ACS OMEGA

Open Access

http://pubs.acs.org/journal/acsodf

Article

Ethanol Concentration Determination in Baijiu by Graph-Regularized PCA and Random Forest-Based Raman Spectroscopy

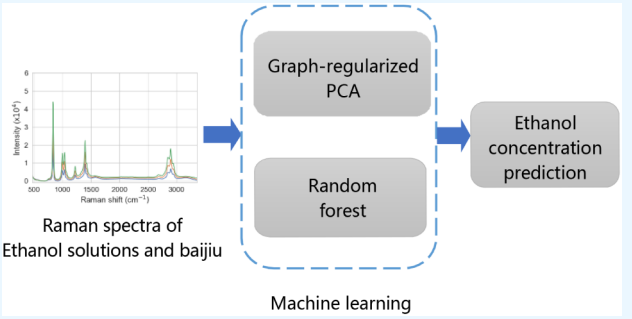
Zhenhao Chen, Zhuangwei Shi,* Jianchen Zi, Chenhui Wang, Hai Bi, and Yunlong Zhu

Cite This: *ACS Omega* 2025, 10, 14373−14381

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Baijiu is a type of traditional Chinese alcoholic beverage with significant economic and cultural value. Ethanol concentration determination through machine learning-based Raman spectroscopy offers the advantages of being contact-free and rapid, and the technique holds considerable potential for baijiu quality control in the industrial manufacturing process. However, current applications of Raman spectroscopy for the quantitative analysis of biochemical materials are restricted by measurement accuracy, as well as the flexibility and robustness of chemometric tools. To address these issues, we propose a method that combines graph-regularized principal component analysis (graph-regularized PCA) and an ensemble learning framework, random forest, to capture effective low-dimensional representations from high-dimensional Raman spectra data while reducing spectra data instability. Furthermore, we propose a protocol that adopts ethanol solutions with various concentrations as the training set for fitting a single regression model to determine the ethanol concentrations of different types of baijiu. In ethanol concentration detection across all three types of baijiu, our proposed method achieves a mean average percentage error (MAPE) of 0.415% on ethanol concentration determination of all three types of baijiu, outperforming all other methods. The results validate the accuracy and robustness of our proposed method.



INTRODUCTION

Baijiu, as a unique traditional Chinese alcoholic beverage with important economic and cultural values, requires precise detection to ensure consistent quality and flavor.[1,2] Therefore, many detection techniques have been applied to baijiu. For instance, He et al.[3] adopted gas chromatography−mass spectrometry (GC-MS) and descriptive sensory analysis to explore and differentiate the chemosensory characteristics of strong-aroma-type Baijiu from various regions. Yan et al.[4] presented the characterization of volatile compounds in baijiu through the application of high-performance liquid chromatography−mass spectrometry (HPLC-MS). Hu and Wang[5] utilized mid-infrared spectroscopy for age discrimination of baijiu. However, as these traditional methods require laborious and expensive preprocessing of the samples, it is necessary to develop time-efficient and cost-effective detection methods for baijiu.

Raman spectroscopy, widely employed in biology,[6,7] chemistry,[8,9] optics,[10,11] and numerous other fields, exhibits significant potential in the noninvasive analysis of food materials. Notably, its efficiency is particularly highlighted in the fingerprint collection of aqueous food products, where the weak Raman effect of water molecules poses challenges for other analytical techniques. Examples of such food products include whiskey,[12] honey,[13] and oil,[14] among others. Since machine learning can significantly improve the efficiency of spectral data analysis, previous studies have demonstrated the feasibility of the quantitative analysis of baijiu through Raman spectroscopy in combination with machine learning algorithms. For instance, Wu et al.[15] utilized Raman spectroscopy and a hybrid model combining soft independent modeling of class analogy (SIMCA) and linear discriminant analysis (LDA) for the classification of Chinese baijiu. Gu et al.[16] adopted deep learning on Raman spectra for baijiu flavor classification. Wang et al.[17] combined LDA and LightGBM (light gradient boosting machine) for baijiu classification.

The precise determination of ethanol concentration is crucial for the quality control of alcoholic beverages. Traditional methods for ethanol concentration determination include the densitometer method, distillation method, HPLC, GC-MS, and so on.[18−20] Since the characteristic Raman peaks of alcoholic beverages are contributed by ethanol and water, Raman spectroscopy can be a promising technique for the rapid and contact-free determination of ethanol concentration.[12,21]

Received: January 20, 2025
Revised: March 26, 2025
Accepted: March 28, 2025
Published: April 3, 2025

Recently, machine learning-based Raman spectroscopy has been applied for baijiu ethanol concentration determination. Liu et al.[22] utilized a PCA-based support vector machine for baijiu quality and ethanol concentration determination using Raman spectroscopy. Zong et al.[23] combined neural networks and a genetic algorithm for analyzing Raman spectral data to predict ethanol concentration, flavor grade, and production year of baijiu.

Generally, these methods of machine learning-based Raman spectroscopy can be summarized into two stages: (1) dimensional reduction algorithms (e.g., PCA) for processing raw spectral data, and (2) machine learning models (e.g., support vector machines) for classification or regression. However, there are several drawbacks to these methods: (1) Dimensional reduction algorithms in previous studies can be categorized into supervised algorithms and unsupervised algorithms. Compared with unsupervised algorithms (e.g., PCA), supervised algorithms (e.g., LDA) require label information for representation learning, which can enhance the robustness of dimensional reduction but may reduce generalization. (2) Although the trace components in baijiu are not directly reflected in its Raman spectrum as characteristic Raman peaks, the minor constituents also subtly influence the Raman spectroscopy. The data complexity makes it difficult for a single machine learning model to sufficiently extract the spectral information for accurate quantification. (3) Current methods adopting machine learning and Raman spectroscopy for ethanol concentration determination commonly conduct cross-validation on a single type of alcoholic beverage, which limits the applicability of the methods to a wider range of products.

Given these challenges, researchers have explored alternative approaches to enhance the robustness and efficiency of dimension reduction algorithms. One such approach is manifold learning,[24] which proposes a manifold constraint to enhance the robustness and efficiency of unsupervised dimension reduction algorithms. The manifold constraint clarifies that samples are distributed on a manifold, and samples with higher feature similarities are closer on the manifold. Manifold regularization provides an optimized way for capturing effective low-dimensional representations from high-dimensional features. The manifold of data can be depicted by a graph structure constructed through feature similarity, which leads to graph regularization. On the other hand, due to the low-rank constraint, dimension reduction algorithms (e.g., PCA) are performed to obtain low-dimensional (i.e., low-rank subspace) representations from the high-dimensional space of features.[25,26] Therefore, graph-regularized PCA can obtain representations following both the manifold constraint and the low-rank constraint simultaneously, which can significantly reduce the effect of noise and improve the performance of machine learning models.[27,28]

Ensemble learning is another approach to tackle these challenges and improve learning performance.[29,30] Representative ensemble learning algorithms include random forest (RF), AdaBoost (adaptive boosting), XGBoost (extreme gradient boosting tree), and LightGBM (light gradient boosting machine). Compared with traditional machine learning algorithms, ensemble learning works by integrating multiple base learners (usually decision trees), which can significantly improve the effectiveness of the model in extracting information. Ensemble learning also has significant advantages in computing speed, model robustness, and parallel deployment.

Therefore, to address the disadvantages of previous methods, we propose an ethanol concentration determination method for baijiu that integrates manifold learning and ensemble learning through graph-regularized PCA and random forest-based Raman spectroscopy. In summary, our main contributions are as follows.

- We adopt graph-regularized PCA to ensure spectral data stability and enhance dimension reduction capability. Compared with PCA, this proposed dimension reduction algorithm is advantageous in terms of accuracy and robustness.
- We adopt random forest for effective representation learning from Raman spectrum data. Compared with nonensemble methods, random forest is superior to nonensemble machine learning algorithms in capturing information from Raman spectra with higher efficiency.
- We propose a protocol utilizing ethanol solutions with various concentrations as the training set for baijiu ethanol concentration determination. The single quantitative model fitted with the training set can therefore be used for multiple baijiu products.

We conducted cross-validation on ethanol solutions with various concentrations for model training. Then, the trained model was directly applied to several types of baijiu with different ethanol concentrations. Our proposed method achieved a mean average percentage error of 0.415% for all three types of baijiu, outperforming all other methods. The results demonstrate that Raman spectra data processed by graph-regularized PCA can achieve better performance during machine learning prediction. The protocol using ethanol solutions as the training set enhances extensibility and flexibility while adopting machine learning and Raman spectroscopy for ethanol concentration determination. Compared with other machine learning algorithms, random forest performs better in the ethanol concentration determination of baijiu.

INPUT PREPARATION

Data Collection. We propose a data collection protocol that employs ethanol solutions with different concentrations of baijiu as the training set and different types of baijiu as the testing set. We conducted cross-validation on a series of aqueous ethanol solutions with various concentrations for model training. Then, the trained model is directly applied to several types of baijiu with different ethanol concentrations.

Using anhydrous ethanol (Zesheng Technology, Anhui, China) and HPLC-grade water, ethanol−water solutions with concentrations ranging from 20 to 60 vol %, at 5 vol % intervals , were prepared for the data collection of the training set. Thus, there are nine types of ethanol solution concentrations . Each type of ethanol solution contains 24 samples. In total, there are 216 samples for the data collection of the training set.

Then, we purchased three types of Chinese baijiu with different ethanol concentrations from the local supplier: (1) Jiujiang Shuangzheng Jingmi 30 (Jinm30), (2) Shiquanyuye (Sqyy), and (3) Zhangyizhai (Zyz). Ethanol concentration concentration values of these three types of baijiu were determined by a densitometer (DMA35, Anton Paar, Shanghai, China), and ethanol concentration concentration of Jinm30, Sqyy, and Zyz were 30.9 vol %, 42.0 vol %, and 53.3 vol %, respectively.

Raman Spectrum Acquisition. The entire Raman spectroscopy system is placed in a temperature-controlled cabinet

https://doi.org/10.1021/acsomega.5c00616
ACS Omega 2025, 10, 14373−14381

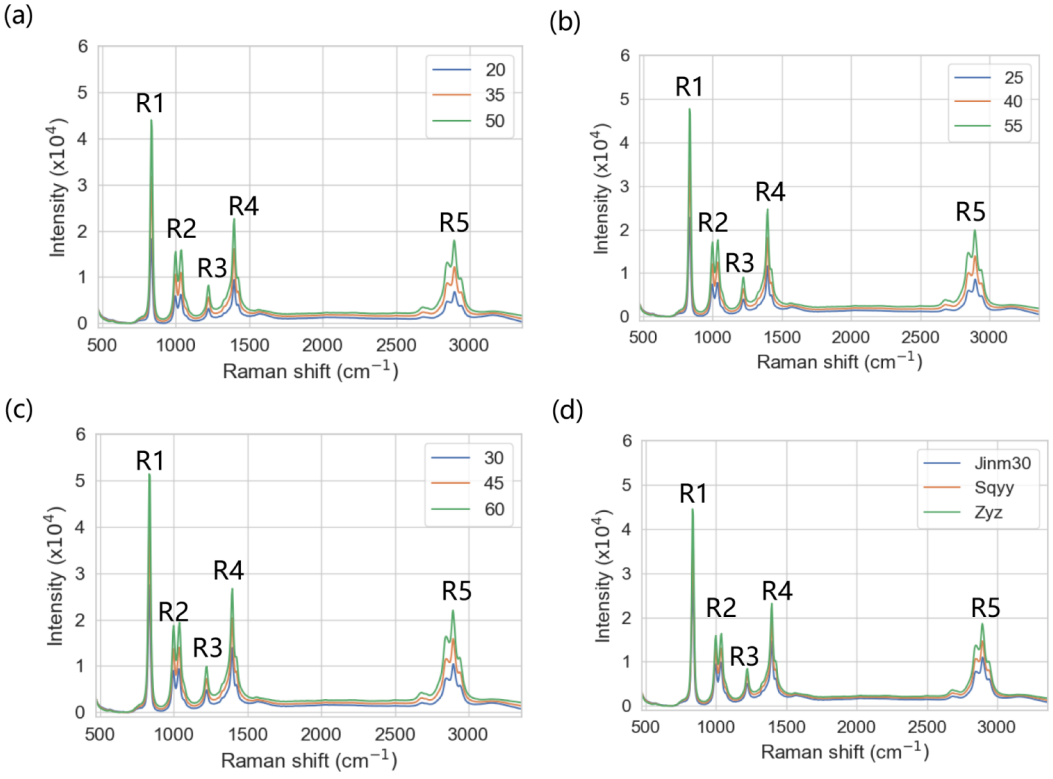


Figure 1. Mean Raman spectra of (a) 20, 35, and 50 vol % ethanol solution samples, (b) 25, 40, and 55 vol % ethanol solution samples, (c) 30, 45, and 60 vol % ethanol solution samples, and (d) baijiu samples.

(SNKM-250L, SNUGEN, China), which maintains a constant temperature of 25 ± 0.1 °C. The Raman spectral acquisition is performed using a NOVA2S Raman spectrometer (7P-785-R, Ideaoptics, China), equipped with a 785 nm fiber laser source (fL-785-T05-R, Ideaoptics, China) whose bandwidth is less than 0.5 nm and power stability is less than 1%, operating at the maximum power of 523 mW. The focal length of the optical probe is 7.5 mm, and the numerical aperture (NA) is 0.22.

The Raman shift, ranging from 470 to 3360 cm^{-1} was recorded at a resolution of 6 cm^{-1}. Each standard ethanol solution and each bottle of baijiu were sampled in triplicate into customized 10 mm path quartz cuvettes with one translucent surface for excitation and collection of Raman signal and three matte surfaces to prevent interference outside the cuvette. Samples in the training set and testing set are put into quartz cuvettes. To avoid interference from external light sources, the detection module is covered with black shading cloth for Raman data acquisition. Each Raman spectrum is preprocessed by background correction using adaptive iterative reweighted penalized least-squares (airPLS).[31] The dimension of each Raman spectrum was 700.

Raman Spectroscopy of Ethanol Solution and Baijiu. The raw mean Raman spectra of ethanol solutions are shown in Figure 1a−c. It can be seen from the figure that the spectra of ethanol solutions with different concentrations share the same characteristic Raman peaks, R1−R5 (Table 1). The spectra contain five characteristic peaks. Baijiu is composed of ethanol, water, and trace components produced during brewing, and the signals of the trace components are difficult to observe directly from the spectrum. Therefore, these peaks in the diagram all originate from the main components of baijiu, i.e., ethanol and water,[32] and the Raman peak height rises with the increase in ethanol concentration. Moreover, the raw mean Raman spectra of baijiu samples are shown in Figure 1d. There is no significant difference in the position and number of Raman peaks between ethanol solution samples of different concentrations and baijiu samples.

Table 1. Raman Peaks of Ethanol Solutions and Baijiu

	Raman shift (cm^{-1})	Explaination
R1	888	C−C−O in-plane stretching
R2	1060−1100	C−C−O out-plane expansion and CH_3 in-plane sway + δ (CHO)
R3	1285	CH_2 torsion + δ (CHO)
R4	1463	Asymmetric deformation of $-CH_3$
R5	2950−3050	−OH stretching vibration of water

MACHINE LEARNING MODELS

Graph-Regularized PCA. *Principal Component Analysis (PCA).* Principal component analysis (PCA) is the most commonly used dimensionality reduction algorithm,[33] and it is widely applied for Raman spectra data analysis.[17,21] Suppose there are n samples, and the feature dimension of each sample is m, then $X \in \mathbb{R}^{n\times m}$ denotes the high-dimensional feature matrix. Dimensionality reduction aims to find an optimal representation matrix (known as the principal component) $U \in \mathbb{R}^{n\times r}$ $(r \ll m)$ to reduce the number of feature dimensions from m to r. PCA optimizes U through maximizing the covariance of X.

$$\max_{U} \quad \mathrm{tr}(U^T X X^T U),$$
$$\text{s.t.} \quad U^T U = I, \tag{1}$$

where tr(·) denotes the trace of a matrix. Hence, PCA can be solved through conducting the singular value decomposition $X =$

https://doi.org/10.1021/acsomega.5c00616
ACS Omega 2025, 10, 14373−14381

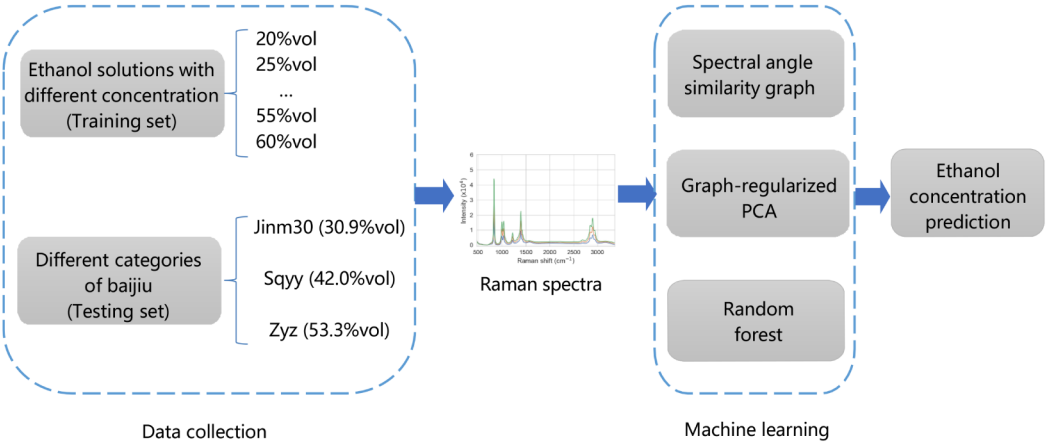


Figure 2. Flowchart of our proposed framework.

USV^T. Here $S = \mathrm{diag}(s_1,s_2,...,s_r)$ is the r-th order diagonal matrix, and s_i is the i-th largest singular value of X. $U \in \mathbb{R}^{n\times r}$ and $V \in \mathbb{R}^{r\times m}$ are orthogonal matrices.

Spectral Angle Similarity Graph. To enhance the robustness and efficiency of PCA, graph regularization is added. Samples are nodes in the graph, and edges in the graph represent the connections and similarities among samples. In this paper, the graph structure among samples is constructed through spectral angle similarity. Spectral angle is widely applied for measuring the similarities among spectral data.[34] Suppose there are two samples, i and j, with spectra x_i and x_j, respectively, the cosine of the spectral angle

$$\cos\theta_{ij} = \frac{x_i^T x_j}{\|x_i\|\cdot\|x_j\|} \tag{2}$$

Since $x_i,x_j > 0$, $\cos\theta_{ij} \in [0,1]$ can measure the similarity between x_i and x_j, larger $\cos\theta_{ij}$ denotes that i and j are more similar. Then, following previous research,[26] the similarity graph can be constructed simply as follows. First, for each node i, sort the spectral angle between i and other nodes, and select the top-10 nodes with the largest spectral angle values, except itself. Second, suppose the set of these nodes for node i is $\mathcal{N}(i)$, matrix C satisfies that $C_{ij} = 1$ if $j \in \mathcal{N}(i)$, otherwise $C_{ij} = 0$. Finally, the adjacency matrix with a self-loop of the constructed graph is

$$A = C^T \odot C + I \tag{3}$$

where $\odot$ denotes the Hadamard product. Then, the normalized Laplacian matrix

$$L = I - D^{-1/2}AD^{-1/2} \tag{4}$$

where degree matrix $D = \mathrm{diag}(d_1,d_2,...,d_n)$, $d_i = \sum_j A_{ij}$.

Graph-Regularized PCA. eq 1 (PCA problem) is solved by conducting singular value decomposition (SVD) as $X = USV^T$. Let $W = SV^T$. SVD can be considered a non-negative matrix factorization problem through the Lee-Seung iteration algorithm,[35]

$$\min_{U,W>0} \|X - UW\|_F^2 + \mu\|U^TU - I\|_F^2 \tag{5}$$

where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix, and μ is a hyperparameter. In this way, since graph regularization is performed by minimizing the quadratic form of the normalized Laplacian matrix, graph-regularized PCA is optimized optimize

$$\min_{U,W>0} \|X - UW\|_F^2 + \lambda\mathrm{tr}(U^TLU) + \mu\|U^TU - I\|_F^2 \tag{6}$$

Following previous research,[36−38] eq 6 can be iteratively solved from random initial values of U and W until convergence, and hyperparameters can be set as $\lambda = \mu = 0.1$. The dimensions of the principal components are $r = 32$.

Random Forest. Random forest integrates multiple decision trees through bootstrap aggregating (bagging).[33] Each decision tree (i.e., base learner in ensemble learning) conducts bootstrap sampling of the training set to obtain a subtraining set. Each tree is trained based on different subtraining sets. The algorithm integrates the prediction values of all base learners by calculating mean values or employing a voting strategy to obtain the final prediction results.

Random forest combines a bagging strategy and a random selection of features to build an uncorrelated forest consisting of decision trees. The feature selection scheme generates a random subset of features, which is capable of learning effective low-dimensional representations from high-dimensional features. In random forest, the random selection of features adopts an adaptive dimension reduction strategy, which can significantly improve the performance of dimension reduction.

The flowchart of our proposed framework is shown in Figure 2. The machine learning algorithms are implemented in the Python 3.9.12 programming environment on a Windows 11 x64 system with an Intel Core i5-12400F 2.50 GHz CPU and 16 GB DDR3 RAM. In this paper, PCA or graph-regularized PCA is used in combination with five machine learning models: k-nearest neighbors (KNN), support vector machine (SVM), decision tree (DT), XGBoost (XGB), and random forest (RF). XGB and RF are ensemble learning models. The machine learning models are implemented using scikit-learn (version 1.0.2). The number of decision trees in the random forest is 50. The other parameters in the machine learning models are consistent with the default values of scikit-learn.

RESULT AND DISCUSSIONS

Ethanol Concentration Determination. To obtain a quantitative analysis model with better prediction accuracy, PCA or graph-regularized PCA is used in combination with five machine learning models: k-nearest neighbors (KNN), support vector machine (SVM), decision tree (DT), XGBoost (XGB), and random forest (RF). Here, SVM-G, KNN-G, DT-G, XGB-G, and RF-G represent five prediction models optimized with

https://doi.org/10.1021/acsomega.5c00616
ACS Omega 2025, 10, 14373−14381

the spectral data processed through graph-regularized PCA. SVM-P, KNN-P, DT-P, XGB-P, and RF-P represent five prediction models optimized with the spectral data processed through PCA.

Cross Validation on Ethanol Solutions. We conduct 4-fold cross-validation on a series of aqueous ethanol solutions with various concentrations for model training. Then, the trained model is directly applied to several types of baijiu with different ethanol concentrations. This regression model is evaluated using these metrics: R^2, mean absolute error (MAE), root-mean-square error (RMSE), and mean percentage error (MAPE). These metrics are defined as follow.

$$R^2 = \frac{\sum_{i=1}^{n} \left\| \hat{y}_i - \bar{y} \right\|^2}{\sum_{i=1}^{n} \left\| y_i - \bar{y} \right\|^2} \tag{7}$$

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |\hat{y}_i - y_i| \tag{8}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2} \tag{9}$$

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{\hat{y}_i - y_i}{y_i} \right| \times 100\% \tag{10}$$

where y_i and $\hat{y}_i$ denote the true value and predicted value of sample i, respectively, and $\bar{Y}$ denotes the mean value of y.

The results of average values in the 4-fold cross-validation are shown in Table 2. Higher R^2, and lower MAE, RMSE, and MAPE illustrate that the performance is better. All five machine learning models combined with graph-regularized PCA perform better than those combined with PCA. Ensemble learning models (random forest and XGBoost) perform better than other models. Our proposed RF-G model (integrating random forest and graph-regularized PCA) achieves the highest R^2 and the lowest MAE, RMSE, and MAPE. The results demonstrate the superiority of our proposed framework integrating graph-regularized PCA and random forest-based Raman spectroscopy for ethanol concentration determination. Figure 3 further illustrates the accuracy of our proposed framework.

Table 2. Results of Average Values in Cross-Validation on Ethanol Solutions[a]

Method	R^2	MAE (vol %)	RMSE (vol %)	MAPE (%)
SVM-G	0.997	0.527	0.660	1.287
KNN-G	0.997	0.520	0.650	1.280
DT-G	0.997	0.505	0.631	1.265
XGB-G	0.998	0.376	0.471	0.988
RF-G	**0.999**	**0.222**	**0.283**	**0.587**
SVM-P	0.996	0.588	0.735	1.535
KNN-P	0.996	0.580	0.726	1.486
DT-P	0.997	0.565	0.707	1.462
XGB-P	0.998	0.452	0.565	1.150
RF-P	0.999	0.301	0.377	0.773

[a]Bold numbers indicate the best results.

Ethanol Concentration Determination on Baijiu. After training models on a series of aqueous ethanol solutions with various concentrations, we evaluated their performance on several types of baijiu with different ethanol concentrations. Table 3 shows the prediction results of ethanol concentration in baijiu using the five models based on Raman spectra processed with graph-regularized PCA or PCA. Note that there are only three true labels, i.e., Jinm30 (30.9 vol %), Sqyy (42.0 vol %), Zyz (53.3 vol %); therefore, it is unnecessary to evaluate the performance using R^2. Hence, lower MAE, RMSE, and MAPE illustrate better performance, which can generally be obtained after applying graph-regularized PCA. For all three types of baijiu, the random forest model outperforms other methods. The random forest model derived from the Raman spectra data processed using graph-regularized PCA is the best method, achieving a MAPE of 0.415%. Tables S1−S3 show the detailed MAE, RMSE, and MAPE for the three types of baijiu, respectively, further illustrating the superiority of our proposed method. Previous methods using gas chromatography,[18] Raman spectroscopy,[12] or high-performance liquid chromatography[20] for ethanol concentration determination can achieve a MAPE of about 1% for different types of alcoholic beverages, but none of them can achieve accuracy comparable to our proposed method.

The main idea of the dimension reduction algorithm is to reduce high-dimensional feature data to low-dimensional space so that representative features can be extracted without the interference of noise.[39,40] This is the low-rank constraint for representation learning. The feature selection scheme in random forest adopts an adaptive dimension reduction strategy, which can flexibly capture important information from high-dimensional Raman spectra data.[41,42] Beyond the low-rank constraint, the manifold constraint performed on the graph based on feature similarity can depict the global consistency of data distribution, while the low-rank constraint pays more attention to the local saliency of each sample. Since the spectral angle graph can accurately depict the similarities among spectra data of samples, our proposed framework, adopting graph-regularized PCA, can model the low-rank constraint and manifold constraint simultaneously and precisely, then optimize both constraints through alternate iteration.[27]

To further evaluate the model's performance, we illustrate the ethanol concentration determination results using box charts in Figure 4. The red dotted lines denote the true concentration values. Table S4 provides the mean values and standard deviations of predictions for all three types of baijiu.

The small letters (a, b, etc.) annotated in Figure 4 represent the grouping results obtained through a one-way analysis of variance (ANOVA) using the Tukey test on the performance of different models under graph-regularized PCA. These groups indicate statistically significant differences in the ethanol concentration determination results among the various models with different letters. Similarly, the capital letters (A, B, etc.) annotated in the figure represent the grouping results obtained by the Tukey test on the performance of different methods under basic PCA. Therefore, the data points with annotation AB indicate that they do not show statistically significant differences compared with both groups A and B.

The results in Figure 4 show that our proposed RF-G method (integrating graph-regularized PCA and random forest) is significantly better than all other methods in Jinm30 prediction; however, it is not significantly better than XGB-G in Sqyy prediction, nor is it significantly better than SVM-G and KNN-G in Zyz prediction. Nonetheless, none of these compared methods can achieve comparable performance with our proposed RF-G in all three prediction tasks, demonstrating the superiority and robustness of our proposed method.

https://doi.org/10.1021/acsomega.5c00616
ACS Omega 2025, 10, 14373−14381

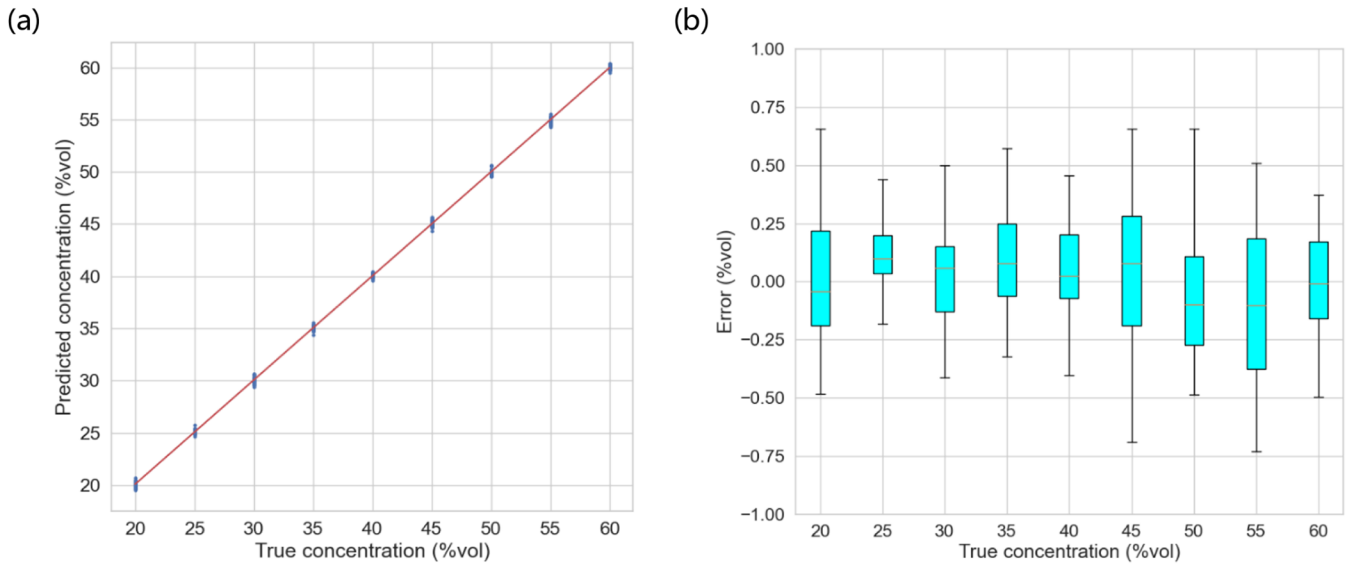


Figure 3. (a) Regression curve of different concentrations of ethanol solutions. (b) Prediction error $e = \hat{y} - y$ on each concentration level.

Table 3. Results of Ethanol Concentration Determination for All Three Types of Baijiu[a]

Method	MAE (vol %)	RMSE (vol %)	MAPE (%)
SVM-G	0.433	0.451	1.177
KNN-G	0.433	0.420	1.185
DT-G	0.468	0.436	1.190
XGB-G	0.278	0.299	0.699
RF-G	**0.178**	**0.210**	**0.415**
SVM-P	0.414	0.470	1.153
KNN-P	0.444	0.531	1.207
DT-P	0.568	0.555	1.439
XGB-P	0.460	0.473	1.192
RF-P	0.454	0.473	0.932

[a]Bold numbers indicate the best results.

Effect of Training Solution Concentration. In the section "Ethanol Concentration Determination," the ethanol concentration predictions for Jinm30 (30.9 vol %), Sqyy (42.0 vol %), and Zyz (53.3 vol %) are based on the same training sample set, which consists of ethanol solutions with concentrations ranging from 20 to 60 vol %. In this section, we attempt to adopt different training sets for ethanol concentration prediction for these three types of baijiu. The training sample sets for the ethanol concentration predictions of Jinm30, Sqyy, and Zyz are each formed using the Raman spectra of ethanol solutions with concentrations ranging from 20 to 40 vol %, 30 to 50 vol %, and 40 to 60 vol %, respectively.

Table S5 reveals that the models based on the standard ethanol solutions with the narrower concentration range lead to slightly more accurate predictions of the ethanol concentrations of Jinm30 and Zyz, while they result in slightly less accurate predictions of the ethanol concentrations of Sqyy. The concentration range of ethanol solutions has a limited impact on the prediction accuracy for the protocol developed in this study. Therefore, building the quantification model on the serial standard ethanol solutions with a wider concentration range would be more convenient for the analysis of various types of baijiu.

Hyperparameters Tuning. Our proposed method integrates graph-regularized PCA and random forest, and the key hyperparameters include the dimension of PCA and the number of trees in the random forest. Hyperparameter tuning experiments are conducted on the cross-validation of ethanol solutions detection (see Table S6), demonstrating that our method performs best when the dimension of PCA $r = 32$ and the number of trees in the random forest is 50.

Model Interpretability. In eq 6, the principal component matrix U and projection matrix W are optimized simultaneously. $W \in \mathbb{R}^{r\times m}$ depict It depicts how the algorithm reduces the dimension of raw spectral data from $m = 700$ to $r = 32$. We define the mean projection vector

$$w = \frac{1}{r}\sum_{i=1}^{r} W_{i:} \tag{11}$$

Since spectra of ethanol solutions with different concentrations share the same characteristic Raman peak at several Raman shifts, the 0−1 normalized intensities of $w \in \mathbb{R}^{1\times m}$ and the mean spectra of a series of aqueous ethanol solutions with various concentrations are visualized and compared in Figure 5. It can be seen that the learning model assigns a higher weight to specific peaks in the spectrum (see Table 1), and this assignment is consistent in both ethanol solutions and baijiu samples. This phenomenon reveals the chemical interpretability of the proposed framework.

It is important to note that ethanol can form extensive hydrogen bonds in aqueous solutions, and the volume reduction due to the formation of ethanol−water binary mixtures leads to a nonlinear correlation between volume concentration and Raman spectroscopy measurements. Furthermore, within the concentration ranges, variations in refractive index can also introduce nonlinearity to Raman spectroscopy measurements. However, as shown in Figure 5, our proposed method can still capture information from specific peaks when analyzing Raman spectra data for baijiu ethanol concentration determination, because it is capable of detecting nonlinear relationships during data analysis.

From the perspective of computer science, graphs and trees are the most typical nonlinear data structures. In fact, manifold learning is nearly equivalent to machine learning on graphs, while ensemble learning (especially random forest learning) is

https://doi.org/10.1021/acsomega.5c00616
ACS Omega 2025, 10, 14373−14381

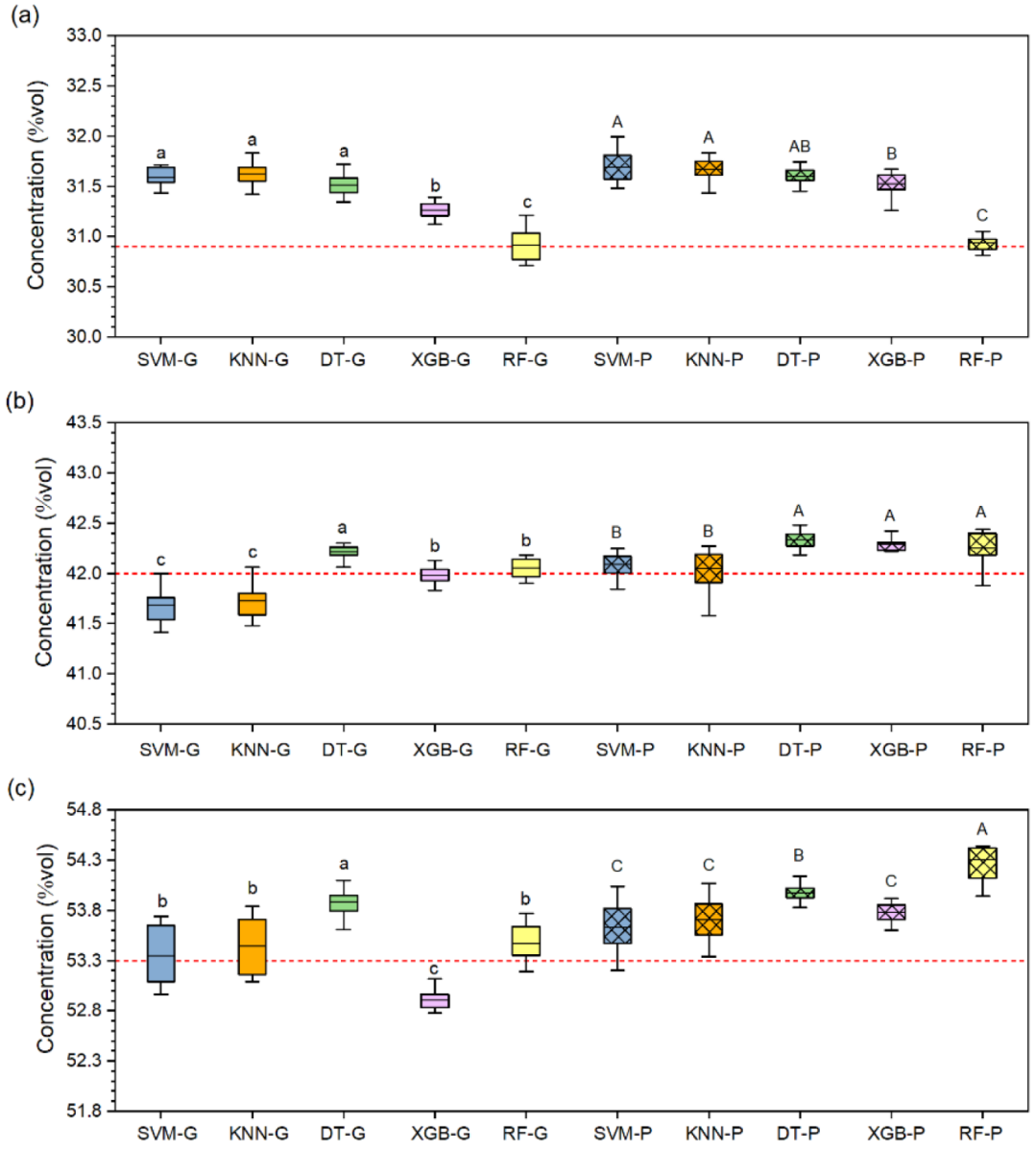


Figure 4. Box charts of baijiu ethanol concentration determination results. (a) Jinm30, (b) Sqyy, and (c) Zyz.

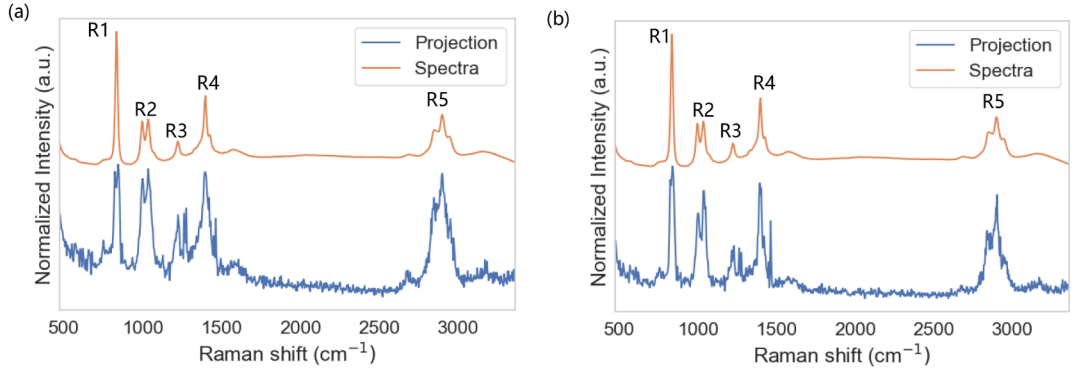


Figure 5. Comparison between mean projection vector and mean spectra of (a) ethanol solutions and (b) baijiu samples.

inspired by the integration of multiple decision trees. Therefore, our proposed ethanol concentration determination method for baijiu, integrating manifold learning and ensemble learning through graph-regularized PCA and random forest-based Raman spectroscopy, can achieve better prediction performance by adopting machine learning on nonlinear data structures to capture nonlinear data relationships while reducing measurement noise.

https://doi.org/10.1021/acsomega.5c00616
ACS Omega 2025, 10, 14373−14381

CONCLUSIONS

Determining ethanol concentration is crucial for improving the quality of baijiu. Machine learning-based Raman spectroscopy is a powerful tool for detecting the Raman spectrum of alcoholic beverages like baijiu. In this article, we propose a method for determining ethanol concentration in baijiu using graph-regularized PCA and random forest-based Raman spectroscopy. We combine graph-regularized PCA and random forest to determine ethanol concentration based on the Raman spectrum. Random forest integrates multiple decision trees and utilizes feature selection, while graph-regularized PCA adds manifold regularization to basic PCA for dimensionality reduction. This combination is highly effective for capturing meaningful information from high-dimensional Raman spectral data. Furthermore, we propose a protocol that adopts ethanol solutions with various concentrations as the training set to develop a quantitative model for detecting a wide range of ethanol concentrations in different baijiu products. The results demonstrate that the joint application of graph-regularized PCA and random forest for Raman spectral data mining contributes to enhanced accuracy and robustness in ethanol concentration determination using machine learning-based Raman spectroscopy. The proposed method is a flexible framework that is time-efficient, cost-effective, noninvasive, and capable of high-throughout determination. Therefore, the proposed method can be extended to machine learning-based Raman spectroscopy applications for other food and biochemical materials.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.5c00616.

Detailed prediction results for Jinm30, Sqyy, and Zyz baijiu (Tables S1−S4), prediction results using a smaller range of ethanol solution concentrations (Table S5), and hyperparameter tuning results (Table S6) (PDF)

AUTHOR INFORMATION

Corresponding Author

Zhuangwei Shi − *XLab, Ji Hua Laboratory, Foshan, Guangdong 528200, China;* orcid.org/0000-0002-8241-7779; Email: shizw@jihualab.ac.cn

Authors

Zhenhao Chen − *Academy for Engineering and Technology, Fudan University, Shanghai 200433, China*

Jianchen Zi − *XLab, Ji Hua Laboratory, Foshan, Guangdong 528200, China*

Chenhui Wang − *XLab, Ji Hua Laboratory, Foshan, Guangdong 528200, China; Personalized Nutrition, Baden-Württemberg Cooperative State University Heilbronn, Heilbronn, Baden-Württemberg 74076, Germany*

Hai Bi − *XLab, Ji Hua Laboratory, Foshan, Guangdong 528200, China;* orcid.org/0000-0002-2017-3668

Yunlong Zhu − *Academy for Engineering and Technology, Fudan University, Shanghai 200433, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.5c00616

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work is supported by the National Key R&D Program of China (No. 2021YFB3600605), the National Natural Science Foundation of China (No. 52173282), and Ji Hua Laboratory (No. X210221TP210).

REFERENCES

(1) Yang, W.; Zou, W.; Shen, C.-H.; Yang, J.-G. Basic flavor types and component characteristics of Chinese traditional liquors: A review. *J. Food Sci.* **2020**, *85*, 4096−4107.

(2) Wang, J.; Chen, H.; Wu, Y.; Zhao, D. Uncover the flavor code of strong-aroma baijiu: Research progress on the revelation of aroma compounds in strong-aroma baijiu by means of modern separation technology and molecular sensory evaluation. *J. Food Compos. Anal.* **2022**, *109*, 104499.

(3) He, Y.; Liu, Z.; Qian, M.; Yu, X.; Xu, Y.; Chen, S. Unraveling the chemosensory characteristics of strong-aroma type Baijiu from different regions using comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry and descriptive sensory analysis. *Food Chem.* **2020**, *331*, 127335.

(4) Yan, Y.; Lu, J.; Nie, Y.; Li, C.; Chen, S.; Xu, Y. Characterization of volatile thiols in Chinese liquor (Baijiu) by ultraperformance liquid chromatography−mass spectrometry and ultraperformance liquid chromatography−quadrupole-time-of-flight mass spectrometry. *Front. Nutr.* **2022**, *9*, 1022600.

(5) Hu, S.; Wang, L. Age Discrimination of Chinese Baijiu Based on Midinfrared Spectroscopy and Chemometrics. *J. Food Qual.* **2021**, *2021*, 5527826.

(6) Wang, Y.; Xu, J.; Cui, D.; Kong, L.; Chen, S.; Xie, W.; Zhang, C. Classification and Identification of Archaea Using Single-Cell Raman Ejection and Artificial Intelligence: Implications for Investigating Uncultivated Microorganisms. *Anal. Chem.* **2021**, *93*, 17012−17019.

(7) Yu, S.; Li, X.; Lu, W.; Li, H.; Fu, Y. V.; Liu, F. Analysis of Raman spectra by using deep learning methods in the identification of marine pathogens. *Anal. Chem.* **2021**, *93*, 11089−11098.

(8) Liu, J.; Osadchy, M.; Ashton, L.; Foster, M.; Solomon, C. J.; Gibson, S. J. Deep convolutional neural networks for Raman spectrum recognition: A unified solution. *Analyst* **2017**, *142*, 4067−4074.

(9) Bi, H.; Palma, C.-A.; Gong, Y.; Hasch, P.; Elbing, M.; Mayor, M.; Reichert, J.; Barth, J. V. Voltage-driven conformational switching with distinct Raman signature in a single-molecule junction. *J. Am. Chem. Soc.* **2018**, *140*, 4835−4840.

(10) Zi, J.; Lobet, M.; Henrard, L.; Li, Z.; Wang, C.; Wu, X.; Bi, H. Effect of near-field optical angular momentum on molecular junctions. *Light Adv. Manuf.* **2023**, *4*, 372−379.

(11) Huang, L.; Fan, X.; He, H.; Yan, L.; He, Z. Single-end hybrid Rayleigh Brillouin and Raman distributed fibre-optic sensing system. *Light Adv. Manuf.* **2023**, *4*, 171−180.

(12) Kiefer, J.; Cromwell, A. L. Analysis of single malt Scotch whisky using Raman spectroscopy. *Anal. Methods* **2017**, *9*, 511−518.

(13) Magdas, D. A.; Guyon, F.; Berghian-Grosan, C.; Molnar, C. M. Challenges and a step forward in honey classification based on Raman spectroscopy. *Food Control* **2021**, *123*, 107769.

(14) Shi, Z.; Su, Y.; Zi, J.; Yang, S.; Li, D.; Luo, Y.; Wang, C.; Bi, H. Intelligent sensory of lard quality by adaptive residual attention networks and Raman spectroscopy. *Microchem. J.* **2025**, *209*, 112680.

(15) Wu, Z.; Long, J.; Xu, E.; Wang, F.; Xu, X.; Jin, Z.; Jiao, A. A feasibility study on the evaluation of quality properties of Chinese rice wine using Raman spectroscopy. *Food Anal. Methods* **2016**, *9*, 1210−1219.

(16) Gu, J.; Liu, H.; Ma, C.; Li, L.; Zhu, C.; Glorieux, C.; Chen, G. Conformal Prediction Based on Raman Spectra for the Classification of Chinese Liquors. *Appl. Spectrosc.* **2019**, *73*, 759−766.

(17) Wang, C.; Shi, Z.; Shen, H.; Fang, Y.; He, S.; Bi, H. Towards robustness and sensitivity of rapid Baijiu (Chinese liquor) discrimination using Raman spectroscopy and chemometrics: Dimension reduction, machine learning, and auxiliary sample. *J. Food Compos. Anal.* **2023**, *118*, 105217.

(18) Wang, M.-L.; Choong, Y.-M.; Su, N.-W.; Lee, M.-S. A rapid method for determination of ethanol in alcoholic beverages using capillary gas chromatography. *J. Food Drug Anal.* **2003**, *11*, 3.
(19) Li, H.; Chai, X.-S.; Deng, Y.; Zhan, H.; Fu, S. Rapid determination of ethanol in fermentation liquor by full evaporation headspace gas chromatography. *J. Chromatogr. A* **2009**, *1216*, 169−172.
(20) Glampedaki, P.; Hatzidimitriou, E.; Paraskevopoulou, A.; Pegiadou-Koemtzopoulou, S. Surface tension of still wines in relation to some of their constituents: A simple determination of ethanol content. *J. Food Compos. Anal.* **2010**, *23*, 373−381.
(21) Fleming, H.; Chen, M.; Bruce, G. D.; Dholakia, K. Through-bottle whisky sensing and classification using Raman spectroscopy in an axicon-based backscattering configuration. *Anal. Methods* **2020**, *12*, 4572−4578.
(22) Liu, W.; Liang, X.; He, S.; Shi, Z.; Cen, B.; Chen, W.; Bi, H.; Wang, C. Rapid and simultaneous quantitative and discriminative analyses of liquor quality parameters with machine learning-assisted batch Raman spectroscopy: Synergistic instrumental upgrade and chemometric optimization. *Food Control* **2024**, *158*, 110242.
(23) Zong, X.; Zhou, X.; Wen, L.; Gan, S.; Li, L. Identification of Baijiu based on the Raman spectroscopy and back-propagation neural network optimized using genetic algorithm. *J. Food Compos. Anal.* **2024**, *126*, 105917.
(24) Belkin, M.; Niyogi, P.; Sindhwani, V. Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples. *J. Mach. Learn. Res.* **2006**, *7*, 2399−2434.
(25) Liu, G.; Lin, Z.; Yan, S.; Sun, J.; Yu, Y.; Ma, Y. Robust recovery of subspace structures by low-rank representation. *IEEE Trans. Patt. Anal. Mach. Intell.* **2013**, *35*, 171−184.
(26) Shi, Z.; Zhang, H.; Jin, C.; Quan, X.; Yin, Y. A representation learning model based on variational inference and graph autoencoder for predicting lncRNA-disease associations. *BMC Bioinform.* **2021**, *22*, 136.
(27) Boyd, S.; Parikh, N.; Chu, E.; Peleato, B.; Eckstein, J. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.* **2010**, *3*, 1−122.
(28) Jin, C.; Shi, Z.; Lin, K.; Zhang, H. Predicting miRNA-Disease Association Based on Neural Inductive Matrix Completion with Graph Autoencoders and Self-Attention Mechanism. *Biomolecules* **2022**, *12*, 64.
(29) Sagi, O.; Rokach, L. Ensemble learning: A survey. *Wiley Interdiscip. Rev.: Data Min. Knowl. Discov.* **2018**, *8*, No. e1249.
(30) Dong, X.; Yu, Z.; Cao, W.; Shi, Y.; Ma, Q. A survey on ensemble learning. *Front. Comput. Sci.* **2020**, *14*, 241−258.
(31) Zhang, Z.-M.; Chen, S.; Liang, Y.-Z.; Liu, Z.-X.; Zhang, Q.-M.; Ding, L.-X.; Ye, F.; Zhou, H. An intelligent background-correction algorithm for highly fluorescent samples in Raman spectroscopy. *J. Raman Spectrosc.* **2010**, *41*, 659−669.
(32) Burikov, S.; Dolenko, T.; Patsaeva, S.; Starokurov, Y.; Yuzhakov, V. Raman and IR spectroscopy research on hydrogen bonding in water-ethanol systems. *Mol. Phys.* **2010**, *108*, 2427−2436.
(33) Zhou, Z.-H. *Machine Learning*; Springer, 2021.
(34) Dennison, P. E.; Halligan, K. Q.; Roberts, D. A. A comparison of error metrics and constraints for multiple endmember spectral mixture analysis and spectral angle mapper. *Remote Sens. Environ.* **2004**, *93*, 359−367.
(35) Lee, D. D.; Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature* **1999**, *401*, 788−791.
(36) Cai, D.; He, X.; Han, J.; Huang, T. Graph Regularized Nonnegative Matrix Factorization for Data Representation. *IEEE Trans. Patt. Anal. Mach. Intell.* **2011**, *33*, 1548−1560.
(37) Yi, H.; Raman, A. T.; Zhang, H.; Allen, G. I.; Liu, Z. Detecting hidden batch factors through data-adaptive adjustment for biological effects. *Bioinformatics* **2018**, *34*, 1141−1147.
(38) Lu, C.; Yang, M.; Li, M.; Li, Y.; Wu, F.-X.; Wang, J. Predicting human lncRNA-disease associations based on geometric matrix completion. *IEEE J. Biomed Health Inform.* **2020**, *24*, 2420−2429.
(39) Wright, J.; Ganesh, A.; Rao, S.; Peng, Y.; Ma, Y.Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization*Advances in Neural Information Processing Systems 22 (NIPS 2009)*NIPS2009222080−2088
(40) Jin, C.; Gao, J.; Shi, Z.; Zhang, H. ATTCry: Attention-based neural network model for protein crystallization prediction. *Neurocomputing* **2021**, *463*, 265−274.
(41) Amjad, A.; Ullah, R.; Khan, S.; Bilal, M.; Khan, A. Raman spectroscopy based analysis of milk using random forest classification. *Vib. Spectrosc.* **2018**, *99*, 124−129.
(42) Li, M.; Xu, Y.; Men, J.; Yan, C.; Tang, H.; Zhang, T.; Li, H. Hybrid variable selection strategy coupled with random forest (RF) for quantitative analysis of methanol in methanol-gasoline via Raman spectroscopy. *Spectrochim. Acta, Part A* **2021**, *251*, 119430.

https://doi.org/10.1021/acsomega.5c00616
ACS Omega 2025, 10, 14373−14381