IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 16, 2020, accepted October 5, 2020, date of publication October 8, 2020, date of current version October 21, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3029714

Identification of Soybean Origin by Terahertz Spectroscopy and Chemometrics

XIAO WEI[1], SHIPING ZHU[1], (Member, IEEE), SHENGLING ZHOU[1], WANQIN ZHENG[2], AND SONG LI[1]

[1]College of Engineering and Technology, Southwest University, Chongqing 400716, China
[2]College of Food Science, Southwest University, Chongqing 400716, China

Corresponding author: Shiping Zhu (zspswu@126.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 31771670 and Grant 62005227, and in part by the Graduate Research Innovation Project of Chongqing under Grant CYB20099.

ABSTRACT Soybeans have the characteristics of balanced amino acid species and high nutritional value. In this article, the feasibility of the identification soybean from three typical origins (Argentina, the United States and China) by interval partial least squares (iPLS) optimized terahertz (THz) spectroscopy combined with chemometrics was investigated. Firstly, the THz frequency-domain spectrum was optimized using iPLS. Then, 168 soybean samples were selected as the correction set, and soybean origin identification models were respectively built using the extreme learning machine (ELM), genetic algorithm support vector machine (GA-SVM), and artificial bee colony algorithm support vector machine (ABC-SVM) combined with 8 pre-processing techniques. Finally, the models were verified through 57 samples of the test set and the comprehensive identification accuracy rate of the ABC-SVM model reached 94.74%. The experimental results showed that after iPLS optimization and appropriate pre-processing technique, THz spectroscopy and chemometrics could accurately identify the origin of soybean.

INDEX TERMS Terahertz, origin, pre-processing technique, iPLS, chemometrics.

I. INTRODUCTION

Soybeans from different origins have large differences in appearance, color, nutritional value, and internal chemical composition [1], [2]. According to the database of the Food and Agriculture Organization of the United Nations, the United States and Argentina are among the top three soybean producing countries in the world in 2019. At the same time, they are also the main sources of imported soybeans for China. According to China Customs, Chinese soybean imports have remained above 80% of Chinese total consumption from 2016-2019. China imports 88 million tons of soybeans and the share of soybean imports in China's domestic consumption is 84.86% in 2019. Imported soybeans are mostly genetically modified soybeans, which are not allowed to be traded privately in China. With the development of globalization, the issue of the protection of the origin of agricultural products and food has attracted more and more attention from researchers in various countries [3], [4]. International scholars in the European Union, the United States and other countries have pioneered a series of exploratory research efforts in this field. These studies mainly focus on the analysis of the chemical composition of products from different regions to find out the specific indicators that can characterize the regional information [5]. For example, Du *et al.* [6] analyzed the geographic origin of rice by mineral elements and characteristic volatile components, with a final average classification accuracy of 93.5%. For example, Latorre *et al.* [7] used mineral element analysis and chemometrics to qualitatively identify potatoes and found this method to be effective. For example, Pilgrim *et al.* [8] applied trace element and stable isotope signatures in determining the origin of tea tree samples and developed a simple method for analyzing and verifying the origin of tea leaves. Other methods, Dittgen *et al.* [9] used liquid chromatography-mass spectrometry (LC-MS) to evaluate the physicochemical characteristics of black rice and select the best genotype that could enhance black rice production. For example, Lim *et al.* [10] used untargeted metabolomics approaches to analyze the geographical origin of rice and proposed a phospholipid-based discrimination method. For example, Wadood *et al.* [11] successfully established a

The associate editor coordinating the review of this manuscript and approving it for publication was Sara Dadras.

classification model based on linear discriminant analysis by gas chromatography to discriminate the geographic origin of Chinese winter wheat. This model could be used to distinguish between geographical origin and variety, but the accuracy rate for origin identification was low. In general, most of these methods are sensitive and accurate. However, these identification methods have disadvantages such as high cost, low efficiency, cumbersome operation, difficult operation by non-specialists and so on. In recent years, spectroscopic techniques such as Near Infrared Spectroscopy (NIR) had been introduced into food origin detection, quality identification, quantitative detection and so on [12], [13]. Although the above-mentioned spectroscopy techniques had been used in agricultural and food research, the use of terahertz (THz) spectroscopy had been received little attention.

THz radiation refers to electromagnetic waves with a frequency between 0.1-10 THz. Compared with other traditional spectroscopy techniques, it has unique advantages. The vibrations of molecules (proteins and amino acids, etc.) and intermolecular interactions are right in the THz frequency range [14]. THz spectroscopy is extremely sensitive to discover subtle differences and changes in the structure of matter, so THz spectroscopy has broad research prospects in biorecognition [15]. At present, there have been some related researches on the detection of agricultural products and food based on THz spectroscopy. For example, Chen *et al.* [16] distinguished transgenic beets by THz spectroscopy, and the final percentage of classification for both transgenic and non-transgenic beets was 100%. For example, Liu *et al.* [17] applied THz spectroscopy and chemometrics to identify transgenic rice seeds, and the accuracy of the prediction set was 96.67%. For example, Liu *et al.* [18] applied THz spectroscopy and chemometrics in the identification of transgenic camellia oil and found that the use of continuous projection arithmetic could improve the classification accuracy of weighted linear discriminant analysis (WLDA). For example, Liu *et al.* [19] used THz spectroscopy combined with chemometrics to distinguish the geographical origin of extra virgin olive oil, and the accuracy of the prediction set was 96.25%. For example, Liu *et al.* [20] used THz spectroscopy to determine the adulterated acacia honey and the final correlation coefficient was 0.985. For example, Liu and Fan [21] used THz spectroscopy to quantify potassium aluminum sulfate dodecahydrate in potato starch and found that potassium aluminum sulfate dodecahydrate had a distinct characteristic absorption peak in the THz band. For example, Zhang *et al.* [22] used THz time-domain spectroscopy and chemometrics to determine amino acid mixtures in cereals and finally proved that THz time-domain spectroscopy could be used for the qualitative and quantitative analysis of amino acids. For example, Peng *et al.* [23] used THz spectroscopy to qualitatively and quantitatively identify the components in the mixture, and the average correlation coefficient for the identification reached 99.135%. But so far, there has been no relevant study report on spectral region-optimized THz spectrum and chemometrics to identify the origin of soybeans.

The purpose of this article was to study the feasibility of THz spectroscopy and interval partial least squares (iPLS) combined with chemometrics to identify soybean origin. At the same time, the soybean origin model identification results after eight pre-processing techniques were compared and the most suitable pre-processing techniques for the three modeling algorithms (extreme learning machine (ELM), genetic algorithm support vector machine (GA-SVM) and artificial bee colony algorithm support vector machine (ABC-SVM)) were separately found.

II. MATERIALS AND RELATED WORK

A. MATERIALS

A total of 75 soybeans from different batches of Argentina, the United States and China were selected as experimental samples. Among them, 25 samples were from Argentina, 9 samples from the US and 41 samples from China. All samples were collected and provided by the Quality Inspection Center of the Tianjin Grain and Oil Wholesale Trading Market in China. Because the soybean samples from these three origins (Argentina, the United States and China) were more convenient to collect by the Quality Inspection Center, and more batches could be provided. Therefore, this article used these three typical origin soybeans for identification. There were 75 different batches of soybean samples and each batch was weighed at 50 g for this experiment. During the experiment, it was found that the origin of different batches of soybeans could not be identified by the naked eye. Therefore, it is impossible to identify the origin of soybeans by naked eyes. Moreover, information on the size, dimensions and weight of soybean samples was not relevant for subsequent origin modeling and prediction. Thus, we did not record the size, dimensions, and weight of the 75 batches of soybeans.

B. THz EXPERIMENTAL DEVICE STRUCTURE

T-SPEC THz time-domain spectroscopy equipment of EKSPLA was used in the experimentation. The optical path was controlled and calibrated by optical lenses. This equipment used low temperature gallium arsenide (LT-GaAs) as a photoconductive antenna. The antenna was made of Ti/Au spray coating. The optical path between the emitter and the detector was about 62.5 cm. The ultrashort pulse laser light source of this system was FF50 femtosecond laser. The ultrashort pulse was divided into strong pulse and weak pulse by the beam splitter after passing through the half-wave plate. After passing through the chopper and mirror, the strong pulse hit the LT-GaAs photoconductive antenna, which generated a THz electromagnetic radiation pulse. Then this THz pulse was focused on the sample to be tested. The THz pulse transmitted from the sample to be tested and the weak pulse merge, and then the merged signal was sent to the amplifier. Finally, the THz time-domain spectrum with the information of the sample to be measured was obtained.

C. SAMPLE PREPARATION AND SPECTRUM ACQUISITION

Firstly, 75 soybean samples were dried in a drying cabinet at 50 °C for 3 h. The dried samples were milled by a grinder. Then, the samples were ground with a mortar and the samples were filtrated by the sieve with the mesh size of 0.074 mm. Secondly, according to the ratio of 3: 7, pure polyethylene powder was added to each sample powder and then the two powders were uniformly mixed. After that, the mixed samples were weighed 135 mg by the precision balance, which were compressed into tablets by the tablet press under a pressure of 20 MPa. Finally, the experimental sample was prepared as a sheet with a thickness of approximately 1 mm and a circular shape. Each soybean sample was weighed into 3 experimental samples, so a total of 225 experimental samples were used in this experiment. Before starting the experiment, the THz spectral instrument was preheated for 30 minutes and simultaneously filled with nitrogen. During the experiment, the relative humidity was controlled below 5% in the instrument and the room temperature of the laboratory was controlled at 25 °C. Each experimental sample was scanned 6 points and each scan point was scanned 256 times. The THz time-domain spectra of this experimental sample were the average of six scan-point spectra. The THz time-domain spectrum was transformed into a THz frequency-domain spectrum through Fourier transform. The frequency range of THz spectra was 0.1-2.5 THz and contained 263 spectral point data.

III. RELATED THEORIES

A. PRE-PROCESSING TECHNIQUE

The THz spectrum can reflect its own data information but it also has some information interference that is not related to the nature of the sample itself, such as sample background, noise, stray light, device response and so on. At the same time, the THz spectrum is also affected by the physical properties of the spectral data, resulting in the baseline drift of the THz spectrum and the non-repeatability of the spectrum. These interferences not only affect the acquisition of useful information of the spectrum, but also affect the establishment of the identification model and the prediction effect of the measured sample. Therefore, before establishing a stable and predictive discrimination model, it has become necessary to use appropriate pre-processing techniques to eliminate irrelevant information from the spectral data. The pre-processing techniques can availably remove the interference of spectral noise and partial physical conditions, making qualitative discrimination more intuitive and reliable [24]. Therefore, this article used 8 pre-processing techniques, including: mean center, auto scaling, standard normal variate (SNV), normalization, multiplicative scatter correction (MSC), first derivative, second derivative, orthogonal signal correction (OSC).

B. iPLS SPECTRAL REGION OPTIMIZATION

Spectral region optimization is extremely significant for improving the accuracy of the qualitative discrimination model. It selects a spectral region with better signal-to-interference ratio for the discrimination model. At the same time, because part of the spectral region with large information interference is removed, the modeling time can be reduced and the identification efficiency can be improved. The position and width of chosen spectral regions have a great important influence on the performance of the identification model [25]. The selected spectral regions are too narrow or the selected location is wrong, resulting in too little information content related to the sample to be tested, so the discrimination effect of the established detection models are also reduced. Selected spectral region is too wide, while the spectral information contained in the spectral region associated with the test sample increases, but the noise content of the respective spectral region also increases. Therefore, the time for establishing the identification model becomes longer and the accuracy of the model becomes lower.

In this article, iPLS [26] was used to select the appropriate spectral interval. iPLS can be used to select spectral regions that are highly relevant to the analyzed components. Firstly, iPLS divides the collected THz spectral interval into n sub-intervals of the same width. Then each sub-interval is established as a local regression model. Finally, by comparing the root mean square error of cross validation (RMSECV) of n subintervals, the appropriate modeling spectral interval is selected.

C. PRINCIPAL COMPONENT ANALYSIS (PCA)

PCA is one of the most widely used data dimensionality reduction methods. It is the process of orthogonally transforming the original data space into a low-dimensional subspace. This transformation reduces the dimensionality of the data set without loss or little loss of the data set [27]. PCA converts the initial variables so that some of the new variables become linear combinations of the initial variables. Usually, the number of new variables selected is less than the initial variables. Meanwhile, these new variables should represent the data information of the initial variables to the greatest extent and they should not lose useful information. The new variables are also named Principal Components (PCs).

D. ELM

ELM is a machine learning method based on the Feedforward Neuron Network (FNN). Compared with traditional neural networks, ELM is faster than traditional learning algorithms under the premise of ensuring learning accuracy and has better generalization ability [28]. ELM randomly gives input weights and thresholds in hidden layer node weights and calculates output weights through sample training. The method does not need to manually give input weights and thresholds during training. It only needs to specify the number of nodes in the hidden layer and chooses the appropriate activation function to obtain the optimal solution.

E. GA-SVM

Support vector machine (SVM) can transform a low-dimensional linearly indistinguishable sample into a high-dimensional feature space by non-linear mapping to make it linearly distinguishable [29]. The kernel functions frequently used by SVM are linear kernel functions, polynomial kernel functions, radial basis functions (RBF) and so on. Among them, the most diffusely applied one is RBF, which can project sample data to a higher dimensional space and require fewer parameters to be ascertained. The paper used genetic algorithm (GA) and artificial bee colony algorithm (ABC) to optimize RBF parameters.

GA is an evolutionary algorithm whose principle is to imitate the evolutionary law in the biological world [30]. It encodes the problem parameter as chromosomes. Then chromosomes use iterative methods to perform selection, crossover, and mutation operations to exchange chromosome information in the population. Finally, chromosomes that meet the optimization goal are generated. In GA, chromosomes correspond to data or arrays, which are usually represented by one-dimensional string structure data. A string of genes is called a chromosome or genotyped individual. A certain number of individuals make up a group. The number of individuals in a group is called the group size, also known as the group size.

F. ABC-SVM

ABC is a new type of swarm intelligence optimization algorithm, which simulates the honey gathering process of bees [31]. It solves the contradiction between expanding new food sources and conducting precise searches around known food sources, to a large extent avoiding falling into local optimal [32]. The ABC algorithm includes three types of bees: leading bee, following bee, and scout bee and the solution of the optimization problem is used as the food source location [33]. An important advantage of the ABC algorithm is the deep switching of information, that is, all bees in the algorithm rely on swinging actions to exchange information. The speed of bees searching for food sources is the speed of solving optimization goals. In this article, the ELM, GA-SVM and ABC-SVM identification models were established and verified in Matlab R2018b. The computer operating system was Windows 10.0.

IV. RESULTS AND DISCUSSION

A. THz FREQUENCY-DOMAIN SPECTRUM

Fig. 1 and Fig. 2 are THz frequency-domain spectral images of 225 samples in the 0.1-2.5 THz and 0.1-1.5 THz range. The differences in the THz frequency-domain spectra of the experimental samples from different origins can be seen in the Fig. 1 and Fig. 2. This might be due to differences in the content of certain chemical components. The chemical component with the highest potential to cause this was the protein content, followed by the moisture content. 0.1-1.5 THz range of the THz frequency domain spectral signal

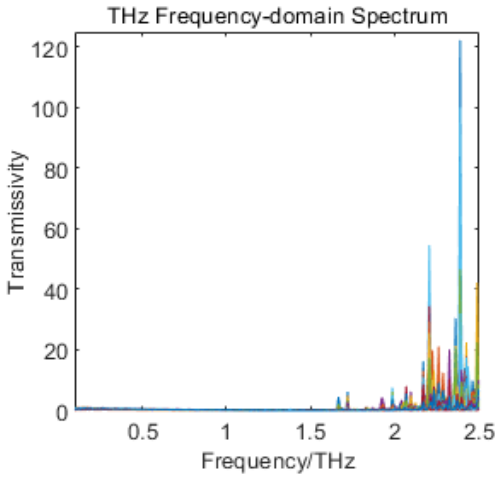

FIGURE 1. THz spectrum in 0.1-2.5 THz range.

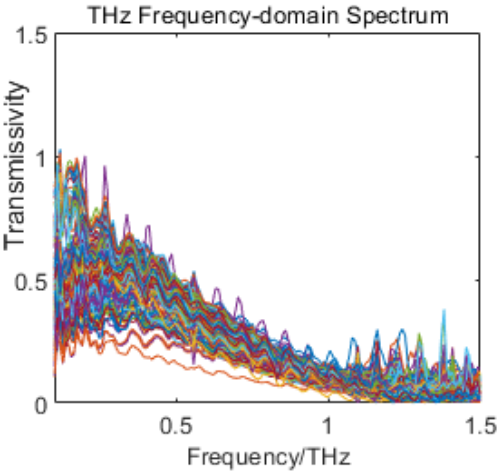

FIGURE 2. THz spectrum in 0.1-1.5 THz range.

gradually decreased. From the THz spectral images, it can be seen that there are significantly obvious interferences in the 1.5-2.5 THz interval. These disturbances had a very large impact on the subsequent identification of soybean origin. Hence, it was essential to optimize the spectral interval of the THz frequency-domain spectrum through iPLS. Fig. 3 shows the average THz frequency-domain spectra of soybean samples from the three origins (USA, China and Argentina). It can be seen from the figure that the spectra of soybean samples from China at 0.1-0.3 THz are higher than those of the other two origins. The THz spectra of the soybean samples from the USA and Argentina were similar.

B. iPLS SELECT THE APPROPRIATE SPECTRAL REGION

Wavelength was used as input when iPLS was used for spectral region optimization of NIR spectra, but the horizontal coordinate of the THz frequency-domain spectrum was frequency and could not be directly input. Therefore, when iPLS is used to optimize the spectral region of the THz spectrum,

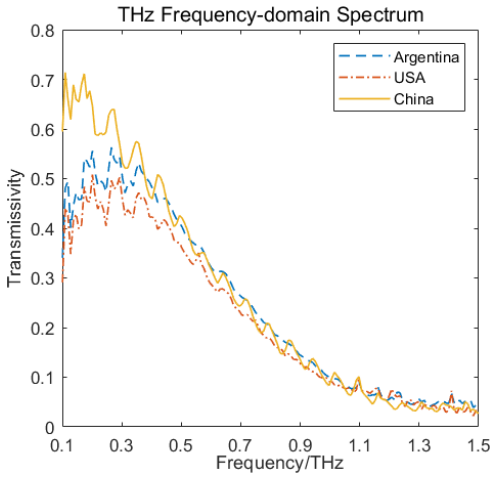

FIGURE 3. Average THz frequency domain spectra of soybean samples from three origins.

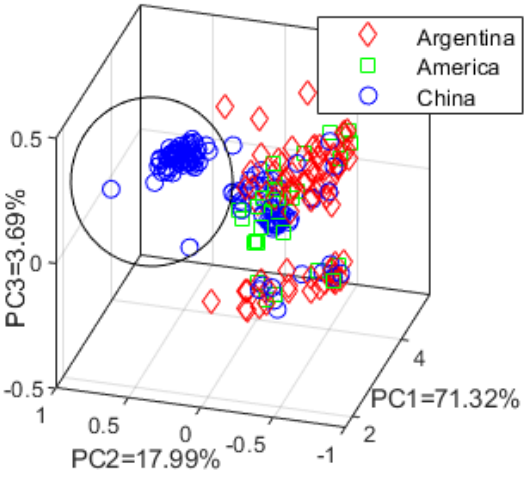

FIGURE 5. PCs score chart.

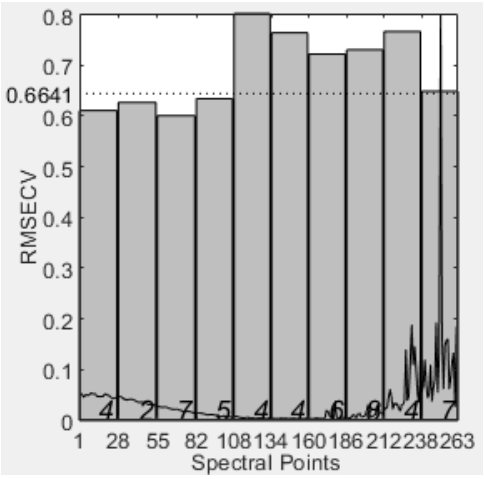

FIGURE 4. RMSECV in the range of 0.1-2.5THz.

the input is replaced by the number of spectral points in the frequency interval. The 263 spectral points (0.1- 2.5 THz) of the THz frequency-domain spectral interval were divided into 10 equal parts and then sub-intervals were selected through iPLS. Fig. 4 shows the RMSECV results for 10 sub-intervals. From the figure, the dotted line indicated that when the PLS components (PLSC) of the full-spectrum PLS model was 6, the RMSECV was 0.6641. In the picture, it could be found that the third sub-interval had the lowest RMSECV, so it was chosen as the first selected sub-interval. Nevertheless, there was very little useful information in an alone sub-interval and it could not achieve the objective of establishing the discrimination models with excellent predictive capability and robustness [34]. Hence, the third sub-interval was taken as the center and then the range of the selected sub-interval was extended bidirectionally, and the sub-interval with RMSECV less than 0.6641 was chose. Finally, the THz frequency-domain spectral interval modeled in this experiment was chosen as the frequency interval of 0.1-1.0792 THz (1-108 spectral points).

C. PCA

The THz spectrum after optimized selection by iPLS was subjected to PCA. The variance contribution rate of the first five PCs were 71.32%, 17.99%, 3.69%, 1.79% and 1.33%, and the cumulative variance contribution rate (CVC) was 96.12%. Fig. 5 is a graph showing the dispersion of the first 3 PCs of 225 soybean samples. It could be seen from the figure that a part of the soybean samples in China and the soybean samples from other places had a more obvious difference, but most of the samples had a serious overlap in three-dimensional space. Soybean samples from the United States were completely scattered between the other two soybeans of origin. Therefore, PCA alone could not effectively distinguish soybean samples from different origins and it was necessary to combine chemometrics to further examine soybean origins.

D. ELM SOYBEAN ORIGIN IDENTIFICATION MODEL

The transmissivity and frequency of the THz frequency-domain spectrum were used as input variables for the ELM soybean origin identification model. Soybean samples were randomly selected according to the 3: 1 ratio between the calibration set and the test set. The 168 calibration set soybean samples contained 56 samples from Argentina, 20 samples from the United States and 92 samples from China. Among the 57 test set soybean samples, there were 19 samples from Argentina, 7 samples from the United States and 31 samples from China. Firstly, THz frequency-domain spectral data was subjected to 8 different pre-processing techniques. After that, the ELM soybean origin identification models were established through the calibration set. Finally, the identification effects of the ELM soybean origin identification models were verified through the test set. The verification results are shown in Table 1. The accuracy cv (Acv) was the identification accuracy of the cross-validation (leave-one-out method) of the calibration set samples. The accuracy 1 (A1) was the correct rate for the identification of Argentine soybean samples. The accuracy 2 (A2) was the identification

TABLE 1. ELM soybean origin identification model verification results.

Spectrum pre-processing technique	Acv%	A1%	A2%	A3%	TA%
(a).None	92.86	73.68	71.43	90.32	82.46
(b).Mean center	**93.45**	**94.74**	**57.14**	**90.32**	**87.72**
(c).Auto scaling	92.26	89.47	28.57	93.55	84.21
(d).SNV	93.45	89.47	42.86	93.55	85.96
(e).Normalization	**94.05**	**94.74**	**42.86**	**93.55**	**87.72**
(f).MSC	93.45	78.95	14.29	96.77	80.70
(g).First derivative	92.86	89.47	57.14	87.10	84.21
(h).Second derivative	92.26	84.21	71.43	87.10	84.21
(i).OSC	97.02	73.68	57.14	83.87	77.19

TABLE 2. GA-SVM and ABC-SVM soybean origin identification model verification results.

Spectrum pre-processing techniques	GA-SVM					ABC-SVM				
	Acv%	A1%	A2%	A3%	TA%	Acv%	A1%	A2%	A3%	TA%
(a).None	100	89.47	71.43	93.55	89.47	100	89.47	71.43	93.55	89.47
(b).Mean center	100	89.47	71.43	93.55	89.47	100	89.47	71.43	93.55	89.47
(c).Auto scaling	**100**	**94.74**	**71.43**	**96.77**	**92.98**	**100**	**94.74**	**85.71**	**96.77**	**94.74**
(d).SNV	98.23	94.74	71.43	93.55	91.23	100	84.21	71.43	93.55	87.72
(e).Normalization	100	94.74	71.43	96.77	92.98	100	94.74	71.43	96.77	92.98
(f).MSC	98.81	84.21	57.14	93.55	85.96	100	89.47	57.14	90.32	85.96
(g).First derivative	83.33	100	28.57	87.10	84.21	99.40	84.21	57.14	93.55	85.96
(h).Second derivative	77.98	100	0	87.10	80.70	100	89.47	57.14	93.55	87.72
(i).OSC	87.50	84.21	0	90.32	77.19	92.26	84.21	42.86	87.10	80.70

accuracy rate of US soybean samples. The accuracy 3 (A3) was the correct rate of Chinese soybean sample identification. The total accuracy (TA) was the correct rate of identification of all experimental samples in the test set.

It could be seen from Table 1 that after pre-processin techniques, the identification effect of the ELM soybean origin identification model was better. This showed that there were definite differences in the soybeans of different origins in the THz frequency and the THz frequency-domain spectrum could be used to detect the soybean origin. Comparing the verification results of the ELM soybean origin identification model after different pre-processing techniques, it could be found that the identification effect of the ELM model after mean center or normalization was the mostaccurate, which was 5.26% higher than the ELM identification model without pre-processing. This might be because the two spectral pre-processing techniques removed the great mass of THz spectral interference and the THz spectral data retained was the most effective. Therefore, these two pre-processing techniques were most suitable for the ELM origin identification model. In terms of accuracy of identification, the identification of soybean samples in Argentina and China was better but the identification of American soybean samples needed to be improved. This was similar to the PCA results. American soybean samples were completely mixed between the other two soybean samples, making accurate identification difficult. Therefore, it was necessary to find a more suitable identification method for American soybean samples.

E. GA-SVM AND ABC-SVM SOYBEAN ORIGIN IDENTIFICATION MODELS

Based on the preceding means, 225 samples were separated into the calibration set and the test set. The GA and ABC were used to optimize the selection of the RBF penalty parameter c and the kernel function parameter g. Firstly, the THz frequency-domain spectrum was subjected to eighprocessing techniques. Secondly, GA-SVM and ABC discriminant models were found by the calibration set. Ultimately, the test set was separately employed for verification. Table 2 shows the verification results of GA-SVM and ABC-SVM soybean origin identification models.

It could be seen from Table 2 that after different pre-processing techniques, the verification result of the ABC-SVM soybean origin identification model was better than the GA-SVM model. This might be because when ABC optimized the parameters of SVM, it simulated the honey collecting process of the bee colony and divided the bee colony into more reasonable functions. Therefore, it showed more superior performance in solving the function optimization problem and finally obtained the optimal discrimination model. After the auto scaling pre-processing technique, the total accuracy rate of the ABC-SVM soybean origin identification model was increased to 94.74% ($c = 9.51\times10^{9}$, $g = 1.73\times10^{-4}$). This further illustrated that THz spectroscopy and chemometrics could accurately identify the origin of soybeans. After the auto scaling pre-processing technique, the GA-SVM and ABC-SVM soybean origin identification models had achieved the best identification effect, which was 3.51% and 5.27% higher than the identification model without pre-processing technique. This might be because auto scaling pre-processing technique deleted redundant spectral data from the THz frequency-domain spectrum, thereby enhancing the differences between the spectral data. In the verification results of the ABC-SVM soybean origin identification model, the identification model after the optimal pre-processing technique (auto scaling) was 14.04% higher than the identification model using the worst pre-processing technique (OSC). This illustrated that the spectral pre-processing technique had a crucial role in the effect of soybean origin identification model.

Comparing Tables 1 and 2, it was found that the ABC-SVM soybean origin identification model had a higher total accuracy than the GA-SVM and ELM models. This indicated that in the actual soybean origin identification, the ABC-SVM soybean origin identification model was better than the other two models. After comprehensively comparing the verification results of the soybean origin identification models of ELM, GA-SVM and ABC-SVM, the paper discovered that after the auto scaling pre-processing technique, the ABC-SVM soybean origin identification model could achieve the best identification effect. In the identification of the origin of the soybean samples in Argentina, America and China, the identification effects of using the auto scaling pre-processing technique and the ABC-SVM identification model were 94.74%, 85.71% and 96.77%.

Comparing the previous related papers on the identification of origin, it could be found that THz spectrum and iPLS combined with chemometrics could be used to identify the origin of soybean. At the same time, it provided better identification, simpler operation and wider application. For example, Dan and Yang [35] based on the L-1-LRC model combined with NIR spectroscopic data to identify orange origin. Although the L-1-LRC model could achieve an accuracy of 92.35% using training samples, the identification performance should be further improved. For example, Zhou *et al.* [36] were able to identify the origin of panax notoginseng samples by combining Fourier transform mid-infrared (FT-MIR) and NIR spectroscopy. Although the origin identification results could reach 95.6%, the model was too complex and it required the use of two spectroscopic techniques. For example, Lai *et al.* [37] applied energy dispersive X-ray fluorescence spectroscopy for the analysis of soybean traceability. The effect of origin identification could reach 96.2%. However, the selected experimental samples were all from China, and the origin identification of foreign soybeans was not performed. Moreover, it needed to measure the nine elements in soybean first, and then carried out the subsequent identification study, which was a complicated experimental operation. Therefore, the identification of soybean origin by THz spectroscopy and iPLS combined with chemometrics is a new method that can replace the traditional soybean origin identification.

V. CONCLUSION

The experimental results showed that THz spectroscopy and iPLS combined with chemometrics could be used to accurately identify soybean origin. After iPLS and auto scaling pre-processing technique, the total accuracy of the ABC-SVM soybean origin identification model reached 94.74%. Mean center or normalization pre-processing technique was the most suitable for ELM soybean origin identification model. Auto scaling pre-processing technique was the best for GA-SVM and ABC-SVM soybean origin identification models. The novelty of this article is the identification of soybean origin by THz spectroscopy. At the same time, a more accurate and scientific method is proposed to select the frequency range of THz frequency-domain spectrum. This article has important reference value for the accurate and rapid identification of other agricultural products and food origins. It can play a great reference role in the research of soybean variety identification, doping and protein content analysis.

ACKNOWLEDGMENT

The authors would like to thank Meidie Hu and Fengxu Li for their precious support in sample acquisition and preparation. They would also like to thank Weiji Wu, Zhiyong Xie, Hua Huang, Xiyu Wu, and Jiaxin Zuo for their helpful discussions on the manuscript.

DISCLOSURES

The authors declare no conflicts of interest associated with this article.

REFERENCES

[1] J. K. Kim, E.-H. Kim, I. Park, B.-R. Yu, J. D. Lim, Y.-S. Lee, J.-H. Lee, S.-H. Kim, and I.-M. Chung, "Isoflavones profiling of soybean [Glycine max (L.) Merrill] germplasms and their correlations with metabolic pathways," *Food Chem.*, vol. 153, pp. 258–264, Jun. 2014.

[2] E. S. Tonello, N. L. Fabbian, D. Sacon, A. Netto, V. N. Silva, and P. M. Milanesi, "Soybean seed origin effects on physiological and sanitary quality and crop yield," *Semin. Agrárias*, vol. 40, no. 5, pp. 1789–1803, May/Oct. 2019.

[3] T. Bosona and G. Gebresenbet, "Food traceability as an integral part of logistics management in food and agricultural supply chain," *Food Control*, vol. 33, no. 1, pp. 32–48, Sep. 2013.

[4] R. Shankar, R. Gupta, and D. K. Pathak, "Modeling critical success factors of traceability for food logistics system," *Transp. Res. E, Logistics Transp. Rev.*, vol. 119, pp. 205–222, Nov. 2018.

[5] S. A. Drivelos and C. A. Georgiou, "Multi-element and multi-isotope-ratio analysis to determine the geographical origin of foods in the European union," *TrAC Trends Anal. Chem.*, vol. 40, pp. 38–51, Nov. 2012.

[6] M. Du, Y. Fang, F. Shen, B. Mao, Y. Zou, P. Li, F. Pei, and Q. Hu, "Multiangle discrimination of geographical origin of rice based on analysis of mineral elements and characteristic volatile components," *Int. J. Food Sci. Technol.*, vol. 53, no. 9, pp. 2088–2096, Sep. 2018.

[7] C. Herrero Latorre, J. Barciela García, S. García Martín, and R. M. Peña Crecente, "Chemometric classification of potatoes with protected designation of origin according to their producing area and variety," *J. Agricult. Food Chem.*, vol. 61, no. 35, pp. 8444–8451, Sep. 2013.

[8] T. S. Pilgrim, R. J. Watling, and K. Grice, "Application of trace element and stable isotope signatures to determine the provenance of tea (Camellia sinensis) samples," *Food Chem.*, vol. 118, no. 4, pp. 921–926, Feb. 2010.

[9] C. L. Dittgen, J. F. Hoffmann, F. C. Chaves, C. V. Rombaldi, J. M. C. Filho, and N. L. Vanier, "Discrimination of genotype and geographical origin of black rice grown in Brazil by LC-MS analysis of phenolics," *Food Chem.*, vol. 288, pp. 297–305, Aug. 2019.

[10] D. K. Lim, C. Mo, J. H. Lee, N. P. Long, Z. Dong, J. Li, J. Lim, and S. W. Kwon, "The integration of multi-platform MS-based metabolomics and multivariate analysis for the geographical origin discrimination of oryza sativa L." *J. Food Drug Anal.*, vol. 26, no. 2, pp. 769–777, Apr. 2018.

[11] S. A. Wadood, G. Boli, Z. Xiaowen, A. Raza, and W. Yimin, "Geographical discrimination of Chinese winter wheat using volatile compound analysis by HS-SPME/GC-MS coupled with multivariate statistical analysis," *J. Mass Spectrom.*, vol. 55, pp. 1–10, Jan. 2020.

[12] X. Feng, Y. Zhao, C. Zhang, P. Cheng, and Y. He, "Discrimination of transgenic maize kernel using NIR hyperspectral imaging and multivariate data analysis," *Sensors*, vol. 17, no. 8, pp. 1–14, Aug. 2017.

[13] P. Wang, X. Wang, Y. Sun, G. Gong, M. Fan, and L. He, "Rapid identification and quantification of the antibiotic susceptibility of lactic acid bacteria using surface enhanced Raman spectroscopy," *Anal. Methods*, vol. 12, no. 3, pp. 376–382, Jan. 2020.

[14] L. Afsah-Hejri, P. Hajeb, P. Ara, and R. J. Ehsani, "A comprehensive review on food applications of terahertz spectroscopy and imaging," *Compr. Rev. Food Sci. Food Saf.*, vol. 18, no. 5, pp. 1563–1621, Sep. 2019.

[15] L. Xie, Y. Yao, and Y. Ying, "The application of terahertz spectroscopy to protein detection: A review," *Appl. Spectrosc. Rev.*, vol. 49, no. 6, pp. 448–461, Aug. 2014.

[16] T. Chen, Z. Li, X. Yin, F. Hu, and C. Hu, "Discrimination of genetically modified sugar beets based on terahertz spectroscopy," *Spectrochimica Acta A, Mol. Biomolecular Spectrosc.*, vol. 153, pp. 586–590, Jan. 2016.

[17] W. Liu, C. Liu, X. Hu, J. Yang, and L. Zheng, "Application of terahertz spectroscopy imaging for discrimination of transgenic rice seeds with chemometrics," *Food Chem.*, vol. 210, pp. 415–421, Nov. 2016.

[18] J. Liu, L. Fan, Y. Liu, L. Mao, and J. Kan, "Application of terahertz spectroscopy and chemometrics for discrimination of transgenic camellia oil," *Spectrochimica Acta A, Mol. Biomol. Spectrosc.*, vol. 206, pp. 165–169, Jan. 2019.

[19] W. Liu, C. Liu, J. Yu, Y. Zhang, J. Li, Y. Chen, and L. Zheng, "Discrimination of geographical origin of extra virgin olive oils using terahertz spectroscopy combined with chemometrics," *Food Chem.*, vol. 251, pp. 86–92, Jun. 2018.

[20] W. Liu, Y. Zhang, M. Li, D. Han, and W. Liu, "Determination of invert syrup adulterated in acacia honey by terahertz spectroscopy with different spectral features," *J. Sci. Food Agric.*, vol. 100, pp. 1913–1921, Mar. 2020.

[21] J. Liu and L. Fan, "Qualitative and quantitative determination of potassium aluminum sulfate dodecahydrate in potato starch based on terahertz spectroscopy," *Microw. Opt. Technol. Lett.*, vol. 62, no. 2, pp. 525–530, Feb. 2020.

[22] X. Zhang, S. Lu, Y. Liao, and Z. Zhang, "Simultaneous determination of amino acid mixtures in cereal by using terahertz time domain spectroscopy and chemometrics," *Chemom. Intell. Lab. Syst.*, vol. 164, pp. 8–15, May 2017.

[23] Y. Peng, C. Shi, M. Xu, T. Kou, X. Wu, B. Song, H. Ma, S. Guo, L. Liu, and Y. Zhu, "Qualitative and quantitative identification of components in mixture by terahertz spectroscopy," *IEEE Trans. THz Sci. Technol.*, vol. 8, no. 6, pp. 696–701, Nov. 2018.

[24] J. Liu, J. Han, J. Xie, H. Wang, W. Tong, and Y. Ba, "Assessing heavy metal concentrations in earth-cumulic-orthic-anthrosols soils using Vis-NIR spectroscopy transform coupled with chemometrics," *Spectrochim. Acta A, Mol. Biomol. Spectrosc.*, vol. 226, Feb. 2020, Art. no. 117639.

[25] Y. He, Y. Zhao, C. Zhang, Y. Li, Y. Bao, and F. Liu, "Discrimination of grape seeds using laser-induced breakdown spectroscopy in combination with region selection and supervised classification methods," *Foods*, vol. 9, no. 2, p. 199, Feb. 2020.

[26] M. G. Nespeca, A. L. Vieira, D. S. Júnior, J. A. G. Neto, and E. C. Ferreira, "Detection and quantification of adulterants in honey by LIBS," *Food Chem.*, vol. 311, May 2020, Art. no. 125886.

[27] E. Scollo, D. C. A. Neville, M. J. Oruna-Concha, M. Trotin, P. Umaharan, D. Sukha, R. Kalloo, and R. Cramer, "Proteomic and peptidomic UHPLC-ESI MS/MS analysis of cocoa beans fermented using the styrofoam-box method," *Food Chem.*, vol. 316, Jun. 2020, Art. no. 126350.

[28] Y. Zhao, F. Qin, F. Xu, J. Ma, Z. Sun, Y. Song, L. Zhao, J. Li, and H. Wang, "Identification of *Tilletia foetida, Ustilago tritici*, and *Urocystis tritici* based on near-infrared spectroscopy," *J. Spectrosc.*, vol. 2019, p. 15, Jul. 2019.

[29] H. Huang, Y. Lan, A. Yang, Y. Zhang, S. Wen, and J. Deng, "Deep learning versus object-based image analysis (OBIA) in weed mapping of UAV imagery," *Int. J. Remote Sens.*, vol. 41, no. 9, pp. 3446–3479, May 2020.

[30] G. Ren, Y. Wang, J. Ning, and Z. Zhang, "Highly identification of keemun black tea rank based on cognitive spectroscopy: Near infrared spectroscopy combined with feature variable selection," *Spectrochimica Acta A, Mol. Biomolecular Spectrosc.*, vol. 230, Apr. 2020, Art. no. 118079.

[31] O. Babaie and M. Nasr Esfahany, "Optimization and heat integration of hybrid R-HIDiC and pervaporation by combining GA and PSO algorithm in TAME synthesis," *Separat. Purification Technol.*, vol. 236, Apr. 2020, Art. no. 116288.

[32] J. Song, G. Li, X. Yang, X. Liu, and L. Xie, "Rapid analysis of soluble solid content in navel orange based on visible-near infrared spectroscopy combined with a swarm intelligence optimization method," *Spectrochimica Acta A, Mol. Biomolecular Spectrosc.*, vol. 228, Mar. 2020, Art. no. 117815.

[33] E. T. Oldewage, A. P. Engelbrecht, and C. W. Cleghorn, "Movement patterns of a particle swarm in high dimensional spaces," *Inf. Sci.*, vol. 512, pp. 1043–1062, Feb. 2020.

[34] X. Wei, W. Zheng, S. Zhu, S. Zhou, W. Wu, and Z. Xie, "Application of terahertz spectrum and interval partial least squares method in the identification of genetically modified soybeans," *Spectrochimica Acta A, Mol. Biomolecular Spectrosc.*, vol. 238, Sep. 2020, Art. no. 118453.

[35] S. Dan and S. X. Yang, "A new L1-LRC based model for oranges origin identification with near infrared spectra data," *Evol. Intell.*, vol. 7, pp. 1–7, Apr. 2020.

[36] Y. Zhou, Z. Zuo, F. Xu, and Y. Wang, "Origin identification of Panax notoginseng by multi-sensor information fusion strategy of infrared spectra combined with random forest," *Spectrochim. Acta A, Mol. Biomol. Spectrosc.*, vol. 226, Feb. 2020, Art. no. 117619.

[37] H. Lai, J. Xi, J. Sun, W. He, Z. Wang, C. Zheng, and X. Mao, "Multi-elemental analysis by energy dispersion X-ray fluorescence spectrometry and its application on the traceability of soybean origin," *At. Spectrosc.*, vol. 41, no. 1, pp. 20–28, Jan./Feb. 2020.

XIAO WEI was born in Liaoning, China, in 1994. He received the master's degree in agricultural mechanization from Southwest University, Chongqing, China, in 2016, where he is currently pursuing the Ph.D. degree in agricultural mechanization engineering. His current research interests include machine learning, pattern recognition, terahertz spectrum detection, and quantitative analysis.

SHIPING ZHU (Member, IEEE) received the Ph.D. degree from China Agricultural University, Beijing, China. He is currently the Vice Dean, a Professor, and a Doctoral Supervisor with the College of Engineering and Technology, Southwest University, Chongqing, China, and the Vice Director of the Chongqing Engineering Experimental Teaching Center. He has published more than 110 articles and ten patents. His current research interests include the rapid detection methods of agricultural products and food quality based on terahertz, near infrared, Raman spectroscopy, hyper spectral, and machine vision.

SHENGLING ZHOU received the Ph.D. degree from Southwest University, Chongqing, China. She is currently a Lecturer with the School of Engineering and Technology, Southwest University. In 2015, she won a China Scholarship Council (CSC) support as a Visiting Scholar under the supervision of Tak W. Kee and Derek Abbott at The University of Adelaide, Adelaide, SA, Australia. Her research interests include signal processing, spectrum analysis, and signal processing for T-ray imaging.

WANQIN ZHENG was born in Yibin, China, in 1995. She is currently pursuing the M.S. degree in food science with the College of Food Science, Southwest University, China. Her current research interests include food chemistry and nutrition.

SONG LI was born in Chongqing, China, in 1990. He received the master's degree in electrical engineering from the Chongqing University of Technology, Chongqing, in 2016. He is currently pursuing the Ph.D. degree with Southwest University, China. He used to work with the Chongqing Aerospace Vocational and Technical College. His current research interest includes gas sensitive detection technology of SF6 circuit breaker.