

Origin identification of *Cornus officinalis* based on PCA-SVM combined model

Yueqiang Jin (✉ Jinyq@niit.edu.cn)

Nanjing Vocational University of Industry Technology

Bing Liu

Nanjing Vocational University of Industry Technology

Chaoning Li

Nanjing Changxingyang Intelligent Home Company Limited

Shasha Shi

Jiangsu Ocean University

Article

Keywords:

Posted Date: July 22nd, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1857964/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Abstract

Infrared spectroscopy can quickly and non-destructively extract analytical information from samples. It can be applied to the authenticity identification of various Chinese herbal medicines, the prediction of the mixing amount of defective products, and the analysis of the origin. In this paper, the spectral information of *Cornus officinalis* from 11 origins was used as the research object, and the origin identification model of *Cornus officinalis* based on mid-infrared spectroscopy was established. First, principal component analysis was used to extract the absorbance data of *Cornus officinalis* in the wavenumber range of 551~3998 cm^{-1} . The extracted principal components contain more than 99.8% of the information of the original data. Second, the extracted principal component information was used as input, and the origin category was used as output, and the origin identification model was trained with the help of support vector machine. In this paper, this combined model is called PCA-SVM combined model. Finally, the generalization ability of the PCA-SVM model is evaluated through an external test set. The three indicators of Accuracy, F1-Score, and Kappa coefficient are used to compare this model with other commonly used classification models such as naive Bayes model, decision trees, linear discriminant analysis, radial basis function neural network and partial least square discriminant analysis. The results show that the PCA-SVM model outperforms several other commonly used models in terms of Accuracy, F1-Score, or Kappa coefficient. Using this model to identify the origin of *Cornus officinalis*, the accuracy rate is 84.8%.

1. Introduction

In terms of climate and geography, China has favorable conditions for the growth of Chinese Herbal Medicines. It is the country with the largest variety of natural medicinal resources and the longest cultivation history in the world. There are many authentic medicinal materials in China that are well-known at home and abroad, and are exported to more than 100 countries and regions [1, 2]. Chinese Herbal Medicines have many unique advantages such as natural materials, stable action, and small toxic and side effects. With the progress of society and the development of productivity, many people are more willing to pay attention from synthetic drugs to natural drugs, hoping to develop safer and more effective new drugs from natural drugs to replace the former [3–6]. However, with the gradual deterioration of the natural ecological environment, the growing environment of wild Chinese Herbal Medicine has been destroyed, and the supply of some wild Chinese Herbal Medicine is in short supply, resulting in chaos in the sales market of Chinese Herbal Medicines. In addition, there are many types of Chinese Herbal Medicines, and different regions have different usage habits, and the confusion of Chinese Herbal Medicine is common, making the identification of Chinese Herbal Medicines a difficult problem [7, 8].

Chinese Herbal Medicines identification methods include character identification, microscopic identification, and physical and chemical identification [9, 10]. These identification methods are relatively inefficient, and it is difficult to obtain accurate identification results for Chinese Herbal Medicines with close kinship and high biological morphological similarity [11, 12]. Thin layer chromatography [13, 14], gas chromatography [15, 16], high performance liquid chromatography [17–20] and capillary electrophoresis [21–23] can achieve extremely high identification accuracy of medicinal materials, but they often require complex pretreatment, and the analysis time is long, the cost is high, and it is difficult to perform non-destructive rapid identification [24–27]. Infrared spectroscopy has the advantages of fast analysis speed, low cost, non-destructive, and simple pretreatment. In recent years, it has been widely used in the field of quality control of Chinese Herbal Medicines [28–31].

Ren, X., et al. et al. used ultra-violet spectroscopy and high performance liquid chromatography combined with chemometric tools to fast discrimination and quantification analysis of Ezhu from three botanical origins. The results show that supervised techniques such as decision tree, Linear Discriminant Analysis (LDA), Support Vector Machine (SVM) and k-nearest neighbor model can accurately identify the origin of Ezhu [32]. Lu, L., et al. used Fourier transform infrared spectra combined with pattern recognition technology for geographic identification of wild *Gentiana rigescens*

[33]. The comparison result showed that the Partial Least Squares Discriminant Analysis (PLS-DA) method is more suitable for geographic origin classification than Principal Component Analysis (PCA).

Each molecule has a unique infrared absorption spectrum determined by its composition and structure, which allows structural analysis and identification of the molecule. Infrared absorption spectrum is generated by the continuous vibration and rotational motion of the molecule. Molecular vibration refers to the relative motion of the atoms in the molecule near the equilibrium position. Polyatomic molecules can form a variety of vibrational images. When the atoms in the molecule make simple harmonic vibrations near the equilibrium position with the same frequency and the same phase, this vibration mode is called normal vibration (such as stretching vibration and variable angle vibration). The energy of molecular vibration corresponds exactly to the photon energy of infrared rays, so when the vibration state of the molecule changes, it can emit infrared spectrum, and it can also generate infrared absorption spectrum due to the vibration of the molecule excited by infrared radiation. The energy of vibration and rotation of molecules is not continuous but quantized. However, because the vibrational transition of molecules is often accompanied by rotational transitions, the vibrational spectrum is banded. Therefore, the infrared spectrum of the molecule is a band spectrum. The mid-infrared spectrum reflects the stretching and deformation vibrational motion of atoms in molecules, and the wavenumber range is between $400 \sim 4000 \text{ cm}^{-1}$. In recent years, mid-infrared combined with chemometrics has been widely used in the identification of origin and species of Chinese Herbal Medicines [34, 35].

Cornus officinalis is a commonly used Chinese medicinal material, and a set of mid-infrared spectral information of it was collected to establish the origin identification model of *Cornus officinalis*. There are 4 species of *Cornus* genus, which are naturally distributed in the junction of subtropical and north temperate zones at 33° - 37° N latitude and 105° - 135° E longitude. They are mainly distributed in central and southern Europe, eastern Asia and eastern North America [36]. There are two species of *Cornus* genus in China, namely *Cornus officinalis* Sieb. et Zucc. and *Cornus chinense* (Wanger) Hutch. The dried and ripe pulp of *Cornus officinalis* has the effects of nourishing the liver and kidney, internal heat and quenching thirst. It is a traditional medicine commonly used in Chinese Herbal Medicines to treat diabetes. *Cornus officinalis* is produced in Shanxi, Jiangsu, Zhejiang, Anhui, Jiangxi, Shandong, Henan, Hunan, Sichuan, Shaanxi and Gansu in China. It grows in forest margins or forests at an altitude of 400-1500m. In this study, based on the infrared spectral data of *Cornus officinalis* from 11 origins, a model for identifying the origin of *Cornus officinalis* was established by principal component analysis and support vector machine. We call this model the PCA-SVM combined model. Compared with other commonly used methods such as naive Bayes model, decision trees, LDA, Radial Basis Function (RBF) neural network and PLS-DA, PCA-SVM performs well in some common evaluation indicators. This model can not only provide a convenient and accurate method for the rapid identification of the origin of *Cornus officinalis*, but also has certain reference significance for the identification of other Chinese Herbal Medicines.

2. Material And Methods

2.1 Data source and preprocessing

Mid-infrared spectroscopy combined with chemometrics can be used for origin identification of Chinese Herbal Medicines. A set of spectral data of *Cornus officinalis* is used in this paper to establish a classification and recognition model of *Cornus officinalis*. This set of data comes from the 2021 China Undergraduate Mathematical Contest in Modeling (http://www.mcm.edu.cn/html_cn/node/4d73a36cc88b35bd4883c276afe39d89.html). In Table 1, the No column is the number of the sample, and the OP column is the origin of the sample. In other columns, the first row of data is the wave number of the spectrum, and the second row to the last row represent the absorbance of the sample under the spectral illumination of the corresponding band. The place of origin (OP 1 ~ 11) of 658 samples are known. The mid-infrared wave number range is $551 \sim 3998 \text{ cm}^{-1}$, with a total of 3448 bands.

Table 1

Mid-infrared spectral data of *Cornus officinalis* from 11 different places of origin.

No	OP	551	552	553	554	...	3996	3997	3998
1	11	0.33846	0.33846	0.33973	0.33973	...	0.05408	0.05408	0.05395
2	1	0.31183	0.31183	0.31221	0.31221	...	0.06327	0.06327	0.06320
3	6	0.35688	0.35688	0.35739	0.35739	...	0.04726	0.04726	0.04716
4	7	0.35823	0.35823	0.35863	0.35863	...	0.06719	0.06719	0.06710
5	8	0.48172	0.48172	0.48201	0.48201	...	0.09169	0.09169	0.09153
6	1	0.44655	0.44655	0.44820	0.44820	...	0.08685	0.08685	0.08665
7	7	0.31368	0.31368	0.31312	0.31312	...	0.06968	0.06968	0.06962
...	...	...	...	...	...	...	...	...	...
652	4	0.25195	0.25195	0.25171	0.25171	...	0.04253	0.04253	0.04244
653	10	0.33162	0.33162	0.33296	0.33296	...	0.07171	0.07171	0.07162
654	9	0.19467	0.19467	0.19549	0.19549	...	0.04188	0.04188	0.04184
655	6	0.28311	0.28311	0.28345	0.28345	...	0.01178	0.01178	0.01170
656	4	0.21455	0.21455	0.21423	0.21423	...	0.05252	0.05252	0.05249
657	5	0.31560	0.31560	0.31631	0.31631	...	0.00553	0.00553	0.00542
658	1	0.37229	0.37229	0.37259	0.37259	...	0.07194	0.07194	0.07183

Data preprocessing is the first step in data modeling. We consider data that is 3 times larger than the average of the left and right nearest neighbors as outliers. Outliers and missing values are interpolated by means of mean interpolation in this paper. After dealing with missing values and outliers, the absorbance data of 3448 corresponding bands under spectral illumination were analyzed and summarized. The range of absorbance after summary is -0.00675 ~ 1.48696 AU. There are some negative values of absorbance in the last 184 bands, and there are 626 groups of data with negative values in total. This is because the absorbance used in the data in this paper is the value corrected by the instrument. Since the absolute value of these negative values is small, and the total amount of negative values is less than 0.001% of the total amount of data, we keep them. The absorbance of some data is greater than 1 AU, but the maximum value of absorbance does not exceed 1.5 AU. We think that they do not deviate from the Lambert-Beer law, so no special treatment is required.

2.2 Data exploratory analysis

Exploratory analysis can give us a preliminary understanding of the spectral data of the sample. Figure 1 shows the mid-infrared spectra of 658 samples. It can be seen from the spectrum comparison that the mid-infrared spectrum of *Cornus officinalis* has obvious similarity, especially in the range of 1700 ~ 2500 cm^{-1} . In the two band ranges of 1000 ~ 1700 cm^{-1} and 3000 ~ 3400 cm^{-1} , there are mainly strong spectral peaks, and the peaks change drastically. This spectral region contains more chemical information. Five strong peaks appeared near the 1070, 1400, 1650, 2950 and 3300 cm^{-1} bands, with average absorbances of 0.734, 0.508, 0.781, 0.455 and 0.827 AU. The spectrum fluctuates greatly in the 1600 ~ 1700 cm^{-1} and 3250 ~ 3350 cm^{-1} bands, indicating that there are certain differences in the mid-infrared spectra

of *Cornus officinalis* from different origins. The difference in absorbance of different origins can be used to identify the origin of *Cornus officinalis* [19, 22].

In order to further compare the differences in mid-infrared spectra from different origins, we classified and summarized 658 samples by category, and averaged the absorbance under different wavelength bands. Figure 2 shows the average absorbance of *Cornus officinalis* from 11 origins in different wavelength bands. It can be seen that the spectral averages of *Cornus officinalis* from the 11 origins are very similar, and the bands where the spectral peaks appear are basically the same. In some wavelength bands, the absorbance of different origins is different, which is due to the different climate and geographical conditions of different origins. These differences lead to different contents of certain elements in *Cornus officinalis*.

The spectral differences reflect the differences in the composition of Chinese Herbal Medicines samples, which can be visually expressed by the degree of overlap of spectral lines in the overlapping spectral graphs. These intuitive differences can be quantitatively expressed by the Euclidean distance between spectral vectors, the cosine of the included angle, and the Pearson linear correlation coefficient [30, 37]. In this paper, the Pearson linear correlation coefficient (Eq. (1)) is used to describe the spectral differences of *Cornus officinalis* from different origins. Table 2 shows the correlation coefficients between spectral vectors of 11 *Cornus officinalis* from different origins. It can be seen that the spectral similarity between the medicinal materials from origin 7 and origin 11 is the highest, and the correlation coefficient is as high as 0.999. It shows that the composition of *Cornus officinalis* samples from the two origins is very similar, which is due to the similar climatic and geographical conditions of the two origins. The identification of these two origins by spectral analysis is the most difficult. The spectral similarity between the medicinal materials from origin 5 and origin 8 was the lowest, and the correlation coefficient was 0.982, which indicated that there were differences in the composition of TCM samples from these two origins. It is easier to identify *Cornus officinalis* from these two origins by spectral analysis. The spectral similarity of all *Cornus officinalis* between the origins exceeded 0.98, and at the significant level of 0.01, they all passed the correlation test, indicating that the spectral vectors of the same type of *Cornus officinalis* were significantly correlated.

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

1

Table 2

Pearson linear correlation coefficients of the spectral vectors of *Cornus officinalis* from 11 different places of origin (with ** denote significant correlation with a significance level of 0.01).

OP	OP 1	OP 2	OP 3	OP 4	OP 5	OP 6	OP 7	OP 8	OP 9	OP 10	OP 11
OP 1	1.000	0.996**	0.998**	0.990**	0.986**	0.997**	0.996**	0.997**	0.989**	0.990**	0.998**
OP 2		1.000	0.995**	0.995**	0.989**	0.997**	0.998**	0.997**	0.992**	0.994**	0.998**
OP 3			1.000	0.990**	0.985**	0.997**	0.996**	0.995**	0.991**	0.991**	0.997**
OP 4				1.000	0.998**	0.995**	0.996**	0.990**	0.998**	0.996**	0.996**
OP 5					1.000	0.992**	0.990**	0.982**	0.998**	0.996**	0.992**
OP 6						1.000	0.998**	0.995**	0.995**	0.993**	0.998**
OP 7							1.000	0.998**	0.994**	0.991**	0.999**
OP 8								1.000	0.986**	0.985**	0.996**
OP 9									1.000	0.997**	0.995**
OP 10										1.000	0.995**
OP 11											1.000

3. Establishment Of Origin Identification Model

3.1 Introduction to basic principles

Principal component analysis was first introduced by Karl Pearson for non-random variables, and then Harold Hotelling extended this method to the case of random vectors. Principal component analysis is performed with minimal loss of data information. It uses the method of mathematical transformation to convert the given multiple index factors into a few principal components, and then replaces the original multi-dimensional related variables with a few principal component factors [38]. Suppose the research on a certain problem involves p indicators, which are represented by $X_1, X_2, \dots, X_p$ respectively, and the p -dimensional random vector composed of these p indicators is $X = (X_1, X_2, \dots, X_p)'$. Let the mean of the random vector X be μ and the covariance matrix be Σ . Linear transformation of X can form a new comprehensive variable, which is represented by Y , that is, the new variable can be linearly represented by the original variable (Eq. (2)).

$$\begin{cases} Y_1 = u_{11}X_1 + u_{21}X_2 + \dots + u_{p1}X_p \\ Y_2 = u_{12}X_1 + u_{22}X_2 + \dots + u_{p2}X_p \\ \vdots \\ Y_p = u_{1p}X_1 + u_{2p}X_2 + \dots + u_{pp}X_p \end{cases} \quad (2)$$

The above linear transformation of the original variable can be carried out arbitrarily, and the statistical characteristics of the comprehensive variable Y obtained by different linear transformations are also different. Therefore, in order to achieve better results, we hope that the variance of $Y_i = u_i'X$ is as large as possible and each Y_i is independent of each other. Since $\text{var}(Y_i) = \text{var}(u_i'X) = u_i'\Sigma u_i$, and for any constant c , there is $\text{var}(cu_i'X) = c^2 u_i'\Sigma u_i$, so when there is no restriction on u_i , $\text{var}(Y_i)$ can be increased arbitrarily, and the problem will become meaningless. We constrain linear transformations to the following principles: (i) $u_i'u_i = 1$ ($i = 1, 2, \dots, p$); (ii) Y_i and Y_j are independent of each other ($i \neq j; i, j = 1, 2, \dots, p$); (iii) Y_1 is the one with the largest variance among all linear combinations of $X_1, X_2, \dots, X_p$ that satisfy the principle (i); Y_2 is the one with the largest variance among all the linear combinations of $X_1, X_2, \dots, X_p$ that are not related to Y_1 ; ...; and Y_p is the one with the largest variance among all linear combinations of $X_1, X_2, \dots, X_p$ that are not related to $Y_1, Y_2, \dots, Y_{p-1}$. The comprehensive variables $Y_1, Y_2, \dots, Y_p$ determined based on the above three principles are called the first, second, ..., p -th principal components of the original variables. In actual research work, only the first few principal components with the largest variance are usually selected, so as to simplify the system structure and grasp the essence of the problem.

$$\begin{aligned} & \max_{w,b} \frac{2}{\|w\|} \\ & s.t. y_i (w^T x + b) \geq 1, i = 1, 2, \dots, m \quad (3) \\ & \min_{w,b} \frac{1}{2} \|w\|^2 \\ & s.t. 1 - y_i (w^T x + b) \leq 0, i = 1, 2, \dots, m \quad (4) \end{aligned}$$

The support vector machine proposed by Vapnik et al. is an algorithm to find a classification plane or hyperplane that separates different types of data in the dataset as much as possible. Figure 3 is its architecture. SVM has the theory of structural risk minimization, and it still has strong robustness in the face of nonlinear datasets and higher-dimensional datasets, so it is widely used in classification algorithms [39]. The essence of SVM is to obtain the optimal parameters w and b to determine an optimal hyperplane, so that as much data as possible is distributed on both sides of this plane to achieve classification. Assuming that the training set is $(x_i, y_i), i = 1, 2, \dots, l, x \in R^n, y \in \{\pm 1\}$, the linear equation $w^T x + b = 0$ is used to divide it, where $w = (w_1; w_2; \dots; w_d)$ is the normal vector and b is the bias term. In order for the hyperplane to have maximum margin (Eq. (3)), it is only necessary to maximize $\|w\|^{-1}$, which is equivalent to minimizing $\|w\|^2$. Therefore, the problem of constructing the optimal hyperplane is transformed into Eq. (4).

Eqs. (5)–(6) are obtained by introducing Lagrange multipliers, where $\alpha = (\alpha_1; \alpha_2; \dots; \alpha_m)$. Eqs. (7)–(8) can be obtained by taking the partial derivatives of w and b of $L(w, b, \alpha)$. Substituting Eq. (6) into Eq. (5), w and b in $L(w, b, \alpha)$ can be eliminated, and then considering the constraints in Eq. (7), the dual problem of Eq. (3) can be obtained (Eq. (9)).

$$\begin{aligned} & \min_{w,b,\alpha} L(w, b, \alpha) \quad (5) \\ & L(w, b, \alpha) = \frac{1}{2} \|w\|^2 + \sum_{i=1}^m \alpha_i (1 - y_i (w^T x + b)) \quad (6) \\ & w = \sum_{i=1}^m \alpha_i y_i x_i \quad (7) \end{aligned}$$

$$0 = \sum_{i=1}^m \alpha_i y_i \quad (8)$$

$$\max_{\alpha} \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j x_i^T x_j$$

$$s. t. \sum_{i=1}^m \alpha_i y_i = 0, \alpha_i \geq 0, i = 1, 2, \dots, m \quad (9)$$

3.2 Mid-infrared spectral feature extraction

Mid-infrared spectroscopy can effectively identify the origin of *Cornus officinalis*. The mid-infrared spectral data collected in this study contains 3448 bands, and the absorbance data of each sample in each band is highly correlated. If all 3448 variables are introduced into the *Cornus officinalis* origin identification model, it will not only make the model training time very long, but also the introduction of highly correlated variables into the model will lead to poor stability and generalization ability. Therefore, it is necessary to use mathematical methods to extract features from the data. Principal component analysis is a common unsupervised analysis technique that is often used to extract features from complex data [32, 40].

Table 3 shows the results of principal component analysis of the mid-infrared spectral data of *Cornus officinalis*. It can be seen that the first principal component contains 80.8% of the information of the original data, and the first three principal components contain more than 95% of the information of the original data. Common methods for the selection of the number of principal components include the cumulative contribution rate criterion and the Kaiser criterion based on eigenvalues greater than 1. In order to introduce more variables into the model so that the model can be fully trained, this paper adopts the Kaiser criterion to select the number of principal components. The eigenvalues corresponding to the first 14 principal components are all greater than 1, and their cumulative contribution rate exceeds 99.8%, containing most of the information of the original variables. According to the Kaiser criterion, the first 14 principal components are selected to replace the original variables to establish the origin identification model of *Cornus officinalis*.

Table 3

Principal component eigenvalues, contribution rate and cumulative contribution rate of the mid-infrared spectral data of *Cornus officinalis*.

Serial number	Principal component	Eigenvalues	Contribution rate (%)	Cumulative contribution rate (%)
1	1st principal component	2785	0.808	0.808
2	2nd principal component	314.31	0.091	0.899
3	3rd principal component	174.91	0.051	0.950
4	4th principal component	77.257	0.022	0.972
5	5th principal component	29.467	0.009	0.981
6	6th principal component	18.32	0.005	0.986
7	7th principal component	15.433	0.004	0.990
8	8th principal component	9.94	0.003	0.993
9	9th principal component	7.689	0.002	0.995
10	10th principal component	2.614	0.001	0.996
11	11th principal component	2.174	0.001	0.997
12	12th principal component	1.761	0.001	0.997
13	13th principal component	1.446	0.000	0.998
14	14th principal component	1.113	0.000	0.998
15	15th principal component	0.928	0.000	0.998

3.3 PCA-SVM combined model construction

After feature extraction, the spectral data of 658 *Cornus officinalis* from 11 different origins are used to establish the origin identification model of *Cornus officinalis*. Since SVM can usually get better results than other algorithms on a small sample training set and has better robustness, it is used to establish a mid-infrared spectroscopy-based identification model for the origin of *Cornus officinalis* [15, 41, 42].

The 658 samples are randomly divided into two parts with a ratio of about 3:1. The first part contains 500 samples for training the model and the second part contains 158 samples for testing the model. Since the sample data comes from 11 different origins, and the number of samples in each origin is different, it is necessary to test the balance of the samples. Figure 4 shows the number of samples included in the training set, test set, and all sets under each origin category. It can be seen that the number of OP 4 in all sets is at most 88, accounting for 13.4% of the total number of samples, and the number of OP 5 is at least 31, accounting for 4.7% of the total number of samples. The number of OP 6 in the training set is at most 72, accounting for 14.4% of the total number of samples, and the number of OP 5 is at least 20, accounting for 4% of the total number of samples. The number of OP 4 in the test set is at most 24, accounting for 15.2% of the total number of samples, and the number of OP 9 is at least 7, accounting for 4.4% of the total number of samples. No matter which set, the ratio of the maximum number and the minimum number of samples does not exceed 4:1, so it is considered that there is no sample imbalance in this study.

A model between the origin identification of *Cornus officinalis* and its discriminant index was established, and the training samples are grouped by K-fold Cross-Validation (KCV) method. K-fold cross-validation is a statistical analysis method used to verify the performance of a classifier. Its basic idea is to group the original data, one part as the training set and the other part as the validation set. First train the classifier with the training set, and then use the validation set to test the trained model, which is used as the performance indicator for evaluating the classifier. KCV divides the original data into K groups, extracts a subset without repetition as a validation set, and combines the remaining K-1 sets of subset data as a training set, as shown in Fig. 5. In this paper, the 10-fold cross-validation method is selected.

The samples are grouped by the K-fold cross-validation method, and then cross-trained by SVM to construct the SVM-based identification model of *Cornus officinalis*. Its specific process is shown in Fig. 6: (i) Organize and collect samples and normalize data. (ii) The sample data is grouped for training by the K-fold cross-validation method, and K = 10 is selected. (iii) Take each subset (50 samples) data as a validation set, and the remaining 9 sets of subset (450 samples) data as a training set, so that 10 training model data will be obtained and brought into the SVM model for training. (iv) When the average accuracy rate of the model is greater than or equal to 80%, it is determined that the model can identify the origin of *Cornus officinalis*, and the optimal result of the training model is determined as the classification model. If the average accuracy rate is less than 80%, the samples will be re-sorted randomly, and return to step (ii) to perform K-fold cross-validation. (v) A K-fold cross-validation-based SVM identification model for the origin of *Cornus officinalis* is obtained. (vi) Input the test sample and get the classification result.

The principal components extracted from the spectral data of *Cornus officinalis* are used as input variables, OP is used as output variables, and 500 divided samples are trained with the help of SVM based on k-fold cross-validation. For SVM based on kernel function, this study compares linear kernel, quadratic kernel, cubic kernel and Gaussian kernel. The box constraint selects the default value of 1. The smaller the box constraint, the larger the margin, which means that the more error samples allowed in training, the more support vectors, and the stronger the generalization ability. Kernel scale mode select auto, set to auto to use a heuristic procedure to select the scale value using subsampling. For the selection of multiclass method, since there is no sample imbalance in this study, we choose the one-vs-all with higher efficiency. The accuracy on the validation set is used to compare different kernel functions. The accuracy rate of the support vector machine with the quadratic kernel function after the experiment is 82.4% on the validation set, which is the highest among the given kernel functions and meets the accuracy requirements set in advance. With the help of Matlab, the PCA-SVM combined model for identifying the origin of *Cornus officinalis* based on mid-infrared spectroscopy is established.

In order to verify whether the model has excellent generalization ability, 158 groups of *Cornus officinalis* test sample data of known origins are input for prediction, and the predicted results are compared with the known results. Figure 7 shows the confusion matrix for the test samples. Each row of the confusion matrix represents the predicted category, and the total number of each row represents the number of data predicted for that category; each column represents the true category, and the total number of data in each column represents the number of data instances of that category. The Precision of each category is shown on the left side of the confusion matrix, the Recall (also called Sensitivity) of each category is shown on the lower side of the confusion matrix, and the Accuracy of the model is shown in the lower right corner of the confusion matrix. Eqs. (10)-(12) is the equations of Precision, Recall and Accuracy, where TP indicates the number of samples whose true value is positive and the model judges as positive; FN indicates the number of samples whose true value is positive and the model judges as negative; FP indicates the number of samples whose true value is negative and the model judges as positive; TN indicates the number of samples whose true value is negative and the model judges as negative [40, 43].

$$Precision_i = \frac{TP_i}{TP_i + FP_i} \quad (10)$$

$$Recall_i = \frac{TP_i}{TP_i + FN_i} (11)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} (12)$$

It can be seen from the confusion matrix that there are 17 *Cornus officinalis* samples predict by the model as OP10, and they are all from OP10. This category has the highest Precision at 100%. There are 18 *Cornus officinalis* samples predict by the model as OP 8, and 12 of them are from OP 8. This category has the lowest Precision at 66.7%. 4 samples of OP 7 are incorrectly predict to be OP 8 by the model. In terms of Sensitivity, all 7 samples of OP 9 are recognized by the model, and the Recall is the highest at 100%. Only 9 of the 16 samples of OP 7 are identified, and the Recall is the lowest at 56.3%, of which 4 samples are misjudged as OP 8. In general, *Cornus officinalis* from OP 7 and OP 8 are easily confused. This is because the climate and geographical conditions of the two places are similar, resulting in similar chemical composition of *Cornus officinalis*. This can also be confirmed by their spectral vector correlation coefficients as high as 0.998. Among the total 158 samples of *Cornus officinalis*, 134 origins are correctly predicted, and the Accuracy is 84.8%, which is similar to the accuracy rate of the validation set, indicating that the model has a strong ability to resist overfitting.

4. Discussion

The spectral characteristics of different Chinese Herbal Medicines are quite different. Even the same Chinese Herbal Medicines from different origins will show different spectral characteristics under the irradiation of near-infrared and mid-infrared spectra due to the differences in the chemical composition of inorganic elements and organic matter. Therefore, these characteristics can be used to identify the species and origin of Chinese Herbal Medicines. Based on mid-infrared spectral data, naive Bayes, decision trees, LDA, RBF and PLS-DA can all identify the origin of *Cornus officinalis*. Table 4 shows the Precision and Recall of each model for the identification of *Cornus officinalis* from different origins [41, 44, 45]. It should be noted that the division of the training set and the test set of the samples of each model, and the method of model verification are the same as the PCA-SVM combined model. It can be seen that each model has a certain ability to identify the origin of *Cornus officinalis*. Decision tree has the highest precision in OP 4; LDA has the highest recall in OP 5; RBF has the highest precision in OP 8 and OP 11, and the highest recall in OP 11; PLS-DA has the highest precision in OP 5 and OP 9, and the highest recall in OP 8 and OP 10; the PCA-SVM combined model presented in this paper has the highest precision in other origins except OP 4, OP 5, OP 8, OP 9 and OP 11, and the highest recall in other origins except OP 5 and OP 11.

Table 4

Precision and Recall (Sensitivity) of each Chinese herbal medicine origin identification model in 11 different origins, where PPV stands for Precision and TPR stands for Sensitivity (values are measured in %).

OP	Naive Bayes		Decision Trees		LDA		RBF		PLS-DA		PCA-SVM	
	PPV	TPR	PPV	TPR	PPV	TPR	PPV	TPR	PPV	TPR	PPV	TPR
OP 1	77.8	77.8	83.3	55.6	83.3	83.3	81.3	72.2	76.5	72.2	89.5	94.4
OP 2	35.7	50.0	23.5	40.0	31.8	70.0	30.0	60.0	55.6	50.0	80.0	80.0
OP 3	80.0	85.7	61.5	57.1	85.7	85.7	73.3	78.6	90.9	71.4	92.9	92.9
OP 4	77.8	87.5	100	75.0	91.3	87.5	86.4	79.2	77.8	87.5	84.6	91.7
OP 5	85.7	66.7	50.0	66.7	80.0	88.9	77.8	77.8	100	66.7	85.7	66.7
OP 6	41.2	46.7	50.0	73.3	80.0	53.3	52.6	66.7	58.8	66.7	92.9	86.7
OP 7	60.0	37.5	35.7	31.3	30.0	18.8	0.0	0.0	75.0	56.2	81.8	56.3
OP 8	76.9	62.5	69.2	56.3	76.9	62.5	92.3	66.7	60.0	75.0	66.7	75.0
OP 9	55.6	71.4	66.7	57.1	87.5	100	44.4	72.7	100	71.4	77.8	100
OP10	92.3	66.7	61.9	72.2	82.4	77.8	52.9	56.2	73.9	94.4	100	94.4
OP11	46.7	63.6	40.0	36.4	50.0	63.6	77.8	100	72.7	72.7	76.9	90.9

In order to comprehensively compare the ability of each model to identify the origin of *Cornus officinalis*, we compared each model from the three indicators of Accuracy, F1-Score and Kappa coefficient. The F1-Score indicator combines the results of Precision and Recall output. Its value ranges from 0 to 1, where 1 represents the best output of the model, and 0 represents the worst output of the model. F1-Score needs to average each category of Precision and Recall (Eqs. (13)-(14)), and then use Eq. (15) to calculate.

$$Precision = \frac{\sum_{i=1}^n Precision_i}{n} \quad (13)$$

$$Recall = \frac{\sum_{i=1}^n Recall_i}{n} \quad (14)$$

$$F1 = 2 \bullet \frac{precision \bullet recall}{precision + recall} \quad (15)$$

$$Kappa = \frac{p_0 - p_e}{1 - p_e} \quad (16)$$

The Kappa coefficient is an indicator used for consistency checks and can also be used to measure the effect of classification. Its calculation is based on the confusion matrix, which takes values between - 1 and 1, usually greater than 0. Eq. (16) is the calculation equation of the Kappa coefficient, where $p_0 = \sum_i p_{ii}$ is called the observation concordance rate, $p_{ii} = \frac{a_{ii}}{N}$, a_{ii} represents the actual observation concordance number, and N represents the total number of samples. $p_e = \sum_i p_{i\bullet} \bullet p_{\bullet i}$ is called the expected concordance rate, that is, the concordance rate of the two test results due to chance, where $p_{i\bullet} = \frac{R_i}{N}$, $p_{\bullet i} = \frac{C_i}{N}$, R_i , C_i are the grand totals for rows and grand totals for columns of the i-th grid point respectively.

Table 5 comprehensively compares the three indicators of each origin identification model. It can be seen that the performance of decision trees, Naive Bayes model and RBF in the three indicators need to be improved compared to other models. LDA and PLS-DA perform well on the three indicators. Regardless of which evaluation indicator is used, the

PCA-SVM combined model proposed in this paper performs the best among all models. Using this model to identify the origin of *Cornus officinalis*, the Accuracy is 84.8%.

Table 5

Comparison results of models for origin identification of *Cornus officinalis* based on mid-infrared spectroscopy.

Evaluation indicators	Naive Bayes	Decision Trees	LDA	RBF	PLS-DA	PCA-SVM
Accuracy (%)	66.5	58.2	70.9	64.6	73.4	84.8
F1-Score	0.657	0.574	0.715	0.635	73.8	0.844
Kappa	0.628	0.538	0.678	0.609	70.3	0.831

5. Conclusions

The origin of Chinese Herbal Medicines is an important part of the quality control of Chinese Herbal Medicines, and it is also of great significance in the exploration and utilization of medicine sources (5, 13). As a non-destructive analysis technique, mid-infrared spectroscopy has the advantages of short analysis time, simple operation, and low analysis cost. In recent years, it has received increasing attention in the identification of Chinese Herbal Medicines. In this study, a method for rapid origin identification of *Cornus officinalis* based on mid-infrared spectroscopy and chemometrics was established using the spectral data of *Cornus officinalis*. The research results showed that although the mid-infrared spectral information images of the same *Cornus officinalis* have strong similarities, they also have certain differences in some parts. The spectral information is fully extracted by principal component analysis, and the classification and identification model established by the support vector machine has a high accuracy. The predictive ability of the model was evaluated by an external test set, and the results showed that the established model could classify and identify 158 *Cornus officinalis* samples from 11 different regions with an accuracy rate of 84.8%. It shows that the model has good predictive ability. In addition, by comparing with other commonly used stoichiometric models such as naive Bayes model, decision trees, LDA, RBF and PLS-DA, the PCA-SVM combined model is the best among the three indicators given in this paper. The method proposed in this paper can effectively shorten the identification time and cost of medicinal materials, and ensure the reliability of identification results. However, the scope of application of any model is limited by the sample space. Although the model established in this experiment shows good accuracy and robustness in both interactive and external tests, there is still much work to be done to promote it as a practical technique. Future studies can collect different classes of Chinese Herbal Medicines for research to improve the generalizability of the model. In addition, mid-infrared spectroscopy provides less information about the content of specific active constituents of the plant, and if more information on the content of specific active constituents of the plant is required, a more sophisticated analysis is required.

Declarations

Acknowledgements

This work was supported by the horizontal scientific research project of Nanjing Vocational University of Industry Technology (No. HK22-38-01).

Author information

Affiliations

1. Public Foundational Courses Department, Nanjing Vocational University of Industry Technology, Nanjing 210023, China

Yueqiang Jin & Bing Liu

2. Research and development department, Nanjing Changxingyang Intelligent Home Company Limited, Nanjing 210023, China

Chaoning Li

3. School of Science, Jiangsu Ocean University, Lianyungang 222005, China

Shasha Shi

*email: Jinyq@niit.edu.cn

Author contributions

Y.J. and B.L. wrote the main manuscript text, C.L. provided funding assistance, and S.S. was responsible for data processing and model validation.

Additional information

Correspondence and requests for materials should be addressed to Yueqiang Jin.

Data availability

The data that support the findings of this study are available from the corresponding author Y.J. upon reasonable request.

References

1. Jiao, L. et al. Cancer chemoprevention and therapy using chinese herbal medicine. *Biol. Proced. Online*. **20**, 1–14 (2018).
2. Lin, C.H. & Hsieh, C.L. Chinese herbal medicine for treating epilepsy. *Front. Neurosci-Switz*. **15**, 1–13 (2021).
3. Ogawa-Ochiai, K. & Kawasaki, K. Panax ginseng for frailty-related disorders: a review. *Front. Nutr*. **5**, 1–8 (2019).
4. Yu, Z.J. et al. Calculus bovis: a review of the traditional usages, origin, chemistry, pharmacological activities and toxicology. *J. Ethnopharmacol*. **254**, 1–60 (2020).
5. Yang, M. et al. Chemical variation of Chenpi (Citrus peels) and corresponding correlated bioactive compounds by LC-MS metabolomics and multibioassay analysis. *Front. Nutr*. **9**, 1–17 (2022).
6. Yao, R., Heinrich, M., Zhao, X., Wang, Q., Wei, J. & Xiao, P. What's the choice for goji: *Lycium barbarum* L. or *L. chinense* Mill.? *J. Ethnopharmacol*. **276**, 1–8 (2021).
7. Baek, S.H., Lim, H.B. & Chun H.S. Detection of melamine in foods using terahertz time-domain spectroscopy. *J. Agric. Food Chem*. **62**, 5403–5407 (2014).
8. Jiang, Y., David, B., Tu, P. & Barbin, Y. Recent analytical approaches in quality control of traditional chinese medicines-A review. *Anal. Chim. Acta*. **657**, 9–18 (2010).
9. Kim, M.K., Kim, J.H., Wang, H., Lee, H.N. & Yang, D.C. Discrimination of Korean ginseng (*Panax ginseng* Meyer) cultivar Chunpoong and American ginseng (*Panax quinquefolius*) using the auxin repressed protein gene. *J. Ginseng. Res*. **40**, 395–399 (2016).

10. Tistaert, C., Dejaegher, B. & Heyden, Y.V. Chromatographic separation techniques and data handling methods for herbal fingerprints: A review. *Anal. Chim. Acta.* **690**, 148–161 (2011).
11. Liu, K. et al. Correlation between macroscopic characteristics and tissue-specific chemical profiling of the root of *Salvia miltiorrhiza*. *Phytomedicine.* **51**, 104–111 (2018).
12. Liu, Y.Y., Wei, J.H., Gao, Z.H., Zhan, Z. & Lyu, J.C. A review of quality assessment and grading for agarwood. *Chin. Herb. Med.* **9**, 22–30 (2017).
13. Zhu, Q.X., Cao, Y.B., Cao, Y.Y. & Lu, F. Rapid detection of four antihypertensive chemicals adulterated in traditional Chinese medicine for hypertension using TLC-SERS. *Spectrosc. Spect. Anal.* **34**, 990–993 (2014).
14. Pozzi, F., Shibayama, N., Leona, M. & Lombardi, J.R. TLC-SERS study of Syrian rue (*Peganum harmala*) and its main alkaloid constituents. *J. Raman. Spectrosc.* **44**, 102–107 (2013).
15. Cui, Z., Ge, N., Zhang, A., Liu, Y., Zhang, J. & Cao, Y. Comprehensive determination of polycyclic aromatic hydrocarbons in Chinese herbal medicines by solid phase extraction and gas chromatography coupled to tandem mass spectrometry. *Anal. Bioanal. Chem.* **407**, 1989–1997 (2015).
16. Cai, H., Cao, G. & Zhang, H.Y. Qualitative analysis of a sulfur-fumigated Chinese herbal medicine by comprehensive two-dimensional gas chromatography and high-resolution time of flight mass spectrometry using colorized fuzzy difference data processing. *Chin. J. Integr. Med.* **4**, 261–269 (2017).
17. Yang, F.Q., Wang, Y.T. & Li, S.P. Simultaneous determination of 11 characteristic components in three species of curcuma rhizomes using pressurized liquid extraction and high-performance liquid chromatography. *J. Chromatogr. A.* **1134**, 226–231 (2006).
18. Sun, F. et al. Quality assessment of different species and differently prepared slices of zedaira rhizome by high-performance liquid chromatography and colorimeter with the aid of chemometrics. *J. Anal. Methods Chem.* **2020**, 1–10 (2020).
19. Obisesan, K.A., Jiménez-Carvelo, A.M., Cuadros-Rodríguez, L., Ruisánchez, I. & Callao, M.P. HPLC-UV and HPLC-CAD chromatographic data fusion for the authentication of the geographical origin of palm oil. *Talanta.* **170**, 413–418 (2017).
20. Schmidt, B., Jaroszewski, J.W., Bro, R. & Witt, M. Combining PARAFAC analysis of HPLC-PDA profiles and structural characterization using HPLC-PDA-SPE-NMR-MS experiments: commercial preparations of St. John's Wort. *Anal. Chem.* **80**, 1978–1987 (2008).
21. Zhao, H. & Chen, Z. Screening of neuraminidase inhibitors from traditional Chinese medicines by integrating capillary electrophoresis with immobilized enzyme microreactor. *J. Chromatogr. A.* **1340**, 139–145 (2014).
22. Zha, X.Q., Luo, J.P. & Wei P. Identification and classification of *Dendrobium candidum* species by fingerprint technology with capillary electrophoresis. *S. Afr. J. Bot.* **75**, 276–282 (2009).
23. Sun, X.H., Gao, C.L., Cao, W.D., Yang, X.R. & Wang, E.K. Capillary electrophoresis with amperometric detection of curcumin in Chinese herbal medicine pretreated by solidphase extraction. *J. Chromatogr. A.* **962**, 117–125 (2002).
24. Park, S.E., Seo, S.H., Lee, K.I., Na, C.S. & Son, H.S. Metabolite profiling of fermented ginseng extracts by gas chromatography mass spectrometry. *J. Gins. Res.* **42**, 57–67 (2018).
25. Sandasi, M., Vermaak, I., Chen, W. & Viljoen, A. The application of vibrational spectroscopy techniques in the quality control of material traded as ginseng. *Planta. Med.* **21**, 472–89 (2016).
26. Yu, C. et al. Adulteration and cultivation region identification of American ginseng using HPLC coupled with multivariate analysis. *J. Pharm. Biomed. Anal.* **99**, 8–15 (2014).
27. Nan, T. et al. Development of a secondary antibody thio-functionalized microcantilever immunosensor and an ELISA for measuring ginsenoside Re content in the herb ginseng, *Anal. Chem.* **84**, 4327–4333 (2012).

28. Esteban-Díez, I., González-Sáiz, J.M., Sáenz-González, C. & Pizarro, C. Coffee varietal differentiation based on near infrared spectroscopy. *Talanta*. **71**, 221–229 (2007).
29. Krähmer, A. et al. Fast and neat–determination of biochemical quality parameters in cocoa using near infrared spectroscopy. *Food Chem.* **181**, 152–9 (2015).
30. Li, W. & Qu, H. Wavelet-based classification and influence matrix analysis method for the fast discrimination of Chinese herbal medicines according to the geographical origins with near infrared spectroscopy. *J. Innov. Opt. Heal. Sci.* **7**, 1–14 (2014).
31. Tolessa, K., Rademaker, M., De, B.B., & Boeckx, P. Prediction of specialty coffee cup quality based on near infrared spectra of green coffee beans. *Talanta*. **150**, 367–374 (2016).
32. Ren, X. et al. Uv spectroscopy and hplc combined with chemometrics for rapid discrimination and quantification of curcuma rhizoma from three botanical origins. *J. Pharmaceut. Biomed.* **202**, 1–12 (2021).
33. Lu, L., Ztz, B., Yzw, B. & Frx A. A fast multi-source information fusion strategy based on ftir spectroscopy for geographical authentication of wild gentiana rigescens. *Microchem. J.* **159**, 1–10 (2020).
34. Qi, L.M., Ma, Y.T., Zhong, F.R. & Shen, C. Comprehensive quality assessment for Rhizoma Coptidis based on quantitative and qualitative metabolic profiles using high performance liquid chromatography, Fourier transform near-infrared and Fourier transform mid-infrared combined with multivariate statistical analysis. *J. Pharm. Biomed. Anal.* **161**, 436–443 (2018).
35. Ma, Y.H., He, H.Q., Wu, J.Z., Wang, C.Y., Chao, K.L. & Huang, Q. Assessment of polysaccharides from mycelia of genus ganoderma by mid-infrared and near-infrared spectroscopy. *Sci. Rep-UK.* **8**, 1–10 (2018).
36. Hou, D.Y., Shi, L.C., Yang, M.M., Li J. & Xu, H.W. De novo transcriptomic analysis of leaf and fruit tissue of cornus officinalis using illumina platform. *Plos One.* **2**, 1–18 (2018).
37. Wu, Z.Y., Zeng, Z.D., Xiao, Z.D., Mok, D.K.W. & Chan, H.Y. Correlation and uncorrelation between hyphenated chromatographic data with chemometric tools. *Chemometr. Intell. Lab.* **171**, 266–276 (2017).
38. Takane, Y. & Hunter, M.A. Constrained principal component analysis: a comprehensive theory. *Appl. Algebr. Eng. Comm.* **12**, 391–419 (2001).
39. Ding, S.F., Zhang, N., Zhang, X.K. & Wu, F.L. Twin support vector machine: theory, algorithm and applications. *Neural. Comput. Appl.* **11**, 3119–3130 (2017).
40. Yang, S. et al. Determination of the geographical origin of coffee beans using terahertz spectroscopy combined with machine learning methods. *Front. Nutr.* **8**, 1–10 (2021).
41. Wang, J.M., Liao, X.Y., Zheng, P.C., Xue, S.W. & Peng, R. Classification of Chinese herbal medicine by laser-induced breakdown spectroscopy with principal component analysis and artificial neural network. *Anal. Lett.* **4**, 575–586 (2017).
42. Liu, J., Li, Z., Hu, F., Chen, T., Du, Y. & Xin, H. Identification of GMOs by terahertz spectroscopy and ALAP–SVM. *Opt. Quantum. Electron.* **47**, 685–695 (2015).
43. Liu, W., Liu, C., Hu, X., Yang, J. & Zheng, L. Application of terahertz spectroscopy imaging for discrimination of transgenic rice seeds with chemometrics. *Food. Chem.* **210**, 415–421 (2016).
44. Zheng, Z.P., Qiu, B., Luo, A.L. & Li, Y.B. Classification for unrecognized spectra in lamost dr6 using generalization of convolutional neural networks. *Publ. Astron. Soc. Pacific.* **132**, 1–13 (2020).
45. Borsato, D., Pina, M., Spacino, K., Scholz, M. & Filho A. Application of artificial neural networks in the geographical identification of coffee samples. *Eur. Food Res. Technol.* **233**, 533–543 (2011).

Figures

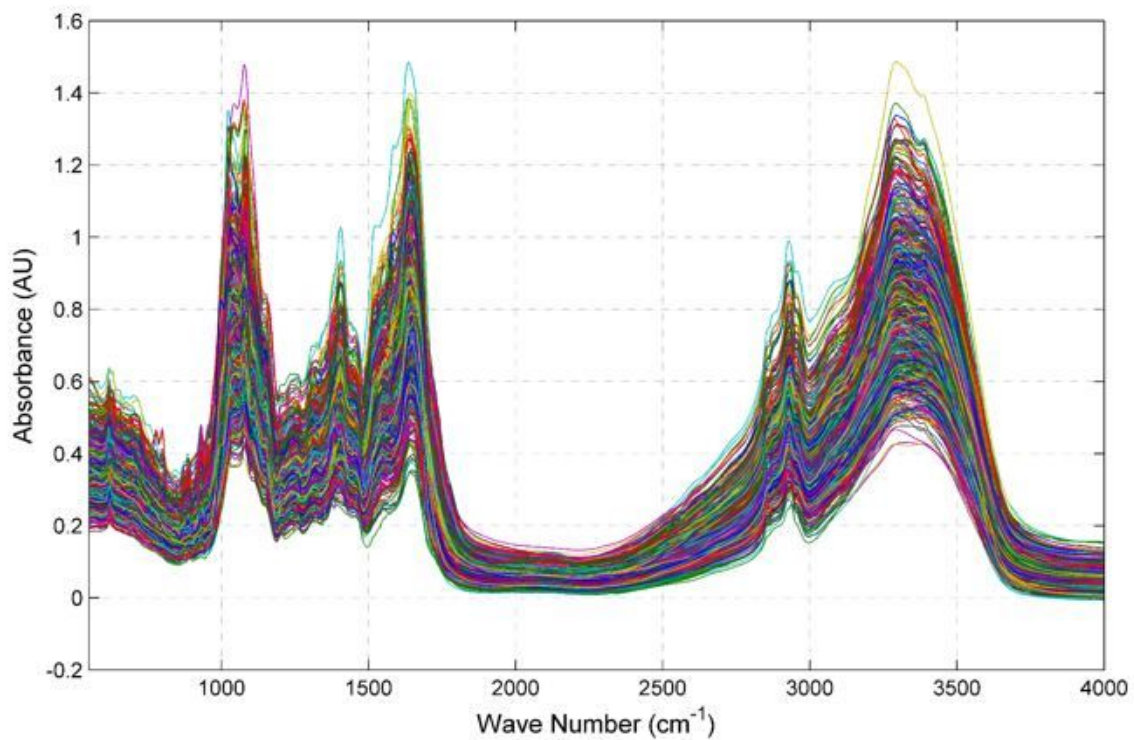


Figure 1

Mid-infrared spectral of 658 *Cornus officinalis* samples from 11 different places of origin. Figures are generated using Matlab (Version R2016a, <https://www.mathworks.com/>) [Software].

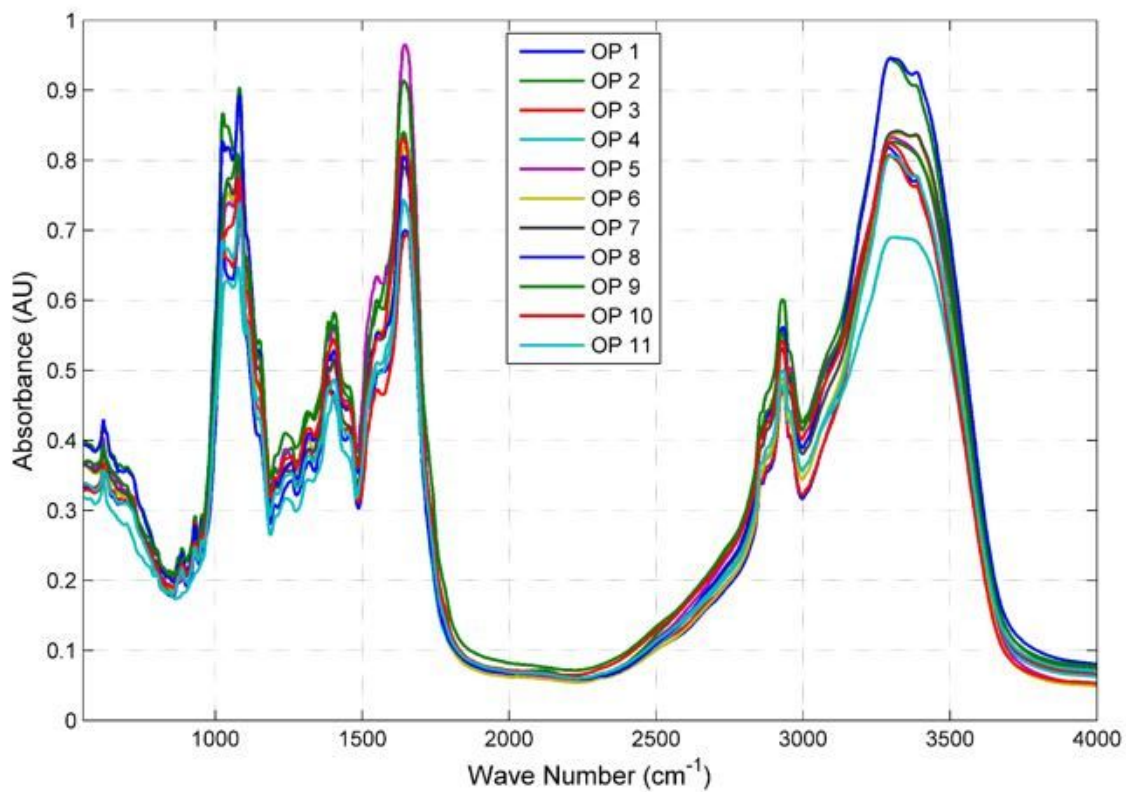


Figure 2

The average mid-infrared spectra of *Cornus officinalis* samples by different places of origin.

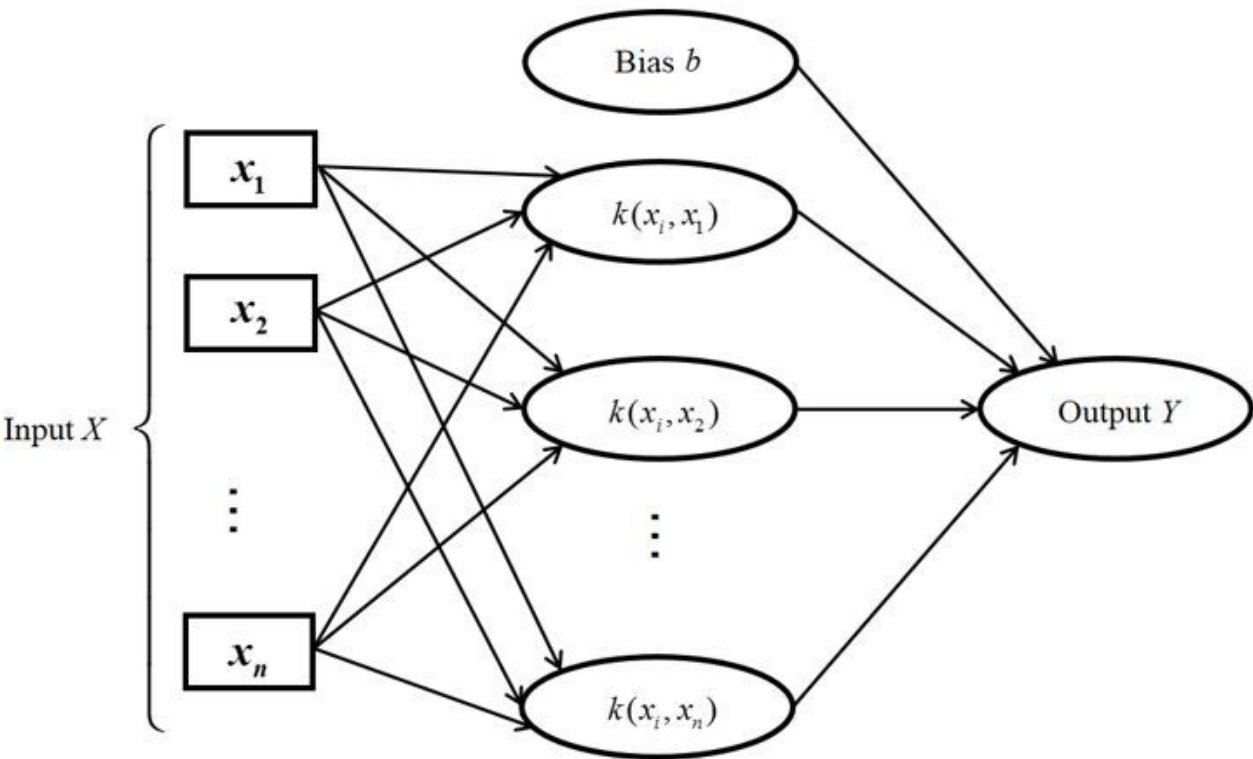


Figure 3

Support vector machine architecture.

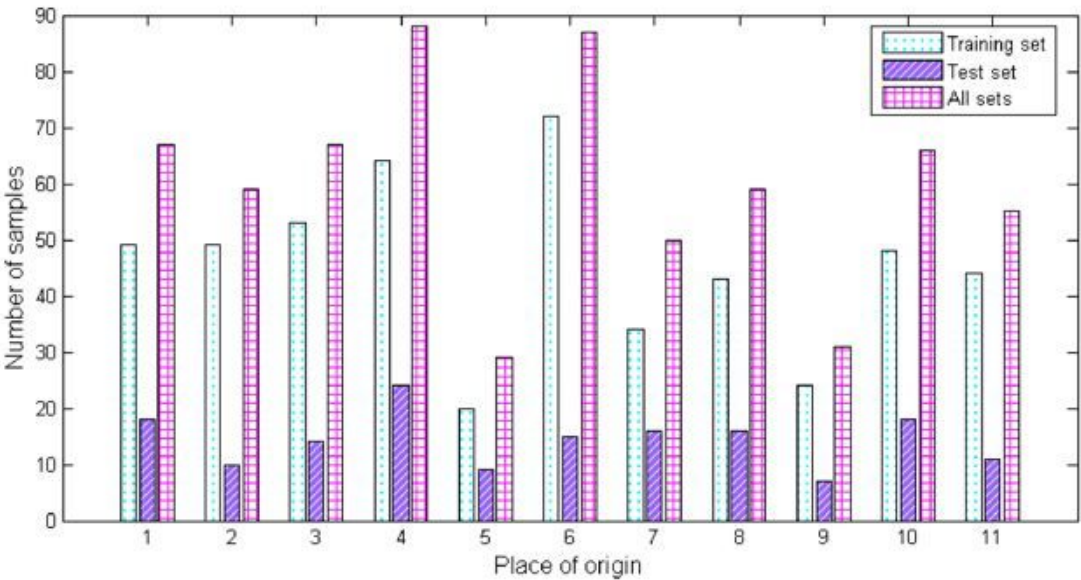


Figure 4

The number of samples in the training set, test set and all sets of *Cornus officinalis* from different origins.

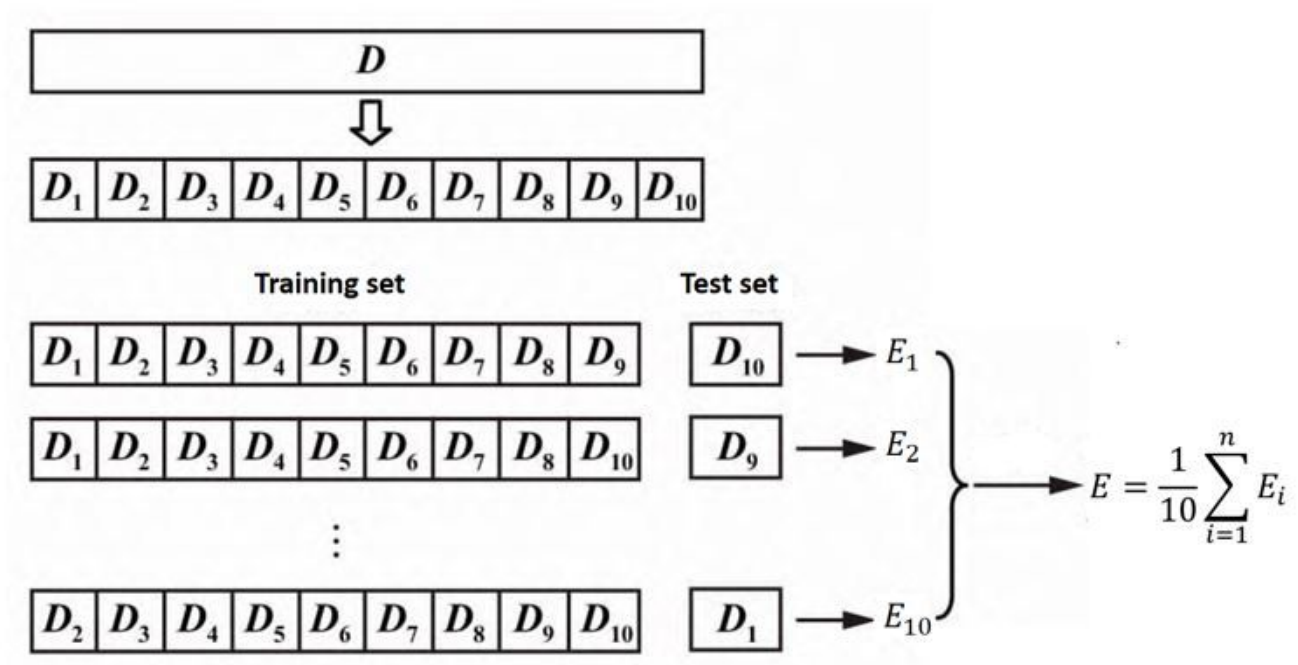


Figure 5

10-fold cross-validation process description and implementation.

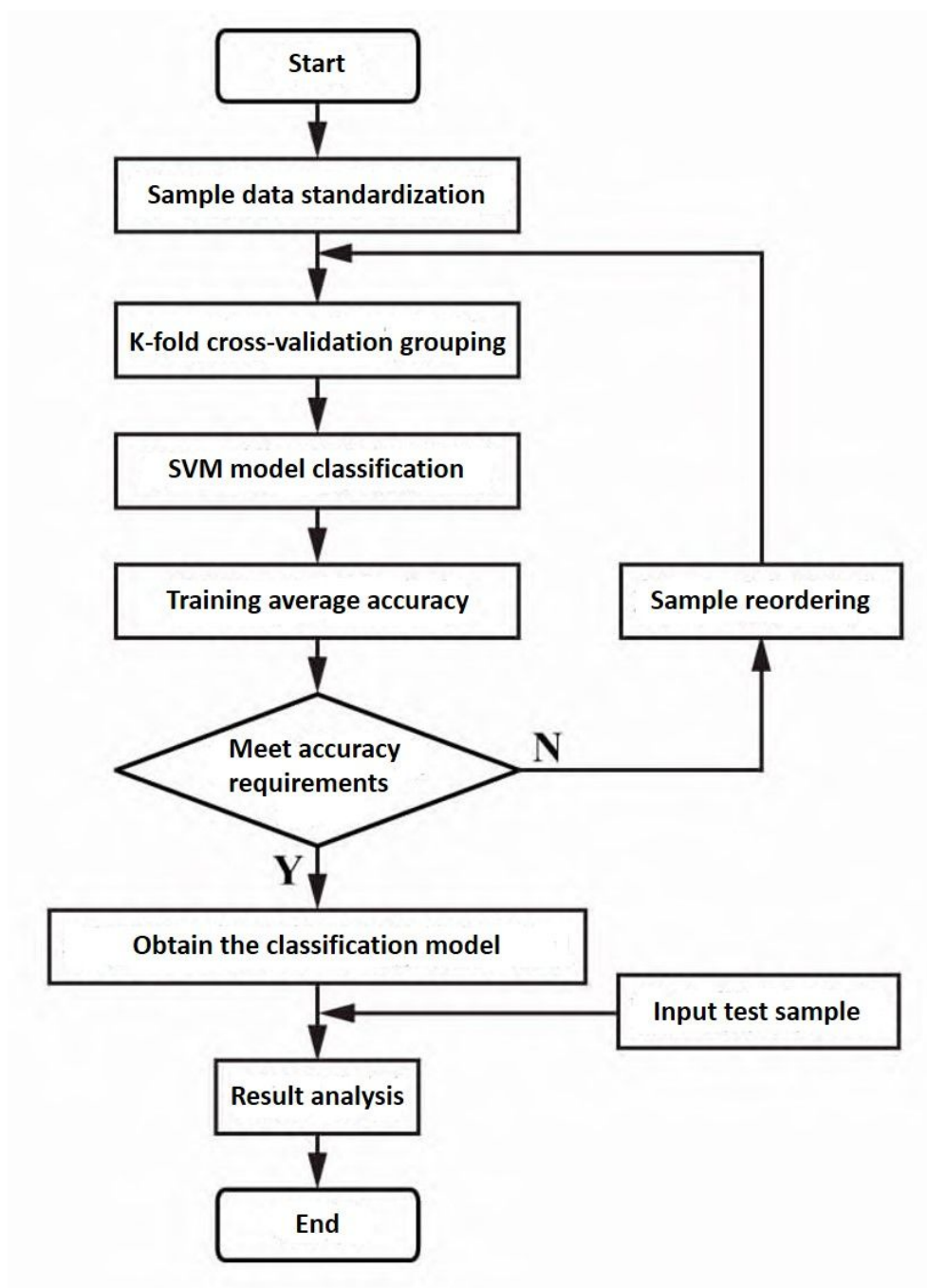


Figure 6

Flow chart of realization of SVM origin identification model of *Cornus officinalis* based on K-fold cross-validation.

Confusion Matrix													
Output Class	1	17 10.8%	0 0.0%	1 0.6%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 0.6%	0 0.0%	0 0.0%	0 0.0%	89.5% 10.5%
	2	0 0.0%	8 5.1%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 0.6%	0 0.0%	0 0.0%	1 0.6%	80.0% 20.0%
	3	0 0.0%	0 0.0%	13 8.2%	0 0.0%	0 0.0%	0 0.0%	1 0.6%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	92.9% 7.1%
	4	1 0.6%	0 0.0%	0 0.0%	22 13.9%	2 1.3%	0 0.0%	1 0.6%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	84.6% 15.4%
	5	0 0.0%	0 0.0%	0 0.0%	0 0.0%	6 3.8%	1 0.6%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	85.7% 14.3%
	6	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	13 8.2%	0 0.0%	0 0.0%	0 0.0%	1 0.6%	0 0.0%	92.9% 7.1%
	7	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	9 5.7%	2 1.3%	0 0.0%	0 0.0%	0 0.0%	81.8% 18.2%
	8	0 0.0%	2 1.3%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	4 2.5%	12 7.6%	0 0.0%	0 0.0%	0 0.0%	66.7% 33.3%
	9	0 0.0%	0 0.0%	0 0.0%	1 0.6%	0 0.0%	0 0.0%	1 0.6%	0 0.0%	7 4.4%	0 0.0%	0 0.0%	77.8% 22.2%
	10	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	17 10.8%	0 0.0%	100% 0.0%
	11	0 0.0%	0 0.0%	0 0.0%	1 0.6%	1 0.6%	1 0.6%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	10 6.3%	76.9% 23.1%
	94.4%80.0%92.9%91.7%66.7%86.7%56.3%75.0%100%94.4%90.9%84.8% 5.6%20.0%7.1%8.3%33.3%13.3%43.8%25.0%0.0%5.6%9.1%15.2%												
	Target Class												
	1	2	3	4	5	6	7	8	9	10	11		

Figure 7

Confusion matrix for *Cornus officinalis* test samples. Each row of the confusion matrix represents the predicted category, and each column represents the true category.