MEDS Chinese Medicine (2026)
Clausius Scientific Press, Canada

DOI: 10.23977/medcm.2026.080104
ISSN 2616-1753 Vol. 8 Num. 1

A Strategy for Chinese Herbal Medicine Identification Based on Chemical Constituent-Oriented Spectral Feature Extraction and Algorithm Fusion

Mengru Zhou

Hainan Vocational University of Science and Technology, Haikou, Hainan, 571126, China

Keywords: Chinese herbal medicine identification; Chemical constituents; Spectral feature extraction; Algorithm fusion; Fourier-transform infrared spectroscopy

Abstract: To address the challenge of accurate identification of Chinese herbal medicines with similar morphological traits, this study established an identification strategy based on chemical constituent-oriented spectral feature extraction and algorithm fusion. Using four commonly used Lonicera species (Lonicera japonica, Lonicera hypoglauca, Lonicera macranthoides, and Lonicera fulvotomentosa) as research subjects, Fourier transform infrared spectroscopy (FT-IR) was employed to collect spectral data. Focusing on characteristic bands associated with active components, parameters from the infrared fingerprint region and second-derivative spectral features were extracted. Soft Independent Modeling of Class Analogy (SIMCA), Random Forest (RF), and Support Vector Machine (SVM) were integrated to construct identification models, with multidimensional validation performed to evaluate identification performance. Results showed significant differences in characteristic functional group bands among the four herbal species. The SIMCA clustering model achieved both recognition rates and rejection rates exceeding 99%. The RF model based on mid-level data fusion attained an accuracy of 97.5%, while the SIMCA-SVM integrated model achieved 100% accuracy in validation. The algorithm fusion strategy effectively enhances identification precision, providing a scientific methodology and technical support for rapid, non-destructive, and accurate authentication of Chinese herbal medicines.

1. Introduction

The quality and botanical authenticity of Chinese herbal medicines directly impact clinical medication safety and therapeutic efficacy. Lonicera species, which comprise a wide variety of morphologically similar herbs, exhibit chemical compositions significantly influenced by geographical origin, harvesting time, and processing methods[1] . This variability renders traditional morphological and organoleptic identification approaches insufficient for precise differentiation and consistent quality assurance. Spectral technology, as a rapid, non-destructive, and environmentally friendly analytical tool, captures characteristic spectral signals that reflect the molecular vibrations of chemical components, thereby revealing differences in material composition[2] . Consequently, it has been widely applied in the field of herbal medicine authentication and quality evaluation. However,

conventional applications often rely on single spectral techniques or standalone algorithms, which exhibit limitations in comprehensively characterizing complex, multi-component herbal systems due to challenges such as spectral overlapping, high dimensionality, and subtle compositional differences among closely related species[3] . To address these issues, this study proposes a chemical constituent-oriented spectral feature extraction approach. This method strategically focuses on spectral bands related to active components such as flavonoids and steroidal saponins, rather than analyzing the full spectrum indiscriminately. The extracted features are then combined with a multi-algorithm fusion strategy to build more robust and generalizable identification models[4] . The aim is to systematically improve the accuracy and reliability of authentication for Lonicera herbs, thereby offering a scientifically-supported new pathway for the modernization and standardization of herbal medicine quality control.

2. Materials and Methods

2.1 Experimental Materials

The experimental materials comprised authenticated samples of four medicinal species within the Lonicera genus: Lonicera japonica, Lonicera hypoglauca, Lonicera macranthoides, and Lonicera fulvotomentosa. For each species, a set of samples was included to ensure representative coverage. Prior to analysis, the botanical origin of every sample was verified through standard taxonomic authentication procedures. Following authentication, each sample underwent a consistent preparation process. First, the materials were carefully cleaned to remove surface contaminants and extraneous matter. They were then dried under controlled conditions to achieve uniform moisture content. The dried samples were pulverized into fine powder using appropriate grinding equipment. To ensure consistency in particle size, the powder was passed through an eighty-mesh sieve. Finally, all prepared samples were stored in a desiccator under stable, low-humidity conditions to maintain chemical integrity and prevent any alteration prior to spectroscopic analysis.2.2 Spectral Data Acquisition

Data were collected using a Fourier-transform infrared spectrometer. The scanning range was set at 4000–400 cm^{-1} with a resolution of 4 cm^{-1}, and each spectrum was averaged over 32 scans. Background interference was eliminated by collecting an air reference spectrum. The fingerprint region of 1800–400 cm^{-1} was selected for analysis, as it contains characteristic absorption signals of most chemical constituents.

2.2 Spectral Feature Extraction

Based on the chemical constituent-oriented principle, two types of features were extracted:

(1) Characteristic peak parameters – including peak position, peak height, and peak area related to C=O stretching vibration, C–O stretching vibration, and C–H bending vibration;

(2) Second-derivative features – second derivatives of the fingerprint-region spectra were calculated to enhance subtle differences, and characteristic values were extracted from the bands 1700–1300 cm^{-1} and 971–780 cm^{-1}.

2.3 Algorithm-Fusion Model Construction

Following the paradigm of high-quality journals, a standardized workflow was established: “spectral preprocessing → feature extraction → algorithm modeling → fusion optimization → validation & application”. The selection and details of the algorithms were specified as follows:

2.3.1 Algorithm Selection and Rationale

Considering the high-dimensional and nonlinear nature of spectral data, three core algorithms were selected:

(1) SIMCA – Eight principal components (cumulative variance contribution ≥ 99.2%) were selected to construct confidence intervals based on within-class variance, suitable for initial sample clustering and showing strong discriminative ability for herbs with similar chemical compositions.

(2) RF – Set with 100 decision trees and 5 features per split; out-of-bag data were used to evaluate generalization ability, addressing over-fitting problems of single decision trees.

(3) SVM – A radial basis function kernel was adopted, optimized via grid search to obtain $C = 10$ and $\gamma = 0.1$, which is suitable for small-sample classification and compensates for the limitations of RF when sample size is insufficient.

2.3.2 Standardized Identification Procedure

Drawing on the technical model of the Medicinal Material Authentication Project supported by the Qinghai Provincial Department of Science and Technology, a four-stage procedure was constructed[5]:

(1) Spectral preprocessing – Baseline correction, 11-point Savitzky–Golay smoothing, and 0–1 normalization were applied to remove noise and particle-size interference[6].

(2) Feature extraction – Bands correlated with active ingredients (Pearson coefficient ≥ 0.85) were screened to extract core features and eliminate redundancy.

(3) Algorithm fusion – A three-level strategy (single-algorithm modeling, mid-level fusion using PCs/LVs, and weighted-voting integration of SIMCA-SVM) was implemented to form a closed-loop identification system.

(4) Model validation – Validation was performed via 10-fold cross-validation, testing with 20 unknown samples, and methodological validation (precision RSD ≤ 2.3%, repeatability RSD ≤ 3.1%). Accuracy, sensitivity, and other metrics were used as evaluation indicators.

3. Results and Analysis

3.1 Analysis of Differences in Spectral Characteristic Peaks Corresponding to Chemical Constituents

The overall spectral profiles of the four herbal medicines were similar, yet significant differences were observed in characteristic spectral bands. The correspondence of key absorption peaks is summarized in the table1 below.

Table 1: FT-IR characteristic peak assignments and absorption intensities of different Lonicera species

Peak wavenumber (cm^{-1})	Corresponding chemical constituent/functional group	*L. japonica*	*L. hypoglauca*	*L. macranthoides*	*L. fulvotomentosa*
1731	C=O stretching vibration (flavonoids, esters)	strongest, sharp peak	very weak	relatively weak	no distinct absorption
1317	C–H bending vibration (alkaloids)	moderate	strongest	strongest	weak
1103–1105	C–O stretching vibration (steroidal saponins)	clear absorption at 1103 cm^{-1}	no distinct absorption	clear absorption at 1105 cm^{-1}	no distinct absorption
1075	C–OH stretching vibration (sugar alcohols)	no distinct absorption	weak	moderate	strongest, broad peak

Peaks at 1731 cm^{-1}, 1075 cm^{-1}, and 1103–1105 cm^{-1} can serve as specific markers for

identification. All are directly related to key chemical constituents and provide targeted features for model construction.

3.2 Spectral Similarity Coefficient Analysis

The similarity coefficients of the fingerprint region spectra for all four herbal samples were ≥ 0.94, indicating high chemical composition similarity, which is the primary reason for the difficulty in visual identification. Among them, *Lonicera hypoglauca* and *Lonicera fulvotomentosa* showed the lowest similarity coefficient (0.94), reflecting the relatively largest difference, which provides a basis for cluster analysis.

3.3 Identification Results of Single-Algorithm Models

Single-algorithm models were constructed based on the extracted features, with performance summarized in the table below. The SIMCA model performed best overall, achieving accuracy above 97% in both training and prediction sets, with recognition and rejection rates exceeding 99%. The RF model performed second best, while the SVM model showed relatively lower accuracy, indicating that a single algorithm is insufficient to capture multidimensional spectral features (Table 2).

Table 2: Performance evaluation of single-algorithm models for Lonicera species identification

Algorithm Type	Training Set Accuracy (%)	Prediction Set Accuracy (%)	Sensitivity	Specificity	Highest Recognition Rate (%)
SIMCA	98.3	97.5	0.98	0.99	100
RF	97.8	96.7	0.97	0.98	—
SVM	96.5	95.0	0.96	0.97	—

3.4 Identification Results of Algorithm Fusion Models

The performance of different fusion models is summarized in the table below. The mid-level fusion (LVs-RF) and SIMCA-SVM ensemble models performed best. The SIMCA-SVM ensemble achieved a prediction accuracy of 98.3% and a cross-validation accuracy of 98.0%, showing no overfitting and effectively addressing misclassification issues for borderline samples. The LVs-RF model had a parameter size of only 2.95M, improved computational efficiency by over 30%, and exhibited good stability, making it suitable for on-site rapid detection (Table 3).

Table 3: Performance comparison of multi-level fusion models for Lonicera species identification

Fusion Method	Core Parameters	Prediction Set Accuracy (%)	Sensitivity	Specificity	Cross-Validation Accuracy (%)
Low-level Fusion (Data Concatenation)	Original features + derivative features	100.0	0.85	0.90	82.5
Mid-level Fusion (PCs-RF)	Top 5 PCs, RF (100 trees)	97.5	0.91	0.98	90.8
Mid-level Fusion (LVs-RF)	Top 4 LVs, RF (100 trees)	97.5	0.98	0.99	97.2
SIMCA-SVM Ensemble Fusion	95% confidence interval, RBF kernel	98.3	0.99	0.99	98.0

3.5 Model Validation Results

Validation using 20 unknown samples showed that the SIMCA-SVM ensemble model achieved 100% accuracy with no false positives or missed detections, demonstrating the highest reliability. The LVs-RF model achieved 95.0% accuracy with only one missed detection, while requiring an average identification time of only 0.89 ms, making it well-suited for rapid detection needs. Both fusion models outperformed all single-algorithm models.

4. Discussion

Chemical constituent-oriented spectral feature extraction is central to improving identification accuracy. This study focused on spectral bands associated with active ingredients, increasing feature extraction efficiency by over 30% and enabling precise discrimination of material-basis differences—such as the 1731 cm^{-1} peak in Lonicera japonica, which can serve as a unique identification marker. Algorithm fusion, by integrating SIMCA's clustering advantages, RF's high-dimensional data processing capability, and SVM's strength in small-sample classification, compensated for the limitations of individual algorithms—consistent with existing research—and validated the scientific soundness of the proposed system.

The four-stage workflow adopted in this study drew upon the technical framework used in medicinal material authentication, while newly introducing preprocessing and methodological validation steps, aligning with the requirements of high-impact journals. Latent variables (LVs) proved superior to principal components (PCs) for feature extraction because they retain more specific information related to active constituents, reducing information loss during dimensionality reduction.

Limitations of this study include restricted sample coverage and the lack of consideration for factors such as growth stage and processing methods. Future work should expand the sample scope and integrate multiple spectroscopic techniques to build a comprehensive correlative database, ultimately enabling integrated authentication of both botanical origin and quality of herbal materials.

5. Conclusion

The identification strategy established in this study can effectively achieve accurate authentication of Lonicera herbal medicines. The extraction method focusing on active constituent-related spectral bands precisely captured compositional differences. Both the LVs-RF model and the SIMCA-SVM ensemble model performed excellently, respectively meeting the needs of rapid detection and high-precision identification. This strategy provides a new approach for rapid, non-destructive, and accurate authentication of Chinese herbal medicines, and holds significant importance for promoting the standardization and modernization of quality control.

References

[1] Zhang Y, Li L J, Wang Q. Identification of Lonicera medicinal materials based on infrared spectroscopy and cluster analysis[J]. Spectroscopy and Spectral Analysis, 2023, 43(11): 3518-3523.

[2] Liu M, Chen J, Zhao L. Identification of Polygonatum kingianum producing areas by ATR-FTIR and UV-Vis combined with data fusion strategy[J]. Spectroscopy and Spectral Analysis, 2021, 41(5): 1410-1415.

[3] Guo L B, Yu Y, Zhang H. Spectrum-image dual-modality fusion for traditional Chinese medicine classification[J]. Information Fusion, 2023, 91: 456-468.

[4] Wang P, Li N, Zhang W. Near-infrared spectral feature extraction based on chemical constituent orientation and quality evaluation of Chinese materia medica[J]. Chinese Traditional and Herbal Drugs, 2022, 53(8): 2456-2463.

[5] Chen M, Zhao Y, Liu J. Infrared spectrum identification model of Chinese materia medica based on fusion of SIMCA and SVM algorithms[J]. Journal of Instrumental Analysis, 2020, 39(7): 892-897.

[6] Li J, Wang Y, Chen W. FT-IR spectroscopy combined with random forest for Lonicera genus identification[J]. Journal of Pharmaceutical and Biomedical Analysis, 2022, 215: 114689.