Support vector regression-based study of interference in absorption spectral lines of mixed gases*

YAN Xiangyu[1,2,3], LI Honglian[1,2,3], WANG Yitong[1,2,3], FANG Lide[1,2,3], and ZHANG Rongxiang[4]****

1. School of Quality and Technical Supervision, Hebei University, Baoding 071002, China

2. National & Local Joint Engineering Research Center of Metrology Instrument and System, Baoding 071002, China

3. Hebei Key Laboratory of Energy Metering and Safety Testing Technology, Hebei University, Baoding 071002, China

4. College of Physics Science and Technology, Hebei University, Baoding 071002, China

(Received 13 April 2022; Revised 29 August 2022)

When measuring the concentration of multi-component gas mixtures based on supercontinuum laser absorption spectroscopy (SCLAS), there are interferences between the absorption spectral lines. For the spectral interference problem of CO_2 and CH_4 at 1 432 nm, a method based on support vector regression (SVR) is proposed in this paper. The SVR model, the k-nearest neighbor (KNN) model and the least squares (LS) model are used to analyze and predict the absorption spectral data, and the prediction accuracies were 96.29%, 88.89% and 85.19%, respectively, with the highest prediction accuracy of the SVR model. The results show that the method can accurately measure the concentration of gas mixtures, realize the detection of mixed gases using a single waveband, and provide a solution to the overlapping spectral line interference of multi-component gas mixtures.

Document code: A **Article ID:** 1673-1905(2022)12-0743-6

DOI https://doi.org/10.1007/s11801-022-2060-3

The supercontinuum laser absorption spectroscopy (SCLAS) is a new type of absorption spectroscopy detection technology with a broad spectrum, easy collimation, high sensitivity, fast response time, and high stability, which can achieve simultaneous measurement of multiple gases. In recent years, with the development of SCLAS technology, it has been gradually applied to optical imaging[1], biomedicine[2,3], gas monitoring[4], and many other fields.

When using the SCLAS technique to detect mixed gas concentrations, the problem of spectral line interference is often encountered. The common solution is to select absorption peak spectral lines at different locations for the measurement of the gas mixture, so that the problem of spectral interference can be avoided. A variety of studies have been conducted to measure multiple components in gas mixtures. MARTIN et al[5] used a supercontinuous spectral detection system to simultaneously detect water vapor at 1 350—1 420 nm and C_2H_2 at 1 510—1 540 nm, and the experimental results were in good agreement with the simulated spectra obtained based on the Hitran database. DONG et al[6] have developed a dual-gas sensor system for CO and CO_2 detection using a single broadband light source, pyroelectric detector and time division multiplexing (TDM) technology. ADAMU et al[7] developed a supercontinuous spectrum laser-based detection system for a wide range of industrial toxic gases from 1 480 nm to 1 700 nm, and the results showed that the system had good responsiveness and selectivity. In addition, the researchers have also worked on experimental data processing methods. By combining genetic algorithm (GA) and a radial basis function neural network (RBFNN), SONG et al[8] proposed an accurate analytical method for the quantitative analysis of multi-component mud logging gases. LI et al[9] proposed a multi-band fusion model to detect multiple concentrations of CO_2 in three wavelength ranges, and improved the performance of SCLAS to detect CO_2 concentration. JAVED et al[10] used machine learning to decode the composition of the unknown gas mixture from the output of the electrochemical sensor array, and accurately predicted the concentration of each gas. It can be seen that many studies that detect gas mixtures use multiple bands. Although this improves the reliability of the detection system, it also increases the complexity of the detection system. In addition, in some cases, the above methods still inevitably encounter the problem of spectral line interference. For example, the absorption spectra of CO_2 and CH_4 have serious overlapping interference, especially when the concentration of CO_2 is

* This work has been supported by the National Natural Science Foundation of China (No.62173122), the Key Projects of Hebei Natural Science Foundation (No.E2021201031), and the Funding Project for Introducing Overseas Students in Hebei Province (No.C20210312).

** E-mails: lihonglian@hbu.edu.cn; zrx@hbu.edu.cn

high, the absorption signal of CH_4 will be annihilated below the absorption spectrum of CO_2, and it is difficult to accurately measure the concentration of CH_4.

In order to improve the reliability of the system and solve the problem of spectral interference in the process of mixed gas measurement, a method based on support vector regression (SVR) was proposed in this paper. In the field of gas detection, SVR has been extensively used. HUANG et al[11] proposed a transformer fault prediction model based on time series and support vector machine for accurate prediction of dissolved gas in oil. MOHAND et al[12] used a temporal support vector machine approach to detect and identify CO, O_3 and NO_2 in gas mixtures. In this paper, an SCLAS detection system was designed and built to detect the absorption spectra of CO_2, CH_4 standard gases and gas mixtures in the 1 420—1 450 nm band. For the problem of spectral interference of CH_4 and CO_2 at 1 432 nm, the SVR model, k-nearest neighbor (KNN) model and the least squares (LS) model were used in this study to analyze and predict the spectral data, and the SVR model had the highest prediction accuracy. The results showed that the SVR model effectively solves the problem, and provides a solution to the spectral line interference problem of gas mixtures.

According to the Lambert-Beer law, when monochromatic light of a specific wavelength passes through a gas medium during spectral absorption, the intensity of the outgoing light is reduced by the absorption of the medium. When a monochromatic laser with a central frequency passes through a gas to be measured, the absorbance is proportional to the concentration and absorption thickness of the gas to be measured[13]. The change in light intensity can be expressed as

$$I_t = I_0 \exp(-\alpha_v) = I_0 \exp[-PLCS(T)g(v-v_0)], \quad (1)$$

where I_0 is the initial light intensity, I_t is the outgoing light intensity, α_v is the absorption coefficient of the gas, P (atm) is the pressure, L (cm) is the light range, C (%) is the concentration of the gas to be measured, $S(T)$ (cm^{-2}·atm^{-1}) is the spectral intensity of the gas at a temperature of T, and $g(v-v_0)$ is a line function.

Supercontinuum (SC) lasers are generated when a high-powered ultrashort pulse laser is coupled through a coupling lens and the non-linear effects occur through a non-linear fiber, causing other frequency components to appear in the spectrum, thus increasing the spectral width of the spectrum[14,15].

A linear relationship between gas concentration and spectral data is known and a regression model is constructed as follows

$$f(\boldsymbol{x}) = \boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x} + \boldsymbol{b}. \quad (2)$$

Introducing the loss function ε, the expression is as follows

$$g(\varepsilon) = \begin{cases} 0, & |\boldsymbol{y}_i - \boldsymbol{\omega}\cdot\boldsymbol{x}_i - \boldsymbol{b}| \le \varepsilon \\ |\boldsymbol{y}_i - \boldsymbol{\omega}\cdot\boldsymbol{x}_i - \boldsymbol{b}| - \varepsilon, & |\boldsymbol{y}_i - \boldsymbol{\omega}\cdot\boldsymbol{x}_i - \boldsymbol{b}| > \varepsilon \end{cases}. \quad (3)$$

After the loss function is substituted, the regression model can be expressed by

$$\min_{\boldsymbol{\omega},\boldsymbol{b}} \frac{1}{2}\|\boldsymbol{\omega}\|^2, \ \text{s.t.} \ |\boldsymbol{y}_i - \boldsymbol{\omega}\cdot\boldsymbol{x}_i - \boldsymbol{b}| \le \varepsilon. \quad (4)$$

Based on Eqs.(3) and (4), the SVR problem is defined as

$$\min_{\boldsymbol{\omega},\boldsymbol{b}} \frac{1}{2}\|\boldsymbol{\omega}\|^2 + C\sum_{i=1}^{n} g(\varepsilon)(f(x_i) - y_i)(i = 1, 2\ldots, n), \quad (5)$$

where C is the penalty factor, and a higher C means a higher penalty on the outlier data.

After the slack variables are added, the regression model can be expressed by

$$\min_{\boldsymbol{\omega},\boldsymbol{b}} \frac{1}{2}\|\boldsymbol{\omega}\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*),$$
$$\text{s.t.} \ \begin{cases} \varepsilon + \xi_i \le y_i - \omega\cdot x_i - b \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \qquad (i = 1, 2, \ldots, n) \end{cases}. \quad (6)$$

When applying the Lagrange multiplier method for Eq.(6), the Lagrange multipliers $\alpha_i \ge 0$, $\acute{\alpha} \ge 0$, $\acute{\alpha}_j \ge 0$ are introduced to obtain the "pairwise problem" for the objective optimization function as follows

$$\max_{\alpha,\acute{\alpha}} \sum_{i=1}^{n} y_i(\acute{\alpha}_i - \alpha_i) - \varepsilon(\acute{\alpha}_i + \alpha_i) -$$
$$\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\acute{\alpha}_i - \alpha_i)(\acute{\alpha}_j - \alpha_j)\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j,$$
$$\text{s.t.} \ \sum_{i=1}^{m}(\acute{\alpha}_i - \alpha_i) = 0, \quad 0 \le \alpha_i, \acute{\alpha}_i \le C. \quad (7)$$

By solving the optimal value of the dual problem, the final linear regression model can be obtained as follows

$$f(\boldsymbol{x}) = \sum_{i=1}^{n}(\acute{\alpha}_i - \alpha_i)\boldsymbol{x}_i \cdot \boldsymbol{x} + \boldsymbol{b}. \quad (8)$$

Considering the nonlinear mapping $\phi(\boldsymbol{x})$ and the kernel function $K(\boldsymbol{x}, \boldsymbol{x}_i)$, the dyadic form of the nonlinear SVR is obtained as follows

$$f(\boldsymbol{x}) = \sum_{i=1}^{n}(\acute{\alpha}_i - \alpha_i)K(\boldsymbol{x}, \boldsymbol{x}_i) + \boldsymbol{b}, \quad (9)$$

where $K(\boldsymbol{x}, \boldsymbol{x}_i) = \phi(\boldsymbol{x})\cdot\phi(\boldsymbol{x}_i)$ is the kernel function.

The commonly used kernel functions include polynomial kernel functions, Gaussian kernel functions, linear kernel functions, etc. The experiments in this study prove that linear kernel functions are more suitable for SVR models of mixed gas data.

When selecting the spectral lines of gas molecules, the line intensities of CO_2 and CH_4 need to be considered. And in the same absorption spectral range, there is spectral line interference phenomenon for two gases. The absorption spectra of CO_2 were found to be concentrated in the range 6 000—7 000 cm^{-1} with a line intensity of 10^{-23}. The absorption intensity of CO_2 reaches its maximum at 6 983 cm^{-1} (wavelength 1 432 nm), and there is also strong absorption of CH_4 at this point, as shown in Fig.1. So finally the 1 420—1 450 nm band is selected for the measurement of the measured gas.

The SCLAS detection system consists of three main parts, namely the light source emission part, the gas chamber part, the signal reception and processing part[16]. The light source part of the system includes the super continuous spectrum laser, the laser line tunable filter (LLTF), and the diaphragm. The gas chamber part contains the gas

cylinder, the dynamic dilution calibrator, and the white type long range absorption cell. The signal receiving and processing part consists of a photodetector, a data acquisition card, and a computer. The schematic diagram of the experimental setup is shown in Fig.2. The laser used in the experiment is a supercontinuous picosecond pulsed laser SC400-4, which has an output range of 400—2 400 nm. The output wavelength range of the LLTF filter is 1 000—2 300 nm with a wavelength tuning resolution of 0.1 nm. The photodetector is a PDA50B germanium tube detector, which can detect the spectral range of 800—1 800 nm with a response time of 50 ns and a gain range of 0—70 dB. The flow measurement accuracy of the dynamic dilution calibrator is ±1.0%, which can match the concentration of the measured gas with high accuracy.

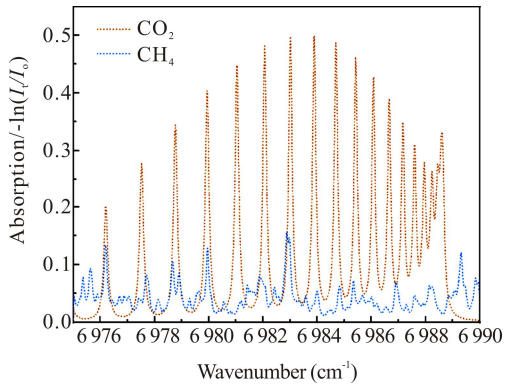

Fig.1 Simulated absorption spectra of 3% CO_2 and CH_4

The experiment was carried out at 296 K and 1 atm. The absorption cell light path was 26.4 m. High purity N_2 was used as a dilution gas and background gas. During the experiment, the concentration of the measured gas was controlled using a dynamic dilution calibrator, and the measured gas passed the absorption cell. The laser was generated by an SC laser and filtered by LLTF with an output band of 1 420—1 450 nm. After the diaphragm filtered out the stray light of the laser, the laser passed the absorption cell. The laser was absorbed by the gas and reflected to the photodetector. The photodetector converted the light signal to an electrical signal, and then the electrical signal was transmitted to the data acquisition card. Data acquisition card transferred data to PC. Before starting a new experiment, N_2 was blown into the absorption cell to ensure that the residual gas is completely exhausted. CH_4, CO_2 and a mixture of two gases at concentrations of 1%, 2% and 3% were tested in turn. Each gas concentration was tested 10 times, and finally 90 sets of signal intensity I_t were obtained, as well as high purity N_2 background signal intensity I_0. The 90 data sets were divided into 9 major groups, as shown in Tab.1.

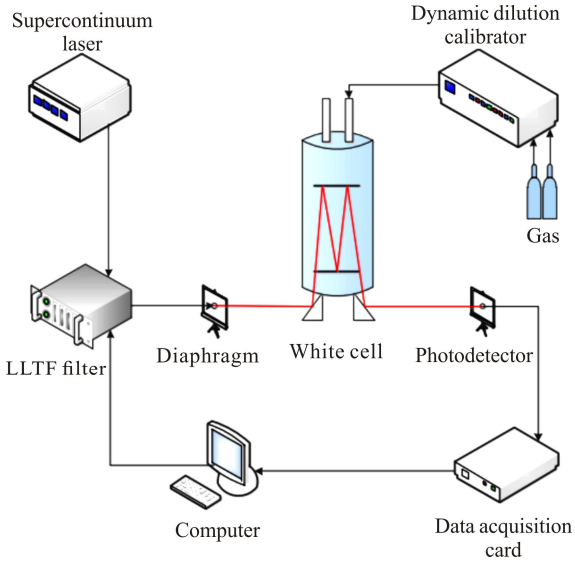

Fig.2 Schematic diagram of the experimental system

Tab.1 Types and concentrations of gases included in the data set

Group	Standard gas category and concentration					
	CH_4			CO_2		
	1%	2%	3%	1%	2%	3%
1	√					
2		√				
3			√			
4				√		
5					√	
6						√
7	√			√		
8		√			√	
9			√			√

The data of group 3, group 6 and group 9 are summed and averaged, and then the absorption signal plots are made. As shown in Fig.3, the absorption signal of CO_2 gas and the absorption signal of mixed gas are similar at 1 432 nm. It is difficult to achieve accurate detection of CH_4 gas in the large signal background of CO_2 gas. Therefore, in this paper, SVR model, KNN model and LS model are used to solve the spectral line interference problem.

The analysis is carried out on the Matlab R2020a platform and adopts the Libsvm software package. In the SVR model, the choice of penalty factor c is very critical, which

indicates the importance of outlier data. The larger the value of c, the more importance is attached to the outlier data, making it more difficult to discard (generalization ability). After the optimization of the parameters by Grid SearchCV, the value of c is 0.031 25. Linear kernel function is used to construct the SVR model.

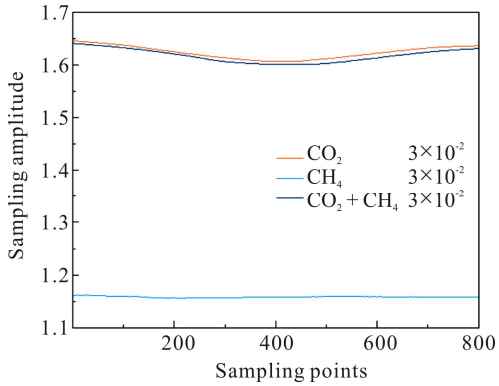

Fig.3 Absorption signal diagrams for three groups of gases

To obtain better prediction results, the data set is processed. A new data set is obtained by adjusting the data in Tab.1 to the same baseline and keeping five hundred data near 1 432 nm. The new data set is normalized to make the regularity of the data set more obvious. The data set is divided into training and prediction sets according to the principle of 7: 3. After modeling on the Matlab platform, setting model parameters, reading data from the training set, and training the model, an SVR model is obtained. Prediction uses prediction set data, and the prediction result of the model on the training set is obtained. The accuracy of the prediction result is 96.29% and the correlation coefficient between the predicted and true values is 0.979 1. The results indicate that the model has good interference immunity and can accurately predict the mixed gas concentration information. A comparison between the actual concentration values of the test set samples and the concentration values predicted by the model is shown in Fig.4.

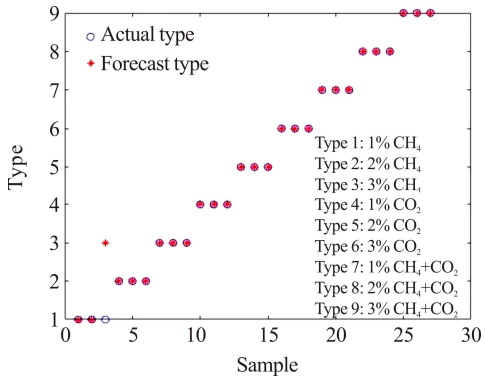

Fig.4 Prediction results of SVR model

KNN algorithm is a simple and commonly used machine learning algorithm. It infers the class to which the data belonged based on the neighbors of a certain data, and classifies the data by measuring the distance between different feature values. It is based on the distance metric to calculate the distance between the samples to be classified and all the samples in the training set, and to find the k samples with the smallest distance from the samples to be classified as the k nearest neighbor samples. Finally, the classification category of the sample to be classified is determined based on the voting of these k nearest neighbor samples, and the predicted sample is classified as the category with the largest number of the k nearest neighbor samples.

The choice of k value has a significant impact on the classification results of KNN algorithm. If the value of k is too small, the phenomenon of overfitting will easily occur, resulting in large prediction errors. If the value of k is too large, the phenomenon of underfitting will occur. In practical applications, the cross-validation method is usually used to select a suitable value of k. The KNN algorithm uses distance to measure the similarity between two samples, and the common distance representation methods are Euclidean distance, Manhattan distance, etc. After a five-fold cross-validation, the value of k taken in this experiment is 3 and the distance representation method taken is Euclidean distance. The model is modeled on the Matlab R2020a platform and the model parameters are set. Modeling is based on the training set above and the prediction set is predicted using the obtained model. The prediction accuracy is 88.89%. Fig.5 shows the graph of the prediction results of the model.

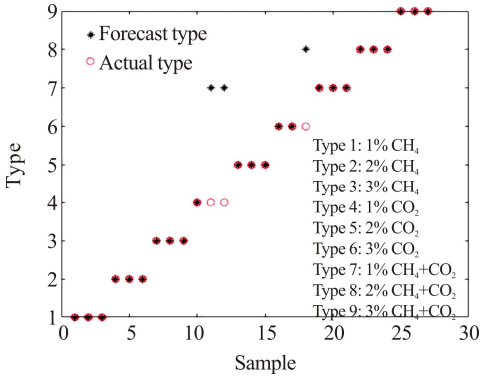

Fig.5 Prediction results of KNN model

Due to the different values of absorption intensity of the gas at different concentrations, the intensity of the gas absorption peak signal is proportional to the concentration of the gas to be measured, so the gas concentration can be inverted based on the signal intensity at different concentration peaks. The LS method is a commonly used fitting algorithm, which uses the least square sum of absolute errors as the evaluation criterion to find the best function match for the data. The basic principle is as follows. For a set of data x_i, y_i (i=1, 2,.... , n), try to find a best-fit curve such that the sum of squares of the difference between the values of the points on this fit

curve and the data values is the smallest among all the fitted curves. In this section, the LS method is used to perform a linear fit to the concentrations to obtain the concentration inversion model.

Each set of data in Tab.1 is divided into training and prediction sets according to the ratio of 7: 3. The training sets are summed and averaged to obtain the new training set. Background deduction is performed on the new training set. The concentrations of the three gases are least squares linearly fitted with their absorbance peaks, and the fitted curves are shown in Fig.6. From Fig.6, it can be seen that the absorbance peak has a good linear relationship with the gas concentrations. The fitting function and the fitting coefficient (R^2) are shown in Tab.2.

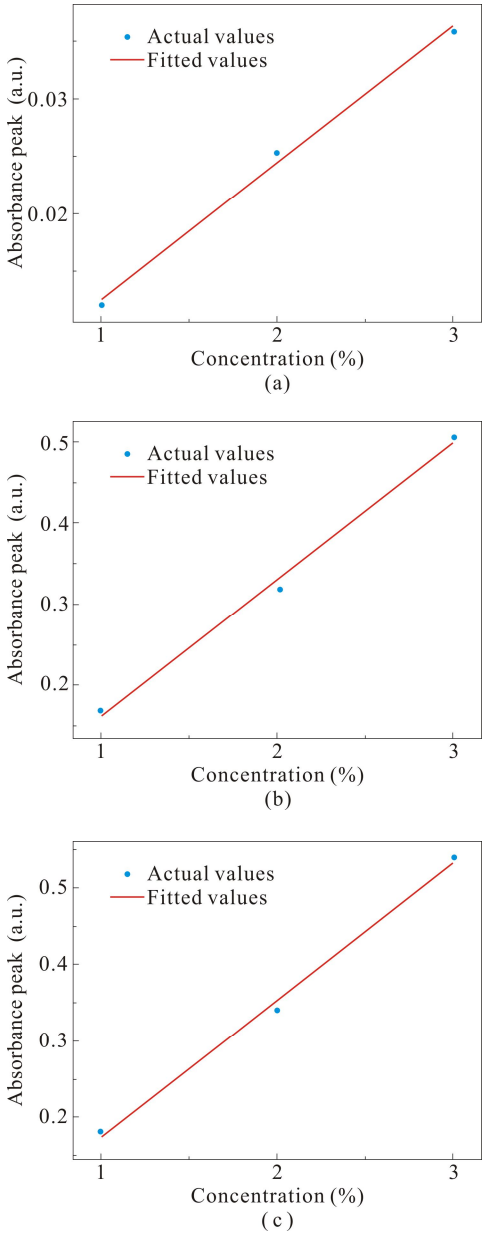

Fig.6 Linear fitting graphs between absorbance peak and concentration of (a) CH_4, (b) CO_2 and (c) the mixed gas

Tab.2 Model performance evaluation

Model	R^2
Y_1=0.011 8X+0.000 7	0.991 3
Y_2=0.168 6X−0.006 9	0.992 5
Y_3=0.179 8X−0.005 8	0.992 8

The concentration inversion is performed for the peak absorption of the three gases in the prediction set. The concentrations of the three gases are predicted according to the fitting function in Tab.2, and the final prediction accuracy obtained is 85.19%.

By comparing the prediction results of these three data processing methods, it can be seen that the SVR model has the highest prediction accuracy. It proves that the SVR model has the advantages of automatic extraction of absorption spectral line features, high prediction accuracy and strong anti-interference.

In this paper, an SVR-based method was proposed to solve the spectral interference problem of CO_2 and CH_4 at 1 432 nm, and realized the simultaneous measurement of CO_2 and CH_4 mixed gas concentration using a single-band laser. In addition to the SVR model, the KNN model and the LS model were also used to predict the gas concentrations. The prediction accuracies of the three models were 96.29%, 88.89% and 85.19%, with the highest prediction accuracy of the SVR model. This study effectively solves the spectral interference problem of CO_2 and CH_4 at 1 432 nm, and provides a reference for the measurement of multi-component gas mixtures using single-band laser, which can be further applied to the measurement of the concentration of other gas mixtures.

Statements and Declarations

The authors declare that there are no conflicts of interest related to this article.

References

[1] TORABZADEH M, STOCKTON P A, TROMBERG B J, et al. Hyperspectral characterization of tissue simulating phantoms using a supercontinuum laser in a spatial frequency domain imaging instrument[J]. Design and quality for biomedical technologies XI, 2018, 10486.

[2] BORONDICS F, JOSSENT M, SANDT C, et al. Supercontinuum-based Fourier transform infrared spectromicroscopy[J]. Optica, 2018, 5(4)：378-381.

[3] JIN W L, CAO N L, ZHU M D, et al. Non-destructive classification of rice seed vigor based on near-infrared ultra-continuous laser spectroscopy[J]. Chinese optics, 2020, 13(05)：1032-1043. (in Chinese)

[4] LI H L, JIA Y Q, DI S, et al. Comparative study of CO_2 measurements with tunable and ultra-continuous spectrum lasers[J]. Laser journal, 2020, 41(05)：23-27. (in Chinese)

[5] MARTIN V J A, GALLEGOS A E, SIERRA H J M, et al. All single-mode-fiber supercontinuum source setup for monitoring of multiple gases applications[J]. Sensors, 2020, 20(11)：3239.

[6] DONG M, ZHONG G Q, MIAO S Z, et al. CO and CO_2 dual-gas detection based on mid-infrared wideband absorption spectroscopy[J]. Optoelectronics letters, 2018, 14(02)：119-123.

[7] ADAMU A I, DASA M K, BANG O, et al. Multispecies continuous gas detection with supercontinuum laser at telecommunication wavelength[J]. IEEE sensors journal, 2020, 20(18)：10591-10597.

[8] SONG L M, GUO S Q, YANG Y G, et al. Quantitative analysis of multicomponent mud logging gas based on infrared spectra[J]. Optoelectronics letters, 2019, 15(04)：312-316.

[9] LI H L, DI S, LV W J, et al. Research on the measurement of CO_2 concentration based on multi-band fusion model[J]. Applied physics B, 2021, 127(1)：1-7.

[10] JAVED U, RAMAIYAN K P, KRELLER C R, et al. Quantification of gas concentrations in $NO/NO_2/C_3H_8/NH_3$ mixtures using machine learning[J]. Sensors and actuators：B chemical, 2022, 359.

[11] HUANG X B, JIANG W T, ZHU Y C, et al. Transformer fault prediction based on time series and support vector machines[J]. High voltage engineering, 2020, 46(07)：2530-2538. (in Chinese)

[12] MOHAND A D, OUSSAMA D, NICOLAS M, et al. A temporal-based SVM approach for the detection and identification of pollutant gases in a gas mixture[J]. Applied intelligence, 2022, 52：6065-6078.

[13] PAYNTER R W. Modification of the Beer-Lambert equation for application to concentration gradients[J]. Surface & interface analysis, 2010, 3(4)：186-187.

[14] ZHANG S Q, CAI L, ZHANG Z. Ultra-continuous spectrum laser beam quality characteristics[J]. Infrared and laser engineering, 2014, 43(05)：1428-1432. (in Chinese)

[15] ZHANG H X, ZHENG L, HUANG C. A wavelength demodulation method for fibre Bragg grating sensors[J]. Journal of Tianjin University, 2012, 45(02)：111-115. (in Chinese)

[16] LI H L, DI S, LI W D, et al. Measurement of CO_2 concentration based on supercontinuum laser absorption spectroscopy[J]. Optoelectronics letters, 2021, 17(03)：176-182.