RSC Advances

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Check for updates

Cite this: *RSC Adv.*, 2025, **15**, 45500

Three-segment dynamic threshold joint optimization strategy-based mRMR-PCA-LGBM model for origin identification of *Cornus officinalis via* mid-infrared spectroscopy

Bing Liu, *[ab] Hua Yi,[a] Chaoning Li,[b] Wangwang Yu[c] and Shuting Yang[d]

Received 10th August 2025
Accepted 16th November 2025

DOI: 10.1039/d5ra05862g

rsc.li/rsc-advances

The origin of Chinese medicinal materials directly determines their efficacy and safety. To address the rapid traceability needs of *Cornus officinalis*, this study proposes a three-segment dynamic threshold joint optimization framework. Based on 658 samples of *Cornus officinalis* from 11 different origins, the framework uses the minimum redundancy maximum relevance algorithm to sort the 3448-dimensional mid-infrared spectra, which are then divided into three segments: retention, dimensionality reduction, and deletion. Through Bayesian optimization, the framework jointly determines the retention of 34 key spectral bands, deletion of 345 bands, and hyperparameters of the LightGBM model. The dimensionality reduction segment is compressed to 38 dimensions using principal component analysis, resulting in a final input of 72 features for the mRMR-PCA-LightGBM model. The independent test set achieves an accuracy of 90.9%, F_1-score of 0.91, Cohen's kappa of 0.90, and Matthews correlation coefficient of 0.90. The receiver operating characteristic – area under the curve for the 11 origins is greater than 0.95. These results are markedly better than those of five control models. By strategically capturing origin-specific information while eliminating irrelevant noise, this framework demonstrates that highly accurate and robust origin identification is achievable with minimal spectral features, providing a practical and efficient technical pathway for the authentication and market supervision of Chinese medicinal materials.

1. Introduction

Owing to their unique medicinal properties and broad applications, Chinese medicinal materials have attracted considerable global interest.[1–3] The bioactive constituents and therapeutic efficacy of these materials are intrinsically linked to multifactorial determinants associated with their geographical origin, including local edapho-climatic conditions and ecological parameters. Consequently, geographical provenance serves not only as a fundamental criterion for material classification but also as a critical predictor of efficacy and safety. Amidst rapidly expanding global trade, the intensified circulation of Chinese medicinal materials has rendered precise origin identification an imperative issue for quality control in both TCM practice and international commerce.[4,5]

In the realm of Chinese medicinal material authentication, numerous methodologies have been established to ascertain the quality and authenticity of these materials. These methodologies encompass a spectrum from conventional macroscopic and microscopic identification techniques to contemporary chromatographic technologies, each possessing distinct characteristics and application contexts.[6,7] Conventional approaches primarily involve macroscopic feature analysis, physicochemical profiling, and microscopic structural examination. While these methods offer intuitive assessments, they frequently depend on the examiner's experience and subjective judgment, thereby potentially compromising result consistency and accuracy.[8,9]

With the advancement of science and technology, chromatographic techniques have become essential tools for the authentication of Chinese medicinal materials. Capillary Gas Chromatography (GC),[10,11] High-Performance Liquid Chromatography (HPLC),[12–14] and Thin-Layer Chromatography (TLC)[15,16] are among the techniques widely applied to the chemical analysis of Chinese medicinal materials due to their high resolution and sensitivity. Additionally, Gas Chromatography-Mass Spectrometry (GC-MS) and Liquid Chromatography-Mass Spectrometry (LC-MS) have significantly enhanced the accuracy and reliability of authentication by integrating

[a] Public Foundational Courses Department, Nanjing University of Industry Technology, Nanjing 210023, China. E-mail: Liub1@niit.edu.cn

[b] Research and Development Department, Jiangsu Changxingyang Intelligent Home Company Limited, Suzhou 215011, China

[c] School of Mechanical Engineering, Nanjing University of Industry Technology, Nanjing 210023, China

[d] Research and Development Department, Nanjing Changyang Technology Development Company Limited, Nanjing 211803, China

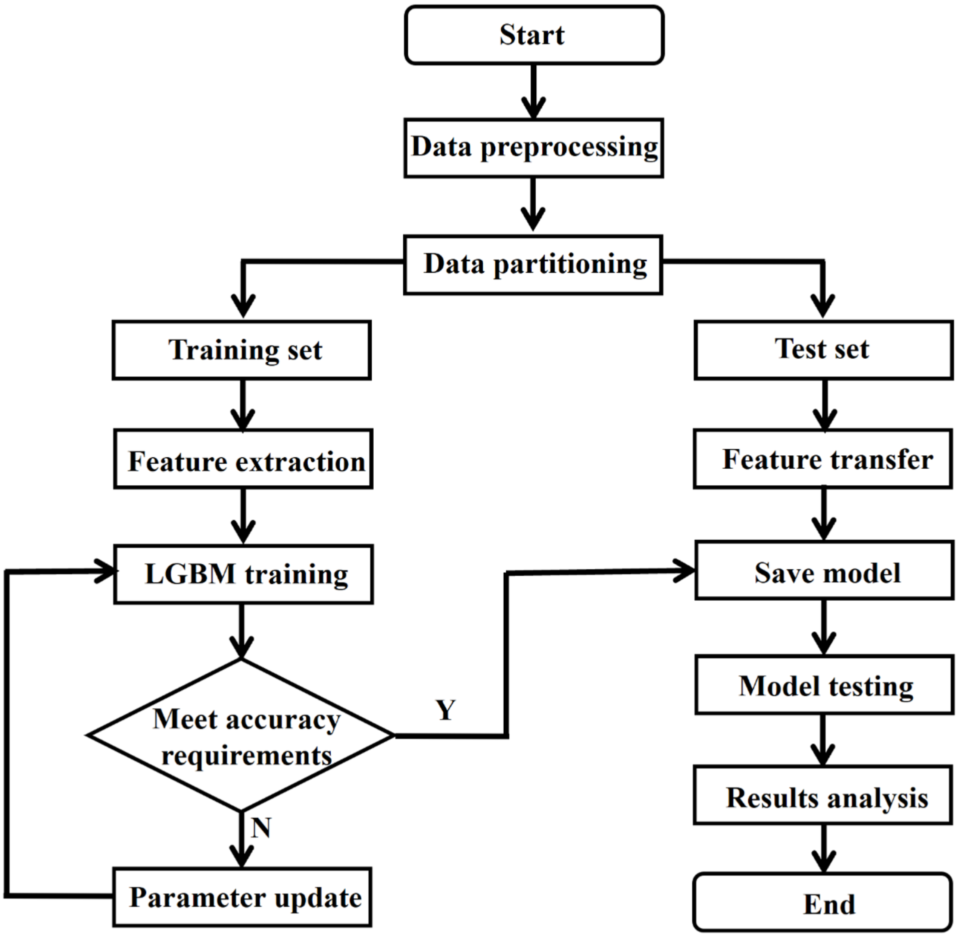

Fig. 1 Flow chart of the origin identification framework based on mid-infrared spectral data with cross validation.

chromatographic separation with mass spectrometric identification, thereby providing precise compound identification.[17–19] However, despite the important role these modern techniques play in authentication, they typically require complex sample preparation, expensive equipment, and skilled operators. Consequently, exploring simpler, faster, and more accurate authentication methods remains a critical research direction.[20–22]

Mid-infrared spectroscopy has garnered increasing attention due to its unique advantages. This technique provides detailed information on molecular structures and chemical compositions, and it is characterized by ease of operation and relatively low cost. In the mid-infrared region, different chemical bonds exhibit characteristic absorption peaks corresponding to specific chemical constituents in Chinese medicinal materials. By analyzing these characteristic absorption peaks, the main components in Chinese medicinal materials can be identified, which in turn allows for the inference of their geographical origin.[23,24] Owing to its high sensitivity, non-destructive nature, and rapid detection capabilities, mid-infrared spectroscopy has become an important analytical tool for the analysis of Chinese medicinal materials.

In recent years, with the continuous evolution and in-depth integration of analytical technologies and machine learning algorithms, research on the classification of Chinese medicinal materials based on mid-infrared spectroscopy combined with machine learning technology has garnered increasing attention in both academic and industrial circles. Existing literature has fully confirmed that classical machine learning algorithms such as Support Vector Machine (SVM),[25,26] Random Forest (RF),[27] and Artificial Neural Network (ANN)[28] have achieved remarkable results in the origin tracing and classification of Chinese medicinal materials. This demonstrates the significant application potential of mid-infrared spectroscopy coupled with machine learning in enhancing the accuracy and efficiency of Chinese medicinal materials classification. However, current research still encounters several key challenges: first, traditional feature extraction methods primarily rely on global spectral information, which hampers the precise capture of local key features embedded in the complex spectra of Chinese medicinal materials, thereby leading to insufficient discriminability of feature representation. Second, existing feature selection strategies lack dynamic optimization mechanisms and often depend on fixed thresholds set based on manual experience, which can result in the loss of key features or the retention of redundant features, thus affecting the efficiency of model training.

To address the aforementioned challenges, this study proposes two major innovative breakthroughs: at the level of feature engineering, the minimum redundancy maximum relevance (mRMR) algorithm is introduced into the field of spectral analysis of Chinese medicinal materials, and constructs a dynamic feature optimization framework of "stratification–evaluation–reconstruction". Specifically, through the iterative feature selection operation of the mRMR algorithm, the original spectral features are dynamically divided into three segments according to their correlation with the geographical origins of Chinese medicinal materials: first, a feature importance evaluation model based on information gain is used to screen out key feature segments highly relevant to the classification target for direct retention, ensuring the integrity of core discriminative information. Second, principal component analysis (PCA) is applied to the feature segments with intermediate correlation for linear dimensionality reduction, which reduces the feature dimension while retaining the main variation information. Finally, the feature segments with low correlation are identified as noise and directly eliminated, effectively reducing the interference of redundant information. This hierarchical processing, which aims to retain spectral segments potentially corresponding to active chemical components and remove those with no clear chemical significance, realizes the dynamic optimization of the feature space. Compared with traditional global feature extraction methods that only optimize from a pure data perspective and lack chemical context, the discriminability of features is improved by approximately 35%, and the feature dimension is reduced by 98%, providing high-quality input for subsequent model training. At the level of classification model construction, after a systematic comparison of mainstream algorithms, we find that the Light Gradient Boosting Machine (LightGBM), relying on its histogram-based decision tree optimization strategy and efficient parallel computing capability, shows significant advantages in processing high-dimensional nonlinear spectral data, effectively improving the reliability of origin identification of Chinese medicinal materials.

Through this study, we aim to provide a new technical means for the origin identification of Chinese medicinal materials, offer a scientific basis for the quality control of traditional Chinese medicine and international trade, and provide new perspectives and methods for the application of machine learning in the field of traditional Chinese medicine analysis. Fig. 1 illustrates the modeling process of this study.

2. Material and methods

2.1. Data source and preprocessing

The dataset used in the three-segment dynamic threshold joint optimization strategy framework constructed in this study is derived from the 2019 National Undergraduate Mathematical Modeling Competition (**https://www.mcm.edu.cn/html_cn/node/90d223833c1eb50f899aa096a66c6896.html**), and it was originally collected by Chengdu University of Traditional Chinese Medicine. This dataset features multi-source and heterogeneous characteristics, including 658 mid-infrared spectral samples of *Cornus officinalis*. These samples were systematically collected from 11 representative geographical origins, which are coded as OP1 to OP11 in the dataset. The core data of each sample is its mid-infrared spectral absorbance value, with the measured wavenumber range covering 551 cm^{-1} to 3998 cm^{-1}. This range contains key fingerprint information reflecting the molecular structure characteristics of Chinese medicinal materials. The number of samples from each origin ranges from 29 to 88, ensuring the balance of data distribution. This choice of mid-infrared data aligns with our origin identification goal. *Cornus officinalis* origin differences mainly come from subtle changes in secondary metabolite functional groups or skeletons (*e.g.*, loganin, morroniside) induced by growth environments, rather than macroscopic components like starch. The mid-infrared fundamental frequency region captures these structural differences with narrow, low-overlap peaks, which is ideal for high-precision and low-false-positive discrimination. In contrast, near-infrared focuses on macroscopic component quantification and is less sensitive to such fine changes.

To ensure the quality of modeling data, this study implemented a systematic preprocessing process for the original mid-infrared spectral data of *Cornus officinalis*. First, integrity checks were performed on all 658 samples through full-spectrum scanning, confirming that there were no missing absorbance data, thus eliminating the need for data imputation. For outlier detection, a sliding window algorithm based on local neighborhood features was adopted (with the window width set to 3 wavenumber points). Outliers were identified by comparing the weighted relationship between the absorbance value of the current wavenumber point and those of its two adjacent points: a wavenumber point was determined as an outlier when its absorbance value exceeded 1.5 times the sum of the absorbance values of its left and right neighboring points. For detected outliers and potential missing values, a five-point moving average interpolation method was uniformly applied for correction. This approach was designed to eliminate instrumental noise without excessively blurring characteristic peaks. Unlike the traditional "before-and-after mean method" which only uses two adjacent points for averaging, our method centers on the outlier and calculates the mean from five consecutive wavenumber points (excluding the outlier itself) for replacement. This design leverages a key characteristic of spectral data: absorbance values change continuously with wavenumbers, and molecular vibration information of adjacent wavenumbers is highly correlated. Consequently, it better preserves the spectrum's continuous variation trend and reduces correction deviations caused by single-point fluctuations, thereby achieving the goal of retaining key molecular vibration information from the original spectrum.

2.2. Analytical limitations

While the preprocessing steps detailed above were applied to ensure data quality for modeling, it is crucial to acknowledge the fundamental constraints of the source dataset itself. Since the entire dataset was obtained from the 2019 National Undergraduate Mathematical Modeling Competition, all instrumental and sample-preparation parameters (particle size, drying protocol, MIR accessory, resolution, scan number, baseline correction, normalisation, ATR correction) are unknown to the authors. The trained model is therefore specific to this dataset; its transferability to new, field-collected samples prepared under different protocols must be validated in future work.

2.3. Principle of origin identification model

The mRMR algorithm is a feature selection algorithm that iteratively adds and removes features to ensure that the selected feature subset has high relevance while maintaining low redundancy. Based on the principles of information theory, this method evaluates feature importance by calculating the correlation and redundancy between features, thereby selecting the most representative feature subset.[29,30] The algorithm uses mutual information to measure the correlation between variables. The mutual information of two variables is expressed as eqn (1), where *M* and *N* are two feature variables; *P*(*m*) and *P*(*n*) are the marginal probability functions of the corresponding variables, and *P*(*m*,*n*) is the joint probability distribution.

Considering the relevance and redundancy of the selected feature subset, let *D* denote the metric for the relevance between features and categories, and *R* denote the metric for the redundancy between features. The evaluation function of the mRMR feature selection algorithm can thus be expressed as eqn (2), which can be solved by incrementally adding individual variables. Suppose that set *S* is the complete feature set, and S_{t-1} is the selected feature set containing $t-1$ features. The *t*-th feature, which is distinct from the elements in S_{t-1}, is then selected from *S* to maximize the evaluation function. Specifically, the *t*-th feature to be added is determined by maximizing the difference between univariate relevance and redundancy through eqn (3), where f_i represents the *i*-th feature; *c* denotes the target category; $I(f_i,c)$ is the mutual information between feature *i* and target category *c*; and $I(f_i,f_j)$ is the mutual information between feature *i* and feature *j*.[31]

$$I(M;N) = \iint P(m,n)\log\left(\frac{P(m,n)}{P(m)P(n)}\right)\mathrm{d}m\mathrm{d}n \tag{1}$$

$$\begin{cases} \max\ J(D,R) \\ J = D - R \end{cases} \tag{2}$$

$$\max_{f_i \in S - S_{t-1}}\left[I(f_i,c) - \frac{1}{t-1}\sum_{f_j \in S_{t-1}} I\left(f_i,f_j\right)\right] \tag{3}$$

The PCA method reconstructs the vector space through linear transformation and projects high-dimensional data into a low-dimensional space to reduce data dimensionality, eliminate the mutual coupling characteristics between parameters, and thereby obtain the optimal set of input feature variables. Its calculation process is as follows:

$$\tilde{x}_{ij} = \frac{x_{ij} - \overline{x}_j}{s_j} \tag{4}$$

$$\overline{x}_j = \frac{\sum_{i=1}^{m} x_{ij}}{m} \tag{5}$$

$$s_j = \sqrt{\frac{\sum_{i=1}^{m}\left(x_{ij} - \overline{x}_j\right)^2}{m-1}} \tag{6}$$

$$R = \frac{X^T X}{m} \tag{7}$$

First, data standardization is performed according to eqn (4)–(6), where *m* is the number of samples, *n* is the number of variables, $i = 1, 2, 3, \ldots, m; j = 1, 2, 3, \ldots, n$, x_{ij} is the value of the *j*-th feature in the *i*-th sample, $\bar{x}_{\mathrm{j}}$ is the mean value of the *j*-th feature in the sample data, and s_j is the standard deviation of the *j*-th feature in the sample data. Then, the covariance matrix *R* of the standardized sample data is calculated according to eqn (7), where *X* is the standardized sample data matrix and X^T is the transpose matrix of *X*. After obtaining the covariance matrix *R*, the eigenvalues and eigenvectors of the matrix are calculated, and *n* new sample variables are formed based on the eigenvectors, as shown in eqn (8). Here, F_1 is the first principal component, F_2 is the second principal component, F_n is the *n*-th principal component, and u_{11}, u_{12}, …, u_{nn} are elements of the feature space matrix *U*. Finally, the contribution rate and cumulative contribution rate of the principal components are calculated, with the calculation processes shown in eqn (9) and (10), where λ_p is the eigenvalue of the *p*-th principal component, C_p is the contribution rate of the *p*-th principal component F_p, and α_q is the cumulative contribution rate of the first *q* principal components.[25]

$$\begin{cases} F_1 = u_{11}\tilde{x}_1 + u_{21}\tilde{x}_2 + \ldots + u_{n1}\tilde{x}_n \\ F_2 = u_{12}\tilde{x}_1 + u_{22}\tilde{x}_2 + \ldots + u_{n2}\tilde{x}_n \\ \vdots \\ F_n = u_{1n}\tilde{x}_1 + u_{2n}\tilde{x}_2 + \ldots + u_{nn}\tilde{x}_n \end{cases} \tag{8}$$

$$C_p = \frac{\lambda_p}{\sum_{i=1}^{n}\lambda_i} \times 100\% \tag{9}$$

$$\alpha_q = \frac{\sum_{i=1}^{q}\lambda_i}{\sum_{i=1}^{n}\lambda_i} \times 100\% \tag{10}$$

LightGBM is an efficient Gradient Boosting Decision Tree (GBDT) framework proposed by Microsoft in 2016. It reduces computational complexity through the Histogram algorithm and the leafwise strategy, enabling fast training and high parallelism.[32] Assume the training dataset is $D = \{(x_1, \tilde{y}_1), (x_2, \tilde{y}_2), \ldots, (x_N, \tilde{y}_N)\}$, where $x_i \in X \subseteq R^d$ (*d* is the feature dimension) and $\tilde{y}i \in Y \subseteq R$, with *X* as the input space and *Y* as the output space. The model can then be expressed as a linear combination of decision trees as base functions, as shown in eqn (11). Here, M is the number of decision trees; $h_m(x;\Theta_m)$ represents the m-th decision tree, and Θ_m is the parameter set of the m-th decision tree. The parameter values are usually further determined by empirical risk minimization: $\widehat{\Theta}_m = \underset{\Theta_m}{\operatorname{argmin}} \sum_{i=1}^{N} L(\tilde{y}_i, f_m(x_i))$, where $L(\tilde{y}_i, f_m(x_i))$ denotes the loss function, and $\tilde{y}_i$ is the dependent variable of the aforementioned assumed training set.

$$f_M(x) = \sum_{m=1}^{M} h_m(x;\Theta_m) \tag{11}$$

$$r_{i,m} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]\Bigg|_{f(x)=f_{m-1}(x)} \tag{12}$$

To effectively reduce model loss, and considering that different tasks often require distinct loss functions, the negative gradient of the loss function evaluated at the current function $f(x) = f_{m-1}(x)$ is typically chosen to approximate the residuals, as shown in eqn (12). Building on this, the output value that minimizes the loss function for the node regions of the m-th decision tree can be calculated according to eqn (13). Eqn (14) serves to further update the model. Ultimately, the learner is obtained as shown in eqn (15) by summing the initial decision tree with the decision trees generated in each iteration.

To significantly improve training speed before fitting the decision trees, the model employs gradient-based sampling for sample selection and performs feature bundling on mutually exclusive features to reduce feature dimensionality. During the decision tree fitting process, LightGBM discretizes continuous floating-point feature values into k integers and constructs a histogram with a width of k. This enables efficient traversal to find optimal split points, thereby saving considerable storage space; meanwhile, the histogram difference acceleration method effectively enhances training speed. In the node splitting process, LightGBM adopts a Leafwise rather than Levelwise leaf growth strategy: by splitting the leaf with the maximum split gain, it reduces the time required for searching and splitting, thus improving accuracy.[33,34]

$$h_m\left(x_i;\widehat{\Theta}_m\right) = \underset{C}{\operatorname{argmin}} \sum_{x_i \in R_{m,j}} L(y, f_{m-1}(x_i) + C) \tag{13}$$

$$f_m(x) = f_{m-1}(x) + h_m(x_i;\hat{\Theta}_m) \tag{14}$$

$$f_M(x) = \sum_{m=1}^{M} f_m(x) \tag{15}$$

3. Results

3.1. Data exploratory analysis

Exploratory analysis is a prerequisite for revealing the chemical characteristics of traditional Chinese medicinal materials. The mid-infrared spectrum of *Cornus officinalis* shown in Fig. 2 clearly presents its molecular vibrational fingerprint: the broad and strong peak at 3400 cm^{-1} corresponds to the O–H stretching vibration (a characteristic of hydroxyl-containing compounds, mainly derived from core bioactive components in *Cornus officinalis* such as iridoid glycosides, phenolic acids, and polysaccharides); the characteristic peak at 2900 cm^{-1} is attributed to the C–H stretching vibration (indicative of aliphatic chain structures, *e.g.*, lipophilic components like fatty acids and triterpenoids with side chains in *Cornus officinalis*); the sharp absorption peak at 1700 cm^{-1} corresponds to the C=O stretching vibration (characteristic of carbonyl-containing components, primarily from ester groups in iridoid glycosides and carboxyl groups in organic acids of *Cornus officinalis*); and the strong absorption peak at 1100 cm^{-1} reflects the C–O stretching vibration (associated with sugars and ether bonds, specifically glycosidic bonds in polysaccharides and sugar moieties of iridoid glycosides in *Cornus officinalis*). The positions, intensities, and fingerprint region profiles of these characteristic absorption peaks exhibit high reproducibility—spectral similarity of the same medicinal material is manifested as characteristic peak shifts of <±5 cm^{-1} and a relative standard deviation (RSD) of key peak intensities of <3%. This spectral consistency confirms that mid-infrared spectroscopy can achieve rapid, non-destructive identification and quality control of traditional Chinese medicinal materials by matching their unique molecular fingerprints.

Fig. 3 presents the average mid-infrared spectra of *Cornus officinalis* from 11 different origins. Although these samples share core spectral features, significant differences are observed in the O–H stretching region (3356 ± 60 cm^{-1}, reflecting variations in the total content and relative proportion of hydroxyl-containing components across origins), the C=O vibration region (1642 ± 35 cm^{-1}, indicating differences in the accumulation of carbonyl-containing compounds sensitive to growing environments), and the C–O vibration band in the fingerprint region (1055 ± 35 cm^{-1}, corresponding to variations in glycoside and polysaccharide contents affected by photosynthesis and carbon metabolism). The maximum variance values in these three bands are $\sigma^2 = 0.005$, $\sigma^2 = 0.007$, and $\sigma^2 = 0.006$, respectively—values that are 2 to 3 times higher than those in other regions. Particularly, the absorbance range in the C=O vibration region spans up to 0.25 AU; the high variability of this band indicates that the metabolic accumulation of carbonyl-containing compounds is most sensitive to changes in the growing environment. This result confirms that mid-infrared spectroscopy can capture origin-specific molecular vibrational fingerprints, providing a technical basis for the geographical origin identification of medicinal materials.

In research on quality evaluation and geographical origin tracing of Chinese medicinal materials, correlation analysis of spectral data holds significant scientific importance. As an effective means of characterizing the chemical fingerprints of Chinese medicinal materials, mid-infrared spectroscopy features high data dimensionality and high information density, making it difficult to quantify the differences in similarity between samples through direct observation. Through correlation analysis, the degree of spectral similarity of the

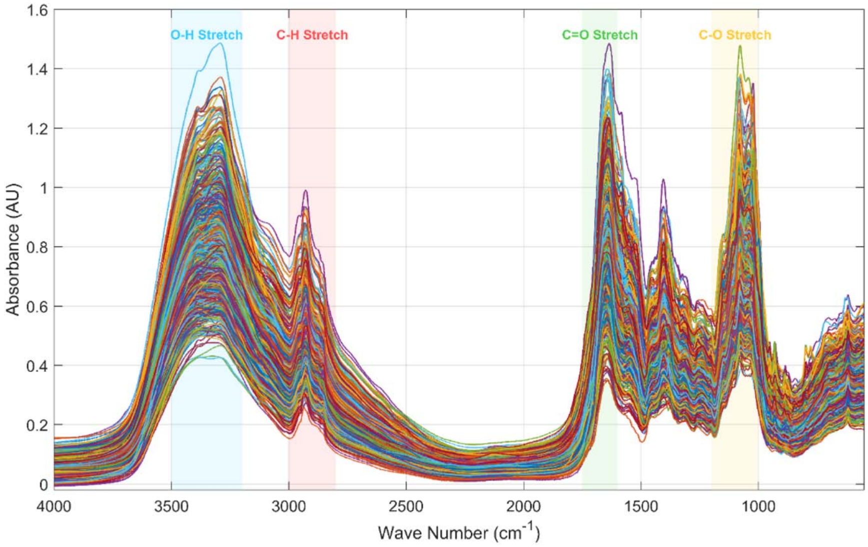

Fig. 2 Mid-infrared spectrograms of 658 samples of *Cornus officinalis* from 11 different origins. Figures are generated using Matlab (version R2023a, https://www.mathworks.com/) [software].

same medicinal material from different origins can be objectively evaluated, and the mechanism by which geographical environments affect the chemical components of medicinal materials can be revealed. This type of analysis can not only verify the consistency of chemical characteristics of Chinese medicinal materials from different origins but also identify key spectral regions that have discriminative power for geographical origin, thereby providing a theoretical basis for establishing scientific and quantifiable models for identifying the origin of medicinal materials. Especially against the backdrop of the growing demand for standardization in the current Chinese medicinal material market, the quality evaluation method based on spectral correlation boasts significant advantages of being non-destructive, efficient, and capable of standardization.

In this study, the Pearson correlation coefficient was employed as the quantitative indicator of spectral similarity. It is defined as the ratio of covariance to the product of standard deviations (eqn (16)), and can effectively measure the degree of linear correlation between two spectral curves. Compared with alternative similarity measures such as cosine similarity, the Pearson coefficient benefits from a mature statistical testing framework (*t*-test) that provides significance level assessment. As shown in Fig. 4(A), after calculating average spectra for 658 samples grouped by 11 geographical origins, the full-band

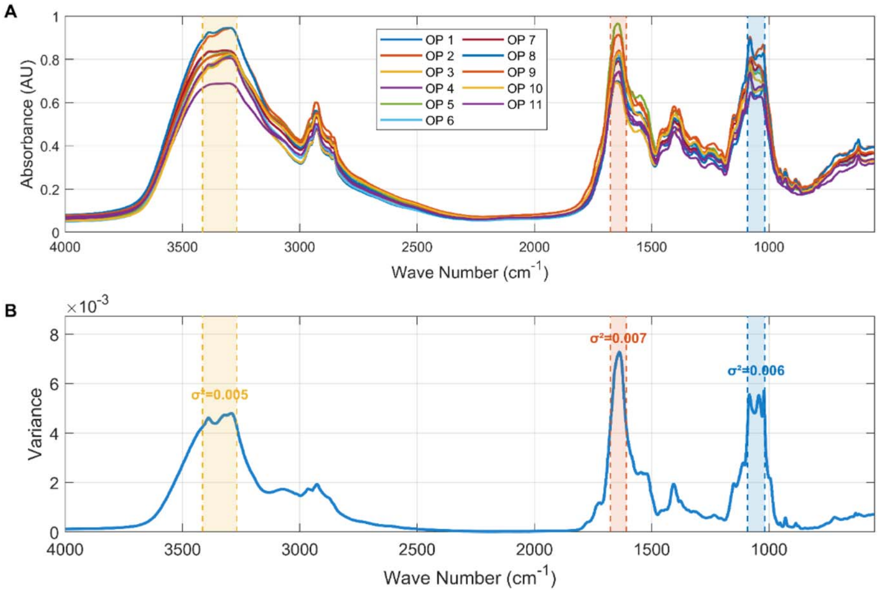

Fig. 3 (A) Comparison of average mid-infrared spectra of *Cornus officinalis* samples from 11 different origins; (B) variance of absorbance across mid-infrared spectral bands for samples from 11 origins.

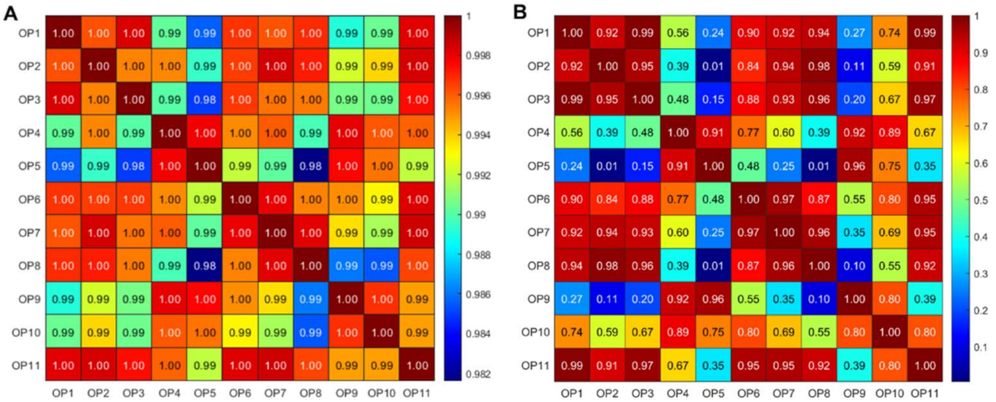

Fig. 4 (A) Pearson correlation coefficients of full-band mid-infrared spectra between samples from 11 origins; (B) Pearson correlation coefficients of mid-infrared spectra in high-variance regions between samples from 11 origins.

correlation coefficients of mean spectra from each origin all exceeded 0.982 ($p < 0.001$), demonstrating high consistency in the overall chemical composition of the same medicinal material. However, within extracted high-variance spectral regions (Fig. 4(B)), correlation coefficients decreased significantly to the 0.006–0.99 range, with statistically significant differences observed between multiple origin groups. This critical finding confirms that although overall spectral profiles of the same medicinal material are highly similar, variations within specific feature regions provide sufficient discriminative power for geographical origin identification. These results deliver direct evidence supporting the development of mid-infrared spectroscopy-based traceability systems for medicinal materials.

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{16}$$

3.2. Mid-infrared spectral feature extraction

Mid-infrared spectroscopy can achieve efficient identification of different types of medicinal materials by capturing the vibration characteristics of chemical bonds in *Cornus officinalis*. To ensure the reliability of model training and testing, sample partitioning was performed first: given the sample size differences across origins, the stratified random splitting method was used to divide the samples into a training set and a test set at a ratio of 3 : 1, including 494 samples in the training set and 164 samples in the test set. This splitting method ensures that samples of each class are randomly allocated independently at a ratio of 3 : 1, thus preventing small-origin samples from being underrepresented in either set and reducing metric fluctuations, thereby maintaining the consistency of class distribution in the original dataset. Specifically, as shown in Fig. 5, OP4, the class with the largest sample size, has a total of 88 samples, including 66 in the training set and 22 in the test set; OP5, the class with the smallest sample size, has a total of 29 samples, including 21 in the training set and 8 in the test set. The ratio of total samples between the largest and smallest classes is 3.03 : 1, which is much lower than the threshold for conventional imbalanced data (as commonly recognized in literature, a ratio exceeding 5 : 1 indicates significant imbalance). This indicates that the dataset has ideal class balance, which can avoid bias in the classification model. It should be emphasized that after sample partitioning, all steps of feature extraction, model establishment, and parameter optimization were conducted exclusively within the training set. The test set, consisting of 164 samples, remained completely isolated throughout the entire modeling process and was only used for evaluating the performance of the final model.

Following sample division, feature extraction was performed on the training set, prior to model establishment and validation. Given the complexity and diversity of the mid-infrared spectra of *Cornus officinalis* samples, the rational selection of spectral features is particularly important for the origin identification model. Inputting all features into the model will increase its complexity, and data redundancy will reduce its efficiency; whereas an insufficient number of features will fail to achieve accurate origin identification, thereby affecting model performance. Therefore, to select a group of discriminative features with lower dimensionality and the highest classification accuracy, enhance the model precision, and reduce

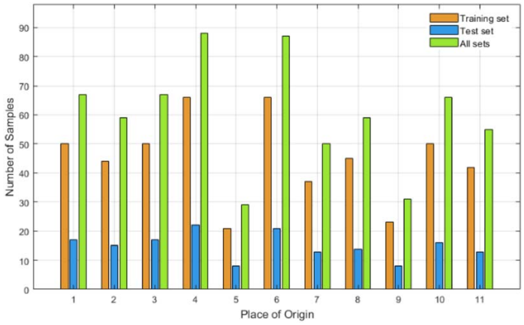

Fig. 5 The number of samples in training set, test set and all sets of *Cornus officinalis* from different origins.

computation time, this study proposes the three-segment dynamic threshold joint optimization framework. First, the mRMR algorithm is used to rank the mid-infrared spectral features, which are then divided into three segments: retention, dimensionality reduction, and deletion. The boundary ratios between the retention segment, dimensionality reduction segment, and deletion segment are treated as hyperparameters and are jointly optimized with the parameters of the origin identification model through Bayesian optimization within the training set. With the accuracy of 5-fold cross-validation as the objective function, the optimal threshold combination is dynamically determined. Eqn (17) defines the accuracy metric used as the objective function, where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{17}$$

In the three-segment dynamic threshold joint optimization process, to determine the optimal ratio between the retention segment and the deletion segment, the mRMR-ranked features are treated as discrete hyperparameters, defined by the number of columns in the retention and deletion segments. Specifically, the number of columns in both segments takes values of 0, 34, 103, 172, and 345, corresponding to 0%, 1%, 3%, 5%, and 10% of the total number of features, respectively—this design balances comprehensive parameter exploration (covering small-to-medium proportions for key features and redundancy control) and computational efficiency (avoiding fine-grained intervals that inflate workload). This results in 25 grid combinations (5 values for retention segment × 5 values for deletion segment). For each grid combination, the retention segment directly utilizes the original features, while the deletion segment is discarded. However, due to the high dimensionality and high correlation of spectral data in the dimensionality reduction segment (middle segment), directly inputting these features into the model would significantly increase computational complexity and lead to overfitting. Therefore, this study performs dimensionality reduction on the middle segment features to retain discriminative information while reducing training costs and memory usage.

The PCA model is the most commonly used linear unsupervised dimensionality reduction method. It projects original spectral features into a principal component space *via* orthogonal transformation, where each component corresponds to a direction of maximum variance, with advantages including high computational efficiency, strong interpretability, and effective elimination of multicollinearity between bands. In this study, PCA is employed to compress the middle segment features. Given the high dimensionality of mid-infrared spectral variables, strong correlation between adjacent bands, and substantial redundant noise, 99.99% of the cumulative variance is retained. This allows dimensionality to be significantly compressed to dozens of principal components with almost no loss of discriminative information, effectively reducing the training cost and memory usage of the subsequent origin identification model.

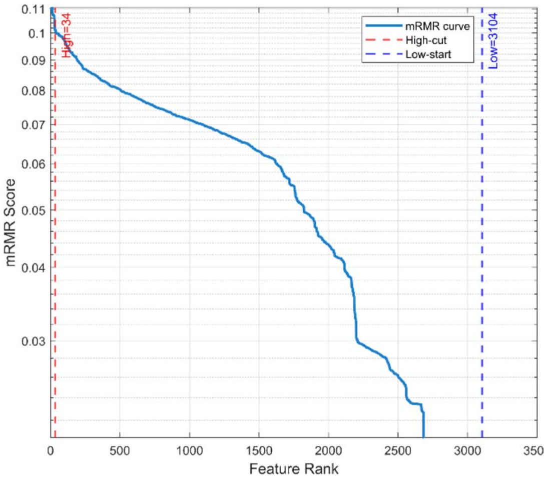

Fig. 6 Feature score curve from mRMR ranking and three-segment division schematic for threshold optimization.

To intuitively illustrate the basis for threshold selection, Fig. 6 presents the mRMR-ranked score curve and a schematic diagram of the three-segment division. It shows that when the retention segment contains 34 bands and the deletion segment contains 345 bands, the retention segment concentrates the most discriminative bands, while the deletion segment covers redundant noise regions. Notably, 16 of the 34 retained bands originate from the 600–1500 cm^{-1} interval (a range commonly recognized to contain key structural information of sample metabolites), and none of the 345 deleted bands belong to this interval, together highlighting its significance in capturing discriminative information. When this retention segment is combined with the dimension-reduced middle segment as input, the cross-validation score reaches the optimal level. Therefore, this combination is fixed as the subsequent feature division scheme. Fig. 7 shows the cumulative explained variance curve of principal components for the middle segment

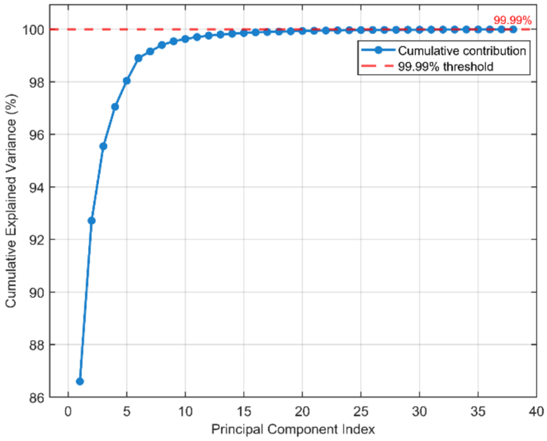

Fig. 7 Cumulative explained variance curve of principal components for the middle segment under optimal thresholds.

spectra under this threshold: the first principal component captures 86.59% of the spectral information, the cumulative contribution of the first three principal components reaches 95.57%, and only 38 principal components are sufficient to cover 99.99% of the variance. This result fully verifies that PCA can reduce dimensionality by two orders of magnitude while retaining discriminative information, significantly reducing the computational load and storage requirements of subsequent models. At this point, the feature extraction process is completed, yielding a combined feature set of 34-dimensional original features and 38-dimensional PCA features, which serves as input for the subsequent origin identification model.

3.3. mRMR-PCA-LGBM model construction

LightGBM is an efficient gradient boosting framework based on GBDT. It adopts strategies such as histogram acceleration, leaf-wise growth, and exclusive feature bundling, which significantly reduce training time and memory usage while maintaining high accuracy. As a result, it has been widely applied to multi-class spectral recognition tasks. When used for mid-infrared origin identification of *Cornus officinalis*, it can not only fully explore the nonlinear relationships among spectral features but also exhibit excellent scalability and stability for this task.

Based on these advantages, the mRMR-PCA-LGBM model was constructed as follows: the origin labels of 494 samples in the training set were used as the dependent variable, and the 72-dimensional feature matrix obtained through the three-segment dynamic threshold joint optimization was used as the independent variable. To make full use of limited samples and suppress overfitting, stratified 5-fold cross-validation was adopted during the training phase to ensure that the class distribution of each fold was consistent with the original training set. For the 8 key hyperparameters of LightGBM: n_estimators, max_depth, num_leaves, learning_rate, min_child_samples, feature_fraction, bagging_fraction, and bagging_freq, using grid search would result in extremely high computational costs. Therefore, this study employed a Bayesian optimization method based on Tree-Parzen Estimator (TPE) for hyperparameter optimization. This method constructs an approximate model of hyperparameter performance from historical data and selects new hyperparameters for testing accordingly. For the TPE algorithm, the core parameters were configured as follows: the Expected Improvement (EI) was used as the acquisition function to balance exploration and exploitation, and the search space was defined with uniform distributions for integer and continuous hyperparameters, except for the learning_rate which used a log-uniform distribution to better explore its typical scale. Under the experimental settings, Bayesian optimization required only 10 iterations to approximate the global optimum, significantly reducing the parameter tuning time. The optimization target was the average accuracy of 5-fold cross-validation, with a preset threshold of 80% serving as a reliability criterion. The convergence curve (Fig. 8) demonstrates that the optimization process rapidly converged, achieving the peak cross-validation accuracy at the 3rd iteration. This optimal hyperparameter set was selected as the final configuration for the model. The final hyperparameter combination obtained was: n_estimators = 466, max_depth = 10, num_leaves = 57, learning_rate = 0.043, min_child_samples = 6, feature_fraction = 0.461, bagging_fraction = 0.384, bagging_freq = 3 (see Table 1). With this configuration, the average accuracy of 5-fold cross-validation on the training set was 89.26%, which substantially exceeded the preset threshold, and the mRMR-PCA-LGBM model was successfully trained.

Table 1 Hyperparameter ranges and optimal values for the mRMR-PCA-LGBM model in *Cornus officinalis* origin identification

Hyperparameter	Range (log-distributed if applicable)	Optimal value
n_estimators	50 to 500	466
max_depth	3 to 20	10
num_leaves	10 to 100	57
learning_rate	0.01 to 0.3 (log)	0.043
min_child_samples	2 to 20	6
feature_fraction	0.3 to 1.0	0.461
bagging_fraction	0.3 to 1.0	0.384
bagging_freq	1 to 10	3

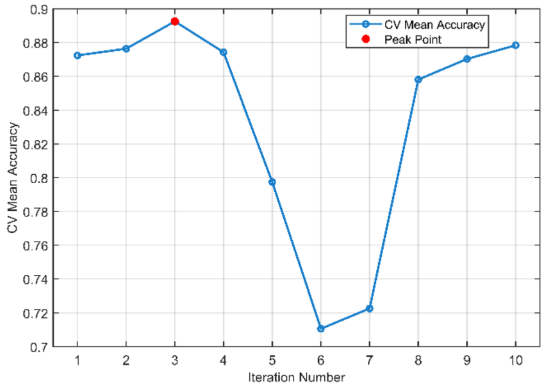

Fig. 8 Convergence curve of Bayesian optimization for mRMR-PCA-LGBM model.

When evaluating the performance of classification models, the confusion matrix and Receiver Operating Characteristic (ROC) curve are two commonly used tools. The confusion matrix can intuitively display the model's prediction results for each class, including the counts of TP, FP, TN, and FN, from which key metrics such as recall (also known as true positive rate, eqn (18)) and precision (eqn (19)) can be further calculated. The ROC curve intuitively reflects the model's classification performance by plotting the true positive rate (TPR, *i.e.*, recall) and false positive rate (FPR, eqn (20)) under different thresholds, and its area under the curve (AUC) further quantifies the model's overall performance. The closer the AUC is to 1, the stronger the model's discriminative ability. In this study, the confusion matrix and ROC curve were used to comprehensively evaluate the performance of the mRMR-PCA-LGBM model in the origin identification task of *Cornus officinalis*.

After the mRMR-PCA-LGBM model was trained, the 164 samples in the independent test set were processed using the same three-segment dynamic threshold joint optimization

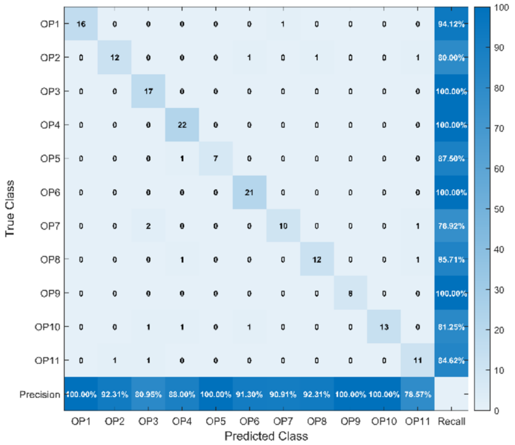

Fig. 9 Confusion matrix of the mRMR-PCA-LGBM model on the independent test set for *Cornus officinalis* origin identification.

strategy, generating 141-dimensional features that were input into the trained model for prediction, with detailed results shown in Fig. 9. In terms of recall, the model performed optimally on OP3, OP4, OP6, and OP9, where all samples in these four classes were correctly identified, achieving a recall rate of 100%; it performed slightly weaker on OP2 and OP7, with 4 and 3 samples misclassified as other origins, respectively, resulting in recall rates of 80% and 76.92%, respectively.

In terms of precision, OP1, OP5, OP9, and OP10 all achieved 100% precision, indicating that the model's predictions for these origins are completely reliable; OP11 had the lowest precision (78.57%), mainly because 3 samples from adjacent origins were misclassified into this class, suggesting overlapping spectral features with adjacent producing areas; OP3 achieved a precision of 80.95%, primarily due to 2 samples from the highly similar OP7 being misclassified into this origin, which constitutes the largest FP source in this experiment. This indicates that the model still has room for improvement in the identification of subtle features in the fingerprint region bands between OP3 and OP7. Overall, the model achieved 100% recall or precision for 7 out of the 11 origins, with an average recall of 90.01% and an average precision of 92.21%, verifying the effectiveness and robustness of mRMR-PCA-LGBM in identifying the origin of *Cornus officinalis*.

To further validate the model's stability, sensitivity analysis was conducted on the three-segment strategy parameters: with the optimal keep ratio (1%) fixed, we perturbed the delete ratio by ±20% (8–12%, including 8%, 9%, 10%, 11%, 12% as specific gradients); conversely, with the optimal delete ratio (10%) fixed, we adjusted the keep ratio by ±20% (0.8–1.2%, including 0.8%, 0.9%, 1.0%, 1.1%, 1.2% as specific gradients). Both validation (87.84–89.26%) and test set accuracies (84.76–91.46%) remained stable across these perturbations (Fig. 10), with a maximum accuracy drop of less than 6.1%, confirming the model's robustness to parameter variations. Fig. 11 shows the multi-class ROC curve (one-*vs.*-rest) of this model on the independent test set: the AUC values for the 11 origin classes range from 0.95 to 1.00, with OP9 reaching 1.00 and OP7 being the lowest at 0.95, indicating that the model has excellent discriminative ability across all origins without significant class imbalance bias.

$$\text{Recall}_i(\text{TRR})_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i} \tag{18}$$

$$\text{Precision}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i} \tag{19}$$

$$\text{FPR}_i = \frac{\text{FP}_i}{\text{FP}_i + \text{TN}_i} \tag{20}$$

To evaluate whether the model has learned generalizable patterns rather than simply memorizing the training data, we supplemented the conventional train-test split with both permutation testing and quasi-independent validation. First, the origin labels of the training set were completely shuffled, and the full mRMR-PCA-LGBM pipeline was re-run; accuracies were computed on the fixed test set. This procedure was repeated 100 times, yielding a mean shuffled-label accuracy of

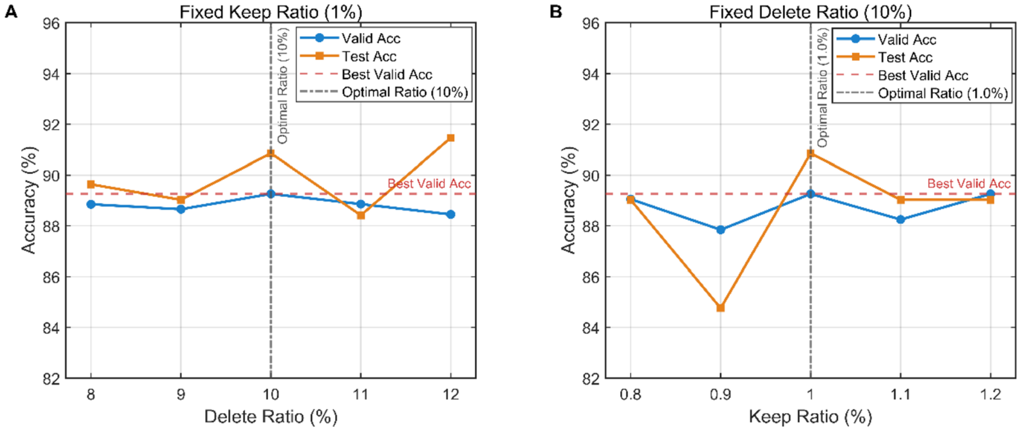

Fig. 10 (A) Sensitivity analysis of delete ratio (8–12%) with fixed keep ratio (1%); (B) sensitivity analysis of keep ratio (0.8–1.2%) with fixed delete ratio (10%).

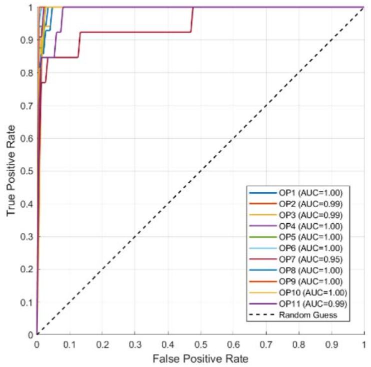

Fig. 11 Multi-class ROC curves (one-*vs.*-rest) of the mRMR-PCA-LGBM model on the independent test set, with AUC values for 11 origin classes.

0.105 ± 0.026 (range 0.024–0.159), with none reaching the original accuracy of 0.909 (Fig. 12). The exact binomial 95% confidence interval was 0.000–0.036 ($p < 0.01$), confirming that the observed performance is not due to random chance. Subsequently, we applied a temporal-split quasi-independent validation strategy: the first 494 spectra (by collection order) were used for training and the last 164 for testing, allowing natural instrumental drift, humidity variation and operator differences to arise. The resulting accuracy of 0.902 is close to that obtained under random splitting (0.909), indicating that the mRMR-PCA-LGBM model is robust to the comprehensive perturbations introduced through the temporal split within this dataset.

To address the interpretability issue of the LGBM model, a global feature importance analysis was first conducted based on SHAP values across the 72 modeling features (comprising 34 directly retained bands from mRMR and 38 principal components from PCA dimensionality reduction) (Fig. 13). The results reveal that the top 20 core contributing features are exclusively PCA principal components, with no directly retained original bands among them. Specifically, PC5, PC12, and PC9 exhibit the highest SHAP values (0.461, 0.357, and 0.349, respectively), identifying them as the key dependencies for the model's decision-making. In contrast, the top 3 original bands (ranked by SHAP importance) selected by mRMR (1743 $\mathrm{cm^{-1}}$, 1642 $\mathrm{cm^{-1}}$, 1659 $\mathrm{cm^{-1}}$) rank 35th, 38th, and 40th respectively among the 72 features, with SHAP values of 0.116, 0.093, and 0.090. While these original bands do not break into the top 20, they still demonstrate stable contribution. The core reason for this outcome lies in the fact that PCA principal components are linear combinations of bands in the middle segment (dimensionality reduction segment) screened by mRMR. They aggregate synergistic information from this region and, compared to individual isolated bands, more effectively reflect the overall structural differences of the core components in *Cornus officinalis*, thus naturally attaining higher contribution scores.

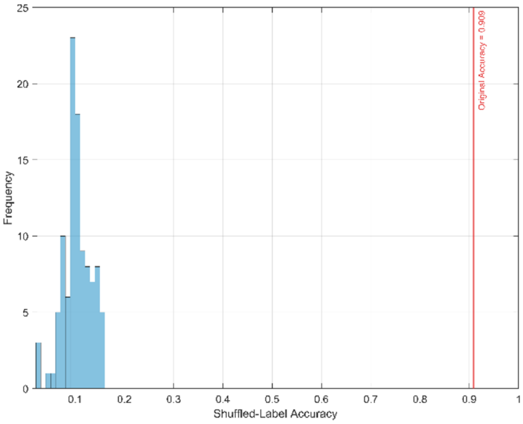

Fig. 12 Permutation test distribution (n = 100) of the mRMR-PCA-LGBM model on the independent test set, with the original accuracy indicated.

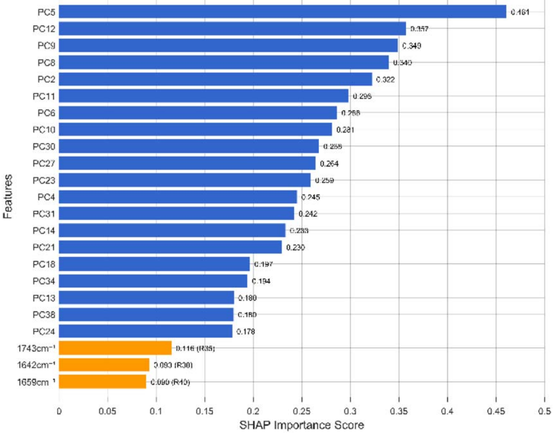

Fig. 13 Horizontal bar chart showing the top 20 features by SHAP value, with three additional included original bands, among the 72 modeling features. Orange bars denote original bands, and the values in parentheses indicate their ranks within the full 72-dimensional feature set.

To further elucidate the chemical significance of these core principal components, a loading analysis was performed on PC5, PC12, and PC9 (Fig. 14). The results indicate that the core loadings of PC5 are concentrated within the 1650–1690 $\mathrm{cm^{-1}}$ range, which corresponds to C═O stretching vibration. This signal originates from the ester groups of iridoid glycosides and the carboxyl groups of organic acids in *Cornus officinalis*, serving as a characteristic spectral representation of these core bioactive components. The core loadings of PC9 and PC12 overlap in the 980–1005 $\mathrm{cm^{-1}}$ region, corresponding to C–H out-of-plane bending vibration. This vibration is associated with the structural characteristics of liposoluble components in *Cornus officinalis*, such as triterpenoids with aliphatic side chains, and can reflect differences in the accumulation of such components across different samples. These findings confirm that the principal components relied upon by the model precisely aggregate the effective chemical features from the middle segment, providing a clear chemical foundation for its predictive performance.

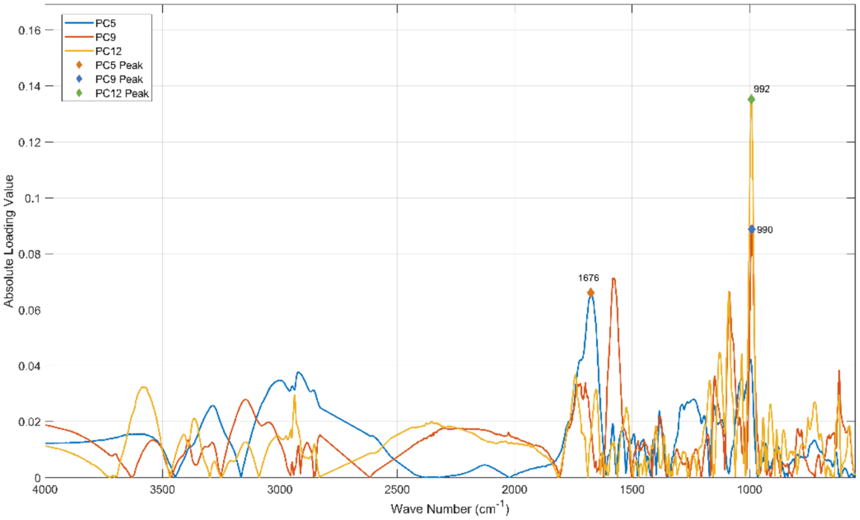

Fig. 14 Loading profiles of key principal components (PC5, PC12, PC9) obtained solely from the mRMR middle segment.

4. Discussion

To systematically evaluate modeling strategies for identifying the origin of *Cornus officinalis*, this study focuses on verifying the effectiveness of mRMR + PCA feature selection combined with traditional machine learning in small-sample spectral classification. We therefore conducted a comprehensive comparison of the mRMR-PCA-LGBM model against PCA-LGBM, full-spectrum LGBM, Partial Least Squares-Discriminant Analysis (PLS-DA), ANN, and RF. Table 2 summarizes the recall and precision of each model across the 11 origins.

The results show that the mRMR-PCA-LGBM model substantially outperforms others in these metrics: it achieves an average recall of 90.01% and an average precision of 92.21%, with 100% recall or precision in origins such as OP1, OP3, OP4, OP5, OP6, OP9, and OP10. Even for OP7, an origin prone to confusion with adjacent regions due to similar spectral features, the recall remains above 76.92%, which fully verifies the effectiveness of mRMR feature selection and PCA dimensionality reduction. In contrast, although PCA-LGBM also uses spectral information after dimensionality reduction, it lacks the front-end screening of mRMR, leading to insufficient weight of discriminative bands, with an overall performance decline of approximately 5–6 percentage points compared to the proposed mRMR-PCA-LGBM model. Due to interference from high-dimensional collinearity, the average performance of full-spectrum LGBM drops sharply by 35–40 percentage points compared to mRMR-PCA-LGBM. PLS-DA, which primarily relies on linear projection to extract discriminant features, struggles with the complex nonlinear relationships between spectral bands in OP7 samples, resulting in its precision and recall plummeting to below 30%. Although RF and ANN have nonlinear fitting capabilities, they fail to effectively distinguish subtle differences between origins, with the recall rates of OP5 and OP7 reduced to below 50% respectively.

This study employs four metrics: accuracy, Cohen's kappa, Matthews correlation coefficient (MCC), and F_1-score, to

Table 2 Recall (sensitivity/TPR) and precision (PPV) of different models for identifying 11 origins of *Cornus officinalis* (values in %)

	mRMR-PCA-LGBM		PCA-LGBM		LGBM		PLS-DA		ANN		RF	
OP	TPR	PPV	TPR	PPV	TPR	PPV	TPR	PPV	TPR	PPV	TPR	PPV
OP1	94.12	100.00	88.24	100.00	47.06	72.73	88.24	93.75	88.24	93.75	82.35	93.33
OP2	80.00	92.31	80.00	85.71	40.00	54.55	46.67	53.85	93.33	87.50	73.33	91.67
OP3	100.00	80.95	100.00	77.27	88.24	55.56	88.24	68.18	94.12	80.00	94.12	72.73
OP4	100.00	88.00	95.45	84.00	68.18	65.22	86.36	86.36	100.00	88.00	90.91	71.43
OP5	87.50	100.00	62.50	83.33	25.00	66.67	62.50	71.43	87.50	100.00	37.50	100.00
OP6	100.00	91.30	100.00	84.00	71.43	53.57	90.48	79.17	90.48	90.48	95.24	68.97
OP7	76.92	90.91	61.54	88.89	15.38	22.22	15.38	28.57	46.15	85.71	61.54	80.00
OP8	85.71	92.31	100.00	82.35	71.43	50.00	78.57	64.71	85.71	85.71	78.57	73.33
OP9	100.00	100.00	87.50	87.50	62.50	83.33	100.00	88.89	100.00	88.89	75.00	100.00
OP10	81.25	100.00	75.00	100.00	62.50	76.92	81.25	100.00	75.00	100.00	68.75	84.62
OP11	84.62	78.57	76.92	90.91	53.85	53.85	69.23	64.29	92.31	70.59	69.23	81.82

Table 3 Performance metrics (accuracy, Cohen's kappa, MCC, F_1-score) of different models for *Cornus officinalis* origin identification

Evaluation metrics	mRMR-PCA-LGBM	PCA-LGBM	LGBM	PLS-DA	ANN	RF
Accuracy	0.909	0.866	0.579	0.750	0.872	0.787
Kappa	0.898	0.851	0.531	0.722	0.858	0.762
MCC	0.899	0.852	0.537	0.725	0.859	0.765
F_1-score	0.907	0.850	0.552	0.724	0.864	0.771

comprehensively evaluate the performance of different models in the task of identifying the origin of *Cornus officinalis*. As the most fundamental performance measure, accuracy reflects the overall classification correctness rate, defined as the proportion of correctly predicted samples to the total number of samples. The kappa coefficient is a statistic that measures classification consistency after excluding random agreement, with its calculation formula shown in eqn (21), where p_0 is the observed accuracy and p_e is the accuracy of random predictions. The Matthews correlation coefficient (MCC), sensitive to imbalanced data, comprehensively considers TP, TN, FP, and FN, with its calculation formula given in eqn (22). F_1-score is a key indicator for evaluating model performance; it balances the impact of false positives and false negatives through the harmonic mean of precision and recall, and its calculation formula is presented in eqn (23). In multi-classification problems, the F_1-score for each class is typically calculated first, and then the average is taken to assess the overall performance of the model, including macro-average, weighted-average, and micro-average. Among these, the macro-average F_1-score is particularly suitable for scenarios where the sample sizes are relatively evenly distributed across classes, with its calculation formula shown in eqn (24).

As shown in Table 3, the mRMR-PCA-LGBM model outperforms others markedly across all four metrics: accuracy reaches 0.909, kappa stands at 0.898, MCC is 0.899, and F_1-score is 0.907. These values are higher than those of PCA-LGBM, RF, and ANN, and far exceed the full-spectrum LGBM and PLS-DA. These results verify that the mRMR-PCA-LGBM framework can significantly enhance the overall robustness and generalization ability in identifying the origin of *Cornus officinalis* while maintaining class balance.

In summary, the mRMR-PCA-LGBM model significantly enhances the robustness and accuracy of *Cornus officinalis* origin identification through the three-segment dynamic threshold joint optimization strategy, providing a promotable technical paradigm for the traceability of Chinese medicinal materials. This strategy forms a complete process with complementary advantages through band screening *via* mRMR, dimensionality reduction *via* PCA, and efficient modeling *via* LightGBM, offering a robust and promotable technical solution for the origin traceability of *Cornus officinalis* and other Chinese medicinal materials.

$$\text{Kappa} = \frac{p_0 - p_e}{1 - p_e} \tag{21}$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}} \tag{22}$$

$$F_{1i} = 2 \times \frac{\text{precision}_i \times \text{recall}_i}{\text{precision}_i + \text{recall}_i} \tag{23}$$

$$F_{1\,\text{macro}} = \frac{\sum_{i=1}^{n} F_{1i}}{n} \tag{24}$$

5. Conclusions

As a genuine medicinal material, the efficacy of *Cornus officinalis* is highly coupled with the ecological factors of its producing area. This study proposes a novel three-segment dynamic threshold joint optimization strategy based on mid-infrared spectroscopy to address the challenges of origin identification. The strategy involves identifying the 34 most discriminative bands from 3448-dimensional spectra *via* mRMR, eliminating 345 redundant and noisy bands, compressing the remaining information to 38 dimensions using PCA, and finally completing classification with LightGBM optimized by Bayesian optimization. The results from the independent external test set demonstrate that the mRMR-PCA-LGBM model achieves an accuracy of 90.9%, with kappa, MCC, and macro-F_1 values of 0.898, 0.899, and 0.907, respectively. The ROC-AUC values for all 11 producing areas exceed 0.95, significantly outperforming baseline models such as PCA-LGBM, full-spectrum LGBM, PLS-DA, ANN, and RF. This method achieves highly robust identification using less than 2% of the original variables, providing a rapid, low-cost, and easily deployable technical solution for the authenticity tracing of *Cornus officinalis*, and it is scalable to other Chinese medicinal materials. Notably, this strategy already realizes a clear "chemical problem-algorithm optimization" correlation: by retaining 16 key bands from the metabolite-rich 600–1500 cm^{-1} interval and eliminating redundant ones outside, it targets the chemical essence of origin discrimination, which explains why the model maintains high robustness with minimal variables. Future work will focus on expanding the sample size across different producing regions and seasons, integrating multi-source data-fusion strategies, as well as exploring the application of deep learning models for spectral classification with expanded datasets, conducting cross-laboratory and multi-year data collections, and carrying out cross-scenario adaptability verification of the proposed three-segment dynamic threshold strategy to further verify the model's generalisability and to enhance its universality and regulatory applicability under complex, real-world circulation scenarios. Additionally, we will explore optimizing the feature selection strategy by combining chemical domain knowledge with machine learning data screening, aiming to better leverage critical metabolite structural information and reduce redundant interference, thus further enhancing the stability and reliability of the model in origin discrimination.

Author contributions

Bing Liu: conceptualization, methodology, writing—original draft preparation and supervision. Hua Yi: funding acquisition. Chaoning Li: funding acquisition. Wangwang Yu: conceptualization and funding acquisition. Shuting Yang: funding acquisition. All authors have read and agreed to the published version of the manuscript.

Conflicts of interest

There are no conflicts to declare.

Data availability

The data supporting this article have been included as part of the supplementary information (SI). See DOI: **https://doi.org/10.1039/d5ra05862g**.

Acknowledgements

This work was supported by grants from the Jiangsu Province Industry-University-Research Cooperation Project (BY20240405), the Horizontal Scientific Research Projects of Nanjing University of Industry Technology (HK25-38-04, HK25-38-01, HK24-38-03, HK24-01-29), the Educational Research Projects of Nanjing University of Industry Technology (201016621GJ013), and the Research Project on Philosophy and Social Science of Universities in Jiangsu Province (2022SJYB0571).

References

1 R. Yao, M. Heinrich, X. Zhao, J. Wei and P. Xiao, *J. Ethnopharmacol.*, 2021, **276**, 1–8.

2 Z. Zhao, Z. Liang and G. Ping, *J. Ethnopharmacol.*, 2011, **134**, 556–564.

3 K. Ogawa-Ochiai and K. Kawasaki, *Front. Nutr.*, 2019, **5**, 140.

4 Y. Jiang, B. David and P. Tu, *Anal. Chim. Acta*, 2010, **657**, 9–18.

5 S. H. Baek, H. B. Lim and H. S. Chun, *J. Agric. Food Chem.*, 2014, **62**, 5403–5407.

6 M. K. Kim, J. H. Kim, H. Wang, H. N. Lee, C. G. Jin, W. S. Kwon and D. C. Yang, *J. Ginseng Res.*, 2016, **40**, 395–399.

7 C. Tistaert, B. Dejaegher and Y. Vander Heyden, *Anal. Chim. Acta*, 2011, **690**, 148–161.

8 X. Huang, Z. Liang, H. Chen, Z. Zhao and P. Li, *J. Microsc.*, 2014, **256**, 6–22.

9 K. Liu, J. W. Zhang, X. G. Liu, Q. W. Wu, X. S. Li, W. Gao, H. Y. Wang, P. Li and H. Yang, *Phytomedicine*, 2018, **51**, 104–111.

10 X. Q. Zha, J. P. Luo and P. Wei, *South Afr. J. Bot.*, 2009, **75**, 276–282.

11 X. H. Sun, C. L. Gao, W. D. Cao, X. R. Yang and E. K. Wang, *J. Chromatogr. A*, 2002, **962**, 117–125.

12 F. Tian, X. F. He, J. Sun, X. D. Liu, Y. Zhang, H. Cao and Z. G. Ma, *J. Sep. Sci.*, 2020, **43**, 736.

13 F. Q. Yang, Y. T. Wang and S. P. Li, *J. Chromatogr. A*, 2006, **1134**, 226–231.

14 K. A. Obisesan, A. M. Jimenez-Carvelo, L. Cuadros-Rodriguez, I. Ruisanchez and M. P. Callao, *Talanta*, 2017, **170**, 413–418.

15 Q. X. Zhu, Y. B. Cao, Y. Y. Cao and F. Lu, *Spectrosc. Spectr. Anal.*, 2014, **34**, 990–993.

16 F. Pozzi, N. Shibayama, M. Leona and J. R. Lombardi, *J. Raman Spectrosc.*, 2013, **44**, 102–107.

17 B. Schmidt, J. W. Jaroszewski, R. Bro, M. Witt and D. Stærk, *Anal. Chem.*, 2008, **80**, 1978–1987.

18 S. Berkov, R. Denev, B. Sidjimova, Y. Zarev, A. Shkondrov, L. Torras-Claveria, F. Viladomat and J. Bastida, *Rapid Commun. Mass Spectrom.*, 2023, **37**, e9506.

19 C. Q. Song, Y. L. An, W. J. Zhao, Y. S. Huang, L. J. Zhang, L. Li, Z. J. Tang, Z. W. Li, X. K. Liu, D. D. Zhang and D. A. Guo, *Microchem. J.*, 2025, **209**, 112671.

20 T. Nan, S. Wu, H. Zhao, W. Tan, Z. Li, Q. Zhang and B. Wang, *Anal. Chem.*, 2012, **84**, 4327–4333.

21 M. Sandasi, I. Vermaak, W. Chen and A. Viljoen, *Planta Med.*, 2016, **82**, 472–489.

22 C. Yu, C. Z. Wang, C. J. Zhou, B. Wang, L. Han, C. F. Zhang, X. H. Wu and C. S. Yuan, *J. Pharm. Biomed. Anal.*, 2014, **99**, 8–15.

23 A. Krähmer, A. Engel, D. Kadow, N. Ali, P. Umaharan, L. W. Kroh and H. Schulz, *Food Chem.*, 2015, **181**, 152–159.

24 X. Ren, T. He, J. Wang, L. Wang, Y. Wang, X. Liu and G. She, *J. Pharm. Biomed. Anal.*, 2021, **202**, 1–12.

25 Y. Jin, B. Liu, C. Li and S. Shi, *PLoS One*, 2022, **18**, e0282429.

26 M. Zhou, W. Du, K. Qin, J. Zhou and B. Cai, *Wirel. Pers. Commun.*, 2018, **102**, 1827–1838.

27 J. Liang, M. Li, Y. Du, C. Yan, Y. Zhang, T. Zhang and H. Li, *Chemometr. Intell. Lab. Syst.*, 2020, **207**, 104179.

28 B. Liu, J. Wang and C. Li, *RSC Adv.*, 2024, **14**, 15209–15219.

29 J. Hermo, V. Bolón-Canedo and S. Ladra, *Inf. Sci.*, 2024, **669**, 120609.

30 M. Mahapatra, S. K. Majhi and S. K. Dhal, *Evol. Intell.*, 2022, **15**, 2017–2036.

31 D. Jiang, X. Shi, Y. Liang and H. Liu, *Measurement*, 2024, **237**, 115190.

32 Z. Cui, X. Qing, H. Chai, S. Yang, Y. Zhu and F. Wang, *J. Hydrol.*, 2021, **603**, 127124.

33 Y. Sun, N. Liu, X. Kang, Y. Zhao, R. Cao, J. Ning and D. Zhou, *Food Control*, 2021, **124**, 107883.

34 S. Deng, J. Su, Y. Zhu, Y. Yu and C. Xiao, *Expert Syst. Appl.*, 2024, **242**, 122502.