Current Research in Food Science 8 (2024) 100782

Contents lists available at ScienceDirect

Current Research in Food Science

journal homepage: www.sciencedirect.com/journal/current-research-in-food-science

Discriminative feature analysis of dairy products based on machine learning algorithms and Raman spectroscopy

Jia-Xin Li [a], Chun-Chun Qing [a], Xiu-Qian Wang [b], Mei-Jia Zhu [c], Bo-Ya Zhang [a], Zheng-Yong Zhang [a,*]

[a] School of Management Science and Engineering, Nanjing University of Finance and Economics, Nanjing, Jiangsu, 210023, PR China
[b] School of Accounting, Nanjing University of Finance and Economics, Nanjing, Jiangsu, 210023, PR China
[c] School of Applied Mathematics, Nanjing University of Finance and Economics, Nanjing, Jiangsu, 210023, PR China

ARTICLE INFO

Handling Editor: Professor Aiqian Ye

Keywords:
Feature analysis
Machine learning
Raman spectroscopy
Classification discrimination
Chemometrics

ABSTRACT

Discriminant analysis of similar food samples is an important aspect of achieving food quality control. The effective combination of Raman spectroscopy and machine learning algorithms has become an extremely attractive approach to develop intelligent discrimination techniques. Feature spectral analysis can help researchers gain a deeper understanding of the data patterns in food quality discrimination. Herein, this work takes the discrimination of three brands of dairy products as an example to investigate the Raman spectral feature based on the support vector machines (SVM), extreme learning machines (ELM) and convolutional neural network (CNN) algorithms. The results show that there are certain differences in the optimal spectral feature interval corresponding to different machine learning algorithms. Selecting the appropriate spectral feature interval can maintain high recognition accuracy and improve the computational efficiency of the algorithm. For example, the SVM algorithm has a recognition accuracy of 100% in the 890-980 cm^{-1}, 1410-1500 cm^{-1} fusion spectral range, which takes about 200 s. The ELM algorithm also has a recognition accuracy of 100% in the 890-980 cm^{-1}, 1410-1500 cm^{-1} fusion spectral range, which takes less than 0.3 s. The CNN algorithm has a recognition accuracy of 100% in the 890-980 cm^{-1}, 1050-1180 cm^{-1}, 1410-1500 cm^{-1} fusion spectral range, which takes about 80 s. In addition, by analyzing the distribution of spectral feature intervals based on Euclidean distance, the distribution of experimental samples based on feature spectra is visually displayed. Through the spectral feature analysis process of similar samples, a set of analysis strategies is provided to deeply reveal the data foundation of classification algorithms, which can provide reference for the analysis of relevant discriminative research patterns.

1. Introduction

As an important component of food, dairy products can be classified into two types of quality and safety risks. One is harmful substances, including illegal additives, heavy metals, harmful toxins, pesticides, veterinary drugs, and antibiotic residues, etc. (Ranveer et al., 2023; Shan et al., 2023) The other is counterfeit products, and all of their testing indicators may meet the requirements of national standards. However, violators can profit from the price differences between various brands or origins (Pan et al., 2024, Zheng-Yong et al., 2017). There are now various detection and control strategies available for these two different forms of risk. For molecules with clear characteristics, such as melamine and sodium thiocyanate, component analysis methods such as chromatography-mass spectrometry and surface enhanced Raman spectroscopy can be employed to identify the target molecule and perform quantitative analysis (Andrey et al., 2023, Zheng-Yong et al., 2015). This strategy mainly aims to identify target molecules in the classification and identification of dairy products. For example, Wang et al. studied the differences in oligosaccharide profiles between caprine and bovine dairy products, detected 27 types of oligosaccharides, and used principal component analysis for sample classification to identify Lacto-N-triose as a potential biomarker for distinguishing caprine milk from bovine milk(Wang et al., 2024). Huang et al. conducted a study on the identification of exogenous protein adulteration in milk powder using laser induced breakdown spectroscopy combined with linear discriminant analysis (LDA), k-nearest neighbor (KNN), random forest

* Corresponding author.
E-mail addresses: zyzhang@nufe.edu.cn, zyzhangnjue@126.com (Z.-Y. Zhang).

https://doi.org/10.1016/j.crfs.2024.100782
Received 7 February 2024; Received in revised form 18 May 2024; Accepted 5 June 2024
Available online 7 June 2024

(RF), support vector machine (SVM), and convolutional neural network (CNN). The CNN model showed good performance, with an average accuracy of 97.8% (Huang et al., 2022). Yang et al. developed a surface enhanced Raman spectroscopy based on magnetic substrates combined with machine learning algorithms (KNN, SVM, decision tree(DT)) to achieve ultra-trace detection of quinolone antibiotics in dairy products (Yang et al., 2023). For similar samples, without target molecules, and belonging to dairy product samples with diverse and complex components, holistic discriminant analysis techniques can be used (Gouvêa et al., 2021). For example, Zheng-Yong et al. established a brand identification method for fresh dairy products based on multidimensional Raman spectroscopy (Zheng-Yong et al., 2021). Singh et al. systematically summarized the comprehensive evaluation application of machine learning models in the pasteurization process of dairy products, pointing out that machine learning methods can effectively monitor the pasteurization process in real time, predict potential equipment failures, improve process efficiency, and evaluate the quality of pasteurization products (Singh et al., 2024). The combination of machine learning algorithms and spectral representation has become one of the most significant research and development directions in this field.

Machine learning algorithms can efficiently utilize the characterization data of dairy products and quickly obtain discrimination results. For example, Yiwei et al. conducted a research on the identification of milk fat cream and non-dairy cream using rapid evolutionary ionization mass spectroscopy combined with machine learning algorithms (DA, DT, SVM, and neural network classifiers). Through hyperparameter optimization and feature engineering, the recognition accuracy reached 98.4–99.6% (Yiwei et al., 2023). Zikang et al. used Raman spectroscopy combined with light gradient boosting machine (LightGBM), SVM, RF, and extreme gradient boosting (XGBoost). The study showed that under single algorithm conditions, the accuracy of brand classification for dairy products exceeded 90%, and when these algorithms were combined and coordinated, the accuracy could reach 99% (Zikang et al., 2024). However, there are also some issues that urgently need improvement, such as spectral feature analysis (Ji et al., 2023; Xue et al., 2023). For machine learning discriminative algorithms, they often act as a black box with limited interpretability. Therefore, it is still necessary to conduct in-depth research on whether the feature intervals representing data have an impact on related algorithms, and how to further improve efficiency to enhance people's understanding of such problems (Giulia et al., 2022). Pu et al. have summarized the commonly used feature construction methods for hyperspectral imaging, and also pointed out that more spectral information analytical methods still need to be continuously explored (Pu et al., 2023). Raman spectroscopy, as a cutting-edge technology for characterizing molecular vibration information, has received widespread attention in the field of dairy product analysis and has shown strong application potential (Wang et al., 2021). For instance, Khan et al. combined Raman spectroscopy with partial least squares regression (PLSR) model to demonstrate its potential in online monitoring of raw milk (Khan et al., 2023).

Herein, this work conducts research on spectral features based on machine learning algorithms. Among them, SVM is an excellent classifier for high-dimensional data classification, extreme learning machine (ELM) is a single hidden layer feedforward neural network algorithm with high learning efficiency and generalization ability, CNN is a new type of neural network algorithm with strong feature extraction ability and excellent generalization ability. Therefore, by combining these three distinctive machine learning algorithms with Raman spectroscopy, the relationship between Raman features and algorithms were investigated and explored.

2. Experimental section

2.1. Samples and equipment

The dairy product samples selected for the experiment were purchased from different manufacturers at Suguo Supermarket in Nanjing, China. Among them, Dingxin dairy products were produced by Heilongjiang Zhaodong Tianlong Dairy Co. Ltd. and labeled as brand 1, Puzhen dairy products were produced by Inner Mongolia Yinuo Halal Food Co. Ltd. and labeled as brand 2, and Xueyuan dairy products were produced by Inner Mongolia Wulanchabu City Jining Xueyuan Dairy Co. Ltd. and labeled as brand 3. There were 40 samples of each brand.

The Raman spectra of dairy samples were collected using a portable laser Raman spectrometer (ProTT-EZRaman-D3, Enwave Optronics, Irvine, CA, USA) and baseline calibration was performed using the instrument's built-in control software. Each dairy sample was placed in a powder state in a small hole of a 96 well plate, and then the Raman spectrometer laser probe was placed closely above the sample for laser irradiation and sample signal collection. The spectral acquisition parameters included laser wavelength of 785 nm, laser power of about 450 mW, charge-coupled device temperature of −85 °C, laser exposure time of 50 s, spectral resolution of 1 cm^{-1}, spectral range of 250–2339 cm^{-1}, and one spectrum was collected for each sample. The ambient temperature was about 20 °C and the humidity was about 50%. Raman spectra were collected directly without any chemical pretreatment.

2.2. Data processing

The SVM, ELM, CNN algorithms, normalization and Euclidean distance calculation involved in the experiment were all implemented using the Matlab platform (MathWorks, Natick, MA, U.S.A.). The relevant MATLAB calculation program can be found in the supporting documents. The computing platform was a personal laptop, configured as central processing unit (CPU) Intel(R) Core(TM) i5-8250U CPU@1.60 GHz 1.80 GHz, and random access memory 24.0 GB.

2.2.1. SVM algorithm

A brief introduction to the relevant algorithms is as follows. For the SVM algorithm, assume a training set (T) at first, $T = \{(x_1, y_2), \cdots, (x_m, y_m)\} \in (X \times Y)^m$, where $x_i \in X = R^n, y_i \in Y = \{1, -1\} (i = 1,2,\cdots,m)$, as well as x_i is the Raman spectral data, and y_i is the brand category label for dairy products. In classification operations, the multi classification problem is transformed into a binary classification problem between any two types of samples for calculation. Choose the appropriate kernel function $K(x, x')$ and parameter C, construct and solve the optimization problem $\min_{\alpha} \frac{1}{2} \sum_{i=1}^{j} \sum_{j=1}^{m} y_i y_j \alpha_i \alpha_j K(x_i, x_j) - \sum_{j=1}^{m} \alpha_j, \ s.t. \sum_{i=1}^{m} y_i \alpha_i = 0,$ $0 \le \alpha_i \le C, \ i = 1, \cdots, m$ (1), and obtain the optimal solution $\alpha^* = (\alpha_1{}^*, \cdots, \alpha_m{}^*)^T$ (2). The radial basis function is used as a kernel function in this work. Choose a positive component of α^* $(0 < \alpha_j{}^* < C)$ and calculate the threshold $b^* = y_j - \sum_{i=1}^{m} y_i \alpha_i{}^* K(x_i - x_j)$ (3). Construct a decision function to calculate the brand discrimination results of experimental samples $f(x) = \mathrm{sgn}\left(\sum_{i=1}^{m} \alpha_i{}^* y_i K(x, x_i) + b^* \right)$ (4) (Xiaofeng et al., 2023; Zheng-Yong, 2020).

In the experimental operation, the "SVMcgForClass" function is used to optimize the grid parameters, and a 5-fold cross validation method is set. The optimization conditions for the kernel function parameter g and penalty coefficient c are $c_{min} = -10$, $c_{max} = 10$, $g_{min} = -10$, $g_{max} = 10$, and the search range is $[2^{-10}, 2^{10}]$. The step values are all 0.5. Finally, the optimal kernel function parameter g is about 0.00097656, and the penalty coefficient c is about 33.3333 for whole Raman spectroscopy of dairy products.

2.2.2. ELM algorithm

For the ELM algorithm, Assuming there are m experimental samples (x_i, y_i), here $x_i = [x_{i1}, x_{i2}, \cdots, x_{in}]^T \in R^n$, $y_i = [y_{i1}, y_{i2}, \cdots, y_{im}]^T \in R^m$.

For a single hidden layer neural network with L hidden layer nodes, there exists

$$\sum_{f=1}^{L}\beta_f g(\lambda_f \cdot x_i + b_f) = o_i \quad (5)$$

here $i = 1, 2, \ldots, n$, $g(\lambda_f \cdot x_i + b_f)$ is the activation function, $\lambda_f = [\lambda_{f1}, \lambda_{f2}, \cdots, \lambda_{fn}]^T$ is the input weight of the f-th hidden layer unit, and b_f is the bias of the f-th hidden layer unit, $\beta_f = [\beta_{f1}, \beta_{f2}, \cdots, \beta_{fm}]^T$ is the output weight of the f-layer. $\lambda_f \cdot x_i$ represents the inner product of λ_f and x_i.

The optimization objective of the neural network is to minimize output errors, expressed as: $\sum_{f=1}^{L}\|o_i - y_i\| = 0$ (6). There are λ_f, x_i, and b_f such that: $\sum_{f=1}^{L}\beta_f g(\lambda_f \cdot x_i + b_f) = W_i$ (7). Additionally, $H \cdot \beta = W$ (8), H is the output of the hidden layer node, β is the output weight, and W is the expected output.

$$\boldsymbol{H} = \begin{bmatrix} g(\lambda_1 \cdot x_1 + b_1) & \cdots & g(\lambda_L \cdot x_1 + b_L) \\ \vdots & \cdots & \vdots \\ g(\lambda_1 \cdot x_n + b_1) & \cdots & g(\lambda_L \cdot x_n + b_L) \end{bmatrix}_{n\times L}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1^T \\ \vdots \\ \beta_L^T \end{bmatrix}_{L\times m} \text{ and } \boldsymbol{W} = \begin{bmatrix} y_1^T \\ \vdots \\ y_n^T \end{bmatrix}_{n\times m}$$

Where, m is the number of outputs, H is the hidden layer output matrix, and W is the training set objective matrix. (Song et al., 2023, Zheng-Yong et al., 2023).

In the experimental operation, the "elmtrain" function is used for ELM creation and training, with the number of hidden layer neurons set to 400 and the activation function of hidden layer neurons using the "Sigmoidal" function.

2.2.3. CNN algorithm

For the CNN algorithm, set each research object category to have m observation samples, where X is the input spectral data and Y is the output data, denoted by the following equation, where c is the category.

$$X = \{x_{i,1}, x_{i,2}, \cdots, x_{i,n}\}_{i=1}^{m}, Y = \{y_{i,1}, y_{i,2}, \cdots, y_{i,c}\}_{i=1}^{m}$$

Convolutional layers learn features from input samples. Perform convolution operation between the input sample and the convolution kernel, shift the convolution results, and use activation function for nonlinear transformation. The calculation method is as follows:

$$x_k^r = f\left(\sum_{i\in R_k} x_i^{r-1} * \omega_{i,k}^r + b_k^r\right) \quad (9)$$

In the formula, r is the sequence number of the layer, x_k^r is the k-th feature output of the r layer, x_i^{r-1} is the output of the $r-1$ layer and the input of the r layer, $\omega_{i,k}^r$ is the convolutional filter of the i-th layer, b_k^r is the deviation, and R_k is the set of input feature maps. $f(\cdot)$ is an activation function (Lu et al., 2023, Dian et al., 2020).

In the experimental operation, taking whole spectral data processing as an example, the main parameters are set as follows: the size of the input layer is set to [2090, 1, 1], the convolutional layer uses 128 convolutional kernels of size [64, 1] for feature extraction, the activation function uses "rectified linear unit (relu)" function, and in the maximum pooling layer, a pooling window and step size of [32,1] are set. The fully connected layer sets three output categories and ultimately transforms the sample dimension into [1, 1, 3], indicating that the data features of the sample are classified into three categories. Subsequently, through the Softmax layer, the probability distribution of samples belonging to each category can be obtained.

2.2.4. Normalization

For normalization, the formula is $z = \frac{(z_{max} - z_{min}) \times (x - x_{min})}{x_{max} - x_{min}} + z_{min}$ (10); here x is the Raman spectral data that needs to be normalized, and z is the normalized data. The x_{min}, x_{max}, and z_{min}, z_{max} are the lower and upper limits of the interval for the raw data and normalized data, respectively. Since this work normalizes the data to $[-1, 1]$, so $z_{min} = -1$, and $z_{max} = 1$.

2.2.5. Euclidean distance

For Euclidean distance, the calculation formula is $d(u, v) = \sqrt{\sum_{i=1}^{n}(u_i - v_i)^2}$ (11). The d is the Euclidean distance between the Raman spectral data of two samples u and v.

3. Results and discussion

3.1. Discriminative analysis of dairy products based on the whole Raman spectroscopy combined with different machine learning algorithms

The Raman spectroscopy of dairy products can characterize the rich component information of the sample and has advantages such as fast, non-destructive, and portable data acquisition. As shown in Fig. 1, the Raman spectral peaks of dairy products are relatively sharp, mainly derived from proteins, fats, and carbohydrates (Weihua et al., 2022). The possible attribution of each peak is shown in Table 1. Dairy products from different brands have high similarity in their spectra, and their peak positions are very close. It can be used as important spectral characterization data and combined with machine learning algorithms to investigate the classification of similar samples.

Directly import the whole Raman spectra into SVM, ELM, and CNN algorithms in sequence. Randomly select 70% of dairy product sample data as the training set, and the remaining 30% of sample data as the test set. Perform 10 operations to obtain the average recognition accuracy. The recognition accuracy calculation results of the whole spectra combined with SVM, ELM, and CNN are 33.3%, 92.5%, and 100%, and the time required for the three algorithms also varies, approximately consuming 2000 s, 1 s, and 700 s respectively. The results show that combining the original whole spectra directly with SVM algorithm cannot achieve effective discrimination of dairy products. The reason for the low recognition accuracy of the SVM algorithm may be that the dimensional difference of the original data is in the range of 0–3000. Such a large dimensional interval results in a large solution error for SVM, which in turn leads to inaccurate classification results. Based on existing reports, subsequent spectral preprocessing is necessary for SVM algorithm to effectively discriminate (Wang et al., 2023). The ELM

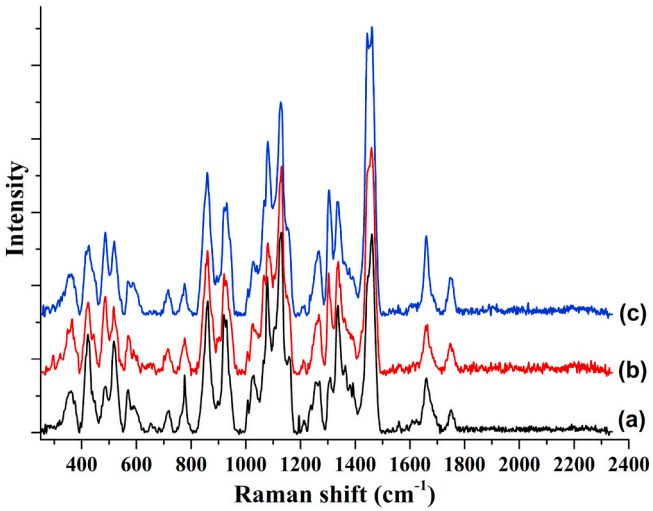

Fig. 1. Raman spectra of dairy products of (a) brand 1, (b) brand 2 and (c) brand 3.

Table 1
The main peak tentative assignments of the Raman spectra of the different brands of dairy products.

Raman shift (cm^{-1}) brand 1	Raman shift (cm^{-1}) brand 2	Raman shift (cm^{-1}) brand 3	Assignment	Possible component attribution
362	365	363	lactose	Lactose
423	425	426	glucose	Glucose
487	487	486	δ(C–C–C) + τ(C–O)	Carbohydrate
518	516	519	glucose	Glucose
570	570	568	δ(C–C–O) + τ(C–O)	Carbohydrate
719	716	717	ν (C–S)	Protein
776	776	775	ν (C–C–O)	Carbohydrate
861	860	859	δ(C–C–H) + δ(C–O–C)	Carbohydrate
920	921	921	δ (C–O–C) + δ(C–O–H) + ν (C–O)	Carbohydrate
1007	1006	1008	Ring-breathing (phenylalanine); ν(C–C)ring	Protein
1029	1025	1026	ν (C–O) + ν (C–C) + δ (C–O–H)	Carbohydrate
1078	1080	1080	ν (C–O) + ν (C–C) + δ (C–O–H)	Carbohydrate
1129	1130	1128	ν (C–O) + ν (C–C) + δ (C–O–H)	Carbohydrate
1270	1266	1267	γ (CH_2)	Carbohydrate
1310	1303	1305	τ (CH_2)	Fat
1337	1338	1336	ν (C–O)+δ (C–O–H)	Carbohydrate
1461	1460	1461	δ (CH_2)	Fat, Carbohydrate
1660	1664	1660	ν (C═O) amide I; ν (C═C)	Fat, Protein
1751	1747	1750	ν (C═O)ester	Fat

ν: stretching vibration; δ: deformation vibration; τ: twisting vibrationg; γ: out-of-plane bending vibration.

algorithm has obvious advantages in computational speed and high recognition accuracy. The CNN algorithm has its own feature extraction ability, the highest recognition accuracy and moderate computation time (Chen et al., 2022).

3.2. Discriminative analysis of dairy products based on feature spectrum combined with different machine learning algorithms

Normalization processing is expected to remove the influence of spectral data dimensionality and improve the discrimination accuracy of the classifier. This work normalizes spectral data to the range of −1 to 1, and then imports the data into three machine learning algorithms. The results show that the SVM algorithm improves the recognition accuracy to 100%, the ELM algorithm improves the recognition accuracy to 94.4%, and the CNN recognition accuracy remains at 100%. The computation time is about 1750s, 1s, and 750s, respectively. It is shown that normalization processing has a good effect on improving the recognition accuracy of experimental data and different machine learning algorithms.

Subsequently, the recognition accuracy changes are studied by combining different spectral characteristic peaks with different machine learning algorithms, as shown in Table 2. The division of feature intervals here is mainly based on the sample characteristic peaks shown in Fig. 1 and Table 1, and the intervals covering each peak are selected. The results clearly show that different feature bands have various contributions to classification algorithms, and at the same time, the same feature band also has certain differences in recognition accuracy corresponding to different classification algorithms. For the SVM algorithm, the top three bands in the Raman feature interval ranked by recognition accuracy are 1410–1500 cm^{-1}, 890-980 cm^{-1}, and 1100-1180 cm^{-1}, respectively. For the ELM algorithm, the top three Raman feature intervals are 1410–1500 cm^{-1}, 1050-1100 cm^{-1}, and 810-890 cm^{-1}, respectively. For the CNN algorithm, the top three Raman feature intervals are 1410–1500 cm^{-1}, 890-980 cm^{-1}, and 810-890 cm^{-1}, respectively.

Table 2
The recognition accuracy of different Raman spectral feature intervals combined with various machine learning algorithms.

Raman shift interval (cm^{-1})	SVM (accuracy, %)	ELM (accuracy, %)	CNN (accuracy, %)
280–390	94.4	87.2	98.3
390–460	89.2	84.2	93.3
460–500	81.1	71.1	96.9
500–550	86.4	71.9	90.3
550–630	97.8	91.1	98.3
630–690	84.2	71.9	84.2
690–745	66.7	52.5	66.8
745–810	85.3	64.7	85.0
810–890	98.9	96.4	98.6
890–980	99.7	93.1	98.9
980–1015	76.9	63.1	89.2
1015–1050	83.1	78.6	92.2
1050–1100	99.2	98.6	98.3
1100–1180	99.4	94.7	96.7
1180–1225	76.4	64.7	76.4
1225–1290	92.2	87.5	95.0
1290–1320	98.3	93.3	95.8
1320–1410	98.6	91.1	96.7
1410–1500	100	99.4	99.7
1500–1630	84.7	70.3	93.6
1630–1710	99.2	89.4	98.3
1710–1780	83.9	63.6	96.1
1780–2339	68.9	59.7	90.8

The feature bands with high recognition accuracy can also provide us with some material information about differences in dairy product classification. The top nine spectral intervals with high recognition accuracy based on SVM algorithm are shown in Fig. 2. The corresponding spectral peak attribution analysis is as follows. The 550-630 cm^{-1} band can mainly be derived from glycosidic ring skeletal deformations, the 810-890 cm^{-1}, 890-980 cm^{-1}, 1050-1100 cm^{-1}, and 1100-1180 cm^{-1} bands can mainly be attributed to glycosidic bonds. The 1290-1320 cm^{-1} band can mainly be derived from lipids, the 1320-1410 cm^{-1} band can mainly be derived from carbohydrate molecules, and the 1410-1500 cm^{-1} band can mainly be derived from fats and carbohydrates. The 1630-1710 cm^{-1} band can mainly originate from the C═C stretching vibration of unsaturated fatty acids and the C═O stretching vibration in the amide I group CONH of proteins (Almeida et al., 2011, Rodrigues

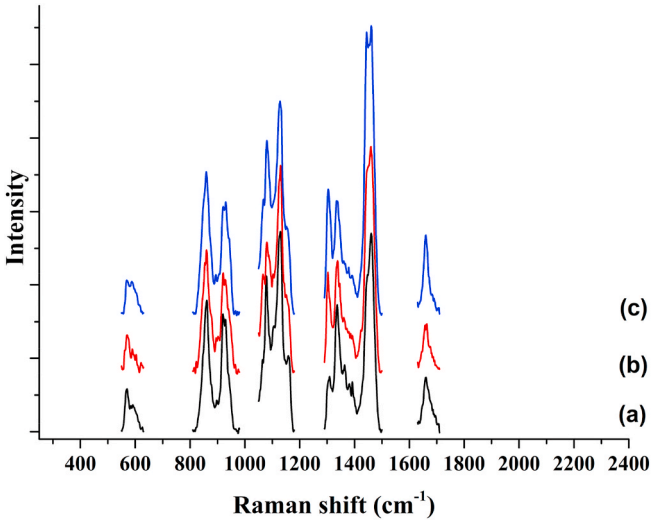

Fig. 2. Raman spectral feature bands of dairy products of (a) brand 1, (b) brand 2 and (c) brand 3.

et al., 2016; Zheng-Yong et al., 2019). There are also some spectral feature regions that exhibit lower recognition accuracy when combined with algorithms, such as 690-745 cm^{-1} and 980-1015 cm^{-1}, which can be derived from proteins, and 1780-2339 cm^{-1}, which mainly belong to the spectral noise region without obvious spectral peaks.

In addition, due to the fact that the feature bands are only a part of the original data, the operation time of each classifier is significantly reduced. Except for a few feature bands (280-390 cm^{-1}, 1500-1630 cm^{-1}, 1780-2339 cm^{-1}) that exceed 100 s, the operation time of all other bands in SVM is within 100 s. The ELM operation time is only within 0.3 s, and the CNN operation time is only within 50 s.

Based on the combination of the single spectral feature interval and the algorithm mentioned above, multiple spectral peaks are selected to make a significant contribution to the recognition accuracy. Further research is conducted on the impact of spectral interval fusion on the algorithm's recognition accuracy. The experiment fuses the spectral intervals of each algorithm's recognition accuracy contribution with the top three bands, and the top nine spectral feature intervals based on SVM algorithm's recognition accuracy contribution in sequence. The change in recognition accuracy is shown in Table 3. It can be seen that different algorithms have achieved optimal recognition in certain fusion segments. For example, the SVM algorithm has a recognition accuracy of 100% in the 890-980 cm^{-1}, 1410-1500 cm^{-1} fusion spectral range, which takes about 200 s. The ELM algorithm also has a recognition accuracy of 100% in the 890-980 cm^{-1}, 1410-1500 cm^{-1} fusion spectral range, which takes less than 0.3 s. The CNN algorithm has a recognition accuracy of 100% in the 890-980 cm^{-1}, 1050-1180 cm^{-1}, 1410-1500 cm^{-1} fusion spectral range, which takes about 80 s. The optimal recognition results based on CNN algorithm are shown in Fig. 3. This research result shows that appropriate feature spectral interval fusion, combined with corresponding machine learning algorithms effectively, can significantly improve recognition accuracy and maintain higher computational efficiency compared to full spectral data.

3.3. Further statistical analysis of spectral feature intervals

To further understand the statistical patterns of the Raman spectral features mentioned above, a quality fluctuation control chart analysis based on the Raman spectral band at 1410-1500 cm^{-1} is conducted firstly. The calculation steps are as follows. The first step is to calculate the spectral mean of dairy product brand 1 as the best estimate of its true value. The second step is to calculate the Euclidean distance between each sample of brand 1 and this mean, and use this result to substitute it into the individual value moving range control chart operation rule to obtain the control limit of individual value and moving range, then draw the corresponding control chart. The third step is to calculate the Euclidean distances of each sample from brand 2 and brand 3 with the mean of brand 1, and plot them on the control chart, as shown in Fig. 4. It can be seen that based on the spectral feature interval that contributes the most to this recognition accuracy, there are quality fluctuations in each sample of brand 1, and they all fluctuate normally around the centerline within the control limit. In the sample of brand 2, there are eleven outliers in the individual value control chart and seven outliers in the moving range control chart. In brand 3, all forty points in the individual value control chart are outliers, and there are fifteen outliers in the moving range control chart. Therefore, it reflects that there is indeed a certain difference in this spectral range among the three brands, but there are also a certain number of samples from other brands that still exist within the control range of brand 1, which intuitively shows a certain statistical fluctuation pattern of the samples.

Table 3
Recognition accuracy of different Raman spectral feature fusion intervals combined with various machine learning algorithms.

Raman shift interval (cm^{-1})	SVM (accuracy, %)	ELM (accuracy, %)	CNN (accuracy, %)
890-980, 1410-1500	100	100	99.4
890-980, 1100–1180, 1410-1500	100	99.4	99.4
1050-1100, 1410-1500	100	99.7	99.7
810-890, 1050–1100, 1410-1500	100	99.7	98.3
810-980, 1410-1500	100	98.9	99.4
890-980, 1050–1180, 1410-1500	100	99.7	100
890-980, 1050–1180, 1410–1500, 1630-1710	100	100	100
810-980, 1050–1180, 1410–1500, 1630-1710	100	100	99.2
810-980, 1050–1180, 1320–1500, 1630-1710	100	100	99.7
810-980, 1050–1180, 1290–1500, 1630-1710	100	100	99.7
550-630, 810–980, 1050–1180, 1290–1500, 1630-1710	100	100	100

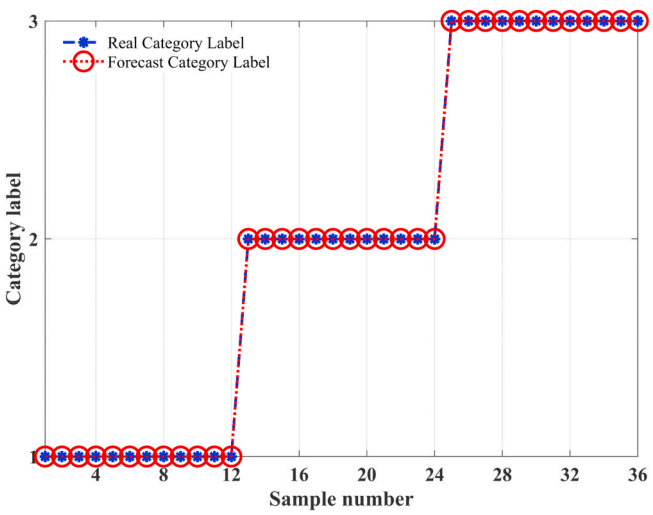

Fig. 3. Recognition results of dairy products based on the CNN (y-axis: label 1 represents brand 1, label 2 represents brand 2, and label 3 represents brand 3).

Secondly, two spectral feature intervals of 1410–1500 cm^{-1} and 890-980 cm^{-1}, as well as the spectral feature interval of 1100–1180 cm^{-1}, are selected to perform Euclidean distance calculations with the spectral mean of brand 1 as the best estimate of truth. Specifically, the Euclidean distance between each sample of brand 1 and its mean under various spectral feature interval conditions is calculated, as well as the Euclidean distance between each sample of brand 2 and brand 3 and the mean of brand 1, then draw the results as shown in Figs. 5 and 6. It can be seen that under the conditions of 1410–1500 cm^{-1} and 890-980 cm^{-1}, there is a certain degree of difference between brand 1 and brand 2, 3 in the two-dimensional space. In the three-dimensional space of 1410-1500 cm^{-1}, 890 980 cm^{-1} and 1100-1180 cm^{-1}, there is also a certain degree of difference between brand 1 and brand 2, 3. Displaying the distribution differences between samples under the conditions of feature spectral intervals, and revealing the spectral feature basis for efficient recognition of machine learning algorithms.

4. Conclusions

In this work, SVM, ELM, CNN algorithms and Raman spectroscopy are combined to investigate the discrimination spectral features of dairy products. It was found that (a) the optimal spectral feature interval of different machine learning algorithms is not the same, (b) a small amount of feature spectral interval fusion can improve the recognition accuracy and computational efficiency of the algorithm to a certain extent, and (c) Raman spectroscopy, as a data input for machine learning

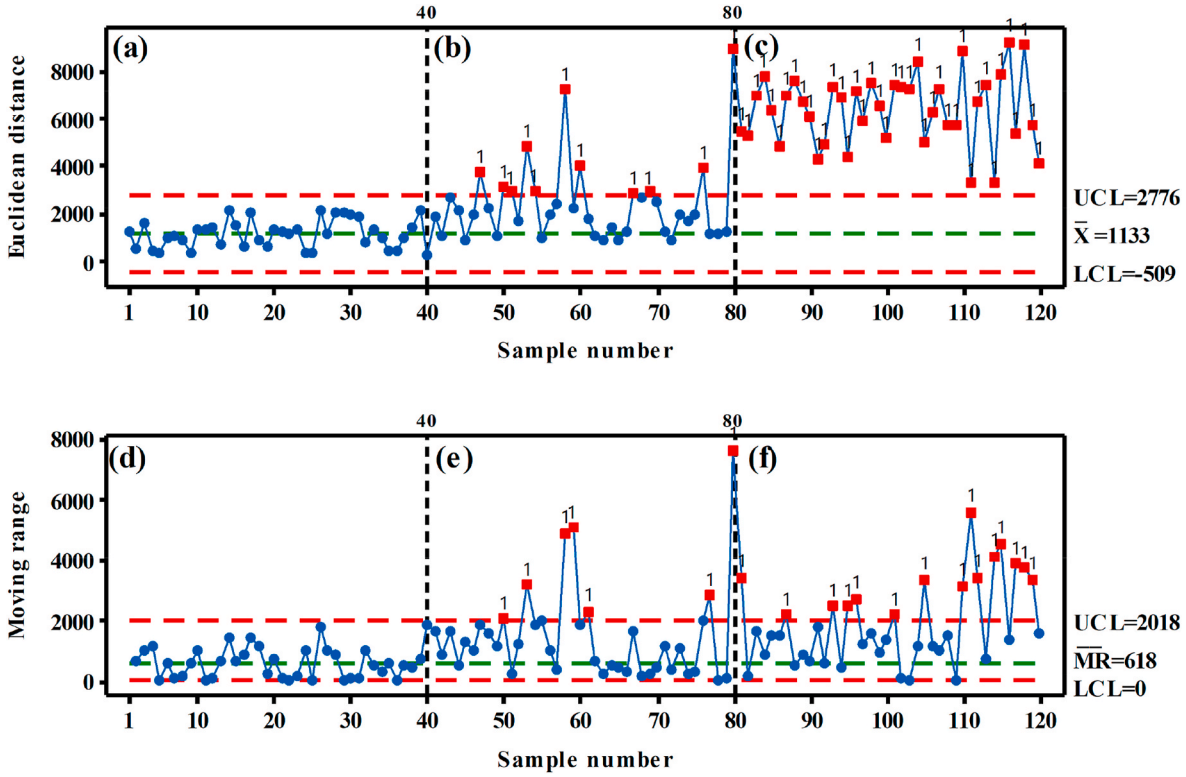

Fig. 4. Quality fluctuation individual value of brand 1 (a), brand 2 (b), and brand 3 (c) and moving range of brand 1 (d), brand 2 (e), and brand 3 (f) based on Euclidean distance of Raman spectral feature interval (1410-1500 cm^{-1}). UCL = upper control limit; LCL = lower control limit; $\overline{MR}$ = the average value of moving range control chart. $\bar{x}$ = the average value of individual control chart.

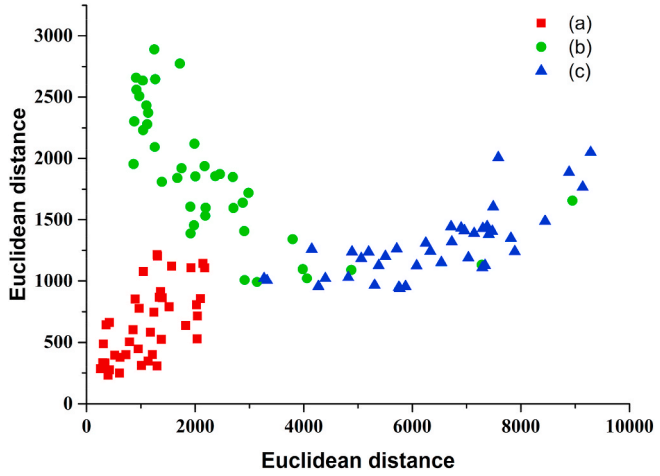

Fig. 5. A Euclidean distance two-dimensional map of dairy products of (a) brand 1, (b) brand 2 and (c) brand 3 based on the Raman spectral feature intervals (x-axis:1410-1500 cm^{-1} and y-axis: 890-980 cm^{-1}).

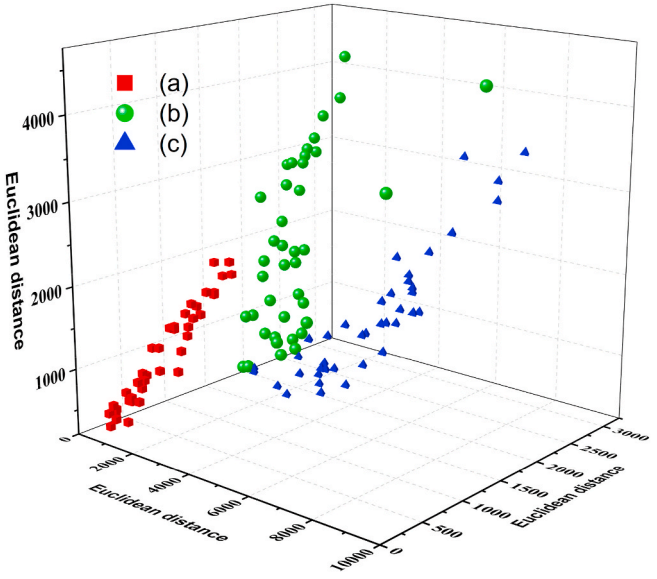

Fig. 6. A Euclidean distance three-dimensional map of dairy products of (a) brand 1, (b) brand 2 and (c) brand 3 based on the Feature Raman spectral feature intervals (x-axis:1410-1500 cm^{-1}, y-axis: 890-980 cm^{-1} and z-axis:1100-1180 cm^{-1}).

algorithms, has different contribution rates among different spectral feature intervals. Spectral feature intervals with high discriminative ability are expected to be visually displayed in terms of sample spatial distribution through statistical distribution analysis. Therefore, the entire set of feature spectral analysis strategies established in this article will help researchers further understand the characterization basis for different categories of dairy products.

Funding

This research was financially supported by the Excellent Young Backbone Teachers of "Blue Project" in Jiangsu Universities in 2021, Jiangsu Province, China, National Natural Science Foundation of China (61602217), and Innovation and Entrepreneurship Training Program for College Students in Jiangsu Province, China (202310327056Z).

CRediT authorship contribution statement

Jia-Xin Li: Conceptualization, Methodology, Investigation, Writing – original draft. **Chun-Chun Qing:** Methodology, Investigation, Formal analysis, Writing – review & editing. **Xiu-Qian Wang:** Investigation, Writing – review & editing. **Mei-Jia Zhu:** Validation, Writing – review & editing. **Bo-Ya Zhang:** Validation, Writing – review & editing. **Zheng-Yong Zhang:** Supervision, Writing – review & editing.

Declaration of competing interest

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

Data availability

Data will be made available on request.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.crfs.2024.100782.

References

Almeida, Mariana R., Oliveira, Kamila De S., Stephani, Rodrigo, Luiz, Fernando C. de Oliveira, 2011. Fourier-transform Raman analysis of milk powder: a potential method for rapid quality screening. J. Raman Spectrosc. 42 (7), 1548–1552.

Andrey, Shishov, Nizov, Egor, Bulatov, Andrey, 2023. Microextraction of melamine from dairy products by thymol-nonanoic acid deep eutectic solvent for high-performance liquid chromatography-ultraviolet determination. J. Food Compos. Anal. 116, 105083.

Chen, Zhuofa, Yousif, Khaireddin, Anna, K. Swan, 2022. Identifying the charge density and dielectric environment of graphene using Raman spectroscopy and deep learning. Analyst 147 (9), 1824–1832.

Dian, Rong, Wang, Haiyan, Ying, Yibin, Zhang, Zhengyong, Zhang, Yinsheng, 2020. Peach variety detection using VIS-NIR spectroscopy and deep learning. Comput. Electron. Agric. 175, 105553.

Giulia, Martini, Bracci, Alberto, Riches, Lorenzo, Jaiswal, Sejal, Corea, Matteo, Rivers, Jonathan, Husain, Arif, Omodei, Elisa, 2022. Machine learning can guide food security efforts when primary data are not available. Nature Food 3 (9), 716–728.

Gouvêa, Silva Monisa, Lima de Paula, Igor, Stephani, Rodrigo, Edwards, Howell G.M., de Oliveira, Luiz Fernando Cappa, 2021. Raman spectroscopy in the quality analysis of dairy products: a literature review. J. Raman Spectrosc. 52 (12), 2444–2478.

Huang, Wei, Fan, Desheng, Li, Wangfang, Meng, Yaoyong, Liu, Timon Cheng-yi, 2022. Rapid evaluation of milk acidity and identification of milk adulteration by Raman spectroscopy combined with chemometrics analysis. Vib. Spectrosc. 123, 103440.

Ji, Huizhuo, Pu, Dandan, Yan, Wenjing, Zhang, Qingchuan, Zuo, Min, Zhang, Yuyu, 2023. Recent advances and application of machine learning in food flavor prediction and regulation. Trends Food Sci. Technol. 138, 738–751.

Khan, H. M. Hussain, McCarthy, Ultan, Esmonde-White, Karen, Casey, Imelda, O'Shea, Norah, 2023. Potential of Raman spectroscopy for in-line measurement of raw milk composition. Food Control 152, 109862.

Lu, Bingxu, Tian, Feng, Chen, Cheng, Wu, Wei, Tian, Xuecong, Chen, Chen, Lv, Xiaoyi, 2023. Identification of Chinese red wine origins based on Raman spectroscopy and deep learning. Spectrochim. Acta Mol. Biomol. Spectrosc. 291, 122355.

Pan, Wei, Liu, Wenjing, Huang, Xiujuan, 2024. Rapid identification of the geographical origin of Baimudan tea using a Multi-AdaBoost model integrated with Raman Spectroscopy. Curr. Res. Food Sci. 8, 100654.

Pu, Hongbin, Yu, Jingxiao, Sun, Da-Wen, Wei, Qingyi, Wang, Zhe, 2023. Feature construction methods for processing and analysing spectral images and their applications in food quality inspection. Trends Food Sci. Technol. 138, 726–737.

Ranveer, Soniya A., Harshitha, C.G., Dasriya, Vaishali, Tehri, Nimisha, Kumar, Naresh, Raghu, H.V., 2023. Assessment of developed paper strip based sensor with pesticide residues in different dairy environmental samples. Curr. Res. Food Sci. 6, 100416.

Rodrigues, Júnior Paulo Henrique, Oliveira, Kamila de Sá, Almeida, Carlos Eduardo Rocha De, Oliveira, Luiz Fernando Cappa De, Stephani, Rodrigo, Pinto, Michele Da Silva, De Carvalho, Antônio Fernandes, Tuler Perrone, Ítalo, 2016. FT-Raman and chemometric tools for rapid determination of quality parameters in milk powder: classification of samples for the presence of lactose and fraud detection by addition of maltodextrin. Food Chem. 196, 584–588.

Shan, JinRui, Shi, Longhua, Li, Yuechun, Yin, Xuechi, Wang, Shaochi, Liu, Sijie, Sun, Jing, Zhang, Daohong, Ji, YanWei, Wang, Jianlong, 2023. SERS-based immunoassay for amplified detection of food hazards: recent advances and future trends. Trends Food Sci. Technol. 140, 104149.

Singh, Poornima, Pandey, Surabhi, Manik, Subhadip, 2024. A comprehensive review of the dairy pasteurization process using machine learning models. Food Control 164, 110574.

Song, Shuai, Wang, Qiaoyun, Zou, Xin, Li, Zhigang, Ma, Zhenhe, Jiang, Daying, Fu, YongQing, Liu, Qiang, 2023. High-precision prediction of blood glucose concentration utilizing Fourier transform Raman spectroscopy and an ensemble machine learning algorithm. Spectrochim. Acta Mol. Biomol. Spectrosc. 303, 123176.

Wang, Kaiqiang, Li, Zonglun, Li, Jinjie, Lin, Hong, 2021. Raman spectroscopic techniques for nondestructive analysis of agri-foods: a state-of-the-art review. Trends Food Sci. Technol. 118, 490–504.

Wang, Jinfeng, Lin, Tenghui, Ma, Siyuan, Ju, Jinyan, Wang, Ruidong, Chen, Guoqing, Jiang, Rui, Wang, Zhentao, 2023. The qualitative and quantitative analysis of industrial paraffin contamination levels in rice using spectral pretreatment combined with machine learning models. J. Food Compos. Anal. 121, 105430.

Wang, Haiyan, Zhang, Xiaoying, Yao, Yu, Huo, Zhenquan, Cui, Xiuxiu, Liu, Mengjia, Zhao, Lili, Ge, Wupeng, 2024. Oligosaccharide profiles as potential biomarkers for detecting adulteration of caprine dairy products with bovine dairy products. Food Chem. 443, 138551.

Weihua, Huang, Guo, Lianbo, Kou, Weiping, Zhang, Deng, Hu, Zhenlin, Chen, Feng, Chu, Yanwu, Wen, Cheng, 2022. Identification of adulterated milk powder based on convolutional neural network and laser-induced breakdown spectroscopy. Microchem. J. 176, 107190.

Xiaofeng, Ni, Jiang, Yirong, Zhang, Yinsheng, Zhou, Ya, Zhao, Yaju, Guo, Fangjie, Wang, Haiyan, 2023. Identification of liquid milk adulteration using Raman spectroscopy combined with lactose indexed screening and support vector machine. Int. Dairy J. 146, 105751.

Xue, Xi, Sun, Hanyu, Yang, Minjian, Liu, Xue, Hu, Hai-Yu, Deng, Yafeng, Wang, Xiaojian, 2023. Advances in the application of artificial intelligence-based spectral data interpretation: a perspective. Anal. Chem. 95 (37), 13733–13745.

Yang, Zichen, Chen, Guoqing, Ma, Chaoqun, Gu, Jiao, Zhu, Chun, Li, Lei, Gao, Hui, 2023. Magnetic Fe_3O_4@COF@Ag SERS substrate combined with machine learning algorithms for detection of three quinolone antibiotics: ciprofloxacin, norfloxacin and levofloxacin. Talanta 263, 124725.

Yiwei, Cui, Lu, Weibo, Xue, Jing, Ge, Lijun, Yin, Xuelian, Jian, Shikai, Li, Haihong, Zhu, Beiwei, Dai, Zhiyuan, Shen, Qing, 2023. Machine learning-guided REIMS pattern recognition of non-dairy cream, milk fat cream and whipping cream for fraudulence identification. Food Chem. 429, 136986.

Zheng-Yong, Zhang, 2020. The statistical fusion identification of dairy products based on extracted Raman spectroscopy. RSC Adv. 10 (50), 29682–29687.

Zheng-Yong, Zhang, Liu, Jun, Wang, Hai-Yan, 2015. Microchip-based surface enhanced Raman spectroscopy for the determination of sodium thiocyanate in milk. Anal. Lett. 48 (12), 1930–1940.

Zheng-Yong, Zhang, Sha, Min, Wang, Hai-Yan, 2017. Laser perturbation two-dimensional correlation Raman spectroscopy for quality control of bovine colostrum products. J. Raman Spectrosc. 48 (8), 1111–1115.

Zheng-Yong, Zhang, Gui, Dong-Dong, Sha, Min, Liu, Jun, Wang, Hai-Yan, 2019. Raman chemical feature extraction for quality control of dairy products. J. Dairy Sci. 102 (1), 68–76.

Zheng-Yong, Zhang, Li, Si-Wei, Sha, Min, Liu, Jun, 2021. Characterization of fresh milk products based on multidimensional Raman spectroscopy. J. Appl. Spectrosc. 87 (6), 1206–1215.

Zheng-Yong, Zhang, Jiang, Min-Qin, Xiong, Huan-Ming, 2023. Optimized identification of cheese products based on Raman spectroscopy and an extreme learning machine. New J. Chem. 47 (14), 6889–6894.

Zikang, Feng, Liu, Dou, Gu, Junyan, Zheng, Lina, 2024. Raman spectroscopy and fusion machine learning algorithm: a novel approach to identify dairy fraud. J. Food Compos. Anal. 129, 106090.