Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 274 (2022) 121133

Contents lists available at ScienceDirect

Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy

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

Identification and quantification of adulterated honey by Raman spectroscopy combined with convolutional neural network and chemometrics

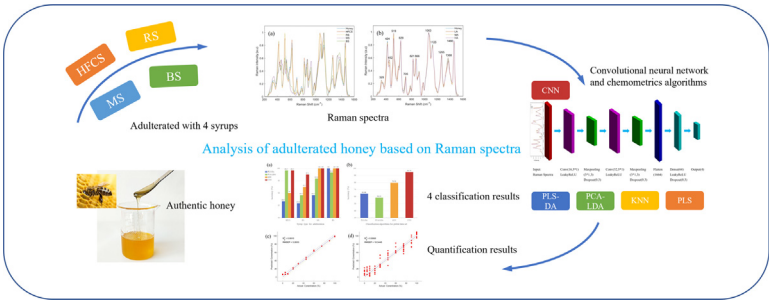
Xijun Wu, Baoran Xu *, Renqi Ma, Yudong Niu, Shibo Gao, Hailong Liu, Yungang Zhang

Measurement Technology & Instrumentation Key Laboratory of Hebei Province, Institute of Electrical Engineering, Yanshan University, Qinhuangdao 066004 China

HIGHLIGHTS

- Honey was blended with each of four common syrups to simulate adulterated honey.
- Two sets of CNNs were used to detect adulteration when the type of syrup was exact or unknown.
- The accuracy of CNN in syrup detection was improved compared to chemometric methods.
- Raman spectra with CNN and chemometrics can identify the authenticity of honey.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 30 November 2021
Received in revised form 23 February 2022
Accepted 7 March 2022
Available online 10 March 2022

Keywords:
Raman spectroscopy
Honey
Adulteration
Convolutional neural network
Partial least squares

ABSTRACT

In this study, Raman spectroscopy combined with convolutional neural network (CNN) and chemometrics was used to achieve the identification and quantification of honey samples adulterated with high fructose corn syrup, rice syrup, maltose syrup and blended syrup, respectively. The shallow CNNs utilized to analyze honey mixed with single-variety syrup classified samples into four categories by the adulteration concentration with more than 97% accuracy, and the general CNN model for simultaneously detecting honey adulterated with any type of syrup obtained an accuracy of 94.79%. The established CNNs had the best performance compared with several chemometric classification algorithms. In addition, partial least square regression (PLS) successfully predicted the purity of honey mixed with single syrup, while coefficients of determination and root mean square errors of prediction were greater than 0.98 and less than 3.50, respectively. Therefore, the proposed methods based on Raman spectra have important practical significance for food safety and quality control of honey products.

1. Introduction

As the oldest natural sweet substance, honey is collected and produced by bees from the nectar or secretions of plants [1]. It mainly consists of carbohydrates (about 80%), water (about 17%) and trace amounts of biologically active substances, such as phenolic compounds, organic acids, vitamins, proteins and minerals. Among them, carbohydrates are mainly composed of monosaccharides (fructose, glucose), small amounts of disaccharides (sucrose, maltose) and other oligosaccharides [2,3]. Honey is rich in biologically active substances. Compared with other types of sweeteners, honey has high nutritional value and medical potential. People consume honey for its antibacterial, anti-inflammatory, anticancer and immunomodulatory properties [4].

* Corresponding author at: Measurement Technology & Instrumentation Key Laboratory of Hebei Province, Institute of Electrical Engineering, Yanshan University, Qinhuangdao 066004, China.
E-mail address: xvbaoran@163.com (B. Xu).

https://doi.org/10.1016/j.saa.2022.121133
1386-1425/

Due to the current increasing consumer demand for honey, some unscrupulous merchants mix cheap syrup into honey to gain illegal economic benefits by lowering the quality and increasing the production [5]. This will not only harm the rights of consumers and lead to potential health problems (e.g., allergies) [6,7], but also can cause unfair competition in the market and affect the healthy development of industry. Therefore, it is very important to develop a robust analytical method that enables the detection of syrup adulteration in honey.

The current methods for detecting sugar adulteration in honey mainly include stable carbon isotope ratio analysis (SCIRA) [8,9], high performance liquid chromatography [10,11], gas chromatography [12,13], ion mobility-mass spectrometry [14], nuclear magnetic resonance (NMR) [1,15], vibrational spectroscopy [16,17], thermographic images [18] and biosensor technology (such as the electronic nose and electronic tongue) [19,20]. Among them, SCIRA is the recommended method for the detection of syrup adulteration in honey by the national standard GB/T18932.1, it can successfully detect honey adulterated with syrup from C4 plants. However, for honey samples mixed with syrup from the C3 plants, it is difficult to excavate the difference in $\delta^{13}C$ value between honey and its protein component by SCIRA, thus there is no assurance that the honey is not contaminated with C3 plant syrup. Therefore, this method can not guarantee the accuracy and stability of honey detection [3,8]. These techniques also require complex sample preprocessing and expensive analysis costs, which is not conducive to large-scale applications and field measurement.

In particular, spectroscopy technology has been widely utilized in quality control and field monitoring of various commodities due to its green, fast, nondestructive and easy-to-implement field detection characteristics [21]. Meanwhile, since the substances to be analysed usually have similar spectral peaks and bands, it is difficult to observe the spectral characteristics or differences only by intuitive visual analysis of the spectra, thus the spectral data analysis is often combined with multivariate statistical methods. For instance, in our previous study [22], we achieved the identification of adulterated sesame oil and successfully predicted the content of sesame flavor by combining three-dimensional fluorescence spectroscopy with Alexnet network and chemometrics. A recent research [23] by Teklemariam showed that the quantitative prediction of several adulterated sugars in commercially available fresh coconut water was achieved by combining Fourier transform infrared spectroscopy (FTIR) with principal component analysis (PCA) and PLS.

Raman spectroscopy obtains fingerprint information about the molecular structure and functional groups of samples through inelastic scattered light from the samples [24], and the Raman measurement is less susceptible to interference from moisture in the samples [23,25], which makes Raman spectroscopy very suitable for detecting honey samples. CNN, one of the most popular deep learning architectures, is increasingly used for the complex food matrix analysis owing to its ability to achieve automatic extraction of spectral features and its outstanding performance in generalization [26]. However, there are few studies using Raman spectroscopy combined with CNN for the detection of honey adulteration.

Based on previous studies [5,27] and local market conditions, four common commercially available syrups were selected for honey adulteration research, including high fructose corn syrup (HFCS), rice syrup (RS), maltose syrup (MS), and commercial blended syrup (BS) containing maltose, fructose, glucose, and sucrose in an unknown concentration ratio. Although the four selected syrups are referred to as four single-variety syrups in this study, their composition does not contain only one type of sugar. HFCS, RS and BS all contain multiple sugars.

Spectroscopic techniques combined with multivariate statistical methods are commonly used to detect honey adulteration of known syrup types. If an identification model trained with the spectral data of honey samples mixed with HFCS is used to detect samples adulterated with RS, the prediction results of the model may exhibit higher detection error. Therefore, it has become increasingly urgent and important to develop a general model capable of detecting honey adulterated with multiple syrups simultaneously. After the general model is established, the spectral data of samples made of honey from more botanical origins or more types of adulterant can be trained in the model, thus recalibrating the model to enhance its robustness and expand its application range. This means that in the actual measurement process, using a trained model with further evolution and expansion potential, when the type of adulterant is unknown, the fast, non-destructive and accurate detection can be achieved, which has important practical significance for honey quality control.

Raman spectroscopy combined with CNN and chemometrics was utilized for qualitative and quantitative detection of honey adulteration. Several chemometric algorithms were used for spectral analysis and performance comparison, including partial least squares discriminant analysis (PLS-DA), principal component analysis-linear discriminant analysis (PCA-LDA), and k-Nearest Neighbors (kNN) algorithm. Based on different spectral datasets, this study developed simple models for analysing four single-variety syrup adulterations as well as a general model capable of simultaneously detecting multiple syrups adulteration in honey. The purpose of this study was to evaluate the application potential of Raman spectroscopy for the detection of exogenous sugars in honey, whether the category of the adulterant was exact or unknown.

2. Materials and methods

2.1. Sample preparation

A common single-flowered lychee honey was obtained from a trusted producer, while samples of four syrups, HFCS, RS, MS, and BS, were purchased from Taobao Internet Mall, syrup manufacturer, and local market, respectively. Three types of syrup were derived from C3 plants as well as one from C4 plants. Details about the natural honey and the four syrups are available in Table S1. All samples were stored at room temperature in the dark before being analysed.

To cover a wide range of adulteration concentrations and obtain detection models with excellent generalization performance, the mixing proportions of pure honey and different adulterants were set as follows: 5%, 10%, 20%, 30%, 45%, 60%, 75%, 90% (w/w). Natural honey (0% adulteration) and pure syrup samples (100% adulteration) were also prepared and analysed. For the qualitative analysis of honey samples, all samples (including pure honey and syrups) were classified into four categories according to the adulteration concentration, namely Honey, low adulteration (LA, 5%-20%), medium adulteration (MA, 30%-60%) and high adulteration (HA, 75%-100%). To avoid the problem of data imbalance during the subsequent data processing, samples of the remaining concentration except pure honey were prepared in five copies separately. Pure honey samples were considered as a separate category and were prepared in 15 copies.

Therefore, 60 samples were prepared for honey mixed with each syrup. A total of 240 samples were prepared in this study, including 60 pure honey samples, 160 adulterated honey samples and 20 pure syrup samples.

All samples were stored in a constant temperature oven at 35 ℃ for at least 24 h to remove air bubbles and crystals in the samples.

Prior to the spectral measurement, the samples were placed in a magnetic stirrer running at a fixed speed for 10 min to ensure that the samples were evenly mixed. All samples were subjected to perform the spectral measurement immediately after completing the above operations.

2.2. Spectra acquisition

All Raman spectra were measured by a micro confocal Raman spectrometer (WITec Alpha300, German) equipped with a 532 nm laser. A 600 g/mm grating and a 50 × 0.75 microscopic objective were utilized for spectra measurement. The integration time of 10 s and the number of accumulations of 3 times were used for spectra acquisition to obtain spectra with outstanding signal-to-noise ratio. The Raman shift range of 250–1530 cm^{-1} was intercepted for subsequent data analysis.

A total of 240 samples were prepared and 10 different positions of each sample were selected to measure spectra. A total of 2,400 Raman spectra were obtained.

2.3. Spectra preprocessing

Although we attempted to keep the external conditions as stable as possible during the spectra collection, slight changes in temperature, humidity, and even cosmic rays can cause small differences in the measured spectra. Therefore, we followed the steps suggested in Yi Xu's research [28] to preprocess the spectra to attenuate noise and interference signals, including Savitzky-Golay (SG) smoothing, standard normal variate (SNV), adaptive iteratively reweighted Penalized Least Squares (airPLS) and min–max normalization.

SG smoothing and SNV were utilized to diminish noise and spectral differences caused by solid particle size and surface scattering, and airPLS was used to eliminate fluorescence background and blackbody radiation. The normalization operation could scale the spectra so that the spectral intensity of all samples was at a similar level, which facilitated subsequent data processing.

2.4. Convolutional neural network

CNN has achieved excellent results for vibrational spectral analysis due to its ability to automatically extract complex feature information from high-dimensional data [26,29]. In general, CNN consists of different functional layers such as convolution layer and pooling layer. The preprocessed Raman spectra are input to the convolution layer, and the convolution operation is performed step by step by sliding the convolution kernels over a specific spectral region to output a set of feature maps with reduced dimensions [30]. The max-pooling layer is located after the convolution layer, and it reduces the feature dimension while preserving local spectral features [31]. In addition, the structural unit combining the convolution and pooling layers can explore and extract the most relevant features from the input data [32], which is beneficial to improve the model performance.

In this study, two groups of CNNs based on Raman spectra were designed. The CNNs belonging to the first group were utilized to analyse the honey samples adulterated with single type of syrup and consisted of four different shallow CNNs. Their architectures and details were presented in Table S2. The second CNN shown in Fig. 1 could identify the honey samples mixed with any type of syrup (four kinds of syrup) based on the spectral data of all samples. For ease of illustration, the two groups of CNNs were abbreviated as CNN_1 and CNN_2.

To avoid overfitting and accelerate network convergence, batch normalization (BN) layer, dropout layer and L2 regularization were applied to two groups of CNNs based on the characteristics of different datasets. The leaky ReLU function implemented the nonlinear transformation of feature maps as an activation function [33]. Afterwards, CNN with the Adam (Adaptive moment estimation) optimization algorithm was used to train the spectral data with category labels. The cross-entropy loss function was defined by Eq. (1), as the error representation of the predicted values of samples.

$$loss = -\sum_{c=1}^{N} \{y_c^t log(y_c^p)\} \quad (1)$$

where c represents a category and N is the total number of categories, y^t and y^p in the form of one-hot vectors respectively denote the true and predicted labels of the samples [34].

The loss function was transferred during the backpropagation process and was utilized to optimize the related parameters including the kernels and weights of different layers. These parameters were updated and modified in the direction of reducing the loss function in successive iterations until the loss function converged to a stable and sufficiently low value, which means that the network training was completed [35].

2.5. Chemometric algorithms

In the identification of honey, the spectral data was divided into calibration and validation sets in a ratio of 8:2 by the Kennard-Stone (K-S) algorithm, which is widely utilized in spectral analysis [36]. A study [37] by Medina showed that PLS-DA and LDA are commonly used classification methods in food fingerprint analysis, and the latter is often applied in conjunction with PCA, while kNN is a simple method suitable for small sample set.

PLS, one of the commonly used chemometric algorithms, was utilized to predict the degree of honey adulteration. It mapped spectral data to a set of low-dimensional latent variables (LVs) and obtained a series of mathematical relationships for predicting the change in adulteration. PLS-DA based on PLS was trained on calibration set with labels in one-hot format, where 1 represented the target category and 0 represented the non-target category. In this study, 0.5 was used as a threshold for classification. In the classification process, PCA-LDA found the best linear projection and acquired the maximum ratio of between-class to within-class variance [37], while kNN achieved the discrimination of samples by calculating the distance between the unknown sample and the known k nearest neighbors.

The optimal parameters for the above chemometric algorithms were determined by 5-fold cross-validation, such as the number of LVs in PLS-DA, the value of k in kNN, etc.

2.6. Characteristic variable selection

When PLS was not effective in predicting the purity of honey under multiple types of syrup adulteration, competitive adaptive reweighted sampling (CARS) based on the "survival of the fittest" principle and variable importance in projection (VIP) were used to select the important spectral variables. CARS can find the combination of key variables with large absolute values of regression coefficients in PLS, and the optimal combination was evaluated and determined by the lowest root mean square error of cross validation [38]. As a well-known method of variable selection, the VIP score reflected the contribution of each variable to the PLS results. Variables corresponding to VIP scores greater than 1.0 were considered as the relevant variables and retained.

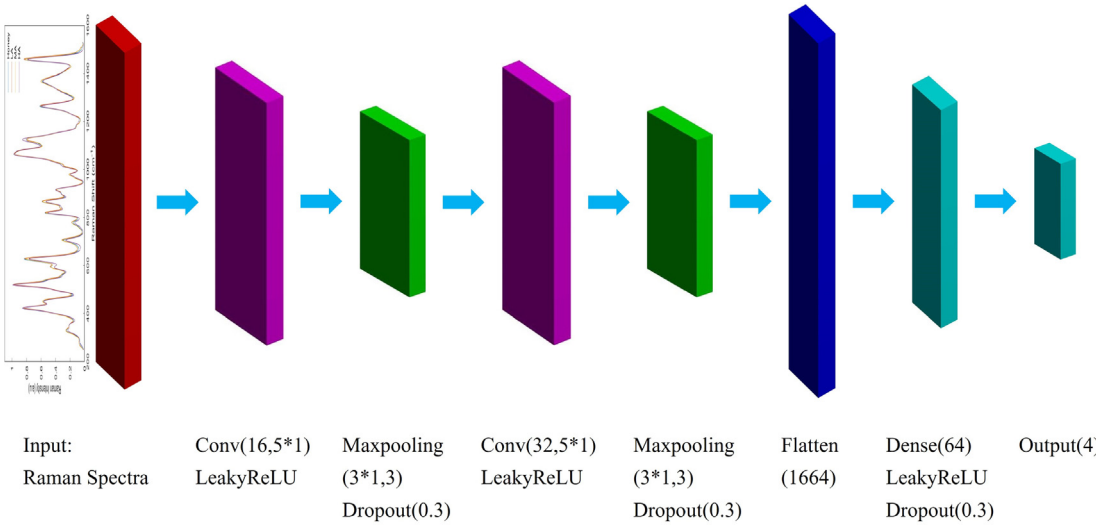


Fig. 1. The architecture of CNN for detecting honey adulterated with any type of syrup.

3. Results and discussion

3.1. Raman spectrum exploratory analysis

The average Raman spectra of natural honey and pure syrup after preprocessing are shown in Fig. 2(a), and the wavenumber range is 250–1530 cm^{-1}. It can be observed that there are very similar spectral characteristics and severe spectral overlap between the spectra of honey and different types of syrup. This is due to the fact that as a supersaturated sugar solution, honey is mainly composed of sugar, and the vibration modes of the functional groups of sugar have a great influence on the characteristic peaks of the Raman spectra of honey [39]. Thus the similar components between honey and syrups used for adulteration lead to obvious resemblance between the spectral peaks.

The Raman spectra of natural honey were mainly composed of bands centered at 328, 424, 452, 519, 705, 629, 821, 866, 1063, 1125, 1265, 1368, 1460 cm^{-1}. The spectral bands in the region of low wavenumbers (including the bands at 328, 424, 452, 519 cm^{-1}) can be attributed to the strong vibration modes of C-C-C, C-C-O, C-O and C-C [25], while the bands at 629, 821, 866 cm^{-1} were associated with ring deformations from fructose, C-O-H bending of carbohydrates and the vibration of C-H and CH_2 deformation, respectively [40]. The remaining two strong signals were located at 1063, 1125 cm^{-1}. The former was related to the vibration modes of C-C, C-O and C-O-H, and the latter was attributed to the combination of vibration modes of C-O and C-O-H in sugars and vibration of C-N in proteins and amino acids [41]. Besides, the existence of the bands centered at 1265, 1368, and 1460 cm^{-1} was associated with the deformation vibration of C-O-H, C-C-H and O-C-H, the symmetrical deformation of CH_2, and the vibration modes of CH_2 and COO^- groups, respectively [39,42].

The average Raman spectra of honey samples adulterated with HFCS are displayed in Fig. 2(b). The Raman peaks at 629, 821, 1265 cm^{-1} in the figure have shown the intensity changes related to the concentration of adulterant. Specifically, the intensity of some Raman peaks increases or decreases as the amount of syrup added increases, which seems to be associated with the introduction of high concentration of fructose in HFCS. At the same time, some similar changes can be observed in the spectra of the samples mixed with other syrups, which can be found in Fig. S1. Fig. S1(d) shows the raw Raman spectra of authentic honey. Compared with samples mixed with the other two syrups, the samples mixed with

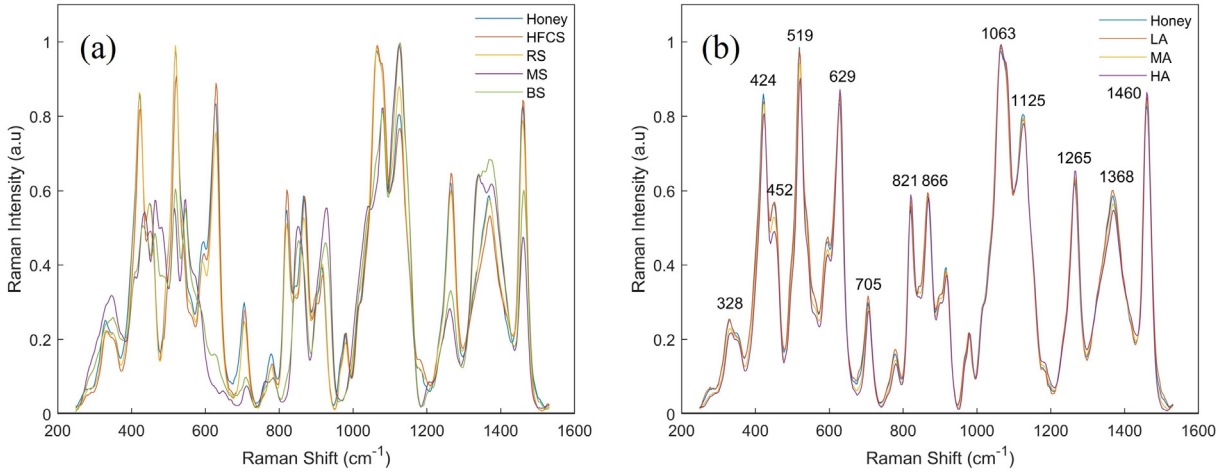


Fig. 2. (a) The preprocessed Raman spectra of natural honey and pure syrups in the range of 250 cm^{-1} to 1530 cm^{-1}. (b) The preprocessed Raman spectra of authentic honey and honey samples adulterated with HFCS.

MS or BS have relatively strong spectral intensity changes and large spectral shifts for different concentration categories, which may be due to the introduction of maltose and sucrose that are less abundant in honey. Although we have noticed that the doping concentration of samples can directly affect the intensity of some spectral bands, these visually observable slight differences alone are not sufficient to ascertain the authenticity of honey. Therefore chemometrics and CNN were introduced for spectral analysis.

The spectral data of honey adulterated with single-variety syrup and the global spectral dataset of all samples were performed PCA simultaneously. Fig. 3(a) and (b) have shown the PCA score plots based on the HFCS-adulterated dataset and the global dataset, and the rest of the score plots are shown in Fig. S2. It is noteworthy that in Fig. 3(a), the data points of samples gradually change from positive to negative along the PC1 axis with the increase of HFCS doping concentration (from Honey, LA, MA to HA), while the score plots of samples mixed with RS or BS also show a similar trend and those corresponding to MS doping samples display an opposite trend along the PC1 axis in Fig. S2. This may be due to the presence of fructose and glucose in HFCS, RS and BS, while the maltose mainly contained in MS is less abundant in honey. The above trends show the great potential of Raman spectroscopy combined with PCA to identify the authenticity of honey.

The above four score plots for samples of single-variety syrup adulteration show outstanding trends of within-class clustering and between-class separation for different concentration categories, but the trends do not seem to apply to the global dataset in Fig. 3 (b). The four classes in the score plot of global dataset are not fully aggregated, and there is also significant overlap of data points between different classes, which may be because it contains all spectral data from multiple syrup adulteration.

When the differences among samples mixed with different types of syrup at the same blending concentration are more obvious than those of samples adulterated with different concentrations of the same syrup, the distribution of samples of the same category becomes more extensive as shown in Fig. 3(b). This is similar to the previous research [43]. This phenomenon indicates that the within-class variance of the global dataset is significantly greater than the between-class variance, which puts forward more stringent requirements for the subsequent data processing algorithms.

3.2. Qualitative analysis of adulterated honey

In this study, we expect to achieve the four-classification task of adulterated honey with high accuracy using one-dimensional CNN models with low complexity. Thus two groups of CNNs (i.e., CNN_1 and CNN_2) were utilized to establish simple models for the detection of honey adulterated with single-variety syrup and a general model that can simultaneously detect samples mixed with any type of syrup, respectively. The performance of CNNs was compared with that of PLS-DA, PCA-LDA and kNN in traditional chemometric algorithms. To comply the 8:2 ratio of the K-S algorithm used to split dataset in chemometrics, 20% of the spectral data were randomly selected as the test set for evaluating the CNNs, and the remaining data were randomly divided into training and validation sets according to the ratio of 8:2 to learn the features in the spectra. Besides, the performance of CNN was evaluated by the overall accuracy of the classification task and the specific accuracy of each adulteration category.

3.2.1. Structure and analysis of CNN models

The structures of CNN_1 and CNN_2 are displayed in Table S2 and Fig. 1. The specific parameters are as follows: the learning rate of all CNNs is 10^{-4} except for that of CNN_1 for HFCS which is set to 10^{-5}. The batch sizes of CNN_1 and CNN_2 are adjusted to 64 and 256, respectively, while the dropout rate and L2 regularization factor of CNN_2 are set to 0.3 and 0.01. These parameters of CNNs are carefully adjusted and determined based on the best classification accuracy.

First, the accuracy curves, loss function curves and confusion matrices of CNN_1 established for the detection of single syrup adulteration can be found in Fig. S3, while the overall accuracy of two groups of CNNs is shown in Table 1. LVs and PCs represent the number of latent variables and principal components, respectively. The Acc represents the classification accuracy. In Fig. S3, it can be observed that the accuracy curves of the training and validation sets of the four CNN models rise steadily until converge to a fixed value close to or equal to 1 with the increase in the number of iterations. The loss function curves show a gradually decreasing trend and finally converge. CNN_1 achieved 100% accuracy for honey doped with MS or BS respectively, which was consistent with the absence of misclassified samples in the corresponding confusion matrices.

The accuracy and loss function curves of the CNN_1 models for HFCS and RS adulteration have stable varying trends and fixed convergence values, despite slight fluctuations. Specifically, accuracy of 99.17% and 97.50% were achieved by CNN_1 for HFCS and RS adulteration on the test set, respectively. The former assigns a HA sample as the MA class, the latter categorizes the two MA samples as LA class and a HA sample as the MA class. The CNN_1 models obtained the accuracy of more than 97% for the adulteration of four types of syrup in honey, which means that CNN combined

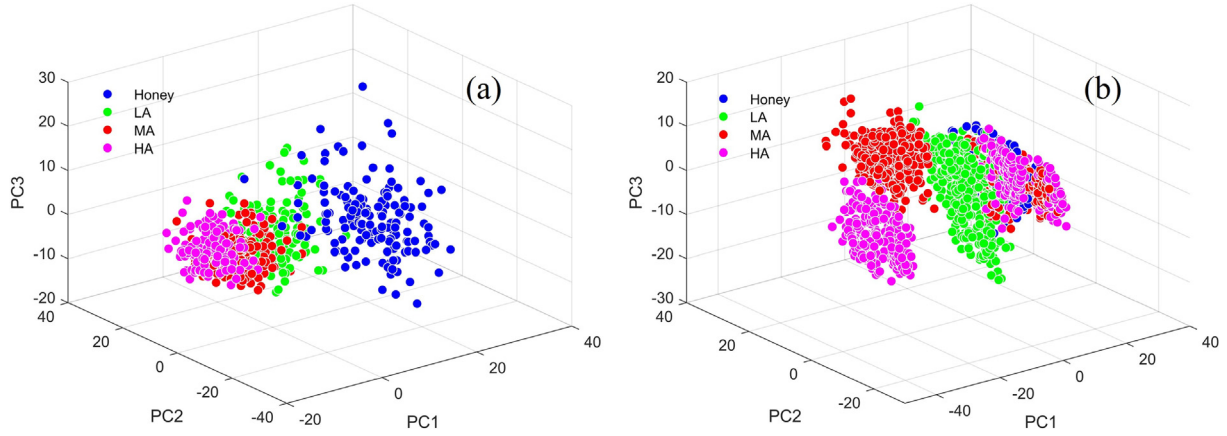


Fig. 3. (a) The PCA score plot of Raman spectra for samples adulterated with HFCS. (b) The PCA score plot of global spectral dataset of all samples.

Table 1
Classification results and corresponding parameters of different algorithms and datasets.

Syrup	PLS-DA		PCA-LDA		kNN		CNN
	LVs	Acc	PCs	Acc	k	Acc	Acc
HFCS	7	86.67%	8	99.17%	7	90.00%	99.17%
RS	5	85.83%	8	89.17%	5	92.50%	97.50%
MS	5	89.17%	5	95.83%	5	100%	100.00%
BS	4	100.00%	3	98.33%	3	100%	100.00%
Global	11	63.96%	13	58.33%	7	79.58%	94.79%

with Raman spectroscopy can identify the honey samples mixed with syrups and indicate the purity level of honey samples to some extent.

In addition, the general model used to identify honey samples adulterated with any variety of syrup, namely CNN_2, its accuracy curve, loss function curve and confusion matrix are shown in Fig. 4. Since the global dataset analysed by CNN_2 have greater within-class variance than the between-class variance, as indicated in the above spectra and PCA score plots, the dropout layers combined with the L2 regularization mechanism were utilized to prevent overfitting while improving model performance. The reason for adopting more dropout layers in the network was that the spectral differences between some samples with the same class label in the global dataset may still be obvious. CNN was more likely to learn some certain features that were only applicable to part of data in the training set, which made it easier to fall into overfitting. However, the existence of dropout layers at different locations can invalidate some of the features passed by the previous layer, and its combination with L2 regularization can force to reduce the weight of some special features, which could well alleviate the overfitting problem and enhance the generalization ability of model.

Although the accuracy and loss function curves shown in Fig. 4 (a) and (b) have slight fluctuations, they maintain stable overall trends and achieve convergence during the 800 iterations. The CNN_2 obtained similar and high accuracy on the training and validation sets, and achieved 94.79% accuracy on the test set, which could be observed in the confusion matrix shown in Fig. 4(c). This indicates that CNN_2 can achieve the identification task of honey adulterated with unknown type of syrup. The main classification errors of CNN_2 for global dataset are: 4 MA samples are classified as pure honey as well as 6 MA samples are assigned to LA class, while 6 LA samples are considered as MA. It can be noted that there are more cases of misclassification between adjacent categories, which is consistent with the common sense that samples are likely to be assigned to similar categories.

Besides, the fluctuations in the accuracy and loss function curves of CNN_2 were likely related to the dropout layers in the network. The dropout layers made the information learned by the network incoherent and incomplete while invalidating some features, which caused the network to oscillate slightly during the training process. But in general, CNN_2 still obtained a prominent accuracy as the general model for honey identification.

3.2.2. The comparison with chemometric algorithms

To illustrate the potential of CNN combined with Raman spectroscopy for honey identification, several common classification algorithms including PLS-DA, PCA-LDA and kNN were utilized for comparison. The classification results of each algorithm for different datasets and the corresponding optimal parameters were compared and displayed in Table 1. Intuitively, each algorithm and CNN performed the best for BS-doped samples, can achieve more than 98% accuracy, while kNN and CNN were also able to achieve 100% discrimination for MS-doped samples. For honey samples adulterated with HFCS, PCA-LDA with nine principal components and CNN obtained an accuracy of 99.17%, while the best PLS-DA and kNN had the accuracy around 90%. The identification of RS-doped honey has always been a difficult task, with PLS-DA and PCA-LDA failing to achieve even 90% correct classification and kNN with k of five only obtaining 92.50% accuracy. However, surprisingly, CNN had the excellent classification ability for this dataset, obtaining an accuracy of 97.50%.

The performance of the above four different algorithms for classifying samples adulterated with single syrup could also be intuitively observed and compared in Fig. S4(a). Although some chemometric algorithms achieved the same accuracy as CNN for different datasets, it is worth noting that not a certain classification algorithm had the comparable performance as CNN for honey samples respectively adulterated with four single-variety syrups. For example, PCA-LDA achieved the best accuracy for HFCS-adulterated samples while the chemometric algorithm that achieved the best classification for RS-doped samples was kNN. However, CNN achieved equivalent or even the best discrimination results for the honey respectively mixed with four single syrups.

Finally, different algorithms were used to analyse the global dataset, and its identification was the most difficult classification

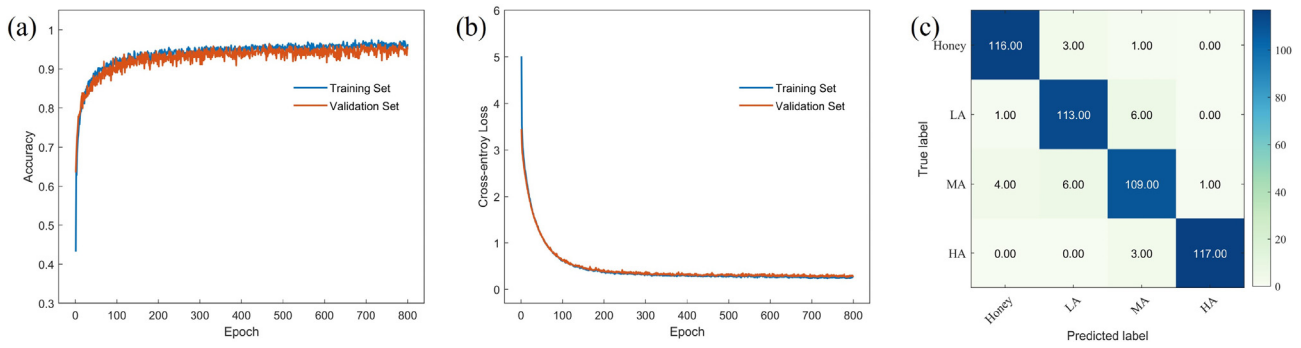


Fig. 4. The accuracy curve (a), loss function curve (b) and confusion matrix (c) of the general model (i.e., CNN_2) used to analyse the global dataset for multiple syrup adulteration.

task. The classification accuracy of chemometric algorithms was illustrated in Table 1 and Fig. S4(b). For the global dataset, kNN with k of seven was the most effective traditional chemometric algorithm and it could only achieve 79.58% accuracy, while CNN_2 improved the accuracy to an outstanding level of 94.79%. Overall, CNN achieved the high accuracy both for the detection of single syrup adulteration and for comprehensive recognition of adulterated honey regardless of the type of syrup. The performance of CNN was greatly superior to other algorithms.

PLS-DA is implemented based on PLS, which maps spectral data to a set of low-dimensional LVs before a linear discriminant classifier is used for classification. Its main advantages are fast calculation speed and high accuracy, while the main limitation of PLS-DA is that it is strongly influenced by different category sizes and requires suitable LVs [44]. It is important to stress that PLS-DA itself is a binary classifier, and the category labels of samples need to be encoded in one-hot form (binary matrix) when performing multi-classification tasks. In addition, LDA can find the best linear projection on the variables and acquire the maximum ratio of between-class to within-class variance. But the main limitation of LDA is that it requires a lower or equal number of spectral variables and samples [37]. Therefore, in this study, PCA was combined with LDA for spectral analysis when the number of spectral variables was significantly larger than the number of samples.

kNN calculates the distance between the unknown samples and known nearest neighbors in the feature space to achieve classification. The main advantage of kNN is that it does not require a particular ratio of quantities between the samples and spectral variables, and does not need to consider the probability distribution of samples during the analysis [45]. However, kNN is also susceptible to the influence of the class size of samples. A sample is more likely to tend to a category with larger number of samples when it is analysed. Besides, the random spectral noise has a significant impact on the classification results of kNN.

In this study, kNN achieved a higher classification accuracy for Raman spectral data of honey compared to the remaining two chemometric algorithms. This may be due to the fact that the number of spectra in the four concentration classes is consistent in this study, and the spectral noise and fluorescence interference are attenuated during the preprocessing process. This makes kNN relatively suitable for the classification task of honey adulteration level, especially for the global dataset. But overall, CNN still achieved the best performance in Table 1 regardless of whether the type of adulterant was exact or unknown.

Moreover, although previous studies [46] have shown that CNN has the ability to extract important features from the original spectra, this is achieved at the cost of adding the network structures such as convolution layers equivalent to the preprocessing steps. In fact, no matter what data analysis algorithm is used, a suitable preprocessing strategy can help the model extract important information to improve its performance, but CNN is relatively less dependent on the preprocessing steps compared to chemometric algorithms [47]. In this study, the preprocessed spectra were fed into the CNN and other chemometric algorithms, mainly because Raman spectra are susceptible to external interference and have a strong fluorescence background [28]. The input of preprocessing spectra can reduce the training resources and model complexity required by CNN, which is beneficial for building shallow CNNs and obtaining high accuracy.

3.3. Quantitative analysis of adulterated honey

In this section, Raman spectroscopy combined with PLS algorithm was used to realize the quantification of honey purity. Five samples at each concentration were prepared. The average spectrum of the ten spectra measured at different positions of each sample was considered as the unique spectral representation of this sample, thus was utilized to analyse the degree of adulteration of each sample. All spectra of selected samples were divided into the calibration and validation sets by the K-S algorithm in a ratio of 3:2. Table 2 combined with Fig. S5 and Fig. S6 have shown the prediction results of simple models with the detection of single syrup adulteration and the general model with simultaneous detection of multiple syrup adulteration. The nVAR represents the number of selected variables input into the PLS models in Table 2.

PLS with different LVs showed excellent results for honey samples doped with single-variety syrup, the determination coefficients of prediction (R_p^2) were all above 0.98, and the root mean square error of prediction (RMSEP) maintained below 3.50 for samples mixed with HFCS or RS, while around 1.0 for samples adulterated with MS or BS. The average error of the validation set also retained within 3.0, which intuitively represented the average absolute value of the sample bias. In comparison, the fitting results of the samples adulterated with HFCS or RS were slightly weaker than those for the MS or BS, which was related to the too similar composition between the first two syrups (HFCS, RS) and authentic honey. The actual and predicted doping concentrations for each sample in the validation set were elucidated in Fig. S4, and a brilliant fitting trend between the actual and predicted values can be observed.

For the general model, the results obtained by PLS of 10 LVs were not satisfactory, with R_p^2 = 0.9068 and RMSEP = 10.5445. The results presented in Fig. S6(a) and Table 2 showed that although the concentration values predicted by PLS had an exact tendency close to the actual values to some extent, the excessive root mean square error and low coefficient of determination proved that PLS was not able to well predict the purity level of adulterated honey without considering the type of syrup. This may be due to the fact that for multi-concentration datasets containing spectra of samples respectively mixed with different types of syrup, PLS may be more susceptible to the influence of different syrup types while modelling the spectrum and concentration label information, thereby reducing the fitting effect. This was similar to the situation where the chemometric algorithms did not perform well in the classification of the global dataset.

Thus CARS and VIP were introduced to attempt to optimize the quantification results of PLS. CARS, as a superior variable selection algorithm, was run 50 times in this study to select the best combination of key wavenumbers, while VIP selected the important variables by calculating the score of each variable. The frequency graph of the key wavenumbers selected by CARS in multiple runs and the indicative plot of the VIP scores were displayed separately in Fig. S7(a) and (b). When certain wavenumbers had higher selected frequencies or larger VIP scores, it indicated that these wavenumbers were more important for the concentration prediction. The spectral peaks centered at 519, 629, 1460 cm^{-1} were considered by both CARS and VIP as important bands for quantifying the doping concentration, and these bands seemed to be very relevant with the concentration of fructose contained in samples. The two algorithms selected some different key variables due to the different operating principles.

However, the results obtained from PLS combined with CARS or VIP shown in Fig. S6(b) and (c) were not superior to the performance of PLS based on continuous spectra in Fig. S6(a), because R_p^2 and RMSEP were not improved. Although the variable selection methods can effectively simplify models, it does not necessarily improve the performance of models [48]. The poor effect of PLS combined with CARS or VIP in this study may be because PLS itself could not well extract the inherent information related to the concentration quantification in spectra under the condition of multiple syrup adulteration, while CARS and VIP were likely to lose some of key information in the process of selecting variables.

Table 2
PLS quantification results for the detection of single and multiple syrup adulteration in honey samples.

Syrup	Models	nVAR	LVs	Calibration		Validation		
				R^2c	RMSEC	R^2p	RMSEP	Average error
HFCS	PLS	486	6	0.9997	0.6270	0.9910	3.2633	2.6410
RS	PLS	486	5	0.9966	1.9949	0.9897	3.4890	2.8930
MS	PLS	486	6	0.9998	0.4708	0.9991	1.0583	0.8817
BS	PLS	486	10	1.0000	0.0308	0.9992	0.9750	0.7031
Global	PLS	486	10	0.9753	5.4111	0.9068	10.5445	8.4751
	CARS-PLS	85	9	0.9693	6.0286	0.9008	10.8683	8.3519
	VIP-PLS	164	9	0.9448	8.0908	0.8774	12.1137	9.4729

In fact, PLS combined with Raman spectroscopy proved to be successful for the quantitative prediction of honey samples adulterated with single syrup. In the quantification of honey samples without considering the type of adulterant, although PLS itself or combined with CARS or VIP performed poorly, the established CNN model was able to successfully classify the adulterated honey into four categories by the adulteration concentration. This can be regarded as an alternative solution to the regression task for the detection of honey adulteration to some extent.

4. Conclusion

In this study, Raman spectroscopy combined with CNN and chemometric algorithms have been proven to be well implemented for the quality control of honey samples. Compared with the PLS-DA, PCA-LDA and kNN algorithms, the established CNN models achieved more than 97% accuracy for honey samples adulterated with single-variety syrup and an outstanding accuracy of more than 94% for honey samples adulterated with any type of syrup. CNN has achieved the best performance in the classification task, whether based on specific accuracy for each category or overall accuracy. When PLS was utilized to predict the adulterant concentration in honey samples mixed with single syrup, the four individual models achieved prominent results, which met the requirements of conventional food quality control (R_p^2 greater than 0.98, RMSEP less than 3.50). Although the PLS model can not predict the purity of honey adulterated with any type of syrup, the CNN_2 model was able to classify the adulterated honey into four categories according to the adulteration concentration, which could be regarded as an alternative solution to the regression task to a certain extent.

The green, rapid, and nondestructive spectroscopy analysis method established in this article can achieve the identification and quantification of honey, which is of great significance for food safety and quality control of honey products. To obtain models with higher accuracy and more robust generalization performance, a wider dataset including the spectra of multiple honey respectively mixed with different types of adulterant will be collected and analysed. Future attempts will be made to utilize the CNN models with different structures to achieve identification and quantification of honey regardless of the type of honey or adulterants.

CRediT authorship contribution statement

Xijun Wu: Project administration, Funding acquisition, Data curation, Conceptualization, Methodology, Writing – original draft, Validation. **Baoran Xu:** Conceptualization, Investigation, Methodology, Software, Formal analysis, Writing – original draft, Writing – review & editing. **Renqi Ma:** Visualization, Formal analysis, Writing – review & editing. **Yudong Niu:** Visualization, Methodology, Writing – review & editing. **Shibo Gao:** Visualization, Methodology, Writing – review & editing. **Hailong Liu:** Supervision. **Yungang Zhang:** Supervision.

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.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (grant number 11674275, 62175208), Natural Science Foundation of Hebei Province (grant number F2020203110, F2021203052), the Central Government Guides Local Science and Technology Development Foundation (grant number 216Z1701G), Science and technology research project of Hebei higher education institutions (grant number QN2018071).

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.saa.2022.121133.

References

[1] A.J. Siddiqui, S.G. Musharraf, M.I. Choudhary, A.U. Rahman, Application of analytical methods in authentication and adulteration of honey, Food Chem. 217 (2017) 687–698.
[2] A.A. Machado De-Melo, L.B.d. Almeida-Muradian, M.T. Sancho, A. Pascual-Maté, Composition and properties of Apis mellifera honey: A review, J. Apic. Res. 57 (1) (2018) 5–37.
[3] Z. Wang, P. Ren, Y. Wu, Q. He, Recent advances in analytical techniques for the detection of adulteration and authenticity of bee products - A review, Food Addit. Contam, Part A Chem. Anal. Control Expo Risk Assess 38 (2021) 533–549.
[4] S.U. Khan, S.I. Anjum, K. Rahman, M.J. Ansari, W.U. Khan, S. Kamal, B. Khattak, A. Muhammad, H.U. Khan, Honey: Single food stuff comprises many drugs, Saudi, J Biol Sci 25 (2018) 320–325.
[5] L. Wu, B. Du, Y. Vander Heyden, L. Chen, L. Zhao, M. Wang, X., Xue, Recent advancements in detecting sugar-based adulterants in honey – A challenge, TrAC, Trends Anal. Chem. 86 (2017) 25–38.
[6] M.J. Aliano-Gonzalez, M. Ferreiro-Gonzalez, E. Espada-Bellido, M. Palma, G.F. Barbero, A screening method based on Visible-NIR spectroscopy for the identification and quantification of different adulterants in high-quality honey, Talanta 203 (2019) 235–241.
[7] M. Petersen, Z. Yu, X. Lu, Application of Raman Spectroscopic Methods in Food Safety: A Review, Biosensors (Basel) 11 (2021).
[8] E.-I. Geană, C.T. Ciucure, D. Costinel, R.E. Ionete, Evaluation of honey in terms of quality and authenticity based on the general physicochemical pattern, major sugar composition and δ13C signature, Food Control 109 (2020) 106919, https://doi.org/10.1016/j.foodcont.2019.106919.
[9] M. Tosun, Detection of adulteration in honey samples added various sugar syrups with 13C/12C isotope ratio analysis method, Food Chem. 138 (2013) 1629–1632.
[10] S. Wang, Q. Guo, L. Wang, L. Lin, H. Shi, H. Cao, B. Cao, Detection of honey adulteration with starch syrup by high performance liquid chromatography, Food Chem. 172 (2015) 669–674.
[11] X. Xue, Q. Wang, Y. Li, L. Wu, L. Chen, J. Zhao, F. Liu, 2-acetylfuran-3-glucopyranoside as a novel marker for the detection of honey adulterated with rice syrup, J. Agric. Food Chem. 61 (2013) 7488–7493.

[12] A.I. Ruiz-Matute, S. Rodríguez-Sánchez, M.L. Sanz, I. Martínez-Castro, Detection of adulterations of honey with high fructose syrups from inulin by GC analysis, J. Food Compos. Anal. 23 (2010) 273–276.
[13] A.I. Ruiz-Matute, M. Weiss, D. Sammataro, J. Finely, M.L. Sanz, Carbohydrate composition of high-fructose corn syrups (HFCS) used for bee feeding: effect on honey composition, J. Agric. Food Chem. 58 (12) (2010) 7317–7322.
[14] S. Yan, M. Song, K. Wang, X. Fang, W. Peng, L. Wu, X. Xue, Detection of acacia honey adulteration with high fructose corn syrup through determination of targeted alphaDicarbonyl compound using ion mobility-mass spectrometry coupled with UHPLC-MS/MS, Food Chem. 352 (2021) 129312.
[15] C. He, Y. Liu, H. Liu, X. Zheng, G. Shen, J. Feng, Compositional identification and authentication of Chinese honeys by (1)H NMR combined with multivariate analysis, Food Res. Int. 130 (2020) 108936.
[16] F. Huang, H. Song, L. Guo, P. Guang, X. Yang, L. Li, H. Zhao, M. Yang, Detection of adulteration in Chinese honey using NIR and ATR-FTIR spectral data fusion, Spectrochim Acta A Mol Biomol Spectrosc 235 (2020) 118297, https://doi.org/10.1016/j.saa.2020.118297.
[17] Q. Li, J. Zeng, L. Lin, J. Zhang, J. Zhu, L.u. Yao, S. Wang, J. Du, Z. Wu, Mid-infrared spectra feature extraction and visualization by convolutional neural network for sugar adulteration identification of honey and real-world application, Lwt 140 (2021) 110856, https://doi.org/10.1016/j.lwt.2021.110856.
[18] M. Izquierdo, M. Lastra-Mejías, E. González-Flores, J.C. Cancilla, M. Pérez, J.S. Torrecilla, Convolutional decoding of thermographic images to locate and quantify honey adulterations, Talanta 209 (2020) 120500, https://doi.org/10.1016/j.talanta.2019.120500.
[19] Z. Gan, Y. Yang, J. Li, X. Wen, M. Zhu, Y. Jiang, Y. Ni, Using sensor and spectral analysis to classify botanical origin and determine adulteration of raw honey, J. Food Eng. 178 (2016) 151–158.
[20] M. Bougrini, K. Tahri, T. Saidi, N. El Alami El Hassani, B. Bouchikhi, N. El Bari, Classification of Honey According to Geographical and Botanical Origins and Detection of Its Adulteration Using Voltammetric Electronic Tongue, Food Analytical Methods, 9 (2016) 2161-2173
[21] S. Lohumi, S. Lee, H. Lee, B.-K. Cho, A review of vibrational spectroscopic techniques for the detection of food authenticity and adulteration, Trends Food Sci. Technol. 46 (2015) 85–98.
[22] X. Wu, Z. Zhao, R. Tian, Z. Shang, H. Liu, Identification and quantification of counterfeit sesame oil by 3D fluorescence spectroscopy and convolutional neural network, Food Chem. 311 (2020) 125882.
[23] T.A. Teklemariam, J. Moisey, J. Gotera, Attenuated Total Reflectance-Fourier transform infrared spectroscopy coupled with chemometrics for the rapid detection of coconut water adulteration, Food Chem. 355 (2021) 129616.
[24] M. Esteki, Z. Shahsavari, J. Simal-Gandara, Use of spectroscopic methods in combination with linear discriminant analysis for authentication of food products, Food Control 91 (2018) 100–112.
[25] B. Ozbalci, I.H. Boyaci, A. Topcu, C. Kadilar, U. Tamer, Rapid analysis of sugars in honey by processing Raman spectrum using chemometric methods and artificial neural networks, Food Chem. 136 (2013) 1444–1452.
[26] Y. Liu, H. Pu, D.-W. Sun, Efficient extraction of deep image features using convolutional neural network (CNN) for applications in detecting and analysing complex food matrices, Trends Food Sci. Technol. 113 (2021) 193–204.
[27] Q. Li, J. Zeng, L. Lin, J. Zhang, J. Zhu, L. Yao, S. Wang, Z. Yao, Z. Wu, Low risk of category misdiagnosis of rice syrup adulteration in three botanical origin honey by ATR-FTIR and general model, Food Chem. 332 (2020) 127356.
[28] Y. Xu, P. Zhong, A. Jiang, X. Shen, X. Li, Z. Xu, Y. Shen, Y. Sun, H. Lei, Raman spectroscopy coupled with chemometrics for food authentication: A review, TrAC, Trends Anal. Chem. 131 (2020).
[29] X. Zhang, T. Lin, J. Xu, X. Luo, Y. Ying, DeepSpectra: An end-to-end deep learning approach for quantitative spectral analysis, Anal. Chim. Acta 1058 (2019) 48–57.
[30] Q. Chai, J. Zeng, D. Lin, X. Li, J. Huang, W. Wang, Improved 1D convolutional neural network adapted to near-infrared spectroscopy for rapid discrimination of Anoectochilus roxburghii and its counterfeits, J. Pharm. Biomed. Anal. 199 (2021) 114035.
[31] J. Yang, X. Wang, R. Wang, H. Wang, Combination of Convolutional Neural Networks and Recurrent Neural Networks for predicting soil properties using Vis–NIR spectroscopy, Geoderma 380 (2020) 114616, https://doi.org/10.1016/j.geoderma.2020.114616.
[32] M. Izquierdo, M. Lastra-Mejías, E. González-Flores, S. Pradana-López, J.C. Cancilla, J.S. Torrecilla, Visible imaging to convolutionally discern and authenticate varieties of rice and their derived flours, Food Control 110 (2020) 106971, https://doi.org/10.1016/j.foodcont.2019.106971.
[33] D. Ma, L. Shang, J. Tang, Y. Bao, J. Fu, J. Yin, Classifying breast cancer tissue by Raman spectroscopy with one-dimensional convolutional neural network, Spectrochim Acta A Mol Biomol Spectrosc 256 (2021) 119732.
[34] M. Fukuhara, K. Fujiwara, Y. Maruyama, H. Itoh, Feature visualization of Raman spectrum analysis with deep convolutional neural network, Anal. Chim. Acta 1087 (2019) 11–19.
[35] D. Rong, H. Wang, Y. Ying, Z. Zhang, Y. Zhang, Peach variety detection using VIS-NIR spectroscopy and deep learning, Comput. Electron. Agric. 175 (2020) 105553, https://doi.org/10.1016/j.compag.2020.105553.
[36] Z. Lan, Y. Zhang, Y. Sun, S. Ji, T. Wang, H. Lu, J.M. Cao, Rapid quantitative detection of the discrepant compounds in differently processed Curcumae Rhizoma products by FT-NIR combined with VCPA-GA technology, J. Pharm. Biomed. Anal. 195 (2021) 113837.
[37] S. Medina, R. Perestrelo, P. Silva, J.A.M. Pereira, J.S. Câmara, Current trends and recent advances on food authenticity technologies and chemometric approaches, Trends Food Sci. Technol. 85 (2019) 163–176.
[38] H. Li, Y. Liang, Q. Xu, D. Cao, Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration, Anal. Chim. Acta 648 (2009) 77–84.
[39] D.P. Aykas, M.-L. Shotts, L.E. Rodriguez-Saona, Authentication of commercial honeys based on Raman fingerprinting and pattern recognition analysis, Food Control 117 (2020) 107346, https://doi.org/10.1016/j.foodcont.2020.107346.
[40] O. Anjos, A.J.A. Santos, V. Paixao, L.M. Estevinho, Physicochemical characterization of Lavandula spp. honey with FT-Raman spectroscopy, Talanta 178 (2018) 43–48.
[41] D.A. Magdas, F. Guyon, C. Berghian-Grosan, C. Muller Molnar, Challenges and a step forward in honey classification based on Raman spectroscopy, Food Control 123 (2021) 107769, https://doi.org/10.1016/j.foodcont.2020.107769.
[42] H.E. Tahir, Z. Xiaobo, L. Zhihua, S. Jiyong, X. Zhai, S. Wang, A.A. Mariod, Rapid prediction of phenolic compounds and antioxidant activity of Sudanese honey using Raman and Fourier transform infrared (FT-IR) spectroscopy, Food Chem. 226 (2017) 202–211.
[43] X. Yang, P. Guang, G. Xu, S. Zhu, Z. Chen, F. Huang, Manuka honey adulteration detection based on near-infrared spectroscopy combined with aquaphotomics, Lwt 132 (2020) 109837, https://doi.org/10.1016/j.lwt.2020.109837.
[44] C.L.M. Morais, M. Paraskevaidi, L. Cui, N.J. Fullwood, M. Isabelle, K.M.G. Lima, P. L. Martin-Hirsch, H. Sreedhar, J. Trevisan, M.J. Walsh, D. Zhang, Y.G. Zhu, F.L. Martin, Standardization of complex biologically derived spectrochemical datasets, Nat. Protoc. 14 (2019) 1546–1577.
[45] C.L.M. Morais, K.M.G. Lima, M. Singh, F.L. Martin, Tutorial: multivariate classification for vibrational spectroscopy in biological samples, Nat. Protoc. 15 (2020) 2143–2162.
[46] X. Zhang, J. Xu, J. Yang, L. Chen, H. Zhou, X. Liu, H. Li, T. Lin, Y. Ying, Understanding the learning mechanism of convolutional neural networks in spectral analysis, Anal. Chim. Acta 1119 (2020) 41–51.
[47] J. Acquarelli, T. van Laarhoven, J. Gerretzen, T.N. Tran, L.M.C. Buydens, E. Marchiori, Convolutional neural networks for vibrational spectroscopic data analysis, Anal. Chim. Acta 954 (2017) 22–31.
[48] L. Li, S. Jin, Y. Wang, Y. Liu, S. Shen, M. Li, Z. Ma, J. Ning, Z. Zhang, Potential of smartphone-coupled micro NIR spectroscopy for quality control of green tea, Spectrochim Acta A Mol Biomol Spectrosc 247 (2021) 119096.