

Comparative analysis of data preprocessing methods and machine learning models for geographical origin prediction in an imbalanced Panax notoginseng dataset using near-infrared spectroscopy

XueFeng Cheng

Chongqing City Management College

Abudhahir Buhari

Infrastructure University Kuala Lumpur

Juan Liu (✉ liujuan@cque.edu.cn)

Chongqing University of Education

Article

Keywords:

Posted Date: January 18th, 2024

DOI: <https://doi.org/10.21203/rs.3.rs-3830586/v1>

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

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

This study explores the application of near-infrared spectroscopy (NIRS) and machine learning to accurately determine the geographical origin of *Panax notoginseng* (*P. notoginseng*), a critical component in traditional Chinese medicine. Given the complexity of *P. notoginseng* geographical origin identification, especially in the face of imbalanced datasets, the study systematically evaluates a range of data preprocessing methods, including autocorrelation, data standardization, Multiplicative Scatter Correction (MSC), Standard Normal Variate (SNV), Savitzky-Golay (S-G) smoothing, first-order derivative (1D), second-order derivative (2D), and Principal Component Analysis (PCA). Furthermore, it assesses various machine learning models such as Gaussian Naive Bayes (GNB), K-Nearest Neighbors (KNN), Classification and Regression Trees (CART), Support Vector Machine (SVM), Linear Regression (LR), and neural networks in this context. First by assembling and preparing a substantial dataset of NIRS of *P. notoginseng* from different geographical locations. The dataset's imbalance, reflective of real-world scenarios, necessitates specialized data handling strategies. The study meticulously applies each preprocessing technique to this dataset, followed by the deployment of different machine learning models. This dual approach allows for an in-depth comparison of how each combination influences the accuracy of geographical origin prediction. Findings of the study reveal that specific combinations of data preprocessing methods and machine learning models yield substantial improvements in predicting the geographical origin of *P. notoginseng*. These combinations are pivotal in addressing the imbalances inherent in the dataset, thereby enhancing the reliability of the predictions. The research contributes significantly to the field by not only providing a solution to the problem of geographical origin prediction in imbalanced datasets but also by laying down a methodological framework that can be adapted for similar challenges in the broader area of herbal medicine research. This study serves as a cornerstone in the intersection of traditional Chinese medicine and modern scientific methods, offering a robust, data-driven approach to ensuring the authenticity and quality of vital medicinal herbs like *P. notoginseng*. Its implications extend beyond this specific application, providing insights and methodologies that could revolutionize quality control and authentication processes in herbal medicine globally.

Introduction

P. notoginseng (Sanqi) belongs to the Araliaceae family and the Panax genus. It is a valuable and historically significant Chinese herb^{1,2}. Since ancient times, *P. notoginseng* has been widely used in traditional Chinese medicine and is considered an important medicinal herb, second only to ginseng. It has a wide range of applications and is extensively used in the treatment of various conditions, including injuries, coronary heart disease, angina pectoris, sequelae of cerebral infarction, and hypertension. *P. notoginseng* is also a key ingredient in major Chinese patent medicines such as Compound Danshen Dripping Pills, Yunnan Baiyao, Xuesaitong, and Pien Tze Huang.

P. notoginseng possesses multiple pharmacological activities, including anti-inflammatory, anti-tumor, anti-atherosclerosis, blood pressure-lowering, and blood glucose-lowering effects. In the Qing Dynasty, the medical expert Zhao Xuemin recorded in "Bengang Mushi" that ginseng is known for its Qi-nourishing

properties, while *P. notoginseng* is revered as the top herb for blood tonification. Their efficacies are considered equivalent, leading to the name "Renshen Sanqi", which represents some of the most precious herbs in traditional Chinese medicine. *P. notoginseng* has a warm nature with a sweet and slightly bitter taste. It enters the liver and stomach meridians, exhibiting the effects of blood tonification, anti-inflammation, and blood circulation promotion. It also possesses antioxidant and anti-aging properties. Furthermore, *P. notoginseng* can dilate blood vessels, reduce coronary resistance, increase coronary blood flow, improve coronary microcirculation, increase nutritive blood flow to the myocardium, lower arterial pressure, slightly reduce heart rate, and significantly decrease myocardial oxygen consumption.

Currently, the market demand for *P. notoginseng* is increasing year by year. Domestic demand in China exceeds 20,000 tons, while international demand is close to 300 tons. There are approximately 400 types and over 1,000 pharmaceutical companies that require *P. notoginseng* as a raw material. The cultivated area of *P. notoginseng* remains at around 500,000 mu (approximately 82,644 acres), with a total production exceeding 35,000 tons. The sales revenue from cultivation reaches 4.8 billion RMB, while the processing industry sales revenue amounts to 1.3 billion RMB, and the circulation industry sales revenue is approximately 7.4 billion RMB.

P. notoginseng can only grow in lands within 23.5° north latitude and at altitudes ranging from 1000 to 2500 meters. It is a shade-loving plant that prefers warm winters and cool summers while being sensitive to extreme cold and heat. It thrives in shaded areas but is susceptible to waterlogging; the optimal soil moisture content ranges from 22–40%. The summer temperature should not exceed 35 degrees Celsius, and the winter temperature should not drop below – 5 degrees Celsius. Loose, well-drained sandy loam soil is preferable, while excessively sticky or sandy soil, as well as low-lying areas prone to water accumulation, are unsuitable for cultivation. Crop rotation is discouraged, and the soil pH should be maintained between 4.5 and 8.

Given these specific requirements, *P. notoginseng* can only flourish in a limited region, primarily in Yunnan's Wenshan area and its surrounding regions. This unique geographic location, combined with favourable light, temperature, water, and soil conditions, makes Yunnan's Wenshan the primary production region for *P. notoginseng*. The convergence of these factors determines why Yunnan's Wenshan has become the ideal habitat for cultivating *P. notoginseng*.

In China, *P. notoginseng* cultivation is widely distributed in provinces such as Yunnan, Guangxi, Sichuan, Guizhou, and Chongqing, the distribution of *P. notoginseng* production areas in China is shown in right side of Fig. 1. The regions of Wenshan, Honghe, Kunming, Qujing, and Yuxi in Yunnan province are recognized as the main production areas of *P. notoginseng*, contributing to 98% of the total national production, Wenshan's *P. notoginseng* production accounts for 60% of these five regions, left side of Fig. 1 shows the distribution of the five majors *P. notoginseng* production areas in Yunnan province.

Due to the generally high medicinal value of *P. notoginseng*, the difficulty of planting, and the high price, a large number of counterfeit and shoddy *P. notoginseng* products appear on the market, which damages the interests of consumers and endangers consumer health. Therefore, innovative research on quality

control methods of *P. notoginseng*, employing modern scientific theories and high-tech means, is necessary. Effective methods for rapid quality identification of *P. notoginseng*, real-time monitoring and analysis of index components, and timely feedback of quality information are needed to achieve optimal control. Improving the quality control system and level of *P. notoginseng* is crucial to enhance the core competitiveness of the traditional Chinese medicinal industry, promote the internationalization and modernization of traditional Chinese medicine, and ensure the quality uniformity and controllability of traditional Chinese medicinal products. This research has become a major scientific frontier in the field of traditional Chinese medicinal materials³⁻¹⁵.

Zhou¹¹ et al. presents a classification model (Random Forest) for the geographical origin identification of *P. notoginseng*, and the classification accuracy is 95.6%. Liu¹⁵ et al. investigated traceability of geographical origins of *P. notoginseng* using NIR data combined with the deep learning model, and the identification accuracy is 100%. Cui⁶ et al. established to develop a rapid method (Partial Least Squares Discriminant Analysis) for the geographical origin identification of *P. notoginseng*, and the accuracy of the training set and the testing set was 97.5% and 98.4%. Zhang¹⁶ et al. reports to identify the origins of *P. notoginseng* using terahertz spectroscopy with machine learning (Improved Whale Optimization Algorithm Optimized Support Vector Machine), and the accuracy of SVM model was 93.75%. Ji¹² et al. proposed to identify the origins of *P. notoginseng* using Mass Spectroscopy with machine learning (SVM), and the best accuracy was 92.59%.

While the studies on *P. notoginseng* geographical origin identification yield promising results, a comprehensive analysis that includes factors like sample distribution and model comparison is imperative. Such scrutiny is essential for a nuanced understanding of how these factors impact prediction models' accuracy and reliability. Regrettably, the referenced studies fall short of providing a thorough analysis of these critical elements, thereby limiting overall comprehension of the models' effectiveness and reliability¹⁷⁻²⁶. Integrating these considerations into future research will significantly contribute to advancing the understanding and practical application of prediction models in the geographical origin identification of *P. notoginseng*.

Sample distribution emerges as a pivotal factor in geographical origin identification, demanding a representative spread of *P. notoginseng* samples from diverse regions. This approach proves crucial to capturing the inherent variations associated with different geographical origins, preventing the emergence of biased or skewed models that may excel in one region but falter in others²⁷⁻³⁸.

The ramifications of imbalanced sample distribution on geographical origin identification within *P. notoginseng* datasets are profound, significantly impacting the performance of machine learning models³⁹⁻⁵². Biased model training in favor of the majority class jeopardizes sensitivity to minority classes, hindering the models' ability to generalize to underrepresented origins. This imbalance elevates the risk of misclassifications and amplifies susceptibility to sampling noise. Furthermore, challenges in

capturing class-specific patterns, coupled with evaluation metrics favoring the majority class, contribute to classification errors and diminished overall model accuracy.

To address these challenges, employing dataset balancing techniques, such as oversampling or undersampling, proves indispensable, enhancing the model's precision in identifying geographical origins. Researchers must also adopt suitable evaluation metrics that provide nuanced insights into model performance across different classes, ensuring a comprehensive understanding of the model's reliability in geographical origin identification studies. In essence, meticulous attention to and management of sample distribution stand as imperative elements for fostering the development of robust machine learning models with practical applicability.

Effective model comparison is essential for identifying the optimal approach in geographical origin identification. Through a thorough evaluation of multiple models on the same dataset, researchers can uncover each model's strengths, weaknesses, and overall performance. This critical step empowers both researchers and practitioners to make informed decisions, selecting the most accurate and reliable model for practical implementation.

In the context of imbalanced datasets, the importance of comparing multiple models becomes even more pronounced. Different models employ unique learning and feature processing strategies. By comparing various models, researchers gain insights into the distinctions in feature recognition across diverse geographical origins, facilitating the selection of the most fitting model. Moreover, comparing models helps assess the effectiveness of different strategies for handling imbalances, determining the optimal approach. This comprehensive approach also allows for the evaluation of models' robustness to noise and outliers, thereby improving overall performance and generalization. Therefore, the process of comparing multiple models equips researchers with a holistic perspective, guiding the choice of the most suitable model to address classification challenges in imbalanced datasets.

In addition, the evolutionary history of geographical origin identification methods for *P. notoginseng* is a complex and diverse process, involving advancements in technology, continuous innovation in research methods, and a deepening understanding of identification issues. In recent years, the application of NIRS technology and machine learning techniques has provided a rapid and non-destructive method for the geographical origin identification of *P. notoginseng*. Through the analysis of NIRS of *P. notoginseng* samples, spectral patterns can be obtained. Machine learning, by training models, can learn and extract patterns from complex datasets, offering an intelligent and efficient approach for geographical origin identification.

This study is based on an imbalanced dataset of *P. notoginseng* NIRS and explores the classification accuracy of geographical origins using different data preprocessing algorithms and machine learning models to address practical challenges. The data preprocessing algorithms encompass autocorrelation, data standardization, MSC, SNV, S-G smoothing, 1D, 2D, and PCA. Machine learning models include GNB, KNN, CART, SVM, LR, Multisource Perception Model (MPM), Backpropagation Neural Network (BPNN), Convolutional Neural Network (CNN), and Deep Neural Network (DNN). Initially, NIRS of *P. notoginseng*

samples were obtained using an infrared spectrometer. Subsequently, the distribution of *P. notoginseng* NIRS data samples were analyzed. Based on this, data preprocessing techniques were applied to preprocess and extract features from the *P. notoginseng* NIRS data. Finally, multiple machine learning models were employed to identify the geographical origins of *P. notoginseng*, and the classification accuracy metric was utilized to evaluate the performance of each model.

Material and methods

Sample Preparation. A total of 98 main root specimens of *P. notoginseng* samples were collected from five different origins in Yunnan province, Wenshan (25), Honghe (27), Kunming (9), Qujing (25), and Yuxi(12). The main roots of all samples were washed, dried, crushed, and passed through 60 mesh screens. The powders were stored at room temperature in plastic sealed pouches for further analysis.

In these samples, the distribution of the number of specimens from each origin is uneven, with more samples from Wenshan, Honghe, and Qujing, relatively fewer samples from Kunming and Yuxi. Specifically, the number of samples from Wenshan, Honghe, and Qujing are 25, 27, and 25, totaling 77 samples, while Kunming has only 9 samples and Yuxi has 12 samples. Such differences in sample distribution result in an imbalanced dataset, where the number of samples from certain origins is significantly lower than others, while other origins have relatively more samples. In the context of machine learning or statistical modeling, this imbalance may impact the performance of models, as models might more easily learn patterns from origins with a larger number of samples, leading to poorer performance on origins with fewer samples. The powders were stored at room temperature in plastic sealed pouches for further analysis.

Spectrum Collection. The NIR spectra of all *P. notoginseng* samples were recorded at interval of 1.0 cm^{-1} over the region of $10,000\text{--}4004\text{ cm}^{-1}$ using a Bruker Matrix-F Fourier transform NIR spectrometer (Bruker Optics Inc., Germany) equipped with an integrating sphere. The NIR spectra were the average of 64 scans with air as the reference standard. Each spectrum was measured in the reflective mode and recorded as the logarithm of the reciprocal reflectance, as shown in Fig. 2.

Theoretical Basis. This paper primarily analyzes the geographical origin of an imbalanced dataset of *P. notoginseng* NIRS from two dimensions: data preprocessing methods and machine learning models. The data preprocessing methods employed include autocorrelation, data standardization, MSC, SNV, S-G smoothing, 1D, 2D, and PCA. The selected machine learning models consist of GNB, KNN, CART, SVM, LR, MPM, BPNN, CNN, and DNN. The following content provides brief descriptions of the aforementioned methods.

In data analysis, autocorrelation is a method used to detect the correlation between observations in time series data. Through the autocorrelation function, one can assess the degree of correlation between observations at different time points within the same time series, revealing trends, periodicity, and model features in the data. In data preprocessing, the roles of autocorrelation include trend detection, periodicity

analysis, model identification, and outlier detection. Analyzing autocorrelation aids in selecting appropriate preprocessing methods for subsequent analysis or modeling, ensuring data stability and adaptability.

Data standardization is a data preprocessing algorithm designed to eliminate scale differences among different features, ensuring a relatively balanced impact of each feature on the model. Through data standardization, the analysis is conducted on data with the same scale, enhancing model convergence speed, stability, and overall accuracy and reliability of data analysis.

MSC is a data preprocessing algorithm primarily employed to eliminate scatter effects in spectral data, enhancing the quality and accuracy of the data. In data analysis, MSC achieves this by applying mathematical transformations to the data, eliminating spectral shifts caused by particle scattering and path differences. This process renders the data more reliable and comparable, playing a crucial role in correcting systematic errors in spectra and improving model performance.

SNV is a commonly used data preprocessing algorithm, primarily designed to remove concentration effects and variability in spectral intensity in spectral data, making the data more stable and comparable. Through SNV, spectral data undergoes standardization, eliminating spectral shifts caused by differences in concentration or instrument variations, thereby improving the accuracy and robustness of models. This method is frequently employed in spectroscopic analysis to ensure better comparability of spectral data acquired under different conditions.

S-G smoothing is a commonly used data preprocessing algorithm primarily employed for smoothing spectral data, reducing noise, and minimizing fluctuations. By applying local polynomial fitting to the data, S-G smoothing helps eliminate high-frequency noise while preserving the overall trend in the data. In data analysis, this smoothing method enhances data readability, reduces sensitivity to noise, and facilitates more accurate extraction of features and patterns from the data. The typical parameter setting is a five-point three-order.

1D and 2D are common data preprocessing algorithms used to highlight or enhance the trends in data analysis. The 1D is employed to detect changes in the slope of the data, emphasizing the overall trend, while the 2D places more emphasis on the curvature of the data, aiding in capturing local details and inflection points. In certain scenarios, both derivative operations can be applied to reduce noise in the data, directing the data analysis towards critical points of change.

PCA is a commonly used data preprocessing algorithm that maps the original data to a new coordinate system through linear transformation, aiming to better capture the main directions of variation in the data. In data analysis, PCA is widely employed for dimensionality reduction, reducing data complexity, and emphasizing key information. By retaining the most significant principal components, PCA aids in improving model efficiency, reducing redundancy, while preserving the essential features of the original data.

GNB algorithm is known for its simplicity and efficiency. This algorithm, based on Bayes' theorem, assumes independence between features and models the probability distribution of features using Gaussian distributions. GNB performs well in handling high-dimensional data, particularly suitable for NIRS data with continuous distribution features. The method is effective for small-sample data and data containing noise.

KNN algorithm is an instance-based learning approach. This algorithm classifies samples by measuring the distance between the sample to be classified and samples in the training set, based on the labels of the nearest neighbors. KNN is known for its simplicity and adaptability to nonlinear and complex data relationships.

CART algorithm is a tree-based machine learning method. CART recursively splits the dataset into subsets with lower impurity, constructing a binary tree structure for classification or regression tasks. The algorithm is suitable for complex nonlinear relationships and exhibits strong robustness, capable of handling mixed-type data and missing values.

SVM algorithm is a powerful supervised learning method. SVM classifies samples by finding the optimal hyperplane that separates different classes. It performs well in handling high-dimensional spaces and nonlinear relationships, and can adapt to different data structures through the use of kernel functions. SVM exhibits strong generalization capabilities and robustness to small sample sizes and noise.

LR algorithm is a widely used regression method. LR aims to establish a linear relationship between input features and output targets by fitting the best-fitting line for prediction. This method is simple, intuitive, and performs well when the data exhibits a linear relationship. In practical applications, LR often requires consideration of feature engineering and outlier handling to improve the model's accuracy.

MPM is a method that integrates information from various sources for data analysis, widely used in the processing of NIRS data. This model combines information from different sensors or data sources to enhance the comprehensiveness and robustness of the data. When dealing with NIRS data, the MPM can effectively capture correlations between different spectral bands, enabling a more comprehensive understanding and utilization of the data.

BPNN is a powerful deep learning approach. Trained through multiple layers of neural networks, BPNN can learn complex nonlinear relationships within the data. It possesses strong fitting capabilities and is suitable for handling highly nonlinear and complex structured data. BPNN effectively extracts features from NIRS data, enabling accurate classification or prediction.

CNN is a deep learning model that has achieved tremendous success in the field of image processing. CNN uses convolutional and pooling layers to automatically learn spatial and spectral features from data, making it suitable for processing NIRS graphs with spatial structures. This model efficiently captures local information in the data and achieves an understanding of global features through multiple

levels of abstraction. CNN demonstrates excellent performance in handling NIRS data but requires a large dataset for training.

DNN is a more complex and deep neural network structure. DNN learns through multiple hidden layers, capturing highly abstract features within the data, making it suitable for processing complex and large datasets. In NIRS analysis, the DNN model can delve deeper into the intrinsic patterns of the data, achieving the learning and representation of complex features.

The proposed methodology integrates data preprocessing methods and machine learning models to tackle the challenges presented by a large and potentially intricate *P. notoginseng* NIRS dataset. This holistic approach, encompassing data preprocessing, feature extraction, dimensionality reduction and classification algorithms, is crafted to compare the accuracy of geographical origin identification.

Results and discussion

To more objectively and accurately reflect the impact of data preprocessing methods and machine learning models on the classification accuracy of the imbalanced *P. notoginseng* NIRS dataset, a step-by-step testing approach was employed during the experimental process. Firstly, the dataset, without any preprocessing, was directly used to train and test nine machine learning models, obtaining baseline indicators for the classification accuracy of *P. notoginseng* geographical origin. Secondly, the imbalanced dataset was subjected to individual data preprocessing methods, followed by training and testing with the same nine machine learning models, resulting in the classification accuracy for *P. notoginseng* geographical origin. Thirdly, a combination of data preprocessing methods was applied to the imbalanced dataset, and the nine machine learning models were utilized for training and testing, yielding the classification accuracy for *P. notoginseng* geographical origin. The detailed results of these tests are presented in Table 1.

Table 1

Geographical origin classification accuracy of imbalanced *P. notoginseng* NIRS dataset under different data preprocessing methods and machine learning models.

Data Preprocessing Methods	Machine Learning Models								
	GNB	KNN	CART	SVM	LR	MPM	BPNN	CNN	DNN
Non-Preprocessing	80%	85%	90%	75%	90%	75%	30%	30%	30%
Autocorrelation	80%	85%	85%	80%	90%	85%	40%	50%	45%
Standardization	80%	85%	80%	85%	90%	90%	35%	90%	90%
MSC	95%	95%	90%	50%	100%	30%	30%	35%	20%
SNV	95%	95%	80%	50%	100%	90%	30%	65%	30%
S-G	80%	85%	75%	75%	90%	75%	30%	30%	20%
1D	100%	95%	95%	90%	90%	95%	50%	20%	35%
2D	100%	95%	80%	90%	90%	95%	60%	30%	40%
PCA	90%	85%	80%	85%	45%	90%	65%	65%	90%
Autocorrelation + Standardization	80%	85%	90%	85%	90%	90%	35%	90%	90%
Autocorrelation + MSC	95%	100%	85%	55%	100%	85%	50%	20%	20%
Autocorrelation + SNV	95%	100%	85%	55%	100%	90%	20%	20%	65%
Autocorrelation + S-G	80%	85%	80%	80%	90%	85%	40%	35%	35%
Autocorrelation + 1D	95%	90%	90%	85%	90%	90%	45%	20%	30%
Autocorrelation + 2D	95%	90%	80%	85%	90%	90%	60%	30%	40%
Autocorrelation + PCA	90%	85%	85%	85%	50%	90%	65%	80%	70%
Standardization + MSC	35%	75%	30%	30%	30%	20%	35%	20%	20%
Standardization + SNV	90%	85%	85%	90%	55%	85%	10%	90%	85%
Standardization + S-G	80%	85%	75%	85%	90%	90%	35%	85%	90%
Standardization + 1D	90%	90%	90%	90%	90%	90%	85%	85%	60%
Standardization + 2D	90%	90%	90%	90%	90%	90%	85%	80%	75%
Standardization + PCA	90%	85%	85%	85%	45%	90%	55%	80%	90%
MSC + S-G	95%	95%	85%	50%	100%	85%	30%	20%	30%
MSC + 1D	100%	100%	90%	90%	100%	85%	50%	20%	30%
MSC + 2D	100%	100%	85%	90%	100%	90%	55%	30%	30%

Data Preprocessing Methods	Machine Learning Models								
	GNB	KNN	CART	SVM	LR	MPM	BPNN	CNN	DNN
MSC + PCA	100%	95%	90%	90%	40%	90%	70%	60%	75%
SNV + S-G	95%	95%	85%	50%	100%	90%	30%	35%	80%
SNV + 1D	100%	100%	80%	90%	100%	90%	90%	60%	60%
SNV + 2D	100%	100%	85%	90%	100%	90%	90%	60%	55%
SNV + PCA	100%	95%	80%	90%	40%	95%	75%	65%	75%
S-G + 1D	100%	95%	85%	90%	90%	95%	50%	30%	30%
S-G + 2D	100%	95%	85%	90%	90%	95%	50%	30%	40%
S-G + PCA	90%	85%	85%	85%	45%	90%	65%	55%	70%
1D + PCA	100%	95%	90%	90%	95%	90%	35%	30%	20%
2D + PCA	100%	95%	90%	90%	95%	90%	45%	20%	30%
Autocorrelation + Standardization + MSC	35%	50%	30%	40%	15%	20%	35%	30%	30%
Autocorrelation + Standardization + SNV	90%	90%	85%	80%	75%	85%	5%	80%	85%
Autocorrelation + Standardization + S-G	80%	85%	80%	85%	90%	90%	35%	90%	80%
Autocorrelation + Standardization + 1D	90%	90%	95%	90%	90%	90%	85%	85%	70%
Autocorrelation + Standardization + 2D	90%	90%	85%	90%	90%	90%	85%	85%	80%
Autocorrelation + Standardization + PCA	90%	85%	80%	85%	50%	90%	65%	80%	90%
Autocorrelation + MSC + S-G	95%	100%	90%	55%	100%	85%	60%	20%	30%
Autocorrelation + MSC + 1D	95%	100%	85%	90%	100%	90%	50%	20%	20%
Autocorrelation + MSC + 2D	95%	100%	85%	90%	100%	90%	55%	30%	40%
Autocorrelation + MSC + PCA	100%	100%	80%	95%	45%	90%	70%	60%	85%
Autocorrelation + SNV + S-G	95%	100%	85%	55%	100%	90%	30%	65%	65%

Data Preprocessing Methods	Machine Learning Models								
	GNB	KNN	CART	SVM	LR	MPM	BPNN	CNN	DNN
Autocorrelation + SNV + 1D	95%	100%	85%	90%	100%	90%	90%	75%	45%
Autocorrelation + SNV + 2D	95%	100%	85%	90%	100%	90%	90%	80%	40%
Autocorrelation + SNV + PCA	100%	100%	80%	95%	40%	90%	80%	80%	85%
Autocorrelation + S-G + 1D	95%	90%	90%	85%	90%	90%	45%	20%	30%
Autocorrelation + S-G + 2D	95%	90%	90%	85%	90%	90%	45%	30%	40%
Autocorrelation + S-G + PCA	90%	85%	85%	85%	50%	90%	65%	75%	80%
Autocorrelation + 1D + PCA	95%	90%	90%	85%	90%	85%	35%	20%	30%
Autocorrelation + 2D + PCA	95%	90%	90%	85%	85%	85%	40%	30%	65%
Standardization + MSC + S-G	35%	75%	30%	30%	30%	20%	35%	20%	30%
Standardization + MSC + 1D	50%	75%	30%	30%	25%	30%	35%	30%	30%
Standardization + MSC + 2D	50%	75%	30%	30%	25%	30%	35%	30%	20%
Standardization + MSC + PCA	25%	70%	30%	30%	15%	20%	35%	30%	20%
Standardization + SNV + S-G	90%	85%	90%	90%	55%	85%	5%	90%	85%
Standardization + SNV + 1D	100%	90%	70%	90%	80%	90%	70%	85%	85%
Standardization + SNV + 2D	100%	90%	75%	90%	80%	90%	65%	90%	80%
Standardization + SNV + PCA	90%	85%	85%	80%	40%	90%	70%	70%	85%
Standardization + S-G + 1D	90%	90%	80%	90%	90%	90%	85%	40%	75%
Standardization + S-G + 2D	90%	90%	90%	90%	90%	90%	85%	45%	85%

Data Preprocessing Methods	Machine Learning Models								
	GNB	KNN	CART	SVM	LR	MPM	BPNN	CNN	DNN
Standardization + S-G + PCA	90%	85%	85%	85%	45%	90%	55%	75%	90%
Standardization + 1D + PCA	90%	90%	85%	90%	60%	90%	60%	45%	70%
Standardization + 2D + PCA	90%	90%	95%	90%	60%	90%	75%	60%	70%
MSC + S-G + 1D	100%	100%	75%	90%	100%	90%	50%	30%	20%
MSC + S-G + 2D	95%	100%	95%	90%	100%	95%	50%	20%	30%
MSC + S-G + PCA	100%	95%	90%	90%	40%	90%	70%	65%	65%
MSC + 1D + PCA	90%	100%	85%	95%	75%	95%	35%	20%	45%
MSC + 2D + PCA	90%	100%	85%	90%	75%	95%	35%	30%	40%
SNV + S-G + 1D	100%	100%	75%	90%	100%	90%	90%	40%	40%
SNV + S-G + 2D	95%	100%	90%	90%	100%	90%	85%	30%	40%
SNV + S-G + PCA	100%	95%	90%	90%	40%	95%	75%	85%	85%
SNV + 1D + PCA	90%	100%	90%	95%	80%	95%	60%	30%	90%
SNV + 2D + PCA	100%	100%	80%	95%	80%	100%	7%	55%	95%
S-G + 1D + PCA	100%	95%	95%	90%	95%	20%	35%	20%	20%
S-G + 2D + PCA	100%	95%	90%	90%	95%	25%	35%	40%	30%
Autocorrelation + Standardization + MSC + S-G	35%	50%	30%	40%	15%	20%	35%	20%	20%
Autocorrelation + Standardization + MSC + 1D	55%	70%	30%	40%	20%	30%	35%	20%	20%
Autocorrelation + Standardization + MSC + 2D	55%	70%	30%	35%	20%	30%	35%	20%	20%
Autocorrelation + Standardization + MSC + PCA	45%	50%	30%	40%	35%	20%	35%	20%	30%
Autocorrelation + Standardization + SNV + S-G	90%	90%	90%	80%	75%	85%	5%	80%	85%

Data Preprocessing Methods	Machine Learning Models								
	GNB	KNN	CART	SVM	LR	MPM	BPNN	CNN	DNN
Autocorrelation + Standardization + SNV + 1D	95%	90%	95%	90%	60%	90%	75%	90%	85%
Autocorrelation + Standardization + SNV + 2D	95%	90%	80%	90%	65%	85%	75%	90%	85%
Autocorrelation + Standardization + SNV + PCA	80%	90%	85%	80%	25%	90%	65%	65%	85%
Autocorrelation + MSC + S-G + 1D	95%	100%	75%	90%	100%	90%	50%	30%	20%
Autocorrelation + MSC + S-G + 2D	95%	100%	75%	90%	100%	90%	50%	20%	30%
Autocorrelation + MSC + S-G + PCA	100%	100%	80%	95%	45%	90%	70%	70%	80%
Autocorrelation + SNV + S-G + 1D	95%	100%	75%	90%	100%	90%	30%	90%	40%
Autocorrelation + SNV + S-G + 2D	95%	100%	75%	90%	100%	90%	85%	40%	30%
Autocorrelation + SNV + S-G + PCA	100%	100%	80%	95%	40%	90%	80%	85%	85%
Autocorrelation + S-G + 1D + PCA	100%	90%	90%	85%	90%	35%	35%	20%	20%
Autocorrelation + S-G + 2D + PCA	95%	90%	90%	85%	90%	85%	35%	30%	20%
Standardization + MSC + S-G + 1D	50%	80%	30%	30%	25%	30%	35%	20%	30%
Standardization + MSC + S-G + 2D	50%	65%	30%	30%	25%	30%	35%	30%	30%
Standardization + MSC + S-G + PCA	25%	70%	30%	30%	15%	20%	35%	30%	20%
Standardization + SNV + S-G + 1D	100%	90%	90%	90%	65%	85%	70%	85%	80%
Standardization + SNV + S-G + 2D	100%	90%	90%	90%	80%	90%	65%	80%	80%
Standardization + SNV + S-G + PCA	90%	85%	85%	80%	40%	90%	70%	55%	70%

Data Preprocessing Methods	Machine Learning Models								
	GNB	KNN	CART	SVM	LR	MPM	BPNN	CNN	DNN
Standardization + S-G + 1D + PCA	95%	90%	75%	90%	55%	90%	75%	35%	85%
Standardization + S-G + 2D + PCA	95%	90%	80%	90%	55%	90%	75%	70%	85%
MSC + S-G + 1D + PCA	90%	100%	85%	95%	75%	20%	35%	20%	30%
MSC + S-G + 2D + PCA	90%	100%	85%	95%	75%	20%	35%	20%	20%
SNV + S-G + 1D + PCA	100%	100%	90%	95%	75%	95%	50%	65%	50%
SNV + S-G + 2D + PCA	100%	100%	80%	95%	80%	95%	50%	30%	40%
Autocorrelation + Standardization + MSC + S-G + 1D	55%	70%	30%	40%	20%	30%	35%	30%	30%
Autocorrelation + Standardization + MSC + S-G + 2D	55%	70%	30%	35%	20%	30%	35%	20%	20%
Autocorrelation + Standardization + MSC + S-G + PCA	45%	50%	30%	40%	35%	20%	35%	30%	20%
Autocorrelation + Standardization + SNV + S-G + 1D	100%	90%	80%	90%	55%	90%	75%	85%	85%
Autocorrelation + Standardization + SNV + S-G + 2D	100%	90%	80%	90%	60%	85%	80%	85%	90%
Autocorrelation + Standardization + SNV + S-G + PCA	80%	90%	85%	80%	25%	90%	65%	55%	80%
Autocorrelation + MSC + S-G + 1D + PCA	90%	100%	85%	95%	85%	20%	35%	20%	20%
Autocorrelation + MSC + S-G + 2D + PCA	90%	100%	75%	90%	85%	20%	35%	30%	30%
Autocorrelation + SNV + S-G + 1D + PCA	90%	100%	85%	95%	85%	95%	50%	55%	55%
Autocorrelation + SNV + S-G + 2D + PCA	90%	100%	75%	90%	85%	95%	50%	30%	40%
Standardization + MSC + S-G + 1D + PCA	25%	70%	30%	25%	20%	35%	35%	30%	30%

Data Preprocessing Methods	Machine Learning Models								
	GNB	KNN	CART	SVM	LR	MPM	BPNN	CNN	DNN
Standardization + MSC + S-G + 2D + PCA	25%	65%	30%	30%	35%	20%	35%	30%	30%
Standardization + SNV + S-G + 1D + PCA	95%	90%	75%	90%	55%	90%	70%	75%	85%
Standardization + SNV + S-G + 2D + PCA	95%	90%	80%	90%	55%	90%	70%	75%	80%
Autocorrelation + Standardization + MSC + S-G + 1D + PCA	45%	75%	30%	40%	30%	20%	35%	30%	30%
Autocorrelation + Standardization + MSC + S-G + 2D + PCA	45%	70%	30%	35%	30%	20%	35%	20%	20%
Autocorrelation + Standardization + SNV + S-G + 1D + PCA	95%	90%	80%	90%	55%	90%	75%	55%	80%
Autocorrelation + Standardization + SNV + S-G + 2D + PCA	95%	90%	85%	90%	55%	90%	80%	65%	80%

Effect analysis of data preprocessing methods. Without data preprocessing methods, the average classification accuracy of machine learning models for the geographical origin is approximately 65%. This result is obtained by directly applying machine learning algorithms to the raw spectral data without employing any data preprocessing methods. This metric serves as a baseline for evaluating the effectiveness of data preprocessing methods in this study.

Although the models achieved a 65% accuracy in distinguishing the geographical origin of *P. notoginseng* even without preprocessing, indicating the presence of useful information related to geographical origin in the raw spectral data, this baseline accuracy is not high, especially considering the practical applications where higher accuracy is crucial for ensuring the reliability of results. For instance, in the fields of medicinal herb identification and quality control, high accuracy is essential as it directly impacts the safety and efficacy of products.

When applying a single data preprocessing method, the average accuracy of the classification models increased to around 70.71%, a significant improvement compared to the 65% baseline accuracy without any preprocessing. This demonstrates that even individual preprocessing techniques can effectively enhance the performance of models in the task of classifying *P. notoginseng* geographical origins. The analysis of different preprocessing methods reveals the importance of this improvement as it uncovers potential distinguishable patterns related to geographical origin in spectral data.

Each method of Single preprocessing methods has its unique strengths and purposes. For example, standardization scales the data to normalize contributions from different features, while MSC attempts to correct scattering effects in spectral data caused by varying particle sizes or changes in light path. S-G smoothing, through polynomial fitting, smoothens data to remove noise without excessively distorting signals, and 1D and 2D can emphasize local variations in spectral curves, enhancing the detectability of minor absorption peaks. PCA aims to reduce data dimensions by extracting the most significant variables.

While single preprocessing methods can improve classification accuracy due to their optimization for specific features or noise in spectral data, they also have limitations. Each method is designed to address particular types of issues and cannot be universally applied. In practical applications, the choice of preprocessing method should be based on data characteristics and the requirements of the classification task. Additionally, excessive or inappropriate preprocessing may lead to the loss of useful information or introduce unnecessary variability, potentially reducing model performance.

When using combined preprocessing methods, the average classification accuracy of the models reaches approximately 69.29%. This result falls between the accuracy without preprocessing (65%) and that with a single preprocessing method (70.71%). It suggests that combined preprocessing methods have a certain effect in improving classification accuracy, although this effect may not be as pronounced as that of single preprocessing methods. This could be attributed to interactions between different preprocessing techniques, potentially offsetting each other's advantages or introducing new interference or loss of crucial information in specific cases.

In some scenarios, combined data preprocessing can significantly enhance model performance. For instance, a combination of derivative processing to enhance features, followed by smoothing techniques to eliminate amplified noise, and finally standardization to ensure features are evaluated on the same scale, may lead to higher accuracy compared to a single preprocessing method. For example, this combined preprocessing approach could elevate the classification accuracy of SVM from 85% to around 90%.

Despite the conceptual belief that combined preprocessing methods can offer the most comprehensive data improvement, their effectiveness depends on careful design and execution of these preprocessing steps. If well-chosen, these methods can synergistically enhance each other, improving the overall performance of the model. However, if chosen inappropriately, they may counteract each other, even diminishing model performance. Therefore, when applying combined preprocessing methods, a deep understanding of data characteristics is crucial, and experimentation is necessary to optimize the selection and sequence of preprocessing steps to ensure a net improvement rather than a reduction in classification accuracy.

Performance analysis of machine learning models. In the geographical origin classification study based on *P. notoginseng* NIRS imbalanced dataset, various machine learning models were employed to assess

their performance. Objective data reveals that different models exhibit varied performance under the influence of preprocessing methods, as shown in Fig. 3.

The GNB model, renowned for its simplicity and speed, is widely employed, particularly when the assumption of feature independence holds. In this dataset, it demonstrated robust classification capabilities, achieving an average accuracy of approximately 85.4%. There were 65 test instances with a classification accuracy greater than or equal to 95%, making it the top-performing model among all. This outcome suggests that even in the presence of complex and imbalanced spectral data, Naive Bayes can capture sufficient statistical patterns for effective classification.

The KNN model, an instance-based learning method that classifies based on distance measurements between new and training samples, outperformed all models with an average accuracy of 89.12%. There were 52 test instances with a classification accuracy greater than or equal to 95%, and 105 instances with accuracy greater than or equal to 85% but less than 95%. This superior performance may be attributed to KNN's adaptability to the local structure of data, especially when the sample size is sufficient, and the feature space is not excessively sparse.

The CART model, utilizing decision tree algorithms to recursively split data into smaller subsets, achieved an average accuracy of 75.72%. However, only 6 test instances had a classification accuracy greater than or equal to 95%, which is lower than GNB and KNN. This discrepancy could reflect limitations of decision trees in handling continuous spectral data, especially without proper preprocessing to address noise and spectral overlap.

The SVM model seeks the optimal separating hyperplane in the sample feature space and achieved an average accuracy of 77.52%. Although the average classification accuracy is close to 80%, only 13 test instances had accuracy greater than or equal to 95%, indicating that its performance is strongly influenced by the choice of kernel function and parameter tuning. While SVM often excels in high-dimensional spaces, appropriate kernel function selection and parameter tuning are crucial for spectral data.

Despite being typically used for regression tasks, linear models can be applied to classification. In this task, the average accuracy of the LR model was 68.8%, with 28 test instances having a classification accuracy greater than or equal to 95%. This might be because the linear boundary assumption for classification problems may not always hold in spectral data, especially without preprocessing.

The MPM, a simple feedforward artificial neural network, achieved an average accuracy of 75.08%. There were 15 test instances with a classification accuracy greater than or equal to 95%, and 80 instances with accuracy greater than or equal to 85% but less than 95%. It can learn nonlinear patterns in input data, but its performance highly depends on network structure and training details.

The BPNN, a more complex artificial neural network trained through backpropagation algorithms, exhibited a performance of 51.86%, the lowest among all models. This may be due to the network

structure failing to adapt to the complexity of the data or insufficient training.

CNNs are designed for data with grid-like structures (such as images) but can be applied to spectral data. In experiments, the average accuracy of CNN was 48.36%, with 19 test instances having a classification accuracy greater than or equal to 85%, indicating that the network structure and parameters may not have adequately captured spatial correlations in spectral data.

Theoretically capable of learning more complex data representations due to its multilayer structure, the DNN achieved an average accuracy of 52.72% in experiments, with only 1 test instance having a classification accuracy greater than or equal to 95%. This might be because of insufficient tuning and alignment with the characteristics of spectral data.

Overall, the KNN model performed the best among these machine learning models, while the BPNN, CNN, and DNN failed to showcase their potential in this study. This could be attributed to the requirement of extensive data and carefully designed network structures for effective deep learning models, which might necessitate more methodological research and experimentation when dealing with smaller, complex spectral data. Furthermore, simple models such as GNB and CART exhibited good performance, emphasizing the importance of choosing models based on data characteristics and available resources when applying machine learning models.

Conclusions

In this paper, the current research hotspot involves the utilization of NIRS combined with machine learning models to identify the geographical origins of Chinese herbal medicines, with a specific focus on *P. notoginseng*, a pivotal species in this field. Despite its significance, the challenge of imbalanced NIRS datasets for *P. notoginseng* has been scarcely addressed in existing literature. The study, therefore, concentrates on evaluating various data preprocessing methods and machine learning models to enhance the accuracy in classifying the geographical origins of *P. notoginseng* under these challenging conditions.

The experimental results demonstrate the efficacy of combining judicious data preprocessing with machine learning models to improve the handling of imbalanced NIRS datasets of *P. notoginseng*, leading to more precise geographical origin classification. This finding is crucial for the advancement of methodologies in herbal medicine authentication and quality control.

In future work, building on the foundation laid by this research, there is a proposal to further refine and optimize machine learning models and develop novel algorithms. The aim is to continually enhance the classification accuracy for the geographical origin of *P. notoginseng* in imbalanced datasets, thus advancing the field of herbal medicine research through technological innovation and data-driven insights.

Declarations

Data Availability

The data used to support the findings of this study are available at GitHub (<https://github.com/gao20030231/paper20231222>) and figshare.

Code availability

The source code of this work can be downloaded from GitHub (<https://github.com/gao20030231/paper20231222>) and figshare.

The authors confirm that *P. notoginseng*, as one of the most commonly used traditional Chinese medicinal herbs in China, was utilized in compliance with relevant institutional, national, and international guidelines and legislation for experimental research. Whether cultivated or wild, the use of this plant material in the study adhered to the appropriate regulations. The research was conducted with due consideration of ethical standards and in alignment with the established guidelines for the collection and use of *P. notoginseng* specimens.

Competing interests

The authors declare no competing interests.

Acknowledgements

This research was funded by Science and Technology Project of Chongqing Municipal Education Commission (Grant No. KJZD-K202303301, KJQN202203309 and KJQN202203308).

Author contributions

Conceptualization, X.F.C; methodology, X.F.C. and J.L.; software, J.L.; validation, X.F.C., and J.L.; formal analysis, X.F.C.; investigation, X.F.C., and J.L.; resources, X.F.C., and J.L.; data curation, J.L.; writing—original draft preparation, J.L.; writing—review and editing, X.F.C.; supervision, X.F.C. and A. B.; project administration and funding acquisition, X.F.C. All authors reviewed the manuscript.

References

1. Shi, L., Li, L., Zhang, F. & Lin, Y. Nondestructive detection of *Panax notoginseng* saponins by using hyperspectral imaging. *Int J Food Sci Technol* **57**, (2022).
2. Zhang, F. *et al.* Application of portable near-infrared spectroscopy technology for grade identification of *Panax notoginseng* slices. *J Food Saf* **43**, (2023).
3. Zhang, F. J., Shi, L., Li, L. X., Zhao, H. R. & Zhu, Y. L. Study on Nondestructive Identification of *Panax Notoginseng* Powder Quality Grade Based on Hyperspectral Imaging Technology. *Guang Pu Xue Yu Guang Pu Fen Xi/Spectroscopy and Spectral Analysis* **42**, (2022).

4. Dong, J. E., Wang, Y., Zuo, Z. T. & Wang, Y. Z. Deep learning for geographical discrimination of *Panax notoginseng* with directly near-infrared spectra image. *Chemometrics and Intelligent Laboratory Systems* **197**, (2020).
5. Zhang, H. *et al.* Rapid determination of *Panax notoginseng* origin by terahertz spectroscopy combined with the machine learning method. *Spectroscopy Letters* **55**, (2022).
6. Cui, Z. Y., Liu, C. L., Li, D. D., Wang, Y. Z. & Xu, F. R. Anticoagulant activity analysis and origin identification of *Panax notoginseng* using HPLC and ATR-FTIR spectroscopy. *Phytochemical Analysis* **33**, (2022).
7. Luo, Z. R., Ouyang, W. H., Zhang, Y. F., Sun, W. Q. & Liu, L. P. Establishment of Quantitative Determination Model of Total Saponins in *Panax Notoginseng* Cream by Near-Infrared Spectroscopy. *Chinese Journal of Pharmaceutical Biotechnology* **25**, (2018).
8. Liu, P. *et al.* Rapid identification and quantification of *Panax notoginseng* with its adulterants by near infrared spectroscopy combined with chemometrics. *Spectrochim Acta A Mol Biomol Spectrosc* **206**, (2019).
9. Li, Y., Zhang, J. Y. & Wang, Y. Z. FT-MIR and NIR spectral data fusion: a synergetic strategy for the geographical traceability of *Panax notoginseng*. *Anal Bioanal Chem* **410**, (2018).
10. Wang, X. Y., Li, W. L. & Qu, H. Bin. Application of feedforward control strategy based on spectra of raw materials to optimize alcohol extraction process of *Panax notoginseng*. *Zhongguo Zhongyao Zazhi* **43**, (2018).
11. Zhou, Y., Zuo, Z., Xu, F. & Wang, Y. Origin identification of *Panax notoginseng* by multi-sensor information fusion strategy of infrared spectra combined with random forest. *Spectrochim Acta A Mol Biomol Spectrosc* **226**, (2020).
12. Ji, C. *et al.* Multi-Element Analysis and Origin Discrimination of *Panax notoginseng* Based on Inductively Coupled Plasma Tandem Mass Spectrometry (ICP-MS/MS). *Molecules* **27**, (2022).
13. Bai, J. *et al.* Identification of geographical origins of *Panax notoginseng* based on HPLC multi-wavelength fusion profiling combined with average linear quantitative fingerprint method. *Sci Rep* **11**, (2021).
14. Yang, J. *et al.* Discovery and verification of bitter components in *Panax notoginseng* based on the integrated strategy of pharmacophore model, system separation and bitter tracing technology. *Food Chem* **428**, (2023).
15. Liu, C., Zuo, Z., Xu, F. & Wang, Y. Study of the suitable climate factors and geographical origins traceability of *Panax notoginseng* based on correlation analysis and spectral images combined with machine learning. *Front Plant Sci* **13**, (2023).
16. Zhang, S. *et al.* Rapid and simultaneous determination of moisture and berberine content in *Coptidis Rhizoma* and *Phellodendri Chinensis Cortex* by near-infrared spectroscopy and chemometrics. *J Innov Opt Health Sci* **13**, (2020).
17. Yan, M. L. *et al.* Rapid Evaluation of Anti-inflammatory Efficacy of *Angelica sinensis* (Oliv.) Diels by Near Infrared Spectroscopy Based on Quality Markers. *Journal of Instrumental Analysis* **39**, (2020).

18. Li, Y. Identification of Traditional Chinese Medicine Based on KNN Algorithm and Random Forest. *Highlights in Science, Engineering and Technology* **60**, (2023).
19. Ma, H. *et al.* Maintaining the predictive abilities of near-infrared spectroscopy models for the determination of multi-parameters in White Paeony Root. *Infrared Phys Technol* **109**, (2020).
20. Hao, Q. *et al.* Prediction the contents of fructose, glucose, sucrose, fructo-oligosaccharides and iridoid glycosides in Morinda officinalis radix using near-infrared spectroscopy. *Spectrochim Acta A Mol Biomol Spectrosc* **234**, (2020).
21. Gong, S. *et al.* Near-Infrared Spectroscopy Combined with Random Forest Algorithm: A Fast and Effective Strategy for Origin Traceability of Fuzi. *Guang Pu Xue Yu Guang Pu Fen Xi/Spectroscopy and Spectral Analysis* **42**, (2022).
22. Li, P., Luan, Z., Li, D., Meng, X. & Sun, J. A novel method to rapidly evaluate the quality of the traditional Chinese medicine Gentiana straminea Maxim. using near infrared spectroscopy coupled with chemometrics. *Spectroscopy Letters* **53**, (2020).
23. Hao, J. W., Chen, Y., Chen, N. D. & Qin, C. F. Rapid Detection of Adulteration in Dendrobium huoshanense Using NIR Spectroscopy Coupled with Chemometric Methods. *J AOAC Int* **104**, (2021).
24. Ma, H. *et al.* Rapid monitoring approaches for concentration process of lanqin oral solution by near-infrared spectroscopy and chemometric models. *Spectrochim Acta A Mol Biomol Spectrosc* **242**, (2020).
25. Li, Q. *et al.* A study on the rapidly non-destructive detection method of Huoxiang Zhengqi oral liquid using near infrared spectroscopy. *Yaoxue Xuebao* **57**, (2022).
26. Lei, L. *et al.* Identification of different bran-fried Atractylodis Rhizoma and prediction of atractylodin content based on multivariate data mining combined with intelligent color recognition and near-infrared spectroscopy. *Spectrochim Acta A Mol Biomol Spectrosc* **262**, (2021).
27. Hu, Y. *et al.* Rapid and simultaneous measurement of praeruptorin A, praeruptorin B, praeruptorin E, and moisture contents in peucedani radix using near-infrared spectroscopy and chemometrics. *J AOAC Int* **103**, (2021).
28. Huang, Z. W. *et al.* Research progress of near-infrared spectroscopy in quality evaluation of valuable Chinese medicinal materials. *Chinese Traditional and Herbal Drugs* **53**, (2022).
29. Yang, Y. *et al.* Rapid and comprehensive quality assessment of Bupleuri Radix through near-infrared spectroscopy combined with chemometrics. *Infrared Phys Technol* **121**, (2022).
30. Shi, T. *et al.* Application of near-infrared spectroscopy analysis technology to total nucleosides quality control in the fermented Cordyceps powder production process. *J Anal Methods Chem* **2020**, (2020).
31. Dai, P. *et al.* Development on near infrared spectroscopy method of in-process quantitative analysis of moisture content during honey refining. *Chinese Traditional and Herbal Drugs* **54**, (2023).
32. Zhan, S. Y., Liang, S. W. & Wang, S. M. Validation and comparison of different near-infrared quantitative methods to rapid determination of the total content of anthraquinones in Rhubarb based on accuracy profiles. *Spectroscopy Letters* **53**, (2020).

33. Chen, H., Tan, C. & Li, H. Untargeted identification of adulterated Sanqi powder by near-infrared spectroscopy and one-class model. *Journal of Food Composition and Analysis* **88**, (2020).
34. Fan, Z., Jiawei, Z. & Jihao, Z. ADULTERATION IDENTIFICATION OF ASTRAGALUS POLYSACCHARIDES BY NIR SPECTROSCOPY COMBINED WITH SIMCA AND PLS-DA. *INMATEH - Agricultural Engineering* **68**, (2022).
35. Gui, X. J. *et al.* Authenticity and species identification of *Fritillariae cirrhosae*: a data fusion method combining electronic nose, electronic tongue, electronic eye and near infrared spectroscopy. *Front Chem* **11**, (2023).
36. Zhao, J., Tian, G., Qiu, Y. & Qu, H. Rapid quantification of active pharmaceutical ingredient for sugar-free Yangwei granules in commercial production using FT-NIR spectroscopy based on machine learning techniques. *Spectrochim Acta A Mol Biomol Spectrosc* **245**, (2021).
37. Wang, F. *et al.* Rapid Identification of *Peucedanum praeruptorum* Dunn and Its Adulterants by Hand-Held Near-Infrared Spectroscopy. *J AOAC Int* **105**, (2022).
38. Zhang, Z. Y. *et al.* Rapid Geographical Origin Identification and Quality Assessment of *Angelicae Sinensis Radix* by FT-NIR Spectroscopy. *J Anal Methods Chem* **2021**, (2021).
39. Chai, Q. *et al.* Improved 1D convolutional neural network adapted to near-infrared spectroscopy for rapid discrimination of *Anoectochilus roxburghii* and its counterfeits. *J Pharm Biomed Anal* **199**, (2021).
40. Ming-Liang, G. *et al.* A gradient-based discriminant analysis method for process quality control of carbonized TCM via Fourier transform near infrared spectroscopy: A case study on carbonized *Typhae Pollen*. *Spectrochim Acta A Mol Biomol Spectrosc* **265**, (2022).
41. Shi, X. *et al.* Rapid detection of *Ganoderma lucidum* spore powder adulterated with dyed starch by NIR spectroscopy and chemometrics. *LWT* **167**, (2022).
42. Wang, F. *et al.* Qualitative classification of *Dendrobium huoshanense* (Feng dou) using fast non-destructive hand-held near infrared spectroscopy. *J Near Infrared Spectrosc* **30**, (2022).
43. Jin, Y., Du, W., Liu, X. & Wu, Y. Application of near infrared spectroscopy and real time release testing combined with statistical process control charts for on-line quality control of industrial concentrating process of traditional Chinese medicine “Jinyinhua”. *Infrared Phys Technol* **123**, (2022).
44. Zhang, J. *et al.* Multi critical quality attributes monitoring of Chinese oral liquid extraction process with a spectral sensor fusion strategy. *Spectrochim Acta A Mol Biomol Spectrosc* **278**, (2022).
45. Wang, X. P. *et al.* Feasibility analysis of near-infrared spectroscopy technology applied to classification and identification of four kinds of taste in traditional Chinese medicine. *Chinese Traditional and Herbal Drugs* **54**, (2023).
46. Wang, Q. *et al.* An integrated strategy of spectrum–effect relationship and near-infrared spectroscopy rapid evaluation based on back propagation neural network for quality control of *Paeoniae Radix Alba*. *Analytical Sciences* **39**, (2023).
47. Ma, H. *et al.* Non-Invasive Detection of Anti-Inflammatory Bioactivity and Key Chemical Indicators of the Commercial Lanqin Oral Solution by Near Infrared Spectroscopy. *Molecules* **27**, (2022).

48. Long, W. *et al.* Accurate identification of the geographical origins of lily using near-infrared spectroscopy combined with carbon dot-tetramethoxyporphyrin nanocomposite and chemometrics. *Spectrochim Acta A Mol Biomol Spectrosc* **271**, (2022).
49. Zhong, Z. *et al.* Evaluation of coating uniformity for the digestion-aid tablets by portable near-infrared spectroscopy. *Int J Pharm* **622**, (2022).
50. Gao, L. *et al.* A new perspective in understanding the processing mechanisms of traditional Chinese medicine by near-infrared spectroscopy with Aquaphotomics. *J Mol Struct* **1284**, (2023).
51. Kajino, A., Bai, W., Yoshimura, N. & Takayanagi, M. Identification of peach and apricot kernels for traditional Chinese medicines using near-infrared spectroscopy. *Vib Spectrosc* **113**, (2021).
52. Liu, N. C., Geng, L. D., Ma, L. J. & Wu, Z. S. Patent technology progress and trend of near-infrared spectroscopy technology in traditional Chinese medicine manufacturing. *Chinese Traditional and Herbal Drugs* **52**, (2021).

Figures

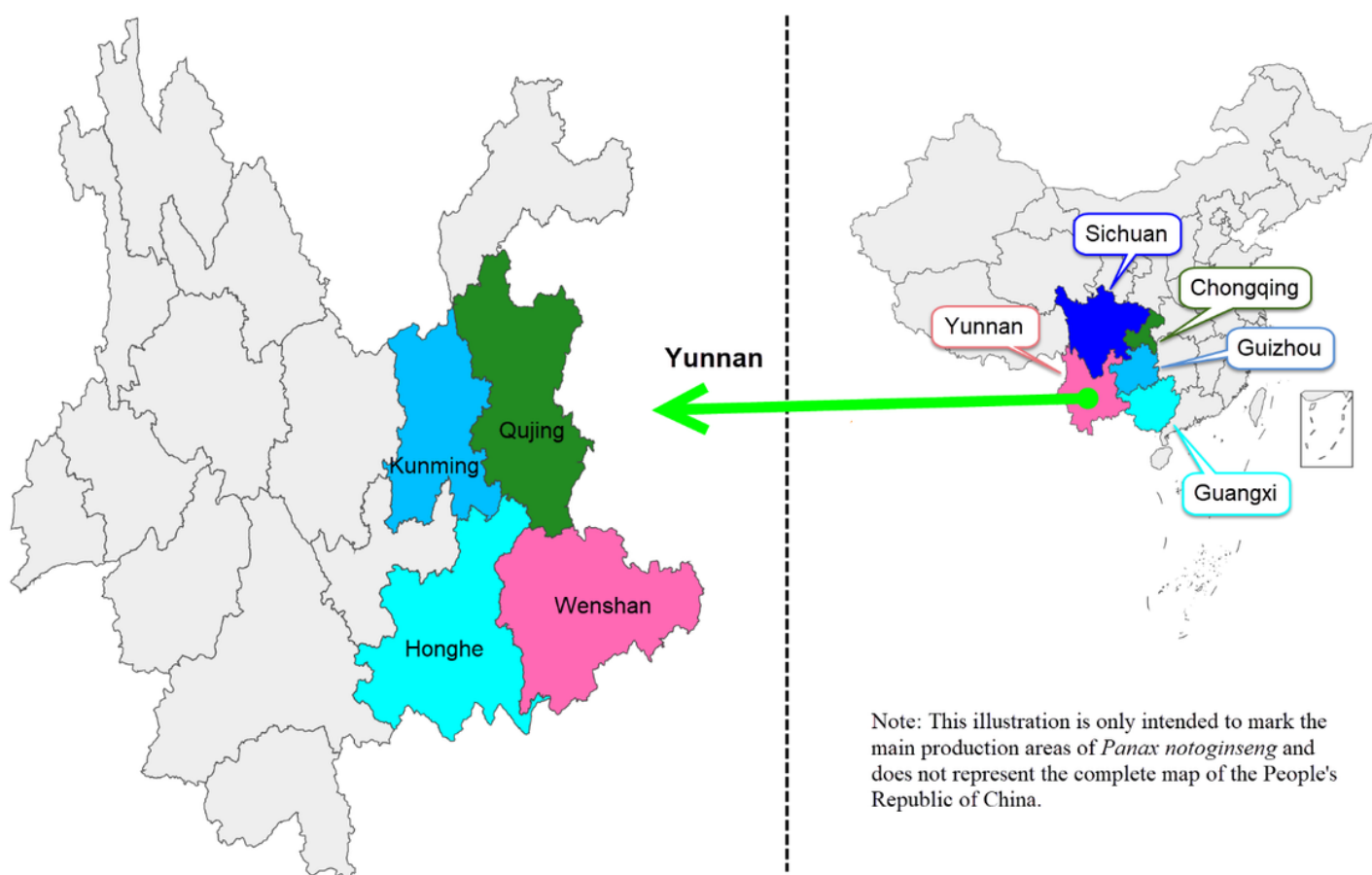


Figure 1

The distribution of *P. notoginseng* production areas in China.

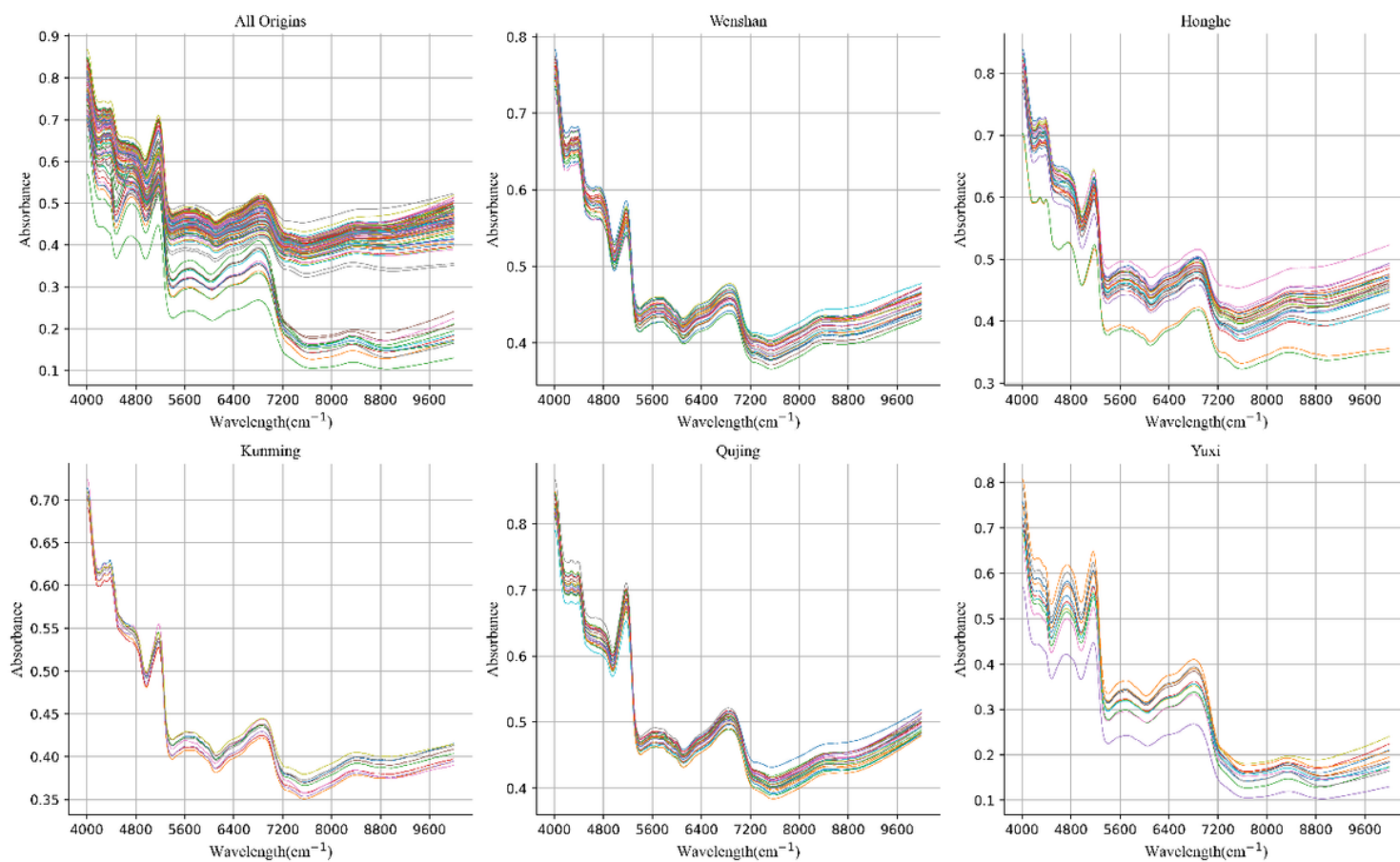


Figure 2

NIRS of the *P. notoginseng* samples.

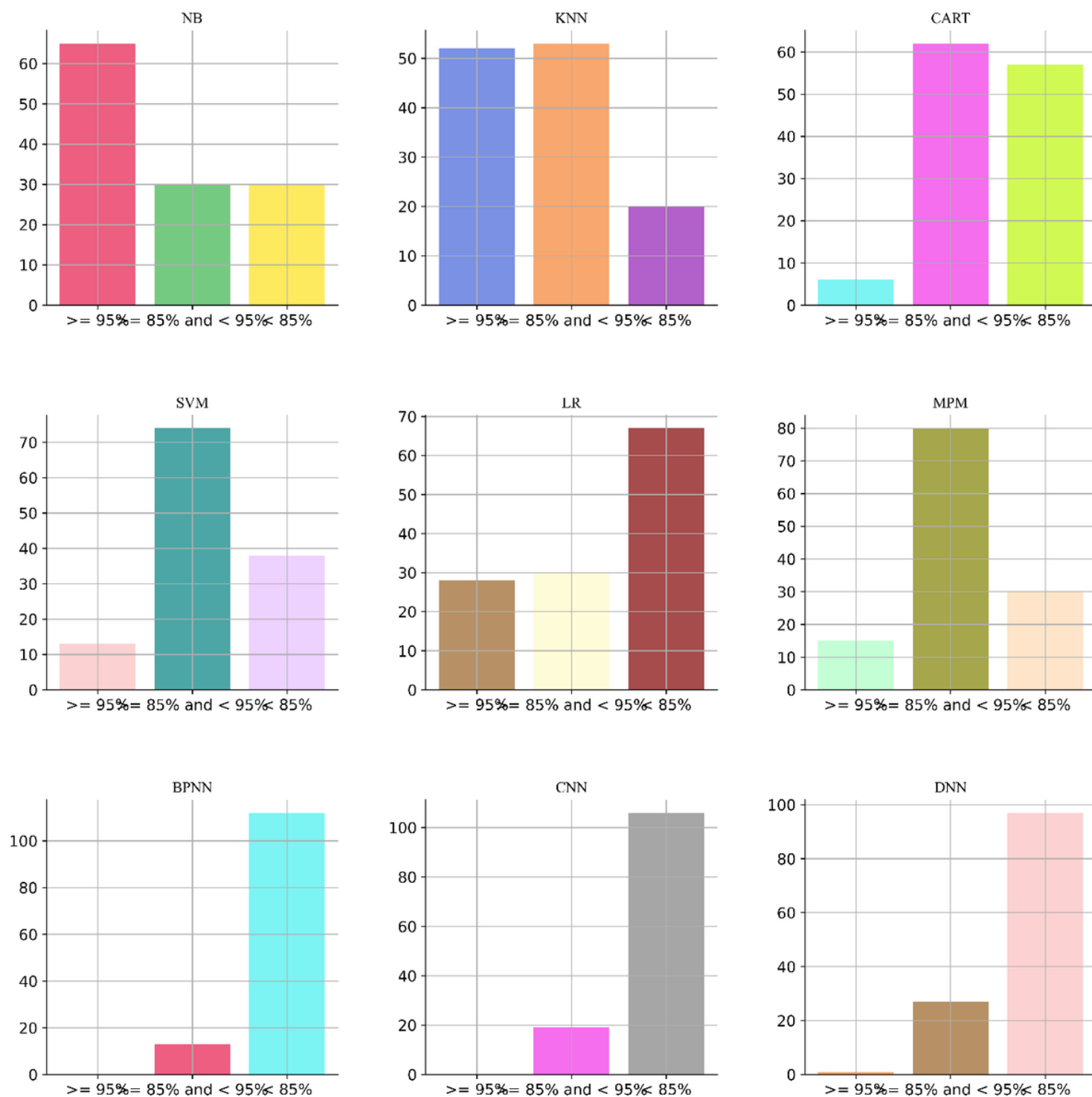


Figure 3

Classification accuracy statistics of the geographical origins of *P. notoginseng* for each machine learning model.