Journal of Pharmaceutical Analysis 13 (2023) 1388–1407

Contents lists available at ScienceDirect

Journal of Pharmaceutical Analysis

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

Review paper

Recent trends of machine learning applied to multi-source data of medicinal plants

Yanying Zhang [a, b], Yuanzhong Wang [a, *]

[a] Medicinal Plants Research Institute, Yunnan Academy of Agricultural Sciences, Kunming, 650200, China
[b] College of Traditional Chinese Medicine, Yunnan University of Chinese Medicine, Kunming, 650500, China

ARTICLE INFO

Article history:
Received 27 April 2023
Received in revised form
17 July 2023
Accepted 19 July 2023
Available online 25 July 2023

Keywords:
Machine learning
Medicinal plant
Multi-source data
Data fusion
Application

ABSTRACT

In traditional medicine and ethnomedicine, medicinal plants have long been recognized as the basis for materials in therapeutic applications worldwide. In particular, the remarkable curative effect of traditional Chinese medicine during corona virus disease 2019 (COVID-19) pandemic has attracted extensive attention globally. Medicinal plants have, therefore, become increasingly popular among the public. However, with increasing demand for and profit with medicinal plants, commercial fraudulent events such as adulteration or counterfeits sometimes occur, which poses a serious threat to the clinical outcomes and interests of consumers. With rapid advances in artificial intelligence, machine learning can be used to mine information on various medicinal plants to establish an ideal resource database. We herein present a review that mainly introduces common machine learning algorithms and discusses their application in multi-source data analysis of medicinal plants. The combination of machine learning algorithms and multi-source data analysis facilitates a comprehensive analysis and aids in the effective evaluation of the quality of medicinal plants. The findings of this review provide new possibilities for promoting the development and utilization of medicinal plants.

1. Introduction

Medicinal plants, which provide raw materials for the preparation of medicinal products of economical and medicinal value, have gained prominence and provide an increasing number of benefits to humans [1]. According to available statistics, medicinal plants have gradually gained more attention in recent years and the utilization rate of medicinal plants in developing countries is as high as 80%, while the market share in developed western countries such as Europe is also gradually increasing [2–4]. The business territory of medicinal plants is restricted to developing countries, but it is being progressively expanded to developed western countries. People are placing a greater emphasis on medical care, which has led to an increasing the demand for medicinal plants. Although millions of people worldwide are engaged in the cultivation, processing, and wild harvest of medicinal plants, the number of medicinal plants available still cannot efficiently satisfy the high demand [4]. Some species are obtained at lower yields owing to their strict requirements in terms of the growth environment and long duration for development. For unscrupulous people, commercial fraudulent events such as adulteration and counterfeits in medicinal plant business are commonplace to obtain greater profits, mainly because of the increasing demand for and profit with medicinal plants, which is one of the factors affecting the quality of medicinal plants. However, the growth of medicinal plants is affected by various environmental factors, such as soil, climate, terrain, and so on, resulting in large variations in active ingredients in the same medicinal plant [5,6]. Therefore, to maintain consistency in the chemical ingredients and to ensure drug efficacy, exploring and evaluating the quality of medicinal plants has become vital. The curative effects of medicinal plants are attributed to the active ingredients, ineffective components, toxic substances, and complex interactions [7]. The evaluation indexes of medicinal plants tend to be a multi-component aspect, but a single analytical technique cannot fully characterize the active ingredients of medicinal plants, which makes it difficult to perform quality evaluation [8].

With advances in modern analytical strategies, multi-source data analysis has shown more advantages for performing a comprehensive evaluation of the quality of medicinal plants [9]. The properties of medicinal plants determine the extent of

Peer review under responsibility of Xi'an Jiaotong University.
* Corresponding author.
E-mail address: boletus@126.com (Y. Wang).

https://doi.org/10.1016/j.jpha.2023.07.012
2095-1779/

advantages in utilizing multi-source data to elucidate changes in chemical information [10]. This is because multi-source data consists of collaborative information, which provides beneficial information for a better understanding or characterizing the quality of medicinal plants. Although the multi-source data strategy has been most recently used in studies of medicinal plants, the increase in data dimensions and quantity also causes analytical issues [11]. As a common approach to processing multi-source data, data fusion strategies offer valuable complementary information by integrating data from multiple sources. This integration leads to enhanced reliability, accuracy, and efficiency compared to relying on a single data source [12]. To acquire this complementary information, it is crucial to carefully select analytical instruments and possess a sound understanding of chemical knowledge [13]. The multi-source data of medicinal plants typically originates from various techniques or medicinal parts, resulting in complex analysis challenges [14]. However, employing data pre-processing, feature extraction techniques can simplify the data fusion process to a certain extent. Moreover, leveraging machine learning to effectively extract available information from multi-source data can significantly improve the robustness and accuracy of the obtained results [15].

The rapid development of artificial intelligence has played a pivotal role in revolutionizing data processing and analysis. Machine learning, as a subset of artificial intelligence, relies on computer algorithms and models to learn complex functions [16]. Over time, it has evolved into a powerful tool for statistical data modeling and mining. In the context of medicinal plants, the proportions of chemical components are sensitive to external factors and exhibit significant variations, which further complicates the quality evaluation process [17]. Thus, the need for techniques to prevent partial data loss has underscored the importance of employing machine learning algorithms with multi-source data of medicinal plants. Machine learning is recognized as a promising tool that can enhance the overall performance of data fusion due to its exceptional computing and analytical capabilities [18,19]. This review provides a comprehensive overview of the process of applying machine learning to multi-source data of medicinal plants for quality evaluation (Fig. 1). The review framework encompasses three essential modules. The first module summarizes the characteristics and applicability of different techniques (Table 1). The second module introduces data processing methods, including pre-processing, feature extraction, and data fusion strategies. The third module provides a specific summary of the application of machine learning in conjunction with multi-source data for the quality evaluation of medicinal plants. Furthermore, the review also discusses the limitations and future development trends in this field. By offering reference methods and improvement measures, this review serves as a valuable resource for enhancing the quality evaluation of medicinal plants.

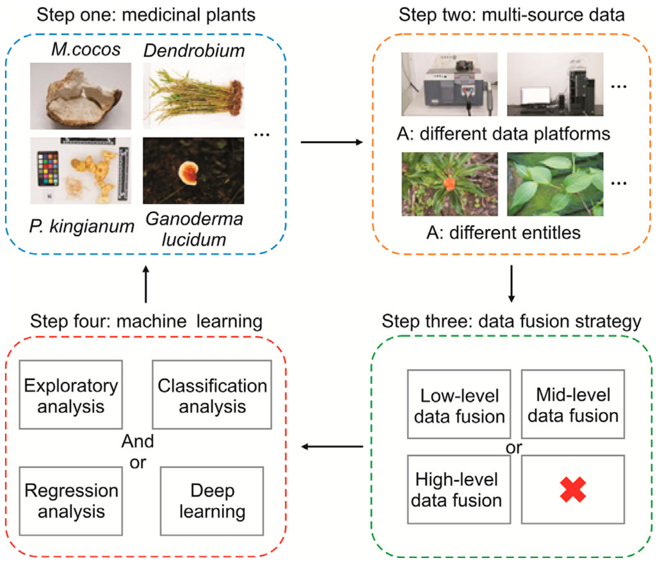

Fig. 1. The complete process of quality evaluation of medicinal plants.

2. Multi-source data of medicinal plants

Multi-source data for medicinal plants can be categorized into two main types. The first type involves detecting the same part of medicinal plants using different analytical techniques. The second type involves detecting different parts of medicinal plants using the same analytical technique. In the subsequent section, a brief introduction is provided for these two types of multi-source data.

2.1. Different detection techniques

2.1.1. Spectroscopy and spectrometry

With the continuous innovation and advancement of techniques, numerous spectroscopy and spectrometry methods have emerged. These methods, including ultraviolet-visible (UV-Vis), infrared, Raman spectroscopy, and nuclear magnetic resonance (NMR), play a crucial role in the quality assessment of medicinal plants. They are widely utilized as qualitative tools to obtain diverse chemical information from medicinal plants.

UV-Vis primarily focuses on analyzing chemical components or groups in medicinal plants that exhibit a tendency to absorb UV-Vis radiation [20]. This technique enables the identification and characterization of compounds. Infrared spectroscopy, operating within a wavelength range of 780–100000 nm, is an important non-invasive and rapid analysis technique [21]. Near-infrared (NIR) spectroscopy and mid-infrared (MIR) spectroscopy, subsets of infrared spectroscopy, provide spectral information about functional groups such as C—H and O—H in medicinal plant components by absorbing incident light energy at different frequency bands [22]. NIR primarily characterizes high-frequency combinations and complex overtones of hydrogen-containing groups, while MIR offers insights into the vibrational forms of organic compound molecules [23]. The development of the Fourier transform (FT) techniques has overcome the limitations of NIR and MIR, significantly improving the signal-to-noise ratio [24,25]. Raman spectroscopy, another widely used vibrational spectroscopic technique, captures information about nonpolar covalent bonds like C=C in medicinal plant components by including vibrational and rotational changes with a laser beam [26]. NMR spectroscopy, on the other hand, detects and quantifies chemical mixtures based on the interaction between the magnetic moments and magnetic fields of different atomic nuclei [27]. The introduction of chemical imaging techniques enables the simultaneous acquisition of spatial and spectral information regarding the chemical components of medicinal plants. Hyperspectral imaging (HSI) creates hypercube images for each wavelength, enhancing the accuracy of distinguishing medicinal plant components and providing detailed fingerprint information [25,28].

Given the complex and diverse nature of the chemical components in medicinal plants, it is essential to utilize qualitative tools sensitive to different chemical characteristics in order to fully identify their constituents. The limitations of relying on a single data source highlight the significance of using multi-source data for the evaluation of medicinal plant quality. Spectroscopy and spectrometry methods, serving as complementary sources, have

Table 1
Application characteristics, advantages, and disadvantages of different techniques in quality evaluation of medicinal plants.

Types	Techniques/ algorithms	Application characteristics	Advantages	Disadvantages
Spectroscopy and spectrometry	UV-Vis	■ Spectral region: 100–780 nm; ■ Mainly used in quantitative analysis of medicinal plants; ■ Less applied in qualitative analysis of medicinal plants	■ Rapid detection; ■ Simple operation; ■ High sensitivity	■ Poor selectivity; ■ Vulnerable to interference from multiple factors
	NIR	■ Spectral region: 780–2500 nm; ■ The most frequently used quality evaluation technique for medicinal plants	■ Rapid detection; ■ Simple operation; ■ Non-destructive and pollution-free	■ Low sensitivity; ■ Severe signal overlap
	MIR	■ Spectral region: 2500–25000 nm; ■ Commonly used in qualitative analysis of medicinal plants	■ High sensitivity; ■ High selectivity	■ Susceptible to moisture; ■ Slow detection speed
	FT-IR	■ Spectral region: 2500–25000 nm; ■ Commonly used in qualitative analysis of medicinal plants	■ High sensitivity; ■ High resolution; ■ High accuracy; ■ Rich spectral information	■ Tablet pressing is easily affected by moisture; ■ Long detection time
	Raman	■ Less applied in the quality evaluation of medicinal plants	■ Not affected by moisture; ■ High sensitivity; ■ Simple operation	■ Susceptible to fluorescence
	NMR	■ Rarely used in the quality evaluation of medicinal plants	■ Good reproducibility; ■ No need for complex preprocessing	■ Severe signal overlap
	HSI	■ A promising technique for evaluating the quality of medicinal plants	■ Simultaneously obtaining spectral and spatial information; ■ Wide scanning wavelength	■ Easy to generate redundant information; ■ Time consuming
Chromatography	HPLC	■ Commonly used for quality control of medicinal plants	■ High separation efficiency; ■ High sensitivity	■ High cost; ■ Environmental pollution
	UHPLC	■ Commonly used for quality control of medicinal plants	■ Low solvent consumption; ■ Quick analysis speed	■ Expensive equipment
	GC	■ Mainly used for the analysis of volatile or semi-volatile components in medicinal plants	■ High specificity; ■ High sensitivity; ■ Strong separation ability	■ Weak qualitative ability; ■ Reference substance required
Elemental analysis	ICP-MS	■ Mainly used for element detection in medicinal plants	■ High sensitivity; ■ Wide linear dynamic range	■ Complex sample preparation; ■ Environmental pollution
	LIBS	■ Mainly used for element detection in medicinal plants	■ No sample preparation required; ■ Real time and fast; ■ Low cost	■ Affected by matrix interference; ■ Self-absorption effect
Artificial sensing technique	E-nose	■ More and more applications for quality evaluation of medicinal plants	■ Quick analysis; ■ Small sample size required;	■ Difficult to quantify
	E-tongue	■ Quality evaluation of medicinal plants mainly used in liquid form	■ Simple and fast; ■ Cross-sensitivity; ■ Non-destructive and pollution-free	■ Susceptible to humidity and temperature
	CV	■ More and more applications in the classification of medicinal plants	■ Simple and fast; ■ Non-destructive and pollution-free	■ Limited information obtained
Traditional machine learning	PCA	■ Commonly used for exploratory analysis and feature extraction in quality evaluation of medicinal plants	■ Less information loss; ■ Simple operation; ■ Low computational cost	■ Poor robustness; ■ Local optimal solution
	t-SNE	■ Rarely used for exploratory analysis in quality evaluation of medicinal plants	■ Preserve local structure of information; ■ Strong ability to visualize high-dimensional data	■ Time-consuming; ■ Slow training speed
	HCA	■ Mainly used for exploratory analysis in quality evaluation of medicinal plants	■ Less restrictive factors; ■ Easy to define proximity	■ Vulnerable to singular value; ■ Computational complexity
	PLS-DA	■ The most commonly used methods in qualitative analysis of medicinal plants	■ Eliminating noise; ■ Strong ability to solve high multi-collinearity	■ Affected by the number of latent variables
	SVM	■ Commonly used for qualitative analysis of medicinal plants	■ Strong robustness; ■ Avoiding dimensional disasters; ■ Strong generalization ability	■ Difficulty in multi-classification; ■ Training time consumption
	LDA	■ Less commonly used in qualitative analysis of medicinal plants	■ Capable of classification and dimensionality reduction	■ Risk of overfitting; ■ Not suitable for non-Gaussian distribution data
	k-NN	■ Less commonly used in qualitative analysis of medicinal plants	■ Simple concept; ■ Low training complexity; ■ Not sensitive to outlier	■ High computational complexity; ■ Weak interpretability

(*continued on next page*)

Table 1 (*continued*)

Types	Techniques/ algorithms	Application characteristics	Advantages	Disadvantages
	RF	▪ Commonly used for qualitative analysis of medicinal plants	▪ Fast training speed; ▪ Excellent ability to handle high-dimensional data; ▪ Strong robustness	▪ Risk of overfitting; ▪ Vulnerable to noise interference
	ELM	▪ Less commonly used in qualitative analysis of medicinal plants	▪ Fast learning speed; ▪ Strong generalization ability	▪ Vulnerable to outlier; ▪ Risk of overfitting
	SIMCA	▪ Less commonly used in qualitative analysis of medicinal plants	▪ Eliminating noise; ▪ Eliminating irrelevant variables; ▪ Maximizing the retention of complete information	▪ Vulnerable to outlier; ▪ Lack of robustness
	PCR	▪ Less commonly used in quantitative analysis of medicinal plants	▪ Simple calculation; ▪ Effectively handle multi-collinearity	▪ Poor robustness; ▪ Poor interpretability
	PLSR	▪ The most commonly used methods in quantitative analysis of medicinal plants	▪ Strong interpretability; ▪ High prediction accuracy; ▪ Effectively handle multi-collinearity	▪ Vulnerable to outlier; ▪ Affected by sample distribution
	MLR	▪ Less commonly used in quantitative analysis of medicinal plants	▪ Simple operation; ▪ Good predictive performance	▪ Vulnerable to outlier; ▪ Vulnerable to factors limitations
	SVR	▪ Commonly used for quantitative analysis of medicinal plants	▪ Minimizing total loss; ▪ Effectively processing high-dimensional data	▪ High computational cost; ▪ Vulnerable to sample size influence
Deep learning	ANN	▪ Less commonly used in qualitative analysis of medicinal plants	▪ Flexible and automatic feature extraction; ▪ Parallel distributed processing; ▪ Robustness and fault tolerance	▪ Lack of interpretability; ▪ Information loss
	CNN	▪ Less commonly used in qualitative analysis of medicinal plants	▪ Flexible and automatic feature extraction; ▪ Effectively avoid overfitting; ▪ Strong ability to handle high-dimensional data	▪ Lack of interpretability; ▪ Gradient disappearance
	ResNet	▪ Less commonly used in qualitative analysis of medicinal plants	▪ Flexible and automatic feature extraction; ▪ Protecting information integrity; ▪ Simplify learning objectives	▪ Lack of interpretability; ▪ Redundant information

UV-Vis: ultraviolet-visible; NIR: near-infrared; MIR: mid-infrared; FT-IR: Fourier transform infrared; NMR: nuclear magnetic resonance; HSI: hyperspectral imaging; HPLC: high performance liquid chromatography; UHPLC: ultra-high performance liquid chromatography; GC: gas chromatography; ICP-MS: inductively coupled plasma-mass spectrometry; LIBS: laser-induced breakdown spectroscopy; E-nose: electronic nose; E-tongue: electronic tongue; CV: computer vision; PCA: principal component analysis; *t*-SNE: *t*-distributed stochastic neighbor embedding; HCA: hierarchical clustering analysis; PLS-DA: partial least squares discriminant analysis; SVM: support vector machine; LDA: linear discriminant analysis; *k*-NN: *k*-nearest neighbors; RF: random forest; ELM: extreme learning machine; SIMCA: soft independent modeling of class analogy; PCR: principal component regression; PLSR: partial least squares regression; MLR: multivariate linear regression; SVR: support vector regression; ANN: artificial neural networks; CNN: convolutional neural networks; ResNet: residual convolutional neural network.

become effective means of qualitative analysis. Complementary sources should be combined based on data attributes, such as the combination of Raman and NIR spectroscopy, which yields excellent results.

2.1.2. *Chromatography*

The advantages and disadvantages of utilizing chromatography for evaluating the quality of medicinal plants are undeniable. This technique requires expensive equipment and time, and its involvement in chemical applications contradicts the concept of environmentally friendly practices. However, due to its powerful separation abilities, chromatography has become the predominant technique in the quantitative analysis of medicinal plants, with even recently developed quantitative techniques unable to challenge its dominant position. This section provides a review of commonly used chromatography methods for evaluating the quality of medicinal plants.

Liquid chromatography (LC), particularly high performance liquid chromatography (HPLC) and ultra-high performance liquid chromatography (UHPLC), is the primary tool employed to determine the secondary metabolites of medicinal plants, thereby facilitating quality control [29]. The high precision and resolution of these techniques make them crucial in quantitative analysis and adulteration detection of medicinal plants [30]. However, LC has limitations in terms of the types of compounds it can analyze, and its detection range is determined by the polarity of the tested substances [31]. Gas chromatography (GC), on the other hand, exhibits robust analytical capabilities for volatile or semi-volatile secondary metabolites in medicinal plants [32]. To further characterize the chemical structure of complex mixtures, continuous advancements and development of chromatographic devices are inevitable trends [33]. Mass spectrometry (MS), a conventional detection technique, enables the elucidation and quantification of secondary metabolites in medicinal plants on a large scale by providing MS^n information and a wide dynamic linear range [29]. The integration of chromatography and MS offers a viable scientific solution for the quality evaluation of medicinal plants, and this hyphenated technique enhances analytical capabilities through coupling, ensuring high accuracy and precision [33,34].

Chromatography serves as an advantageous platform for both quantitative and qualitative analysis of medicinal plants. When combined with spectroscopy as complementary analysis methods, chromatography has demonstrated exceptional outcomes in predicting the content of medicinal plants. However, for the long-term development of chromatography, it is crucial to prioritize research and development efforts toward environmentally friendly solvents. Additionally, it should be noted that there is a lack of existing data processing methods for chromatography, which hinders the acquisition of excellent fingerprints. In future research, greater attention should be given to the data processing and analysis of chromatography to obtain more accurate results.

2.1.3. Elemental analysis

Trace elements, as one of the fundamental components of medicinal plants, make significant contributions to various aspects such as disease prevention, human growth and development, and other aspects. The data on trace elements provide much available information for evaluating the quality of medicinal plants, including indications of pesticide or fertilizer usage [35]. Consequently, elemental analysis holds great research significance in the quality evaluation of medicinal plants. Inductively coupled plasma-mass spectrometry (ICP-MS) occupies a prominent position in elemental analysis [36]. However, it requires a complex sample pretreatment process hampers its development due to the challenges of real-time detection [37]. In contrast, laser-induced breakdown spectroscopy (LIBS) utilizes an ultra-strong laser beam to generate plasma, enabling real-time elemental composition analysis of the sample [38]. LIBS offers a crucial advantage by directly measuring elements in medicinal plants with high speed and resolution under atmospheric conditions, eliminating the need for time-consuming pre-processing [39]. Both ICP-MS and LIBS are employed techniques for elemental analysis in medicinal plants. Overall, elemental analysis has received limited attention in the context of medicinal plants, and further research in this area is warranted. However, it is crucial to recognize the importance of multi-element monitoring as a standard for assessing the quality of medicinal plants. Monitoring multiple elements is essential to obtain a comprehensive understanding of the plant's quality.

2.1.4. Artificial sensing technique

The color, morphology, and flavor are important properties for evaluating the quality of medicinal plants. Traditionally, sensory evaluation and modern analytical techniques such as spectroscopy and chromatography have been employed for this purpose. However, some of these methods exhibiting subjective differences are time-consuming and costly, making real-time monitoring challenging in the market [40]. The emergence of artificial sensing techniques has opened up new possibilities for the quality evaluation of medicinal plants. Artificial sensing techniques such as electronic nose (e-nose), electronic tongue (e-tongue), and computer vision (CV) systems are commonly used to replace human olfactory, gustatory, and visual senses in analyzing the sensory properties of complex mixtures in medicinal plants.

In many studies, the combination of e-nose and e-tongue is utilized as an analytical tool, especially for assessing the flavor information in medicinal plants. These sensor systems consist of a sensor array, data processing units, and pattern recognition systems [41]. Although they share similar working principles, the sensor arrays used in e-nose and e-tongue differ due to the distinct objects being studied. The e-nose employs gas sensor arrays to identify volatile or semi-volatile components in medicinal plants inspired by the human olfactory nervous system [42,43]. Metal oxide sensors are widely used in e-nose as a key component of the gas sensor array [44]. On the other hand, the e-tongue employs liquid sensor arrays driven by different measurement principles, such as optics and electrochemistry, to analyze specific chemical components by generating electrical signals [45]. Among the measurement methods, the electrochemical approach takes the lead in the analysis of liquid substances [46]. Both e-nose and e-tongue employ data processing units to mitigate the impact of environmental factors on sensors, while the pattern recognition systems recognize sets of response outputs from sensor arrays and determine their categorical attributes [47]. CV is a rapidly developing electronic sensor technique that accurately describes the color, size, and surface structure of medicinal plants by capturing, processing, and analyzing images [48]. Proper illumination plays a crucial role in obtaining high-quality images, as a good light source effectively reduces interference factors such as reflections, shadows, and noise, thus reducing the workload of image processing [49]. However, the CV system has limitations, as it can only obtain chemical composition information of medicinal plants at specific spectral wavelengths, making it challenging to obtain additional information [50].

The highly complex chemical composition of medicinal plants necessitates the integration of multiple artificial sensing techniques for comprehensive evaluation. Combining multiple artificial sensor techniques has become a popular trend in the quality evaluation of medicinal plants. While the above-mentioned three artificial sensing techniques are excellent qualitative tools, achieving accurate quantification remains a challenge. Ongoing research in artificial sensing techniques is necessary for further improvement and refinement.

2.2. Different medicinal parts

As is well known, the clinical efficacy of traditional Chinese medicine results from its synergistic effect of multi-component, multi-target, and multi-channel. The quality of medicinal plants, as the material foundation of traditional Chinese medicine, significantly influences their safety. Holistic quality evaluation of medicinal plants primarily focuses on analyzing the chemical information of the same medicinal part using various analytical techniques. Previous studies have demonstrated that the accumulation and distribution of secondary metabolites in medicinal plants vary across different medicinal parts [51,52]. The environment plays a crucial role in regulating the chemical composition content of medicinal plants, and different medicinal parts exhibit distinct responses to the environment, as well as internal and external influences that contribute to these variations [53]. However, the evaluation of medicinal plants using chemical information from different medicinal parts as multi-source data is rare and often overlooked in existing research. The chemical information from different medicinal parts can provide robust support for the comprehensive evaluation of medicinal plant quality and better reflect their response and adaptability to the environment [54].

3. Data processing for multi-source data

Data processing forms the core component of multi-source data analysis. The volume of information in multi-source data is enormous. Data processing methods can eliminate redundant information, extract key variables, and greatly accelerate the analysis process. Therefore, it is considered an effective measure to improve the results and predictive performance of the model.

3.1. Data pre-processing

Various data pre-processing methods are available for multi-source data obtained from different techniques. In the case of spectral data, one key challenge is dealing with artifacts, including baseline offset, noise, and multiplicative effects. There are various methods for baseline correction, such as data-driven and coarse to fine baseline correction scheme based on empirical mode decomposition, adaptive iteratively reweighted penalized least squares (airPLS), and other algorithms can effectively handle baseline drift introduced by spectroscopy, MS, chromatography, and other multivariate analysis techniques. This drift can otherwise blur the signal and lead to undesirable results [15,55–57]. Baseline correction produces sharper and more accurate outcomes. Standard normal variate (SNV) and multiplicative scatter correction (MSC) are powerful means to reduce the influence of scattering effects, solid particle size, and optical path variation in spectral data [58]. Normalization is primarily used to scale multi-source data proportionally, aligning them to the same range and interval. This normalization process is beneficial for reducing the impact of distribution differences, scales, and features on data fusion or modeling results. Orthogonal signal correction (OSC) demonstrates satisfactory performance in improving the model's prediction ability and simplifying the analysis process. OSC can filter out spectral changes that are orthogonal and uncorrelated with the response variables to a certain extent [15]. Derivative combined with Savitzky-Golay (S-G) smoothing, as a pre-processing method for spectral data, enhances spectral resolution and sensitivity without introducing additional noise, while preserving valuable signal characteristics such as height, width, and shape [15,59]. Finding the optimal pre-processing method typically involves using a trial and error approach to explore all available options, which can be time-consuming [60]. However, the emergence of ensemble pre-processing methods provides a fast and effective solution for determining the best approach. The main advantage of ensemble pre-processing is its ability to eliminate artifacts by combining multiple pre-processing methods, as opposed to using a single method [61]. A design of experiments (DOE)-based approach evaluates the performance of corresponding models based on the order in which pre-processing strategies are applied, enabling the identification of the optimal combination of pre-processing [62]. With the growing recognition of complementary information, multi-block analysis based on sequential methods is gaining popularity. For example, sequential pre-processing through orthogonalization (SPORT) can sequentially extract different pre-processed information from various blocks, thus, elucidating the maximum variability of the response variables [63].

Data pre-processing plays a crucial role in chromatographic analysis as well. The complex signals present in chromatographic data, along with standard baseline drift and small time offset, increase the challenges of extracting important information [64]. Assisted baseline estimation and denoising using sparsity (BEADS) has demonstrated favorable results in addressing baseline drift and noise in chromatographic data [65]. Another common pre-processing method for chromatographic data is the correlation-optimized warping (COW) algorithm, primarily employed to correct retention time shifts between samples [66]. Furthermore, the application of image recognition in the quality evaluation of medicinal plants has emerged as a prominent research direction. Various factors, such as inadequate lighting, low resolution, and the distance of imaging equipment, can lead to imperfect image acquisition [67]. Therefore, image pre-processing becomes an essential step to ensure the smooth progress of the image analysis process. Common image pre-processing operations include grayscale adjustment, geometric correction, noise reduction, contrast or sharpness enhancement, and defocus correction [49,68].

Selecting appropriate pre-processing methods based on the attributes of different types of multi-source data is essential for improving the final data analysis process. Ensemble pre-processing methods offer significant advantages in selecting the best pre-processing approaches and acquiring complementary information, making them highly appealing. While ensemble preprocessing is commonly applied to spectral data and suitable for situations where determining a single optimal option is challenging, it is worth noting that integrated pre-processing is rarely utilized in other types of data [61]. Exploring the integration of preprocessing techniques in different data types could serve as a potential breakthrough for future research.

3.2. *Feature extraction and variable selection*

Feature extraction and variable selection are essential steps in simplifying data analysis and mid-level data fusion. Feature extraction involves extracting variables from vector combinations, while variable selection focuses on selecting raw data variables that significantly contribute to the target attributes [42]. Principal component analysis (PCA) and partial least squares discriminant analysis (PLS-DA) are the simplest and most commonly used feature extraction methods that extract principal components (PCs) and latent variables (LVs) from data, respectively. However, these methods are primarily designed for small datasets and face challenges when dealing with large volumes of data. On the other hand, orthogonal total variation component analysis (OTVCA) has been proven superior to several commonly used feature extraction methods as it can preserve the spatial structure of features and greatly reduce data dimensions [69]. Autoencoder, as a nonlinear feature extraction tool, can effectively enhance model performance, particularly when spectral features are not apparent [70]. Convolutional neural networks (CNN), benefiting from convolutional and pooling layers, demonstrates strong feature extraction capabilities, especially for large datasets [71]. Variable selection methods include successful projections algorithm (SPA), competitive adaptive reweighed sampling (CARS), uninformative variable elimination (UVE), variable importance in projection (VIP), and interval partial least squares (iPLS). When faced with multiple options, the selection should be based on the characteristics of the data.

Implementing feature extraction and variable selection leads to improved predictive results in the assessment of medicinal plant quality. There is an increasing emphasis on the interpretability of extracted features or selected variables in evaluating the quality of medicinal plants. Future attention should focus more on detailed descriptions of variables or features.

3.3. *Data fusion strategy*

Data fusion strategies play a significant role in comprehensively evaluating the quality of medicinal plants by integrating multi-source data. These strategies can be categorized into three types: low-level data fusion, mid-level data fusion, and high-level data fusion. Fig. 2 showed the flow chart of three data fusion strategies, each with distinct characteristics. A deep understanding of these strategies is necessary to effectively apply them to multi-source data of medicinal plants.

The three types of data fusion integrate multi-source data in different ways, but their common goal is to obtain a more comprehensive response than a single data source. Low-level fusion involves simply concatenating data from different analysis platforms or biological entities to create a new matrix. It is also known as data-level fusion. Given that multi-source data may have different scales, variable scaling before concatenation is necessary to combine the data effectively [10,72]. Low-level fusion has the lowest operational difficulty and simplest principle among the three fusion strategies, as it directly fuses the data while retaining the raw data [73]. Although the direct combination method may introduce redundant information, which can slow down analysis speed and affect result accuracy. Pre-processing is strongly

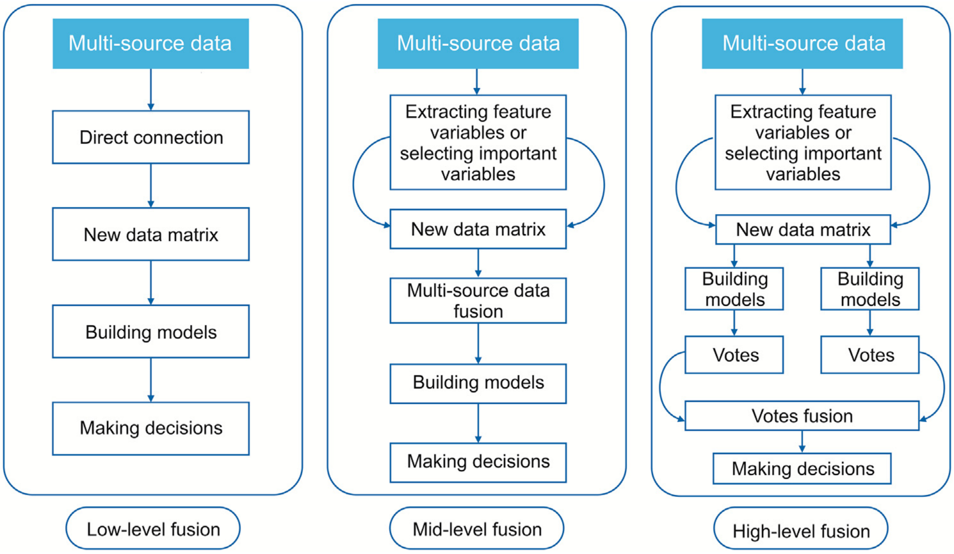

Fig. 2. The flow chart of three types of data fusion strategy.

recommended to improve the accuracy of qualitative or quantitative analysis results. Mid-level fusion, also performed at the data level, implements feature extraction or variable selection methods within each dataset. The fused data comprises the extracted features or selected variables from each dataset. It is referred to as feature-level fusion and overcomes the drawback of low-level fusion by reducing the volume of data. Mid-level fusion enhances result interpretability by visualizing the contributions of each dataset [74]. It is the most popular fusion strategy in the quality evaluation of medicinal plants. However, mid-level fusion may still encounter issues such as noise data or collinearity between different data sets, which can impact the model's predictive performance [75]. The application of high-level fusion in the field of medicinal plants is relatively limited. High-level fusion is typically used to address multi-classification problems and differs from low-level and mid-level fusion as it is performed at the prediction level [76]. In high-level fusion, each dataset independently establishes a classification or regression model, and the final decision result is obtained by fusing all the outputs. High-level fusion (decision-level fusion) is less susceptible to interference, resulting in stronger robustness [77]. However, it should be noted that directly fusing decision results can lead to information loss, and careful data processing is required [13]. What is more, high-level fusion is time-consuming and labor-intensive, as it involves establishing models for each dataset. Due to these reasons, its application in the quality evaluation of medicinal plants is relatively limited.

When choosing a data fusion strategy, the crucial aspect to consider is the relationship between samples and variables. It is essential to select an appropriate strategy based on the underlying data structure [11]. As technical analysis platforms continue to diversify, existing data fusion strategies are also influenced and may need to adapt accordingly. Looking ahead, an important focus will be on enhancing the generalization ability of the three data fusion strategies.

4. Machine learning

4.1. Traditional machine learning

Traditional machine learning is particularly well-suited for addressing learning problems that arise in situations with limited sample sizes. There exists a wide range of methods within the realm of traditional machine learning. Notably, unsupervised learning and supervised learning have found extensive application in the evaluation of medicinal plants' quality. These methods encompass three fundamental types of analysis: Exploratory analysis, classification analysis, and regression analysis [67]. The summarization of these analyses is provided below.

4.1.1. Exploratory analysis

Exploratory analysis, grounded in statistical principles, uncovers the relationships between samples, variables, and sample-variable associations [78]. Each of these three types of information carries distinct meanings. Utilizing the inter-sample information, one can visualize sample distribution trends and detect outliers. The inter-variable relationships aid in discerning complementary, redundant, and similar information. Moreover, the sample-variable associations reflect the significant contribution of variables to sample classification.

Within the realm of medicinal plant quality evaluation, three commonly employed exploratory techniques are PCA, *t*-distributed stochastic neighbor embedding (*t*-SNE), and hierarchical clustering analysis (HCA). PCA serves the purposes of data exploration, feature extraction, and dimensionality reduction. It achieves orthogonal projection, transforming spatial data into a set of mutually orthogonal principal components (PCs), thereby eliminating redundant information. Given the abundant information in multi-source medicinal plant data, PCA is frequently employed for data visualization and dimension reduction before qualitative and quantitative analysis [79]. By utilizing multiple PCs to express high-dimensional data, PCA maximizes the preservation of original information while mitigating the challenges of high-dimensional multi-objective optimization [80,81]. As a nonlinear dimensionality reduction algorithm, *t*-SNE excels in visualizing high-dimensional data and compensates for PCA's limitation in preserving the local data structure [79]. HCA, on the other hand, adopts a connectivity-based model that generates a hierarchical clustering structure by assessing the proximity between observations [82]. Various proximity measures, such as Mahalanobis distance, Euclidean distance, Manhattan distance, and *D* value, can be employed. The resulting clustering outcomes are easily comprehensible as they present the multi-level organization of clusters through a dendrogram. HCA and PCA are often regarded as complementary approaches, and their combined usage in medicinal plant quality evaluation is not uncommon [83]. The significance of exploratory analysis is frequently undervalued, despite its crucial role in comprehending data relationships before engaging in classification and regression analysis. It indirectly guides model selection by providing a deeper understanding of the interplay among various data elements.

4.1.2. Classification analysis

In the context of medicinal plant quality evaluation, classification models play a vital role in conducting qualitative analysis. Specifically, they are employed for geographical traceability, adulteration identification, and variety identification. These models fall under the domain of supervised learning, necessitating the training of the model using prior knowledge to distinguish unknown samples. Classification models can be categorized into two types based on the specific problems they aim to address. The first type is discriminant analysis, which evaluates predefined categories of target substances. The second type is class modeling, which assesses the assignment of the target object to a selected category of interest [84].

Several discriminant analysis methods are frequently utilized in the quality evaluation of medicinal plants, including PLS-DA, support vector machine (SVM), linear discriminant analysis (LDA), *k*-nearest neighbors (*k*-NN), random forest (RF), and extreme learning machine (ELM). Among these, PLS-DA has emerged as the most commonly employed method shown in relevant research papers. PLS-DA is a linear discriminant analysis technique that establishes correlations between independent and dependent variables to assign objects to known groups [85]. It exhibits exceptional performance when applied to quality evaluation tasks involving limited sample sizes of medicinal plants. Its ability to handle highly collinearity and noise data is nearly flawless, which explains its status as a mainstream discriminant analysis method [86]. Another popular classification analysis method is SVM, which constructs an optimal hyperplane in high-dimensional space to separate positive and negative classes [87]. While SVM demonstrates excellent generalization and recognition capabilities, its extensive computational requirements and time-consuming nature make it a less preferred choice. In contrast, soft independent modeling of class analogy (SIMCA) represents the most classical class modeling method. It evaluates the distance between a sample and the model by performing PCA on each predefined class [88]. Unlike discriminant analysis methods, SIMCA has the ability to assign objects to multiple classes or to none at all [89].

The aforementioned classification analysis methods heavily depend on manual labeling, and their performance in classifying large sample datasets requires enhancement to remain competitive in market applications. Various approaches for improving these methods have been proposed, but they often consist of partial optimizations. The crucial aspect lies in enhancing their ability to analyze big data and increase their applicability in the market.

4.1.3. Regression analysis

The regression analysis method employs mathematical calculations to examine the relationship between data properties and corresponding continuous variables, such as the chemical composition content and adulteration rate of medicinal plants. It plays a crucial role in the quality evaluation of medicinal plants, primarily for content prediction and quantitative analysis of adulterants [90]. The regression analysis methods commonly used in medicinal plant quality evaluation include principal component regression (PCR), partial least squares regression (PLSR), multivariate linear regression (MLR), and support vector regression (SVR). Among these methods, PLSR, a linear regression analysis method, is widely utilized in quantitative analysis of medicinal plants. PLSR divides the data into scores and loadings and performs least squares regression on the extracted scores, which exhibit maximum covariance with the response [91]. By extracting important variables to explain the dependent variable, PLSR mitigates the risk of overfitting [92]. On the other hand, SVR, the most prevalent nonlinear regression analysis method, employs rigorous mathematical derivation to conduct regression analysis. It delivers excellent prediction results for low-dimensional data with small sample sizes [93]. However, when confronted with regression analysis involving large sample sizes, SVR's prediction outcomes are not satisfactory, and a practical solution to this limitation has yet to be identified.

It should be noted that the aforementioned regression analysis methods are solely applicable to the analysis of single modal data. However, the multi-source data of medicinal plants is not only multivariate but also multimodal. Consequently, the trend in subsequent applications for quantitative analysis of medicinal plants is to employ multimodal prediction modeling. Such an approach effectively enhances prediction performance and facilitates the interpretation of shared and distinct information within multi-source data.

4.2. Deep learning

Deep learning, also known as deep neural networks, possesses inherent strengths in capturing data-driven patterns and automatically extracting feature information and hidden data structures from large datasets [94]. It relies on nonlinear information processing to extract multi-level features for data exploration and prediction, making it a valuable tool for addressing qualitative and quantitative challenges in medicinal plant quality evaluation involving large sample sizes [95]. Artificial neural networks (ANN) and CNN are two popular deep learning architectures extensively employed in the quality evaluation of medicinal plants. ANN exhibits a relatively simple structure, comprising three distinct layers: input layer, hidden layer, and output layer. Through the interconnection of numerous simple components, ANN forms a highly nonlinear complex network capable of performing intricate logical operations and establishing nonlinear relationships [96]. It excels in handling qualitative and quantitative information, and its generalization ability and fault tolerance can be enhanced by adjusting weights [97]. In contrast, CNN possesses a more intricate structure, consisting of five main components: input layer, convolutional layer, pooling layer, fully connected layer, and output layer. While ANN learns feature information through input and input mapping, CNN extracts feature information by performing convolution operations on the output of each layer [98]. The convolution layer, a crucial element in the CNN structure, exhibits the advantage of sharing and creating sparse connections in filters to mitigate model overfitting [99]. However, CNN has a notable drawback that as the network deepens, it is prone to issues such as gradient vanishing or explosion [100]. Residual convolutional neural network (ResNet) addresses this problem by introducing residual modules within the traditional CNN structure, utilizing skip connections to overcome gradient vanishing [101,102]. This approach simplifies learning objectives while preserving information integrity, significantly enhancing data analysis efficiency.

The application of deep learning in the quality evaluation of medicinal plants in scenarios involving large samples provides unique advantages, bridging the gap left by traditional machine learning in analyzing massive data. Nevertheless, the interpretability of the features extracted by deep learning remains a challenging issue that warrants further research.

5. The application of machine learning combine with multi-source data of medicinal plants

5.1. Geographical traceability

Geographical traceability plays a crucial role in controlling the quality of medicinal plants, as the chemical composition of medicinal plants can vary due to diverse growth conditions such as terrain, soil, and climate. For instance, Lu et al. [103] discovered significant variations in the contents of ganoderic acid A and B, polysaccharide, and triterpenoids in *Ganoderma lucidum* sourced from different geographical regions. Therefore, to ensure the quality and safety of clinical medication, it is imperative to employ a reliable and suitable method for tracing the origin of medicinal plants. However, due to the systematism, multi-target and synergistic effects of medicinal plants, identifying their origin based on a single or multiple chemical components, proves challenging. Relying solely on a single analytical technique has its limitations and fails to provide comprehensive information regarding the chemical composition of medicinal plants [104]. Establishing a robust geographical traceability system for medicinal plants poses a considerable challenge. Nevertheless, leveraging multi-source data can bridge the gap between various analytical instruments and reveal the holistic chemical profile, thus, serving as an effective measure to trace the geographical origin of medicinal plants [105]. Table 2 [54,106–122] summarized some examples of using machine learning algorithms in combination with multi-source data to identify the geographical origin of medicinal plants.

The combination of different spectral techniques as data sources, coupled with machine learning, has shown promising results in the geographical traceability of medicinal plants. In a study by Li et al. [115], FT-MIR and NIR data were collected from cultivated *Panax notoginseng* samples originating from Wenshan and four other regions in Yunnan. Three data fusion strategies, in conjunction with machine learning algorithms (SVM and RF), were employed to analyze the preprocessed data sets using SNV and S-G filter to identify the geographical origin of *Panax notoginseng*. The models based on a single data source exhibited incomplete accuracy in classifying *Panax notoginseng* samples from the five geographical origins, with classification accuracy ranging from 91% to 94%. However, the use of three data fusion strategies enhanced the classification accuracy, with SVM models based on high-level data fusion demonstrating the best performance, achieving classification accuracy ranging from 98% to 100% for both the training and test sets. Another study also demonstrated the efficacy of a collaborative strategy involving FT-MIR and NIR as multi-source data, along with data fusion strategies, for authenticating the origin of *Panax notoginseng* [113]. The results revealed that the random forest Boruta (RF-Bo) and random forest variable selection (RF-Vs) models, based on the data fusion strategy, exhibited favorable classification performance, with RF-Bo being faster in data analysis and achieving an accuracy rate of 95.6%. The growth environment significantly influences the quality of *Paris*, making its origin a crucial factor. In the study of Wu et al. [110], the feasibility of employing FT-MIR and UV-Vis alone or in combination for the geographical traceability of wild *Paris polyphylla* var. *yunnanensis* (Franch.) Hand.-Mazz (PPY) was evaluated. The spectral data were processed using PLS-DA and support vector machines grid search (SVM-GS) supplemented by a low- and mid-level data fusion strategy. The study found that the classification models based on the data fusion strategy outperformed those based on a single data source. For PPY samples from the southeast and northwest regions of Yunnan, China , PLS-DA and SVM-GS models utilizing intermediate data fusion could accurately classify them, as feature extraction and variable selection eliminated redundant information and enhanced model performance. In addition to spectral methods, chromatography can also serve as a multi-source data approach for the geographical traceability of medicinal plants. In the case of *Gentiana rigescens*, five variable selection methods (CARS, RF, genetic algorithm (GA), MC-UVE, and SPA) were utilized to extract characteristic fingerprint information from HPLC and FT-IR. Subsequently, PLS-DA models were developed to identify the geographical origin of *Gentiana rigescens* [123]. The results indicated that the GA-PLS-DA model was robust in identifying *Gentiana rigescens*, and the HPLC and FT-IR results were consistent. Moreover, FT-IR demonstrated faster detection speed compared to HPLC.

In summary, the integration of data fusion strategies with machine learning algorithms is increasingly prevalent in the geographical traceability of medicinal plants. It has been demonstrated that this approach yields significantly higher classification accuracy compared to using a single analytical method. Consequently, it offers a novel solution for the geographical traceability of medicinal plants. Additionally, due to various factors, such as the nature of chromatographic methods, there have been fewer applications in the field of geographical traceability of medicinal plants. As a result, spectroscopy methods have emerged as the dominant choice in this domain.

5.2. Adulterate identification

Adulteration represents a significant factor that compromises the quality of medicinal plants. Adulterants typically do not alter the taste or chemical composition of the medicinal plants, making their identification challenging through visual inspection alone. Due to the limited annual production of certain medicinal plants, such as *Panax notoginseng* and *Paris*, the market demand often surpasses the supply. Unscrupulous merchants exploit this gap by resorting to adulteration to meet the market's quantity requirements, resulting in severe consequences for consumer interests. Therefore, employing reliable methods to detect adulteration in medicinal plants and ensure their quality is imperative. The process of identifying adulteration in medicinal plants is highly intricate, and conventional target analytical methods can only analyze a limited number of labeled components [124]. However, the data fusion strategy has proven to be effective in enhancing the accuracy of adulteration detection and has found widespread application in the domains of food and medicinal plants. Table 3 [125–137] summarized some examples of applying machine learning algorithms in combination with multi-source data to identify medicinal plant adulteration.

Uncaria tomentosa (UT), known for its higher content of active ingredients, is susceptible to adulteration with the low-value *Uncaria guianensis* (UG). Kaiser et al. [126] applied classification (KNN, SIMCA) and regression (PCR, PLS) machine learning algorithms to analyze liquid chromatography-photo diode array (LC-PDA), UV, and FT-IR data. This allowed them to identify and quantify adulteration in UT. Comparison of discrimination and regression models built using different data sources revealed that UV and LC-PDA analysis of polyphenols demonstrated excellent identification and prediction results. Specifically, SIMCA performed well in recognition, while PLS excelled in quantification. *Panax notoginseng*, valued for its economic worth and pharmacological effects, frequently falls victim to challenging and recurrent adulteration. Infrared spectroscopy offers non-destructive, rapid, and cost-effective advantages for qualitative analysis. Yang et al. [137] trained SVM models using manually extracted NIR and FT-MIR characteristic wavelengths as input data to identify adulteration in *Panax notoginseng* powder. They employed data fusion and particle swarm optimization (PSO) as auxiliary methods to improve model performance. The PSO-SVM model

Table 2
Summary of references for geographic traceability using machine learning combined with multi-source data.

Year	Object	Multi-source	Strategy	Data processing	Machine learning	Result	Refs.
2019	*Gentiana rigescens* Franch.	HPLC fingerprint data sets of different medicinal parts	LLF, MLF	VIP	RF, OPLS-DA, PCA	Acc (LLF-OPLS-DA): 97.87%–100.00%, SEN: 0.96–1.00, SPE: 0.98–1.00, MCC: 0.95–1.00, EFF: 0.97–1.00	[54]
2019	*Paris polyphylla* var. *yunnanensis* (Franch.) Hand.-Mazz	UV, FT-IR	NM	SNV, OSC, FD, SD, S-G, CARS	PLSR	$R^2 > 0.9$, RPD = 3.3515	[106]
2022	*Gentiana rigescens* Franch.	FT-IR, HPLC	NM	NM	ResNet	Acc (synchronous 2DCOS model) = 100%	[107]
2016	*Cuminum cyminum* L.	E-nose, E-tongue, SPME-GC-MS	NM	Feature extraction	HCA, PCA, SVM	Acc (E-nose) = 94.44%, Acc (E-tongue) = 100%	[108]
2022	*Abrus precatorius* L. leaves	UPLC-MS, HPLC, NMR	NM	NM	HCA, PCA, OPLS-DA	$R^2X = 0.61$, $R^2Y = 0.848$, $Q^2 = 0.717$	[109]
2018	*Paris polyphylla* var. *yunnanensis* (Franch.) Hand.-Mazz	UV-Vis, FT-IR	LLF, MLF	SNV, derivatives, S-G filter and different combinations, PCA	PLS-DA, SVM-GS	MLF-PLS-DA: Q^2(central) = 0.74, Acc (MLF-SVM-GS) = 98%	[110]
2019	*Paris polyphylla* var. *yunnanensis* (Franch.) Hand.-Mazz	ATR-FTIR spectra data sets of Paridis rhizome and leaf tissues	MLF	MSC, S-G filter, PCA	*t*-SNE, PLS-DA, RF	Acc (PLS-DA) = 100%, RF: parameters values are large than 0.85	[111]
2021	*Amomum tsaoko* Crevost & Lemaire	FT-NIR, UV-Vis	MLF	PCA, SO-PLS, VIP, SO-CovSel	PLS-DA	Acc (SO-PLS-PLS-DA) = 100%	[112]
2020	*Panax notoginseng* (Burkill) F. H. Chen	NIR, FT-IR	LLF, MLF, HLF	SNV, S-G, derivative and different combinations	RF-Bo, RF-Vs	Acc (HLF-RF-Bo) = 95.6%	[113]
2019	*Wolfiporia cocos* (F.A. Wolf) Ryvarden & Gilb.	ATR-FTIR, UFLC	LLF, MLF	correlation optimized warping algorithm, SD, VIP, PCA	PLS-DA	Acc (LLF-ATR-FTIR) = 100%	[114]
2017	*Panax notoginseng* (Burkill) F. H. Chen	FT-IR, NIR	LLF, MLF, HLF	SNV, S-G filter	RF, PCA	Acc (HLF): 98%–100%	[115]
2022	*Wolfiporia cocos* (F.A. Wolf) Ryvarden & Gilb.	FT-IR, FT-NIR	LLF, MLF	CARS, SPA, FD, SD, SNV, MSC, S-G filter	*t*-SNE, PLS-DA, SVM, ELM, HCA	Acc = 100%	[116]
2022	*Astragalus membranaceus* (Fisch.) Bunge, *Astragalus membranaceus* var. *mongholicus* (Bunge) P. K. Hsiao	IR, LIBS	LLF, MLF	SNV, MSC, FD, WT, normalization, VIP	RF, PCA	MLF-RF: SEN = 0.9667, SPE = 0.9833, Acc = 0.9778	[117]
2022	*Lycium barbarum* L. leaves	Vis-NIR-HSI, textural data	LLF, MLF	CARS, iVISSA, UVE, IRIV, GLCM	2D-CNN, PCA	iVISSA-MLF: Acc = 97.34%, mean F1: 100%	[118]
2022	*Corylus avellana* L.	FT-NIR, NMR	LLF	MSC, bucketing, mean averaging	Discriminant classifier, PCA	Acc = 96.6% ± 2.8%	[119]
2022	*Paris polyphylla* var. *yunnanensis* (Franch.) Hand.-Mazz	NIR, MIR	LLF, MLF, HLF	SO-PLS, SO-CovSel, SPA, CARS, FD, SD, MSC, SNV, VSN, VIP	PLS-DA, SVM	PLS-DA: Acc = 96.03%, F1 = 96.01%	[120]
2019	*Dendrobium* Sw.	NIR, MIR	LLF, MLF	MSC, SD	PLS-DA, SVM-GS, RF	LLF: Acc (PLS-DA, SVM-GS) = 100%	[121]
2020	*Eucommia ulmoides* Oliv. leaves	FT-NIR, ATR-FTIR	LLF, MLF, HLF	MSC, SD	RF, PLS-DA, HCA, *t*-SNE, PCA	HLF-RF: Acc (calibration) = 92.86%, Acc (validation) = 93.44%	[122]

NM: no mention; UV: ultraviolet; FT-IR: Fourier transform infrared; PLSR: partial least squares regression; SNV: standard normal variate; OSC: orthogonal signal correction; FD: first derivative; SD: second derivative; S-G: Savitzky-Golay; CARS: competitive adaptive reweighted sampling; RPD: residual predictive deviation; HPLC: high-performance liquid chromatography; ResNet: residual convolutional neural network; E-nose: electronic nose; E-tongue: electronic tongue; SPME-GC-MS: solid phase micro-extraction gas chromatography-mass spectrometry; HCA: hierarchical cluster analysis; PCA: principal component analysis; SVM: support vector machine; UPLC-MS: ultra performance liquid chromatography-mass spectrometry; NMR: nuclear magnetic resonance; OPLS-DA: orthogonal projections to latent structures discriminant analysis; UV-Vis: ultraviolet-visible; LLF: low-level fusion; MLF: mid-level fusion; PLS-DA: partial least squares discrimination analysis; SVM-GS: support vector machines grid search; ATR-FTIR: attenuated total reflectance-Fourier transform infrared spectrometry; *t*-SNE: *t*-distributed stochastic neighbor embedding; MSC: multiplicative scatter correction; RF: random forest; FT-NIR: Fourier transform near infrared; SO-PLS: sequential and orthogonalized partial-least squares; VIP: variable importance in projection; SO-CovSel: sequential and orthogonalized covariance selection; NIR: near infrared; RF-Bo: random forest Boruta; RF-Vs: random forest variable selection; UFLC: ultra-fast liquid chromatography; HLF: high-level fusion; ELM: extreme learning machine; SPA: successive projection algorithm; LIBS: laser induced breakdown spectroscopy; WT: wavelet transform; Vis-NIR-HSI: visible and near infrared hyperspectral; 2D-CNN: two-dimensional convolutional neural network; iVISSA: interval variable iterative space shrinking analysis; UVE: uninformative variable elimination; IRIV: iterative retained information variable; GLCM: grey-level co-occurrence matrix; VSN: variables sorting for normalization; SEN: sensitivity; SPE: specificity; EFF: efficiency; Acc: accuracy.

achieved impressive accuracy rates of 96.65% and 96.97% for identifying two levels of adulteration in *Panax notoginseng* powder, surpassing the performance of the unoptimized model. Optimization algorithms prove effective in addressing poor model performance. Similarly, PLS-DA models were developed by fusing NIR and MIR data to differentiate genuine and adulterated *Pinellia ternata* [127]. Data fusion improved the model's classification ability, with the recognition accuracy of the PLS-DA model based on mid-level data fusion reaching 100%. Modeling a single spectrum as a data source lacked the ability to correctly distinguish adulterated samples. Infrared spectroscopy-based data fusion has demonstrated remarkable achievements in detecting adulteration in medicinal plants. Beyond spectroscopic techniques, emerging non-destructive techniques such as electronic sensor systems can serve as powerful tools for adulteration detection in medicinal plants. Saffron, a highly valued and pharmacologically diverse spice, is prone to adulteration

Table 3
Summary of references for adulterate identification using machine learning combined with multi-source data.

Year	Object	Multi-source	Strategy	Data processing	Machine learning	Result	Refs.
2021	*Crocus sativus* L.	NIR, MIR	NM	Mean-centering, SD	PCA, PLS-DA, PLSR	R^2 = 0.95−0.99	[125]
2020	*Uncaria tomentosa* (Willd.) DC., *Uncaria guianensis* (Aubl.) Gmel.	HPLC-PDA, UV, FT-IR	NM	correlation optimized warping algorithm	SIMCA, *k*-NN, PCR, PLSR	SIMCA: SEN = 100%, SPE = 100%; PLSR: $R > 0.99$	[126]
2019	*Pinellia ternata* (Thunb.) Breit.	NIR, MIR	LLF, MLF	SNV, MSC, S-G smoothing, derivative and different combinations	PLS-DA	MLF: Acc = 100%, SEN = 1, SPE = 1	[127]
2014	*Harpagophytum procumbens* DC. ex Meisn.	^{1}H-NMR, UHPLC-MS	NM	NM	PCA, OPLS-DA	UHPLC-MS: R^2X = 0.258, R^2Y = 0.957, Q^2 = 0.934; ^{1}H-NMR: R^2X = 0.830, R^2Y = 0.865, Q^2 = 0.829	[128]
2017	*Crocus sativus* L.	CVS, E-nose	NM	NM	PCA, HCA, SVM, ANN-MLP	SVM: Acc(training) = 100%, Acc(validation) = 89%; ANN-MLP: $R^2_{\text{Color analysis}} >$ 0.95, $R^2_{\text{Aroma analysis}} >$ 0.97	[129]
2021	*Lonicera japonica* Thunb.	NIR, FT-IR	NM	VIP, CARS, SPA, UVE, SD, S-G smoothing	PLS-DA, PLSR	PLS-DA: Acc = 100%; SiPLS-VIP: RMSEP = 1.02%	[130]
2020	*Carthamus tinctorius* L.	NIR, CVS, HPLC	NM	VIP, SNV, MSC, S-G, baseline, S-G + FD, S-G + SD	PLS-DA, PLSR	PLS-DA: Acc = 100%; PLS: RPD(HSYA) = 2.5046, RPD(water extract) = 5.6195	[131]
2016	*Crocus sativus* L.	ATR-FTIR, Raman, LIBS	NM	S-G derivatization, mean-centering, normalization	PCA, PLSR	R^2 = 0.999, LOD = 1.86%, LOQ = 9.32%	[132]
2021	*Vanilla planifolia* Jacks. ex Andrews	NIR, MIR, Raman	NM	PCA	PLS-DA, PCA, SIMCA, SVM	PLS-DA: Acc (Raman) = 0.9, SEN (Raman, MIR) = 1, SPE (Raman, MIR) = 1, EFF (Raman, MIR) = 1, PRE (Raman, MIR) = 1	[133]
2023	*Cuminum cyminum* L.	FT-IR, portable NIR	NM	S-G smoothing, S-G + FD, SNV	PCA, PLSR, DD-SIMCA	DD-SIMCA: SEN (FT-IR) = 94.1%, SPE (FT-IR) = 91.7%; PLSR: RPD (FT-IR) = 8.9	[134]
2019	*Corydalis yanhusuo* W. T. Wang	FT-NIR, MIR	Data fusion	NM	PCA, MWPLS-DA, LDA, PLS-DA	MWPLS-DA-fusion: Acc (training) = 100%, Acc (prediction) = 100%	[135]
2021	*Origanum vulgare* subsp. *hirtum* (Link) Ietsw.	Four blocks of DART-HRMS data	MLF	NM	SVM, HCA, PLS-DA	MLF-SVM: SEN > 90%, SPE > 90%	[136]
2019	*Panax notoginseng* (Burkill) F. H. Chen	NIR, MIR	MLF	SNV, baseline correction, FD, PCA	PSO-SVM	L14: Acc = 96.65%; L15: Acc = 96.97%	[137]

NM: no mention; NIR: near-infrared; MIR: mid-infrared; PLSR: partial least squares regression; PCA: principal component analysis; PLS-DA: partial least squares discrimination analysis; SD: second derivative; HPLC-PDA: high performance liquid chromatography-photo diode array; UV: ultraviolet; FT-IR: Fourier transform infrared; SIMCA: soft independent modeling class analogy; *k*-NN: *k*-nearest neighbors; PCR: principal component regression; LLF: low-level fusion; MLF: mid-level fusion; HLF: high-level fusion; SNV: standard normal variate; MSC: multiplicative scatter correction; S-G: Savitzky-Golay; 1H-NMR: proton nuclear magnetic resonance; UHPLC-MS: ultra-high performance liquid chromatography coupled to mass spectrometry; OPLS-DA: orthogonal projections to latent structures discriminant analysis; CVS: computer vision system; E-nose: electronic nose; HCA: hierarchical cluster analysis; SVM: support vectors machine; ANN-MLP: two multilayer artificial neural network; VIP: variable importance for projection; CARS: competitive adaptive reweighted sampling; SPA: successive projection algorithm; UVE: uninformative variable elimination; SiPLS: synergy interval PLS; HSYA: hydroxy safflower yellow pigment A; FD: first derivative; ATR-FTIR: attenuated total reflectance-Fourier transform infrared spectrometry; LIBS: laser induced breakdown spectroscopy; SEN: sensitivity; SPE: specificity; EFF: efficiency; PRE: precision; Acc: accuracy; RPD: residual predictive deviation; LDA: linear discriminant analysis; MWPLS-DA: moving window partial least-squares discriminant analysis; DART-HRMS: direct analysis in real time-high resolution mass spectrometry; PSO: particle swarm optimization; FT-NIR: Fourier transform near infrared.

due to economic motivations. Kiani et al. [129] developed an integrated system based on computer vision and electronic nose to extract color and aroma characteristic variables from saffron samples for adulteration detection. SVM, PCA, and HCA based on color and aroma data exhibited strong consistency and effectively identified adulterated saffron samples. The SVM model achieved accuracies of 89% and 100% in identifying different types of adulterated saffron. The results were further corroborated by two multilayer artificial neural network (ANN-MLP) models, indicating that color and aroma could serve as indicators for detecting saffron adulteration. Specifically, the aroma characteristic variables exhibited outstanding ability in identifying saffron adulteration.

In conclusion, the application of machine learning algorithms in conjunction with multi-source data for statistical analysis holds significant promise in addressing fraud detection. This method offers an effective solution for accurately combating fraudulent practices in the medicinal plants industry. However, it is important to acknowledge that this approach still has certain limitations and cannot be readily applied for rapid and widespread market fraud detection in the domain of medicinal plants. Therefore, it is crucial to further enhance and refine the method to expedite its widespread adoption. Such advancements are of utmost importance in effectively controlling the quality of medicinal plants.

5.3. Variety identification

Medicinal plants belonging to the same genus often exhibit similar morphology and pharmacological effects. However, variations in active ingredients among species contribute to significant fluctuations in the quality of medicinal plants. While some differences can be visually identified, this approach relies on experienced professionals and is time-consuming and labor-intensive.

Therefore, the development of a rapid and scientific method for identifying different varieties of medicinal plants becomes imperative. Table 4 [64,138–151] summarized some examples of using machine learning algorithms in combination with multi-source data to identify the variety of medicinal plants.

The identification of different species within the genus *Mentha* L. is challenging due to variations in compounds. In an attempt to authenticate various mint species, UV-Vis and attenuated total reflectance-Fourier transform infrared (ATR-FTIR) spectroscopy, combined with machine learning, were employed [139]. The authentication process involved two steps. Initially, SIMCA was used to filter data for species other than spearmint and pepper. However, the classification effectiveness of SIMCA did not meet the requirements, necessitating the second step. In the second step, PLS-DA and SVM models were utilized to enhance classification accuracy. The results indicated that both PLS-DA and SVM showed potential as authentication tools, with UV-Vis making a more prominent contribution to the process. Rhubarb authenticity poses challenges due to variations in efficacy between official and unofficial sources, raising concerns about mixing and misuse in clinical medication. Sun et al. [138] distinguished rhubarb using three data fusion strategies of NIR and MIR. The fused datasets were subjected to PLS-DA, SIMCA, SVM, and ANN analyses. The modeling results revealed high classification accuracy for the four models using mid-level data fusion, with iPLS and wavelet compression (WC) as beneficial feature extraction methods. However, single spectral data-based classification models were inadequate for distinguishing official and unofficial rhubarb. NIR-HSI, combining machine vision and NIR advantages, facilitated the acquisition of spectral and spatial information. In the case of *Cinnamomum verum* and *Cinnamomum cassia*, which are commonly sold as powder and sticks, certification complexity arises. Cruz-Tirado et al. [141] conducted a quantitative analysis of active components using HPLC, and highlighting phenolic compounds as potential chemical markers. NIR-HSI information from cinnamon samples was collected and employed in combination with PLS-DA and SVM models for classifying *Cinnamomum verum* and *Cinnamomum cassia*. The category input was proposed based on the distribution of phenolic compounds. Both PLS-DA and SVM demonstrated excellent classification results, surpassing 90% accuracy. However, the models were species-specific, with PLS-DA performing better for *Cinnamomum verum* authentication and SVM being more suitable for *Cinnamomum cassia*. Integration of data from different chromatographic techniques, coupled with machine learning, also exhibits robust capabilities in identifying medicinal plant varieties. By collecting information on the active ingredients of genuine and non-genuine *Magnolia officinalis* leaves using ultra-high performance liquid chromatography-quadrupole time-of-flight tandem mass spectrometry (UHPLC-Q-TOF-MS/MS) and GC-MS, data integration via PLS-DA and heat map analysis demonstrated the potential of combining the two chromatographic techniques for identifying *Magnolia officinalis* leaves [150].

The process of identifying different varieties of medicinal plants is highly intricate and cannot be effectively accomplished using a single technique alone. Numerous studies have made significant progress in variety identification by leveraging multi-source data in conjunction with machine learning, particularly through the integration of qualitative and quantitative analysis. In recent years, rapid advancements in electronic sensors and chemical imaging techniques have opened up promising avenues in this field. These technologies hold considerable potential and can serve as alternatives to traditional macroscopic identification methods.

5.4. Content prediction

The potency of medicinal plants is predominantly determined by the levels of their active ingredients. Consequently, content prediction plays a crucial role in assessing the quality of medicinal plants. Content prediction is carried out using a calibration model. Crocin I and II contents serve as a key indicator for evaluating saffron quality, with strict regulations outlined in trade standards and the Chinese Pharmacopoeia. Notably, the crocin content of can be influenced by multiple factors, making content prediction an important measure for saffron quality control. Table 5 [152–171] summarized some examples of using machine learning algorithms in combination with multi-source data to predict the content of medicinal plants. Li et al. [152] developed PLSR models to establish correlations between NIR spectra and HPLC data, enabling the prediction of crocin I and II content. Different spectral ranges, pre-processing methods, and their combinations were considered to determine the calibration model's performance. The optimal crocin I content prediction model utilized the spectral range of 9403.3–7498.0 cm^{-1} and 6101.8–4246.5 cm^{-1}, along with vector normalization for spectral pre-processing. Similarly, the crocin II prediction model employed the spectral range, 9403.3–7498.0 cm^{-1}, 6101.8–5449.9 cm^{-1}, and 4601.4–4246.5 cm^{-1}, with vector normalization as the pre-processing method. The best prediction model achieved root mean square error of cross-validation (RMSECV) values of 1.4 and 0.3 for the two content types, demonstrating that selecting appropriate spectral ranges and pre-processing methods can significantly improve the calibration model's prediction performance. For predicting the vital active ingredient, puerarin, in Radix puerariae, PLS models were proposed using both low- and mid-level data fusion strategies [153]. The combination of NIR and UV yielded complementary effects, and iPLS was employed to extract feature variables and filter redundant information. The PLS model based on low-level data fusion exhibited superior prediction performance compared to single spectral data and mid-level data fusion. It achieved a root mean square error of prediction (RMSEP) and a residual predictive deviation (RPD) of 0.418 and 4.295, respectively. In a study by Song et al. [154], UV-Vis and ultra-high performance liquid chromatography-quadrupole-time of flight-mass spectrometry (UHPLC/Q-TOF-MS) were used for mid-level data fusion. A partial least squares regression model was established to predict the antioxidant capacity and total phenol content of bear fruit leaves. Although the PLS model based on independent UV spectra yielded reliable results, data fusion further improved the model's predictive ability. The mid-level data fusion PLS model demonstrated the best prediction performance for total phenol content and DPPH, with RPD values of 6.258 and 6.699, respectively. These studies highlight the power of combining data fusion and machine learning as a potent tool for accurately predicting single components.

In summary, content prediction plays a critical role in ensuring the quality control of medicinal plants. Enhancing the performance of calibration models through effective methods can yield significant improvements. It is worth noting that the current research focus is primarily on the geographical traceability of medicinal plants, with limited attention given to content prediction. However, content prediction holds a significant position in the quality evaluation of medicinal plants and merits greater attention in research endeavors.

5.5. Other applications

The use of machine learning algorithms in analyzing multi-source data of medicinal plants should extend beyond

Table 4
Summary of references for variety identification using machine learning combined with multi-source data.

Year	Object	Multi-source	Strategy	Data processing	Machine learning	Result	Refs.
2023	Soothing herbs	UV–Vis, HPLC	MLF	COW, PCA	PCA, PLS-DA	MLF: Acc (PLS-DA) = 87.5%	[64]
2017	*Rheum palmatum* L., *Rheum tanguticum* Maxim. ex Balf., *Rheum offificeinale* Baill.	NIR, MIR	LLF, MLF, HLF	WC, iPLS, MSC, FD, SD, S-G smoothing	PLS-DA, SVM SIMCA, ANN	MLF: Acc (PLS-DA) = 97.14%, Acc (SIMCA) = 94.73%, Acc (SVM) = 100%, Acc (ANN) = 100%	[138]
2021	Genus *Mentha* L.	ATR-FTIR, UV-Vis	NM	S-G smoothing, PQN	PLS-DA, SVM, SIMCA	Acc: 60%–80%	[139]
2023	*Amaranthus cruentus* L. and *Chenopodium quinoa* Will. seeds	UV-Vis, HPLC, GC	NM	NM	PCA, CDA	Probability (CDA) = 100%	[140]
2023	*Cinnamomum verum* J. Presl, *Cinnamomum cassia* (L.) J. Presl	NIR-HSI, HPLC	NM	SNV, ROI, S-G derivative	PCA, PLS-DA, SVM	Pixel-wise: Acc = 93.8%; sample-wise: Acc = 100%	[141]
2021	*Amomum tsaoko* Crevost & Lemaire, *Amomum paratsao-ko* S. Q. Tong & Y. M. Xia	GC-MS, FT-NIR	NM	Normalization, VIP, MSC, FD, S-G	PCA, OPLS-DA	$R^2 = 0.995$, $Q^2 = 0.961$, Acc = 100%	[142]
2019	*Citrus reticulata* Blanco, *Citrus reticulata* 'Chachi'	HPLC, HPTLC	NM	NM	OPLS-DA	Can effectively distinguish	[143]
2021	*Curcuma phaeocaulis* Val., *Curcuma kwangsiensis* S. G. Lee et C. F. Liang, *Curcuma wenyujin* Y. H. Chen et C. Ling	HPLC, HS-GC-MS	NM	VIP	PCA, LDA, *k*-NN, BPNN, OPLS-DA	HS-GC-MS: Acc (OPLS-DA) = 100%	[144]
2022	*Curcuma phaeocaulis* Val., *Curcuma kwangsiensis* S. G. Lee et C. F. Liang, *Curcuma wenyujin* Y. H. Chen et C. Ling	Spectrophotometry, flash GC E-nose	NM	VIP	PCA, PLS-DA, LDA	Acc (LDA) = 100%	[145]
2017	*Humulus lupulus* L.	NIR, MIR	NM	SNV, S-G filter	PCA, HCA, PLS-DA	Acc (NIR) = 94.2%, Acc (MIR) = 96.6%	[146]
2021	*Dendrobium* Sw.	FT-NIR, ATR-FTIR	NM	SDD	ResNet	Acc = 100%	[147]
2011	*Pelargonium sidoides* DC., *Pelargonium reniforme* Curt.	FT-NIR, FT-IR	NM	FD, SD, MSC, SNV	PCA, OPLS-DA	NIR: variation = 5.79%, R^2X = 0.962, $Q^2 = 0.918$; MIR: variation = 9.22%, R^2X = 0.497, $Q^2 = 0.658$	[148]
2014	Wood species	color, texture and spectral feature	MLF	GLCM	BPNN	Accuracy to approximately 90%	[149]
2022	Leaves of *Magnolia officinalis* Rehder & E. H. Wilson and *Magnolia officinalis* var. biloba Rehder & E. H. Wilson	GC-MS, UHPLC-Q-TOF-MS/MS	NM	VIP	PLS-DA, heat map analysis	UHPLC-Q-TOF-MS/MS: R^2X = 0.638, R^2Y = 0.929, $Q^2 = 0.649$; GC-MS: R^2X = 0.655, R^2Y = 0.979, $Q^2 = 0.934$	[150]
2014	*Harpagophytum procumbens* DC. ex Meisn., *Harpagophytum zeyheri* Decne.	MIR, SWIR	NM	NM	OPLS-DA, PCA	MIR: R^2X = 0.86, $Q^2 = 0.63$; SWIR: R^2X = 0.99, $Q^2 = 0.78$	[151]

NM: no mention; LLF: low-level fusion; MLF: mid-level fusion; HLF: high-level fusion; PLS-DA: partial least squares discrimination analysis; SVM: support vectors machine; SIMCA: soft independent modeling class analogy; ANN: artificial neural network; WC: wavelet compression; iPLS: interval partial least squares; MSC: multiplicative scatter correction; S-G: Savitzky-Golay; FD: first derivative; SD: second derivative; NIR: near-infrared; MIR: mid-infrared; ATR-FTIR: attenuated total reflectance-Fourier transform infrared spectrometry; UV-Vis: ultraviolet-visible; PQN: probabilistic quotient normalization; HPLC: high-performance liquid chromatography; GC: gas chromatography; CDA: canonical discriminant function analysis; NIR-HSI: near infrared-hyperspectral imaging; ROI: region of interest; SNV: standard normal variate; GC-MS: gas chromatography-mass spectrometry; OPLS-DA: orthogonal projections to latent structures discriminant analysis; FT-NIR: Fourier transform near infrared; VIP: variable importance in projection; HPTLC: high-performance thin-layer chromatography; HS-GC-MS: headspace gas chromatography-mass spectrometry; LDA: linear discriminant analysis; BPNN: back propagation neural network; *k*-NN: *k*-nearest neighbors; E-nose: electronic nose; HCA: hierarchical cluster analysis; ResNet: residual convolutional neural network; SDD: spectrum standard deviation; COW: correlation optimized warping; GLCM: grey-level co-occurrence matrix; UHPLC-Q-TOF-MS/MS: ultra-high performance liquid chromatography-quadrupole time-of-flight tandem mass spectrometry; SWIR: short wave infrared hyperspectral imaging.

geographical traceability and adulteration identification. Other factors, such as medicinal parts and growth years, are also crucial in assessing the quality and commercial value of medicinal plants. Therefore, the combination of multi-source data and machine learning in these areas holds significant practical importance. The applications of multi-source data combined with machine learning in other fields are summarized in Table 6 [59,172–181].

For instance, the content of polysaccharides in *Polygonatum kingianum*, a key active ingredient, is subject to fluctuations influenced by growth years, resulting in variations in quality. Zhang et al. [175] investigated the efficacy of a mid-level data fusion-based PLS-DA model in identifying *Polygonatum kingianum* samples with different growth years. ATR-FTIR and UV-Vis data were fused after pre-processing and feature variable extraction. The data fusion approach yielded superior modeling results (100% accuracy) compared to using a single data source. This study provides a valuable reference method for evaluating the quality of medicinal plants. Furthermore, a data fusion strategy combined with machine learning was employed to discriminate Vietnamese ginseng samples based on their parts and growth years [181]. By integrating ATR-FTIR and UPLC-QTOF/MS data through low- and mid-level data fusion, orthogonal projections to latent structures discriminant analysis (OPLS-DA) and SVM models were established to analyze the multi-source data. The two classification models demonstrated different

Table 5
Summary of references for content prediction using machine learning combined with multi-source data.

Year	Object	Multi-source	Strategy	Data processing	Machine learning	Result	Refs.
2017	*Crocus sativus* L.	NIR, HPLC	NM	Vector normalization, MSC, FD, SD, SLS, min-max normalization	PLSR	crocin I: RMSECV = 1.40; crocin II: RMSECV = 0.30	[152]
2020	*Pueraria lobata* (Willd.) Ohwi	NIR, UV, HPLC	LLF, MLF	Min-max normalization, iPLS	PLSR	LLF: RMSEP = 0.418, RPD = 4.295	[153]
2020	*Arctostaphylos uva-ursi* (L.) Spreng. leaves	UV-Vis, UHPLC/Q-TOF-MS	LLF, MLF	SNV, MSC, S-G, smoothing + FD, PCA	PLSR	MLF: RPD (TPC) = 6.258, RPD (DPPH) = 6.699	[154]
2020	*Morinda officinalis* F. C. How	NIR, UHPLC	NM	SNV, MSC, FD, SD, S-G smoothing	PCA, PLSR	RPD (SNV) = 9.30	[155]
2020	*Magnolia biondii* Pamp.	NIR, HPLC, GC-MS	NM	MSC, SNV, FD, SD, S-G smoothing	PLSR	RPD > 5.80, $R^2 > 0.90$	[156]
2021	Bran-fried *Atractylodes lancea* (Thunb) DC.	NIR, intelligent color recognition, HPLC	NM	Vector normalization, MSC, FD, SD	PLSR	$R^2 = 0.9717$, RMSE = 0.026	[157]
2019	*Rheum palmatum* L.	NIR, HPLC	NM	NM	PLSR, PSO-LSSVM	PSO-LSSVM: RSEP (free anthraquinone) = 10.66%, RSEP (total anthraquinone) = 4.95%	[158]
2021	*Lonicera japonica* Thunb.	NIR, HPLC	NM	FD, SD, MSC, SLS, vector normalization, min-max normalization, S-G smoothing, constant offset elimination and different combinations	PLSR, ANN	RMSEP > 1.20, $R > 0.98$	[159]
2018	*Potentilla erecta* subsp. *strictissima* (Zimmeter) A. J. Richards	ATR and DRIFTS, Raman, HPLC	NM	S-G filter	PLSR	RSEP (Raman): 2.0–4.9%; RSEP (NIR): 2.7–6.5%	[160]
2016	Essential oils of *Lavandula angustifolia* Mill., *Lavandula latifolia* Medik.	GC-MS, GC-FID, Raman, MIR	NM	NM	PLSR, PCA	$R^2 \geq 0.87$, REP ≥ 2.14	[161]
2022	*Ginkgo biloba* L. leaves	NIR, HPLC	NM	FD, SD, SNV, MSC, OSC, S-G smoothing, autoscaling, mean-centering, Par and different combinations, iPLS, SiPLS, CARS, SPA	PLSR, SVR	PLSR: $R^2 > 0.95$, RESECV <0.30; SVR: $R^2 > 0.96$, RESECV <0.50	[162]
2019	*Ziziphus jujuba* Mill.	NIR, HPLC	NM	MSC, derivative, smoothing, normalization and their combinations	PLSR, SVR	SVR: R^2c = 0.9192, RMSEC = 23.6116, R^2p = 0.7859, RMSEP = 33.4818	[163]
2017	*Gardenia jasminoides* J. Ellis	MIR, NIR, HPLC	NM	NM	PLSR	$R^2 > 0.95$, RSEP <7%	[164]
2023	*Abelmoschus esculentus* L.	NIR, HPLC	NM	SNV, MSC	PLSR, PCR, SMLR	R^2: 0.818–0.931, RPD: 2.036–2.702	[165]
2022	Stir-fried *Paeonia suffruticosa* Andr.	NIR, HPLC	NM	S-G smoothing, MSC, FD, SD, SNV, CARS, VCPA, IRIV, GA	PLSR, SVR	PLSR: R^2c > 0.82, RMSEC < 1.60, R^2p > 0.82, RMSEP < 1.70, RPD > 2.40	[166]
2018	*Epimedium brevicornu*Maxim.	NIR, HPLC	NM	CARS	PLSR	R^2c = 0.9314, RMSEC = 0.0408, R^2p = 0.9269, RMSEP = 0.0480	[167]
2021	*Plantago asiatica* L.	NIR, HPLC, UV-Vis	NM	SNV, MSC, S-G smoothing, derivative and their combinations; GA, PSO, CARS	PLSR	$R^2 > 0.80$, RPD > 2	[168]
2016	*Chrysanthemum morifolium* Ramat.	NIR, HPLC/Q-TOF-MS	NM	MSC, SNV, DT, S-G smoothing, FD, SD and their combinations; SiPLS	BPNN, RF, SVR	BPNN: R = 0.89	[169]
2018	*Coptis chinensis* Franch., *Coptis deltoidea* C. Y. Cheng & P. K. Hsiao, *Coptis omeiensis* (Chen) C. Y. Cheng, *Coptis teeta* Wall.	HPLC, FT-IR, FT-NIR	LLF, MLF	PCA, VIP, FD, SD, MSC, SNV, smoothing (11 points)	PLSR	MLF-VIP: R^2c = 0.916, RMSEC = 6.144, R^2p = 0.942, RMSEP = 4.877, RPD = 3.768	[170]
2020	*Wolfiporia cocos* (F.A. Wolf) Ryvarden & Gilb.	FT-IR, UV, HPLC	MLF	SNV, SD, FD, iPLS, VIP	PLSR	UV1-VIP: RPD = 3.4, RMSECV = 0.04, RMSEP = 0.03	[171]

NM: no mention; LLF: low-level fusion; MLF: mid-level fusion; NIR: near-infrared; HPLC: high-performance liquid chromatography; PLSR: partial least squares regression; FD: first derivative; SD: second derivative; MSC: multiplicative scatter correction; RMSECV: root mean square error of cross-validation; UHPLC: ultra high-performance liquid chromatography; PCA: principal component analysis; SNV: standard normal variate; RPD: residual predictive deviation; HPLC: high-performance liquid chromatography; GC-MS: gas chromatography-mass spectrometry; S-G: Savitzky-Golay; PSO-LSSVM: particle swarm optimization based least square support vector machines; RSEP: relative standard error of prediction; ANN: artificial neural networks; SLS: subtract straight line; DRIFTS: diffuse reflectance mid- and near-infrared spectra; GC-FID: gas chromatography-flame ionisation detector; SVR: support vector regression; OSC: orthogonal signal correction; Par: Pareto scaling; iPLS: interval partial least square; SiPLS: synergy interval partial least square; CARS: competitive adaptive reweight sampling; SPA: successive projections algorithm; PCR: principal component regression; SMLR: stepwise multiple linear regression; RPD: residual predictive deviation; VCPA: variable combination population analysis; IRIV: iteratively retaining informative variables; GA: genetic algorithm; HPLC/Q-TOF-MS: performance liquid chromatography-quadrupole-time of flight-mass spectrometry; BPNN: back propagation neural network; DT: de-trend; FT-NIR: Fourier transform near infrared; FT-IR: Fourier transform infrared; UV: ultraviolet; RMSEP: root mean square error of prediction.

Table 6
Summary of references for other applications using machine learning combined with multi-source data.

Year	Object	Multi-source	Strategy	Data processing	Machine learning	Result	Refs.
2020	*Curcuma phaeocaulis* Val., *Curcuma kwangsiensis* S. G. Lee et C. F. Liang and *Curcuma wenyujin* Y. H. Chen et C. Ling	FT-NIR, E-nose, colorimeter, HPLC	MLF	FD, MSC, SNV, min-max normalization, GA, IRIV, CARS	PCA, PLSR, PLS-DA	MLF: Acc = 100%	[59]
2011	Different parts of *Panax ginseng* C.A. Meyer	ATR-FTIR, DR-NIR	NM	NM	PCA	NM	[172]
2020	Wild and cultivated *Wolfiporia cocos* (F.A. Wolf) Ryvarden & Gilb.	UHPLC, ATR-FTIR	NM	S-G filter, VIP	PLSR, PLS-DA	Acc >95.14%, RPD = 2.494	[173]
2022	Different parts and harvest time of *Dendrobium officinale* Kimura & Migo	ATR-FTIR, FT-NIR	LLF	SD, MSC, SNV	PLS-DA, SVM, PCA, *t*-SNE	Different parts (PLS-DA): Acc (ATR-FTIR) = 1.00, harvest time (PLS-DA): Acc (ATR-FTIR) = 1.00	[174]
2021	different growth ages of *Polygonatum kingianum* Collett & Hemsl.	ATR-FTIR, UV-Vis	MLF	PCA, FD, SD, S-G	PLS-DA, PCA, HCA	MLF: Acc (PLS-DA) = 100%	[175]
2023	Heavy metals in *Lilium brownii* var. *viridulum* Baker	NIR, LIBS, ICP-MS	LLF, MLF	SNV, MSC, FD, VIP, CARS, LASSO	PLSR, LSSVR	MLF-PLSR: R^2 (Zn) = 0.9858, R^2 (Cu) = 0.9811, R^2 (Pb) = 0.9460, RMSEP (Zn) = 4.3047 mg/kg, RMSEP (Cu) = 4.9592 mg/kg, RMSEP (Pb) = 8.3881 mg/kg	[176]
2023	*Polygonum multiflorum* Thunb.	E-eye, HPLC	LLF	VIP	OPLS-DA, PLSR	LLF: R^2 (OPLS-DA) = 0.753, Q^2 (OPLS-DA) = 0.490; Q^2 (PLSR) = 0.872, R^2 (PLSR) = 0.67	[177]
2021	*Abelmoschus esculentus* L.	Vis-NIR hyperspectral imaging (texture features, effective wavelengths)	MLF	SPA, GLCM	LIBSVM, MLR	MLF: Acc (LIBSVM) = 91.7%, RMSECV (MLR) = 1.348%, R^2 (MLR) = 0.816, RPD (MLR) = 2.33	[178]
2022	Cultivation methods and growth years of *Dendrobium huoshanense* C. Z. Tang & S. J. Cheng	Nano-effect near and mid infrared spectra	MLF	VIP	PLS-DA, OPLS-DA, PLSR	MLF: Acc (OPLS-DA) = 100%, RMSEC (PLSR) = 0.1478, R^2c (PLSR) = 0.913, RMSEP (PLSR) = 0.1951, R^2p (PLSR) = 0.984	[179]
2021	Wild and cultivated *Wolfiporia cocos* (F.A. Wolf) Ryvarden & Gilb.	HPLC, ATR-FTIR	LLF, MLF, HLF	Boruta, PCA, COW, S-G	PLS-DA, RF	MLF: Acc (Boruta-PLS-DA) = 97.50%	[180]
2021	Age and parts of *Panax vietnamensis* Ha & Grushv.	ATR-FTIR, UPLC-QTOF/MS	LLF, MLF	RFE, CARS, FD, SD, MSC, S-G smoothing and their combinations	OPLS-DA, SVM	Different growth years (OPLS-DA): Acc (LLF) = 100%; different parts (RFE-SVM): Acc (MLF) = 83.33%	[181]

NM: no mention; LLF: low-level fusion; MLF: mid-level fusion; DR-NIR: diffuse reflectance near-infrared spectroscopy; ATR-FTIR: attenuated total reflectance-Fourier transform infrared spectrometry; PCA: principal component analysis; UHPLC: ultra high-performance liquid chromatography; PLSR: partial least squares regression; PLS-DA: partial least squares discrimination analysis; S-G: Savitzky-Golay; VIP: variable importance in projection; *t*-SNE: *t*-distribute stochastic neighbor embedding; SVM: support vectors machine; SD: second derivative; MSC: multiplicative scatter correction; SNV: standard normal variate; UV-Vis: ultraviolet-visible; HCA: hierarchical cluster analysis; FD: first derivative; LIBS: laser-induced breakdown spectroscopy; ICP-MS: inductively coupled plasma-mass spectrometry; LSSVR: least squares support vector regression; CARS: competitive adaptive reweighted sampling; LASSO: least absolute shrinkage and selection operator; RMSEP: root mean square error of prediction; E-eye: electronic eye; HPLC: high-performance liquid chromatography; OPLS-DA: orthogonal projections to latent structures discriminant analysis; E-nose: electronic nose; IRIV: iteratively retaining informative variables; GA: genetic algorithm; SPA: successive projections algorithm; GLCM: grey-level co-occurrence matrix; LIBSVM: library for support vector machines; MLR: multiple linear regression; RF: random forest; COW: correlation-optimized warping algorithm; UPLC-QTOF/MS: ultra-performance liquid chromatography quadrupole time-of-flight mass spectrometry; SVM: support vectors machine; RFE: recursive feature elimination; Acc: accuracy.

suitability: combining OPLS-DA with low-level data achieved 100% accuracy in identifying different parts, while the SVM model based on mid-level data fusion was more suitable for classifying different growth years, achieving an accuracy rate of 83.33%. The study also compared different feature extraction methods, finding that variables extracted using REF were more effective in improving model performance compared to those extracted using CARS. In previous studies, differences in the content of secondary metabolites between wild and cultivated medicinal plants, have been observed due to varying growth environments. Wang et al. [173] conducted the qualitative and quantitative analysis using ATR-FTIR and UHPLC to explore the disparities between wild and cultivated *Macrohyporia cocos*. The combination of spectroscopy and PLS-DA proved to be a rapid identification technique for distinguishing between wild and cultivated samples, with an accuracy rate exceeding 95.14%. Additionally, PLSR was employed to correlate ATR-FTIR and UHPLC data, confirming that spectroscopy and chromatography provide complementary sources for effective content prediction.

In conclusion, although the application of machine learning combined with multi-source data in these specific areas is limited, it still offers valuable insights for research in related fields. The quality evaluation of medicinal plants is a complex process, and focusing solely on aspects, such as geographical origin or adulteration is insufficient for comprehensive quality certification. Developing a comprehensive quality evaluation platform, including the integration of blockchain technique, holds great practical significance for assessing the quality of medicinal plants.

6. Conclusion and prospect

Medicinal plants possess their therapeutic effects through the combined action of various compounds. However, ensuring their high-quality is crucial to maximize their clinical efficacy. Traditional analytical methods may not fully elucidate the mechanism of action or identify all medicinal components. In contrast, utilizing multi-source data provides a more comprehensive understanding. This review examines the application of machine learning algorithms, which have demonstrated effectiveness in analyzing and processing the multi-source data of medicinal plants. When combined with data fusion strategies, these algorithms can integrate diverse data sources and enhance the accuracy of classification and prediction models. Numerous studies have demonstrated that machine learning algorithms, in conjunction with multi-source data, enable comprehensive quality evaluation of medicinal plants. Their performance surpasses that of single analytical methods; thus, positively impacting the clinical application of traditional Chinese medicine and providing a theoretical basis for market supervision of medicinal plants.

Although significant progress has been made in combining machine learning with multi-source data for evaluating the quality of medicinal plants, there are still areas that require improvement. Firstly, there should be a focus on acquiring complementary information through pathways that generate synergistic data, rather than arbitrary combinations. This aspect has received limited attention in current research and warrants further consideration. Secondly, the selection of data pre-processing methods often relies on traditional trial and error approaches, which are time-consuming and may overlook valuable complementary information. There is a need for clear guidelines and regulations regarding the selection of pre-processing methods. Ensemble pre-processing methods show promise but still have room for improvement. Selecting and developing appropriate pre-processing methods will be a major challenge. Thirdly, data fusion offers significant advantages in processing high-dimensional data of medicinal plants and can enhance inference accuracy by integrating multi-source information. Among the three data fusion strategies, mid-level data fusion is the most commonly used, closely followed by low-level data fusion. While many studies have shown that mid-level data fusion performs well, blind adherence to this strategy should be avoided. Choosing the most suitable approach based on specific objectives and data properties is more conducive to obtaining scientifically sound results. Feature extraction plays a critical role in data fusion to prevent overfitting and dimensionality issues effectively. The reasonable selection of feature extraction or variable selection methods should be the focal point of data analysis. Lastly, there is no universal machine learning algorithm that can be applied to all datasets. The selection of machine learning algorithms should be based on the specific problems to be solved and the characteristics of the dataset. Machine learning is a powerful tool for analyzing multi-source data of medicinal plants, but the constant evolution of analysis technology necessitates continuous the innovation in machine learning algorithms. The four areas mentioned above (multi-source data, pre-processing and feature extraction, data fusion, and machine learning) require further research to overcome existing challenges. This review provides constructive suggestions for improving the quality evaluation methods of medicinal plants and enhancing the applicability of market monitoring.

CRediT author statement

Yanying Zhang: Writing - Original draft preparation, Software; Yuanzhong Wang: Resources, Funding acquisition, Supervision.

Declaration of competing interest

The authors declare that there are no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No.: U2202213), and the Special Program for the Major Science and Technology Projects of Yunnan Province, China (Grant Nos.: 202102AE090051-1-01, and 202202AE090001).

References

[1] S. Chen, H. Yu, H. Luo, et al., Conservation and sustainable use of medicinal plants: Problems, progress, and prospects, Chin. Med. 11 (2016), 37.

[2] J. He, B. Yang, M. Dong, et al., Crossing the roof of the world: Trade in medicinal plants from Nepal to China, J. Ethnopharmacol. 224 (2018) 100–110.

[3] T. Shen, H. Yu, Y. Wang, Assessing the impacts of climate change and habitat suitability on the distribution and quality of medicinal plant using multiple information integration: Take Gentiana rigescens as an example, Ecol. Indic. 123 (2021), 107376.

[4] A.C. Hamilton, Medicinal plants, conservation and livelihoods, Biodivers. Conserv. 13 (2004) 1477–1517.

[5] W.L. Applequist, J.A. Brinckmann, A.B. Cunningham, et al., Scientists' warning on climate change and medicinal plants, Planta Med. 86 (2020) 10–18.

[6] M. Yang, Z. Li, L. Liu, et al., Ecological niche modeling of *Astragalus membranaceus* var. *mongholicus* medicinal plants in Inner Mongolia, China, Sci. Rep. 10 (2020), 12482.

[7] W. Kong, J. Wang, Q. Zang, et al., Fingerprint-efficacy study of artificial *Calculus bovis* in quality control of Chinese materia medica, Food Chem. 127 (2011) 1342–1347.

[8] M. He, Y. Zhou, How to identify "Material basis-Quality markers" more accurately in Chinese herbal medicines from modern chromatography-mass spectrometry data-sets: Opportunities and challenges of chemometric tools, Chin. Herb. Med. 13 (2020) 2–16.

[9] C. Assis, H.V. Pereira, V.S. Amador, et al., Combining mid infrared spectroscopy and paper spray mass spectrometry in a data fusion model to predict the composition of coffee blends, Food Chem. 281 (2019) 71–77.

[10] A. Sanaeifar, X. Li, Y. He, et al., A data fusion approach on confocal Raman microspectroscopy and electronic nose for quantitative evaluation of

pesticide residue in tea, Biosyst. Eng. 210 (2021) 206–222.
[11] S. Azcarate, R. Ríos-Reina, J. Amigo, et al., Data handling in data fusion: Methodologies and applications, Trends Anal. Chem. 143 (2021), 116355.
[12] P. Zhang, T. Li, Z. Yuan, et al., A data-level fusion model for unsupervised attribute selection in multi-source homogeneous data, Inf. Fusion 80 (2022) 87–103.
[13] E. Borràs, J. Ferré, R. Boqué, et al., Data fusion methodologies for food and beverage authentication and quality assessment - a review, Anal. Chim. Acta 891 (2015) 1–14.
[14] X. Zhou, X. Li, B. Zhao, et al., Discriminant analysis of vegetable oils by thermogravimetric-gas chromatography/mass spectrometry combined with data fusion and chemometrics without sample pretreatment, LWT 161 (2022), 113403.
[15] H. Wang, P. Chen, J. Dai, et al., Recent advances of chemometric calibration methods in modern spectroscopy: Algorithms, strategy, and related issues, Trac Trends Anal. Chem. 153 (2022), 116648.
[16] N. Taoufik, W. Boumya, M. Achak, et al., The state of art on the prediction of efficiency and modeling of the processes of pollutants removal based on machine learning, Sci. Total Environ. 807 (2022), 150554.
[17] D.V. Nazarenko, P.V. Kharyuk, I.V. Oseledets, et al., Machine learning for LC–MS medicinal plants identification, Chemom. Intell. Lab. Syst. 156 (2016) 174–180.
[18] T. Meng, X. Jing, Z. Yan, et al., A survey on machine learning for data fusion, Inf. Fusion 57 (2020) 115–129.
[19] I. Magnus, M. Virte, H. Thienpont, et al., Combining optical spectroscopy and machine learning to improve food classification, Food Contr. 130 (2021), 108342.
[20] Q. Li, Y. Huang, J. Zhang, et al., A fast determination of insecticide deltamethrin by spectral data fusion of UV-vis and NIR based on extreme learning machine, Spectrochim. Acta Part A Mol. Biomol. Spectrosc. 247 (2021), 119119.
[21] L. Zhou, C. Zhang, Z. Qiu, et al., Information fusion of emerging non-destructive analytical techniques for food quality authentication: A survey, Trac Trends Anal. Chem. 127 (2020), 115901.
[22] N. Hussain, D. Sun, H. Pu, Classical and emerging non-destructive technologies for safety and quality evaluation of cereals: A review of recent applications, Trends Food Sci. Technol. 91 (2019) 598–608.
[23] D. Cozzolino, Foodomics and infrared spectroscopy: From compounds to functionality, Curr. Opin. Food Sci. 4 (2015) 39–43.
[24] L. Yin, J. Zhou, D. Chen, et al., A review of the application of near-infrared spectroscopy to rare traditional Chinese medicine, Spectrochim. Acta Part A Mol. Biomol. Spectrosc. 221 (2019), 117208.
[25] N. Modupalli, M. Naik, C.K. Sunil, et al., Emerging non-destructive methods for quality and safety monitoring of spices, Trends Food Sci. Technol. 108 (2021) 133–147.
[26] J. Wang, Q. Chen, T. Belwal, et al., Insights into chemometric algorithms for quality attributes and hazards detection in foodstuffs using Raman/surface enhanced Raman spectroscopy, Compr. Rev. Food Sci. Food Saf. 20 (2021) 2476–2507.
[27] M.M. Oliveira, J.P. Cruz-Tirado, D.F. Barbin, Nontargeted analytical methods as a powerful tool for the authentication of spices and herbs: A review, Compr. Rev. Food Sci. Food Saf. 18 (2019) 670–689.
[28] G. Wan, S. Fan, G. Liu, et al., Fusion of spectra and texture data of hyperspectral imaging for prediction of myoglobin content in nitrite-cured mutton, Food Contr. 144 (2023), 109332.
[29] K. Feng, S. Wang, L. Han, et al., Configuration of the ion exchange chromatography, hydrophilic interaction chromatography, and reversed-phase chromatography as off-line three-dimensional chromatography coupled with high-resolution quadrupole-Orbitrap mass spectrometry for the multicomponent characterization of *Uncaria sessilifructus*, J. Chromatogr. A 1649 (2021), 462237.
[30] A.M. Mustafa, S. Angeloni, D. Abouelenein, et al., A new HPLC-MS/MS method for the simultaneous determination of 36 polyphenols in blueberry, strawberry and their commercial products and determination of antioxidant activity, Food Chem. 367 (2022), 130743.
[31] Z. Liu, M.Q. Yang, Y. Zuo, et al., Fraud detection of herbal medicines based on modern analytical technologies combine with chemometrics approach: A review, Crit. Rev. Anal. Chem. 52 (2022) 1606–1623.
[32] J. Schripsema, S.M. da Silva, D. Dagnino, Differential NMR and chromatography for the detection and analysis of adulteration of vetiver essential oils, Talanta 237 (2022), 122928.
[33] P.H. Stefanuto, A. Smolinska, J.F. Focant, Advanced chemometric and data handling tools for GC×GC-TOF-MS, Trac Trends Anal. Chem. 139 (2021), 116251.
[34] Y. Picó, Chromatography–mass spectrometry: Recent evolution and current trends in environmental science, Curr. Opin. Environ. Sci. Health 18 (2020) 47–53.
[35] H.A. Gad, S.H. El-Ahmady, M.I. Abou-Shoer, et al., Application of chemometrics in authentication of herbal medicines: A review, Phytochem. Anal. 24 (2013) 1–24.
[36] C. Suárez-Oubiña, P. Herbello-Hermelo, P. Bermejo-Barrera, et al., Exploiting dynamic reaction cell technology for removal of spectral interferences in the assessment of Ag, Cu, Ti, and Zn by inductively coupled plasma mass spectrometry, Spectrochim. Acta B 187 (2022), 106330.
[37] Y. Huang, A. Bais, A novel PCA-based calibration algorithm for classification of challenging laser-induced breakdown spectroscopy soil sample data, Spectrochim. Acta B 193 (2022), 106451.
[38] D. Stefas, N. Gyftokostas, P. Kourelias, et al., Honey discrimination based on the bee feeding by Laser Induced Breakdown Spectroscopy, Food Contr. 134 (2022), 108770.
[39] S. Müller, J.A. Meima, Mineral classification of lithium-bearing pegmatites based on laser-induced breakdown spectroscopy: Application of semi-supervised learning to detect known minerals and unknown material, Spectrochim. Acta B 189 (2022), 106370.
[40] Q. Chen, C. Sun, Q. Ouyang, et al., Classification of different varieties of Oolong tea using novel artificial sensing tools and data fusion, LWT Food Sci. Technol. 60 (2015) 781–787.
[41] W. Zheng, Y. Shi, Y. Ying, et al., Olfactory-taste synesthesia model: An integrated method for flavor responses of electronic nose and electronic tongue, Sens. Actuat. A 350 (2023), 114134.
[42] Y. Xu, J. Zhang, Y. Wang, Recent trends of multi-source and non-destructive information for quality authentication of herbs and spices, Food Chem. 398 (2023), 133939.
[43] G. Wei, M. Dan, G. Zhao, et al., Recent advances in chromatography-mass spectrometry and electronic nose technology in food flavor analysis and detection, Food Chem. 405 (2023), 134814.
[44] W. Wojnowski, T. Majchrzak, T. Dymerski, et al., Electronic noses: Powerful tools in meat quality assessment, Meat Sci. 131 (2017) 119–131.
[45] P. Vahdatiyekta, M. Zniber, J. Bobacka, et al., A review on conjugated polymer-based electronic tongues, Anal. Chim. Acta 1221 (2022), 340114.
[46] J.X. Leon-Medina, M. Anaya, D.A. Tibaduiza, Yogurt classification using an electronic tongue system and machine learning techniques, Intell. Syst. Appl. 16 (2022), 200143.
[47] T. Wasilewski, D. Migoń, J. Gębicki, et al., Critical review of electronic nose and tongue instruments prospects in pharmaceutical analysis, Anal. Chim. Acta 1077 (2019) 14–29.
[48] M. Modzelewska-Kapituła, S. Jun, The application of computer vision systems in meat science and industry - A review, Meat Sci. 192 (2022), 108904.
[49] A. Taheri-Garavand, S. Fatahi, M. Omid, et al., Meat quality evaluation based on computer vision technique: A review, Meat Sci. 156 (2019) 183–195.
[50] S. Ma, Y. Li, Y. Peng, Spectroscopy and computer vision techniques for noninvasive analysis of legumes: A review, Comput. Electron. Agric. 206 (2023), 107695.
[51] H. Yang, J. Zhang, Y. Wang, et al., Content determination of total saponins in different parts of plant *Paris polyphylla* var. *chinensis*, Adv. Mater. Res. 926-930 (2014) 969–974.
[52] H. Yang, J. Liu, S. Chen, et al., Spatial variation profiling of four phytochemical constituents in *Gentiana straminea* (Gentianaceae), J. Nat. Med. 68 (2014) 38–45.
[53] Y. Guo, X. Li, Z. Zhao, et al., Predicting the impacts of climate change, soils and vegetation types on the geographic distribution of *Polyporus umbellatus* in China, Sci. Total Environ. 648 (2019) 1–11.
[54] T. Shen, H. Yu, Y. Wang, Assessing geographical origin of *Gentiana rigescens* using untargeted chromatographic fingerprint, data fusion and chemometrics, Molecules 24 (2019), 2562.
[55] X. Liu, Z. Zhang, Y. Liang, et al., Baseline correction of high resolution spectral profile data based on exponential smoothing, Chemom. Intell. Lab. Syst. 139 (2014) 97–108.
[56] Z. Zhang, S. Chen, Y. Liang, Baseline correction using adaptive iteratively reweighted penalized least squares, Analyst 135 (2010) 1138–1146.
[57] X. Xu, X. Huo, X. Qian, et al., Data-driven and coarse-to-fine baseline correction for signals of analytical instruments, Anal. Chim. Acta 1157 (2021), 338386.
[58] Y. Dai, Z. Dai, G. Guo, et al., Nondestructive identification of rice varieties by the data fusion of Raman and near-infrared (NIR) spectroscopies, Anal. Lett. 56 (2023) 730–743.
[59] Z. Lan, Y. Zhang, Y. Sun, et al., A mid-level data fusion approach for evaluating the internal and external changes determined by FT-NIR, electronic nose and colorimeter in Curcumae Rhizoma processing, J. Pharm. Biomed. Anal. 188 (2020), 113387.
[60] S. Wu, L. Wang, G. Zhou, et al., Strategies for the content determination of capsaicin and the identification of adulterated pepper powder using a handheld near-infrared spectrometer, Food Res. Int. 163 (2023), 112192.
[61] P. Mishra, A. Biancolillo, J.M. Roger, et al., New data preprocessing trends based on ensemble of multiple preprocessing techniques, Trac Trends Anal. Chem. 132 (2020), 116045.
[62] J. Gerretzen, E. Szymańska, J.J. Jansen, et al., Simple and effective way for data preprocessing selection based on design of experiments, Anal. Chem. 87 (2015) 12096–12103.
[63] P. Mishra, J.M. Roger, D.N. Rutledge, et al., SPORT pre-processing can improve near-infrared quality prediction models for fresh fruits and agro-materials, Postharvest Biol. Technol. 168 (2020), 111271.
[64] C. Pérez-Ràfols, N. Serrano, J.M. Díaz-Cruz, Authentication of soothing herbs by UV–vis spectroscopic and chromatographic data fusion strategy, Chemom. Intell. Lab. Syst. 235 (2023), 104783.
[65] J.A. Navarro-Huerta, J.R. Torres- Lapasió, S. López-Ureña, et al., Assisted baseline subtraction in complex chromatograms using the BEADS algorithm, J. Chromatogr. A 1507 (2017) 1–10.
[66] T. Skov, F. van den Berg, G. Tomasi, et al., Automated alignment of chromatographic data, J. Chemome. 20 (2006) 484–497.
[67] L. Zhu, P. Spachos, E. Pensini, et al., Deep learning and machine vision for food processing: A survey, Curr. Res. Food Sci. 4 (2021) 233–249.

[68] D.I. Patrício, R. Rieder, Computer vision and artificial intelligence in precision agriculture for grain crops: A systematic review, Comput. Electron. Agric. 153 (2018) 69–81.

[69] S. Lorenz, P. Seidel, P. Ghamisi, et al., Multi-sensor spectral imaging of geological samples: A data fusion approach using spatio-spectral feature extraction, Sensors 19 (2019), 2787.

[70] S. Jo, W. Sohng, H. Lee, et al., Evaluation of an autoencoder as a feature extraction tool for near-infrared spectroscopic discriminant analysis, Food Chem. 331 (2020), 127332.

[71] K. Cai, H. Chen, W. Ai, et al., Feedback convolutional network for intelligent data fusion based on near-infrared collaborative IoT technology, IEEE Trans. Ind. Inform. 18 (2022) 1200–1209.

[72] Y. Sun, Y. Wang, H. Xiao, et al., Hyperspectral imaging detection of decayed honey peaches based on their chlorophyll content, Food Chem. 235 (2017) 194–202.

[73] Q. Zhou, Z. Dai, F. Song, et al., Monitoring black tea fermentation quality by intelligent sensors: Comparison of image, e-nose and data fusion, Food Biosci. 52 (2023), 102454.

[74] R. Ríos-Reina, R.M. Callejón, F. Savorani, et al., Data fusion approaches in spectroscopic characterization and classification of PDO wine vinegars, Talanta 198 (2019) 560–572.

[75] J. Cheng, J. Sun, K. Yao, et al., A decision fusion method based on hyperspectral imaging and electronic nose techniques for moisture content prediction in frozen-thawed pork, LWT 165 (2022), 113778.

[76] H. Yu, L. Qing, D. Yan, et al., Hyperspectral imaging in combination with data fusion for rapid evaluation of tilapia fillet freshness, Food Chem. 348 (2021), 129129.

[77] Y. Li, Y. Huang, J. Xia, et al., Quantitative analysis of honey adulteration by spectrum analysis combined with several high-level data fusion strategies, Vib. Spectrosc. 108 (2020), 103060.

[78] M.P. Callao, I. Ruisánchez, An overview of multivariate qualitative methods for food fraud detection, Food Contr. 86 (2018) 283–293.

[79] M. Boubchir, R. Boubchir, H. Aourag, The Principal Component Analysis as a tool for predicting the mechanical properties of Perovskites and Inverse Perovskites, Chem. Phys. Lett. 798 (2022), 139615.

[80] Y. Xu, Z. Wu, Parameter identification of unsaturated seepage model of core rockfill dams using principal component analysis and multi-objective optimization, Structures 45 (2022) 145–162.

[81] Q. Liu, Z. Gui, S. Xiong, et al., A principal component analysis dominance mechanism based many-objective scheduling optimization, Appl. Soft Comput. 113 (2021), 107931.

[82] J. Yang, E. Grunsky, Q. Cheng, A novel hierarchical clustering analysis method based on Kullback-Leibler divergence and application on dalaimiao geochemical exploration data, Comput. Geosci. 123 (2019) 10–19.

[83] C. Liu, Z. Zuo, F. Xu, et al., Authentication of herbal medicines based on modern analytical technology combined with chemometrics approach: A review, Crit. Rev. Anal. Chem. (2022) 1–26.

[84] D. Granato, P. Putnik, D.B. Kovačević, et al., Trends in chemometrics: Food authentication, microbiology, and effects of processing, Compr. Rev. Food Sci. Food Saf. 17 (2018) 663–677.

[85] A.S. Wilde, S.A. Haughey, P. Galvin-King, et al., The feasibility of applying NIR and FT-IR fingerprinting to detect adulteration in black pepper, Food Contr. 100 (2019) 1–7.

[86] M.N. Mohamad Asri, R. Verma, N.A. Mahat, et al., Discrimination and source correspondence of black gel inks using Raman spectroscopy and chemometric analysis with UMAP and PLS-DA, Chemom. Intell. Lab. Syst. 225 (2022), 104557.

[87] W.F. Lamberti, Blood cell classification using interpretable shape features: A comparative study of SVM models and CNN-Based approaches, Comput. Meth. Programs Biomed. Update 1 (2021), 100023.

[88] D. Duca, M. Mancini, G. Rossini, et al., Soft Independent Modelling of Class Analogy applied to infrared spectroscopy for rapid discrimination between hardwood and softwood, Energy 117 (2016) 251–258.

[89] R. Brendel, S. Schwolow, N. Gerhardt, et al., MIR spectroscopy versus MALDI-ToF-MS for authenticity control of honeys from different botanical origins based on soft independent modelling by class analogy (SIMCA) - A clash of techniques? Spectrochim. Acta Part A Mol. Biomol. Spectrosc. 263 (2021), 120225.

[90] Y. Li, Y. Shen, C. Yao, et al., Quality assessment of herbal medicines based on chemical fingerprints combined with chemometrics approach: A review, J. Pharm. Biomed. Anal. 185 (2020), 113215.

[91] P. Mishra, J.M. Roger, D. Jouan-Rimbaud-Bouveresse, et al., Recent trends in multi-block data analysis in chemometrics for multi-source data integration, Trac Trends Anal. Chem. 137 (2021), 116206.

[92] H. Liu, H. Liu, J. Li, et al., Review of recent modern analytical technology combined with chemometrics approach researches on mushroom discrimination and evaluation, Crit. Rev. Anal. Chem. (2022) 1–24.

[93] L. Zhu, X. Zhou, W. Liu, et al., Total organic carbon content logging prediction based on machine learning: A brief review, Energy Geosci. 4 (2023), 100098.

[94] Z. Ma, G. Mei, Deep learning for geological hazards analysis: Data, models, applications, and opportunities, Earth Sci. Rev. 223 (2021), 103858.

[95] B. Debus, H. Parastar, P. Harrington, et al., Deep learning in analytical chemistry, Trac Trends Anal. Chem. 145 (2021), 116459.

[96] Z. Yuan, M. Niu, H. Ma, et al., Predicting mechanical behaviors of rubber materials with artificial neural networks, Int. J. Mech. Sci. 249 (2023), 108265.

[97] X. Jin, S. Zheng, Measurement and calibration of optical instruments based on metrological calibration method and artificial neural network, Optik (2022), 170479.

[98] W. Ma, Z. Liu, Z.A. Kudyshev, et al., Deep learning for the design of photonic structures, Nat. Photonics 15 (2021) 77–90.

[99] X. Zhang, J. Yang, T. Lin, et al., Food and agro-product quality evaluation based on spectroscopy and deep learning: A review, Trends Food Sci. Technol. 112 (2021) 431–441.

[100] Z. Liu, L. Jin, J. Chen, et al., A survey on applications of deep learning in microscopy image analysis, Comput. Biol. Med. 134 (2021), 104523.

[101] K. He, X. Zhang, S. Ren, et al., Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). June 27-30, 2016, Las Vegas, IEEE, NV, USA, 2016, pp. 770–778.

[102] J. Dong, Z. Zuo, J. Zhang, et al., Geographical discrimination of *Boletus edulis* using two dimensional correlation spectral or integrative two dimensional correlation spectral image with ResNet, Food Contr. 129 (2021), 108132.

[103] J. Lu, J. Qin, P. Chen, et al., Quality difference study of six varieties of *Ganoderma lucidum* with different origins, Front. Pharmacol. 3 (2012), 57.

[104] P. Wang, Z. Yu, Species authentication and geographical origin discrimination of herbal medicines by near infrared spectroscopy: A review, J. Pharm. Anal. 5 (2015) 277–284.

[105] L. Qi, F. Zhong, Y. Chen, et al., An integrated spectroscopic strategy to trace the geographical origins of emblic medicines: Application for the quality assessment of natural medicines, J. Pharm. Anal. 10 (2020) 356–364.

[106] Y. Yang, Y. Zhao, Z. Zuo, et al., Determination of total flavonoids for *Paris polyphylla* var. *Yunnanensis* in different geographical origins using UV and FT-IR spectroscopy, J. AOAC Int. 102 (2019) 457–464.

[107] C. Liu, T. Shen, F. Xu, et al., Main components determination and rapid geographical origins identification in *Gentiana rigescens* Franch. based on HPLC, 2DCOS images combined to ResNet, Ind. Crops Prod. 187 (2022), 115430.

[108] K. Tahri, C. Tiebe, N. El Bari, et al., Geographical provenience differentiation and adulteration detection of cumin by means of electronic sensing systems and SPME-GC-MS in combination with different chemometric approaches, Anal. Methods 8 (2016) 7638–7649.

[109] C. He, W. Huang, X. Xue, et al., UPLC-MS fingerprints, phytochemicals and quality evaluation of flavonoids from *Abrus precatorius* leaves, J. Food Compos. Anal. 110 (2022), 104585.

[110] X. Wu, Z. Zuo, Q. Zhang, et al., FT-MIR and UV-vis data fusion strategy for origins discrimination of wild *Paris Polyphylla* Smith var. *yunnanensis*, Vib. Spectrosc. 96 (2018) 125–136.

[111] X. Wu, Q. Zhang, Y. Wang, Traceability the provenience of cultivated *Paris polyphylla* Smith var. *yunnanensis* using ATR-FTIR spectroscopy combined with chemometrics, Spectrochim. Acta A Mol. Biomol. Spectrosc. 212 (2019) 132–145.

[112] Z. Liu, S. Yang, Y. Wang, et al., Multi-platform integration based on NIR and UV-Vis spectroscopies for the geographical traceability of the fruits of *Amomum tsao-ko*, Spectrochim. Acta Part A Mol. Biomol. Spectrosc. 258 (2021), 119872.

[113] Y. Zhou, Z. Zuo, F. Xu, et al., Origin identification of *Panax notoginseng* by multi-sensor information fusion strategy of infrared spectra combined with random forest, Spectrochim, Acta Part A Mol. Biomol. Spectrosc. 226 (2020), 117619.

[114] Q. Wang, H. Huang, Y. Wang, Geographical authentication of *Macrohyporia cocos* by a data fusion method combining ultra-fast liquid chromatography and Fourier transform infrared spectroscopy, Molecules 24 (2019), 1320.

[115] Y. Li, J. Zhang, Y. Wang, FT-MIR and NIR spectral data fusion: A synergetic strategy for the geographical traceability of *Panax notoginseng*, Anal. Bioanal. Chem. 410 (2018) 91–103.

[116] L. Li, S. Zhang, Z. Zuo, et al., Data fusion of multiple-information strategy based on Fourier transform near infrared spectroscopy and Fourier-transform mid infrared for geographical traceability of *Wolfiporia cocos* combined with chemometrics, J. Chemom. 36 (2022), e3436.

[117] Y. Wang, M. Li, T. Feng, et al., Discrimination of *Radix* Astragali according to geographical regions by data fusion of laser induced breakdown spectroscopy (LIBS) and infrared spectroscopy (IR) combined with random forest (RF), Chin. J. Anal. Chem. 50 (2022), 100057.

[118] J. Hao, F. Dong, Y. Li, et al., Investigation of the data fusion of spectral and textural data from hyperspectral imaging for the near geographical origin discrimination of wolfberries using 2D-CNN algorithms, Infrared Phys. Technol. 125 (2022), 104286.

[119] N. Shakiba, A. Gerdes, N. Holz, et al., Determination of the geographical origin of hazelnuts (*Corylus avellana* L.) by Near-Infrared spectroscopy (NIR) and a Low-Level Fusion with nuclear magnetic resonance (NMR), Microchem. J. 174 (2022), 107066.

[120] S. Li, C. Liu, C. Cai, et al., Geographical traceability of germplasm resources of *Paris polyphylla* var. *yunnanensis* based on multi-block information integration platform, J. Appl. Res. Med. Aromat. Plants 31 (2022), 100440.

[121] Y. Wang, Z. Zuo, H. Huang, et al., Original plant traceability of *Dendrobium* species using multi-spectroscopy fusion and mathematical models, R. Soc. Open Sci. 6 (2019), 190399.

[122] C. Wang, L. Tang, L. Li, et al., Geographic authentication of *Eucommia ulmoides* leaves using multivariate analysis and preliminary study on the compositional response to environment, Front. Plant Sci. 11 (2020), 79.
[123] Y. Zhao, T. Yuan, L. Wu, et al., Identification of *Gentiana rigescens* from different geographical origins based on HPLC and FTIR fingerprints, Anal. Methods 12 (2020) 2260–2271.
[124] H. Fu, Q. Yin, L. Xu, et al., A comprehensive quality evaluation method by FT-NIR spectroscopy and chemometric: Fine classification and untargeted authentication against multiple frauds for Chinese *Ganoderma lucidum*, Spectrochim. Acta Part A Mol. Biomol. Spectrosc. 182 (2017) 17–25.
[125] A. Amirvaresi, N. Nikounezhad, M. Amirahmadi, et al., Comparison of near-infrared (NIR) and mid-infrared (MIR) spectroscopy based on chemometrics for saffron authentication and adulteration detection, Food Chem. 344 (2021), 128647.
[126] S. Kaiser, Â.R. Carvalho, V. Pittol, et al., Chemical differentiation between *Uncaria tomentosa* and *Uncaria guianensis* by LC-PDA, FT-IR and UV methods coupled to multivariate analysis: A reliable tool for adulteration recognition, Microchem. J. 152 (2020), 104346.
[127] F. Sun, Y. Chen, K. Wang, et al., Identification of genuine and adulterated *Pinellia ternata* by mid-infrared (MIR) and near-infrared (NIR) spectroscopy with partial least squares-discriminant analysis (PLS-DA), Anal. Lett. 53 (2020) 937–959.
[128] N.P. Mncwangi, A.M. Viljoen, J. Zhao, et al., What the devil is in your phytomedicine? Exploring species substitution in *Harpagophytum* through chemometric modeling of ^{1}H-NMR and UHPLC-MS datasets, Phytochemistry 106 (2014) 104–115.
[129] S. Kiani, S. Minaei, M. Ghasemi-Varnamkhasti, Integration of computer vision and electronic nose as non-destructive systems for saffron adulteration detection, Comput. Electron. Agric. 141 (2017) 46–53.
[130] H. Yang, L. Bao, Y. Liu, et al., Identification and quantitative analysis of salt-adulterated honeysuckle using infrared spectroscopy coupled with multi-chemometrics, Microchem. J. 171 (2021), 106829.
[131] L. Lin, M. Xu, L. Ma, et al., A rapid analysis method of safflower (*Carthamus tinctorius* L.) using combination of computer vision and near-infrared, Spectrochim. Acta Part A Mol. Biomol. Spectrosc. 236 (2020), 118360.
[132] S. Varliklioz Er, H. Eksi-Kocak, H. Yetim, et al., Novel spectroscopic method for determination and quantification of saffron adulteration, Food Anal. Meth. 10 (2017) 1547–1555.
[133] A.M. Jiménez-Carvelo, M. Tonolini, O. McAleer, et al., Multivariate approach for the authentication of vanilla using infrared and Raman spectroscopy, Food Res. Int. 141 (2021), 110196.
[134] J.P. Cruz-Tirado, R.L. de França, M. Tumbajulca, et al., Detection of cumin powder adulteration with allergenic nutshells using FT-IR and portable NIRS coupled with chemometrics, J. Food Compos. Anal. 116 (2023), 105044.
[135] H. Fu, Q. Shi, L. Wei, et al., Rapid recognition of geoherbalism and authenticity of a Chinese herb by data fusion of near-infrared spectroscopy (NIR) and mid-infrared (MIR) spectroscopy combined with chemometrics, J. Spectrosc. 2019 (2019) 1–9.
[136] A. Massaro, A. Negro, M. Bragolusi, et al., Oregano authentication by mid-level data fusion of chemical fingerprint signatures acquired by ambient mass spectrometry, Food Contr. 126 (2021), 108058.
[137] X. Yang, J. Song, L. Peng, et al., Improving identification ability of adulterants in powdered *Panax notoginseng* using particle swarm optimization and data fusion, Infrared Phys. Technol. 103 (2019), 103101.
[138] W. Sun, X. Zhang, Z. Zhang, et al., Data fusion of near-infrared and mid-infrared spectra for identification of rhubarb, Spectrochim. Acta A Mol. Biomol. Spectrosc. 171 (2017) 72–79.
[139] K. Kucharska-Ambrożej, A. Martyna, J. Karpińska, et al., Quality control of mint species based on UV-VIS and FTIR spectral data supported by chemometric tools, Food Contr. 129 (2021), 108228.
[140] M.R. Gómez, I. Maestro-Gaitán, P.C. Magro, et al., Unique nutritional features that distinguish *Amaranthus cruentus* L. and *Chenopodium quinoa* Willd seeds, Food Res. Int. Ott. Ont 164 (2023), 112160.
[141] J.P. Cruz-Tirado, Y.L. Brasil, A.F. Lima, et al., Rapid and non-destructive cinnamon authentication by NIR-hyperspectral imaging and classification chemometrics tools, Spectrochim. Acta Part A Mol. Biomol. Spectrosc. 289 (2023), 122226.
[142] H. Qin, Y. Wang, W. Yang, et al., Comparison of metabolites and variety authentication of *Amomum tsao-ko* and *Amomum paratsao-ko* using GC-MS and NIR spectroscopy, Sci. Rep. 11 (2021), 15200.
[143] S. Li, X. Guan, Z. Gao, et al., A simple method to discriminate Guangchenpi and Chenpi by high-performance thin-layer chromatography and high-performance liquid chromatography based on analysis of dimethyl anthranilate, J. Chromatogr. B 1126-1127 (2019), 121736.
[144] Y. Wang, T. He, J. Wang, et al., High performance liquid chromatography fingerprint and headspace gas chromatography-mass spectrometry combined with chemometrics for the species authentication of Curcumae Rhizoma, J. Pharm. Biomed. Anal. 202 (2021), 114144.
[145] J. Zhang, C. Fei, W. Zhang, et al., Rapid identification for the species discrimination of Curcumae Rhizoma using spectrophotometry and flash gas chromatography e-nose combined with chemometrics, Chin. J. Anal. Chem. 50 (2022), 100167.
[146] J.C. Machado Jr., M.A. Faria, I.M. Ferreira, et al., Varietal discrimination of hop pellets by near and mid infrared spectroscopy, Talanta 180 (2018) 69–75.
[147] Y. Ding, Q. Zhang, Y. Wang, A fast and effective way for authentication of *Dendrobium* species: 2DCOS combined with ResNet based on feature bands extracted by spectrum standard deviation, Spectrochim. Acta A Mol. Biomol. Spectrosc. 261 (2021), 120070.
[148] J.E. Maree, A.M. Viljoen, Fourier transform near- and mid-infrared spectroscopy can distinguish between the commercially important Pelargonium sidoides and its close taxonomic ally P. reniforme, Vib. Spectrosc. 55 (2011) 146–152.
[149] P. Zhao, G. Dou, G. Chen, Wood species identification using feature-level fusion scheme, Optik 125 (2014) 1144–1148.
[150] G. Huang, L. Lin, M. Zhang, et al., Discrimination of genuine and non-genuine Magnolia officinalis leaves based on multi-technique data fusion of ultra-high performance liquid chromatography-quadrupole time-of-flight tandem mass spectrometry, gas chromatography-mass spectrometry, and chemometrics, Sep. Sci. Plus 6 (2023), 2200074.
[151] N. Mncwangi, I. Vermaak, A.M. Viljoen, Mid-infrared spectroscopy and short wave infrared hyperspectral imaging—a novel approach in the qualitative assessment of *Harpagophytum* procumbens and *H. zeyheri* (Devil's Claw), Phytochem. Lett. 7 (2014) 143–149.
[152] S. Li, Q. Shao, Z. Lu, et al., Rapid determination of crocins in saffron by near-infrared spectroscopy combined with chemometric techniques, Spectrochim. Acta A Mol. Biomol. Spectrosc. 190 (2018) 283–289.
[153] Y. Wang, Y. Yang, H. Sun, et al., Application of a data fusion strategy combined with multivariate statistical analysis for quantification of puerarin in Radix puerariae, Vib. Spectrosc. 108 (2020), 103057.
[154] X. Song, E. Canellas, E. Asensio, et al., Predicting the antioxidant capacity and total phenolic content of bearberry leaves by data fusion of UV-Vis spectroscopy and UHPLC/Q-TOF-MS, Talanta 213 (2020), 120831.
[155] Q. Hao, J. Zhou, L. Zhou, et al., Prediction the contents of fructose, glucose, sucrose, fructo-oligosaccharides and iridoid glycosides in *Morinda officinalis* radix using near-infrared spectroscopy, Spectrochim. Acta Part A Mol. Biomol. Spectrosc. 234 (2020), 118275.
[156] J. Li, J. Wen, G. Tang, et al., Development of a comprehensive quality control method for the quantitative analysis of volatiles and lignans in *Magnolia biondii* Pamp. by near infrared spectroscopy, Spectrochim. Acta Part A Mol. Biomol. Spectrosc. 230 (2020), 118080.
[157] L. Lei, C. Ke, K. Xiao, 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 Part A Mol. Biomol. Spectrosc. 262 (2021), 120119.
[158] S. Zhang, H. Ma, H. Pan, et al., Quantitative real-time release testing of rhubarb based on near-infrared spectroscopy and method validation, Vib. Spectrosc. 104 (2019), 102964.
[159] J. Xue, Q. Yang, C. Li, et al., Rapid and simultaneous quality analysis of the three active components in *Lonicerae* Japonicae *Flos* by near-infrared spectroscopy, Food Chem. 342 (2021), 128386.
[160] S. Mazurek, I. Fecka, M. Węglińska, et al., Quantification of active ingredients in *Potentilla tormentilla* by Raman and infrared spectroscopy, Talanta 189 (2018) 308–314.
[161] S. Lafhal, P. Vanloot, I. Bombarda, et al., Identification of metabolomic markers of lavender and lavandin essential oils using mid-infrared spectroscopy, Vib. Spectrosc. 85 (2016) 79–90.
[162] S. Zhang, X. Gong, H. Qu, Near-infrared spectroscopy and HPLC combined with chemometrics for comprehensive evaluation of six organic acids in *Ginkgo biloba* leaf extract, J. Pharm. Pharmacol. 74 (2022) 1040–1050.
[163] C. Chen, H. Li, X. Lv, et al., Application of near infrared spectroscopy combined with SVR algorithm in rapid detection of cAMP content in red jujube, Optik 194 (2019), 163063.
[164] L. Tao, Z. Lin, J. Chen, et al., Mid-infrared and near-infrared spectroscopy for rapid detection of Gardeniae Fructus by a liquid-liquid extraction process, J. Pharm. Biomed. Anal. 145 (2017) 1–9.
[165] Y. Cui, J. Wu, Y. Chen, et al., Optimization of near-infrared reflectance models in determining flavonoid composition of okra (*Abelmoschus esculentus* L.) pods, Food Chem. 418 (2023), 135953.
[166] Z. Lan, Y. Zhang, H. Lin, et al., Efficient monitoring for the nutrient changes in stir-fried Moutan Cortex using non-destructive near-infrared spectroscopy sensors, Microchem. J. 183 (2022), 107972.
[167] Y. Yang, Y. Wu, W. Li, et al., Determination of geographical origin and icariin content of Herba Epimedii using near infrared spectroscopy and chemometrics, Spectrochim. Acta A Mol. Biomol. Spectrosc. 191 (2018) 233–240.
[168] Y. Guan, T. Ye, Y. Yi, et al., Rapid quality evaluation of *Plantaginis Semen* by near infrared spectroscopy combined with chemometrics, J. Pharm. Biomed. Anal. 207 (2022), 114435.
[169] G. Ding, B. Li, Y. Han, et al., A rapid integrated bioactivity evaluation system based on near-infrared spectroscopy for quality control of *Flos Chrysanthemi*, J. Pharm. Biomed. Anal. 131 (2016) 391–399.
[170] L. Qi, Y. Ma, F. Zhong, et al., Comprehensive quality assessment for Rhizoma Coptidis based on quantitative and qualitative metabolic profiles using high performance liquid chromatography, Fourier transform near-infrared and Fourier transform mid-infrared combined with multivariate statistical analysis, J. Pharm. Biomed. Anal. 161 (2018) 436–443.
[171] Q. Wang, H. Huang, Y. Wang, FTIR and UV spectra for the prediction of triterpene acids in *Macrohyporia cocos*, Microchem. J. 158 (2020), 105167.

[172] Y. Wu, Y. Zheng, Q. Li, et al., Study on difference between epidermis, phloem and xylem of Radix Ginseng with near-infrared and infrared spectroscopy coupled with principal component analysis, Vib. Spectrosc. 55 (2011) 201–206.
[173] Q. Wang, Z. Zuo, H. Huang, et al., Comparison and quantitative analysis of wild and cultivated *Macrohyporia cocos* using attenuated total refection-Fourier transform infrared spectroscopy combined with ultra-fast liquid chromatography, Spectrochim. Acta Part A Mol. Biomol. Spectrosc. 226 (2020), 117633.
[174] L. Li, Y. Zhao, Z. Li, et al., Multi-information based on ATR-FTIR and FT-NIR for identification and evaluation for different parts and harvest time of *Dendrobium officinale* with chemometrics, Microchem. J. 178 (2022), 107430.
[175] J. Zhang, Y. Wang, M. Yang, et al., Identification and evaluation of *Polygonatum kingianum* with different growth ages based on data fusion strategy, Microchem. J. 160 (2021), 105662.
[176] Q. Zhao, Y. Yu, N. Hao, et al., Data fusion of Laser-induced breakdown spectroscopy and Near-infrared spectroscopy to quantitatively detect heavy metals in lily, Microchem. J. 190 (2023), 108670.
[177] P. Zhang, Y. Xu, F. Qu, et al., Rapid quality evaluation of four kinds of Polygoni Multiflori *Radix* Praeparata by electronic eye combined with chemometrics, Phytochem. Anal. 34 (2023) 301–316.
[178] G. Xuan, C. Gao, Y. Shao, et al., Maturity determination at harvest and spatial assessment of moisture content in okra using Vis-NIR hyperspectral imaging, Postharvest Biol. Technol. 180 (2021), 111597.
[179] C. Hai, W. Long, Y. Suo, et al., Nano-effect multivariate fusion spectroscopy combined with chemometrics for accurate identification the cultivation methods and growth years of *Dendrobium huoshanense*, Microchem. J. 179 (2022), 107556.
[180] L. Wang, Q. Wang, Y. Wang, et al., Comparison of geographical traceability of wild and cultivated *Macrohyporia cocos* with different data fusion approaches, J. Anal. Meth. Chem. 2021 (2021), 5818999.
[181] L. Liu, W. Li, Z. Zuo, et al., Multisource information fusion strategies of mass spectrometry and Fourier transform infrared spectroscopy data for authenticating the age and parts of Vietnamese ginseng, J. Chemom. 35 (2021), e3376.