Journal of Animal & Plant Sciences, 35(4): 2025, Page: 889-899
ISSN (print): 1018-7081; ISSN (online): 2309-8694
https://doi.org/10.36899/JAPS.2025.4.0077

DETECTING ADULTERANTS IN TEA USING MID-INFRARED SPECTROSCOPY: A COMPARATIVE STUDY OF DEEP LEARNING AND MACHINE LEARNING

W. Liu[1,2], Y. Lin[1,3], Y. Cai[1], Honghao Cai[1*] and Hui Ni[4]

Department of Physics, School of Science, Jimei University, Xiamen, Fujian Province, China
[2]School of Materials Science and Engineering, Guilin University of Technology, Guangxi Province, China
[3]School of Electronic Science and Engineering, Xiamen University, Xiamen, Fujian Province, China
[4]Xiamen Ocean Vocational College, Xiamen, Fujian Province, China
*Corresponding author's E-mail: hhcai@jmu.edu.cn
ORCID ID: 0000-0002-1870-8061

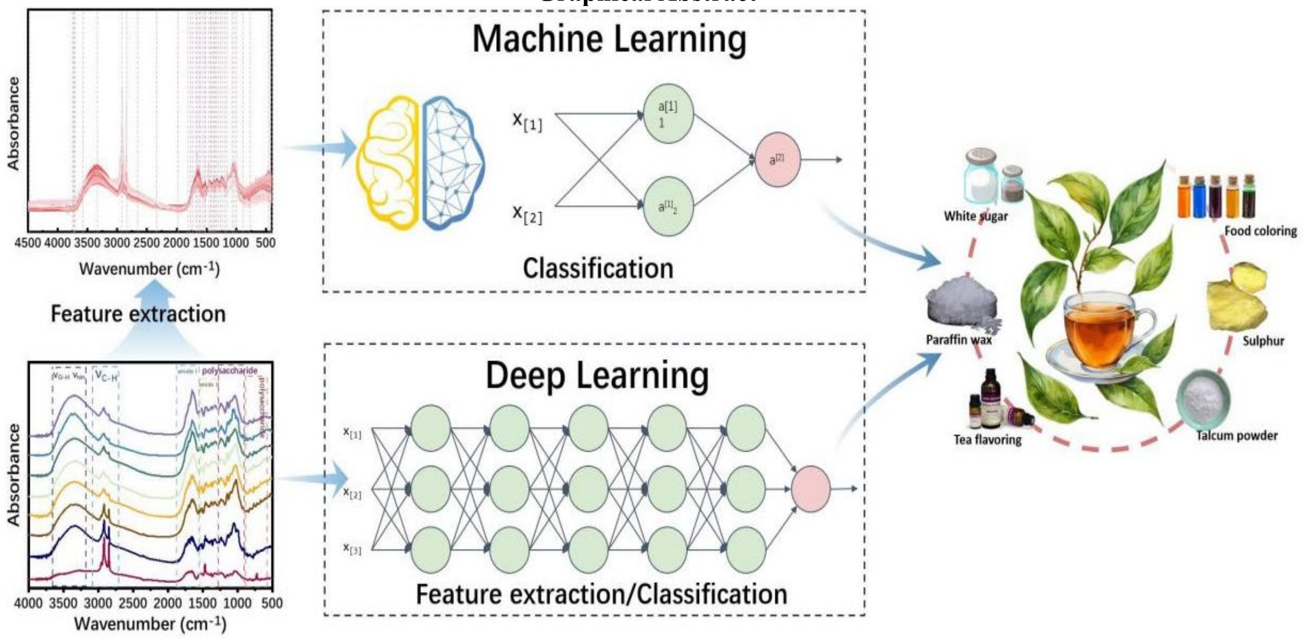
ABSTRACT

The detection of adulterants in tea using infrared spectroscopy has gained prominence. However, there has yet to be a systematic comparison of the performance of traditional machine learning methods versus deep learning in the context of modelling infrared data for tea quality. This study compares machine learning and deep learning for modeling spectral data. Machine learning methods like Random Forest, K-Nearest Neighbors (KNN), Support Vector Classification, and Gaussian Naive Bayes used the Successive Projections Algorithm (SPA) for feature extraction, while deep learning models automatically extracted features. SPA-KNN showed superior performance with 0.950 accuracy, 0.953 macro-precision, 0.950 macro-recall, and 0.950 macro-F1 score on the test set (n=80), processing in 1.33 seconds. Deep learning models such as ResNet achieved a lower accuracy of 0.688 and required a longer processing time of 335.60 seconds. This may be partly due to the limited generalization ability caused by the small sample size. Additionally, the complex structure of ResNet, such as its depth, may contribute to the longer processing time. This study offers insights for selecting appropriate methods in tea quality assessment. Machine learning methods, especially with spectral preprocessing and SPA-based feature extraction, remain effective, while deep learning may not excel in limited data scenarios due to its higher computational demands.

Keywords: Adulteration, Food quality, 1D CNN, Resnet, LSTM, Classification algorithm, Feature selection

Published first online June 10 , 2025 Published final July 29, 2025

Graphical Abstract



INTRODUCTION

The combination of spectroscopy and machine learning algorithms has significantly advanced various applications (Chen *et al.*, 2023; Gao *et al.*, 2024a; Qiu *et al.*, 2024; Xiao *et al.*, 2024). In the field of food science, it effectively facilitates quantitative component analysis (Yan *et al.*, 2022; Liao *et al.*, 2024), the detection of adulterants (Lin *et al.*, 2025; Liu *et al.*, 2025) and pesticide residues (Ye *et al.*, 2022), quality classification (Magnus *et al.*, 2021), origin tracing (Zhang *et al.*, 2023), and the assessment of fermentation levels (Wu *et al.*, 2024). In tea testing, the primary spectroscopic techniques include in tea quality testing and control include infrared (IR) spectroscopy (Bai *et al.*, 2022), ultraviolet-visible (UV-Vis) spectroscopy (Wang *et al.*, 2025), Raman spectroscopy (Li *et al.*, 2023), mass spectrometry (Peng *et al.*, 2023), and nuclear magnetic resonance (NMR) spectroscopy (Hou *et al.*, 2024). IR spectroscopy is rapid, non-destructive, and versatile, yet it has low resolution and is sensitive to moisture. UV-Vis spectroscopy is sensitive for detecting specific components but has a limited range. Raman spectroscopy is highly sensitive and selective, though its signal is weak and detection speed is slow. Mass spectrometry offers precise structural information, but it requires complex and costly sample preparation. NMR spectroscopy provides non-destructive structural data, but it is expensive, slow, and operationally demanding. Overall, IR spectroscopy is the most suitable method for mass production and quality control due to its speed, non-destructiveness, and wide application.

In recent years, deep learning (Le, 2020; Zhang *et al.*, 2021; Luo *et al.*, 2022; Cai *et al.*, 2025; Lin *et al.*, 2025) has begun to revolutionize mathematical modelling based on spectroscopy. Unlike traditional machine learning methods, which often rely on handcrafted features and linear models, deep learning networks excel at automatically extracting relevant features from complex data sets. This enables them to capture intricate patterns and relationships within the spectral data that traditional methods may miss. Infrared spectra arise from both linear and non-linear combinations of molecular vibrations, including combination bands and overtones. Additionally, these spectra can exhibit signals from physical effects, such as baseline variations due to light scattering, multiplicative influences from pathlength discrepancies, and temperature-induced shifts in peak positions. Traditional machine learning methods, particularly those involving spectra preprocessing (such as smoothing, normalization, and baseline correction) and feature selection algorithms (such as principal component analysis (PCA), successive projections algorithm (SPA) and competitive adaptive reweighted sampling (CARS)), have effectively modelled near-infrared (NIR) data. However, to enhance model performance in addressing non-linear data characteristics, non-linear techniques have also been employed. That said, non-linear approaches, such as those utilizing kernel support vector machines, may encounter issues with overfitting, largely due to the absence of built-in regularization mechanisms (Melssen *et al.*, 1994; Smits *et al.*, 1994; Anderson *et al.*, 2020).

In tea research, both machine learning and deep learning (Kamrul *et al.*, 2020; Yang *et al.*, 2021; Hu *et al.*, 2025) have achieved significant advancements and applications. Issues related to tea adulterants occasionally arise in China. Despite the government implementing a series of strict food safety regulations, unscrupulous merchants still take risks. Prolonged consumption of certain adulterants, such as sulphur and talcum powder, can lead to kidney burdens and even cancer (Li *et al.*, 2016). Even seemingly harmless adulterants like white sugar can adversely affect diabetic patients. Therefore, rapid and effective detection methods for tea adulterants are crucial for protecting the rights of legitimate manufacturers and consumers. However, research applying machine learning or deep learning in conjunction with spectroscopy for this purpose is still relatively sparse. Amsaraj and Mutturi combined spectral data with chemometric analysis to classify and quantify multiple adulterants in black tea (Amsaraj and Mutturi, 2024). Another study used ultra-high performance liquid chromatography-tandem mass spectrometry with QuEChERS purification to detect 27 pyrrolizidine alkaloids in tea (Yao *et al.*, 2024). In the realm of infrared (IR) spectroscopy, a study assessed the validity and redundancy of spectral data for detecting sucrose-doped tea (Liu *et al.*, 2022), while Fourier transform infrared (FT-IR) transmission spectroscopy was employed to detect talcum powder in green tea (Li *et al.*, 2017). Additionally, a rapid method using FT-IR spectroscopy and chemometrics was developed to identify sunset yellow adulteration in tea powder (Amsaraj and Mutturi, 2023). These works collectively advance tea adulteration detection methods. However, the diversity of adulterants is still insufficient, and there is a lack of research on the simultaneous adulteration of multiple substances.

This paper focuses on tea adulterants as the subject of study, employing both machine learning and deep learning techniques to develop models for detecting seven types of tea adulterants. We will comprehensively evaluate and analyse the performance of these models from accuracy, speed, computational resources, generalization ability and model transparency and interpretability. This work not only provides the scientific community with guidelines for the application of machine learning and deep learning with IR spectroscopy but also identifies the most suitable model for detecting tea adulterants. This model will facilitate more reliable and efficient screening processes, thereby enhancing food safety standards and consumer confidence in tea products.

MATERIALS AND METHODS

Sample preparation: The experimental materials included: Wuyi rock tea obtained from First Class Teas Inc. in Wuyishan City, Fujian, China; white sugar sourced from Sanlvyuan Food Factory; talcum powder from Guilin Guiguang Talc Development Co., Ltd.; food coloring (Sunset Yellow) supplied by Dongguan Jinjiahe Food Co., Ltd.; paraffin wax from Penglei Chemical Division; tea flavoring also provided by Dongguan Jinjiahe Food Co., Ltd. All products were securely sealed in plastic bags containing a desiccant and stored in a laboratory maintained at 25 °C. Due to the legal prohibition of these unauthorized adulterants, it is difficult to find documented ratios in literature or books, so we had to consult some tea manufacturers that have previously used these adulterants. Based on their experience, the typical ratio of fresh tea leaves to adulterants during the tea processing is around 1:50. This ratio effectively improves the flavor, appearance, and shelf life of the tea. In our study, we chose this ratio to ensure the authenticity and representativeness of the experimental results. Additionally, similar ratios have been used in previous studies on tea extraction and analysis, where a 1:50 tea-to-water ratio is often employed to prepare tea infusions for the determination of various components such as minerals and bioactive compounds. For instance, Vuong *et al.* (2011) investigated the optimum conditions for the water extraction of L-theanine from green tea and found that a tea-to-water ratio of 1:50 (g/ml) at pH <6 yielded the maximum extraction efficiency. Similarly, Banerjee and Chatterjee (2015) studied the extraction of bioactive components from tea and found that a 50:1 water-to-tea ratio for 40 minutes to extract polyphenols from black tea is more suitable. These studies support the validity of using a 1:50 ratio in our research, as it aligns with established practices in the field of tea extraction and analysis. Accordingly, we prepared seven different types of tea samples: tea with sugar, tea with talcum powder, tea with sunset yellow, tea with paraffin wax, tea with tea flavoring, tea with sugar and paraffin wax, and tea with talcum powder, paraffin wax, and tea flavoring.

Before the IR spectroscopy experiments, both the tea samples containing adulterants and the pure tea samples were prepared into 3 mg tea tablets using the potassium bromide pellet technique (Ingebrigtson and Smith, 1954). Each sample type consisted of 60 tablets, resulting in a total of 8 types and 480 tablets (training set: 400 and test set: 80). Preparation of tablets: Tea and tea adulterant samples were first dried in a drying chamber (Shangcheng Co., Ltd, China) at 43°C for 2 hours, then precisely weighed using a loading balance with an accuracy of 0.0001g (Lichen Co., Ltd, China). The tea samples were subsequently ground into fine particles and sieved through a 100-mesh sieve to ensure uniform particle size. For tablet preparation, the potassium bromide (KBr) pellet technique was employed. This involved mixing the finely ground sample with dried KBr powder at a ratio of approximately 2% sample to KBr. The mixture was placed into a KBr pellet die and compressed under high pressure (10 tons) using a hydraulic press for about 2 minutes to form transparent and homogeneous pellets. Finally, the quality of the formed pellets was carefully inspected to ensure they were free from defects and suitable for the intended analysis.

Acquisition of spectra: IR spectral acquisition was conducted at a stable temperature of 25 °C utilizing a WQF-500 Fourier transform infrared spectrometer (Beifen-Ruili Analytical Instrument Co., Ltd., China) paired with Main FTOS Suite software. The spectra, captured in absorbance mode, spanned from 4500 to 400 cm^{-1}, comprising 256 scans that produced a total of 4253 variables. The potassium bromide background was subtracted from the sample spectra. For each sample, three measurements were performed and subsequently averaged for further analysis.

Spectra preprocessing and evaluation metrics: In this study, we aimed to compare the performance of machine learning on preprocessed data and deep learning on raw data. The raw IR spectra were not subjected to any preprocessing, while the preprocessed data underwent min-max normalization, baseline correction, standard normal variate transformation, Savitzky-Golay smoothing, and first and second derivative calculations. These preprocessing steps were implemented in Python using the Spyder compiler software (Anaconda, Inc., USA). To evaluate the models developed, we used several metrics, including accuracy, marco-precision, marco-recall, and marco-F1 score. These metrics provide a comprehensive assessment of the classification performance. Accuracy measures the overall proportion of correctly classified instances, while precision focuses on the correctness of positive predictions. Recall assesses the model's ability to identify positive samples, and the F1 score, as the harmonic mean of precision and recall, offers a balanced evaluation of performance. We chose macro-averaged metrics because they treat each class equally, which is suitable when all classes are equally important and you want to assess model performance across them uniformly. This choice allows us to thoroughly evaluate our model's effectiveness in detecting different types of tea adulterants, ensuring that we don't overlook the performance on less frequent but potentially significant classes.

Algorithm: In our study, we employed four traditional machine learning classifiers (Random Forest (RF), K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), and Gaussian Naive Bayes (GNB)) and four deep learning algorithms (one-dimensional Convolutional Neural Networks (1D CNN), Residual Networks (ResNet), MobileNetV3, and Long Short-Term Memory (LSTM) networks) for the classification task.

RF, another ensemble method using multiple decision trees, provides robust performance and handles high-dimensional data well, reducing overfitting risk through feature and sample randomness. KNN, a simple instance-based learning algorithm, offers a non-parametric approach, useful for small-scale data with distinct clusters. SVC, based on support vector machines, is powerful for high-dimensional and non-linearly separable data by finding optimal hyperplanes with kernel tricks. GNB, based on Bayes' theorem with feature independence assumptions, is computationally efficient and suitable for high-dimensional features, though performance drops with correlated features.

The deep learning algorithms, including 1D CNN, ResNet, MobileNetV3, and LSTM networks, were chosen for their unique strengths in processing complex data. 1D CNN can capture local patterns and hierarchical structures in sequential data, making them well-suited for analyzing the complex and high-dimensional nature of IR spectroscopy data. ResNet, with its residual connections, helps mitigate the vanishing gradient problem in deep networks, allowing for the training of deeper models that can capture more complex patterns. MobileNetV3 offers a good balance between computational efficiency and model accuracy, making it suitable for real-world applications with limited computational resources. LSTM networks are capable of modelling temporal dependencies, which is particularly useful if the spectral data has a sequential aspect, such as time-series measurements.

This diverse selection of models allows for a comprehensive comparison of performance, strengths, and limitations in the context of tea adulteration detection. By including both traditional machine learning and deep learning approaches, we leverage the benefits of interpretability, efficiency, and powerful feature extraction. The combination ensures that we can draw meaningful conclusions about the most effective approaches for detecting tea adulterants, offering valuable insights for future research and applications in the field.

RESULTS

Spectral band assignment: The absorption band in the range of 3000 to 2700 cm^{-1} is linked to C-H stretching vibrations from proteins and amino acids. The band between 1690 and 1640 cm^{-1} corresponds to the amide I band of proteins, which arises from C=O vibrations of polyphenolic compounds. Meanwhile, the band from 1540 to 1500 cm^{-1} represents the amide II band, associated with the bending vibrations of the peptide bond (-CONH-) linked to N-H from polyphenols. Furthermore, the spectral region between 1200 and 950 cm^{-1} is indicative of polysaccharide components (Fig. 1a).

When various adulterants are introduced, the tea samples exhibit altered chemical compositions and molecular environments. For instance, paraffin wax, which contains alkyl groups, may show absorption bands corresponding to the stretching vibrations of hydrocarbon bonds (C-H) within the wavenumber range of 3000 to 2750 cm^{-1}. The presence of wax can also influence the vibrational frequencies of the molecules, resulting in shifts in the positions of certain absorption bands observed between 1750 to 1500 cm^{-1} and 1200 to 1000 cm^{-1}.

Figure 1b illustrates the spectra of six pure adulterants: sugar, talcum powder, food coloring, food flavoring, and paraffin. These adulterants exhibit distinct spectral characteristics in comparison to pure tea samples. Sugar, for example, displays characteristic absorption bands due to C-O stretching vibrations in the 1000-1200 cm^{-1} region, which are distinct from the broader O-H stretching vibrations observed in tea polyphenols and polysaccharides. Talcum powder, primarily composed of magnesium silicate, shows absorption bands related to Si-O stretching vibrations around 1000-1100 cm^{-1} and O-H bending vibrations near 1600-1800 cm^{-1}, which are different from the amide I and II bands of tea proteins. Food coloring and flavoring agents, which often contain aromatic structures and functional groups such as carbonyls and conjugated double bonds, exhibit specific absorptions in the 1600-1800 cm^{-1} and 3000-3100 cm^{-1} regions, features not typically observed in pure tea spectra.

Feature selection: IR spectroscopy typically covers a wide range of wavenumbers, some of which are irrelevant to the model. Therefore, it is necessary to eliminate these interfering wavenumbers and select the most relevant ones to improve model accuracy and reduce modelling time. This study applies SPA to select the most relevant frequencies for the classification model. The results show that the SPA identified 41 valid characteristic wavenumbers, which are highlighted by dashed lines in Figure 2.

Model comparison: Machine Learning. The KNN model adjusts its accuracy by modifying the number of neighbors, denoted as n_neighbors, which determines the influence of each sample point. Smaller values increase model complexity, leading to a higher risk of overfitting, while larger values reduce complexity and may result in underfitting. The complexity of the SVC model is controlled by the regularization parameter C. Smaller values of C simplify the decision boundaries, reducing the risk of overfitting, while larger values create more complex boundaries, which may cause overfitting.

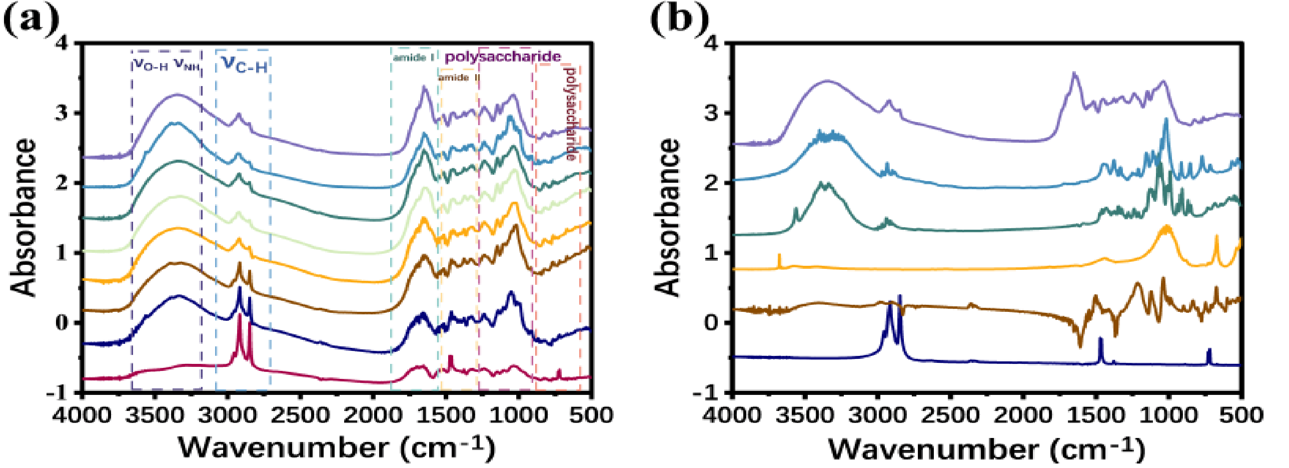


Figure 1. (a) Mid-infrared spectra of tea with wavenumber identifiers: — = tea + paraffin wax, — = pure tea, — = tea + white sugar, — = tea + talcum powder, — = tea + food coloring, — = tea + tea flavoring, — = tea + talcum powder + paraffin wax + tea flavoring, — = tea + white sugar + paraffin wax; (b) Mid-infrared spectra of adulterants: — = paraffin wax, — = food coloring, — = talcum powder, — = white sugar, — = tea flavoring, — = pure tea.

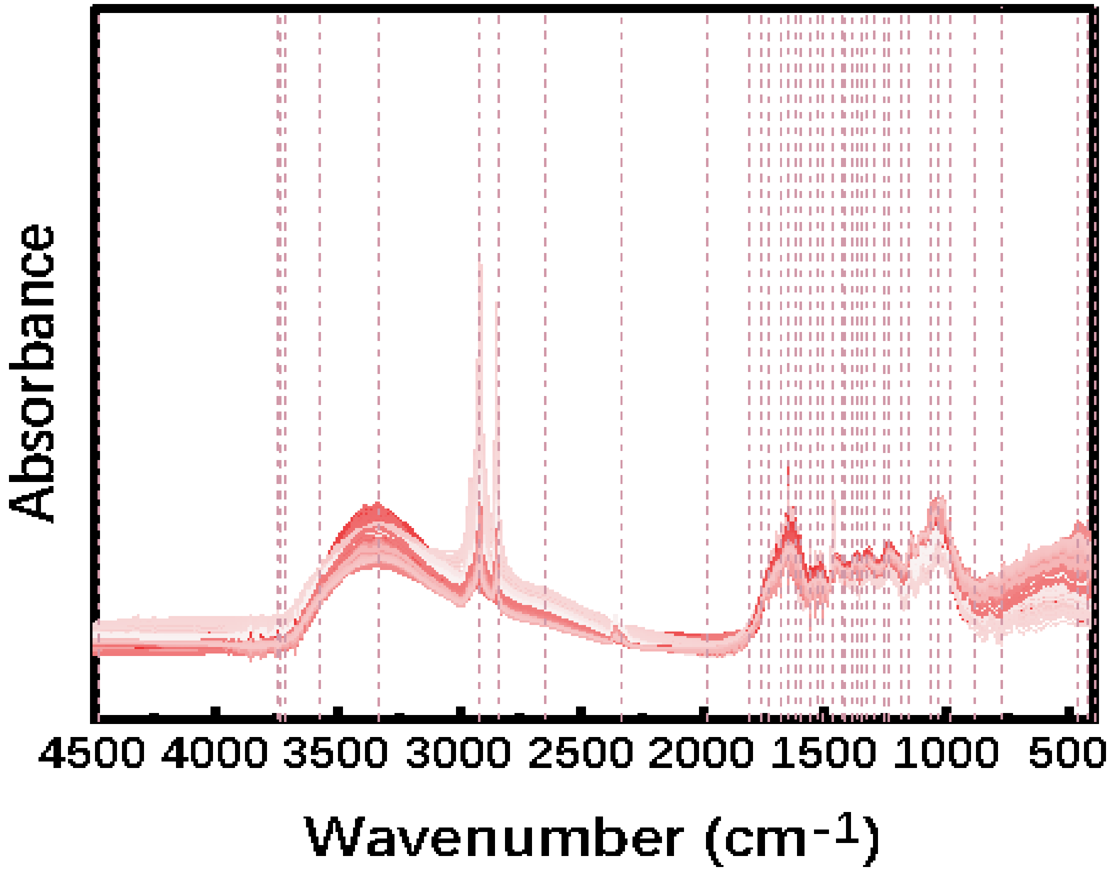


Figure 2. The wavenumbers selected by SPA

The appropriate C value can be selected through cross-validation. The GNB model uses the smoothing parameter var_smoothing, added to the variance to prevent zero variance issues. Smaller values increase model complexity, leading to potential overfitting, while larger values reduce complexity and may lead to underfitting. The RF model is controlled by parameters such as the number of trees (n_estimators), maximum tree depth (max_depth), minimum samples required to split internal nodes (min_samples_split), and the maximum number of features to consider when splitting (max_features). Increasing the number of trees typically improves performance but increases computational cost. The optimal number of trees can be determined via cross-validation. A larger max_depth increases model complexity, leading to overfitting, while a smaller max_depth limits model expressiveness, potentially causing underfitting. Smaller min_samples_split values allow for more complex tree structures, which may result in overfitting, while larger values reduce complexity. Similarly, smaller max_features values limit tree growth, helping to avoid overfitting, whereas larger values increase complexity and may lead to overfitting. Cross-validation and grid search were used to determine the optimal parameters.

The model was tested on 80 samples of pure tea and tea with various adulterants. The test dataset consists of 10 spectra of pure tea and 70 spectra representing various adulterants: white sugar, talcum powder, food coloring, paraffin, tea flavoring, white sugar + paraffin, and talcum powder + paraffin + tea flavoring. Table 1 summarizes the model's performance on the test set. As shown in Table 1, when sorted by accuracy on the test set, the order is SPA-KNN > SPA-RF = SPA-SVC > SPA-GNB. Most algorithms experienced a decline in accuracy, likely due to differences in data quality and distribution between the training and test sets. Among them, SPA-KNN achieved the best overall performance, with the highest accuracy (0.950).

Table 1. Comparison of the prediction results of different machine learning models on the training set (n=400) and test set (n=80).

Classification models	Cross-validation scores	Accura-cy	Macro-precision	Macro-recall	Macro-F1	Processing time (s)
SPA-KNN	0.970 ± 0.027	0.950	0.953	0.950	0.950	1.33
SPA-SVC	0.925 ± 0.034	0.938	0.948	0.938	0.937	1.10
SPA-GNB	0.880 ± 0.040	0.750	0.775	0.750	0.750	0.39
SPA-RF	0.980 ± 0.019	0.938	0.950	0.938	0.936	2.98

Note: Cross-validation scores, Accuracy, Macro-precision, Macro-recall and Macro-F1 data were based on the pre-processed spectral data.

Deep Learning. Increasing the number of convolutional kernels in a 1D CNN allows the model to extract more features, enhancing its expressiveness and data fitting ability. Pooling methods such as max pooling or average pooling can downsample the feature map, reducing data and computational complexity while preserving key feature information. This improves the model’s robustness and generalization ability. ResNet introduces residual connections to deepen the network, mitigate vanishing gradients, and enable learning of deeper feature representations, thereby improving model accuracy. Adjusting parameters such as the convolution kernel size, stride, and padding can also alter the feature map's size and receptive field, affecting feature extraction and representation capabilities. MobileNetV3 relies heavily on depthwise separable convolution, decomposing convolution into depthwise and 1×1 pointwise convolutions. This significantly reduces computational load and parameter count while maintaining high performance, making the model efficient and accurate in resource-constrained environments like mobile devices. The addition of the Squeeze-and-Excitation (SE) module, an attention mechanism, dynamically adjusts the weights of each feature map channel to highlight important features and suppress less relevant ones, further enhancing feature representation and model accuracy. Optimizations to the block structure, such as removing shortcut connections in certain cases, reduce computational overhead without compromising accuracy. Additionally, modifying expensive layers at the network's beginning and end, such as reducing the number of kernels in the first convolutional layer and simplifying the last stage, reduces model delay and computation while maintaining good accuracy. LSTM can improve sequence data prediction and classification accuracy by increasing the number of hidden units, allowing the model to learn more features and complex sequence patterns. Tuning the forgetting gate parameters, such as the weight matrix and bias terms, helps the model better retain important information and discard irrelevant data, optimizing its ability to capture long-term dependencies. The choice of activation functions, such as sigmoid and tanh, plays a crucial role in information processing. Selecting or adjusting the activation function based on the specific task and data characteristics can further improve model performance. Proper data preprocessing, such as normalization and standardization, helps achieve a more uniform data distribution, facilitating faster convergence and higher accuracy. Additionally, optimizing the sequence length ensures the model retains sufficient information without making training more difficult due to overly long sequences.

Table 2 shows the performance of four deep learning models on raw data. Sort the test set in order of accuracy, ResNet > LSTM > MobileNetV3 > 1D CNN. Among them, ResNet has the best comprehensive performance and the highest accuracy (0.688), but the processing time is long (335.6 s), which is suitable for scenes with high classification accuracy requirements. LSTM has acceptable performance in cross-validation scores, accuracy, accuracy, recall and F1 scores, and has the shortest processing time (11.67 s), which is suitable for scenes that require both time and classification effect. 1D CNN has a relatively poor performance in various indicators, so it may need to further optimize the model structure or adjust the hyperparameters. Some processing of the data will make these models perform better.

Table 2. Comparison of the prediction results of different deep learning models on the training set (n=400) and test set (n=80).

Classification models	Cross-validation scores	Accura-cy	Macro-precision	Macro-recall	Macro-F1	Processing time (s)
1D-CNN	0.619 ± 0.073	0.513	0.523	0.518	0.520	383.95
LSTM	0.835 ± 0.091	0.600	0.563	0.563	0.563	11.67
MobileNetV3	0.348 ± 0.183	0.580	0.563	0.563	0.563	39.32
ResNet	0.558 ± 0.096	0.688	0.683	0.688	0.685	335.60

Note: Cross-validation scores, Accuracy, Macro-precision, Macro-recall and Macro-F1 data were based on were based on the raw spectra.

DISCUSSION

The results of this study demonstrate the effectiveness of combining mid-infrared (MIR) spectroscopy with both machine learning and deep learning algorithms for the detection of tea adulterants. Among the tested models, the SPA-KNN model achieved the highest prediction accuracy (0.950), macro-precision (0.953), macro-recall (0.950), and macro-F1 score (0.950), with a processing time of 1.33 seconds. This indicates that SPA-KNN is the most effective model among those tested for this specific application. Among the deep learning models, ResNet achieved the highest accuracy (0.688), macro-precision (0.683), macro-recall (0.688), and macro-F1 score (0.832), albeit with a long processing time of 335.60 seconds. Recently, Yang *et al.* (2021) combined diffuse reflection MIR spectroscopy with chemometrics to quickly identify adulteration in Radix Astragali, using the KNN classification method, achieving 100% prediction accuracy. Both this and previous studies demonstrate that KNN, a non-parametric method based on locality-based classification, is particularly well-suited for handling IR spectral data with complex, unknown distributions and similar spectral characteristics.

When trained on large and diverse datasets, deep learning models tend to generalize better (Gao *et al.*, 2024b, c). They can automatically extract relevant features from the data, making them highly flexible and effective for complex tasks like NIR spectroscopy. However, if the dataset is too small, as in our current study, or if the models are not properly regularized, overfitting can become a significant issue. In contrast, traditional machine learning models may perform better on smaller datasets, especially when feature engineering is effectively applied. Yet, they often struggle with the complexity of high-dimensional data. The generalization gap between training and testing performance is typically more pronounced in deep learning models, but this gap can be reduced through techniques such as dropout, batch normalization, and data augmentation.

Deep learning models are often seen as "black-boxes" due to their lack of transparency. Their decision-making processes are hard to trace because features are abstract and spread across layers. Traditional machine learning models like decision trees and linear regression are more interpretable, with simpler, visualizable decision processes (e.g., tree paths or linear weights). While deep learning may perform better, its transparency is lower, making it harder to understand decisions or debug. In our machine learning study, using SPA for feature selection helps identify features linked to specific adulterants, offering insights into tea adulteration. SPA focuses on peak regions, stepwise selecting the features most strongly correlated with the adulterant target (Esteki *et al.*, 2016), and extracts features by calculating the area under the peak in the spectrum, aiding in the accurate analysis and identification of adulterants. For example, as shown in Figure 1b, paraffin, which is composed of long alkane chains, exhibits prominent C-H stretching absorptions in the range of 2800-3000 cm^{-1} and C-C stretching absorptions around 700-900 cm^{-1}. These absorption features contrast with the more complex and varied patterns observed in the components of tea (Al-Mokhalelati *et al.*, 2023). These distinct spectral differences facilitate the identification and differentiation of various adulterants in tea samples using infrared spectroscopy. Likewise, Sun *et al.* (2019) proposed using

SPA combined with stepwise regression to select characteristic wavelengths, which effectively improved the correlation coefficient of the prediction set in the established multiple linear regression model.

Previous studies have used machine learning and IR spectroscopy to examine tea adulterants. For instance, Amsaraj and Mutturi (2023) reported the quantification of sunset yellow in tea powder using a random forest based on FT-IR, achieving high R^2 and RMSE values. Another study (Li *et al.*, 2017) utilized FT-IR transmission spectroscopy for the detection of talcum powder in green tea, demonstrating the effectiveness of FT-IR techniques in identifying specific adulterants. Our study advances the field by offering a broader detection scope of tea adulterants and the ability to detect multiple adulterants simultaneously. Unlike previous studies that focused on specific adulterants, our approach can identify a wider range of contaminants.

Conclusions: The MIR spectroscopy combined with deep learning and machine learning had potential for differentiating tea adulterated with various adulterants, which could be particularly useful in regulatory enforcement and industry quality control settings. Among the tested classification models, the SPA-KNN model delivered the most comprehensive test-set results, achieving a prediction accuracy of 0.950, macro-precision of 0.953, macro-recall of 0.950, macro-F1 score of 0.950, and a detection time of 1.33 s. Moreover, the selected features can help identify adulterant types.

Author contributions: All authors contributed equally to the concept idea and design of the study. All authors have read and agreed to the published version of the manuscript.

Conflicts of interest: All of the authors declares that there is no conflict of interest regarding the publication of this paper.

Data availability statement:

Some or all data, models, or code that support the findings of this study are available from the corresponding author.

Funding: This work was supported by the Natural Science Foundation of Fujian Province of China [grant number 2022J01821, 2022J05163 and 2021J011169];

the Natural Science Foundation of China [grant number 32172339]; the Natural Science Foundation of China [grant number 11705068].

Conflict of interest disclosure: The authors declare no competing interests.

Permission to reproduce material from other sources: All the materials in this manuscript are original.

REFERENCES

Al-Mokhalelati, K., F. Karabet, A. Allaf, M. Naddaf, B. Assfour and A. Al Lafi (2023). Silicone oils aided fabrication of paraffin wax coated super-hydrophobic sand: A spectroscopic study. Heliyon 9(10): e20874. https://doi.org/10.1016/j.heliyon.2023.e20874

Amsaraj, R. and S. Mutturi (2023). Rapid detection of sunset yellow adulteration in tea powder with variable selection coupled to machine learning tools using spectral data. J. Food Sci. Tech. 60(5): 1530-1540. http://doi.org/10.1007/s13197-023-05694-3

Amsaraj, R. and S. Mutturi (2024). Classification and quantification of multiple adulterants simultaneously in black tea using spectral data coupled with chemometric analysis. J. Food Compos. Anal. 125: 105715. https://doi.org/10.1016/j.jfca.2023.105715

Anderson, N., K. Walsh, P. Subedi and C. Hayes (2020). Achieving robustness across season, location and cultivar for a NIRS model for intact mango fruit dry matter content. Postharvest Biol. Tec. 168: 111202. https://doi.org/10.1016/j.postharvbio.2020.111202

Banerjee, S. and J. Chatterjee (2015). Efficient extraction strategies of tea (*Camellia sinensis*) biomolecules. J. Food Sci. Tech. 52: 3158-3168. https://doi.org/10.1007/s13197-014-1487-3

Bai, X., L. Zhang, C. Kang, B. Quan, Y. Zheng, X. Zhang, J. Song, T. Xia and M. Wang (2022). Near-infrared spectroscopy and machine learning-based technique to predict quality-related parameters in instant tea. Sci. Rep. 12(1): 3833. https://doi.org/10.1038/s41598-022-07652-z

Chen, S., Y. Wang, Q. Zhu, H. Ni and H. Cai (2023). Fast recognition of the harvest period of *Porphyra haitanensis* based on mid-infrared spectroscopy and chemometrics. J. Food Meas. Charact. 17(5): 5487-5496. https://doi.org/10.1007/s11694-023-01999-1

Cai, Y., Y. Lin, H. Cai and H. Ni (2025). Deep learning in vibrational spectroscopy: Benefits, limitations, and recent progress. J. Chin. Chem. Soc. 72(6): 611 – 626. https://doi.org/10.1002/jccs.70031

Esteki, M., S. Nouroozi and Z. Shahsavari (2016). A fast and direct spectrophotometric method for the simultaneous determination of methyl paraben and hydroquinone in cosmetic products using successive projections algorithm. Int. J. Cosmet. Sci. 38(1): 25-34. https://doi.org/10.1111/ics.12241

Gao, Z., Q. Lin, Q. He, C. Liu, H. Cai and H. Ni (2024a). Rapid detection of spoiled apple juice using electrical impedance spectroscopy and data augmentation-based machine learning. Chiang Mai J. Sci. 51(5): e2024071. https://doi.org/10.12982/CMJS.2024.071

Gao, Z., S. Chen, J. Huang and H. Cai (2024b). Real-time quantitative detection of hydrocolloid adulteration in meat based on Swin Transformer and smartphone. J. Food Sci. 89(7): 4359-4371. https://doi.org/10.1111/1750-3841.17159

Gao, Z., J. Huang, J. Chen, T. Shao, H. Ni and H. Cai (2024c). Deep transfer learning-based computer vision for real-time harvest period classification and impurity detection of *Porphyra haitnensis*. Aquacult. Int. 32(4): 5171-5198. https://doi.org/10.1007/s10499-024-01422-6

Hou, Z., Y. Jin, Z. Gu, R. Zhang, Z. Su and S. Liu (2024). ^{1}H NMR spectroscopy combined with machine-learning algorithm for origin recognition of Chinese famous green tea Longjing tea. Foods 13(17): 2702. https://doi.org/10.3390/foods13172702

Hu, Y., W. Sheng, S. Y. S. S. Adade, J. Wang, H. Li and Q. Chen (2025). Comparison of machine learning and deep learning models for detecting quality components of vine tea using smartphone-based portable near-infrared device. Food Control 174: 111244. https://doi.org/10.1016/j.foodcont.2025.111244

Ingebrigtson, D. N. and A. L. Smith (1954). Infrared analysis of solids by potassium bromide pellet technique. Anal. Chem. 26(11): 1765-1768. https://doi.org/10.1021/ac60095a023

Kamrul, M. H., M. Rahman, M. R. I. Robin, M. S. Hossain, M. H. Hasan and P. Paul. (2020). *A deep learning based approach on categorization of tea leaf.* Paper presented at the Proceedings of the International Conference on Computing Advancements. https://doi.org/10.1145/3377049.3377122

Le, B. T. (2020). Application of deep learning and near infrared spectroscopy in cereal analysis. Vib. Spectrosc. 106: 103009. https://doi.org/10.1016/j.vibspec.2019.103009

Li, F., Y. Huang, X. Wang, D. Wang and M. Fan (2023). Surface-enhanced Raman scattering integrating with machine learning for green tea storage time identification. Luminescence 38(3): 302-307. https://doi.org/10.1002/bio.4449

Li, X., Y. Zhang and Y. He (2016). Rapid detection of talcum powder in tea using FT-IR spectroscopy coupled with chemometrics. Sci. Rep. 6(1): 30313. https://doi.org/10.1038/srep30313

Li, X., Y. Zhang and Y. He (2017). Study on detection of talcum powder in green tea based on fourier transform infrared (FTIR) transmission spectroscopy. Spectrosc. Spect. Anal. 37(4): 1081-1085. https://doi.org/10.3964/j.issn.1000-0593(2008)10-2428-03

Liao, W., H. Cai and H. Ni (2024). Sesame seed metabolism during germination under auxin: An in vivo NMR study. J. Plant Growth Regul. 44: 2764–2777. https://doi.org/10.1007/s00344-024-11574-7

Lin, Y., H. Cai, S. Lin and H. Ni (2025). Mid-infrared-spectroscopy-based method for identifying single and multiple vegetable protein adulterants in whey protein. J. Appl. Spectrosc. 91: 1378-1386. https://doi.org/10.1007/s10812-025-01863-8

Lin, Y., Y. Cai, H. Chen, Y. Cai, Z. Lin, H. Cai and H. Ni (2025). Adaptive feedback cross-loop for preserving and robust spectral information optimization without spectral processing in few-shot learning. Meas. Sci. Technol. 36: 075503. https://doi.org/10.1088/1361-6501/aded2a

Liu, M., Q. Wu, X. Wang, Q. Chen, Y. Zhang, S. Huang and J. Fang (2022). Validity and redundancy of spectral data in the detection algorithm of sucrose-doped content in tea. Spectrosc. Spect. Anal. 42(11): 3647-3652. https://doi.org/10.3964/j.issn.1000-0593(2008)10-2428-03

Liu, W., Y. Lin, C. Liu, H. Cai and H. Ni (2025). Rapid detection of common additives in tea for quality assurance via mid-infrared spectroscopy and machine learning. Acta Sci. Polon.-Techn. 24(1): 27-45. https://doi.org/10.17306/J.AFS.001277

Luo, R., J. Popp and T. Bocklitz (2022). Deep learning for Raman spectroscopy: A review. Analytica 3(3): 287-301. https://doi.org/10.3390/analytica3030020

Magnus, I., M. Virte, H. Thienpont and L. Smeesters (2021). Combining optical spectroscopy and machine learning to improve food classification. Food Control 130: 108342. https://doi.org/10.3390/analytica3030020

Melssen, W., J. Smits, L. Buydens and G. Kateman (1994). Using artificial neural networks for solving chemical problems: Part II. Kohonen self-organising feature maps and Hopfield networks. Chemom. Intell. Lab. Syst. 23(2): 267-291. https://doi.org/10.1016/0169-7439(93)E0036-4

Peng, Y., C. Zheng, S. Guo, F. Gao, X. Wang, Z. Du, F. Gao, F. Su, W. Zhang and X. Yu (2023). Metabolomics integrated with machine learning to discriminate the geographic origin of Rougui Wuyi rock tea. npj Sci. Food 7(1): 7. https://doi.org/10.1038/s41538-023-00187-1

Qiu, J., Y. Lin, J. Wu, Y. Xiao, H. Cai and H. Ni (2024).

Rapid beef quality detection using spectra pre-processing methods in electrical impedance spectroscopy and machine learning. Int. J. Food Sci. Tech. 59(3): 1624-1634. https://doi.org/10.1111/ijfs.16915

Smits, J., W. Melssen, L. Buydens and G. Kateman (1994). Using artificial neural networks for solving chemical problems: Part I. Multi-layer feed-forward networks. Chemom. Intell. Lab. Syst. 22(2): 165-189. https://doi.org/10.1016/0169-7439(93)E0035-3

Sun, J., X. Zhou, Y. Hu, X. Wu, X. Zhang and P. Wang (2019). Visualizing distribution of moisture content in tea leaves using optimization algorithms and NIR hyperspectral imaging. Comput. Electron. Agric. 160: 153-159. https://doi.org/10.1016/j.compag.2019.03.004

Vuong, Q. V., C. E. Stathopoulos, J. B. Golding, M. H. Nguyen and P. D. Roach (2011). Optimum conditions for the water extraction of L-theanine from green tea. J. Sep. Sci. 34(18): 2468-2474. https://doi.org/10.1002/jssc.201100401

Wang, X., H. Chen, R. Ji, H. Qin, Q. Xu, T. Wang, Y. He and Z. Huang (2025). Detection of carmine in black tea based on UV-Vis absorption spectroscopy and machine learning. Food Anal. Methods 18(2): 149-160. https://doi.org/10.1007/s12161-024-02705-7

Wu, D., Y. Xu, F. Xu, M. Shao and M. Huang (2024). Machine learning algorithms for in-line monitoring during yeast fermentations based on Raman spectroscopy. Vib. Spectrosc. 132: 103672. https://doi.org/10.1016/j.vibspec.2024.103672

Xiao, Y., H. Cai and H. Ni (2024). Identification of geographical origin and adulteration of Northeast China soybeans by mid-infrared spectroscopy and spectra augmentation. J. Consum. Prot Food S. 19(1): 99-111. https://doi.org/10.1007/s00003-023-01471-8

Yan, H., W. Fan, X. Chen, H. Wang, C. Qin and X. Jiang (2022). Component spectra extraction and quantitative analysis for preservative mixtures by combining terahertz spectroscopy and machine learning. Spectrochim. Acta, Part A 271: 120908. https://doi.org/10.1016/j.saa.2022.120908

Yang, J., C. Yin, X. Miao, X. Meng, Z. Liu and L. Hu (2021). Rapid discrimination of adulteration in Radix Astragali combining diffuse reflectance mid-infrared Fourier transform spectroscopy with chemometrics. Spectrochim. Acta, Part A 248: 119251. https://doi.org/10.1016/j.saa.2020.119251

Yao, L., Y. Chen, H. Lin, L. Wang, P. Shi, Y. Zhang, T. Huang, J. Song, Y. Wang, Q. Dai and C. Liu (2024). Simultaneous determination of 27 pyrrolizidine alkaloids in tea by ultra-high performance liquid chromatography-tandem mass spectrometry with integrated QuEChERS purification. J. Tea Sci. 44(5): 831-842. https://doi.org/10.13305/j.cnki.jts.2024.05.005

Ye, W., T. Yan, C. Zhang, L. Duan, W. Chen, H. Song, Y. Zhang, W. Xu and P. Gao (2022). Detection of pesticide residue level in grape using hyperspectral imaging with machine learning. Foods 11(11): 1609. https://doi.org/10.3390/foods11111609

Zhang, L., H. Dai, J. Zhang, Z. Zheng, B. Song, J. Chen, G. Lin, L. Chen, W. Sun and Y. Huang (2023). A study on origin traceability of white tea (White Peony) based on near-infrared spectroscopy and machine learning algorithms. Foods 12(3): 499. https://doi.org/10.3390/foods12030499

Zhang, X., J. Yang, T. Lin and Y. Ying (2021). Food and agro-product quality evaluation based on spectroscopy and deep learning: A review. Trends Food Sci. Tech. 112: 431-441. https://doi.org/10.1016/j.tifs.2021.04.008.