achieved [3]. The total number of samples and the division into training and test sets are shown

ISSN 2616-5767 Vol.7, Issue 3: 84-88, DOI: 10.25236/AJETS.2024.070312

in Table 1:

Table 1: Dataset Division

Total samples	Training set	Test set
276	220	56

4. Model Construction and Result Analysis

4.1. Partial Least Squares Discriminant Analysis (PLS-DA)

Partial Least Squares Discriminant Analysis is a supervised learning method that builds a classification model by finding the linear combination of independent variables (spectral data) and dependent variables (sample categories) that maximizes covariance. Based on the framework of Partial Least Squares Regression (PLSR), PLS-DA seeks latent variables that maximize the covariance between independent variables (X) and dependent variables (Y) to achieve data classification. This model first encodes the classification targets into numerical data (such as one-hot encoding), where each category is represented by an independent variable. Then, using the PLSR method to analyze the independent variables X and the encoded dependent variables Y, it identifies the latent variables (LVs) that can explain the covariance between X and Y[4]. Finally, the model classifies samples based on its output, typically comparing the prediction results of the PLSR model with a threshold to determine the sample categories.

4.2. One-Dimensional Convolutional Neural Network (1DCNN)

One-Dimensional Convolutional Neural Network (1D CNN) is an innovative technology in the field of deep learning, designed specifically for sequence data. By applying one-dimensional convolution operations on data sequences, 1D CNN can effectively capture local dependencies and pattern features, significantly enhancing the understanding and analysis of sequence data. This model type automates the extraction of important features using convolutional layers, avoiding the cumbersome feature engineering steps common in traditional machine learning methods. Subsequently, by introducing nonlinear activation functions and pooling layers, the model not only enhances its nonlinear expression ability but also effectively reduces the feature dimension, improving computational efficiency [5]. After a series of convolutional and pooling operations, the fully connected layer maps the extracted features to the final output categories or values, completing the classification or regression tasks. These advantages make 1D CNN significantly effective in processing near-infrared spectral data.

In this article, we design and use a 1D CNN network consisting of four convolutional blocks (CONV1 to CONV4) and two fully connected layers, each convolutional block contains a convolutional layer, batch normalization layer, ReLU activation function, and max pooling layer.

(1)Convolutional Blocks

CONV1: Uses Conv1d, input channel is 1, output channel is 32, kernel size is 3, stride is 1. Followed by batch normalization, ReLU activation, and 2x2 max pooling.

CONV2: Uses Conv1d, input channel is 32, output channel is 64, kernel size is 2, stride is 1. Followed by batch normalization, ReLU activation, and 2x2 max pooling.

CONV3: Uses Conv1d, input channel is 64, output channel is 128, kernel size is 2, stride is 1. Followed by batch normalization, ReLU activation, and 2x2 max pooling.

CONV4: Uses Conv1d, input channel is 128, output channel is 256, kernel size is 1, stride is 1. Followed by batch normalization, ReLU activation, and 2x2 max pooling.

(2)Fully Connected Layers

The first fully connected layer maps from 512 nodes to nls nodes (nls is a dynamic parameter representing the number of nodes in the output layer, set to 3 in this case, dividing apple samples into three categories), using Dropout to reduce overfitting.

The second fully connected layer maintains the mapping from nls to nls, also applying Dropout to prevent overfitting.

ISSN 2616-5767 Vol.7, Issue 3: 84-88, DOI: 10.25236/AJETS.2024.070312

4.3. Result Analysis

Accuracy is a method of measuring the performance of classification models, defined as the proportion of correctly classified samples to the total number of samples. Mathematically expressed as:

$$\mathrm{ACC} = \frac{\mathrm{NC}}{\mathrm{N}} \tag{1}$$

where NC represents the number of correctly classified samples, and N represents the total number of samples.

Based on this model evaluation method, the results obtained for the two models in this article are shown in Table 2:

Table 2: Classification Accuracy of Each Model

Preprocessing Method\Modeling Method	PLS_DT	1DCNN
SG+SPXY	0.7963	0.8856
SNV+SPXY	0.7532	0.7866
SG+KS	0.6954	0.6832
SNV+KS	0.5634	0.6259

Analyzing the accuracy in the table, it can be concluded that the SG+SPXY+1DCNN method can accurately perform the task of grading and sorting apples.

5. Conclusion

This study, by employing Near-Infrared Spectroscopy technology combined with two classification models, Partial Least Squares Discriminant Analysis (PLS-DA) and One-Dimensional Convolutional Neural Network (1D CNN), successfully conducted non-destructive detection of apple quality and accurately graded its sweetness level. Among these, the 1D CNN model achieved an accuracy rate of 0.8856 after processing with SG smoothing filtering and SPXY data partitioning algorithms.

Through this research, we further confirm the practical value and broad prospects of Near-Infrared Spectroscopy technology in the fields of food safety and quality control. In the future, these technologies can not only be applied to the quality grading of single fruit varieties like apples but are also expected to be extended to a wider range of agricultural product quality assessment and classification, thus providing consumers with safer and higher quality food options. Moreover, with the continuous advancement of artificial intelligence technologies such as deep learning, intelligent quality detection combining spectroscopy technology will become an important development direction in the field of food safety.

References

[1] CX,Zhang,YH,et al.Tobacco Plant Parts Similarity Analysis Based on Near-Infrared Spectroscopy and SIMCA Algorithm[J].SPECTROSC SPECT ANAL, 2011, 2011,31(4)(-):924-927.

[2] Xu L , Liu J , Wang C ,et al.Rapid determination of the main components of corn based on near-infrared spectroscopy and a BiPLS-PCA-ELM model[J].Applied optics, 2023.

[3] Yu-Miao L , Hai-Peng L I , Kun S ,et al.Recent Advances in the Application of Near-Infrared Spectroscopy in Beef Quality Grading Systems[J].Meat Research, 2012.

[4] Zhu X , Hu C , Wang W ,et al.Nondestructive detection of water content of jujubes based on visible-near infrared spectroscopy[C]//2017 Spokane, Washington July 16 - July 19, 2017.2017. DOI:10.13031/aim.201700800.

[5] Li L , Li B , Jiang X ,et al.A Standard-Free Calibration Transfer Strategy for a Discrimination Model of Apple Origins Based on Near-Infrared Spectroscopy[J].Agriculture, 2022, 12.