Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 12, No. 3, 2024

Grade Identification of Strong-flavor Raw Liquor Based on GC-MS Combined with Spearman-KPCA Feature Extraction

Ni Fu[1, 2], Xianguo Tuo[1, 2, *], Wei Zhang[1, 2]

[1]School of Automation and Information Engineering Sichuan University of Science and Engineering, Yibin, China
[2]Artificial Intelligence Key Laboratory of Sichuan Province, Yibin 644000, China
[*] Corresponding author: Xianguo Tuo (Email: puckio98@163.com)

Abstract: Taking different grades of strong-flavor raw liquor as the research object, gas chromatography-mass spectrometry (GC-MS) technology was used to obtain the volatile components mapping data of raw liquor, and Spearman correlation coefficient (Spearman) combined with principal component analysis (PCA) and kernel principal component analysis (KPCA) was used to realize the secondary feature extraction of the GC-MS data, and then combined with the support vector machine (SVM), extreme gradient boosting (XGBoost), and BP neural network to establish the raw liquor grades identification model, respectively. The results show that the prediction accuracy of the grade identification model based on Spearman-KPCA dimensionality reduction data is better, in which the Spearman-KPCA-BP neural network model has the best classification effect, and the accuracy of the correction set and prediction set reaches 99.44% and 96.10%, respectively. Research shows that the principal components extracted based on Spearman-KPCA secondary features can better characterize the characteristic information of different grades of raw liquor. Combined with the BP neural network model, it can effectively realize the identification of different grades of raw liquor. It is an effective method for identifying the grade of raw liquor.

Keywords: Strong-flavor base liquor; gas chromatography-mass spectrometry; Spearman's rank correlation coefficient; kernel principal component analysis; grade identification model.

1. Introduction

The main components of liquor are water and ethanol, the content of which accounts for about 98% of the total amount, and the remaining 1% to 2% are acids, esters, aldehydes, alcohols and other trace components, and the content of these trace components and the quantitative relationship between them determines the aroma, quality and style of liquor [1,2]. According to the different microcomponents in the main body of liquor, liquor can be divided into twelve types of aroma, such as strong aroma, soy sauce aroma, clear aroma, rice aroma, etc. Strong aroma liquor is representative of one of the four basic aroma, which is based on ethyl caproate as the main aroma substance, with colorless and transparent, cellar aroma, aroma and coordination, and the end of the net refreshing and so on [3,4]. As an intermediate product from grain to finished wine, the accurate classification of raw wine has a significant impact on the quality of graded storage and final finished wine [5].

In recent years, fingerprint technology has been widely used in the field of food quality and safety, and the fingerprint obtained by spectral or chromatographic instruments can reflect the internal changes of liquor [6].Gas chromatography-massspectrometry (GC-MS) is an online technology that combines gaschromatograph with qualitative mass spectrometer with separation capability. The chromatographic and mass spectrometry data of the sample to be tested can be obtained in a relatively short time for qualitative and quantitative analysis of the multi-component mixture, which has the advantages of low detection limit, high sensitivity, strong stability and good separation [7,8].Liu Qingru et al.[9]obtained the fingerprint of volatile components of base wine by GC-MS, and combined with the extreme gradient lifting algorithm, established a regression model to realize the identification of storage time of base wine.Qian Yu et al. [10] realized effective differentiation of Luzhou-flavor liquor of different brands by using GC-MS fingerprint of volatile components of liquor combined with chemometrics.Zhu Kaixian et al. [11] combined with chemometrics to process the GC-MS spectrum data and analyze the volatile components of five flavored baijiu to achieve accurate classification of different flavored baijiu.Fan Shanshan et al. [12] applied GC-MS spectrum technology to the identification of Xiaoqu pure flavor raw wine of different grades, and established a grade discrimination model of Xiaoqu pure flavor raw wine in combination with PL-DA to achieve accurate identification of raw wine samples.YanboLiu et al. [13] used GC-MS to detect volatile components in wine samples, combined with fuzzy mathematics and principal component analysis to establish a comprehensive quality evaluation system for Luzhou-flavor liquor.It can be seen that the application of GC-MS spectrum technology in the classification and recognition of liquor can achieve the purpose of accurate detection and analysis of liquor.

Raw wine is a complex multi-component system, and changes in the content of some trace components have little or no impact on the quality of raw wine. Therefore, the GC-MS atlas data of raw wine contains a large number of redundant characteristic compounds, which will introduce errors in the modeling process, thus affecting the accuracy of the model. The dimensionality reduction of data can filter out the irrelevant features in the data and reduce the complexity of the data.Reduce the errors caused by redundant information and improve the prediction accuracy [14,15].At present, the single feature extraction method is used to reduce the dimensionality of GC-MS atlas data, and it is difficult to

combine the advantages of different methods to extract the essential structure of the data. Therefore, in order to improve the accuracy of model discrimination, Spearman grade correlation coefficient (Spearman) was first used for feature screening to filter out redundant feature compounds. Then using principal component analysis (principalcomponentanalysis, PCA) and kernel principal component analysis (kernelprincipalcomponentanalysis, KPCA) for feature extraction, Finally combining support vector machines (supportvectormachine, SVM), maximum gradient increase tree (ExtremeGradientBoostingTree, XGBoost) and BP neural network, this paper compares and analyzes three kinds of machine learning modelA Spearman-KPCA-BP neural network method for grade identification of Luzhou-flavor raw wine was proposed, which provided a new method for grade identification and quality control of Luzhou-flavor raw wine.

2. Materials and Methods

2.1. Materials and Reagents

2.1.1. Liquor samples

The samples for this experiment were selected from a series of strong-flavored white wines of a wine industry in Sichuan, and the collection process of the original wine samples was completed by wine pickers with more than 10 years of experience in picking wines, and after the completion of the sample collection, the tasting panel of the wine enterprise (five professional sommeliers) carried out the grading of the original wines, and the distribution of the grades of the samples was as shown in Table 1.

Table 1. Distribution information of original wine samples

tab	Grade of Raw Liquor	Sample size	peculiarity
1	First drink	72	High alcohol, aldehydes, poor quality
2	Mid-course wine	107	Strong aroma, outstanding style, clear and transparent wine
3	Tail wine	77	The alcohol content is low, the taste is not mellow, the wine is cloudy, and there is a mixed taste

2.1.2. Chemical Reagents

N-amyl acetate, tert-amyl alcohol, 2-ethylbutyric acid (all chromatographic pure), anhydrous ethanol (purity 99.5%) : Shanghai Maclin Biochemical Technology Co., LTD.Methanol (purity 99.9%) : Shanghai Adamas Reagent Co., LTD.C7~C40 normal alkanes (chromatographically pure) : Beijing Manhag Biotechnology Co., LTD.

2.2. Main instruments and equipment

Gas chromatograph (Model: 7890B), mass spectrometer (model: G7000D): Both from the United States Agilent company.

2.3. Method

2.3.1. Determination of volatile compounds in raw wine samples by GC-MS

(1) Sample pretreatment

Measure 1ml liquor sample solution with a micro pipette gun into the sample bottle, add 10μL prepared internal standard solution, make label records, mix well, and use GC-MS for detection and analysis.

(2) GC-MS Parameter setting

The GC column was Agilent DB-WAX (60m×0.25mm ×0.25μm).The injection volume was 1μL, the shunt ratio was 20:1, and the inlet temperature was 250℃.The carrier gas was high purity helium (He) with a flow rate of 1mL/min.The initial column temperature was 40℃, and was heated to 120℃ at 10℃/min for 2min, then heated to 200℃ at 10℃/min for 2min, and then heated to 250℃ at 10℃/min for 10min.The interface temperature of MS is 280℃.EI ion source 70eV ionization;Ion source temperature 230℃, quadrupole temperature 150℃;Full scanning mode: Scanning range 30 to 540m/z.

(3) Qualitative analysis

Qualitative analysis is to determine the type of complex components in the sample to be tested [16].The chromatographic peaks were automatically identified and compared with the 12 mass spectrometry library of the national institute of standards and technology (NIST) in combination with the retention index.RI) The flavor components in the original wine were qualitatively analyzed.

(4) Quantitative analysis

The specific content of volatile compounds in the sample to be tested is determined as quantitative analysis [17].The internal standard method was used to calculate the content of each trace component by referring to the peak area method in GB/ T10345-2007 "Liquor Analysis Method". The internal standard compounds must be selected as substances that are relatively stable, weak in volatility, do not react with the sample itself, have similar chemical properties with the substance to be measured and do not exist in the sample. Based on the basic characteristics of liquor and the peak time distribution of its trace components, 2-ethylbutyric acid, n-amyl acetate and tert-amyl alcohol were selected as the internal standard compounds of acids, esters, alcohols and other flavor substances.

2.3.2. Spearman rank correlation coefficient

Spearman's rank correlation coefficient is used to evaluate the correlation of two variables. Different from Pearson's correlation coefficient and Kendall's correlation coefficient, which are strict on sample data, Spearman's correlation coefficient is calculated based on the difference of the corresponding series of two pairs of ranks [18].The idea of solving Spearman's rank correlation coefficient is as follows:

Define $X=\{X_1,X_2,\cdots,X_n\},Y=\{Y_1,Y_2,\cdots,Y_n\}$ Is two sets of variables, The data of the two groups of variables are arranged in the same order of size $X=\{x_1,x_2,\cdots,x_n\},Y=\{y_1,y_2,\cdots,y_n\}$, x_i,y_i Represents the rank of the I-th element in the last two sets of data, hen the correlation coefficient of X and Y is calculated as shown in (1) :

$$r_s=1-\frac{6\sum d_i^2}{n\left(n^2-1\right)} \quad (1)$$

Where di is the grade difference between xi and yi, and n represents the amount of data;The value range of rs is [-1,1], where rs>0 indicates that the two variables are positively correlated,rs<0 indicates that the two variables are negatively correlated, and |rs| the larger the value is, the stronger the correlation between the two variables.

2.3.3. PCA

Principal component analysis is a linear dimensionality reduction algorithm. The idea of the algorithm is to convert original feature data into a set of linearly unrelated feature vectors through orthogonal transformation to achieve the purpose of dimensionality reduction [19].The algorithm steps are as follows:

(1) Standardization of sample data

(2) calculation of correlation coefficient matrix

(3) calculation of eigenvalues and eigenvectors

(4) calculation of principal component contribution rate and cumulative contribution rate

(5) Selection of principal components Generally

the more the cumulative contribution rate is closer to 1, the better the characteristic information of the original data can be characterized. The number of target principal components can be determined by specifying the magnitude of the cumulative contribution rate.

2.3.4. KPCA

Kernel principal component analysis introduces kernel function on the basis of principal component analysis so that it can handle linearly indivisible data sets [20].KPCA kernel functions commonly used include linear kernel function, polynomial kernel function, Sigmoid kernel function and Gaussian kernel function, this paper selects Gaussian kernel function.The basic idea of KPCA is to map linearly indivisible raw data to a high-dimensional feature space through kernel function, and then use PCA to reduce dimensionality in the high-dimensional space to extract nonlinear information in the data [21,22].The algorithm steps are as follows:

The sample matrix X composed of n samples' M-dimensional features is mapped to the high-dimensional feature space through the mapping function $\phi(X)$, and the covariance matrix of the centralized data in the high-dimensional space is obtained as follows:

$$C = \frac{1}{n}\tilde{\phi}(X)\tilde{\phi}\left(X\right)^{T} \tag{2}$$

Where, $\tilde{\phi}\left(x_i\right)$ is the centralized data vector.

Therefore, the characteristic equation of the covariance matrix is:

$$\frac{1}{n}\tilde{\phi}(X)\tilde{\phi}\left(X\right)^{T}\mu = \lambda\mu \tag{3}$$

Bringing (2) into (3) yields:

$$\begin{aligned} \mu &= \tilde{\phi}\left(X\right)\alpha \\ \alpha &= \frac{1}{\lambda n}\tilde{\phi}\left(X\right)^{T} \end{aligned} \tag{4}$$

By introducing the kernel matrix $K = \tilde{\phi}\left(X\right)^{T}\tilde{\phi}\left(X\right)$, the eigenequation of the kernel matrix obtained by combining (3) and (4) is as follows:

$$K\alpha = \lambda n\alpha \tag{5}$$

The eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ and the corresponding eigenvectors $\alpha_1, \alpha_2, \cdots, \alpha_n$ of the kernel matrix are computed and the principal components are determined based on the magnitude of the cumulative contribution of the desired eigenvalues.

2.3.5. Raw Liquor grade identification model

In order to explore a more suitable model for the grade identification of Luzhou-flavor liquor, this study compared the effectiveness of three classification models, namely support vector machine, extreme gradient lifting tree and BP neural network, on the basis of data dimensionality reduction.

(1) Support vector machine

Support vector machine is a supervised learning algorithm based on statistical theory, which has good generalization ability [23].Its goal is to minimize structural risk, so it can achieve good results on small samples and non-linear data.The main idea is to find an optimal hyperplane (decision boundary) in high-dimensional space through kernel function, and then separate different classes of samples [24].

(2) Extreme gradient lifting tree

The ultimate gradient lifting tree is based on the idea of gradient lifting tree, and the final prediction result is obtained by summing up the results of multiple decision trees.The model generates new trees through continuous iteration, and the trees generated by each iteration can fit the residual predicted by the previous tree, and iterates many times to form a strong classifier composed of multiple weak classifiers [25].

(3) BP neural network

BP neural network is a kind of multi-layer feedforward neural network, which consists of input layer, hidden layer and output layer, each layer is composed of several neurons, and the neurons in the same layer do not interfere with each other and are independent of each other. The neurons between different layers are nonlinear connected through parameters such as threshold and weight value, and have strong nonlinear mapping ability [26,27].The process of BP neural network is mainly divided into two stages: data forward propagation and error back propagation [28].

1) Data forward propagation

The data enters the network from the input layer, carries out nonlinear changes to the input data through the neurons of the hidden layer and the output layer, and carries out error analysis on the predicted value and expected value. If the error is large, it needs to be transferred to the error backpropagation.

The output of the JTH node of the hidden layer is:

$$\alpha_j = f(\sum_{i=1}^{n} u_{ij}x_i + b_j) \tag{6}$$

Where u_{ij} is the weight between the i-th node and the j-th node b_j is the threshold of node j.

The output of thek-th node of the output layer is:

$$y_k = f\left(\sum_{i=1}^{m} w_{ik}\alpha_i + d_k\right) \tag{7}$$

2) Error backpropagation

Error backpropagation is to transmit the output error layer by layer to the input layer through the hidden layer, and then adjust the network weight and threshold of each layer to make the error function decline along the gradient direction, so that the BP neural network prediction results are constantly approaching the expected output.The error function of BP neural network is:

$$E = \frac{1}{2}\sum_{i=1}^{l}\left(t_i - y_i\right)^2 \tag{8}$$

Where, ti is the expected value of the i-th node of the output layer;yi is the predicted value of the i node in the output layer.

3. Results and Analysis

3.1. Analysis of volatile compounds in Raw Liquor by GC-MS

Volatile compounds in raw wine samples were detected by the GC-MS detection method mentioned in 1.3.1 above, and the total ion flow chromatogram was shown in Figure 1. The names and contents of volatile compounds are shown in Table 2.

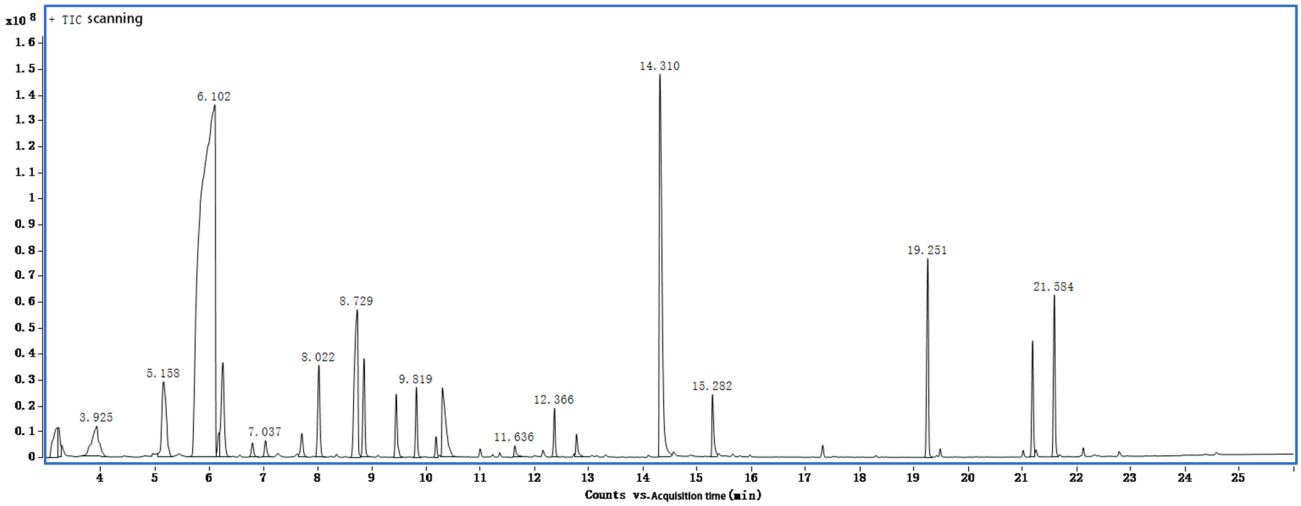


Figure 1. Total ion flow chromatogram of representative sample of Luzhou-flavor raw wine

Table 2. Volatile Compounds Content Information (mg/L)

Serial number	Compound name	Contentrange range
1	1,1-Diethoxy-3-methylbutane	0~163.67
2	Ethyl laurate	0~19.38
3	Ethyl nonanoate	0~17.92
4	Ethyl 9-hexadecenoate	0~111.94
5	(2,2-diethoxyethyl)-benzene	0~10.86
6	Ethyl decanoate	3.22~27.79
7	Butyl Lactate	0~14.7
8	Hexyl acetate	0~53.62
9	n-Pentanoic acid	1.57~47.21
10	Isoamyl Lactate	0~22.32
11	Propyl caproate	0~90.93
12	Ethyl Tetradecanoate	3.54~204.84
13	Isoamyl Caproate	4.47~74.04
14	Ethyl valerate	1.35~246.06
15	2-Methylbutanol	0.68~65.73
16	Ethyl trans-oleate	17.34~855.02
17	Ethyl heptanoate	17~274.63
18	Hexyl caproate	20.92~284.69
19	Ethyl 2-hydroxy-4-methyl-pentanoate	0~87.1
20	Butyric acid	23.01~188.34
21	Ethyl hexadecanoate	24.86~1454.88
22	Ethyl caprylate	22.65~404.03
23	2-Ethylbutyric acid	181.82~181.82
24	Hexanoic acid	62.61~387.76
25	Isoamyl alcohol	0~181.19

26	Acetic acid	68.48~295.48
27	Ethyl hexanoate	114.49~3684.09
28	Ethyl L(-)-lactate	151.17~720.82
29	Ethyl heptadecanoate	0~16.43
30	Ethyl Pentadecanoate	0~37.96
31	Hexyl formate	0~156.14
32	Ethyl octadecanoate	0~99.91
33	Ethyl linolenate	0~199.55
34	Ethyl oleate	0~4.57
35	Butyl caproate	12.4~189.14
36	Ethyl linoleate	28.19~1298.21
37	Palmitic acid	0~179.98
38	2-Heptanol	0~4.29
39	Isobutyl caproate	0~11.23
40	Hexamethylcyclotrisiloxane	0~6.26
41	Pentyl caproate	0~42.87
42	Ethyl 3-phenylpropionate	0~73.5
43	Ethyl phenylacetate	0~31.09
44	3-Methyl-2-butanol	0~4.2
45	Butyl Butyrate	0~30.21
46	Butyl isovalerate	0~5.98
47	6,10,14-Trimethyl-2-pentadecanone	0~10.78
48	(2R,3R)-(-)-2,3-Butanediol	0~30.66
49	Pentyl butyrate	0~10.61
50	3-Methyl-2-Hexanol	0~4.03
51	p-Cresol	0~17.27
52	Isobutyric acid	0~16.1
53	Phenethyl alcohol	0~13.52
54	Isoamyl Butyrate	0~9.45
55	Isoamyl acetate	0~18.5
56	Nonanal	0~3.15
57	Isoamyl valerate	0~2.92
58	3-Hydroxy-2-butanone	0~4.81
59	Heptyl valerate	0~5.53
60	2-tert-butyl-3-methyl-oxirane	0~204.9
61	Diethyl succinate	0~15.5
62	Octanoic acid	0~57.66
63	1,2-Propanediol	0~41.66
64	Ethyl 10-undecenoate	0~12.95
65	Hexanal	0~3.36
66	Heptanoic acid	0~25.78
67	cis-8,11,14-eicosatrienoic acid	0~9
68	Ethyl methylpentanoate	0~8.03
69	Benzaldehyde	0~2.76
70	Ethyl benzoate	0~2.42
71	Octyl formate	0~9.51
72	Ethyl Butyrate	0~24.58
73	Butyl Valerate	0~6.91
74	2-Ethyl-2-methyl-1,3-propanediol	0~3.22
75	Phenethyl acetate	0~3.42
76	2-Phenylethyl Hexanoate	0~17.8

As can be seen from Table 2, a total of 76 volatile substances were isolated and identified from all the original wine samples, of which the esters were the most abundant, totaling 47; there were also 9 acid compounds (except for 2-ethylbutyric acid); 9 alcohol compounds; 5 aldehydes and ketones; and 5 other compounds, which is in line with the research result that ester compounds have the richest variety and the most abundant content in the strong-flavored white wine [29].

3.2. Feature extraction and original wine grade identification modeling

Stratified sampling was used to divide the 256 raw wine samples into a training set and a test set in the ratio of 7:3. The training set was used for feature selection and model construction, and the test set was used for model prediction. The 76 feature compounds in Table 2 were feature extracted using PCA and KPCA. In general, the principal components

with cumulative variance contribution rate >95% were selected to fully characterize the structural information of the original data, so the principal components with cumulative variance contribution rate of 95% were retained, and the variance contribution rate and cumulative variance contribution rate of each principal component are shown in Fig. 2.

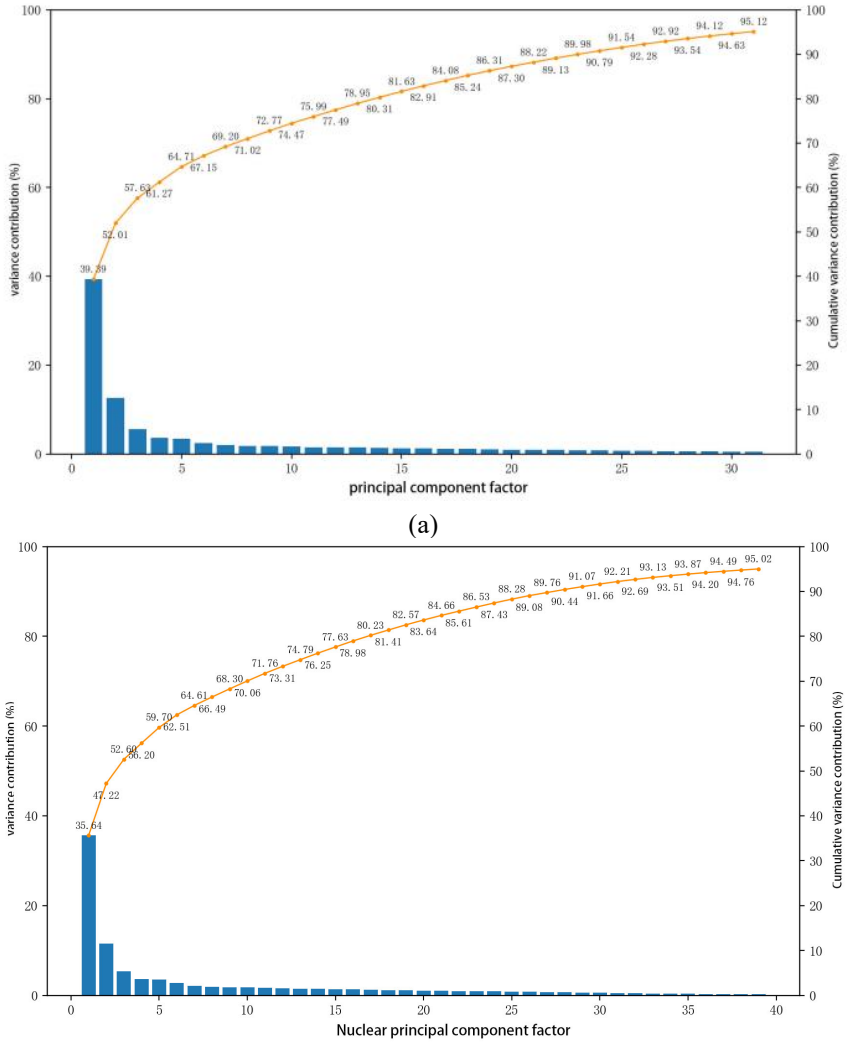


Figure 2. Pareto chart of principal components based on raw data

As can be seen from Fig. 2, the cumulative variance contribution rate of the first 31 principal components of PCA is 95.12%, and the cumulative variance contribution rate of the first 39 principal components of KPCA is 95.02%. Therefore, the first 31 principal components of PCA and the first 39 principal components of KPCA are selected as the inputs, and the grade of the original wine is taken as the output, and the original wine grade is established based on the SVM, XGBoost, and BP neural network, respectively. grade identification model, and its modeling results are shown in Table 3.

Table 3. Identification results based on feature extraction of raw data combined with different models

model	**original**	
	training set%	test set%
SVM	98.88	89.61
XGBoost	100	89.61
BP neural network	100	88.31
model	**PCA**	
	correction set%	prediction set%
SVM	100	84.42
XGBoost	100	79.22
BP neural network	100	72.73
model	**KPCA**	
	correction set%	prediction set%
SVM	98.32	85.71
XGBoost	100	81.82
BP neural network	100	84.42

As can be seen from Table 3, the prediction accuracies of the classification models established by direct feature extraction on the original data all decreased on the basis of the original data, which may be due to the fact that the changes in the content of some compounds in the original wine do not have a positive effect on the grade identification of the original wine, which belongs to the redundant features, and the existence of the redundant features may affect the condensation of the feature information by the PCA and the KPCA, so it is proposed to carry out a feature screening process before feature extraction. Therefore, it is proposed to perform feature screening before feature extraction to retain the features that have a greater impact on the grade classification of the original wine. Correlation analysis can determine the degree of correlation between the compounds and the grade of the original wine, so as to determine which feature compounds are more important for the identification of the grade of the original wine. Therefore, the Spearman rank correlation coefficient was used to analyze the correlation between the 76 detected feature compounds and the grades of the original wines before feature extraction by PCA and KPCA, and the features with a higher correlation with the grades of the original wines were selected by setting the correlation coefficient threshold at 0.5 for the subsequent analysis, and the correlation coefficients with a correlation coefficient of more than 0.5 for the grades of the substances are shown in Fig. 3. Power diagram.

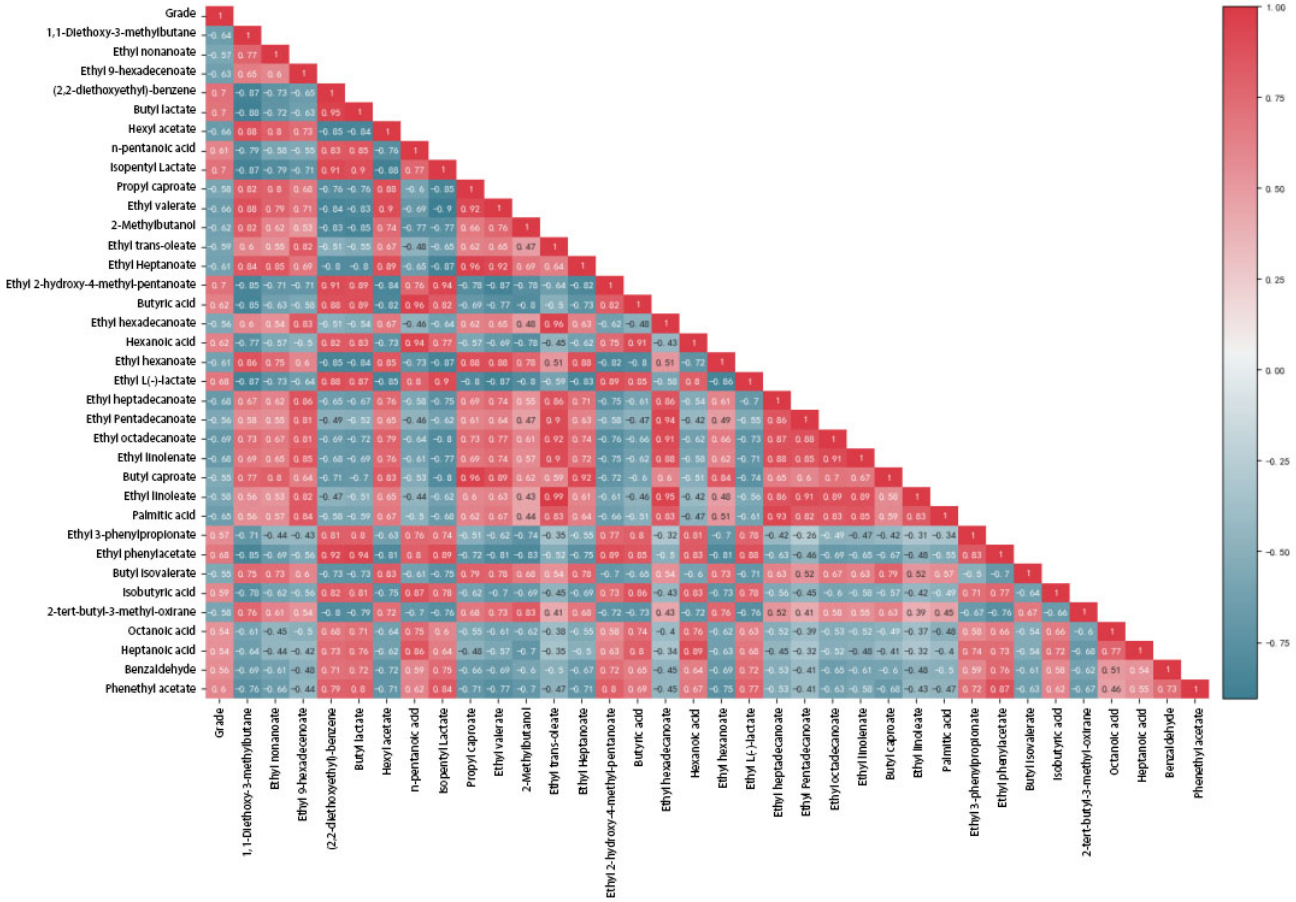
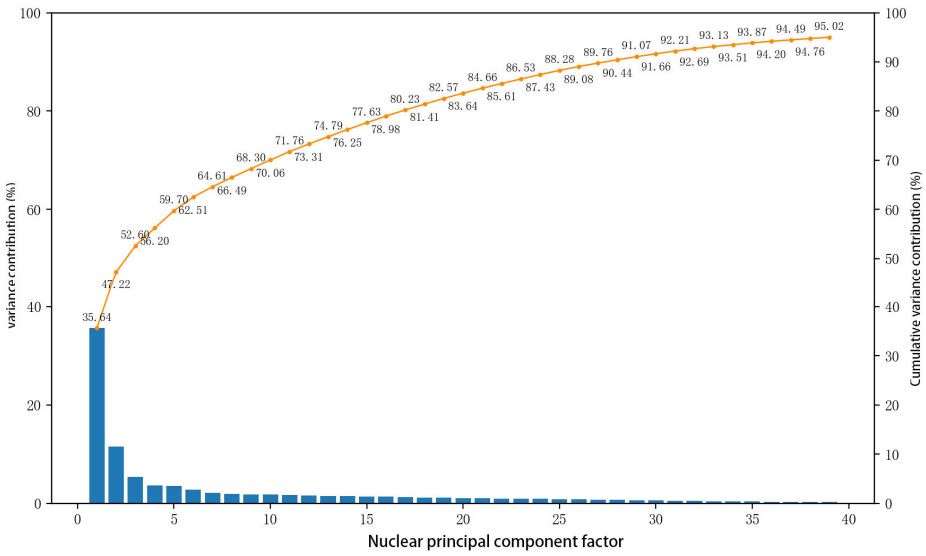


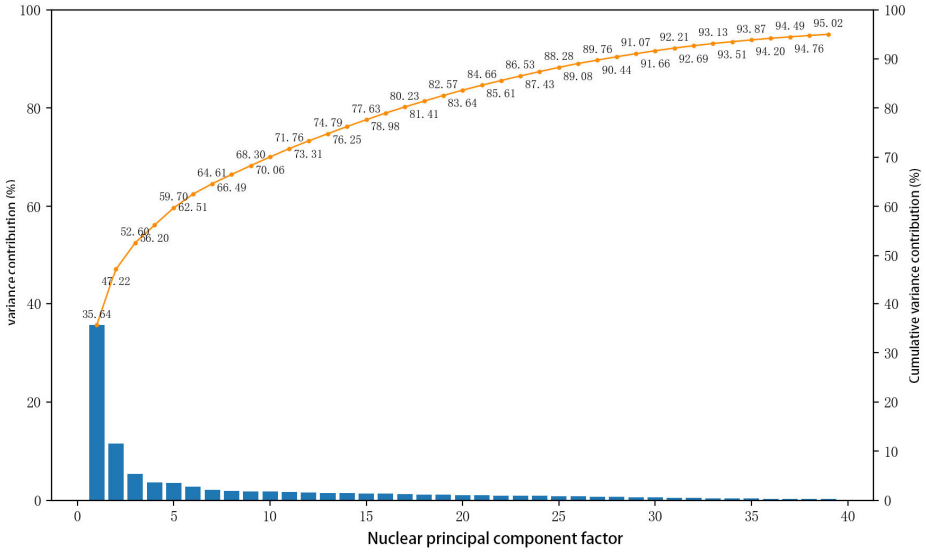
Figure 3. Heat map of Spearman's rank correlation coefficient (rs>0.5)

As can be seen in Fig. 3, there were 35 feature compounds with greater correlation with the grade of the original wine selected by setting the Spearman correlation coefficient threshold to 0.5, and then the 35 feature compounds screened were subjected to secondary feature extraction by using PCA and KPCA, which transformed the feature compounds with stronger correlation into a set of linearly uncorrelated feature variables, and obtained the cumulative variance contribution rate of >95%. The results of feature extraction are shown in Figure 4.

As can be seen from Fig. 4, using PCA and KPCA for secondary feature extraction of the 35 feature compounds selected by spearman, the number of principal components with a cumulative variance contribution rate of 95% were 9 and 12, respectively, which retained fewer feature dimensions compared to the direct feature extraction of the original data, and the model prediction results of the original wine grade identification model established based on the retained principal components are shown in Table 6.



(a)



(b)

Figure 4. Pareto chart for secondary feature extraction based on Spearman feature screening

Figure Notes: (a) Principal Component Pareto Chart; (b) Nuclear Principal Component Pareto Chart

Table 4. Identification results based on secondary feature extraction combined with different classification models

model	Spearman	
	correction set%	prediction set%
SVM	98.88	92.21
XGBoost	99.44	88.31
BP neural network	100	89.61
model	**Spearman-PCA**	
	correction set%	prediction set%
SVM	100	87.01
XGBoost	99.44	90.91
BP neural network	100	90.91
model	**Spearman-KPCA**	
	correction set%	prediction set%
SVM	98.32	92.21
XGBoost	100	90.91
BP neural network	99.44	96.10

As can be seen from Table 4, the prediction effect of the classification model established by PCA and KPCA combined with Spearman secondary feature extraction is better than that of direct feature extraction of the original data by PCA and KPCA, which is due to the fact that the feature compounds that have a greater impact on the grading of the original wine are retained through the correlation analysis of Spearman, so that the principal components of the secondary feature extraction are better able to characterize the quality of the original wine. This is because after Spearman correlation analysis, the characteristic compounds that have a greater influence on the classification of raw wine grades are retained, so that the principal components after secondary feature extraction can better characterize the change of the quality of raw wine; the prediction accuracy of the model established by Spearman-KPCA combined with SVM, XGBoost and BP neural network is higher than that of Spearman-PCA, and this may be due to the fact that the change of the content of the substances in the raw wines of different grades is not simple linear increase or decrease. KPCA introduces a kernel function on the basis of PCA, which can better extract the nonlinear information in the data set; among them, the accuracy of the correction set and prediction set of the original

wine grade identification model established by Spearman-KPCA in combination with BP neural network is 99.44% and 96.10%, respectively, and the prediction effect is the best, and the specific discriminative results of the model are shown in Fig. 5, and the confusion matrix of the classification results of the model is shown in Table 5.

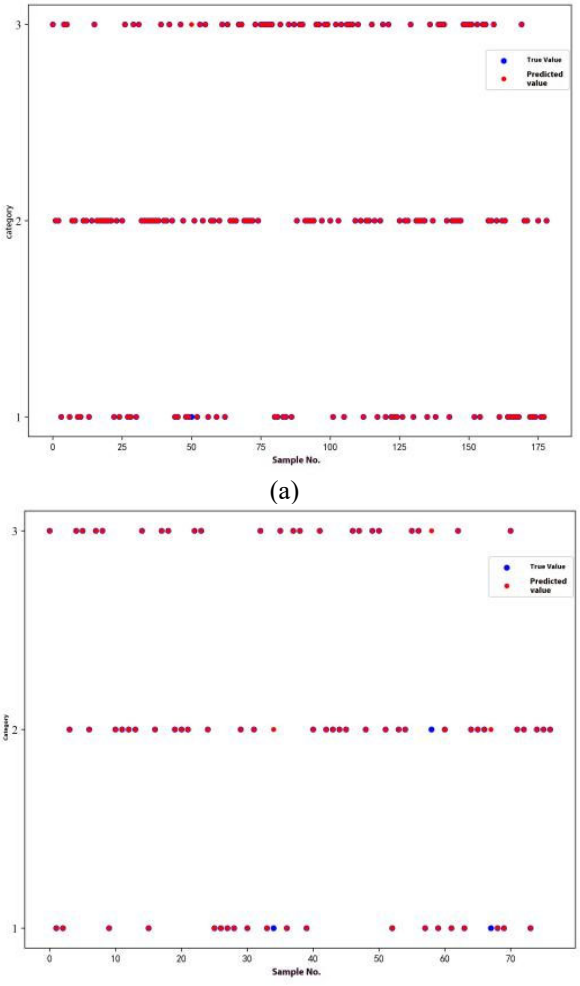


Figure 5. Discriminative results of Spearman-KPCA-BP neural network model correction set and prediction set

Table 5. Confusion matrix for discriminatory results of Spearman-KPCA-BP neural network models

true category	correction set			
	First drink	Mid-course wine	Tail wine	Accuracy rate%
First drink	49	0	1	98
Mid-course wine	0	75	0	100
Tail wine	0	0	54	100
true category	**prediction set**			
	First drink	Mid-course wine	Tail wine	Accuracy rate%
First drink	20	2	0	90.91
Mid-course wine	0	31	1	96.88
Tail wine	0	0	23	100

As can be seen from Table 5, there were 179 samples in the correction set of the Spearman-KPCA-BP neural network model, and only one sample was misclassified, and the classification accuracies of its head, middle, and tail wines were 98%, 100%, and 100%, respectively; in addition, there were 77 samples in the prediction set, and three samples were misclassified, and the classification accuracies of its head, middle, and tail wines were 90.91%, 96.88%, and 100%. This result indicates that the Spearman-KPCA-BP neural network model is able to accurately predict the grades of full-flavored original wine.

4. Conclusion

In this study, we took different grades of strong-flavored original wine as the research object, obtained the volatile components mapping data of original wine through GC-MS technology, and used Spearman combined with PCA and KPCA to downscale the original data, and then established the grade identification model of strong-flavored original wine by combining with SVM, XGBoost and BP neural network, respectively. The main conclusions of the experiments are: (1) Spearman feature screening of the GC-MS raw data set can effectively improve the condensation of feature information by PCA and KPCA; (2) the prediction accuracy of the grade identification model established after the secondary feature extraction by KPCA is higher, which indicates that KPCA is better able to excavate the hidden difference information in the data of different grades of raw wines compared to PCA;(3) The grade discrimination model established by Spearman-KPCA combined with BP neural network is the most effective in discriminating different grades of raw wines, with the prediction accuracy of 99.44% and 96.10% in the calibration set and test set, which has strong practical application value. The results show that the Spearman-KPCA-BP neural network model can well realize the classification of different grades of raw wine, which provides a new idea and theoretical basis for the quality control and grade identification of raw wine, and also provides a valuable reference means for other quality identification of liquor.

References

[1] SUN J, ZHAO D, ZHANG F, et al. Joint direct injection and GC–MS chemometric approach for chemical profile and sulfur compounds of sesame-flavor Chinese Baijiu (Chinese liquor)[J/OL]. European Food Research and Technology, 2018, 244(1): 145-160.

[2] HONG J, TIAN W, ZHAO D. Research progress of trace components in sesame-aroma type of baijiu[J/OL]. Food Research International, 2020, 137: 109695.

[3] LIU Q R, ZHANG X J, ZHENG L, et al. Machine learning based age-authentication assisted by chemo-kinetics: Case study of strong-flavor Chinese Baijiu[J/OL]. Food Research International, 2023, 167: 112594.

[4] QIAN Y, ZHANG L, SUN Y, et al. Differentiation and classification of Chinese Luzhou-flavor liquors with different geographical origins based on fingerprint and chemometric analysis[J/OL]. Journal of Food Science, 2021, 86(5): 1861-1877.

[5] Zhou Xuan. Research on volatile composition analysis and grade identification of base wine of strong aromatic liquor [D/OL]. Jiangsu University, 2019.

[6] CAMARA J S, MEDINA S, PERESTRELO R. Recent Developments in the Applications of Fingerprinting

Technology in the Food Field [J/OL]. FOODS, 2022, 11(14): 2006.

[7] LIU Fei. Discussion on the application of gas chromatography-mass spectrometry in food analysis[J/OL]. Modern Food, 2020(11): 167-168.

[8] Han Yuncui. Aroma modeling and automated wine picking for strong-flavored base wines [D/OL]. Qilu University of Technology, 2023.

[9] Liu Qingru, Meng Lianjun, Zhang Xiaojuan, et al. Identification of storage time of Lu-type base wine based on GC-MS fingerprinting and XGBoost machine learning[J]. Food Science, 2022, 43(24): 310-317.

[10] QIAN Yu, HU Xue, SUN Yue, et al. Classification of strongly flavored liquor based on fingerprinting and chemometrics[J]. China Brewing, 2021, 40(6): 152-156.

[11] ZHU Kaixian, HU Xue, DENG Jing, et al. Discriminative analysis of different aromatic liquors based on GC-MS technology[J]. China Brewing, 2023, 42(1): 213-218.

[12] FAN Sanshuan, TANG Jie, LUO Xianxuan, et al. Classification of small-square clear-flavored wine based on HS-SPME-Arrow-GC-MS and chemometrics[J/OL]. Food and Fermentation Industry, 2021, 47(13): 254-260.

[13] LIU Y, QIAO Z, ZHAO Z, et al. Comprehensive evaluation of Luzhou-flavor liquor quality based on fuzzy mathematics and principal component analysis[J/OL]. FOOD SCIENCE & NUTRITION, 2022, 10(6): 1780-1788.

[14] YAO Y, MENG H, GAO Y, et al. Linear dimensionality reduction method based on topological properties[J/OL]. Information Sciences, 2023, 624: 493-511.

[15] ELHENAWY M, MASOUD M, GLASER S, et al. A New Approach to Improve the Topological Stability in Non-Linear Dimensionality Reduction[J/OL]. IEEE ACCESS, 2020, 8: 33898-33908.

[16] REN Yulan, TIAN Mi, LI Chunyan, et al. Gas chromatographic analysis of trace components in liquor[J]. China Brewing, 2011(7): 177-179.

[17] YAN Y, CHEN S, NIE Y, et al. Quantitative Analysis of Pyrazines and Their Perceptual Interactions in Soy Sauce Aroma Type Baijiu[J/OL]. Foods, 2021, 10(2): 441.

[18] LAN Wenbao, CHE Chang, TAO Chengyun. Spearman rank correlation-based single-acting spectral component selection and its application to SAR target identification[J/OL]. Journal of Radio Science, 2020, 35(3): 414-421.

[19] CHEN X, HOU Y, XI P. Parameter estimation of the structured illumination pattern based on principal component analysis (PCA): PCA-SIM[J/OL]. LIGHT-SCIENCE & APPLICATIONS, 2023, 12(1): 41.

[20] ZHAI Shuang, TOU Xiangguo, ZHANG Guiyu, et al. Rapid discrimination of white spirit base wine based on FT-NIR spectroscopy combined with KPCA-MD-SVM[J/OL]. Modern Food Science and Technology, 2022, 38(4): 248-253.

[21] HE Y, YE L, ZHU X, et al. Feature extraction based on PSO-FC optimizing KPCA and wear fault identification of planetary gear[J/OL]. Journal of Mechanical Science and Technology, 2021, 35(6): 2347-2357.

[22] ZHANG K, ZHANG K, BAO R. Prediction of gas explosion pressures: A machine learning algorithm based on KPCA and an optimized LSSVM[J/OL]. Journal of Loss Prevention in the Process Industries, 2023, 83: 105082.

[23] CHEN H, TAN C, WU T, et al. Discrimination between authentic and adulterated liquors by near-infrared spectroscopy and ensemble classification[J/OL]. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2014, 130: 245-249.

[24] LIU G, WANG L, LIU D, et al. Hyperspectral Image Classification Based on Non-Parallel Support Vector Machine[J/OL]. Remote Sensing, 2022, 14(10): 2447.

[25] GÜNDOĞDU S. Efficient prediction of early-stage diabetes using XGBoost classifier with random forest feature selection technique[J/OL]. Multimedia Tools and Applications, 2023.

[26] CHEN R, JIA B, MA L, et al. Marine Radar Oil Spill Extraction Based on Texture Features and BP Neural Network[J/OL]. JOURNAL OF MARINE SCIENCE AND ENGINEERING, 2022, 10(12): 1904.

[27] YANG Y, LIU H, GU Y. A Model Transfer Learning Framework With Back-Propagation Neural Network for Wine and Chinese Liquor Detection by Electronic Nose[J/OL]. IEEE ACCESS, 2020, 8: 105278-105285.

[28] Tang, Jianqing. Quantitative investment based on BP neural network [D/OL]. Soochow University, 2019.

[29] XU Y, ZHAO J, LIU X, et al. Flavor mystery of Chinese traditional fermented baijiu: The great contribution of ester compounds[J/OL]. Food Chemistry, 2022, 369: 130920.