Hindawi
Journal of Spectroscopy
Volume 2019, Article ID 9636823, 11 pages
https://doi.org/10.1155/2019/9636823

Research Article

Rapid Identification and Quality Evaluation of Medicinal Centipedes in China Using Near-Infrared Spectroscopy Integrated with Support Vector Machine Algorithm

Sihe Kang,[1,2] **Haiying Deng**,[3] **Long Chen**,[1,4] **Xiaoxuan Zeng**,[1] **Yimei Liu**,[1] and **Keli Chen**[1]

[1] *Key Laboratory of Ministry of Education on Traditional Chinese Medicine Resource and Compound Prescription, Hubei University of Chinese Medicine, Wuhan 430065, China*
[2] *NMPA Key Laboratory of Quality Control of Chinese Medicine, Hubei Institute for Drug Control, Wuhan 430075, China*
[3] *Medical College of Wuhan University of Science and Technology, Hubei Province Key Laboratory of Occupational Hazard Identification and Control, Wuhan 430065, China*
[4] *Xiangyang Central Hospital, Xiangyang 441021, China*

Correspondence should be addressed to Yimei Liu; liuyimei1971@126.com and Keli Chen; kelichen@126.com

Received 14 May 2019; Revised 23 July 2019; Accepted 2 August 2019; Published 22 September 2019

Academic Editor: Eugen Culea

To investigate the feasibility of rapid identification and quality evaluation of Chinese medicinal centipedes using NIR spectroscopy, the qualitative and quantitative analysis models were explored. A PCA-SVC model was optimized to differentiate five species of the genus *Scolopendra*. When the model was validated with the calibration and prediction sets, the prediction accuracy was 100% and 81.82%, respectively; it can meet the requirement for rapid and preliminary identification. Based on nitrogen content detected by the chemical method, and the dimensionality of spectral data reduced with PLS, the quantitative analysis models were successfully built by PLSR and SVR algorithms. After spectra were pretreated and parameters were optimized, the performance, rationality, and prediction ability of the models were validated and evaluated with RMSECV, RMSEP, RMSEE, R^2, and RPD. Compared with the features and advantages of these two models, the PLS-SVR model had better performance and stronger prediction capacity, and it was finally regarded as the optimal quantitative analysis model to predict nitrogen content. The relative deviation between the predictive value and the reference was 2.69%, and the average recovery was 99.02%, which indicated it has potential for rapid prediction and evaluation of the quality of medicinal centipedes. This research suggested that NIR spectroscopy can be used as a rapid detection method to identify species and evaluate the quality of medicinal centipedes in China.

1. Introduction

Animals of the genus *Scolopendra* are widely distributed in the world, especially in tropical and subtropical areas [1]. In China, there are 14 species which are mainly distributed in the southern region [2]. Recently, the medicinal value of centipedes had become a research hotspot; the venom was reported to be used for relieving pain and anticoagulation [3, 4]. As an important source of Chinese medicinal materials, five species of *Scolopendra* are commonly used in China [2], which were reported to possess analgesic [3], anti-inflammatory [5], and antitumor [6, 7] activities and to improve blood rheology [8]. However, *Scolopendra mutilans* is the only species recorded in Chinese Pharmacopoeia 2015 (ChP 2015), and the other four species are just used in local regions; for instance, *S. multidens* is used in Guangxi and *S. mojiangica* in Yunnan [2]. As the animals of *Scolopendra* are poisonous, the venom and toxic ingredients of some species will bring high risk to humans [9], and the activity and quality of species are different. Therefore, in order to ensure safety and clinical efficacy, a simple, rapid, and accurate method is needed to identify the species and control the quality of medicinal centipedes.

Previously, medicinal centipedes were mostly identified using morphological description, but some similar characteristics were probably shown among closely related species.

If samples were damaged or powdered, they were difficult to be identified, and confusion and misuse would be unavoidable. Presently, molecular methods are gradually applied to identify *Scolopendra* species [10]. However, the complexity of operation and high technical requirements make it difficult to obtain rapid and accurate results, especially in mixed samples. Proteins or amino acids are recognized as the main active ingredient of medicinal centipedes [11–13], and their content is usually used for quality evaluation. Because of the diversity of ingredients, and the complexity of chemical determination methods, this measurement is usually cumbersome [14].

Near-infrared (NIR) spectroscopy combined with chemometrics is a fast, nondestructive, and environmentally friendly analysis technique that can realize multicomponent analysis. Nowadays, it is widely used in agriculture and medicine [15–17]. NIR spectroscopy mainly reflects the absorption of overtone and combination peaks containing hydrogen bonds of C-H, O-H, and N-H [16]. Lipids and proteins are considered to be important medicinal components, which are rich in centipedes and show characteristic absorption in the near-infrared region. However, the application of NIR spectroscopy on medicinal centipedes has not yet been reported. In this study, the NIR spectroscopy analysis methods were investigated, a PCA-SVC model was explored to identify the species of *Scolopendra*, and in light of nitrogen content determined by the chemical method, a quantitative model was established for quality prediction using regression algorithms.

2. Materials and Methods

2.1. Instruments and Software. The nitrogen content of samples was determined with the DK 20 Heating Digester (VELP, Italy) and UDK 149 Automatic Distillation Unit (VELP, Italy). Spectra were collected with an MPA FT-NIR spectrometer (Bruker Optics Co., Ltd., Germany) and analyzed using the OPUS 7.5 spectrum analysis software (Bruker), MATLAB R2014a data analysis software (MathWorks, Inc., USA), and Unscrambler 9.7 data analysis software (CAMO Software AS, Norway).

2.2. Samples and Identification. A total of 64 samples from 28 batches have been collected from field surveys or market commodity in China since 2015. All samples were identified into five nominal species according to characteristics recorded by Siriwut et al. [1], Kang et al. [2], Song et al. [18], and Zhang and Wang [19]. The sample information is summarized in Table 1. All samples were kept below −20°C and housed in Hubei University of Chinese Medicine, Wuhan, China.

2.3. Content Determination. After being scanned with a near-infrared spectrometer, the nitrogen content of 50 mg powder of each sample was determined with the semimicro quantitative nitrogen determination method referring to the guideline of ChP 2015. The samples were digested using the DK 20 Heating Digester with a program as follows: 200°C for 5 min, then up to 260°C sustaining for 5 min, 340°C for 5 min, and 420°C for 40 min, and at last cooled down to 200°C. The sample solution was measured using the UDK 149 Automatic Distillation Unit with a program as follows: 50 ml H_2O and 20 ml 40% NaOH were added to the digested solution, 20 ml 2% H_3PO_4 was used for receiving, the steam quantity was 50%, the distillation time was 4 min, and then titration was done with 0.025 mol/L H_2SO_4 standard solution (Metrological Testing Technology Research Institute of Shanghai; Batch number 150901).

2.4. Spectra Acquisition. After samples were smashed and dried at 55°C for 24 h, the powder of 2 g of individuals was scanned using the MPA FT-NIR spectrometer with a diffuse reflection integral sphere. The spectra were obtained in a range of 12000~4000 cm^{-1} by the coaddition of 32 scans at a resolution of 8 cm^{-1}. Each sample was scanned three times, and the average of three spectra was used for analysis. The spectra diagram is shown in Figure 1.

2.5. Spectral Pretreatment Method. Usually, the raw spectrum includes a lot of irrelevant information or noise, which would lead to baseline drift and instability. Therefore, spectrum pretreatment is a critical step in spectral analysis. There are many pretreatment methods, and each has advantages to improve model performance. For instance, vector normalization (VN) can be used to eliminate influences of the optical path change on the spectrum. The derivative methods including the first derivative (FD) and second derivative (SD) are always employed to eliminate spectral difference from baseline [20], while multiple scattering correction (MSC) is commonly performed to process diffuse reflection spectra [21]. In this study, methods such as VN, FD, SD, and MSC or combined pretreatments were employed by OPUS to optimize model performance.

2.6. Spectral Data Compression Method

2.6.1. Principal Component Analysis (PCA) Method. PCA is a commonly used method for data compression. It performs dimensionality reduction of a high-dimensional dataset, while retaining its variation as much as possible. This method can transform a number of possibly correlated variables (the original data matrix) into one or a few important variables (principal components (PCs)) to reveal the internal structure. Each PC is a linear combination of the original data. The new variables are not related to each other, which can eliminate the overlapped part of information. Moreover, these new variables include the most informative dimensions of the original variables without losing too much information. Commonly, the number of PCs is determined by the contribution rate to original variables. When the cumulative contribution rate is more than 85%, the main components can represent most of the information provided by the original variable [22]. In our identification research, the PCA was used to reduce the dimension of original spectral data.

Table 1: Sample information of medicinal centipedes.

Number	Species	Batch no.	Nitrogen content (%)	Origin
1	*S. mutilans* L. Koch	WG 002-1	10.09	Suizhou, Hubei
2	*S. mutilans* L. Koch	WG 003-1	11.36	Suizhou, Hubei
3	*S. mutilans* L. Koch	WG 004-1	11.82	Jingmen, Hubei
4	*S. mutilans* L. Koch	WG 004-2	10.60	Jingmen, Hubei
5	*S. mutilans* L. Koch	WG 005-1	10.77	Xiangyang, Hubei
6	*S. mutilans* L. Koch	WG 005-2	10.49	Xiangyang, Hubei
7	*S. mutilans* L. Koch	WG 006-1	9.47	Yichang, Hubei
8	*S. mutilans* L. Koch	WG 013-1	10.43	Suizhou, Hubei
9	*S. mutilans* L. Koch	WG 014-1	9.25	Jinshan, Hubei
10	*S. mutilans* L. Koch	WG 014-2	10.22	Jinshan, Hubei
11	*S. mutilans* L. Koch	WG 016-1	10.10	Suizhou, Hubei
12	*S. mutilans* L. Koch	WG 016-2	9.83	Suizhou, Hubei
13	*S. mutilans* L. Koch	WG 017-1	8.20	Anlu, Hubei
14	*S. mutilans* L. Koch	WG 017-2	10.15	Anlu, Hubei
15	*S. mutilans* L. Koch	WG 018-1	11.02	Yichang, Hubei
16	*S. mutilans* L. Koch	WG 019-1	10.05	Nanzhang, Hubei
17	*S. mutilans* L. Koch	WG 019-2	10.16	Nanzhang, Hubei
18	*S. mutilans* L. Koch	WG 020-1	9.06	Anhui
19	*S. mutilans* L. Koch	WG 020-2	11.74	Anhui
20	*S. mutilans* L. Koch	WG 027-1	10.55	Henan
21	*S. mutilans* L. Koch	WG 027-2	11.10	Henan
22	*S. mutilans* L. Koch	WG 032 -1	9.96	Machang, Hubei
23	*S. mutilans* L. Koch	WG 032-2	10.01	Machang, Hubei
24	*S. mutilans* L. Koch	WG 045-1	9.68	Machang, Hubei
25	*S. mutilans* L. Koch	WG 045-2	9.68	Machang, Hubei
26	*S. multidens* Newport	WG 012-1	11.27	Yulin, Guangxi
27	*S. multidens* Newport	WG 012-2	10.83	Yulin, Guangxi
28	*S. multidens* Newport	WG 012-3	10.54	Yulin, Guangxi
29	*S. multidens* Newport	WG 012-4	11.74	Yulin, Guangxi
30	*S. multidens* Newport	WG 012-5	11.33	Yulin, Guangxi
31	*S. multidens* Newport	WG 012-6	9.77	Yulin, Guangxi
32	*S. multidens* Newport	WG 012-7	11.27	Yulin, Guangxi
33	*S. multidens* Newport	WG 012-8	11.94	Yulin, Guangxi
34	*S. multidens* Newport	WG 021-1	9.85	Guangxi
35	*S. multidens* Newport	Wg039-1	11.92	Guangxi
36	*S. multidens* Newport	Wg039-2	10.98	Guangxi
37	*S. multidens* Newport	Wg040-1	10.25	Mengzi, Yunnan
38	*S. multidens* Newport	Wg040-2	9.81	Mengzi, Yunnan
39	*S. dehaani* Brandt	WG 028-1	12.31	Yunnan
40	*S. dehaani* Brandt	WG 038-1	12.62	Yunnan
41	*S. dehaani* Brandt	WG 038-2	12.17	Yunnan
42	*S. dehaani* Brandt	WG 038-3	12.58	Yunnan
43	*S. dehaani* Brandt	WG 038-4	12.30	Yunnan
44	*S. dehaani* Brandt	WG 038-5	12.36	Yunnan
45	*S. dehaani* Brandt	WG 038-6	11.67	Yunnan
46	*S. dehaani* Brandt	WG 038-7	12.46	Yunnan
47	*S. dehaani* Brandt	WG 038-8	11.93	Yunnan
48	*S. mojiangica* Zhang et Chi	WG 007-1	8.37	Mojiang, Yunnan
49	*S. mojiangica* Zhang et Chi	WG 007-2	8.04	Mojiang, Yunnan
50	*S. mojiangica* Zhang et Chi	WG 007-3	8.16	Mojiang, Yunnan
51	*S. mojiangica* Zhang et Chi	WG 007-4	7.47	Mojiang, Yunnan
52	*S. mojiangica* Zhang et Chi	WG 007-5	8.18	Mojiang, Yunnan
53	*S. mojiangica* Zhang et Chi	WG 008-1	8.92	Mojiang, Yunnan
54	*S. mojiangica* Zhang et Chi	WG 008-2	9.06	Mojiang, Yunnan
55	*S. mojiangica* Zhang et Chi	WG 041-2	8.54	Bixi, Yunnan
56	*S. mojiangica* Zhang et Chi	WG 041-3	8.37	Bixi, Yunnan
57	*S. negrocapitis* Zhang et Wang	WG 022-1	10.48	Suizhou, Hubei
58	*S. negrocapitis* Zhang et Wang	WG 022-2	10.07	Suizhou, Hubei
59	*S. negrocapitis* Zhang et Wang	WG 022-3	10.52	Suizhou, Hubei
60	*S. negrocapitis* Zhang et Wang	WG 022-4	10.65	Suizhou, Hubei

Table 1: Continued.

Number	Species	Batch no.	Nitrogen content (%)	Origin
61	*S. negrocapitis* Zhang et Wang	WG 022-5	11.45	Suizhou, Hubei
62	*S. negrocapitis* Zhang et Wang	WG 015-1	10.86	Chaohu, Anhui
63	*S. negrocapitis* Zhang et Wang	WG 015-2	11.71	Chaohu, Anhui
64	*S. negrocapitis* Zhang et Wang	WG 015-3	10.82	Chaohu, Anhui

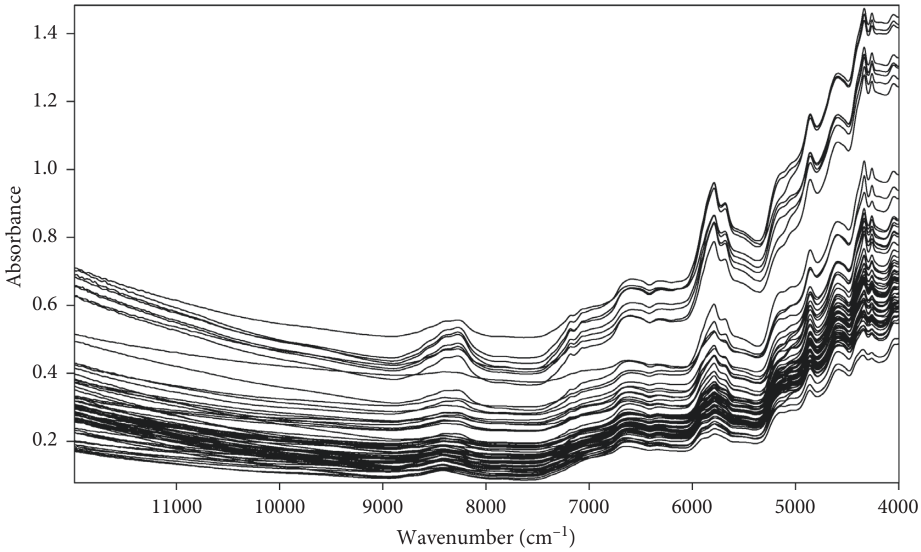


Figure 1: NIR spectra diagram of samples.

2.6.2. Partial Least-Squares (PLS) Method. The PLS is a new multivariate statistical analysis method. It attempts to recombine the original variables (mainly continuous variables) into a group of new independent comprehensive variables and extracts a few comprehensive variables to reflect the information on the original variables as much as possible. The extracted new variables have good interpretation ability for the dependent variables. During modeling, it not only considers factors of the independent variable matrix (spectral matrix) but also takes the "response" matrix (content matrix) into account. The principal component scores extracted by dimension reduction are used as input variables to avoid multicollinearity, improve stability, and simplify the model. Therefore, the PLS has the ability to simplify the model and characteristics of quick calculation and strong prediction ability, and as one of the most classical data processing tools in multiple correlation regression, it is widely applied in NIR spectroscopy quantitative analysis [23–25].

2.7. Support Vector Machine (SVM) Algorithm. SVM is a powerful supervised learning algorithm that was first proposed by Vapnik [26] and successfully extended by researchers in recent years. It is based on the principle of minimization of structural risk in constructing an optimally separating hyperplane that separates different classes of data. In the process, input vectors are mapped to a newly constructed high-dimensional space, and then parallel hyperplanes are constructed to maximize the interplane distance which separates the data. The SVM can solve nonlinear problems in a higher-dimensional space based on radical basis function (RBF) to construct a linear function. Usually, the SVM algorithm includes both support vector machine for classification (SVC) and support vector machine for regression (SVR); the former is used to solve problems of classification, while the latter is used for regression analysis.

RBF is a commonly used kernel function in the SVM algorithm. It has a strong ability to deal with nonlinear problems. It can be expressed as follows:

$$K(x, y) = \exp\left(-g|x - y|^2\right), \quad g > 0. \tag{1}$$

RBF has two important parameters in the SVM algorithm, i.e., penalty factor "C" and kernel function parameter "g," which have a great influence on model prediction ability, and the values should be determined during the model optimization process. Commonly, the optimization methods include the grid search (GS), particle swarm optimization (PSO), and genetic algorithm (GA). The PSO algorithm simulates the flight foraging behaviors of bird clusters through collaboration among birds to achieve the best objective. The GA is an operation based on biological natural selection and genetic mechanism to realize the optimal result, while the GS iterates through every intersection in the grid to find the best combination of parameters (C, g) and makes cross-validation most accurate. The RMSECV was used to guide the optimization of internal parameters.

During modeling of the SVM algorithm, the input data need to be mapped to a higher-dimensional space to realize dimension reduction and regression fitting. So, the data should be firstly pretreated and compressed.

2.8. Model Validation and Evaluation

2.8.1. Qualitative Model. In the PCA-SVC qualitative model, the model performance was evaluated by 3-fold cross-validation (3-CV) of the calibration set. The internal parameters C and g were optimized with the GS method. When the accuracy of 3-CV reached maximum values, the optimal C and g were determined. After models were established, the calibration set and prediction set were input to the model, and the prediction accuracies were used as indexes to evaluate the prediction ability.

2.8.2. Quantitative Model. During the process of modeling, the calibration set was used for internal cross-validation to validate model performance, the internal cross-validation adopted 6-fold cross-validation, and the root mean square error of internal cross-validation (RMSECV), coefficient of determination (R^2), and ratio of performance to deviation (RPD) were taken to guide the model optimization process. The predication set was used for external validation to evaluate the model, with the root mean square error of prediction (RMSEP), R^2, and RPD taken as indexes to evaluate prediction ability. Generally, the smaller the RMSECV and the larger the R^2 are, the better the model performance would be; the smaller the RMSEP and the greater the R^2 are, the stronger the model prediction ability is. Moreover, when RPD >2, it indicates that the model has excellent reliability. After the model was established, the calibration set used for full cross-validation was input to the model again, and the root mean square error of evaluation (RMSEE) was used as an index to further evaluate the reasonability of the model. Theoretically, the RMSECV value was higher than the RMSEE value, which indicated that the modeling process was reasonable and feasible.

3. Results and Discussion

3.1. Determination of Nitrogen Content. The content of 64 samples was measured. Samples used in the analysis are as follows: 25 specimens of *S. mutilans*, 13 of *S. multidens*, 9 of *S. mojiangica*, 8 of *S. negrocapitis*, and 9 of *S. dehaani*. The nitrogen content of species was between 8.19% and 12.27%, the total mean was 10.0%, and the mean value of each species was 10.25%, 11.04%, 8.19%, 10.58%, and 12.27%, respectively. The results are shown in Table 1.

3.2. Analysis of NIR Spectra. The NIR spectra of samples were scanned in the range of 12000–4000 cm^{-1}; the spectra diagram is shown in Figure 1. It indicated that the characteristic wavenumber was mainly in the range of 9000 to 4000 cm^{-1}, while the spectral characteristics showed high similarity among samples that it was difficult to distinguish species from peak data. Hence, the chemometric method was needed for spectral pretreatment and characteristic information extraction in qualitative and quantitative analysis.

3.3. Qualitative Model Based on PCA-SVC Algorithm

3.3.1. Sample Classification. The spectra of 64 samples were randomly classified into calibration and prediction sets in a proportion of approximately 2 : 1. Finally, 42 samples of the calibration set were used for model establishment, and 22 samples of the prediction set were used for model evaluation. The species were represented with category label numbers 1 to 5. The classification information is shown in Table 2.

3.3.2. Optimization of Pretreatment. In this qualitative analysis, the three methods VN, FD, and SD were used to pretreat the raw spectra. The PCA method was used to reduce dimensions of raw and three pretreated spectra. The accumulative contribution rates of PCs were calculated. The result showed that the contribution rates of the first two PCs (PC1 and PC2) were more than 85%, which can represent most of the spectrum information [22]. Hence, the PC1-PC2 correlation diagram was attempted for a preliminary investigation to differentiate the samples. However, most species overlapped together in space and cannot be discriminated, except *S. mojiangica* in the FD and SD.

To further investigate the influence of different pretreatments, a group of PCA-SVC models was established using the scores of the first 2 PCs as input variables and category labels as output variables. The model performance was evaluated by 3-fold cross-validation (3-CV) of the calibration set. The internal parameters of the SVC algorithm were optimized with the GS method. The values of best C and g were obtained based on the initial optimization in a range of $\log 2\,c \in [0, 50]$ and $\log 2\,g \in [0, 50]$ and in steps of 5 and then the second fine optimization in an adjusted narrow scope. When the 3-CV accuracy reached the maximum, the optimal values of C and g were determined. After models were established, the calibration set and prediction set were input to models and predicted, and the prediction accuracies were used to evaluate the prediction ability. As shown in Table 3, it can be seen that "overfitting" and mismatching existed in the raw spectrum model for high accuracy (90.48%) in the calibration set and low accuracy (59.09%) in the prediction set. In contrast, the other three models VN, FD, and SD had nearer accuracies between calibration and prediction sets to possess a relatively rational structure, although the values were not even high. Also, the model of the SD had the highest accuracy among all pretreatments, whether in the calibration or prediction set. Therefore, the SD was regarded as the best pretreatment for its better prediction ability.

3.3.3. Optimization of the Number of Principal Components (NPC). Although the SD was determined as the optimal pretreatment in a preliminary investigation, the accuracy in the model with scores of the first 2 PCs as input variables was just about 70%, which did not meet the requirement of discrimination. Hence, the best NPC still needs to be optimized. In light of the modeling and SD pretreatment method mentioned above, 10 PCA-SVC models (SVC-5 to SVC-14) were established using the scores of the first 1, 2, 3, . . ., 10 PCs of the calibration set as input variables. As shown in Table 4,

Table 2: Classified information of the qualitative model of medicinal centipedes.

Sample set	*S. multidens*	*S. dehaani*	*S. negrocapitis*	*S. mojiangica*	*S. mutilans*	Total
Calibration set	8	6	5	6	17	42
Prediction set	5	3	3	3	8	22
Label value	1	2	3	4	5	—

Table 3: Different spectral pretreatments of PCA-SVC models.

Model number	Pretreatment	NPC	C	g	Accuracy rate (%)		
					3-fold cross-validation	Calibration set	Prediction set
SVC-1	Raw	2	524288	0.03125	54.7619	90.4762	59.0909
SVC-2	VN	2	$6.71089 * 10^7$	0.0078125	64.2857	66.6667	63.6364
SVC-3	FD	2	16	32768	59.5238	64.2857	63.6364
SVC-4	SD	2	$3.35544 * 10^7$	32768	66.6667	71.4286	68.1818

Table 4: Comparison on PCA-SVC models established with different NPCs.

Model number	NPC	C	g	Accuracy rate (%)		
				3-fold cross-validation	Calibration set	Prediction set
SVC-5	1	64	$4.29497 * 10^9$	59.5238	80.9524	63.6364
SVC-6	2	$3.35544 * 10^7$	32768	66.6667	71.4286	68.1818
SVC-7	3	$4.1943 * 10^6$	262144	66.6667	85.7143	81.8182
SVC-8	4	2521.38	$1.27148 * 10^7$	71.4286	90.4762	77.2727
SVC-9	5	23170.5	$3.65135 * 10^6$	73.8095	97.6190	72.7273
SVC-10	6	26615.9	794672	73.8095	90.4762	77.2727
SVC-11	7	26615.9	$1.2045 * 10^6$	78.5714	97.6190	81.8182
SVC-12	8	$5.93164 * 10^6$	5792.62	83.3333	100	81.8182
SVC-13	9	131072	262144	80.9524	100	81.8182
SVC-14	10	$1.04858 * 10^6$	65536	80.9524	100	77.2727

the accuracy improved with the increase of the NPC. When the NPC was 8, the accuracy in the calibration set was 100% and in the prediction set was 81.82%. When the NPC was higher than 8, the accuracy in the calibration set was 100%, while the accuracy in the prediction set did not increase or even decreased. Therefore, number 8 was considered the best NPC, and model SVC-12 was the optimal qualitative model.

3.3.4. Validation and Evaluation of PCA-SVC Model. According to the research above, SVC-12 was determined as the best qualitative analysis model. After the full spectrum was pretreated with the SD and the dimension was reduced with PCA, the model was established using the scores of the first 8 PCs as input variables and category labels as output variables. The internal parameters of best C and g were optimized with the GS. The initial search was in a range of $\log 2\, c \in [0, 50]$ and $\log 2\, g \in [0, 50]$ and in steps of 5, and then the fine search was in a range of $\log 2\, c \in [20, 23]$ and $\log 2\, g \in [12, 16]$ and in steps of 0.5; when $C = 5.93164 * 10^6$ and $g = 5792.62$, the accuracy of 3-CV was 83.33%. When the model was predicted with the calibration set and prediction set, the accuracy was 100% (42/42) and 81.82% (18/22), respectively, which might be accepted for rapid identification. The optimizations of internal parameters are shown in Figure 2, and the predictive results are shown in Figure 3.

3.4. Quantitative Model Based on PLSR and PLS-SVR Algorithms

3.4.1. Partition of Sample Set. In this quantitative analysis, the Kennard–Stone (K-S) algorithm was used to divide 64-sample spectra into the calibration set and prediction set in a proportion of 2 : 1 in the MATLAB R2014a software; 42 samples of the calibration set were used for validation, while 22 samples of the prediction set were used for prediction.

3.4.2. PLSR Model. The partial least-squares regression (PLSR) model is one of the multiple linear regression (MLR) models; it can easily realize the ideal linear relationship between input variables (spectral information) and output variables (ingredient contents) after high dimensions are compressed by PLS. PLSR has the desirable property to analyze data that are strongly collinear (correlated), noisy, and independent variables and also simultaneously model several response variables; now, it has been developed as a standard tool in chemometrics [27].

The full spectral data (12000~4000 cm^{-1}) were used for modeling. To eliminate noise and other factors, they need to be firstly pretreated. The pretreatments including Raw, VN, FD, FD + VN, MSC, and FD + MSC were applied. After the dimensions were reduced with PLS, the treated spectral data were used as input variables and nitrogen content was the

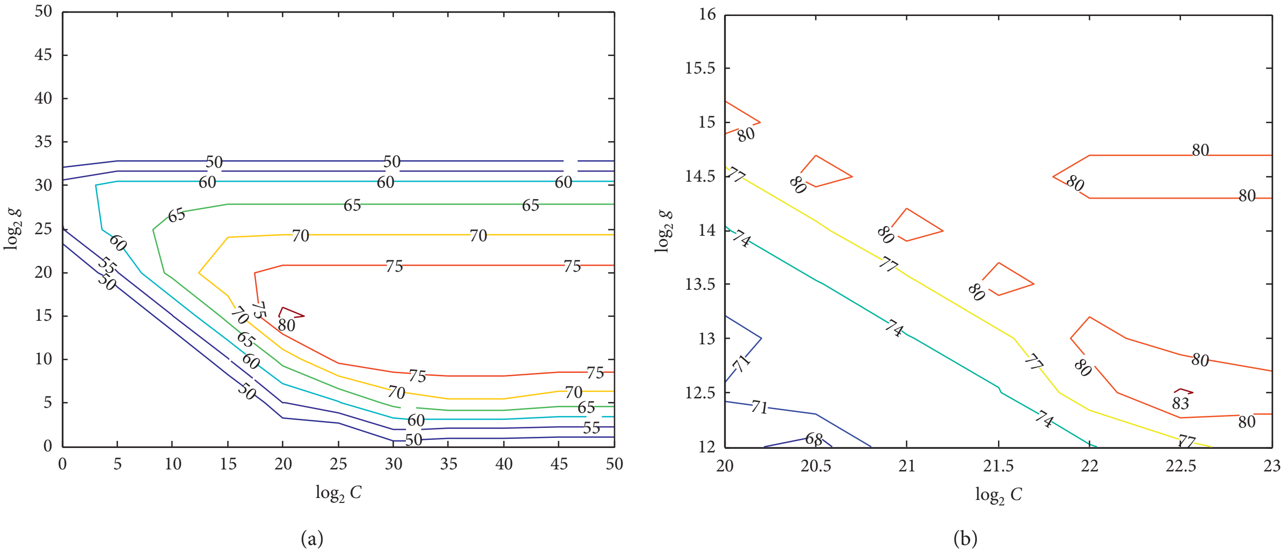


Figure 2: Optimization of internal parameters with the grid search of the PCA-SVC model. (a) Initial grid search. (b) Fine search.

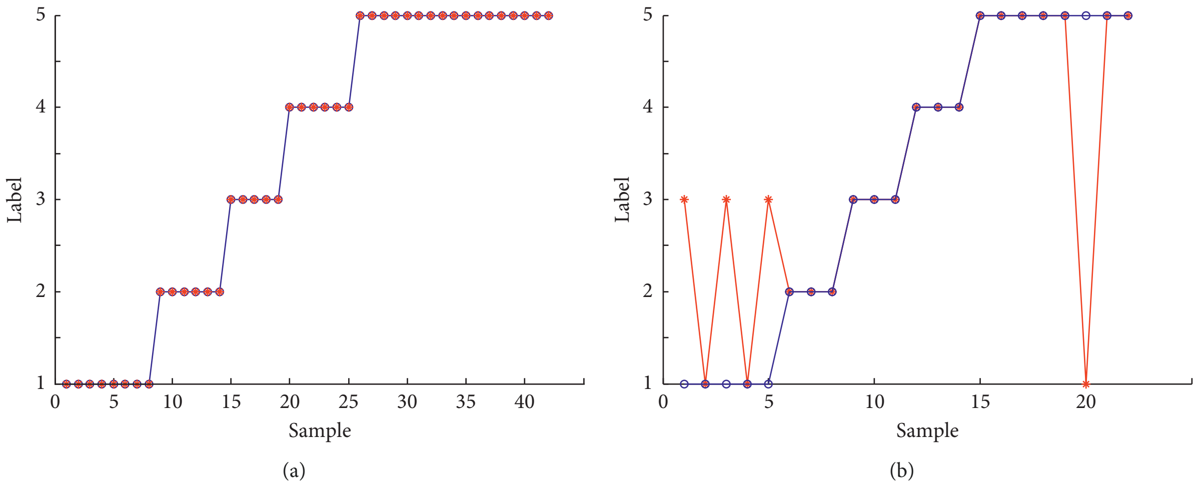


Figure 3: Validation results of the PCA-SVC model for medicinal centipedes: (a) calibration set; (b) prediction set. The red points represent validation results, and the blue points represent reference values.

output reference, and a series of PLSR models were established with the Unscrambler 9.7 software.

During the process, the model was validated and evaluated. As shown in Table 5, the RPDs were all over 2 to possess model reliability. All the values of RMSECV were higher than those of RMSEE, which indicated the feasibility of the models. Models PLSR-1, PLSR-4, and PLSR-6 had lower RMSECV and higher R^2 to present good performance, while PLSR-1 had great RMSEP, minimum R^2, in external validation, and the largest NPC, and its structure is unreasonable. There was no significant difference in performance and prediction ability between PLSR-4 and PLSR-6, but considering the rationality of pretreatment, RMSECV, and RMSEE, the PLSR-6 model was considered the best model.

The optimization of the NPC is an important step during modeling. It can be obtained from the RMSECV-NPC diagram. For instance, in the PLSR-6 model, with the change of the NPC, the RMSECV had different values; when the NPC was 5, the RMSECV had a minimum value, and the model had the best performance. Therefore, the optimal NPC was determined as 5. The optimization is shown in Figure 4. The regression equation of the PLSR-6 model between the principal component scores (SPL_i, $i = 1, 2, \ldots, 5$) and the nitrogen content is expressed as follows:

$$Y = 803.21\,\text{SPL}_1 + 661.59\,\text{SPL}_2 + 589.71\,\text{SPL}_3 + 595.33\,\text{SPL}_4 + 476.86\,\text{SPL}_5 + 10.26,\ R^2 = 90.95\%. \tag{2}$$

As described above, the best PLSR model was finally determined, the optimized pretreatment was determined as FD + MSC, and the NPC was 5. During modeling, 6-fold cross-validation was used as internal validation to validate

Table 5: Validation and predictive results of PLSR models with different pretreatment methods.

Model number	Pretreatment	6-fold cross-validation			External validation			RMSEE (%)	NPC
		RMSECV (%)	R^2 (%)	RPD	RMSEP (%)	R^2 (%)	RPD		
PLSR-1	Raw	0.42	90.50	2.85	0.51	80.78	2.30	0.27	9
PLSR-2	VN	0.47	87.22	2.51	0.43	84.3	2.51	0.34	5
PLSR-3	FD	0.46	88.14	2.53	0.46	84.41	2.39	0.31	6
PLSR-4	FD + VN	0.41	90.71	3.04	0.43	85.84	2.53	0.32	5
PLSR-5	MSC	0.44	89.63	2.69	0.50	81.72	2.39	0.26	8
PLSR-6	FD + MSC	0.40	90.95	2.96	0.44	85.61	2.51	0.31	5

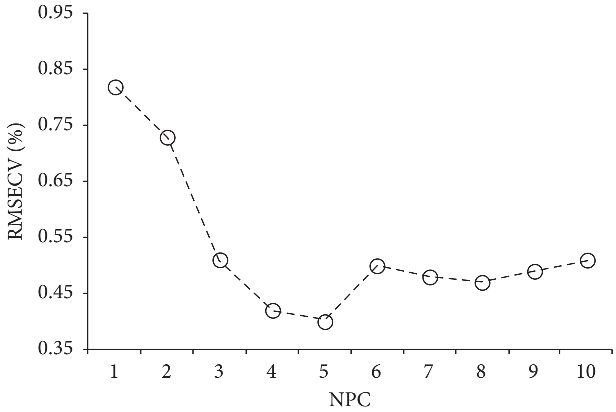


Figure 4: RMSECV-NPC diagram of the PLSR-6 model.

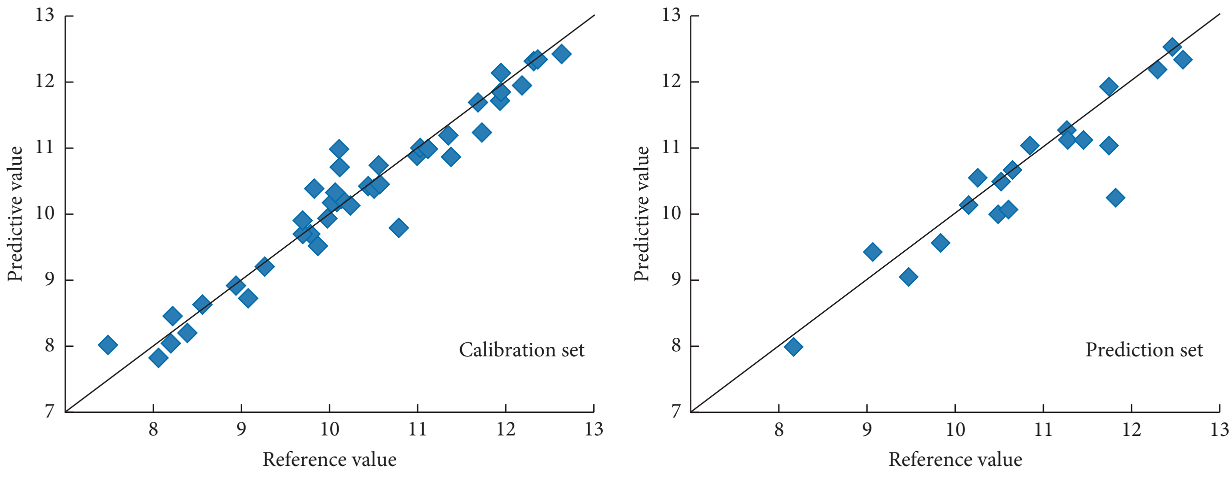
the performance, and the predictive ability was evaluated with external validation using the prediction set. The predictive results are shown in Figure 5. The average relative deviation between the predictive value and the reference was 2.71%, and the average recovery was 98.77%.

3.4.3. PLS-SVR Model. Besides ingredient information, the NIR spectroscopy also contains much other information, such as physical and chemical information, which often causes spectral bands seriously overlapped. Actually, in most cases, it shows nonlinear relationship between sample spectra and content. With the development of application of chemometrics, modern intelligent algorithms have attached more attention to NIR spectroscopy analysis for its strong nonlinear fitting ability and obtained preliminary exploration and application. The SVM algorithm is based on statistics to allow obtain a good fitting effect and stable structure. As a result, it becomes a commonly used nonlinear regression algorithm. Compared with the ANN algorithm which is suitable for solving problems of complex mapping and large sample size [28], the SVR model has undergone much application to become a relatively mature model, and it is suitable for small sample size.

In this study, an SVR algorithm combined with dimensions reduced by the PLS was used to establish a nonlinear regression model. When the parameters determined in the PLSR model (the pretreatment was FD + MSC, and dimensions reduced with PLS and NPC were 5) were introduced into the SVM algorithm, the SVR models were performed in the MATLAB R2014a software. The GS and GA were adopted to optimize the internal parameters (C, g). The model performance was validated with 6-fold cross-validation using the calibration set, and the prediction ability was evaluated with external validation using the prediction set. As shown in Table 6, the RMSECV in PLS-SVR-2 was 0.34, which is less than 0.4 in PLS-SVR-1, and the R^2 in PLS-SVR-2 was 93.29, which is larger than 91.54 in PLS-SVR-1. Therefore, the PLS-SVR-2 model with internal parameters (C, g) optimized with the GS had relatively excellent performance, and it was regarded as the suitable SVR model. The optimization of internal parameters (C, g) and predictive results are shown in Figure 6.

3.4.4. Analysis and Evaluation of Quantitative Models. In this study, the linear regression model of PLSR and nonlinear regression model of PLS-SVR were successfully established. As shown in Tables 5 and 6, the prediction ability had no significant difference between the two models. The relative deviations between the predictive value and the reference were 2.71% and 2.69%, respectively, and the average recoveries were 98.77% and 99.02%. Totally, the two optimized models had a reasonable structure and good prediction ability. Both of them could meet the requirements of accuracy and precision of quantitative analysis and could be used for nitrogen content analysis and quality evaluation of medicinal centipedes.

However, the PLSR model was built based on a linear regression algorithm to have characteristics of fast fitting and simple calculation, when the analysis requirements were not too high, and it would be widely used. In contrast, the SVR model was established based on the nonlinear regression algorithm, and it had the strong nonlinear fitting ability. It



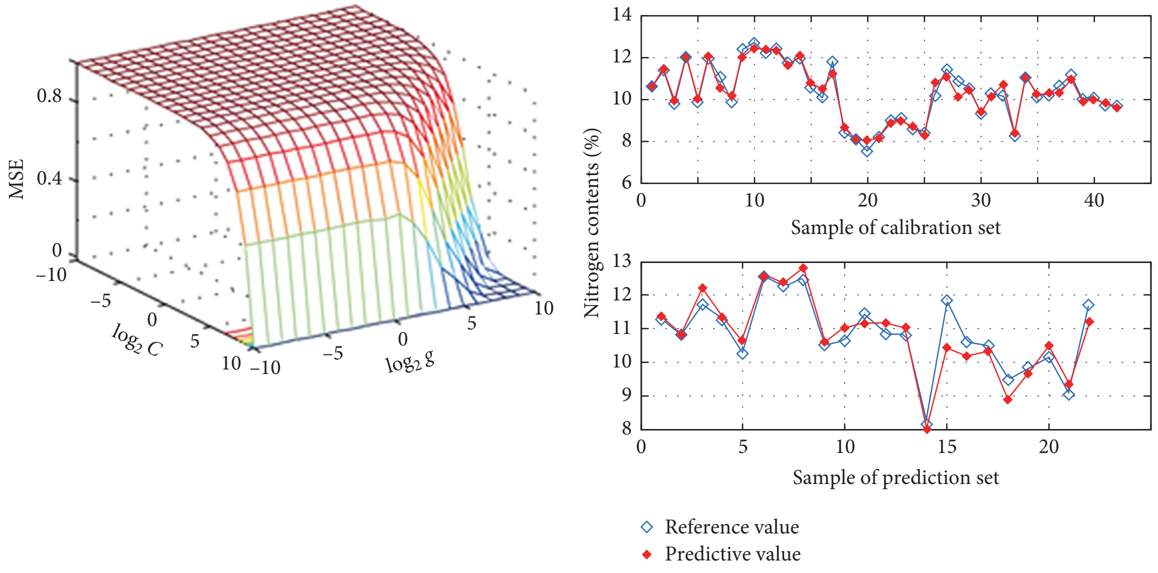
(a)

(b)

Figure 5: Predictive results in the calibration set (a) and prediction set (b) of the PLSR-6 model.

Table 6: Validation and evaluation results of SVR models.

Model number	Optimization method	C	g	6-fold cross-validation			External validation			RMSEE (%)
				RMSECV (%)	R^2 (%)	RPD	RMSEP (%)	R^2 (%)	RPD	
PLS-SVR-1	GA	99.99	997.03	0.4	91.54	2.91	0.41	85.89	2.55	0.34
PLS-SVR-2	GS	512	1024	0.34	93.29	3.72	0.43	85.5	2.54	0.32



(a)

(b)

Figure 6: Parameter optimization (a) and predictive results (b) of the PLS-SVR-2 model.

was shown from Tables 5 and 6 that the values of RMSECV and RMSEP of SVR models were generally less than those of PLSR models, and the values of R^2 of SVR models were commonly higher than those of PLSR models. It was indicated that the PLS-SVR model had perfect performance and better prediction effect than the PLSR model. For this reason, the PLS-SVR model now becomes the most widely used regression model in NIR spectroscopy analysis. Therefore, the PLS-SVR model with internal parameters (C, g) optimized with the GS was considered the most suitable NIR spectroscopy quantitative model for nitrogen content analysis of medicinal centipedes, and the PLSR model can act as supplement and verification for the analysis.

4. Conclusions

This study was carried out to explore the feasibility of using the NIR spectroscopy method to rapidly differentiate species and evaluate the quality of Chinese medicinal centipedes. In the qualitative analysis, after spectra were pretreated with the SD, dimensions were reduced with PCA, and internal parameters were optimized with the GS algorithm, a PCA-SVC model was set up using the scores of the first 8 PCs as input variables and category labels as output variables. The optimal model (SVC-12) was validated and evaluated, which could identify five species of medicinal centipedes with an accuracy of 100% (42/42) in the calibration set and 81.82% (18/22) in the prediction set. It could be accepted as an objective, rapid, and auxiliary method for identifying the species of medicinal centipedes. Through the spectra pretreated with FD + MSC, data dimension reduced with PLS, and NPC determined as 5, two best quantitative models of PLSR and PLS-SVR were also successfully determined. During the process of modeling, the RMSECV, R^2, and RPD of 6-fold internal cross-validation in the calibration set indicated the better performance and stronger modeling capacity. The RMSEP, R^2, and RPD of external validation in the prediction set proved stronger prediction ability. In addition, to investigate the reasonability of the model, the calibration set used for full cross-validation was input to the models again, and the RMSEE was used as an index. Comparing the characteristics and advantages of two different regression algorithms, the PLS-SVR-2 model had excellent performance and strong prediction capacity, and it was finally considered the most suitable quantitative model of NIR spectroscopy for nitrogen content analysis of medicinal centipedes.

Meanwhile, the pretreatment methods were also optimized in this paper; although the SD was determined in the qualitative model, MSC or its combined methods were applied to pretreat the spectra in quantitative models. The MSC had advantages of weakening or eliminating interference caused by the uneven grain size of solid powder in the diffuse reflection spectrum [29]. In this research, all samples were smashed into powder, and the NIR spectra were obtained with a diffuse reflection spectrum, so the application of FD + MSC was proved to be reasonable.

This study indicated that NIR spectroscopy combined with chemometric algorithms could be successfully used to differentiate species and evaluate the quality of medicinal centipedes in China, which was characterized with rapid, nondestructive, and environmentally friendly properties. However, this study just represented preliminary exploratory research; although 28 batch samples and 64 individuals were conducted, the sample size was still limited. In the future, more samples will be used to improve the prediction ability, and other algorithms will also be considered to simplify the model and improve performance. This study also provided a reference for rapid identification and quality analysis of other animal medicinal materials using NIR spectroscopy.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by a grant from the major drug discovery projects of the National Ministry of Science and Technology of China (no. 2014ZX09304307001).

References

[1] W. Siriwut, G. D. Edgecombe, C. Sutcharit, and S. Panha, "The centipede genus Scolopendra in mainland southeast asia: molecular phylogenetics, geometric morphometrics and external morphology as tools for species delimitation," *PLoS One*, vol. 10, no. 8, Article ID e0135355, 2015.

[2] S. H. Kang, H. Y. Deng, Z. Y. Jiang, Y. M. Liu, J. Li, and K. L. Chen, "Taxonomy and distribution of Chinese medicinal centipedes," *Journal of Chinese Medicinal Materials*, vol. 39, pp. 727–731, 2016.

[3] S. Yang, Y. Xiao, D. Kang et al., "Discovery of a selective NaV1.7 inhibitor from centipede venom with analgesic efficacy exceeding morphine in rodent pain models," *Proceedings of the National Academy of Sciences*, vol. 110, no. 43, pp. 17534–17539, 2013.

[4] Y. Kong, "Cytotoxic and anticoagulant peptide from Scolopendra subspinipes mutilans venom," *African Journal of Pharmacy and Pharmacology*, vol. 7, no. 31, pp. 2238–2245, 2013.

[5] I.-J. Jo, G. S. Bae, K. C. Park et al., "Scolopendra subspinipes mutilans protected the cerulein-induced acute pancreatitis by inhibiting high-mobility group box protein-1," *World Journal of Gastroenterology*, vol. 19, no. 10, pp. 1551–1562, 2013.

[6] W. Ma, D. Zhang, L. Zheng, Y. Zhan, and Y. Zhang, "Potential roles of Centipede Scolopendra extracts as a strategy against EGFR-dependent cancers," *American Journal of Translational Research*, vol. 7, no. 1, pp. 39–52, 2015.

[7] W. Ma, R. Liu, J. Qi, and Y. Zhang, "Extracts of centipede Scolopendra subspinipes mutilans induce cell cycle arrest and apoptosis in A375 human melanoma cells," *Oncology Letters*, vol. 8, no. 1, pp. 414–420, 2014.

[8] Y. Kong, S.-L. Huang, Y. Shao, S. Li, and J.-F. Wei, "Purification and characterization of a novel antithrombotic peptide from Scolopendra subspinipes mutilans," *Journal of Ethnopharmacology*, vol. 145, no. 1, pp. 182–186, 2013.

[9] M. Stankiewicz, A. Hamon, R. Benkhalifa et al., "Effects of a centipede venom fraction on insect nervous system, a native Xenopus oocyte receptor and on an expressed Drosophila muscarinic receptor," *Toxicon*, vol. 37, no. 10, pp. 1431–1445, 1999.

[10] H. Y. Zhang, J. Chen, J. Jia et al., "Identification of Scolopendra subspinipes mutilans and its adulterants using DNA barcode," *China Journal of Chinese Materia Medica*, vol. 39, p. 2208, 2014.

[11] H. Fang, F. Deng, and K. Q. Wang, "Chemical analysis of Scolopendra multidens newport," *Chinese Pharmaceutical Journal*, vol. 32, pp. 202–204, 1997.

[12] H. Fang, F. Deng, Y. C. Yan, and K. Q. Wang, "Chemical constituents of Scolopendra negrocapitis," *Journal of Chinese Medicinal Materials*, vol. 22, no. 5, pp. 226–228, 1999.

[13] X. Chen, H. M. Wen, R. Liu et al., "Analysis of extracted proteins of Scolopendra by nanoflow reversed phase liquid

chromatography-tandem mass spectrometry," *Chinese Journal of Analytical Chemistry*, vol. 42, pp. 239–243, 2014.

[14] J. Wang, Y. P. Chu, S. S Yang et al., "Influences of sterilization method on the nitrogen content of Scolopendra," *Asia-Pacific Traditional Medicine*, vol. 12, pp. 27-28, 2016.

[15] Y. He, X. L. Li, and Y. N. Shao, "Discrimination of varieties of apple using near infrared spectra based on principal component analysis and artificial neural network model," *Spectroscopy and Spectral Analysis*, vol. 26, no. 5, pp. 850–853, 2006.

[16] C. Xie, N. Xu, Y. Shao, and Y. He, "Using FT-NIR spectroscopy technique to determine arginine content in fermented Cordyceps sinensis mycelium," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 149, pp. 971–977, 2015.

[17] Y. Sun, L. Chen, B. Huang, and K. Chen, "A rapid identification method for calamine using near-infrared spectroscopy based on multi-reference correlation coefficient method and back propagation artificial neural network," *Applied Spectroscopy*, vol. 71, no. 7, pp. 1447–1456, 2017.

[18] Z. S. Song, D. X. Song, and M. S. Zhu, "Systematic classification of chilopoda and the order scolopendromorpha (myriapoda)," *Journal of Liaoning Normal University*, vol. 27, no. 1, pp. 69–72, 2004.

[19] C. Z. Zhang and K. Q. Wang, "A new centipede, Scolopendra negrocapitis sp. nov. from Hubei province, China (Chilopoda: Scolopendromorpha: Scolopendridae)," *Acta Zootaxonomica Sinica*, vol. 24, no. 2, pp. 136-137, 1999.

[20] X. L. Chu, H. F. Yuan, and W. Z. Lu, "Progress and application of spectral data pretreatment and wavelength selection methods in NIR analytical technique," *Progress in Chemistry*, vol. 16, no. 4, pp. 528–542, 2004.

[21] Q. Kang, Q. Ru, Y. Liu et al., "On-line monitoring the extract process of Fu-fang Shuanghua oral solution using near infrared spectroscopy and different PLS algorithms," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 152, pp. 431–437, 2016.

[22] W. Z. Lu, *Near Infrared Spectroscopy Instrument*, Chemical Industry Press, Beijing, China, 2010.

[23] G. Bázár, R. Romvári, A. Szabó, T. Somogyi, V. Éles, and R. Tsenkova, "NIR detection of honey adulteration reveals differences in water spectral pattern," *Food Chemistry*, vol. 194, pp. 873–880, 2016.

[24] Z.-S. Wu, L.-W. Zhou, S.-Y. Dai, X.-Y. Shi, and Y.-J. Qiao, "Evaluation of the value of near infrared (NIR) spectromicroscopy for the analysis of glycyrrizhic acid in licorice," *Chinese Journal of Natural Medicines*, vol. 13, no. 4, pp. 316–320, 2015.

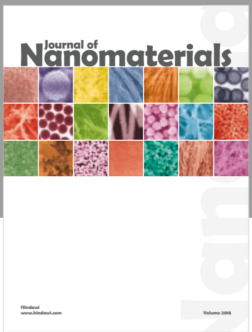
[25] M. Y. Yuan, B. S. Huang, C. Yu, Y. M. Liu, and K. L. Chen, "A NIR qualitative and quantitative model of 8 kinds of carbonate-containing mineral Chinese medicines," *China Journal of Chinese Materia Medica*, vol. 39, pp. 267–272, 2014.

[26] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer Berlin Heidelberg, New York, NY, USA, 1995.

[27] S. Wold, M. Sjöström, and L. Eriksson, "PLS-regression: a basic tool of chemometrics," *Chemometrics and Intelligent Laboratory Systems*, vol. 58, no. 2, pp. 109–130, 2001.

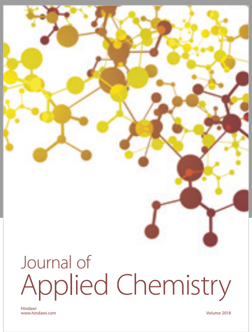
[28] L. Chen, J. Wang, Z. Ye et al., "Classification of Chinese honeys according to their floral origin by near infrared spectroscopy," *Food Chemistry*, vol. 135, no. 2, pp. 338–342, 2012.

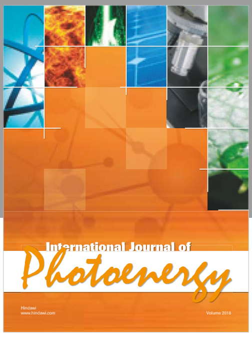
[29] W. M. Wang, D. M. Dong, W. G. Zheng, X. D. Zhao, L. Z. Jiao, and M. F. Wang, "Pretteatment method of near-infrared diffuse reflection spectra for sugar content prediction of pears," *Spectroscopy and Spectral Analysis*, vol. 33, no. 2, pp. 359–362, 2013.

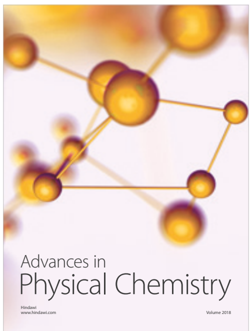

Journal of
Nanomaterials

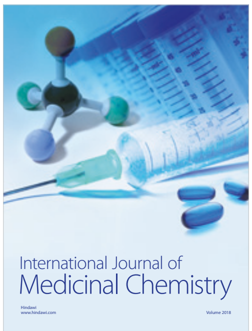
Journal of
Analytical Methods
in Chemistry

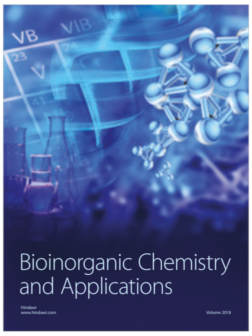
The Scientific
World Journal


Journal of
Applied Chemistry

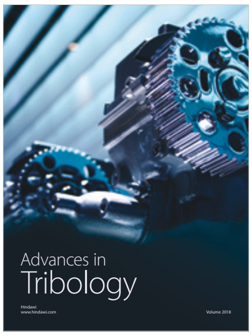

International Journal of
Photoenergy


Advances in
Physical Chemistry


International Journal of
Medicinal Chemistry


Bioinorganic Chemistry
and Applications

Journal of
Materials

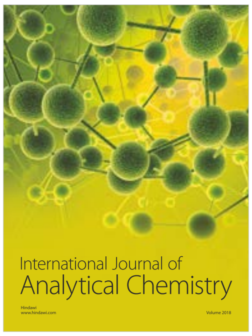

Advances in
Tribology

Journal of
Chemistry

BioMed
Research International

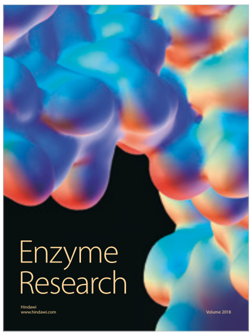
International Journal of
Spectroscopy

International Journal of
Electrochemistry


International Journal of
Analytical Chemistry

Journal of
Spectroscopy

Journal of
Nanotechnology


Enzyme
Research

Biochemistry
Research International