International Journal on Semantic Web and Information Systems
Volume 18 • Issue 1

NIR Spectroscopy Oranges Origin Identification Framework Based on Machine Learning

Songjian Dan, Chongqing University of Education, China*

ABSTRACT

According to the characteristics of NIR spectral data, a complete general framework for origin identification is proposed. It includes steps such as data preprocessing, feature selection, model building, and cross validation. The authors compare multiple preprocessing algorithms and multiple machine learning algorithms under the framework. Based on NIR spectroscopy to identify the origin of orange, a good identification result was obtained. They improve the accuracy of orange origin identification and obtain the best origin identification accuracy of 92.8%.

KEYWORDS

Machine Learning, NIR Spectroscopy, Origin Identification

1. INTRODUCTION

As a fast, accurate, convenient and non-destructive analysis technology, near-infrared spectroscopy analysis technology has been widely used in agricultural product quality detection and origin identification. It is considered to be a non-destructive testing method that is expected to replace traditional chemical analysis (Caligiani, Palla, Acquotti, Marseglia & Palla, 2014; Sádecká, Jakubíková, Májek & Kleinová, 2016; Li, Nunes, Wang, Williams, Zheng, Zhang & Zhu, 2013; Chen, Lin, Wu, Wang, Wu & Tan, 2015). At present, the identification technology of orange origin based on near-infrared spectroscopy is still relatively time-consuming, laborious and not accurate enough. There is still a big gap between its completeness, system and operability and actual application. How to establish an effective technical system that can quickly identify the origin of orange has an important role in the healthy development of the orange industry in my country (Cao, Liang, Xu, Hu, Zhang & Fu, 2017; Diniz, Gomes, Pistonesi, Band & Araújo, 2014).

2. MACHINE LEARNING-BASED NIR SPECTROSCOPY ORANGE ORIGIN IDENTIFICATION FRAMEWORK

In this paper, a universal framework for rapid and non-destructive identification of the origin of orange is established by the spectral analysis technology based on machine learning. The specific process is

DOI: 10.4018/IJSWIS.297039

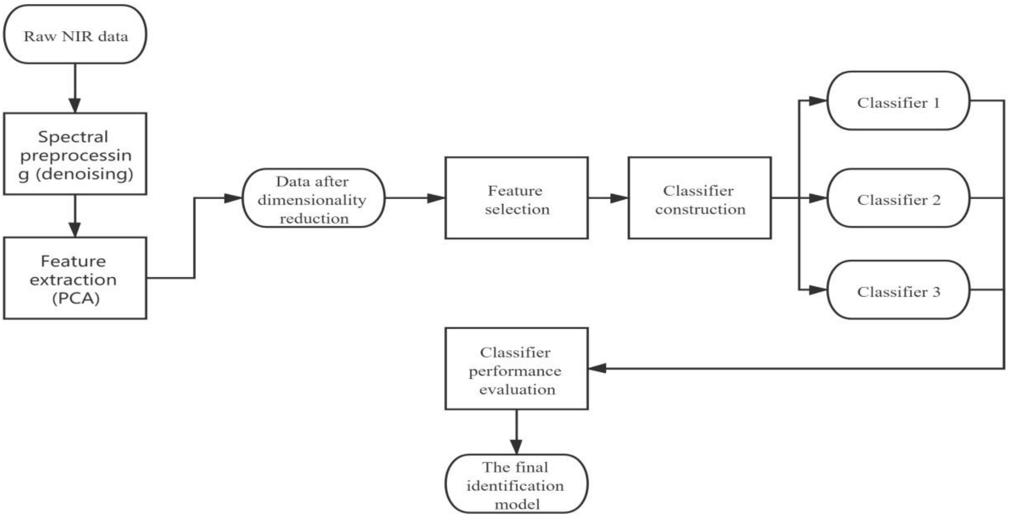
*Corresponding Author

shown in Figure 1. First, the preprocessing algorithm is used to shape the spectrum to reduce noise, thereby reducing the interference of the noise in the original data to the classifier; Secondly, the PCA method is used to extract the features of the denoised NIR spectrum, so as to reduce the dimension of the high-dimensional data to an appropriate dimension; Then, use the feature selection algorithm to perform proper feature selection on the reduced-dimensional spectral data to facilitate faster and more accurate learning of the classifier; Finally, choose different classifiers, and select the best classifier to build a spectral recognition model under a unified training framework and performance evaluation index (Ren, Wang, Ning, Xu, Wang, Xing, Wan & Zhang, 2013; Zhao, Guo, Wei & Bo, 2013; Shen, Zou, Shi, Li, Huang & Xu, 2015; Wang, Yan & Yang, 2015; Asir, Appavu & Jebamalar, 2016; El-Bendary, El, Hassanien & Badr, 2015).

2.1 Spectral Denoising

The near-infrared spectrum data is usually collected manually by a spectrum analyzer. It is inevitable that there will be a lot of noise. Their sources include background noise, stray light, light scattering, internal response of the instrument, manual operation errors, and so on. These noises will cause a certain amount of interference to the real data, leading to the loss of data characteristics and even the baseline drift of the near-infrared spectrum. Therefore, denoising the collected original near-infrared spectra is a very important part in the process of spectral analysis and testing.Common spectral denoising algorithms include smoothing, N-order derivatives, numerical normalization (maximum and minimum normalization), Standard Normal Variants (SNV)(Magwaza, Opara, Terry, Landahl & Bart, 2016), Multiplicative Scatter Correction (MSC))(Teye, Huang, Sam-Amoah, Takrama, Boison, Botchway & Kumi, 2015) and Orthogonal Signal Correction (OSC)(Xudong, Hailiang & Yande, 2016) and so on. Among them, the smoothing and derivative methods are more conducive to removing background noise and baseline interference to improve the signal-to-noise ratio. Normalization is often used for the processing of samples with different mixing degrees. Standard normal variable transformation and multivariate scattering correction are mainly used to eliminate the influence of solid particle size, surface scattering and optical path changes on the diffuse reflectance spectrum (Bao, Liu, Kong, Sun, He & Qiu, 2017; Hue, Gunata, Bergounhou, Assemat, Boulanger & Sauvage, 2014; Wang, Hu & Xie, 2014; Srivichien, Terdwongworakul & Teerachaichayut, 2015).

Figure 1. Frame diagram of NIR spectrum origin identification and identification based on machine learning


Raw NIR data
Spectral preprocessin g (denoising)
Feature extraction (PCA)
Data after dimensionality reduction
Feature selection
Classifier construction
Classifier 1
Classifier 2
Classifier 3
Classifier performance evaluation
The final identification model

In this article, we mainly use the Savitzky-Golay convolution smoothing method to preprocess the collected orange spectra (Sánchez, De, Serrano & Pérez-Marín, 2016). Savitzky-Golay convolution is a polynomial regression algorithm implemented by using a local sliding window. The principle is to realize by replacing the original interfered spectrum value with the average value. The mean value is not simply obtained by averaging, but a polynomial fitting that is most similar to the real signal for the value to be replaced through a local moving window, so that the originally disturbed spectral value can be corrected to a more reasonable signal value. The essence of the approximation process can be regarded as a weighted average method (Pissard, Fernández, Baeten, Sinnaeve, Lognay, Mouteau, Dupont,Rondia & Lateur, 2016; Moscetti, Haff, Stella, Contini, Monarca, Cecchini & Massantini, 2017; Lu, Castillo, Chiang & Edgar, 2014; Kucheryavskiy & Lomborg, 2015).

In this article, we use three denoising algorithms: SG smoothing method, first-order and second-order SG smoothing derivatives, to obtain approximate denoised orange NIR spectral data, so that the identified model can express the characteristics of the NIR spectrum of orange more accurately.

2.2 Feature Extraction Based on Principal Component Analysis

In chemical and spectral analysis, Principal Component Analysis (PCA) is one of the common feature extraction methods used in high-dimensional data (de Oliveira, Bureau, Renard, Pereira-Netto & de Castilhos, 2014). In this article, through PCA operation, we can replace thousands of original NIR spectral data with a small amount of data Principal Components (PCs). These new features are linear combinations of the previous original data. At the same time, these linear combinations also maximize the variance of the sample as much as possible, so that the new features are not correlated with each other, so that the discriminatory classifier can learn various samples more accurately. In order to retain more data, a total of 1500 data of the original spectral distribution from the 1000-2500nm spectral range were collected and used to establish the identification model.

In fact, the value of the principal component can be regarded as the result of projecting the points on the data set in the direction of the feature vector. The feature vector determines the degree of correlation between the principal component and the original data set. After projection, the information in the original feature space is concentrated in the larger part of the PCA result. At the same time, the noise information is also averaged into each feature vector. Therefore, the PCA method reduces the dimensionality of the data while removing redundant information. The larger the feature value corresponding to the principal component, the more important the information it expresses. The principal components corresponding to the smaller eigenvalues can be ignored in practical applications. After selecting the number of principal components, the discriminant classifier can be fully trained without being disturbed by too many principal components with noise. In this article, we will build different models for different principal components, so as to select the best number of principal components to achieve the highest recognition accuracy.

2.3 Feature Selection Based on Information Entropy

After PCA dimensionality reduction, we have obtained the most effective feature for expressing the original spectral data. However, the PCA method does not have classification information, which means that it only considers the projection direction that can retain the most original data. When these different types of data are projected in one direction, the distinguishing features may be confused. For example, in the commonly used character recognition, when we use the PCA method to recognize the letters “Q” and “O”, because PCA more considers the commonality of the two types of letters. Therefore, when determining the projection direction, the “tail” feature of the key distinguishing part of Q may be discarded, which reduces the recognition degree(Park, Liu, Philip, Jeong & Jeong, 2016; El-Bendary, El Hariri, Hassanien & Badr, 2015).

The purpose of feature selection is to select the most discriminative features in the feature space, thereby improving the performance of the classifier. Feature selection can be defined as: from the N features of a given sample, select a subset containing M features, M<N. When using the selected

feature subset M for classification training and prediction, under a certain evaluation index, the classification accuracy is higher than the previous classifier using N features(Fayoumi & Hajjar,2020; Cheng, Zhang, Wang, Zhao, Yu, Wang & de Pablos, 2021).

Feature selection can be roughly divided into four basic steps: ① Generation of feature subset: determine how to generate the next feature subset to be selected; ,Evaluation criteria: judge whether the selected feature is better than the existing feature set; ƒStop condition: decide when to terminate the feature selection process; „Verify feature set: verify whether the selected subset is valid. The process is shown in Figure 2.

In the original feature set, there are usually N different features, so the total number of candidate feature subsets can be obtained as 2^N (Fiorini, 2020). If verification is performed on each subset, even if N is not large, it often takes a lot of time. Therefore, the generation of feature subsets and evaluation indicators are the key points in feature selection. In this article, we use a feature selection algorithm based on information entropy to filter the spectral information after dimensionality reduction, so as to select the data that is most conducive to the training of the classifier.

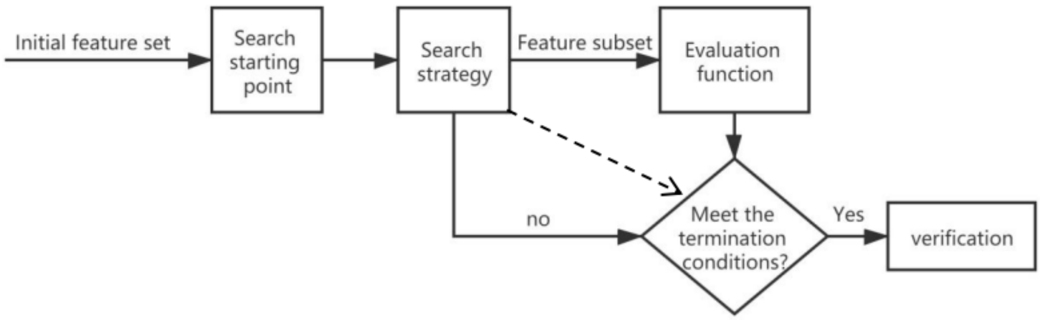
2.4 Discriminative Classifier

The construction of classifiers is the core content of machine learning theory and the main means to realize near-infrared spectroscopy analysis based on machine learning(Pandey & Banerjee, 2019; Gangadhar, Hota, Rao & Rao, 2019). Through the learning of the selected features, the classifier can construct a suitable identification model of the origin, so as to realize the classification of the origin of orange. Different classifiers often have different characteristics(Al-Qerem, Alauthman, Almomani & Gupta, 2020). Even if the same data set is used for training, the results are often different. In this article, we mainly use common classification algorithms in data mining to build our origin identification model. These include naive Bayes, decision trees, K-nearest neighbors, and linear discriminant methods(Wang, Li, Li, Gupta & Choi, 2020). In the experimental results stage, we will compare the performance of these classifiers.

2.5 Classifier Cross-Validation and Performance Evaluation

In order to test the performance of the classifier, we randomly divide the sample set into two sets of training samples and test samples. For the same classifier, using different data sets for training may produce different results. In order to judge the performance of each classifier more accurately, in the origin identification framework proposed in this paper, we adopt the M×N cross-validation method. That is, first divide the data set into N equal parts, and then use the data of N-1 data subsets for training. Use the remaining subset for testing. The above process traverses M times to ensure that each data set participates in training and testing. At the same time, the entire experimental process has to be repeated M times to ensure that the impact of sample division is minimal. Finally, when

Figure 2. Basic process of feature selection



the test on the test set ends, we can train M×N models and get M×N test results. In the performance comparison, the average value of M×N results is taken as the final performance index of the model. In the experiment, this article sets M=10, N=5, that is, each model has to perform 50 different tests on each data set to obtain the most accurate evaluation, thereby reducing the error caused by sample division.

In the selection of performance indicators, we use the correct rate to evaluate the overall performance of the identification model:

$$P_a = \frac{n_r}{N_t}$$

Among them, n_r is all samples that are correctly classified, and N_t is the total number of test samples.

Since in practice the distribution of samples is unknown, except for the correct rate. Therefore, we also need some other indicators to summarize the performance of the classifier. One of the most commonly used methods is to list the confusion matrix of the classification results, and consider a dichotomous problem. That is, the instance is divided into positive or negative samples. The four possible situations are shown in Table 1.

Among them, n_{TP} is the number of samples that correctly classify positive samples; n_{FN} is the number of all positive samples that are classified into negative samples. n_{FP} is the number of negative samples divided into positive samples. n_{TN} is the number of correctly identified negative samples.

According to the above confusion matrix, we can get different evaluation criteria. The main indicators used in this article include:

$$TPR = \frac{n_{TP}}{n_{TP} + n_{FN}}$$

$$FPR = \frac{n_{TN}}{n_{FP} + n_{TN}}$$

$$F_1 = \frac{2n_{TP}}{2n_{TP} + n_{FP} + n_{FN}}$$

Among them, TPR is true positive rate, also known as sensitivity. It represents the ratio of the positive samples correctly identified by the discrimination classifier to the total number of all positive samples; FPR is the false positive rate, also called specificity. It calculates the ratio of the number of samples that the classifier misrecognizes negative samples as positive instances to the total number of all negative samples; the F1 value is a comprehensive indicator of the two, which can be understood as the weighted average of the two indicators. When the classifier can correctly identify

Table 1. Confusion matrix for classification results

		Forecast Result	
		Positive Sample	**Negative Sample**
Actual results	**Positive sample**	n_{TP}	n_{FN}
	Negative sample	n_{FP}	n_{TN}

all test samples, the above indicators all get the maximum value of 1. When the classifier classifies all samples wrongly, the above indicators are all 0.

In the experiments of this article, we will select the most suitable classification model for the identification of orange origin after multiple cross-validation based on the above performance indicators.

3. EXPERIMENTAL RESULTS AND ANALYSIS

In the experiment, we selected the common naive Bayes, nearest neighbor classification (KNN) and decision tree algorithm as the test classifier [13-14]. The origin of orange collected from 16 different regions in 6 provinces was identified. The original near-infrared spectrum ranges from 1000-2499nm. The original feature dimension is 1500 dimensions. Approximately 100 orange samples were collected from each region. The total sample size is 1558. According to the identification framework, the original spectral data is preprocessed, feature extraction, feature selection, and model cross-validation to get the final performance evaluation.All simulation experiments are implemented using Matlab 2008b under the Windows 7 platform. Used statistical toolbox and data mining toolbox.

3.1 Original Spectra and Preprocessing Results

Taking into account the inevitable noise caused by near-infrared spectroscopy instruments, experimental environment and operating errors, it is very necessary to preprocess the original data to remove noise interference. We use the SG smoothing method to reshape the spectrum. The SG smoothing is performed under a window of 121 size, and the original SG smoothing and the first and second derivatives derived on this basis are used. The three denoising methods and the information of the original spectrum are shown in Figure 3.

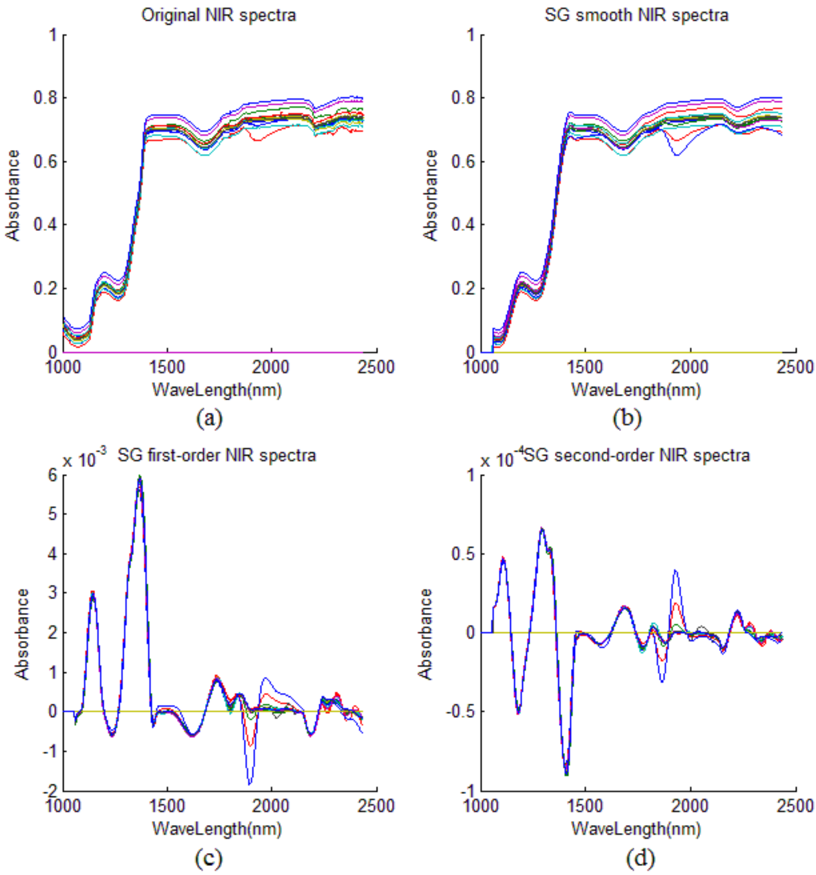
The original orange NIR spectrum absorbance is shown in Figure 3(a). The remaining three smoothing methods are shown in Figure 3(b)-(d). It can be seen that after SG smoothing, the original spectrogram becomes smoother. After performing the first derivative operation, the spectral range is compressed from [0,1] to [-0.002,0.006]. At the same time, the spectral signal is further smoothed. From the results of the second derivative, the smoothing effect is close to that of the first derivative. But the data is further compressed. Its range is reduced to [-0.00009] to [0.00007]. Although the derivative operation can further smooth the data, it may also lose some discriminative details. Therefore, the denoising preprocessing operation needs to be appropriately selected. The classifier performance obtained by different denoising algorithms will be discussed in the following subsections. It is worth noting that, as can be seen from Figure 3, the spectra of orange samples from 16 regions have great overlap. It will be very challenging to directly use these data (1500 dimensions) for identification.

3.2 Feature Extraction Results

It can be seen from the experiment in the previous section that the denoised data is not suitable for direct training of the classifier. They need to perform proper feature extraction in order to extract the main information and remove unnecessary redundant information. The PCA method is used in the recognition framework to extract the principal components of the spectrum. Because there is not enough evidence to show that a certain spectrum has a strong degree of discrimination, the principal component extraction is performed on the entire spectrum (1000-2499nm) to obtain the most representative spectrum information. The results sorted by the contribution of the principal components are shown in Figure 4.

Generally speaking, the number of principal components required to establish a model is often determined by the proportion of the spectral information occupied by the first few most representative principal components. As shown in Figure 4, the histogram represents the contribution of the principal component (that is, the proportion of the information contained in the entire data set). The red dots represent the cumulative contribution of the first N principal components. It can be seen from the figure that the first three principal components occupy a large proportion. For example, in Figure 4(a),

Figure 3. The original spectrum of orange and the effect after denoising. Among them, from left to right: (a) Original NIR spectrum; (b) SG smoothed effect diagram; (c) SG smoothed first derivative effect diagram; (d) SG smoothed second derivative effect diagram

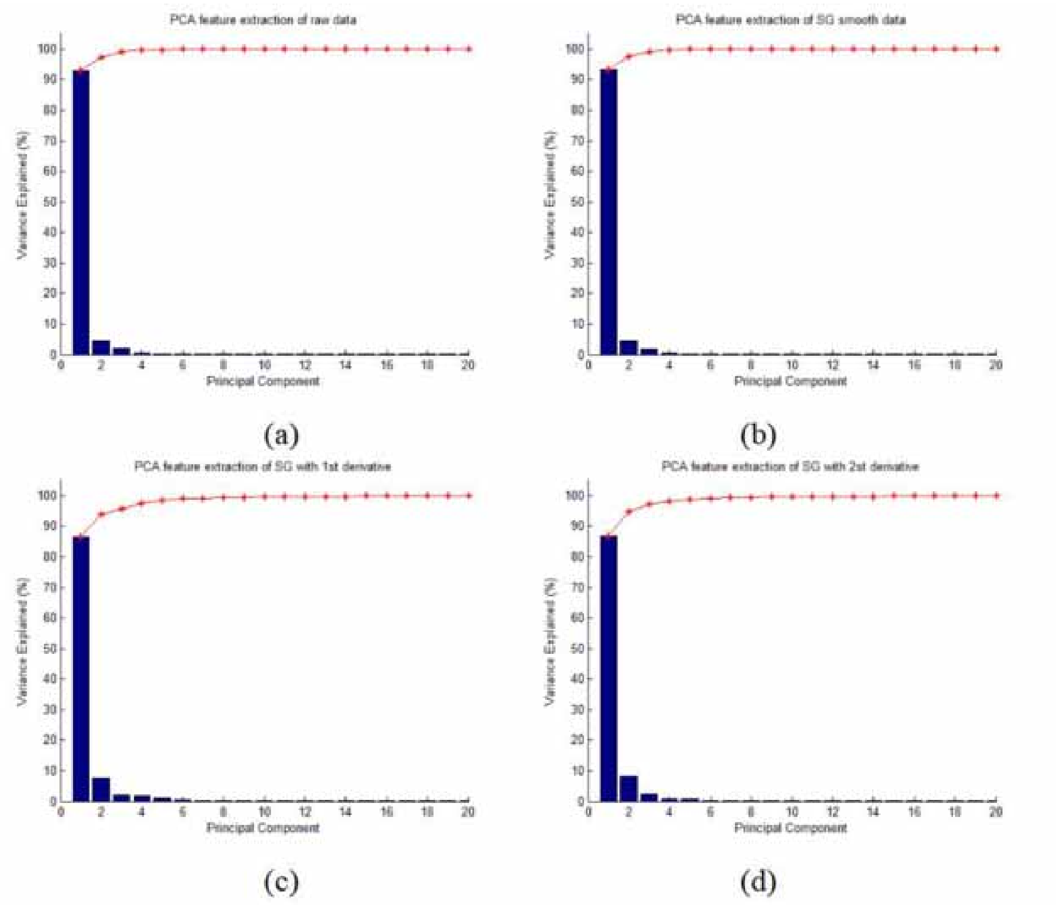


PCA dimensionality reduction is performed on the original spectral data. The first three principal components accounted for 98.98% of the information. The principal components are extracted from the SG smoothed data. The first three principal components account for 99.11% of the information. The principal components of the smoothed data after the first and second derivatives are extracted, and the first three principal components account for 95.17% and 97.16%, respectively.

Although the first three principal components can effectively represent the previous original data set, for the classifier, the information it represents may not be distinguishable. For example, the joint distribution of the original data and the first two principal components after using different smoothing algorithms. We use a scatter plot to represent. As shown in Figure 5. In order to better show its distribution characteristics, only 20 orange spectral samples from 5 different regions are drawn here. They include Wusheng in Sichuan, Linhai in Zhejiang, Wushan in Chongqing, Fengjie and Beibei.

It can be seen from Figure 5 that after PCA dimensionality reduction is performed on the original spectrum and the SG smoothed spectrum data, the PC distribution between different provinces has a certain degree of discrimination. The samples from the three different origins in Chongqing are close to each other. The growth environment of orange is similar. So there is a certain degree of overlap. After using SG smoothing combined with the first and second derivative methods, the distribution space of the sample is expanded. Thereby increasing the degree of dispersion between samples. But

Figure 4. Principal component contribution degree of orange NIR spectrum data after PCA feature extraction. Among them, from left to right: (a) PCA feature extraction from the original NIR spectrum; (b) SG smoothed data for PCA feature extraction; (c) Perform PCA feature extraction on the first derivative data after SG smoothing; (d) PCA feature extraction is performed on the second derivative data after SG smoothing



(a) (b) (c) (d)

it also further increases the area where the samples overlap. No matter which method is used, it is difficult to directly identify the first two PCs of orange samples. Therefore, more PC features can be appropriately added for training to increase its recognition. We take the first 20 PCs as training features and input them into the classifier.

3.3 Feature Selection and Classifier Performance Results

After data smoothing and principal component extraction, common classifiers in machine learning algorithms are mainly used. They include Decision Tree Algorithm (DT), Bayes Classifier (NB), K Nearest Neighbor Classifier (KNN) and Linear Discriminant Classifier (LDA). The origin identification model was established for orange samples from 16 regions in 6 provinces. According to the proposed origin identification framework, all classifiers have been cross-validated 5×10 times. And get the average recognition rate after 50 runs as the output result. The performance of each classifier is shown in Table 2.

First, in the absence of feature selection, Table 2 counts the average accuracy Pa of the four classifiers tested.

Figure 5. After PCA feature extraction is performed on orange NIRs in five regions, the distribution diagrams of the first and second principal components of the contribution. (a) Original NIR spectrum; (b) Use SG smoothing method; (c) Use the first derivative method after SG smoothing;(d) Second derivative method after SG smoothing

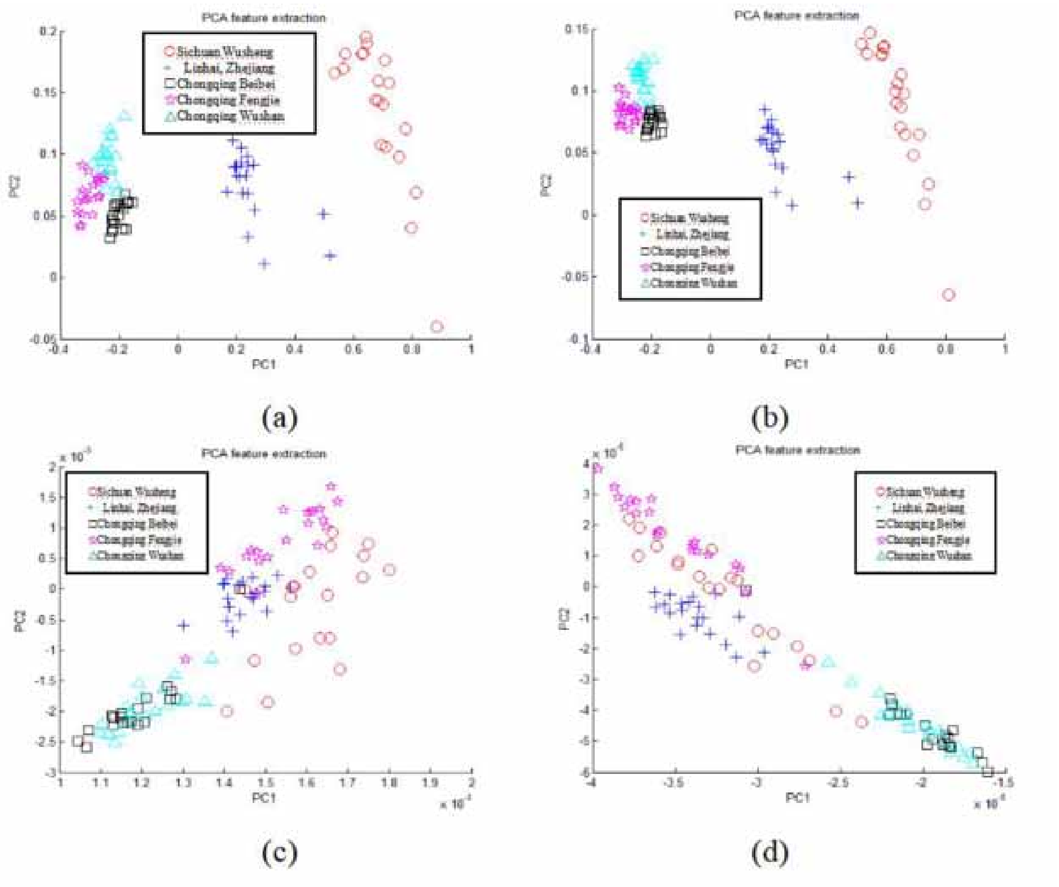


Table 2. When there is no feature selection, the average accuracy of origin identification of the four classifiers tested Pa(%)

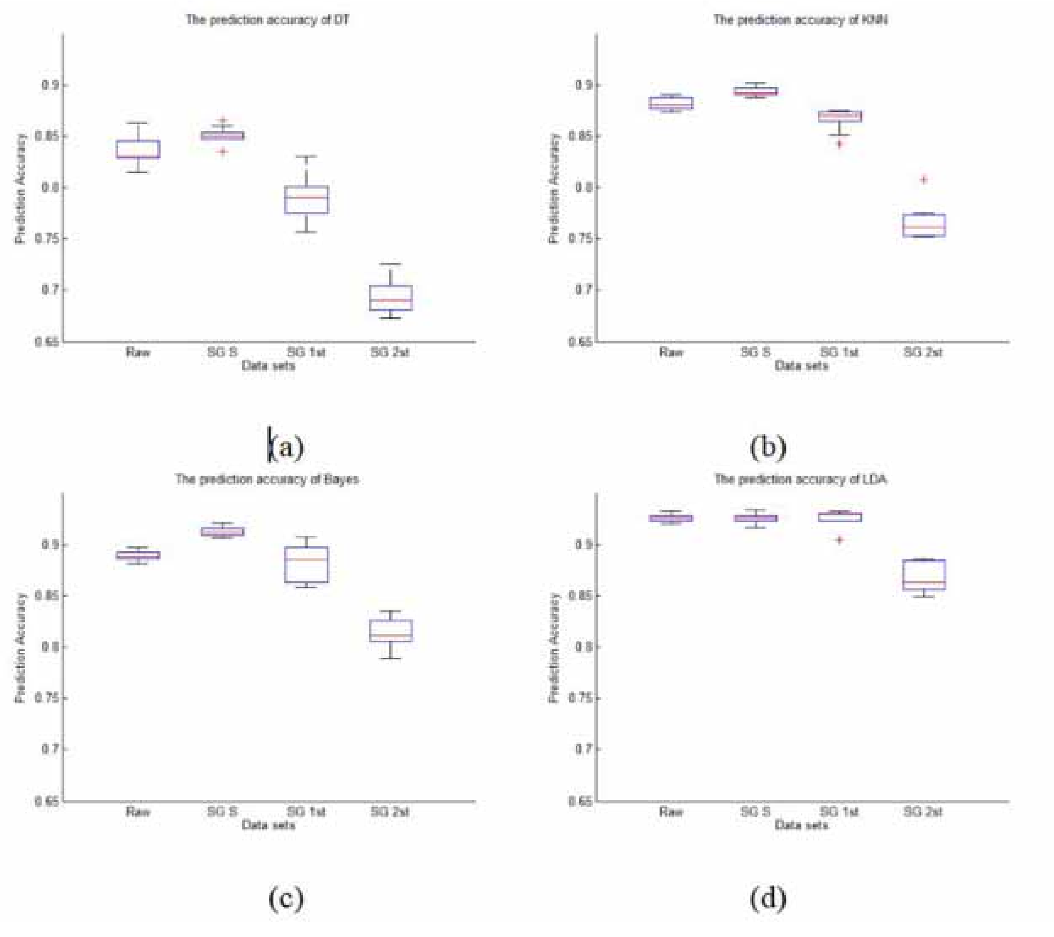
Classifier	Raw data	SG smooth	SG smooth 1st derivative	SG smooth 2nd derivative
DT	83.5	85.1	79.0	69.4
KNN	88.2	89.3	86.6	76.6
NB	88.8	91.3	88.2	81.3
LDA	92.6	92.5	92.6	86.7

Note: DT stands for decision tree, NB stands for Bayesian classifier, KNN is the nearest neighbor; LDA is linear discrimination.

Among them, the bold part in the table represents the best classifier result in the test. As can be seen from Table 2, the LDA classifier performs best on each data set. The highest average accuracy rate is 92.6%. Followed by KNN and NB classifiers. In terms of data smoothing algorithm, compared with the original data set, the performance of DT, NB and KNN classifiers have been significantly improved on the data set smoothed by SG. The LDA algorithm hardly changes. However, the method of combining the derivative after SG smoothing actually reduces the recognition accuracy. In particular, the more derivative orders, the worse the effect. The reason may be that too much smoothing leads to the loss of distinguishing features.

In order to further show the performance and stability of the 50 test classifiers in cross-validation, this paper draws the box plots of the four classifiers under different smoothing algorithms. As shown in Figure 6. In the box plot[14], the distribution of the data is expressed by five numerical points. They are the minimum (min), the lower quartile (Q1), the median (median), the upper quartile (Q3) and the maximum (max). As shown in the figure, the lower edge of the box represents Q1, and the upper edge represents Q3. The box part represents the interquartile range (IQR), that is, the middle 50% value of the observation data, and the middle horizontal line is the average value. The straight part extending from the edge of the box is called the tentacles. The outward extension of the tentacles represents the largest and smallest in the data set. The asterisk (*) in the figure represents an abnormal value, which is defined as a point that deviates from the box (greater than Q3 or less than Q1) and is more than 1.5 times away from the corresponding boundary. The length of the box and the number of abnormal points reflect the degree of dispersion of the data. It can also reflect the distribution and stability of the predicted results of the tested classifier after different training sets and test sets are used in cross-validation.

Figure 6. Box plots of the accuracy of the four classifiers tested using different smoothing algorithms. (a) Decision tree model; (b) KNN model; (c) Bayesian model; (d) LDA model.



It can be seen from Figure 6 that after using SG smoothing, most of the classifiers generally achieve the highest prediction accuracy (except for LDA and the original data are the same), and the most stable. After the first and second derivatives are used, the data is excessively smoothed, which affects its stability.

In addition to accuracy, this article also counts other performance indicators. Such as: sensitivity (TPR), specificity (FPR) and comprehensive index F1. The results are shown in Table 3.

The results in Table 3 are similar to those in Table 2. In terms of performance indicators, LDA still has the highest recognition rate. The recognition results of DT, KNN and NB classifiers on the SG smoothed data set are better.

After further selection of the features after PCA dimensionality reduction, we cross-validated the same classifier and data set. The results are shown in Table 3. After feature selection, the LDA model still obtains the highest recognition accuracy. But the improvement is not obvious compared to before feature selection. The reason is that LDA has considered the characteristic projection direction with the greatest discrimination when seeking the best projection direction.

Compared with other models, the performance before feature selection has been significantly improved. The specific comparison diagram is shown in Figure 7. It can be seen from Figure 7 that after adopting feature selection, both KNN and NB have reached a high degree of recognition (^{3}90%). Especially on the data set smoothed by the second derivative method, the four classifiers tested have been greatly improved. The most improved are the DT and KNN models. The average accuracy rates have increased from 69.4% and 76.6% to 80.4% and 88.0%, respectively.

Finally, Table 5 shows the results based on sensitivity (TPR), specificity (FPR) and comprehensive index F1 after feature selection. It can be seen that after feature selection, KNN and NB have achieved similar performance to LDA. The recognition effect of the DT model has also been significantly improved. However, LDA has not improved much, and the performance difference of each data set is not obvious.

Table 3. When there is no feature selection, the average sensitivity, specificity and comprehensive index F1 value of the four classifiers tested

Data set	Evaluation index	DT	KNN	NB	LDA
Raw data	TPR	78.5	83.3	83.9	87.0
	FPR	78.5	82.6	83.6	87.0
	F1	78.3	82.8	83.6	87.0
SG smooth	TPR	79.9	84.3	85.9	86.9
	FPR	79.9	83.7	85.9	87.0
	F1	79.7	83.9	85.7	86.9
SG smooth 1st derivative	TPR	75.0	81.9	83.6	87.0
	FPR	74.3	81.5	83.0	87.0
	F1	74.4	81.5	83.0	86.9
SG smooth 2nd derivative	TPR	66.3	74.2	77.3	82.2
	FPR	65.8	72.2	77.0	81.8
	F1	65.6	72.8	76.9	81.8

Note: DT stands for decision tree, NB stands for Bayesian classifier, KNN stands for nearest neighbor and LDA stands for linear discriminant classifier.

Table 4. After feature selection, the average accuracy of origin identification of the four classifiers tested Pa

Classifier	Raw data		SG smooth		SG smooth 1st derivative		SG smooth 2nd derivative	
DT	88.8	+5.3	89.1	+4.0	86.3	+7.3	80.4	+11.0
KNN	91.5	+3.4	91.5	+2.1	90.4	+3.8	88.0	+11.4
NB	91.3	+2.5	92.0	+0.8	91.7	+3.5	90.2	+9.0
LDA	92.5	-0.1	92.7	+0.2	92.8	+0.2	92.4	+5.6

Note: The data on the right is the result of comparing the classifiers without feature selection. The "+" sign indicates an improvement compared to before. "-" means that the recognition rate has decreased.

Table 5. After feature selection, the average sensitivity, specificity and comprehensive index F1 value of the four classifiers tested

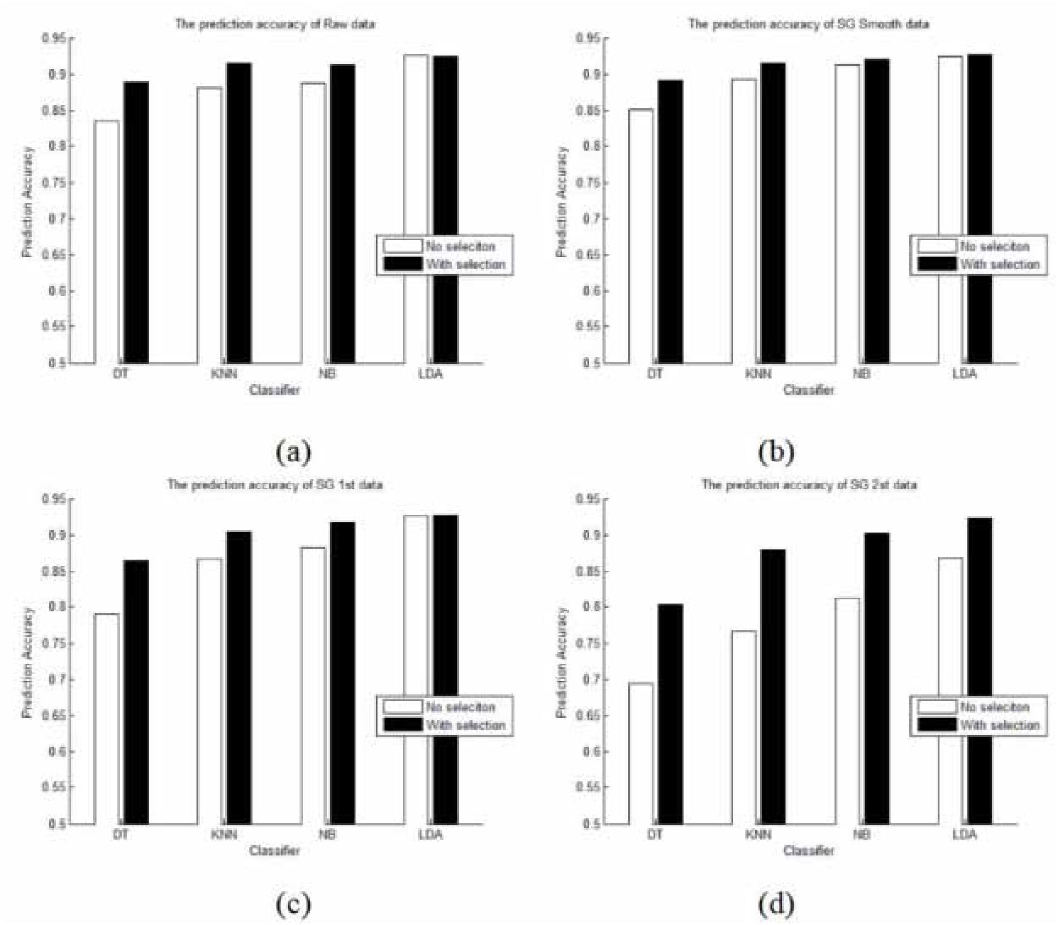
Data set	Evaluation index	DT	KNN	NB	LDA
Raw data	TPR	83.6	86.1	85.8	86.8
	FPR	83.4	85.7	85.8	86.9
	F1	83.4	85.8	85.8	86.8
SG smooth	TPR	83.9	86.0	86.6	87.1
	FPR	83.7	85.6	86.6	87.1
	F1	83.7	85.7	86.4	87.1
SG smooth 1st derivative	TPR	81.3	85.2	86.1	87.0
	FPR	81.0	85.1	86.3	87.2
	F1	81.0	85.0	86.1	86.9
SG smooth 2nd derivative	TPR	76.1	83.2	84.9	86.8
	FPR	75.8	82.8	84.9	86.8
	F1	75.8	82.9	84.8	86.8

Note: DT stands for decision tree, NB stands for Bayesian classifier, KNN stands for nearest neighbor, and LDA stands for linear discrimination.

4. SUMMARY

Aiming at the identification of orange spectral origin, this paper proposes a general identification framework and identifies the origin of orange samples under this framework. First, use SG smoothing method and SG smoothing combined with the first and second derivative methods to smooth the data. And use PCA to reduce the dimensionality of the data to extract the most representative features. After that, the feature selection algorithm of information entropy is used to further select the extracted features with the most discriminative degree. Finally, the decision tree, nearest neighbor, naive Bayes and linear discriminant analysis model are used to establish the origin identification model based on the orange data of 16 regions. Experimental results show that the SG smoothing algorithm can increase the recognition results of most classifiers. Feature selection algorithms also have a positive effect on the classification results of orange origins. Among the tested classifiers, the performance of LDA is the most stable. And obtained the best origin identification accuracy of 92.8%.

Figure 7. A histogram of the comparison of the effects of the four classifiers tested before and after feature selection. (a) Original NIR spectral data; (b) Use SG smoothing method; (c) Use SG smoothed first derivative method; (d) SG smoothed second derivative method.



(a)

(b)

(c)

(d)

ACKNOWLEDGMENT

This research was supported by Scientific and Technological Research Program of Chongqing Municipal Education Commission (Grant No. KJQN20191620).

REFERENCES

Al-Qerem, A., Alauthman, M., Almomani, A., & Gupta, B. B. (2020). IoT transaction processing through cooperative concurrency control on fog–cloud computing environment. *Soft Computing*, *24*(8), 5695–5711. https://doi.org/10.1007/s00500-019-04220-y

Asir, D., Appavu, S., & Jebamalar, E. (2016). Literature Review on Feature Selection Methods for High-Dimensional Data. *International Journal of Computer Applications,* 136. 10.5120/ijca2016908317

Bao, Y., Liu, F., Kong, W., Sun, D.-W., He, Y., & Qiu, Z. (2017). Measurement of soluble solid contents and pH of white vinegars using VIS/NIR spectroscopy and least squares support vector machine. *Food and Bioprocess Technology*, *7*(1), 54–61. https://doi.org/10.1007/s11947-013-1065-0

Caligiani, A., Palla, L., Acquotti, D., Marseglia, A., & Palla, G. (2014). Application of 1 H NMR for the characterisation of cocoa beans of different geographical origins and fermentation levels. *Food Chemistry*, *157*, 94–99. doi:10.1016/j.foodchem.2014.01.116 PMID:24679756

Cao, D.-S., Liang, Y.-Z., Xu, Q.-S., Hu, Q.-N., Zhang, L.-X., & Fu, G.-H. (2017). Exploring nonlinear relationships in chemical data using kernel-based methods. *Chemometrics and Intelligent Laboratory Systems*, *107*(1), 106–115. doi:10.1016/j.chemolab.2011.02.004

Chen, H., Lin, Z., Wu, H., Wang, L., Wu, T., & Tan, C. (2015). Diagnosis of colorectal cancer by near-infrared optical fiber spectroscopy and random forest. *Spectrochimica Acta. Part A: Molecular and Biomolecular Spectroscopy*, *135*(0), 185–191. doi:10.1016/j.saa.2014.07.005 PMID:25064501

Cheng, Y., Zhang, X., Wang, X., Zhao, H., Yu, Y., Wang, X., & de Pablos, P. O. (2021). Rethinking the Development of Technology-Enhanced Learning and the Role of Cognitive Computing. *International Journal on Semantic Web and Information Systems*, *17*(1), 67–96. https://doi.org/10.4018/IJSWIS.2021010104

de Oliveira, G. A., Bureau, S., Renard, C. M.-G. C., Pereira-Netto, A. B., & de Castilhos, F. (2014). Comparison of NIRS approach for prediction of internal quality traits in three fruit species. *Food Chemistry*, *143*, 223–230. https://doi.org/10.1016/j.foodchem.2013.07.122

Diniz, P. H. G. D., Gomes, A. A., Pistonesi, M. F., Band, B. S. F., & Araújo, M. C. U. D. (2014). Simultaneous Classification of Teas According to Their Varieties and Geographical Origins by Using NIR Spectroscopy and SPA-LDA. *Food Analytical Methods*, *7*(8), 1712–1718. doi:10.1007/s12161-014-9809-7

El-Bendary, N., El Hariri, E., Hassanien, A. E., & Badr, A. (2015). Using machine learning techniques for evaluating tomato ripeness. *Expert Systems with Applications*, *42*(4), 1892–1905. https://doi.org/10.1016/j.eswa.2014.09.057

El-Bendary, N., El Hariri, E., Hassanien, A. E., & Badr, A. (2015). Using machine learning techniques for evaluating tomato ripeness. *Expert Systems with Applications*, *42*(4), 1892–1905. https://doi.org/10.1016/j.eswa.2014.09.057

Fayoumi, A. G., & Hajjar, A. F. (2020). Advanced learning analytics in academic education: Academic performance forecasting based on an artificial neural network. *International Journal on Semantic Web and Information Systems*, *16*(3), 70–87. https://doi.org/10.4018/IJSWIS.2020070105

Fiorini, R. A. (2020). Computational intelligence from autonomous system to super-smart society and beyond. *International Journal of Software Science and Computational Intelligence*, *12*(3), 1–13. https://doi.org/10.4018/IJSSCI.2020070101

Gangadhar, P., Hota, A. K., Rao, M. V., & Rao, V. V. (2019). Performance of memory virtualization using global memory resource balancing. *International Journal of Cloud Applications and Computing*, *9*(1), 16–32. https://doi.org/10.4018/IJCAC.2019010102

Hue, C., Gunata, Z., Bergounhou, A., Assemat, S., Boulanger, R., & Sauvage, F. X. (2014). Near infrared spectroscopy as a new tool to determine cocoa fermentation levels through ammonia nitrogen quantification. *Food Chemistry*, *148*, 240–245. https://doi.org/10.1016/j.foodchem.2013.10.005

Kucheryavskiy, S., & Lomborg, C. J. (2015). Monitoring of whey quality with NIR spectroscopy—A feasibility study. *Food Chemistry*, *176*(0), 271–277. https://doi.org/10.1016/j.foodchem.2014.12.086

Li, G., Nunes, L., Wang, Y., Williams, P. N., Zheng, M., Zhang, Q., & Zhu, Y. (2013). Profiling the ionome of rice and its use in discriminating geographical origins at the regional scale. *Journal of Environmental Sciences (China)*, *25*(1), 144–154. doi:10.1016/S1001-0742(12)60007-2 PMID:23586309

Lu, B., Castillo, I., Chiang, L., & Edgar, T. F. (2014). Industrial PLS model variable selection using moving window variable importance in projection. *Chemometrics and Intelligent Laboratory Systems*, *135*(0), 90–109. https://doi.org/10.1016/j.chemolab.2014.03.020

Magwaza, L. S., Opara, U. L., Terry, L. A., Landahl, S., & Bart, M. N. (2016). Evaluation of Fourier transform-NIR spectroscopy for integrated external and internal quality assessment of Valencia oranges. *Journal of Food Composition and Analysis, 31*(1), 144-154. 10.1016/j.jfca.2013.05.007

Moscetti, R., Haff, R. P., Stella, E., Contini, M., Monarca, D., Cecchini, M., & Massantini, R. (2017). Feasibility of NIR spectroscopy to detect olive fruit infested by Bactrocera oleae. *Postharvest Biology and Technology*, *99*(0), 58–62. https://doi.org/10.1016/j.postharvbio.2014.07.015

Pandey, A., & Banerjee, S. (2019). Test suite optimization using firefly and genetic algorithm. *International Journal of Software Science and Computational Intelligence*, *11*(1), 31–46. https://doi.org/10.4018/IJSSCI.2019010103

Park, J. I., Liu, L., Philip Ye, X., Jeong, M. K., & Jeong, Y.-S. (2016). Improved prediction of biomass composition for switchgrass using reproducing kernel methods with wavelet compressed FT-NIR spectra. *Expert Systems with Applications*, *39*(1), 1555–1564. https://doi.org/10.1016/j.eswa.2011.05.012

Pissard, A., Fernández Pierna, J. A., Baeten, V., Sinnaeve, G., Lognay, G., Mouteau, A., Dupont, P., Rondia, A., & Lateur, M. (2016). Non-destructive measurement of vitamin C, total polyphenol and sugar content in apples using near-infrared spectroscopy. *Journal of the Science of Food and Agriculture*, *93*(2), 238–244. https://doi.org/10.1002/jsfa.5779

Ren, G., Wang, S., Ning, J., Xu, R., Wang, Y., Xing, Z., Wan, X., & Zhang, Z. (2013). Quantitative analysis and geographical traceability of black tea using Fourier transform near-infrared spectroscopy (FT-NIRS). *Food Research International*, *53*(2), 822–826. doi:10.1016/j.foodres.2012.10.032

Sádecká, J., Jakubíková, M., Májek, P., & Kleinová, A. (2016). Classification of plum spirit drinks by synchronous fluorescence spectroscopy. *Food Chemistry*, *196*, 783–790. doi:10.1016/j.foodchem.2015.10.001 PMID:26593555

Sánchez, M.-T., De la Haba, M.-J., Serrano, I., & Pérez-Marín, D. (2016). Application of NIRS for nondestructive measurement of quality parameters in intact oranges during on-tree ripening and at harvest. *Food Analytical Methods*, *6*(3), 826–837. https://doi.org/10.1007/s12161-012-9490-7

Shen, T., Zou, X., Shi, J., Li, Z., Huang, X., & Xu, Y. (2015). Determination Geographical Origin and Flavonoids Content of Goji Berry Using Near-Infrared Spectroscopy and Chemometrics. *Food Analytical Methods*, *9*(1), 1–12. doi:10.1007/s12161-015-0175-x

Srivichien, S., Terdwongworakul, A., & Teerachaichayut, S. (2015). Quantitative prediction of nitrate level in intact pineapple using Vis–NIRS. *Journal of Food Engineering*, *150*(0), 29–34. https://doi.org/10.1016/j.jfoodeng.2014.11.004

Teye, E., Huang, X., Sam-Amoah, L. K., Takrama, J., Boison, D., Botchway, F., & Kumi, F. (2015). Estimating cocoa bean parameters by FT-NIRS and chemometrics analysis. *Food Chemistry*, *176*(0), 403–410. https://doi.org/10.1016/j.foodchem.2014.12.042

Wang, A., Hu, D., & Xie, L. (2014). Comparison of detection modes in terms of the necessity of visible region (VIS) and influence of the peel on soluble solids content (SSC) determination of navel orange using VIS–SWNIR spectroscopy. *Journal of Food Engineering*, *126*, 126–132. https://doi.org/10.1016/j.jfoodeng.2013.11.011

Wang, H., Li, Z., Li, Y., Gupta, B. B., & Choi, C. (2020). Visual saliency guided complex image retrieval. *Pattern Recognition Letters*, *130*, 64–72. https://doi.org/10.1016/j.patrec.2018.08.010

Wang, J.J., Yan, S.M., & Yang, B. (2015). Determination of Ginsenosides Amount and Geographical Origins of Ginseng by NIR Spectroscopy. *Guang pu xue yu guang pu fen xi = Guang pu, 35*(7). 10.3964/j.issn.1000-0593(2015)07-1885-04

Xudong, S., Hailiang, Z., & Yande, L. (2016). Nondestructive assessment of quality of Nanfeng mandarin fruit by a portable near infrared spectroscopy. *International Journal of Agricultural and Biological Engineering*, *2*(1). https://doi.org/10.3965/ijabe.v2i1.42

Zhao, H., Guo, B., Wei, Y., & Bo, Z. (2013). Near infrared reflectance spectroscopy for determination of the geographical origin of wheat. *Food Chemistry, 138*(2–3), 1902-1907. 10.1016/j.foodchem.2012.11.037

Songjian Dan is an associate professor at Chongqing University of Education of China. He received the Ph.D. degree in Communication Engineering from the Chongqing University, China in 2017. His research interests include machine learning and data mining.