Food Chemistry: X 19 (2023) 100860

Contents lists available at ScienceDirect

Food Chemistry: X

journal homepage: www.sciencedirect.com/journal/food-chemistry-x

Machine learning applications for multi-source data of edible crops: A review of current trends and future prospects

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

Keywords:
Machine learning
Edible crops
Multi-source data
Data fusion strategy
Quality evaluation

Chemical compounds studied in this article::
3-O-Caffeoylquinic acid (PubChem CID: 1794427)
3,5-Dicaffeoylquinic acid (PubChem CID: 6474310)
4,5-Dicaffeoylquinic acid (PubChem CID: 6474309)
(+)-catechin (PubChem CID: 9064)

ABSTRACT

The quality and safety of edible crops are key links inseparable from human health and nutrition. In the era of rapid development of artificial intelligence, using it to mine multi-source information on edible crops provides new opportunities for industrial development and market supervision of edible crops. This review comprehensively summarized the applications of multi-source data combined with machine learning in the quality evaluation of edible crops. Multi-source data can provide more comprehensive and rich information from a single data source, as it can integrate different data information. Supervised and unsupervised machine learning is applied to data analysis to achieve different requirements for the quality evaluation of edible crops. Emphasized the advantages and disadvantages of techniques and analysis methods, the problems that need to be overcome, and promising development directions were proposed. To monitor the market in real-time, the quality evaluation methods of edible crops must be innovated.

1. Introduction

The edible crop industry has spread worldwide, especially in developing countries, and has attracted the attention of researchers and the public (Su et al., 2016). A series of active ingredients contained in edible crops are the basis for ensuring their quality and are closely related to human health and nutrition (Wen et al., 2018). The world's demand for edible crops has been rising, but their quality and yield are affected by the natural environment and social factors (Delpeuch & Leblois, 2014; Su et al., 2016). To maximize the benefits obtained, the edible crop industry has fallen into a crisis of trust, which has greatly consumed consumer confidence. Consumers tend to consume edible crops that have positive effects on metabolic activities. In order to increase the competitiveness of market consumption, it is imperative to certify and control the quality of edible crops (Wen et al., 2018). However, the evaluation indicators of edible crops are increasingly inclined to be multi-component, and it is difficult to achieve quality evaluation because only a single analysis technology cannot fully characterize its active components (He & Zhou, 2021).

There are two main sources of multi-source data for edible crops: One uses the same data platform to analyze different biological entities; the other uses different data platforms to analyze the same biological entity (Stavropoulos et al., 2021). Multi-source data can reflect more comprehensive information compared to a single data source. Due to the complexity and diversity of the chemical components of edible crops, it is necessary to fully characterize them with multi-source data (Pei et al., 2020). Data fusion is typically used as a method to deal with imperfect raw data. Integrating multi-source data through a data fusion strategy can reveal a variety of reliable and accurate feature information and obtain better decision-making, prediction, and classification results (Zhang et al., 2022). And it has significant advantages over single analytical technology, enabling higher precision and predictive accuracy. For example, active ingredients with low content can be detected through data fusion, and the detection range is more extensive than that of a single analytical technique (Li et al., 2021). Nowadays, data fusion has been more and more applied to civilian applications. However, because its data come from multiple complementary sources, it has led to cumbersome experiments and increased computation costs (Zhou et al., 2022). Only by using data mining can the target information in multi-source data be identified more quickly and accurately (He & Zhou,

* Corresponding author.
E-mail address: boletus@126.com (Y. Wang).

https://doi.org/10.1016/j.fochx.2023.100860
Received 6 July 2023; Received in revised form 23 August 2023; Accepted 31 August 2023
Available online 3 September 2023

2021).

Machine learning is one of the powerful tools for data mining, which can extract meaningful information from a large amount of data (Salcedo-Sanz et al., 2020). Machine learning imitates human recognition ability and provides reliable solutions to problems through computation (Greener et al., 2022). Because of its robust estimation and prediction ability, it can compensate for the lack of data fusion, and combining the two significantly improves the ability to process multi-source data (Meng et al., 2020). The multi-source data of edible crops often contain too much complex feature information, and the dataset is also relatively large, machine learning has significant advantages in dealing with such problems, which can effectively shorten the data analysis time (Greener et al., 2022). Therefore, the purpose of this review is to comprehensively summarize the research status of multi-source data combined with machine learning in edible crop quality evaluation in the past decade. It briefly introduced multi-source data types, machine learning, and their strengths and weaknesses. It has clarified the significance of using data fusion strategies to process multi-source data and the differences between different data fusion strategies. The applicability of different data combinations and models in the quality evaluation of edible crops was discussed in depth. Further analysis of the limitations of existing evaluation methods and prospects for future research directions. Fig. 1 showed the application of multi-source data/data fusion strategy combined with machine learning in edible crops.

2. Multi-source data of edible crops

2.1. Modern analytical technique

Some studies, such as edible crops fraud, showed that traditional methods had been gradually replaced by modern analytical techniques, including spectroscopy, chromatography, mass spectrometry (MS), nuclear magnetic resonance spectroscopy (NMR), and electronic sensor system (Liu et al., 2022b). As one of the sources of multi-source data on edible crops, these five analytical techniques provide a scientific basis for their quality evaluation. The following sections would briefly introduce each of the aforementioned analytical techniques.

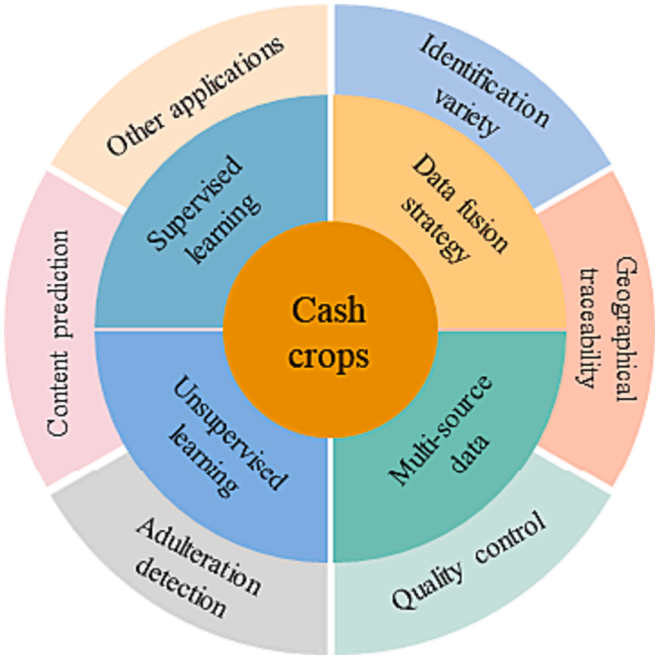

Fig. 1. The application of multi-source data/data fusion strategy combined with machine learning in edible crops.

2.1.1. Spectroscopic techniques

With the pursuit of green environmental protection, spectroscopy has been increasingly used in the research of edible crops. Different spectroscopic techniques can obtain information on different functional groups, and combining different spectral information is helpful to comprehensively study the overall chemical profile of edible crops. Ultraviolet–visible (UV–Vis) spectroscopy, as an electronic spectroscopy technique, is able to absorb ultraviolet radiation to cause energy level transition of electrons, thereby providing the information of different chromophores such as carboxyl groups, double bonds, triple bonds, and conjugated double bonds, etc. in the wavelength range of 200–780 nm (Rajput et al., 2022; Włodarska et al., 2021; Zaroual et al., 2022).

One of the most popular green chemical t techniques is infrared (IR) spectra, which can be divided into three sub-regions: far-infrared region, mid-infrared (MIR) region, and near-infrared (NIR) region. The chemical ingredients and related compounds of the analyte can be detected by the energy provided by the molecular bond movement (such as vibration) and the energy absorbed when the mid-infrared light irradiates the analyte (De Marchi et al., 2014). In order to solve the problems of low resolution and slow speed of MIR spectrometer, Fourier transform infrared (FT-IR) spectroscopy makes up for the lack of MIR spectra, and is widely used because of its wide detection spectral range and high resolution (Chen et al., 2022). Similar to the principle of MIR spectra, the signal of NIR spectra is closely related to molecular vibrations, especially with overtones and combinations of fundamental vibrations (Abbas et al., 2018). The absorption of electromagnetic radiation is the basis for the formation of the NIR spectrum, and the use of different measurement modes of the NIR instrument is able to capture the NIR radiation that interacts with the samples (Shafiee & Minaei, 2018). Every edible crop possesses its unique fingerprint in the NIR spectrum, so combining machine learning with NIR spectra is necessary for quality control (Oliveira et al., 2019).

Raman spectroscopy is also an emerging molecular vibrational spectroscopy technique. On the basis of inelastic scattering that occurs between photons and molecules, there are two types of Raman scattering: Stokes and the other is anti-Stokes shifts (Allakhverdiev et al., 2022; Yang et al., 2022). Compared with vibrational spectroscopy, the hyperspectral imaging (HSI) technique can obtain relevant spectral features and provide spatial information to generate hyperspectral images of samples at different wavelengths (Peng et al., 2022). Nevertheless, atomic spectroscopy mainly provides information about the internal structure of atoms. In the analysis of edible crops, inductively coupled plasma mass spectrometry (ICP-MS) and inductively coupled plasma-optical emission spectrometric (ICP-OES) are two common methods used to detect elemental profiles. ICP-MS combined with machine learning is a proven strategy for identifying the geographical origin of grapes and wines in China (Gao et al., 2022). ICP-OES is a hyphenated technique that has developed rapidly in recent years and has been widely used in the analysis of edible crops (Khan et al., 2021). The traditional method needs to reserve enough knowledge and experience, which is time-consuming and labor-intensive, while Laser induced plasma spectroscopy (LIBS) and X-ray fluorescence analysis (XRF) do not require sample pretreatment, and have the advantages of fast and non-destructive (Quackatz et al., 2022).

The active ingredients and nutrients in edible crops cover a wide range. Different spectroscopic techniques can represent different chemical properties, so the multi-spectral technique is usually used as a complementary source in many links, such as adulteration identification, geographical origin identification, and quality control of edible crops, and has achieved remarkable results. Spectroscopic technique plays a vital role in the quality assessment of edible crops. In addition, terahertz spectroscopy can be used to identify the physical and chemical properties of edible crops. However, its application in combination with other techniques in edible crops is very few, which is a direction with excellent development space.

2.1.2. Chromatography

Chromatography can provide data with good precision and accuracy, among which gas and liquid chromatography are widely used. These two chromatography methods use the compounds' size and the stationary phase's affinity to separate and detect compounds. Both Liquid chromatography (LC) and UV can be used as means of quantitative analysis, liquid chromatography-tandem mass spectrometry (LC-MS) has the advantages of high sensitivity, high detection selectivity, and high qualitative ability over UV (Abdel Razeq et al., 2021). High-performance liquid chromatography (HPLC) is a relatively common LC as a reliable method for the comprehensive evaluation of food and plant extracts, its fingerprints have high precision and resolution. Because of the increasingly stringent requirements of the experiment, the pursuit of a fast, sensitive, and specific method, and the concern of environmental impact, supercritical fluid chromatography (SFC) has more advantages than LC and can shorten the analysis time (Toribio et al., 2021).

Gas chromatography (GC) outperforms LC in terms of theoretical plate number and cost performance (Ichihara et al., 2021). GC mainly uses the differences in the boiling point, polarity, and adsorption properties of substances to separate mixtures. It has been widely used to detect the separation, identification, and quantification of most edible crops and aroma compounds containing chemical composition, such as volatile compositions and lipids (Guo et al., 2022b; Yao et al., 2022). Solid-phase microextraction gas chromatography-mass spectrometry (SPME-GC–MS) is gradually replacing traditional analytical methods applied to the detection of aroma compositions (Arslan et al., 2022). Meanwhile, comprehensive two-dimensional GC–MS (2D GC–MS) is also an effective tool for analyzing complex volatile and semi-volatile compositions (Sudol et al., 2022).

One of the key indicators to evaluate the economic value of edible crops is the content of active ingredients. Chromatography is a recognized quantitative analysis technique, and although some emerging techniques can serve as substitutes, its position is still unshakable. The most common way to predict the content of edible crops is to use regression models to correlate spectral and chromatographic data. At the same time, chromatography also has unique advantages in identifying adulteration of edible crops, which are highly sensitive to compounds with low concentrations.

2.1.3. Nuclear magnetic resonance spectroscopy

The generation of NMR spectroscopy is based on the energy exchange between nuclei with spin quantum numbers not equal to 0, such as ^{1}H, ^{1}C, ^{1}F, and alternating magnetic fields under the influence of two magnetic fields (constant magnetic field and alternating magnetic field) (Cao et al., 2021a). It is an indispensable quantitative analysis technique to obtain structural information on organic compounds. NMR combined with chemometrics enables the comprehensive characterization of complex compounds, which is beneficial for quality control and authenticity studies of edible crops (Monakhova et al., 2018). NMR is an essential tool to analyze the structure of compounds. It is mainly used for qualitative analysis in the research of edible crops, but it is rarely used at present. It is undeniable that NMR also has specific development potential in the quantitative analysis of chemical mixtures. In future edible crop research, NMR can be widely considered for use.

2.1.4. Mass spectrometry and its hyphenated techniques

Mass spectrometry (MS) is hyphenated with separation techniques such as LC and GC to help take advantage of both atmospheric-pressure chemical ionization (APCI) and electrospray ionization (ESI) are usually used as ionization sources. GC–MS is the earliest-developed chromatography-MS technique and is commonly used for the determination of complex compounds due to its high accuracy and sensitivity (Feizi et al., 2021). LC-MS is regarded as an extension of GC–MS, but compared with GC–MS, LC-MS is more difficult to couple due to the liquid eluent in the ion source (Famiglini et al., 2021). MS generally plays an auxiliary role in the qualitative and quantitative analysis of edible crops. It is mainly used in conjunction with techniques with separation capabilities to complement each other's strengths and weaknesses. With the emergence of mass spectrometry imaging, MS has a new application form in edible crops (Jiang et al., 2022). Nevertheless, only some studies have reported applying this technique to edible crops. Mass spectrometry imaging has high resolution and sensitivity while maintaining the integrity of samples, which can be used as a potential tool for the quality assessment of edible crops.

2.1.5. Electronic sensor systems

Traditional sensory evaluation relies on the experience of professionals, has intense subjective color, and the evaluation results are easily affected by many factors. With the rapid development of the era of artificial intelligence, electronic sensing technique has emerged at the historical moment and gradually become a substitute for traditional analytical techniques and sensory evaluation. Three kinds of electronic sensor techniques are widely used in the quality evaluation of edible crops: Electronic nose (E-nose), electronic tongue (E-tongue), and computer vision.

The design inspiration for the E-nose and E-tongue comes from the behavior pattern of mammalian recognition targets. The sensors of the E-nose and E-tongue are non-selective or semi-selective, combined in the array to output the response to the target. Another important part is the pattern recognition system, which analyzes the sensor's perception of smell or taste (Huang et al., 2019). Computer vision simulates human visual functions and comprises lighting equipment, a camera, and a personal computer. Its primary purpose is to obtain the color, shape, and texture information of the target through the camera and lighting equipment. In addition, the digital attributes of the target or imaging scene can also be collected by computer vision. The personal computer is the core part, used to analyze and process the valuable information in the image. These three sensors are mainly used for qualitative analysis and are challenging to carry out accurate quantitative analysis (Chen et al., 2015). It is worth mentioning that a single sensor still has certain defects in detection and needs to be combined with other analysis techniques to obtain more accurate data.

2.2. Data fusion

2.2.1. Pre-processing

Selecting appropriate pre-processing methods according to the types and characteristics of multi-source data is necessary to improve the model's performance. Spectral or chromatographic analysis is affected by factors such as the external environment (humidity, temperature), measurement mode, and instrument during sample collection, which may easily lead to unfavorable changes such as baseline drift (Lan et al., 2020; Mishra et al., 2020). For example, spectral data not only provides rich functional group information but also may introduce noise, scattering effect, and other information that is not conducive to data fusion or modeling. The existence of redundant information highlights the need for pre-processing. Common pre-processing methods include baseline correction, standard normal variate (SNV), normalization, orthogonal signal correction (OSC), multiplicative scattering correction (MSC), derivative, and Savitzky-Golay (S-G) smoothing. Compared with a single pre-processing method, different complementary information can be obtained from the combination of multiple pre-processing methods. Pre-processing ensembles with response-oriented sequential alternative calibration (PROSAC) is also a multi-block data modeling technique, which is helpful for analyzing multi-block data of NIR spectrum data with different pre-processing methods (Mishra et al., 2022). This technique has superior performance and shows more advantages than sequential (SPORT) and parallel (PORTO) orthogonalized partial least squares regression, for example, it does not need to consider the order of pre-processing and data scaling after pre-processing.

In most studies, the purpose of pre-processing multi-source data is to obtain the optimal pre-processing method to develop high-precision

models. The difficulty in selecting appropriate pre-processing methods to optimize the corresponding data needs to be clearly stated which pre-processing method should be chosen for the problems in the dataset. The use of trial-and-error method filtering in pre-processing methods often overlooks the complementary information carried by different methods, which could be more conducive to obtaining reliable results. The essence of edible crops determines the complexity of its multi-source data. The ensemble pre-processing method can retain complementary information while achieving the optimization effect. It helps to improve the classification and prediction results of edible crops.

2.2.2. Feature extraction/selection

Feature extraction/selection is crucial before modeling with multi-source data or performing mid-level data fusion. The massive amount of data generated by multi-source data and the direct fusion of high-dimensional feature datasets increase the complexity of analysis and other issues determining the superiority of feature extraction/selection (Buchaiah & Shakya, 2022). Feature extraction/selection can improve model accuracy, solve model overfitting, reduce data processing time, etc., by reducing dimensionality (eliminating redundant information and noise to retain important features). Some commonly used feature extraction/selection methods such as principal component analysis (PCA), independent component analysis (ICA), wavelet transform (WT), Fourier transform (FT), partial least squares-discriminant analysis (PLS-DA), genetic algorithm (GA), competitive adaptive reweighed sampling (CARS), etc. However, the development of science and technology leads to the continuous increase of data dimension, which makes the feature extraction/selection of multi-source data particularly difficult, and the existing feature extraction/selection methods face unprecedented challenges. Therefore, improving and developing feature extraction/selection methods are necessary. The public gradually recognized the convolutional neural network (CNN) composed of a large number of neurons in the feature extraction of multi-source data fusion (Xiao et al., 2020).

2.2.3. Data fusion strategies

With the development of emerging instruments and techniques, the physicochemical information of edible crops can be easily obtained from various sources. Due to the complexity and diversity of the components of edible crops, it is difficult for a single data source to express their complex and complete chemical information. Analyzing them by a single method is not systematic, and it is easy to introduce noise and interference information. Data fusion is a complex process of detecting, combining, correlating, and estimating multi-source data (Lin et al., 2022). By optimizing the information obtained, more reliable inference can be generated than a single data source, so the quality of edible crops can be accurately and in real-time evaluated. There is a desirability to use data fusion to increase and improve the quantity and quality of information available about edible crops. In order to obtain more complementary information, most multi-source data are provided by different instruments. In addition to the instrumental analysis technique, information obtained from different sensor techniques and physical and chemical analysis can be used as the source of data fusion. The following was a brief introduction to these three data fusion strategies.

Low-level data fusion concatenates raw data with similar variance and quantity from different platforms to form a new data matrix with the number of rows equal to the number of samples and the number of columns equal to the number of variables. Although the operation of this data fusion strategy is more straightforward than the other two, it also has more variables and takes longer to calculate. This means that there is still a large amount of redundant information in the newly formed data matrix. Mid- and low-level data fusion are similar in that they are both carried out at the data level, the difference is that the former needs to extract features from the raw data before fusion. Moreover, mid-level data fusion can solve some problems existing in low-level data fusion, such as reducing redundant information and eliminating noise and interference. Mid-level data fusion compresses the amount of information to a certain extent, which is conducive to real-time analysis. High-level data fusion is the independent analysis of raw data from each source, and then the fusion of their results to produce a final decision. High-level data fusion has the significant advantage that each data source does not interact with each other and suffers less interference (Li et al., 2020b). However, data fusion in this way has a high probability of causing information loss, so care must be taken when processing the raw data (Borras et al., 2015). In the current study, low- and mid-level data fusion is mainly used for quantitative analysis of edible crops, while high-level data fusion is usually used for qualitative analysis.

3. Machine learning

3.1. Unsupervised learning

Unsupervised learning is a training method of machine learning for statistical analysis. Its main goal is to discover the inherent hidden properties of the dataset by calculating the commonalities between unlabeled samples, so as to avoid the trouble of labeling samples in supervised learning (Su et al., 2022). Due to the nature of unsupervised learning, it is often used as a powerful tool for label-expensive analysis or irrelevant applications (Wang & Biljecki, 2022). Although unsupervised learning cannot directly perform classification and regression, it has significant advantages in real-time data analysis (Cao et al., 2021b). Unsupervised learning has become one of the important solutions to problems such as detection, denoising, and recognition. Next, we will briefly introduce several unsupervised learning algorithms. Fig. 2 showed different unsupervised learning algorithms' advantages, shortcomings, and development.

3.1.1. Clustering

Clustering is an unsupervised machine learning for data mining that divides datasets into different clusters based on similarity to reveal the inherent properties of data (Ay et al., 2023). The choice of distance metric is important for quantifying similarity between data and affects the shape and configuration of the formation of clusters (Tarnutzer & Weber, 2022). Due to the strong practicality of clustering, it has become popular in applications such as pattern recognition and multivariate data analysis of edible crops. Hierarchical clustering is able to provide different resolution results without knowing a predetermined number of clusters, but it has the defects of inaccuracy and high time cost. In response to the above problems, Varshney et al. (2022) came up with the Probabilistic Intuitionistic Fuzzy Hierarchical Clustering Algorithm (PIFHC), which used probabilistic Euclidean distance as a distance measure to eliminate uncertainty in the data through intuitive fuzzy sets. The results showed that the clustering accuracy of PIFHC is significantly better than that of other clustering algorithms.

3.1.2. Dimensionality reduction

Because of the characteristics of high-dimensional data, it is difficult to understand and analyze directly, and the use of all variables that contain high correlation leads to confusion of information (Xu & Wu, 2022). Data dimensionality reduction is achieved by mapping high-dimensional data to low-dimensional feature space, which reduces redundant information without losing important data structures (Flexa et al., 2021). Dimensionality reduction is divided into two types: linear and nonlinear. PCA is the most classical and popular linear dimensionality reduction algorithm, which aims to minimize the mean square error to achieve the goal of dimensionality reduction (Li et al., 2022). PCA uses linear transformation to convert a group of multivariable into several comprehensive variables. The newly generated variable is the principal component derived from the variance–covariance matrix (Aidoo et al., 2021). Because the principal components are orthogonal to each other, and each principal component represents different

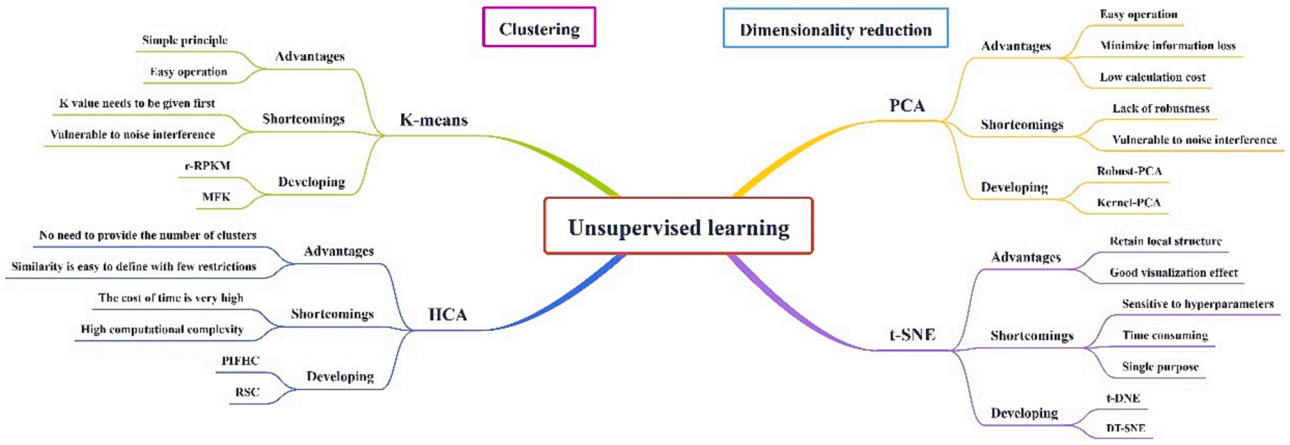

Fig. 2. The advantages, shortcomings and developing of different unsupervised learning algorithms.

information, they can explain most of the original data information, significantly improving the efficiency of data processing. In this algorithm, the reduction of dimensions promotes the visualization of hidden attributes and the correlation of data in PC space (Boubchir et al., 2022). At present, PCA is also increasingly used to explore edible crops' different components and geographical origins. However, traditional PCA is vulnerable to noise interference and lacks robustness, Qiao et al. (2022) proposed robust PCA to solve the above problems. The results showed that the improved algorithm has advantages and effectiveness. PCA cannot retain the local structure of the dataset, *t*-Distributed Stochastic Neighbor Embedding (*t*-SNE) emerges as the times require. *t*-SNE, as the most popular nonlinear dimensionality reduction algorithm, has a strong ability to capture manifold structure in high-dimensional data, attracting much attention in machine learning (Maaten &

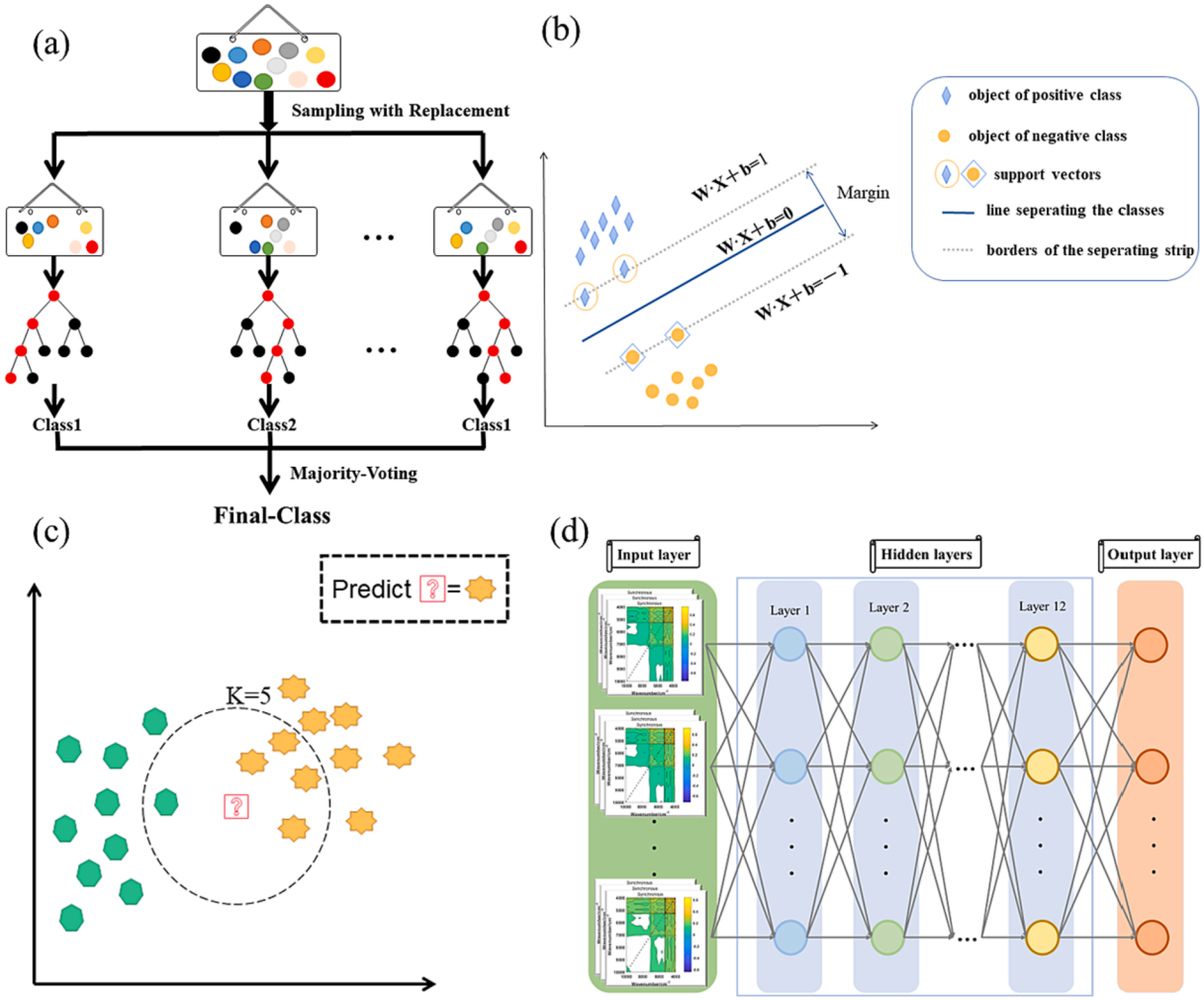

Fig. 3. The schematic diagram of four supervised learning algorithms.

Hinton, 2008). The main purpose of *t*-SNE is to balance two conditional probability distributions (input data and low dimensional representation) using the gradient descent method (Flexa et al., 2021). Compared with other nonlinear dimensionality reduction algorithms, *t*-SNE is considered the best tool to convert high-dimensional data into 2D or 3D, which is conducive to visualizing edible crops' multi-source data sets (Maaten & Hinton, 2008). It is sensitive to the selected super parameters and distance measurement. In order to make up for its shortcomings, Flexa et al. (2021) proposed a new *t*-SNE algorithm, that is, a polygonal coordinate system-based deterministic strategy combined with *t*-SNE, called *t*-Distributed deterministic Neighbor Embedding (*t*-DNE). The results of Friedman's significance test showed that a polygonal coordinate system has significant advantages in data embedding.

3.2. Supervised learning

Supervised learning is a branch of machine learning algorithms, and is the most widely used algorithm at present. It is mainly applied to two problems: One is the regression problem when the variables are continuous; the other is the classification problem when the sample data is a category. Supervised learning is trained by the sample data of an artificial marker, which requires a large sample size to achieve the expected model performance. Although this algorithm requires high calculation and time cost, it is undeniable that the performance of supervised learning is much better than unsupervised learning (Yakimovich et al., 2021). Table S1 compared different supervised learning algorithms' advantages, limitations, and development. This review mainly introduced the following supervised learning algorithms. Fig. 3 showed the schematic diagrams of four supervised learning algorithms.

3.2.1. Regression

3.2.1.1. Partial least squares regression (PLSR). PLSR is a mature analytical chemistry tool, that plays an important role in multivariate statistical analysis, so it is often used in data analysis (Metz et al., 2021). It combines three basic algorithms: PCA, canonical correlation analysis, and multivariate linear regression analysis. The purpose of PLSR is to minimize the variance of the prediction and maximize the covariance of matrices X and Y (He et al., 2015). The algorithm can extract the PCs of the input and output matrices at the same time, so it has a good effect in dealing with multicollinearity problems and noise data, especially when the internal variables are highly linearly correlated (Ma et al., 2022; Zhang et al., 2018). However, the PLSR model is susceptible to outliers, which reduces the model's predictive power (Metz et al., 2021). The proposal of domain adaption regularization-based kernel partial least squares regression (DarKPLS) is beneficial to improving the adaptive ability of the model, which has also been confirmed in subsequent research (Shan et al., 2023).

3.2.1.2. Support vector regression (SVR). SVR is developed from a support vector classifier and is widely used in nonlinear regression problems (Liu et al., 2020). This method is an excellent regularization algorithm based on kernel because it can estimate the distance between the hyperplane and boundary line (Rezaei et al., 2023). The advantage of SVR lies in its excellent handling of high-dimensional regression tasks, and its ability to minimize empirical risk and confidence interval to solve the problem of model overfitting (Shen et al., 2023). Moreover, in the case of a limited sample size, the global optimal solution can still be obtained from the existing knowledge and experience. Traditional SVR still has some drawbacks, which rely on relevant parameters and priori knowledge, and have high computational costs during the training phase (Shen et al., 2023). In order to overcome the shortcomings of traditional SVR, Peng (2010) developed a twin SVR, which has a faster training speed and stronger regression ability.

3.2.2. Classification

3.2.2.1. PLS-DA. PLS-DA is a multivariate statistical analysis method that reduces dimensionality by extracting features (Ma et al., 2020). PLS-DA successfully divides the space into two regions by a straight line to classify samples. Its discriminant rule is to compare the predicted response values of Y to a fixed scalar threshold (usually 0.5) (Jimenez-Carvelo et al., 2021a). Correctly selecting the number of latent variables is critical to constructing a PLS-DA model because too many latent variables will lead to overfitting of the model, while an insufficient number of latent variables will lead to underfitting of the model. Generally, the root mean square error of cross-validation (RMSECV) determines the number of latent variables selected. PLS-DA has apparent advantages in dealing with highly collinear and noise. Therefore, PLS-DA is widely used in classification and adulteration identification of edible crops (Vieira et al., 2021).

3.2.2.2. Support vector machine (SVM). SVM, as an effective classification method, is mainly based on statistical theory and has attracted more and more attention in related fields, such as data mining and machine learning (Cortes & Vapnik, 1995). Recently, SVM has been widely used in pattern recognition and classification analysis of edible crops (Park et al., 2015). Essentially, the SVM model achieves the purpose of separating the positive and negative classes by constructing the best-fitting hypersurface in a high-dimensional space (Lamberti, 2021). The important parameters of the support vector machine model are the penalty parameter (*c*) and kernel function (*g*). Suppose the parameter *c* is too high or too low. In that case, it will affect the model's performance, so the model's complexity is controlled by adjusting the parameter c. In contrast, the parameter g determines the complexity of the data distribution in the high-dimensional feature space (Wu et al., 2018). These two parameters determine the accuracy and generalization ability of the model to a certain extent. Cross-validation is usually used to determine the best parameters of the model. The optimization algorithm of parameters adopts GA, Grid Search (GS), Particle Swarm Optimization (PSO), etc. (Huang et al., 2020). However, compared with the Grey Wolf Optimization (GWO) algorithm, the above SVM optimization algorithms are prone to premature convergence and local optimization, thus GWO has more advantages in function optimization, simple operation, and few parameters (Liang et al., 2022). These optimization algorithms can make up for the defects of SVM to a certain extent.

3.2.2.3. Random forest (RF). RF integrates the decision tree and uses statistical estimation to predict and classify, while the decision tree is generated by selecting training samples and variable subsets through replacement (a bagging approach) (Liang et al., 2020). The RF model is constructed by a recursive partition containing hundreds of decision trees. The value stored in each node of the decision tree is related to the variables of classification information (Lovatti et al., 2019). There are two important parameters in random forest, one is Ntree, and the other is Mtry. The model's performance is usually evaluated using Ntree, and the best split node is determined by the best binary split result of Mtry. The split stops when the training samples on all decision trees have the same category on their corresponding nodes (Liang et al., 2020). Due to the large number of decision trees in the RF model, the processing of the model is more complex. Therefore, it is necessary to use the out-of-bag error rate (OOB) to optimize the parameters, adjust the number of decision trees and improve the model's performance (Wang et al., 2020). Compared with other statistical analysis methods, RF is a relatively new pattern recognition method to solve the problem of recognition and prediction, which can overcome the difficulties of traditional chemometrics in classifying data with many variables and few samples, and its risk of overfitting is lower and more robust than a single decision tree. However, before establishing the model of RF, some super parameters need to be established artificially, which increases the calculation cost

and even affects the accuracy of the results due to some errors in the selection process. Feng et al. (2021) used an improved artificial bee colony algorithm (IABC) combined with RF to optimize the super parameters of the model, which greatly improved the calculation efficiency.

3.2.2.4. Soft independent modeling of class analogy (SIMCA). SIMCA is a supervised pattern recognition technique, which is not only a powerful tool for processing, distinguishing, and classifying complex data, but also one of the commonly used classification methods in chemometrics (Gomez-de Anda et al., 2012). In each category, SIMCA performs PCA on it, builds a multi-dimensional space around the calculated PCs, and then classifies unknown samples by the distance between samples and models (Duca et al., 2016). The results of SIMCA are expressed in terms of sensitivity and specificity, and the higher the efficiency of cross-validation, the better performance of the model (Firmani et al., 2020). Although SIMCA is very sensitive to outliers and will affect the classification results to a certain extent, the advantage of SIMCA is that the unknown classification object will only be assigned to the category with the highest possibility, and if the residual variance of the sample exceeds the upper limit of each class in the modeling data set, it will not be assigned (Khanmohammadi et al., 2013). Because SIMCA has a strong ability to capture spectral signal differences caused by different chemical components, it is often used to process spectral data of edible crops. Aiming at the shortcomings of traditional SIMCA, robust-SIMCA, and data driven-soft independent modeling of class analogy (DD-SIMCA) improve its performance from different perspectives (Adenan et al., 2020; Rodionova et al., 2016).

3.2.2.5. K-nearest neighbor algorithm (KNN). KNN is one of the most common algorithms in artificial intelligence and is widely used in data mining and pattern recognition (Xiong & Yao, 2021). The KNN algorithm does not require training data, so it is the simplest supervised learning algorithm. It is based on the majority voting rule. It compares the Euclidean distance to determine the nearest K points in the attribute space and then categorizes the unknown samples into the most frequent category among the K points. The value of K represents the number of nearest neighbors to the unknown sample. The KNN algorithm has been used in the classification and regression prediction of edible crops. Choosing appropriate K values can improve accuracy, which is advantageous when applied to samples with little or no experience (Guo et al., 2022a; Kiran Naik et al., 2021). The KNN algorithm does not preprocess the data before classification, it takes time and space in the processing process, and it is difficult to select the value of K (Wang et al., 2022a). Even, the KNN algorithm ignores the correlation between sample features, resulting in high computational complexity and slow classification speed (Ji et al., 2022). Romero-del-Castillo et al. (2022) presented a new method for KNN optimization, which introduced a local K value into a multi-label label-KNN classifier to find the optimal K value.

3.2.2.6. CNN. CNN is considered the most popular deep learning structure, it is a deep feedforward neural network and a powerful feature extraction tool (Liu et al., 2021a). It is only sparsely connected to the neurons of the next layer according to the position of the neurons in the convolution layer, is different from ordinary neural networks, and relies on the convolution kernel to realize the mutual conversion of weights, which improves the learning time (Debus et al., 2021).

Residual neural network (ResNet) is one of the most advanced CNN models, which optimizes the problem of vanishing or exploding gradient in deep learning models through shortcut connection. At the same time, the residual module is introduced to improve the problem of model degradation caused by the increase in the number of layers in the CNN. Using shortcut connection has obvious advantages, neither adding extra parameters nor increasing computational costs (He et al., 2016). CNN and ResNet can classify edible crops by identifying two-dimensional spectral images or hyperspectral images. The above are only commonly used fields at present, and this algorithm is increasingly used for disease and pest detection or yield prediction.

3.2.2.7. Artificial neural network (ANN). ANN is a product of technological progress, which occupies a place with strong computing power and ultra-fast data processing speed. The mechanism by which the human brain nervous system processes complex information is the source of inspiration for generating ANN, where information is transformed and transmitted within neurons (Ng et al., 2019). ANN is interconnected by neurons to form complex neural networks, so the number of neurons in the hidden layer is important to ensure the model's accuracy and avoid overfitting (Wang et al., 2019). Benefiting from its powerful data processing capabilities, it can be used to solve the classification problem of edible crops and the regression problem. However, ANN relies excessively on large sample sizes and requires sufficient training data to maintain the stability of the model (Kholi et al., 2023). To improve the ANN model, Yuan et al. (2023) proposed a self-adjusting particle swarm optimization (APSO) algorithm to optimize the weights, thresholds, and number of neurons of the ANN.

4. Applications

In the context of economic globalization, edible crops have become one of the hot objects in the world, and their quality and safety issues have caused public concern. There are many kinds of edible crops, covering a wide range. The origin, species, content of main chemical components, storage period, and artificial adulteration are all crucial factors affecting their quality. Therefore, it is urgent to find a scientific and effective method to evaluate the quality of edible crops. Using machine learning to mine chemical information to represent quality has become one of the most popular methods at present. There are two main ways: One is to directly use multi-source data in combination with machine learning for evaluation, and the other is to process multi-source data through data fusion strategy and then conduct quality authentication with machine learning. Table 1 emphasized the application of machine learning in the quality evaluation of edible crops. Tables S2 and S3 summarized the applications of the above two methods in the quality evaluation of edible crops in detail.

4.1. Application of multi-source data to quality evaluation of edible crops

4.1.1. Qualitative perspective

Multi-source data is conducive to representing the quality of edible crops from different perspectives. Better quality evaluation methods can be obtained by comparing the results of machine learning modeling based on two or more single data sources. Amirvaresi et al. (2021) compared the ability of NIR and MIR to detect the geographical origin and adulteration of Iranian saffron. First of all, the PCA models of NIR and MIR spectra based on mean-centering and second derivative pretreatment were established to visually analyze the authentic saffron samples from different regions in Iran. Comparing the modeling results, it was found that the PCA based on NIR had a better ability to predict the distribution of saffron samples. Secondly, PLS-DA and PLSR models were established using NIR and MIR spectra after pretreatment to detect authentic saffron and adulterated samples. The qualitative analysis showed that the discrimination models based on NIR and MIR spectra had good classification performance, but the NIR showed more accurate and excellent identification results. Quantitative analysis also showed the same results, only NIR can effectively quantify the adulteration rate in authentic saffron samples. In conclusion, NIR spectroscopy can be used as a powerful tool for geographical traceability and adulteration detection of saffron. In a similar method, a series of machine learning algorithms named PCA, HCA, and PLS-DA were constructed with NIR and MIR spectra as input data to distinguish 33 kinds of hops for

Table 1
The application of machine learning in the quality evaluation of edible crops.

ML	Intention	Object	Data sources	Characteristics	Major results	Ref
HCA	A	Hop pellets	NIR, MIR	Both are rapid, non-destructive, low-cost, but MIR has a higher specificity than NIR	Based on NIR and MIR data, hop pellets were divided into two clusters, but the distribution of MIR-HCA was not as clear as NIR	(Machado et al., 2018)
	B	Black pepper	UHPLC-Q-Orbitrap-HRMS, ^{1}H NMR, GC-HRMS	High-throughput analysis, high sensitivity, and high resolution, and NMR-UHPLC-GC data fusion approach can reduce the dimensionality of ^{1}H NMR and improve accuracy	Based on UHPLC-Q-Orbitrap-HRMS HCA model, the clustering of samples from Brazil and Vietnam was close, and the clustering of sterilized samples and corresponding unsterilized samples was also very close	(Rivera-Perez et al., 2021)
	C	Rhizoma Coptidis	HPLC, FT-NIR, FT-IR	Spectra can not only can it be used for qualitative analysis, but it can also assist in quantitative analysis; Data fusion can obtain more comprehensive chemical information	Based on FT-NIR, when the distance was 23, the samples of *C.omeiensis* and *C. chinensis* are clustered into one class, and the distance was 24, which involves *C. deltoidea* samples; Based on FT-IR, *C. deltoidea* and *C. teeta* samples are clustered into one class; Based on data fusion, *C. deltoidea*, *C. chinensis* and *C. teeta* samples are clustered into one class when the distance was 12	(Qi et al., 2018)
PCA	A	Oolong tea	Gustatory sensors, olfactory sensors	Sensor technology is fast and accurate, and can make non-specific reactions to relevant chemical components, but only a single sensor technology cannot fully characterize samples	PCA was established based on taste sensor data and olfactory sensor data respectively, and the two showed similar results. The four types of samples showed a trend of separation but were not clear. Relatively speaking, the clustering effect of taste sensor data was better. The clustering trend of PCA models based on data fusion was also unsatisfactory	(Chen et al., 2015)
	B	Palm oil	HPLC-UV, HPLC-CAD	Both can provide sample information in a non-selective manner, and the fingerprint can serve as a complete analytical data	Using PCA to visualize samples of HPLC-CAD and HPLC-UV, two outliers were found in HPLC-CAD, while there were no outliers in HPLC-UV	(Obisesan et al., 2017)
	D	Saffron	NIR, MIR	Both are easy to operate, fast, and environmentally friendly, but they are selective, so in order to overcome their shortcomings, chemometrics is needed	NIR-PCA showed two trends in the distribution of saffron samples, and MIR-PCA showed no significant distribution trend compared to NIR	(Amirvaresi et al., 2021)
PLSR	D	Olive oil	NIR, MIR	–	NIR: $R^2 = 0.896$, RMSEP = 7.09; MIR: $R^2 = 0.966$, RMSEP = 4.04; LLF: $R^2 = 0.975$, RMSEP = 3.44; HLF: $R^2 = 0.988$, RMSEP = 2.86 (Best)	(Li et al., 2019)
	E	*Ziziphus jujuba*	NIR, MIR	–	NIR: $R^2 = 0.9312$, RPD = 2.82 MIR: $R^2 = 0.8951$, RPD = 2.28 LLF: $R^2 = 0.9475$, RPD = 2.10 MLF: $R^2 = 0.9621$, RPD = 2.44 (Best)	(Arslan et al., 2019)
	E	Cottonseed	NIR, GC–MS	NIR is high-throughput, simple and low-cost	R^2cal > 0.7	(Zhuang et al., 2023)
SVR	E	Yuezhou Longjing tea	NIR, HPLC	NIR has the advantages of non-destructive testing, fast testing speed, and high efficiency	Sensory quality: RPD(PLSR) = 1.888,RPD (RF) = 2.033, RPD(SVR) = 2.485 (Best); Catechins: RPD(PLSR) = 1.857, RPD(SVR) = 2.088, RPD(RF) = 2.584 (Best); Caffeine: RPD(PLSR) = 2.076, RPD(SVR) = 2.799, RPD(RF) = 2.873 (Best)	(Jia et al., 2022)
	E	*Ginkgo biloba* leaf extract	NIR, HPLC	Due to the characteristics of weak absorption peaks and wide peaks in NIR, obtaining reliable information requires the use of chemometrics	PLSR: $R^2 > 0.95$, RESECV < 0.30; SVR: $R^2 > 0.96$, RMSECV < 0.50	(Zhang et al., 2022b)
	E	Red jujube	NIR, HPLC	NIR has the advantages of being fast, simple, and environmentally friendly	PLSR: R^2c = 0.9076, RMSEC = 25.2625, R^2p = 0.8323, RMSEP = 29.0407; SVR: R^2c = 0.9850, RMSEC = 11.1233, R^2p = 0.9388, RMSEP = 13.0739 (Best)	(Chen et al., 2019)
PLS-DA	B	*Amomum tsao-ko*	FT-NIR, UV–Vis	Both have the characteristics of low cost, speed, and convenience, but the information they can provide is limited	MLF: Acc = 100%	(Liu et al., 2021b)
	B	Cocoa bean shells	ATR-FTIR, NIR, ICP-OES	The process of collecting spectral information is simple and has chemical specificity	MLF: Acc = 0.84	(Mandrile et al., 2019)
	D	Vanilla	NIR, MIR, Raman	Vibration spectroscopy has the advantages of high sensitivity, ease of use, and low cost	PLS-DA: SEN (NIR) = 0.82, SPE (NIR) = 0.72, PRE (NIR) = 0.76; Acc (Raman) = 0.9, SEN (Raman, MIR) = 1, SPE (Raman, MIR) = 1, EFF (Raman, MIR) = 1, PRE (Raman, MIR) = 1 (Best)	(Jimenez-Carvelo et al., 2021)

(*continued on next page*)

Table 1 (*continued*)

ML	Intention	Object	Data sources	Characteristics	Major results	Ref
SVM	A	Cocoa bean	NIR, E-tongue	Both have the advantages of speed and simplicity.	NIR, ET: Acc = 83%-93%; MLF: Acc = 100% (Best)	(Teye et al., 2014)
	C	Black tea	FT-NIR, CVS		NIR: Acc (LDA) = 86.30%-89.19%, Acc (KNN) = 89.19%-90.41%, Acc (SVM) = 89.19%-97.26%;CVS: Acc (LDA) = 89.19%-91.78%, Acc (KNN) = 65.75%-89.19%, Acc (SVM) = 89.19%-97.26%; MLF: Acc (LDA) = 91.89%-98.63%, Acc (KNN) = 75.68%-91.89%, Acc (SVM) = 100% (Best)	(Jin et al., 2020)
	D	Extra virgin olive oil	LC-MS, GC-IMS, FGC-Enose	–	LC-(+/-)MS: Acc = 0.94, SEN = 0.93, SPE = 0.95, AUC = 0.97;LC-(+/-) MS + GC-IMS + FGC-Enose: Acc = 0.96, SEN = 0.93, SPE = 0.96, AUC = 0.98; GC-IMS + FGC-Enose: Acc = 0.96, SEN = 0.93, SPE = 0.97, AUC = 0.99 (Best)	(Tata et al., 2022)
RF	B	*Panax notoginseng*	FT-IR, NIR	Spectroscopic technique has the advantages of simplicity, speed, and ease of use, the disadvantage is that a single Spectroscopic technique expresses limited chemical information	FT-IR: Acc = 91.2%; NIR: Acc = 92.6%; LLF: Acc = 95.6%; MLF: Acc (RF-Vs) = 94.1%, Acc (RF-Bo) = 97.1%; HLF: Acc (RF-Vs) = 97.1%, Acc (RF-Bo) = 95.6% (Best)	(Zhou et al., 2020)
	B	Radix Astragali	LIBS, MIR	LIBS has the advantages of fast, real-time, and no complex preprocessing, and infrared spectroscopy can be used to obtain molecular vibration information	LIBS: SEN = 0.9411, SPE = 0.9716, Acc = 0.9624, TIME = 32.2 s; MIR: SEN = 0.9722, SPE = 0.9864, Acc = 0.9829, TIME = 8.4 s; LLF: SEN = 0.9889, SPE = 0.9948, Acc = 0.9932, TIME = 35.7 s; MLF: SEN = 0.9900, SPE = 0.9951, Acc = 0.9932, TIME = 6.9 s (Best)	(Wang et al., 2022b)
	B	*Eucommia ulmoides* leaves	FT-NIR, ATR-FTIR	Spectroscopic technique can reflect the overall chemical profile of a sample, but it cannot perform quantitative analysis	LLF: Acc (calibration) = 85.71%;MLF: Acc (calibration) = 81.75%, Acc (validation) = 88.52%;HLF: Acc (calibration) = 92.86%, Acc (validation) = 93.44% (Best)	(Wang et al., 2020)
SIMCA	D	*Uncaria tomentosa*, *Uncaria guianensis*	LC-PDA, FT-IR, UV	–	FT-IR: SEN = 100%, SPE = 100%; UV: SEN = 100%, SPE = 100%; UV + KOH: SEN = 100%, SPE = 100%; UV + $AlCl_3$: SEN = 100%, SPE = 100%;LC (PPH) : SEN = 100%, SPE = 100%	(Kaiser et al., 2020)
	A	Olive oil	HPLC-DAD, HPLC-FID	–	HPLC-DAD, HPLC-FID: Acc > 94.59%; LLF: Acc = 100% (Best); MLF: Acc = 97.30%	(Bajoub et al., 2017)
	A	Rhubarb	NIR, MIR	–	NIR: Acc (PLS-DA) = 94.12%, Acc (SIMCA) = 82.35%, Acc (SVM) = 94.12%, Acc (ANN) = 100%; MIR: Acc (PLS-DA) = 82.35%, Acc (SIMCA) = 82.35%, Acc (SVM) = 94.12%, Acc (ANN) = 76.47%%; LLF: Acc (PLS-DA) = 94.12%, Acc (SIMCA) = 88.24%, Acc (SVM) = 94.12%, Acc (ANN) = 100%; MLF (Best): Acc (PLS-DA) = 94.12%, Acc (SIMCA) = 94.12%, Acc (SVM) = 100%, Acc (ANN) = 100%;	(Sun et al., 2017)
KNN	D	Guava pulp	NIR, MIR	Infrared spectroscopy has the advantages of non-destructive, efficient, and cost-effective analysis, and allows for the simultaneous analysis of multiple types of chemical components	NIR: Acc = 100%, EFF = 97.5%-99.5%; MIR: Acc = 98.9%-100%, EFF = 99.1%-100% (Best); LLF: Acc = 95.1%-100%, EFF = 93.3%-100%;	(Alamar et al., 2020)
	A	Curcumae kwangsiensis, Curcumae phaeocaulis, Curcumae wenyujin	HPLC, HS-GC–MS	HPLC has good separation and detection capabilities, while HS-GC–MS mainly focuses on the identification and quantification of volatile components without the need for standard substances	HPLC: Acc (LDA) = 90.91%, Acc (KNN) = 100%, Acc (BPNN) = 100%; HPLC, HS-GC–MS: Acc (LDA) = 100%, Acc (KNN) = 100%, Acc (ANN) = 100%, Acc (OPLS-DA) = 100%	(Wang et al., 2021a)
	B	Herba Epimedii	NIR, HPLC	NIR has significant advantages in detection speed, accuracy, and cost	Acc (KNN) = 90.74%, Acc (DA) = 79.63%, Acc (BPNN) = 87.04%, Acc (SVM) = 94.44%	(Yang et al., 2018)
CNN	B	*Gentiana rigescens*	FT-IR, HPLC	Spectroscopic technique can obtain complete chemical profiles and is widely used in the field of quality evaluation	Acc (synchronous 2DCOS) = 100%	(Liu et al., 2022a)
	B	Wolfberries	Vis-NIR-HSI, textural data	Spectroscopic technique has the advantages of simple operation, fast speed, non-destructive and low-cost	Full spectral wavelengths: Acc = 95.21%, mean F1 = 95.17%; MLF: Acc = 97.34%, mean F1 = 100%	(Hao et al., 2022)
	B	*Panax notoginseng*	NIR, HPLC	NIR has the advantages of green, pollution-free, fast detection speed, and simplicity, but its original spectral signals are prone to	Acc (synchronous 2DCOS) = 100%	zhuomian

(*continued on next page*)

Table 1 (*continued*)

ML	Intention	Object	Data sources	Characteristics	Major results	Ref
				overlap, making available information limited.		
ANN	E	Lonicerae Japonicae Flos	NIR, HPLC	–	Chlorogenic acid: RMSEP (PLSR) = 1.15, R (PLSR) = 0.9940 (Best), RMSEP (ANN) = 1.99, R (ANN) = 0.9842; Isochlorogenic acid A: RMSEP (PLSR) = 0.93, R (PLSR) = 0.9892 (Best), RMSEP (ANN) = 1.05, R (ANN) = 0.9862; Isochlorogenic acid C: RMSEP (PLSR) = 0.27, R (PLSR) = 0.9692, RMSEP (ANN) = 0.18, R (ANN) = 0.9868 (Best)	(Xue et al., 2021)
	D	Olive oil	FT-IR, Vis-NIR, EEMs	Spectroscopic technique can provide chemical fingerprints of samples, not limited to a specific component	PLS-DA: Acc (FT-IR, Vis-NIR) = 100% (Best); BPNN: Acc (EEMs) = 100% (Best)	(Meng et al., 2023)
	C	*Flos Chrysanthemi*	NIR, HPLC-qTOF-MS	NIR has the advantages of simple operation, fast speed, and low-cost	R = 0.89	(Ding et al., 2016)

"-": no mention; A: identification variety; B: geographical traceability; C: quality control; D: adulteration detection; E: content prediction; HCA: hierarchical cluster analysis; PCA: principal component analysis; PLSR: partial least squares regression; SVR: support vector regression; PLS-DA: partial least squares-discriminant analysis; SVM: support vector machine; RF: random forest; SIMCA: soft independent modeling of class analogy; KNN: K-nearest neighbors; CNN: convolutional neural network; ANN: artificial neural network NIR: near infrared; MIR: mid infrared; UHPLC-Q-Orbitrap-HRMS: ultra-high performance liquid chromatography-quadrupole-Orbitrap-high-resolution mass spectrometry; ^{1}H NMR: proton nuclear magnetic resonance; GC-HRMS: gas chromatography-high-resolution mass spectrometry; HPLC: high performance liquid chromatography; FT-NIR: Fourier transform-near infrared spectroscopy; FT-IR: Fourier transform mid-infrared spectroscopy; HPLC-UV: high performance liquid chromatography-ultraviolet; HPLC-CAD: high performance liquid chromatography-charged aerosol; GC–MS: gas chromatography-mass spectrometry; UV–Vis: ultraviolet–visible; ATR-FTIR: attenuated total reflectance-Fourier transform mid-infrared spectroscopy; ICP-OES: inductively coupled plasma-optical emission spectroscopy; CVS: computer vision system; LC-MS: liquid chromatography-high resolution mass spectrometry; GC-IMS: gas chromatography-ion mobility spectrometry; FGC-Enose: flash gas-chromatography electronic nose; LIBS: laser induced breakdown spectroscopy; LC-PDA: liquid chromatography- photo diode array; HPLC-DAD: high-performance liquid chromatography-diode array detector; HPLC-FID: high-performance liquid chromatography-flame ionization detector; HS-GC–MS: headspace gas chromatography-mass spectrometry; Vis-NIR-HIS: visible-near infrared-hyperspectral imaging; EEMs: excitation-emission matrix fluorescence spectroscopy; HPLC-qTOF-MS: high liquid chromatography-quadrupole-time of flight-mass spectrometry.

commercial purposes (Machado et al., 2018). The research showed that the accuracy of NIR and MIR modeling after feature band extraction and standard normal variate (SNV) and combined Savitzky-Golay filter smoothing pretreatment was 96.6% and 94.2%, respectively, which proved that NIR and MIR spectroscopy could be used as a green, convenient and non-destructive method to identify commercial hop varieties. The quality of white teas mainly depends on the maturity of fresh leaves, Li et al. (2020a) collected the chemical fingerprint information (NIR and HPLC) of white tea samples produced from fresh leaves with different maturity. The results showed that the clustering effect of PCA based on the concentration of 39 compounds was not satisfactory, while the PCA based on NIR could quickly identify white tea samples with different maturity, which illustrated the advantages of the NIR technique applied to the quality evaluation of white tea.

4.1.2. Quantitative perspective

The combination of data from different sources is also one of the popular methods in the field of edible crop quality evaluation, mainly used for the content prediction and analysis of chemical components. Cottonseed is one of the important oilseed crops, and fatty acid and protein are important quality evaluation indicators. Cottonseed is one of the important oil-bearing crops, and its important quality evaluation indicators are fatty acid and protein. Zhuang et al. (2023) obtained the phenotypic data of 17 fatty acids, oil, and proteins in shell-intact upland cottonseed using GC–MS and Soxhlet extraction methods, and correlated the content with the preprocessed NIR datasets through PLSR. The NIR spectral region of 950–1650 nm was selected as the model input data because fatty acids and proteins have strong absorption peaks in this range. The report demonstrated that the established method could accurately predict the content of 14 fatty acids, oils, and proteins. Unfortunately, the calibration model performance of three fatty acids still needs to be improved. This study laid a foundation for the quality evaluation of cottonseed in practical application. The linear regression method-PLSR was used in the above study, but in some special cases, the linear regression method is not applicable. The development of nonlinear regression methods such as ANN provides a solution to this problem. Xue et al. (2021) used HPLC to detect the contents of three active components in Lonicerae Japonicae Flos and combined with NIR, and implemented two types of calibration models, namely PLSR and ANN, to predict chlorogenic acid, isochlorogenic acid A and isochlorogenic acid C in Lonicerae Japonicae Flos. The results showed that in the PLSR model, the best pretreatment methods of near-infrared spectra of chlorogenic acid, isochlorogenic acid A, and isochlorogenic acid C were first derivative, first derivative + straight line subtraction (SLS), and first derivative + vector normalization (VN), respectively. At the same time, the selection of spectral region greatly affects the results of the model. The spectral region suitable for the prediction of chlorogenic acid was 12000–4250 cm^{-1}, and that of isochlorogenic acid A and isochlorogenic acid C was 7500–4250 cm^{-1}. In the ANN model, the best spectral pretreatment of chlorogenic acid was still the first derivative and the best pretreatment of isochlorogenic acid A and isochlorogenic acid C were SLS and VN, respectively. Comparing the results of six calibration models, it is found that the model suitable for the prediction of chlorogenic acid and isochlorogenic acid A content was PLSR, and ANN was the best strategy for the prediction of isochlorogenic acid C. Proof by facts, NIR spectroscopy is an effective technique for quality evaluation of Lonicerae Japonicae Flos.

To sum up, spectral pretreatment and selection of characteristic regions are very important, which lays a solid foundation for establishing a robust model and obtaining accurate results. More accurate inference is generated through information complementation between multi-source data than that of a single data source. However, analysis efficiency is a concern due to the huge amount of information. The proposed data fusion strategy not only simplifies the process of multi-source data processing and improves the efficiency of data analysis but also can obtain more complete and unified information, which is conducive to strengthening the robustness of decision-making.

4.2. Application of data fusion strategy in edible crops

4.2.1. Identification variety

There are many varieties of edible crops, and their appearance and

morphology are similar, but there are differences in chemical components, sensory, and other aspects. Therefore, in order to protect the rights and interests of consumers and market order, variety identification is an essential step. Teye et al. (2014) studied the feasibility of data fusion of NIR spectroscopy and electronic tongue for distinguishing five cocoa bean varieties. After SNV pretreatment, the spectral area of 9500–7500 cm^{-1} was selected as the modeling data for NIR, the data of the two sensors were combined through PCA, and the best variables were selected. Using a single data source and fused data as input to establish the SVM classification model, the results showed that the classification accuracy of SVM based on electronic tongue and NIR was 92%–93% and 80%–81%, respectively, while the performance of the SVM model for data fusion was greatly improved, and the recognition accuracy was 100%. Dankowska & Kowalewski (2019) collected UV–Vis, synchronous fluorescence (SF), and NIR information on different types of tea and carried out low-level data fusion on different data combinations (SF + UV–Vis, NIR + UV–Vis, SF + NIR, and SF + UV–Vis + NIR). Used PCA to realize data dimensionality reduction and then built Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Regularized Discriminant Analysis (RDA), and SVM models for the single data source and four low-level data fusion. By comparing and analyzing the modeling results, the classification effect of data fusion was significantly better than that of a single data source, with an error of less than 3%. The research has proved that different spectral information can complement each other, and improve classification accuracy, and the developed method can effectively prevent tea fraud. In another study, the gustatory and olfactory sensor systems were combined with machine learning (PCA, LDA) to quickly identify different varieties of Oolong tea (Chen et al., 2015). The feature extraction of the data points of the Cyclic voltammetry in the gustatory sensors system and the average of the center of each dye point in the olfactory sensors system were performed. The mid-level data fusion was performed after the above operations were completed. The author found that it is difficult to accurately evaluate different kinds of Oolong tea by using an olfactory or gustatory sensor system alone. The use of a data fusion strategy can provide more comprehensive information and improve classification accuracy. The accuracy of the LDA model based on data fusion reached 100%. The results indicated that the data fusion strategy has a promising prospect in classifying Oolong tea varieties.

4.2.2. Geographical traceability

The natural environment is an important factor affecting edible crops' quality, resulting in uneven product quality from different geographical origins. The proposal of geographical indication protection makes the public pay more attention to the source of products. Mandrile et al. (2019) identified cocoa shell samples from different geographical origins by fusing data from NIR, ATR-FT-IR, and inductively coupled plasma-optical emission spectroscopy (ICP-OES) and combining the PLS-DA model. Compared with the modeling results of a single data source, the model effect of data fusion was more satisfactory, and the classification accuracy was higher. In order to identify palm oil from different geographical sources, a data fusion strategy and PLS-DA were used to carry out research (Obisesan et al., 2017). Feature variables were extracted through interval partial least squares (iPLS) and PCA, respectively, to implement further low- and high-level data fusion strategies for two groups (HPLC-UV, HPLC-CAD) of data sources. The results showed that the data fusion strategy significantly improved the classification accuracy of a single data source, and the accuracy of iPLS as a feature variable extraction method was 100%. *Amomum tsao-ko* is easily affected by the geographical environment, resulting in different qualities of different origins. Liu et al. (2021b) used the mid-level data fusion strategy to integrate the information of FT-NIR and UV–Vis to identify the origin of *Amomum tsao-ko*, adopted four methods of feature variable extraction: PCA, VIP, sequential and orthogonalized partial-least squares (SO-PLS), sequential and orthogonalized covariance selection (SO-CovSel). Spectral preprocessing was a combination of S-G, variables sorting for normalization (VSN), and first derivative. PLS-DA models based on four feature extraction methods were established respectively. The results showed that the classification ability of the model was excellent when so-pls was used as the feature extraction method of data fusion, which laid a theoretical foundation for the geographical traceability and even the quality and safety of *Amomum tsao-ko*.

4.2.3. Quality control

In addition to natural factors, the quality of edible crops is also affected by many factors, such as product processing, cultivation methods, and storage period. Therefore, it is very necessary to find a fast and scientific quality control method applied to the industrial chain of edible crops. In the existing research, data fusion combined with machine learning has been widely used as a powerful tool for quality control. The quality of black tea is affected by the degree of fermentation, and changes in the content of tea polyphenols will accompany it. For this reason, tea polyphenols content is an important indicator to determine the classification of different degrees of fermentation of black tea. FT-NIR, computer vision system, and mid-level data fusion were employed to detect the fermentation degree of black tea so as to achieve the purpose of quality control (Jin et al., 2020). Feature extraction is an important part of mid-level data fusion. This study implemented two feature extraction strategies, one was to use PCA to extract the features of FT-NIR and computer vision system, and the other was to extract the features of FT-NIR and computer vision system, respectively, through SPA and Pearson correlation for subsequent research. KNN, LDA, and SVM were used to analyze the above data sets. The results obtained by the SVM model of data fusion were better than that of single data source and other data fusion models. Among them, the best feature extraction method was PCA.

4.2.4. Adulteration detection

In order to obtain greater benefits, illegal businesses fill the gap in the market by adulterating. The adulterants usually do not cause changes in taste and chemical composition, which makes it difficult to identify with the naked eye. It is urgent to adopt reliable and convenient methods to detect adulteration in edible crops. Vibration spectroscopy is widely used in adulterating olive oil due to its non-destructive, low-cost, and fast advantages. Li et al. (2019) applied three data fusion strategies to integrate NIR and MIR and combined them with the PLSR model to quantify adulterants in olive oil. Taking RESEP as the model performance evaluation index, it can be concluded that the prediction accuracy of low-level (3.44) and high-level data fusion (2.86) is higher than that of NIR (7.09), MIR (4.04), and mid-level data fusion (6.09), respectively. It is worth pondering that SPA, as a feature extraction method for intermediate data fusion, did not improve the model's performance, which showed the importance of selecting appropriate feature extraction for fusion results. Tata et al. (2022) used completely different techniques from the above research to detect the adulteration of olive oil. Performed low and mid-level data fusion for the four data sets of LC – (+/−) MS, gas-chromatography ion mobility spectrometry (GC-IMS), and flash gas-chromatography electronic nose (FGC-E-nose), and then analyzed the fused data using PLS-DA-SVM. Low-level data fusion based on GC-IMS and FGC-E-nose has the highest classification accuracy (0.96) and may provide a new way to detect olive oil adulteration. The results of both studies demonstrated that a data fusion strategy combined with machine learning has a bright future in adulteration detection.

4.2.5. Content prediction

Chemical composition is the key index to measure the quality of edible crops. Accurate evaluation of chemical composition content is one of the ways to improve the quality of edible crops. The correlation between spectroscopy and chemical component content is one of the most popular content prediction methods at present. MIR and NIR data

fusion was proposed to predict the polyphenol content in Chinese dates (Arslan et al., 2019). Pre-processed NIR and MIR with SNV and detrending, and used Si-PLS to extract their characteristic variables to establish the PLS model of two fusion strategies (NIR-MIR fusion, GA-fusion). GA-fusion-PLS model has the best prediction ability, with an RPD of 2.44. Polyphenol and catechin undergo different degrees of oxidation with the change in the processing degree of black tea. Wang et al. (2021b) collected the data of NIR and computer vision of black tea in different processing steps on the spot, and used low- and mid-level data fusion strategies to jointly analyze them. It is worth noting that in order to improve the performance of the calibration model, the following two operations were carried out before modeling: (1) Pearson correlation analysis and CARS were used as the feature variable extraction methods of computer vision and NIR respectively; (2) Select S-G smoothing as the pre-processing method of NIR. Comparing the results of PLS models based on a single data source and two data fusion strategies, it was concluded that mid-level data fusion could make up for the deficiency of single data and improve the prediction accuracy of the model. Unfortunately, low-level data fusion cannot overcome the low prediction accuracy of single data. In general, the combination of spectroscopy and imaging systems can be used as an effective content prediction method.

5. Challenges and prospects

The combination of multi-source data and machine learning has been widely used in edible crops, and many studies have proven that they are powerful quality evaluation tools with promising application prospects. However, it has to be acknowledged that there are still many undiscovered development opportunities for this method that are worth further exploration. Multi-source data, machine learning, and applications are the most important components of this review, and their challenges and future prospects will be discussed in this section.

(1) Multi-source data. Every step, from data acquisition to multi-source data processing, is crucial. Spectroscopy, chromatography, and new sensor systems are the leading techniques for obtaining multi-source data. First of all, these instruments and equipment are suitable for analysis in specific locations, which could be more conducive to real-time detection on the market. The quality evaluation of edible crops requires more convenient analytical instruments, and the emergence of a portable NIR spectrometer is a good start. However, it is regrettable that its accuracy cannot meet the desktop standard, and infrared detection has special requirements for illumination, moisture, etc., which limits its application in the market. The development of portable instruments is a prerequisite for further market use while improving the accuracy of instrument testing is a problem that must be addressed in development. Sensor systems are non-destructive and intelligent. Currently, sensor systems have been used as a substitute for traditional analytical techniques, and they are a class of techniques with great development space. They exhibit excellent performance in qualitative analysis, and accurate quantitative analysis is still a problem that needs to be tackled by such techniques. Quantitative analysis of edible crops relies on chromatographic techniques, but their use runs counter to the concept of green environmental protection. The development of environmentally friendly reagents is particularly important. Secondly, the type of data combination is also a way to improve the accuracy of results. For example, the results obtained by combining different data sources are inconsistent (Dankowska & Kowalewski, 2019). Wang et al. (2023) also found that combining Raman and NIR spectroscopy can improve prediction accuracy. Therefore, selecting suitable complementary sources as data combinations can obtain more complete information and improve the accuracy and stability of data analysis. However, the large amount of information in multi-source data makes data processing difficult, and the fusion strategy is an effective means of data processing. Many studies have shown that fusion strategies improve the performance of classification or regression models by integrating multi-sources of information to obtain more comprehensive information. High-level data fusion applications in this area are not as good as expected, and only a few have met the requirements. However, high-level data fusion still has great potential in the application of multi-source data for edible crops, so it is necessary to further explore it in order to improve accuracy and robustness. Last but not least, both multi-source data and data fusion have factors such as redundant data or noise that affect the model's performance, pre-processing and feature extraction are the most critical steps in data processing accordingly. Notably, each pre-processing method solves different problems, and selecting appropriate pre-processing based on the attributes of the data can achieve the goal of optimization. Similarly, selecting feature extraction methods should also be based on multiple perspectives. Currently, there are no reasonable solutions to the determination of pre-processing combinations and multi-dimensional data feature extraction, and further in-depth research is needed to address these issues.

(2) Machine learning and applications. Although unsupervised learning algorithms can reduce time costs, their accuracy, and robustness are not as good as supervised learning methods. Supervised learning methods are limited to a certain extent by manual tagging, which requires a greater amount of manual tagging when applied to more complex systems. Both are subject to factors such as sample size or hyperparameters. With the development of artificial intelligence, neural networks such as ANN, CNN, and ResNet are gradually applied to the research of edible crops due to their powerful performance. They require large amounts of data to support their training process. Most existing research is based on small sample sizes, which poses a significant challenge to neural networks. Hence, selecting machine learning based on the sample size and the problem to be solved is conducive to obtaining better decision results. It is necessary to optimize unsupervised further and supervised learning algorithms or develop a more powerful machine learning algorithm to address the complex and volatile requirements of the field of edible crop research. In the investigated papers, the application of machine learning combined with multi-source data or data fusion strategies in edible crops mainly focuses on: Quality control, content prediction, and geographical identification. It can be seen that this method also has a certain potential in other research fields of edible crops and needs further investigation.

6. Conclusions

The research on edible crops has been a hot topic in food and agriculture. Although the existing research has made some progress in the quality evaluation of edible crops, it is difficult to characterize the quality with a single analytical technique fully. An accurate, fast, reliable, and robust quality evaluation method is urgently needed. Machine learning is a gift of rapid development in the era of science and technology. Using it to process and analyze multi-source data has great potential in the application of edible crops. This review summarized the recent application of machine learning combined with multi-source data in edible crops. In addition, the limitations and future application prospects of the method were also discussed. Multi-source data is not a supplement to a single data source. It can absorb different data information characteristics, obtaining more accurate and reliable results than a single data source. Data fusion is a commonly used strategy for processing multi-source data, which can extract valuable and accurate target information from complementary data. Low- and mid-level data

fusion is favored in classification and regression applications, especially mid-level data fusion, which has unique advantages in filtering noisy data and reducing data dimensionality. Moreover, machine learning is a very critical part of the process of edible crop quality evaluation, and selective use based on data attributes and purposes can achieve ideal results. Data pre-processing and feature extraction are also one of the focuses and difficulties of attention. The above four parts (multi-source data, data fusion, machine learning, pre-processing, and feature extraction) all require further research to break through bottlenecks. This review can provide constructive suggestions for the quality evaluation methods of edible crops so as to improve the applicability of market monitoring.

Declaration of Competing Interest

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

Data availability

No data was used for the research described in the article.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant number: U2202213), and the supported by Special Program for the Major Science and Technology Projects of Yunnan Province, China (Grant numbers: 202102AE090051-1-01, 202202AE090001).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.fochx.2023.100860.

References

Abbas, O., Zadravec, M., Baeten, V., Mikus, T., Lesic, T., Vulic, A., et al. (2018). Analytical methods used for the authentication of food of animal origin. *Food Chemistry, 246*, 6–17. https://doi.org/10.1016/j.foodchem.2017.11.007

Abdel Razeq, S. A., Abdel Aziz, S. E., & Ahmed, N. S. (2021). Stability-indicating UPLC, TLC-densitometric and UV-spectrophotometric methods for alcaftadine determination. *Journal of Chromatography B-Analytical Technologies in the Biomedical and Life Sciences, 1177*, Article 122804. https://doi.org/10.1016/j.jchromb.2021.122804

Adenan, M. N. H., Moosa, S., Muhammad, S. A., Abrahim, A., Jandric, Z., Islam, M., et al. (2020). Screening Malaysian edible bird's nests for structural adulterants and geographical origin using Mid-Infrared – Attenuated Total Reflectance (MIR-ATR) spectroscopy combined with chemometric analysis by Data-Driven – Soft Independent Modelling of Class Analogy (DD-SIMCA). *Forensic Chemistry, 17*, Article 100197. https://doi.org/10.1016/j.forc.2019.100197

Aidoo, E. N., Appiah, S. K., Awashie, G. E., Boateng, A., & Darko, G. (2021). Geographically weighted principal component analysis for characterising the spatial heterogeneity and connectivity of soil heavy metals in Kumasi, Ghana. *Heliyon, 7*(9), e08039.

Alamar, P. D., Caramês, E. T. S., Poppi, R. J., & Pallone, J. A. L. (2020). Detection of fruit pulp adulteration using multivariate analysis: Comparison of NIR, MIR and data fusion performance. *Food Analytical Methods, 13*(6), 1357–1365. https://doi.org/10.1007/s12161-020-01755-x

Allakhverdiev, E. S., Khabatova, V. V., Kossalbayev, B. D., Zadneprovskaya, E. V., Rodnenkov, O. V., Martynyuk, T. V., et al. (2022). Raman spectroscopy and its modifications applied to biological and medical research. *Cells, 11*(3), 386. https://doi.org/10.3390/cells11030386

Amirvaresi, A., Nikounezhad, N., Amirahmadi, M., Daraei, B., & Parastar, H. (2021). Comparison of near-infrared (NIR) and mid-infrared (MIR) spectroscopy based on chemometrics for saffron authentication and adulteration detection. *Food Chemistry, 344*, Article 128647. https://doi.org/10.1016/j.foodchem.2020.128647

Arslan, M., Zareef, M., Tahir, H. E., Guo, Z., Rakha, A., He, X. T., et al. (2022). Discrimination of rice varieties using smartphone-based colorimetric sensor arrays and gas chromatography techniques. *Food Chemistry, 368*, Article 130783. https://doi.org/10.1016/j.foodchem.2021.130783

Arslan, M., Zou, X. B., Tahir, H. E., Zareef, M., Hu, X. T., & Rakha, A. (2019). Total polyphenol quantitation using integrated NIR and MIR spectroscopy: A case study of Chinese dates (Ziziphus jujuba). *Phytochemical Analysis, 30*(3), 357–363. https://doi.org/10.1002/pca.2818

Ay, M., Özbakır, L., Kulluk, S., Gülmez, B., Öztürk, G., & Özer, S. (2023). FC-Kmeans: Fixed-centered K-means algorithm. *Expert Systems With Applications, 211*, Article 118656. https://doi.org/10.1016/j.eswa.2022.118656

Bajoub, A., Medina-Rodriguez, S., Gomez-Romero, M., Ajal el, A., Bagur-Gonzalez, M. G., Fernandez-Gutierrez, A., et al. (2017). Assessing the varietal origin of extra-virgin olive oil using liquid chromatography fingerprints of phenolic compound, data fusion and chemometrics. *Food Chemistry, 215*, 245–255. https://doi.org/10.1016/j.foodchem.2016.07.140

Borras, E., Ferre, J., Boque, R., Mestres, M., Acena, L., & Busto, O. (2015). Data fusion methodologies for food and beverage authentication and quality assessment - A review. *Analytica Chimica Acta, 891*, 1–14. https://doi.org/10.1016/j.aca.2015.04.042

Boubchir, M., Boubchir, R., & Aourag, H. (2022). The Principal Component Analysis as a tool for predicting the mechanical properties of Perovskites and Inverse Perovskites. *Chemical Physics Letters, 798*, Article 139615. https://doi.org/10.1016/j.cplett.2022.139615

Buchaiah, S., & Shakya, P. (2022). Bearing fault diagnosis and prognosis using data fusion based feature extraction and feature selection. *Measurement, 188*, Article 110506. https://doi.org/10.1016/j.measurement.2021.110506

Cao, R. G., Liu, X. R., Liu, Y. Q., Zhai, X. Q., Cao, T. Y., Wang, A. L., et al. (2021a). Applications of nuclear magnetic resonance spectroscopy to the evaluation of complex food constituents. *Food Chemistry, 342*, Article 128258. https://doi.org/10.1016/j.foodchem.2020.128258

Cao, Z. Y., Li, X. R., Feng, Y. M., Chen, S. H., Xia, C. Q., & Zhao, L. (2021b). ContrastNet: Unsupervised feature learning by autoencoder and prototypical contrastive learning for hyperspectral imagery classification. *Neurocomputing, 460*, 71–83. https://doi.org/10.1016/j.neucom.2021.07.015

Chen, C., Li, H. Y., Lv, X. Y., Tang, J., Chen, C., & Zheng, X. X. (2019). Application of near infrared spectroscopy combined with SVR algorithm in rapid detection of cAMP content in red jujube. *Optik, 194*, Article 163063. https://doi.org/10.1016/j.ijleo.2019.163063

Chen, Q. S., Sun, C. C., Ouyang, Q., Wang, Y. X., Liu, A. P., Li, H. H., et al. (2015). Classification of different varieties of Oolong tea using novel artificial sensing tools and data fusion. *LWT-Food Science and Technology, 60*(2), 781–787. https://doi.org/10.1016/j.lwt.2014.10.017

Chen, Z. X., Zeng, J. F., He, M. H., Zhu, X. S., & Shi, Y. W. (2022). Portable ppb-level carbon dioxide sensor based on flexible hollow waveguide cell and mid-infrared spectroscopy. *Sensors and Actuators B-Chemical, 359*, Article 131553. https://doi.org/10.1016/j.snb.2022.131553

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*, 273–297. https://doi.org/10.1007/BF00994018

Dankowska, A., & Kowalewski, W. (2019). Tea types classification with data fusion of UV-Vis, synchronous fluorescence and NIR spectroscopies and chemometric analysis. *Spectrochimica Acta Part A-Molecular and Biomolecular Spectroscopy, 211*, 195–202. https://doi.org/10.1016/j.saa.2018.11.063

De Marchi, M., Toffanin, V., Cassandro, M., & Penasa, M. (2014). Invited review: Mid-infrared spectroscopy as phenotyping tool for milk traits. *Journal of Dairy Science, 97*(3), 1171–1186. https://doi.org/10.3168/jds.2013-6799

Debus, B., Parastar, H., Harrington, P., & Kirsanov, D. (2021). Deep learning in analytical chemistry. *TrAC-Trends in Analytical Chemistry, 145*, Article 116459. https://doi.org/10.1016/j.trac.2021.116459

Delpeuch, C., & Leblois, A. (2014). The elusive quest for supply response to cash-crop market reforms in sub-Saharan Africa: The case of cotton. *World Development, 64*, 521–537. https://doi.org/10.1016/j.worlddev.2014.06.007

Ding, G. Y., Li, B. Q., Han, Y. Q., Liu, A. N., Zhang, J. R., Peng, J. M., et al. (2016). A rapid integrated bioactivity evaluation system based on near-infrared spectroscopy for quality control of *Flos Chrysanthemi*. *Journal of Pharmaceutical and Biomedical Analysis, 131*, 391–399. https://doi.org/10.1016/j.jpba.2016.09.008

Duca, D., Mancini, M., Rossini, G., Mengarelli, C., Foppa Pedretti, E., Toscano, G., et al. (2016). Soft Independent Modelling of Class Analogy applied to infrared spectroscopy for rapid discrimination between hardwood and softwood. *Energy, 117*, 251–258. https://doi.org/10.1016/j.energy.2016.10.092

Famiglini, G., Palma, P., Termopoli, V., & Cappiello, A. (2021). The history of electron ionization in LC-MS, from the early days to modern technologies: A review. *Analytica Chimica Acta, 1167*, Article 338350. https://doi.org/10.1016/j.aca.2021.338350

Feizi, N., Hashemi-Nasab, F. S., Golpelichi, F., Saburouh, N., & Parastar, H. (2021). Recent trends in application of chemometric methods for GC-MS and GC×GC-MS-based metabolomic studies. *TrAC-Trends in Analytical Chemistry, 138*, Article 116239. https://doi.org/10.1016/j.trac.2021.116239

Feng, T. G., Wang, C. R., Zhang, J., Wang, B., & Jin, Y. F. (2021). An improved artificial bee colony-random forest (IABC-RF) model for predicting the tunnel deformation due to an adjacent foundation pit excavation. *Underground Space, 7*(4), 514–527. https://doi.org/10.1016/j.undsp.2021.11.004

Firmani, P., La Piscopia, G., Bucci, R., Marini, F., & Biancolillo, A. (2020). Authentication of P.G.I. Gragnano pasta by near infrared (NIR) spectroscopy and chemometrics. *Microchemical Journal, 152*, Article 104339. https://doi.org/10.1016/j.microc.2019.104339

Flexa, C., Gomes, W., Moreira, I., Alves, R., & Sales, C. (2021). Polygonal Coordinate System: Visualizing high-dimensional data using geometric DR, and a deterministic version of t-SNE. *Expert Systems with Applications, 175*, Article 114741. https://doi.org/10.1016/j.eswa.2021.114741

Gao, F. F., Hao, X. Y., Zeng, G. H., Guan, L. X., Wu, H., Zhang, L., et al. (2022). Identification of the geographical origin of Ecolly (Vitis vinifera L.) grapes and wines from different Chinese regions by ICP-MS coupled with chemometrics. *Journal of*

Food Composition and Analysis, 105, Article 104248. https://doi.org/10.1016/j.jfca.2021.104248

Gomez-de Anda, F., Gallardo-Velazquez, T., Osorio-Revilla, G., Dorantes-Alvarez, L., Calderon-Dominguez, G., Nogueda-Torres, B., et al. (2012). Feasibility study for the detection of Trichinella spiralis in a murine model using mid-Fourier transform infrared spectroscopy (MID-FTIR) with attenuated total reflectance (ATR) and soft independent modelling of class analogies (SIMCA). *Veterinary Parasitology, 190*(3–4), 496–503. https://doi.org/10.1016/j.vetpar.2012.07.004

Greener, J. G., Kandathil, S. M., Moffat, L., & Jones, D. T. (2022). A guide to machine learning for biologists. *Nature Reviews Molecular Cell Biology, 23*(1), 40–55. https://doi.org/10.1038/s41580-021-00407-0

Guo, K. M., Yan, H., Huang, D. W., & Yan, X. J. (2022a). Active learning-based KNN-Monte Carlo simulation on the probabilistic fracture assessment of cracked structures. *International Journal of Fatigue, 154*, Article 106533. https://doi.org/10.1016/j.ijfatigue.2021.106533

Guo, Q. Y., Adelina, N. M., Hu, J. T., Zhang, L. G., & Zhao, Y. H. (2022b). Comparative analysis of volatile profiles in four pine-mushrooms using HS-SPME/GC-MS and E-nose. *Food Control, 134*. https://doi.org/10.1016/j.foodcont.2021.108711

Hao, J., Dong, F. J., Li, Y. L., Wang, S. L., Cui, J. R., Zhang, Z. F., et al. (2022). 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 Physics & Technology, 125*, Article 104286. https://doi.org/10.1016/j.infrared.2022.104286

He, K. M., Zhang, X. Y., Ren, S. Q., & Sun, J. (2016). *Deep residual learning for image recognition*. Paper presented at the. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

He, M., & Zhou, Y. (2021). 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. *Chinese Herbal Medicines, 13*(1), 2–16. https://doi.org/10.1016/j.chmed.2020.05.006

He, Y. L., Geng, Z. Q., Xu, Y., & Zhu, Q. X. (2015). A robust hybrid model integrating enhanced inputs based extreme learning machine with PLSR (PLSR-EIELM) and its application to intelligent measurement. *ISA Transactions, 58*, 533–542. https://doi.org/10.1016/j.isatra.2015.06.007

Huang, F. R., Song, H., Guo, L., Guang, P. W., Yang, X. H., Li, L. Q., et al. (2020). Detection of adulteration in Chinese honey using NIR and ATR-FTIR spectral data fusion. *Spectrochimica Acta Part A-Molecular and Biomolecular Spectroscopy, 235*, Article 118297. https://doi.org/10.1016/j.saa.2020.118297

Huang, X. Y., Yu, S. S., Xu, H. X., Aheto, J. H., Bonah, E., Ma, M., et al. (2019). Rapid and nondestructive detection of freshness quality of postharvest spinaches based on machine vision and electronic nose. *Journal of Food Safety, 39*(6), e12708.

Ichihara, K., Kohsaka, C., & Yamamoto, Y. (2021). Determination of proteinaceous free amino acids by gas chromatography. *Analytical Biochemistry, 633*, Article 114423. https://doi.org/10.1016/j.ab.2021.114423

Ji, L. L., Lin, M., Jiang, W. B., Cao, G. H., Xu, Z. P., & Hao, F. (2022). An improved rock typing method for tight sandstone based on new rock typing indexes and the weighted fuzzy kNN algorithm. *Journal of Petroleum Science and Engineering, 210*, Article 109956. https://doi.org/10.1016/j.petrol.2021.109956

Jia, J. M., Zhou, X. F., Li, Y., Wang, M., Liu, Z. Y., & Dong, C. W. (2022). Establishment of a rapid detection model for the sensory quality and components of Yuezhou Longjing tea using near-infrared spectroscopy. *LWT-Food Science and Technology, 164*, Article 113625. https://doi.org/10.1016/j.lwt.2022.113625

Jiang, H. Y., Zhang, Y. X., Liu, Z. G., Wang, X. Y., He, J. M., & Yin, H. T. (2022). Advanced applications of mass spectrometry imaging technology in quality control and safety assessments of traditional Chinese medicines. *Journal of Ethnopharmacology, 284*, Article 114760. https://doi.org/10.1016/j.jep.2021.114760

Jimenez-Carvelo, A. M., Martin-Torres, S., Ortega-Gavilan, F., & Camacho, J. (2021a). PLS-DA vs sparse PLS-DA in food traceability. A case study: Authentication of avocado samples. *Talanta, 224*, Article 121904. https://doi.org/10.1016/j.talanta.2020.121904

Jimenez-Carvelo, A. M., Tonolini, M., McAleer, O., Cuadros-Rodriguez, L., Granato, D., & Koidis, A. (2021b). Multivariate approach for the authentication of vanilla using infrared and Raman spectroscopy. *Food Research International, 141*, Article 110196. https://doi.org/10.1016/j.foodres.2021.110196

Jin, G., Wang, Y. J., Li, L. Q., Shen, S. S., Deng, W. W., Zhang, Z. Z., et al. (2020). Intelligent evaluation of black tea fermentation degree by FT-NIR and computer vision based on data fusion strategy. *LWT-Food Science and Technology, 125*, Article 109216. https://doi.org/10.1016/j.lwt.2020.109216

Kaiser, S., Carvalho, Â. R., Pittol, V., Peñaloza, E. M., de Resende, P. E., Soares, F. L. F., et al. (2020). 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. *Microchemical Journal, 152*, Article 104346. https://doi.org/10.1016/j.microc.2019.104346

Khan, S. R., Sharma, B., Chawla, P. A., & Bhatia, R. (2021). Inductively coupled plasma optical emission spectrometry (ICP-OES): A powerful analytical technique for elemental analysis. *Food Analytical Methods, 15*(3), 666–688. https://doi.org/10.1007/s12161-021-02148-4

Khanmohammadi, M., Garmarudi, A. B., Ramin, M., & Ghasemi, K. (2013). Diagnosis of renal failure by infrared spectrometric analysis of human serum samples and soft independent modeling of class analogy. *Microchemical Journal, 106*, 67–72. https://doi.org/10.1016/j.microc.2012.05.006

Kholi, F. K., Park, S., Yang, J. S., Ha, M. Y., & Min, J. K. (2023). A detailed review of pulsating heat pipe correlations and recent advances using Artificial Neural Network for improved performance prediction. *International Journal of Heat and Mass Transfer, 207*, Article 124010. https://doi.org/10.1016/j.ijheatmasstransfer.2023.124010

Kiran Naik, B., Chinthala, M., Patel, S., & Ramesh, P. (2021). Performance assessment of waste heat/solar driven membrane-based simultaneous desalination and liquid desiccant regeneration system using a thermal model and KNN machine learning tool. *Desalination, 505*, Article 114980. https://doi.org/10.1016/j.desal.2021.114980

Lamberti, W. F. (2021). Blood cell classification using interpretable shape features: A comparative study of SVM models and CNN-Based approaches. *Computer Methods and Programs in Biomedicine Update, 1*, Article 100023. https://doi.org/10.1016/j.cmpbup.2021.100023

Lan, Z. W., Zhang, Y., Sun, Y., Ji, D., Wang, S. M., Lu, T. L., et al. (2020). 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. *Journal of Pharmaceutical and Biomedical Analysis, 188*, Article 113387. https://doi.org/10.1016/j.jpba.2020.113387

Li, C. L., Zong, B. Z., Guo, H. W., Luo, Z., He, P. M., Gong, S. Y., et al. (2020a). Discrimination of white teas produced from fresh leaves with different maturity by near-infrared spectroscopy. *Spectrochimica Acta Part A-Molecular and Biomolecular Spectroscopy, 227*, Article 117697. https://doi.org/10.1016/j.saa.2019.117697

Li, P., Zhang, W. L., Lu, C. J., Zhang, R., & Li, X. L. (2022). Robust kernel principal component analysis with optimal mean. *Neural Networks, 152*, 347–352. https://doi.org/10.1016/j.neunet.2022.05.005

Li, Q. Q., Huang, Y., Zhang, J. X., & Min, S. G. (2021). A fast determination of insecticide deltamethrin by spectral data fusion of UV-vis and NIR based on extreme learning machine. *Spectrochimica Acta Part A-Molecular and Biomolecular Spectroscopy, 247*, Article 119119. https://doi.org/10.1016/j.saa.2020.119119

Li, Y., Huang, Y., Xia, J. J., Xiong, Y. M., & Min, S. G. (2020b). Quantitative analysis of honey adulteration by spectrum analysis combined with several high-level data fusion strategies. *Vibrational Spectroscopy, 108*, Article 103060. https://doi.org/10.1016/j.vibspec.2020.103060

Li, Y., Xiong, Y. M., & Min, S. G. (2019). Data fusion strategy in quantitative analysis of spectroscopy relevant to olive oil adulteration. *Vibrational Spectroscopy, 101*, 20–27. https://doi.org/10.1016/j.vibspec.2018.12.009

Liang, J., Li, M. G., Du, Y., Yan, C. H., Zhang, Y., Zhang, T. L., et al. (2020). Data fusion of laser induced breakdown spectroscopy (LIBS) and infrared spectroscopy (IR) coupled with random forest (RF) for the classification and discrimination of compound salvia miltiorrhiza. *Chemometrics and Intelligent Laboratory Systems, 207*, Article 104179. https://doi.org/10.1016/j.chemolab.2020.104179

Liang, Y., Hu, S. S., Guo, W. S., & Tang, H. Q. (2022). Abrasive tool wear prediction based on an improved hybrid difference grey wolf algorithm for optimizing SVM. *Measurement, 187*, Article 110247. https://doi.org/10.1016/j.measurement.2021.110247

Lin, T., Wu, P., & Gao, F. M. (2022). Information security of flowmeter communication network based on multi-sensor data fusion. *Energy Reports, 8*, 12643–12652. https://doi.org/10.1016/j.egyr.2022.09.072

Liu, C. L., Shen, T., Xu, F. R., & Wang, Y. Z. (2022a). Main components determination and rapid geographical origins identification in *Gentiana rigescens* Franch. based on HPLC, 2DCOS images combined to ResNet. *Industrial Crops and Products, 187*, Article 115430. https://doi.org/10.1016/j.indcrop.2022.115430

Liu, X. Y., Jin, J., Wu, W. N., & Herz, F. B. (2020). A novel support vector machine ensemble model for estimation of free lime content in cement clinkers. *ISA Transactions, 99*, 479–487. https://doi.org/10.1016/j.isatra.2019.09.003

Liu, Y., Pu, H. B., & Sun, D. W. (2021a). Efficient extraction of deep image features using convolutional neural network (CNN) for applications in detecting and analysing complex food matrices. *Trends in Food Science & Technology, 113*, 193–204. https://doi.org/10.1016/j.tifs.2021.04.042

Liu, Z. M., Yang, M. Q., Zuo, Y. M., Wang, Y. Z., & Zhang, J. Y. (2022b). Fraud detection of herbal medicines based on modern analytical technologies combine with chemometrics approach: A review. *Critical Reviews in Analytical Chemistry, 52*(7), 1606–1623. https://doi.org/10.1080/10408347.2021.1905503

Liu, Z. M., Yang, S. B., Wang, Y. Z., & Zhang, J. Y. (2021b). Multi-platform integration based on NIR and UV-Vis spectroscopies for the geographical traceability of the fruits of *Amomum tsao-ko*. *Spectrochimica Acta Part A-Molecular and Biomolecular Spectroscopy, 258*, Article 119872. https://doi.org/10.1016/j.saa.2021.119872

Lovatti, B. P. O., Nascimento, M. H. C., Neto, Á. C., Castro, E. V. R., & Filgueiras, P. R. (2019). Use of Random forest in the identification of important variables. *Microchemical Journal, 145*, 1129–1134. https://doi.org/10.1016/j.microc.2018.12.028

Ma, H. L., Wang, T., Li, B. L., Cao, W. Y., Zeng, M., Yang, J. H., et al. (2022). A low-cost and efficient electronic nose system for quantification of multiple indoor air contaminants utilizing HC and PLSR. *Sensors and Actuators B-Chemical, 350*, Article 130768. https://doi.org/10.1016/j.snb.2021.130768

Ma, L., Gao, R., Han, H. J., Chen, C., Yan, Z. W., Zhao, J. Y., et al. (2020). Efficient identification of Bachu mushroom by flourier transform infrared (FT-IR) spectroscopy coupled with PLS-GS-SVM. *Optik, 224*. https://doi.org/10.1016/j.ijleo.2020.165712

Maaten, L. V. D., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research, 9*, 2579–2605.

Machado, J. C., Jr., Faria, M. A., Ferreira, I., Pascoa, R., & Lopes, J. A. (2018). Varietal discrimination of hop pellets by near and mid infrared spectroscopy. *Talanta, 180*, 69–75. https://doi.org/10.1016/j.talanta.2017.12.030

Mandrile, L., Barbosa-Pereira, L., Sorensen, K. M., Giovannozzi, A. M., Zeppa, G., Engelsen, S. B., et al. (2019). Authentication of cocoa bean shells by near- and mid-infrared spectroscopy and inductively coupled plasma-optical emission spectroscopy. *Food Chemistry, 292*, 47–57. https://doi.org/10.1016/j.foodchem.2019.04.008

Meng, T., Jing, X. Y., Yan, Z., & Pedrycz, W. (2020). A survey on machine learning for data fusion. *Information Fusion, 57*, 115–129. https://doi.org/10.1016/j.inffus.2019.12.001

Meng, X. R., Yin, C. L., Yuan, L. B., Zhang, Y., Ju, Y., Xin, K. H., et al. (2023). Rapid detection of adulteration of olive oil with soybean oil combined with chemometrics by Fourier transform infrared, visible-near-infrared and excitation-emission matrix fluorescence spectroscopy: A comparative study. *Food Chemistry, 405*, Article 134828. https://doi.org/10.1016/j.foodchem.2022.134828

Metz, M., Abdelghafour, F., Roger, J. M., & Lesnoff, M. (2021). A novel robust PLS regression method inspired from boosting principles: RoBoost-PLSR. *Analytica Chimica Acta, 1179*, Article 338823. https://doi.org/10.1016/j.aca.2021.338823

Mishra, P., Biancolillo, A., Roger, J. M., Marini, F., & Rutledge, D. N. (2020). New data preprocessing trends based on ensemble of multiple preprocessing techniques. *TrAC-Trends in Analytical Chemistry, 132*, Article 116045. https://doi.org/10.1016/j.trac.2020.116045

Mishra, P., Roger, J. M., Marini, F., Biancolillo, A., & Rutledge, D. N. (2022). Pre-processing ensembles with response oriented sequential alternation calibration (PROSAC): A step towards ending the pre-processing search and optimization quest for near-infrared spectral modelling. *Chemometrics and Intelligent Laboratory Systems, 222*, Article 104497. https://doi.org/10.1016/j.chemolab.2022.104497

Monakhova, Y. B., Holzgrabe, U., & Diehl, B. W. K. (2018). Current role and future perspectives of multivariate (chemometric) methods in NMR spectroscopic analysis of pharmaceutical products. *Journal of Pharmaceutical and Biomedical Analysis, 147*, 580–589. https://doi.org/10.1016/j.jpba.2017.05.034

Ng, W., Minasny, B., Montazerolghaem, M., Padarian, J., Ferguson, R., Bailey, S., et al. (2019). Convolutional neural network for simultaneous prediction of several soil properties using visible/near-infrared, mid-infrared, and their combined spectra. *Geoderma, 352*, 251–267. https://doi.org/10.1016/j.geoderma.2019.06.016

Obisesan, K. A., Jimenez-Carvelo, A. M., Cuadros-Rodriguez, L., Ruisanchez, I., & Callao, M. P. (2017). HPLC-UV and HPLC-CAD chromatographic data fusion for the authentication of the geographical origin of palm oil. *Talanta, 170*, 413–418. https://doi.org/10.1016/j.talanta.2017.04.035

Oliveira, M. M., Cruz-Tirado, J. P., & Barbin, D. F. (2019). Nontargeted analytical methods as a powerful tool for the authentication of spices and herbs: A review. *Comprehensive Reviews in Food Science and Food Safety, 18*(3), 670–689. https://doi.org/10.1111/1541-4337.12436

Park, B., Seo, Y., S.C., Y. (2015). Hyperspectral microscope imaging methods to classify gram-positive and gram-negative foodborne pathogenic bacteria. *Transactions of the Asabe, 58*, 5-16. doi:10.13031/trans.58.10832.

Pei, Y. F., Zhang, Q. Z., & Wang, Y. Z. (2020). Application of authentication evaluation techniques of ethnobotanical medicinal plant Genus *Paris*: A review. *Critical Reviews in Analytical Chemistry, 50*(5), 405–423. https://doi.org/10.1080/10408347.2019.1642734

Peng, W., Beggio, G., Pivato, A., Zhang, H., Lü, F., & He, P. J. (2022). Applications of near infrared spectroscopy and hyperspectral imaging techniques in anaerobic digestion of bio-wastes: A review. *Renewable & Sustainable Energy Reviews, 165*, Article 112608. https://doi.org/10.1016/j.rser.2022.112608

Peng, X. J. (2010). TSVR: An efficient Twin Support Vector Machine for regression. *Neural Networks, 23*(3), 365–372. https://doi.org/10.1016/j.neunet.2009.07.002

Qi, L. M., Ma, Y. T., Zhong, F. R., & Shen, C. (2018). 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. *Journal of Pharmaceutical and Biomedical Analysis, 161*, 436–443. https://doi.org/10.1016/j.jpba.2018.09.012

Qiao, Z. K., Yuan, P., Wang, L. F., Zhang, Z. H., Huang, Y. S., Zhang, J. J., et al. (2022). Research on aeromagnetic compensation of a multi-rotor UAV based on robust principal component analysis. *Journal of Applied Geophysics, 206*, Article 104791. https://doi.org/10.1016/j.jappgeo.2022.104791

Quackatz, L., Griesche, A., & Kannengiesser, T. (2022). Spatially resolved EDS, XRF and LIBS measurements of the chemical composition of duplex stainless steel welds: A comparison of methods. *Spectrochimica Acta Part B-Atomic Spectroscopy, 193*, Article 106439. https://doi.org/10.1016/j.sab.2022.106439

Rajput, J. M., Nandre, D. S., & Pawar, B. G. (2022). A comprehensive review on advanced chromatographic techniques and spectroscopic techniques in pharmaceutical analysis. *International Journal of Pharmaceutical Research and Applications, 7*(3), 53–62. https://doi.org/10.35629/7781-07035362

Rezaei, I., Amirshahi, S. H., & Mahbadi, A. A. (2023). Utilizing support vector and kernel ridge regression methods in spectral reconstruction. *Results in Optics, 11*, Article 100405. https://doi.org/10.1016/j.rio.2023.100405

Rodionova, O. Y., Titova, A. V., & Pomerantsev, A. L. (2016). Discriminant analysis is an inappropriate method of authentication. *TrAC Trends in Analytical Chemistry, 78*, 17–22. https://doi.org/10.1016/j.trac.2016.01.010

Romero-del-Castillo, J. A., Mendoza-Hurtado, M., Ortiz-Boyer, D., & García-Pedrajas, N. (2022). Local-based k values for multi-label k-nearest neighbors rule. *Engineering Applications of Artificial Intelligence, 116*, Article 105487. https://doi.org/10.1016/j.engappai.2022.105487

Salcedo-Sanz, S., Ghamisi, P., Piles, M., Werner, M., Cuadra, L., Moreno-Martínez, A., et al. (2020). Machine learning information fusion in Earth observation: A comprehensive review of methods, applications and data sources. *Information Fusion, 63*, 256–272. https://doi.org/10.1016/j.inffus.2020.07.004

Shafiee, S., & Minaei, S. (2018). Combined data mining/NIR spectroscopy for purity assessment of lime juice. *Infrared Physics & Technology, 91*, 193–199. https://doi.org/10.1016/j.infrared.2018.04.012

Shan, P., Bi, Y. M., Li, Z. G., Wang, Q. Y., He, Z. H., Zhao, Y. H., et al. (2023). Unsupervised model adaptation for multivariate calibration by domain adaptation-regularization based kernel partial least square. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 292*, Article 122418. https://doi.org/10.1016/j.saa.2023.122418

Shen, H. Y., Ma, Y. Z., Lin, C. L., Zhou, J., & Liu, L. J. (2023). Hierarchical Bayesian support vector regression with model parameter calibration for reliability modeling and prediction. *Reliability Engineering & System Safety, 229*, Article 108842. https://doi.org/10.1016/j.ress.2022.108842

Stavropoulos, G., van Vorstenbosch, R., Jonkers, D., Penders, J., Hill, J. E., van Schooten, F. J., et al. (2021). Advanced data fusion: Random forest proximities and pseudo-sample principle towards increased prediction accuracy and variable interpretation. *Analytica Chimica Acta, 1183*, Article 339001. https://doi.org/10.1016/j.aca.2021.339001

Su, G. S., Huang, J. H., Xu, H. J., & Qin, Y. Z. (2022). Extracting acoustic emission features that precede hard rock instability with unsupervised learning. *Engineering Geology, 306*, Article 106761. https://doi.org/10.1016/j.enggeo.2022.106761

Su, S. L., Zhou, X. C., Wan, C., Li, Y. K., & Kong, W. H. (2016). Land use changes to cash crop plantations: Crop types, multilevel determinants and policy implications. *Land Use Policy, 50*, 379–389. https://doi.org/10.1016/j.landusepol.2015.10.003

Sudol, P. E., Galletta, M., Tranchida, P. Q., Zoccali, M., Mondello, L., & Synovec, R. E. (2022). Untargeted profiling and differentiation of geographical variants of wine samples using headspace solid-phase microextraction flow-modulated comprehensive two-dimensional gas chromatography with the support of tile-based Fisher ratio analysis. *Journal of Chromatography A, 1662*, Article 462735. https://doi.org/10.1016/j.chroma.2021.462735

Sun, W. J., Zhang, X., Zhang, Z. Y., & Zhu, R. H. (2017). Data fusion of near-infrared and mid-infrared spectra for identification of rhubarb. *Spectrochimica Acta Part A-Molecular and Biomolecular Spectroscopy, 171*, 72–79. https://doi.org/10.1016/j.saa.2016.07.039

Tarnutzer, A. A., & Weber, K. P. (2022). Pattern analysis of peripheral-vestibular deficits with machine learning using hierarchical clustering. *Journal of the Neurological Sciences, 434*, Article 120159. https://doi.org/10.1016/j.jns.2022.120159

Tata, A., Massaro, A., Damiani, T., Piro, R., Dall'Asta, C., & Suman, M. (2022). Detection of soft-refined oils in extra virgin olive oil using data fusion approaches for LC-MS, GC-IMS and FGC-Enose techniques: The winning synergy of GC-IMS and FGC-Enose. *Food Control, 133*, Article 108645. https://doi.org/10.1016/j.foodcont.2021.108645

Teye, E., Huang, X. Y., Takrama, J., & Gu, H. Y. (2014). Integrating NIR spectroscopy and electronic tongue together with chemometric analysis for accurate classification of cocoa bean varieties. *Journal of Food Process Engineering, 37*(6), 560–566. https://doi.org/10.1111/jfpe.12109

Toribio, L., Bernal, J., Martín, M. T., & Ares, A. M. (2021). Supercritical fluid chromatography coupled to mass spectrometry: A valuable tool in food analysis. *TrAC-Trends in Analytical Chemistry, 143*, Article 116350. https://doi.org/10.1016/j.trac.2021.116350

Varshney, A. K., Muhuri, P. K., & Danish Lohani, Q. M. (2022). PIFHC: The probabilistic intuitionistic fuzzy hierarchical clustering algorithm. *Applied Soft Computing, 120*, Article 108584. https://doi.org/10.1016/j.asoc.2022.108584

Vieira, L. S., Assis, C., de Queiroz, M., Neves, A. A., & de Oliveira, A. F. (2021). Building robust models for identification of adulteration in olive oil using FT-NIR PLS-DA and variable selection. *Food Chemistry, 345*, Article 128866. https://doi.org/10.1016/j.foodchem.2020.128866

Wang, C. Y., Tang, L., Li, L., Zhou, Q., Li, Y. J., Li, J., et al. (2020a). Geographic authentication of *Eucommia ulmoides* leaves using multivariate analysis and preliminary study on the compositional response to environment. *Frontiers in Plant Science, 11*, 79. https://doi.org/10.3389/fpls.2020.00079

Wang, H. Y., Xu, P. D., & Zhao, J. H. (2022a). Improved KNN algorithms of spherical regions based on clustering and region division. *Alexandria Engineering Journal, 61*(5), 3571–3585. https://doi.org/10.1016/j.aej.2021.09.004

Wang, J., & Biljecki, F. (2022). Unsupervised machine learning in urban studies: A systematic review of applications. *Cities, 129*, Article 103925. https://doi.org/10.1016/j.cities.2022.103925

Wang, X. H., Li, B., Yan, Y. Y., Gao, N., & Chen, G. M. (2019). Predicting of thermal resistances of closed vertical meandering pulsating heat pipe using artificial neural network model. *Applied Thermal Engineering, 149*, 1134–1141. https://doi.org/10.1016/j.applthermaleng.2018.12.142

Wang, Y., He, T., Wang, J. J., Wang, L., Ren, X. Y., He, S. H., et al. (2021a). High performance liquid chromatography fingerprint and headspace gas chromatography-mass spectrometry combined with chemometrics for the species authentication of Curcumae Rhizoma. *Journal of Pharmaceutical and Biomedical Analysis, 202*, Article 114144. https://doi.org/10.1016/j.jpba.2021.114144

Wang, Y., Li, M. G., Feng, T., Zhang, T. L., Feng, Y. Q., & Li, H. (2022b). 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). *Chinese Journal of Analytical Chemistry, 50*(3), Article 100057. https://doi.org/10.1016/j.cjac.2022.100057

Wang, Y. J., Li, L. Q., Liu, Y., Cui, Q. Q., Ning, J. M., & Zhang, Z. Z. (2021b). Enhanced quality monitoring during black tea processing by the fusion of NIRS and computer vision. *Journal of Food Engineering, 304*, Article 110599. https://doi.org/10.1016/j.jfoodeng.2021.110599

Wang, Z. Q., Liu, J. M., Zeng, C. H., Bao, C. H., Li, Z. J., Zhang, D. J., et al. (2023). Rapid detection of protein content in rice based on Raman and near-infrared spectroscopy fusion strategy combined with characteristic wavelength selection. *Infrared Physics & Technology, 129*, Article 104563. https://doi.org/10.1016/j.infrared.2023.104563

Wen, C. T., Zhang, J. X., Zhang, H. H., Dzah, C. S., Zandile, M., Duan, Y. Q., et al. (2018). Advances in ultrasound assisted extraction of bioactive compounds from cash crops – A review. *Ultrasonics – Sonochemistry, 48*, 538–549. https://doi.org/10.1016/j.ultsonch.2018.07.018

Włodarska, K., Piasecki, P., Lobo-Prieto, A., Pawlak-Lemańska, K., Górecki, T., & Sikorska, E. (2021). Rapid screening of apple juice quality using ultraviolet, visible, and near infrared spectroscopy and chemometrics: A comparative study. *Microchemical Journal, 164*. https://doi.org/10.1016/j.microc.2021.106051

Wu, X. M., Zhang, Q. Z., & Wang, Y. Z. (2018). Traceability of wild *Paris polyphylla* Smith var. *yunnanensis* based on data fusion strategy of FT-MIR and UV-Vis combined with SVM and random forest. *Spectrochimica Acta Part A-Molecular and Biomolecular Spectroscopy, 205*, 479–488. https://doi.org/10.1016/j.saa.2018.07.067

Xiao, Q. L., Bai, X. L., Gao, P., & He, Y. (2020). Application of convolutional neural network-based feature extraction and data fusion for geographical origin identification of Radix Astragali by visible/short-wave near-infrared and near infrared hyperspectral imaging. *Sensors, 20*(17), 4940. https://doi.org/10.3390/s20174940

Xiong, L., & Yao, Y. (2021). Study on an adaptive thermal comfort model with K-nearest-neighbors (KNN) algorithm. *Building and Environment, 202*, Article 108026. https://doi.org/10.1016/j.buildenv.2021.108026

Xu, Y. P., & Wu, Z. Y. (2022). Parameter identification of unsaturated seepage model of core rockfill dams using principal component analysis and multi-objective optimization. *Structures, 45*, 145–162. https://doi.org/10.1016/j.istruc.2022.09.020

Xue, J. T., Yang, Q. W., Li, C. Y., Liu, X. L., & Niu, B. X. (2021). Rapid and simultaneous quality analysis of the three active components in Lonicerae Japonicae Flos by near-infrared spectroscopy. *Food Chemistry, 342*, Article 128386. https://doi.org/10.1016/j.foodchem.2020.128386

Yakimovich, A., Beaugnon, A., Huang, Y., & Ozkirimli, E. (2021). Labels in a haystack: Approaches beyond supervised learning in biomedical applications. *Patterns, 2*(12). https://doi.org/10.1016/j.patter.2021.100383

Yang, W., Knorr, F., Latka, I., Vogt, M., Hofmann, G. O., Popp, J., et al. (2022). Real-time molecular imaging of near-surface tissue using Raman spectroscopy. *Light-Science & Applications, 11*(1), 90. https://doi.org/10.1038/s41377-022-00773-0

Yang, Y., Wu, Y. J., Li, W. L., Liu, X. S., Zheng, J. Y., Zhang, W. T., et al. (2018). Determination of geographical origin and icariin content of Herba Epimedii using near infrared spectroscopy and chemometrics. *Spectrochimica Acta Part A-Molecular and Biomolecular Spectroscopy, 191*, 233–240. https://doi.org/10.1016/j.saa.2017.10.019

Yao, C., Qi, L. M., Zhong, F. R., Li, N., & Ma, Y. T. (2022). An integrated chemical characterization based on FT-NIR, GC-MS and LC-MS for the comparative metabolite profiling of wild and cultivated agarwood. *Journal of Chromatography B-Analytical Technologies in the Biomedical and Life Sciences, 1188*, Article 123056. https://doi.org/10.1016/j.jchromb.2021.123056

Yuan, Z. R., Niu, M. Q., Ma, H. T., Gao, T., Zang, J., Zhang, Y. W., et al. (2023). Predicting mechanical behaviors of rubber materials with artificial neural networks. *International Journal of Mechanical Sciences, 249*, Article 108265. https://doi.org/10.1016/j.ijmecsci.2023.108265

Zaroual, H., Chene, C., El Hadrami, E. M., & Karoui, R. (2022). Application of new emerging techniques in combination with classical methods for the determination of the quality and authenticity of olive oil: A review. *Critical Reviews in Food Science and Nutrition, 62*(16), 4526–4549. https://doi.org/10.1080/10408398.2021.1876624

Zhang, P. F., Li, T. R., Yuan, Z., Luo, C., Wang, G. Q., Liu, J., et al. (2022). A data-level fusion model for unsupervised attribute selection in multi-source homogeneous data. *Information Fusion, 80*, 87–103. https://doi.org/10.1016/j.inffus.2021.10.017

Zhang, S. J., Gong, X. C., & Qu, H. B. (2022b). Near-infrared spectroscopy and HPLC combined with chemometrics for comprehensive evaluation of six organic acids in *Ginkgo biloba* leaf extract. *Journal of Pharmacy and Pharmacology, 74*(7), 1040–1050. https://doi.org/10.1093/jpp/rgab177

Zhang, X. H., Zhu, Q. X., Jiang, Z. Y., He, Y. L., & Xu, Y. (2018). A novel ensemble model using PLSR integrated with multiple activation functions based ELM: Applications to soft sensor development. *Chemometrics and Intelligent Laboratory Systems, 183*, 147–157. https://doi.org/10.1016/j.chemolab.2018.10.016

Zhou, X., Li, X. Q., Zhao, B., Chen, X. T., & Zhang, Q. H. (2022). Discriminant analysis of vegetable oils by thermogravimetric-gas chromatography/mass spectrometry combined with data fusion and chemometrics without sample pretreatment. *LWT-Food Science and Technology, 161*, Article 113403. https://doi.org/10.1016/j.lwt.2022.113403

Zhou, Y. H., Zuo, Z. T., Xu, F. R., & Wang, Y. Z. (2020). Origin identification of *Panax notoginseng* by multi-sensor information fusion strategy of infrared spectra combined with random forest. *Spectrochimica Acta Part A-Molecular and Biomolecular Spectroscopy, 226*, Article 117619. https://doi.org/10.1016/j.saa.2019.117619

Zhuang, T., Xin, M., Wang, Q. K., Wang, Y. M., Saeed, M., Xing, H. X., et al. (2023). Determination of protein and fatty acid composition of shell-intact upland cottonseed using near-infrared reflectance spectroscopy. *Industrial Crops and Products, 191*, Article 115909. https://doi.org/10.1016/j.indcrop.2022.115909

Further reading

Liu, C., Zuo, Z., Xu, F., & Wang, Y. (2023). Study of the suitable climate factors and geographical origins traceability of *Panax notoginseng* based on correlation analysis and spectral images combined with machine learning. *Frontiers in Plant Science, 13*, 1009727. https://doi.org/10.3389/fpls.2022.1009727