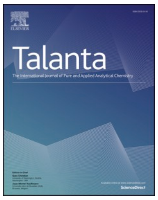
Contents lists available at ScienceDirect

Talanta

journal homepage: www.elsevier.com/locate/talanta



Identification of *Panax notoginseng* origin using terahertz precision spectroscopy and neural network algorithm

Hongyu Gu [a,1], Shengfeng Wang [a,1], Songyan Hu [a], Xu Wu [a], Qiuye Li [b], Rongrong Zhang [b], Juan Zhang [b], Wenbin Zhang [c], Yan Peng [a,*]

[a] University of Shanghai for Science and Technology, Terahertz Technology Innovation Research Institute, Shanghai Key Lab of Modern Optical System, Shanghai Institute of Intelligent Science and Technology, Shanghai, 200093, China

[b] Wenshan Institute for Food and Drug Control, Yunnan, 663099, China

[c] Wenshan Sanqi Institute of Science and Technology, China

ARTICLE INFO

Keywords:
Convolutional neural networks
Terahertz spectra
Panax notoginseng origin
Permutation variable importance

ABSTRACT

Panax notoginseng (*P. notoginseng*), a Chinese herb containing various saponins, benefits immune system in medicines development, which from Wenshan (authentic cultivation) is often counterfeited by others for large demand and limited supply. Here, we proposed a method for identifying *P. notoginseng* origin combining terahertz (THz) precision spectroscopy and neural network. Based on the comparative analysis of four qualitative identification methods, we chose high-performance liquid chromatography (HPLC) and THz spectroscopy to detect 252 samples from five origins. After classifications using Convolutional Neural Networks (CNNs) model, we found that the performance of THz spectra was superior to that of HPLC. The underlying mechanism is that there are clear nonlinear relations among the THz spectra and the origins due to the wide spectra and multi-parameter characteristics, which makes the accuracy of five-classification origin identification up to 97.62%. This study realizes the rapid, non-destructive and accurate identification of *P. notoginseng* origin, providing a practical reference for herbal medicine.

1. Introduction

Panax notoginseng (*P. notoginseng*) is a perennial herb of the family Araliaceae, mainly produced in Yunnan Province, China [1]. As a traditional precious Chinese medicinal material, *P. notoginseng* exerts significant positive effects on the human cardiovascular and immune systems [2,3]. *P. notoginseng* contains various bioactive components, including saponins [4–7], flavonoids [8], amino acids [3], polysaccharides [9], fatty acids [10], peptides [3], etc. Among them, saponins are the main active constituents [11], with ginsenoside Rb_1 (content 30–36% of saponins), ginsenoside Rg_1 (content 20–40% of saponins), and notoginsenoside R_1 (content 7–10% of saponins) being the standardized compounds used for evaluating the quality of *P. notoginseng*. Modern pharmacological research validates the anticancer, antitumor, anti-inflammatory, and hypoglycemic effects of saponins [12–15]. The accumulation of saponins in *P. notoginseng* is highly susceptible to geographical and climatic influences, demanding stringent temperature, humidity, and soil requirements. Therefore, *P. notoginseng* cultivated in diverse geographical environments exhibits varying quality attributes. Notably, the saponins content of *P. notoginseng* produced in Wenshan Autonomous Prefecture (Wenshan) of Yunnan Province is much higher than that in Kunming and Yuxi city [16]. Therefore, the *P. Notoginseng* produced in Wenshan is also called the geological authentic cultivation (Daodi in Chinese). However, due to problems such as continuous cropping obstacles [17] and rampant diseases [18], the yield of *P. Notoginseng* in Wenshan was insufficient, which led to the problem of origin falsification.

Traditional methods for *P. notoginseng* identification mainly involve morphological and microscopic identification. Morphological identification relies on visual observation, tactile examination, taste, and olfactory perception to discern the shape, size, surface features, and aroma of the herb. This method requires substantial experience and is limited to the original state of herb. However, when the herb is processed into slices or powders, they lose essential morphological characteristics, limiting the identification capability of this method [19]. Microscopic identification is to analyze and identify the tissue structure, cell

[*] Corresponding author.
E-mail address: py@usst.edu.cn (Y. Peng).

[1] Co-first author.

https://doi.org/10.1016/j.talanta.2024.125968
Received 9 January 2024; Received in revised form 11 March 2024; Accepted 20 March 2024
Available online 31 March 2024
0039-9140/© 2024 Elsevier B.V. All rights reserved.

morphology, and intracellular components of herb. It is suitable for identifying powdered materials. However, microscopic features of herbs within the same genus can be similar, making it difficult to identify them [20]. In modern times, high performance liquid chromatography (HPLC) is often used for qualitative identification and quantitative analysis of herbal medicine [21]. HPLC method has very high accuracy, but it also has many shortcomings such as lengthy analysis time, high equipment cost, complex operation, frequent maintenance requirements. Zhou et al. [22] were able to identify the origin of *P. notoginseng* samples by combining Fourier transform mid-infrared (FT-MIR) and near infrared (NIR) spectroscopy. Although the origin identification results could reach 95.6%, it required the use of two spectroscopic techniques, which were limited to the detection of surface substances. Zhang et al. [23] employed terahertz technology and the whale optimization algorithm to identify the production origins of *P. notoginseng*. Although they achieved an accuracy rate of 98.44%, the samples were collected from four provinces (Guizhou, Hunan, Guangxi and Yunnan) in China, and the analysis was focused on the low-frequency band, resulting in limited geographical precision and reference. The rapid and precise identification of the origin of *P. notoginseng* has always been a difficulty in this field. Hence, researchers have been actively seeking a more efficient solution.

Terahertz (THz) waves refer to electromagnetic waves with a wavelength of 0.3–3.0 mm. Compared with other wavebands, THz waves offer numerous advantages for substance detection. 1) Fingerprint spectroscopy: When THz waves are transmitted through biochemical substances, resonance absorption occurs if the molecular vibration and rotation frequencies coincide with certain terahertz frequencies, forming a unique fingerprint spectrum. This characteristic has made significant achievements in the field of biomedical detection [24–35]. 2) Non-destructive detection: operating at millielectron volt energy levels, THz waves do not induce ionization or disrupt the intrinsic properties of the detected substances, making them ideal for physical probing [36]. 3) Transient nature: typical THz pulses exhibit picosecond-level pulse widths, enabling the display of material changes and dynamic processes at the picosecond level [37]. Therefore, compared to conventional detection methods, THz spectroscopy has great potential in accurately identifying the properties and origin of *P. notoginseng*.

In this paper, we aimed to achieve accurate and interpretable identification of *P. notoginseng* from different origins using terahertz precision spectroscopy combined with neural network algorithms. Firstly, HPLC, ultraviolet (UV), Raman and terahertz spectroscopy were compared and analyzed for the identification of *P. notoginseng* saponins. In addition, we performed HPLC and THz detection on 252 *P. notoginseng* samples from five different origins and constructed a Convolutional Neural Networks (CNNs) model to train the data. Binary classification of *P. notoginseng* from Wenshan and other origins was performed based on HPLC and THz spectra, and then five-classification of *P. notoginseng* from five origins (Wenshan, Chuxiong, Honghe, Kunming and Qujing) in Yunnan was performed. Furthermore, we analyzed the importance of frequency bands using the Permutation Variable Importance (PVI) method to find the nonlinear relations among THz spectra and P. notoginseng origins, and then extracted additional identification information from the THz spectra into CNNs, constructing the Feature Enhancement Convolutional Neural Networks (FE-CNNs) model for five-classification origin identification.

2. Materials and methods

2.1. Experimental materials

We purchased Ginsenoside R_1 (>98%, CAS: 80,418-24-2), ginsenoside Rb_1 (>98%, CAS: 41,753-43-9) and ginsenoside Rg_1 (>98%, CAS: 22,427-39-0) from PureChem Standard in Chengdu, China, and Cyclic Olefin Copolymer (COC) powder from *Sigma-Aldrich* in Shanghai, China. In addition, we used 8 batches of different *P. notoginseng* produced in Wenshan Autonomous Prefecture, Chuxiong City, Kunming City, Qujing City, and Honghe Autonomous Prefecture in Yunnan Province, China, which were used in the form of block roots. No further purification was performed on the samples.

2.2. Sample preparation

For the THz spectroscopy analysis, we used the tablet pressing method, which needs grinding and sieving before tablet pressing. Firstly, *P. notoginseng* samples were ground into powder for 3 min at a vibration frequency of 90 Hz by MM400 ball mill (*Retsch*, Germany). The powdered samples with a particle size of 40 μm were kept dry under an infrared lamp. After being sieved, they were mixed in an agate vessel with COC powder, which has extremely low loss for THz frequency signals [38]. Samples were then pressed into 0.7 mm thick and 13 mm diameter tablets by a tablet machine with 8 tons of pressure.

For the HPLC Analysis, the dried *P. notoginseng* powder (0.6 g) was added to a 10 mL round-bottomed flask followed by 50 mL methanol (99.99%, CAS: 67-56-1). The mixture then was incubated overnight. The sample was refluxed in a water bath at 80 °C for 2 h, cooled, weighed, and made up the lost weight with methanol. The mixture was then shaken and strained to get the filtrate.

For UV spectroscopy and Raman spectroscopy, the milled sample powder was mixed with COC and pressed into 0.7 mm thick and 13 mm diameter tablets by a tablet machine with 8 tons of pressure.

2.3. Experimental instruments

The equipment for THz spectroscopy analysis is Fourier transform infrared spectrometer (FTIR vectex80v, *Bruker Optics*). The detector is a deuterated L-alanine triglycine sulfate detector. The far-infrared (IR) light source is a self-cooled mercury lamp. The effective coverage of the spectral region is 30–680 cm^{-1}, whose range is 1.5–16 THz and signal-noise rate is better than 10,000: 1. The resolution is 2 cm^{-1}, the scanning times and speed is 128 s and 5 kHz. In order to minimize the impact of water vapor during the experiment, the measurement of all spectra is conducted under vacuum conditions at room temperature (~22 °C).

The equipment for HPLC analysis is Agilent 1200 high-performance liquid chromatography (MA, USA). The reagents used is methanol (99.99%, CAS: 67-56-1), ultrapure water (Milli-Q50 SP pure water system) and others (analytic pure).

The equipment for Ultraviolet–Visible spectroscopy (UV-VIS) analysis is recorded using a Lambda 1050 spectrophotometer (PerkinElmer, USA). The range and the accuracy of wavelength is 200 nm–400 nm and ±1 nm.

The equipment for Raman analysis is LabRAM HR Evolution, (HORIBA, Japan), The wavelength of the laser source is 532 nm. The spectral region is 200–1800 cm^{-1}.

2.4. Data processing and neural network

2.4.1. Data processing

To bolster the network's generalization ability and achieve the precise classification of spectra collected in practical scenarios, we applied ten distinct sets of noise to each sample in the input spectrum data, creating 2520 new spectra from the original 252 samples. This approach aims to train the model to withstand noise, intensify the complexity of network training, and induce a certain regularization effect.

Subsequently, 2520 spectra are randomly scrambled as a training set, with 10% as a validation set. The training and validation datasets are independent and drawn from the same distribution.

By employing this methodology, our model is poised to exhibit greater robustness and improved generalization, enabling it to effectively handle noisy spectra commonly encountered in practical scenarios.

2.4.2. Z-score standardization

In order to address the issue of varying numerical ranges and units among different features, which may lead to certain features dominating the model training process, we employ the z-score normalization technique on the input data. This normalization method allows us to bring all features to a similar scale, thereby avoiding undue influence from specific features and facilitating a more comprehensive exploration of inter-feature relations.

The z-score normalization involves subtracting the mean and dividing by the standard deviation for each data point along rows, columns, or other attributes. The resulting transformation ensures that all data for each attribute or column is centered around zero and has a variance of 1. The input spectra $x_1, x_2, \ldots, x_n$ are subjected to this transformation as shown in the following equation:

$$\mathrm{y_i} = \frac{\mathrm{x_i\text{-}\overline{x}}}{\mathrm{s}} \tag{1}$$

In this equation $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$, $y_1, y_2, \ldots, y_n$ represent the standardized spectra after normalization. The resulting transformation ensures that all data for each attribute or column is centered around zero and has a variance of 1. This normalization step is instrumental in promoting a more balanced and effective learning of the interrelations among features, ultimately leading to improved model performance and better generalization capabilities.

2.4.3. Convolutional Neural Networks

CNNs have gained extensive utilization across diverse domains, including data classification, image recognition, and natural language processing, rendering them an indispensable instrument [39–44]. Their salient attribute lies in their capacity to learn and extract features from input data, which has culminated in remarkable achievements in tasks such as classification and regression. Here, we use 1 dCNN architecture to classify *P. notoginseng* Origin.

2.4.4. Permutation variable importance

PVI [45] is a feature selection technique employed to evaluate the significance of each feature in a model. Its underlying concept involves randomizing the values of each feature and subsequently assessing the model's performance based on the shuffled feature set. When shuffling a feature's values results in an increase in model error, the feature is deemed "important.' This approach aids in quantifying the impact of each feature on the model's performance, enabling the elimination of irrelevant features and facilitating the development of interpretable neural networks. As a result, it allows for the explanation and optimization of intricate models. Hence, in order to improve the accuracy of the model, we learn the nonlinear relations between dataset and origins using PVI.

2.4.5. Evaluation index

The common evaluation metrics for neural networks include Accuracy, Precision, and Recall. These metrics offer different aspects of the model's performance in a classification task.

Accuracy is a prevalent classification metric that represents the proportion of correctly classified samples among all samples. It is calculated by dividing the number of correct predictions by the total number of samples, as shown in the following formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

where, TP (True Positive) is the number of instances correctly predicted as positive. FN (False Negative) is the number of instances positive but predicted as negative. FP (False Positive) is the number of instances negative but predicted as positive. TN (True Negative) is the number of instances correctly predicted as negative. Accuracy provides an overall assessment of the model's classification performance. However, it may be biased in the case of imbalanced class distributions.

Precision focuses on the accuracy of the model's positive predictions. It measures the proportion of true positive predictions among all positive predictions, helping us understand the reliability of the model in correctly detecting samples belonging to a particular class. The calculation formula for Precision is as follows:

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

Recall, also known as sensitivity or true positive rate, measures the proportion of correctly predicted positive samples among all actual positive samples. It helps us understand the model's ability to correctly detect samples from a specific class. The calculation formula for Recall is as follows:

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

3. Results and discussion

3.1. Spectroscopy analysis of saponins with different detection methods

R_1, Rb_1 and Rg_1 are the main bioactive components of *P. notoginseng*. We observed their chromatogram and spectra using four detection techniques: HPLC, UV, Raman and THz spectroscopy. The resulting chromatogram and spectra are depicted in Fig. 1. The standards R_1, Rb_1 and Rg_1 were first detected by HPLC. The difference of their retention times is used to differentiate the three saponins, as shown in Fig. 1(a). R_1 is appeared at 25.7 min, Rg_1 at 29.0 min, and Rb_1 at 52.4 min. According to the positions of the measured peaks, the subsequent identification of *P. notoginseng* samples is determined, and the saponin contents are calculated using corresponding peak areas. The results of UV spectroscopy are displayed in Fig. 1(b), revealing that the peaks of these three saponins are not distinct and cannot be differentiated from each other due to interference from noise. Similarly, Raman spectroscopy results in Fig. 1(c) show that the absorption peaks of the three saponins are all concentrated at 1467 cm^{-1} and 1530 cm^{-1}, which cannot be distinguished. Through THz detection, due to the resonance absorption between the THz wave and the vibration and rotation of the functional groups in different molecules, the absorption peaks of the three saponins are clearly different, as shown in Fig. 1(d). The absorption peaks of R_1 are mainly located in regions I (7.6–8.9 THz) and III (10.6–12.1 THz), and those of Rb_1 and Rg_1 are mainly distributed in regions I, II (9.0–10.6 THz) and region III. Obviously, both UV and Raman detection are unable to effectively differentiate the three saponins of *P. notoginseng*. Therefore, HPLC and THz methods are mainly used for the later analysis of origin identification of subsequent *P. notoginseng* samples.

3.2. THz and HPLC detection and analysis

After confirming the ability of HPLC and THz techniques for qualitative identification and quantitative analysis of saponins, we began to analyze the source of *P. notoginseng*. A total of 252 batches of samples were collected in five major producing areas of *P. notoginseng*: Honghe Autonomous Prefecture (73), Kunming City (46), Qujing City (28), Wenshan Autonomous Prefecture (92) and Chuxiong City (13) in Yunnan Province, China. The corresponding geographical distribution is displayed in Fig. 2(a) –2(e). The farming conditions is detailed in Fig. S2.

Firstly, we tested all samples with HPLC equipment (see Section 2.3 for the processing steps). Prior to analysis, each sample underwent several pre-processing steps to extract the corresponding saponin eluates. Subsequently, saponin standard solutions were injected into the HPLC system to establish the standard retention times. This process was repeated three times for consistency. Next, each sample was injected and analyzed twice. The entire process took 28 h in total. Representative

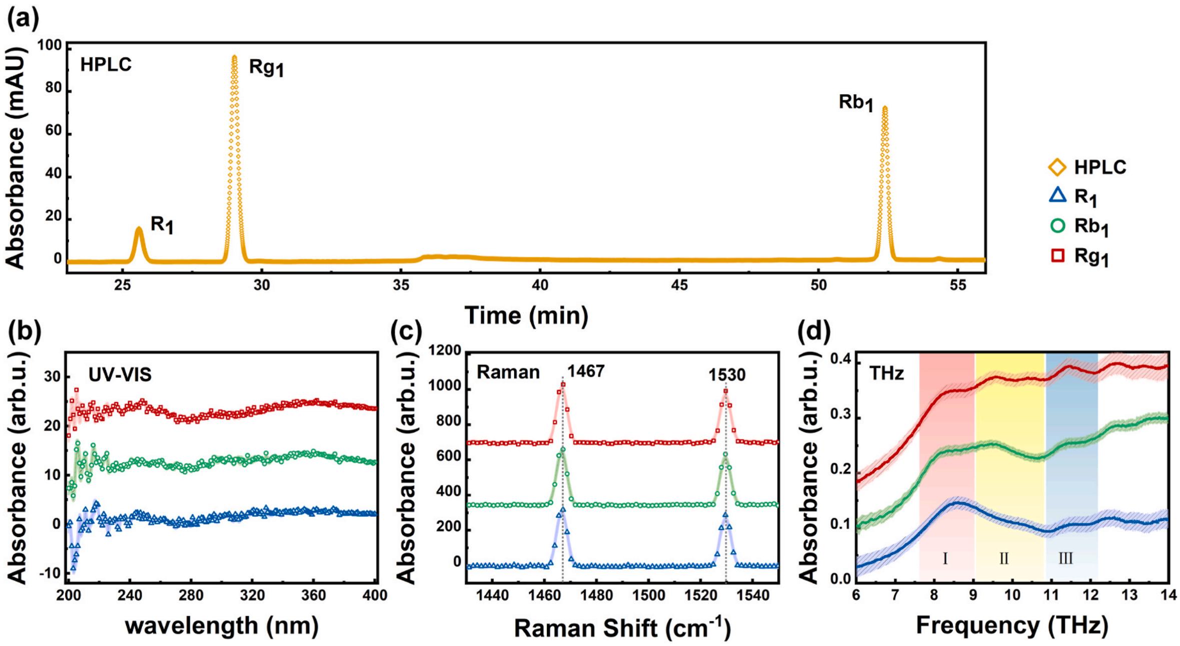


Fig. 1. The resulting chromatogram and spectra: Chromatograms of R_1, Rb_1 and Rg_1 under HPLC detection (Orange diamond), (b) UV spectra, (c) Raman spectra and (d) THz of R_1 (blue triangle)、Rb_1 (Green circle) and Rg_1 (Red square). The error bar has been labelled on each curve.

chromatograms of each origin tested by HPLC are presented in Fig. 2(f) to 2(j). It can be seen that the elution time of saponins R_1, Rb_1 and Rg_1 is 25.8–26.5 min, 29.1–30.1 min, and 52.6–53.6 min, respectively. These intervals show certain distinctions in terms of retention times, peak areas, and tailing factors.

Next, for the THz detection, we tested all samples with FTIR equipment (see Section 2.2 for the processing). Each sample was tested four times, each time for 3 min, a total of 12 min. The obtained spectra of *P. notoginseng* from each origin are shown in Fig. 2(k) - 2(o). It can be observed that there is an absorption peaks in each range including I (7.9 THz-9.5 THz), II (9.5 THz-11.3 THz) and III (11.3 THz-12.6 THz). All spectra of 252 samples, and spectra-to-spectra variance for each origin are shown in Fig. S1. The absorption peaks were basically the same as those of the three saponins in Fig. 1(d), and the slight changes in the width and shape were due to the overlap of absorption peaks caused by other substances in *P. notoginseng*, such as polysaccharides and amino acids. In addition, due to the unknown information introduced by the environments of different origins, the absorption peaks of *P. notoginseng* exhibited certain disparities in terms of frequency, amplitude, relative proportion, and peak areas.

The spectral differences among samples from the five origins obtained by HPLC and THz methods may be caused by a series of factors, the first of which is extrinsic factors, such as variations in instrument components, measurement surroundings and operation procedures, which have little influence under standardized operation. The second category is intrinsic factors, such as differences in the contents of various components of *P. notoginseng*, including R_1, Rb_1, Rg_1, other saponins, glucose and amino acids, etc, which occupy the main influence. The third category encompasses some unknown elements introduced by the local environment, such as trace elements in the soil.

3.3. Traditional deep learning method

Based on the above spectral differences among *P. notoginseng* of different origins, we can combine the algorithm to extract the relevant information for origin identification. Considering that the overall information of the samples involves nonlinear variations resulting from multiple factors, conventional analytical methods prove insufficient for analyzing such intricate samples. Hence, we use a CNNs model to utilize and enhance the information within the spectra, enabling effective origin discrimination.

3.3.1. Modeling of CNNs

Our CNNs model is set up with two convolutional layers, two pooling layers, one dropout layer, and one fully connected Layers. The convolutional layers carry out element-wise multiplication and summation on the input data using convolutional kernels, with the intention of extracting local features from the input data and capturing spatial relations and specific patterns [46]. Since our input datasets are composed of one-dimensional spectra, both convolutional layers are designed according to a 1 dCNN architecture. The dimensions of the convolutional kernels, along with the specifications for padding and stride, are defined as 3, 1, and 1 respectively. To diminish data dimensionality and the quantity of parameters while retaining significant features, we choose max-pooling layers. The parameters of the two max-pooling layers are standardized, featuring a kernel size and stride of 1 and 2 respectively. Furthermore, we incorporate Dropout layers to counteract potential neural network overfitting, with a dropout rate set at 0.5. Ultimately, the neural network comprehends the relations and weights among features through fully connected layers. In this process, the employment of the softmax activation function facilitates the mapping of the ultimate feature to their corresponding output categories.

Given the challenges posed by traditional optimization algorithms during model training, such as the difficulty in selecting appropriate learning rates, issues of gradient instability, and managing parameters of varying scales, we have opted for the Adam optimizer as the optimization method for our model. By employing strategies such as adaptive learning rates and dynamic momentum, the Adam optimizer effectively expedites model convergence and yields improved results, thereby alleviating the burden of hyperparameter tuning. Ultimately, the model's batch size is set at 64, signifying that each iteration incorporates 64 spectral samples as inputs for the model. Additionally, the model's learning rate is fixed at 0.0001.

3.3.2. Binary classification of the CNNs model for origin identification

Once the CNNs model was established, the spectral data of *P. notoginseng* samples were initially categorized into two major groups: Yunnan Wenshan and Other Origins. Subsequently, a binary

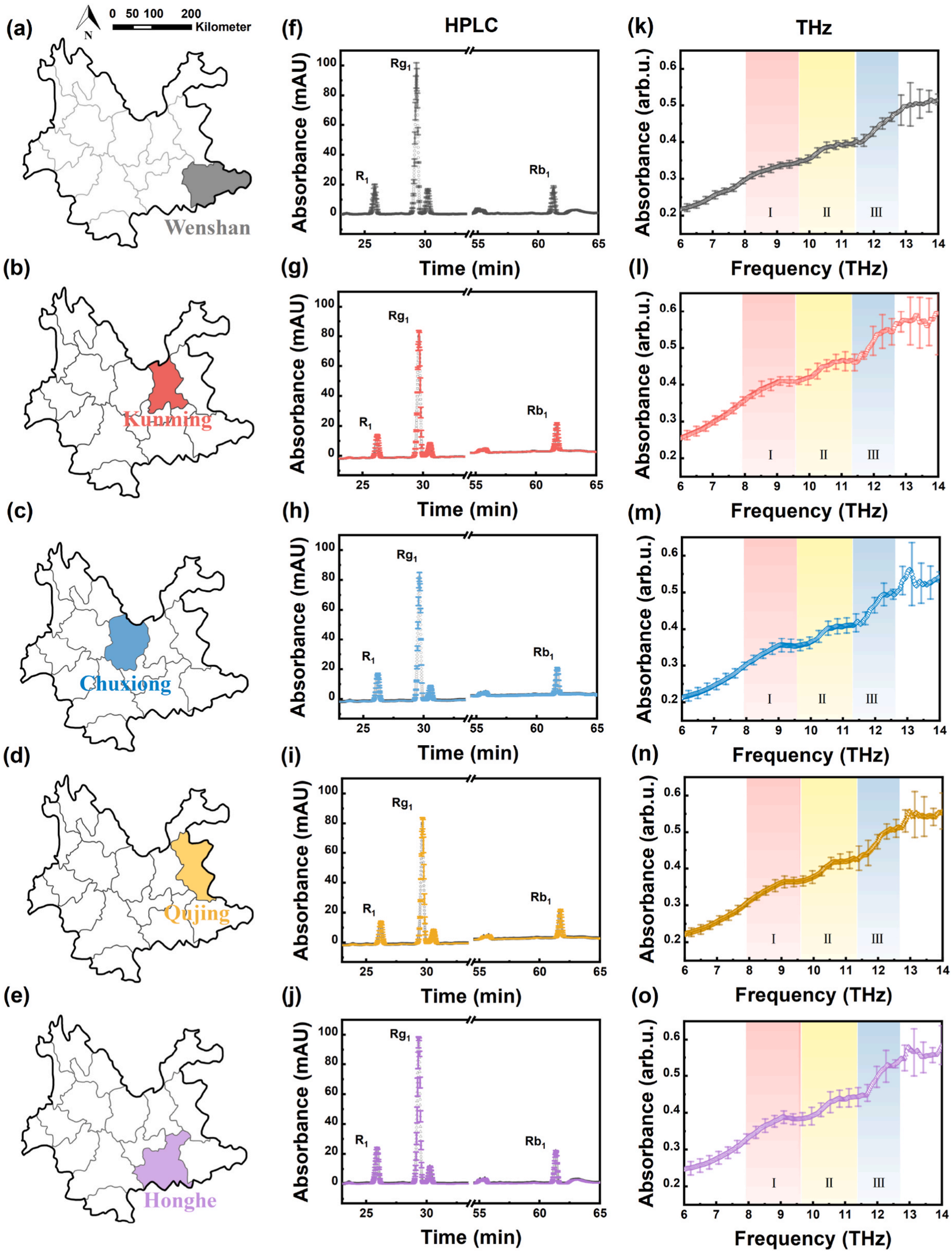


Fig. 2. Geographical distribution of *P. notoginsengs* from five origins of Yunnan Province, China: (a) Wenshan, (b) Kunming, (c)Chuxiong, (d) Qujing and (e) Honghe, (f)–(j) Representative chromatograms of five origin through HPLC detection, (k)–(o) Representative spectra of five origins through THz detection. The error bar has been labelled on each curve.

classification identification model was trained based on HPLC and THz spectra training dataset.

The binary classification results of the CNNs Model for HPLC and THz spectral validation datasets are exhibited in Fig. 3. In the HPLC analysis, out of the 252 samples, 12 samples that actually originated from Yunnan Wenshan were misclassified as other regions, resulting in an accuracy of 95.24%, as shown in Fig. 3(a). In terms of THz spectra, 5 of the 252 samples were incorrectly judged to be from other regions, with an accuracy of 98.02%, as shown in Fig. 3(b). It can be seen from the observation results that the accuracy of binary classification of the

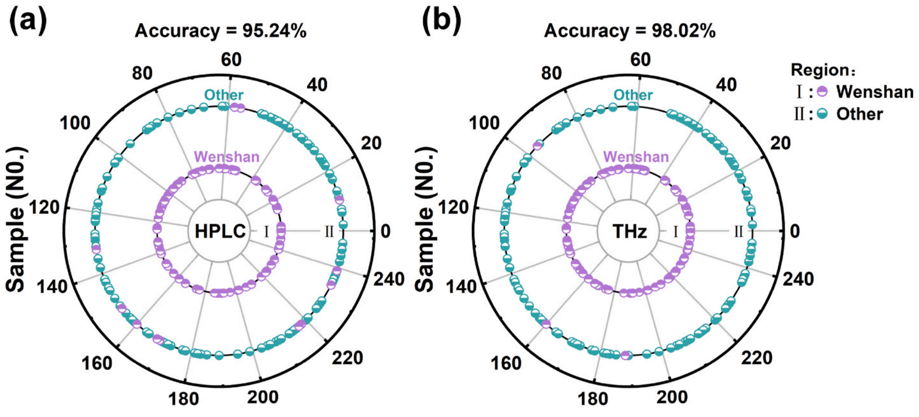


Fig. 3. Binary classification radar map of CNNs model (252 samples): (a) HPLC, (b) THz.

model is always above 95%, which meets the identification requirements of *P. notoginseng* from Wenshan. Subsequently, we proceed to further enhance our analysis by employing the CNNs model to perform a comprehensive five-classification origin identification.

3.3.3. Five-classification of the CNNs model for origin identification

According to the origins of the samples, we divided the spectral data into five categories: Wenshan Autonomous Prefecture, Honghe Autonomous Prefecture, Kunming City, Qujing City and Chuxiong City. Correspondingly, we built a five-classification identification model using HPLC and THz spectral training data.

The five-classification results of the CNNs Model for HPLC and THz spectral validation datasets are exhibited in Fig. 4. In HPLC analysis, 22 out of 252 samples were misclassified with an accuracy of 91.27%, as shown in Fig. 4(a). In terms of the THz analysis, 10 out of 252 samples were misjudged with an accuracy of 96.03%, as shown in Fig. 4(b). It can be observed that, compared to the binary classification model, the accuracy of the five-classification model has slightly decreased. Notably, the accuracy of the five-classification based on THz spectra is still more than 95%, while the accuracy of the five-classification based on HPLC is significantly decreased.

To analyze the reasons for the decrease of accuracy, we used a confusion matrix to visualize the categorization of the model. As shown in Fig. 4(c), in HPLC analysis, the classified recall rate of Chuxiong *P. notoginseng* by CNNs model was pretty low, only 57.1%. This signifies that the model has limited ability to accurately detect Chuxiong *P. notoginseng*, resulting in a propensity to misidentify *P. notoginseng* from Chuxiong as originating from other regions. Furthermore, the CNNs model exhibits precision as low as 68.2% when classifying *P. notoginseng* from Qujing. This indicates a weak reliability of the model in correctly identifying samples as sourced from Qujing, which consequently leads to an elevated tendency of mislabeling *P. notoginseng* from

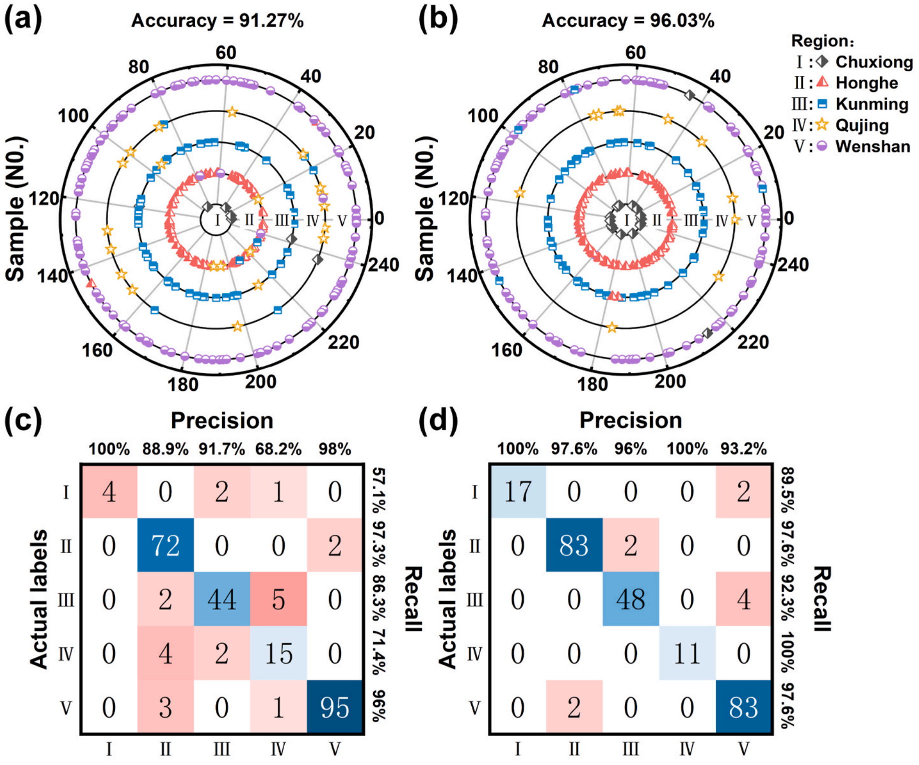


Fig. 4. Five-classification radar map of CNNs model (252 samples): (a) HPLC, (b) THz. Five-classification confusion matrix of CNNs model (252 samples): (c) HPLC, (d) THz.

other origins as originating from Qujing. This is due to the limited quantity of *P. notoginseng* samples obtained from Chuxiong and Qujing, accounting for only 5.16% and 11.11% of the total dataset, respectively. This scarcity of samples increases the challenge of the model to learn the intricate nonlinear relations between the dataset and the origins of Chuxiong and Qujing.

The five-classification confusion matrix based on THz spectroscopy is visualized in Fig. 4(d). For limited quantity of *P. notoginseng* samples, the model demonstrates a recall rate of 89.5% for identifying *P. notoginseng* originating from Chuxiong, while the precision and recall rates for the other origins both surpass the 90% threshold. Evidently, compared to HPLC, the combination of THz spectroscopy and the CNNs model has

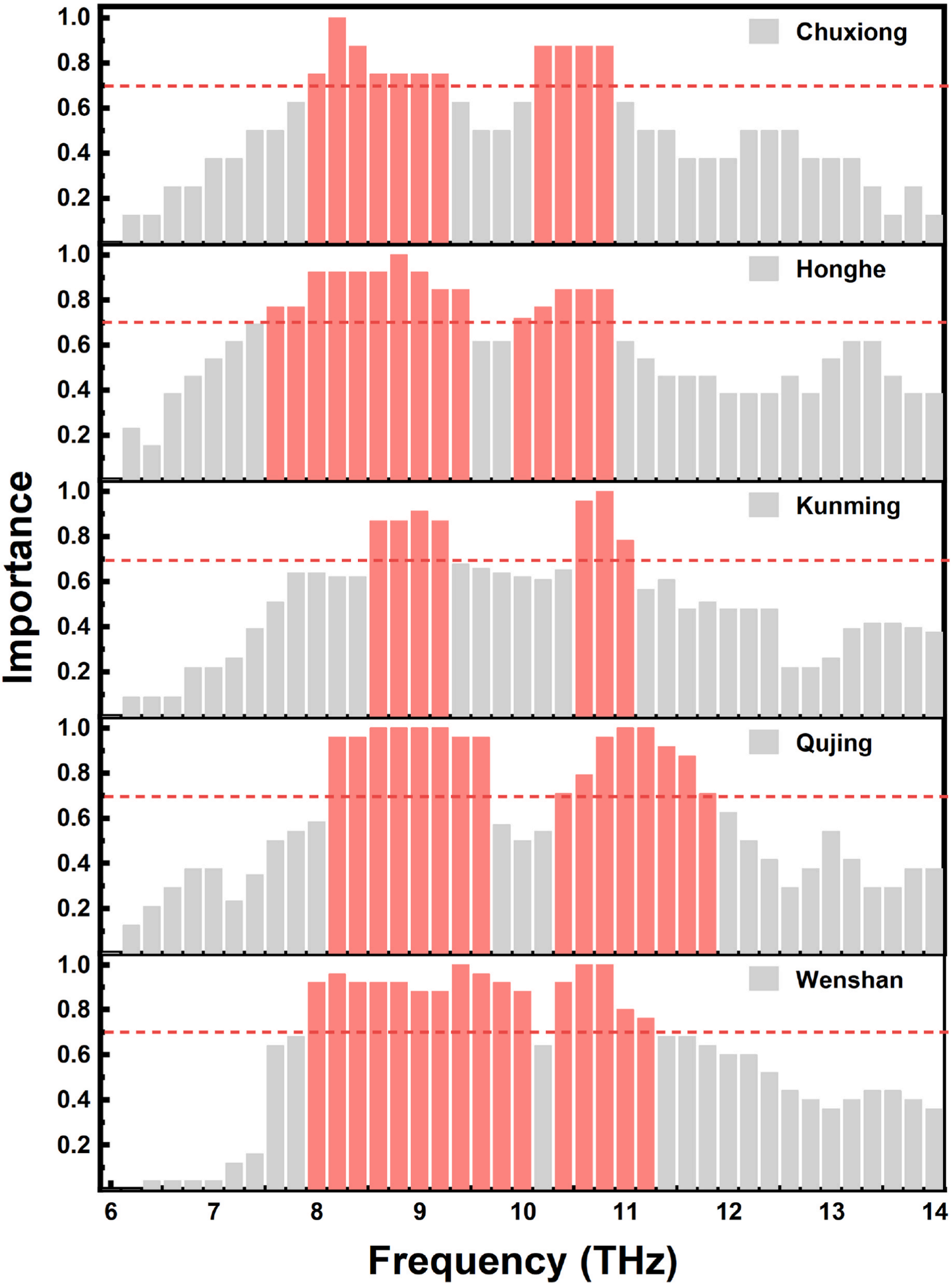


Fig. 5. Permutation Feature Importance (bar charts): (a) Chuxiong, (b) Honghe, (c) Kunming, (d) Qujing and (e) Wenshan.

better performance in the field of geographic five-classification of *P. notoginseng.*

3.4. Nonlinear relations between spectra and origins

By comparing the results of binary and five-class classifications, it is evident that under the same model, THz spectroscopy has excellent performance in origin identification. To explain this phenomenon, we integrate Permutation Variable Importance (PVI) to analyze the important features of THz spectra. This approach aims to analyze the underlying mechanism information behind origin identification.

As shown in Fig. 5, important features with importance greater than 0.7 are highlighted in red. For *P. notoginseng* from different origin, the ranges of important features depicted in Fig. 5(a)-5(e) are as follows: 7.9–9.3 THz and 10.1–10.9 THz (Chuxiong), 7.6–9.5 THz and 9.9–10.8 THz (Honghe), 8.5–9.3 THz and 10.5–11.1 THz (Kunming), 8.1–9.6 THz and 10.3–11.9 THz (Qujing), and 7.9–10.1 THz and 10.3–11.3 THz (Wenshan). These ranges correspond closely to the absorption peaks I (7.91 THz-9.55 THz) and II (9.55 THz-11.3 THz) of THz spectra, detailed in Fig. 2(f) - 2(o). We can see that these important features are manifested in different frequency points, including the amplitude of a single frequency point, the relative amplitude ratio between frequency points, the spectrum width and the corresponding area of continuous coverage by frequency points. A combination of these parameters can be used as the multi-dimensional information for the origin identification.

Furthermore, in order to illustrate the relations among THz spectra, importance, and origins intuitively, we present the contour map for visualization as depicted in Fig. 6(a). The importance of five origins displays a clear contrast as it varies with frequency. In order to extract trends in feature importance as a function of frequency, we performed polynomial regression analysis, which revealed that the data from different origins conformed to different nonlinear function curves, as shown in the following formula:

$$y = a_0 + a_1x + a_2x^2 + a_3x^3 + a_4x^4 + a_5x^5 + a_6x^6 + a_7x^7 + a_8x^8 + a_9x^9 \quad (5)$$

Table 1 below lists the polynomial coefficients and the goodness of fit (R^2) for each origin. Significant differences in the coefficients were observed across the different origins. The corresponding distinct nonlinear functional is depicted in Fig. 6(b), which can reflect the uniqueness of each origin.

Based on these specific nonlinear relations, we can correspondingly enhance the feature extraction of these regions to further improve the accuracy of origin identification. Attributes such as frequency, amplitude, relative proportion, peak areas, and the full width at half maximum (FWHM) of peaks I and II are all correlated with the characteristics of their respective origins. Therefore, we combine these features with the spectral data as inputs to construct the FE-CNNs model. The results of FE-CNNs model are shown in Fig. 7(a). It can be seen that the accuracy of five-classification is up to 97.62%, only 6 out of the 252 samples were incorrectly originated. The precision and recall rates for Chuxiong and Qujing reached 100% due to the extraction of important information from the THz absorption peaks, as depicted in Fig. 7(b). These results highlight that FE-CNNs model can provide effective origin identification even with a limited sample size and demonstrates an elevated capacity in terms of identification performance indicators compared with CNNs model. Therefore, the combination of FE-CNNs model with THz spectroscopy confers a robust advantage in accurately classifying the five origins of *P. notoginseng*, fulfilling the demand of precise identification of *P. notoginseng* origin.

4. Conclusion

In this paper, we proposed a method combining terahertz precision spectroscopy and CNNs algorithm to identify the origin of *P. notoginseng*. We focused on HPLC and THz detection to obtain chromatograms and spectra of 252 *P. notoginseng* samples from five origins and constructed a CNNs model to train the datasets, resulting in binary origins classification accuracy rates of 95.24% and 98.02%, respectively. Upon distinguishing Wenshan from other origins in the initial step, a five-classification of origins was conducted for *P. notoginseng* samples from Wenshan, Chuxiong, Honghe, Kunming, and Qujing, Yunnan, achieving accuracy rates of 96.03% using THz spectra higher than HPLC (91.27%). Upon analyzing the reasons, we found that there are clear nonlinear relations between the origin and THz spectra, where the relatively broad absorption peaks encompass equivalent informative contents such as frequency, amplitude, relative proportion, peak areas and full width at half maximum. Extracting these contents into CNNs model, the accuracy of five-classification reached 97.62%. This study achieves rapid, nondestructive, and precise detection of the origin of *P. notoginseng*, offering potential applications to other herbal medicines and playing a crucial role in classification and identification within the herbal medicine domain.

CRediT authorship contribution statement

Hongyu Gu: Writing – original draft, Investigation. **Shengfeng Wang:** Formal analysis. **Songyan Hu:** Investigation. **Xu Wu:** Supervision. **Qiuye Li:** Resources. **Rongrong Zhang:** Resources. **Juan Zhang:** Resources. **Wenbin Zhang:** Resources. **Yan Peng:** Supervision, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

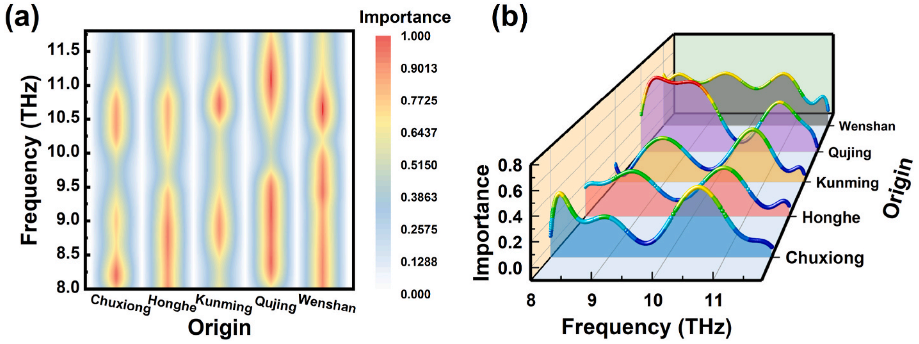


Fig. 6. The relations among THz spectra, importance and origins. (a) Contour map of relations (b) 3D diagram of the nonlinear relations between important frequency and five origins.

Table 1
The polynomial coefficients and the goodness of fit (R^2) for five origins.

	Chuxiong	Honghe	Kunming	Qujing	Wenshan
a_0	−32,000,000	−5640,111	16,363,300	15,842,100	37,551,100
a_1	28,873,500	4,965,104	−15,000,000	−15,000,000	−35,000,000
a_2	−12,000,000	−1929,068	6,394,064	6,341,432	14,477,400
a_3	2,724,770	433917.7	−1549,741	−1551,814	−3482,989
a_4	−409243	−62233.1	240962.8	243251.7	537353.5
a_5	40862.13	5897.06	−24923.8	−25330.7	−55132.6
a_6	−2712.39	−368.822	1714.847	1752.399	3761.819
a_7	115.4214	14.66273	−75.6767	−77.667	−164.602
a_8	−2.85717	−0.33567	1.94358	2.00115	4.1911
a_9	0.03135	0.00336	−0.02213	−0.02284	−0.04731
R^2	0.97762	0.96146	0.89354	0.984	0.819

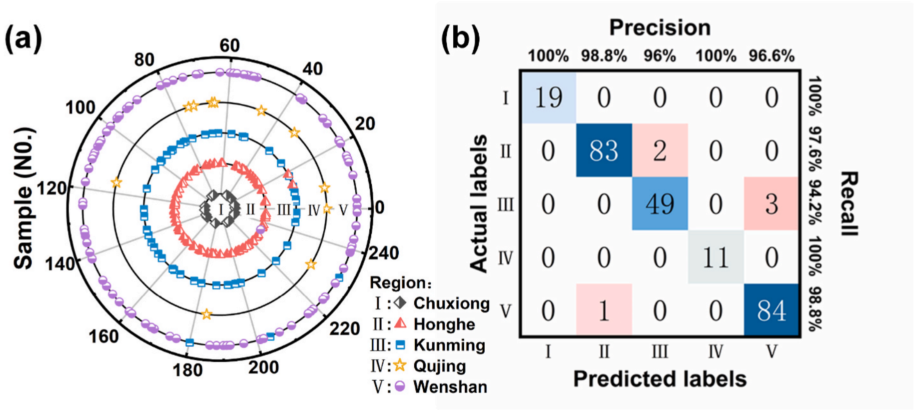


Fig. 7. Predictive Results of the FE-CNNs Model for THz spectra: (a) radar map, (b) confusion matrix.

Data availability

Data will be made available on request.

Acknowledgements

Funding, This work was supported by the National Natural Science Foundation of China (grant numbers 61988102, 62335012), National Key Research and Development Program of China (grant number 2022YFA1404004), Yunnan Science and Technology Talents and Platform Plan (grant number 202005AF150076).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.talanta.2024.125968.

References

[1] H.J. Park, D.H. Kim, S.J. Park, J.M. Kim, J.H. Ryu, Ginseng in traditional herbal Prescriptions, J. Ginseng Res 36 (3) (2012), https://doi.org/10.5142/jgr.2012.36.3.225, 225-241.

[2] X. Yang, X. Xiong, H. Wang, J. Wang, Protective effects of panax notoginseng saponins on cardiovascular diseases: a comprehensive overview of experimental studies, Evid. base Compl. Alternative Med. 2014 (2014) 204840, https://doi.org/10.1155/2014/204840.

[3] T. Wang, R.X. Guo, G.H. Zhou, X.D. Zhou, Z.Z. Kou, F. Sui, C. Li, L.Y. Tang, Z. J. Wang, Traditional uses, botany, phytochemistry, pharmacology and toxicology of Panax notoginseng (Burk.) FH Chen: a review, J. Ethnopharmacol. 188 (2016) 234–258, https://doi.org/10.1016/j.jep.2016.05.005.

[4] M. Yoshikawa, T. Morikawa, Y. Kashima, K. Ninomiya, H. Matsuda, Structures of new dammarane-type triterpene saponins from the flower buds of Panax notoginseng and hepatoprotective effects of principal ginseng saponins, J. Nat. Prod. 66 (7) (2003) 922–927, https://doi.org/10.1021/np030015l.

[5] J.B. Wan, Q.W. Zhang, S.J. Hong, J. Guan, W.C. Ye, S.P. Li, M.Y.S. Lee, Y.T. Wang, 5,6-Didehydroginsenosides from the roots of panax notoginseng, Molecules 15 (11) (2010) 8169–8176, https://doi.org/10.3390/molecules15118169.

[6] P.Y. Liao, D. Wang, Y.J. Zhang, C.R. Yang, Dammarane-type glycosides from steamed notoginseng, J. Agric. Food Chem. 56 (5) (2008) 1751–1756, https://doi.org/10.1021/jf073000s.

[7] L. Qiu, Y. Jiao, G.K. Huang, J.Z. Xie, J.H. Miao, X.S. Yao, New dammarane-type saponins from the roots of panax notoginseng, Helv. Chim. Acta 97 (1) (2014) 102–111, https://doi.org/10.1002/hlca.201300155.

[8] J.T.T. Zhu, W.K.W. Leung, J.K.H. Cheung, K.J. Zhao, T.T.X. Dong, K.W.K. Tsim, A flavonol glycoside, isolated from roots of Panax notoginseng, protects the beta-amyloid-induced neurotoxicity in cultured PC12 cells, Neurosignals 15 (3) (2006), 150-150.

[9] P. Wang, L. Zhang, J. Yao, Y. Shi, P. Li, K. Ding, An arabinogalactan from flowers of Panax notoginseng inhibits angiogenesis by BMP2/Smad/Id1 signaling, Carbohydr. Polym. 121 (2015) 328–335, https://doi.org/10.1016/j.carbpol.2014.11.073.

[10] N. Komakine, M. Okasaka, Y. Takaishi, K. Kawazoe, K. Murakami, Y. Yamada, New dammarane-type saponin from roots of Panax notoginseng, J. Nat. Med. 60 (2) (2006) 135–137, https://doi.org/10.1007/s11418-005-0016-0.

[11] J. Liu, Y. Wang, L. Qiu, Y. Yu, C. Wang, Saponins of Panax notoginseng: chemistry, cellular targets and therapeutic opportunities in cardiovascular diseases, Expet Opin. Invest. Drugs 23 (4) (2014) 523–539, https://doi.org/10.1517/13543784.2014.892582.

[12] C.-y. Yang, J. Wang, Y. Zhao, L. Shen, X. Jiang, Z.-g. Xie, N. Liang, L. Zhang, Z.-h. Chen, Anti-diabetic effects of Panax notogineng saponins and its major anti-hyperglycemic components, J. Ethnopharmacol. 130 (2) (2010) 231–236, https://doi.org/10.1016/j.jep.2010.04.039.

[13] B. Sun, J. Xiao, X.-B. Sun, Y. Wu, Notoginsenoside R1 attenuates cardiac dysfunction in endotoxemic mice: an insight into oestrogen receptor activation and PI3K/Akt signalling, Br. J. Pharmacol. 168 (7) (2013) 1758–1770, https://doi.org/10.1111/bph.12063.

[14] W. Zhang, J. Wojta, B.R. Binder, Effect of notoginsenoside R1 on the synthesis of tissue-type plasminogen activator and plasminogen activator inhibitor-1 in cultured human umbilical vein endothelial cells, Arteriosclerosis and thrombosis : a, j. vascular biology 14 (7) (1994) 1040–1046, https://doi.org/10.1161/01.Atv.14.7.1040.

[15] Y. Zhao, W. Wang, L. Han, E.R. Rayburn, D.L. Hill, H. Wang, R. Zhang, Isolation, structural determination, and evaluation of the biological activity of 20(S)-25-methoxyl-dammarane-3 beta, 12 beta, 20-triol 20(S)-25-OCH3-PPD , a novel natural product from Panax notoginseng, Med. Chem. 3 (1) (2007) 51–60, https://doi.org/10.2174/157340607779317508.

[16] J. Sibik, K. Lobmann, T. Rades, J.A. Zeitler, Predicting crystallization of amorphous drugs with terahertz spectroscopy, Mol. Pharm. 12 (8) (2015) 3062–3068, https://doi.org/10.1021/acs.molpharmaceut.5b00330.

[17] J.T. Gong, K.M. Li, M. Sun, H.J. Liu, Z.L. Zhang, Allelopathy and soil sickness in continuous cropping of Panax medicinal plants, Allelopathy J. 39 (1) (2016) 1–17.

[18] J. Li, Y. Bao, Z. Wang, Q. Yang, X. Cui, Research progress in diseases of Panax notoginseng, Physiological and, Mol. Plant Pathol. 121 (2022) 101878, https://doi.org/10.1016/j.pmpp.2022.101878.

[19] Z. Zhao, Z. Liang, G. Ping, Macroscopic identification of Chinese medicinal materials: traditional experiences and modern understanding, J. Ethnopharmacol. 134 (3) (2011) 556–564, https://doi.org/10.1016/j.jep.2011.01.018.

[20] C. Ji, Q. Zhang, R. Shi, J. Li, X. Wang, Z. Wu, Y. Ma, J. Guo, X. He, W. Zheng, Determination of the authenticity and origin of panax notoginseng: a review, J. AOAC Int. 105 (6) (2022) 1708–1718, https://doi.org/10.1093/jaoacint/qsac081.

[21] J. Zhu, X. Fan, Y. Cheng, R. Agarwal, C.M.V. Moore, S.T. Chen, W. Tong, Chemometric analysis for identification of botanical raw materials for pharmaceutical use: a case study using panax notoginseng, PLoS One 9 (1) (2014) e87462, https://doi.org/10.1371/journal.pone.0087462.

[22] Y. Zhou, Z. Zuo, F. Xu, Y. Wang, Origin identification of Panax notoginseng by multi-sensor information fusion strategy of infrared spectra combined with random forest, Spectrochim. Acta Mol. Biomol. Spectrosc. 226 (2020) 117619, https://doi.org/10.1016/j.saa.2019.117619.

[23] H. Zhang, L. Huang, C. Xu, Z. Li, X. Yin, T. Chen, Y. Wang, Rapid determination of Panax notoginseng origin by terahertz spectroscopy combined with the machine learning method, Spectrosc. Lett. 55 (9) (2022) 566–578, https://doi.org/10.1080/00387010.2022.2125017.

[24] B. Liu, Y. Peng, Z. Jin, X. Wu, H. Gu, D. Wei, Y. Zhu, S. Zhuang, Terahertz ultrasensitive biosensor based on wide-area and intense light-matter interaction supported by QBIC, Chem. Eng. J. 462 (2023), https://doi.org/10.1016/j.cej.2023.142347.

[25] H. Gu, C. Shi, X. Wu, Y. Peng, Molecular methylation detection based on terahertz metamaterial technology, Analyst 145 (20) (2020) 6705–6712, https://doi.org/10.1039/d0an01062f.

[26] X. Wu, L. Wang, Y. Peng, F. Wu, J. Cao, X. Chen, W. Wu, H. Yang, M. Xing, Y. Zhu, Y. Shi, S. Zhuang, Quantitative analysis of direct oral anticoagulant rivaroxaban by terahertz spectroscopy, Analyst 145 (11) (2020) 3909–3915, https://doi.org/10.1039/d0an00268b.

[27] Y. Peng, J. Huang, J. Luo, Z. Yang, L. Wang, X. Wu, X. Zang, C. Yu, M. Gu, Q. Hu, X. Zhang, Y. Zhu, S. Zhuang, Three-step one-way model in terahertz biomedical detection, Photonix 2 (1) (2021), https://doi.org/10.1186/s43074-021-00034-0.

[28] Y. Peng, C. Shi, Y. Zhu, M. Gu, S. Zhuang, Terahertz spectroscopy in biomedical field: a review on signal-to-noise ratio improvement, Photonix 1 (1) (2020), https://doi.org/10.1186/s43074-020-00011-z.

[29] Z. Zhou, J. Huang, X. Li, X. Gao, Z. Chen, Z. Jiao, Z. Zhang, Q. Luo, L. Fu, Adaptive optical microscopy via virtual-imaging-assisted wavefront sensing for high-resolution tissue imaging, Photonix 3 (1) (2022), https://doi.org/10.1186/s43074-022-00060-6.

[30] C. Liu, W. Eschen, L. Loetgering, D.P.S. Molina, R. Klas, A. Iliou, M. Steinert, S. Herkersdorf, A. Kirsche, T. Pertsch, F. Hillmann, J. Limpert, J. Rothhardt, Visualizing the ultra-structure of microorganisms using table-top extreme ultraviolet imaging, Photonix 4 (1) (2023), https://doi.org/10.1186/s43074-023-00084-6.

[31] S. Zhang, X. Chen, K. Liu, H. Li, Y. Xu, X. Jiang, Y. Xu, Q. Wang, T. Cao, Z. Tian, Nonvolatile reconfigurable terahertz wave modulator, Photonix 3 (1) (2022), https://doi.org/10.1186/s43074-022-00053-5.

[32] X.Z. Yang Zhu, Haoxiang Chi, Yiwen Zhou, Yiming Zhu, Songlin Zhuang, Metasurfaces designed by a bidirectional deep neural network and iterative algorithm for generating quantitative field distributions, Light: Adv. Manuf. 4 (9) (2023) 104–114, https://doi.org/10.37188/lam.2023.009.

[33] X. Zang, B. Yao, L. Chen, J. Xie, X. Guo, A.V. Balakin, A.P. Shkurinov, S. Zhuang, Metasurfaces for manipulating terahertz waves, Light: Adv. Manuf. 2 (2) (2021) 10.

[34] Q. Li, T. Lei, Y. Cheng, X. Wei, D.-W. Sun, Predicting wheat gluten concentrations in potato starch using GPR and SVM models built by terahertz time-domain spectroscopy, Food Chem. 432 (2024) 137235, https://doi.org/10.1016/j.foodchem.2023.137235.

[35] Q. Ma, Y. Teng, C. Li, L. Jiang, Simultaneous quantitative determination of low-concentration ternary pesticide mixtures in wheat flour based on terahertz spectroscopy and BPNN, Food Chem. 377 (2022) 132030, https://doi.org/10.1016/j.foodchem.2021.132030.

[36] T. Li, H. Ma, Y. Peng, X. Chen, Z. Zhu, X. Wu, T. Kou, B. Song, S. Guo, L. Liu, Y. Zhu, Gaussian numerical analysis and terahertz spectroscopic measurement of homocysteine, Biomed. Opt Express 9 (11) (2018) 5467–5476, https://doi.org/10.1364/boe.9.005467.

[37] X. Zhang, THz time-domain spectroscopy technology, Laser Optron. Prog. 42 (7) (2005) 35–38.

[38] N.R. Han, Z.C. Chen, C.S. Lim, B. Ng, M.H. Hong, Broadband multi-layer terahertz metamaterials fabrication and characterization on flexible substrates, Opt Express 19 (8) (2011) 6990–6998, https://doi.org/10.1364/oe.19.006990.

[39] U. Erdenebayar, H. Kim, J. Park, k. dongwon, K.J. Lee, Automatic prediction of atrial fibrillation based on convolutional neural network using a short-term normal electrocardiogram signal, J. Kor. Med. Sci. 34 (7) (2019) 1–10.

[40] S. Harbola, V. Coors, One dimensional convolutional neural network architectures for wind prediction, Energy Convers. Manag. 195 (2019) 70–75, https://doi.org/10.1016/j.enconman.2019.05.007.

[41] D. Han, J. Chen, J. Sun, A parallel spatiotemporal deep learning network for highway traffic flow forecasting, Int. J. Distributed Sens. Netw. 15 (2) (2019) 155014771983279, https://doi.org/10.1177/1550147719832792.

[42] Q. Zhang, D. Zhou, X. Zeng, HeartID: a multiresolution convolutional neural network for ECG-based biometric human identification in smart health applications, IEEE Access 5 (2017) 11805–11816, https://doi.org/10.1109/access.2017.2707460.

[43] O. Abdeljaber, O. Avci, S. Kiranyaz, M. Gabbouj, D.J. Inman, Real-time vibration-based structural damage detection using one-dimensional convolutional neural networks, J. Sound Vib. 388 (2017) 154–170, https://doi.org/10.1016/j.jsv.2016.10.043.

[44] O. Abdeljaber, S. Sassi, O. Avci, S. Kiranyaz, A.A. Ibrahim, M. Gabbouj, Fault detection and severity identification of ball bearings by online condition monitoring, IEEE Trans. Ind. Electron. 66 (10) (2019) 8136–8147, https://doi.org/10.1109/tie.2018.2886789.

[45] Y.X. Liu, Y.F. Cheng, W. Wang, A Survey of the Application of Deep Learning in Computer Vision, Global Intelligent Industry Conference (GIIC), PEOPLES R CHINA, Beijing, 2018.

[46] X. Cao, L. Zhang, Z. Wu, Z. Ling, J. Li, K. Guo, Quantitative analysis modeling for the ChemCam spectral data based on laser-induced breakdown spectroscopy using convolutional neural network, Plasma Sci. Technol. 22 (11) (2020) 115502, https://doi.org/10.1088/2058-6272/aba5f6.