DOI: 10.13382/j.jemi.B2407781

# 基于波簇区间的挥发性有机气体红外光谱光谱波长选择算法*

严　玥[1,2]　许世豪[1]　何海星月[1]　周　雪[1]
(1. 重庆工商大学人工智能学院　重庆　400067;2. 重庆工商大学检测控制集成系统重庆市市级工程实验室　重庆　400067)

**摘　要**:以特征波长点簇类和吸收峰区间筛选串联选择模式,提出了一种基于波簇区间的波长选择算法用于挥发性有机气体红外光谱波长选择。首先进行簇类聚集,在保留足够特征吸收峰特性同时避免算法波长区间机械划分或随机不确定性,接着设计改进移动窗口方式对同一簇类中的波长点进行再次筛选,保留最能代表光谱特征的波长区间用于后期各种模型预测。用苯乙烯、对二甲苯和邻二甲苯近红外光谱数据在偏最小二乘法、偏最小二乘、岭回归、支持向量机 4 种模型上进行了验证分析,结果表明在不影响模型精度前提下,数据集可缩小至原来的 43.71%~36.35%;以 3 种气体各 2 种浓度全排列组合混合气体为数据集,通过 3 种不同结构卷积神经网络(CNN)模型上光谱波形选择前后实验对比,在保证预测精度的同时验证了算法在降低机器学习模型复杂度上的有效性,波长选择前后在 3 种 CNN 预测模型上运行效率提升 90%。

**关键词**: 红外光谱;波长选择算法;波簇区;神经网络;预测精度
**中图分类号**: TP212.2;TN911.72　　**文献标识码**: A　　**国家标准学科分类代码**: 510.40

# Wavelength selection algorithm for infrared spectra of volatile organic gases based on wave-cluster interval

Yan Yue[1,2]　Xu Shihao[1]　He Haixingyue[1]　Zhou Xue[1]
(1. School of Artificial Intelligence, Chongqing Technology and Business University, Chongqing 400067, China; 2. Chongqing Key Laboratory of Intelligent Perception and Block Chain Technology, Chongqing Technology and Business University, Chongqing 400067, China)

**Abstract**: A novel wavelength selection algorithm, based on wave cluster interval, for infrared spectroscopy in the detection of volatile organic gases is presented. The algorithm employs a series selection mode, utilizing characteristic wavelength point cluster classification and absorption peak interval screening. To begin with, cluster analysis is conducted to retain prominent absorption peak features while minimizing the potential for algorithmic over splitting or random uncertainty in wavelength intervals. Subsequently, an improved moving window method is devised, and a greedy algorithm is employed to re-screen wavelength points within the same cluster class. This process ensures the retention of the optimal wavelength range, crucial for representing spectral characteristics and facilitating subsequent model predictions. Experimental validation was conducted using infrared spectral data of styrene, para-xylene, and o-xylene, employing four models: partial least squares, ridge regression, support vector machine. The results demonstrate that, while maintaining model accuracy, the dataset can be reduced to 43.71%~36.35% of its original size. Additionally, utilizing a dataset comprising three gases (two concentrations each), as well as fully arranged and combined mixed gases, we conducted comparative experiments on three different CNN structures. The effectiveness of the proposed algorithm in reducing machine learning model complexity while ensuring prediction accuracy was validated through experimental comparisons before and after spectral waveform selection, with the CNN prediction models demonstrating a 90% increase in operational efficiency post-wavelength selection.

**Keywords**: infrared spectroscopy; wavelength selection algorithm; wave cluster area; neural network; prediction accuracy

收稿日期: 2024-08-22　Received Date: 2024-08-22
*基金项目:国家自然科学基金(12305041)、重庆市自然科学基金面上项目(CSTB2022NSCQ-MSX1370)、重庆市教委科学技术研究项目(KJZD-K202200803)项目资助

## 0　引　言

红外光谱(infrared spectroscopy,IR)尤其是近红外光谱检测技术因其高效、成本低、无损性以及无耗材等优势被广泛用于环境检测、化学化工、食品检测等各个领域[1-5]。挥发性有机气体(volatile organic compounds,VOCs)在生产生活中普遍存在且危害身体健康,因此需要大量成本适中的便携式检测设备。随着光谱硬件设备技术和应用范围不断延展,一种物质的红外光谱通常包含几万个波长变量,且通常具有多组分、干扰因素复杂以及包含电气、环境噪音等特点,大幅增加了红外线光谱定量预测难度。因此近年来有关红外光谱的定量分析性能的研究领域的热点和难点主要集中在光谱特征选择、高精度定量预测改进上。

在定量模型建立及改进上,以线性定量回归模型为核心的各种模型被广泛研究和运用,这些模型主要有最小二乘(partial least squares, PLS)、岭回归(ridge regression, RR)、主成分回归(principal component regression,PCR)[6-8]等。高分辨率光谱也同时意味着光谱信息中包含大量潜在变量,并呈现出越来越多的非线性特征[9],这些模型对潜在变量解释能力相对较弱,且无法跟随待测样品组分或者环境条件发生变化进行更新。因此近几年深度学习被逐渐引入到红外线光谱分析领域并逐渐成为了目前的重点研究方向[10],各种独立模型、组合模型相继被提出并运用,卷积神经网络[11](convolutional neural network, CNN)、反向(back propagation, BP)神经网络[12]、注意力机制[13-14](attention mechanisms)、EfficientNet 神经网络[15]、基于“指纹谱线”的神经网络架构[16]、粒子群优化[17-18]等。然而一方面实际环境中的近红外光谱有别于图形图像等深度学习的传统研究领域,不但具有非线性以及高纬等特性,哪怕是同一种气体的不同浓度红外线光谱结构差别也非常大,而每一个全谱段光谱通常都会有几万个波长变量;另外一方面越复杂深度学习模型就意味着需要越强大算力支持和越高昂硬件成本,这显然不能满足新型光谱检测设备的应用需求,因此需要更具有针对光谱特性的深度学习模型。

基于选择可以改善模型预测效果的假设,针对红外光谱特性在一定波长范围甚至全谱段范围内采取某种策略,找到更有利于后续模型的波长变量就被称为波长选择。目前根据波长选择算法策略上的不同,大体可以分为基于回归模型的波长选择算法,比如前向区间最小二乘法(forward interval partial least squares, FiPLS)、基于数据特征分析的波长选择算法,比如随机蛙跳[19](random frog jumping, RFR)、蒙特卡洛无信息消除法[20](Monte Carlo uninformative variable elimination, MC-UVE)、鲸鱼优化算法[21](whale optimization algorithm,WOA)等等,很多研究已经表明各种波长选择算法的组合在一些领域也具有非常不错的效果[22-23]。总体来看,波长选择算法可以非常有效地去除光谱信号中的噪音,增强模型预测能力[24-25]。实际上波长选择不仅为模型筛选出了有效波长变量同时也为尽量避免单纯依靠增加神经网络复杂程度来提升性能提供了有效途径。

有别于深度学习领域常用的图形、自然语言等数据大模型领域,红外光谱分析的关键是建立合理模型,核心是模型精度以及环境适应能力。本文首先根据特殊波长点上的吸收峰反映物质特性,同时又通常存在一定连续区间的特点,提出一种波长点和波长区间相集合的波长选择算法,以吸收峰波长点为核心建立了一种新簇类聚集算法,再采取一种改进的移动窗口机制筛选出合理波长区间,以波长区间做为波长选择的筛选结果,进一步在此基础上构建了多个不同结构卷积神经网络模型,进行了全谱段和经波长选择后的实验对比,由此建立了更为准确且数据量更小适应能力更强的预测模型,达到获取有效波长变量以及减少深度学习模型运算量的目的。

## 1　波长选择及预测模型建立

### 1.1　红外线光谱特征分析及算法提出

分子的振动和转动引起红外线吸收从而产生红外线吸收光谱,但每种气体由于分子结构不相同往往涉及多个振动和转动模式,因此不同物质光谱很难通过单一理论建模来进行分析。苯乙烯(styrene,C8H8)、对二甲苯(P-xylenes,C8H10)、邻二甲苯(O-xylenes,C8H10)为日常生活中非常常见的有机挥发气体,其中对二甲苯和邻二甲苯同是二甲基苯的同分异构体之一,其化学式同为C8H10。3 种物质的结构区别主要在于甲基基团的位置不同,这不仅导致它们的物理性质和化学性质有所不同,同时红外线吸收光谱迥然不同。以 EPA 光谱数据库中的 3 种气体红外线吸收光谱为例,起始波数 1 500 $cm^{-1}$ 至终止波数 1 515 $cm^{-1}$ 以及起始波数 1 535 $cm^{-1}$ 至终止波数 1 550 $cm^{-1}$ 的部分吸收光谱图如图 1 所示,可以看出 3 种气体吸收光谱在很多波长区间光谱中存在明显偏差,而在另一些波长区间中又可能存在相当大的重叠。受实际工况条件和经济成本限制不可能使用波长范围跨度很大的测量仪器,而如果使用单一理论模型分析它们的光谱又具有一定难度,因此很难得到较为满意的检测精度。

吸收光谱波长位置与物质特性相关。每种物质在全谱段上的表现并无规律,但无论吸收峰峰值是多少,位置在哪里,都能体现物质特性,且都能表现出来一定的区间

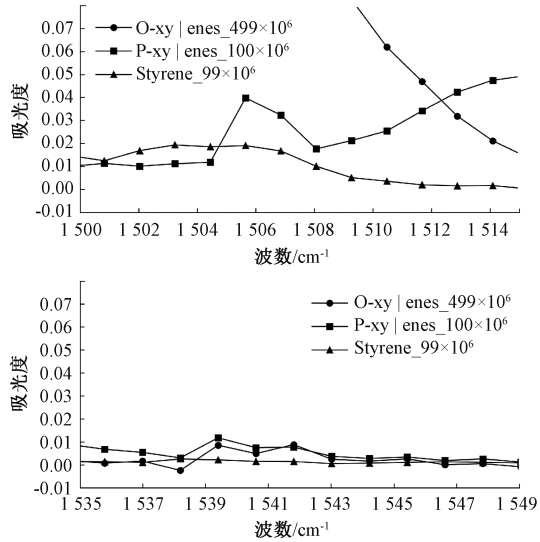



图 1 苯乙烯、对二甲苯、邻二甲苯吸收光谱图

Fig. 1 Absorption spectra of toluene, Styrene, P-xylene, and O-xylene

性。也就是说并不存在孤立的吸收峰。本研究提出了基于波簇区间的波长选择算法(wave cluster-based interval selection,WBIS)。算法首先采取类似于聚类的算法,确定能够代表物质特性的吸收峰大致位置;此时每个簇类中都会保留一定的吸收峰信息;进一步分析每个簇类中的波长点,以在簇类中寻找“差异最大化”的算法思想,采取移动窗口方式查找簇类的波长点,保留包含若干波长点之间起伏最大也就是最能体现吸收峰特性的窗口作为波长选择的结果,进入后面的模型预测。

### 1.2 波形选择算法模型设计

为分析每种的整个光谱图中吸收峰位置信息,设有某种浓度一定气体的全吸收光谱,其光谱吸收强度数据集 $P=\{p_1,p_2,\cdots,p_n\}$,令 $C=\{c_1,c_2,\cdots,c_k\}$ 为吸收峰中心,将数据集分成 $K$ 个簇,设定目标函数为:

$$D(C,\{S\})=\sum_{K=1}^{K}\sum_{i=1}^{n}S_{ik}\cdot\|p_i-c_k\|^2 \tag{1}$$

其中,$S=\{s_{ik}\}$ 表示由每个波长点以及对应的吸收峰强度组成的二元矩阵,并用 $s_{ik}=1$ 表示数据点 $p_i$ 属于簇 $k$,$s_{ik}=0$ 表示数据点 $p_i$ 不属于簇 $k$。$n$ 为全光谱图中吸收强度总数。以目标函数 $D$ 最小化为原则,通过不断迭代,找到 $C$ 和 $P$ 的簇分配关系 $R$。收敛条件设定为:

$$\forall j,\lim_{i\to\infty}\|C_{i+1}^{(j)}-C_i^{(j)}\|^n\leqslant\varepsilon \tag{2}$$

式中:$C_{i+1}^{(j)}$ 表示第 $i$ 次迭代完成后第 $j$ 个簇中心;$\varepsilon$ 为设定容差。这种做法不仅不会机械地将全光谱分割成若干个区间,又可以避免将吸收强度孤立化,找到气体全吸收光谱中最有利于分析吸收峰的位置区间。

波长点簇类 $CW=\{cw_1,cw_2,\cdots,cw_k\}$,$cw_j$ 起始区间为 $[a,b]$,采取移动窗口机制对 $cw_j$ 中的吸收强度进行进一步分析,$w_j$ 是 $cw_i$ 中的某一个移动窗口,计算 $n$ 个吸收峰强度彼此之间的相邻程度值 $d_j$,$d_j$ 计算方法如式(3)所示。

$$\max_j\left|\max_{(x_p,y_p,),(x_q,y_q)\in w_j}\left(\frac{x_p\cdot x_q+y_p\cdot y_q}{\sqrt{x_p^2+y_p^2}\cdot\sqrt{x_q^2+y_q^2}}\right)\right| \tag{3}$$

将 $D=\{d_1,d_2,\cdots,d_k\}$ 做为最能表达区间内吸收峰起伏的波长点,也就是波长选择的结果,用于后续的定量或者定性分析。

### 1.3 波长选择算法验证模型设计

为验证 WBIS 波长选择算法在波长筛选上的效果,设计了最小二乘(least squares,LS)、偏最小二乘(partial least squares,PLS)、岭回归(ridge regression,RR)、支持向量机(support vector machine,SVM)4 种模型。为保证验证的准确性,所有模型均不进行归一化处理。

PLS 模型成分数量为 1,未防止模型计算前后数值计算和实际光谱的偏差,不对数据进行缩放,模型迭代最大次数为 100,迭代收敛容差为 $1\times10^{-6}$。

RR 模型正则化强度为 1,收敛容差为 $1\times10^{-3}$,并允许计算截距项,采用 Cholesky 分解法作为求解方法,由于经过波长选择的区间不具有连续性,不使用随机种子。

由分子振动和旋转引起的吸收峰通常表现出非线性关系。径向基函数(RBF)核能更好地捕捉这种非线性关系。此外,RBF 内核不仅对异常值表现出鲁棒性,而且允许通过带宽参数调谐来调整高斯函数的宽度,使 RBF 内核能够适应不同光谱数据集的特性。因此,在 SVR 模型中,采用高斯径向基函数作为核函数,指定“poly”和“sigmoid”作为核函数的系数,并根据特征数的倒数自动计算 gamma。模型停止准则容差设置为 $1\times10^{-3}$,模型误差容差范围为 0.1。

### 1.4 不同结构的卷积模型设计

为验证经波长选择后波长变量对深度学习模型预测精度的影响,构建 3 种不同结构的卷积神经网络模型。一个典型的卷积神经网络模型由输入层、卷积层、池化层、全连接层、输出层组成。卷积神经网络在捕捉形状的轮廓特征和图像中的局部特征方面表现出色[26-28]。因此,如果红外光谱被视为一种“图像”形式,适当配置的多层卷积结构可以帮助捕获跨波长间隔的位置信息,通常更深层次、更复杂的卷积模型可以更好地捕捉光谱数据中的特征,更强的表征能力,从而提高模型预测能力,但在红外光谱预测中还意味着算力提升对应更高硬件造价要求,这并不有利于红外光谱检测仪器实际各种复杂应用环境,因此一个好的波长选择算法应该能在保证预测精度的前提下降低模型复杂程度。

为更客观比对模型预测结果,所有卷积模型均使用 ReLU 激活函数,使用 Adam 优化器和均方误差作为损失函数,并均包含全连接层和输出层,但在能直接决定卷积模型复杂程度的卷积层数量、滤波器数量和大小、池化层的数量上有所不同,3 种 CNN 模型的结构比较如表 1 所示。

**表 1　3 种不同结构 CNN 结构对比**

**Table 1　Comparison of three different CNN architectures**

| | Mode1 | Mode 2 | Mode 3 |
|---|---|---|---|
| 卷积层数 | 1 | 2 | 4 |
| 滤波器数量(每层) | 32 | 32、64 | 32、64、128、256 |
| 滤波器大小(每层) | 3 | 3、3 | 3、3、3、3 |
| 最大池化层数量 | 1 | 2 | 4 |
| 全连接层数量 | 1 | 1 | 2 |
| 输出层 | 1 | 1 | 1 |

## 2　实验部分

### 2.1　实验数据

采用污染气体红外吸收光谱研究领域常用的美国环境保护署(environmental protection agency,EPA)光谱数据库作为数据来源[29-32]。使用该数据库中苯乙烯、对二甲苯、邻二甲苯 3 种气体,光谱测量温度 25 ℃,测量光程 3 m,起始波数 400 $cm^{-1}$,终止波数 4 000 $cm^{-1}$,光谱分辨率 0.125 $cm^{-1}$。这 3 种物质分子结构类似,吸收峰重叠性较高并且在涂料、油漆和橡胶等常见物质中广泛存,因此仿真实验更具有典型性和实际运用价值,实验样本数据情况如表 2 所示。

**表 2　样本数据集气体及浓度**

**Table 2　Gas and concentration of sample dataset**

**($\times 10^{-6}$)**

| O-xylene | P-xylene | Styrene |
|---|---|---|
| 499 | 100 | 99 |
| 99 | 502 | 500 |

数据集中一种气体每一种浓度全谱段的光谱数据量高达 33 185,原始数据精确到小数点后 4 位,总数据量为 199 110。对单一气体进行波长选择阶段,分别使用浓度 $99\times10^{-6}$ 的苯乙烯、浓度 $100\times10^{-6}$ 对二甲苯、浓度 $499\times10^{-6}$ 邻二甲苯 3 种气体独立完成波长选择并进行效果分析。

为了考察在波长选择前后在不同 CNN 模型中复杂光谱预测性能,1.4 节中的 3 种 CNN 模型的输入数据集并非单一气体,而是根据 Lambert-Beer 定律,在不考虑气体之间的相互作用的情况下模拟多组分的混合气体,将表 2 的 3 种气体各自 2 种浓度共 8 种组合,在全光谱条件下得到 265 480 个样本数据。显然若将全谱作为模型的输入数据,意味着非常高的计算量,将极大限制模型通用性能和执行效率。

### 2.2　实验过程及验证方法

所有实验均在 pycharm 2020.3.2 上进行。首先对 3 种气体全谱段进行波长筛选,用筛选前后的数据集在 LS、PLS、RR、SVM 4 种模型上进行预测,为了能更准确地评估 4 种模型预测性能对比性以及泛化性能,4 种模型均采用交叉验证机制,4 种模型所使用到的交叉验证数学模型为:

$$K(D,K)=\frac{1}{K}\sum_{k=1}^{K}RMSE_k \tag{4}$$

其中,$RMSE_k$ 表示第 $k$ 次交叉验证的均方误差。全谱段数据集和经过波长选择后的数据集都会被划分为 5 个互斥的子集,然后进行 5 折交叉验证。每次训练时会使用 4 个子集作为训练集,剩下的 1 个子集作为测试集,然后重复 5 次。模型的性能指标将基于这 5 次交叉验证的结果进行计算。

上述波长筛选仅针对单一气体,每种气体经过波长筛选后留下的波长变量不尽相同,因此根据 Lambert-Beer 定律在生成混合气体数据集时,按照以下算法模型进行保留。

物质单一全谱段经波长筛选后为独立数据集 $D_1$,$D_2,\cdots,D_n$,其中 $D_i=\{\lambda_i,I_i,i=1,2,\cdots,m\}$,$D_i\in C_i$。其中 $C_i$ 为对应物质全谱段数据集。构造新的数据集 $S=\{S_1,S_2,\cdots,S_n\}$,其中 $S_i$ 满足下面条件:

$$S_k=\{(\lambda_k,I_{1,k},I_{2,k},I_{3,k},\cdots,I_{n,k}),\lambda_k\varepsilon D_1\cap D_2\cap\cdots,D_n\} \tag{5}$$

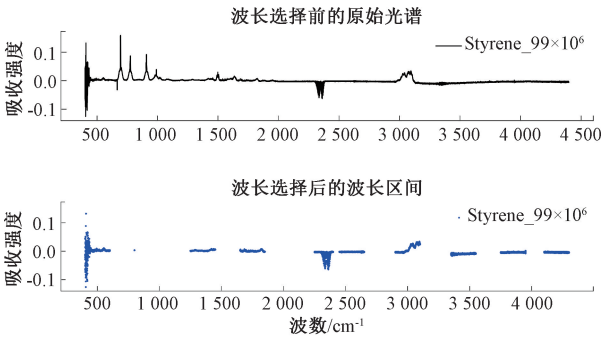
在不考虑物质间分子结构相互影响前提下,完成 3 种气体共 8 种混合气体全谱段以及波长筛选完成后的数据集,并进行随机化。

完成波长筛选前后 3 种 CNN 模型的预测验证。

## 3　结果与讨论

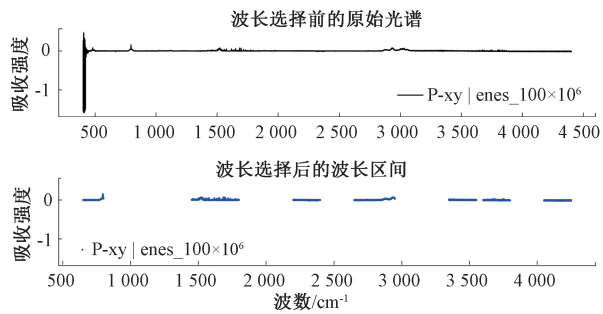
### 3.1　波长选择算法改进效果分析

WBIS 算法筛选出来的波长点和 $k$ 值大小有关,在深度学习模型中过少的数据量并不有利于复杂模型学习和预测[33]。在 WBIS 算法选择设置 $k=10$,窗口大小为 200,步长为大小为 50 时,3 种单一气体全光谱和经过波长选择后留下的波长区间位置如图 2 所示。显然波长选择算法并非只保留了吸收强度值最大、吸收峰值最明显的单个区间,而是尽可能地保留了全谱段中存在的多个吸收峰位置,对于那些吸收峰强度值比较小的区间,根据式(3)也尽可能地保留了能反映出吸收峰的区间,这对于后续多种气体混合环境中的定量测试是非常有利的。



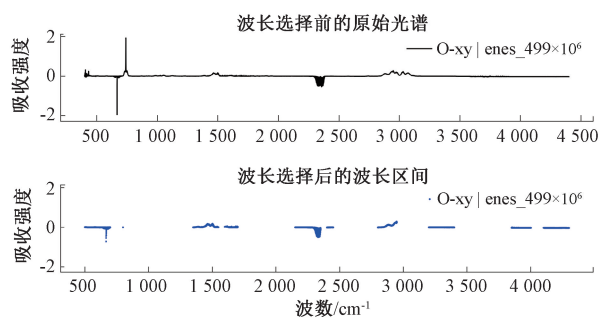

(a) 苯乙烯原始光谱与波长选择后波长区间的对比

(a) Comparison of the original spectrum of Styrene and the selected wavelength range after wavelength selection




(b) 对二甲苯原始光谱与波长选择后波长区间的对比

(b) Comparison of the original spectrum of Para-xylene and the selected wavelength range after wavelength selection




(c) 邻二甲苯原始光谱与波长选择后波长区间的对比

(c) Comparison of the original spectrum of O-xylene and the selected wavelength range after wavelength selection

图 2 苯乙烯、对二甲苯、邻二甲苯原始光谱与波长选择后波长区间对比

Fig. 2 Comparison of the original spectra of Styrene, Para-xylene, and Ortho-xylene with the selected wavelength range after wavelength selection

经波长选择算法后,苯乙烯、对二甲苯、对二甲苯保留的波长点数量分别为 14 505、13 250、12 063。分别占原始全光谱数据的 43.71%、39.92%、36.35%。通过图 2 可以看到 3 种气体经波长选择后的光谱数据量大量减少的同时整体波长区间被保留的范围并不相同,取 3 种气体都保留有光谱数据的 1 400~1 480 $cm^{-1}$ 波数区间,随机抽取 1/10 的数据对比如图 3 所示,对比图 1 不难看出选择出来的波长区间各自充分保留了物质的吸收峰特性,这在对二甲苯上尤为明显;3 种气体选择出来的波长区间有些位置重叠,原因是 3 种物质本身分子结构类似导致在红外线光谱上具有一定的相似性,但结合选出的波长点数量上不难看出,3 种气体选出的波长区间并不受吸收峰高低影响且又有比较好的区分度。

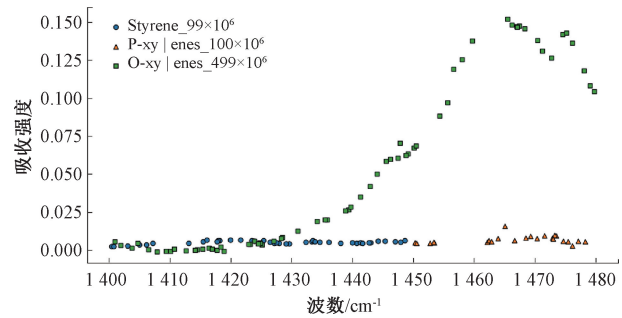



图 3 苯乙烯、对二甲苯、邻二甲苯经波长选择后 1 400 $cm^{-1}$-1 480 $cm^{-1}$ 波长区间保留的光谱数据对比

Fig. 3 Comparison of spectral data retained in the wavelength range of 1 400 $cm^{-1}$-1 480 $cm^{-1}$-for Styrene, P-xylene, and O-xylene after wavelength selection.

在 LR、PLS、RR、SVR 4 种模型上,以预测集均方根误差(RMSEP)、以及和校准集均方根误差(RMSEC)、校准集相以及模型均方根误差(RMSE)进行对比,此时单一气体波长变量数量从 33 185 降至 12 063,筛选后的波长变量约为原始数据集的 36.35%。

4 种模型的 RMSE 性能表现如表 3 所示,3 种气体分别在 LR、PLS、RR 模型整体性能依次逐渐表现更好。

**表 3 波长选择前后 4 种模型的 RMSE 性能**

**Table 3 RMSE performance of four models before and after wavelength selection**

| Data set | Model | Before | After |
|---|---|---|---|
| Styrene | LR | 0.009 2 | 0.009 2 |
| | PLS | 0.009 2 | 0.009 2 |
| | RR | 0.009 2 | 0.009 2 |
| | SVR | 0.035 5 | 0.013 0 |
| P-xylenes | LR | 0.061 8 | 0.013 0 |
| | PLS | 0.061 8 | 0.013 0 |
| | RR | 0.061 8 | 0.013 0 |
| | SVR | 0.088 2 | 0.060 1 |
| O-xylenes | LR | 0.068 7 | 0.065 7 |
| | PLS | 0.068 7 | 0.065 7 |
| | RR | 0.068 7 | 0.065 7 |
| | SVR | 0.128 9 | 0.156 5 |

由表 3 看出虽然 PLS 和 RR 模型本身具有更好的非线性表现力有关,但在 SVR 模型中苯乙烯、对二甲苯降幅明显,表明这 2 种气体更适合 SVR 模型,而邻二甲苯性能不但没有降低反而提高,表明邻二甲苯并不适合

SVR 模型。对比 3 种气体的 RMSE 可以发现,3 种气体在同一模型上的预测精度差别也比较大,这充分说明了不同物质在红外光谱中数据特点迥异,并不具有规律性,因此波形选择对复杂情况下多组分气体预测显得尤为重要。

另外预测集误差表现情况,在数据集减少 63.75%的情况下对比波形选择前后,苯乙烯在 LR、PLS、RR、SVR 模型中 RMSEP 的降幅分别为 0.32%、-0.17%、0.32%、63.39%;对二甲苯在 4 种模型中 RMSEP 降幅分别为 79.03%、80.21%、79.03%、31.86%;邻二甲苯 LR、PLS、RR 模型中的 RMSEP 降幅分别为 4.36%、3.62%、4.36%。由此表明在波长选择前后对模型预测精度没有影响或性能有所提升。

从校准数据上的拟合程度上来看,苯乙烯在 LR、PLS、RR 模型上,RMSEC 在波形选择前后没有明显变化但在 SVR 模型行有明显大幅度降低;对二甲苯在 4 种模型上 RMSEC 均有明显大幅度降低;邻二甲苯在适合它的 LR、PLS、RR 模型上均有一定幅度的降低。由此表明模型对经过选择的数据集具有更好地拟合能力。

从模型整体性能上来看,除了邻二甲苯在 SVR 模型上的表现以外,在数据集减少 63.75%的情况下对比波形选择前后,3 种气体分别在 4 种模型上的 RMSE 均有不同程度的降低,降幅因气体种类不同而不同,同种气体在不同模型上均有降幅且表现稳定,

由此经 WBIS 算法波长选择后的数据集不但不会降低模型预测精度,反而对预测模型的稳定性和精度均有所帮助,由此表明 WBIS 算法在波长选择上的有效性。

### 3.2　模型预测效果改进分析

波长选择根本目的是为后续模型处理提供更优质的数据集。以最小二乘法为代表的传统吸收光谱处理模型通常在单一气体预测上有较好表现,但现实环境中基本不可能存在单一气体,同时电器噪音、仪器老化、水分、温度、多组分气体分子间相互影响等都会影响气体红外线光谱,同一仪器不同外部环境下得到的吸收光谱也不尽相同,从而表现出了非常复杂的非线性特性,因此近几年各种深度学习模型被逐渐引入到了吸收光谱定量和定性分析中,特别在多组分气体精度检测提升上取得了不错的效果[34-35]。深度学习模型在提升精度的同时也带来了更大的算力需求和硬件要求,这对于挥发性有机气体仪器的快速响应、成本控制以及便携式产品设计非常不利。因此需要在不损失检测精度的同时缩小数据集,尽可能采用更简单的模型结构。

为进一步验证经 WBIS 波长筛选前后数据集对深度学习模型预测能力的影响,本研究并未简单将波长选择前后的单一气体数据集在 CNN 模型中进行对比,而是将波形选择前后的 3 种浓度气体均采用全排列组合方式构成共 8 种不同混合气体组分形成输入,以 3 种气体浓度作为输出。数据集中随机抽取 80%为训练集,20%为测试集,epochs 设置均为 30,构造了表 1 所述 3 种不同的 CNN 模型进行了实验对比。

WBIS 算法中设置 $k$ 为 10 时,单一气体经波形选择之后的波长点为 3 312 个,但由于每个单一数据集波长点位置并不完全相同,采用式(5)的方法保留了 3 种气体共同 831 个波长点数据进行后续 3 个 CNN 模型测试。

表 4 为 3 种气体波长选择前后在 3 个卷积网络中的预测效果。从整体来看 RMSE 和平均绝对误差(mean absolute error,MAE)值比较小,表明模型每个输出的预测效果较好,预测结果与真实值之间的平均偏差较小,说明模型预测效果良好。

**表 4　波长选择前后在 3 个 CNN 模型中实验预测**

**Table 4　Experimental predictions in three CNN models before and after wavelength selection**

| Model | RMSE | |
|---|---|---|
| | Before | After |
| Model 1 | 0.108 43 | 0.000 55 |
| Model 2 | 0.013 61 | 0.000 41 |
| Model 3 | 0.001 75 | 0.000 59 |
| **Model** | **MAE** | |
| | Before | After |
| Model 1 | 0.088 56 | 0.000 41 |
| Model 2 | 0.011 00 | 0.000 33 |
| Model 3 | 0.001 60 | 0.000 53 |

3 种模型的复杂程度依次递增,从表 4 得知随着 CNN 网络模型复杂程度增加,RMSE 预测精度逐渐提升,MAE 也出现了明显的下降趋势,这表明采用 CNN 模型的合理性,同时也表明 CNN 对于该数据集有足够好的预测效果且误差精度合理;比较波长选择前后的 RMSE 和 MAE 可以看到,在数据集仅为原始数据量的 36.75%的前提下,经过波长选择的数据集表现出了更好的性能,体现出了更好的预测效果,说明了波长选择的有效性;进一步分析得知,波长选择后在 Model 3 上 RMSE 和 MAE 和 Model 2 对比发现略有升高,这表明选择后的数据集已经在 Model 3 开始出现过拟合,模型复杂程度提升并未带来更好的性能,Model 2 更适合该数据集;可见经 WBIS 算法波长选择后的数据集在同一结构的 CNN 模型中可提升了模型的预测精度,且在保证检测误差前提下不再需要更复杂 CNN 模型。

对比模型运行消耗情况,表 5 为同一实验系统环境下(2.5 GHz 酷睿 i7,1 G 内存,64 位 Win10)波形选择前后 3 种 CNN 模型运行时间增长情况(不包含波形选择算法运行时间)。首先由于经波长选择算法之后数据量远远低于选择前,运行时间会出现较大幅度,运算时间约为

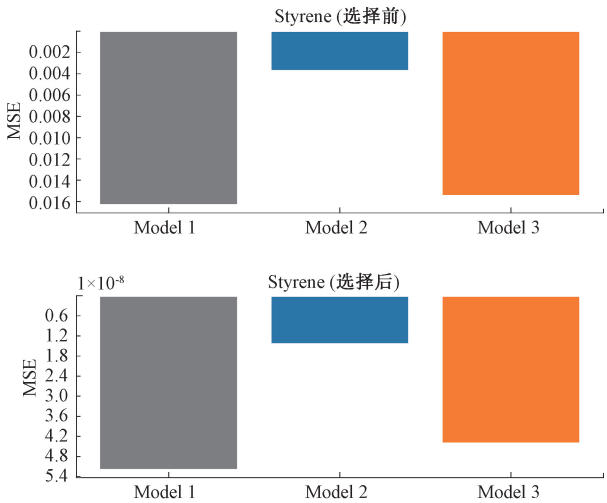
原来的 7.6%、10.06%以及 9.08%，进一步分析运行时间增长幅度，当模型复杂程度成倍提升以后（3 个模型的每层滤波器数量分别为 32、32，64、32，64，128，256），选择前 Model2、Model3 与 Model 相比运行时间增幅依次为 1.86%、31.63%，而选择后的模型运行时间虽较少但增长比例很大，Model2、Model3 与 Model 相比运行时间增幅分别为 35.09%、57.46%。结合前面分析可见 Model 2 是最合适的模型结构。因此在保证足够预测精度的前提下，结构更简单的模型才更适合近红外光谱分析。

**表 5 波形选择前后 3 种模型运行时间对比**

**Table 5 Comparison of runtime growth for 3 CNN models before and after waveform selection (s)**

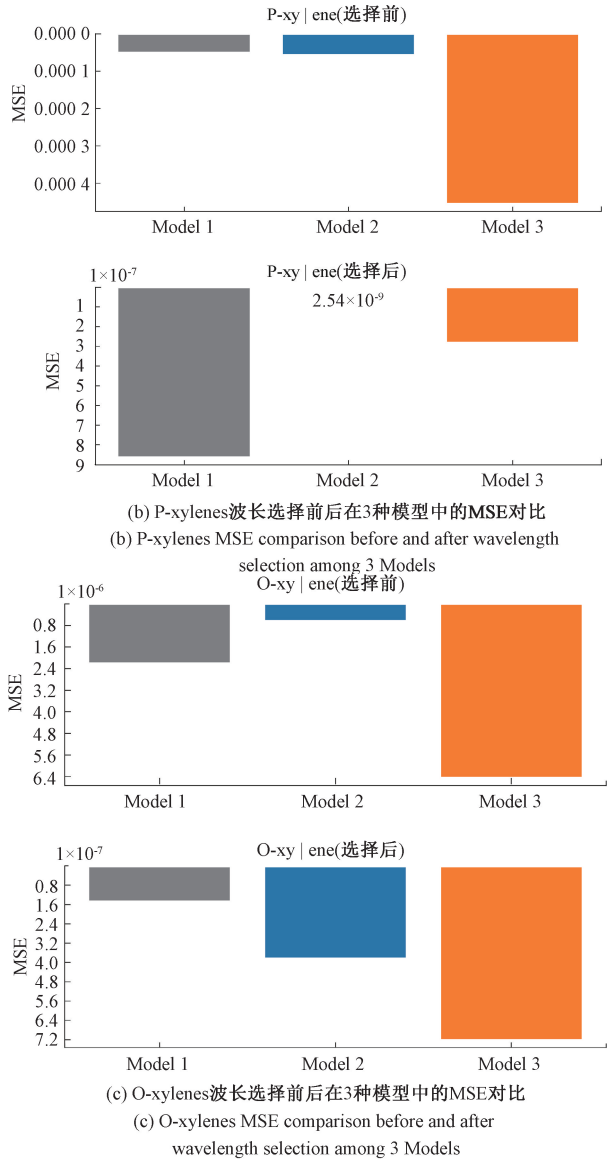
| Waveform Selection | Before | After |
|---|---|---|
| Model 1 | 30.03 | 2.28 |
| Model 2 | 30.59 | 3.08 |
| Model 3 | 39.53 | 3.59 |

图 4 所示为 3 种模式下 3 种气体在波长选择前后的 MSE 值情况。经对比 3 种气体预测精度在很大程度上取决于全光谱数据集中的模型复杂性。随着模型复杂度的增加，预测精度显著提高。比如苯乙烯 MSE 从最大值 0.016 248 降至最小值 $2.17\times10^{-6}$，主要原因是在全谱段数据集下数据量达到 265 480，需要更复杂的网络结构才能取得更好的效果，但这同时也意味着更高的硬件要求、更低的应用环境适应能力；经波长选择后 3 种气体的预测精度首先表现出了更好的预测精度，比如苯乙烯在 Model1MSE 值为 $5.17\times10^{-8}$，其次随着模型复杂程度的提升误差改变明显收窄，也就是说模型复杂度提升对预测精度提升帮助并不大，波长选择后的数据集在简单模型上表现出了非常好的预测效果，这无疑提升了模型的适应能力，有利于降低实际设备造价。




(a) Styrene波长选择前后在3种模型中的MSE对比

(a) Styrene MSE comparison before and after wavelength selection among 3 Models




(b) P-xylenes波长选择前后在3种模型中的MSE对比

(b) P-xylenes MSE comparison before and after wavelength selection among 3 Models

(c) O-xylenes波长选择前后在3种模型中的MSE对比

(c) O-xylenes MSE comparison before and after wavelength selection among 3 Models

图 4 Styrene、P-xylenes、O-xylenes 波长选择前后在 3 种模型中的 MSE 对比

Fig. 4 Styrene, P-xylenes, O-xylenes MSE comparison before and after wavelength selection among 3 models

3 种气体在同一模型下的预测精度表现不尽相同。其中对二甲苯波长选择前后的预测精度都是最高，且明显优于苯乙烯和邻二甲苯，同时在波长选择前后，Model1 中对二甲苯的预测精度不仅没有降低，而是有所提高。但在 Model2 中，它表现出了出色的预测性能，这表明 CNN 模型参数对这种气体数据集特别敏感。这种差异可说明对二甲苯数据集与 CNN 网络结构具有更好的兼容性；相比之下邻二甲苯的误差在选择前后都是最高的，模型复杂性对误差精度的影响最小。这表明与其他两种气体相比，邻二甲苯的数据集与 CNN 结构的兼容性较差；在波形选择之前，苯乙烯的预测性能随着模型复杂性

的增加而不断提高。虽然波形选择后的整体预测效果比之前好,但预测性能仍然随着模型复杂度的增加而提高。

由此可见深度学习模型在红外线光谱数据集中的分析有别于常见神经网络研究领域,其模型表现力受物质分子特性的影响比较大,不同结构物质表现出来的特性各不相同,在进行模型设计时并不能单一通过增加模型复杂度来提升预测效果,更应考虑物质个性和复杂外部环境。

## 4　结　论

本文提出了一种波长点和吸收峰区间相结合的近红外光谱波长选择方法,即 WBIS 算法,算法核心思想是根据光谱吸收特性设置簇中心并利用移动窗口相结合的方式找到最有利于深度学习模型的多个吸收峰区间,从而形成新的数据集。利用在 3 种浓度的苯乙烯、对二甲苯、邻二甲苯上进行了波长筛选后,波长点数量从原来的 33 185 降至 3 321 个,并在传统的 LR、PLS、RR、SVR 模型上进行了对比分析,模型预测性能优于选择前。为验证波长选择算法的适应性,用每种气体各自 2 种浓度的数据集构造了混合气体数据集,并采用 3 种不同结构的 CNN 模型进行了对比,实验表明在 2 层卷积结构(Model 2)时,3 种气体均取得了最好的误差效果和运行效率。本研究表明在将深度学习模型相关研究引用到红外线光谱分析中时,WBIS 波长选择算法对缩小数据集规模、降低网络模型复杂度上积极有效,这为深度学习机制在红外光谱中的应用改进以及成本控制提供了参考,有利于挥发性有机气体便携式检测设备设计;同时也表明在深度学习模型设计时应对光谱种类及自身特性予以足够关注。希望今后能在更多种类红外吸收光谱中做进一步研究。

## 参考文献


[1] IBRAHIM E A, ALHAITHLOUL H A S, SHAMSELDIN S A M, et al. Morphological, biochemical, and molecular diversity assessment of egyptian bottle gourd cultivars[J]. Genetics Research, 2024(1): 4182158.

[2] BODDAPATI V, FERRIS A M, HANSON R K. Predicting the physical and chemical properties of sustainable aviation fuels using elastic-net-regularized linear models based on extended-wavelength FTIR spectra [J]. Fuel, 2024, 356: 129557.

[3] 王昱麒,李斌,朱明旺,等. 应用最小角回归索套算法优选苹果糖度预测模型的建模样本和波长[J]. 光谱学与光谱分析,2023,43(5):1419-1425.
WANG Y Q, LI B, ZHU M W, et al. Optimizations of sample and wavelength for apple brix prediction model based on LASSO Lars algorithm[J]. Spectroscopy and Spectral Analysis, 2023, 43(5): 1419-1425.

[4] 赵海龙,甘淑,袁希平,等. 基于多尺度连续小波分解的土壤氧化铁反演[J]. 光学学报,2022,42(22): 209-216.
ZHAO H L, GAN SH, YUAN X P, et al. Inversion of soil iron oxide based on multi-scale continuous wavelet decomposition[J]. Acta Optica Sinica, 2022, 42(22): 209-216.

[5] LIN H, SHU G, XIPING Y, et al. Spatial differentiation analysis of water quality in dianchi lake based on GF - 5 NDVI characteristic optimization [J]. Journal of Spectroscopy, 2021(1): 5542126.

[6] DIAZ V F, DE KETELAERE B, AERNOUTS B, et al. Cost-efficient unsupervised sample selection for multivariate calibration[J]. Chemometrics and Intelligent Laboratory Systems, 2021, 215: 104352.

[7] MIAO X X, MIAO Y, LIU Y, et al. Measurement of nitrogen content in rice plant using near infrared spectroscopy combined with different PLS algorithms[J]. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2023, 284: 121733.

[8] 胡会强,位云朋,徐华兴,等. 基于高光谱成像技术和主成分分析对粉葛年限的鉴别[J]. 光谱学与光谱分析,2023,43(6):1953-1960.
HU H Q, WEI Y P, XU H X, et al. Identification of the age of puerariae thomsonii radix based on hyperspectral imaging and principal component analysis [J]. Spectroscopy and Spectral Analysis, 2023, 43(6): 1953-1960.

[9] 李忠兵,袁章雨,梁海波,等. 多层非线性局部感受野极限学习机方法用于录井气体分析[J]. 仪器仪表学报,2024,45(3):157-169.
LI ZH B, YUAN ZH Y, LIANG H B, et al. Multi-layer nonlinear local receptive field extreme learning machine method for logging gas analysis[J]. Chinese Journal of Scientific Instrument, 2024, 45(3): 157-169.

[10] GONG W, HU J, WANG Z, et al. Recent advances in laser gas sensors for applications to safety monitoring in intelligent coal mines[J]. Frontiers in Physics, 2022, 10: 1058475.

[11] LI L Q, PAN X P, FENG Y C, et al. Deep convolution network application in identification of multi-variety and multi-manufacturer pharmaceutical [J]. Spectroscopy and Spectral Analysis, 2019, 39(11): 3606-3613.

[12] 雷蕾,滕亚君,刘汗青,等. 基于分子光谱的翡翠不同产地快速鉴别研究[J]. 激光与光电子学进展,2021,58(12):516-521.
LEI L, TENG Y J, LIU H Q, et al. Rapid discrimination

of jade origins based on molecular spectra[J]. Laser & Optoelectronics Progress, 2021, 58(12): 516-521.

[13] 陈广秋,温奇璋,尹文卿,等. 用于红外与可见光图像融合的注意力残差密集融合网络[J]. 电子测量与仪器学报,2023,37(8):182-193.
CHEN G Q, WEN Q ZH, YIN W Q, et al. Attentional residual dense connection fusion network for infrared and visible image fusion[J]. Journal of Electric Measurement and Instrumentation, 2023,37(8):182-193.

[14] 漆建环,倪波,周晓彦,等. 基于注意力密集网络的伪彩色红外与可见光图像融合[J]. 国外电子测量技术,2024,43(5):84-91.
QI J H, NI B, ZHOU X Y, et al. Pseudo-color infrared and visible image fusion based on attention-dense network[J]. Foreign Electric Measurement Technology, 2024,43(5): 84-91.

[15] 钟扬,吴黎明,温腾腾,等. 基于深度神经网络的液体视觉识别研究[J]. 电子测量技术,2022,45(11): 22-29.
ZHONG Y, WU L M, WEN T T,et al. Research on liquid vision recognition based on deep neural network [J]. Electric Measurement Technology, 2022,45(11):22-29.

[16] 徐鹏,杨根,王寅,等. 基于红外吸收光谱的 NEPE 推进剂贮存寿命无损监测[J]. 固体火箭技术,2023,46(3):439-446.
XU P, YANG G, WANG Y, et al. Non-destructive monitoring of storage life for NEPE propellant based on infrared absorption spectroscopy [J]. Journal of Solid Rocket Technology. 2023,46(3):439-446.

[17] 闫格,张磊,于玲,等. 面向天然气泄漏检测的中红外甲烷传感系统与应用[J]. 中国激光,2022, 49(18): 118-126.
YAN G, ZHANG L, YU L, et al. Mid-infrared methane sensor system for natural gas leakage detection and its application [J]. Chinese Journal of Lasers, 2022, 49(18): 118-126.

[18] 吴旭阳,管港云,刘志伟,等. 基于改进的粒子群优化-反向传播神经网络的 CO2 红外吸收光谱定量分析[J]. 光学学报,2024,44(11): 313-322.
WU X Y, GUAN G Y, LIU ZH W, et al. Quantitative analysis of CO2 infrared absorption spectrum based on improved particle swarm optimization-back propagation neural network[J]. Acta Optica Sinica, 2024, 44(11): 313-322.

[19] 李伟,谭峰,张伟,等. 改进随机蛙跳算法在大豆品种快速鉴别中的应用[J]. 光谱学与光谱分析,2023,43(12):3763-3769.
LI W, TAN F, ZHANG W, et al. Application of improved random frog algorithm in fast identification of soybean varieties [J]. Spectroscopy and Spectral Analysis, 2023, 43(12): 3763-3769.

[20] 王怡淼,朱金林,张慧,等. 基于 MC-UVE、GA 算法及因子分析对葡萄酒酒精度近红外定量模型的优化研究[J]. 发光学报,2018,39(9):1310-1316.
WANG Y M, ZHU J L, ZHANG H, et al. Optimization of near infrared quantitative model for wine alcohol content based on MC-UVE, GA algorithm and factor analysis. [J]. Chinese Journal of Luminescence, 2018, 39(9): 1310-1316.

[21] 王仲雨,高美凤. 基于改进鲸鱼优化算法的近红外光谱波长变量选择方法及其应用[J]. 分析测试学报,2023,42(1):37-44.
WANG ZH Y, GAO M F. Selection of near infrared spectral wavelength variables based on improved whale optimization algorithm and its application[J]. Journal of Instrumental Analysis,2023,42(1):37-44.

[22] 孙代青,谢丽蓉,周延,等. 基于近红外光谱的 SG-MSC-MC-UVE-PLS 算法在全血血红蛋白浓度检测中的应用[J]. 光谱学与光谱分析, 2021, 41(9): 2754-2758.
SUN D Q, XIE L R, ZHOU Y, et al. Application of SG-MSC-MC-UVE-PLS algorithm in whole blood hemoglobin concentration detection based on near infrared spectroscopy[J]. Spectroscopy and Spectral Analysis, 2021, 41(9): 2754-2758.

[23] BALABIN R M, SMIRNOV S V. Variable selection in near-infrared spectroscopy: Benchmarking of feature selection methods on biodiesel data [J]. Analytica Chimica Acta, 2011,692(12):63-72.

[24] YUN Y H, LI H D, DENG B C, et al. An overview of variable selection methods in multivariate analysis of near-infrared spectra [J]. Trends in Analytical Chemistry, 2019, 113: 102-115.

[25] LIU J, LUO X, ZHANG D,et al. Rapid determination of rice protein content using near-infrared spectroscopy coupled with feature wavelength selection[J]. Infrared Physics & Technology, 2023, 135: 104969.

[26] LI Y, WANG J, CHEN Z, et al. Artificial neural network for the quantitative analysis of air toxic VOCs [J]. Analytical Letters, 2001, 34(12): 2203-2219.

[27] NEBAUER C. Evaluation of convolutional neural networks for visual recognition[J]. IEEE Transactions on Neural Networks, 1998, 9(4): 685-696.

[28] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[29] TRIPATHY B, DASH A, DAS A P. Detection of environmental microfiber pollutants through vibrational

spectroscopic techniques: Recent advances of environmental monitoring and future prospects [J]. Critical Reviews in Analytical Chemistry, 2024, 54(7): 1925-1935.

[30] SEESAARD T, KAMJORNKITTIKOON K, WONGCHOOSUK C. A comprehensive review on advancements in sensors for air pollution applications [J]. Science of The Total Environment, 2024(915): 175696.

[31] EP R, KOCH M. On-site detection of volatile organic compounds (VOCs)[J]. Molecules, 2023, 28(4): 1598.

[32] LAN H, HARTONEN K, RIEKKOLA M L. Miniaturized air sampling techniques for analysis of volatile organic compounds in air [J]. TrAC Trends in Analytical Chemistry, 2020, 126: 115873.

[33] 刘建伟,赵会丹,罗雄麟,等. 深度学习批归一化及其相关算法研究进展[J]. 自动化学报,2020,46(6): 1090-1120.
LIU J W, ZHAO H D, LUO X L, et al. Research progress on batch normalization of deep learning and its related algorithms[J]. Acta Automatica Sinica, 2020, 46(6): 1090-1120.

[34] 黄文彪,夏滑,王前进,等. 基于 BP 神经网络模型的呼出气 δ13C、δ18O 同位素丰度测量方法研究[J]. 光谱学与光谱分析,2024,44(10):2761-2767.
HUANG W B, XIA H, WANG Q J, et al. Research on measurement method of δ 13C and δ 18O isotopes abundance in exhaled gas based on the BP neural network model[J]. Spectroscopy and Spectral Analysis, 2024, 44(10): 2761-2767.

[35] 阚玲玲,朱富海,梁洪卫. 基于 1D-WCWKCNN 的痕量甲烷气体浓度检测[J]. 光谱学与光谱分析,2024,44(3): 829-835.
KAN L L, ZHU F H, LIANG H W. Detection of trace methane gas concentration based on 1D-WCWKCNN[J]. Spectroscopy and Spectral Analysis, 2024, 44(3): 829-835.

## 作者简介

**严玥**,1998 年于西南大学获学士学位,2005 年于重庆大学获得硕士学位,现任重庆工商大学副教授,主要研究方向为信号处理及运用、传感器技术方向研究。
E-mail: 1193173573@qq.com

**Yan Yue** received her B. Sc. degree from Southwest University in 1998, M. Sc. degree from Chongqing University in 2005. Now she is an associate professor at Chongqing University of Technology and Business. Her main research interests include signal processing and application, and sensor technology.

**许世豪**,2022 年于淮阴工学院获学士学位,现为重庆工商大学硕士研究生,主要研究方向为红外线信号处理及运用、传感器技术方向。
E-mail: yycq@ctbu.edu.cn

**Xu Shihao** received his B. Sc. degree from Huaiyin Institute of Technology in 2022. Now he is a M. Sc. candidate at Chongqing University of Technology and Business. His main research interests include infrared signal processing and application, and sensor technology.