Solving the origin traceability problems of *Boletus bainiugan* based on multiple data processing

Shuai Liu\textsuperscript{a,b}, Honggao Liu\textsuperscript{c}, Jieqing Li\textsuperscript{a*}, and Yuanzhong Wang\textsuperscript{b*}

\textsuperscript{a} College of Agronomy and Biotechnology, Yunnan Agricultural University, Kunming, 650201, China

\textsuperscript{b} Medicinal Plants Research Institute, Yunnan Academy of Agricultural Sciences, Kunming 650200, China

\textsuperscript{c} Yunnan Key Laboratory of Gastrodia and Fungi Symbiotic Biology, Zhaotong University, Zhaotong 657000, Yunnan China

*Corresponding authors:

Mr: Jieqing Li, College of Agronomy and Biotechnology, Yunnan Agricultural University, Kunming, 650201, China. E-mail: lijieqing2008@126.com

Mr: Yuanzhong Wang, Medicinal Plants Research Institute, Yunnan Academy of Agricultural Sciences, Kunming 650200, China. E-mail: Boletus@126.com
Abstract

In this study, various data processing methods were used for model building with the aim of finding a suitable method to solve the origin traceability problems of *Boletus bainiugan*. The effects of partial least squares discriminant analysis (PLS-DA) and support vector machine (SVM) models are compared. In addition, a three-dimensional correlation spectroscopy combined with residual network (3DCOS-ResNet) model was established for the first time under the data fusion strategy. The aim is to provide new options for the selection of data processing methods in qualitative research in related fields. The results show that: 1. PLS-DA and 3DCOS-ResNet models built based on the original dataset are more advantageous; 2. Variable importance in the projection (VIP)-PLS-DA and Competitive Adaptive Reweighted Sampling (CARS)-SVM models based on second-order derivative (2D), Savitzky-Golay (SG) smoothing + 2D, multiplicative scattering correction (MSC) + 2D, and standard normal transform (SNV) + 2D preprocessing methods are the most effective; 3. PLS-DA and SVM models show better performance based on mid-level data fusion strategies. In summary, the choice of data processing methods before building the model is particularly important. Not a random selection of data processing methods can improve the model performance and achieve the expected results.

**Keywords:** *Boletus bainiugan*, Infrared spectroscopy, Chemometrics, Deep learning, Data fusion
Introduction

As a food with medicinal properties, consumers choose to consume mushrooms because of their different characteristics. On the one hand, people choose to eat mushrooms because of their delicious taste and their protein content as a substitute for meat and fish; on the other hand, people choose to eat them because of their medicinal value [1]. Mushrooms provide essential amino acids that humans cannot synthesize on their own, and contain many essential trace elements in the human body [2,3]. Porcini mushrooms are one of the most valuable species of mushrooms, and edible wild boletes are favored in various countries. In addition, it has high antioxidant and anti-microbial activity [4]. Wild porcini mushroom resources in China are mainly concentrated in Yunnan Province, which is a suitable place for wild porcini mushroom growth due to its climate and geographical location. The protein content and other nutrients of porcini mushrooms are influenced by environmental factors, and there are differences in the nutrient content of porcini mushrooms from different origins [5]. Secondly, porcini mushrooms collected from origins with excessive soil background values are most likely to be enriched in large amounts of heavy metals [2]. Therefore, the identification of the origin of Bolete is extremely important, and the clarification of the origin of Bolete can also make a judgment on whether there is a possibility of excessive heavy metals in its fruit bodies. At present, the best technique for solving the identification problem of Bolete is infrared spectroscopy. The use of infrared spectroscopy to obtain information on porcini mushrooms does not require the use of chemical reagents and the instrument is simple and efficient to operate. Infrared spectroscopy has been shown
in several studies to be a non-destructive, fast, convenient and accurate analytical 
technique, thus enabling a wide range of applications in important analytical fields [6].

For example, Esteves et al. [7] solved the identification problem of magic mushrooms 
by using FTIR-ATR and OPLS-DA to build a discriminative model that could identify 
the target object accurately and achieve loss lessness and efficiency in the experimental 
process; Dong et al. [8] used Fourier transform near infrared (FT-NIR) combined with 
deep learning to solve the species identification problem of wild porcini mushrooms 
with a model accuracy of 100%; Li et al. [9] developed a PLS-DA model to solve the 
rice adulteration problem using seven pre-processing methods and CARS to process 
the IR spectral data and successfully improve the model performance.

Infrared spectroscopy has the characteristics of non-destructive, non-polluting, 
simple operation and accuracy compared with other wet chemical methods, but it needs 
to be combined with suitable modeling methods to take greater advantage. The use of 
IR spectroscopy combined with machine learning has great advantages, but the choice 
of data processing methods prior to modeling is particularly important, as can be 
derived from the results of Yan et al. [10].

The purpose of this study is to build an origin identification model of Boletus 
bainiugan based on different data processing and compare the model effects of PLS- 
DA, SVM and 3DCOS-ResNet after different data processing. For the first time, 
3DCOS combined with deep learning model is established under data fusion strategy. 
Then, based on this, the performance effects of different data processing methods on 
the model are explored. The optimal data processing method is derived for this study,
which provides valuable reference for data processing in qualitative research.

2. Materials and methods

2.1 Information of samples

_Boletus bainiugan_ were collected from six different origins (1: Chuxiong, 2: Dali, 3: Kunming, 4: Nujiang, 5: Qujing and 6: Yuxi) in Yunnan Province, China between 2021 and 2022, with the number of samples collected being 85, 26, 24, 55, 21 and 61, respectively. Samples were collected at a number of different sampling points in six sampling areas, with the sole aim of increasing the biological variability of the samples to improve model performance (Phuangsombut, Meetim, & Terdwongworakul, 2023).

Species of all experimental samples were identified by Yuanzhong Wang (a researcher at the Institute of Medicinal Plants, Yunnan Academy of Agricultural Sciences) and detailed information of samples are shown in Table 1. The collected fresh samples were dried with a hot blower (45°C) until the samples were at a constant weight to end the drying process. The dried intact fruit bodies were pulverized using a pulverizer, filtered through a 100-mesh sieve, stored in polyethylene bags sealed at room temperature, and protected from light for subsequent experiments. The operating procedure is shown in Figure 1.

2.2 Acquisition of infrared spectral data

2.2.1 Acquisition of Fourier transform near infrared (FT-NIR) spectral data

NIR spectra were collected using FT-NIR spectrometer (Thermo Science Inc., Antaris II) and samples were scanned after the instrument had warmed up for 30 minutes. 2 g of _Boletus bainiugan_ powder was loaded into a transparent device and held
upside down in the light port of the FT-NIR instrument waiting to be scanned. The instrument measures the spectral range of 10,000-4,000 cm\(^{-1}\) with a resolution of 8 cm\(^{-1}\).

A total of 64 scans are made and the samples are repeated three times before the average spectrum is taken. The instrument background is first acquired before the sample is scanned to remove any interference from CO\(_2\) and H\(_2\)O in the air.

2.2.2 Acquisition of Fourier transform middle infrared (FT-MIR) spectral data

Spectral data was collected using a FT-MIR Spectrometer (Thermo Fisher Scientific 5225 Verona Rd.) equipped with a diamond crystal attenuated total reflection (ATR) accessory and the instrument was warmed up for 30 minutes before scanning the samples. Take 1 g of *Boletus bainiguan* powder and place it in the round hole of the ATR attachment and wait for scanning. The spectra of clean and dry ATR crystals against air were recorded and used as background spectra. Once the background spectra have been acquired, the formal scanning of the sample proceeds. A cleaning procedure was carried out by wiping the ATR crystals with ethanol (99%) before and after each scan of the sample. The spectra were kept at room temperature throughout the experiment and the spectral range was 4000-400 cm\(^{-1}\) with a resolution of 4 cm\(^{-1}\). 64 scans were accumulated and the average spectrum was taken after three repetitions.

2.3 Principal component analysis (PCA)

The main reason for performing principal component analysis on samples is that it is an unsupervised method. Firstly, it is possible to determine whether there is a difference between samples by the characteristics of the samples themselves without relying on human influence; secondly, it is possible to extract the characteristic...
variables of the samples while reducing the dimensionality of the data [11,12]. In this study, the principal component analysis score plot can visually show whether the samples from different origins can be clustered. After the principal component analysis, it can also help the researcher to indicate how the subsequent research work should be carried out and to be clearer about the characteristics among the samples studied. The score plots of the principal components of *Boletus bainiugan* from the six origins are shown in Figure 2. After deriving the grouping trends of different origins of *Boletus bainiugan* through the principal component analysis, it was then decided whether to choose a supervised approach to try to solve the problems of *Boletus bainiugan* origins identification.

### 2.4 Advantages of data processing

The total number of variables collected in the raw FT-NIR spectra was 1557 and the total number of variables collected in the raw FT-MIR spectra was 7468. Due to a large amount of data and redundant information in the mid-infrared, the model can take long time to train. At the same time, the model performance can be affected. In particular, modeling with low-level data fusion strategies will be time-consuming. Therefore, these problems require pre-processing of the raw spectral data before formal model building. SG smoothing can effectively reduce noise and improve signal-noise ratio; MSC and SNV are used to remove the effect of scattering on the spectrum due to uneven distribution or size of sample particles; 2D can solve the problem of unstable spectral baseline [13,14]. The use of feature variable extraction methods enables the acquisition of important variables while reducing the amount of data and increasing the
speed of model training [15]. Preprocessing methods such as 2D, SNV, SG smoothing (third-order polynomial, 11-point windows) and MSC and their combinations are used in this study to reduce the adverse effects. The mid-level data fusion is used to solve the problem of long training time and poor performance of the model caused by many variables and small contributions, which is extremely important in practical applications because the time cost has to be taken into account; on the other hand, the fusion of information from different sources can reflect the information of sample characteristics more comprehensively thus improving the performance of the model. In this study, the model is built on the basis of the original data and the pre-processed data for feature variable extraction, and the model performance is compared to find the optimal combination methods. Using the Kernnard-Stone (KS) algorithm, 2/3 of the data are used as the training set and 1/3 as the test set for subsequent model building.

2.4.1 Single-spectrum data processing before modeling

The FT-NIR and FT-MIR spectral data were preprocessed by 2D, MSC, SNV, SG and their combinations, and the models were built after different preprocessing in order to compare the performance improvement of the models under different data preprocessing. Three feature variable extraction methods were used based on different preprocessing, namely VIP, SPA and CARS. The purpose is to compare the models built based on different pre-processing followed by feature variable extraction. In turn, it goes to explore whether the degree of performance variation of the same model differs significantly from that of the model built under different pretreatments, and whether the effect of the same model is substantially improved after removing the influence of
redundant information of the samples as much as possible; and whether the performance difference between different models is significant under the same data processing.

2.4.2 Low-level data fusion processing before model building

The two datasets are simply joined and then the data are processed using different pre-processing methods, and then a model is built for comparison. The difference between building a model after low-level data fusion and building a model with a single spectrum is that the model obtains more complete sample information, larger amount of data, longer training time and slower training speed. Therefore, after comparing the performance of model building with low-level data fusion under different pre-processing, three feature variables are extracted from the low-level data fusion. The use of feature variable extraction can achieve the elimination of non-feature information while reducing the amount of data, thus improving the model training speed. Ensuring the integrity of sample information while reducing the time to train the model using non-feature information. Based on this, the performance comparison of the same model after different data processing and different models after the same data processing is performed.

2.4.3 Mid-level data fusion processing before model building

Mid-level data fusion is the use of feature variable extraction to filter out variables in different data sets that can express information about sample characteristics and cascade the data. A total of three feature variable extraction methods (VIP, CARS, SPA) are used in this study and are divided into two ways for feature-level fusion. The first
one: using the same feature variable extraction method to extract feature variables for FT-NIR and FT-MIR datasets respectively, and then fusing the obtained feature variables with the data. Selection criteria were that the preprocessing and feature variable extraction methods that performed best in mono spectral modeling were selected to perform mid-level data fusion; the second one: feature variable extraction using different feature variable extraction methods for both datasets, followed by feature-level data fusion, with the same method chosen for pre-processing and feature variable extraction that performs optimally in a single spectral modeling. In this study, the VIP values of the variables were obtained by PLS-DA, and the variables with VIP values greater than or equal to 1 were screened as the characteristic variables of the studied samples. When using CARS to extract feature variables, the feature variables are extracted by CARS several times, and a model is built for comprehensive evaluation based on the feature variables extracted each time to finally determine the optimal number of feature variables. The basis for CARS and SPA to determine the number of characteristic variables is shown in Figure S1.

2.5 Machine Learning

2.5.1 Chemometrics model

Using the KS algorithm, 2/3 of 272 samples from 6 origins were used as the training set (57, 17, 17, 37, 14, 41) and the remainder as the test set (28, 9, 7, 18, 7, 20). Based on the different processed data to build models, PLS-DA and SVM models under single-spectrum data and multi-spectrum data are established respectively. The detailed PLS-DA model types are shown in Table S1-6. The model was built with 5 cross
validations and passed 200 permutations, and the results are shown in Figure S2 in the Supplementary Material. $R^2$ represents the fitting ability of the model and $Q^2$ represents the predictive ability of the model. The relationship between the $R^2$ and $Q^2$ values of the model is used to indicate whether there is overfitting. To judge whether the model is overfitted, as shown in Figure S2. If the left-hand value is smaller than the right-hand true value and the intercept of the dashed line extending from the true value to the line $X=0$ is smaller than the vertical coordinate of the true value, the model is not overfitted, otherwise the model is overfitted. The process of PLS-DA model building is all done independently by the software SIMCA 14.1.

The types of SVM models are shown in Tables S7-12, and the training set hyperplane diagram and test set classification of the model are shown by Figure S3. The SVM model is built by using the grid search method to find the penalty parameter $c$ and the kernel parameter $g$ of the model. The range of $c$ and $g$ values is determined as the most appropriate range based on the prediction set results of the final model [16]. The value of $c$ is higher or lower than the reasonable range, the generalization ability of the model becomes lower; the value of $g$ is not in the reasonable range and the smaller value indicates that the number of support vectors of the model is larger, and the larger value indicates that the number of support vectors of the model is smaller. Too many support vectors cause the training speed of the model to decrease and take longer time. Therefore, the SVM model can better reveal the model advantage when dealing with classification problems with small samples.

2.5.2 3DCOS combined with ResNet
In this study, a total of 3264 3DCOS images were generated, of which 816 images each were generated from near-infrared spectral data, mid-infrared spectral data, low-level data fusion and mid-level data fusion. Each dataset is divided into synchronous, asynchronous and integrated 3DCOS images, with a total of 272 images in each category. There are 54 images in the training set, 63 images in the test set, and 155 images in the external validation set. The full band of FT-NIR spectra was selected for 3DCOS map generation, while 1750-400 cm\(^{-1}\) in FT-MIR spectra was selected for 3DCOS image generation. The reason for this is that the peak is mainly in this band and the software workload can be reduced. Compared with 1DCOS and 2DCOS images, 3DCOS images can solve the problem of overlapping and masking information in 1DCOS and the absorption peaks can be more clearly observed in 3D stereo images [17]. The generation of 3DCOS images is done by Matlab R2017 software, and the generated images are automatically classified and saved in JPEG format to prepare for the next step of deep learning model building. The formula utilized for spectral image generation is shown below.

\[
S(v) = \begin{cases} 
  s(v, t_1) \\
  s(v, t_2) \\
  \vdots \\
  s(v, t_m)
\end{cases}
\]  

(1)

\(S\) represents the spectral intensity at variable \(v\), \(v\) represents any variable in the spectrum, \(t\) represents the time interval of the spectral perturbation, and \(m\) represents the number of measurements of the spectral intensity [18].
\[ \Phi(v_1, v_2) = \frac{1}{m-1} S(v_1)^T S(v_2) \]  
\[ \Psi(v_1, v_2) = \frac{1}{m-1} S(v_1)^T N S(v_2) \]  
\[ I(v_1, v_2) = \Phi(v_1, v_2) \cdot \Psi(v_1, v_2) = \frac{1}{(m-1)^2} [S(v_1)^T S(v_2)] [S(v_1)^T N S(v_2)] \]  

The principles of acquiring synchronous 3DCOS(\(\Phi\)), asynchronous 3DCOS(\(\Psi\)) and integrated 3DCOS(\(I\)) images are shown in Equations (2), (3) and (4), respectively [19,20].

\[ N_{jk} = \begin{cases} 0, & j = k \\ \frac{1}{\pi(k-j)}, & j \neq k \end{cases} \]  

\(N\) for the Hilbert-Noda matrix in Equation (5).

ResNet models are also built in four categories, based on FT-NIR, FT-MIR, low-level data fusion and mid-level data fusion datasets. Unlike the data types used in the PLS-DA, SVM model, the ResNet model built under single spectrum is based on the original spectral data; while the ResNet model built by low-level data fusion is based on low-level data fusion followed by feature variable extraction using VIP; the ResNet model built by mid-level data fusion is based on VIP combined with data fusion under different pre-processing to build the model. Detailed information is shown by Table 2. The flow chart of the 3DCOS image generation method is shown in Figure S4. The model building and external validation are done by the software Spyder, and the loss values, the accuracy and the confusion matrix plots of the external validation of the
models are shown in Figure S5.

2.6 Evaluation methods of model performance

A total of two chemometric methods namely PLS-DA and SVM models are used in this study. An additional deep learning model (3DCOS combined with ResNet) was used. Each model was evaluated for sensitivity (Sen), specificity (Spe) and accuracy (Acc) of the test set. Besides, the PLS-DA model was evaluated by root mean square error of evaluation (RMSEE), root mean square error of cross-validation (RMSECV), and root mean square error of prediction (RMSEP); the SVM model was evaluated by parameters c and g; and the ResNet model was evaluated for the number of iterations epoch and loss values. The Sen, Spe and Acc of the model are calculated as follows:

\[
\text{Sen} = \frac{TP}{TP + FN} \\
\text{Spe} = \frac{TN}{TN + FP} \\
\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN}
\]

The TP of equations (6), (7), (8) are the true positive, which means the number of samples correctly classified in each category; TN is true negative, which means the sum number of classes other than the current target class that can be classified correctly; FP is a false positive, which means the number of classes other than the target class
misclassified to the column in which the target class is located; FN is false negative, which means the number of classes other than the target class misclassified to the row in which the target class is located. Sen, Spe and Acc values of the models are calculated from the confusion matrix plot of the model classification. The magnitude of these three metrics of the model are compared to indicate the good or bad performance of the model.

3. Results and Discussion

3.1 PCA

PCA is based on fusion datasets of the original FT-NIR, FT-MIR and low-level data, and the results are shown in Figure 2. Figures a, b, and c show the principal component analysis score plots for the three original datasets. The relationship plots for PC1 and PC2 explain 83.8% and 13.6% of the total sample variance respectively; b plot of PC1 and PC2 explained 58.6% and 32.7% respectively; c plot of PC1 and PC2 explained 27.2% and 49.4% respectively. From the distribution of Boletus bainiugan among the six origins in Figs. a, b and c, there was no significant clustering effect among the origins. On the one hand, PCA could not distinguish Boletus bainiugan from six origins, probably because the species studied were all Boletus bainiugan and all came from Yunnan. The samples were grown in similar environments, so the differences in their chemical composition were not obvious; on the other hand, as an unsupervised chemometric method, its effect on the classification of such samples was less than that of the supervised method.

3.2 Analysis of spectral images

The spectra of FT-NIR and FT-MIR are shown in A and B of Figure 3, respectively.
By observing the two spectra, it was found that the peak shapes of *Boletus bainiugan* spectra of the six origins were consistent, with some differences only in the absorbance values. The differences in absorbance values reflect the differences in the chemical composition content of *Boletus bainiugan* grown in different origins. The analysis of the possible chemical composition of each wave number band was first carried out from the A-plot, with the main peaks in the range of 8600 to 4000 cm\(^{-1}\). The peaks in the range of 8379 to 8000 cm\(^{-1}\) may be associated with \(-\text{CH}_3\) [21]. Absorption peaks in the band range of 6800-6300 cm\(^{-1}\) are associated with O-H stretching of starch in mushrooms and N-H stretching of amines [21,22]. In the vicinity of 5765 cm\(^{-1}\) is the stretching vibration of C-H [23]. The absorption peaks corresponding to the wave number band from 5149 to 5130 cm\(^{-1}\) may be related to the combination of O-H stretching and O-H-O deformation; the absorption peaks corresponding to the spectral bands in the range of 4594-4326 cm\(^{-1}\) are associated with a combination of stretching and deformation of the protein [24].

The B-plot shows the FT-MIR spectrum, and the peak near 3274 cm\(^{-1}\) is associated with the stretching vibration of the O-H bond [25]. The peak near 2929 cm\(^{-1}\) is associated with a combination of symmetric and asymmetric vibrations of C-H in the CH\(_3\) group and CH\(_2\) asymmetric vibrations [26]. The absorption peak at 1625 cm\(^{-1}\) is associated with the C=O bond [27]. The peaks in the 1570-1554 cm\(^{-1}\) may correspond to the amide II bands of fat and protein in *Boletus bainiugan* [28,29]. The absorption peak at 1399 cm\(^{-1}\) is related to the stretching vibration of the C-N bond [30]. The absorption peak near 1021 cm\(^{-1}\) may be related to the C-O bond stretching vibration of
aromatic ethylene [31]. It is difficult to observe the differences between the six origin samples with overlapping spectrograms. Therefore, different models were subsequently used to solve the problem of difficulty in differentiating between origins.

3.3 Comparative analysis of three models under raw spectral data

As shown in Table S1, the model of Low data fusion-RAW-PLS-DA works best. The Acc of the model was 96.39% and the RMSEP value was 0.21884 and there was no risk of overfitting the model. After two hundred cross-validations of the model, the relationship between $R^2$ and $Q^2$ is shown in Figure S2. Acc, Sen and Spe values the model classification are relatively lower for FT-NIR-RAW-PLS-DA and FT-MIR-RAW-PLS-DA. Table S2 shows the comparative results of model building using VIP on the basis of raw spectral data. The low data fusion-RAW-VIP-PLS-DA is the best. The other indicators were different, but the differences were not significant. Table S3 shows the models built after CARS processing, and Low data fusion-RAW-CARS-PLS-DA has the best classification results. In summary, the reason may be due to the sample information obtained by the model is more comprehensive under low-level data fusion, which leads to the improved performance of the model. Table S4 is a comparison table of the effects of the models after the SPA treatment. Under the comparison of the models with the three types of spectral data, the results were found to be different from the previous ones. The test set classification accuracy of low data fusion-RAW-SPA-PLS-DA was only 59.18%. The $R^2$ and $Q^2$ values were 0.09399 and 0.07094, respectively, and the RMSEE, RMSEP and RMSECV were higher than 0.3. The FT-MIR-RAW-SPA-PLS-DA model works best and the Low data fusion-RAW-
SPA-PLS-DA works worst. In general, except for Low data fusion-RAW-SPA-PLS-DA, the classification results of the models built are more satisfactory. However, all models are built on the basis of the original data, and there is still room to improve the RMSEE, RMSEP, RMSECV, $R^2$ and $Q^2$ values of the models.

According to Tables S7-S12 of the supplementary materials, the optimal ranges of $c$ and $g$ for the SVM model with different data processing are $2-2^6$ and $10^{-4}-0.9$, respectively. If the SVM model works well, the values of parameters $c$ and $g$ should be within a reasonable range, while the value of $c$ should be as large as possible to make the model generalization ability stronger, and the value of $g$ should also be as large as possible, so that the number of support vectors becomes smaller, thus achieving an increase in model training speed. Table S7 shows the results of comparing the SVM models built based on three different raw spectral datasets, the FT-NIR-RAW-SVM has the highest Acc, Sen and Spe values. However, the parameters $c$ and $g$ of the models built on the three datasets are out of reasonable range. The $c$ value of FT-NIR-RAW-SVM is too large, and the model has the risk of overfitting; the $c$ value of FT-MIR-RAW-SVM is too large and the $g$ value is too small beyond reasonable range, which leads to the risk of overfitting and the number of support vectors is too large, which reduces the model training speed; $g$-value of low data fusion-RAW-SVM is too small, which reduces the model training speed, takes a lot of time and has poor model performance. The three models in Table S8 are FT-NIR-RAW-VIP-SVM, FT-MIR-RAW-VIP-SVM and low data fusion-RAW-VIP-SVM. The third model is the least effective and the values of $c$ and $g$ are not within a reasonable range. The first two
models had better model classification results, the $c$ and $g$ values were likewise not within a reasonable range. The $c$-values of FT-NIR-RAW-CARS-SVM, FT-MIR-RAW-CARS-SVM and Low data fusion-RAW-CARS-SVM in Table S9 are out of reasonable range and the models are at risk of overfitting. The accuracies of FT-NIR-RAW-SPA-SVM, FT-MIR-RAW-SPA-SVM and Low data fusion-RAW-SPA-SVM in Table S10 are above 90%, but the $c$-values of all three models are beyond reasonable range. The comparison of the above various models shows that using the feature variable extraction method to build a model based on the original data set also does not allow the model to achieve satisfactory results. In summary, the parameters $c$ and $g$ of the model do not meet the requirements of a reasonable range even if the feature extraction methods are used on the basis of the three original data sets. It is also necessary to add data preprocessing methods again to improve the range of parameters $c$ and $g$ of the model.

Build deep learning models based on two kinds of raw spectral data separately. Compare the synchronous, asynchronous and integrated 3DCOS-ResNet models under two kinds of raw spectral data. The performance comparison results of the models are shown in Table 2, and the synchronous 3DCOS-ResNet model built based on the FT-NIR raw spectral data works best. The accuracy of the model reaches 100% on the training and test sets as well as 99.12% on the external validation set. In the case of modeling based on raw spectral data, the 3DCOS-ResNet model gives the best results compared to PLS-DA and SVM. The advantage is that the spectral data do not need to be pre-processed or feature variables extracted, and the model can achieve better results.
The high efficiency, low loss value, high accuracy, good sensitivity and specificity of the model are sufficient to make the model the preferred technical approach in identification studies.

3.4 Improvement of PLS-DA models under different data processing

3.4.1 Comparative analysis of the performance of PLS-DA under different pretreatment

The performance comparison of PLS-DA models built under three different spectral datasets is shown in Table S1, and the performance of the models built based on different preprocessing methods differed significantly. The model built by preprocessing the FT-NIR spectral dataset using 2D was the best, with a model accuracy of 99.11%. Compared to the models built after other pre-processing, the model with 2D treatment has the smallest RMSEP value and the $R^2$ and $Q^2$ are 0.87768 and 0.62862, respectively. Modeling based on the FT-MIR dataset was found to work best using the SNV+2D method. Although it is slightly inferior to the model built based on the 2D processed FT-NIR data, the model works well overall. After low-level data fusion, the Acc, Sen and Spe of the model built by pre-processing with SNV+2D reached 100%. The RMSEE, RMSECV and RMSEP values were the lowest compared to the previous two models and the $R^2$ and $Q^2$ values of the model reached 0.94019 and 0.71481. The advantage of low-level data fusion is that the information obtained when the model is built is more comprehensive and thus can improve the model performance. All models were not overfitted, and the results of the models after 200 permutation tests are shown in Figure S2. With the above conclusions, we conclude that preprocessing...
the spectral data can improve the RMSEE, RMSEP, PMSEP, $R^2$ and $Q^2$ values of the model. And there is a significant improvement compared with the model built based on the original data.

3.4.2 Performance analysis comparison of different pre-processing combined with different feature variable extraction methods for building PLS-DA models

The comparison between Tables S1 and S2-S4 shows that modeling was performed based on three different datasets. Except for the VIP method, the models built using CARS and SPA extraction methods are worse than those built using only different preprocessing. Two feature variable extraction methods, namely CARS and SPA, were used on the basis of different pretreatments. The $R^2$ values of the models ranged from 0.32 to 0.75 and 0.09 to 0.57, respectively, and the $Q^2$ values ranged from 0.24 to 0.59 and 0.07 to 0.40, respectively. The $R^2$ values of the models built based on different pretreatments ranged from 0.48 to 0.94, and the $Q^2$ values ranged from 0.28 to 0.73. By comparison, it is found that the above two feature variable extraction methods do not improve the model performance and make the model $R^2$ and $Q^2$ values decrease. The VIP algorithm led to some improvements in the performance of the models based on the FT-NIR spectral dataset, but the overall effect did not change significantly; the models based on the FT-MIR dataset showed some degradation in Sen, Spe and Acc, but the $R^2$ and $Q^2$ values of the models built using the SNV+2D method were improved. A comprehensive comparative analysis concluded that the use of three different feature variable extraction methods on the basis of different preprocessing did not bring much improvement to the model. A fairly satisfactory result can be achieved.
by pre-processing the data on the basis of low-level data fusion. At the same time, the models did not appear to be over-fitted.

3.4.3 Performance analysis comparison of PLS-DA models for mid-level data fusion with the same feature variable extraction and different feature variable extraction

As shown in Tables S5 and S6, after comparing two different mid-level data fusion strategies, it is concluded that the performance of the model exhibits satisfactory results. Sen, Spe and Acc of the model built under the second method in Table S5 reached 100%, and the $R^2$ and $Q^2$ values of the model were 0.778337 and 0.61902, respectively. The classification accuracies of the other models are also above 97%, and there is no major difference in the performance of the two mid-level fusion models. As observed in Table S6, the best modeling results were obtained by fusing the data of FT-NIR-2D-VIP and FT-MIR-SNV+2D-CARS.

According to the model comparison results, the PLS-DA model built by using CARS for feature variable extraction of FT-NIR and FT-MIR spectral data is better than the models built based on the other two feature variable extraction methods. The low-level data fusion strategy is not worse than the mid-level fusion strategy, and the results of the model built are equally satisfactory. Comparing only the mid-level fusion strategies based on the extraction of two different feature variables, Sen, Spe and Acc of the different models did not reach 100%, but all were above 90% and the models were not overfitted. The overall effect of the models is superior under the mid-level fusion strategy.
3.5 Improvement of SVM models under different data processing

3.5.1 Comparative analysis of the performance of SVM under different pretreatment

As shown in Table S7, the models achieved satisfactory results after the raw FT-NIR spectral data were processed by four pre-processing methods. The Sen, Spe and Acc of the models were 98.88%, 99.77% and 99.62%, respectively. And the values of parameters $c$ and $g$ of the model are within reasonable limits. Comparison of the final results shows that the model built using the preprocessing method performs better and gives satisfactory results than the model built based on the original spectral dataset. The effect of the model built under FT-MIR spectral dataset is consistent with FT-NIR, and the model classification effect is better and the $c$ and $g$ values are within a reasonable range. The dataset obtained after low data fusion was subjected to four preprocessing processes, and the classification results of the other three models were satisfactory, except for the model with low data fusion-2D-SVM, which had poor performance. The Sen, Spe, and Acc of low data fusion-SG+2D-SVM and low data fusion-MSC+2D-SVM all reach 100%. It follows that the use of appropriate preprocessing methods during model building in general does improve the performance of the model. The model results are still unsatisfactory after data preprocessing, and a discussion of whether the use of feature variable extraction can improve the performance of the model is carried out in section 3.5.2.

3.5.2 Performance analysis comparison of different pre-processing combined with different feature variable extraction methods for building SVM models
As shown in Table S8, the Sen, Spe, and Acc values of FT-NIR-2D-VIP-SVM were 100%, which improved the model performance compared with FT-NIR-2D-SVM. The other three models worked well, and the performance of the models did not change significantly compared to the models built without the VIP. The FT-MIR-2D-VIP-SVM model has poor performance and $c$ and $g$ values are beyond reasonable range, and the FT-MIR-2D-SVM has better performance. The performance of the other three models built based on FT-MIR spectra did not vary significantly. Under the low data fusion strategy, the Sen, Spe, and Acc values of low data fusion-2D-VIP-SVM are 100%. Compared to low data fusion-2D-SVM, the model performance is greatly improved. However, the performance of the other three models deteriorated and the $c$ and $g$ values were out of reasonable range. The effects of the models built by CARS processing on the basis of different preprocessing are shown in Table S9. The model performance of FT-NIR-SNV+2D-CARS-SVM, FT-MIR-2D-CARS-SVM, FT-MIR-SNV+2D-CARS-SVM and low data fusion-2D-CARS-SVM are improved. The classification effects of the other models did not significantly change the model performance due to the use of CARS. Table S10 shows the performance comparison results of the models built with SPA. The model performance of FT-MIR-2D-SPA-SVM, FT-MIR-MSC+2D-SPA-SVM and low data fusion-SG+2D-SPA-SVM is reduced to some extent compared to Table S7, with the third model showing the most significant reduction in performance effect. The model performance of low data fusion-2D-SPA-SVM is greatly improved compared to low data fusion-2D-SVM.

According to the above comparison results, the model performance of low data
fusion-2D-SVM can be improved by using feature variable extraction methods. On the basis of better model performance, there is no need for feature extraction, otherwise the result may appear to degrade the original model performance.

3.5.3 Performance analysis comparison of SVM models for mid-level data fusion with the same feature variable extraction and different feature variable extraction

Table S11 shows the comparison results of the mid-level fusion models built based on the same feature variable extraction method. The performance of the mid-level fusion models built by the three feature variable extraction methods is better, and the c and g values of the models are within a reasonable range. Table S12 shows the comparative results of the mid-level fusion models built based on different feature variable extraction methods. The performance of models is better, and c and g values of the models were in a reasonable range. The comprehensive comparison enables to conclude that the SVM models built with the same or different feature variable extraction methods can obtain satisfactory classification results under the mid-level fusion strategy.

3.6 Deep learning models under different spectral data processing

The first row in Table 2 shows the 3DCOS-ResNet model built for the FT-NIR spectral dataset. 3DCOS images are classified as synchronous, asynchronous and integrated. The models built from different 3DCOS images combined with ResNet vary greatly in their effectiveness. The FT-NIR-RAW-Synchronous 3DCOS-ResNet model works best. The epoch of the model is only 57, the loss value is 0.048 and the Sen, Spe and Acc are above 97%. The models built from asynchronous and integrated 3DCOS
images combined with ResNet were less effective, with an accuracy of less than 64%
for the external validation set. The external validation confusion matrix plots for the
three models are shown in Figure S4. As shown in Table 2, the synchronous model has
the best results among the three types of FT-MIR-RAW-3DCOS-ResNet models. The
epochs of the model are 58, the loss value is 0.048, and the Sen, Spe and Acc values of
the model are 89.68%, 98.12% and 96.19%, respectively. The synchronous model
based on FT-MIR is less effective than the model based on FT-NIR. The asynchronous
and integrated models are less effective compared to the synchronous models. The
external validation confusion matrix diagrams for the three models are shown in Figure
S4. The accuracy of the synchronous model of low data fusion-2D-VIP-3DCOS-
ResNet is 97.75%, and its Sen and Spe are both above 90%. Due to the high number of
variables under low-level data fusion, the epoch of the model will be elevated. The
model loss value reaches a minimum of 0.081 when the epochs of the synchronous
model are 89. The three-evaluation metrics of the model are above 93%, and the model
effect is superior. Data fusion was performed after 2D-VIP processing for NIR and MIR,
respectively, and the best results were obtained for mid-data fusion-asynchronous-
3DCOS-ResNet based on the comparison of the mid-level fusion models. The accuracy
of the model was 87.43%, but the Sen value was only 69.87%. The other two models
were less effective, with Sen values below 60%. The epochs of all three models were
above 100, but the models did not show better results. The classification accuracy of
the 3DCOS-ResNet model built under the mid-level fusion strategy was lower than that
of the model built by single-spectrum and low-level data fusion.
A comparison of the models built from different spectral datasets shows that NIR-RAW-Synchronous-3DCOS-ResNet is the best model, followed by MIR, low data fusion and mid-data fusion. The model effect did not achieve the expected result, probably due to the result of the small sample size, which did not allow the model to be better trained.

3.7 Results

The two chemometric methods were compared on the basis of mono spectral and multispectral data under different data processing methods. It was found that the PLS-DA model built after different pre-processing of the spectral data was better than the SVM model. And the overall performance of building SVM models based on the spectral data under CARS and SPA feature variable extraction methods is better than that of PLS-DA models. In addition, for the PLS-DA model, VIP-PLS-DA has the best performance among the models built after extracting three feature variables based on four pre-processing; for the SVM model, CARS-SVM has the best performance among the models built after extracting three feature variables based on four pre-processing.

The advantage of the deep learning model is that based on the original single-spectrum data modeling, the accuracy of the model can reach more than 96%.

4. Conclusion

In this study, four pre-processing methods (2D, SG+2D, SNV+2D, MSC+2D) and three feature variable extraction methods (VIP, CARS, SPA) were used to process the data of NIR, MIR and NIR+MIR spectral, based on which two chemometric models (PLS-DA, SVM) models were built for comparison. In addition, 3DCOS-ResNet
models are built under the original NIR, MIR spectral datasets and low-level and mid-level data fusion for single-spectrum and multi-spectral deep learning model comparison. The following conclusions are drawn: 1. Among the models built on the basis of different original datasets, the low data fusion-RAW-PLS-DA model works best, followed by the NIR-RAW-3DCOS-ResNet model. Modeling based on the original spectral data, the $R^2$ and $Q^2$ values of the model are low; 2. The model performance of low data fusion-SNV+2D-PLS-DA, low data fusion-SG+2D-SVM and low data fusion-MSC+2D-SVM built with different preprocessing methods is the best and the $R^2$ and $Q^2$ values of the models are improved; 3. Among the models built using VIP on the basis of different preprocessing, low data fusion-MSC+2D-VIP-PLS-DA, NIR-2D-VIP-SVM and low data fusion-SNV+2D-VIP-SVM models work best; 4. The PLS-DA models built using CARS and SPA were not satisfactory, with higher values of RMSEE, RMSECV and RMSEP, and lower values of $R^2$ and $Q^2$. NIR-SNV+2D-CARS-SVM, MIR-2D-CARS-SVM and NIR-SG+2D-SPA-SVM work best; 5. Both PLS-DA and SVM models show excellent model results under the two mid-level fusion methods. On the contrary the 3DCOS-ResNet model under mid-level fusion is slightly less effective, and the 3DCOS+ResNet model under low-level fusion is better.

In summary, the use of appropriate preprocessing methods and feature variable extraction methods when processing single-spectrum data can improve the performance of the model to some extent, while reducing the number of variables to enhance the model training speed and improve the accuracy of the model. The model always works satisfactorily under different mid-level data fusion strategies, which is the better choice.
among the methods for processing data. In this study, the origin identification problem of Boletus bainiuger was solved as an example, and the model was built after different data processing for three data sets. By comparing the model effects, the better data processing method are derived to provide some reference for qualitative research in related fields.

Acknowledgements

The study was supposed by the National Natural Science Foundation of China (Grant number:32160735); Special Program for the Major Science and Technology Projects of Yunnan Province (202002AA100007); and Special Program for the Major Science and Technology Projects of Yunnan Province (202102AE090051-1-01)

References


[12] C. P. Sharma, S. Sharma, V. Sharma, and R. Singh, Rapid and non-destructive


[26] A. No and A. M. Committee, Fourier transform infrared spectroscopic analysis of


<table>
<thead>
<tr>
<th>Specie</th>
<th>Origin</th>
<th>Number</th>
<th>Longitude</th>
<th>Latitude</th>
<th>Elevation(m)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Boletus bainiugan</em></td>
<td>Huangtushan Mountain, Lufeng County, Chuxiong Prepecture</td>
<td>21</td>
<td>102.088</td>
<td>25.183</td>
<td>1573</td>
</tr>
<tr>
<td></td>
<td>Shizi Mountain, Wuding County, Chuxiong Prepecture</td>
<td>11</td>
<td>102.3781</td>
<td>25.5492</td>
<td>2368</td>
</tr>
<tr>
<td></td>
<td>Xiaobaita Mountain, Wuding County, Chuxiong Prepecture</td>
<td>9</td>
<td>102.38</td>
<td>25.4933</td>
<td>1891</td>
</tr>
<tr>
<td></td>
<td>Yuntaishan Mountain, Nanhua County, Chuxiong Prepecture</td>
<td>14</td>
<td>101.3091</td>
<td>25.3102</td>
<td>2516</td>
</tr>
<tr>
<td></td>
<td>Xishansi Mountain, Nanhua County, Chuxiong Prepecture</td>
<td>11</td>
<td>101.2501</td>
<td>25.1932</td>
<td>1984</td>
</tr>
<tr>
<td></td>
<td>Shuangdian Town, Nanhua County, Chuxiong Prepecture</td>
<td>18</td>
<td>101.2702</td>
<td>25.2701</td>
<td>2051</td>
</tr>
<tr>
<td></td>
<td>Laojunshan Mountain, Jianchuan County, Dali City</td>
<td>12</td>
<td>99.7094</td>
<td>26.6308</td>
<td>2789</td>
</tr>
<tr>
<td></td>
<td>Bijiashan Mountain, Jianchuan County, Dali City</td>
<td>14</td>
<td>99.6331</td>
<td>26.4233</td>
<td>2502</td>
</tr>
<tr>
<td></td>
<td>Near the Shilin County, Kunming City</td>
<td>14</td>
<td>103.34</td>
<td>24.64</td>
<td>1902</td>
</tr>
<tr>
<td></td>
<td>Kedudanguihongjun Village, Xudian County, Kunming City</td>
<td>10</td>
<td>102.849</td>
<td>25.5687</td>
<td>2035</td>
</tr>
<tr>
<td></td>
<td>Shuiyanchang Village, Lanping County, Nujiang Prepecture</td>
<td>20</td>
<td>99.1888</td>
<td>26.6593</td>
<td>2631</td>
</tr>
<tr>
<td></td>
<td>Shuiyanchang Village, Lanping County, Nujiang Prepecture</td>
<td>15</td>
<td>99.238</td>
<td>26.6662</td>
<td>2631</td>
</tr>
<tr>
<td></td>
<td>Shuiyanchang Village, Lanping County, Nujiang Prepecture</td>
<td>20</td>
<td>99.2049</td>
<td>26.6622</td>
<td>2631</td>
</tr>
<tr>
<td></td>
<td>Shenjiashan Village, Malong County, Qujian City</td>
<td>21</td>
<td>103.5711</td>
<td>25.4303</td>
<td>2045</td>
</tr>
<tr>
<td></td>
<td>Luzhi Town, Yimen County, Yuxi City</td>
<td>20</td>
<td>101.9724</td>
<td>24.695</td>
<td>1568</td>
</tr>
<tr>
<td></td>
<td>Lijiaao Village, Yimen County, Yuxi City</td>
<td>10</td>
<td>102.029</td>
<td>24.6454</td>
<td>1969</td>
</tr>
<tr>
<td></td>
<td>Lehei Mountain, Yuxi City</td>
<td>11</td>
<td>102.3706</td>
<td>24.3335</td>
<td>2192</td>
</tr>
<tr>
<td></td>
<td>Gaolushan Mountain, Yuxi City</td>
<td>10</td>
<td>102.3567</td>
<td>24.3115</td>
<td>2362</td>
</tr>
<tr>
<td></td>
<td>Laohe Village, Yuxi City</td>
<td>10</td>
<td>102.373</td>
<td>24.332</td>
<td>2192</td>
</tr>
</tbody>
</table>
Table 2 Performance comparison of 3DCOS-ResNet models built on different datasets.

<table>
<thead>
<tr>
<th>Types of data</th>
<th>Preprocessing methods/Methods for extraction</th>
<th>Types of 3DCOS</th>
<th>epoch</th>
<th>Train set</th>
<th>Test set</th>
<th>Loss value</th>
<th>Sen</th>
<th>Spe</th>
<th>Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>FT-NIR</td>
<td>RAW</td>
<td>Synchronous</td>
<td>57</td>
<td>100</td>
<td>100</td>
<td>0.048</td>
<td>97.42</td>
<td>99.47</td>
<td>99.12</td>
</tr>
<tr>
<td></td>
<td>Asynchronous</td>
<td>Synchronous</td>
<td>80</td>
<td>100</td>
<td>52</td>
<td>0.048</td>
<td>32.26</td>
<td>70.42</td>
<td>58.82</td>
</tr>
<tr>
<td></td>
<td>Asys</td>
<td>Synchronous</td>
<td>79</td>
<td>100</td>
<td>48</td>
<td>0.067</td>
<td>36.77</td>
<td>74.41</td>
<td>63.57</td>
</tr>
<tr>
<td>FT-MIR</td>
<td>RAW</td>
<td>Synchronous</td>
<td>58</td>
<td>100</td>
<td>95</td>
<td>0.048</td>
<td>89.68</td>
<td>98.12</td>
<td>96.19</td>
</tr>
<tr>
<td></td>
<td>Asynchronous</td>
<td>Synchronous</td>
<td>84</td>
<td>100</td>
<td>52</td>
<td>0.044</td>
<td>42.58</td>
<td>81.65</td>
<td>72.19</td>
</tr>
<tr>
<td></td>
<td>Asys</td>
<td>Synchronous</td>
<td>82</td>
<td>100</td>
<td>49</td>
<td>0.08</td>
<td>41.29</td>
<td>77.86</td>
<td>67.84</td>
</tr>
<tr>
<td>FT-NIR+FT-MIR</td>
<td>2D; VIP</td>
<td>Synchronous</td>
<td>89</td>
<td>100</td>
<td>100</td>
<td>0.081</td>
<td>93.55</td>
<td>98.64</td>
<td>97.75</td>
</tr>
<tr>
<td></td>
<td>Asynchronous</td>
<td>Synchronous</td>
<td>89</td>
<td>100</td>
<td>37</td>
<td>0.177</td>
<td>40</td>
<td>76.92</td>
<td>66.67</td>
</tr>
<tr>
<td></td>
<td>Asys</td>
<td>Synchronous</td>
<td>92</td>
<td>74</td>
<td>21</td>
<td>0.662</td>
<td>16.67</td>
<td>50</td>
<td>37.5</td>
</tr>
<tr>
<td>FT-NIR+FT-MIR</td>
<td>2D; VIP</td>
<td>Synchronous</td>
<td>104</td>
<td>100</td>
<td>89</td>
<td>0.047</td>
<td>58.06</td>
<td>87.38</td>
<td>80.60</td>
</tr>
<tr>
<td></td>
<td>Asynchronous</td>
<td>Synchronous</td>
<td>119</td>
<td>100</td>
<td>76</td>
<td>0.051</td>
<td>69.87</td>
<td>92.06</td>
<td>87.43</td>
</tr>
<tr>
<td></td>
<td>Asys</td>
<td>Synchronous</td>
<td>119</td>
<td>100</td>
<td>62</td>
<td>0.081</td>
<td>58.06</td>
<td>87.38</td>
<td>80.60</td>
</tr>
</tbody>
</table>

Notes: FT-NIR: Fourier transform near infrared; FT-MIR: Fourier transform near infrared; 3DCOS: three-dimensional correlation spectroscopy; 2D: second order derivative; VIP: Variable Importance in Projection; Sen: sensitivity; Spe: specificity; Acc: accuracy.
Figure 1. Sample (*Boletus bainiugan*) preparation process demonstration.

Figure 2. Score plots of PCA (a, b, c are the score plots of PCA under FT-NIR, FT-MIR and low-level data fusion, respectively).
Figure 3. A, B are the spectral images of FT-NIR and FT-MIR of the six origin samples, respectively.