ISBN: 978-981-18-3959-7
2022 the 12th International Workshop on Computer Science and Engineering (WCSE 2022)
doi: 10.18178/wcse.2022.06.032

Nondestructive Identification of Cyperus Esculentus Based on Machine Learning and Vis-NIR Hyperspectral Information

Jingyi Zhao [1,2,+], Tao Sha [1,2], Jiahao Wang [1,2] and Wanlin Gao [1,2,+]

[1] Key Laboratory of Agricultural Information Standardization, Ministry of Agriculture and RuralAffairs, China Agricultural University, Beijing 100083, China

[2] College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China

Abstract. At present, the detection of Cyperus esculentus is limited to manual testing and testing based on physical and chemical properties, which is time-consuming and damaging, so it is of great significance to establish a non-destructive testing model for the classification of Cyperus esculentus on the market. Vis-NIR hyperspectral images were used to identify three kinds of Cyperus esculentus from different producing areas. The Vis-NIR Hyperspectral system (382.19nm~1026.66 nm) was used to collect 600 samples and extract the region of interest (ROI). Taking the ROI of a single seed as a sample, a total of 600 sample data were obtained. Firstly, the Savitzky- Golay (SG) algorithm is used to preprocess the sample data. Then Competitive Adapative Reweighted Sampling (CARS) and Successive Projections Algorithm (SPA) are used to reduce the dimensionality of the preprocessed data. Finally, the reduced-dimension features are input into Support Vector Machine (SVM), Extreme Learning Machine (ELM) and Softmax classifier for training. The results show that the reduced-dimensional data can greatly reduce the data redundancy and then reduce the complexity of the model. The experimental results show that the verification accuracy based on SVM, ELM and Softmax model is up to 85.1%, which can effectively achieve rapid non-destructive testing.

Keywords: cyperus esculentus, Vis-NIR hyperspectral, dimensionality reduction, machine learning

1. Introduction

Cyperus esculentus, also known as tiger nut, is a perennial herb of Cyperaceae, which has the characteristics of drought resistance, water saving, rich nutrition and high oil yield. The oil content of oil Cyperus esculentus is 20%-32%, and the per unit yield is higher than that of oil crops such as Cyperus esculentus, rape and peanuts. The underground dry fruit yield is 2 times that of peanuts and 3-4 times that of rape, which is called the king of oil crops [1-3]. Cyperus esculentus is not only a source of high-quality energy intake for human beings, but also a high-quality feed for herbivores.

There are many kinds of Cyperus esculentus, and the existing traditional methods such as chemical detection are used to measure the main components and oil physical and chemical indexes of Cyperus esculentus, which can intuitively use the statistical data of moisture, ash, crude fiber, crude protein and other statistical data to clarify the differences between different producing areas. At present, the testing methods for the quality and variety of Cyperus esculentus are generally traditional testing methods. Li Guanghui et al. [4] analyzed the main nutrient contents of different varieties of Cyperus esculentus according to the national standard, so as to realize the correspondence between different varieties of Cyperus esculentus and their component contents. Liu Yulan et al. [5] tested the impurities, particle size, bulk density, 1000-grain weight, seed coat color, kernel color, kernel color, crude fat, crude protein, crude fiber, moisture, ash and acid value, peroxide value and fatty acid composition of Cyperus esculentus oil according to the national standard. the quality differences of Cyperus esculentus from different areas were analyzed. Zhao Yongguo et al. [6] in view of the advantages of simple primer design and good polymorphism of SRAP molecular markers, 14 molecular fingerprints of different producing areas and different appearance and structure were constructed and their genetic diversity was analyzed.

+ Corresponding author. Tel.: +86-13031891818; fax: +86-10-62738536.
E-mail address: 592753744@qq.com.

Although the traditional detection methods solve the problem of classification accuracy to a certain extent, the aging problem appears for the huge sample data. at the same time, the traditional detection of agricultural products generally needs chemical pretreatment, and the indicators are mostly determined by chemical methods. these detection methods based on chemical methods are not only time-consuming, but also cause damage to the samples. Therefore, there is an urgent need to develop a rapid, non-destructive and green detection method for Cyperus esculentus [7].

2. Related Work

With the arrival of the era of big data, the importance and advantages of machine learning and even deep learning have been highlighted. The improvement of hardware and the continuous optimization of algorithm performance improve the control of data, and then through the massive data for more effective training to obtain more accurate classification or prediction. Tian Qiong et al. [8] determined the element content in Cyperus esculentus samples by X-ray fluorescence spectroscopy, screened and filtered abnormal samples based on box map correction method, then reduced the dimension of Cyperus esculentus sample data by principal component analysis, and established Cyperus esculentus origin identification models by support vector machine, RUSBoosted tree and artificial neural network respectively. The experimental results show that the identification accuracy of artificial neural network modeling is the best, and the accuracy can reach 95.8%. Yan Weimin et al. [9] obtained quinoa data by mid-infrared spectrum irradiation, and principal component analysis (PCA), linear discriminant analysis (LDA) and confusion matrix were used in the stage of dimension reduction modeling and analysis. The experimental results show that the classification accuracy of quinoa in different producing areas can reach 96.25% by using PCA-LDA classification model. Zhang Lixin et al. [10] based on near infrared spectroscopy and intelligent learning algorithm, red Fuji apples from different producing areas were identified and preprocessed, characteristic bands were selected and modeled in a variety of ways. The results show that the feature wavelength extracted by CARS algorithm is the best, and the correct rates of verification sets on ELM and SVM models are 98.75% and 100%, respectively. Hyperspectral technology has been widely used in quality testing, origin traceability and other fields because of its advantages of unified map, high speed and no pollution[11,12]. Hyperspectral technology has been widely used in quality testing, origin tracing and other fields because of its advantages such as unified map, high speed, no pollution and so on. In the research of agricultural products, it is of great significance to the classification of Cyperus esculentus which only stay in the traditional way of detection.

Based on Vis-NIR hyperspectral imaging technology, this paper will carry out non-destructive testing of three kinds of Cyperus esculentus, which provides an experimental basis for the establishment of non-destructive testing model of Cyperus esculentus. The hyperspectral images of Cyperus esculentus at 382.19nm~1026.66nm wavelength were collected, and the spectral data were preprocessed by SG. In order to establish a more accurate model, CARS and SPA were used to reduce the dimensionality of the preprocessed data, and a non-destructive testing method based on visible-near infrared hyperspectroscopy was proposed based on SVM, ELM and Softmax analysis.

3. Experimental Part

3.1. Sample Preparation

In this paper, the varieties of Cyperus esculentus obtained from the origin of Yunnan, Hunan and Hebei provinces in China are selected, and the varieties are shown in Table 1.

Table 1: Experimental sample data

Variety	Sample Source	Total
YN	Acquisition of Origin in Baoshan City, Yunnan Province	200
HN	Acquisition of Origin in Changsha City, Hunan Province	200
BD	Acquisition of Origin in Baoding City, Hebei Province	200

3.2. Vis-NIR hyperspectral Imaging System

In this paper, the Gaiasorter hyperspectral sorter is used for data acquisition, and the enhanced spectral camera Image- λ -V10E-LU is used for shooting. The data acquisition software uses the SpecView software of the hyperspectral imaging system, and the software interface is shown in the figure.

The software can set basic parameters such as exposure time debugging, mobile platform speed setting, hyperspectral image resolution setting, system black-and-white correction, hyperspectral image and access, hyperspectral image analysis and other follow-up processing and analysis. The spectral range is 400nm ~ 1000nm, the wavelength interval is 0.84nm, and there are 728 wavelengths. The parameters of data acquisition include that the distance between the camera lens and Cyperus esculentus is 25cm, and the moving speed of the platform is 0.7cm/s.

The structure of the hyperspectral imaging system is shown in figure 1.

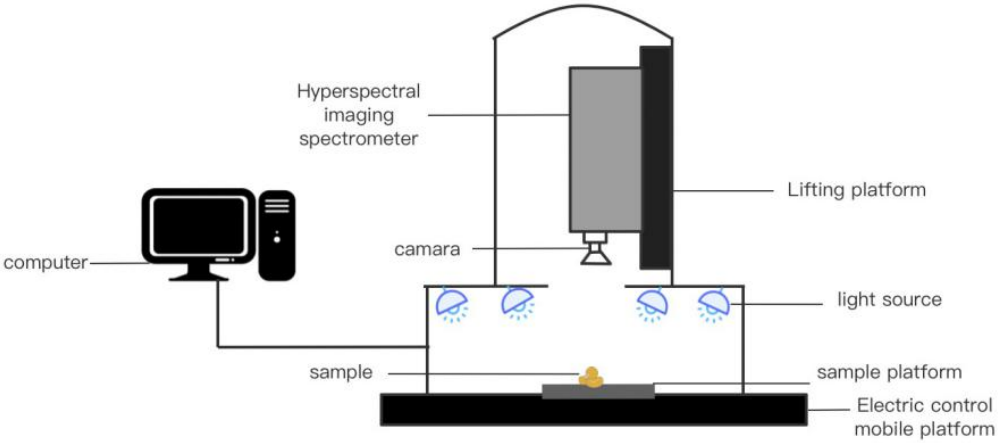

Fig. 1: Schematic diagram of hyperspectral imaging system structure

3.3. Image Acquisition and Correction

Considering that the long-term use of the instrument will produce large noise, dark current and uneven light source intensity distribution, in order to reduce the impact of the above interference on the image signal, and for the effectiveness of follow-up experiments, black-and-white correction provides a good solution to ensure the stability of hyperspectral image shooting. Black and white correction can be done in SpecView, covering the lens cap to take a black calibration image with a reflectivity close to 0%, and then taking off the lens cap to capture a white calibration image with a reflectivity close to 100% of the standard reflection whiteboard.

The formula (1) is used to correct the black and white of the original hyperspectral image of Cyperus esculentus.

$$I_N = \frac{I_R - I_D}{I_W - I_D} \tag{1}$$

Among them, all-white calibration data and all-black calibration data. The original data of the collected samples are corrected in black and white, and the data after reflectivity correction are obtained.

3.4. Spectral Data Extraction

The region of interest of the loaded Cyperus esculentus hyperspectral image is manually selected, and a single Cyperus esculentus sample is selected as a unit. 25 ROI are selected for each image. The average value of spectral reflectance of each sample within the selected pixel range is taken as the feature set of the sample. The ROI of the selected sample is shown in figure 2.

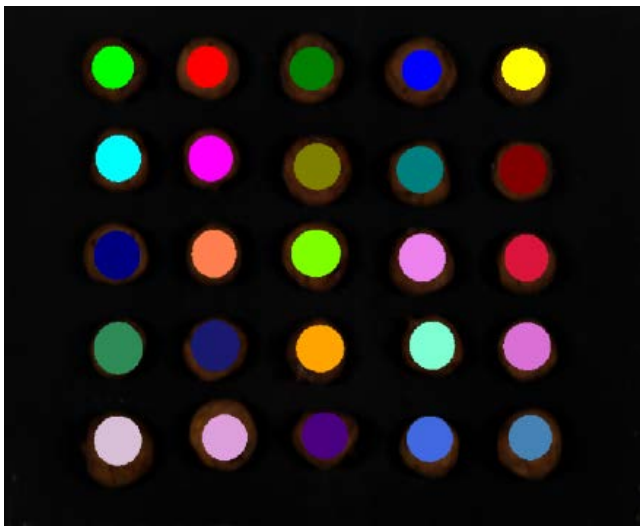

Fig. 2: Selection of ROI of Cyperus esculentus

3.5. Spectral data Preprocessing

The algorithm process is as follows:

$$\overline{X}_i = \frac{1}{c}\sum_{k=-m}^{n} x_{i+k} h_k \tag{2}$$

In formula (2), it represents the average wavelength at the point after convolution smoothing, and C represents the normalized constant and the convolution smoothing coefficient. In this experiment, the 5.2 smoothing formula was used to preprocess the visible near-infrared hyperspectral data of different types of Cyperus esculentus, and five smoothing formulas(3),(4),(5),(6),(7) were obtained by polynomial least square fitting to the moving window (the highest power of the polynomial is 2 and the filter width is 5).

$$\overline{y}_{-2} = 1/35(31y_{-2} + 9y_{-1} - 3y_0 - 5y_1 + 3y_2) \tag{3}$$

$$\overline{y}_{-1} = 1/35(9y_{-2} + 13y_{-1} + 12y_0 + 6y_1 - 5y_2) \tag{4}$$

$$\overline{y}_0 = 1/35(-3y_{-2} + 12y_{-1} + 17y_0 + 12y_1 - 3y_2) \tag{5}$$

$$\overline{y}_1 = 1/35(-5y_{-2} + 6y_{-1} + 12y_0 + 13y_1 + 9y_2) \tag{6}$$

$$\overline{y}_2 = 1/35(3y_{-2} - 5y_{-1} - 3y_0 + 9y_1 + 31y_2) \tag{7}$$

3.6. Spectral Feature Selection

3.6.1. Competitive Adaptive Reweighted Sampling（CARS）

The steps of the algorithm are as follows:

(1) Monte Carlo sampling: randomly select a certain proportion of sample data from the sample set to establish the PLS model, according to the formula.

formula (8) calculate the weight of the first variable:

$$W_i = \frac{|\beta_i|}{\sum_{i=1}^{p}|\beta_i|}, i = 1,2,\ldots, p \tag{8}$$

In the formula, it is the regression coefficient of the PLS model, and p is the number of wavelength variables of the original sample set.

(2) filter the variable value of weight wavelength, eliminate the smaller wavelength based on the exponential attenuation function, retain the larger wavelength, and the retention rate is calculated by the exponential function formula (9).

$$r_i = ae^{-ki} \tag{9}$$

(3) N wavelength variable quantum sets are obtained after N times sampling, and the variable quantum set with the minimum Root mean square error of cross validation (RMSECV) of cross-verification in each sampling process is taken as the optimal wavelength variable quantum set.

3.6.2.Successive projections algorithm（SPA）

SPA uses the projection of the vector for analysis, compares the size of the projection vector through the wavelength projection, takes the wavelength of the largest projection vector as the wavelength to be selected, and sends it to the correction model to screen and determine the feature wavelength combination with less redundant features and the least collinearity.

The original texture feature data space is J represents the number of texture features, and the final number of features extracted is N algorithm as follows:

The initial value n is 0, and when n < N, the loop is carried out, and each cycle carries on the n=n+1 operation. A column vector is selected from J column vectors, and the footer of the unselected column vector is added to the set S.

$$\mathrm{S} = \{\mathrm{j}, 1 <= \mathrm{j} <= \mathrm{J},\ j \in \{k_0, ..., k_{n-1}\}\} \tag{10}$$

(1) calculate the projection in the unselected column vector.

$$P_{X_j} = x_j - (x_j^T x_{k(n-1)}) x_{k(n-1)} (x_{k(n-1)}^T x_{k(n-1)})^{-1},\ j \in s \tag{11}$$

(2) extract the features of the maximum projection vector.

$$K(\mathrm{n}) = \arg(\max(\|P_{X_j}\|)),\ j \in s \tag{12}$$

3.7. Acquisition of Spectral Data of Cyperus Esculentus

The feature combinations selected by different band selection methods are modeled around SVM, ELM and Softmax classifiers, and the optimal discrimination model is proposed.

4. Results and Discussions

4.1. Acquisition of Spectral Data of Cyperus Esculentus

The average spectral data is obtained by calculating the average reflectivity of pixels in the sample ROI, as shown in figure 3. The same steps were repeated for three kinds of samples, and the 600-728 spectral data matrix of 600 samples was obtained. The samples were randomly divided into 3:1 training set and test set, of which 150 samples were in the training set and 50 samples were in the test set.

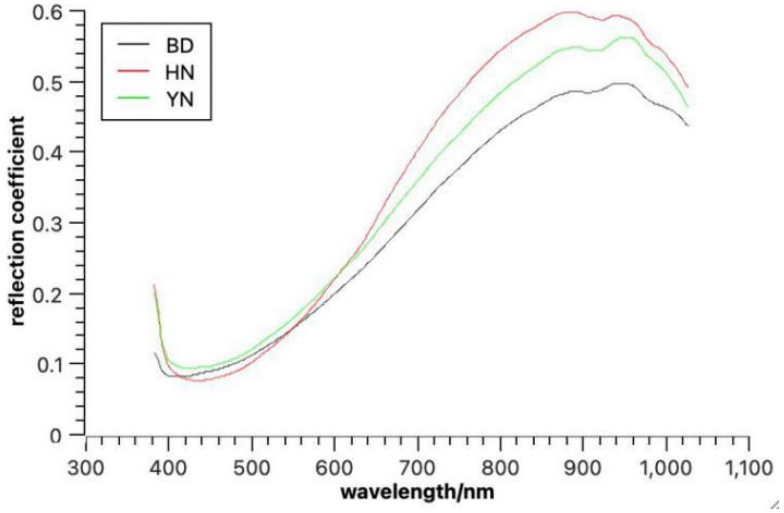

Fig. 3: Average reflectivity curve of three kinds of Cyperus esculentus.

4.2. Pre-processing of Spectral data of Cyperus Esculentus

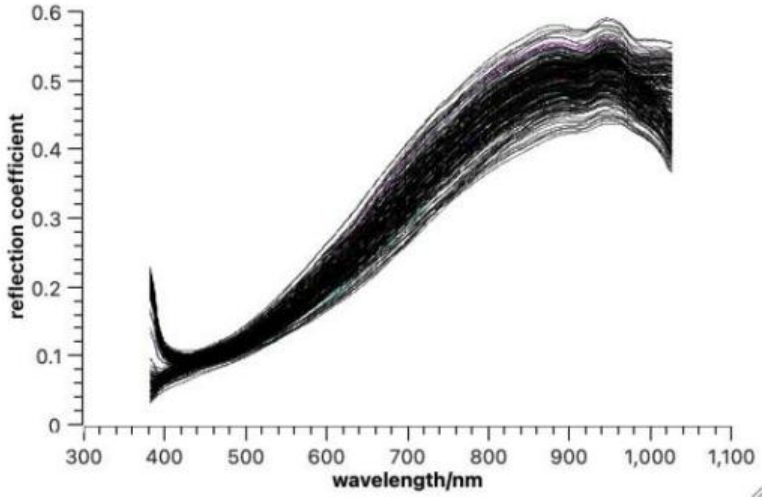

Fig. 4 (a): Reflectivity curve of original spectrum and pretreated spectrum.

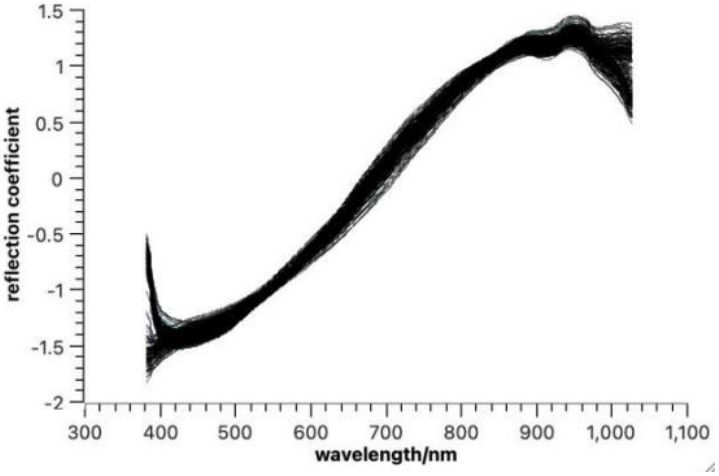

Fig. 4 (b): Reflectivity curve of original spectrum and pretreated spectrum.

In the stage of hyperspectral image data acquisition, the influence of some external factors is reduced by black-and-white correction. The original spectral reflectivity curve is shown in figure 4 (a). The original spectral data of 728D are preprocessed by SG, and the reflectivity curve is obtained as shown in figure 4 (b). However, there are still various interference factors in the obtained spectral data, and these noises will lead to some errors in the predicted values. From the original spectral reflectance in figure 4 (a), it can be seen that the data was greatly disturbed at the beginning and end of the acquisition. Therefore, the first 10 wavelengths and the last 20 wavelengths are manually eliminated in this paper, that is, the spectrum with a wavelength range of 390.63-1009.95nm is selected for analysis.

4.3. Selection of Spectral Band of Cyperus Esculentus

The dimensionality of the preprocessed data is reduced. CARS sets the number of MCS to 50 times. MCS sampling is random. Therefore, repeated iterations are used to compare the RMSECV values to determine the number of characteristic wavelengths. Finally, 74 characteristic wavelengths corresponding to the minimum RMSECV value of 0.403 are determined. When SPA is used for feature selection, the minimum number of feature wavelengths is set to 50, and when the minimum RMSECV value is 0.397, 50 feature wavelengths are obtained.

It can be seen from Table 2 that different dimensionality reduction algorithms can get different dimensionality reduction results. The original 698-dimensional data can be reduced to dozens of dimensions while ensuring the amount of information.

Table 2: Classification results of different models

algorithm	Number of variables	Characteristic wavelength (nm)
CARS	74	401.63 417.73 464.60 465.45 477.45 479.16 480.88 481.73 493.75 495.47 535.13 582.88 585.48 587.23 595.96 612.57 613.45 615.20 616.07 616.95 617.83 618.70 619.58 644.15 645.03 645.90 646.79 647.67 648.55 649.44 650.32 651.20 652.08 664.41 666.17 667.95 668.83 669.71 671.47 674.13 675.01 676.78 693.59 694.47 697.14 698.90 700.69 720.22 722.90 723.78 747.85 748.76 749.65 750.53 826.91 751.44 752.33 753.21 782.78 783.69 785.47 786.38 787.28 827.82 828.72 834.15 873.15 893.17 896.83 921.51 928.84 949.03 958.22 995.14
SPA	50	390.63 393.17 394.85 396.54 399.08 425.38 427.92 438.14 440.70 449.23 482.58 490.32 493.75 498.91 504.07 510.10 516.13 522.16 537.72 585.48 601.20 613.45 648.55 660.89 690.94 717.55 738.94 745.17 754.11 809.77 826.01 835.96 846.82 858.60 876.78 884.07 890.45 900.47 905.04 919.67 924.26 928.84 933.42 948.11 954.54 957.30 989.59 1000.69 1003.46 1021.08

4.4. Comparison of Identification Models of Cyperus Esculentus

The feature combinations selected by different band selection methods are modeled based on SVM, ELM and Softmax classifiers. The results of different models are shown in Table 3.

Table 3: Classification results of different models

Model	Method	Number of variables	Calibration (%)	Validation (%)
SVM	SG +CARS	74	74.7	70.7
	SG+SPA	50	71.9	71.3
	SG	698	70.7	70.1
ELM	SG +CARS	74	81.8	80.3
	SG+SPA	50	82.1	81.1
	SG	698	81.3	79.5
Softmax	SG +CARS	74	85.1	82.2
	SG+SPA	50	86.3	85.1
	SG	698	83.3	81.5

By comparing the experimental results, it is found that: From the prediction accuracy of three classification models combined with different data processing methods, the verification accuracy of Softmax classifier under any combination method is higher than that of the other two models. From the prediction results of two band selection combination methods in different models, the model using dimensionality reduction method can improve the accuracy of the model. The 677 features of the original data can be simplified to dozens of features, which also shows that there is a lot of redundant information in the original data.

The results show that the non-destructive testing model established by SPA dimensionality reduction method is feasible.

5. Conclusion

It can be seen that the Vis-NIR Hyperspectral image can realize the non-destructive acquisition of sample data. Through the combination of theoretical analysis and experimental verification, and through the combination of different dimensionality reduction methods and modeling methods, an effective method for nondestructive identification of Cyperus esculentus is proposed. First of all, SG prepossessing is used to reduce the impact of data acquisition process, and then CARS and SPA are used for band selection to reduce the complexity of the model. Finally, the input features are classified by SVM, ELM and Softmax. Among them, the combined model based on SG, SPA and Softmax has achieved the best results in the current application scene of non-destructive identification of Cyperus esculentus from different producing areas, and the accuracy of verification is 85.1%. The results show that the combined model of SG, SPA and Softmax is effective for non-destructive identification of Cyperus esculentus in China. Band selection instead of full-wavelength modeling based on SPA can effectively improve the accuracy of classification. Modeling based on Softmax can effectively identify Cyperus esculentus from different areas.

6. References

[1] Prince Zichen, Zhang Bing, Guan Yongxiang, et al. Preliminary study on cultivation Physiology of Cyperus esculentus Saline soil. *Soil*. 2017. 49 (6): 1126-1131

[2] Zhang Ming, Wu Chengdong, Geng Anhong, et al. Feasibility Analysis of developing Cyperus esculentus Industry in Saline-alkali Land of Yancheng City. *Modern Agricultural Science and Technology*. 2015 (11): 333,337

[3] Chen Xing, Chen Di, Liu Lei, et al. Analysis of total components of Cyperus esculentus. *Food Science and Technology*. 2009. 34 (03): 165-168

[4] Li Guanghui, Wang Xingjun, Zhang Bin, et al. Study on yield and quality of different Cyperus esculentus varieties planted in Shandong. *Shandong Agricultural Sciences*. 2021. 53 (03): 61-64

[5] Liu Yulan, Wang Xiaoning, Shu Yi, et al. Study on the character and composition of Cyperus esculentus from different producing areas. *China Fats and Oils*. 2020. 545 (08): 125-129

[6] Zhao Yongguo, Guo Ruixing, Luo Lixia, et al. Construction of SRAP fingerprinting and analysis of genetic diversity of Cyperus esculentus. *Journal of Plant genetic Resources*. 2013. 14 (02): 222-225

[7] Liu Tao, Sun Xudong, Liu Yande, et al. Progress in non-destructive testing of crop quality by near-infrared spectroscopy. *Food and Machinery*. 2010. June 26 (3): 161-166

[8] Tian Qiong, Hong Wuxing, Lu Yunyu, et al. Identification of the origin of imported Cyperus esculentus based on X-ray fluorescence spectroscopy. *China Port Science and Technology*. 2021. 3 (11): 48-57

[9] Yan Weimin, Liu Gang, Tian Xue, et al. Identification of producing area of quinoa by infrared spectrum . *Chemical agent*. 2022

[10] Zhang Lixin, Zhang Nannan, Zhang Xiao, et al. Discriminant analysis of apple producing area based on machine learning algorithm . *Advances in Laser and Optoelectronics*. 2022. 59 (04): 451-457

[11] Li Shengyang, Liu Zhiwen, Liu Kang, et al. Advances in the application of aerospace hyperspectral remote sensing [J]. Infrared and Laser Engineering, 2019, 48 (03): 9-23.

[12] Liu Jiamin, Yang Song, Huang Hong, et al. Hyperspectral remote sensing image classification based on locally reconstructed Fisher analysis [J]. China Laser, 2020, 47 (07): 390,401.