牛黄类药材 REIMS 图谱的快速识别研究 *

石岩,荆文光,程显隆,魏锋 **
(中国食品药品检定研究院,北京 102629)

摘要 **目的:**使用快速蒸发离子化质谱(REIMS)技术对牛黄、培植牛黄、人工牛黄和体外培育牛黄进行测定,并使用机器学习技术对样品的 REIMS 谱图进行快速识别预测。**方法:**干法灼烧将样品组分引入 REIMS 中,质谱扫描范围 *m/z* 50~1 200;扫描模式为灵敏模式;扫描时间为 0.2 s。负离子模式采集,数据记录为 continuum 模式。通过对样品 REIMS 谱图的数据进行聚类分析和主成分分析,分析样品数据概况。分别建立偏最小二乘判别分析、逻辑回归、决策树、随机森林、自适应提升(分别以逻辑回归和决策树为弱评估器)模型,并通过 GaussianCopula、CTGAN、CopulaGAN 和 TVAE 算法仿真合成数据,然后与原训练集数据组成新的训练集用于模型的训练。**结果:**使用新训练集训练得到的以决策树为弱评估器的自适应提升模型对 4 种牛黄类药材识别预测能力最好,对测试集识别的准确率为 0.97,精确率为 0.90,召回率为 0.97,F1 得分为 0.93,ROC 面积为 1.00。使用模型输出的概率还可以根据药品监管的实际应用场景通过调整概率阈值灵活地使用。**结论:**使用 REIMS 技术与机器学习技术联用可以实现牛黄类药材的快速而准确的识别。

关键词:牛黄;培植牛黄;人工牛黄;体外培育牛黄;快速蒸发离子化质谱;人工智能;机器学习;真伪鉴别;偏最小二乘判别分析;逻辑回归;决策树;随机森林;自适应提升;仿真数据合成;识别概率

中图分类号:R 917 **文献标识码:**A **文章编号:**0254-1793(2025)02-0350-12
doi:10.16155/j.0254-1793.2024-0312

Study of rapid identification of cow-bezoar and its substitutes medicinal herbs using REIMS*

SHI Yan, JING Wen-guang, CHENG Xian-long, WEI Feng**
(National Institutes for Food and Drug Control, Beijing 102629, China)

Abstract **Objective:** To study the rapid identification of cow-bezoar and its substitutes medicinal herbs using the technique of rapid evaporative ionization mass spectrometry (REIMS) couple with machine learning. **Methods:** The samples were ionized and determined by REIMS with *m/z* 50-1 200 as scanning range in sensitive mode and negative ion mode, 0.2 s as scanning time, and using dry burning method. REIMS data of samples was recorded as continuous mode. Then the general situation of REIMS data distribution was studied and analyzed through the methods of cluster analysis and principal component analysis. Some models or algorithms, such as partial least squares discriminant analysis (PLS-DA), logistic regression (LR), decision tree (DT), random forest (RF) and

* 中国食品药品检定研究院关键技术研究基金项目(GJJS-2022-10-2);中国食品药品检定研究院学科带头人培养基金项目(2023X10)
** 通信作者 Tel:(010)53852020;E-mail:weifeng@nifdc.org.cn
第一作者 Tel:(010)53852081;E-mail:shiyan@nifdc.org.cn

adaptive boosting (AdaBoost, with LR and DT as base estimator respectively) were established. In the models training procedure, simulation synthesis data generated by algorithms of GaussianCopula, CTGAN, CopulaGAN and TVAE joined the original training set data as the new training set. **Results:** AdaBoost (DT as base estimator) trained with the new training set was the best model which could accurately predict cow-bezoar and its substitutes medicinal herbs. The accuracy for identifying the test set was 0.97, the precision was 0.90, the recall was 0.97, the F1 score was 0.93, and the AUC of ROC was 1.00. The probability output from the model could also be flexibly used by adjusting the probability threshold according to the actual application scenarios of drug regulation. **Conclusion:** The combination of REIMS technology and machine learning technology can achieve fast and accurate recognition of cow-bezoar and its substitutes medicinal herbs.

Keywords: cow-bezoar; cultured cow-bezoar; artificial cow-bezoar; cow-bezoar cultured *in vitro*; REIMS; artificial intelligence; machine learning; authenticity identification; PLS-DA; logistic regression; decision tree; random forest; adaptive boosting; simulation data synthesis; identification probability

牛黄为牛科动物牛 *Bos taurus domesticus* Gmelin 的干燥胆结石，是我国传统的名贵中药材，其临床使用历史悠久，功效显著，在我国有广泛的使用[1-3]。牛黄药材需求旺盛，但资源稀缺，且价格昂贵，难以满足人民群众用药需求，为缓解牛黄日益紧张的供需矛盾，陆续出现了培植牛黄、体外培育牛黄以及人工牛黄作为牛黄的代用品[2-5]。这些牛黄的代用品，在研发阶段就是尽可能接近牛黄的化学成分以获得与牛黄近似的药理活性，因此与牛黄具有一定的相似性，但同时这些代用品在化学成分和药理活性方面与牛黄也存在较大的差异[2,6]。

4 种牛黄类药材从来源属性看，有自然产物和工业产品的差别，价格差异巨大，而且其药理活性、临床疗效和国家药品监督管理局对其临床使用的相关规定也有不同[2,7]，因此有必要开展这 4 种牛黄类药材的鉴别方法研究。但是，目前极少有能够同时鉴别 4 种牛黄类药材的文献报道，且对于部分牛黄类药材的相关鉴别研究[8-9]都存在停留在难以实际应用的无监督学习模型阶段，数据预处理过于烦琐，模型判别准确率仍有待提升等问题。

本研究基于以上现实情况，为了实现这 4 种牛黄类药材的准确识别，并且切实解决掣肘研究成果投入实际应用的复杂的数据处理问题，开展了快速蒸发离子化质谱（rapid evaporative ionization mass spectrometry，REIMS）分析技术和多角度多方位的机器学习相关技术联合应用研究。其中干法 REIMS 技术是在之前该技术成功应用于植物来源中药材研究[10]基础之上的又一次尝试。在对 REIMS 数据进行的数据分析和机器学习研究中，本研究从样品数据实际出发，充分利用麻省理工学院 Data to AI Lab 的合成数据研究成果，使用 4 种方式对样品的 REIMS 实验数据进行数据增强处理，然后通过准确率指标，评估、比较并优选出数据增强前后和不同初始模型训练得到的最优模型。以上研究的大致流程，以及对于最优模型在药物分析中的实际应用见图 1。本研究建立了能够简便、准确快速识别 4 种牛黄类药材 REIMS 图谱的模型，为牛黄类药材的药品市场监管工作提供了技术支撑，同时对于其他品种药材的研究与监管工作也具有一定的参考价值。

1 仪器与试药

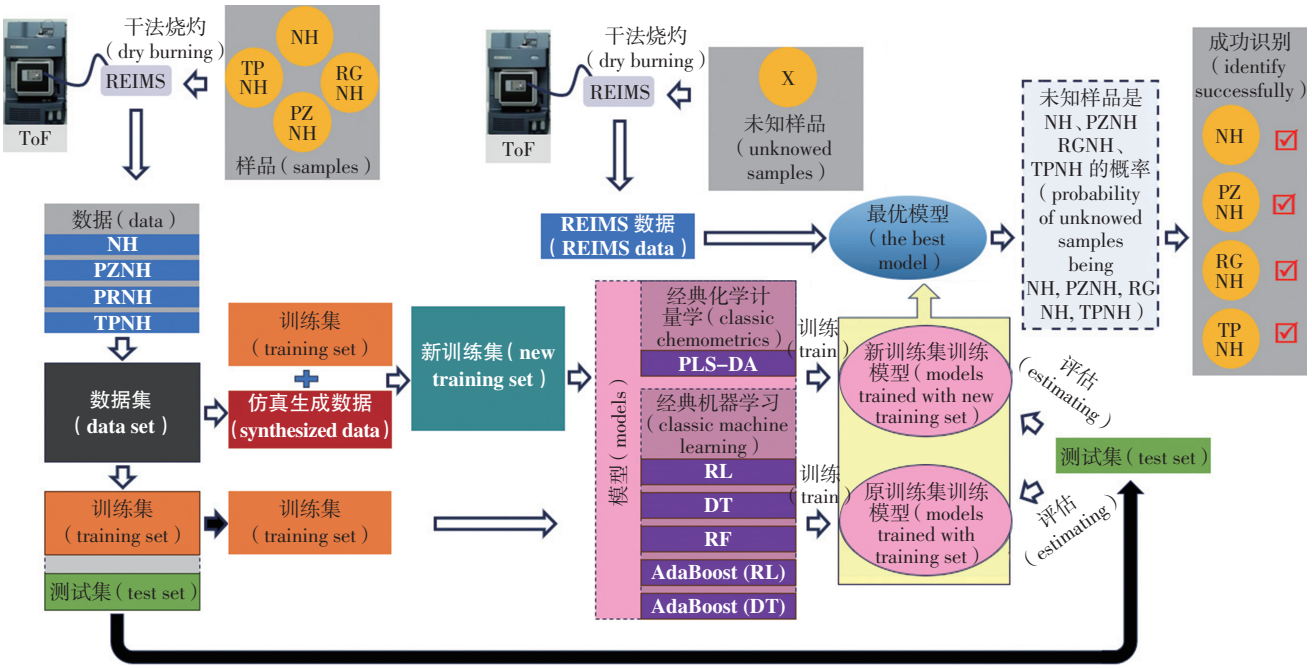
1.1 仪器

SYNAPT G2-S 四极杆飞行时间质谱仪，配置 REIMS 离子源（Waters 公司）；Pump 11 Elite 注射泵（Harvard Apparatus 公司）；WSD71 可调温恒温数显焊台（Weller 公司）。

1.2 试药

亮氨酸脑啡肽（Sigma-Aldrich 公司，批号 W13012301，质谱级纯度）。

牛黄样品 20 批（编号 nh_001~nh_020）、培植牛黄样品 187 批（编号 pznh_001~pznh_187）、人工牛黄样品 135 批（编号 rgnh_001~rgnh_135）和体外培育牛黄样品 95 批（编号 tpnh_001~tpnh_095）。牛黄样品收集自各药材市场，培植牛黄样品和体外培育牛黄样品均为各自独家生产厂家生产，人工牛黄样品为药品抽检样品，经检验均符合现行标准。样品均经中国食品药品检定研究院石岩研究员鉴定为正品。



NH. 牛黄 (cow-bezoar)　PZNH. 培植牛黄 (cultured cow-bezoar)　TPNH. 体外培育牛黄 (artificial cow-bezoar)　RGNH. 人工牛黄 (cow-bezoar cultured *in vitro*)

图 1　研究及应用流程图

Fig. 1　Research and application flowchart

1.3　编程语言和软件

Python 编 程 语 言 (Python Software Foundation, version:3.8.8)。

2　方法与结果

2.1　质谱检测数据采集

干法 REIMS 焊台烙铁头温度 450 ℃，100 ng·mL^{-1} 亮氨酸脑啡肽异丙醇溶液，流速 100 μL·min^{-1}。质谱扫描范围 *m/z* 50~1 200；扫描模式为灵敏模式；扫描时间 0.2 s。负离子模式采集，数据以 continuum 模式记录。样品的质谱图见图 2。

2.2　数据概况分析

2.2.1　数据概况　质谱扫描范围 *m/z* 50~1 200，样品总批数为 437 批，数据最终构成 437 × 67 的向量矩阵，导出 csv 格式备用。

2.2.2　聚类分析　以欧氏距离为度量，根据 Ward 法对各样品的 REIMS 图谱进行系统聚类分析，并按照质谱 *m/z* 通道信号强弱绘制聚类分析热图，见图 3。

2.2.3　主成分分析　将各批样品的 REIMS 图谱数据进行主成分降维解构，结果前 5 个主成分 (分别为 PC1、PC2、PC3、PC4、PC5) 的累计方差水平为 87.2%，4 种牛黄样品前 5 个主成分分别对应的二维空间的得分分布情况见图 4，对角为各类样品在对应主成分上的分布图。

2.3　数据集的划分处理

数据集样本总数为 437 个，随机打乱次序并按照 7∶3 比例随机划分训练集和测试集。数据集划分后，训练集中样本总数为 305 个，其中牛黄、培植牛黄、人工牛黄和体外培育牛黄分别为 16 个、133 个、89 个和 67 个；测试集中样本总数为 132 个，其中牛黄、培植牛黄、人工牛黄和体外培育牛黄分别为 4 个、54 个、46 个和 28 个。

本研究所采用数据集划分及机器学习相关算法中，所涉及到随机情况的，随机数种子均设为 837。

2.4　偏最小二乘判别分析算法的建立及评估

原数据中表示类别的牛黄、培植牛黄、人工牛黄和体外培育牛黄均为字符串格式，在进行偏最小二乘判别分析 (partial least squares discriminant analysis, PLS-DA) 时须对其进行数字表示，本研究使用独热编码方法对 4 种牛黄的类别进行了变换。主成分数在 2~9 范围内，PLS-DA 模型对测试集样品识别准确率为 0.82~0.92，当主成分数为 4 时，PLS-DA 模型的准确率最高，即 0.92，所以选择 PLS-DA 主成分数为 4。

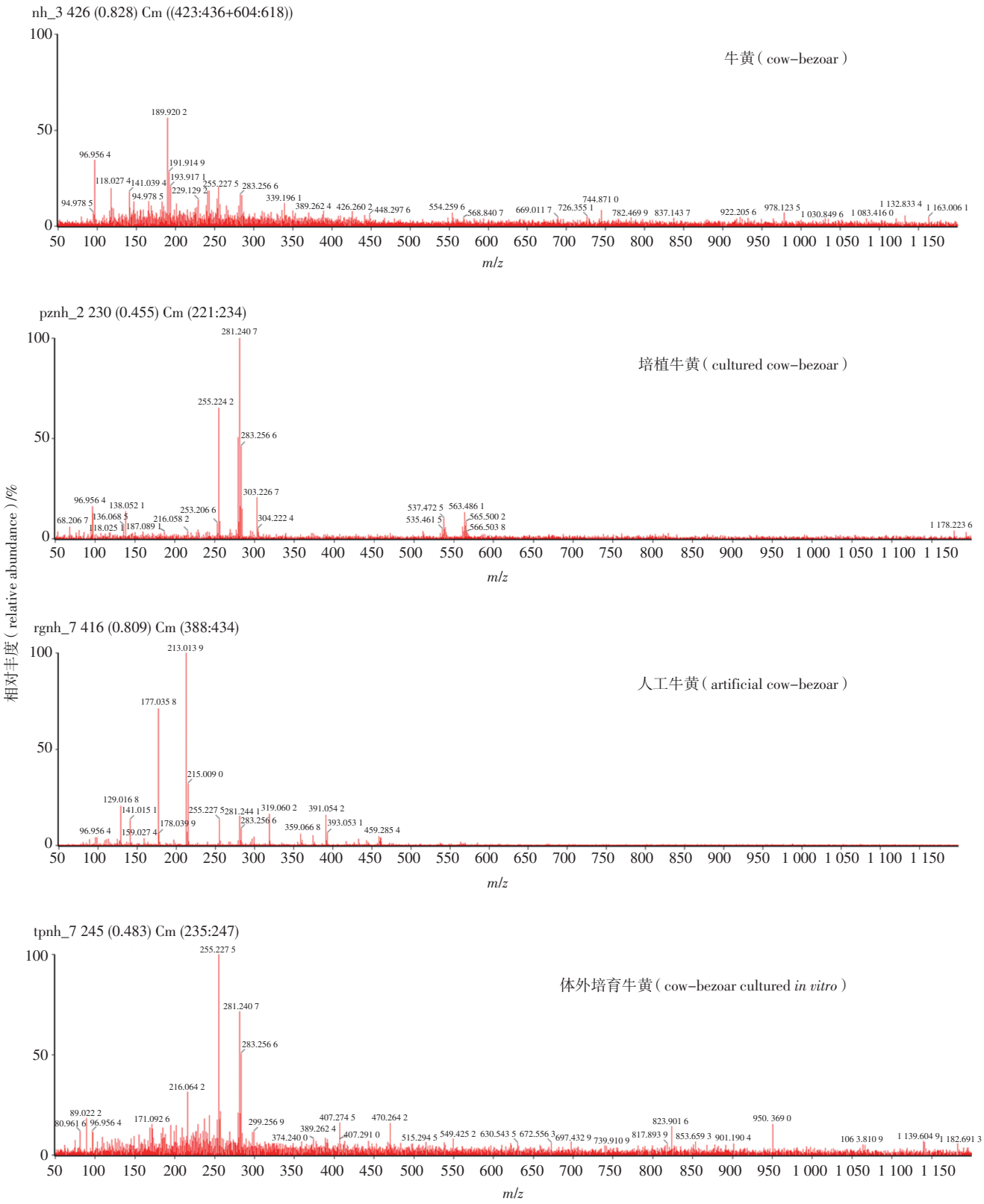


图 2　4 种牛黄类样品的 REIMS 图

Fig. 2　REIMS spectra of 4 kinds of cow-bezoar samples

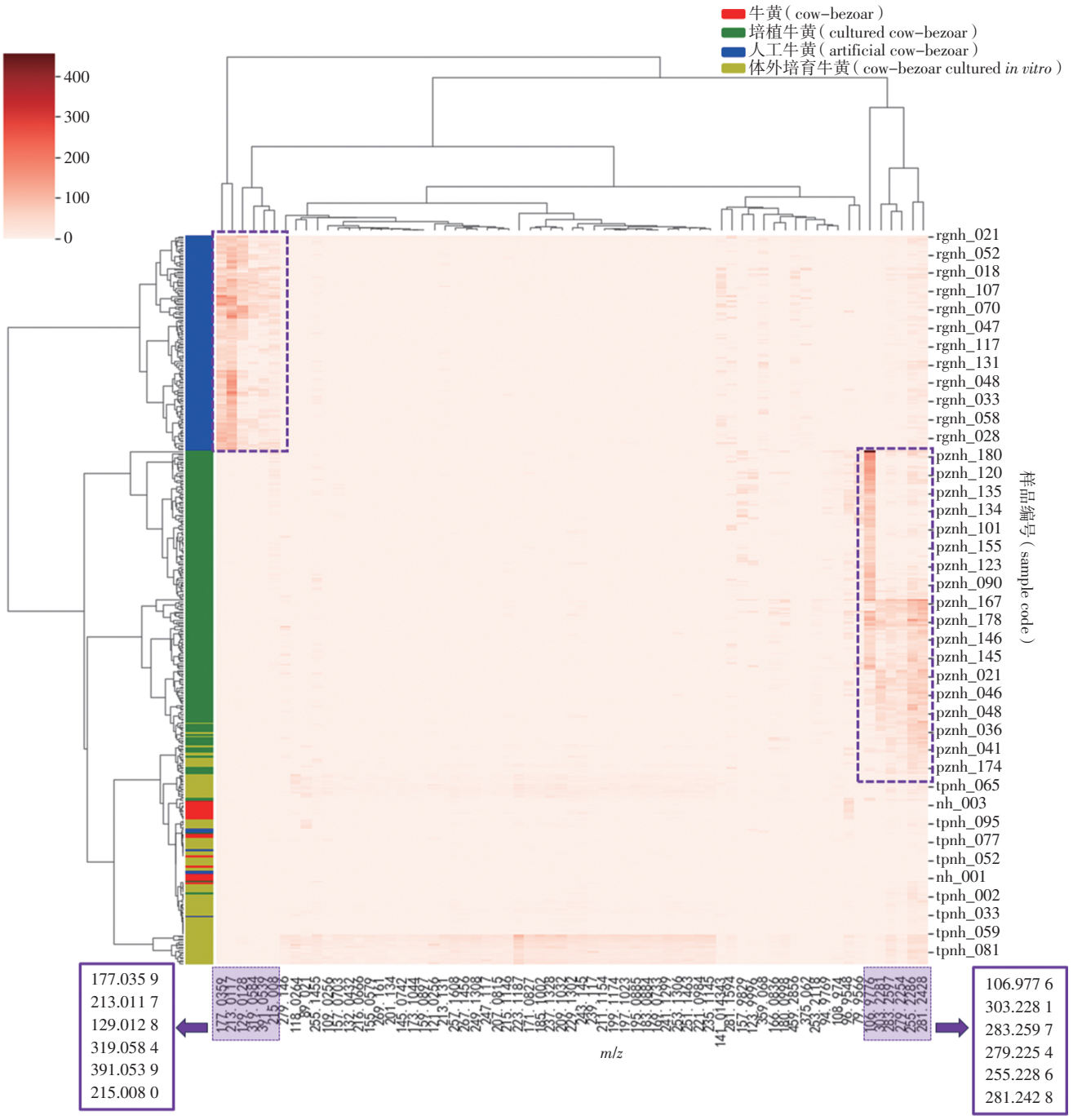


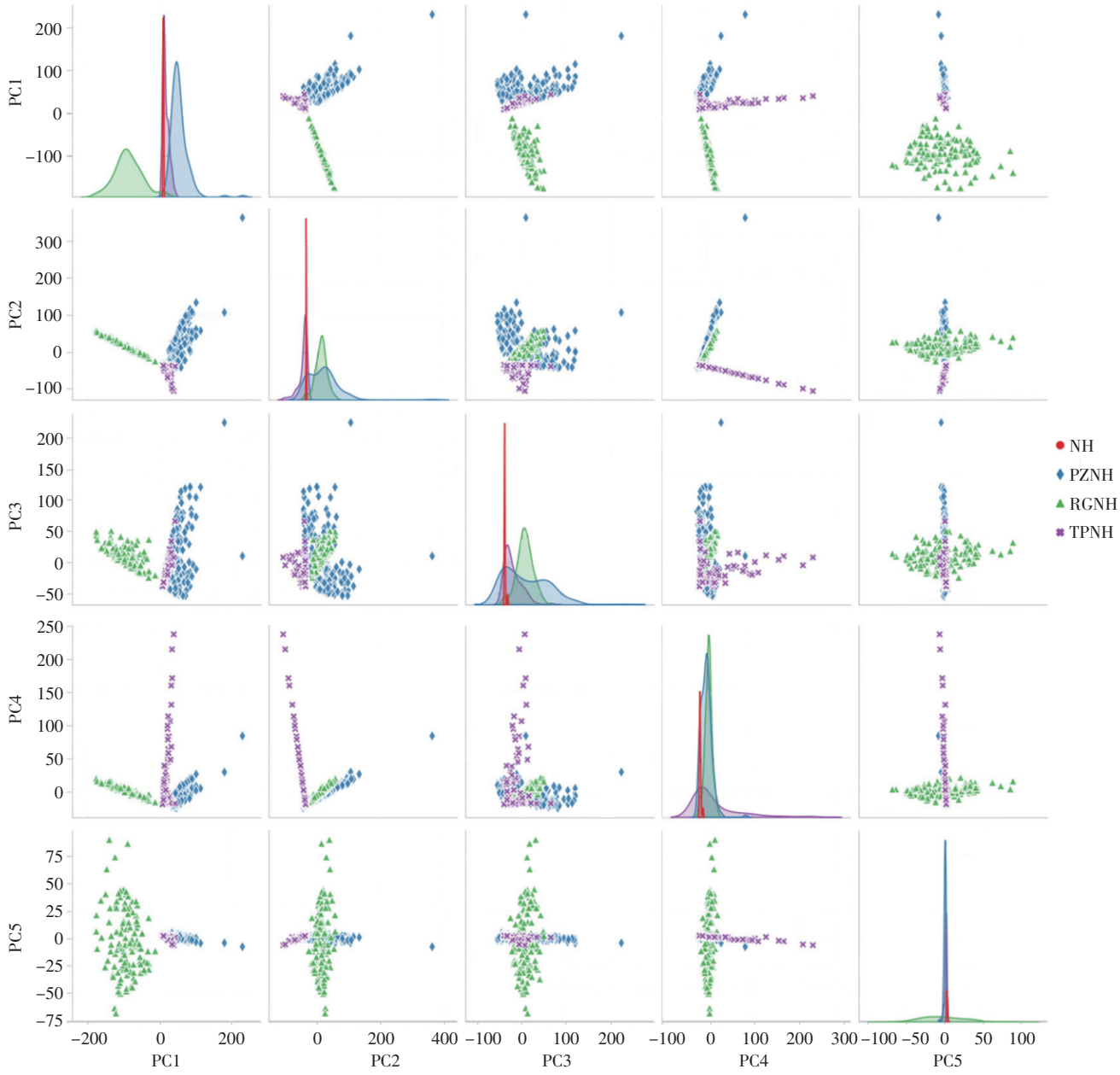
图 3　聚类分析热图

Fig. 3　Cluster analysis heatmap

2.5　机器学习预测模型的建立及评估

2.5.1　逻辑回归模型　逻辑回归（logistic regression，LR）模型惩罚项选择 L2 法（岭回归法），求解器选择 saga，最大迭代次数为 1 500，正则化强度倒数 C 为 0.01。以训练集对初始化模型进行训练，然后使用测试集对模型进行评估，准确率为 0.94。

2.5.2　决策树模型　决策树（decision tree，DT）特征选择标准（criterion）选择 log-loss 法，DT 最大深度（max_depth）为 7，子节点下划分最小样本数（min_samples_split）为 3，叶节点最小样本数（min_samples_leaf）为 1。以训练集对初始化模型进行训练，然后使用测试集对模型进行评估，准确率为 0.96。



NH. 牛黄（cow-bezoar） PZNH. 培植牛黄（cultured cow-bezoar） TPNH. 体外培育牛黄（artificial cow-bezoar） RGNH. 人工牛黄（cow-bezoar cultured *in vitro*）

图 4 样品主成分得分分布图

Fig. 4 Principal component score plots of samples

2.5.3 随机森林模型 随机森林（random forest，RF）模型中弱分类器决策树（n_estimators）为 29 个，DT 最大深度（max_depth）为 8。以训练集对初始化模型进行训练，然后使用测试集对模型进行评估，准确率为 0.94。

2.5.4 自适应提升模型 自适应提升（adaptive boosting，AdaBoost）是弱分类器组成的集成算法，本研究分别使用逻辑回归和决策树作为弱分类器，组成 2 种 AdaBoost 模型。

以 LR 为弱分类器的 AdaBoost 模型[AdaBoost（LR）]，其中逻辑回归参数设置同本研究中的逻辑回归模型，弱分类器数为 69 个，学习率为 0.037。以训练集对初始化模型进行训练，然后使用测试集对模型进行评估，准确率为 0.95。

以 DC 为弱分类器的 AdaBoost 模型[AdaBoost（DT）]，其中决策树参数设置同本研究中的决策树模

型,弱分类器数为 44 个,学习率为 2.492。以训练集对初始化模型进行训练,然后使用测试集对模型进行评估,准确率为 0.96。

2.5.5 数据增强 由于原数据集中 4 种牛黄的数据极不平衡,如牛黄只有 20 批,远少于培植牛黄的 187 批,不利于机器学习算法训练,因此本研究使用 GaussianCopula、CTGAN、CopulaGAN 和 TVAE 算法分别对牛黄、人工牛黄及体外培育牛黄类别的数据进行了学习及数据增强,按照每类样本的总数量达到约 200 批为目标,牛黄、人工牛黄及体外培育牛黄分别仿真虚拟生成了与原数据具有相同格式和统计特征的 180 批、60 批和 100 批数据。

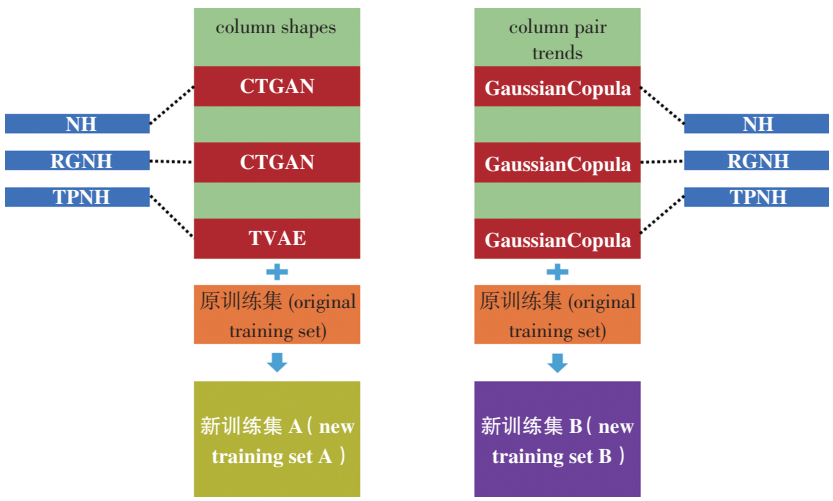
对 GaussianCopula、CTGAN、CopulaGAN 和 TVAE 算法所生成的数据质量分别从 column shapes 和 column pair trends 2 个方面进行统计相似性评价,结果见表 1。

表 1 生成数据评价

Tab. 1 Evaluation of generate data

算法(algorithm)	牛黄(cow-bezoar)		人工牛黄(cow-bezoar cultured *in vitro*)		体外培育牛黄(artificial cow-bezoar)	
	column shapes	column pair trends	column shapes	column pair trends	column shapes	column pair trends
GaussianCopula	77.23%	91.34%	74.63%	93.75%	41.89%	93.39%
CTGAN	88.15%	78.56%	86.67%	91.45%	70.40%	69.18%
CopulaGAN	87.61%	79.37%	76.05%	91.24%	46.93%	69.25%
TVAE	86.67%	79.82%	84.04%	91.48%	71.74%	88.54%

2.5.6 新训练集训练各类模型 依次按照牛黄、人工牛黄及体外培育牛黄分别选取 column shapes 和 column pair trends 得分最高值的生成数据与原训练集数据拼接组成 2 组新的训练集数据,具体新训练集组成方案示意图见图 5。



NH. 牛黄(cow-bezoar) PZNH. 培植牛黄(cultured cow-bezoar) TPNH. 体外培育牛黄(artificial cow-bezoar) RGNH. 人工牛黄(cow-bezoar cultured *in vitro*)

图 5 新的训练集组成示意

Fig. 5 Schematic diagram of the new training set composition

使用 2 组新的训练集对“2.4”和“2.5”项中的各类模型从初始状态开始进行训练,模型超参数设置与原训练集训练时完全相同。使用“2.3”项所划分的原测试集分别对 2 组新的训练集训练得到的模型进行准确率评估,结果与原训练集训练得到的模型准确率评估结果汇总见表 2。表 2 显示,所有训练完毕的各类模型中,以使用新训练集 B(column pair trends 得分最高的仿真生成数据与原训练集组成)训练得到的 AdaBoost(DT)对测试集识别准确率最高,可达 0.97。

表 2 识别测试集准确率汇总

Tab. 2 Accuracy summary of identification of test set

算法（algorithm）	准确率（accuracy）		
	原训练集训练（trained with original training set）	新训练集 A 训练（trained with new training set A）	新训练集 B 训练（trained with new training set B）
PLS-DA	0.92	0.74	0.80
LR	0.94	0.91	0.94
DT	0.96	0.96	0.96
RF	0.94	0.94	0.95
AdaBoost（LR）	0.95	0.92	0.95
AdaBoost（DT）	0.96	0.96	0.97

3 讨论

3.1 REIMS 技术特点及采样测定方式

REIMS 是一种新型的原位质谱分析技术，分析过程快速且无需样品的预处理，最早应用于医学领域，目前食品领域也有了越来越多的应用[14-17]。通常情况下，REIMS 需借助于前端的 iKnife 装置对分析对象进行气溶胶化，导入 REIMS 离子源进行电离，然而对于中药材而言，由于含水量较低，无法直接使用 iknife 对分析对象进行处理。牛黄类样品也属于这种情况，鉴于此，本研究采用了电烙铁加热的方式对牛黄类药材进行加热处理，而后导入质谱进行分析测定。

4 种牛黄类药材样品一般为类球形、卵形或粉末等形态，为了使 REIMS 数据具有代表性和普遍性，并兼顾实际使用情形，在测定时采用多部位多角度综合取样，使测得的 REIMS 数据能够包括和代表一般情况下对牛黄类药材不同取样部位的采样。因此，在目测情况下同类的牛黄类药材样品的 REIMS 谱图会有较明显的组内差异。

REIMS 技术通常需要借助数据分析技术，而本研究实际干法灼烧采样方式及较明显的组内差异，更使得必须使用先进的数据分析和处理方法对 REIMS 谱图进行多学科交叉研究，因此在后续过程中采用了多种机器学习模型和算法，并与传统的化学计量学方法进行了对比。

3.2 数据预处理方式的选择

对于本研究涉及的数值型表格数据来说，建模前通常需要进行标准化预处理。标准化处理虽然并不困难，但是仍然会增加分析过程的烦琐程度，无疑也增加了不熟悉数据分析的相关人员的操作难度。为了更加方便模型的实际应用与现场操作，在研究之初就以最简化流程为目标进行设计，因此对数据实行了无标准化，直接应用于模型的处理，从模型对测试集识别准确率的评估结果来看，结果令人满意。

3.3 数据概况分析

依托聚类分析和主成分分析对样品的 REIMS 谱图概况进行探索，结果表明这 4 种牛黄类样品的 REIMS 谱图具有一定的组间差异。图 3 是聚类分析热图，该图显示，与牛黄聚类最接近的是体外培育牛黄及部分培植牛黄，而人工牛黄与它们有较明显的聚类距离，但是在体外培育牛黄聚类簇中夹杂有其他 3 种牛黄样品，这 4 种牛黄类样品之间聚类界限不清晰，聚类簇中多见存在混杂种类的情况。此外，图 3 中显示培植牛黄与人工牛黄各自有部分响应较高的离子存在，有特征性离子的可能，但在进一步地对这些离子数据分析时发现效果不佳，有待进一步地深入研究。图 4 展示了经过主成分分析降维后，样品在前 5 个主成分空间的分布情况，从该图可知人工牛黄在第 1 主成分上与其他种类牛黄样品有较明显差异，但从第 1 主成分分布曲线上可见，仍有一定数量的人工牛黄样品与其他牛黄分布有交集，这些牛黄类药材在另外 4 个主成分的分布上则存在大量的重叠情况。因此，总体来说 4 种牛黄类样品的 REIMS 图谱虽然有一定组间差异，但是差异并未达到能够准确识别的程度。

3.4 模型超参数的优化

机器学习模型中超参数的选取可以直接影响所建模型的经验风险和结构风险，选取适当的超参数进行学习，可以使得模型具有优良的泛化能力，因此在机器学习领域内，超参数优化至关重要。

在本研究中，除了所应用到的传统化学计量学 PLS-DA 外，其余的机器学习算法 LR、DT、RF、AdaBoost（LR）和 AdaBoost（DT）都有较大量的超参数需要优化。在优化超参数的过程中，同时使用了绘制学习曲线和先进的超参数优化库（optuna）2 种方法，如 LR 模型的惩罚项选择、求解器等超参数是由学习曲线确定的，而 DT、RF、AdaBoost（LR）和 AdaBoost（DT）中的与树相关、与弱分类器数目和学

习率等相关的超参数优化则是通过 optuna 实现的。

学习曲线的绘制，采用的是训练集样品 10 折交叉验证的平均准确率进行评估，以 LR 模型的求解器（solver）的选择为例，其学习曲线见图 6。

optuna 是一个基于贝叶斯优化算法、先进超参数采样算法和剪枝策略的机器学习库，在明确超参数种类、超参数空间以及优化方向后便可以实现自动搜索超参数最优组合，相比网格搜索法具有极高的超参数优化效率，同时也可以提供优化过程的可视化，如图 7 为 RF 模型的 50 轮 trial 的优化过程。

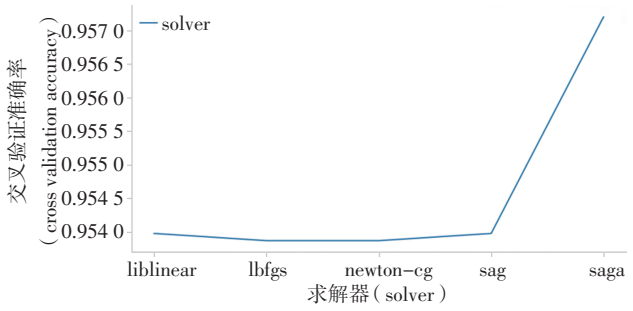


图 6 求解器选择学习曲线

Fig. 6 Solver selection learning curve

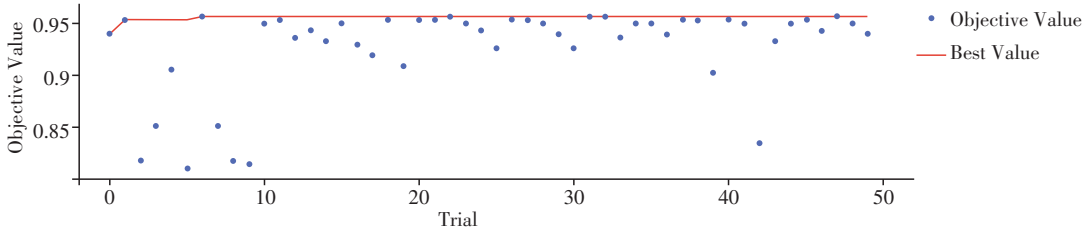


图 7 optuna 对 RF 模型优化过程

Fig. 7 RF model optimization process with optuna

训练集交叉验证结果和测试集识别结果均表明，本研究中的模型超参数选择和优化均十分有效，基本可以保证模型的泛化能力。

3.5 数据增强算法应用的效果与意义

数据是模型建立的基础，数据分布的不平衡不利于模型从样本量少的分类中学习提取出足够的规律，进而影响模型预测识别的准确率。本研究数据以体外培育牛黄为最多，而以牛黄为最少，差异接近 10 倍。为了解决此类数据问题，常采用欠采样、过采样、着重拟合等方法，但并不适合本研究实际情况。除以上方法外，还有基于原型的数据仿真生成方法，即学习原型数据的特点，针对性地进行生成，如本研究采用的 GaussianCopula、CTGAN、CopulaGAN 和 TVAE 算法。以 CTGAN 算法为例，该算法是依据生成对抗网络（cenerative adversarial network，GAN）模型为处理生成表格型数据衍生得来。GAN 是 2014 年由 Goodfellow 等[11]提出的，目前仍是计算机视觉领域与图像转换工作中最常用的神经网络之一，在图像数据增强领域有着重要作用[12-13]。GAN 的基础结构是由生成器与鉴别器组成，生成器对真实数据学习并由初始噪音逐步生成与真实数据相仿的虚拟数据，鉴别器则负责对真实数据进行学习并对生成器生成的虚拟数据进行识别判断，二者不断迭代，相互促进并优化，最终生成可以假乱真的仿真数据。其他的 GaussianCopula、CopulaGAN 和 TVAE 算法也有各自的原理及应用，通常为概率图形建模和深度学习技术，在此不再赘述。

基于仿真的合成数据主要用于训练机器学习模型时，对真实数据的补充或扩充，在某些领域也用于真实数据的泄露。本研究在应用合成数据对模型进行训练的评估结果（表 2）表明，加入合成数据后的训练集对复杂的模型的训练效果较好，对于传统化学计量学算法的 PLS-DA 反而产生了一定的反作用。这种结果，可能是因为越是复杂的机器学习模型，其充分的训练就越是依赖充足的样本数据，而简单的模型算法虽对样本量要求不高，但对于未知样本的预测识别准确率明显偏低。

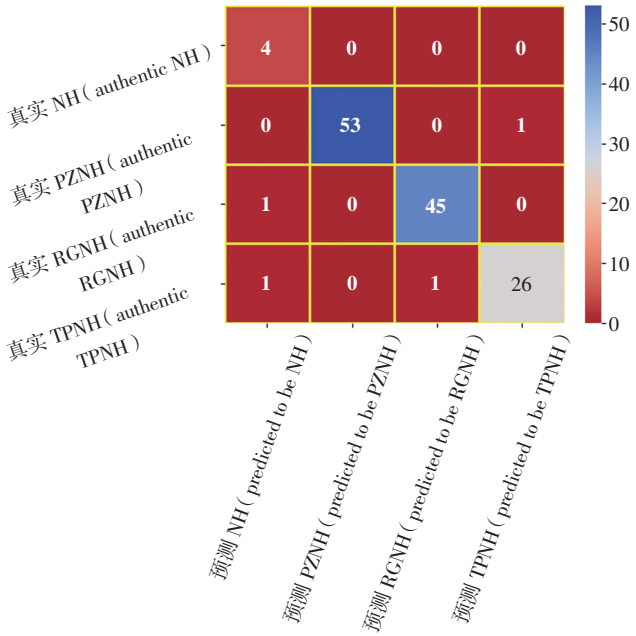
column shapes 和 column pair trends 是真实数据与生成数据的分别在单独列和成对列的统计相似性，分别代表单独列的边际分布和列的相关性。由表 2 可知，对于表征合成数据质量的 2 个衡量指标而言，column shapes 和 column pair trends 得分的选择会导致不同的模型训练结果，在本研究中，使用了 column pair trends 得分最高的新训练集 B 相较于使用了

column shapes 得分最高的新训练集 A 而言，模型对未知样本的预测识别能力的提升效果更好。可能是由于 column shapes 仅评估了单独列的边际分布，而列的相关性对于模型训练而言可能更加重要。

3.6 最优模型的选择及识别能力评估

本研究建立并优化了 6 种不同的模型和算法，通过不同训练集训练得到 18 个可以对牛黄类药材 REIMS 图谱进行识别的模型和算法，其中使用新训练集 B 训练得到的 AdaBoost（DT）对测试集的识别准确率可达到 0.97，是 18 个模型和算法中准确率最高的，因此可以作为最优模型以 AdaBoost（DT）表示。

最优模型 AdaBoost（DT）对 132 批测试集数据的识别，结果 128 批数据识别正确，混淆矩阵见图 8。



NH. 牛黄（cow-bezoar） PZNH. 培植牛黄（cultured cow-bezoar） TPNH. 体外培育牛黄（artificial cow-bezoar） RGNH. 人工牛黄（cow-bezoar cultured *in vitro*）

图 8 最优模型识别测试集样品混淆矩阵

Fig. 8 Confusion matrix of test set predicted by the best model

该图显示了测试集中各类样品真实类别和模型识别预测的类别，可见牛黄样品识别完全正确；54 批培植牛黄中有 1 批错误识别为体外培育牛黄；46 批人工牛黄中有 1 批错误识别为牛黄；28 批体外培育牛黄中有 2 批分别错误识别为牛黄和人工牛黄。

以 132 批测试集数据对最优模型 AdaBoost（DT）进行全面系统的识别预测能力的评估，结果各项数据表现良好，准确率为 0.97，精确率为 0.90，召回率为 0.97，F1 得分为 0.93，ROC 的曲线下面积（AUC）为 1.00。

这些模型评价指标中，识别的精确率结果相对较低。精确率的概念是对于识别预测为正样本中实际为正样本的比例，反映了模型对于正样本识别能力。由图 8 混淆矩阵可知，由于牛黄样本量较少，在模型对于牛黄的识别预测中出现 2 个错误，因此拉低了模型对于整体各牛黄类别的平均精确率。

3.7 关于模型识别预测概率的问题

最优模型 AdaBoost（DT）对测试集识别预测错误的 4 批样品所属各牛黄类别的概率预测结果见表 3。模型识别预测未知样品的类别是基于最大概率类别为准则，即未知样品被模型分别计算得出的各类别的概率值，最终识别预测的类别是其中概率最大的类别。表 3 中，除了 pznh_044 样品被模型以极大概率识别预测为体外培育牛黄外，tpnh_073、tpnh_068、rgnh_004 的各类别概率几乎没有出现极大概率的情况。因此，若将模型对未知样品所被识别预测的类别概率设定为不得低于 7.30×10^{-1}，则所有类别的概率都小于该值设为无法确定，表 3 中有 3 批样品为无法确定类别，而 128 批识别预测正确的样品中有 2 批也为无法确定类别，那么 132 批测试集数据中则有 1 批为识别预测错误，5 批无法确定。可见，若使用类别概率进行牛黄种类的识别判断，可以更加精确地识别预测结果，对于模型不能十分确切识别预测的样品，可以进一步使用其他技术进行佐证判断。此外，概率阈值的设定还可以结合实际使用需要进行灵活调节。

表 3 识别预测错误样品的各类别概率

Tab. 3 Identify the probabilities of each category for predicting incorrect samples

样品（sample）	概率（probability）			
	牛黄（cow-bezoar）	培植牛黄（cultured cow-bezoar）	人工牛黄（cow-bezoar cultured *in vitro*）	体外培育牛黄（artificial cow-bezoar）
pznh_044	7.00×10^{-10}	5.12×10^{-15}	1.44×10^{-9}	1
tpnh_073	4.83×10^{-1}	3.52×10^{-10}	4.34×10^{-1}	8.32×10^{-2}
tpnh_068	4.11×10^{-3}	4.42×10^{-14}	7.28×10^{-1}	2.68×10^{-1}
rgnh_004	4.83×10^{-1}	3.52×10^{-10}	4.34×10^{-1}	8.32×10^{-2}

4 结论

本研究使用 REIMS 技术对牛黄、培植牛黄、人工牛黄和体外培育牛黄进行测定，并应用传统化学计量学以及机器学习领域的多种模型和算法对样品的 REIMS 谱图进行快速识别预测研究，采用先进的理念与算法技术对模型进行超参数优化及不平衡数据的仿真合成，取得了较好的结果。本研究的流程是药物分析研究的实际与机器学习理论的结合，结果表明机器学习相关技术的应用可极大提高药物分析数据的利用率，这一研究流程模式对于药物分析研究具有一定参考意义，尤其对于贵细药物的研究和分析实验成本较高的情况更加具有意义。此外，机器学习模型输出的概率，可以使数据的预测更精确、更科学，可以灵活地在药品实际监管场景下进行使用，更加有利于药品监管工作的优化和细化。

参考文献

[1] 中华人民共和国药典 2020 年版. 一部[S]. 2020: 72
ChP 2020. Vol Ⅰ[S]. 2020: 72

[2] 黄漠然，赵文靖，李晋生，等. 牛黄及其代用品化学成分、分析方法和药理作用研究进展[J]. 药物分析杂志，2018，38(7): 1116
HUANG MR, ZHAO WJ, LI JS, *et al*. Research advance of chemical constituents, analytical methods and pharmacological effects of cow-bezoar and its substitutes[J]. Chin J Pharm Anal, 2018, 38(7): 1116

[3] 胡晓茹，倪景华，孙磊，等. 牛黄及代用品的红外指纹图谱鉴别研究[J]. 中国现代中药，2022，24(3): 438
HU XR, NI JH, SUN L, *et al*. Identification of Bovis Calculus and its substitutes by infrared fingerprint[J]. Mod Chin Med, 2022, 24(3): 438

[4] 石岩，孙冬梅，魏锋，等. 柱前衍生 HPLC 法测定体外培育牛黄中主要胆汁酸类成分[J]. 药物分析杂志，2016，36(11): 2046
SHI Y, SUN DM, WEI F, *et al*. Quantification of main bile acids in *in vitro* cultivated Calculus Bovis by high-performance liquid chromatography with pre-column derivatization[J]. Chin J Pharm Anal, 2016, 36(11): 2046

[5] SHI Y, XIONG J, SUN DM, *et al*. Simultaneous quantification of the major bile acids in artificial Calculus Bovis by high-performance liquid chromatography with precolumn derivatization and its application in quality control[J]. J Sep Sci, 2015, 38(16): 2753

[6] 张程亮，向东，刘东. 牛黄的现代研究(一)：回顾与展望[J]. 医药导报，2017，36(1): 1
ZHANG CL, XIANG D, LIU D. Modern research of Calculus Bovis (First): retrospect and prospect[J]. Her Med, 2017, 36(1): 1

[7] 李喜平，张程亮，刘东. 牛黄的现代研究(四)：药理作用[J]. 医药导报，2017，36(4): 355
LI XP, ZHANG CL, LIU D. Modern research of Calculus Bovis (Fourth): pharmacological effects[J]. Her Med, 2017, 36(4): 355

[8] XIONG J, ZHENG TJ, SHI Y, *et al*. Analysis of the fingerprint profile of bioactive constituents of traditional Chinese medicinal materials derived from animal bile using the HPLC-ELSD and chemometric methods: an application of a reference scaleplate[J]. J Pharm Biomed Anal, 2019, 179: 50

[9] 石岩，王晓伟，魏锋，等. 基于机器学习鉴别牛黄类药材红外光谱的研究[J]. 中国药物警戒，2023，20(2): 140
SHI Y, WANG XW, WEI F, *et al*. Succession medicinal substances of Calculus Bovis with infrared spectroscopy coupled with machine learning methods[J]. Chin J Pharmacovigil, 2023, 20(2): 140

[10] 石岩，李宁，魏锋. 机器学习算法在不同形态浙贝母与湖北贝母的干法 REIMS 指纹图谱鉴别分析中的应用研究[J]. 药物分析杂志，2024，44(1): 134
SHI Y, LI N, WEI F. Research of machine learning in the application of authenticity discrimination of Fritillariae Thunbergii Bulbus and Fritillariae Hupehensis Bulbus in different form with dry-process REIMS fingerpring[J]. Chin J Pharm Anal, 2024, 44(1): 134

[11] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, *et al*. Generative adversarial nets[C]// 27th NeurIPS Montreal Canada, 2014

[12] 常晓，蔡昕，杨光，等. 生成对抗网络在医学图像转换领域的应用[J]. 波谱学杂志，2022，39(3): 366
CHANG X, CAI X, YANG G, *et al*. Applications of generative adversarial networks in medical image translation[J]. Chin J Magn Resonance, 2022, 39(3): 366

[13] 于森，许铮铧. 生成对抗网络医学图像去噪研究综述[J]. 中国生物医学工程学报，2022，41(6): 724
YU M, XU ZH. A review on generative adversarial networks in medical image[J]. Chin J Biomed Eng, 2022, 41(6): 724

[14] STRITTMATTER N, REBEC M, JONES E A, *et al*. Characterization and identification of clinically relevant microorganisms using rapid evaporative ionization mass spectrometry[J]. Anal Chem, 2014, 86(13): 6555

[15] BALOG J, SACHEEN K, ALEXANDER J, *et al*. *In vivo* endoscopic tissue identification by rapid evaporative ionization mass spectrometry (REIMS)[J]. Angew Chem Int Ed Engl, 2015, 54(38): 11059

[16] SONG G, CHEN K, WANG H, *et al*. *In situ* and real-time authentication of *Thunnus* species by iKnife rapid evaporative ionization mass spectrometry based lipidomics without sample pretreatment[J]. Food Chem, 2020, 318: 126504

[17] 刘鸣畅，林继红，刘哲硕，等. 快速蒸发电离质谱技术(REIMS)鉴别肉品[J]. 质谱学报，2020，41(5): 470
LIU MC, LIN JH, LIU ZS, *et al*. Identification of meat by rapid evaporation ionization mass spectrometry (REIMS)[J]. J Chin Mass Spectrom Soc, 2020, 41(5): 470

(本文于 2024 年 5 月 10 日收到)