ACS OMEGA

Open Access

http://pubs.acs.org/journal/acsodf

Article

Development and Application of a Quantitative Model for Proximate and Ultimate Analysis of Flue-Cured Tobacco Based on Near-Infrared Spectroscopy

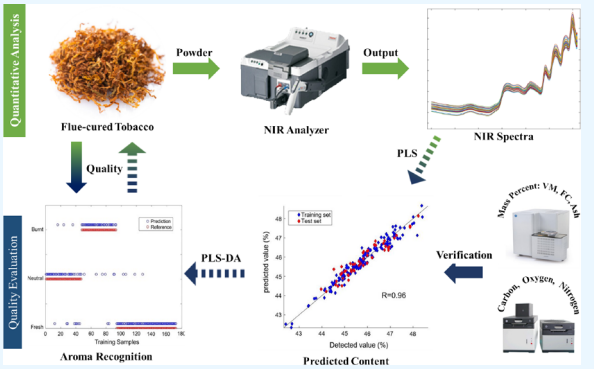
Yuhan Peng, Jiaxu Xia, Qingxiang Li, Yiming Bi, Shitou Li, and Hui Wang*

Cite This: *ACS Omega* 2024, 9, 48196−48204

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: A methodology for predicting proximate and ultimate analysis data was developed by using near-infrared spectroscopy (NIR) combined with chemometric methods. The quantitative model has high accuracy, as evidenced by low root-mean-square-error of prediction (RMSEP) values (e.g., 0.41% for volatile matter and 0.29% for carbon). The model was further applied to tobaccos with distinct aroma profiles, and the predicted ultimate and proximate data lead to aroma classification with 86.6% accuracy. This methodology can be expanded to the aroma discrimination of imported tobaccos from Brazil, the United States, Canada, and Zimbabwe, demonstrating its broad reliability. Compared with traditional analyses, this NIR-based approach offers a fast and accurate method for large-scale tobacco evaluation, highlighting its potential for enhancing tobacco quality characterization through a quantifiable, digital, and high-throughput process.



1. INTRODUCTION

Thermochemical technologies such as combustion, gasification, pyrolysis, and carbonization have been developed to convert coal, biomass, and other organic materials into heat, fuel, and other high-value products.[1,2] The properties of these organic materials vary significantly depending on the species and origins, leading to distinct characteristics during the thermal conversion process. Thus, investigation of the feedstock properties is highly necessary. At present, proximate and ultimate analysis is frequently used to evaluate the fuel quality of coal and biomass for energy applications.[3−5] Proximate data refer to the contents of moisture (M), volatile matter (VM), fixed carbon (FC), and ash (A). Ultimate analysis data include the contents of carbon (C), hydrogen (H), oxygen (O), nitrogen (N), and sulfur (S), which are the main elements of the organic components in feedstock. These indexes are closely related to conversion efficiency, heat production, and product formation in the thermal conversion process. The results of the proximate and ultimate analyses of coal and biomass can offer a guideline for setting and adjusting thermal conversion process parameters. These results can improve the process operation flow according to the composition characteristics of different raw materials.[6−9]

As a special lignocellulose biomass, tobacco mainly undergoes two chemical reactions during smoking, namely, pyrolysis and combustion. Hence, the cigarette itself can be regarded as a miniature reactor where pyrolysis and combustion reaction occurs.[10,11] Proximate and ultimate analysis data have great application prospects in the prediction of smoke generation and combustion state such as temperature field distribution of a combustion cone.[12] Meanwhile, thermochemical processing technologies, especially pyrolysis, have shown potential for the utilization of tobacco wastes. Proximate and ultimate analysis of tobacco feedstock has significant instructions for the construction of this system.

In addition to the application for energy purposes, proximate and ultimate analysis data are also closely associated with the quality of feedstock.[3,13,14] For example, proximate and ultimate data can be used to predict the mechanical durability (MD) and main chemical constituents of biomass.[3,13,14] The results of proximate and ultimate analyses are the most basic parameters for the quality evaluation of coal.[14,15] For example, the contents of VM and element H on a dry, ash-free basis can reflect the degree of coalification and are an important basis for coal classification. The quality of tobacco has many dimensions, including external attributes such as origin, position, and grade as well as internal quality attributes such

Received: June 11, 2024
Revised: September 19, 2024
Accepted: November 19, 2024
Published: November 26, 2024

https://doi.org/10.1021/acsomega.4c05472
ACS Omega 2024, 9, 48196−48204

as aroma and sensory. Its evaluation involves multiple procedures including purchasing, processing, cigarette product design, maintenance, etc., which is of pivotal importance to the tobacco industry.[16,17] Currently, the quality characterization of tobacco mainly relies on manual grading and sensory evaluation. These methods have the disadvantages of strong subjectivity and difficulty in quantification. In fact, the results of proximate and ultimate analyses are essentially data reflecting the overall chemical composition of tobacco, which is the basis of its quality. Therefore, there should be a theoretical correlation between the proximate and ultimate results and tobacco quality. These data have considerable potential for the digital prediction and characterization of tobacco quality. However, the wide variety of tobacco and the high fluctuation of its quality as an agricultural product make it difficult and costly to acquire tobacco proximate and ultimate data, limiting the full exploitation of these data. Establishing an efficient, rapid, and low-cost detection method for proximate and ultimate data is a prerequisite for the in-depth application of these data in the tobacco field.

Near-infrared (NIR) spectroscopy is a valuable tool in tobacco analysis due to its rapid and nondestructive nature, allowing for the comprehensive assessment of chemical compositions without damaging the samples.[18] In recent years, the integration of chemometric techniques with NIR spectroscopy has significantly advanced the field of tobacco analysis. Studies have shown that chemometrics can enhance the accuracy and efficiency of NIR spectral analysis, allowing for a more comprehensive understanding of tobacco's chemical composition.[19−21] For example, an adaptive strategy for selecting representative calibration samples in the continuous wavelet domain was developed to improve the performance of NIR spectral analysis.[19]

Although proximate and ultimate analysis data have been widely used in the fields of coal and biomass, the application of these data in tobacco has not been reported. Herein, we put forward the application of the rapid detection method for proximate and ultimate analysis data based on near-infrared (NIR) spectra to the tobacco field for the first time. Combined with chemometrics methods, the quantitative analysis models of proximate analysis data, namely, VM, FC, A, and the elements C, O, and N in tobacco, were established, realizing the simultaneous, rapid, and accurate analysis of these six parameters used for energy purposes. Among them, C and O are the two elements with the highest content in tobacco, and the content of N is an important index to evaluate the quality of tobacco. The main nitrogen compounds in tobacco, such as nicotine and amino acids, play a decisive role in sensory characteristics. Due to the strong hygroscopic property of tobacco, samples are easily affected by environmental moisture in the process of pretreatment and storage. To avoid the interference of water absorption on the results of proximate including ultimate analysis and ensure that these data can better reflect the characteristics of tobacco, we choose dry basis data to characterize the results of tobacco proximate and ultimate analysis. Furthermore, the potential of proximate and ultimate analysis data in tobacco quality characterization was discussed. Especially, tobacco aroma can be accurately discriminated using these data combined with PLS-DA algorithms. In recent years, pattern recognition algorithms such as support vector machine (SVM),[22−24] k-nearest neighbor (KNN),[23,24] extreme learning machine (ELM),[22] and partial least-squares-discriminant analysis (PLS-DA)[22,25] have been proven to be an effective method for the quality analysis of agricultural products. In this work, PLS-DA algorithms were adapted for its efficacy in classifying samples into predefined groups based on complex data sets, as it combines dimensionality reduction with classification to enhance interpretability.[26]

2. MATERIALS AND METHODS

2.1. Near-Infrared Spectroscopy. All tobaccos used in this work were obtained from the Technology Center of China Tobacco Zhejiang Industrial Co., Ltd. (Hangzhou, China). All tobacco samples were ground and sieved. The tobacco powder with a size ranging from 40 to 60 mesh sieves was collected in a sealed valve bag before subsequent measurements.

NIR diffuse reflectance spectra (1000−2500 nm) were collected from dry powder samples using an Antaris II FT-NIR analyzer (Thermo Fisher Scientific, USA), equipped with an integrating sphere operated at an 8 cm^{-1} resolution (wavenumber range of 10,000−3800 cm^{-1}). All tobacco powders were placed in a rotating cup over a water-free 50 mm diameter quartz window. Instrument performance was verified before analysis using instrumental self-examination. Individual spectra represented an average of 64 scans.

2.2. Proximate and Ultimate Analysis. After the tobacco powder samples were air-dried, their contents of M, VM, FC, and A were determined for 199 samples by an automatic industrial analyzer (5E-MAG6700, Changsha Kai-Yuan Hongsheng Technology Co., Ltd.), according to the national standard GB/T 212-2008. The contents of the element C, H, and N in tobacco powder samples were analyzed by an element analyzer (5E-CHN2200, Changsha Kai-Yuan Hongsheng Technology Co., Ltd.) based on GB/T 476-2008 and GB/T19227-2008. The O element content of the oxazole is calculated by subtracting the content of C, H, N, M, and A from 100%. To avoid the influence of water absorption on the test results during the sample placement and pretreatment, the proximate and ultimate analysis data were converted into the dry base.

2.3. Quantitative Model of Proximate and Ultimate Analysis Data. The quantitative model was developed using the partial least squares (PLS) method based on NIR spectroscopy and proximate and ultimate analysis of 199 samples. Before modeling, the Monte Carlo cross-validation (MCCV) method was performed as the outlier detection step. Then, the data set was randomly divided into a calibration set (149 samples) and a test set (50 samples). The PLS method was used to build a calibration model. Five-fold cross validation was performed for the calibration set to calculate the root-mean-square error of the cross-validation (RMSECV) value. An F test based on the results of cross validation was used to select the optimal number of latent variables. The significance level was set to 0.25 as previously suggested.[27] Prior to building the PLS model, all data were mean-centered.

The root-mean-square error (RMSE) and determination coefficient (R^2) are used as a measure of model performance. RMSE is defined as follows,

$$\mathrm{RMSE} = \sqrt{\frac{\sum (y_{\mathrm{pre}} - y_{\mathrm{ref}})^2}{N}} \quad (1)$$

where y_{pre} is the predicted value, y_{ref} is the actual value, and N is the number of samples. The root-mean-square error of prediction calibration (RMSEC) is the RMSE calculated from

https://doi.org/10.1021/acsomega.4c05472
ACS Omega 2024, 9, 48196−48204

Table 1. Statistics of Proximate and Ultimate Analysis Data for Tobacco Samples

	full set			calibration set			validation set		
parameter	range	mean	SD	range	mean	SD	range	mean	SD
VM (d%)	71.55−75.63	73.93	0.87	71.55−75.63	73.91	0.88	72.24−75.51	73.99	0.84
FC (d%)	14.75−20.36	17.81	1.05	14.75−20.36	17.81	1.08	15.90−20.20	17.81	0.96
A (d%)	6.28−13.59	8.26	1.12	6.28−13.59	8.28	1.17	6.58−10.72	8.20	0.98
C (d%)	42.42−48.42	45.78	1.11	42.42−48.42	45.78	1.15	43.97−48.23	45.79	0.98
O (d%)	35.35−40.64	38.05	1.15	35.66−40.64	38.03	1.16	35.35−40.39	38.11	1.12
N (d%)	1.56−2.95	2.18	0.28	1.66−2.95	2.18	0.27	1.56−2.86	2.16	0.29

the calibration samples, serving as a measure of fit. RMSECV is calculated from the cross-validated samples. The root-mean-square error of prediction (RMSEP) is calculated from test samples.

R^2 is defined as follows,

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (y_{\mathrm{ref},i} - y_{\mathrm{pre},i})^2}{\sum_{i=1}^{N} (y_{\mathrm{ref},i} - \overline{y}_{\mathrm{ref}})^2} \quad (2)$$

where $\overline{y}_{\mathrm{ref}}$ is the mean of the actual values. R^2c is the R^2 calculated from the calibration samples. R^2p is the R^2 calculated from the test samples.

2.4. Classification Model of Aroma Styles. An additional 172 tobacco samples with typical characteristic aromas and 67 imported tobacco samples with unclear aroma labels were selected. The NIR-based model was adapted to predict the proximate and ultimate analysis data for these samples.

The aforementioned 172 tobacco samples with typical characteristic aromas was used to develop a partial least-squares-discriminant analysis (PLS-DA) model to classify the tobacco aroma. According to ecological and sensory evaluations, the aroma of flue-cured tobaccos in China can be divided into three styles: fresh, neutral, and burnt. Based on PLS-DA, a three-dimensional system was performed to represent aroma styles. The labels are defined as fresh [1 0 0], neutral [0 1 0], and burnt [0 0 1]. The aroma type of any sample can be quantified as [*a*, *b*, and *c*], where *a*, *b*, and *c* represent the significance degree of fresh, neutral, and burnt aroma, respectively. The prediction aroma was determined as the maximum values of *a*, *b*, and *c*. The model developed a method for calculating the values of [*a*, *b*, *c*] based on proximate and ultimate analysis data predicted from 172 samples, enabling the classification of the aroma style. In addition, the aroma classification model was then applied to determine the aroma style of 67 imported tobacco samples.

As for the classification model, the accuracy of the training model was reported to evaluate the quality of the PLS-DA model. The accuracy is calculated as

$$\text{accurary} = \frac{N_{\mathrm{cor}}}{N} \times 100\% \quad (3)$$

where N_{cor} is the number of correctly classified samples. This provides a direct measure of how well the model classifies the data.

3. RESULTS AND DISCUSSION

3.1. Analysis of Samples. Table 1 lists the ranges, mean values, and standard deviations (SD) for proximate and ultimate analysis data of 199 tobacco samples. As shown in these reference data, the dry basis contents of VM, FC, and A and the elements C, O, and N in 199 tobacco samples fell within the following ranges: VM (71.55−75.63%), FC (14.75−20.36%), A (6.28−13.59%), C (42.42−48.42%), O (35.66−40.64%), and N (1.66−2.95%). The correlation between these six parameters was also analyzed using Pearson coefficient, as shown in Table 2. The correlation between FC and A is relatively significant with Pearson coefficients of −0.68. As for the relationship between ultimate analysis data, there is a moderate negative correlation of the O element with C and N elements, with the Pearson coefficients of −0.59 and −0.60, respectively. In terms of the relationship between ultimate and proximate data, C was moderately correlated with FC and A, as listed in Table 2.

Table 2. Pearson Correlation Coefficient between Six Parameters

parameter	VM (d%)	FC (d%)	A (d%)	C (d%)	O (d%)	N (d%)
VM (d%)	1.00	−0.33	−0.47	−0.10	0.40	0.26
FC (d%)	−0.33	1.00	−0.68	0.66	0.01	−0.03
A (d%)	−0.47	−0.68	1.00	−0.54	−0.33	−0.18
C (d%)	−0.10	0.66	−0.54	1.00	−0.59	0.53
O (d%)	0.40	0.01	−0.33	−0.59	1.00	−0.60
N (d%)	0.26	−0.03	−0.18	0.53	−0.60	1.00

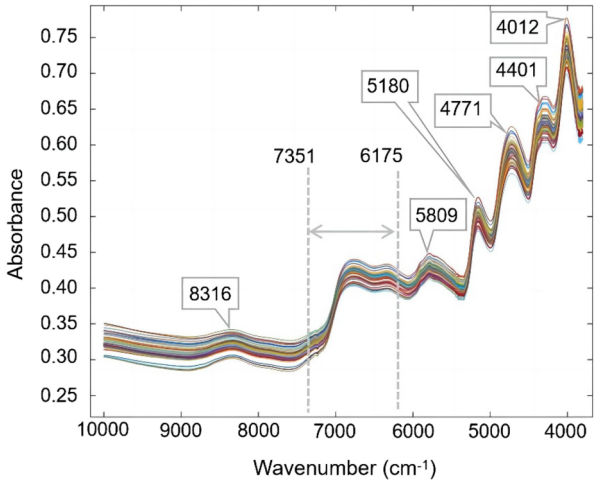


Figure 1. Spectra of the flue-cured tobacco samples.

NIR spectra of the 199 types of tobacco samples are illustrated in Figure 1. Seven major principle bands were presented for all of the samples. The peaks from 6175 to 7351 cm^{-1} corresponded to the 2 × C−H stretching + C−H deformation.[28−31] The peaks at around 8300 and 5800 cm^{-1} were ascribed to the third and the second overtone of C−H bonds, respectively.[28] The peak around 5180 cm^{-1} was attributed to the O−H stretching and H−O−H deformation of moisture.[28] The peaks at 4771 and 4401 cm^{-1} were mainly

https://doi.org/10.1021/acsomega.4c05472
ACS Omega 2024, 9, 48196−48204

Table 3. Result of PLS Models for Determination of the VM, FC, and A Contents in Tobacco

parameter	wavenumber (cm^{-1})	LV	RMSEC	RMSECV	RMSEP	R^2c	R^2p	mean (Y)
VM d%	3800−10000 cm^{-1}	11	0.39	0.69	0.41	0.80	0.76	73.91
FC d%	3800−10000 cm^{-1}	15	0.40	0.70	0.41	0.86	0.82	17.81
A d%	3800−10000 cm^{-1}	13	0.35	0.56	0.32	0.91	0.89	8.28

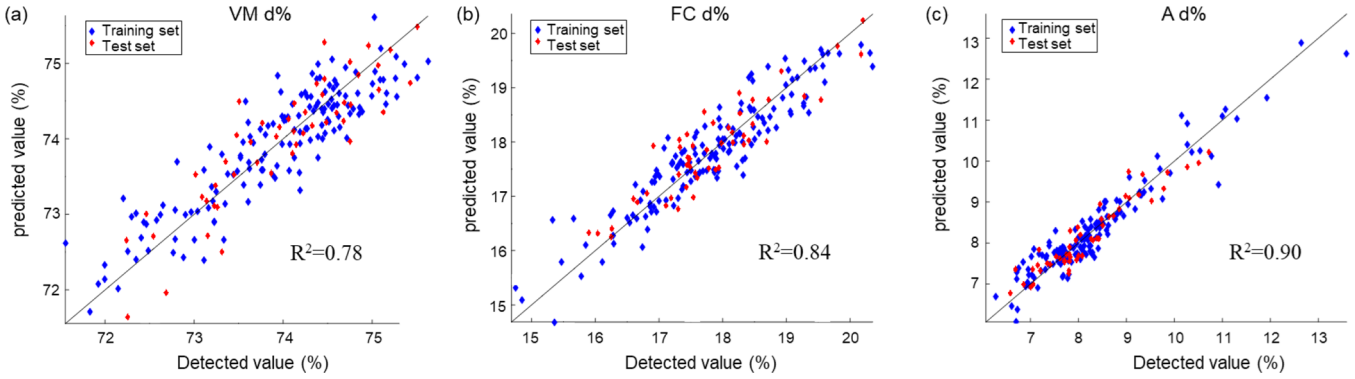


Figure 2. Correlation between the predicted content obtained by the quantitative analysis models and actual one detected by an industrial analyzer of (a) VM, (b) FC, and (c) A.

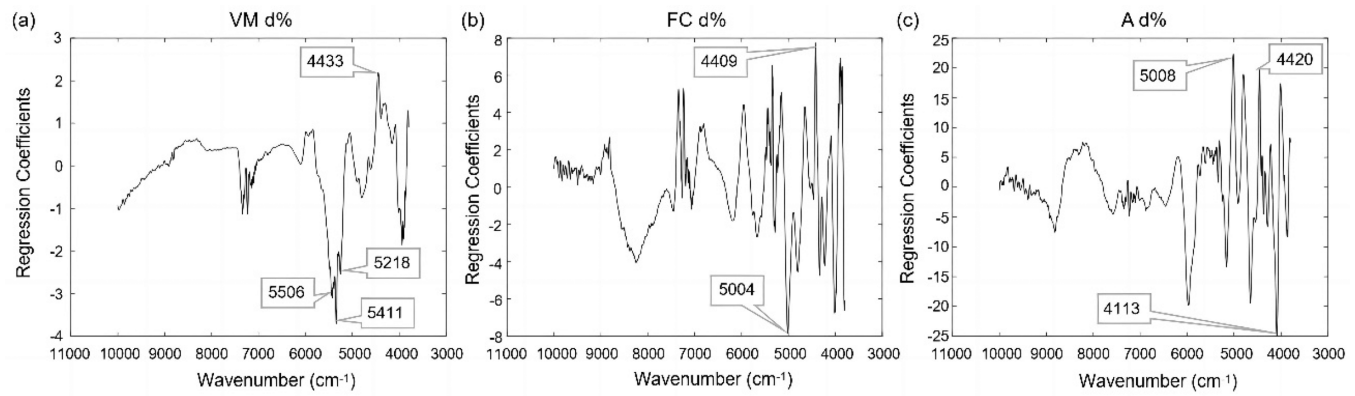


Figure 3. Regression coefficients of the quantitative analysis models for (a) VM, (b) FC, and (c) A.

Table 4. Result of PLS Models for Determination of C, O, and N Contents in Tobacco

parameter	wavenumber (cm^{-1})	LV	RMSEC	RMSECV	RMSEP	R^2c	R^2p	mean (Y)
C d%	3800−10000 cm^{-1}	17	0.25	0.52	0.29	0.96	0.90	45.78
O d%	3800−10000 cm^{-1}	8	0.48	0.70	0.48	0.88	0.84	38.03
N d%	3800−10000 cm^{-1}	16	0.03	0.06	0.05	0.99	0.97	2.18

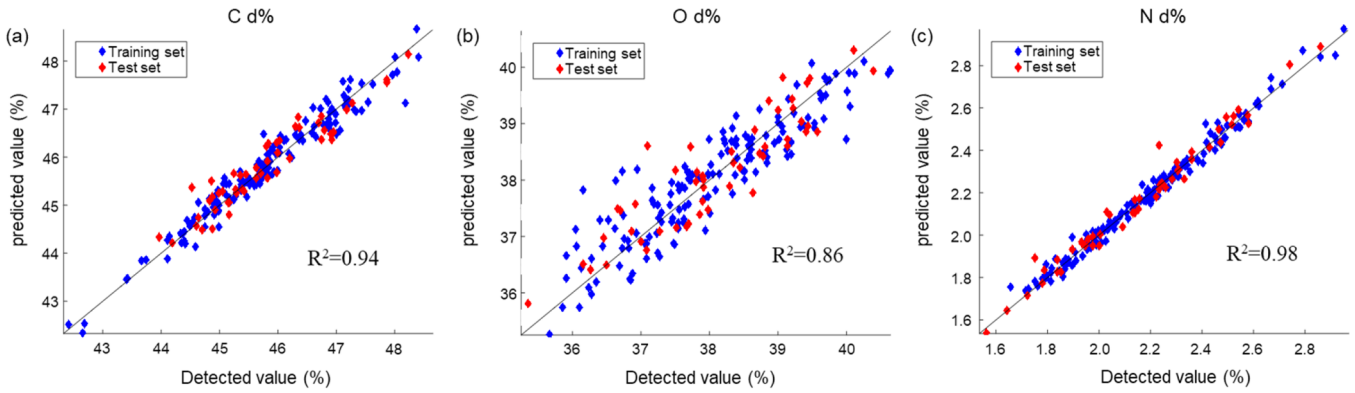


Figure 4. Correlation between the predicted content obtained by the quantitative analysis models and actual one detected by an element analyzer of (a) C, (b) O, and (c) N.

https://doi.org/10.1021/acsomega.4c05472
ACS Omega 2024, 9, 48196−48204

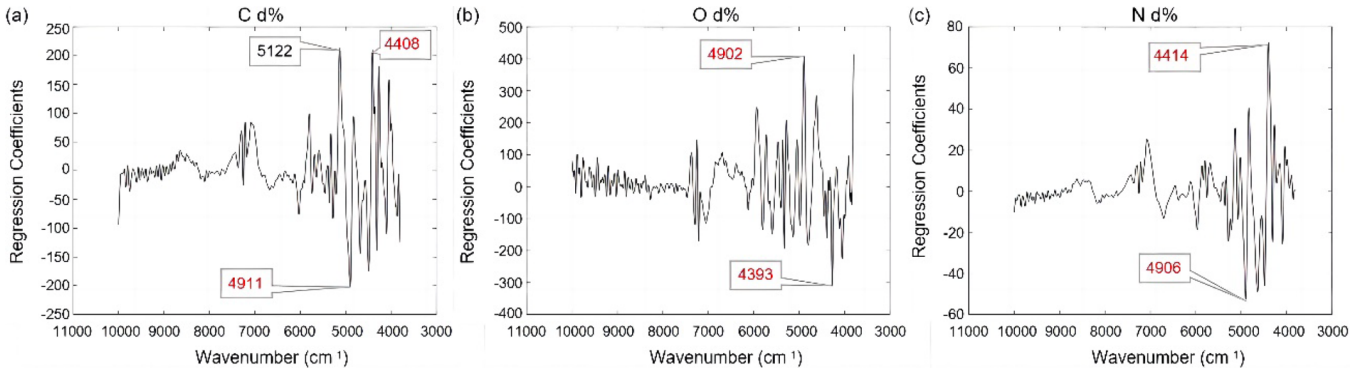


Figure 5. Regression coefficients of quantitative analysis models for (a) C, (b) O, and (c) N.

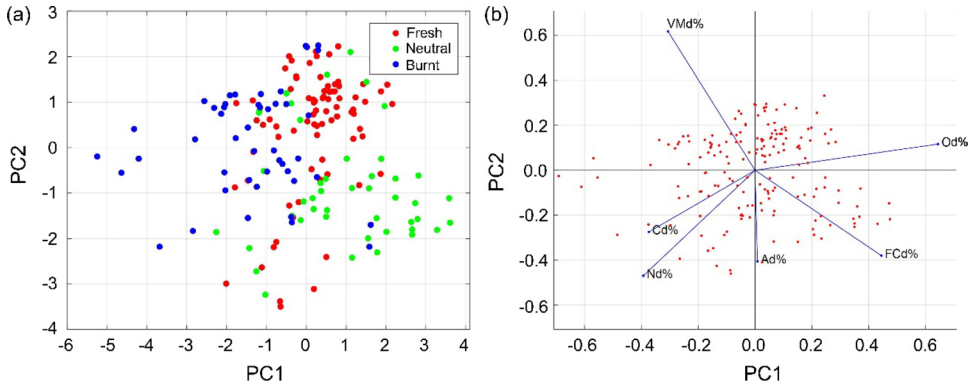


Figure 6. (a) Score plot of the two first PCs for 172 tobacco samples. (b) Orthonormal principal component coefficients for each variable of proximate and ultimate data.

Table 5. Statistical Results (Standard Variation) of 172 Flue-Cured Tobacco Samples

parameter	sample size	statistic	V d%	FC d%	A d%	C d%	O d%	N d%
fresh	79	mean	60.67	16.74	6.35	2.37	33.07	51.10
		std	2.51	0.23	0.34	0.26	0.78	0.63
neutral	47	mean	57.56	16.26	6.98	2.29	33.65	51.11
		std	2.99	0.31	0.37	0.28	1.35	0.76
burnt	46	mean	60.59	15.70	7.51	2.46	31.53	51.07
		std	2.35	0.35	0.43	0.29	1.38	0.61

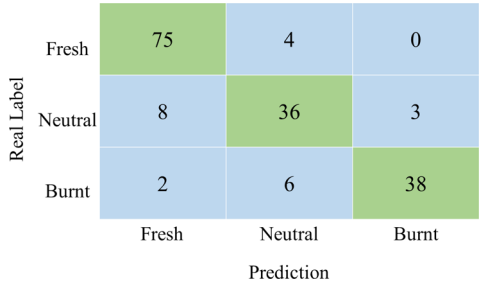


Figure 7. Confusion matrix for aroma recognition results of the model on 172 tobacco samples.

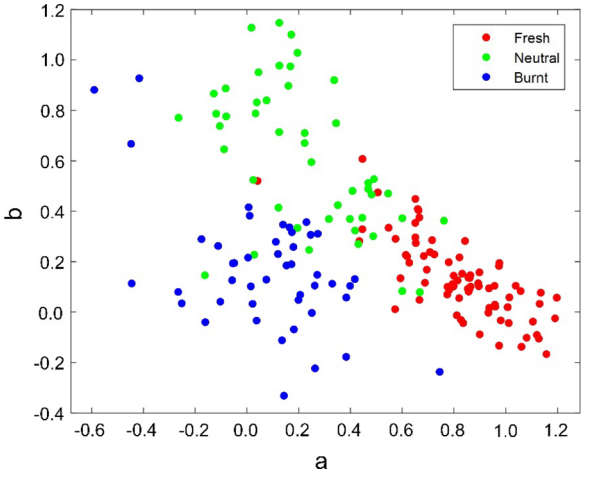


Figure 8. Values of *a* and *b* predicted by the PLS-DA model for 172 tobacco samples.

assigned to 2 × O−H deformation + 2 × C−O stretching and O−H stretching + C−C/C−O stretching, respectively.[29] The absorption peak for the asymmetrical C−O−O stretch in the third overtone of cellulose was also located around 4770 cm^{-1}.[30] The peak at 4012 cm^{-1} corresponded to the C−H stretching + C−C stretching of starch. This peak was also related to C−H stretching and deformation combination of the aromatic structure.[31] Overall, the NIR spectra of 199 tobaccos did not appear to be significantly different.

https://doi.org/10.1021/acsomega.4c05472
ACS Omega 2024, 9, 48196−48204

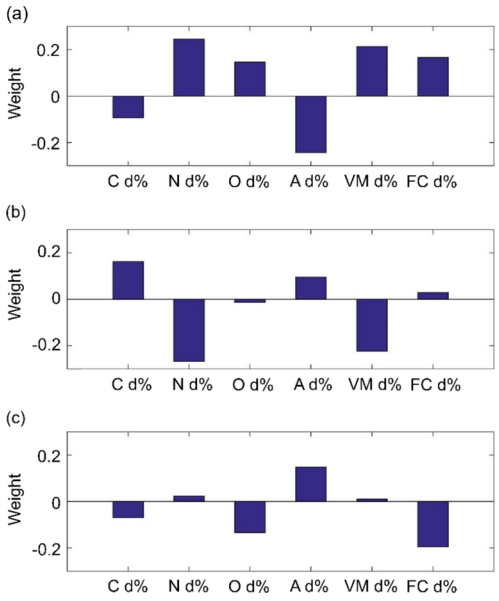


Figure 9. Coefficients of the PLS-DA model for the calculations of (a) *a*, (b) *b*, and (c) *c*.

Table 6. Aroma Recognition Results of the Model for 67 Tobacco Samples

parameter	sample size	statistic	fresh	neutral	burnt	aroma recognition
BZ	30	mean	0.25	0.09	0.66	burnt
		std	0.19	0.21	0.24	
CA	4	mean	1.14	0.07	−0.20	fresh
		std	0.27	0.09	0.19	
US	4	mean	−0.44	0.39	1.05	burnt
		std	0.16	0.13	0.11	
ZW	29	mean	0.12	0.55	0.33	neutral
		std	0.11	0.12	0.13	

3.2. Establishment of Quantitative Analysis Models for Proximate Analysis Data. The 199 tobacco samples were numbered from 1 to 199 and randomly divided into 149 calibration samples and 50 test samples according to the ratio (3:1) of the calibration set to test set. Partial least-squares algorithm written by MATLAB software was used to correlate the contents of VM, FC, and A in 149 calibration samples with the corresponding NIR spectra. The quantitative analysis models of VM, FC, and A of tobacco based on NIR spectrum

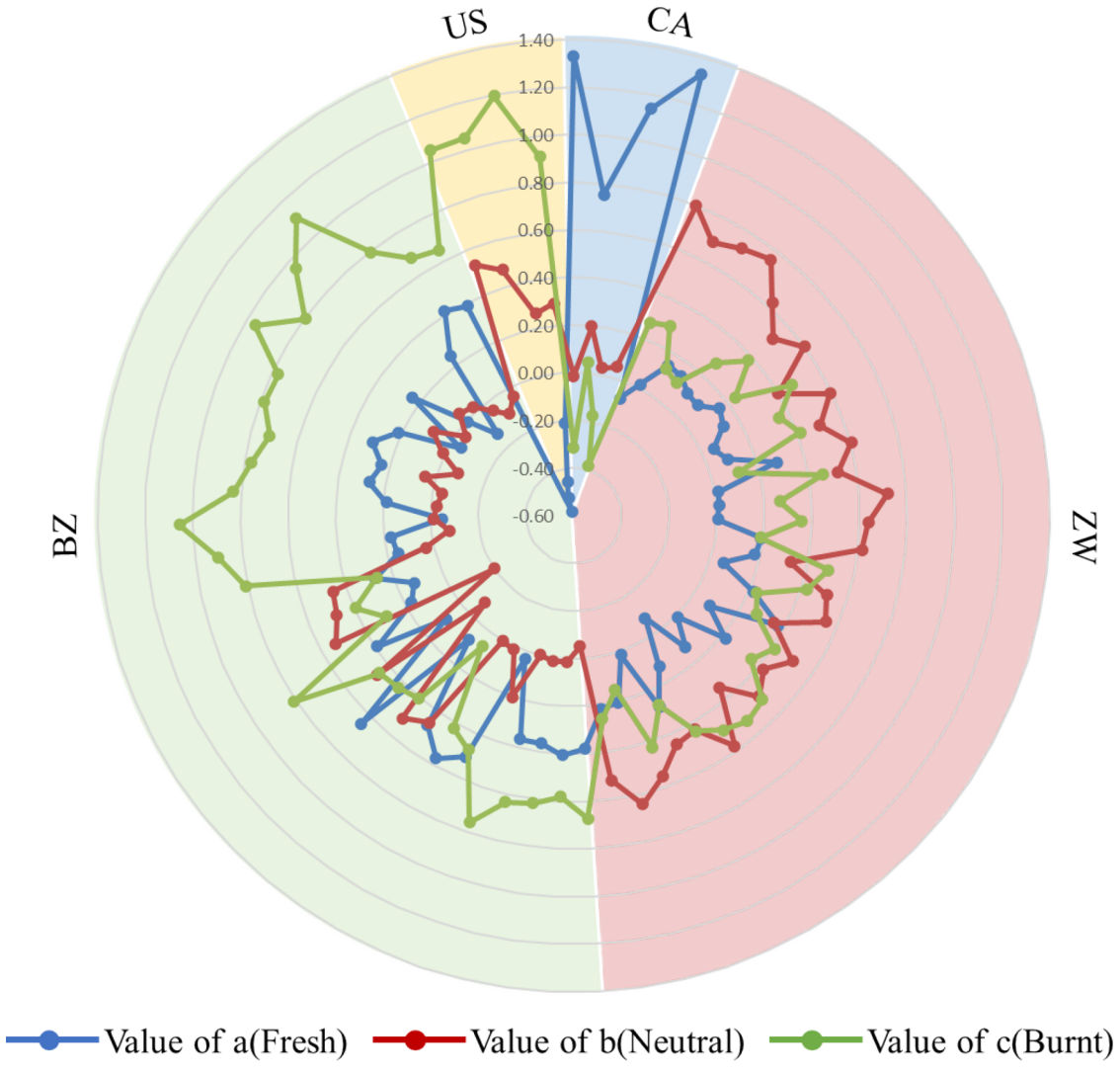


Figure 10. Aroma styles of 67 imported tobaccos recognized by the model.

https://doi.org/10.1021/acsomega.4c05472
ACS Omega 2024, 9, 48196−48204

were finally established via RMSEC and RMSECV. Table 3 shows the wavenumber range for developing the model and the number of latent variables (LV), RMSEC, RMSECV, RMSEP, R^2p and R^2c of the model. The mean contents (mean (Y)) of VM, FC, and A for 199 tobacco samples are also listed in Table 3. Quantitative analysis of these six parameters relied on the full band spectrum, i.e., 11,000−3800 cm^{-1}. NIR spectra were preprocessed with first-order derivative.

Figure 2 shows the scatter plots of VM, FC, and A contents predicted by a quantitative analysis model and measured by an industrial analyzer for calibration and test set samples. Therefore, the established models can be used for the accurate detection of tobacco proximate analysis data. The regression coefficient of the model was used to analyze the contribution of different wavenumbers in the NIR spectra to VM, FC, and A contents, as shown in Figure 3. The contents of VM exhibited strong correlation with the absorption intensity of NIR spectroscopy at 5000−5500 cm^{-1}, namely, 5506, 5411, and 5218 cm^{-1}. This band mainly corresponds to the combination of O−H stretching and other modes of vibration, such as C−H bending and C−O stretching in polyhydric alcohols such as glucose and cellulose.[18] There also existed a strong correlation between the VM content and NIR at 4433 cm^{-1}, which may be attributed to the C−H bending and C−H stretching combination.[32] As shown in Figure 3b,c, the waveband of 3800−6000 cm^{-1} is crucial for the quantitative analysis of both FC and A. Especially, FC contents were strongly correlated with NIR spectroscopy at around 5004 and 4409 cm^{-1}, among which the peak around 5004 cm^{-1} was related to the stretching and deformation of the O−H bands.[33] Ash is mainly composed of a series of inorganic salts. Although inorganic salts themselves have no infrared activity and cannot be characterized by NIR, their content can be detected indirectly by forming chelates with hydrogen-containing organic groups, thus changing the shape of the NIR spectrum. The NIR absorption intensity at wavenumber 4113 cm^{-1} has an important contribution to the quantitative analysis of ash, as shown in Figure 3c.

3.3. Establishment of Quantitative Analysis Models of C, O, and N Elements. In accordance with the same analysis method of proximate analysis data aforementioned, we established quantitative analysis models of C, O, and N elements. Table 4 shows the wavenumber, LV, RMSEC, RMSECV, RMSEP, R^2p, and R^2c of the model. RMSEP values of C, O, and N were 0.29, 0.48, and 0.05, respectively, which were relatively small compared with mean (Y), indicating high accuracy of the quantitative analysis models of C, O, and N. Figure 4 shows the scatter plots of predicted and true values of C, O, and N contents, indicating high accuracy of the quantitative analysis models established for the prediction of C, O, and N contents.

Figure 5 shows the regression coefficients of the quantitative analysis models of C, O, and N based on the NIR spectra. It can be concluded from Figure 5a that wavenumbers from 4000 to 6000 cm^{-1} were optimal for the carbon content prediction model. This range contained a variety of vibrational modes of carbon-containing groups, for example, the combination of C−H stretching and deformation in structures such as CH_2, CH_3, aromatics, −CHO, and the first overtone of C−H stretching.[29,34] Especially, the content of element C was positively correlated with NIR spectra at 5122 cm^{-1}, which is mainly ascribed to second overtone of C=O stretching in $-CO_2H$.[29] As shown in Figure 5a−c, the C, O, and N contents are strongly correlated with the absorption intensity of NIR spectroscopy at around 4900 and 4400 cm^{-1}. The band of 4900 cm^{-1} corresponds to the second overtone of C=O stretching in −CONH−.[34] The peak of 4400 cm^{-1} is mainly related to the combination of C−H bending and C−H stretching; in addition, the N−H stretching and C=O stretching combination of amino acid is also in this position.[34] For different elements, the positive and negative characteristics of the model coefficients in these two positions are different. Specifically, the C and O model coefficients at around 4900 and 4400 cm^{-1} are negative and positive, respectively, which is reversed for N.

3.4. Application Potential of Proximate and Ultimate Analysis Data in Tobacco Quality Characterization. For the tobacco industry, the formula of cigarettes is composed of flue-cured tobaccos with different aroma styles. In China, based on sensory evaluations, the aroma of flue-cured tobaccos can be divided into three styles: fresh, neutral, and burnt.[35,36] The aroma style represents the aroma characteristics of tobacco. The fresh style is characterized by a light, fresh, and green aroma. The burnt style is characterized by heavy, burnt, and nutty aroma. The neutral style presents a smooth aroma that is excluded from both fresh and burnt styles. A practical problem is that the aroma style is determined by only the formulator through sensory evaluation, which is relatively subjective. Especially, for imported flue-cured tobacco, different evaluators have great dispute on the evaluation of its flavor style. At present, there is still a lack of objective methods to evaluate the aroma of imported tobacco in China.

This work focuses on the correlation of proximate and ultimate analysis data with an aroma type of domestic flue-cured tobacco and establishes an aroma classification model based on NIR-predicted proximate and ultimate analysis data. The model can be applied to the aroma identification of imported tobacco from Brazil, the United States, Canada, and Zimbabwe.

Specifically, 172 tobacco samples from China with three aroma styles were collected. The proximate and ultimate information were obtained by the aforementioned models through NIR spectra. Principal component analysis (PCA) was performed to investigate the variability of proximate and ultimate data and search for differentiation and groupings among tobacco with different aroma styles. Figure 6a reports the score plot of the two first PCs for 172 tobaccos.

Figure 6b shows the orthonormal principal component coefficients for each variable, which refers to the six indicators of proximate and ultimate analyses in this work. All six variables are represented in this biplot by a vector. The direction and length of the vector indicate how each variable contributes to the two principal components in the plot.[36] The first principal component (PC1), on the horizontal axis, has positive coefficients for element O, FC, and VM with the vectors directed into the right half of the plot. The largest absolutely value of coefficients in the first principal component is element O, which indicates a larger contribution of O content to PC1. Statistical results of proximate and ultimate data for 172 flue-cured tobaccos are listed in Table 5. Burnt-style tobacco has the lowest average content of the O element. This is consistent with the results in Figure 6, which shows that burnt-style tobacco tends to be distributed on the left side of the score plot, corresponding to lower PC1 values.

On the whole, PCA results showed that there was no obvious regularity in the distribution of 172 samples. Except

https://doi.org/10.1021/acsomega.4c05472
ACS Omega 2024, 9, 48196−48204

for the weak regularity that burnt-style tobaccos were prone to distribute on the more negative side along the PC1 direction, all samples were basically uniformly distributed along the PC2 direction, as shown in Figure 6a.

The PLS-DA model was developed to classify the aroma based on NIR-predicted proximate and ultimate data of tobacco. A three-dimensional system was performed to represent aroma styles. As mentioned in Section 2.4, the labels are defined as fresh [1 0 0], neutral [0 1 0], and burnt [0 0 1]. The aroma type of any sample is quantified as [*a b c*], where *a*, *b*, and *c* represent the significance of fresh, neutral, and burnt aroma, respectively. The prediction aroma was determined as the maximum values of *a*, *b*, and *c*.[37] Using the PLS-DA algorithm, the model developed a method for calculating the values of [*a*−*c*] based on proximate and ultimate analysis data predicted from 172 samples, enabling the classification of the aroma style.

The confusion matrix in Figure 7 shows the classification performance of the model. Each number in the confusion matrix represents the number of samples classified into a specific category; a correctly classified sample appears along the diagonal of the matrix, while misclassified samples appear off-diagonal. For example, there existed 79 fresh aroma-style tobacco samples, among which 75 samples were correctly classified, while the remaining four samples were misclassified as the neutral aroma style. As shown in Figure 7, only 23 samples were predicted with wrong classification. The accuracy of the model was 86.6%. This result is close to the misclassification error of artificial sensory evaluation, indicating that the model had reliable accuracy.

Figure 8 shows the values of (*a*) and (*b*) predicted by the PLS-DA model for each sample. It can be observed that fresh-style tobaccos were almost located on the positive part of horizontal axis while the vertical axis separate the neutral type from fresh and burnt types. Figure 9 shows the regression coefficients of the PLS-DA model for the calculations of *a*, *b*, and *c*. It can be concluded that the fresh type is positively correlated with N, O, VM, and FC and negatively correlated with C and A. The neutral type is positively correlated with C, A, and negatively correlated with N and VM.

The model was then applied to flue-cured tobacco from Brazil (BZ), Canada (CA), the United States (US), and Zimbabwe (ZW). Figure 10 shows the prediction values of aroma parameters a−c for 67 imported tobaccos. Table 6 lists the number of samples from different origins and the corresponding aroma identification results. The tobaccos were classified into burnt, fresh, burnt, and neutral styles, respectively. This result is consistent with the preference and formulation orientation of China Tobacco Industry Company for imported tobacco, providing a reliable quantitative basis and strong proof for the aroma classification of imported tobacco for the first time.

4. CONCLUSIONS

In this study, we developed quantitative models using NIR spectroscopy combined with chemometrics to rapidly and accurately measure the proximate analysis data (e.g., volatile matter, fixed carbon, and ash) and ultimate analysis data (e.g., C, O, and N elements) of tobacco samples. We also found significant correlations between these data and tobacco aroma styles, offering a pathway to more objective quality evaluation. Different from traditional, subjective methods, our approach provides a scientific basis for digital characterization of biomass intrinsic quality. This work supports the shift from experiential to data-driven quality assessment and paves the way for future research in this area.

AUTHOR INFORMATION

Corresponding Author

Hui Wang − *Technology Center, China Tobacco Zhejiang Industrial Co., Ltd, Hangzhou 310012, China*; orcid.org/0009-0006-5286-788X; Email: wxhg2021@126.com

Authors

Yuhan Peng − *Technology Center, China Tobacco Zhejiang Industrial Co., Ltd, Hangzhou 310012, China*

Jiaxu Xia − *Key Laboratory of Refrigeration and Cryogenic Technology of Zhejiang Province, Zhejiang University, Hangzhou 310027, China*

Qingxiang Li − *Technology Center, China Tobacco Zhejiang Industrial Co., Ltd, Hangzhou 310012, China*

Yiming Bi − *Technology Center, China Tobacco Zhejiang Industrial Co., Ltd, Hangzhou 310012, China*; orcid.org/0000-0002-4368-3783

Shitou Li − *Technology Center, China Tobacco Zhejiang Industrial Co., Ltd, Hangzhou 310012, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.4c05472

Author Contributions

Yuhan Peng: conceptualization, methodology, investigation, data curation, writing—original draft. Jiaxu Xia: methodology, investigation, writing—review and editing. Qingxiang Li: methodology, investigation, part of experiments, data curation, visualization. Yiming Bi: visualization, validation, sample pretreatment. Shitou Li: data curation. Hui Wang: conceptualization, resources, visualization.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This research was supported by the Science Foundation of China Tobacco Zhejiang Industrial (grant no. ZJZY2022B006).

REFERENCES

(1) Wang, S.; Dai, G.; Yang, H.; Luo, Z. Lignocellulosic Biomass Pyrolysis Mechanism: A State-of-the-Art Review. *Prog. Energy Combust. Sci.* **2017**, *62*, 33−86.

(2) Patel, M.; Zhang, X.; Kumar, A. Techno-Economic and Life Cycle Assessment on Lignocellulosic Biomass Thermochemical Conversion Technologies: A Review. *Renew. Sustain. Energy Rev.* **2016**, *53*, 1486−1499.

(3) Gillespie, G. D.; Everard, C. D.; Fagan, C. C.; McDonnell, K. P. Prediction of Quality Parameters of Biomass Pellets from Proximate and Ultimate Analysis. *Fuel* **2013**, *111*, 771−777.

(4) Qian, C.; Li, Q.; Zhang, Z.; Wang, X.; Hu, J.; Cao, W. Prediction of Higher Heating Values of Biochar from Proximate and Ultimate Analysis. *Fuel* **2020**, *265*, No. 116925.

(5) Güleç, F.; Pekaslan, D.; Williams, O.; Lester, E. Predictability of Higher Heating Value of Biomass Feedstocks via Proximate and Ultimate Analyses − A Comprehensive Study of Artificial Neural Network Applications. *Fuel* **2022**, *320*, No. 123944.

(6) Skvaril, J.; Kyprianidis, K. G.; Dahlquist, E. Applications of Near-Infrared Spectroscopy (NIRS) in Biomass Energy Conversion Processes: A Review. *Appl. Spectrosc. Rev.* **2017**, *52* (8), 675−728.

https://doi.org/10.1021/acsomega.4c05472
ACS Omega 2024, 9, 48196−48204

(7) Singh, Y. D.; Mahanta, P.; Bora, U. Comprehensive Characterization of Lignocellulosic Biomass through Proximate, Ultimate and Compositional Analysis for Bioenergy Production. *Renew. Energy* **2017**, *103*, 490−500.
(8) Noushabadi, A. S.; Dashti, A.; Ahmadijokani, F.; Hu, J.; Mohammadi, A. H. Estimation of Higher Heating Values (HHVs) of Biomass Fuels Based on Ultimate Analysis Using Machine Learning Techniques and Improved Equation. *Renew. Energy* **2021**, *179*, 550−562.
(9) Zhu, S.; Preuss, N.; You, F. Advancing Sustainable Development Goals with Machine Learning and Optimization for Wet Waste Biomass to Renewable Energy Conversion. *J. Clean. Prod.* **2023**, *422*, No. 138606.
(10) Peng, Y.; Hao, X.; Qi, Q.; Tang, X.; Mu, Y.; Zhang, L.; Liao, F.; Li, H.; Shen, Y.; Du, F.; Luo, K.; Wang, H. The Effect of Oxygen on *in-Situ* Evolution of Chemical Structures during the Autothermal Process of Tobacco. *J. Anal. Appl. Pyrolysis* **2021**, *159*, No. 105321.
(11) Wei, H.; Xing, J.; Luo, K.; Peng, Y.; Fan, J.; Zhang, K.; Wang, H. Predicting Tobacco Pyrolysis Based on Chemical Constituents and Heating Conditions Using Machine Learning Approaches. *Fuel* **2023**, *335*, No. 126895.
(12) Senneca, O.; Chirone, R.; Salatino, P.; Nappi, L. Patterns and Kinetics of Pyrolysis of Tobacco under Inert and Oxidative Conditions. *J. Anal. Appl. Pyrolysis* **2007**, *79* (1), 227−233.
(13) Nimmanterdwong, P.; Chalermsinsuwan, B.; Piumsomboon, P. Prediction of Lignocellulosic Biomass Structural Components from Ultimate/Proximate Analysis. *Energy* **2021**, *222*, No. 119945.
(14) Yan, J.; Lei, Z.; Li, Z.; Wang, Z.; Ren, S.; Kang, S.; Wang, X.; Shui, H. Molecular Structure Characterization of Low-Medium Rank Coals via XRD, Solid State 13C NMR and FTIR Spectroscopy. *Fuel* **2020**, *268*, No. 117038.
(15) Chelgani, S. C.; Mesroghli, Sh.; Hower, J. C. Simultaneous Prediction of Coal Rank Parameters Based on Ultimate Analysis Using Regression and Artificial Neural Network. *Int. J. Coal Geol.* **2010**, *83* (1), 31−34.
(16) Xiang, B.; Cheng, C.; Xia, J.; Tang, L.; Mu, J.; Bi, Y. Simultaneous Identification of Geographical Origin and Grade of Flue-Cured Tobacco Using NIR Spectroscopy. *Vib. Spectrosc.* **2020**, *111*, No. 103182.
(17) Bi, Y.; Li, S.; Zhang, L.; Li, Y.; He, W.; Tie, J.; Liao, F.; Hao, X.; Tian, Y.; Tang, L.; Wu, J.; Wang, H.; Xu, Q. Quality Evaluation of Flue-Cured Tobacco by near Infrared Spectroscopy and Spectral Similarity Method. *Spectrochim. Acta. A. Mol. Biomol. Spectrosc.* **2019**, *215*, 398−404.
(18) Bokobza, L. Near Infrared Spectroscopy. *J. Infrared Spectrosc.* **1998**, *6* (1), 3−17.
(19) Chen, D.; Cai, W.; Shao, X. An Adaptive Strategy for Selecting Representative Calibration Samples in the Continuous Wavelet Domain for Near-Infrared Spectral Analysis. *Anal. Bioanal. Chem.* **2007**, *387* (3), 1041−1048.
(20) Li, Y.; Shao, X.; Cai, W. A Consensus Least Squares Support Vector Regression (LS-SVR) for Analysis of near-Infrared Spectra of Plant Samples. *Talanta* **2007**, *72* (1), 217−222.
(21) Shao, X.; Bian, X.; Cai, W. An Improved Boosting Partial Least Squares Method for Near-Infrared Spectroscopic Quantitative Analysis. *Anal. Chim. Acta* **2010**, *666* (1), 32−37.
(22) Liu, P.; Wang, J.; Li, Q.; Gao, J.; Tan, X.; Bian, X. Rapid Identification and Quantification of *Panax Notoginseng* with Its Adulterants by near Infrared Spectroscopy Combined with Chemometrics. *Spectrochim. Acta. A. Mol. Biomol. Spectrosc.* **2019**, *206*, 23−30.
(23) Chen, Y.; Sun, W.; Jiu, S.; Wang, L.; Deng, B.; Chen, Z.; Jiang, F.; Hu, M.; Zhang, C. Soluble Solids Content Binary Classification of Miyagawa Satsuma in Chongming Island Based on Near Infrared Spectroscopy. *Front. Plant Sci.* **2022**, *13*, No. 841452.
(24) Siano, D. B.; Abdullakasim, W.; Terdwongworakul, A.; Phuangsombut, K. Improving the Performance of the Model Developed from the Classification of Adulterated Honey with Different Botanical Origins Based on Near-Infrared Hyperspectral Imaging System and Supervised Classification Algorithms. *Infrared Phys. Technol.* **2023**, *131*, No. 104692.
(25) Wang, F.; Jia, B.; Dai, J.; Song, X.; Li, X.; Gao, H.; Yan, H.; Han, B. Qualitative Classification of Dendrobium Huoshanense (Feng Dou) Using Fast Non-Destructive Hand-Held near Infrared Spectroscopy. *J. Infrared Spectrosc.* **2022**, *30* (3), 147−153.
(26) Fordellone, M.; Bellincontro, A.; Mencarelli, F. Partial Least Squares Discriminant Analysis: A Dimensionality Reduction Method to Classify Hyperspectral Data. *Stat. Appl.* **2020**, *31* (2), 181−200.
(27) Haaland, D. M.; Thomas, E. V. Partial Least-Squares Methods for Spectral Analyses. 1. Relation to Other Quantitative Calibration Methods and the Extraction of Qualitative Information. *Anal. Chem.* **1988**, *60* (11), 1193−1202.
(28) Sirisomboon, P.; Funke, A.; Posom, J. Improvement of Proximate Data and Calorific Value Assessment of Bamboo through near Infrared Wood Chips Acquisition. *Renew. Energy* **2020**, *147*, 1921−1931.
(29) Zhang, K.; Zhou, L.; Brady, M.; Xu, F.; Yu, J.; Wang, D. Fast Analysis of High Heating Value and Elemental Compositions of Sorghum Biomass Using Near-Infrared Spectroscopy. *Energy* **2017**, *118*, 1353−1360.
(30) Posom, J.; Sirisomboon, P. Evaluation of Lower Heating Value and Elemental Composition of Bamboo Using near Infrared Spectroscopy. *Energy* **2017**, *121*, 147−158.
(31) Posom, J.; Shrestha, A.; Saechua, W.; Sirisomboon, P. Rapid Non-Destructive Evaluation of Moisture Content and Higher Heating Value of *Leucaena Leucocephala* Pellets Using near Infrared Spectroscopy. *Energy* **2016**, *107*, 464−472.
(32) *Application of NIR Spectroscopy to Agricultural Products*; Burns, D. A.; Ciurczak, E. W., Eds.; CRC Press, 2007; pp 365−404.
(33) Xu, F.; Yu, J.; Tesso, T.; Dowell, F.; Wang, D. Qualitative and Quantitative Analysis of Lignocellulosic Biomass Using Infrared Techniques: A Mini-Review. *Appl. Energy* **2013**, *104*, 801−809.
(34) Fagan, C. C.; Everard, C. D.; McDonnell, K. Prediction of Moisture, Calorific Value, Ash and Carbon Content of Two Dedicated Bioenergy Crops Using near-Infrared Spectroscopy. *Bioresour. Technol.* **2011**, *102* (8), 5200−5206.
(35) Li, S. T.; Fu, L.; He, W. M.; Zhang, L. L.; Tie, J. X.; Li, Y. S.; Hao, X. W.; Tian, Y. N.; Bi, Y. M.; Wu, J. Z.; Wang, H.; Xu, Q. Q. Tobacco Substitution and Cigarette Blend Maintenance Based on near Infrared Spectral Similarity. *Tob. Sci. Technol.* **2020**, *53*, 88−93.
(36) Mancini, M.; Rinnan, Å. Classification of Waste Wood Categories According to the Best Reuse Using FT-NIR Spectroscopy and Chemometrics. *Anal. Chim. Acta* **2023**, *1275*, No. 341564.
(37) Liao, F.; Li, Y.; He, W.; Tie, J.; Hao, X.; Tian, Y.; Li, S.; Zhang, L.; Tang, L.; Wu, J.; Wang, H.; Xu, Q.; Bi, Y. Evaluation of Aroma Styles in Flue-Cured Tobacco by near Infrared Spectroscopy Combined with Chemometric Algorithms. *J. Infrared Spectrosc.* **2020**, *28*, 93.

https://doi.org/10.1021/acsomega.4c05472
ACS Omega 2024, 9, 48196−48204