

FT-MIR combined with 3DCOS-ResNet model for rapid identification of wild and cultivated *Gastrodia elata*

Liu Shuai

Yunnan Agricultural University

Honggao Liu

Yunnan Key Laboratory of Gastrodia and Fungi Symbiotic Biology, Zhaotong University, Zhaotong, Yunnan China

Jieqing Li

Yunnan Agricultural University

Yuanzhong Wang (✉ Boletus@126.com)

Medicinal Plants Research Institute, Yunnan Academy of Agricultural Sciences

Research Article

Keywords: *Gastrodia elata*, Chemometrics, Three-dimensional correlated spectral (3DCOS), Deep learning, Identification

Posted Date: April 27th, 2023

DOI: <https://doi.org/10.21203/rs.3.rs-2855573/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

FT-MIR combined with 3DCOS-ResNet model for rapid identification of wild and cultivated *Gastrodia elata*

Shuai Liu^{1,3}, Honggao Liu², Jieqing Li¹, and Yuanzhong Wang³

¹ College of Agronomy and Biotechnology, Yunnan Agricultural University, Kunming, 650201, China

² Yunnan Key Laboratory of Gastrodia and Fungi Symbiotic Biology, Zhaotong University, Zhaotong 657000, Yunnan China

³ Medicinal Plants Research Institute, Yunnan Academy of Agricultural Sciences, Kunming 650200, 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

Wild *Gastrodia elata* resources are in short supply and the market is dominated by cultivated *Gastrodia elata* resources and the medicinal value of both is very different. It is significant to find a highly accurate and stable technique to identify wild and cultivated *Gastrodia elata*. It could prevent market fraud and protect the rights of consumers. In this context, this study is the first to combine three-dimensional correlated spectral (3DCOS) images with deep learning to identify wild and cultivated *Gastrodia elata*. Also, partial least squares discrimination analysis (PLS-DA) and support vector machine (SVM) models are compared with this model. The PLS-DA and SVM models are built based on Fourier transform mid-infrared (FT-MIR) spectral data after nine different preprocessing. The PLS-DA model with second-order derivatives (2D) gives the best results when comparing the effects of the models with different preprocessing. the SVM model with parameters c , g in a reasonable range also gives satisfactory model results. The advantage of the deep learning model over them is that no processing of the original spectral data is required. With only 46 iterations, the accuracy of the model is stable at 100% for the training set, test set and external validation set. The excellent performance of the model allows it to be used as a technical reference to solve studies on the qualitative aspects of *Gastrodia elata*.

Keywords: *Gastrodia elata*, Chemometrics, Three-dimensional correlated spectral (3DCOS), Deep learning, Identification

1. Introduction

Gastrodia elata is a genus of plants in the orchid family Gastrodia, whose earliest use in Chinese history was recorded a thousand years ago. The dried and steamed root is used as a folk medicine and is known as "TianMa" (Zhao et al., 1999). It improves nervousness, relieves insomnia and headaches, has sedative, anti-inflammatory and sleeping effects (Wang et al., 2014; Qiu et al., 2014; Wang et al., 2014). The shortage of wild *Gastrodia elata* resources has caused the wild species to be extremely valuable. The shortage of wild resources in the market has led to the rapid development of the cultivated *Gastrodia elata* industry. Market cultivated *Gastrodia elata* accounts for a huge proportion of the market and is the main source of *Gastrodia elata*, but there are differences in the medicinal value of wild and cultivated *Gastrodia elata* in terms of medicinal use (Pei, 2022). The value of wild *Gastrodia elata* is much higher than cultivated, plus the shortage of resources. Its price is also high, which allows unscrupulous traders to see huge profits (Chen et al., 2021). It is difficult to identify cultivated and wild *Gastrodia elata* by the naked eye, and the traditional way of identification is by morphology, which is more subjective and difficult to ensure accuracy by empiricism (Liu, 2012). And for inexperienced consumers, it is difficult to distinguish between wild and cultivated *Gastrodia elata* differences. Illegal traders take advantage of this, the cultivation of *Gastrodia elata* packaged as wild *Gastrodia elata*, sold to consumers for profit. Therefore, the market and consumers urgently need a fast, accurate and effective technical method to identify wild and cultivated *Gastrodia elata*. In order to protect the rights and interests of consumers and businesses, and maintain

market order.

At present, the main methods for species identification are electronic nose, chromatography, mass spectrometry, spectroscopy, molecular marker technology and genome sequencing technology (Wei et al., 2022). Spectroscopy is an emerging technology that has gained a lot of recognition in recent years. The benefits are extremely obvious compared to other technical methods. The preparation of experimental samples is simple and has the advantages of being non-destructive, green, efficient and convenient. The operation of the spectroscopic instrument is simple and does not require a professional to operate; secondly, the process of analyzing the obtained spectral data is not complicated and does not require a lot of time to process the data. Fourier infrared spectroscopy (FTIR) combined with chemometrics, models built by deep learning have shown excellent results in qualitative and quantitative studies of species. Traditional chemometrics combined with infrared spectroscopy has solved a variety of qualitative and quantitative problems and achieved good results. However, pre-processing of spectral data is essential before building a chemometric model. Problems such as light scattering and baseline instability caused by the sample itself, the instrument, etc. need to be reduced. The accuracy and stability of the model can be improved by selecting the appropriate preprocessing method to minimize some adverse effects. This is a necessary step for traditional chemometrics combined with spectroscopic data to build a model (Chen et al., 2023; Zhang et al., 2021; Wang et al., 2022). There are many problems in the process of resolving one-dimensional spectra that are difficult to solve, such as overlapping of spectral information for large sample

sizes and difficulty in obtaining detailed information. In this case, two-dimensional correlation spectroscopy (2DCOS) images have emerged, which has the advantage of solving the problem of overlapping information in one-dimensional spectra and better displaying the spectral absorption peaks of the samples. At the same time, 2DCOS can express the intensity relationship of the spectral signal and improve the visibility of the spectral information (Noda, 2018). In solving the species characterization problem using 2DCOS combined with deep learning models, Yue et al. (2021) directly used the raw spectral data to generate 2DCOS images to build a residual convolutional neural network (ResNet) model with 100% accuracy and concluded that the accuracy of the model would not be improved using the preprocessing method. ResNet is the most widely used of the CNN networks, with the advantage of being able to eliminate the problem of gradient explosion and disappearance (Dong et al., 2023). Recently, Chen et al. (2022) used the 2DCOS-ResNet model to solve the origin tracing problem of a variety of wild *Boletus* species, and the classification accuracy of the model was satisfactory; Yan et al. (2022) successfully identified a variety of wild *Boletus* species by building a 2DCOS-ResNet model and compared it with the traditional chemometric model, demonstrating that the 2DCOS-ResNet model outperformed the traditional chemometric models; Chen et al. (2022) successfully identified different storage periods of wild *Boletus* using 2DCOS-ResNet model with 100% accuracy of the model. Li et al. (2022) proposed a 3DCOS combined with a deep learning model for the species identification problem of wild porcini mushrooms. The advantages of 3DCOS images are shown in the planar to spatial changes that make the spectral absorption peaks more

visible, and the deep learning models that give a more comprehensive access to the information of the images, making the performance of the models better.

In this paper, 3DCOS images are combined with deep learning models for the first time to solve the identification problem of wild and cultivated *Gastrodia elata*, and PLS-DA and SVM models are developed for comparison with deep learning models. Before building these two chemometric models, the raw spectral data were processed by first-order derivative (1D), 2D, multiplicative scattering correction (MSC), standard normal variate transformation (SNV) and Savitzky-Golay (SG) smoothing, and their combinations. To compare the effects of two chemometric models under different preprocessing methods. To build 3DCOS-ResNet models on the basis of raw spectral data, evaluate the model performance and compare with PLS-DA and SVM models. Through comparison, a most suitable identification model was found for the identification of wild and cultivated *Gastrodia elata* to provide a valuable technical reference for future practical application in the market.

2. Materials and methods

2.1 Collection and treatment of samples

All samples of *Gastrodia elata* were collected from Yunnan Province of China. There are 172 samples include 109 cultivated and 63 wild *Gastrodia elata*. Details information are shown at Table S1. The samples studied were certified by Wang Yuanzhong, a researcher at the Institute of Medicinal Plants, Yunnan Academy of Agricultural Sciences. After removing the soil and other impurities from the surface of the collected samples, they were rinsed with pure water and dried. Place all samples in

the oven at 60°C until they reach a constant weight and then remove. The samples were made into powder using a DFY-500 high-speed crusher (Wenling Linda Machinery Co., Ltd.), filtered through a 100-mesh stainless steel sieve tray and then stored in polyethylene bags at room temperature and protected from light. The flow chart for the preparation of the experimental samples is shown in Figure 1.

2.2 Acquisition and preprocessing of FT-MIR spectrum data

The samples were scanned using a FT-MIR spectrometer (PerkinElmer, USA), and the instrument was equipped with a detector and a diamond single reflection universal attenuated total reflection (ATR) sampling accessory. The used software is SpectrumTM 10 ES. The spectra were acquired in the range 4000-450 cm^{-1} with a resolution of 4 cm^{-1} . Each sample was scanned 3 times with each scan being 64 and the average was taken as the final spectrum. The original mean spectra of cultivated and wild *Gastrodia elata* are shown in Figure 2.

Spectroscopic instruments are subject to various interferences in the actual use process. The samples have different particle sizes and uneven distribution, as well as problems with light scattering and baseline instability (Robert & Gosselin, 2022). The result is that there is some redundant information or invalid information in the obtained spectral data. The absence of spectral data processing increases the time to train the model and reduces the model accuracy. Therefore, it is necessary to pre-process the spectral data before building the model. Yan et al. (2022) used a variety of preprocessing methods before building the model with the aim of minimizing such physical effects. The use of pretreatment methods can reduce the adverse effects of the physical

properties of the sample on the spectroscopic measurement process. MSC, SNV can reduce the scattering effect of the spectra; the use of derivatives can separate overlapping peaks; SG can smooth out spurious peaks (Caused by noise) (Kademi et al., 2019; Robert et al., 2022).

In this study, each sample has 1841 variables, and the Kennard Stone algorithm uses 70% of the total sample as the training set and 30% as the test set for subsequent model building. Data preprocessing is performed using 1D, 2D, MSC, SNV, SG (third-order derivatives with a window size of 15) and combinations of them.

2.3 Principal component analysis (PCA)

Principal component analysis, as an exploratory analysis technique, enables the classification of samples using information about their own characteristics. The visualization of the analysis results enables visualization of whether there are significant differences between the samples under study. For example, Da Costa Filho et al. (2023), in their study of adulteration of skim milk powder, used PCA to conclude that milk powder adulterated with lactose was similar to pure milk powder. Visualization of results by unsupervised PCA can point the way to subsequent experiments. In cases where it is difficult to differentiate between samples, researchers need to use supervised methods to solve the problem. The PCA score plots of cultivated and wild *Gastrodia elata* are shown in Figure 3.

2.4 PLS-DA and SVM

PLS-DA and SVM models, corresponding to linear and nonlinear problems, are frequently used in traditional chemometrics. PLS-DA is capable of solving

multicollinearity problems between variables. It is a supervised method to artificially classify and label the studied samples before model building. The building of model is based on an independent variable X matrix (spectral data) and a dependent variable Y matrix (classification labels) (Wang et al., 2022). SVM is a nonlinear supervised method that has excellent performance in solving classification problems (Dankowska & Kowalewski, 2019; Teye et al., 2014). It demonstrates some advantages in solving nonlinear and high-dimensional problems with small samples (Li et al., 2018).

These two models use the same preprocessing method before building, on the basis of which the model performance is compared. The performance comparison of the two models is shown by Table 3. The test set confusion matrix plot of the PLS-DA model is shown in Figure S1, and the training set hyperplane and test set classification results of the SVM model are shown in Figure S2.

2.5 Acquisition of 3DCOS images

The 3DCOS image categories generated using the software Matlab R2017 are divided into synchronous, asynchronous and integrated. There are 172 images in each category, for a total of 516 images. Of these, 84 images per category were used as the training set, 36 as the test set, and the remaining 52 images were used as the external validation set. The generated images are saved in JPEG format for use in the 3DCOS-ResNet model building. The spectral band of 4000-425 cm^{-1} utilized in this study was used for image generation, and secondly the spectral data used before image generation was not pre-processed. The reason for this is that there is no change in the performance of the model under preprocessing (Yue et al., 2021). The 3DCOS images are shown in

Figure 4.

$$S(v) = \begin{cases} s(v, t1) \\ s(v, t2) \\ \dots \\ s(v, tm) \end{cases} \quad (1)$$

The spectral intensity is represented by the column vector S , the variable is represented by V , the perturbation time interval is represented by t , and m denotes the number of times (Noda, 2014).

$$\Phi(v_l, v_2) = \frac{1}{m-1} S(v_l)^T \cdot S(v_2) \quad (2)$$

$$\Psi(v_l, v_2) = \frac{1}{m-1} S(v_l)^T \cdot N \cdot S(v_2) \quad (3)$$

$$I(v_l, v_2) = \Phi(v_l, v_2) \cdot \Psi(v_l, v_2) = \frac{1}{(m-1)^2} [S(v_l)^T \cdot S(v_2)] \cdot [S(v_l)^T \cdot N \cdot S(v_2)] \quad (4)$$

Φ , Ψ and I are the expressions for the synchronous, asynchronous and integrated three-dimensional correlation strengths, respectively.

$$\begin{cases} N_{jk} = 0, & j = k \\ N_{jk} = \frac{1}{\pi(k-j)}, & j \neq k \end{cases} \quad (5)$$

N is the expression in the case of equal or unequal j rows and k columns of the Hilber Noda matrix. The evaluation of the 3DCOS-ResNet model is shown in Table 4.

2.6 Evaluation methods of the models

After the model has been built, it is particularly important to evaluate the performance of the model. The goodness or badness of a model cannot be judged by the accuracy rate of the training and testing sets alone. The evaluation of a model should be comprehensive and scientific, and cannot be evaluated by only one metric. Therefore, this study evaluates the PLS-DA, SVM and 3DCOS-ResNet models with multiple metrics to comprehensively evaluate the overall effectiveness of the models.

The evaluation metrics of PLS-DA model are R^2 , Q^2 , root mean square error of evaluation (RMSEE), root mean square error of cross-validation (RMSECV), root mean square error of prediction (RMSEP) and accuracy of the model, respectively. The R^2 represents the fitting ability of the model and Q^2 represents the prediction ability of the model, and the closer the value is to 1, the better. The results of the model after 200 permutation tests are shown in Figure S3. The results of the evaluation metrics of the PLS-DA model are shown in Table 1.

The evaluation metrics of the SVM model are penalty parameter c , kernel parameter g , sensitivity (Sen), specificity (Spe), effectiveness (Eff), and accuracy (Acc). The values of parameters c and g are obtained by the grid search method. The optimal range of c is 2 to 2^{-2} and the range of g is 2×10^{-4} to 2×10^{-2} . The choice of parameters c and g for the SVM model is particularly important because the values of c and g vary considerably during the model building process. Within a reasonable range, the larger the value of c , the stronger the generalization ability of the model; the larger the value of g , the smaller the number of support vectors. Therefore, the size of g value indirectly

affects the training speed of the model. The results of the evaluation metrics of the SVM model are shown in Table 2.

The evaluation metrics of 3DCOS-ResNet model are loss value, number of iterations (epochs), Acc, Sen, Spe and Eff. The closer the loss value is to 0, the closer the other values are to 1 indicates that the model works better.

$$Acc = \frac{TP+TN}{TP+FP+TN+FN} \quad (6)$$

$$Sen = \frac{TP}{TP+FN} \quad (7)$$

$$Spe = \frac{TN}{TN+FP} \quad (8)$$

$$Eff = \sqrt{Sen \times Spe} \quad (9)$$

TP represents true positive samples; TN represents true negative samples; FP represents false positive samples; FN represents false negative samples. Sen is the ability of the model to identify samples from the target class, and Spe is the ability of the model to identify samples from the non-target class. Equation (9) is the relationship equation between the two, representing the efficiency of the model (Oliveri & Downey, 2012).

2.7 Introduction of the used software

PCA and PLS-DA model was built by the software SIMCA 14.1; SVM model was

built by the Matlab R2017a; 3DCOS images were generated by the software Matlab R2017a; 3DCOS-ResNet model was built by the software Spyder (Anaconda3).

3. Result and discussion

3.1 Chemical composition analysis based on FT-MIR spectral images

The raw average spectral images of wild and cultivated *Gastrodia elata* are shown in Figure 2. From the spectral images, the peak shapes of the two samples were basically the same, except for some differences in absorbance. From Figure 2, it was inferred that the content values of the chemical components were different between wild and cultivated *Gastrodia elata*. The specific wave number bands correspond to which chemical bonds need to be further analyzed. The C-O, C-C stretching and C-O-H bending at 1014 cm^{-1} may belong to the vibrational band of C-O on deoxyribose in DNA (Bureau et al., 2019). The absorption peak near 1220 cm^{-1} may be the phosphodiester functional group of DNA/RNA polysaccharide (Naumann, 2001). The absorption peak at 1240 cm^{-1} may be the PO_2^- of nucleic acids, phospholipids and phosphorylated proteins, caused by the asymmetric stretching vibration of phosphodiester PO_2^- (Bellisola & Sorio, 2012). The complex absorption band corresponding to $1300\text{-}1500\text{ cm}^{-1}$ is mainly caused by lipids, CH_2 and CH_3 bending modes of proteins and cyclic vibrations of nucleic acids (Naumann, 2001). The absorption peak at 1635 cm^{-1} is probably due to the C=O stretching of the amide I band (Bellisola et al., 2012). $3000\text{-}2800\text{ cm}^{-1}$ may be the spectral band corresponding to C-H of aliphatic or cyclic groups, and 2921 cm^{-1} contains aliphatic methylene bands, which are asymmetric stretches of CH_2 groups (Ribeiro et al., 2001; Rosa et al., 2015;

Sisouane et al., 2017). The chemical composition corresponding to the characteristic peak near 3276 cm^{-1} is the hydroxyl group and the N-H structure in the amide, and the absorption peak is related to the N-H and O-H stretching (Cao et al., 2022; Hell et al., 2016).

The entire spectral band was analyzed to decipher the specific chemical composition corresponding to the $4000\text{-}450\text{ cm}^{-1}$ range of wild and cultivated *Gastrodia elata*. Models were developed to distinguish between the two based on the differences in absorbance and characteristic fingerprint profiles.

3.2 PCA

As shown in Figure 3, PC1 and PC2 explained 77.1% and 19% of the total sample variance, respectively. Wild *Gastrodia elata* samples were mainly concentrated in the negative half-axis of PC1, and cultivated *Gastrodia elata* were mainly concentrated in the positive half-axis of PC1. However, there were still overlapping indistinguishable parts between the two samples, from which we can also conclude that there is some similarity between the two samples. This is reflected in Figure 2, where the spectral images show that the peak shapes of the two types of samples are similar, but there are differences in absorbance. Indirectly, the difference in the content of each chemical component between wild and cultivated *Gastrodia elata* was indicated, which was confirmed by Pei et al (2022). PCA, as an unsupervised chemometric method, relies on the sample's own characteristics that make it difficult to fully cluster two classes of samples. This is also a step of exploratory analysis to help us provide ideas for subsequent problem solving. Therefore, we later used a supervised chemometric

approach and deep learning to solve this problem.

3.3 Analysis and comparison of chemometric models

The PLS-DA model was built on the basis of the datasets obtained under nine different preprocessing methods, and the results are shown in Table 1. The results of the PLS-DA model under the nine pretreatments after 200 permutation tests are shown in Figure S3, and none of the models showed overfitting (the left-hand side measurements are all lower than the right-hand side true values, and the y-values of Q^2 are all less than zero). The PLS-DA model built under 2D treatment gave the best results, and the model under MSC treatment gave the worst results, followed by SNV. The model built based on the raw spectral data works satisfactorily instead, better than any other pre-processing methods except 2D. The reason may be that due to the relatively obvious differences existing between wild and cultivated *Gastrodia elata*, a satisfactory result can be achieved using a supervised linear analysis technique. In the choice of preprocessing methods, except for 2D, none of the preprocessing methods could reduce the three root mean square error values of the model, nor could they improve the R^2 and Q^2 of the model.

The SVM models were built in the same process as PLS-DA, all under nine different data sets. The results of the models are shown in Table 2, and the best results are finally obtained under 2D, MSC+2D and SG+2D. The parameters c and g of the model are within reasonable range. Although the c and g of the model built by the SNV+2D method were within a reasonable range, the accuracy of the test set was only 63.46%. The c and g values of the model built under other preprocessing are not within

a reasonable range. There is a risk that the model is overfitted or that too many support vectors reduce the model training speed.

Table 3 shows the comparison results of PLS-DA model and SVM model. Taking the best modeling results of SVM as a reference, only three methods, 2D, MSC+2D and SG+2D, are selected for comparison. Because the model effects of SNV+2D and MSC+2D are the same, only one of them is selected for comparison. The Sen, Spe, Acc and Eff of both models reach 100%. And the combined comparison of Tables 1, 2 and 3 concludes that the PLS-DA model gives better results. The accuracy of the test set is 100% regardless of the preprocessing method or the original data. Except for SNV and MSC, the model has good results for all the metrics under the other methods.

3.4 Deep learning model

The evaluation results of the three 3DCOS-ResNet models are shown in Table 4. In terms of the accuracy of the external validation set of the models, the synchronous-3DCOS-ResNet model gives the best results followed by the integrated and asynchronous models. The loss values, accuracy and external validation set results for the synchronous, asynchronous and integrated models are shown in Figure 5. When building the synchronous model, the performance of the model is unstable until the number of iterations is 32, and the test set accuracy fluctuates. After the number of iterations exceeds 35, the model's training set and test set accuracies stabilize at 100%. To stabilize the model performance, the epoch is finally set to 46. The model works better with 100% accuracy by external validation test. The asynchronous and integrated models were not satisfactory, and the test set accuracy still did not trend to a stable

value when the number of iterations reached 79. During this process, the accuracy of the training set was able to stabilize at 100%, but the accuracy of the test set fluctuated widely.

A comprehensive comparison shows that the synchronous-3DCOS-ResNet model works best. Firstly, it builds the model based on the original spectral data compared to PLS-DA and SVM, and does not require pre-processing of the data to achieve a satisfactory result; secondly, the loss value, number of iterations, accuracy, Sen, Spe and Eff of the model are at a satisfactory level. The 3DCOS-ResNet model was established for the first time for the identification of wild and cultivated *Gastrodia elata*. The performance of the model was evaluated by various indexes, and the preliminary judgment was that the model can play a fast and accurate identification effect.

4. Conclusion

In this study, 3DCOS combined with a deep learning model was applied for the first time to the identification of wild and cultivated *Gastrodia elata*, and a comparison of synchronous, asynchronous and integrated models yielded the best results for the synchronous-3DCOS-ResNet model. All evaluation metrics of the synchronous model are satisfactory, while the performance of the asynchronous and integrated models is unstable, and the accuracy of the test set fluctuates to a large extent that is difficult to stabilize. Before the establishment of PLS-DA and SVM models, the spectral data need to be preprocessed, and the evaluation index of the models can be improved under the appropriate preprocessing methods. However, the advantages are not significant compared with the 3DCOS-ResNet model. The synchronous-3DCOS-ResNet model

achieves a stable level of accuracy for the training and test sets at a lower number of iterations (46), and 100% accuracy for the external validation set. Mao et al. used Backed-propagation (BP) neural network technology to solve the shelf-life prediction problem for channel catfish fillets, and the number of iterations of the model was over 8700 to achieve such results (Mao et al., 2023). At the same time, no pre-processing of spectral data is required for building deep learning models. This saves a lot of time cost for studies with large samples and makes the advantages of the model more obvious. Thus, the overall results of the synchronous-3DCOS-ResNet model are satisfactory as concluded from the results of this paper. It is able to solve the identification problem of wild and cultivated *Gastrodia elata*, but the ability of asynchronous and integrated deep learning models needs to be improved and further experiments are needed to optimize the performance of the model.

Conflicts of interest

There is no conflict of interest between the authors throughout this study.

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:

- Bellisola, G., & Sorio, C. (2012). Infrared spectroscopy and microscopy in cancer research and diagnosis. *American journal of cancer research*, 2(1), 1.
- Bureau, S., Cozzolino, D., & Clark, C. J. (2019). Contributions of Fourier-transform mid infrared (FT-MIR) spectroscopy to the study of fruit and vegetables: A review. *Postharvest Biology and Technology*, 148, 1-14.
<https://doi.org/https://doi.org/10.1016/j.postharvbio.2018.10.003>.
- Cao, H., Gu, Y., Fang, J., Hu, Y., Ding, W., He, H., & Chen, G. (2022). Application of stacking ensemble learning model in quantitative analysis of biomaterial activity. *Microchemical Journal*, 183, 108075.
<https://doi.org/https://doi.org/10.1016/j.microc.2022.108075>.
- Chen, H., Tan, C., & Li, H. (2021). Discrimination between wild-grown and cultivated *Gastrodia elata* by near-infrared spectroscopy and chemometrics. *Vibrational Spectroscopy*, 113, 103203.
<https://doi.org/https://doi.org/10.1016/j.vibspec.2020.103203>.
- Chen, J., Li, J. Q., Li, T., Liu, H. G., & Wang, Y. Z. (2022). Rapid identification of the storage duration and species of sliced boletes using near-infrared spectroscopy. *Journal of Food Science*, 87(7), 2908-2919.
<https://doi.org/https://doi.org/10.1111/1750-3841.16220>.
- Chen, J., Liu, H. G., Li, T., & Wang, Y. Z. (2023). Edibility and species discrimination of wild bolete mushrooms using FT-NIR spectroscopy combined with DD-SIMCA and RF models. *LWT*, 114701.

<https://doi.org/https://doi.org/10.1016/j.lwt.2023.114701>.

Chen, X., Liu, H. G., Li, J. Q., & Wang, Y. Z. (2022). A geographical traceability method for *Lanmaoa asiatica* mushrooms from 20 township-level geographical origins by near infrared spectroscopy and ResNet image analysis techniques. *Ecological Informatics*, 71, 101808.

<https://doi.org/https://doi.org/10.1016/j.ecoinf.2022.101808>.

Da Costa Filho, P. A., Chen, Y., Cavin, C., & Galluzzo, R. (2022). Mid-infrared spectroscopy: Screening method for analysis of food adulterants in reconstituted skimmed milk powder. *Food Control*, 136, 108884.

<https://doi.org/https://doi.org/10.1016/j.foodcont.2022.108884>.

Dankowska, A., & Kowalewski, W. (2019). Tea types classification with data fusion of UV–Vis, synchronous fluorescence and NIR spectroscopies and chemometric analysis. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 211, 195-202.

<https://doi.org/https://doi.org/10.1016/j.saa.2018.11.063>.

Dong, J. E., Li, J. Q., Liu, H. G., & Wang, Y. Z. (2023). Machine learning and deep learning based on the small FT-MIR dataset for fine-grained sampling site recognition of *boletus tomentipes*. *Food Research International*, 167, 112679.

<https://doi.org/https://doi.org/10.1016/j.foodres.2023.112679>.

Hell, J., Prückler, M., Danner, L., Henniges, U., Apprich, S., Rosenau, T., & Böhmendorfer, S. (2016). A comparison between near-infrared (NIR) and mid-infrared (ATR-FTIR) spectroscopy for the multivariate determination of

- compositional properties in wheat bran samples. *Food Control*, 60, 365-369.
<https://doi.org/https://doi.org/10.1016/j.foodcont.2015.08.003>.
- Kademi, H. I., Ulusoy, B. H., & Hecer, C. (2019). Applications of miniaturized and portable near infrared spectroscopy (NIRS) for inspection and control of meat and meat products. *Food Reviews International*, 35(3), 201-220.
- Li, J. Q., Wang, Y. Z., & Liu, H. G. (2022). Application of spectral image processing with different dimensions combined with large-screen visualization in the identification of boletes species. *Frontiers in Microbiology*, 13. <https://doi.org/10.3389/fmicb.2022.1036527>.
- Li, Y., Li, F., Yang, X., Guo, L., Huang, F., Chen, Z., & Zheng, S. (2018). Quantitative analysis of glycated albumin in serum based on ATR-FTIR spectrum combined with SiPLS and SVM. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 201, 249-257.
<https://doi.org/https://doi.org/10.1016/j.saa.2018.05.022>.
- Mao, S., Zhou, J., Hao, M., Ding, A., Li, X., Wu, W., & Shi, L. (2023). BP neural network to predict shelf life of channel catfish fillets based on near infrared transmittance (NIT) spectroscopy. *Food Packaging and Shelf Life*, 35, 101025.
<https://doi.org/https://doi.org/10.1016/j.fpsl.2023.101025>.
- Minghai, L. (2012). Identification and quality comparison of main characters between wild *Gastrodia elata* and planted *Gastrodia elata*. *China Pharmaceuticals* 21(06), 70-71.
- Naumann, D. (2001). FT-infrared and FT-Raman spectroscopy in biomedical research.

- Applied spectroscopy reviews*, 36(2-3), 239-298.
<https://doi.org/https://doi.org/10.1081/ASR-100106157>.
- Noda, I. (2014). Frontiers of two-dimensional correlation spectroscopy. Part 1. New concepts and noteworthy developments. *Journal of Molecular Structure*, 1069, 3-22. <https://doi.org/https://doi.org/10.1016/j.molstruc.2014.01.025>.
- Noda, I. (2018). Two-dimensional correlation and codistribution spectroscopy (2DCOS and 2DCDS) analyses of time-dependent ATR IR spectra of d-glucose anomers undergoing mutarotation process in water. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 197, 4-9.
<https://doi.org/https://doi.org/10.1016/j.saa.2017.09.052>.
- Oliveri, P., & Downey, G. (2012). Multivariate class modeling for the verification of food-authenticity claims. *TrAC Trends in Analytical Chemistry*, 35, 74-86.
<https://doi.org/https://doi.org/10.1016/j.trac.2012.02.005>.
- Pei, W. H., Li, S. Y., Song, H. J., Yu, X. J., & Zhang, H. (2022). Content Determination of Nine Components in Gastrodiae Rhizoma and Analysis of Different Components Between Wild and Cultivated Species. *Chin J Mod Appl Pharm*, 39(18), 2347-2353. <https://doi.org/10.13748/j.cnki.issn1007-7693.2022.18.008>.
- Qiu, F., Liu, T. T., Qu, Z. W., Qiu, C. Y., Yang, Z., & Hu, W. P. (2014). Gastrodin inhibits the activity of acid-sensing ion channels in rat primary sensory neurons. *European Journal of Pharmacology*, 731, 50-57.
<https://doi.org/https://doi.org/10.1016/j.ejphar.2014.02.044>.
- Ribeiro, J. S., Ok, S. S., Garrigues, S., & De la Guardia, M. (2001). FTIR tentative

- characterization of humic acids extracted from organic materials. *Spectroscopy letters*, 34(2), 179-190. <https://doi.org/https://doi.org/10.1081/SL-100002007>.
- Robert, G., & Gosselin, R. (2022). Evaluating the impact of NIR pre-processing methods via multiblock partial least-squares. *Analytica chimica acta*, 1189, 339255. <https://doi.org/https://doi.org/10.1016/j.aca.2021.339255>.
- Rosa, F., Sales, K. C., Cunha, B. R., Couto, A., Lopes, M. B., & Calado, C. R. (2015). A comprehensive high-throughput FTIR spectroscopy-based method for evaluating the transfection event: estimating the transfection efficiency and extracting associated metabolic responses. *Analytical and bioanalytical chemistry*, 407, 8097-8108.
- Sisouane, M., Cascant, M., Tahiri, S., Garrigues, S., Krati, M. E., Boutchich, G. E. K., & de La Guardia, M. (2017). Prediction of organic carbon and total nitrogen contents in organic wastes and their composts by infrared spectroscopy and partial least square regression. *Talanta*, 167, 352-358. <https://doi.org/https://doi.org/10.1016/j.talanta.2017.02.034>.
- Teye, E., Huang, X., Takrama, J., & Haiyang, G. (2014). Integrating NIR spectroscopy and electronic tongue together with chemometric analysis for accurate classification of cocoa bean varieties. *Journal of Food Process Engineering*, 37(6), 560-566. <https://doi.org/https://doi.org/10.1111/jfpe.12109>.
- Wang, H. Y., Chen, Q., & Li, J. (2014). Comparision of Different Method for Sample Preparation and Extraction of Gastradin in Gastradia elata Bl . . *Journal of Xinyang College of Agriculture and Forestry*, 24(03), 127-130.

<https://doi.org/10.16593/j.cnki.41-1433/s.2014.03.041>.

Wang, H., Zhang, R., Qiao, Y., Xue, F., Nie, H., Zhang, Z., & Tan, Q. (2014). Gastrodin ameliorates depression-like behaviors and up-regulates proliferation of hippocampal-derived neural stem cells in rats: involvement of its anti-inflammatory action. *Behavioural Brain Research*, 266, 153-160. <https://doi.org/https://doi.org/10.1016/j.bbr.2014.02.046>.

Wang, L., Liu, H. G., Li, T., Li, J. Q., & Wang, Y. Z. (2022). Verified the rapid evaluation of the edible safety of wild porcini mushrooms, using deep learning and PLS-DA. *Journal of the Science of Food and Agriculture*, 102(4), 1531-1539. <https://doi.org/https://doi.org/10.1002/jsfa.11488>.

Wei, Y., Li, L., Liu, Y., Xiang, S., Zhang, H., Yi, L., & Xu, W. (2022). Identification techniques and detection methods of edible fungi species. *Food Chemistry*, 374, 131803.

Yan, Z. Y., Liu, H. G., Li, T., Li, J. Q., & Wang, Y. Z. (2022). Two dimensional correlation spectroscopy combined with ResNet: Efficient method to identify bolete species compared to traditional machine learning. *LWT*, 162, 113490. <https://doi.org/https://doi.org/10.1016/j.lwt.2022.113490>.

Yue, J. Q., Huang, H. Y., & Wang, Y. Z. (2021). A practical method superior to traditional spectral identification: Two-dimensional correlation spectroscopy combined with deep learning to identify Paris species. *Microchemical Journal*, 160, 105731. <https://doi.org/https://doi.org/10.1016/j.microc.2020.105731>.

Zhang, J., Wang, Y. Z., Yang W. Z., & Zhang, J. Y. (2021). Data Fusion of ATR-FTIR

and UV-Vis Spectra to Identify the Origin of Polygonatum Kingianum.
Spectroscopy and Spectral Analysis, 41(05), 1410-1416. [https://doi.org/10 .
3964k / j . issn 1000-0593\(2021\)05-1410-07](https://doi.org/10.3964k/j.issn1000-0593(2021)05-1410-07).

Zhao, Y., Cao, Q. E., Xiang, Y., & Hu, Z. (1999). Identification and determination of active components in *Gastrodia elata* Bl. by capillary electrophoresis. *Journal of Chromatography A*, 849(1), 277-283.
[https://doi.org/https://doi.org/10.1016/S0021-9673\(99\)00534-8](https://doi.org/https://doi.org/10.1016/S0021-9673(99)00534-8).

Table 1 The results of each evaluation index of PLS-DA model.

Model	Preprocessing	Data set	Number	Acc of	Acc	RMSEE	RMSECV	RMSEP	R ²	Q ²
				train set	of test set					
PLS-DA	RAW	Train set	Cultivated (76) Wild (44)	100%	100%	0.0766	0.1777	0.1251	0.9777	0.8918
	1D			100%	100%	0.1009	0.1788	0.1745	0.9591	0.8498
	2D			100%	100%	0.0959	0.1669	0.1289	0.9624	0.9069
	SNV			99.17%	100%	0.1725	0.3135	0.1638	0.8836	0.6046
	MSC			99.17%	100%	0.1828	0.3206	0.1778	0.8681	0.5854
	SG			100%	100%	0.1571	0.24703	0.1664	0.9035	0.7608
	SNV+2D			100%	100%	0.1088	0.1829	0.1278	0.9511	0.8672
	MSC+2D			100%	100%	0.1089	0.1829	0.1278	0.9511	0.8671
	SG+2D			100%	100%	0.0931	0.1723	0.1489	0.9649	0.8904

Notes: PLS-DA: partial least squares discrimination analysis; 1D: first-order derivative; 2D: second-order derivative; SNV: standard normal variate transformation; MSC: multiplicative scattering correction; SG: Savitzky-Golay; Acc: accuracy; RMSEE: root mean square error of evaluation; RMSECV: root mean square error of cross-validation; RMSEP: root mean square error of prediction.

Table 2 The results of each evaluation index of SVM model.

Model	Preprocessing	c	g	Acc of train set (120)	Acc of test set (52)
SVM	RAW	8.19×10^3	6.10×10^{-5}	98.33%	63.46%
	1D	1.81×10^2	3.45×10^{-4}	99.17%	100%
	2D	2	7.81×10^{-3}	100%	100%
	SNV	3.46×10^5	8.76×10^{-6}	97.5%	63.46%
	MSC	5.24×10^5	2.70×10^{-6}	98.33%	63.46%
	SG	8.19×10^3	1.22×10^{-4}	98.33%	100%
	SNV+2D	5.66	3.91×10^{-3}	100%	100%
	MSC+2D	5.66	3.91×10^{-3}	100%	100%
	SG+2D	2.83	7.81×10^{-3}	100%	100%

Notes: SVM: support vector machine; 1D: first-order derivative; 2D: second-order derivative; SNV: standard normal variate transformation; MSC: multiplicative scattering correction; SG: Savitzky-Golay; Acc: accuracy; c : penalty parameter; g : nuclear parameter.

Table 3 The results of comparing each evaluation index of PLS-DA and SVM models.

Models	Preprocessing	Sen	Spe	Eff	Acc
PLS-DA	2D	1	1	1	1
	MSC+2D	1	1	1	1
	SG+2D	1	1	1	1
SVM	2D	1	1	1	1
	MSC+2D	1	1	1	1
	SG+2D	1	1	1	1

Notes: PLS-DA: partial least squares discrimination analysis; SVM: support vector machine; 2D: second-order derivative; MSC: multiplicative scattering correction; SG: Savitzky-Golay; Sen: sensitivity; Spe: specificity; Eff: effectivity; Acc: accuracy.

Table 4 Comparison of 3DCOS-ResNet models.

Model	Classifications	Epochs	Acc of train set	Acc of test set	Acc of EV	Loss value	Sen	Spe	Eff
3DCOS- ResNet	Synchronous	46	100%	100%	100%	0.005	1	1	1
	Asynchronous	79	100%	61%	63.46%	0.027	63.46%	63.46%	63.46%
	Integrated	79	100%	86%	88.46%	0.007	88.46%	88.46%	88.46%

Notes: 3DCOS-ResNet: three-dimensional correlated spectral images combined with residual convolutional neural networks; Acc: accuracy; Sen: sensitivity; Spe: specificity; Eff: effectivity; EV: external verification.

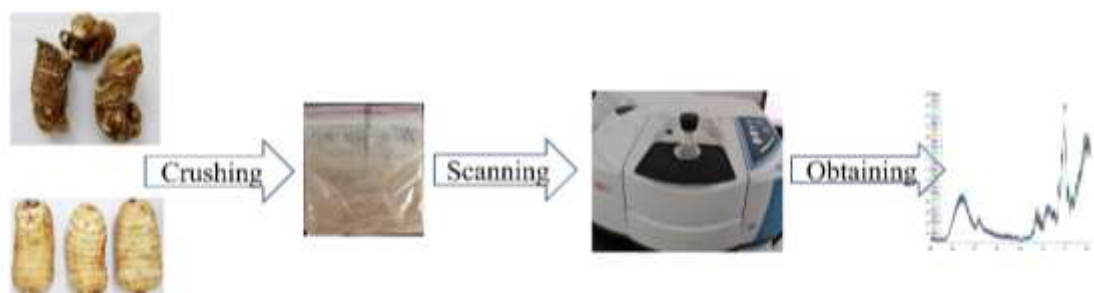


Figure 1. Flow chart of sample preparation.

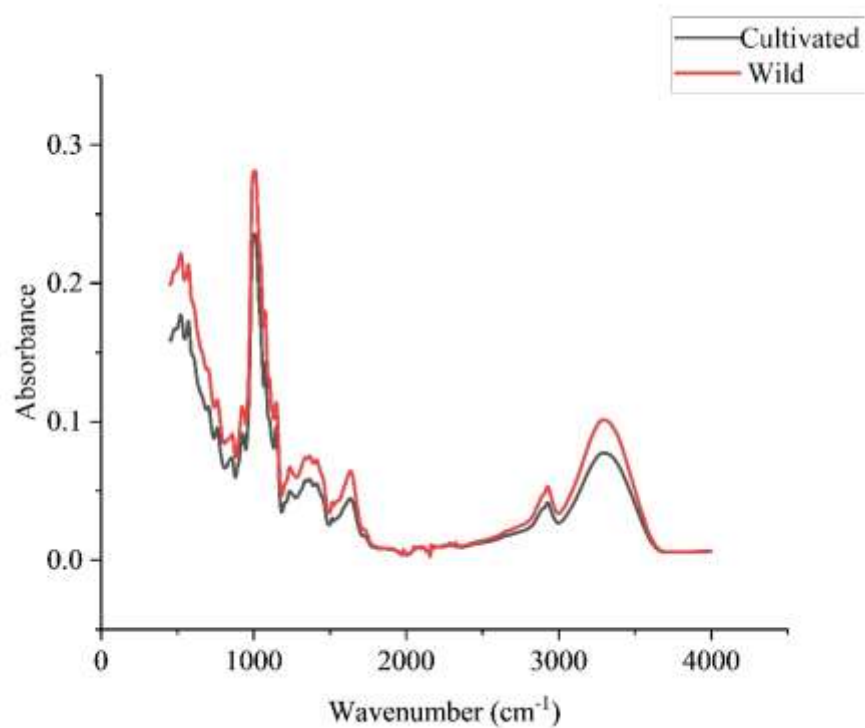


Figure 2. Raw mean spectral images of wild and cultivated *Gastrodia elata*.

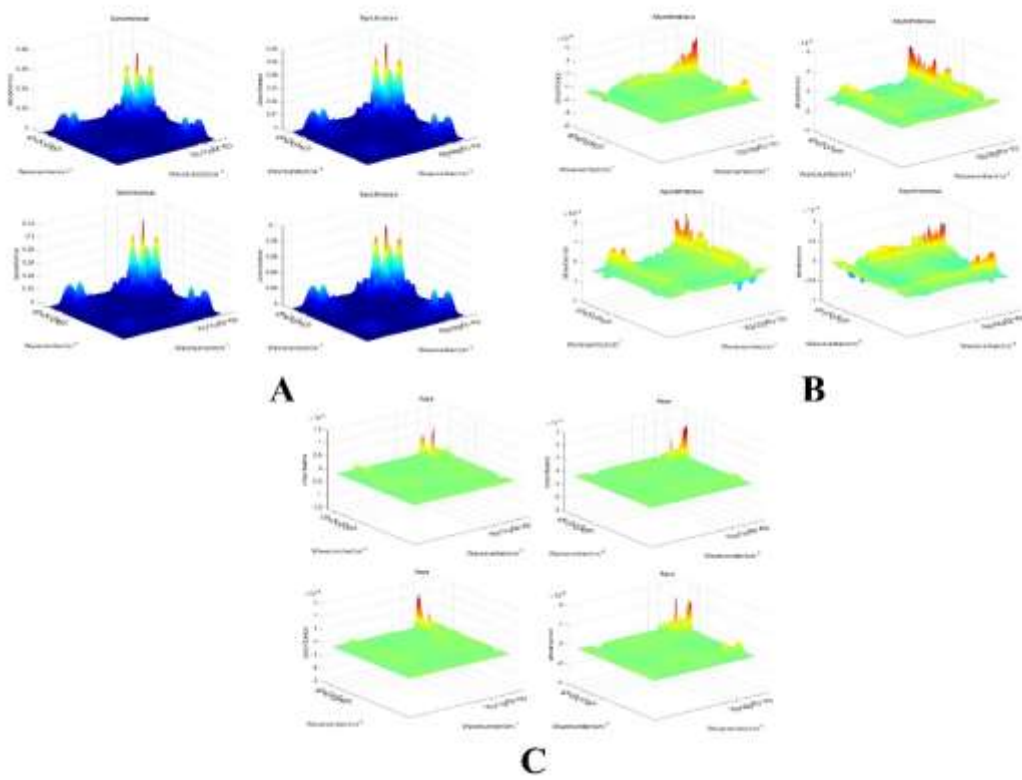


Figure 4. Three-dimensional correlation spectral images (A: synchronous three-dimensional spectral images; B: asynchronous three-dimensional spectral images; C: integrated three-dimensional spectral images).

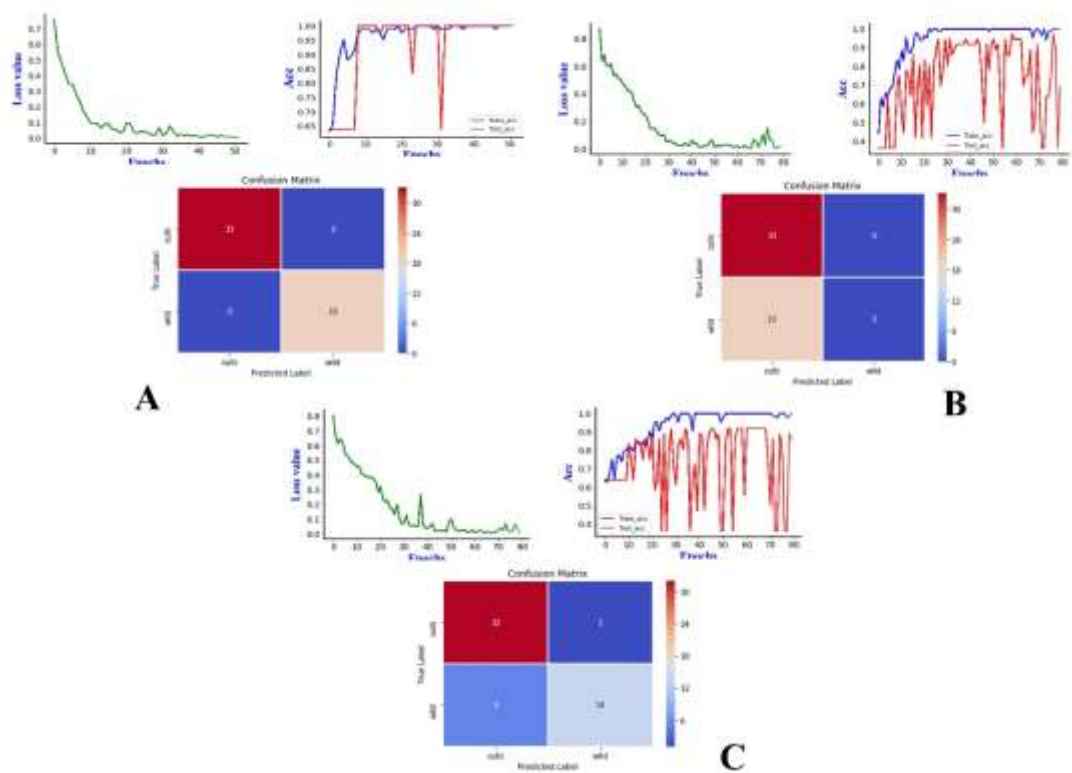


Figure 5. Plots of results for the training set, test set, loss values, and external validation set for the three models (A: synchronous model; B: asynchronous model; C: integrated model).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supplementarymaterials.docx](#)