A Method of Distinguishing Tea varieties Based on Hyperspectral Imaging

Zhong Yanqi1st, Kang Zhiliang2nd*, Wang Peng3rd, Luo Xiong4th
Sichuan Agricultural University, College of Mechanical and Electrical Engineering, Ya’an, China
*12200@sicau.edu.cn

Abstract—In order to realize the rapid and non-destructive identification of tea varieties, this paper based on hyperspectral imaging technology to find the optimal discrimination model of tea varieties. This article is mainly divided into three aspects: the discriminant model of tea varieties based on spectral characteristics, the discriminant model of tea varieties based on image features, and the discriminant model of tea varieties based on spectral-image fusion features. The experimental results show that Model 1 uses the full-spectrum feature combined with support vector machine (SVM) model, which can distinguish the accuracy of different tea varieties up to 100%. Model 2 is based on the GLCM texture feature based on the characteristic gray image combined with the SVM model, and the discrimination accuracy of different tea varieties reaches 100%. Model 3 discusses the impact of different preprocessing methods on the accuracy of classification under the fusion of two information features, determines Minmax as the best preprocessing method, and obtains 100% classification accuracy in the test set.

1. Introduction
Tea contains bioactive substances that resist various diseases, such as catechins, flavonols, anthocyanins, which make consumers pay more attention to the quality and nutrition of tea in recent years. The type and grade of tea have a direct impact on the taste and market value of tea. However, in the current tea market, there is still a phenomenon of shoddy and genuine tea, which not only has a negative impact on the establishment of tea brands, but also greatly damages the rights and interests of consumers [1, 2]. Therefore, it is of great significance to study a rapid method for tea quality identification.

Under the influence of the acquisition and processing speed of hyperspectral data, hyperspectral imaging technology is mainly used in the basic research of nondestructive testing of quality and safety of agricultural products. Compared with near-infrared spectroscopy, on-line detection and direct use of hyperspectral imaging technology for on-line or real-time detection applications are relatively few. There have been many achievements in the field of tea research in near infrared spectroscopy. Guangxin Ren et al. used near infrared spectroscopy combined with PLS model to establish regression models for the four components of black tea. In the correction set, the root mean square error RMSE reached 0.102%, 0.654%, 0.552% and 0.248%, respectively, and the correlation coefficients R reached 0.955, 0.962, 0.954, 0.927, respectively. Then, the origin traceability models of different types of black tea were established with an accuracy of 94.3%[3]. There are also different results for studies based on hyperspectral. A method of origin identification of Lushan Yunwu tea based on hyperspectral
technology was proposed by Ai Shirong et al. First, the hyperspectral imaging system was used to collect the hyperspectral data of Lushan Yunwu Tea and Yunwu Tea in three other producing areas, namely Guangxi, Sichuan and Fujian. Principal component analysis method was used to select three characteristic wavelengths from the original hyperspectral data block and extract the texture features of the grayscale image at each characteristic wavelength respectively. The BP neural network method was used to establish the Lushan Yunwu tea origin identification model. The recognition rate of response in model training is 97.25%, and that in prediction is 95% [4].

In short, it is rare to combine spectral information and image information for agricultural products, especially tea. Hyperspectral images obtained by hyperspectral imaging system have the characteristic of "Image spectrum integration", that is, they contain image information and spectral information at the same time. Image information can be used to detect the external quality of samples, while spectral information can be used to detect their internal quality and security [5]. This article aims to explore the use of hyperspectral imaging technology to detect various aspects of tea, and establish a multi-index fusion recognition model for different types of tea.

2. Materials and Methods

2.1 The experimental device

In this study, "Gaia Sorter" hyperspectral separator (Beijing • ZOLIX INSTRUMENTS CO., LTD) was selected for the acquisition of spectral data and image information. The core components of this hyperspectral acquisition system include hyperspectral camera, uniform light source, electrically controlled mobile platform, computer and control software, etc. It is powered by an AC220V power supply. Among them, the hardware that plays the main imaging function is a hyperspectral camera. The hyperspectral camera used in the hyperspectral imaging system uses the Image-λ "spectral image" series of hyperspectral cameras from ZOLIX INSTRUMENTS CO., LTD, which is mainly composed of an imaging spectrometer and a CCD. In the actual process of acquiring hyperspectral images, the error will be caused by the weather, external lights, etc. that will irradiate the sample to be measured. In order to eliminate this error, the light source system, the electronic control platform system and the hyperspectral camera are combined into a dark box with light-shielding materials. After adjusting the various parameters of the system, in the process of acquiring images, the interference of light sources outside the dark box environment should be avoided. The overall structure of the hyperspectral imaging system is shown in Figure 1.

As the equipment control and data storage terminal of the whole system, computer plays a leading role in the imaging system. SpectraVIEW software provided by Beijing ZOLIX INSTRUMENTS CO., LTD is mainly used to set up the basic imaging system, including the starting position, forward short distance, forward speed and backward speed of the electric mobile platform; it also includes the
camera pixels, exposure time, Gain and other parameters. It can also realize the preliminary correction of hyperspectral image data, such as black-and-white correction, uniform correction, envelope removal and other analysis functions.

2.2 The experimental materials
The tea samples used in this experiment were purchased from Ya'an Hongling Tea Factory, including three different varieties of Mengshan Huangya, Zhuyeqing and Ganlu in Ya'an City, Sichuan Province. Immediately after the purchase, they were imaged using a hyperspectral imaging system. Tea samples were placed in a 100×20 mm glass petri dish and randomly and evenly spread around the bottom of the dish for a sample about 1cm. Each sample was about 10g, and 50 samples were collected for each variety, a total of 150 samples. All purchased teas comply with the provisions of the national standard GB/T 18665-2008.

2.3 Hyperspectral image acquisition
Before image collection, the brightness of the light source and the height of the lens should be adjusted to the appropriate position. Through Spec-View software, we set the exposure time of the hyperspectral imaging system to 1.5s, the gain factor to 1.1, the moving speed of the motorized translation stage to 3.5mm/s, the maximum measurement stroke to 12cm, and the lens to the sample height to 50cm. Finally, a total of 300 tea samples with 520 bands were collected in the range of 380 nm to 1040nm. After performing black-and-white correction and MNF noise removal on the hyperspectral image, through the ROI (Region of Interests Function) tool of ENVI4.8 software, manually extract a 300×300×520 data cube from the original hyperspectral image for further analysis.

2.4 Data processing and model evaluation
After the acquisition of hyperspectral images, the acquired images need to be uniformly corrected in black and white [6]. In order to further eliminate noise and redundancy, the black and white corrected hyperspectral image is positively transformed, and the first 10 bands with larger eigenvalues are extracted for MNF inverse transformation [7-10]. Pre-processing methods also include data standardization, data centralization, and normalization. The common strategy of raw data redundancy is data dimensionality reduction.

In order to explore the robust analysis model, three analysis methods were used to compare tea classification. By using spectral information, texture and color characteristics using hyperspectral image features, spectral characteristics and image characteristics, and establishing the corresponding identification models of three tea varieties for comparison, the final preferred model is obtained.

- Model 1: Using spectral information. By comparing the accuracy between the original spectral feature modeling and the spectral feature modeling filtered by SPA algorithm, the screening effect of principal component analysis (PCA) algorithm [12] can be discussed. Next, the Support Vector Machine (SVM) [13] and BP neural network [14-17] were respectively used to model the original spectral features (None) without any treatment and the spectral features screened by SPA, and the models of None-SVM, None-BP, SPA-SVM and SPA-BP were established respectively.

- Model 2: Using the texture and color features of hyperspectral image features. The color features and texture information are analyzed as image features, and the original image feature (None) and the image feature filtered by SPA are modeled by SVM and BP neural network. Because the extracted feature levels are inconsistent, the data is first Autoscaled, and the models of None-SVM, None-BP, SPA-SVM, and SPA-BP are established respectively. The SPA algorithm sets the variable selection range to 1-20.

- Method 3: The spectral characteristics and image characteristics are fused to establish the corresponding tea variety identification model. It is not comprehensive to use any unilateral characteristics to identify tea varieties, so the two characteristics are combined to establish the corresponding tea variety identification model. This method mainly studies different data preprocessing methods to determine the influence of model accuracy: Autoscaling, Center, Minmax and None.
The above contents mainly use the accuracy and precision of correction set and test set as well as modeling time to evaluate and verify the model. The analysis process was completed in ENVI4.8 and MATLAB2014. The technical route of the research process is shown in Figure 2.

3. Results and Discussion

3.1 Spectral analysis of tea
The above mentioned preprocessing of the original data includes black-and-white correction and MNF transformation, which can only eliminate the influence of certain noises. After the MNF transformation, the true-color composite image as shown in Figure 3 and the spectral curve of the same pixel are obtained. (The abscissa of the line graph is the wavelength, and the ordinate is the reflectance) It can be seen from the image that the random noise is significantly reduced and the spectral curve tends to be smooth, which is suitable for the next analysis of hyperspectral images.
Using the ROI (Regin of Interests Function) tool of ENVI4.8 software, a 300×300×520 data cube was manually extracted from the original hyperspectral image, and the average spectral curve of the samples of the three types of tea in the range of 300×300 pixels was calculated, as shown in Figure 4 below. Through observation, it can be found that the three types of tea have strong spectral absorption between 400 and 750nm. The three types of tea have similar trends in spectral absorption, but have different reflectance values at different light wavelengths. The spectral absorption curves of the three types of tea near 430nm and 660nm are significantly different, which may be related to the content of chlorophyll-a in the three types of tea [18].

3.2 Feature extraction and modeling results
- The above three models are adopted in this study. First, the spectral information of hyperspectral images was used to model the classification of tea internal quality. A method of variable selection was adopted to reduce the complexity of the model: Successive Projections Algorithm (SPA) [19-21]. Through comparison with the original data, the above four models are obtained for comparison, and the discrimination results are shown in Table 1. It can be seen from the table that for the spectral data, the variety discrimination model using the SVM algorithm is better than the model using the BP neural network overall in accuracy. Regarding the effect of SPA filtering characteristic wavelengths, the
results show that modeling the spectral features after SPA screening does not increase the accuracy of the discriminant model, but decreases it somewhat.

**TABLE I. IDENTIFICATION RESULTS OF TEA VARIETIES BASED ON SPECTRAL CHARACTERISTICS**

| Mode | Variable selection | Varieties | Calibration set | The test set |
|------|--------------------|-----------|-----------------|--------------|
|      |                    |           | Miscalculation n/a | Accurate rate/% | Miscalculation n/a | Accurate rate/% |
| SVM  | None               | Ganlu     | 0               | 100           | 0               | 100            |
|      | SPA                | Ganlu     | 0               | 100           | 0               | 100            |
|      |                    | Huangya   | 0               | 100           | 0               | 100            |
|      |                    | Zhuyeqing | 0               | 100           | 0               | 100            |
|      |                    |           | 100             |               | 100            |                |
| SPA  | None               | Ganlu     | 0               | 100           | 0               | 100            |
|      | SPA                | Ganlu     | 0               | 100           | 0               | 100            |
|      |                    | Huangya   | 0               | 100           | 0               | 100            |
|      |                    | Zhuyeqing | 0               | 100           | 0               | 100            |
|      |                    |           | 100             |               | 100            |                |
| BPN  | None               | Ganlu     | 0               | 100           | 0               | 100            |
|      | SPA                | Ganlu     | 0               | 100           | 0               | 100            |
|      |                    | Huangya   | 0               | 100           | 0               | 100            |
|      |                    | Zhuyeqing | 0               | 100           | 0               | 100            |
|      |                    |           | 100             |               | 100            |                |

**TABLE II. TEA SPECIES DISCRIMINATION MODEL BASED ON HYPERSPECTRAL IMAGE FEATURES**

| Mode | Variable selection | Varieties | Calibration set | The test set |
|------|--------------------|-----------|-----------------|--------------|
|      |                    |           | Miscalculation n/a | Accurate rate/% | Miscalculation n/a | Accurate rate/% |
| SVM  | None               | Ganlu     | 0               | 100           | 0               | 100            |
|      | SPA                | Ganlu     | 0               | 100           | 0               | 100            |
|      |                    | Huangya   | 0               | 100           | 0               | 100            |
|      |                    | Zhuyeqing | 0               | 100           | 0               | 100            |
|      |                    |           | 100             |               | 100            |                |
| SPA  | None               | Ganlu     | 0               | 100           | 0               | 100            |
|      | SPA                | Ganlu     | 0               | 100           | 0               | 100            |
|      |                    | Huangya   | 0               | 100           | 0               | 100            |
|      |                    | Zhuyeqing | 0               | 100           | 0               | 100            |
|      |                    |           | 100             |               | 100            |                |
| BPN  | None               | Ganlu     | 0               | 100           | 0               | 100            |
|      | SPA                | Ganlu     | 0               | 100           | 0               | 100            |
|      |                    | Huangya   | 0               | 100           | 0               | 100            |
|      |                    | Zhuyeqing | 0               | 100           | 0               | 100            |
|      |                    |           | 100             |               | 100            |                |

- The second is the modeling research of tea internal quality classification based on image characteristics. To obtain the texture features of the feature images of each tea variety, PCA method was used to extract the feature wavelength, and then the gray-level co-occurrence Matrix (GLCM) method was used to extract the texture features. Principal component analysis of hyperspectral images is performed to remove the linear correlation between image data and the dimension reduction of the original data. The grayscale co-occurrence matrix was used to extract texture features [22]. The above four models are also compared, and the discrimination results are shown in Table 2. The discriminant model based on BP neural network and image features is better than the discriminant model based on SVM, achieving 100% classification accuracy in both correction set and test set. However, two Zhuyeqing samples were misclassified as Huangya in the SVM classification model. In general, the use of image features combined with pattern recognition methods can also effectively discriminate tea types.
The third is feature fusion, which combines spectral features with image features. The fusion feature matrix was obtained by combining 36 image-based features with 320 spectral based features, and the discrimination model of different tea varieties was established based on the fusion features. Since the magnitude of color feature and spectral feature are not on the same dimension, data preprocessing of fusion matrix is needed. Feature extraction methods were extracted according to the above two methods. Due to the superiority of the SVM algorithm in modeling and analysis on the basis of small sample data, the SVM model between the fusion feature and the tea category is established, and the results are shown in Table 3. By observing the above table, the SVM tea variety classification model obtained by using the Minmax-None pretreatment method has obtained the best performance, and achieved 100% classification accuracy on both the calibration set and the test set, which can be considered verified The feasibility of identifying tea varieties based on fusion features.

| pretreatment | Autoscaling | Center | Minmax | None |
|--------------|-------------|--------|--------|------|
| Correction set accuracy/% | 100 | 100 | 100 | 100 |
| Test set accuracy/% | 96.67 | 93.33 | 100 | 93.33 |
| Modeling time/s | 4.29 | 4.20 | 3.96 | 4.19 |

3.3 Discussions
It can be found that the method of fusion of spectral characteristics and image characteristics is more comprehensive than that of one method. Spectral data characterize tea internal information, image data characterize tea external information. Due to the large differences between the varieties, the unilateral research has also achieved good classification results. From the variety classification alone, the results obtained by the three methods are relatively good.

4. Conclusions
In the tea varieties identification model based on the fusion feature, the spectral characteristics of the fusion with GLCM texture feature, and then discuss the four different impact on the SVM discriminant model of data pretreatment method, and determine the optimal pretreatment method of the fusion feature for normalization (Minmax), at this point on the calibration set and testing set were 100% accurate, after all.

Although there are still many shortcomings, such as the use of feature extraction and band analysis method is relatively single, it still needs to make some references for the real application of hyperspectral technology in production practice through constantly enriching the database. In the future, the existing methods can be presented in the form of human-computer interaction software, and the data in the analysis process can be visualized to make some preparations for online detection. But the experimental results show that the tea variety discrimination method based on hyperspectral imaging can basically realize the fast and nondestructive classification of tea.

Author
Zhiliang Kang* (Corresponding author) is a professor in College of Mechanical and Electrical Engineering, Sichuan Agriculture University, Ya’an, 625014, China. His research interests include agricultural engineering, automatic control technology and information processing. (Email: 12200@sicau.edu.cn)

REFERENCES
[1] Zheng Guojian, GAO Haiyan. Analysis of current situation of tea product quality safety in China [J]. Journal of Food Safety and Quality Inspection, 2015(7):2869-2872.
[2] Tong Xiaolin. Current situation and Countermeasures of Tea Quality safety Management standardization in China: the 9th China Standardization Forum, Liuzhou, Guangxi, 2012[C].

[3] Guangxin Ren, Shengpeng Wang, Jingming Ning, et al. Quantitative analysis and geographical traceability of black tea using Fourier transform near-infrared spectroscopy (FT-NIRS)[J]. Food Research International, 2013,2(53):822-826.

[4] Ai Shi-rong, WU Rui-mei, Wu Yan-hong, et al. Identification of lushan Cloud Tea producing area by hyperspectral imaging technology [J]. Journal of Jiangxi Agricultural University, 2014(02):428-433.

[5] Zhang Baohua, LI Jiangbo, Fan Shuxiang, et al. Principle and application of hyperspectral imaging technology in nondestructive testing of fruit and vegetable quality and safety [J]. Spectroscopy and Spectral Analysis, 2014(10):2743-2751.

[6] WU D, CAO F, FENG S, et al. Application of infrared spectroscopy technique to protein content fast measurement in milk powder based on support vector machines[J]. Spectroscopy & Spectral Analysis, 2008,5(28):1071-1075.

[7] Wei Xinhua, Wu Shu, XU Laqi, et al. Hyperspectral image recognition of grain cotton hetero-fiber based on minimal noise separation [J]. Chinese Journal of Agricultural Engineering, 2014(9):243-248.

[8] WU D, HE Y, FENG S, et al. Study on infrared spectroscopy technique for fast measurement of protein content in milk powder based on LS-SVM[J]. Journal of Food Engineering, 2008,1(84):124-131.

[9] Corinna Cortes, Vladimir Vapnik. Support-vector networks. Machine Learning[M]. Springer, 1995: 273-297.

[10] Jiang chun-lei, zhang shu-qing, zhang ce, et al. Modeling and prediction of MODIS leaf area index time series based on sarima-bp neural network combination method [J]. Spectroscopy and spectral analysis, 2017,37(1):189-193.

[11] Lu D Y, CUI X T, HUANG J R, et al. Observation of chlorophyll absorption spectra [J]. Journal of University Physics, 2006(01):50-53.

[12] Shi Hui. Study on rapid detection method of prawn quality Information based on hyperspectral imaging technology [D]. Hangzhou: Zhejiang University, 2013.

[13] HARALICK R. Statistical and structural approaches to texture[J]. Proceedings of the IEEE, 2005,5(67):786-804.