# Rapid Detection of Alcohol Content in Beer Based on Near-infrared Spectroscopy (NIRS)

**Zhiyu ZHANG** *

College of Information and Electrical Engineering, Heilongjiang Bayi Agricultural University, Daqing 163319, China

**Abstract** [**Objectives**] This study was conducted to establish a new method for rapid determination of alcohol content in beer using near-infrared spectroscopy (NIRS). [**Methods**] Genetic algorithm (GA) was used to select characteristic wavelengths from the original spectra, and 158 key wavelength variables were identified. The Kennard-Stone (KS) algorithm was adoped to divide 216 beer samples into a calibration set and a prediction set at a ratio of 7∶3. Partial least squares regression (PLSR) and support vector regression (SVR) models were subsequently established. [**Results**] The GA-SVR model constructed after GA screening exhibited the best performance, with a coefficient of determination ($R_P^2$) of 0.963, a root mean square error of prediction (RMSEP) of 0.318% vol, and a residual predictive deviation (RPD) of 5.633 for the prediction set. [**Conclusions**] This method enables rapid, non-destructive, and high-precision prediction of alcohol content in beer, providing an effective technical approach for quality control in beer production.

**Key words** Near-infrared spectroscopy; Alcohol content; Genetic algorithm; Non-destructive testing

**DOI**:10.19759/j.cnki.2164-4993.2026.02.009

As a popular alcoholic beverage worldwide, the accurate and rapid determination of its core quality indicator, alcohol content, is of great significance for quality control during production, tax administration, and consumer protection[1]. Traditional methods for alcohol content determination are limited by the need for chemical reagents and sample destruction, making it difficult to meet the urgent demands of the modern brewing industry for rapid, online, and non-destructive screening of large sample batches.

Near-infrared spectroscopy (NIRS) is applied to the detection of food and agricultural products because of its advantages of fast, non-destructive and multi-component simultaneous analysis[2-3]. This technique is based on the overtone and combination absorption of hydrogen-containing groups, reflecting the chemical composition of the sample. However, the spectral data are characterized by high dimensionality, redundancy, and strong collinearity, and noise and irrelevant information in the full spectra increase model complexity. Therefore, effective preprocessing and characteristic wavelength selection are critical steps for improving model performance.

Genetic algorithm (GA), an optimization algorithm that simulates natural selection and genetic mechanisms, searches globally for optimal combinations of characteristic wavelengths through selection, crossover, and mutation operations. It identifies the subset of features most relevant to the target attribute, thereby enhancing model interpretability and predictive capability[4]. In terms of sample set partitioning, the Kennard-Stone (KS) algorithm selects a representative calibration set by maximizing the uniform distribution of samples in the spectral space[5].

For quantitative modeling, partial least squares regression (PLSR) has become the dominant algorithm in NIRS analysis due to its ability to handle high-dimensional collinear data[6]. Meanwhile, support vector regression (SVR), a machine learning method based on the principle of structural risk minimization, offers unique advantages in handling small sample sizes and nonlinear problems. Combining and comparing these two approaches facilitates the development of a more optimized predictive model for alcohol content in beer[7].

In this study, a rapid detection method for alcohol content in beer was established based on NIRS. After spectral acquisition, GA was adopted to select characteristic wavelengths, and the KS algorithm was employed to divide the samples into calibration and prediction sets at a ratio of 7∶3. PLSR and SVR regression models were then constructed. By comparing model performance, this study aimed to establish an efficient and accurate quantitative analysis model and provide technical support for beer quality monitoring and process optimization.

## Materials and Methods

### Preparation of beer samples

The craft beer and industrial beer samples used in this study were all purchased from local market suppliers. A total of 138 craft beer samples were included, covering 23 different styles and involving multiple common brands. For each style, two independent production batches were collected, and three bottles were taken from each batch as parallel samples. A total of 78 industrial beer samples were collected, covering 13 commercially available styles from different brands. Similarly, two production batches were collected for each style of each brand, with three bottles per batch. The interval between the production date and the collection date did not exceed two months for all samples, ensuring freshness and representativeness. The physicochemical index for alcohol content of both craft beer and industrial beer was based on the labeled values on the product packaging. The alcohol content of all

Received: December 22, 2025 Accepted: March 3, 2026

Zhiyu ZHANG (2005-), male, P. R. China, major: computer science and technology.

* Corresponding author.

beer samples ranged from ≥3.1% vol to ≥12% vol. Representative beer samples are shown in Fig. 1.

**Fig. 1 Some beer samples**

### Spectral data acquisition

Spectral acquisition was performed using a NIR-F210 near-infrared miniature fiber optic spectrometer (PYNECT, Shenzhen, China) based on digital light processing (DLP) technology. Before the experiment, the instrument was preheated for 30 min to ensure signal stability. Spectral acquisition was carried out in transmission mode, with air used as the background reference, and the background spectrum was re-scanned once every hour. The spectral scanning range was set to 900 – 1 700 nm, resulting in a total of 228 wavelength variables. Each beer sample was measured only once, with each measurement consisting of three consecutive scans. The average of the three scan results was taken as the raw spectral data for that sample. Ultimately, a total of 216 raw spectra were obtained.

### Feature extraction method

In this study, GA was used to select characteristic wavelengths, and the fitness function was based on the root mean square error of cross-validation (RMSECV) of the PLSR model. The smaller the error, the higher the fitness. Through multiple generations of evolution, the algorithm converges to an optimal individual, whose corresponding wavelength subset minimizes the RMSECV. This effectively eliminates redundant and noisy wavelengths, thereby enhancing the prediction accuracy, robustness, and computational efficiency of the regression model.

### Modeling method and evaluation indicators

In this study, two algorithms, PLSR and SVR, were employed to establish quantitative prediction models between the alcohol content of beer and the near-infrared spectral data. To objectively evaluate and compare the predictive performance of the models, two statistical indicators were used for comprehensive assessment. The first one was the coefficient of determination ($R^2$), which measures the models' ability to explain data variation. The closer its value is to 1, the stronger the models' fitting or predictive capability. The second was the root mean square error (RMSE), which reflects the deviation between the predicted and actual values. The smaller its value, the higher the models' prediction accuracy. Additionally, the residual predictive deviation (RPD), defined as the ratio of the standard deviation of the samples to the RMSE, was introduced to further evaluate the models' stability and practicality. Generally, an RPD > 3.0 indicates that the model has excellent predictive capability[8].

## Results and Analysis

### Data snalysis

In this study, Savitzky-Golay (SG) smoothing, multivariate scattering correction (MSC), standard normal variate (SNV), and first derivative (FD) were used for spectral preprocessing of beer samples. Based on the 10-fold cross-validation coefficient of determination ($R^2_{cv}$), root mean square error of cross-validation (RMSECV), and residual predictive deviation of cross-validation (RPDCV) of the PLSR model, the optimal preprocessing method for different models was determined. After calculation and comparison, SNV was determined as the optimal preprocessing method for predicting alcohol content in beer. The results of different spectral preprocessing methods are shown in Table 1.

**Table 1 The results of different spectral preprocessing methods**

| Preprocessing method | $R^2_{cv}$ | RMSECV | RPDCV | LVs |
|---|---|---|---|---|
| N/A | 0.911 | 0.509 | 3.356 | 8 |
| SG | 0.932 | 0.443 | 3.860 | 8 |
| MSC | 0.929 | 0.451 | 3.799 | 8 |
| SNV | 0.934 | 0.438 | 3.902 | 8 |
| DF | 0.933 | 0.440 | 3.887 | 5 |

$R^2_{cv}$, RMSECV, and RPDCV represent the cross-validated coefficient of determination, root mean square error, and residual predictive deviation, respectively.

When the number of samples is limited, an effective sample set partitioning method can be used to select highly representative samples for constructing the calibration set and establishing the spectral regression quantitative model[9]. For regression modeling sample set division, the dataset was partitioned at a ratio of 7 ∶ 3 using the KS algorithm into a calibration set containing 151 samples and a validation set containing 65 samples.

### Spectral feature extraction

To minimize the impact of randomness inherent in the GA algorithm on the final modeling results, ten rounds of the GA algorithm were performed using the full-spectrum data of the calibration set to construct an initial set of characteristic wavelengths for predicting alcohol content in beer samples. The higher the selection frequency of a wavelength variable in the initial set, the stronger its correlation with the detection target. When the number of repeated selections reached three, the RMSECV of the PLSR model built was 0.433, and the number of selected wavelength variables was 158. Among these, the wavelength variable with the highest selection frequency was 1 392.22 nm, corresponding to the second overtone of $CH_3$ and the second overtone of $CH_2$.

### Model construction and evaluation

PLSR and SVR models were constructed using the GA characteristic wavelength selection method, and their performance in predicting alcohol content was evaluated using the coefficient of

*determination* $R^2$, RMSE, and RPD for both the calibration and validation sets. As shown in Table 2, the $R^2$ values of the regression models for alcohol content in beer samples established using GA-selected characteristic wavelengths were all greater than 0.9, the RMSE values were all less than 0.4% vol, and the RPD values were all greater than 3. These results demonstrate successful modeling and superior performance compared with models built using the full spectra. The 158 wavelength variables selected by GA accounted for 69.29% of the full spectral wavelength variables, and 70 weakly correlated redundant wavelength variables were eliminated. The GA characteristic wavelength selection significantly improved the regression performance of the calibration model for alcohol content in beer samples, and its modeling performance was clearly superior to that of models built using the full spectra.

**Table 2 Evaluation indicators of different regressive models**

| Model | Wavelengths | $R_c^2$ | $R_p^2$ | RMSEC//% vol | RMSEP//% vol | RPD |
|---|---|---|---|---|---|---|
| Full-PLSR | 228 | 0.978 | 0.917 | 0.276 | 0.398 | 3.496 |
| GA-PLSR | 158 | 0.984 | 0.936 | 0.255 | 0.384 | 3.606 |
| Full-SVR | 228 | 0.951 | 0.945 | 0.358 | 0.366 | 4.312 |
| GA-SVR | 158 | 0.979 | 0.963 | 0.274 | 0.318 | 5.633 |

$R_c^2$ and $R_p^2$ represent the coefficients of determination for the calibration set and the prediction set, respectively, while RMSEC and RMSEP denote the root mean square error for the calibration set and the prediction set, respectively.

The optimal regression model for predicting alcohol content in beer was the SVR model constructed after feature extraction using GA, with $R^2$ values of 0.979 and 0.963 for the calibration and validation sets, respectively. The corresponding RMSE values were 0.274 and 0.318, and the RPD was 5.633. The performance of this model was significantly superior to other regression models. The favorable prediction accuracy for alcohol content can be attributed to the strong and specific absorption bands of ethanol ($-CH_3$, $-CH_2$, and -OH groups) in the near-infrared region, as well as the minimal spectral overlap with other beer components in the 900 – 1 700 nm range. This enables GA to readily extract ethanol-related features, and the resulting regression model can meet the rapid detection requirements for alcohol content in beer.

## Conclusion and Discussion

This study successfully applied NIRS combined with chemometric methods to achieve rapid quantitative detection of alcohol content in beer. Characteristic wavelength selection using GA effectively removed redundant spectral information, and the 158 selected wavelength variables significantly improved model performance. Comparing the PLSR and SVR modeling approaches, the GA-SVR model constructed after GA screening exhibited the best predictive performance ($R_p^2$ = 0.963, $R^2$ = 0.963, RMSEP = 0.318% vol, RPD = 5.633). This method fully leverages the advantages of NIRS for rapid and non-destructive analysis. The established model demonstrates high accuracy and strong robustness, providing a reliable technical solution for online quality control and rapid screening of alcohol content in the beer industry.

## *References*

[1] WEI Y, LIU J, XI G, *et al.* Quantitative detection of key parameters and authenticity verification for beer using near-infrared spectroscopy[J]. Foods, 2025, 14(22): 3936.

[2] XUE Z, BAI J, LUO Y, *et al.* Rapid quality evaluation of Saposhnikoviae Radix powders by NIRS combined with multivariate intelligent algorithms[J]. Microchemical Journal, 2025, 215: 114559.

[3] SHARMA S, GOYAL P, DEVI J, *et al.* Using near-infrared reflectance spectroscopy (NIRS) to predict the nitrogen levels in the stem and root tissues of *Brassica juncea* (Indian mustard)[J]. Spectrochimica acta Part A, Molecular and biomolecular spectroscopy, 2024, 322: 124755.

[4] YANG Y, WANG S, ZHANG G, *et al.* GA-OMTL: Genetic algorithm optimization for multi-task neural architecture search in NIR spectroscopy [J]. Expert Systems with Applications, 2025, 290: 128517.

[5] REN J, XIONG Y, CHEN X, *et al.* Comparative analysis of machine learning and deep learning algorithms for assessing agricultural product quality using NIRS[J]. Sensors (Basel, Switzerland), 2024, 24(16).

[6] ZENG S, ZHANG Z, CHENG X, *et al.* Prediction of soluble solids content using near-infrared spectra and optical properties of intact apple and pulp applying PLSR and CNN[J]. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2024, 304: 123402.

[7] ZHONG K, LI Y, HUAN W, *et al.* A novel near infrared spectroscopy analytical strategy for soil nutrients detection based on the DBO-SVR method[J]. Spectrochimica acta Part A, Molecular and biomolecular spectroscopy, 2024, 315: 124259.

[8] YE T, ZHENG Y, GUAN Y, *et al.* Rapid determination of chemical components and antioxidant activity of the fruit of *Crataegus pinnatifida* Bunge by NIRS and chemometrics[J]. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2023, 289: 122215.

[9] CHEN M, SONG J, HE H, *et al.* Quantitative analysis of high-price rice adulteration based on near-infrared spectroscopy combined with chemometrics[J]. Foods, 2024, 13(20): 3241.

Editor: Yingzhi GUANG Proofreader: Xinxiu ZHU