doi：10.12452/j.fxcsxb.24032201

基于机器学习的中药制剂中间体微生物限度快速预判研究

赵堉文[1,2,3]，李芷瑶[1,2]，刘艺丹[1,2]，李　正[1,2,4]，王海霞[1,2,4*]

（1. 天津中医药大学　中药制药工程学院，天津　301617；2. 天津市中药智能制药与绿色制药重点实验室，天津　301617；3. 中药制药过程控制与智能制造技术全国重点实验室，江苏康缘药业股份有限公司，江苏　连云港　222001；4. 现代中医药海河实验室，天津　301617）

摘　要：通过比较随机森林(RF)、支持向量机(SVM)、主成分分析-支持向量机(PCA-SVM)与卷积神经网络(CNN)的模型性能，获得了基于表面增强拉曼光谱(SERS)技术的微生物限度快速预判最优模型，为中药制剂中间体的微生物限度快速预判提供了新方法。首先合成Au@Ag@SiO_2复合纳米材料作为SERS增强基底，随后使用双层膜过滤法制备中药制剂中间体待测样本，并对样本抗菌活性进行考察。最后采集30批中药制剂中间体样本的SERS光谱，并分别建立RF、SVM、PCA-SVM与基于ResNet架构的CNN快速预判模型。结果表明，所建立的CNN模型的准确度、精确度、召回率均为100.0%，F1分数为1.0，受试者操作特征曲线(ROC)显示CNN模型对中药制剂中间体需氧菌总数(TAMC)、霉菌和酵母菌总数(TYMC)的快速预判能力均高于其他3种算法，能对待测样品微生物限度进行有效预判，对不合格样品进行有效风险预警，从而提高对中药生产过程中间体微生物的质量控制水平。

关键词：表面增强拉曼散射(SERS)；机器学习；微生物限度检测；快速预判

中图分类号：O657.3；TB9　　**文献标识码：**A　　**文章编号：**1004-4957(2024)11-1725-10

Research on Rapid Prediction of Microbial Limit of Intermediates in Traditional Chinese Medicine Formulations Based on Machine Learning

ZHAO Yu-wen[1,2,3]，LI Zhi-yao[1,2]，LIU Yi-dan[1,2]，LI Zheng[1,2,4]，WANG Hai-xia[1,2,4*]

（1. College of Pharmaceutical Engineering of Traditional Chinese Medicine，Tianjin University of Traditional Chinese Medicine，Tianjin　301617，China；2. Tianjin Key Laboratory of Intelligent and Green Pharmaceuticals for Traditional Chinese Medicine，Tianjin　301617，China；3. State Key Laboratory on Technologies for Chinese Medicine Pharmaceutical Process，Control and Intelligent Manufacture，Jiangsu Kanion Pharmaceutical Co.，Ltd，Lianyungang　222001，China；4. HaiHe Laboratory of Modern Chinese Medicine，Tianjin　301617，China）

Abstract：By comparing the model performance of random forest(RF)，support vector machine(SVM)，principal component analysis support vector machine(PCA-SVM)，and convolutional neural network(CNN)，the optimal model for rapid microbial limit prediction based on surface enhanced Raman spectroscopy(SERS) technology was obtained，providing a new method for rapid microbial limit prediction of intermediates in traditional chinese medicine (TCM) formulations. First，Au@Ag@SiO_2 composite nanomaterials were synthesized as SERS reinforcement substrates. Secondly，a double-layer membrane filtration method was used to prepare the intermediate test sample of TCM preparations，and the antibacterial activity of the sample was investigated. Finally，SERS spectra of 30 batches of intermediate samples from TCM preparations were collected，and RF，SVM，PCA-SVM，and CNN fast prediction models based on ResNet architecture were established，respectively. The results showed that the accuracy，precision，and recall of the established CNN model were all 100.0%，with an F1 score of 1.0. The receiver operating characteristic(ROC) curve showed that the CNN model had higher rapid prediction ability for the total aerobic bacterial count(TAMC)，total mold and yeast count(TYMC) of TCM intermediates than the other three algorithms. It can effec-

收稿日期：2024-03-22；**修回日期：**2024-05-11
基金项目：国家自然科学基金项目(82003944)；2023年天津市高等学校研究生教育改革研究计划项目[TJYG110]
*通讯作者：王海霞，博士，副研究员，研究方向：微生物快速检测，E-mail：whxtcm@tjutcm.edu.cn

tively predict the microbial limit of the test sample, provide effective risk warning for unqualified samples, and improve the quality control level of intermediate microorganisms in the production process of TCM.

Key words: surface enhanced Raman scattering(SERS); machine learning; microbial limit testing; rapid prediction

中药制剂中间体的质量与饮片、提取物、制剂终产品等密切相关，充分了解中间体的质量概貌，对确保最终中药制剂产品的安全、有效、质优有重要作用。中药制剂中间体质量监控包括物理化学分析、微生物限度检测、重金属和有害物质检测以及稳定性研究等方面，其中微生物负载量是一项重要的监测指标，与最终产品的安全性与有效性密切相关。2020版《中国药典》规定的微生物限度检查项目包括需氧菌总数(TAMC)、霉菌和酵母菌总数(TYMC)、耐胆盐革兰阴性菌、大肠埃希菌和沙门菌[1]，用以考察药品整体微生物负载水平，以避免药品受到微生物污染，确保药品的有效性与安全性。

目前，传统微生物培养计数法仍广泛应用，但其往往需要3~7天甚至更长的检测时间，免疫、分子生物学技术存在检测成本高和对相似菌种判别分析难度大的缺点[2-3]。表面增强拉曼散射(SERS)技术作为一种指纹光谱技术，可获得生物体中蛋白质、多糖、脂质和碳水化合物等成分信息，近年来已成为生物学、药学检测的有效工具[4]。这是因为构成物质的分子振动具有唯一性，故不同物质散射光的频率也各不相同，因此拉曼光谱可以作为一种指纹光谱用于表征不同物质的分子结构。构成微生物细胞膜、细胞质和细胞核的成分可以在SERS光谱上表现出独特的谱峰信息[5-7]，从而形成“微生物指纹图谱”，这使得SERS技术成为微生物检测的热点技术。目前，SERS技术不仅可以对空白及简单基质中的微生物样品进行检测[8]，还可对血液[9]、食品[10]、中药样品[11]中的微生物进行有效检测。但目前报道的中药基质样品多为提取液，对中药固体基质样品中微生物的检测鲜有涉及。这是因为中药固体样品颜色往往过深，且成分复杂，为微生物光谱检测带来困难。因此需要进一步提高对复杂基质样品中微生物SERS光谱数据的分析精度和准确度。以机器学习为代表的人工智能技术飞速发展，助力从光谱数据中探究新模式、新规律，能大大提高分析结果的可信度与准确度，为微生物的复杂SERS数据分析提供了新方法[12]。常见的用于分类的机器学习算法有：随机森林(RF)[13]、支持向量机(SVM)[14]、卷积神经网络(CNN)[15]等。RF作为符号主义学习的代表性方法，融合了决策树和投票策略，具有良好的预测准确度，对异常值具有很好的容忍度，并且不易出现过拟合。SVM是统计学习的代表性算法，专门针对二分类任务设计，核函数直接决定了SVM与核方法的最终性能，具有泛化能力强、超参数较少、适用于处理小样本数据和高维数据的特点。主成分分析(PCA)-SVM则是在进行数据分析之前，先使用PCA对数据降维，提取原始光谱中的特征向量代替原始光谱数据作为模型的输入变量，以降低模型的复杂性。此外，随着计算机算力的增强，掀起了连接主义学习的热潮，以“深度学习”为代表的复杂模型开始受到关注。其中，CNN通过改变网络结构来缓解由于网络中隐含层数量的增加所带来的梯度弥散问题与过拟合情况，成为典型的深度学习模型。

本研究以中药制剂中间体为检测对象，对其微生物负载量进行检测。共收集30个不同批次的中药制剂中间体样本，首先通过联合使用稀释法和薄膜过滤法去除样本抗菌活性的干扰，依据2020版《中国药典》四部通则“1108 中药饮片微生物限度检查法”，对样本进行TAMC与TYMC检测，精确判定样本中不同类别微生物的污染情况。其中TAMC可接受的最大菌数为2×10^3 CFU/g，TYMC可接受的最大菌数为2×10^2 CFU/g，据此精确划定微生物限度合格限，对符合微生物限度检测的样本赋予“合格”标签，对不符合的样本赋予“不合格”标签。同时分别采集“合格”与“不合格”样本的SERS光谱，构建样本数据集。通过评估RF、SVM、PCA-SVM和CNN等机器学习算法建立的快速预判模型的性能，选择预测性能好、泛化能力强的模型用于中药制剂中间体TAMC与TYMC值的快速预判。

1 实验部分

1.1 试剂与仪器

氯金酸三水合物($HAuCl_4\cdot3H_2O$)、柠檬酸钠三水合物($C_6H_5Na_3O_7\cdot3H_2O$)、4-巯基苯基硼酸(4-MPBA)、硅酸钠溶液、3-氨基丙基-三乙氧基硅烷(APTES)和抗坏血酸购自 Sigma-Aldrich(上海)。磷酸

盐缓冲溶液（PBS，pH 7.4）购自北京索莱宝科技有限公司；硝酸银（$AgNO_3$，纯度99.9%）购自天津市北方天医化学试剂厂；盐酸（HCl）购自天津大茂化学试剂厂（天津）；Luria－Bertani（LB）肉汤、营养琼脂和胰蛋白胨大豆琼脂购自上海生工生物工程技术服务有限公司（上海）。实验用水均为超纯水，由Millipore 超纯水机（德国默克公司）制得。所有SERS光谱均通过（DXR2，Thermo Fisher，USA）显微拉曼光谱仪采集。

1.2　实验方法

1.2.1　中药制剂中间体抗菌活性去除与微生物限度检查　实验用中药制剂中间体样本为由川芎醇提物、当归醇提物、红花水提物等混合组成的深棕色干燥粉末。有研究报道其部分成分具有抗菌活性[16]，应首先去除抗菌活性以避免对微生物限度检查结果造成干扰。联合使用稀释法与薄膜过滤法去除样本抗菌活性。具体操作为：将样品稀释100倍后过500目滤布，通过在供试品对照组中添加典型微生物，模拟微生物污染过程，以回收率作为指标评价方法适用性。当试验组病原菌回收率为50%~200%时，表明可用常规方法进行微生物限度检测；若试验组病原菌回收率小于50%，则需要采取适宜方法消除供试品的抗菌活性。本研究以枯草芽孢杆菌与大肠埃希菌为典型微生物，回收率计算公式如下：

$$\text{试验组病原菌回收率}=\frac{\text{试验组平均菌落数}-\text{供试品对照组平均菌落数}}{\text{菌液对照组平均菌落数}}$$

去除抗菌活性后的样本参照2020版《中国药典》四部“1105非无菌产品微生物限度检查法”，使用平皿法对中药制剂中间体粉末进行TAMC和TYMC限度检查。

1.2.2　SERS基底的合成与待测样品制备　通过Lee-Meisel方法[17]制备金纳米粒子（AuNPs），随后以柠檬酸钠作为稳定剂，用抗坏血酸在AuNPs表面还原$AgNO_3$制得金包银（Au@Ag）纳米粒子。之后在Au@Ag外包覆SiO_2外壳，制得Au@Ag@SiO_2纳米材料。具体步骤如下：首先，在圆底烧瓶中加入20 mL超纯水，依次加入0.7 mL AuNPs，1 mL（1%，质量分数）的柠檬酸钠溶液和3 mL的10 mmol/L抗坏血酸溶液，于室温下搅拌5 min。之后用微量注射器以0.15 mL/min的速度匀速滴加0.7 mL浓度为10 mmol/L的$AgNO_3$溶液，溶液由淡紫色转为淡金黄色即得Au@Ag。在100 mL圆底烧瓶中加入30 mL Au@Ag与1 mmol/L APTES溶液0.4 mL，室温下搅拌15 min后加入3.2 mL硅酸钠溶液（pH 10.3，0.54%，质量分数），继续搅拌3 min，最后将反应温度升至90 ℃，保持45 min。将获得的热反应物在冰水浴中冷却以停止反应，所得的Au@Ag@SiO_2纳米溶液在5 500 r/min转速下离心15 min，使用超纯水洗涤两次后，再次离心，弃去上清液即得到干净的Au@Ag@SiO_2纳米粒子，于4 ℃储存，并以其作为SERS基底材料。

取适量待测中药制剂中间体样品粉末，使用无菌1×PBS配制成1∶100供试液。以双层膜过滤方法对供试液进行处理，截留供试液中的微生物于滤纸上用于检测。首先，使用500目滤布过滤供试液，除去供试液中的难溶或不溶颗粒，收集滤液，随后使用孔径为0.45 μm的滤膜将病原菌截留在滤纸表面，最后将1 mL的SERS基底溶液注入过滤装置中，与截留在滤纸表面的病原菌充分作用5 min后，去除未结合的纳米溶液，取出滤纸，采集SERS光谱。

1.2.3　光谱数据采集　所有SERS光谱均由Thermo Fisher DXR2显微拉曼光谱仪测得，使用785 nm激光器，在10×物镜下进行mapping模式采集。具体采集参数如下：激光能量30 mW，25 μm狭缝，曝光次数：4；曝光时间：5 s；扫描次数：16，选定200 μm × 100 μm的区域，以20 μm为步长采集光谱。

依据2020版《中国药典》微生物限度标准对TAMC和TYMC的限度要求，将样本的微生物限度检查结果划分为“合格”与“不合格”。其中，参与建模的合格样本和不合格样本均为15批，每批次样本均按照“1.2.2”流程制备4个滤纸片，设置显微拉曼光谱仪采样步长，在每个滤纸片上均匀采集50条光谱，即每个批次获得200条光谱。共采集合格样本光谱3 000条，不合格样本光谱3 000条，总计6 000条。

1.2.4　判别模型的建立与评价　使用Scikit-learn框架的train_test_split函数打乱所有原始数据顺序，以随机抽样方法，按照7∶3比例将数据集划分为校正集与测试集。其中，校正集包含“合格”样本与“不合格”样本SERS光谱各2 100条，测试集包含“合格”样本与“不合格”样本SERS光谱各900条。以相同的校正集分别建立RF、SVM、PCA-SVM和CNN模型，RF与SVM模型基于Scikit-learn框架构建，CNN模型基于Pytorch 2.0深度学习框架搭建，模型训练使用Tesla V100 GPU（美国NVIDIA公司）

完成。

使用网格搜索法与交叉验证对RF、SVM以及PCA-SVM模型的超参数进行优化，CNN模型超参数的优化以Loss和准确度为指标。以准确度、精确度、召回率、F1分数、受试者操作特征曲线(ROC)以及AUC(Area under curve)值为评价指标对各模型用于测试集TAMC和TYMC微生物限度值的快速预判模型性能进行评估，计算公式如下：

$$\text{准确度} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$\text{精确度} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{召回率} = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{4}$$

其中，TP、FP、TN、FN分别为真正例、假正例、真负例、假负例。

2 结果与讨论

2.1 抗菌活性去除评价

依据2020版《中国药典》规定方法，使用两个不同批次样品进行回收率实验，枯草芽孢杆菌的回收率分别为82%和98%；大肠埃希菌的回收率分别为80%和87%。结果表明，联合使用稀释法与薄膜过滤法后试验组病原菌的回收率均高于80%，表明该条件下的供试液无抑菌作用或抑菌作用很弱，可采用常规方法进行病原菌限度检测。

2.2 样本微生物限度检测结果

按照2020版《中国药典》四部“1108 中药饮片微生物限度检查法”中的限度规定，根据样本TAMC和TYMC的检测结果，将样本划分为“合格”与“不合格”类别。其中，TAMC与TYMC均符合限度标准即为“合格”，TAMC或TYMC指标有一项不符合限度标准即为“不合格”。不合格样本存在3种情况：仅TAMC值不符合限度标准、仅TYMC值不符合限度标准、TAMC与TYMC值均不符合限度标准。

为使所建立的模型具备较强的“泛化”能力，更好地适用于整个样本空间，训练集样本应尽量反映出样本空间的特性，因此，使用表1中的30批样本采集SERS光谱构建数据集，30批样本中包括“合格”与“不合格”样本各15批，“不合格”样本中包括TYMC值不符合限度标准的样本(序号16~20)、TAMC值不符合限度标准的样本(序号21~25)、TAMC值与TYMC值均不符合限度标准样本(序号26~30)各5批。

表1 基于平板计数法的样本微生物限度检测结果

Table 1 Microbial limit test results of samples

No.	Batch number	TAMC/(CFU·g^{-1})	TYMC/(CFU·g^{-1})	Result
1	T220420	0	0	合格
2	T220421	0	0.33×10^{2}	
3	T220502	0	0	
4	T220503	0	0	
5	T220504	0	0	
6	T220505	0	0	
7	T220509	0.33×10^{2}	0	
8	T220612	0	0.33×10^{2}	
9	T220701	0	0.67×10^{2}	
10	T220702	0	0.33×10^{2}	
11	T220703	0	0	
12	T220704	0.33×10^{2}	0	
13	T220802	0.33×10^{2}	0.33×10^{2}	
14	T220803	0	0	
15	T220804	0	0	
16	T220417	0	3.50×10^{2}	TAMC合格
17	T220418	0	2.67×10^{2}	TYMC不合格
18	T220419	0.67×10^{2}	2.00×10^{2}	

（续表1）

No.	Batch number	TAMC/(CFU·g^{-1})	TYMC/(CFU·g^{-1})	Result
19	T220523	2.67×10^2	4.3×10^3	
20	T220705	0	2.67×10^2	
21	T220801	2.67×10^4	0.33×10^1	TAMC不合格
22	T220805	3.72×10^4	0	TYMC合格
23	T220603	1.3×10^4	0	
24	T220807	6.5×10^4	0.67×10^1	
25	T220806	7.2×10^4	0	
26	T220510	7.2×10^4	2.14×10^2	TAMC不合格
27	T220610	6.34×10^4	3.3×10^2	TYMC不合格
28	T220524	3.7×10^5	1.8×10^3	
29	T220703	4.49×10^4	1.1×10^3	
30	T220803	5.9×10^4	4.3×10^2	

2.3　样本SERS光谱分析

图1为各类“合格”与“不合格”样本的SERS光谱，不同颜色代表同类样本中不同批次的样本数据。可以看出“合格”样本与“不合格”样本之间无法直接体现显著差异用于微生物限度值快速预判，因此采用机器学习算法对SERS光谱数据进行分类处理。

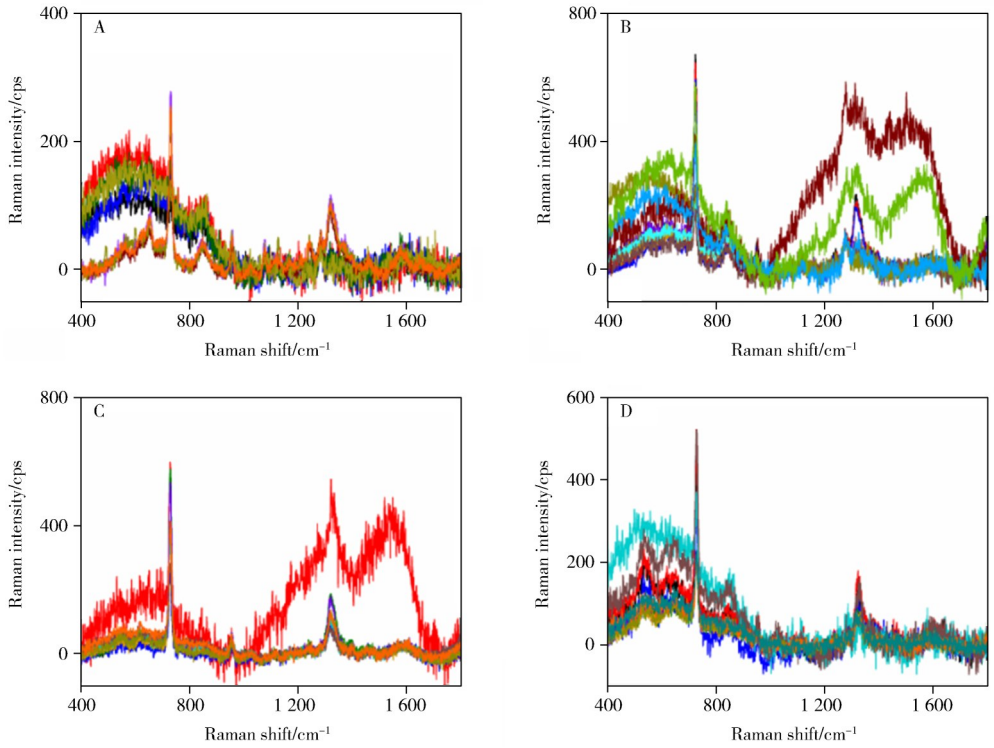

图1　“合格”样本与“不合格”样本的SERS光谱

Fig. 1　SERS spectra of “pass” and “fail” samples

A：pass，B：TAMC pass TYMC fail，C：TAMC fail TYMC pass，D：TAMC fail and TYMC fail

2.4　不同算法的快速预判模型性能

2.4.1　基于RF的快速预判模型　网格搜索法是穷举所设定超参数的所有情况下的模型，同时对各个模型的交叉验证结果进行排序。在RF的模型训练过程中，使用交叉验证和网格搜索法对参数进行优化，包括n_estimators、criterion和max_depth。其中，n_estimators为弱学习器的最大迭代次数；criterion是划分决策树时对特征的评价标准，通常默认为Gini指数；max_depth表示决策树的最大深度，其超参数优化区间为n_estimators={100，200，300}、max_depth={2，10，20}、criterion=“Gini”。结果显示，当参数n_estimators=200，max_depth=20时，模型的准确度最高且较为稳定，校正集样本经十折交叉验证后，准确度为99.67%，精准度为99.34%，召回率为100.0%，F1分数为99.67%，存在过拟合现象。

如图2所示，RF模型在外部测试集的准确度为70.67%，精确度为100.0%，召回率为41.33%，F1

分数为58.49%。不存在将“不合格”样本误判为“合格”样本的假阳性情况，但将“合格”样本误判为“不合格”样本的假阴性情况较为严重。

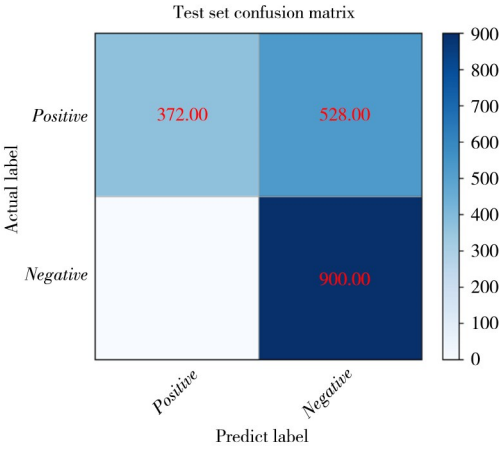

图2 RF模型外部测试集的混淆矩阵

Fig. 2 Confusion matrix for external test set of RF model

2.4.2 基于SVM的快速预判模型 在进行SVM建模分析时，考察了惩罚系数C与核函数参数设置对模型效果的影响。C即对误差的宽容程度，C越高，说明越不能容忍出现误差，容易过拟合；反之C越小，容易欠拟合。C通常在一个对数尺度上进行选择，设置范围为10^0~10^2，以便在参数调整中进行合理搜索。SVM模型中常用的核函数有线性(Linear)、径向基函数(RBF)、sigmoid和多项式(POLY)。这些核函数可以将线性不可分的数据重新映射到线性可分的高维空间，进而将SVM的建模问题转换为通过核函数对原始数据的映射问题。本文在进行SVM建模分析时，考察了这4个核函数的建模分析效果。如表2所示，以F1分数降序排列，校正集中SVM最优模型的C为1、核函数为Linear，该模型在校正集的准确度为99.21%、精确度为100.0%、召回率为98.42%、F1分数为99.20%，拟合时间2.112 0 ms，模型具有较好的性能。

表2 SVM在校正集的模型性能

Table 2 Model performance of SVM in validation set

C	Kernel function	Model evaluation metric				
		Accuracy/%	Precision/%	Recall/%	F1	Fit time/ms
1	Linear	99.21	100.0	98.42	0.992 0	2.112 0
10	Linear	99.21	100.0	98.42	0.992 1	2.067 3
100	Linear	99.21	100.0	98.42	0.992 1	2.289 6
100	RBF	98.78	100.0	97.57	0.987 7	3.174 0
10	RBF	97.04	100.0	94.10	0.969 5	6.527 6
1	RBF	73.81	100.0	47.62	0.645 1	10.432 9
1	Sigmoid	61.33	63.13	54.90	0.586 7	7.307 1
10	Sigmoid	57.90	58.11	57.05	0.575 3	6.709 1
100	Sigmoid	57.21	57.32	57.19	0.572 0	6.428 7
100	Poly	69.88	100.0	39.76	0.568 8	11.693 6
10	Poly	66.74	100.0	33.48	0.501 4	12.699 0
1	Poly	62.19	100.0	24.38	0.391 3	12.763 7

在测试集中以混淆矩阵可视化精度和召回率，如图3所示，最优SVM模型在测试集中的准确度为78.50%、精确度为99.42%、召回率为57.33%、F1分数为72.69%，包含384个假反例和3个假正例样本。模型在测试集的表现不佳，过拟合情况较为严重。

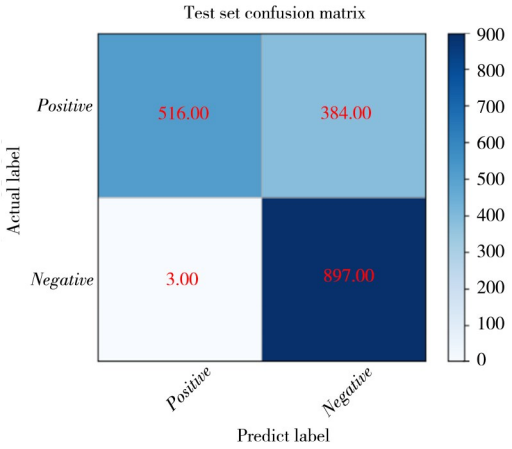

图3 SVM模型外部测试集的混淆矩阵

Fig. 3 Confusion matrix for external test set of SVM model

2.4.3 基于PCA-SVM的快速预判模型 使用PCA算法抽取原始SERS光谱中最具代表性的特征向量，选择方差解释率总计大于95%的特征向量代替原始数据作为SVM模型的输入变量，建立PCA-SVM模型。如图4所示，当保留6个主成分时，特征贡献率超过99.8%，之后每增加一个主成分，贡献率增加不足0.5%，因此，将6个主成分作为SVM模型的输入进行训练。

PCA-SVM模型的超参数优化结果如表3所示，最

优模型中的 C 为 100、核函数为 RBF，该模型在校正集中的准确度为 98. 95%、精确度为 100. 0%、召回率为 97. 90%、F1 分数为 0. 989 4，模型性能良好。经 PCA 处理后的数据，各核函数间的拟合时间和预测时间无明显差别，计算性能良好。

在模型的最优超参数下，使用外部测试集检验模型性能，如图 5 所示，PCA-SVM 模型在测试集的准确度为 54. 89%、精确度为 74. 44%、召回率为 14. 89%、F1 分数为 0. 248 2，过拟合情况严重。共包含 46 个假正例，766 个假反例，将大量“合格”样本误判为“不合格”样本，误判比例超出样本总数 5% 以上，模型分类性能差。这可能是由于经过 PCA 处理虽然降低了 SERS 光谱的维度并以正交变量建立模型，但却丢失了有用的光谱信息。

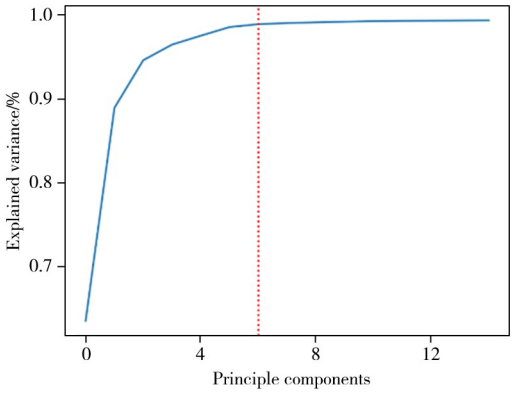

图 4　PCA 主成分帕累托图

Fig. 4　Pareto chart of PCA principal components

表 3　PCA-SVM 在校正集的模型性能

Table 3　Model performance of PCA-SVM in the validation set

C	Kernel function	Model evaluation metric				
		Accuracy/%	Precision/%	Recall/%	F1 scores	Fit time/ms
100	RBF	98. 95	100. 0	97. 90	0. 989 4	0. 091 7
10	RBF	97. 10	100. 0	94. 19	0. 970 1	0. 139 3
10	Linear	96. 88	98. 86	94. 86	0. 968 1	0. 157 6
100	Linear	96. 83	98. 14	95. 48	0. 967 9	0. 297 2
1	Linear	96. 33	99. 50	93. 14	0. 962 1	0. 185 4
1	RBF	94. 71	100. 0	89. 43	0. 944 1	0. 288 7
100	Poly	85. 02	100. 0	70. 05	0. 823 8	0. 227 5
10	Poly	78. 17	100. 0	56. 33	0. 720 6	0. 333 4
1	Sigmoid	63. 14	64. 16	59. 90	0. 619 3	0. 253 2
10	Sigmoid	60. 81	60. 99	60. 29	0. 606 1	0. 246 9
100	Sigmoid	60. 29	60. 35	60. 29	0. 603 0	0. 207 0
1	Poly	63. 81	99. 04	27. 90	0. 435 0	0. 416 7

2. 4. 4　基于 CNN 的快速预判模型　网络架构指定了网络的拓扑结构和层次，控制着神经元和权重之间的连接方式以及它们的数量，这对 CNN 而言十分重要。He 等[18]开发的 ResNet 通过残差结构巧妙解决了网络退化问题，极大地提升了网络的深度并保证了网络性能。本研究设计并使用基于 ResNet18 经典架构的 CNN 网络结构（图 6），该结构共包含 39 层：1 个输入层、17 个卷积层、3 个 BN 层、14 个 ReLU 激活函数层，1 个平均池化层，1 个全连接层（FC 层），1 个 Softmax 层和 1 个输出层。其中卷积层与 FC 层包含权重和偏置，共计 18 层。表 4 为校正集样本 CNN 模型超参数优化的范围与结果，包括批处理大小、卷积核尺寸、学习率、动量、Epoch 数量。通过交叉熵计算模型预测值与真实值的误差，使用 SGD 优化器初始化权重和偏置、计算梯度以及更新模型参数，使用 BN 层加速模型训练，减少对网络初始化参数的敏感性，避免模型过拟合。

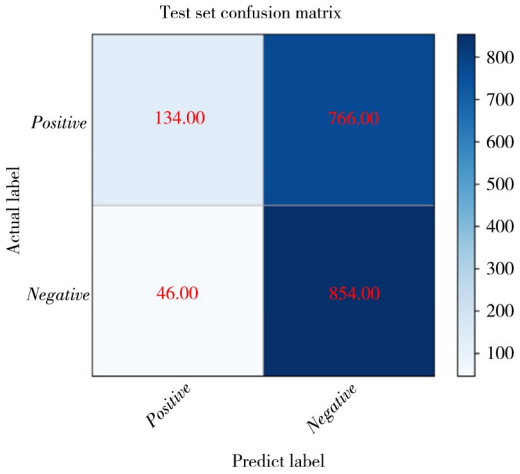

图 5　PCA-SVM 模型外部测试集的混淆矩阵

Fig. 5　Confusion matrix for external test set of PCA-SVM model

准确度是评估模型性能的重要指标，随着 Epoch 的增加，模型不断通过训练数据进行学习，逐渐提高预测能力，如图 7 所示，校正集模型准确度随着 Epoch 的增加逐渐升高，当 Epoch 为 200 时，模型

准确度达到较高水平，并且随着 Epoch 的继续增加，准确度趋于稳定且没有出现持续的波动与下降，模型性能较好，表明当前的超参数设置较为理想。

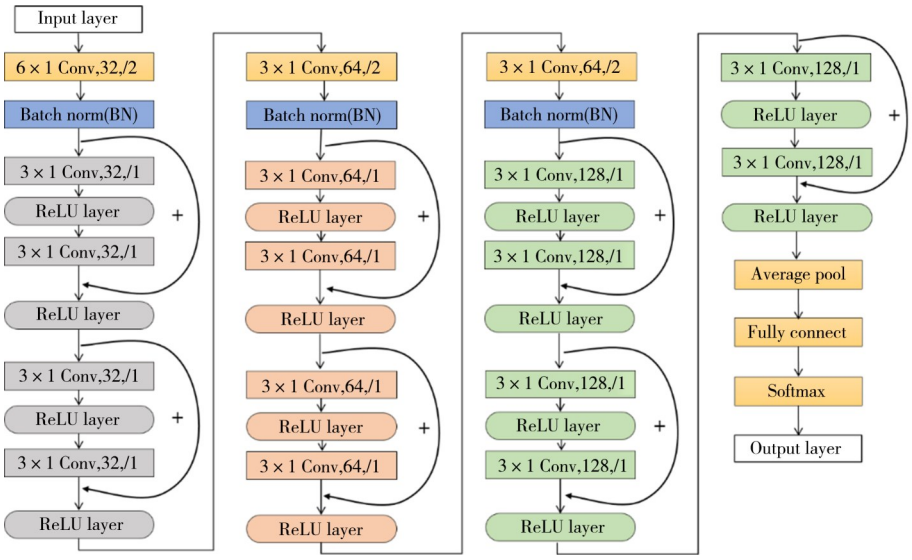

图 6　CNN 网络结构

Fig. 6　CNN network structure

表 4　CNN 的超参数范围和优化值

Table 4　Hyperparameter range and optimized values for CNN

Parameters	Range	Result
Batch size	{50，100，150，200}	100
Convolution kernel size(1D)	{2，3，4，5}	3
Learning rate	{0. 1，0. 01，0. 001，0. 000 1}	0. 000 1
Momentum	{0. 3，0. 6，0. 9}	0. 9
Epochs	{500，1 000，1 500}	500
Optimizer	Stochastic Gradient Descent(SGD)	
Loss function	Cross Entropy	

Loss 是衡量模型预测值与真实值之间差异的指标，如图 8 所示，随着 Epoch 的增加，Loss 值逐渐降低，表明模型学习了更多关于训练样本内在结构的信息，提高了模型的拟合能力，使其能够更加准确地预测训练数据的标签。Epoch 为 400 时，loss 已经降到一个较低值，继续增加 Epoch 反而会引起 Loss 的波动，因此，Epoch 为 400 时，模型性能最佳。

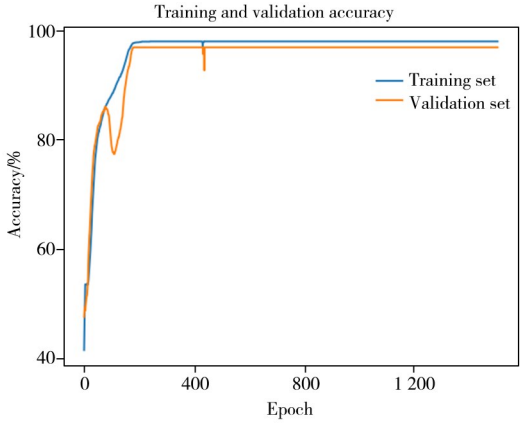

图 7　模型准确度随 Epoch 的变化情况

Fig. 7　Variation of model accuracy with epoch

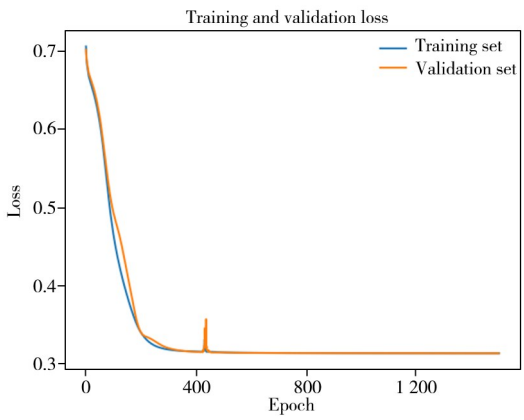

图 8　模型损失函数随 Epoch 的变化情况

Fig. 8　Variation of model loss function with epoch

在确定了CNN模型的最优超参数后，使用最优模型对测试集样本进行考察，结果如图9所示。CNN模型在测试集中的准确度为100.0%、精确度为100.0%、召回率为100.0%、F1分数为1，包含0个假正例和0个假反例样本，模型在测试集的表现优异。

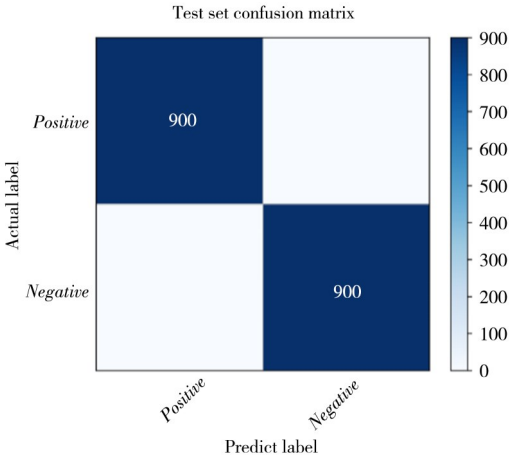

图9　CNN模型外部测试集的混淆矩阵

Fig. 9　Confusion matrix for external test set of CNN model

2.5　模型性能对比分析

RF、SVM、PCA-SVM、CNN各模型在测试集的最佳性能汇总见表5，由表可知CNN模型在准确度、精确度、召回率和F1分数各项指标上均表现最佳。以准确度排序，CNN最优，SVM、RF次之，PCA-SVM最差。但在实际生产中，若假正例偏多，即将“不合格”样本误判为“合格”，会导致不合格的中间体被进一步加工为药物，从而对企业和患者造成极大的风险隐患；若假反例偏多，即将“合格”样本误判为“不合格”，会增加企业的检测成本。因此本研究以召回率和F1分数影响权重为主要指标。

表5　各模型在外部测试集中的性能比较

Table 5　Comparison of the performance of each model in the external test set

Model	Model evaluation metric			
	Accuracy/%	Precision/%	Recall/%	F1 score
RF	70.67	100	41.33	0.584 9
SVM	78.50	94.42	57.33	0.726 9
PCA-SVM	54.89	52.72	94.89	0.677 8
CNN	100.0	100.0	100.0	1.0

如图10所示，计算RF、SVM、PCA-SVM、CNN四种模型的ROC曲线和AUC值。ROC曲线横坐标为假正例率，纵坐标为真正例率，能够体现所有可能的分类阈值下分类器的性能。ROC曲线越靠近左上角，说明分类器效果越好，该曲线越能最大可能地将正样本和负样本分开。4个模型的AUC值从大到小排序依次为CNN、SVM、RF、PCA-SVM，其中CNN的AUC为1.0，表明可以将测试集所有样本正确分类，预测性能和泛化能力均远超于其他机器学习方法，适合作为中药制剂中间体TAMC和TYMC的快速预判模型。

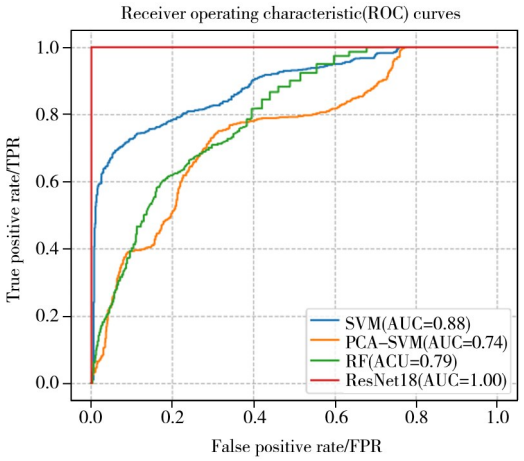

图10　4种模型的ROC曲线

Fig. 10　ROC curves for the four models

3　结　论

本研究以中药制剂中间体的TAMC与TYMC值为检测目标，建立了基于RF、SVM、PCA-SVM和CNN的中药制剂中间体样本中微生物限度的快速预判模型。在相同的外部测试集下，RF、SVM、PCA-SVM和CNN各模型的准确度分别为70.67%、78.50%、54.89%和100.0%。基于ResNet18架构的CNN模型表现出良好的分类性能，丰富了CNN在SERS光谱分析中的应用。相对于传统的模式识别方法，CNN对光谱预处理方法和波长变量的选择要求不高，具备很强的学习能力，在处理非线性、大样本数据时常常具有优异的性能表现。但CNN也具有一些缺点：如严重依赖训练样本数量，数据量越大，质量越优，CNN的表现性能越好；且CNN网络模型设计复杂，具有大量超参数，参数优化过程复杂且时间成本高。此外，CNN对硬件的计算能力要求高，普通的硬件设备难以满足巨大计算量的速度，耗费代价高。最需要注意的是，CNN工作过程缺乏清晰的理论对作用机制进行解释，算法分析相对困难，难以回归实验数据本身，可解释性有待进一步提高。但总体而言，

本文为中药制剂中间体两大类微生物限度快速预判提供了有效方法，可对不合格样品进行有效风险预警，对提高中药生产过程中间体微生物的质量控制水平大有裨益。

参考文献：

[1] Pharmacopoeia Committee. *Pharmacopoeia of the People's Republic of China* (*2020*). 1107 Microbial Limit Standards for Non Sterile Drugs. Beijing: China Medical Science and Technology Press(国家药典委员会. 中国药典. 1107 非无菌药品微生物限度标准. 北京：中国医药科技出版社), **2020**.

[2] Cui Z, Ojaghian M R, Tao Z, Kakar K U, Zeng J, Zhao W, Duan Y, Vera Cruz C M, Li B, Zhu B, Xie G. *J. Appl. Microbiol.*, **2016**, 120(5): 1357-1367.

[3] Marjan M, Akhtar H, Louis M J. *TrAC*, **2018**, 107: 60-77.

[4] Zhu Y T, Li Z Y, Xie M M, Yan Y L, Zhang T, Wang H X. *J. Instrum. Anal.*(朱怡亭，李芷瑶，谢茂梅，颜月玲，张桐，王海霞. 分析测试学报), **2024**, 43(1): 138-146.

[5] Gao W C, Li B, Yao R Z, Li Z P, Wang X W, Dong X L, Qu H, Li Q X, Li N, Chi H, Zhou B, Xia Z P. *Anal. Chem.*, **2017**, 89(18): 9836-9842.

[6] Prakash O, Sil S, Verma T, Umapathy S. *J. Phys. Chem. C*, **2019**, 124(1): 861-869.

[7] Cui L, Chen P, Chen S, Yuan Z, Yu C, Ren B, Zhang K. *Anal. Chem.*, **2013**, 85(11): 5436-5443.

[8] Zhu X Y, Zhao Y W, Zhang Z S, Wang H X, Liu B S, Li Z, Wang M F. *Microchim. Acta*, **2021**, 188: 1-11.

[9] Zhou S Y, Guo X J, Huang H Q, Huang X Q, Zhou X, Zhang Z B, Sun G D, Cai H H, Zhou H B, Sun P H. *Anal. Chem.*, **2022**, 94(15): 5785-5796.

[10] Liu H B, Du X J, Zang Y X, Li P, Wang S. *J. Agric. Food Chem.*, **2017**, 65: 10290-10299.

[11] Zhu X Y, Ning Y, Zhang Z, Wen Y, Zhao Y W, Wang H X. *Anal. Bioanal. Chem.*, **2023**, 415(8): 1529-1543.

[12] Zhao Y W, Zhang Z S, Ning Y, Miao P Q, Li Z, Wang H X. *Spectrochim. Acta A*, **2023**, 293: 122510.

[13] Wang Q, Zeng W D, Xia Z P, Li Z P, Qu H. *Chin. J. Laser*(王其，曾万聃，夏志平，李志萍，曲晗. 中国激光), **2021**, 48(3): 136-144.

[14] Seju K, Inyoung K, Peter J V. *Anal. Chem.*, **2021**, 93(27): 9319-9328.

[15] Cheng N T, Chen D J, Lou B, Fu J, Wang H Y. *Biosens. Bioelectron.*, **2021**, 186: 113246.

[16] Cheng J L, Ma J, Xiao A H, Zhu Z L, Sang Z Y, Ma L. *J. Beijing Fores.Univ.*(程嘉莉，马江，肖爱华，朱仲龙，桑子阳，马履. 北京林业大学学报), **2020**, 42(2): 96-105.

[17] Lee P, Meisel D. *J. Phys. Chem. C*, **1982**, 86(17): 3391-3395.

[18] Xie S N, Girshick R, Dollar P, Tu Z W, He K M. *IEEE CVPR*, **2017**, 7: 1492-1500.

（责任编辑：盛文彦）