## IDENTIFICATION OF QUALITY CHARACTERISTICS OF FLUE-CURED TOBACCO BASED ON RAMAN SPECTROSCOPY**

**Feng-Feng Liu [1], Yun-Lan Shen [2], Si-Wen Zhan [3], Yue Wang [4], Yi Mou [1], Shi-Liang Dong [1], Jie-Wang He [1*]**

*[1] China Tobacco Hubei Industrial Co. Ltd., Hubei Wuhan, China; e-mail: 1652193082@qq.com*
*[2] School of Mathematics and Statistics at Zhongnan University of Economics and Law, Hubei Wuhan, China*
*[3] School of Chemistry, Chemical Engineering and Life Science at Wuhan University of Technology, Hubei Wuhan, China*
*[4] Hubei Tobacco Gold Leaf Redrying Ltd. Liability Co., Hubei Xiangyang, China*

*A rapid identification method for flue-cured tobacco quality was proposed based on Raman spectroscopy. Considering the critical quality factors of flue-cured tobacco-like oil content, softness, and glossiness, four statistical methods, random forest, K-nearest neighbor, logistic regression, and partial least squares, can effectively improve the accuracy of quality identification. We randomly collected 149 flue-cured tobacco samples from multiple producing areas in China. After Raman spectroscopy analysis, Savitzky–Golay convolution smoothing and multi-scatter correction were done. The functional groups were analyzed to select characteristic peaks as features for discriminant analysis. The results show that the Raman spectroscopic information can distinguish the quality of flue-cured tobacco with an accuracy greater than 95%, whereas the partial least-squares approach delivers an accuracy of 100%. We conclude that Raman spectroscopy can be considered a vital avenue for identifying the quality of flue-cured tobacco.*

***Keywords:*** *Raman spectroscopy, flue-cured tobacco quality, discriminant analysis.*

## ОПРЕДЕЛЕНИЕ КАЧЕСТВА ТАБАКА ДЫМОВОЙ СУШКИ НА ОСНОВЕ СПЕКТРОСКОПИИ КОМБИНАЦИОННОГО РАССЕЯНИЯ СВЕТА

**F.-F. Liu [1], Y.-L. Shen [2], S.-W. Zhan [3], Y. Wang [4], Y. Mou [1], S.-L. Dong [1], J.-W. He [1*]**

*УДК 535.375.5:663.97*

*[1] China Tobacco Hubei Industrial Co. Ltd., Хубэй, Ухань, Китай; e-mail: 1652193082@qq.com*
*[2] Школа математики и статистики Чжуннаньского университета экономики и права, Хубэй, Ухань, Китай*
*[3] Школа химии, химической инженерии и естественных наук Уханьского технологического университета, Хубэй, Ухань, Китай*
*[4] Hubei Tobacco Gold Leaf Redrying Co. Ltd., Хубэй, Сяньян, Китай*

*(Поступила 11 апреля 2022)*

*Предложен метод быстрой идентификации качества табака дымовой сушки, основанный на спектроскопии комбинационного рассеяния света (КР). С учетом критических факторов качества (содержания табачного масла, мягкости и глянцевитости) четыре статистических метода — случайного леса, k-ближайших соседей, логистической регрессии и частичных наименьших квадратов — могут эффективно повысить точность определения качества. После анализа 149 образцов табака из нескольких районов Китая методом КР-спектроскопии выполнены сглаживание свертки Савицкого–Голея и коррекция множественного рассеяния. Функциональные группы проанализирова-*

**Full text is published in JAS V. 90, No. 1 (http://springer.com/journal/10812) and in electronic version of ZhPS V. 90, No. 1 (http://www.elibrary.ru/title_about.asp?id=7318; sales@elibrary.ru).

*ны для выбора характерных пиков в качестве признаков для дискриминантного анализа. Показано, что КР-спектроскопия позволяет определить качество табака дымовой сушки с точностью >95%, в то время как метод частичных наименьших квадратов обеспечивает точность 100%. КР-спектроскопия является важным методом определения качества табака дымовой сушки.*

***Ключевые слова:*** *спектроскопия комбинационного рассеяния света, качество табака дымовой сушки, дискриминантный анализ.*

**Introduction.** The flue-cured tobacco has substantial economic value as one of the essential economic crops in China. In practice, it is crucial to classify and rank the quality of flue-cured tobacco. The quality of flue-cured tobacco is affected by many factors, such as production area, climate, and processing technology [1, 2]. The appearance and the internal qualities of flue-cured tobacco are closely related. The appearance quality of flue-cured tobacco is often indicated by internal quality [3, 4]. Specifically, the appearance qualities of flue-cured tobacco related to oil content, softness, and glossiness are intricately linked with maturity, physicochemical properties, and gas quality [5]. Usually, professional personnel conduct the appearance grading for the oil content, softness, and glossiness with a visual, tactile, aural, and olfactory inspection. Such classifications for flue-cured tobacco based on empirical perceptions lack objective authenticity and are susceptible to external factors. Furthermore, the implementation of this procedure is time consuming and cost intensive for industrialization.

Raman spectroscopy refers to scattering, in which a substance vibrates after being irradiated by light. Raman spectroscopy can reveal information about the structure and content of the substance under inspection [6]. Raman spectroscopy is widely used in agriculture and animal husbandry [7–9], food [10, 11], chemicals [12, 13], medicine [14, 15], and other industries owing to its wide detection range, high sensitivity, and nondestructive testing of samples.

Recently, spectral technology combined with stoichiometry has been applied to identify flue-cured tobacco varieties and grades. For example, Wang et al. [16] realized the rapid discrimination of flue-cured tobacco aroma type by using visible-near-infrared spectroscopy combined with the principal component analysis and the partial least-squares discriminant analysis; the discrimination accuracy reached 100%. Bin et al. [17] established an automatic identification method for tobacco grading based on near-infrared spectroscopy and the extreme learning machine algorithm. It was found that the extreme learning machine algorithm had the best prediction performance, achieving an overall accuracy of sample classification greater than 90%. Also, Marcelo et al. [18] analyzed standard tobacco bunches through near-infrared imaging and support vector machine-discriminant analysis (SVM-DA). Their results showed that the model's prediction accuracy for tobacco types "smoking Virginia" and "air smoking Burley" was 80.4 and 88.1%, respectively. Additionally, the model's prediction accuracy for the "smoked Virginia" and "air-smoked Burley" types was 95.9 and 96.5%, respectively. The overall prediction accuracy of tobacco quality was between 61.5 and 100.0%. Finally, the prediction accuracy for "air-dried white leaves" was between 78.8 and 100.0% [19] established a nil-CNN model for tobacco leaf area classification with a discrimination accuracy of 95% based on applying a convolutional neural network to NIR spectral data.

Although there have been many discriminant studies on flue-cured tobacco based on near-infrared spectroscopy, this work is aimed at adopting Raman spectroscopy for identifying and classifying flue-cured tobacco and thus providing an essential alternative. In this study, Raman spectroscopy combined with stoichiometry was used to effectively identify different grades as per the oil content, softness, and gloss intensity of flue-cured tobacco.

**Materials and methods.** A total of 149 tobacco samples were collected from multiple areas such as Hubei, Henan, Hunan, and others. The main varieties include Yunyan 87, K326, Longjiang 911, and Hongda. The grade distribution is mainly B2F upper limit, B2F lower limit, C2F upper limit, and C2F lower limit. More details are shown in Fig. 1. All tobacco samples were collected from 2019 to 2020 and then redried by China Tobacco Hubei Industry Co., Ltd. to make flue-cured tobacco.

According to the current national flue-cured tobacco grading standards, an officer responsible for grading scored the oil content, softness, and glossiness of 149 test flue-cured tobacco samples. The total oil content score was 20 points, the softness score was 18 points, and the glossiness score was 10 points. Based on scores, flue-cured tobacco was divided into class 1, class 2, and class 3 with respect to oil content, softness, and glossiness, reflecting the different grades of flue-cured tobacco quality. The specific classification criteria and the distributions are shown in Table 1.

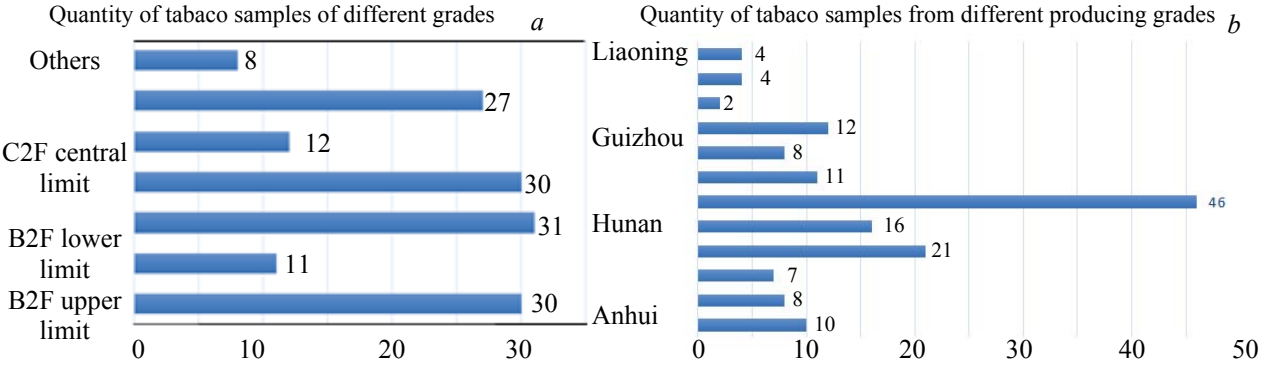


Fig. 1. Information on different production areas and grades of tobacco.

TABLE 1. Classification Criteria and Distributions of Flue-Cured Tobacco Samples

| Class | Oil content | | Softness | | Glossiness | |
|---|---|---|---|---|---|---|
| | Score | Sample size | Score | Sample size | Score | Sample size |
| 1 | 18–20 | 20 | 13–18 | 84 | 9–10 | 46 |
| 2 | 15–17 | 57 | 7–12 | 51 | 7–8 | 77 |
| 3 | 1–14 | 72 | 1–6 | 14 | 1–6 | 26 |

Nexus intelligent Fourier transform infrared spectrometer was used in the experiment. This instrument is equipped with an Nd:$YVO_4$ laser that has a 1064-nm wavelength as the excitation source. The laser power is 0.3 W. The scanning beam is within the range 1/3000 to 1/100. Each sample is scanned 400 times. The Raman spectra of various samples are shown in Fig. 2.

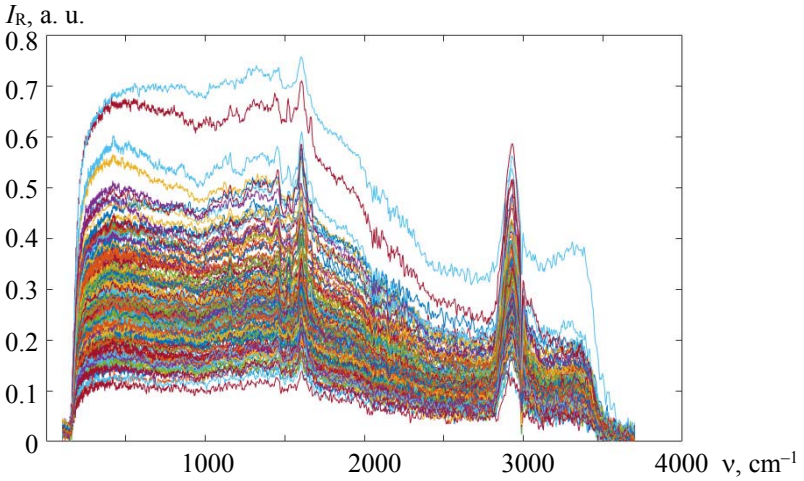


Fig. 2. Original Raman spectra.

Matlab R2016a was used to preprocess the original Raman spectra, the Savizky–Golay (SG) convolution smoothing, and multivariate scattering correction. The random forest (RF), *K*-nearest neighbor (KNN), and logistic regression (LR) algorithms used for discriminant analysis were performed in Python 3.7, and the partial least-squares (PLS) discrimination analysis was performed in R 4.0.2. Samples for each class were randomly divided into training and test data sets in a ratio of 7:3, and the value of accuracy was calculated for the evaluation of discriminations, which is the proportion of the number of correctly discriminated samples to the total number of samples.

**Results and discussion.** *Raman spectroscopy preprocessing.* The spectrum data are often affected by many factors, such as the stability of the spectrum collection instrument, electrostatic noise, sample shape background, and light scattering. In Fig. 2, the Raman spectra of flue-cured tobacco of different quality types have similar trends and peaks, but the signal-to-noise ratio is low. Therefore, it is not easy to analyze their

characteristics and classification directly. In this study, the SG convolution smoothing on the original flue-cured tobacco Raman spectrum is implemented to reduce noise interference. The selection of smoothing window width and polynomial order in SG convolution smoothing have an incredible effect on smoothing results. For a single sample, for example, smoothing results for different window widths and cubic polynomials are shown in Fig. 3a. Finally, a window width of 31 ($W$ = 31) with the cubic polynomial was selected to smooth all the raw Raman spectra, and the results after smoothing are shown in Fig. 3b.

The multivariate scattering correction (MSC) can effectively eliminate the light scattering caused by physical factors such as uneven distribution of samples, different particle sizes, and ambient temperatures. The multivariate scattering correction can significantly improve the spectral signal-to-noise ratio. This study establishes the discriminant models by the Raman spectral data processed with multivariate scattering correction. It is based on using the linear least-squares technique to fit the linear model between a reference spectrum, which is generally the mean of the data set and other spectra of the data set. A corrected spectrum is constructed by changing the scale and the offset of the sample spectrum to get as close to the reference spectrum as possible. The Raman spectrum after processing with multivariate scattering correction for each class is shown in Fig. 3.

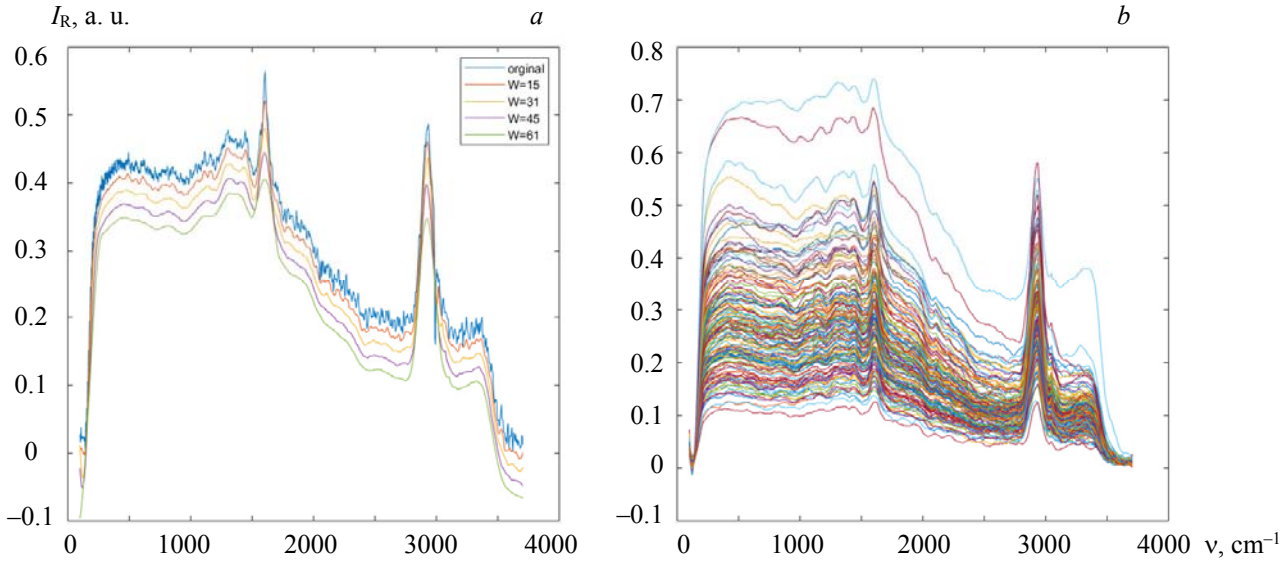


Fig. 3. Process of smoothing the raw Raman spectra for different window widths.

*Extraction and analysis of characteristic wavelengths.* We collected 934 spectral data points for each sample in the Raman spectra. It is challenging to build a model efficiently with these data points. Therefore, the preprocessing of the spectral data is necessary. Extracting the characteristic wavelengths of the spectra can significantly reduce the complexity of the original high-dimensional data, speed up the modeling speed, and increase the robustness of the model.

Raman spectra often contain much information about functional groups at the position of characteristic peaks, reflecting the chemical composition and chemical content in flue-cured tobacco to a certain extent. The tobacco leaves of different quality levels have different chemical substances and contents. Therefore, their corresponding vibrational spectra will also be different [20] and selecting only the specific Raman characteristic peaks in each category is necessary for dimensionality reduction. The characteristic wavelengths selected based on the oil content, softness, and glossiness are shown in Table 2.

There are 16 selected characteristic wavelengths for each category of oil content, softness, and glossiness. The extracted Raman characteristic wavelengths in each category seem to overlap. However, there are some differences between them in terms of peak shape, intensity, and width. In the oil content spectra, there are normal peaks at 1599, 2104, 2200, and 2933 cm$^{-1}$. However, at 2933 cm$^{-1}$, the intensities of the Raman spectra for the low oil content are stronger than that for the high and medium oil content.

TABLE 2. Extracted Characteristic Wavelengths Based on Oil Content, Softness, and Glossiness

| Class | Selected wavelengths | No. |
|---|---|---|
| *Oil content* | | |
| 1 | 411, 519, 596, 843, 1136, 1290, 1344, 1429, 1599, 2104, 2200, 2763, 2933, 3045, 3242, 3342 | 16 |
| 2 | 426, 511, 596, 850, 1159, 1298, 1344, 1429, 1599, 2104, 2200, 2748, 2933, 3045, 3226, 3327 | 16 |
| 3 | 411, 519, 580, 827, 1159, 1290, 1336, 1437, 1599, 2104, 2200, 2756, 2933, 3045, 3203, 3350 | 16 |
| Total | 411, 426, 511, 519, 580, 596, 827, 843, 850, 1136, 1159, 1290, 1298, 1336, 1344, 1429, 1437, 1599, 2104, 2200, 2748, 2756, 2763, 2933, 3045, 3203, 3226, 3242, 3327, 3342, 3350 | 31 |
| *Softness* | | |
| 1 | 411, 519, 588, 835, 1128, 1282, 1344, 1437, 1599, 2104, 2200, 2740, 2933, 3045, 3242, 3342 | 16 |
| 2 | 418, 519, 580, 850, 1151, 1290, 1336, 1429, 1599, 2104, 2200, 2733, 2933, 3045, 3242, 3330 | 16 |
| 3 | 426, 519, 596, 850, 1151, 1298, 1344, 1429, 1599, 2104, 2193, 2744, 2933, 3045, 3276, 3319 | 16 |
| Total | 411, 418, 426, 519, 580, 588, 596, 835, 850, 1128, 1151, 1282, 1290, 1298, 1336, 1344, 1429, 1437, 1599, 2104, 2193, 2200, 2733, 2740, 2744, 2933, 3045, 3242, 3276, 3319, 3330, 3342 | 32 |
| *Glossiness* | | |
| 1 | 426, 519, 596, 850, 1136, 1290, 1344, 1429, 1599, 2104, 2200, 2756, 2933, 3045, 3242, 3342 | 16 |
| 2 | 426, 519, 596, 835, 1144, 1290, 1336, 1429, 1599, 2104, 2196, 2763, 2933, 3045, 3234, 3319 | 16 |
| 3 | 418, 534, 627, 820, 1151, 1290, 1340, 1437, 1599, 2104, 2200, 2748, 2933, 3045, 3203, 3334 | 16 |
| Total | 418, 426, 519, 534, 596, 627, 820, 835, 850, 1136, 1144, 1151, 1290, 1336, 1340, 1344, 1429, 1437, 1599, 2104, 2196, 2200, 2748, 2756, 2763, 2933, 3045, 3203, 3234, 3242, 3319, 3334, 3342 | 33 |

It may be due to more alkanes in low-oil flue-cured tobacco than in high or medium oil content. The peak at 1599 $cm^{-1}$ is presumed to be from the C=C skeleton vibration of nicotine. The peaks at 2104 and 2200 $cm^{-1}$ are generally because of the vibrations of alkynes. The Raman peak at 2933 $cm^{-1}$ is due to the C-H vibrations of alkanes, most probably the straight-chain alkanes such as *n*-eicosane and *n*-tetracosane. All three kinds of flue-cured tobacco containing high, medium, and low oil contain nicotine and straight-chain alkanes. Nicotine itself has a unique aroma of tobacco at high-temperature decomposition, and it can produce a variety of tobacco resin fragrances. Consequently, it closely relates to the oil content of flue-cured tobacco and affects its color, aroma, and taste.

The 16 characteristic peaks at 411, 519, 596, 843, 1136, 1290, 1344, 1429, 1599, 2104, 2200, 2763, 2933, 3045, 3242, and 3342 $cm^{-1}$ wavelengths were extracted from high-oil tobacco leaves. Peaks 411, 519, and 596 $cm^{-1}$ are supposed to be from the skeletal stretching vibrations of reducing sugar, 843 $cm^{-1}$ resulting from C-C vibrations of the volatile component of the benzyl alcohol branch chain, 1136 $cm^{-1}$ originates from C-C ring vibrations of nicotine. Moreover, peaks at 1344 and 1429 $cm^{-1}$ are assigned to C-C ring vibrations of pinane and 3045 $cm^{-1}$ to C=C-H vibrations of cibai trienediol. Finally, the peak occurring at 3242 and 3342 $cm^{-1}$ may be attributed to C-OH vibrations of cibai trienediol or benzyl alcohol. The main components of flue-cured tobacco oil are nicotine, cipertrienol, pinane, and benzyl alcohol.

*Discriminant analysis with Raman spectroscopy.* Currently, two main discrimination methods are based on Raman spectroscopy data. The first discrimination method contains linear methods, such as principal component analysis (PCA) and PLS. The second discrimination method comprises nonlinear methods such as KNN and BP neural network. This paper presents the combination of RF, KNN, LR, and PLS discrimina-

tion analyses to identify the class of oil content, softness, and glossiness of flue-cured tobacco. Selecting the optimal model provides rapid and accurate identification of flue-cured tobacco quality.

The RF method is an extended model based on decision trees and self-sampling integration. A decision tree is a tree structure that can manage classification problems. For an input sample, each decision tree is a classifier. Hence, *N* trees will have *N* classifiers. The RF method integrates all classification results; the most frequent category is output as the final result. During high-dimensional data processing, the model can select feature subsets randomly, making its training speed faster. For unbalanced data, it can effectively balance errors. However, when the sample data are too noisy, over-fitting can occur [21].

The KNN method finds the sample category by voting on the categories of the *K* sample points closest to the sample in the feature space. Generally, the category with the most votes is regarded as the final category of the sample. The *K* value is usually determined by cross-validation as a critical factor of prediction accuracy. The model is simple to use, fast to train, and highly tolerant of outliers. However, when the sample distribution is unbalanced, the model prediction error is relatively large [22].

LR is a machine-learning algorithm mainly used to solve two classification problems, but its extension can be used to solve multiple classification problems. It uses the logistic function to measure the relationship between the predicted label and multiple feature variables and the maximum likelihood method to estimate the model's parameters [23]. It does not need to scale the input features or require too much calculation, so the model is very efficient. However, the data need to be linearly separable, and for nonlinear features, some transformations of the data are required.

PLS discrimination analysis is a linear discriminant method based on partial least-squares regression. It combines the advantages of principal component analysis and canonical correlation analysis. PLS discrimination analysis eliminates redundant information in the matrix by decomposing the spectrum matrix and the concentration matrix, thereby obtaining variables with solid explanations. It can also process high-dimensional data [24] and is suitable for situations where the correlation between variables is strong and the noise is considerable. It has good robustness and solid predictive ability.

*Discrimination results and discussion.* As mentioned above, samples for each class are randomly divided into training and test sets with a ratio of 7:3. Therefore, different samples in the training and test sets cause slight differences in the discrimination results. In order to ensure the stability of the model, the experiment is repeated five times for each model, and the discriminant accuracy obtained in each experiment is recorded separately. Then, the average is used to measure the final discrimination result of the model.

As shown in Table 3, when Raman spectral data are used with KNN, RF, LR, and PLS discrimination analysis to identify the class of tobacco oil content, the discriminant accuracy of oil obtained by SG preprocessing is about 40–50%. However, with SG & MSC preprocessing, the overall accuracy of four models on the training and test sets is greater than 95%. Moreover, the accuracy of each class is more than 90%. However, the discrimination accuracy of class 1 samples is slightly lower than that of other samples in the training set and the test set, mainly because of the small number of class 1 tobacco samples, which only accounted for 13.4% of the total number of samples. After comparing the classification results of the above four models, it can be found that Raman spectra data with PLS discrimination analysis gives the best results for oil content class identification, reaching 100%.

From the discrimination results of the softness in Table 3, the Raman spectral data with four models can identify the softness of flue-cured tobacco. The discriminant accuracy of softness obtained by SG preprocessing is about 50–60%. However, for SG & MSC preprocessing, the overall accuracy for the training and the test sets is greater than 95%. The accuracy of most of the classes remains above 85%. The accuracy of class 3 in the test set is only 55% in the logistic regression model. It is because there are only 4 samples in the test set of class 3. Hence, it has little influence on the overall accuracy. It is also evident that partial least-squares discrimination analysis with the Raman spectra data yields an accuracy of 100% for softness identification.

Table 3 shows the classification results for the glossiness of flue-cured tobacco with the Raman spectra data and four models. The discriminant accuracy for glossiness obtained by SG preprocessing is about 45–55%. However, with SG & MSC preprocessing, training accuracy goes up to 100%. For the test set, the overall accuracy of the logistic model reached 99%, and the remaining models reached 100%. Therefore, the Raman spectroscopy combined with these four types of models can effectively distinguish the class of glossiness of flue-cured tobacco.

TABLE 3. Discrimination Results of Oil Content, Softness, and Glossiness

| Data preprocessing | Model | Training set | | | | Test set | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Class 1 | Class 2 | Class 3 | Three overall | Class 1 | Class 2 | Class 3 | Three overall |
| *Oil content* | | | | | | | | | |
| SG & MSC | RF | 1.00 | 1.00 | 1.00 | 1.00 | 0.90 | 1.00 | 1.00 | 0.99 |
| | KNN | 0.96 | 1.00 | 1.00 | 0.99 | 0.97 | 1.00 | 1.00 | 1.00 |
| | LR | 0.99 | 1.00 | 1.00 | 1.00 | 0.90 | 0.99 | 1.00 | 0.98 |
| | PLS | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| SG | RF | 0.87 | 0.97 | 0.98 | 0.96 | 0.03 | 0.43 | 0.52 | 0.42 |
| | KNN | 0.37 | 0.63 | 0.69 | 0.63 | 0.03 | 0.36 | 0.52 | 0.39 |
| | LR | 0.00 | 0.12 | 0.95 | 0.50 | 0.00 | 0.08 | 0.87 | 0.46 |
| | PLS | 0.00 | 0.19 | 0.88 | 0.50 | 0.00 | 0.14 | 0.87 | 0.47 |
| *Softness* | | | | | | | | | |
| SG & MSC | RF | 1.00 | 1.00 | 1.00 | 1.00 | 0.98 | 1.00 | 1.00 | 0.99 |
| | KNN | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.95 | 1.00 |
| | LR | 1.00 | 1.00 | 0.88 | 0.99 | 1.00 | 0.98 | 0.55 | 0.95 |
| | PLS | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| SG | RF | 1.00 | 0.98 | 0.90 | 0.97 | 0.76 | 0.29 | 0.05 | 0.53 |
| | KNN | 0.93 | 0.54 | 0.22 | 0.73 | 0.89 | 0.32 | 0.00 | 0.60 |
| | LR | 0.90 | 0.43 | 0.00 | 0.65 | 0.84 | 0.30 | 0.00 | 0.57 |
| | PLS | 0.90 | 0.24 | 0.00 | 0.59 | 0.90 | 0.23 | 0.00 | 0.59 |
| *Glossiness* | | | | | | | | | |
| SG & MSC | RF | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| | KNN | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| | LR | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.98 | 1.00 | 0.99 |
| | PLS | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| SG | RF | 0.93 | 0.99 | 0.91 | 0.95 | 0.21 | 0.71 | 0.20 | 0.46 |
| | KNN | 0.85 | 0.86 | 0.51 | 0.80 | 0.21 | 0.81 | 0.07 | 0.49 |
| | LR | 0.21 | 0.95 | 0.17 | 0.59 | 0.03 | 0.85 | 0.05 | 0.45 |
| | PLS | 0.02 | 0.99 | 0.09 | 0.53 | 0.00 | 0.96 | 0.03 | 0.52 |

In conclusion, the PLS discrimination method performs best among the other four models when using the Raman spectroscopy to discriminate the class of oil content, softness, and glossiness of flue-cured tobacco. Therefore, compared with manual classification, the Raman spectroscopy combined with partial least-squares discrimination analysis can accurately identify flue-cured tobacco quality.

**Conclusions.** In this study, 149 flue-cured tobacco samples were collected from different regions in China, in which oil content, softness, and glossiness were taken as the critical quality factors. Based on the Raman spectroscopy, a rapid quality identification method for flue-cured tobacco quality was established. First, through the analysis and comparison, the raw Raman spectra of all samples were very similar, suggesting the similarity of the major chemical components of flue-cured tobacco. Owing to the lousy signal-to-noise ratio, it was difficult to directly analyze its characteristics and identification features. Second, the Savitzky–Golay convolution smoothing and multivariate scattering correction were applied to reduce noise.

Moreover, to reduce the complexity of the high-dimensional data and speed up the algorithm, we selected the relevant Raman characteristic peaks of the samples for each class as the characteristic features through the analysis of the functional groups. We found differences in peak shape, intensity, and width at the extracted Raman characteristic wavelengths. For example, in the spectrum based on oil content, the Raman spectrum intensity at 2933 $cm^{-1}$ for low oil content is more substantial than that for high and medium oil content. Third, four statistical approaches, random forest, *K*-nearest neighbor, logistic regression, and partial least squares, were adopted for discriminant analysis. The results showed that the Raman spectroscopic information could distinguish the quality of flue-cured tobacco very well, with the overall accuracy of the four discrimination methods reaching more than 95%. However, owing to the unbalanced distribution of samples, the accuracy of some classes was slightly lower. From the perspective of accuracy for classification, it is

recommended to use the partial least-squares approach with the Raman spectral data to identify the quality of flue-cured tobacco.

**REFERENCES**

1. Y. Liu, J. Que, L. J. Yu, et al., *Chin. Agric. Sci. Bull*., **29**, No. 22, 83–89 (2013).
2. Q. Bao, Y. Zhang, A. G. Wang, et al., *Tobacco Sci. Technol*., **48**, No. 7, 14–19 (2015).
3. G. F. Li, *Study on the Material Basis of Tobacco Oil*, Zhengzhou University (2015).
4. H. L. Jiang, K. Peng, Y. Zhang, et al., *Chin. Tobacco Sci*., **40**, No. 2, 80–86 (2019).
5. China Tobacco Production, Purchase and Marketing Corporation. Training Materials for Mational Standards of Flue-cured Tobacco Grading, Beijing, China Standards Press, 39–40 (2005).
6. W. T. Wang, H. Zhang, Y. Yuan, et al., *AAPS Pharm. Sci. Tech.*, **19**, No. 7, 2921–2928 (2018).
7. J. N. Chen, S. Ye, D. M. Dong, *Laser J*., **39**, No. 9, 25–29 (2018).
8. B. Ranjan, Y. Saito, P. Verma, *Appl. Phys. Express*, **9**, No. 3, 032401 (2016).
9. R. Galli, G. Pressse, C. Schnabel, et al., *PloS One*, **13**, No. 2, 334 (2018).
10. Y. L. Jy, H. P. Jin, M. Hyoyoung, et al., *Food Chem*., **254**, 109–114 (2018).
11. M. M. Hassan, M. Zareef, Y. Xu, et al., *Food Chem*., **344**, 128652 (2020).
12. C. K. A. Nyamekye, J. M. Bobbitt, Q. Zhu, et al., *Anal. Bioanal. Chem*., **412**, No. 24, 6009 (2020).
13. L. Zhang, Application of Raman Spectroscopy in Qualitative and Quantitative Analysis of Raw Materials and Products in Chemical Production, Shanghai: East China University of Science and Technology (2016).
14. R. E. Kast, S. C. Tucker, K. Killian, et al., *Cancer Metastasis Rev*., **33**, No. 2–3, 673–693 (2014).
15. Q. Gao, Z. H. Zhang, F. Lu, *Spectrosc. Spectr. Anal.*, **32**, No. 12, 3258–3261 (2012).
16. Y. D. Wang, M. Q. Zhao, B. Fu, et al., *Chin. Tobacco Sci*., **36**, No. 6, 88–93 (2015).
17. J. Bin, J. H. Zhou, W. Fan, et al., *Acta Tabac. Sin.*, **23**, No. 2, 60–68 (2017).
18. C. A. Marcelo, F. L. F. Soares, J. A. Ardila, C. Jailson et al., *Anal. Methods*, **11**, No. 14, 1–10 (2019).
19. M. Y. Lu, K. Yang, P. F. Song, et al., *Spectrosc. Spectr. Analysis*, **38**, No. 12, 3724–3728 (2018).
20. F. M. Tian, H. Y. Yu, F. Tan, et al., *Oxidation Comm*., **4**, No. 2, 3273–3283 (2016).
21. Breiman Leo, *Machine Learning*, **45**, 5–32 (2001).
22. L. Wang, M. Han, X. J. Li, et al., *IEEE ACCESS*, **9**, 64606–64628 (2021).
23. H. W. Luo, X. B. Pan, Q. S. Wang, et al., *IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*, 916–917 (2019).
24. C. L. Lee, C. Y. Liong, et al., *Analyst*, **143**, No. 15, 3526–3539 (2018).