Review of Scientific Instruments ARTICLE pubs.aip.org/aip/rsi

Qualitative analysis of terpenoid esters based on near-infrared spectroscopy and machine learning

Cite as: Rev. Sci. Instrum. 96, 023701 (2025); doi: 10.1063/5.0243298
Submitted: 10 October 2024 • Accepted: 3 January 2025 •
Published Online: 3 February 2025

Haiyi Bian,[1,a)] Ling Huang,[1] Qinxin Xu,[1] Rendong Ji,[1] and Jun Wang[2]

AFFILIATIONS

[1] Jiangsu Engineering Research Center of Lake Environment Remote Sensing Technologies, Huaiyin Institute of Technology, Huaian 223003, China
[2] Jiangsu Guoyao Cloud Technology Co., Ltd., Suzhou 215008, China

[a)] **Author to whom correspondence should be addressed:** bianhaiyi@163.com

ABSTRACT

This study delves into the method of qualitative analysis of terpenoid esters using near-infrared spectroscopy technology. Terpenoid esters are bioactive compounds widely used in the pharmaceutical and cosmetics industries. Near-infrared spectroscopy technology enables rapid and accurate component analysis without compromising the integrity of the sample, which is particularly important for valuable samples that need to be preserved intact or require subsequent analysis. This research combines machine learning techniques, such as K-Nearest Neighbors (K-NN) classifier, Random Forests algorithm, and Back Propagation Neural Networks (BPNN), to analyze terpenoid ester samples extracted from different concentrations of eluents, and compares and evaluates these algorithms. This study results show that in the test set, the prediction accuracy of the K-NN classifier is 96.154% and BPNN is 94.231%, and the Random Forest algorithm performs the best with a prediction accuracy of 100%. Additionally, this study utilizes the Random Forest algorithm to predict the characteristic spectra of terpenoid esters, demonstrating the effectiveness of feature spectrum extraction by ensuring a prediction accuracy of 100% while reducing the number of spectral features.

I. INTRODUCTION

Terpenoid esters are a class of compounds with significant biological activity and wide-ranging applications, possessing diverse functionalities. Widely utilized in the pharmaceutical and cosmetics industries, terpenoid esters exhibit antibacterial, anti-inflammatory, and antioxidant properties, finding applications in the production of medications, perfumes, and skincare products, and, notably, holding a prominent position in traditional Chinese medicine research. In industrial production, column chromatography techniques are employed to enrich the active ingredients from Ginkgo biloba leaf extracts while simultaneously removing impurities and allergenic components.[1]

Near-infrared spectroscopy technology holds immense potential across various sectors, such as chemical engineering,[2] pharmaceuticals,[3] food,[4] agriculture,[5] and particularly in pharmaceuticals, where it is utilized for analyzing drug components and determining their concentrations during manufacturing processes, as well as in medical diagnostics and biomedical research. Near-infrared spectroscopy offers several advantages: (1) its non-destructive nature preserves the integrity of samples, which is crucial for valuable or samples requiring subsequent analysis; (2) it is fast and efficient, enabling the acquisition of large amounts of data in a short time for real-time monitoring and control; (3) it has the capability of multi-component analysis, simultaneously detecting various compounds, thus reducing analysis time and costs; and (4) data processing and interpretation are relatively simple, allowing for rapid and accurate analysis results through model building or comparison with standard samples. Leveraging these advantages, many researchers in the field have utilized near-infrared spectroscopy to achieve qualitative identification of traditional Chinese medicines. Congshi *et al.*,[6] for instance, established a near-infrared pattern

recognition model for six types of resins and other traditional Chinese medicines, achieving rapid identification of the six medicines. Gong *et al.*[7] successfully differentiated Angelica sinensis medicinal herbs from their counterfeits using near-infrared spectroscopy technology. Guo *et al.*[8] conducted a qualitative analysis on the origins of Panax notoginseng slices using near-infrared spectroscopy, providing reference for the quality control and rational utilization of Panax notoginseng powder slices. Wang *et al.*[9] studied the model identification of the four tastes (bitter, sweet, sour, and salty) of traditional Chinese medicines using near-infrared spectroscopy technology combined with principal component analysis-discriminant analysis, partial least squares-discriminant analysis, and K-nearest neighbor algorithm. Chen *et al.*[10] combined near-infrared spectroscopy technology with convolutional neural networks for discriminant analysis of three different degrees of carbonization of *Puerariae carbonisata*.

Extensive research has been conducted on terpenoid esters in academia. For instance, He *et al.*[11] combined near-infrared spectroscopy technology with partial least squares regression to achieve rapid detection of terpenoid ester content in Ginkgo biloba leaves. Liuwei *et al.*[12] used near-infrared spectroscopy and chemometrics to develop a consistency verification model for terpene lactones in Ginkgo leaves by employing partial least squares regression. In addition, Ni *et al.*[13] used near-infrared spectroscopy combined with a genetic algorithm to establish a quantitative model for active components in Ginkgo leaves. In studying terpenoid esters in Ginkgo biloba extract products, real-time monitoring for both quantitative analysis and qualitative analysis are necessary. In this study, considering that liquid chromatography is commonly used in industrial production to process samples, requiring the addition of eluents with different concentrations, this may result in differences in drug concentration and medicinal value in the final product. Therefore, qualitative analysis of terpenoid esters is particularly important. Especially during the bottling process, mechanical malfunctions and other issues may lead to bottling errors. Conducting qualitative analysis helps to avoid these problems, facilitating the recovery and treatment of terpenoid esters, thereby reducing waste.

In this context, this study selects several common machine learning algorithms, such as K-NN, RF, and BPNN, as tools for the qualitative analysis of terpenoid esters. The rationale for choosing these algorithms is based on their advantages in handling multidimensional data, complex feature extraction, and pattern recognition. For example, the K-NN algorithm can perform simple and effective classification by calculating the distance between samples, making it suitable for preliminary analysis of multidimensional data.[14,15] The RF algorithm excels in dealing with high-dimensional data, feature selection, and preventing overfitting, which is especially beneficial for near-infrared spectroscopy data that contains high collinearity and redundant features. For instance, Jiao *et al.*[16] used near-infrared spectroscopy and the Random Forest algorithm to establish a rapid identification model for four types of honey (linden honey, acacia honey, jujube honey, and rapeseed honey), achieving a classification accuracy of 97.58%, thus improving the accuracy and efficiency of honey variety identification. BPNN, as a classical neural network method, possesses strong adaptability and generalization ability in classifying complex nonlinear data.[17] For example, Sun *et al.*[18] used BPNN to establish a near-infrared spectral identification model for different varieties of cigar leaves after aging, achieving a high accuracy rate. Through comparative analysis of these machine learning algorithms, this study not only enables the qualitative analysis of terpenoid esters but also further explores effective methods for spectral feature extraction.

Therefore, the selection of specific machine learning algorithms is aimed at fully utilizing their individual advantages to overcome the challenges posed by NIR spectroscopy data in the qualitative analysis of terpenoid esters, ultimately improving the accuracy and efficiency of the analysis.

II. MATERIALS AND METHODS

A. Sample preparation and spectral acquisition

In order to obtain statistically significant results, 50–100 samples from different regions were selected. The standardized sample preparation and spectral acquisition scheme referenced the quality control methods maintained by Shanghai Shangyaoxingling Pharmaceutical Technology Co., Ltd., during the production of traditional Chinese medicine. During the experimental process, strict adherence to established sample collection and preparation protocols was followed. Samples were ground into fine powder and subjected to column chromatography separation techniques. In the first stage of column chromatography, varying concentrations of active ingredient eluents (18%, 30%, 50%, and 95%) were added to the sample powder. The samples were separated by adsorbents packed in the column and eluted with flowing solvent or buffer solutions. Different components moved at different rates within the column based on their interactions with the stationary phase and mobile phase, thus achieving separation.

During the spectral data acquisition process, prepared sample solutions using high-performance liquid chromatography were placed in a near-infrared spectrometer for spectral acquisition. The spectrometer was operated in reflectance mode with a spectral resolution set to 4 cm^{-1} and a wavelength range set from 900 to 1700 nm.

B. Instruments and equipment

The near-infrared spectrometer (model Patux) used in the experiment was manufactured by the American company VIAVI Solutions Inc., equipped with a deuterium–halogen lamp as the light source and an InGaAs detector, operating in the wavelength range of 900–1700 nm. The high-performance liquid chromatography (HPLC) system utilized the latest Agilent 1290 system produced by Agilent Technologies Inc. from the United States, equipped with a diode array detector (DAD). The modeling process was conducted using Matlab R2021b.

C. Data preprocessing

Data preprocessing is a necessary step to enhance the quality of signals and improve the predictive performance of models as raw near-infrared spectroscopy data are often affected by various issues, such as low signal-to-noise ratio, baseline drift, spectral overlap, and instrument drift. In this study, the multiple scatter correction (MSC) technique was employed to mitigate the influence of baseline drift and high-frequency noise.[19] MSC is a widely used spectral preprocessing method that mathematically transforms the spectra

of each sample to effectively eliminate scattering effects, thus stabilizing the spectral data. This preprocessing technique significantly reduces interference from baseline drift and high-frequency noise, providing a more reliable foundation for subsequent data analysis and modeling.

MSC aims to correct spectral deviations caused by light scattering effects. These scattering effects are common across different samples or measurement conditions and can result in shifts and distortions in the spectral signal, affecting data accuracy and model predictive capability. The goal of MSC is to eliminate variations caused by scattering, enabling spectra from different samples to have a consistent baseline after processing, thereby enhancing the reliability of data analysis and modeling.

The core idea of MSC is to standardize each spectrum to align with a reference spectrum, which is typically the mean spectrum of all spectra. Suppose there are n spectra, each containing m wavelength points, the mean spectrum is calculated using the following formula:

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}. \tag{1}$$

In the equation, $\bar{x}_j$ is the mean spectral value at the jth wavelength point and x_{ij} is the value of the ith spectrum at the jth wavelength point.

The next step requires calculating the mean $\bar{x}_r$ and standard deviation σ_r of the reference spectrum. Each spectrum to be processed, x_i, is then standardized to eliminate scattering effects, with the detailed calculation shown in the following equation:

$$\bar{x}_r = \frac{1}{m}\sum_{j=1}^{m} \bar{x}_j \quad \sigma_r = \sqrt{\frac{1}{m}\sum_{j=1}^{m} (\bar{x}_j - \bar{x}_r)^2} \Downarrow MSC_{(x_{ij})} = \frac{x_{ij} - \bar{x}_j}{\sigma_j} \cdot \frac{\sigma_r}{\bar{x}_r}. \tag{2}$$

In the equation, σ_j is the standard deviation at the jth wavelength point and $\bar{x}_j$ is the mean at the jth wavelength point.

D. Research methodology

In this study, the Random Forest algorithm serves as the primary machine learning tool, which combines the Bagging ensemble theory and the random subspace method. It constructs a classification model by integrating multiple decision trees, with each tree being independently built. When classifying new samples, each tree makes its own judgment, and the final classification result is determined by a voting process. In building each decision tree, a method of random sampling of rows and columns is employed to reduce the risk of overfitting, without conducting pruning operations. Each decision tree is grown using a complete split method to ensure that the leaf nodes are either pure in classification or unable to be further split. The metric for splitting attributes in the decision tree is measured using Gini Impurity, which quantifies the uncertainty of samples in the dataset during classification. A smaller Gini Impurity indicates higher purity of the node, meaning that the samples within the node are more inclined to belong to the same category. The calculation formula is shown in Eq. (3). For near-infrared spectroscopy data containing numerous variables, the Random Forest algorithm effectively addresses issues of collinearity and redundancy, making it an ideal classification algorithm,

$$G(D) = 1 - \sum_{k=1}^{K} \left(\frac{|C_k|}{|D|}\right)^2. \tag{3}$$

In Eq. (1), K represents the number of categories, $|C_k|$ denotes the number of samples belonging to category k, and $|D|$ represents the total number of samples in dataset D.

BPNN utilizes an error backpropagation technique for training, demonstrating flexibility and powerful generalization ability in classification problems, suitable for handling various types of data. BPNN consists of an input layer, a hidden layer, and an output layer, with the hidden layer connecting the input and output layers, which is relatively important. The calculation process of the hidden layer is shown as follows:

$$y_i = f\left(\sum_{j=1}^{n} v_{ij}x_i - a_j\right) = f\left(\sum_{j=0}^{n} v_{ij}x_i\right), i = 1, 2, \ldots, m, \tag{4}$$

$$f(x) = \frac{1}{1 + e^{-x}}. \tag{5}$$

In the equation, $f(\cdot)$ represents the transfer function of the hidden layer, as shown in Eq. (5); m represents the number of nodes in the hidden layer; v_{ij} represents the connection weight; a represents the threshold of the hidden layer; and x_i represents the input.

The K-NN classifier is based on the principle of proximity for classification, which is an intuitive and straightforward method, particularly suitable for handling multidimensional data. The most important aspect of the K-NN classifier is to measure similarity by calculating the distance between samples. In this study, the distance metric used is the Euclidean distance, as shown in the following equation:

$$r = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}. \tag{6}$$

In the equation, p_i and q_i represent the values of two data points on the i feature, respectively.

This study employed the Random Forest algorithm, BPNN algorithm, and K-NN classifier as qualitative models and utilized the Least Absolute Shrinkage and Selection Operator (LASSO) method for key variable selection to optimize model performance.

E. Evaluation metrics

In this study, model performance was evaluated using training accuracy and prediction accuracy. The calculation method for accuracy (A) is shown in the following equation:

$$A = \frac{N_c}{N_t} \times 100\%. \tag{7}$$

In the equation, N_c represents the number of correctly identified samples and N_t represents the total number of samples.

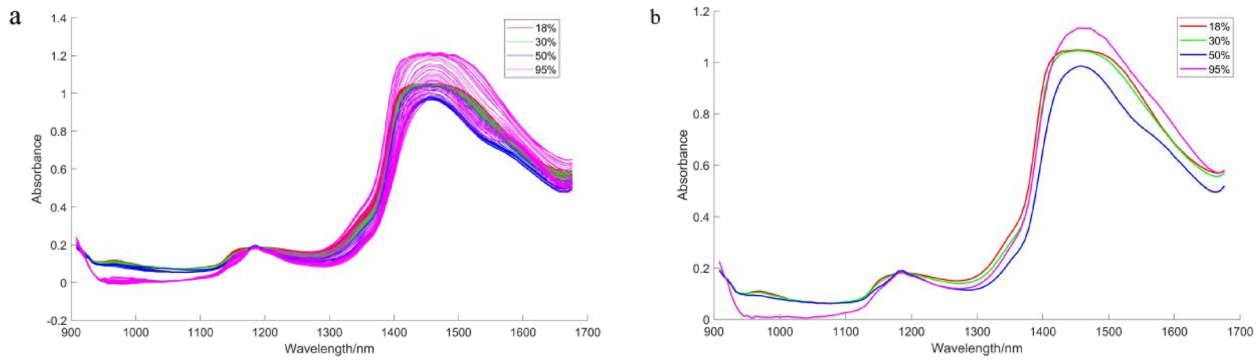

FIG. 1. Near-infrared spectral plot: (a) raw spectral plot and (b) average spectral plot.

III. RESULTS AND ANALYSIS

A. Terpenoid esters near-infrared spectral curve

As shown in Fig. 1, the near-infrared spectroscopy (NIRS) data reveal that the trends and characteristic absorptions of terpenoid esters NIRS curves obtained from four different concentrations of eluents are similar, making them difficult to distinguish directly. From the average spectral graph [Fig. 1(b)], variations in absorbance can be observed, with the order being 95% > 18% > 30% > 50%, providing some basis for classification. Due to the limited wavelength range measured by the instrument, mainly aromatic hydrocarbons, methyl, methylene, ethyl, water, and amine combination and overtone absorption peaks were obtained. The low signal intensity and broad peak spectra resulted in overlapping sample information and spectral peaks, making direct discrimination difficult, especially for 18% concentration and 30% concentration where the spectral overlap is more severe. Therefore, further analysis using chemometric algorithms is necessary.

B. Dataset partitioning

According to the Kennard–Stone (K–S) criterion, the data are partitioned into training and testing sets in a 7.5:2.5 ratio, ensuring that the training set contains 75% of the total data, while the testing set contains the remaining 25%. The number of samples in the training set is 159, and the number of samples in the test set is 52. This ensures that the model has sufficient data during training to learn patterns and features while also having enough data during testing to evaluate the model's performance and generalization ability. This helps reduce the risk of overfitting, thereby enhancing the reliability and applicability of the model.

C. Classification based on full spectral data

Using the full spectral data consisting of 125 features as input and concentration levels as classification labels, the K-NN classifier, Random Forest algorithm, and BP neural network are employed to train and predict the training and testing dataset. The confusion matrix is used to reflect the actual prediction accuracy of each component in the training and testing sets as evaluation metrics.

For the K-NN classifier, the value of K is determined using cross-validation. After multiple experiments, the highest cross-validation accuracy is achieved when K = 1. The Random Forest algorithm determines the optimal number of decision trees based on the out-of-bag (OOB) error. As shown in Fig. 2, the number

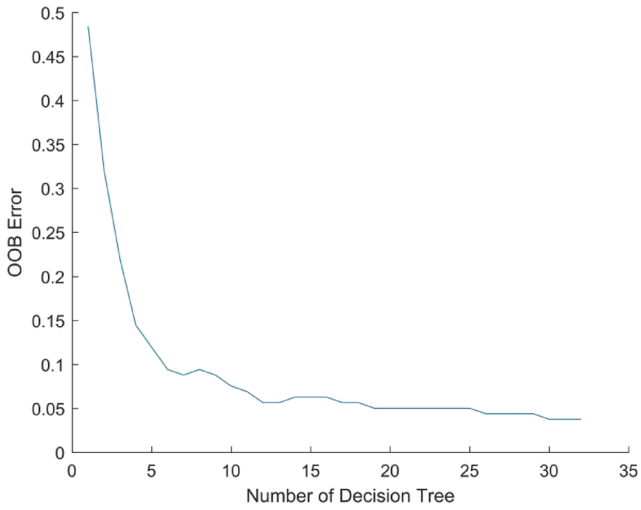

FIG. 2. Random forest out-of-bag (OOB) error.

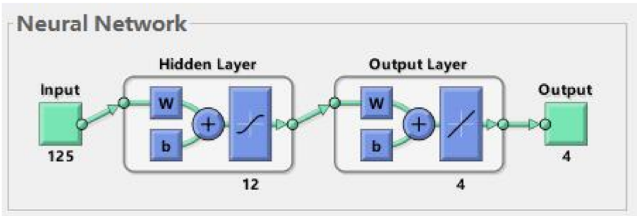

FIG. 3. BPNN structure diagram.

TABLE I. Accuracy of the K-NN classification model.

	Discriminant grouping									
	Training set					Testing set				
Actual grouping (%)	18 (%)	30 (%)	50 (%)	95 (%)	Accuracy (%)	18 (%)	30 (%)	50 (%)	95 (%)	Accuracy (%)
18	46				100	23	2			92
30		32			100		8			100
50			32		100			10		100
95				49	100				9	100
Overall accuracy	100							96.154		

TABLE II. Accuracy of the random forest classification model.

	Discriminant grouping									
	Training set					Testing set				
Actual grouping (%)	18 (%)	30 (%)	50 (%)	95 (%)	Accuracy (%)	18 (%)	30 (%)	50 (%)	95 (%)	Accuracy (%)
18	46				100	25				100
30		32			100		8			100
50			32		100			10		100
95				49	100				9	100
Overall accuracy	100							100		

TABLE III. Accuracy of the BPNN classification model.

	Discriminant grouping									
	Training set					Testing set				
Actual grouping (%)	18 (%)	30 (%)	50 (%)	95 (%)	Accuracy (%)	18 (%)	30 (%)	50 (%)	95 (%)	Accuracy (%)
18	46				100	23	2			92
30	1	30	1		93.75	1	7			87.5
50		1	31		96.875			10		100
95				49	100				9	100
Overall accuracy	98.113							94.231		

of decision trees is set to 32, and the minimum leaf size is set to 1. The BP neural network adopts an experimental method called trial and error to determine the most suitable number of hidden layer nodes. This process starts with setting a small number of nodes and then gradually increasing the number of nodes based on feedback from network training. The possible range of hidden layer node numbers is defined based on an empirical formula (8) as a reference.[20] In formula (6), m represents the number of hidden layer nodes, n represents the number of input layer nodes, c represents the number of output layer nodes, and a is a constant in the [0,8] interval,

$$m = \sqrt{n + c} + a. \tag{8}$$

The full spectral input variables consist of 125 features (N = 125), while the output variables comprise four categories (C = 4). According to Eq. (8), the reasonable range for the number of hidden layer nodes is determined to be from 12 to 20. In the

experiment, the maximum number of iterations is set to 1000, the training accuracy reaches 1×10^{-7}, and the learning rate is set to 0.01. Through a series of experiments, the optimal number of hidden layer nodes is determined to be 12, and the detailed network structure is shown in Fig. 3. This approach ensures high precision and learning efficiency in the experiment while validating the feasibility of the network structure.

Tables I–III present the prediction accuracy results of the three classification algorithms. On the training set, both the K-NN classifier and Random Forest model achieved an average training accuracy of 100%. In comparison, the training performance of the BP neural network model was slightly lower, with an accuracy of 98.113%. On the testing set, the Random Forest algorithm exhibited the highest classification prediction accuracy, reaching 100%. This was followed by the K-NN classifier with an accuracy of 96.154%, while the BP neural network had the lowest accuracy at 94.231%. Figure 4 illustrates the corresponding confusion matrices. Overall, the prediction accuracy of these three algorithm models is satisfactory, particularly in the classification prediction of terpenoid esters, meeting the expected targets, with the Random Forest algorithm demonstrating outstanding performance.

To prevent model overfitting, this study further reports an average cross-validation loss value of 0.0375 for the model. The low average loss in cross-validation indicates that the model's generalization ability has been enhanced. The ten-fold cross-validation process is shown in Fig. 5.

To verify the effectiveness of MSC preprocessing, this section compares the prediction results without MSC preprocessing using the best-performing random forest model as an example. The results show that without MSC preprocessing, the test set accuracy is

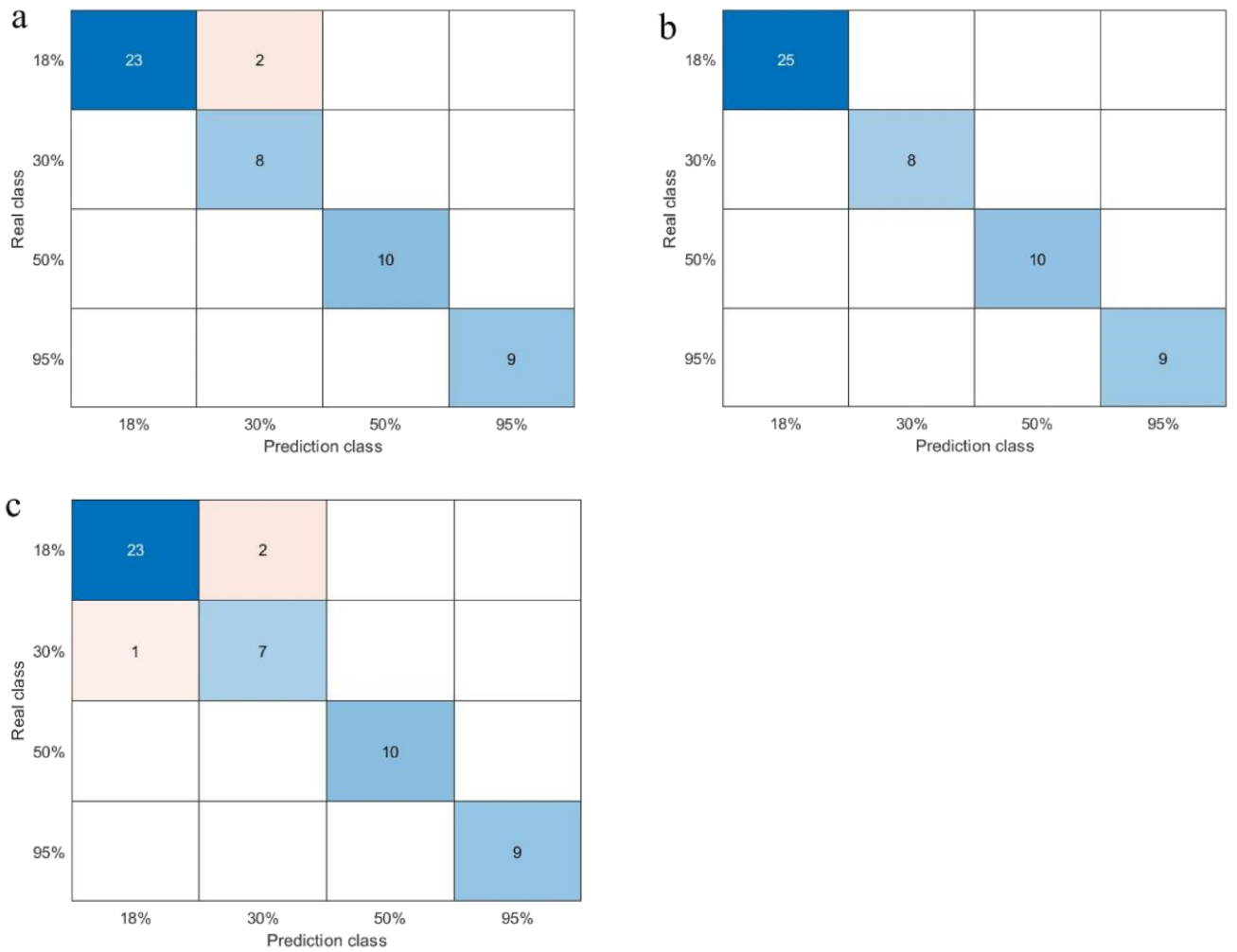

FIG. 4. Confusion matrices of three prediction models: (a) confusion matrix of K-NN classifier, (b) confusion matrix of Random Forest, and (c) confusion matrix of BPNN.

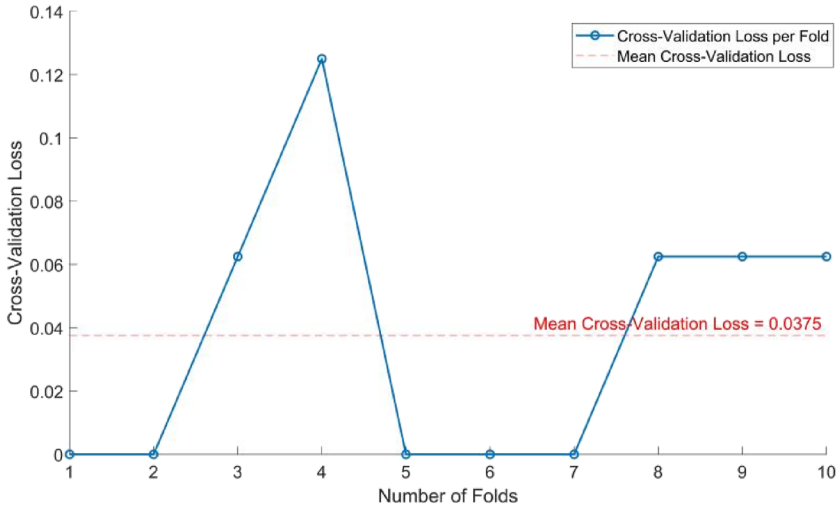

FIG. 5. Average cross-validation loss.

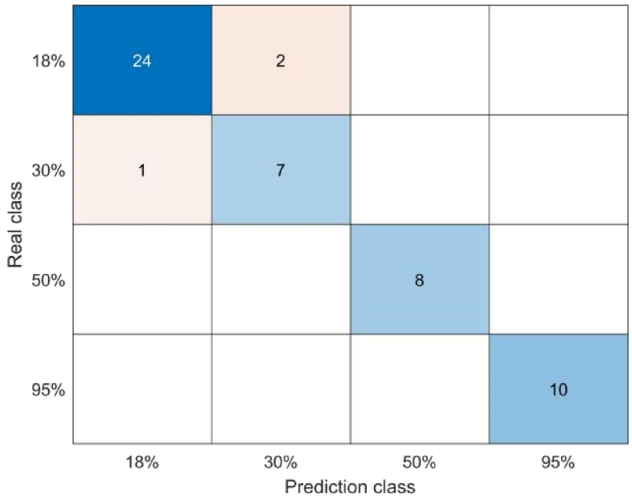

FIG. 6. Confusion matrix for predictions without MSC preprocessing.

0.942 31, with three misclassified predictions. The specific prediction results are shown in Fig. 6.

D. Feature selection

The LASSO (Least Absolute Shrinkage and Selection Operator) method is a statistical model used for regression analysis, proposed by Robert Tibshirani in 1996. It serves as both an estimation technique and a variable selection technique, particularly suitable for handling cases with multicollinearity or when the data dimensionality far exceeds the sample size. The essence of LASSO lies in its introduction of L1 regularization into the regression model. The core formula is shown in the following equation:

$$\min_{\beta}\left\{\frac{1}{2n}\sum_{i=1}^{n}\left(y_i-\beta_o-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|\right\}. \tag{9}$$

In Eq. (7), n represents the number of samples, p represents the number of features, y_i represents the observations, x_{ij} represents the value of the jth feature of the ith observation, β_0 represents the intercept, β_j represents the coefficient of the jth feature, and λ represents the regularization parameter.

The LASSO algorithm was chosen for feature selection, and the model performance was evaluated through ten-fold cross-validation to assess the performance of different values of A. Ultimately, the A value that minimizes the mean squared error (MSE) of the model was selected as the optimal regularization parameter. By analyzing the coefficient values of various feature wavelengths in the LASSO model, 32 wavelengths significantly influencing the model were selected. The distribution of these wavelengths is depicted in Fig. 7. This feature selection process revealed the crucial role of skewness and kurtosis in enhancing model accuracy. Furthermore, the research results indicated differences in spectral skewness and kurtosis of terpenoids in eluents of different concentrations, highlighting these differences as important features for classification. The larger the absolute value of the coefficient of a feature wavelength, the greater its contribution to the model. The specific feature wavelengths and their coefficient values are detailed in Table IV.

E. Classification based on feature spectra

To construct a discriminant model for terpenoid classification based on near-infrared feature spectra, the obtained 32 feature wavelengths were used as inputs, and the Random Forest algorithm was employed for modeling. In the modeling process, the output feature importance bar chart and heatmap are shown in Fig. 8. High-importance features are also present in the feature wavelengths, showing strong correlation with prediction accuracy, and can be considered key factors for optimizing model performance. The discriminant effect of the model is shown in Table V, while detailed test set classification predictions are displayed in Fig. 9. Compared to using original spectral data, the feature spectrum discriminant model significantly reduces the use of spectral variables, simplifying the modeling process. Despite the reduction in input variables,

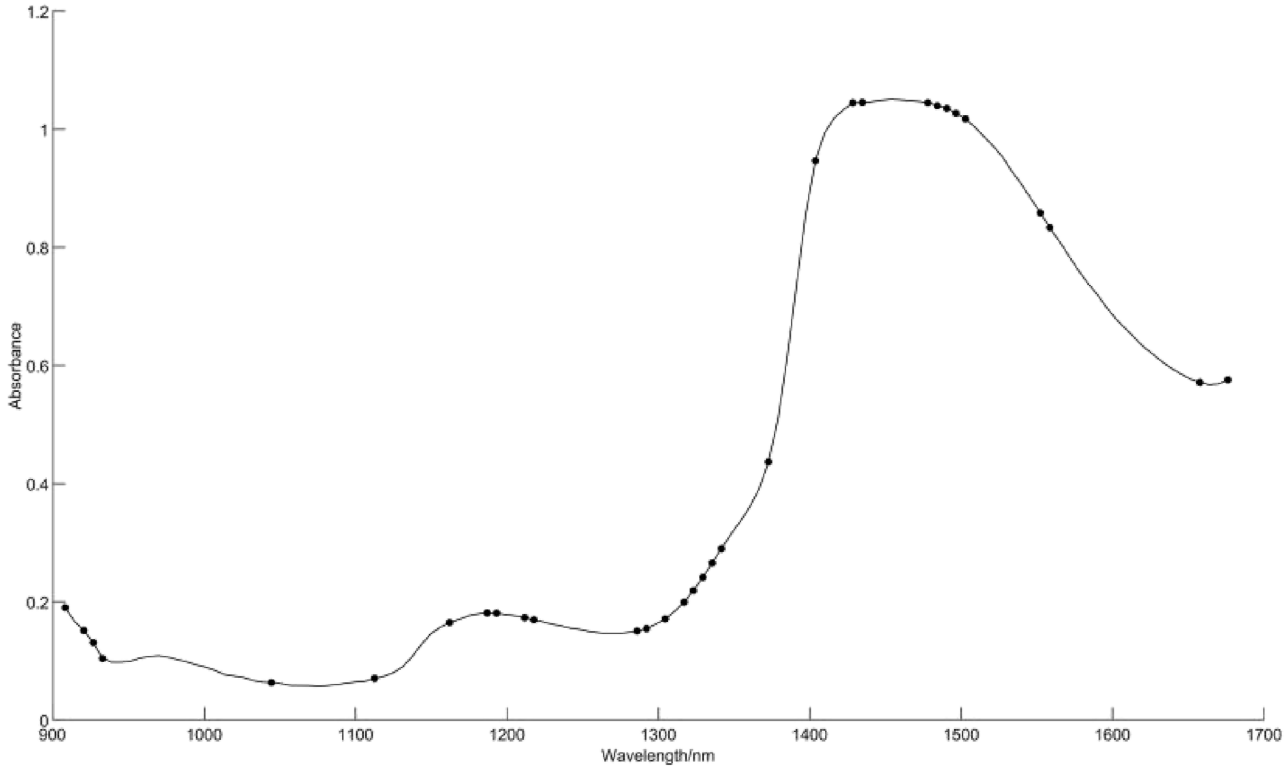

FIG. 7. Distribution of feature points.

TABLE IV. Feature wavelengths and their coefficient values.

Serial number	Wavelength (nm)	Coefficient	Serial number	Wavelength (nm)	Coefficient
1	908.1	−0.6737	17	1329.316	0.3584
2	920.489	0.5148	18	1335.51	0.4294
3	926.683	0.5196	19	1341.705	0.4368
4	932.877	1.5819	20	1372.677	0.3166
5	1044.376	0.4798	21	1403.648	−1.2018
6	1112.514	−3.5377	22	1428.426	0.248
7	1162.069	0.212	23	1434.62	1.5467
8	1162.069	1.9179	24	1477.981	−0.6885
9	1193.04	−0.6584	25	1484.175	−1.0463
10	1211.623	−2.3445	26	1490.369	−0.6386
11	1217.818	−0.639	27	1496.564	−1.238
12	1285.956	−2.3445	28	1502.758	−0.2757
13	1292.15	−0.639	29	1552.313	1.2327
14	1304.539	0.4517	30	1558.507	2.698
15	1316.927	0.7282	31	1657.617	−3.4577
16	1323.122	0.7042	32	1676.2	2.2511

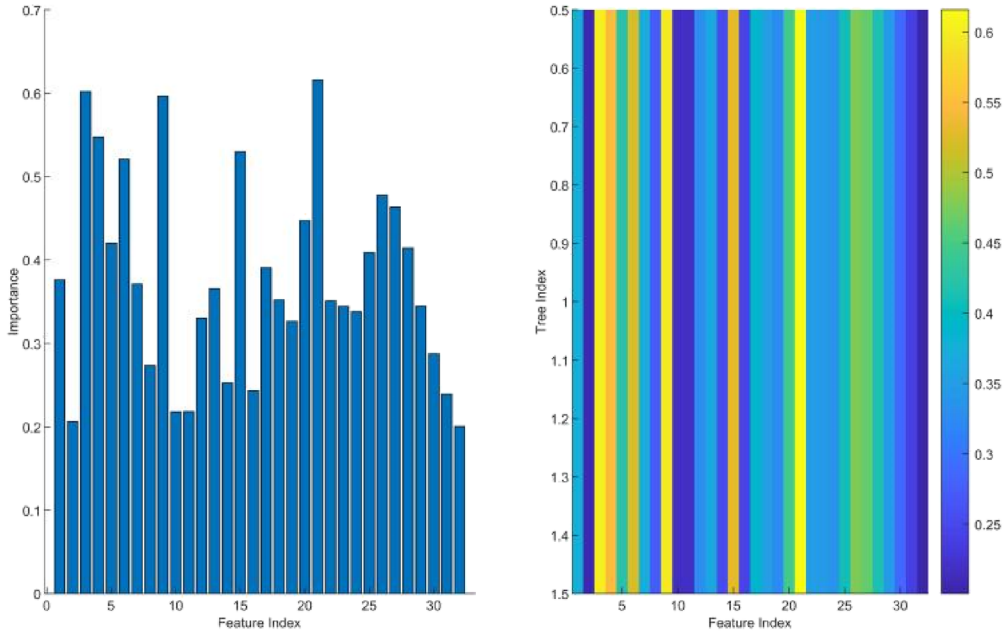

FIG. 8. Feature importance and heatmap.

TABLE V. Accuracy of the random forest classification model.

	Discriminant grouping									
	Training set					Testing set				
Actual grouping (%)	18 (%)	30 (%)	50 (%)	95 (%)	Accuracy (%)	18 (%)	30 (%)	50 (%)	95 (%)	Accuracy (%)
18	46				100	25				100
30		32			100		8			100
50			32		100			10		100
95				49	100				9	100
Overall accuracy	100							100		

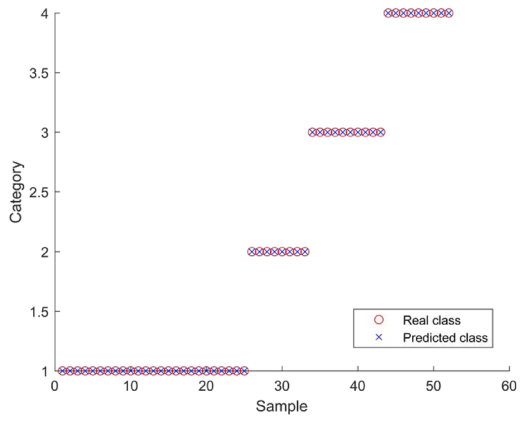

FIG. 9. Test set classification prediction results.

the model maintains a discriminant accuracy of 100% on both the training and test sets, demonstrating excellent stability. Thus, this feature spectrum discriminant model effectively achieves accurate classification and identification of terpenoids.

IV. CONCLUSION

This study investigates the combined use of near-infrared spectroscopy (NIRS) technology and machine learning algorithms for the qualitative analysis of terpenoids. Spectral data acquisition and analysis were conducted on terpenoid samples treated with different concentrations of eluent, and experiments were performed to identify and classify them using Random Forest, BP neural network, and K-NN algorithms. The experimental results show that the Random Forest algorithm performs best in all tests, with an accuracy of up to 100%. This is attributed to the effective management of

feature interactions by the Random Forest algorithm when dealing with large-scale datasets while reducing the risk of overfitting by constructing multiple decision trees.

In this study, the LASSO method was used for feature selection, effectively reducing the number of spectral features used for modeling. Through the regularization process of LASSO, 32 key wavelengths were selected from the original spectrum. This not only significantly reduced the model complexity but also maintained a classification performance comparable to that of the full-spectrum model. The advantage of this feature selection method lies in its ability to improve model computational efficiency while reducing the interference of redundant information in model predictions.

In practical applications, especially in industrial or pharmaceutical quality control scenarios, reducing the number of features offers multiple benefits. First, it lowers computational and storage costs, which is particularly important in production lines requiring real-time detection and control. In the pharmaceutical industry, rapidly and accurately monitoring the concentration of drug components is a key task for quality control. By using the few key wavelengths selected by LASSO, detection time can be significantly reduced while maintaining accuracy. Furthermore, reducing the number of features decreases the requirements for data acquisition and processing, making it suitable for portable or embedded near-infrared detection devices, thereby enhancing the scalability and flexibility of the technology.

This study demonstrates that the combination of near-infrared spectroscopy technology and advanced machine learning methods can effectively conduct rapid and accurate qualitative analysis of terpenoids. The successful application of this approach demonstrates its potential value in drug analysis and quality control, among other fields. Future research could further explore the application of these techniques in the analysis of other bioactive compounds and optimize them for more diverse sample processing and data analysis challenges.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (Grant Nos. 62141502, 62205120), the Project of State Key Laboratory of Radiation Medicine and Protection, Soochow University (Grant No. GZK1202217), and the Postgraduate Research & Practice Innovation Program of Jiangsu Province (Grant No. SJCX24_2149).

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Haiyi Bian: Conceptualization (equal); Data curation (equal); Formal analysis (equal); Funding acquisition (equal); Investigation (equal); Methodology (equal); Project administration (equal); Resources (equal); Software (equal); Supervision (equal); Validation (equal); Visualization (equal); Writing – original draft (equal); Writing – review & editing (equal). **Ling Huang**: Conceptualization (equal); Data curation (equal); Formal analysis (equal); Funding acquisition (equal); Investigation (equal); Methodology (equal); Project administration (equal); Resources (equal); Software (equal); Supervision (equal). **Qinxin Xu**: Conceptualization (equal); Data curation (equal); Formal analysis (equal); Funding acquisition (equal); Investigation (equal); Methodology (equal); Project administration (equal); Resources (equal); Software (equal); Supervision (equal); Validation (equal); Visualization (equal); Writing – original draft (equal); Writing – review & editing (equal). **Rendong Ji**: Conceptualization (equal); Data curation (equal); Formal analysis (equal); Funding acquisition (equal); Investigation (equal); Methodology (equal). **Jun Wang**: Data curation (equal); Formal analysis (equal).

DATA AVAILABILITY

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

REFERENCES

[1]Y. X. Zhang, W. Y. Liu, Y. Jun, and D. Yue, "Optimization of the preparation process and quality control of *Ginkgo biloba* leaf extract," Chin. Pharm. **24**(1), 75–80 (2021).
[2]Y. Wang, Y. W. Peng, and Y. Lu, "Application of near-infrared spectroscopy technology in fine chemical production," Chem. Fiber Text. Technol. **52**(5), 41–43 (2023).
[3]Y. Q. Ouyang, L. Xia, Y. Zhang, F. Zhang, C. Z. Mo, Y. Zeng, and Y. Huang, "Detection of total flavonoids in Epimedium based on near-infrared spectroscopy combined with chemometrics," Hans J. Med. Chem. **12**(1), 61–67 (2024).
[4]Z. Li, H. G. Qi, Y. Yu, C. Liu, R. Cong, L. Li, and G. Zhang, "Near-infrared spectroscopy method for rapid proximate quantitative analysis of nutrient composition in Pacific oyster *Crassostrea gigas*," J. Oceanol. Limnol. **41**(1), 342–351 (2023).
[5]I. R. Rukundo, M. G. C. Danao, J. C. Macdonald, R. L. Wehling, and C. L. Weller, "Performance of two handheld NIR spectrometers to quantify crude protein of composite animal forage and feedstuff," AIMS Agric. Food **6**(2), 463–478 (2021).
[6]C. S. Wei, F. H. Lei, W. X. Ai, J. Feng, H. Zheng, D. Ma, and X. H. Shi, "Rapid identification of six resins and other traditional Chinese medicines based on NIRS technology and PCA-SVM algorithm," Chin. J. Exp. Tradit. Med. Formulae **23**(9), 17–23 (2017).
[7]J. T. Gong, L. Li, H. Q. Zou, D. Xu, D. Q. Wang, Y. Cong, and C. L. Liu, "Establishment of qualitative discrimination model of *Angelica sinensis* and counterfeits based on near-infrared spectroscopy and gradient boosting decision tree," World Sci. Technol.: Modernization Tradit. Chin. Med. **21**(10), 7 (2019).
[8]H. Guo, Y. C. Liang, J. Y. Liu, R. Tan, Y. P. Liu, and H. P. Chen, "Study on rapid qualitative discrimination of the source of marketed Panax notoginseng powder slices based on near-infrared spectroscopy," Chin. Med. Mater. **40**(11), 4 (2017).
[9]X. P. Wang, L. Zhang, P. J. Chen, Y. L. Wang, L. Han, X. J. Gui, R. X. Liu, and X. L. Li, "Feasibility analysis of near-infrared spectroscopy technology applied to the classification and identification of four types of taste in traditional Chinese medicine," Chin. Herb. Med. **54**(4), 11 (2023).
[10]C. W. Chen, T. S. Wang, K. F. Hu, B. H. Bao, H. Yan, and X. C. Yang, "Near-infrared discrimination method of processed Puhuang products based on convolutional neural networks and voting mechanism," Spectrosc. Spectral Anal. **42**(11), 3361–3367 (2022).
[11]Y. Q. He, C. H. Zong, J. Wang, Q. Li, J. Wang, Y. J. Wu, Y. Chen, and X. S Liu, "Prediction of multiple lactone components during chromatographic separation

of Ginkgo leaves using near-infrared spectroscopy," China J. Chin. Mater. Medica **47**(5), 1293–1299 (2022).
[12]L. W. Li, W. S. Zeng, and Y. Han, "Consistency verification model for identifying Ginkgo biloba leaf extract tablets based on near-infrared spectroscopy," Drug Eval. **21**(6), 670–673 (2024).
[13]H. F. Ni, L. T. Si, J. P. Huang, Q. Zan, Y. Chen, L. J. Luan, Y. J. Wu, and X. S. Liu, "Rapid determination of active ingredients during the purification process of Ginkgo biloba leaves using near-infrared spectroscopy optimized by a genetic algorithm with extreme learning machine," China J. Chin. Mater. Med. **46**(1), 110–117 (2021).
[14]L. Cai, S. Zhao, F. Meng, and T. Zhang, "Adaptive K-NN metric classification based on improved Kepler optimization algorithm," J. Supercomput. **81**, 66 (2025).
[15]S. Uddin, I. Haque, H. Lu, M. Moni, and E. Gide, "Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction," Sci. Rep. **12**, 6256 (2022).
[16]N. Qiao, M. Rao, X. Y Huang, Y. Xiao, and C. Zhang, "Rapid identification of honey varieties based on portable near-infrared spectroscopy and random forest method," China Port Sci. Technol. **6**(8), 75–80 (2024).
[17]Z. Wang, Z. Zhang, R. A. Williams, and Y. Li, "NIR inversion model of Larch Wood density at different moisture contents based on MVO-BPNN," J. Appl. Spectrosc. **91**, 472–479 (2024).
[18]L. Sun, Y. Zhang, G. Y. Meng, Y. Yu, F. Gao, and L. Wang, "Research on the identification model of cigar leaf varieties based on near-infrared spectroscopy after aging," Tianjin Agric. Sci. **30**(4), 82–90 (2024).
[19]Y. T. Yang, M. Jiang, R. L. Zhang, H. Xia, Y. A. Cao, Y. Wang, Y. H. Ren, and L. X. Peng, "Establishment and optimization of a rapid detection model for characteristic nutritional components of highland barley," Sci. Technol. Food Ind. **45**(18), 228–238 (2024).
[20]F. Wang, J. Lu, Z. B. Fan, C. J. Ren, and X. Geng, "Continuous motion estimation of lower limbs based on deep belief networks and random forest," Rev. Sci. Instrum. **93**(4), 044106 (2022).