Modeling and Simulation 建模与仿真, 2024, 13(3), 3247-3261
Published Online May 2024 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2024.133296

Hans汉斯

基于机器学习对泌尿类疾病标志物气体识别模式研究

孙宇帆[1*]，黄志健[1]，俞志超[1]，韩雨彤[1]，曹　明[2]，朱志刚[1#]

[1]上海理工大学健康科学与工程学院，上海
[2]上海交通大学医学院附属仁济医院泌尿科，上海

收稿日期：2024年4月24日；录用日期：2024年5月22日；发布日期：2024年5月31日

摘　要

泌尿类疾病，例如膀胱癌和前列腺癌，对全球健康构成严重威胁。本研究针对泌尿类疾病标志性VOC气体(甲苯，乙苯，异丙醇，戊醛)，通过选取电子鼻采集的多传感器信号与这四类气体的关联特征，采用常规气体传感器特征，建立四分类VOC分类预测模型。采用主成分分析(PCA)对样本点降维，分别使用三种分类算法：K-邻近(KNN)，支持向量机(SVM)和随机森林(RF)进行分类预测，三者准确率分别达到了88%，85%和91%。最后使用Stacking集成方式，分别对KNN和SVM，KNN和RF，SVM和RF进行两两集成，集成后的准确率有明显提升，其中效果最佳的集成方式是SVM和RF，其准确率达到了97%。研究表明stacking集成的SVM和RF模型成功地预测四种标志物VOC，为泌尿类关键疾病的早期筛查和无创检测打下坚实基础。

关键词

气体分类，电子鼻，机器学习，集成算法

Research on Gas Recognition Pattern of Urinary Disease Markers Based on Machine Learning

Yufan Sun[1*], Zhijian Huang[1], Zhichao Yu[1], Yutong Han[1], Ming Cao[2], Zhigang Zhu[1#]

[1]School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai
[2]Department of Urology, Renji Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai

Received: Apr. 24th, 2024; accepted: May. 22nd, 2024; published: May. 31st, 2024

[*]第一作者。
[#]通讯作者。

文章引用: 孙宇帆, 黄志健, 俞志超, 韩雨彤, 曹明, 朱志刚. 基于机器学习对泌尿类疾病标志物气体识别模式研究[J]. 建模与仿真, 2024, 13(3): 3247-3261. DOI: 10.12677/mos.2024.133296

Abstract

Urinary diseases, such as bladder cancer and prostate cancer, pose a serious threat to global health. This study focuses on the landmark VOC gases (toluene, ethylbenzene, isopropanol, and glutaraldehyde) of urinary diseases. By selecting the correlation characteristics between the multi-sensor signals collected by the electronic nose and these four gases, and using conventional gas sensor features, a four class VOC classification prediction model is established. Principal Component Analysis (PCA) was used to reduce the dimensionality of sample points, and three classification algorithms were used: K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Random Forest (RF) for classification prediction, with accuracy rates of 88%, 85%, and 91%, respectively. Finally, using the Stacking integration method, KNN and SVM, KNN and RF, SVM and RF were integrated pairwise, and the accuracy was significantly improved after integration. The best integration method was SVM and RF, with an accuracy of 97%. Research has shown that the SVM and RF models integrated with stacking have successfully predicted four biomarkers of VOC, laying a solid foundation for early screening and non-invasive detection of key urological diseases.

Keywords

Gas Classification, Electronic Nose, Machine Learning, Integrated Algorithm

Copyright © 2024 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

泌尿类疾病，例如膀胱癌和前列腺癌，对全球健康构成严重威胁。然而，传统的泌尿类疾病检测方法存在需要大型仪器或有创伤的弊端。在癌症中，疾病早期筛查和发现能够有效地提高后续患者针对性治疗的成功率，以便提高生存率。因此，初期预防和早期筛查仍然是减轻全球癌症日益加重负担的关键战略。

电子鼻技术作为一种新兴的无创诊断工具备受关注，且在不同领域有广泛应用，如食品工业[1] [2] [3]、环境监测[4] [5]和医疗诊断[6] [7] [8]。电子鼻技术的关键在于算法研究，目前主要集中在传统机器学习和基于神经网络的深度学习两个方向。长沙理工大学的喻璐学者采用商用的金属氧化物气体传感器构成传感器阵列，使用支持向量机算法[9]，实现对丙酮和乙醇的快速检测。他们采用了分段拟合的方式提取特征，并取得了令人满意的结果，最终准确率达到了 92.5%。然而，需要指出的是，大部分特征工程都受到人为先验知识的影响，这导致了这些特殊的特征工程仅适用于特定情况，缺乏普遍适用性。Cong Fang 学者采用深度学习算法，具体包括卷积神经网络和递归神经网络，对整体原始数据进行分析，以实现对不同白酒的鉴定[10]。尽管深度学习以其多层神经网络和特殊层对原始数据进行全面分析而闻名，但其对于电子鼻技术的应用存在一些限制。深度学习算法的确无需进行繁琐的特征工程，因为它能够直接从原始数据中学习到隐藏特征。然而，由于其复杂的网络结构，深度学习模型需要大量的样本数据来支持训练，否则可能会产生不确定性，而且计算成本也会非常高，这使得其在商业应用方面不太适用。因此，本文设计一种仅基于常规传感器特征的机器学习算法，对泌尿系统疾病进行筛查的气体识别电子鼻系统

具有重要意义。

对于通过挥发性有机化合物(Volatile Organic Compounds, VOCs)诊断疾病的方法已经在国内外得到了广泛研究，许多疾病的气体标志物已被发现，并部分已用于临床诊断。例如，丙酮被发现与糖尿病[11]、一氧化氮与哮喘[12]、硫化氢与口臭[13]、异戊二烯与心脏病[14]等相关联。本文选取了膀胱癌最大关联性的标志物乙苯[15]与异丙醇[16]，以及前列腺癌最大关联性的标志物甲苯[17]和戊醛[18]作为研究对象。我们采用了 8 种不同的金属氧化物气体传感器对这四种目标气体进行数据采集。具体地，我们通过传感器响应值的导数判断传感器阵列的实时状态，并提取相应时间段的特征。最后，我们采用了传统的机器学习算法，包括 K-最近邻(K-Nearest Neighbor)、支持向量机(Support Vector Machine, SVM)和随机森林(Random Forest, RF)进行建模。为了提高分类模型的预测性能，我们采用了 stacking 方法对前述三种基本分类器进行两两集成。通过比较六种方法的实验结果，我们致力于利用常规的气体传感器特征，寻找出最佳模型。

2. 实验部分

2.1. 传感器阵列设计和选取

半导体(MOS)气体传感器主要以其灵敏度高、成本低、稳定性好等特点而闻名。针对甲苯、乙苯、异丙醇和戊醛等四类挥发性有机化合物(VOCs)，我们选用了 8 种商用 MOS 气体传感器。这些传感器的型号、厂家和主要目标气体如表 1 所示。为了评估传感器阵列的性能，图 1 展示了采用静态配气法获取的数据，显示了各传感器对目标气体的实时响应。静态配气法是一种常用的实验方法，用于精确控制目标气体的浓度和流量，以便研究传感器对特定气体的响应特性。

Table 1. Sensors selected for electronic noses and their corresponding detection gases
表 1. 电子鼻选用的传感器与他们对应的检测气体

序号	传感器型号	主要目标气体
1	KQ-2801 (深圳市博丰盛电子)	甲苯
2	TGS-2600 (日本费加罗)	乙苯
3	TGS-260 2(日本费加罗)	异丙醇
4	MP-135 (郑州炜盛)	乙苯
5	TGS-2603 (日本费加罗)	甲苯
6	MP-503 (郑州炜盛)	戊醛
7	TGS-2620 (日本费加罗)	戊醛
8	WSP-2110 (郑州炜盛)	异丙醇

2.2. 数据的获取

实验采用静态气体制备法，在气测试箱中的蒸发皿上注入目标挥发性有机化合物并加热蒸发，以获得特定浓度的气体环境。VOCs 气体浓度计算方法如下：

$$Q = (V \times C \times M)/(22.4 \times d \times \rho) \times 10^{-9} \times (273 + TR)/(273 + TB)$$

Q 为应取的液体体积(mL)；V 为测试瓶体积(mL)；M 为物质分子量(g)；ρ 为液体的纯度；C 为所要配制气体的浓度(1×10^{-6}, ppm)；d 为液体密度(g/cm^3)；TR 为测试环境温度(℃)；TB 为测试瓶内的温度(℃)。

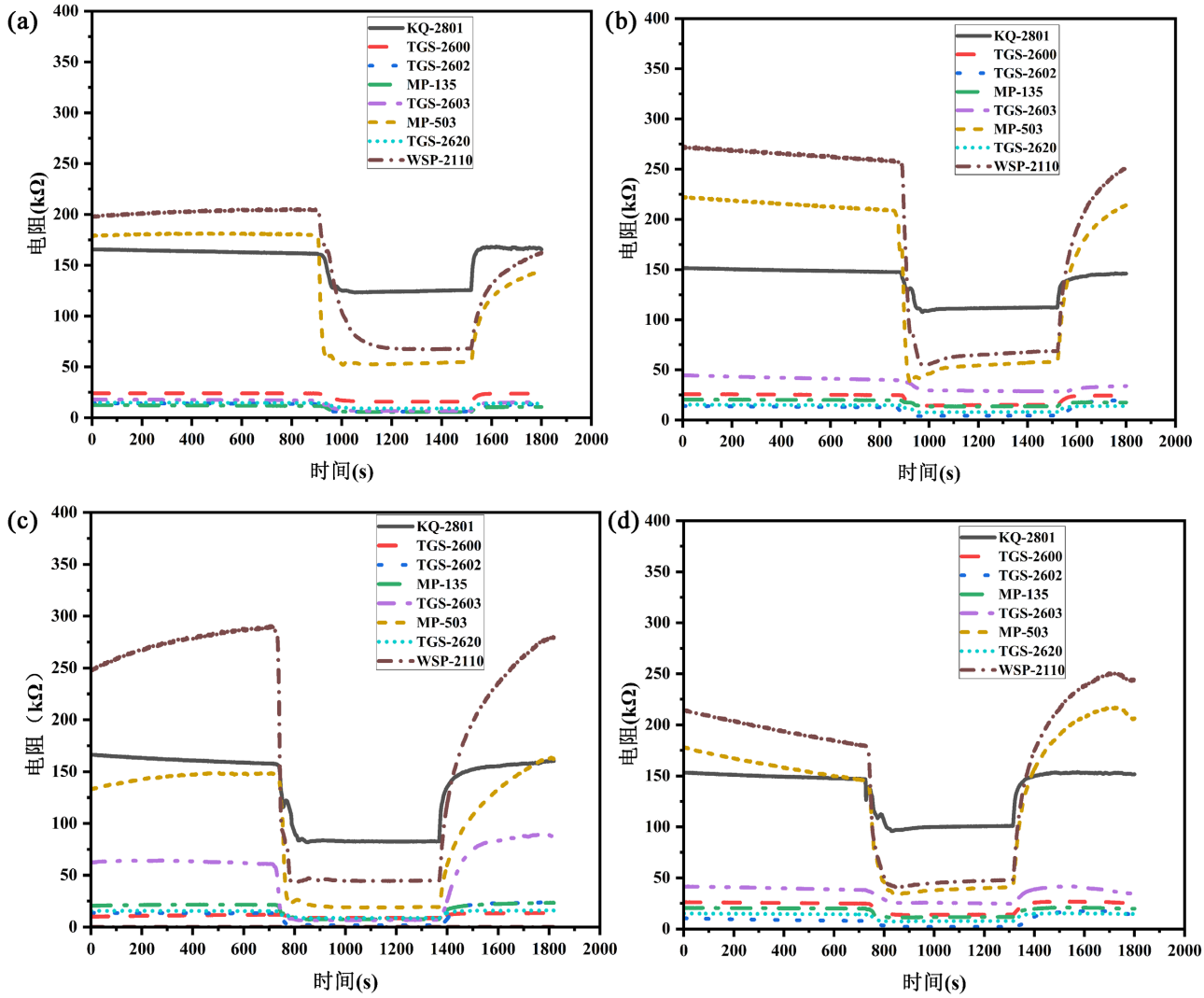

Figure 1. Test diagram of sensor array for target VOCs (a) Sensor array response to 10 ppm isopropanol; (b) Sensor array responds to 10 ppm toluene; (c) Sensor array responds to 10 ppm glutaraldehyde; (d) Sensor array response to 10 ppm ethylbenzene
图 1. 传感器阵列对目标 VOCs 的测试图 (a) 传感器阵列对 10 ppm 异丙醇响应；(b) 传感器阵列对 10 ppm 甲苯响应；(c) 传感器阵列对 10 ppm 戊醛响应；(d) 传感器阵列对 10 ppm 乙苯响应

系统实验测试示意图如图 2 所示，对甲苯、乙苯、异丙醇和戊醛的浓度分别设置为 10、20、30、50、70、90 ppm。在实验过程中，甲苯和戊醛每种浓度重复五次，乙苯和异丙醇每种浓度重复四次，因此总的样本点数量为 108 个。

2.3. 特征工程

为了避免人为提取数据所带来的干扰，我们利用导数判断状态，并自动从原始数据中提取稳定状态时间段的平均电阻值。

具体而言，在图 3 中，注入了 6 次甲苯气体，浓度梯度分别为 10, 20, 30, 50, 70, 90 ppm。通过对导数的分析，我们推导出 6 个时序相邻大于 300 秒的最小导数值，分别为[834, 2037, 3555, 5148, 6909, 8707]。同理，我们还得到了对应的最大导数时间点，分别为[1293, 2594, 4237, 5809, 7691, 9471]。其中最小导数时间点确定了进入响应阶段的时间段，而最大导数时间点确定了进入恢复阶段的时间段。

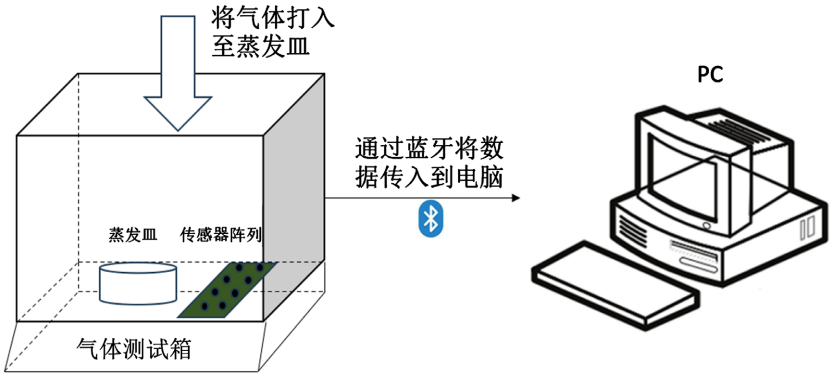

Figure 2. Schematic diagram of gas collection process
图 2. 气体采集流程示意图

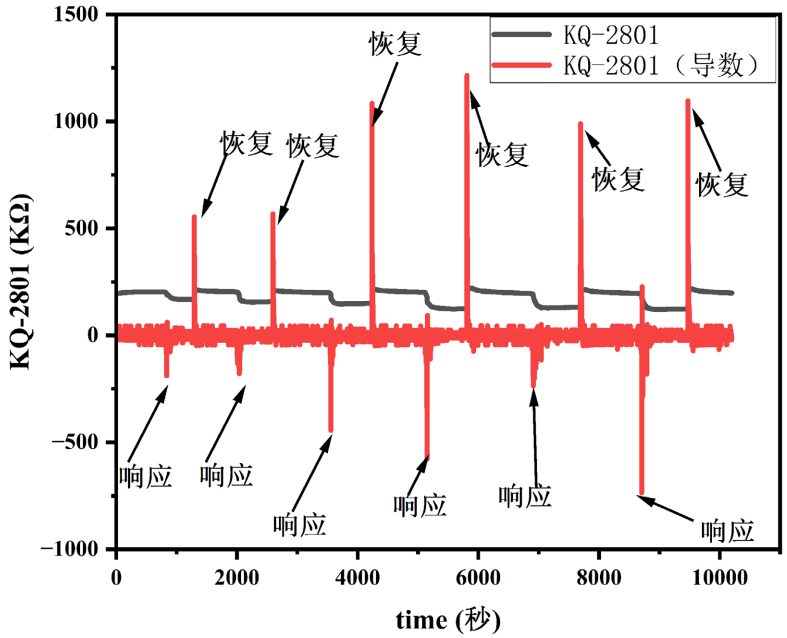

Figure 3. Real time response and derivative diagram of sensor KQ-2801 to toluene
图 3. 传感器 KQ-2801 对甲苯的实时响应与其导数图

在导数确定传感器状态后，我们先对数据标准化，使用 $X_i = \frac{X_i - \mu}{\sigma}$，式中 μ 为所有样本下某一特征的均值，σ为所有样本下某一特征的标准差。随后，采用传感器响应 $\frac{R_g - R_a}{R_a}$ 作为特征，其中 R_a 为基线状态下的平均响应电阻值，R_g 为响应稳定状态下的平均电阻值。这一转换将连续时间段的信号离散为 6 个样本点信号。

2.4. 机器学习模型

本研究为了鉴别泌尿类疾病标志物 VOC：甲苯，乙苯，戊醛与异丙醇，我们先对原始数据进行自动化提取特征，将数据集转换为[108, 8]的矩阵，108 表示样本个数，8 表示各传感器对于样本的特征值。

108 个数据集划分为 33 个样本的测试集，另外的 75 个样本用来作为 5 折交叉验证，60 个样本作为训练集，15 个样本作为验证集。图 4 和图 5 分别是各算法的示意图和数据选择流程图。

2.4.1. PCA

为了实现数据的可视化和数据的降维，首先对原始数据进行了主成分分析(Principal Component Analysis, PCA)，以达到聚类的效果。PCA 是一种采用的降维和数据压缩技术，通过线性变换将原始数据转换为一组新的坐标系，使得在新的坐标系下数据的方差最大化。这些新的坐标，代表着最有效解释数据变异性的方向[19] [20] [21]。

我们选择采用降维的原因是，许多气体传感器存在交叉敏感性，导致传感器阵列的原始数据中存在大量冗余信息。通过 PCA 降维，我们可以更有效地捕获数据中的主要变化，从而减少特征的维度，并保留数据的关键信息。

2.4.2. KNN

K 最近邻方法(K-Nearest Neighbor)是一种基于实例的学习算法，用于解决分类和回归问题。在这种方法中，根据测试集中的自变量观测值与训练集中自变量观测值的距离，选择最近的 k 个点。然后，对这 k 个最近邻点的因变量进行加权平均，以预测测试集的因变量[22]。

2.4.3. SVM

支持向量机(Support Vector Machine, SVM)是一种强大的监督学习算法，用于分类和回归任务。其核心思想是通过找到一个最佳的超平面来对不同类别的数据进行分离[23] [24]。

2.4.4. SVM

随机森林(Random Forest)是一种基于集成学习的强大机器学习算法，它通过组合多个决策树来进行分类或回归任务。其基本原理是通过对训练集的随机采样，以及对特征的随机选择，构建多棵决策树，并且通过投票或平均的方式来获得最终的预测结果[25] [26]。

2.4.5. Stacking 集成

Stacking 集成是一种模型集成技术，它通过将多个基本模型的预测结果作为新的特征，然后再训练一个元模型(Meta Model)来进行最终的预测。这种方法能够有效地提高模型的泛化能力和预测性能[27]。

使用 stacking 集成方法将上面三种模型进行两两集成，将模型 A 和模型 B 的预测结果输入给逻辑回归模型(Logistic Regression, LR)，逻辑回归模型学习如何为 A 和 B 的预测结果分配适当的权重，从而在最终的集成预测中有效地捕捉每个基本模型输出的重要性。

2.5. 评分标准

分类问题是监督学习的核心问题，为了评估分类模型的性能如何，我们使用混淆矩阵(Confusion matrix)来评估性能指标，直观判断模型的性能。在混淆矩阵中，有四个参数，分别是真阳性(true positives; TP)；假阳性(false positives; FP)；假阴性(false negatives; FN)；真阴性(true negatives; TN)。正确地归类为阳类的例子称为真阳性，正确地归为阴类的例子称为真阴性，而被错误归类为阴性的阳性例子称为假阴性，被错误归类为阳性的阴性例子称为假阳性。

混淆矩阵提供了在测试数据上模型对各类别的估计与真实情况，因此由以下指标判断模型分类效果是否成功。以下为指标其计算公式：

准确率(Accuracy)：

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \times 100$$

精确率(Precision)：

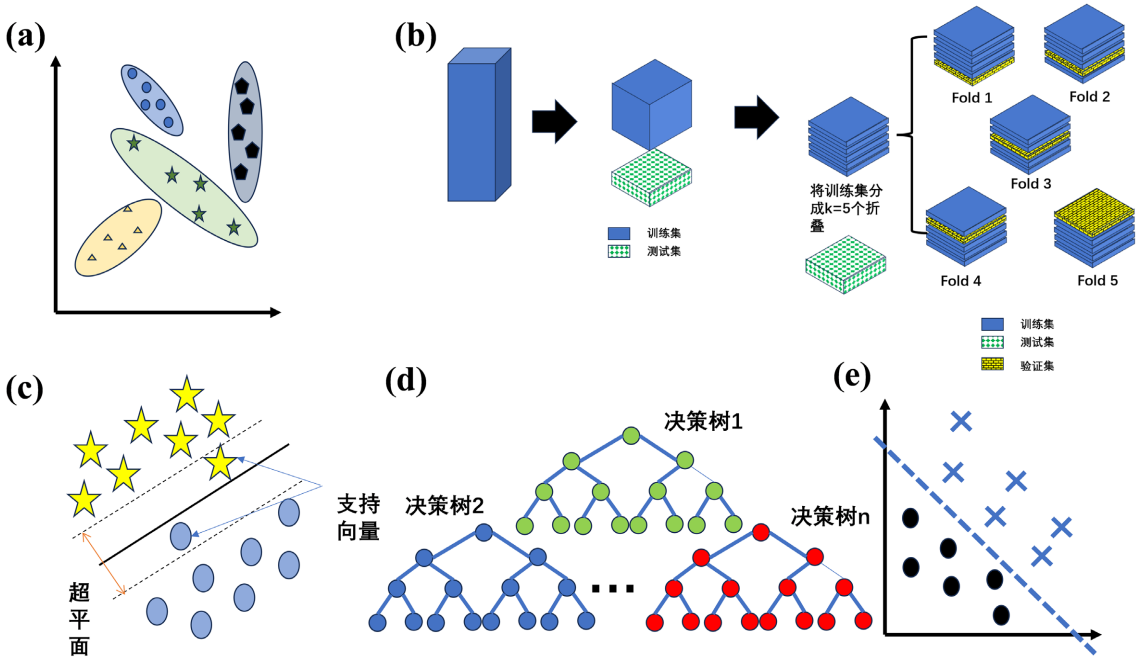

Figure 4. Schematic diagram of the algorithm designed in the article: (a) PCA schematic diagram; (b) Data partitioning; (c) SVM schematic diagram; (d) RF schematic diagram; (e) KNN schematic diagram

图 4. 文章设计算法的示意图：(a) PCA 示意图；(b) 数据的划分；(c) SVM 示意图；(d) RF 示意图；(e) KNN 示意图

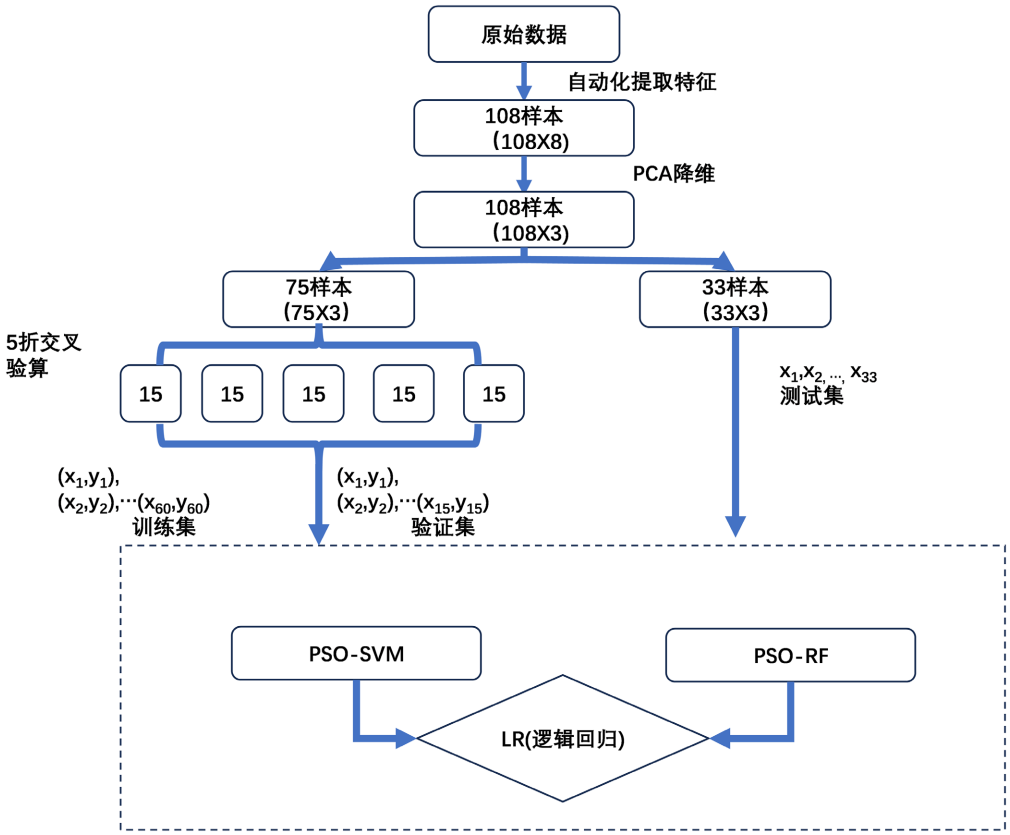

Figure 5. Data selection flowchart (Taking PSO-SVM and PSO-RF Integration as an example)

图 5. 数据选择流程图(以 PSO-SVM 和 PSO-RF 集成为例)

$$\text{Precision} = \frac{\text{TP}}{\text{FP}+\text{TP}} \times 100$$

召回率(Recall)：

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FP}} \times 100$$

F1 值：

$$\text{F1} = \frac{2\text{TP}}{2\text{TP}+\text{FP}+\text{FN}}$$

3. 结果与讨论

本研究采用 Python 语言，利用 Scikit-learn 框架进行实验。实验结果显示，各分类器的准确率如图 6 所示。在单一分类器方面，随机森林表现最佳，其准确率达到了 91%。而在集成算法模型中，由于数据集样本有限，仅考虑了两两集成。其中，SVM 和 RF 的集成效果最显著，准确率达到了 97%。相比之下，SVM 和 KNN 的集成准确率略有下降，仅为 82%。

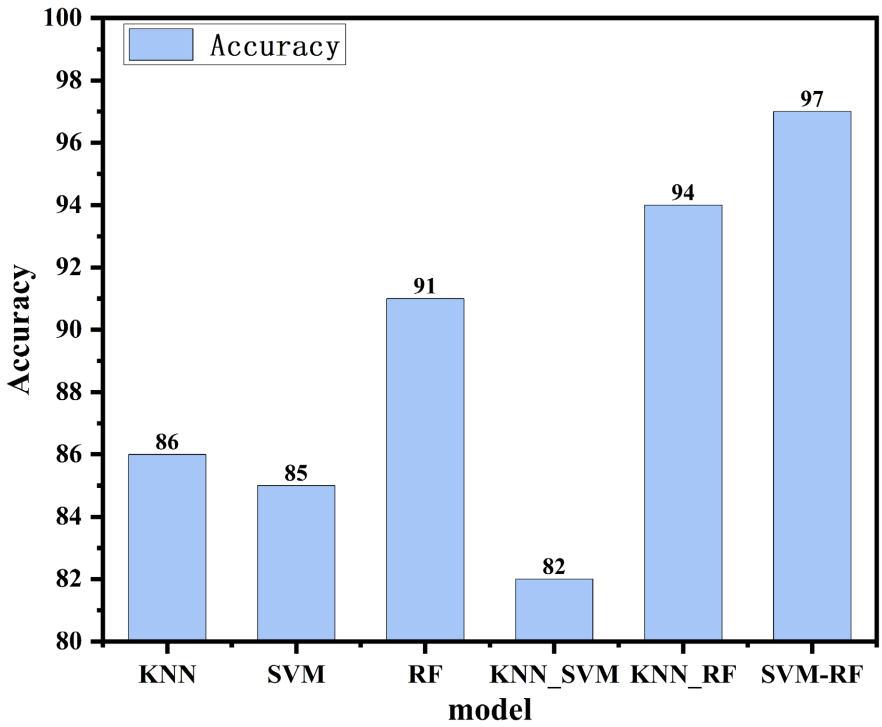

Figure 6. Accuracy of classification models for each algorithm
图 6. 各算法分类模型的准确率

3.1. PCA 聚类结果

图 7 展示了前两个主成分 PC1 和 PC2 组成的二维图，其中前两个主成分的贡献率分别为 73.3%和 15.7%，累计贡献率达到 89%，置信区间为 0.95。PCA 反映了四种气体的聚类情况，异丙醇和戊醛具有较高的识别度，而甲苯与乙苯则纠缠在一起。这可能是因为两者具有相似的分子结构，导致传感器阵列在它们的响应上无法很好地区分开。

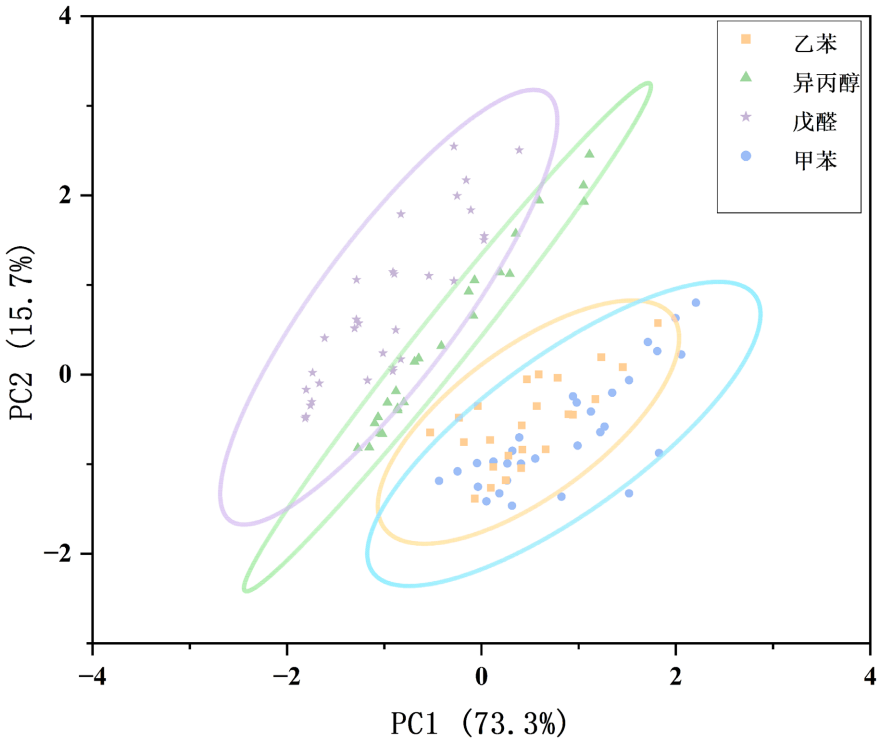

Figure 7. PCA clustering effect diagram for four types of gases
图 7. 四类气体 PCA 聚类效果图

3.2. 单分类器和集成分类器实验结果分析

3.2.1. KNN 的实验结果分析

大多数算法模型容易受其超参数影响，从而限制了模型的性能。对于 KNN 算法而言，由于其超参数 K 为离散的整数类型，因此无需优化算法进行超参数调整。我们直接对其三种距离度量和 K 值范围在[3] [10]内进行多次拟合，以选择最佳的 K 值和距离度量。从图 8 可以看出，当 K = 3 时，欧氏距离取得最佳效果，准确率达到了 88%。在后续的集成算法中，我们确定了 K = 3 和欧氏距离作为最佳超参数，用作 stacking 集成算法的基分类器。

本文的目标是进行四类气体的分类，其中期望输出为：0 → 甲苯；1 → 乙苯；2 → 异丙醇；3 → 戊醛。图 10(c)展示了 KNN 模型在测试集上的混淆矩阵，可以观察到，该模型将 11 个甲苯样本中的 3 个误分类为乙苯；将 6 个乙苯样本中的 1 个误分类为异丙醇。这一结果可能是由于它们都属于苯系物，导致在传感器阵列上的响应相似，从而在 PCA 降维后的特征空间中，甲苯与乙苯的区分度不高。最终，KNN 模型的整体准确率为 88%。然而，针对乙苯的精确度和甲苯的召回率分别为 0.62 和 0.73，如图 11(a)和图 11(c)所示。

3.2.2. SVM 和 RF 的实验结果分析

与 KNN 不同的是，SVM 的超参数是连续的变化，所以需要采用优化算法搜索最佳超参数；而 RF 的超参数虽然是离散的整数型，但是空间变化很大，穷举地选取最佳值耗时耗力，所以仍然需要采用优化算法。因此，我们决定引入粒子群优化算法(Particle Swarm Optimization, PSO)对其超参数进行调优[28]。在粒子群算法中，SVM 的优化主要涉及惩罚系数 C 和核函数常数项系数 gamma；而 RF 的优化则包括决策树数量和决策树深度。这些超参数的范围与作用如表 2 和表 3 所示。选择准确率(Accuracy)作为优化参

数，通过五折交叉验证，计算五次准确率的平均值和标准差，并使用加权求和，得到一个得分系数：0.7*avg~0.3*var (其中 avg 是五次交叉验证准确率的平均，var 是五次交叉验证准确率的方差)。该得分系数用于评估优化结果的质量，以提高模型的稳定性和泛化性。

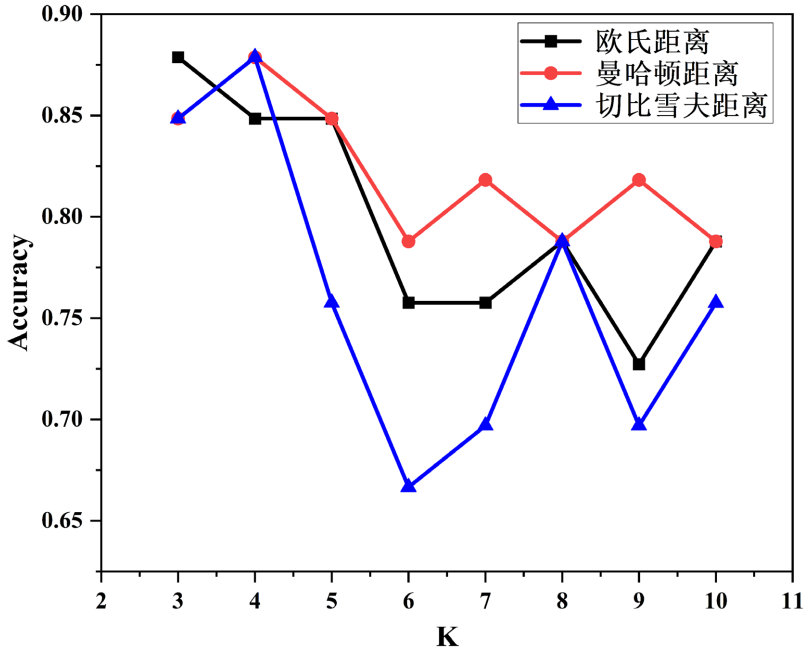

Figure 8. Accuracy of different K values under three distance schemes
图 8. 三种距离方案下，不同 K 值的准确率

Table 2. Superparameter range of PSO optimized SVM
表 2. PSO 优化 SVM 的超参数范围

超参数	搜索范围	作用
惩罚参数(C)	Range (1, 100)	决定拟合效果
核函数参数(gamma)	Range (0.01, 1)	支持向量的个数

Table 3. Superparameter range of PSO optimized RF
表 3. PSO 优化 RF 的超参数范围

超参数	搜索范围	作用
决策树的数量(n_estimators)	Range [10, 200]	数量越多，误差越小，最后会趋于收敛，然而计算量也随之增加
决策树的深度(max_depth)	Range [2, 10]	决定拟合效果

粒子群优化的对比如图 9 所示，其中(a)展示了未优化超参数的 SVM 特征空间示意图，而(b)展示了经过优化后的 SVM 特征空间示意图。通过粒子群优化算法，成功将 SVM 的准确率从 54%提升到 85%。图 10(a)和图 10(b)分别是 SVM 和 RF 混淆矩阵，可观察到，PSO-SVM 算法模型在甲苯与乙苯之间的鉴别方面仍有欠缺，这与 KNN 的结果类似，这可能是因为 SVM 是要在特征空间绘画出最优超平面的原因，而特征空间中，甲苯与乙苯纠缠不清，这才导致了 SVM 对于甲苯与乙苯的误分类情况。然而，在甲苯与乙苯的纠缠问题中，PSO-RF 有着出乎意料的优异性能。

结果而言，超参数优化后的最佳模型，SVM 的准确率达到了 85%，然而针对乙苯的精确度和针对甲

苯的召回率都只有 55%，图 11(a)和图 11(c)所示；RF 的整体准确率为 91%，但是针对异丙醇的精确度仅有 78%，图 11(b)所示。

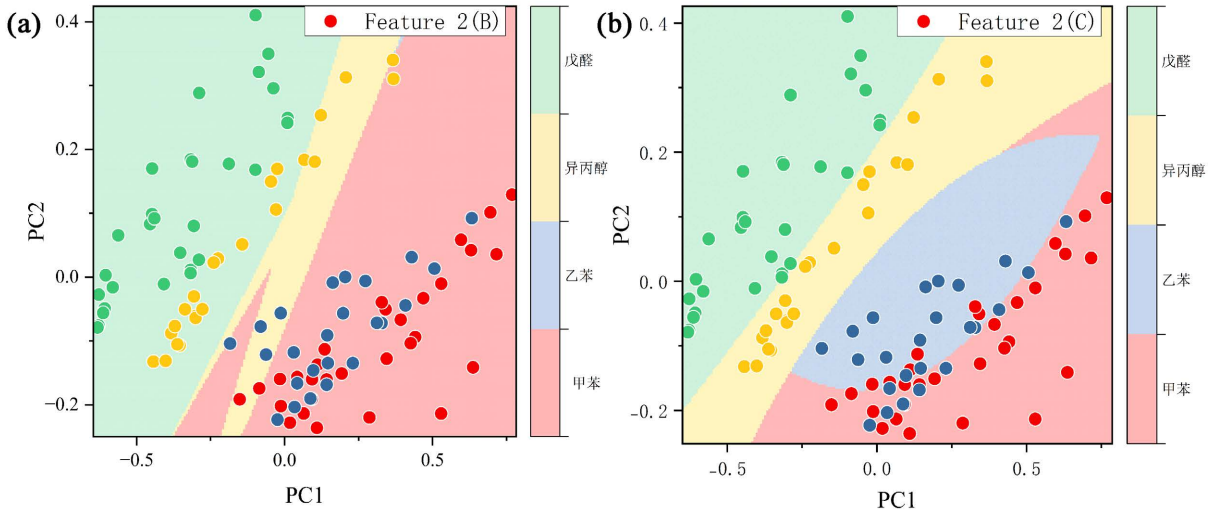

Figure 9. (a) The SVM model with an accuracy of 54%, and its hyperplane is illustrated in the feature space; (b) The SVM model optimized by PSO (with an accuracy of 85%) has a hyperplane in the feature space schematic diagram
图 9. (a) 准确率为 54%的 SVM 模型，其超平面在特征空间示意图；(b) PSO 优化后的 SVM 模型(准确率为 85%)，其超平面在特征空间示意图

3.2.3. 集成方法的实验结果分析

为了进一步优化分类器性能，决定采用 Stacking 方法，将 SVM，KNN 和 RF 模型进行两两集成，以充分发挥它们各自的优势，以提高整体模型的综合性能。

整体实验结果如图 10(d)~(f)所示，SVM 与 RF 的集成效果最好，其次是 KNN 和 RF，两者准确率分别达到了 97%和 94%，相对于基分类器的准确率有明显提升。然而，在 SVM 和 RF 的集成中，模型效果反而不如之前的基分类器。这可能是因为 SVM 和 KNN 都是基于特征空间中的样本点距离来进行分类判断；而 RF 是一种集成算法模型，它的输出是根据多个决策树的预测结果进行综合，其中每棵决策树的训练样本存在差异。因此，只要大多数决策树都能做出正确的判断，即使个别决策树存在误判，也不会对整体模型的性能产生太大影响。

如图 10(d)所示，作为 RF 与 SVM 的集成模型，其完美地继承了两个模型的优势，SVM 在处理乙苯与异丙醇上有着巨大优势，而 RF 能更好地甲苯与乙苯之间的纠缠；KNN 和 RF 同样如此，集成模型继承了基分类器在不同类别判断中的优点，如图 10(f)；然而 SVM 和 KNN 的集成，则是产生了冲突，如图 10(e)所示。

图 11 是六种模型对各个类别预测的评分标准。其中六种模型对戊醛的检测最佳，如图 11(d)所示，这符合原始数据在 PCA 降维后在特征空间上的分布。其中甲苯与乙苯由于其本身都是苯系物的原因，它们在传感器阵列上的响应类似，因此误分类点大部分出自于它们。

4. 结论

本研究使用电子鼻技术，利用常规气体传感器的响应值作为特征，采用 KNN、SVM 和 RF 等三种分类算法，并通过 Stacking 集成模式对它们进行了两两集成。对比了六种模型的性能，结果显示，SVM 和 RF 集成模型表现出色，达到了 97%的最终分类准确率。尤其在甲苯与乙苯等互相纠缠的情况下，集成模

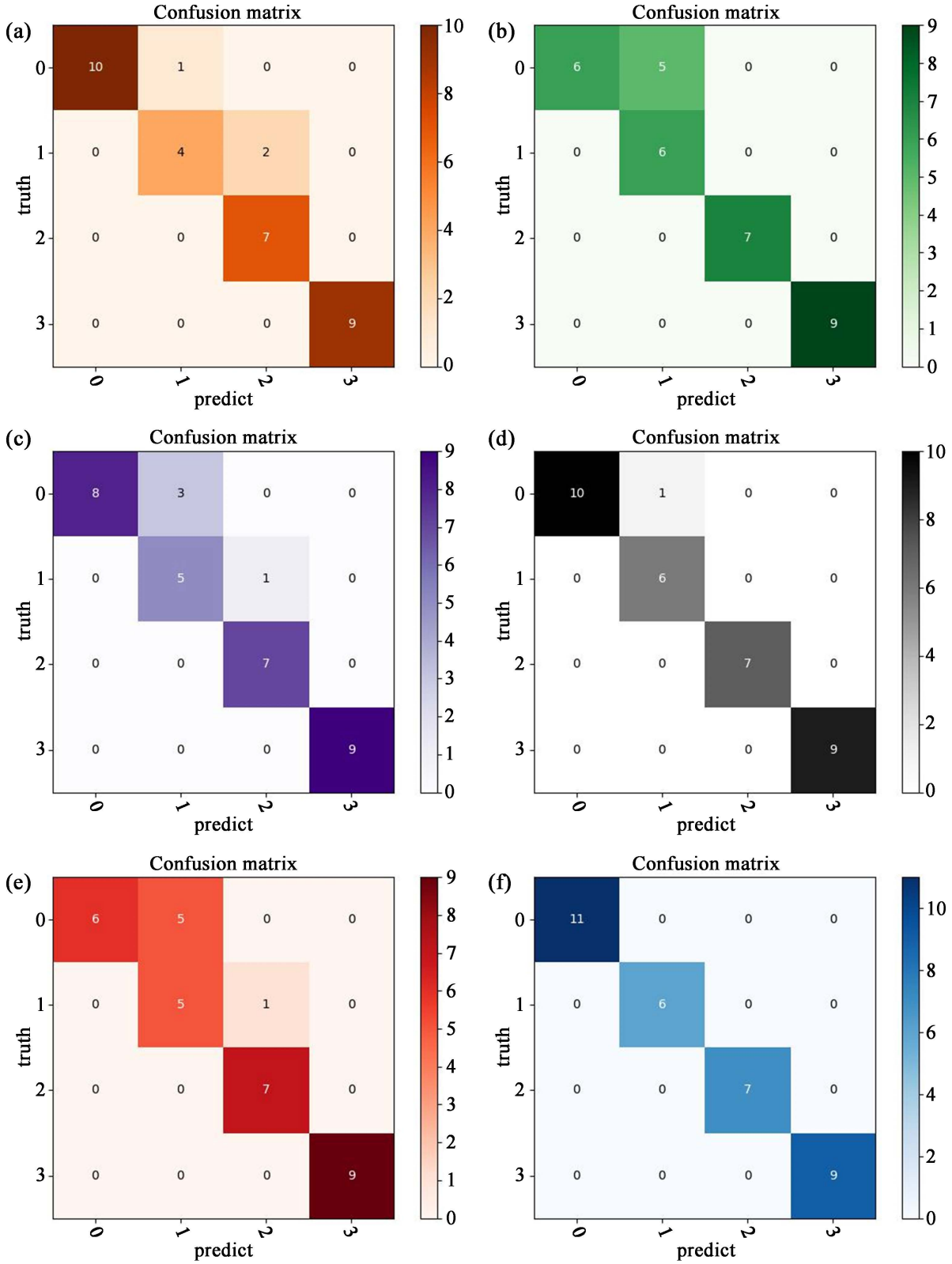

Figure 10. Confusion matrices of each algorithm model for the test set: (a) RF; (b) SVM; (c) KNN; (d) RF-SVM; (e) SVM_KNN; (f) RF_KNN

图 10. 各个算法模型对于测试集的混淆矩阵：(a) RF；(b) SVM；(c) KNN；(d) RF_SVM；(e) SVM_KNN；(f) RF_KNN

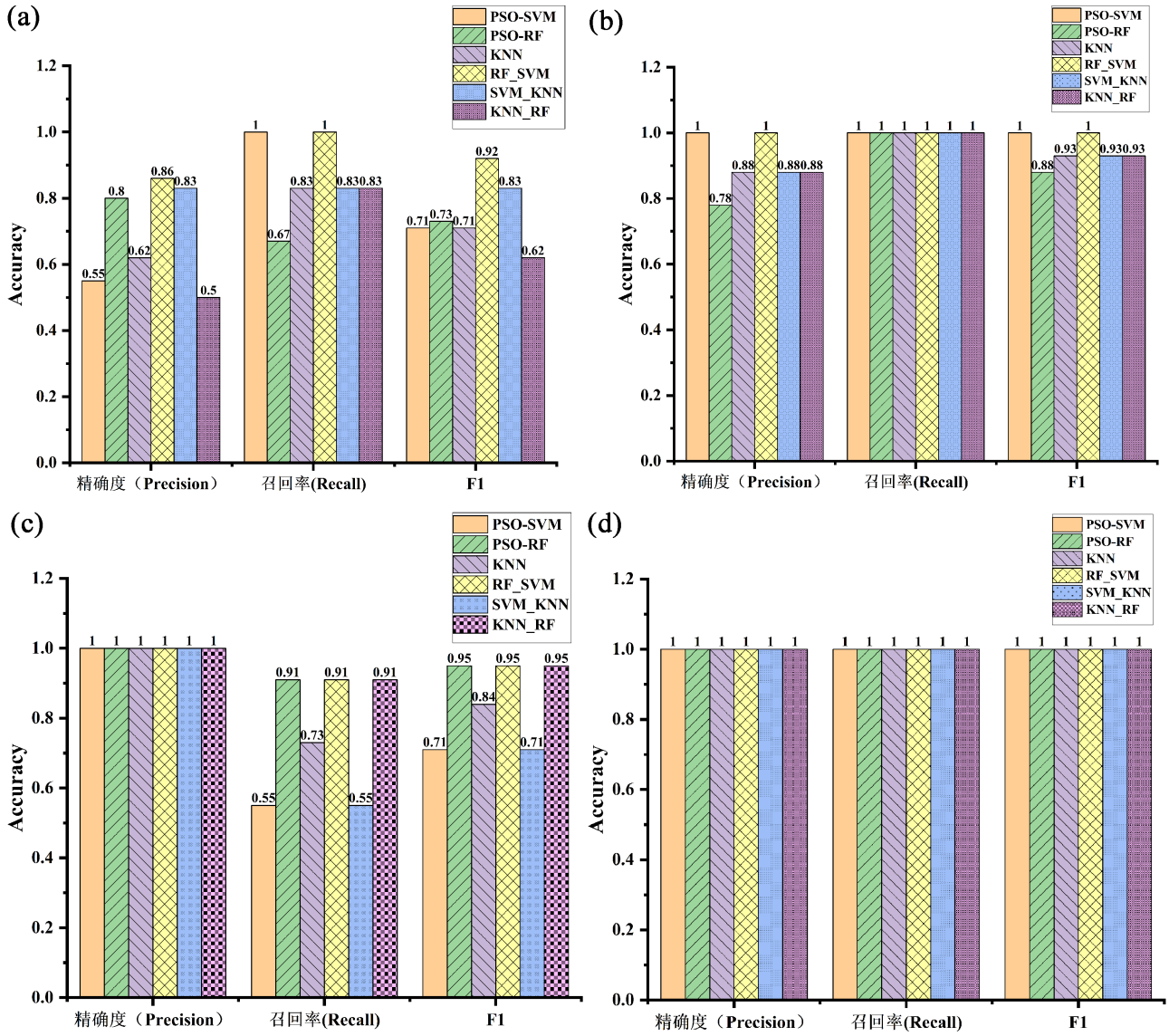

Figure 11. Performance evaluation indicators for four types of gases using six models. (a) Ethylbenzene; (b) Isopropanol; (c) Toluene; (d) Glutaraldehyde

图 11. 六种模型对四类气体的性能评价指标。(a) 乙苯；(b) 异丙醇；(c) 甲苯；(d) 戊醛

型的分类准确率也达到了 94%。这表明，仅使用常规气体传感器特征就能实现较好的单一气体检测。不过，尿液挥发物并非仅限于单一 VOC，因此，后续实验将进一步扩展到多元气体的预测。

基金项目

上海市科委高校能力建设项目(21010502800)。

参考文献

[1] Anwar, H., Anwar, T. and Murtaza, S. (2023) Review on Food Quality Assessment Using Machine Learning and Electronic Nose System. *Biosensors and Bioelectronics*: *X*, **14**, Article ID: 100365. https://doi.org/10.1016/j.biosx.2023.100365

[2] 庞林江，王俊，路兴花，等. 基于电子鼻技术的山核桃陈化指标预测模型研究[J]. 传感技术学报，2019，32(9):

1303-1307.

[3] Anwar, H., Anwar, T. and Murtaza, M.S. (2023) Applications of Electronic Nose and Machine Learning Models in Vegetables Quality Assessment: A Review. 2023 *IEEE International Conference on Emerging Trends in Engineering, Sciences and Technology* (*ICES&T*), Bahawalpur, 9-11 January 2023, 1-5. https://doi.org/10.1109/ICEST56843.2023.10138839

[4] Gardner, J.W., Shin, H.W., Hines, E.L., *et al.* (2000) An Electronic Nose System for Monitoring the Quality of Potable Water. *Sensors and Actuators B*: *Chemical*, **69**, 336-341.

[5] Attallah, O. and Morsi, I. (2022) An Electronic Nose for Identifying Multiple Combustible/Harmful Gases and Their Concentration Levels *via* Artificial Intelligence. *Measurement*, **199**, Article ID: 111458. https://doi.org/10.1016/j.measurement.2022.111458

[6] Wilson, A.D. (2013) Diverse Applications of Electronic-Nose Technologies in Agriculture and Forestry. *Sensors*, **13**, 2295-2348. https://doi.org/10.3390/s130202295

[7] 郭泽尚, 王磊, 常志勇. 电子鼻在肠道疾病诊断中应用的研究进展[J]. 吉林大学学报(医学版), 2022, 46(6): 1332-1337.

[8] Filianoti, A., Costantini, M., Bove, A.M., *et al.* (2022) Volatilome Analysis in Prostate Cancer by Electronic Nose: A Pilot Monocentric Study. *Cancers*, **14**, Article 2927. https://doi.org/10.3390/cancers14122927

[9] 喻璐, 谭志文, 邹望辉. 基于传感器阵列的气体检测与分析系统设计[J]. 电子设计工程, 2022, 30(10): 129-133, 138.

[10] Fang, C., Li, H.Y., Li, L., *et al.* (2022) Smart Electronic Nose Enabled by an All-Feature Olfactory Algorithm. *Advanced Intelligent Systems*, **4**, Article ID: 2200074. https://doi.org/10.1002/aisy.202270032

[11] Righettoni, M., Tricoli, A. and Pratsinis, S.E. (2010) Si: WO_3 Sensors for Highly Selective Detection of Acetone for Easy Diagnosis of Diabetes by Breath Analysis. *Analytical Chemistry*, **82**, 3581-3587. https://doi.org/10.1021/ac902695n

[12] Smith, A.D., Cowan, J.O., Filsell, S., *et al.* (2004) Diagnosing Asthma: Comparisons between Exhaled Nitric Oxide Measurements and Conventional Tests. *American Journal of Respiratory and Critical Care Medicine*, **169**, 473-478. https://doi.org/10.1164/rccm.200310-1376OC

[13] Choi, S.J., Jang, B.H., Lee, S.J., *et al.* (2014) Selective Detection of Acetone and Hydrogen Sulfide for the Diagnosis of Diabetes and Halitosis Using SnO_2 Nanofibers Functionalized with Reduced Graphene Oxide Nanosheets. *ACS Applied Materials & Interfaces*, **6**, 2588-2597. https://doi.org/10.1021/am405088q

[14] Mendis, S., Sobotka, P.A. and Euler, D.E. (1995) Expired Hydrocarbons in Patients with Acute Myocardial Infarction. *Free Radical Research*, **23**, 117-122. https://doi.org/10.3109/10715769509064026

[15] Dragonieri, S., Schot, R., Mertens, B.J.A., *et al.* (2007) An Electronic Nose in the Discrimination of Patients with Asthma and Controls. *Journal of Allergy and Clinical Immunology*, **120**, 856-862. https://doi.org/10.1016/j.jaci.2007.05.043

[16] Zhu, S., Corsetti, S., Wang, Q., *et al.* (2019) Optical Sensory Arrays for the Detection of Urinary Bladder Cancer-Related Volatile Organic Compounds. *Journal of Biophotonics*, **12**, e201800165. https://doi.org/10.1002/jbio.201800165

[17] Jian, Y., Zhang, N., Liu, T., *et al.* (2022) Artificially Intelligent Olfaction for Fast and Noninvasive Diagnosis of Bladder Cancer from Urine. *ACS Sensors*, **7**, 1720-1731. https://doi.org/10.1021/acssensors.2c00467

[18] Tyagi, H., Daulton, E., Bannaga, A.S., *et al.* (2021) Urinary Volatiles and Chemical Characterisation for the Non-Invasive Detection of Prostate and Bladder Cancers. *Biosensors*, **11**, Article 437. https://doi.org/10.3390/bios11110437

[19] Gao, Q., Su, X., Annabi, M.H., *et al.* (2019) Application of Urinary Volatile Organic Compounds (VOCs) for the Diagnosis of Prostate Cancer. *Clinical Genitourinary Cancer*, **17**, 183-190. https://doi.org/10.1016/j.clgc.2019.02.003

[20] Karamizadeh, S., Abdullah, S.M., Manaf, A.A., *et al.* (2013) An Overview of Principal Component Analysis. *Journal of Signal and Information Processing*, **4**, 173-175. https://doi.org/10.4236/jsip.2013.43B031

[21] Abdi, H. and Williams, L.J. (2010) Principal Component Analysis. *WIREs Computational Statistics*, **2**, 433-459. https://doi.org/10.1002/wics.101

[22] Kumar, N.S. and Arun, M. (2017) Genetic Algorithm-Based Feature Selection for Classification of Land Cover Changes Using Combined LANDSAT and ENVISAT Images. *International Journal of Bio-Inspired Computation*, **10**, 172-187. https://doi.org/10.1504/IJBIC.2017.086700

[23] Pardo, M. and Sberveglieri, G. (2005) Classification of Electronic Nose Data with Support Vector Machines. *Sensors and Actuators B*: *Chemical*, **107**, 730-737. https://doi.org/10.1016/j.snb.2004.12.005

[24] Sinju, K.R., Bhangare, B.K., Debnath, A.K. and Ramgir, N.S. (2023) Quick Classification and Prediction of CO_2, NH_3, H_2S, and NO_2 Gases from Their Mixture Using a ZnO Nanowire-Based Electronic Nose. *Journal of Electronic Materials*, **52**, 4686-4698. https://doi.org/10.1007/s11664-023-10419-5

[25] Cutler, A., Cutler, D.R. and Stevens, J.R. (2012) Random Forests. In: Zhang, C. and Ma, Y., Eds., *Ensemble Machine Learning*, Springer, New York, 157-175. https://doi.org/10.1007/978-1-4419-9326-7_5

[26] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. https://doi.org/10.1023/A:1010933404324

[27] Zhang, H., Li, J.L., Liu, X.M., *et al.* (2021) Multi-Dimensional Feature Fusion and Stacking Ensemble Mechanism for Network Intrusion Detection. *Future Generation Computer Systems*, **122**, 130-143. https://doi.org/10.1016/j.future.2021.03.024

[28] Poli, R., Kennedy, J. and Blackwell, T. (2007) Particle Swarm Optimization: An Overview. *Swarm Intelligence*, **1**, 33-57. https://doi.org/10.1007/s11721-007-0002-0