https://dx.doi.org/10.21577/0103-5053.20210143

J. Braz. Chem. Soc., Vol. 33, No. 3, 251-259, 2022
©2022 Sociedade Brasileira de Química

Article

Rapid Recognizing the Producing Area of a Tobacco Leaf Using Near-Infrared Technology and a Multi-Layer Extreme Learning Machine Algorithm

Ruidong Li,[a] Wenyong Huang,[a] Guanlan Shang,[a] Xiaobing Zhang,[b] Xin Wang,[b] Jianguo Liu,[b] Yong Wang,[c] Junfeng Qiao,[c] Xing Fan,[c] Kai Wu*,[c] and Wenhua Zi *,[c]

[a]Yunnan Leaf Tobacco Redrying Co., Ltd., 650000 Kunming, China

[b]China Tobacco Zhejiang Industrial Co., Ltd., 310008 Hangzhou, China

[c]College of Energy and Environment Science, Yunnan Normal University, 650000 Kunming, China

A novel recognition method was put forward to identify the producing areas of the flue-cured tobacco leaves rapidly and non-destructively by using a near-infrared (NIR) spectrometer and a multi-layer-extreme learning machine (ML-ELM) algorithm. In contrast to traditional linear discriminant analysis (LDA) and extreme learning machine (ELM) algorithms, the accuracy, sensitivity and specificity were the highest for the proposed ML-ELM algorithm. The ML-ELM models for different producing areas of Yunnan tobacco leaves had the best generalization ability and prediction results. Besides, the above three algorithms were also identified by using the chemical index data. The experimental results indicated that the NIR spectroscopy technology together with ML-ELM algorithm achieved the best prediction performance both using the NIR spectral data and chemical index data. It indicates that the combination of NIR and ML-ELM can recognize different producing areas of Yunnan tobacco leaves rapidly, accurately, and non-destructively.

Keywords: NIR spectroscopy, ML-ELM algorithm, tobacco leaves, producing area identification

Introduction

Tobacco is a high economic crop in China. The quality of cigarette product is significantly affected by the intrinsic attribute of the tobacco leaf itself. The partition for tobacco leaves producing area and quality level play crucial roles in the final cigarette products quality management. Nowadays, the producing area recognition of a tobacco leaf is mainly dependent on the chemical analysis and human sensory responses.[1,2] The chemical analysis method is expensive, time-consuming and cannot be synchronized to the tobacco producing and processing. Meanwhile, the results reliability of human sensory responses is not quantified and sometimes it is subjective.[3] Herein, it is very important and necessary to develop a new method which is rapid, cheap, highly efficient and more objective.

As a useful analytical chemistry tool, near infrared (NIR) spectroscopy exhibits advantages such as non-destructive, cheap, accurate, and fast.[4,5] It has been widely used in the fields of agriculture,[6] medicine,[7] food,[8-10] traditional Chinese medicine[11-13] and so on. In the previous research,[14] the different producing areas of tobacco leaves were classified by artificial neural networks together with NIR spectroscopy. The pattern recognition for tobacco leaves planted in different producing areas, positions and levels was carried out by Mahalanobis distance criterion based on principal components of the leaves characterized by NIR.[15] Du *et al*.[16] has built 115 models of destination tobacco leaves of different producing areas, levels and varieties with soft independent modeling of class analogy method and NIR. Besides, NIR with least-support vector machines was applied to determine producing areas of tobacco leaves.[17] The previous reports cited above were mainly focused on the recognition of the producing areas of tobacco leaves that were planted in different provinces of China. There were few researches concerning on the recognition of the producing areas of tobacco leaves that were cultivated in different cities of one province, especially in Yunnan Province. As the largest tobacco leaf planting area in China, the tobacco leaf production of Yunnan Province accounted for 45% of China's total production in 2020. In

*e-mail: kaiwu8868@hotmail.com; kgclxwh@163.com
Editors handled this article: Eduardo Carasek and Ivo M. Raimundo Jr. (Associate)

fact, the chemical and style characteristics of the tobacco leaves are very different like the climate and altitude variation of different cities in Yunnan province. Therefore, it is also necessary study to determine differences of the chemical and style characteristics of the tobacco leaves from different cities using a rapid method.

Extreme learning machine (ELM) put forward by Huang *et al*.[18] has been widely used in classification,[19] regression,[20] feature selection[21] and so on. However, ELM still has some key problems to be solved, especially processing the high-dimensional spectral data. Deep learning is used to analyze the characteristics of the spectral data. Multi-layer extreme learning machine (ML-ELM) is one of the unsupervised learning methods using both deep learning and extreme learning machine. This learning process of the method is layer by layer. Comparing with the traditional ELM algorithm, ML-ELM algorithm can not only obtain the essential features of the original spectral data, but also can reduce the dimensionality. Meanwhile, as a kind of deep neural network, ML-ELM algorithm can approximate the complicated function and does not need iteration when building the calibration model. Comparing with the traditional ELM algorithm and other machine learning algorithms, the generalization performance and speed of ML-ELM algorithm are better. It has several advantages over ELM algorithm in processing the spectral data. ML-ELM is already been used in image recognition, hyperspectral data classification, speech recognition and so on.[22] However, there are few applications in the classification of NIR spectral data. Considering the above discussion and analysis, ML-ELM is very suitable for the processing of NIR spectral data.

In the study, a novel classification method using NIR technology and ML-ELM algorithm was put forward to recognize the producing area of flue-cured tobacco leaves rapidly and non-destructively. The experimental results showed that the combination of NIR spectroscopy and ML-ELM algorithm is a promising tool for identifying the different producing areas of Yunnan tobacco leaves accurately and non-destructively.

Experimental

Sample preparation and test

The NIR spectrometer should be firstly preheated with one hour. Then, the tobacco leaves were scanned after the tests of the NIR spectrometer performed successfully. In the scanning process, all the tobacco leaves should be ground into the powder which was then put into the rotating sample cup. The absorbance spectra of the tobacco leaves samples were acquired by using a NIR-Antaris II (Thermo Fisher Scientific America, Massachusetts, USA). The range of the wavenumber was 10,000-4000 cm^{-1}. The spectral resolution was 4 cm^{-1} and 64 scans were co-added. A polytetrafluoroethylene (PTFE) background disc was used as the spectral reference. Each sample was recorded with three spectrums and the means of the three spectrums were calculated as the final spectrum of each sample.[23,24]

In the following research, two C1F and C2F classes experimental sets from different producing areas were chosen. The first C1F experimental set has 501 samples and the samples were harvested in 2019 from Jinggu, Yaoan, Xinping and Luliang cities, Yunnan Province of China. The second C2F experimental set has 643 samples and the samples were also harvested in 2019. It contains 4 different producing areas: Xuanwei, Luxi, Jingdong and Malong cities, Yunnan Province of China. The distribution of the 8 producing areas are shown in Figure 1. It can be seen from Figure 1 that some locations of 8 producing areas are very close. As the result, it may be difficult to recognize the producing areas of tobacco leaves. In the experimental process, the samples were divided into 3 parts. It contained calibration, validation and testing samples. The above 3 types of samples were chosen randomly. 345 samples were chosen as the calibration set, 89 samples as the validation samples and 67 samples as the test set for data set 1. 430 samples were chosen as the calibration set, 129 samples as the validation samples and 84 samples as the test set for data set 2. The results were showed in Table 1 for the details of data sets 1 and 2. The NIR spectral data of the two sets collected by the NIR spectrometer was shown in Figure 2.

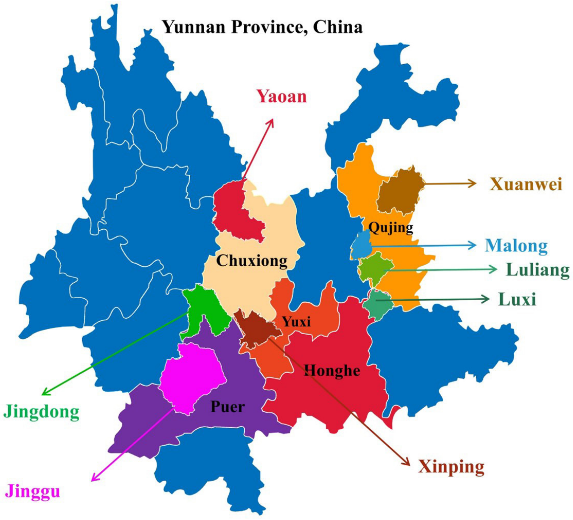


Figure 1. Original spectral data of the two data sets.

The content of total sugar, reducing sugar, potassium, total plant alkali, chlorine, and total nitrogen are 6 routine

Table 1. Details of data set 1 and data set 2

Data	Year	Producing area	Class	No. features	No. samples	No. calibration samples	No. validation samples	No. testing samples
Data set 1	2019	Jinggu, Puer	C1F	1609	152	105	27	20
	2019	Yaoan, Chuxiong	C1F	1609	114	79	20	15
	2019	Xinping, Yuxi	C1F	1609	136	92	24	20
	2019	Luliang, Qujing	C1F	1609	99	69	18	12
Data set 2	2019	Xuanwei, Qujing	C2F	1609	114	79	20	15
	2019	Luxi, Honghe	C2F	1609	223	146	47	30
	2019	Jingdong, Puer	C2F	1609	195	130	40	25
	2019	Malong, Puer	C2F	1609	111	75	22	14

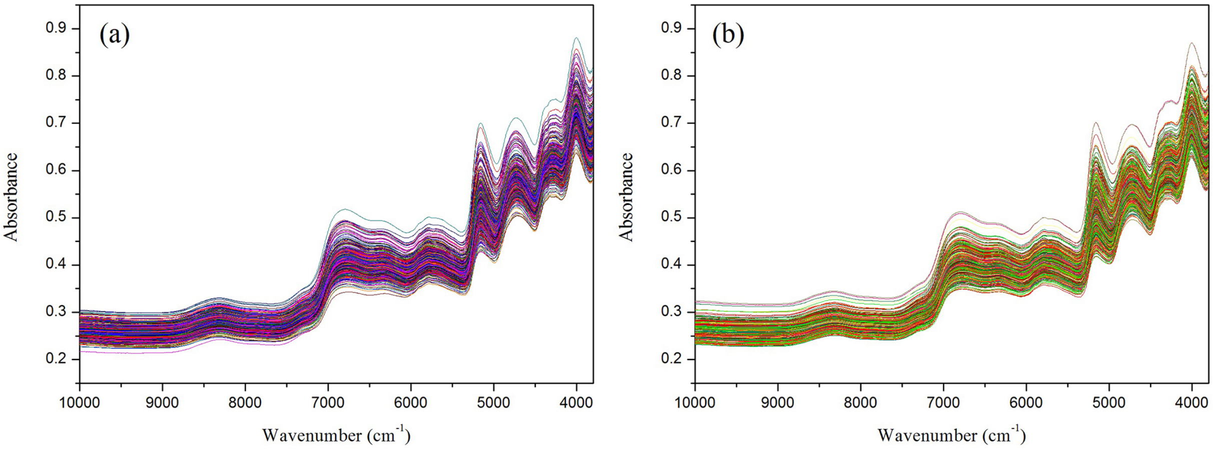


Figure 2. Original spectral data of the two data sets: (a) data set 1; (b), data set 2.

chemical indexes of a tobacco leaf. The values of 6 indexes can reflect the quality of a tobacco leaf to some degree and they have also been used for recognizing the producing areas of tobacco leaves in the previous research.[15,16] As the result, the above 6 routine chemical indexes of all the tobacco leaf samples were also detected by using continuous flow analytical method with Skalar SANPWS flow analyzer (Breda, Netherlands).[25] The results of Table 2 showed the average values and standard deviations of the 6 routine chemical indexes of tobacco leaves in 8 different producing areas. It can be seen from Table 2 that 6 routine chemical indexes of a tobacco leaf exhibited some difference although the locations of some producing areas are close to each other. For example, the maximum average value of total sugar, potassium, reducing sugar, total plant alkali, chlorine, total nitrogen were 20.29, 52.03, 55.27, 56.41, 222.88, 15.50% higher than that of the minimum value in 6 different producing areas, respectively. The differences of the 6 routine chemical indexes of a tobacco leaf in different producing areas are huge. Therefore, it is necessary to recognize the tobacco leaves producing areas in different cities of Yunnan province.

Theory of ELM algorithm

ELM algorithm was put forward by Huang *et al.*[18] and the hidden nodes of ELM algorithm were usually performed randomly. If the input data is mapped to L dimensional ELM random feature space, then the network output can be defined as equation 1:

$$f_L(x) = \sum_{i=1}^{L} \beta_i h_i(x) = \mathbf{h}(x)\beta \tag{1}$$

where $\beta = [\beta_1, \dots, \beta_L]^T$ is the output weight matrix, $\mathbf{h}(x) = [g_1(x), \dots, g_L(x)]$ are the hidden node outputs and $g_i(x)$ is the output of the i-th hidden node. Given N training samples $\{(x_i, t_i)\}_{i=1}^{N}$, ELM is to resolve the following learning problems:

$$\mathbf{H}\beta = \mathbf{T} \tag{2}$$

where $\mathbf{T} = [t_1, \dots, t_N]^T$ are the target labels and $\mathbf{H} = [h^T(x_1), \dots, h^T(x_N)]^T$. The output weights β can be calculated by equation 3:

$$\beta = \mathbf{H}^{\dagger}\mathbf{T} \tag{3}$$

Table 2. Average values and standard deviations of routine chemical indexes with different tobacco leaves producing areas

Chemical index		Data set 1 (C1F)				Data set 2 (C2F)			
		Jinggu	Yaoan	Xinping	Luliang	Xuanwei	Luxi	Jingdong	Malong
Total sugar / %	average value	31.315	37.469	34.395	36.562	33.088	31.150	33.047	34.466
	standard deviation	4.189	3.350	4.469	3.527	4.133	3.640	3.619	3.881
Potassium / %	average value	2.285	1.503	1.662	1.729	1.693	1.711	2.063	1.822
	standard deviation	0.305	0.308	0.350	0.291	0.264	0.306	0.311	0.365
Reducing sugar / %	average value	24.935	28.053	37.080	27.545	23.881	25.533	23.940	24.447
	standard deviation	3.082	2.509	3.258	2.545	3.373	2.794	2.846	3.022
Total plant alkali / %	average value	2.717	1.958	2.373	1.967	1.755	2.745	2.348	1.755
	standard deviation	0.624	0.561	0.691	0.388	0.467	0.472	0.504	0.379
Chlorine / %	average value	0.433	0.762	0.566	0.483	0.236	0.676	0.394	0.346
	standard deviation	0.189	0.461	0.240	0.249	0.131	0.288	0.177	0.271
Total nitrogen / %	average value	2.148	1.922	2.112	2.073	1.987	2.220	2.038	2.079
	standard deviation	0.280	0.293	0.289	0.260	0.237	0.260	0.274	0.261

where $\mathbf{H}^{\dagger}$ is the Moore-Penrose generalized inverse of matrix $\mathbf{H}$.

If the solution wants to be more robust and has better generalization performance, a regularization term needs to be added and it is shown in equation 4.

$$\beta = \left(\frac{1}{C} + \mathbf{H}^{T}\mathbf{H}\right)^{-1} \mathbf{H}^{T}\mathbf{T} \qquad (4)$$

where C is the regularization coefficient and the values of this parameter will be assigned randomly after the appropriate hidden layer numbers are set.

Theory of ML-ELM algorithm

Multi layer neural networks perform poorly when trained with back propagation (BP) only. Hence hidden layer weights in a deep network are initialized using layer wise unsupervised training and the whole neural network is fine-tuned using BP further. Similar to deep networks, each ML-ELM hidden layer weights are initialized using extreme learning machine auto encoder (ELM-AE) which performs layer wise unsupervised training. However, in contrast to deep networks, ML-ELM does not require fine tuning.

The activation functions of ML-ELM hidden layer can be either linear or nonlinear piecewise. If the number of nodes L^{k} in k-th hidden layer is equal to the number of nodes L^{k-1} in the (k – 1)-th hidden layer, g could be linear, otherwise, g could be nonlinear piecewise, e.g., sigmoidal function.

$$\mathbf{H}^{k} = g\left(\left(\beta^{k}\right)^{T} \mathbf{H}^{k-1}\right) \qquad (5)$$

where $\mathbf{H}^{k}$ represents the outputs of ML-ELM k-th hidden layer. If k-1= 0, this layer is the input layer, and $\mathbf{H}^{k}$ represents the inputs of ML-ELM. β^{k} represents the output weights of ELM-AE, and the inputs of ELM-AE are $\mathbf{H}^{k}$ at this time. The output weights β^{k} of ML-ELM can be analytically calculated using regularized least squares.

The flow chart is shown in Figure 3 for recognizing the producing area of a flue-cured tobacco leaf by using ML-ELM algorithm. The calibration and test samples were serial treated by spectral pre-processing, feature extraction using principal component analysis (PCA), parameters determination and classification of ML-ELM.

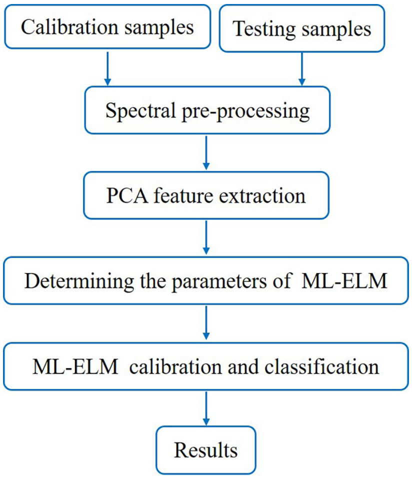
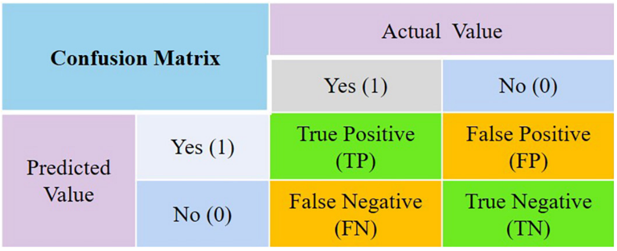


Figure 3. Diagram of recognizing the producing area on tobacco leaves by using ML-ELM algorithm.

Measures of classification performance

Confusion matrix is a concept from machine learning, and it contains information about actual and predicted classifications done by a classification system. A confusion matrix has two-dimensions, one dimension is indexed by the actual class of an object, the other is indexed by the class that the classifier predicts. Figure 4 presents the basic form of confusion matrix for a classification task.



Confusion Matrix		Actual Value	
		Yes (1)	No (0)
Predicted Value	Yes (1)	True Positive (TP)	False Positive (FP)
	No (0)	False Negative (FN)	True Negative (TN)

Figure 4. Confusion matrix.

A number of measures of classification performance can be defined based on the confusion matrix. Some common measures are given as follows.

Accuracy is the proportion of the total number of predictions that were correct:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{6}$$

where TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

Sensitivity is a measure of the ability of a prediction model to select instances of a certain class from a data set, which is defined by the formula:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{7}$$

Specificity is the proportion of actual negatives measured that were correct:

$$\text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}} \tag{8}$$

Results and Discussion

According to the pre-processing and PCA operation methods used in previous reports,[23-25] firstly Savitzky-Golay derivative pre-processing operation is performed on the spectral data to smooth and remove the noise.[23] Here first derivative with a 11-point number of smoothing points and two polynomial order methods were chosen. Figure 5 shows the spectra after pre-processing. The band assignments are: total sugar, 5050, 5200, and 7194 cm^{-1}; potassium, 5050 and 5200 cm^{-1}; reducing sugar, 4194, 4444, and 4789 cm^{-1}; total plant alkali, 4444 and 4664 cm^{-1}; chlorine, 4194, 4789, and 5200 cm^{-1}; total nitrogen, 4789 and 7194 cm^{-1}. It could also be seen from Figure 5 that the resolution of the spectral data has been improved after the pre-processing operation.[23,24] However, the dimension of the spectral data is still huge after pre-processing operation. Therefore, PCA was used for reducing the dimension of the data.[25] Here, mean centering operation was used in the PCA analysis. The result of PCA showed that the first six principal components are 99.32 and 99.78% for data sets 1 and 2, respectively. This means that the first six principal components contained vast majority of information for both data sets 1 and 2.

Then, linear discriminant analysis (LDA), ELM, and ML-ELM algorithms were applied using the loadings and scores obtained by PCA operation. Accordingly, they were also used to identify the spectral data of different producing areas of tobacco leaves. To achieve the fair comparison and avoid the randomness in test results, all the calibration and testing samples were randomly chosen and the three algorithms ran on the same calibration and test splits for each calculation. For ML-ELM, the number of layers was an important parameter. Therefore, the first work was to define the number of hidden layers of the ML-ELM algorithm in order to achieve the better performance with less parameters. Here, sigmoid was set as the activation function and the number of hidden nodes was set as 10 and 500. It can be seen from Figure 6 that the overall accuracies of both data sets were increasing firstly and then decreasing with the number of the hidden layers increasing. The accuracy was the highest when the number was 3. Three

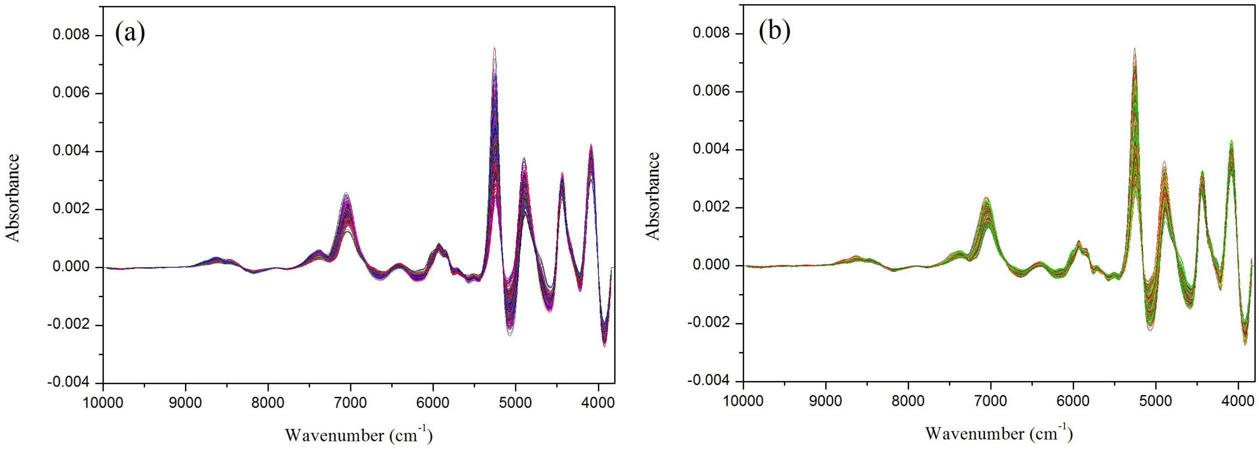


Figure 5. Pre-processing results by using Savitzky-Golay derivative and two polynomial order methods. (a) Data set 1; (b) data set 2.

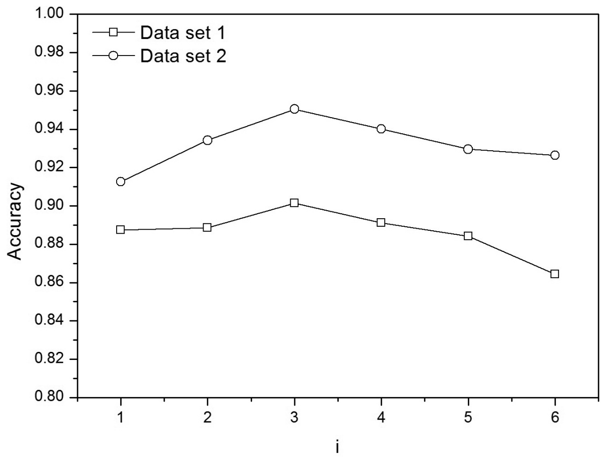


Figure 6. Overall accuracy via different hidden layer numbers.

hidden layers of the ML-ELM algorithm will be chosen in the following experiment.

As mentioned above, the samples were divided into 3 parts. It contained calibration, validation and testing samples. The above 3 types of samples were chosen randomly. Here we use accuracy, sensitivity, and specificity to evaluate the performances of the calibration models, validation results and testing results for each producing area and each algorithm. In order to reflect the performance of the different predictors faithfully and to avoid over-fitting, the experiment is performed and verified using a ten-fold cross validation. It means all the three algorithms were calculated 10 times. For the sake of comparison, the performance of LDA, ELM and ML-ELM algorithms are shown in Tables 3 and 4 for data sets 1 and 2 in the form of a confusion matrix. As shown in Tables 3 and 4, the accuracy, sensitivity, specificity of the ML-ELM algorithm were the highest for the calibration, validation and prediction samples compared with the LDA and ELM algorithms. The above results show the ML-ELM algorithm has the best performance to build the different calibration models for different producing areas of Yunnan tobacco leaves with NIR spectral data. Besides, the calibration models built by the ML-ELM algorithm also have the better prediction performance than the other LDA and

Table 3. Performance of calibration and validation of various models

Data set	Algorithm	Producing area	Sample		Cal			Val		
			Cal	Val	AC / %	Sn / %	Sp / %	AC / %	Sn / %	Sp / %
Data set 1 (C1F)	LDA	Jinggu	105 × 10	27 × 10	83.96	78.95	86.17	85.28	83.33	86.12
		Yaoan	79 × 10	20 × 10	77.96	38.23	89.85	80.00	41.5	91.16
		Xinping	92 × 10	24 × 10	80.55	62.93	87.01	83.26	68.75	88.61
		Luliang	69 × 10	18 × 10	83.24	66.57	87.28	83.15	62.78	88.31
	ELM	Jinggu	105 × 10	27 × 10	96.15	93.23	97.44	91.68	85.92	94.19
		Yaoan	79 × 10	20 × 10	88.78	79.24	91.63	78.76	58.00	84.78
		Xinping	92 × 10	24 × 10	94.69	94.02	94.94	84.27	78.75	86.31
		Luliang	69 × 10	18 × 10	89.24	63.66	95.47	82.25	48.89	90.70
	ML-ELM	Jinggu	105 × 10	27 × 10	100	100	100	95.28	94.44	95.64
		Yaoan	79 × 10	20 × 10	99.88	99.75	99.92	95.06	91.00	96.23
		Xinping	92 × 10	24 × 10	100	100	100	93.26	85.00	96.31
		Luliang	69 × 10	18 × 10	99.88	99.71	99.92	93.26	81.11	96.34
Data set 2 (C2F)	LDA	Xuanwei	79 × 10	20 × 10	78.35	51.39	84.41	83.33	55.00	88.53
		Luxi	146 × 10	47 × 10	73.56	56.09	82.53	76.27	68.30	80.85
		Jingdong	130 × 10	40 × 10	73.95	44.77	86.60	75.89	57.25	84.27
		Malong	75 × 10	22 × 10	78.97	59.60	83.07	82.63	46.36	90.09
	ELM	Xuanwei	79 × 10	20 × 10	89.32	70.00	93.67	87.44	65.00	91.56
		Luxi	146 × 10	47 × 10	87.27	87.05	87.39	82.01	77.66	84.51
		Jingdong	130 × 10	40 × 10	89.56	74.61	96.03	83.49	66.00	91.35
		Malong	75 × 10	22 × 10	96.29	93.20	96.96	91.08	77.27	93.92
	ML-ELM	Xuanwei	79 × 10	20 × 10	100	100	100	96.43	91.00	97.43
		Luxi	146 × 10	47 × 10	99.86	99.79	99.89	96.04	93.83	97.32
		Jingdong	130 × 10	40 × 10	99.97	99.92	100	95.74	93.00	96.97
		Malong	75 × 10	22 × 10	99.84	99.60	99.89	96.28	88.64	97.85

Cal: calibration; Val: validation; AC: accuracy; Sn: sensitivity; Sp: specificity; LDA: linear discriminant analysis; ELM: extreme learning machine; ML-ELM: multi-layer extreme learning machine.

Table 4. Testing results using different algorithms with chemical and NIR spectral data

Data set	Algorithm	Producing area	Sample (Pred)	NIR spectral data			Chemical data		
				AC / %	Sn / %	Sp / %	AC / %	Sn / %	Sp / %
Data set 1 (C1F)	LDA	Jinggu	20	83.58	80.00	85.11	82.09	75.00	85.11
		Yaoan	15	79.10	40.00	90.38	77.61	40.00	88.46
		Xinping	20	82.09	70.00	87.23	80.59	70.00	85.11
		Luliang	12	83.58	58.33	89.09	82.09	50.00	89.09
	ELM	Jinggu	20	92.53	85.00	95.74	86.57	70.00	93.62
		Yaoan	15	82.09	55.33	90.38	77.61	53.33	84.61
		Xinping	20	88.06	85.00	89.36	86.56	85.00	87.23
		Luliang	12	83.58	58.33	89.09	80.59	41.67	89.09
	ML-ELM	Jinggu	20	97.01	95.00	97.87	94.03	90.00	95.74
		Yaoan	15	95.52	86.67	98.08	92.54	86.67	94.23
		Xinping	20	94.03	90.00	95.74	92.53	85.00	95.74
		Luliang	12	92.54	83.33	94.54	89.55	66.67	94.54
Data set 2 (C2F)	LDA	Xuanwei	15	83.13	57.14	88.40	79.76	46.67	86.96
		Luxi	30	77.11	66.67	83.02	75.00	66.67	79.63
		Jingdong	25	78.31	60.00	86.21	76.19	56.00	84.74
		Malong	14	84.33	57.14	89.85	80.95	42.86	88.57
	ELM	Xuanwei	15	86.90	66.67	91.30	86.90	66.67	91.30
		Luxi	30	84.52	76.67	88.89	85.71	76.67	90.74
		Jingdong	25	84.52	72.00	89.83	85.71	76.00	89.83
		Malong	14	89.28	71.43	92.86	91.67	78.57	94.28
	ML-ELM	Xuanwei	15	96.43	93.33	97.12	92.86	80.00	95.65
		Luxi	30	97.62	96.67	98.15	92.86	90.00	94.44
		Jingdong	25	96.43	92.00	98.30	92.86	88.00	94.91
		Malong	14	95.24	85.71	97.14	92.85	78.57	95.71

NIR: near-infrared; AC: accuracy; Sn: sensitivity; Sp: specificity; LDA: linear discriminant analysis; ELM: extreme learning machine; ML-ELM: multi-layer extreme learning machine.

ELM algorithms. This is because they could be the result that the minimum Euclidean distance was used for LDA algorithm to classify the spectral data and it was ineffective when the dimensional spectral data was high. For ELM algorithm, the amounts of hidden nodes of ELM algorithm were randomly set. However, the ML-ELM classification algorithm picked up the best number of hidden nodes by using the unsupervised learning, thus learning more abstract features of the NIR spectral data.

In order to verify the experimental results as described above, different tobacco leaves producing areas were also recognized after using 6 routine chemical indexes. The classification and cross-validation methods were the same as the above experiment using NIR spectral data. The experimental results with chemical indexes are shown in the last three columns of Table 4. The results showed that the accuracy, sensitivity, specificity of ML-ELM algorithm was the highest among the three algorithms when using chemical indexes. However, the above evaluation results of ML-ELM algorithm using chemical index data were much lower than using NIR spectral data for each producing area. Besides, NIR spectroscopy technology was cheap, low-cost, and effective compared with the chemical index detection method. Considering experimental results and consequences as described above, the NIR spectroscopy technology together with ML-ELM algorithm could be the most effective tool for recognizing different producing areas of tobacco leaves among all the above methods tested.

The averaged elapsed execution time are usually used to estimate the performance of an algorithm. Here the averaged elapsed execution time contains the calibration model building and the prediction process. All the experiments were performed on the same computer. The parameters of the computer are Core TM i7-8700h, 3.20GHz, CPU with 8GB RAM, with Windows 7 Professional operation system. All the algorithms were calculated by using the language of Matlab.[26] The results of Figure 7 showed the computing times of LDA, ELM, and ML-ELM algorithms on the two data sets using NIR spectral data. It was obvious that ELM and ML-ELM algorithms were much more efficient than LDA algorithm. Although the computing time of ML-ELM algorithm was a bit slower than the ELM algorithm, considering the classification accuracy, ML-ELM algorithm was also the best option to classify the NIR spectral data of the tobacco leaves from different producing areas.

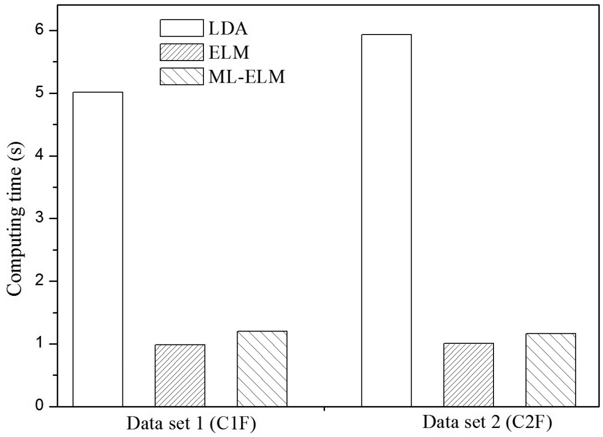


Figure 7. Execution time of LDA, ELM, ML-ELM algorithms based on NIR spectral data.

Conclusions

Our study proposed a novel method using NIR spectroscopy technology together with ML-ELM algorithm to identify the different producing areas of tobacco leaves cultivated in Yunnan province. The results showed that the method put forward was an alternative strategy to discriminate different producing areas of tobacco leaves rapidly, accurately, and non-destructively. Besides, the ML-ELM algorithm performed much better than traditional LDA and ELM algorithms based on both NIR spectral data and chemical indexes data. The results indicated that application of the NIR spectroscopy technology together with ML-ELM algorithm could be useful for determining different plantation areas of Yunnan tobacco leaves.

Acknowledgments

We really want to thank Dr Zhang Jianqiang from Yunnan Reascend Tobacco Technology (Group) Co., Ltd. for the help in collecting the NIR spectral data and giving the advice of pre-processing and PCA operation methods. This work is financially supported by the China Tobacco Corporation (110201901031(FK-02)) and Yunnan Tobacco Company (2019530000241042-1). The authors would like to thank the Yunnan Leaf Tobacco Redrying Co., Ltd. for providing the experimental samples.

Author Contributions

Ruidong Li, Wenyong Huang, Guanlan Shang and Xiaobing Zhang contributed to the resources; Xin Wang, Jianguo Liu, Yong Wang and Junfeng Qiao conducted data curation and investigation; Xin Fang and Kai Wu wrote original draft; Wenhua Zi was responsible for the conceptualization and writing-review.

References

1. Chen, B.; Xing, W.; Lu, D.; Qi, X.; *J. Jiangsu Univ.* **2015**, *36*, 545.
2. Sun, J. G.; He, J. W.; Wu, F. G.; Tu, S. X.; Yan, T. J.; Hui, S. L.; Xie, H.; *Agric. Sci. China* **2011**, *10*, 1222.
3. Bin, J.; Ai, F. F.; Fan, W.; Zhou, J. H.; Yun, Y. H.; Liang, Y. Z.; *RSC Adv.* **2016**, *6*, 30353.
4. Chauchard, F.; Cogdill, R.; Roussel, S.; Roger, J. M.; Bellon-Maurel, V.; *Chemom. Intell. Lab. Syst.* **2004**, *71*, 141.
5. Li, B.; Wang, C.; Xi, L.; Wei, Y.; Duan, H.; Wu, X.; *Anal. Methods* **2014**, *6*, 9691.
6. Nicolaï, B. M.; Beullens, K.; Bobelyn, E.; Peirs, A.; Saeys, W.; Theron, K. I.; Lammertyn, J.; *Postharvest Biol. Technol.* **2007**, *46*, 99.
7. Colak, S. B.; van der Mark, M. B.; Hooft, G. W.; Hoogenraad, H. H.; Can, D. L. E. S.; Kuijpers, F. A.; *IEEE J. Sel. Top. Quantum Electron.* **1999**, *5*, 1143.
8. Chen, Q. S.; Zhao, J. W.; Liu, M. H.; Cai, J.; Liu, J.; *J. Pharm. Biomed. Anal.* **2008**, *46*, 568.
9. Guo, X.; Cai, R.; Wang, S.; Tang, B.; Li, Y.; Zhao, W.; *R. Soc. Open Sci.* **2018**, *5*, 170714.
10. Urickova, V.; Sadecka, J.; *Spectrochim. Acta, Part A* **2015**, *148*, 131.
11. Beć, K. B.; Grabska, J.; Huck, C. W.; *J. Pharm. Biomed. Anal.* **2021**, *193*, 113686.
12. Zhang, Z. Y.; Wang, Y. J.; Yan, H.; Chang, X. W.; Duan, J.; *J. Anal. Methods Chem.* **2021**, 887586.
13. Pan, W.; Wu, M.; Zheng, Z.; Guo, L.; Qiu, B.; *J. Food Sci.* **2020**, *85*, 2004.
14. Hana, M.; McClure, W. F.; Whitaker, T. B.; White, M.; Bahler, D.; *J. Near Infrared Spectrosc.* **1997**, *5*, 19.
15. Shu, R. X.; Wang, G. D.; Zhang, J. P.; Ni, L. J.; *Tob. Sci. Technol.* **2006**, *39*, 12.
16. Du, W.; Yi, J. H.; Tan, X. L.; *Acta Tab. Sin.* **2009**, *15*, 1.
17. Shi, F. C.; Li, D. L.; Feng, G. L.; Song, G.; *Tob. Sci. Technol.* **2013**, *46*, 56.
18. Huang, G. B.; Zhu, Q. Y.; Siew, C. K.; *Neurocomputing* **2006**, *70*, 489.
19. Mao, Y. C.; Dong, X.; Cheng, J. F.; Jiang, J. H.; Tuan, L. B.; Liu, S. J.; *Spectrosc. Spectral Anal.* **2017**, *37*, 89.
20. Yu, Q.; Heeswijk, M. V.; Miche, Y.; Nian, R.; He, B.; Séverin, E.; Lendasse, A.; *Neurocomputing* **2014**, *129*, 153.

21. Benoít, F.; Heeswijk, M. V.; Miche, Y.; Verleysen, M.; Lendasse, A.; *Neurocomputing* **2013**, *102*, 111.
22. Deng, C. W.; Huang, G. B.; Xu, J.; Tang, J. X.; *Sci. China Inf. Sci.* **2015**, *58*, 1.
23. Zhang, J. Q.; Liu, W. J.; Hou, Y.; Qiu, C. G.; Yang, S. Y.; Li, C. Y.; Nie, L. R.; *Anal. Lett.* **2017**, *51*, 1029.
24. Zhang, J. Q.; Liu, W. J.; Zhang, H. H.; Hou, Y.; Yang, P. P.; Li, C. Y.; Yang, Y. M.; Li, M.; *J. Near Infrared Spectrosc.* **2018**, *26*, 101.
25. Shuangyan, Y.; Ying, H.; Lingchun, Y.; Jianqiang, Z.; Weijuan, L.; Changgui, Q.; Ming, L.; Yanmei, Y.; *J. Braz. Chem. Soc.* **2018**, *29*, 1480.
26. *Matlab*, 2016b; The MathWorks Inc., Natick, Massachusetts, USA, 2016.

Submitted: April 6, 2021
Published online: October 21, 2021