molecules

MDPI

Article

Rapid Identification of Wild *Gentiana Genus* in Different Geographical Locations Based on FT-IR and an Improved Neural Network Structure Double-Net

Pan Zeng [1], Xiaokun Li [1], Xunxun Wu [1], Yong Diao [2], Yao Liu [3] and Peizhong Liu [1,4,*]

[1] School of Medicine, Huaqiao University, Quanzhou 362021, China
[2] School of Biomedical Science, Huaqiao University, Quanzhou 362021, China
[3] College of Science and Engineering, National Quemoy University, Kinmen 89250, Taiwan
[4] College of Engineering, Huaqiao University, Quanzhou 362021, China
* Correspondence: pzliu@hqu.edu.cn

Abstract: *Gentiana genus*, a herb mainly distributed in Asia and Europe, has been used to treat the damp heat disease of the liver for over 2000 years in China. Previous studies have shown significant differences in the compositional contents of wild *Gentiana genus* samples from different geographical origins. Therefore, the traceable geographic locations of the wild *Gentiana genus* samples are essential to ensure practical medicinal value. Over the last few years, the developments in chemometrics have facilitated the analysis of the composition of medicinal herbs via spectroscopy. Notably, FT-IR spectroscopy is widely used because of its benefit of allowing rapid, nondestructive measurements. In this paper, we collected wild *Gentiana genus* samples from seven different provinces (222 samples in total). Twenty-one different FT-IR spectral pre-processing methods that were used in our experiments. Meanwhile, we also designed a neural network, Double-Net, to predict the geographical locations of wild *Gentiana genus* plants via FT-IR spectroscopy. The experiments showed that the accuracy of the neural network structure Double-Net we designed can reach 100%, and the F1_score can reach 1.0.

Keywords: wild *Gentiana genus*; FT-IR spectroscopy; deep learning; Double-Net; geographical location identification

Citation: Zeng, P.; Li, X.; Wu, X.; Diao, Y.; Liu, Y.; Liu, P. Rapid Identification of Wild *Gentiana Genus* in Different Geographical Locations Based on FT-IR and an Improved Neural Network Structure Double-Net. *Molecules* **2022**, *27*, 5979. https://doi.org/10.3390/molecules27185979

Academic Editor: Maria Paula Marques

Received: 14 August 2022
Accepted: 31 August 2022
Published: 14 September 2022

1. Introduction

Gentiana genus is an herb found mainly in Asia and Europe that has been used for treating liver damp heat disease for more than 2000 years in China [1]. Recent studies have demonstrated the efficacy of Gentiana in treating diabetes, with liver protection and anti-inflammatory properties [2–5]. The TCM classic herbal formula with Gentiana as the 'Jun herb' (critical herb), Longdanxiegan, has been widely prescribed in treating hypertension by TCM physicians in China [6]. The literature data on the chemical composition suggests that the medicinal substances of Gentiana are iridoids (gentiopicroside, swertidine), flavonoids (isoorientin), xanthones, and polysaccharides [7–9]. Differing from chemical drugs, herbal products exhibit their curative efficacy on the basis of multi-components and multi-targets, and the medicinal substances are closely related to the soil, climate, harvest season, growth age, and other factors [10,11]. As an important medicinal plant, Gentiana is widely distributed in the temperate mountainous regions of China [12]. Previous studies have shown significant differences in the contents of wild Gentiana samples from different geographical sources [13,14]. Hence, tracing the geographical origin of wild Gentiana samples is crucial to ensure their valid medicinal value, which will help to ensure the potency of the herb.

Several advanced spectral and chromatographic techniques have been successfully utilized to identify the authenticity and quality of various herbal medicines, including Fourier transform infrared (FT-IR), high-performance liquid chromatography (HPLC), ultra-performance liquid chromatography (UPLC), nuclear magnetic resonance (NMR),

and Raman spectroscopy techniques [15–17]. However, some of the above methods require complicated sample pre-processing procedures and generate a considerable amount of organic solvent waste solution during the experiment. In contrast, FT-IR spectroscopy is more widely used because of its advantage of allowing rapid and nondestructive measurements [18,19]. Furthermore, the sample volume required for FT-IR testing is very small (down to a few milligrams). As a result, it is more sensitive and possibly more appropriate to obtain valuable information from the sample profile. FT-IR spectroscopy focuses on the MIR region of the electromagnetic spectrum, which could provide information about the foundational vibration (from the stable vibrational state to the first excited vibrational state) of the chemical functional group [20].

With the rise of machine learning and deep learning, more and more people are using machine learning and deep learning algorithms to accomplish scientific research tasks related to spectra. Zareef et al. used an improved machine science collaborative interval partial least squares algorithm combined with competitive adaptive reweighted sampling (Si-CARS-PLS) to predict the antioxidant activity of walnuts with good results [21]. The prediction of gentian from various geographical sources using high-performance liquid chromatography (HPLC) and Fourier transform infrared spectroscopy (FT-IR) was also achieved by Zhao et al. [22]. In the work by Wu et al. [13], support vector machines (SVM) and the related PLS algorithm were used in combination with FT-IR spectroscopy and HPLC to evaluate the quality of wild Gentiana rigescens from different regions, and it was shown experimentally that the improved PLS and SVM algorithms can be used as an alternative method for the qualitative identification and quantitative evaluation of the quality of Gentiana rigescens. In the work by Pei et al. [23], they fused mid-infrared (MIR) and near-infrared (NIR) data and used random forest (RF) and partial least squares discriminant analysis (PLS-DA) machine learning models to predict wild *Paris polyphylla* populations in Yunnan. They also used a principal component analysis (PCA) and certain algorithms for the essential feature selection to extract important features, and finally also achieved a good experimental result with 100% accuracy. There has also been plenty of related work using machine learning algorithms [24,25].

In addition to machine learning algorithms [26], deep learning algorithms [27,28], such as artificial neural networks, also show better performance. There have also been many better research studies using deep learning models to process data related to spectra. In the work by Mutlu et al. [29], FT-IR spectra were used to predict the quality parameters of wheat, and they performed a chemical analysis of flour samples from 79 different wheat varieties grown in different regions of Turkey and used NIR spectra to train the artificial neural network (ANN), and finally achieved a better result. In the work by Gonzalez-Viejo et al. [30], they used PLS and ANN correlation algorithms to predict four chemometrics of beer, such as the pH, alcohol, brix, and maximum volume of foam. After the experiments, they showed that the artificial neural networks could predict these four chemometrics well and that the R^2 of their model built with neural networks reached 0.95. Neural network models have been used in many scientific tasks with good results, and in addition to the application of neural networks in FT-IR spectroscopy, there are many applications of neural networks in other fields. For example, Qie et al. used a BP neural network to study the trajectory planning of redundant robotic arms during upper limb rehabilitation [31], and they demonstrated the feasibility of their method through relevant experiments. Fatigue driving has been a hot topic for a long time, and Chen et al. combined a BP neural network with a time-cumulative effect to detect drowsy driving [32]. They used three features related to fatigue (the longest time a driver closes their eyes continuously, the number of yawns in a period, and the time their eyes are closed) as inputs to the neural network, and finally built an effective model for driver fatigue detection.

The aim of this study was to build a fast, nondestructive, and efficient method for the identification of the geographical origin of wild *Gentiana genus* using FT-IR combined with chemometrics. Meanwhile, for the identification of the geographic location of the wild *Gentiana genus*, we designed a well-performing neural network structure, Double-Net. Due

to its good performance, we believe that Double-Net can be applied for the geographical origin identification of other similar herbs.

2. Results

2.1. Results of Data Pre-Processing

2.1.1. Pre-Processing Results for FT-IR Spectroscopy

The results of the different pre-processing methods for the FT-IR spectra are shown in Figure 1. For the pre-processing of wild *Gentiana genus* FT-IR spectral data, in addition to using a single pre-processing method in Figure 1, we also combined different pre-processing methods in our experiments to observe the experimental results of the different models.

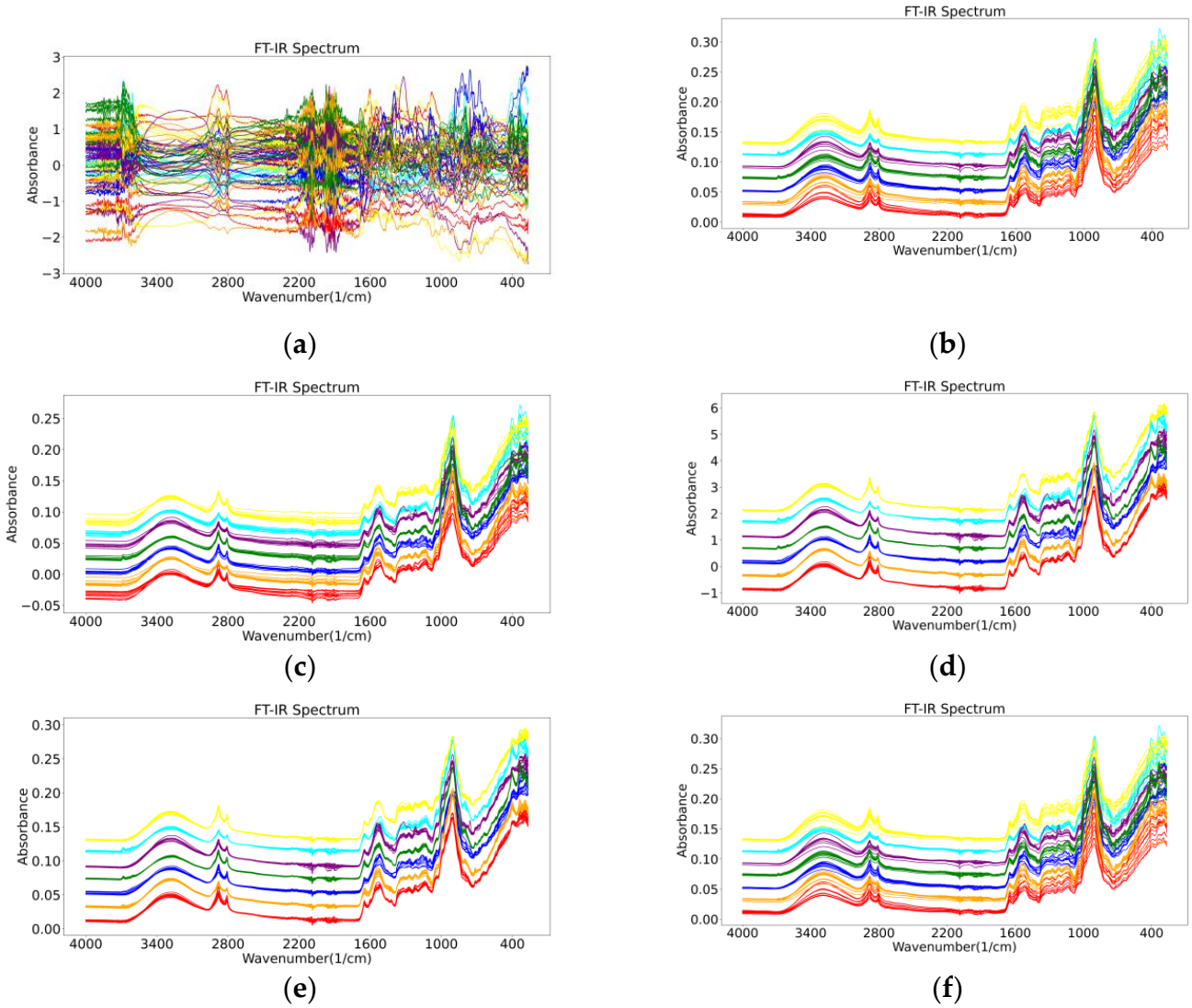
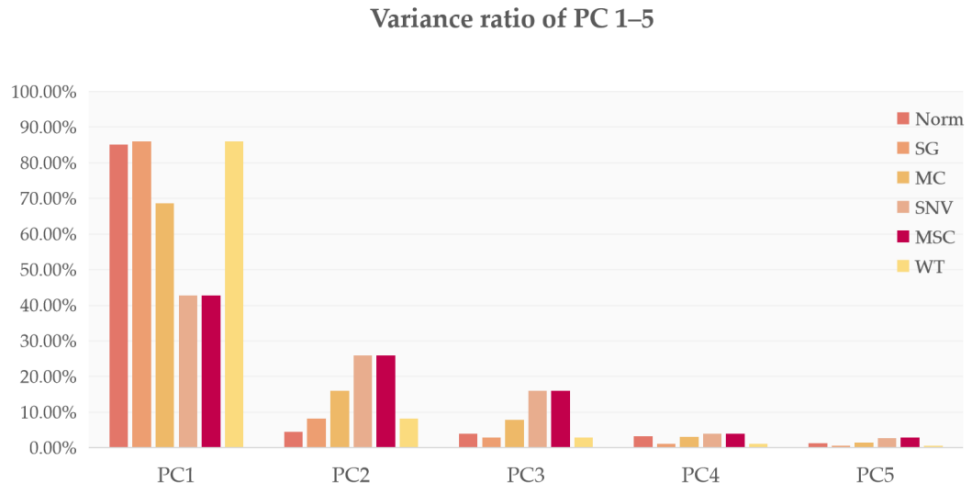


Figure 1. The results of six different pre-processing methods for FT-IR spectra of wild *Gentiana genus*: (**a**) normalization (Norm); (**b**) Savitzky–Golay (SG); (**c**) mean centralized (MC); (**d**) standard normalized variate (SNV); (**e**) multivariate scatter correction (MSC); (**f**) wavelet transform (WT).

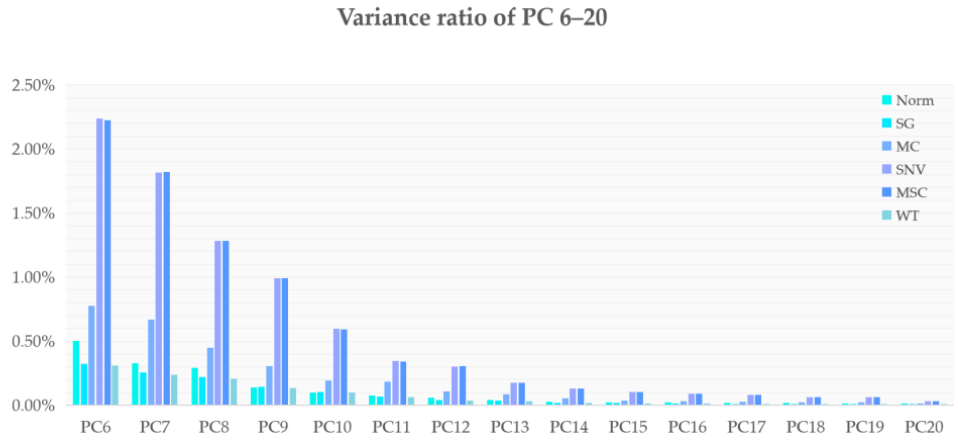
2.1.2. Results of PCA Processing

The percentages of each of the 20 principal components in the FT-IR spectra processed in six different ways are shown in Figure 2. It can be seen from Figure 2 that when we used 20 principal components, the first 15 principal components had the largest contribution and the last few principal components contributed almost 0 (variance ratio < 0.1%). In order to

make the machine learning model, we used more features for learning and we input all of the 20 obtained principal components into the model.



(**a**)



(**b**)

Figure 2. Twenty principal component contributions for six different data pre-processing methods: (**a**) percentage contributions of the 1st to 5th principal components for the six different data pre-processing methods; (**b**) percentage contributions of the 6th to 20th principal components for the 6 different data pre-processing methods.

2.2. Macroscopic Chemistry Components in IR Spectra

The roots of gentian are rich in iridoids (gentiopicroside, swertiamarin, loganin), flavonoids (luteolin, isoorientin, apigenin, and chrysoeriol), xanthones (mangiferin, gentisin, and its glycosides), and polysaccharides [15–17]. The raw FT-IR spectrum of one of the wild *Gentiana genus* samples is shown in Figure 3. The 2500–3700 cm^{-1} region is called the hydrogen stretching zone, as the vibration frequencies of C-H, N-H, and O-H appear in this area. The figure shows that the first C-H stretching vibration peaks appear at 2923 and 2850 cm^{-1}. The region of 2000–2300 cm^{-1} is referred to as the triple bond stretching region (C≡C and C≡N), with almost no peaks in the IR spectrum of the Gentiana sample. The 1600–2000 cm^{-1} region is known as the double bond stretching region (C=C, C=N, and C=O). The peaks around 1736 cm^{-1} represent the CO stretching vibration of the

ester, and the peak at 1615 cm^{-1} suggests the presence of free carboxyl groups or carbohydrates. The peaks at 1421 and 1375 cm^{-1} represent the asymmetric bending vibration of the methyl group, which is the result of the esters. The presence of intense bands at 1025 cm^{-1} is considered to be caused by the glucose skeleton. The last interval in the FT-IR spectra is the fingerprint region, from 1300 to 400 cm^{-1}, which could exhibit more detailed functional group information for the sample [33]. The spectrograms of Gentiana samples from different regions were very similar, so we needed to rely on chemometrics for the analysis [34].

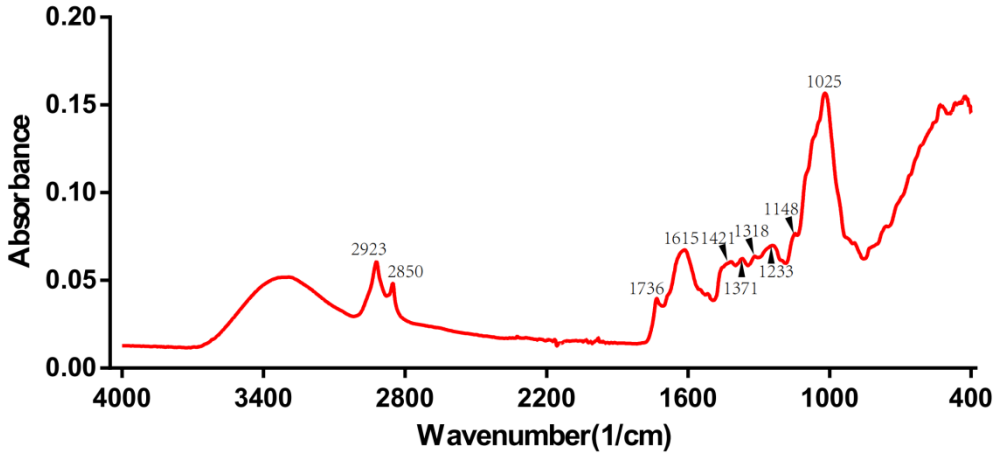


Figure 3. The raw FT-IR spectrum of one of the wild *Gentiana genus* samples.

2.3. Dataset Description

In terms of deep learning, since the deep learning model can go through many features to select those that are important for the current task, there is no need for us to manually extract the sample features artificially, so for the BP neural network and our own designed neural network Double-Net, the data we input for each sample are all the features of the FT-IR spectrum.

In order to make the model show the best performance and also to better test the performance of the model, 20% of the wild *Gentiana genus* data are divided into the test data set, and 80% are divided into the training data set. Meanwhile, to make the data for the wild *Gentiana genus* more reasonable, we used stratified sampling for FT-IR spectral data of the wild *Gentiana genus* according to the different locations. The first column shows the different pre-processing methods for FT-IR spectra, where NO_OP indicates the input for the raw FT-IR data (without any pre-processing of FT-IR spectra).

2.4. Models Verification

The specific accuracy and F1_score values for different pre-processing methods for FT-IR spectral data of wild *Gentiana genus* genera are shown in Table 1. Acc in the table denotes accuracy, and F1_score in the table denotes the F1_score of the model. The first column of the table shows the different pre-processing methods for FT-IR spectra of the wild *Gentiana genus*, and the NO_OP means that the FT-IR spectral data input to the model are raw (without any pre-processing of the FT-IR spectra).

Table 1. Performance of each model.

Model	Decision Tree		Naive Bayes		SVM		BP Neural Network		Double-Net (Ours)	
Evaluation Metrics	Acc	F1_Score	Acc	F1_Score	Acc	F1_Score	Acc	F1_Score	Acc	F1_Score
NO_OP	80.00%	0.80	86.67%	0.84	82.20%	0.75	95.56%	0.94	97.78%	0.97
Norm	82.22%	0.79	84.44%	0.85	88.90%	0.86	95.56%	0.93	95.56%	0.94
SG	77.78%	0.78	86.67%	0.84	82.20%	0.75	95.56%	0.94	**100.00%**	**1.00**
MC	80.00%	0.81	88.89%	0.88	84.40%	0.76	95.56%	0.95	95.56%	0.92
WT	80.00%	0.78	82.22%	0.81	82.20%	0.75	95.56%	0.95	95.56%	0.94
MSC	75.56%	0.75	88.89%	0.91	82.20%	0.81	55.56%	0.38	95.56%	0.96
SNV	77.78%	0.76	84.44%	0.87	91.10%	0.88	**97.78%**	**0.97**	**100.00%**	**1.00**
Norm + SG	80.00%	0.75	84.44%	0.85	88.90%	0.86	95.56%	0.93	95.56%	0.94
Norm + MC	82.22%	0.78	88.89%	0.86	88.90%	0.86	91.11%	0.87	95.60%	0.94
Norm + WT	**84.44%**	**0.81**	84.44%	0.85	88.90%	0.86	**97.78%**	**0.97**	**100.00%**	**1.00**
Norm + MSC	26.67%	0.06	64.44%	0.61	26.70%	0.06	26.67%	0.06	95.60%	0.91
Norm + SNV	64.44%	0.59	88.89%	0.87	91.10%	**0.91**	93.33%	0.91	**100.00%**	**1.00**
SG + MC	77.78%	0.81	88.89%	0.88	84.40%	0.76	**97.78%**	**0.97**	97.78%	0.97
SG + WT	77.78%	0.76	82.22%	0.81	82.20%	0.75	95.56%	0.94	**100.00%**	**1.00**
SG + MSC	77.78%	0.77	88.89%	0.91	82.20%	0.81	73.33%	0.65	97.78%	0.97
SG + SNV	80.00%	0.78	84.44%	0.87	91.10%	0.88	95.56%	0.96	95.56%	0.96
MC + WT	71.11%	0.70	**91.10%**	0.88	82.20%	0.73	95.56%	0.94	95.60%	0.94
MC + MSC	75.56%	0.75	88.89%	0.91	82.20%	0.81	26.67%	0.06	95.56%	0.95
MC + SNV	80.00%	0.78	84.44%	0.87	91.10%	0.88	95.56%	0.96	**100.00%**	**1.00**
WT + MSC	71.11%	0.67	88.89%	**0.92**	82.20%	0.81	73.33%	0.67	**100.00%**	**1.00**
WT + SNV	75.56%	0.70	84.44%	0.89	**93.30%**	**0.91**	93.33%	0.94	**100.00%**	**1.00**
MSC + SNV	80.00%	0.79	84.44%	0.87	91.10%	0.88	**97.78%**	**0.97**	95.60%	0.96
Max	84.44%	0.81	91.10%	0.92	93.30%	0.91	97.78%	0.97	100.00%	1.00
Min	26.67%	0.06	64.44%	0.61	26.70%	0.06	26.67%	0.06	95.56%	0.91
Avg	75.35%	0.72	85.45%	0.86	83.62%	0.79	85.46%	0.81	97.48%	0.97

2.4.1. Machine Learning Models

As presented in Table 1, the highest accuracy of the decision tree model for predicting the geographic location of the wild *Gentiana genus* was 84.44%, and the highest F1_score was 0.81; the highest accuracy of the naive Bayes model for predicting the geographic location of the wild *Gentiana genus* was 91.10%, and the highest F1_score was 0.91; the highest accuracy of the SVM model for predicting the geographic location of the wild *Gentiana genus* was 83.3%, and the highest F1_score was 0.91. The accuracy and F1_score values of these machine learning models with different pre-processing methods are shown in Figure 4. Since the FT-IR spectral data we obtained had some noise, after using different spectral pre-processing methods, we achieved different degrees of denoising. For decision tree models, Norm + WT is a better pre-processing method; for the naive Bayes model, WT + MSC and MC + WT are the best pre-processing methods to make the model perform better. For the SVM model, the best pre-processing method is WT + SNV.

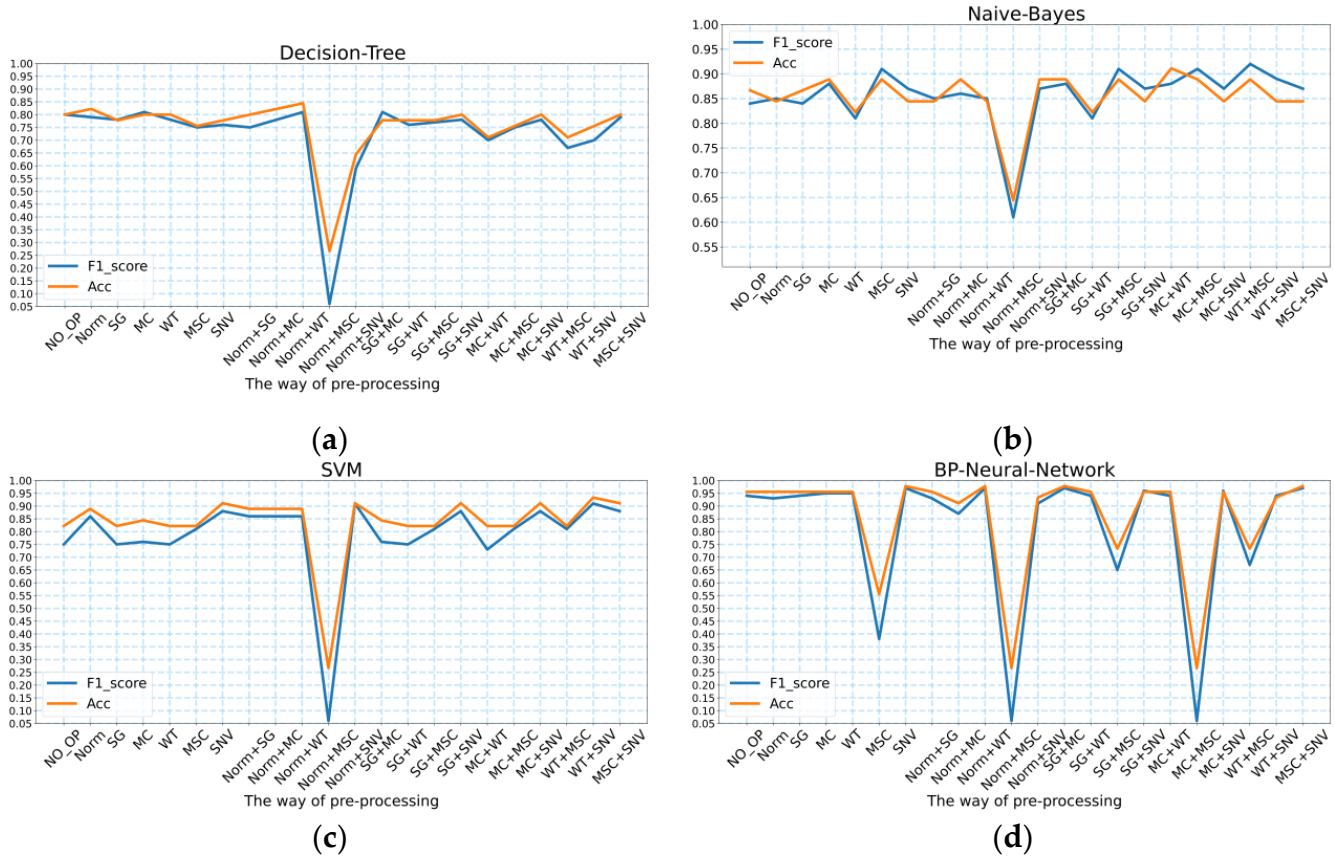


Figure 4. The accuracy and F1_score values of the four models under different pre-processing methods: (**a**) decision tree; (**b**) naive Bayes; (**c**) SVM; (**d**) BP neural network.

2.4.2. BP Neural Network

As shown in Figure 4d and Table 1, the BP neural network model predicted the geographic location of the wild *Gentiana genus* with the highest accuracy value of 97.78% and the highest F1_score of 0.97. For the BP neural network, the effective pre-processing methods for FT-IR spectral data of the wild *Gentiana genus* were SNV, Norm + WT, SG + MC, and MSC + SNV. Compared to the machine learning models, the BP neural networks improved in performance. However, there was a decrease in the BP neural network performance for some pre-processing data methods (e.g., MSC, Norm + MSC, and MC + MSC).

2.4.3. Double-Net

Based on the performance of the BP neural network in predicting the geographic location of the wild *Gentiana genus*, we believe that there is much potential for improvement, and after many experiments, we designed a neural network structure called Double-Net that performs better. As shown in Figure 5 and Table 1, we designed the neural network Double-Net for the geographic location prediction of the wild *Gentiana genus* with the highest accuracy of 100% and the highest F1_score of 1.0. Based on our experiments, for different methods of data pre-processing, the average accuracy of Double-Net can reach 94.48% and the average F1_score can reach 0.97. Compared with the models used in this experiment, the neural network structure Double-Net that we designed performs better than other models for various data processing methods. This proves the effectiveness of our designed neural network structure Double-Net.

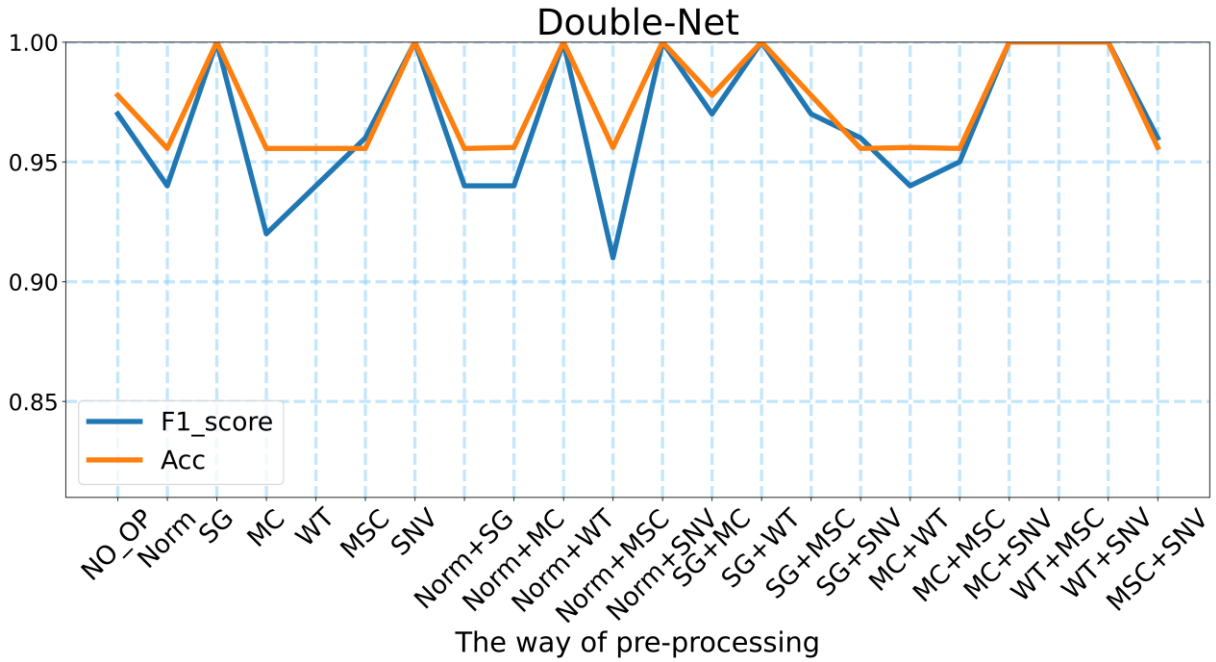


Figure 5. The accuracy and F1_score of Double-Net under different pre-processing methods.

For the neural network Double-Net that we designed, there are 8 FT-IR spectral pre-processing methods that make the model perform the best. These eight methods are SG, SNV, Norm + WT, Norm + SNV, SG + WT, MC + SNV, WT + MSC, and WT + SNV. This means that the pre-processing methods are essential for the FT-IR spectral data. It was also found that our neural network Double-Net performs better in predicting the geographic location of the wild *Gentiana genus* by pre-processing the FT-IR spectral data. In addition, we also found from Table 1 that the accuracy of both the BP neural network and Double-Net can reach more than 95% without FT-IR spectral pre-processing. There are 7468 features being fed into the deep learning model (far more than the number of features fed into the machine learning models). There is no doubt that there are many unimportant features among the 7468 features, which means that the deep learning model needs to select the major features among the 7468 features to predict the geographic location of the wild *Gentiana genus*. This is a powerful reflection of the advantage of deep learning models, which can select the significant features among many features to achieve better performance.

3. Discussion

In previous studies [13,18–22], most of the methods used to accomplish herb location prediction were machine learning algorithms. Machine learning methods (e.g., PLS, SVM) work well for data with few sample features. However, when there are more features in the sample to be processed, the following solutions are commonly used. First, based on human a priori knowledge, the important features are selected among many features and then processed via machine learning algorithms. Second, high-dimensional data are first changed into easy-to-process low-dimensional data by using a dimensionality reduction algorithm, and then processed using certain machine learning methods. Third, an important feature selection algorithm is used to select certain critical features, which to a certain extent achieves the dimensionality reduction, and then certain machine learning algorithms are used for processing. All of these algorithms contain a tedious step, meaning they have to find certain crucial features among the many features to perform the following steps.

However, with the improvement of scientific research, the algorithms related to deep learning have performed well in many fields and solved many challenges. Among the many deep learning algorithms, combined with the relevant data for this experiment, the neural network is a very applicable network model. A neural network is an algorithm that

simulates the human brain, and the use of activation functions such as ReLU, Tanh, and Sigmoid in the neural network can mean the relevant model of the neural network has stronger nonlinear characterization ability [35]. When using an artificial neural network, we can hand over all of the features of the sample to the artificial neural network, which removes the need to manually extract the relevant variable features of the sample. All features of the sample are input into the model of the artificial neural network, which also enables the artificial neural network to have more features to learn, so that the model has a better basis for making decisions. At the same time, this can also avoid the occurrence of poor model results due to feature extraction errors, so that the model can achieve better results. For the task of predicting the geographic location of the wild *Gentiana genus* using FT-IR spectra, because all data from the FT-IR spectra of the wild *Gentiana genus* were input into the deep learning model, such that the deep learning model was able to make full use of the relevant information in the TF-IR spectral data of the wild *Gentiana genus*, which is perhaps one of the reasons for the improved performance of the deep learning model.

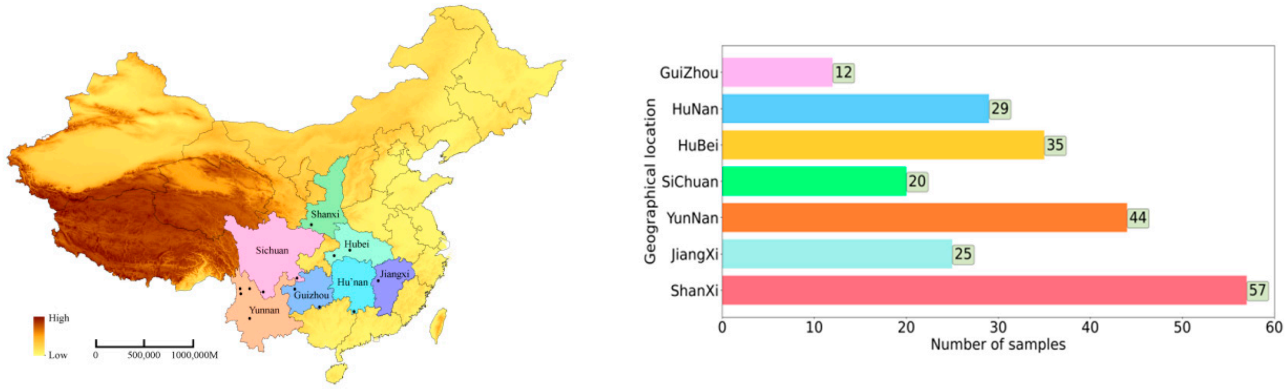
Differing from chemical drugs, herbal products have multi-component and multi-target efficacy, and the resulting drugs are closely related to the soil, climate, harvesting season, and growing age, among other factors. Meanwhile, these factors also affect the contents of herbal ingredients. The information related to specific chemical functional groups (from the stable vibrational state to the first excited state) can be found in FT-IR spectra. For instance, the information related to groups such as C-H, O-H, N-H, and C=O is able to be reflected in the FT-IR spectra. In order to establish a rapid, harmless, efficient, and combined chemometric method to identify the geographic origin of the wild *Gentiana genus*, we collected wild *Gentiana genus* herbs from seven provinces, and we obtained the FT-IR spectral data of wild Gentiana using FT-IR spectroscopy instruments. Finally, we designed a neural network called Double-Net that performed well on the wild *Gentiana genus* dataset, and the network performed well in the geographic location prediction of the wild *Gentiana genus*.

After relevant experiments, we validated the use of FT-IR spectral information to efficiently, rapidly, and nondestructively predict the geographic location of the wild *Gentiana genus* as a workable solution. Given the excellent performance of Double-Net, we believe that this neural network model we designed can be used to roughly identify the geographical location of the wild *Gentiana genus*. In addition to the wild *Gentiana genus*, we believe that Double-Net can also be used for the geographical location identification of other medicinal herbs. The use of FT-IR spectroscopy data enables the nondestructive and rapid identification of the geographical locations of herbs, which is a good way to avoid time-consuming, laborious, and costly losses of herbs. This is good news for the geographical identification of valuable herbs. Moreover, we believe that more work like this can be carried out in the future based on our experiments.

4. Materials and Methods

4.1. Samples Preparation

The 222 roots of wild *Gentiana genus* samples were collected from seven geographical locations in P.R China (Jiangxi, Sichuan, Yunnan, Guizhou, Hubei, Hunan, and Shanxi), as shown in Figure 6a, and the distribution of the number of wild *Gentiana genus* samples at each location is shown in Figure 6b. All wild samples were identified as *Gentiana genus* by Professor Xianxiang Xu (School of medicine, Huaqiao University). All root samples were washed with tap water and were dried in a drying oven at 50 °C, then milled and sifted through 80 mesh sieves. All samples were packed in polyethylene zip-lock bags and stored in a dry environment for a further analysis.



(**a**) (**b**)

Figure 6. The data sources and data distribution: (**a**) seven geographic sampling locations of wild *Gentiana genus* samples; (**b**) the distribution of wild *Gentiana genus* samples in 7 provinces.

4.2. Fourier Transform Infrared (FT-IR) Spectroscopy Analysis

The infrared absorption spectra of samples were recorded using an FT-IR spectrometer equipped with a deuterated triglycine sulfate (DTGS) detector and an ATR (attenuated total reflection) accessory (NICOLET iS 50, Thermo Fisher, Waltham, MA, USA). Typically, the accumulation spectra of 16 scans per sample were collected and averaged. The absorption spectra in the area between 4000 and 400 cm^{-1} with a 4 cm^{-1} resolution were obtained. Three analytical replicates of FT-MIR spectral data of all wild *Gentiana genus* samples were obtained.

4.3. Data Pre-Processing

4.3.1. Raw Spectrum and Its Processing

The raw FT-IR spectra of the wild *Gentiana genus* samples are presented in Figure 7, and the FT-IR spectra of the wild *Gentiana genus* from different regions do not vary much. For some classification models, particularly for machine learning models, it is difficult to find subtle differences to distinguish the wild *Gentiana genus* samples from different areas. In order to improve the robustness and performance of the model we built, we needed to use some data enhancement algorithms to process the raw data before we built the model, such as removing noise from the data, reducing random errors in the data, and eliminating baseline drift interference. For data processing, we used the data processing methods used by Rinnan et al. and Shao et al. in the processing of the FT-IR spectra, such as normalization (Norm), Savitzky–Golay (SG), and wavelet transform (WT) methods, to pre-process our wild *Gentiana genus* spectral data [36,37].

4.3.2. Exploratory Analysis of PCA

Since the machine learning models used in our experiments require the manual extraction of some features to be input to the models, in order to make our input to the machine learning models more objective and make our machine learning models perform better, we used the principal component analysis (PCA) algorithm to reduce the dimensionality of the 7468-dimensional data.

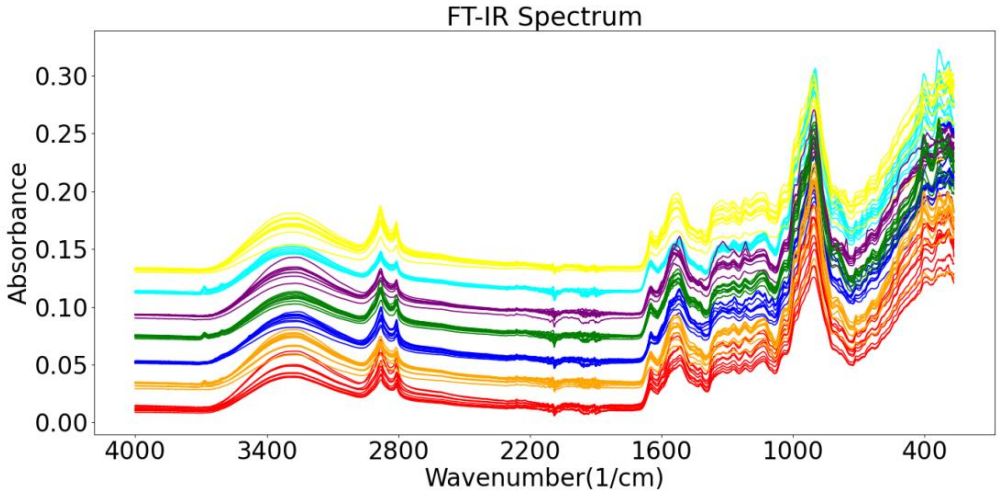


Figure 7. Raw FT-IR spectral data of wild *Gentiana genus*.

4.4. Models

4.4.1. Machine Learning Models

In the experiments, to evaluate the models more objectively, we used models with good results in performance sharing among machine learning methods [38], such as decision trees, plain Bayesian, and support vector machine (SVM) classification models, and compared them with deep learning algorithms and BP neural networks that have good performance in classification. The decision tree is a tree structure built using the entropy of information or the Gini index. For the input features of the decision tree, the more important features will be distributed at the top of the tree structure, and the decision tree model will judge these important input features first and then the other less important input features. After the decision tree model, each sample can be assigned to a specific category. The naive Bayes classification model has a long history and is widely used for spam filtering and news classification. Of course, the naive Bayes classification algorithm can be used not only for the classification of textual data, but also for different classification tasks depending on the target task. In the case of an input sample, the naive Bayes model will use all input features. In prediction, it will predict the probability for each class in the current classification task, and in the prediction result the current sample is predicted for the class with the highest classification probability. The support vector machine algorithm is a supervised learning algorithm. For simple binary classification data, the SVM can learn a maximum-margin hyperplane to achieve the classification of binary data. Meanwhile, the SVM can also classify the data nonlinearly using the kernel method. Based on the principle of the binary classification of the data via SVM, the SVM can also perform certain multiclassification tasks using pairwise classification methods.

4.4.2. BP Neural Networks

Artificial neural networks (ANN), also known as neural networks (NN), are types of algorithms that mimic the structure and function of biological neural networks. There are three neural network layers (input, output, and hidden layers) in a classical neural network. The structure of a classical neural network is shown in Figure 8a, where each blue circle represents a neuron, which is also called a perceptron, and the neurons in a neural network are an imitation of the neurons in the human brain. The structure of a neuron in a neural network is shown in Figure 8b, where X_1, X_2, ..., X_n represents the input to the neuron and W_1, W_2, ..., W_n represents the weight of the current neuron on these inputs to X_1, X_2, ..., X_n; b represents the bias; $F(u)$ represents the activation function (used to complete the nonlinear transformation of the data); u represents the output of the neuron when it is not activated by the activation function; and Y represents the final output of the neuron. The calculation of these variables is shown in Equations (1) and (2).

The data from the neurons in the hidden layers and the output layer of a neural network are calculated from the input data and then they are output. The neurons in a neural network are connected between layers but not connected between neurons in the same layer of the neural network. The final layer, the output layer, is also usually called the fully connected layer.

$$u = \sum_{i=1}^{n} (W_i * X_i + b) \quad (1)$$

$$Y = F(u) \quad (2)$$

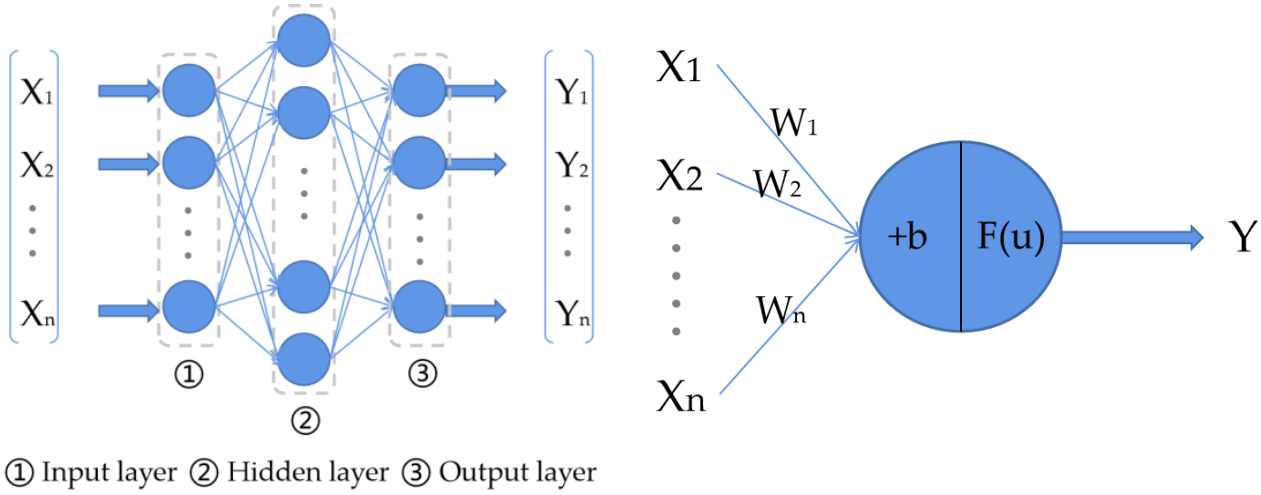


Figure 8. The structure of a classical neural network and a neuron: (**a**) classical neural network with three layers; (**b**) the structure of a neuron in a neural network.

BP neural networks usually have multiple layers of perceptrons between the input and output layers [39], which can fit any linear and nonlinear function. The concept of a neural network was proposed by scientists led by Rumelhart and McClelland in 1986, which was back-propagated according to the error (usually using a gradient descent algorithm to optimize the network). Together with training on a given dataset, it is possible to optimize a BP neural network that can be used for different tasks.

4.4.3. Improved Neural Network Structure (Double-Net)

In order to build a good model for predicting the geographic location of the wild *Gentiana genus*, we used algorithms with good performance in machine learning, such as decision tree, naive Bayes, and support vector machine methods. After experimentation, these models use the FT-IR spectral data of the wild *Gentiana genus* to predict the geographic location of the wild *Gentiana genus*. When the sample data contain a large number of features, the processing performance during deep learning will be better. BP neural networks have achieved good results in many fields through their excellent performance, especially in classification tasks. Considering that every wild *Gentiana genus* sample has 7468 spectral datapoints (the step size is 0.5), we used the BP neural network to complete the prediction of the geographic location of the wild *Gentiana genus*. The BP neural network greatly enhanced the performance of predicting the geographic location of the wild *Gentiana genus*, but the BP neural network did not achieve the desired effect. Consequently, we tried to design a neural network structure to predict the geographical location of the wild *Gentiana genus*.

Inspired by the network structure of the Siamese neural network [40], we designed a neural network structure (Double-Net) with exceptional performance in the task of

predicting the geographic location of the wild *Gentiana genus*. Using our designed neural network structure Double-Net, the accuracy of predicting the geographic locations of wild *Gentiana genus* samples using FT-IR spectroscopy can reach 100%, and our neural network Double-Net structure is shown in Figure 9, where the blue and yellow circles represent the neurons of the neural network, the red circle indicates the summation operation of the parameters of the two neural networks, $X_1, X_2, \ldots, X_n$ indicates the model of the input (FT-IR spectral data of wild *Gentiana genus* samples), and Y indicates the output of the model (geographic location of the wild *Gentiana genus*).

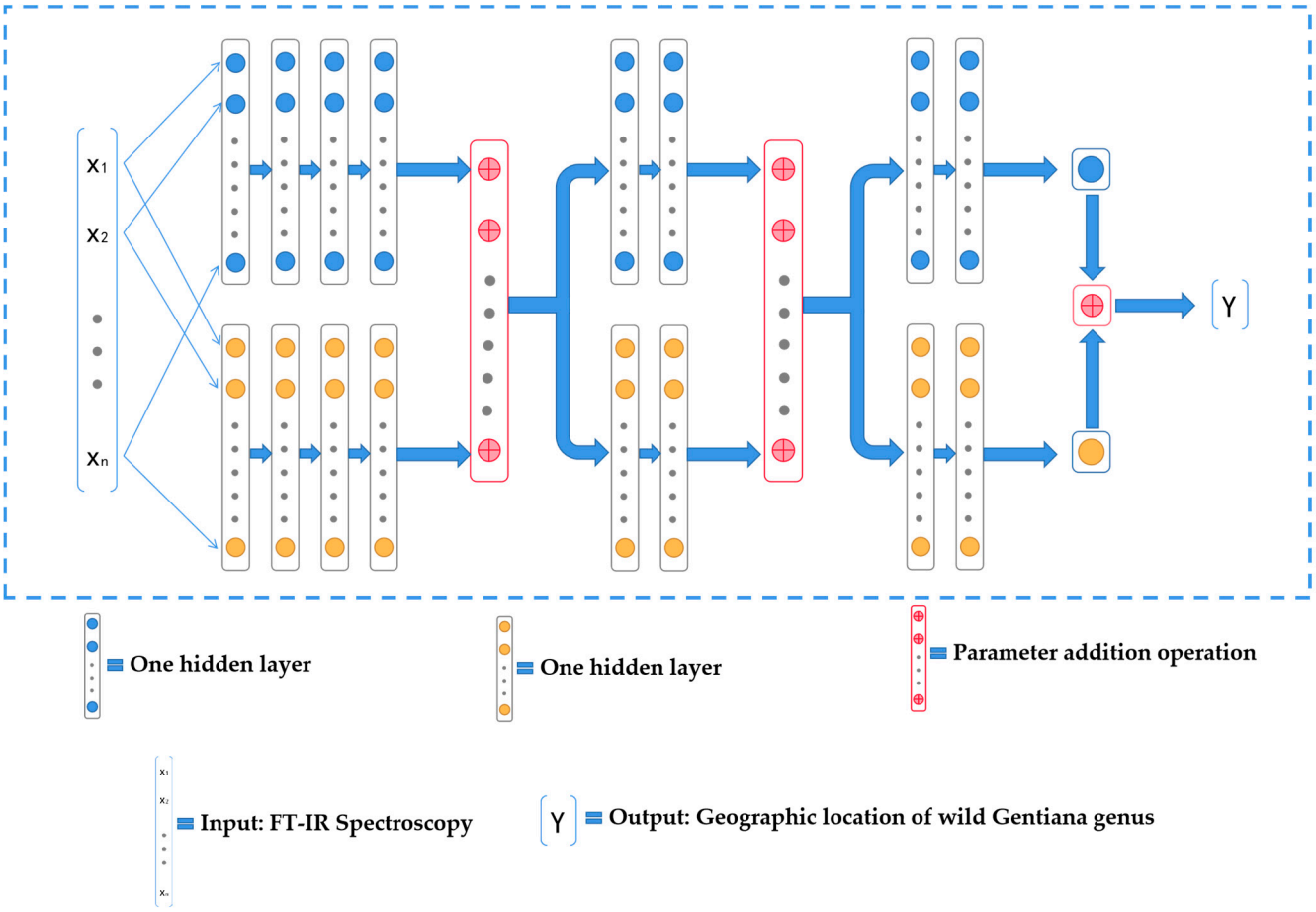


Figure 9. The structure of the neural network Double-Net.

4.5. Evaluation of the Model Performance

We applied multiple model evaluation metrics to evaluate the performance of the models. For the evaluation of the classification task models, the metrics used are usually the precision, recall, F1_score, and accuracy of the model prediction. Through the precision results, we can judge the performance of the model classification process. The closer the accuracy is to 100%, the more efficient the model is. The recall rate indicates whether the model can classify positive samples as positive samples, reflecting the capacity of the model to distinguish between each type of sample. The calculation formulas for precision and recall are shown in Equations (3) and (4), where TP (true positive) indicates the number of correct predictions by the model for the true location of the sample. FP (false positive) represents the number of incorrect predictions by the model for the location of the negative sample, whereby the sample is incorrectly predicted as a positive sample from a certain location, but the sample is actually a sample from another location. TN (true negative) is the number of negative samples that the model predicts properly. FN (false negative) represents the number of errors in the model's prediction of positive samples, whereby the

sample is predicted to be a positive sample from other locations, but the sample is actually a positive sample from the current location:

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

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

The F1_score is a comprehensive metric of the precision and recall, which is a better metric of the model performance than the precision and recall, so we adopted the F1_score as one of the metrics of our model (the formula of F1_score is Equation (5). Furthermore, the accuracy is also a measure of the model performance for classification models, and the accuracy indicates the number of correct predictions among all samples (including positive and negative samples). The accuracy is used in our model evaluation metrics, and the calculation of the accuracy is shown in Equation (6):

$$\text{F1\_Score} = \frac{\text{Precison} * \text{Recall}}{2 * (\text{Precision} + \text{Recall})} \tag{5}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{6}$$

4.6. Software

This experiment was based on the Windows 10 operating system. The FT-IR spectra were processed using Omnic (Version 8.2, Thermo Fisher Scientific, Madison, WI, USA). All models were created using PyCharm (version 2021 professional), and the pre-processing of the FT-IR data was also done using PyCharm (version 2021 professional). The programming environment was Python 3.7, and the deep learning framework used in our study was PyTorch 1.7.

5. Conclusions

In this study, the geographical location of the wild *Gentiana genus* was predicted using benchtop FT-IR spectroscopy coupled with an improved neural network structure Double-Net. Here, 21 FT-IR spectral data pre-processing methods and 5 efficient algorithms were used for comparison and evaluation. The experiments showed that our improved neural network structure, Double-Net, is the optimal model for predicting the geographic locations of wild *Gentiana genus* samples. Our improved neural network structure, Double-Net, achieved 100% accuracy and an F1_score of 1.0 on the test dataset of the wild *Gentiana genus*. This means that it can be used to establish a rapid, nondestructive, and efficient method for the identification of the geographical locations of wild *Gentiana genus* plants combined with chemometrics. Given that this experiment is a preliminary study, we believe that FT-IR spectroscopy can be used to explore the geographical locations of more traditional Chinese medicines in the future.

Author Contributions: Conceptualization, Y.D. and X.W.; methodology, Y.D. and X.W.; software, P.Z. and Y.L.; validation, P.Z.; formal analysis, X.L.; resources, X.W., Y.D. and P.L.; data curation, Y.L.; writing—original draft preparation, P.Z. and X.L.; writing—review and editing, X.W. and P.L.; visualization, P.Z. and Y.L.; supervision, Y.D. and P.L.; project administration, P.L.; funding acquisition, P.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Fujian Provincial Science and Technology Major Project (No.2020HZ02014), the Quanzhou Science and Technology Major Project (No.2021GZ1), grants from the National Natural Science Foundation of Fujian (2021J011404, 2021J01408), and the Quanzhou Scientific and Technological Planning Projects (2021C037R, 2019C028R).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Jiang, M.; Cui, B.W.; Wu, Y.L.; Nan, J.X.; Lian, L.H. Genus Gentiana: A review on phytochemistry, pharmacology and molecular mechanism. *J. Ethnopharmacol.* **2021**, *264*, 113391. [CrossRef] [PubMed]
2. Wan, Z.; Li, H.; Wu, X.; Zhao, H.; Wang, R.; Li, M.; Liu, J.; Liu, Q.; Wang, R.; Li, X. Hepatoprotective effect of gentiopicroside in combination with leflunomide and/or methotrexate in arthritic rats. *Life Sci.* **2021**, *265*, 118689. [CrossRef] [PubMed]
3. Xiao, H.; Sun, X.; Lin, Z.; Yang, Y.; Zhang, M.; Xu, Z.; Liu, P.; Liu, Z.; Huang, H. Gentiopicroside targets PAQR3 to activate the PI3K/AKT signaling pathway and ameliorate disordered glucose and lipid metabolism. *Acta Pharm. Sin. B* **2022**, *12*, 2887–2904. [CrossRef] [PubMed]
4. Jia, N.; Ma, H.; Zhang, T.; Wang, L.; Cui, J.; Zha, Y.; Ding, Y.; Wang, J. Gentiopicroside attenuates collagen-induced arthritis in mice via modulating the CD147/p38/NF-κB pathway. *Int. Immunopharmacol.* **2022**, *108*, 108854. [CrossRef] [PubMed]
5. Zheng, Y.; Fang, D.; Huang, C.; Zhao, L.; Gan, L.; Chen, Y.; Liu, F. *Gentiana scabra* Restrains Hepatic Pro-Inflammatory Macrophages to Ameliorate Non-Alcoholic Fatty Liver Disease. *Front. Pharmacol.* **2022**, *12*, 816032. [CrossRef]
6. Xiong, X.J.; Yang, X.C.; Liu, W.; Duan, L.; Wang, P.Q.; You, H.; Li, X.K.; Wang, S. Therapeutic Efficacy and Safety of Traditional Chinese Medicine Classic Herbal Formula *Longdanxiegan* Decoction for Hypertension: A Systematic Review and Meta-Analysis. *Front. Pharmacol.* **2018**, *9*, 466. [CrossRef] [PubMed]
7. Olennikov, D.N.; Kashchenko, N.I.; Chirikova, N.K.; Koryakina, L.P.; Vladimirov, L.N. Bitter Gentian Teas: Nutritional and Phytochemical Profiles, Polysaccharide Characterisation and Bioactivity. *Molecules* **2015**, *20*, 20014–20030. [CrossRef]
8. Wang, Z.; Wang, C.; Su, T.; Zhang, J. Antioxidant and immunological activities of polysaccharides from Gentiana scabra Bunge roots. *Carbohydr. Polym.* **2014**, *112*, 114–118. [CrossRef]
9. Guedes, L.; Reis, P.B.P.S.; Machuqueiro, M.; Ressaissi, A.; Pacheco, R.; Serralheiro, M.L. Bioactivities of *Centaurium erythraea* (Gentianaceae) Decoctions: Antioxidant Activity, Enzyme Inhibition and Docking Studies. *Molecules* **2019**, *24*, 3795. [CrossRef]
10. Dai, W.; Yang, Y.; Patch, H.M.; Grozinger, C.M.; Mu, J. Soil moisture affects plant-pollinator interactions in an annual flowering plant. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **2022**, *377*, 20210423. [CrossRef]
11. Hou, Q.Z.; Ur Rahman, N.; Ali, A.; Wang, Y.P.; Shah, S.; Nurbiye, E.; Shao, W.J.; Ilyas, M.; Sun, K.; Li, R.; et al. Range expansion decreases the reproductive fitness of Gentiana officinalis (Gentianaceae). *Sci. Rep.* **2022**, *12*, 2461. [CrossRef] [PubMed]
12. Zhang, J.; Wang, Y.Z.; Gao, H.K.; Zuo, Z.T.; Yang, S.B.; Cai, C.T. Different strategies in biomass allocation across elevation in two Gentiana plants on the Yunnan-Guizhou Plateau, China. *J. Mt. Sci.* **2020**, *17*, 2750–2757. [CrossRef]
13. Wu, Z.; Zhao, Y.; Zhang, J.; Wang, Y. Quality Assessment of Gentiana rigescens from Different Geographical Origins Using FT-IR Spectroscopy Combined with HPLC. *Molecules* **2017**, *22*, 1238. [CrossRef] [PubMed]
14. Zhang, M.; Jiang, D.; Yang, M.; Ma, T.; Ding, F.; Hao, M.; Chen, Y.; Zhang, C.; Zhang, X.; Li, M. Influence of the Environment on the Distribution and Quality of Gentiana dahurica Fisch. *Front. Plant Sci.* **2021**, *12*, 706822. [CrossRef] [PubMed]
15. Sasaki, N.; Nemoto, K.; Nishizaki, Y.; Sugimoto, N.; Tasaki, K.; Watanabe, A.; Goto, F.; Higuchi, A.; Morgan, E.; Hikage, T.; et al. Identification and characterization of xanthone biosynthetic genes contributing to the vivid red coloration of red-flowered gentian. *Plant J.* **2021**, *107*, 1711–1723. [CrossRef]
16. Pan, Z.; Xiong, F.; Chen, Y.-L.; Wan, G.-G.; Zhang, Y.; Chen, Z.-W.; Cao, W.-F.; Zhou, G.-Y.J.M. Traceability of geographical origin in Gentiana straminea by UPLC-Q exactive mass and multivariate analyses. *Molecules* **2019**, *24*, 4478. [CrossRef]
17. Khalil, A.; Kashif, M. Nuclear Magnetic Resonance Spectroscopy for Quantitative Analysis: A Review for Its Application in the Chemical, Pharmaceutical and Medicinal Domains. *Crit. Rev. Anal. Chem.* **2021**, 1–15. [CrossRef]
18. Wu, X.M.; Zhang, Q.Z.; Wang, Y.Z. Traceability of wild *Paris polyphylla* Smith var. yunnanensis based on data fusion strategy of FT-MIR and UV-Vis combined with SVM and random forest. *Spectrochim. Acta A Mol. Biomol. Spectrosc.* **2018**, *205*, 479–488. [CrossRef]
19. Yao, S.; Li, T.; Li, J.; Liu, H.; Wang, Y. Geographic identification of Boletus mushrooms by data fusion of FT-IR and UV spectroscopies combined with multivariate statistical analysis. *Spectrochim. Acta A Mol. Biomol. Spectrosc.* **2018**, *198*, 257–263. [CrossRef]
20. Mousa, M.A.A.; Wang, Y.; Antora, S.A.; Al-Qurashi, A.D.; Ibrahim, O.H.M.; He, H.J.; Liu, S.; Kamruzzaman, M. An overview of recent advances and applications of FT-IR spectroscopy for quality, authenticity, and adulteration detection in edible oils. *Crit. Rev. Food Sci. Nutr.* **2021**, 1–19. [CrossRef]
21. Zareef, M.; Arslan, M.; Mehedi Hassan, M.; Ali, S.; Ouyang, Q.; Li, H.; Wu, X.; Muhammad Hashim, M.; Javaria, S.; Chen, Q. Application of benchtop NIR spectroscopy coupled with multivariate analysis for rapid prediction of antioxidant properties of walnut (*Juglans regia*). *Food Chem.* **2021**, *359*, 129928. [CrossRef]
22. Zhao, Y.; Yuan, T.; Wu, L.; Zuo, Z.; Wang, Y. I Identification of *Gentiana rigescens* from different geographical origins based on HPLC and FTIR fingerprints. *Anal. Methods* **2020**, *12*, 2260–2271. [CrossRef]
23. Pei, Y.F.; Zuo, Z.T.; Zhang, Q.Z.; Wang, Y.Z. Data Fusion of Fourier Transform Mid-Infrared (MIR) and Near-Infrared (NIR) Spectroscopies to Identify Geographical Origin of Wild *Paris polyphylla* var. *yunnanensis*. *Molecules* **2019**, *24*, 2559. [CrossRef] [PubMed]

24. Liu, W.J.; Li, W.J.; Qin, H.; Li, H.G.; Ning, X. Research on identifying maize haploid seeds using near infrared spectroscopy based on kernel locality preserving projection. *Spectrosc. Spect. Anal.* **2019**, *39*, 2574–2577. [CrossRef]
25. Liu, W.J.; Li, W.J.; Li, H.G.; Qin, H.; Xin, N. Research on the method of identifying maize haploid based on KPCA and near infrared. *Spectrosc. Spect. Anal.* **2017**, *37*, 2024–2027. [CrossRef]
26. Hillel, T.; Bierlaire, M.; Elshafie, M.Z.; Jin, Y. A systematic review of machine learning classification methodologies for modelling passenger mode choice. *J. Choice Modell.* **2021**, *38*, 100221. [CrossRef]
27. Ding, S.; Li, H.; Su, C.; Yu, J.; Jin, F. Evolutionary artificial neural networks: A review. *Artif. Intell. Rev.* **2013**, *39*, 251–260. [CrossRef]
28. Sarker, I.H. Deep cybersecurity: A comprehensive overview from neural network and deep learning perspective. *SN Comput. Sci.* **2021**, *2*, 1–16. [CrossRef]
29. Mutlu, A.C.; Boyaci, I.H.; Genis, H.E.; Ozturk, R.; Basaran-Akgul, N.; Sanal, T.; Evlice, A.K. Prediction of wheat quality parameters using near-infrared spectroscopy and artificial neural networks. *Eur. Food Res. Technol.* **2011**, *233*, 267–274. [CrossRef]
30. Gonzalez Viejo, C.; Fuentes, S.; Torrico, D.; Howell, K.; Dunshea, F.R. Assessment of beer quality based on foamability and chemical composition using computer vision algorithms, near infrared spectroscopy and machine learning algorithms. *J. Sci. Food Agric.* **2018**, *98*, 618–627. [CrossRef]
31. Qie, X.; Kang, C.; Zong, G.; Chen, S. Trajectory Planning and Simulation Study of Redundant Robotic Arm for Upper Limb Rehabilitation Based on Back Propagation Neural Network and Genetic Algorithm. *Sensors* **2022**, *22*, 4071. [CrossRef] [PubMed]
32. Chen, J.; Yan, M.; Zhu, F.; Xu, J.; Li, H.; Sun, X. Fatigue Driving Detection Method Based on Combination of BP Neural Network and Time Cumulative Effect. *Sensors* **2022**, *22*, 4717. [CrossRef] [PubMed]
33. Zojaji, I.; Esfandiarian, A.; Taheri-Shakib, J. Toward molecular characterization of asphaltene from different origins under different conditions by means of FT-IR spectroscopy. *Adv. Colloid Interface Sci.* **2021**, *289*, 102314. [CrossRef] [PubMed]
34. Liu, L.; Zuo, Z.T.; Xu, F.R.; Wang, Y.Z. Study on Quality Response to Environmental Factors and Geographical Traceability of Wild Gentiana rigescens Franch. *Front. Plant Sci.* **2020**, *11*, 1128. [CrossRef] [PubMed]
35. Dubey, S.R.; Singh, S.K.; Chaudhuri, B.B. Activation Functions in Deep Learning: A comprehensive Survey and Benchmark. *Neurocomputing* **2022**, *503*, 92–108. [CrossRef]
36. Rinnan, Å.; Van Den Berg, F.; Engelsen, S.B. Review of the most common pre-processing techniques for near-infrared spectra. *Trac-Trend Anal. Chem.* **2009**, *28*, 1201–1222. [CrossRef]
37. Shao, X.; Zhuang, Y. Determination of chlorogenic acid in plant samples by using near-infrared spectrum with wavelet transform preprocessing. *Anal. Sci.* **2004**, *20*, 451–454. [CrossRef]
38. Soofi, A.A.; Awan, A. Classification techniques in machine learning: Applications and issues. *J. Basic Appl. Sci.* **2017**, *13*, 459–465. [CrossRef]
39. Rumelhart, D.E.; Durbin, R.; Golden, R.; Chauvin, Y. Backpropagation: The basic theory. In *Backpropagation: Theory, Architectures and Applications*; Chauvin, Y., Rumelhart, D.E., Eds.; Lawrence Brlbaum Associates: Hillsdale, NJ, USA, 1995; pp. 1–34.
40. Chopra, S.; Hadsell, R.; LeCun, Y. Learning a similarity metric discriminatively, with application to face verification. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; pp. 539–546.