

doi: 10.12452/j.fxcxb.26041604

基于X-射线衍射技术与机器学习的朱砂硫化汞含量预测及模型可解释性研究

李佳薇^{1,2}, 任小英^{1,2}, 谭乐俊², 隋肖肖^{1,2}, 张梓晗^{1,2}, 林林^{1,2*}

(1. 天津中医药大学 中药学院, 天津 301617; 2. 山东省食品药品检验研究院, 山东省中药标准创新与质量评价工程实验室, 产业技术基础公共服务平台, 山东 济南 250101)

摘要: 朱砂中硫化汞(HgS)含量是质量评价的核心指标。针对传统测定法耗时长、破坏样本且难以满足快速质控需求的问题, 该研究将X-射线衍射(XRD)与多维度机器学习算法结合, 构建了偏最小二乘回归(PLSR)、随机森林(RF)、粒子群优化支持向量回归(PSO-SVR)、长短期记忆网络(LSTM)及卷积神经网络(CNN)预测模型体系。系统考察了多种光谱预处理方法及特征筛选算法, 并引入沙普利加性解释(SHAP)框架解析模型决策逻辑。结果表明, 标准正态变量变换(SNV)预处理结合竞争性自适应重加权采样(CARS)特征筛选后, PSO-SVR与LSTM模型的综合性能最佳。LSTM模型校正集与验证集决定系数(R^2_c 、 R^2_v)分别为0.992 6与0.879 4, 均方根误差(RMSE_c、RMSE_v)为0.120 5与0.411 5, 剩余预测偏差(RPD)达2.879 7; PSO-SVR模型 R^2_v 为0.865 0, RMSE_v为0.435 4, RPD为2.721 4。10批独立样本外部验证 R^2 均大于0.80, 证实模型具备良好的泛化能力与稳健性。SHAP分析进一步揭示, $2\theta=24^\circ\sim 35^\circ$ 与 $50^\circ\sim 60^\circ$ 区间的衍射强度为模型决策的核心特征, 其中 27.92° 、 45.59° 等模型识别的关键衍射角对应的特征与HgS含量呈显著正相关, 算法决策与HgS晶体衍射特征高度一致。XRD联用智能算法可实现朱砂中硫化汞含量的无损、快速、准确预测, 为矿物药快速质量评价及智能化质控体系建设提供了方法学支撑。

关键词: X-射线衍射; 朱砂; 硫化汞; 长短期记忆网络(LSTM); 粒子群优化支持向量回归(PSO-SVR); 快速检测

中图分类号: O657.3 文献标识码: A 文章编号: 1004-4957(2026)09-1953-12

Prediction of Mercury Sulfide Content in Cinnabar Based on X-ray Diffraction Technology and Machine Learning with Model Interpretability

LI Jia-wei^{1,2}, REN Xiao-ying^{1,2}, TAN Le-jun², SUI Xiao-xiao^{1,2}, ZHANG Zi-han^{1,2}, LIN Lin^{1,2*}

(1. School of Chinese Materia Medica, Tianjin University of Traditional Chinese Medicine, Tianjin 301617, China; 2. Shandong Institute for Food and Drug Control, Shandong Engineering Laboratory for Innovation and Quality Evaluation of Traditional Chinese Medicine Standards, Public Service Platform for Industrial Technology Foundation, Jinan 250101, China)

Abstract: The mercury sulfide(HgS) content is the core indicator for the quality evaluation of cinnabar. To address the limitations of traditional determination methods, which are time-consuming, destructive, and inadequate for rapid quality control, this study combined X-ray diffraction(XRD) with multiple machine learning algorithms to develop a predictive modeling system comprising partial least squares regression(PLSR), random forest(RF), particle swarm optimization-support vector regression(PSO-SVR), long short-term memory(LSTM), and convolutional neural network(CNN). Various spectral preprocessing methods and feature selection algorithms were systematically investigated, and the shapley additive explanations(SHAP) framework was introduced to elucidate the decision-making logic of the models. The results indicated that after applying standard normal variate(SNV) preprocessing coupled with competitive adaptive reweighted sampling(CARS) for feature selection, the PSO-SVR and LSTM models exhibited the best overall performance. For the LSTM mod-

收稿日期: 2026-04-16; 修回日期: 2026-05-12

基金项目: 国家重点研发计划-中医药现代化专项(2023YFC3504102)

*通讯作者: 林林, 硕士, 主任药师, 研究方向: 中药质量控制与安全性评价, E-mail: linlinsda@163.com

网络首发日期: 2026-07-03

el, the coefficients of determination for the calibration and validation sets (R_c^2 , R_v^2) were 0.992 6 and 0.879 4, respectively, with root mean square errors ($RMSE_c$, $RMSE_v$) of 0.120 5 and 0.411 5, and a residual predictive deviation (RPD) of 2.879 7. The PSO-SVR model achieved an R_v^2 of 0.865 0, an $RMSE_v$ of 0.435 4, and an RPD of 2.721 4. External validation on 10 independent sample batches yielded R^2 values all exceeding 0.80, confirming the models' strong generalization ability and robustness. SHAP analysis further revealed that the diffraction intensities within the 2θ ranges of 24° – 35° and 50° – 60° were the core features driving model decisions. Specifically, the features corresponding to key diffraction angles identified by the models, such as 27.92° and 45.59° , showed a significant positive correlation with HgS content, demonstrating a high consistency between algorithmic decision-making and the crystal diffraction characteristics of HgS. The integration of XRD with intelligent algorithms enables non-destructive, rapid, and accurate prediction of HgS content in cinnabar, providing methodological support for the rapid quality evaluation of mineral medicines and the construction of intelligent quality control systems.

Key words: X-ray diffraction; cinnabar; HgS; long short-term memory (LSTM); particle swarm optimization-support vector regression (PSO-SVR); rapid detection

朱砂(*Cinnabaris*), 又称丹砂或辰砂, 来源于硫化物类矿物辰砂族辰砂, 为常用的一味矿物药, 主要药理作用为“安神”, 常用于治疗失眠、焦虑、癫痫^[1-2]。朱砂的主要化学成分为硫化汞(HgS)^[3], 其含量是评价朱砂质量的关键指标。尽管 HgS 在胃肠道中溶解度极低, 但朱砂的毒性风险主要源于原料中残留的可溶性汞盐、游离汞等杂质, 以及长期服用后经肠道微生物转化生成的可溶性汞形态^[4]。因此, 严格控制可溶性汞等杂质限量, 确保 HgS 高纯度, 是降低毒副风险、保障临床安全用药的关键措施。

《中国药典》(2025 版)采用滴定法作为朱砂中硫化汞含量测定的方法^[5], 不仅操作繁琐、耗时长, 还存在一定的安全隐患, 已经难以满足现代药品质量控制对“快速、无损、便捷”的迫切需求。随着现代分析技术的不断进步, 光谱技术以其高效、无损等优势, 在药品、食品等领域展现出巨大的应用潜力。X 射线衍射(XRD)技术能够直接获取物质的晶体结构信息, 在复杂多相矿物体系的物相鉴定和含量测定中具有独特优势, 已经成为分析矿物药的重要工具^[6-7], 并广泛用于其质量控制^[8-11]。XRD 图谱本身具有多维度、高信息密度的特征, 其数据潜力的深度释放亟需智能化分析手段的支撑。

近年来, 机器学习方法在复杂数据解析与特征挖掘方面展现出强大能力^[12-13], 为矿物药目标物含量的快速、精准预测提供了新的技术途径。矿物药基体复杂、批次差异显著, 若模型决策逻辑不透明, 其预测结果在不同样品种类、含量区间或数据处理流程下易产生适应性偏差, 难以直接转化为通用质控标准^[14]。为此, 引入可解释性模块成为提升算法实用价值的关键环节。基于合作博弈论的沙普利加性解释(SHAP)方法, 是目前公认的模型可解释性分析的有效方法, 已成功应用于机器学习和深度学习的模型解释研究^[15-17], 为矿物药无损检测模型的可靠性验证与标准化应用提供了新的思路。

为建立朱砂中硫化汞(HgS)含量的快速、精准、无损定量检测方法, 本研究将机器学习算法与 XRD 技术联用, 系统构建了覆盖传统线性、非线性至深度学习架构的多维度预测模型体系。研究系统考察了光谱预处理与特征波长筛选策略的协同效应, 横向对比了偏最小二乘回归(PLSR)、粒子群优化支持向量回归(PSO-SVR)、随机森林(RF)、长短期记忆网络(LSTM)及卷积神经网络(CNN)等算法的泛化性能与预测精度。在此基础上, 引入 SHAP 可解释性分析框架, 定量解析优选模型的内部决策机制与关键衍射特征贡献规律。本研究旨在突破传统 XRD 定量分析的效率与可解释性瓶颈, 为朱砂药材的智能化质量评价及矿物药无损检测体系的标准化建设提供可靠的方法学支撑与理论依据。上述完整分析流程详见图 1。

1 实验部分

1.1 样本收集情况

共收集到朱砂 105 批, 产地涉及贵州、陕西、湖南、广东、甘肃。其中包括贵州铜仁万山区、湖南凤凰、沅陵等朱砂道地产区, 另采集吉尔吉斯斯坦流通样本 2 份。所采集的样本经山东省食品药品检验研究院中药科主任药师林林鉴定, 均为硫化物类矿物辰砂族辰砂。将收集到的朱砂使用玛瑙研钵

研磨, 过100目筛, 置于避光干燥处密封保存。

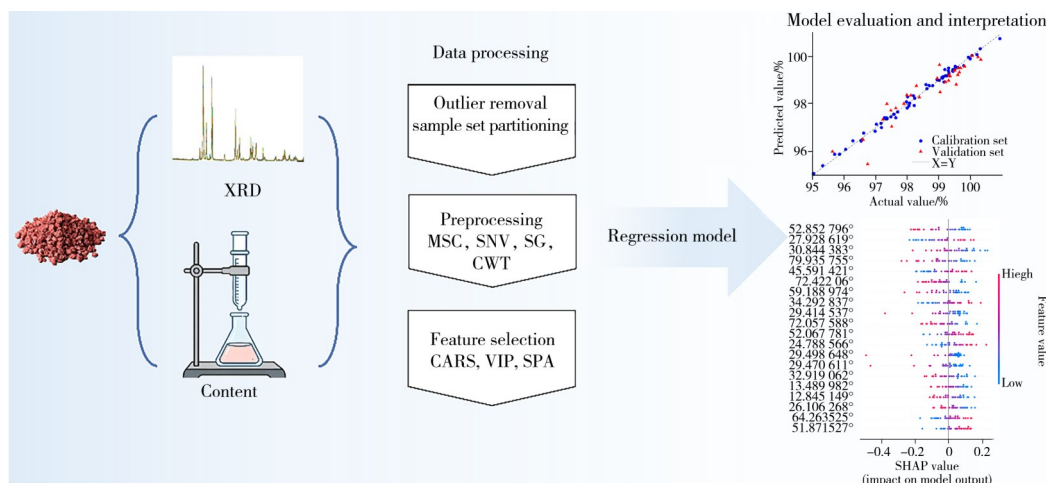


图1 基于XRD与机器学习联用的朱砂硫化汞定量分析建模流程图

Fig. 1 Flowchart of the modeling workflow for quantitative analysis of HgS in cinnabar based on XRD combined with machine learning

1.2 硫化汞含量测定

根据《中国药典》2025版中朱砂项下的含量测定方法, 取朱砂粉末约0.03 g, 用硫氰酸铵滴定液(0.1 mol/L)滴定。每1 mL硫氰酸铵滴定液(0.1 mol/L)相当于11.63 mg的硫化汞。为确保实验数据的准确性, 每批样本平行测定两次, 取平均值作为最终含量参考值。

1.3 XRD光谱数据采集

使用EVASTAR E2便携式X射线衍射仪(苏州伊凡智通科技有限公司)进行样品光谱信息采集。取朱砂样品粉末少量, 置于载样板凹槽内, 使用载玻片轻压, 确保样品在凹槽内均匀紧密分布。X射线衍射仪采集参数设置: 电压35 kV, 电流1 mA, 扫描速度 $10^{\circ} \cdot \text{min}^{-1}$, 扫描角度为 $10^{\circ} \sim 80^{\circ}$, 采用Cu-K α 辐射源, 步长为 0.02° , 静态扫描时间为0.07 s。每个光谱产生2 497个数据点。

1.4 数据处理

1.4.1 异常样本处理与数据划分 光谱采集过程中会因仪器本身噪音或操作问题产生一定的实验误差, 从而导致某个样品光谱偏离整体分布, 若将其纳入模型构建过程, 会对模型训练过程产生影响, 最终导致模型的预测精度受损^[18-19]。因此, 本研究采用基于最小协方差行列式(MCD)思想的迭代加权稳健主成分分析(Robust PCA)进行异常样本诊断: 首先以中位数进行稳健中心化, 然后通过迭代加权估计稳健协方差矩阵, 计算各样本的稳健马氏距离, 并与F分布临界值比较以识别异常; 最终在PC1-PC2得分图上绘制95%置信椭圆, 将椭圆外的样本判定为异常值并予以剔除。

为了确保划分的样本具有代表性, 本研究采用基于联合X-Y距离的样本集划分法(SPHY), 将样本以2:1划分为校正集和验证集。SPHY是一种融合X-Y双维度信息的样本划分方法, 通过标准化X的欧氏距离与Y的绝对差值构建综合距离矩阵, 迭代优选最具代表性的样本构建校正集, 可以有效减小小样本研究中因Y值分布不均导致的模型评估偏差^[20]。

1.4.2 光谱数据预处理方法 合理的数据预处理能有效消除噪声、基线漂移、光散射等无关信号的干扰, 并增强光谱有效特征^[21]。本研究分别考察了多元散射校正(MSC)、标准正态变量变换(SNV)、Savitzky-Golay平滑微分(SG)和连续小波变换(CWT)4种预处理方法, 从不同角度对原始光谱进行优化, 其中SG设置窗口为15, 通过3阶多项式加一阶导数进行运算^[22]; CWT使用sym2小波, 在尺度30下对光谱信号进行小波变换。

1.4.3 特征波长选择方法 光谱数据通常具有超高维度, 直接将其用于模型训练会导致训练效率降低、模型解释性变差等问题, 并且原始光谱中还存在与样本属性无关的吸收波段, 这些冗余的波段会导致模型学习的虚假关联, 最终导致预测出现偏差^[23]。因此光谱在经预处理后还需进行特征波长选择, 以增强“信号-目标”关联性, 规避“维度灾难”, 提升模型的可解释性。本实验采取3种特征选择方

法,包括竞争性自适应重加权采样(CARS)、连续投影算法(SPA)和变量重要性投影(VIP),具体原理数学基础以及相关参数设置见表 1。

表 1 3 种特征选择方法对比及参数设置

Table 1 Comparison and parameter settings of three feature selection method

Model	Basic principle	Mathematical foundation	Parameter settings
CARS	模拟达尔文进化论中的“适者生存”原理,通过自适应重加权 和指数衰减函数竞争性选择特征	蒙特卡洛采样+PLS 回归+ 指数衰减函数	蒙特卡洛次数 $N=30$; 频率阈值 threshold=0.2
SPA	基于向量投影分析,最小化变量间的共线性,选择信息量最大 且冗余度最小的特征子集	向量投影运算+ 共线性最 小化	阈值设置 0.75; 5 折交叉验证; 初始 化次数为 1 000 次
VIP	基于 PLS 模型,衡量每个变量对模型解释能力的贡献度	PLS 回归+变量权重计算	PLS 主成分数为 2; 阈值设置为 1; 迭代次数为 1 500 次

1.5 定量预测模型构建

预测模型的合理选取,是实现基于 XRD 光谱的样本成分精准定量预测的核心环节。为筛选出最适配本研究数据特征与预测目标的最优算法,本研究选取 PLSR、PSO-SVR、RF、LSTM 和 CNN 五种代表性模型开展对比分析。

1.5.1 PLSR 模型 PLSR 最大潜在变量数 $LV=15$,交叉验证折数 $K=15$,通过计算交叉验证的均方根误差,并寻找误差最小时对应的潜变量数,自动确定 PLSR 模型的最佳复杂度。

1.5.2 PSO-SVR 模型 SVR 的预测性能高度依赖惩罚因子 C 、核函数参数 γ 与不敏感损失系数 ε 的取值,为获取最优超参数组合,本文采用粒子群优化算法(PSO)对上述 3 个关键参数进行全局寻优。粒子群优化算法是一种模拟鸟群觅食行为的群体智能优化算法^[24]。在 PSO 寻优过程中,每组待优化的超参数对应搜索空间中的一个“粒子”,所有粒子以特定速度在搜索空间内迭代更新,并通过跟踪个体历史最优位置与种群全局最优位置调整自身运动轨迹。为均匀化参数搜索空间、提升算法收敛效率,本文采用对数变换对超参数进行编码;以校正集 5 折交叉验证得到的平均均方根误差(RMSE)作为适应度函数,评价粒子对应参数组合的优劣,适应度值越小,代表该组超参数对应的模型泛化性能越优。本文 PSO 算法的关键参数设置如下:粒子种群规模 $N=30$,最大迭代次数 $T=100$;超参数的对数空间搜索范围设定为: $\log C \in [0, 1]$, $\log \varepsilon \in [-4, -1]$, $\log \gamma \in [-4, -1]$ 。

1.5.3 RF 模型 RF 模型采用 100 棵 CART 决策树为基学习器构建集成回归模型,通过 Bootstrap 自助采样与特征随机选择机制防止过拟合,以均方误差为分裂准则进行训练,设置随机种子为 42 以保证结果可复现。

1.5.4 深度学习模型 LSTM 模型采用单层 50 神经元结构,使用 0.2 的 Dropout 防止过拟合,以 Adam 优化器和 MSE 损失函数进行训练,设置批量大小为 32 并采用早停机制(耐心值 10)控制训练过程。

CNN 模型由 3 个卷积块、1 个最大池化层、1 个自适应平均池化层及两层全连接层构成。训练采用 AdamW 优化器与均方误差损失,结合学习率衰减、梯度裁剪及 Dropout 正则化,并根据验证集损失保存最优模型。

1.6 模型评估

基于决定系数(R^2)、均方根误差(RMSE)、剩余预测偏差(RPD)评估定量模型的准确性和鲁棒性。评估指标的定义如下:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (1)$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (2)$$

$$\text{RPD} = \frac{\text{SD}(y)}{\text{RMSE}} \quad (3)$$

其中 y_i 表示第 i 样本的真实值, $\hat{y}_i$ 表示第 i 个样本的预测值, $\bar{y}$ 表示所有样本真实值的平均值, n 代表样本数量, $\text{SD}(y)$ 表示所有样本真实值的标准差,反映真实数据本身的离散程度; RMSE 反映预测值与真

实值的偏差大小; RPD 数值越大, 通常代表模型的预测能力越强, 一般认为 $RPD > 2$ 时模型性能较优。

1.7 模型可解释性分析方法

可解释机器学习技术, 为深度解析模型的建模逻辑与决策机制提供了体系化的实用工具。其中, SHAP 作为经典的可解释性分析方法, 以博弈论中的夏普利值为核心理论与计算基础, 通过量化各特征对模型预测结果的贡献程度, 将黑箱模型的复杂预测过程拆解为各特征贡献值的线性加和形式, 为模型提供可解释性支撑^[25]。SHAP 方法具备优异的跨模型适配性, 可兼容各类机器学习算法, 目前已被广泛应用于复杂模型决策过程的解析与归因分析。本研究基于 SHAP 分析框架, 依托 Python SHAP 算法库, 以 SHAP 值为核心量化指标, 评估了入选波数在最优预测模型中的贡献程度, 并对贡献度最显著的前 20 个特征波数开展了重点分析。

2 结果与讨论

2.1 原始光谱分析

105 批朱砂样品的 X 射线衍射原始叠加图谱如图 2 所示。采用 Jade 9.0 软件对样品 XRD 图谱进行物相定性分析, 通过与 ICDD PDF 标准卡片 (97-007-0054) 比对 (图 3), 完成硫化汞物相的指认, 并确定其对应的特征衍射峰位。

2.2 数据预处理

为了验证模型是否具有良好的泛化能力, 本研究将 105 批样本划分为校正集 (R_c^2)、验证集 (R_v^2)、独立预测集 (R_p^2) 3 部分。首先, 随机抽取 10 批样本作为独立预测集, 用于模型训练完成后的外部验证; 剩余 95 批样本用于回归模型的构建。建模前, 采用稳健 PCA 对 95 批样本进行异常值检测, 共识别出 5 个离群样本 (Z9、Z15、Z44、Z76、Z81), 如图 4 所示, 离群样本易导致模型在训练过程中出现回归系数偏移或过拟合现象, 进而降低预测精度与稳定性; 剔除后可有效提升模型的鲁棒性与泛化性能。异常样本剔除后, 剩余 90 批样本按 2:1 的比例划分校正集 (60 批) 和验证集 (30 批)。由表 2 可知, 校正集 HgS 的含量范围为 95.04%~100.94%, 验证集 HgS 的含量范围为 95.64%~100.34%, 校正集的样本范围包含了验证集。图 5 为样本主成分空间的分布情况, 从图中可以看出, 验证集样本均匀分布于校正集所覆盖的特征空间内, 未见明显的分布偏差或区域缺失。由此可表明该方法划分的数据合理, 为建立良好的模型奠定了基础。

2.3 光谱预处理方法筛选

本研究比较了原始光谱与 MSC、SNV、SG 及 CWT 四种预处理方法对预测模型的影响。表 3 展示了每个模型最优的预处理结果, 结果表明, 不同光谱预处理方法对 5 种模型拟合效果与泛化性能的影响

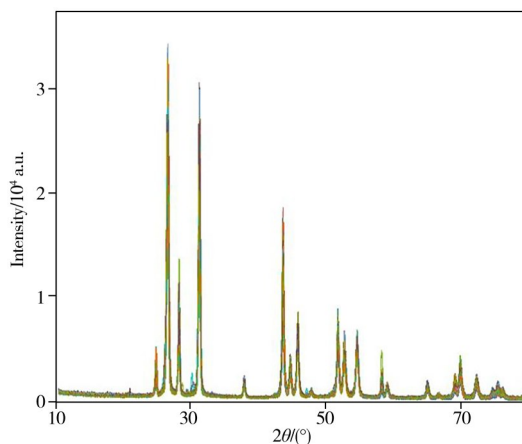


图2 105批朱砂样品的原始XRD图谱

Fig. 2 Original XRD patterns of 105 batches of cinnabar samples

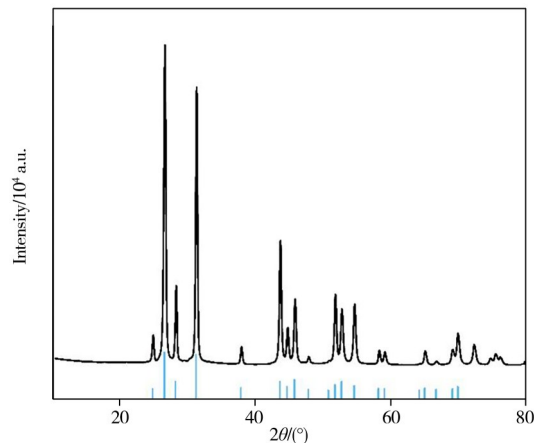


图3 基于XRD的硫化汞物相表征及ICDD PDF #97-007-0054标准卡片(蓝色线)比对

Fig. 3 Phase characterization of mercuric sulfide by XRD and comparison with ICDD standard PDF card 97-007-0054 (blue line)

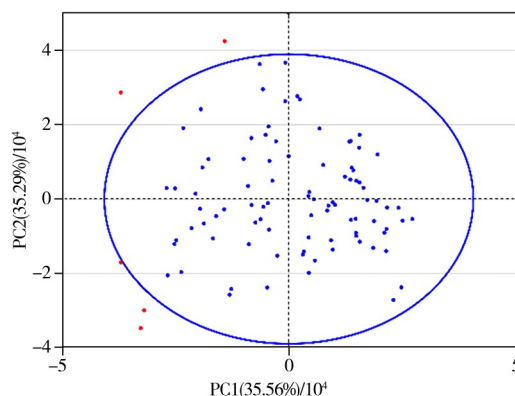


图4 主成分分析图

Fig. 4 Principal component analysis (PCA) score plot

存在显著差异, 未出现可适配所有模型、实现预测性能稳定提升的预处理方案。整体来看, 无论是原始光谱直接建模, 还是经上述预处理后建模, 模型的预测精度与泛化能力均存在明显的提升空间, 仅依靠光谱预处理难以充分剔除光谱数据中的冗余信息、消除非目标因素与无关变量的干扰, 仍需通过特征选择方法进一步筛选核心有效特征。

表2 朱砂样本划分情况
Table 2 Division of cinnabar samples

Data	Number of sample	Maximum/%	Minimum/%
Calibration set	60	100.94	95.04
Validation set	30	100.34	95.64
Test set	10	99.52	95.33

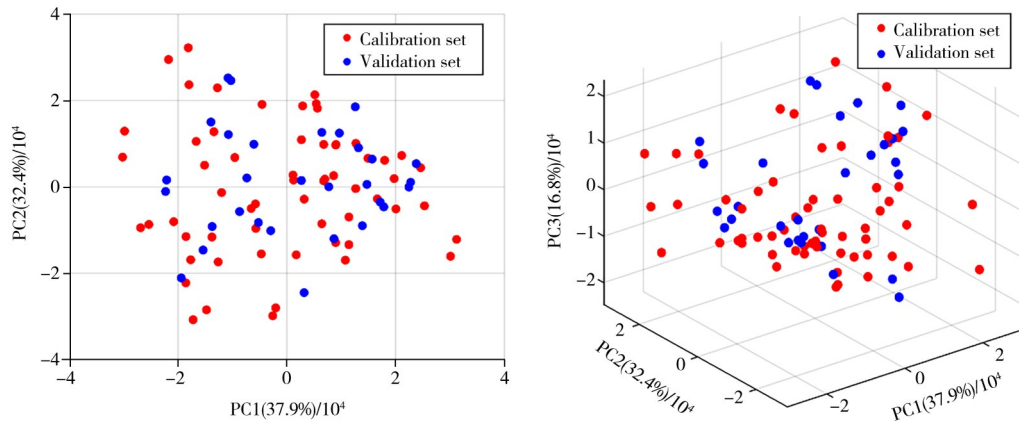


图5 SPXY 样本划分在主成分空间的分布验证

Fig. 5 Distribution verification of SPXY sample partitioning in the principal component space

表3 不同模型的最优预处理结果
Table 3 Modeling results of different preprocessing methods

Model	Preprocessing	Calibration set		Validation set		RPD
		R_c^2	RMSE _c	R_v^2	RMSE _v	
PLSR	RAW	0.629 4	0.851 4	0.517 1	0.861 9	1.636 3
	SNV	0.745 3	0.705 9	0.525 3	0.895 4	1.575 0
SVR	RAW	0.995 1	0.098 2	0.524 6	0.817 1	1.711 5
	SNV	0.995 1	0.098 1	0.610 5	0.739 6	1.891 0
RF	RAW	0.909 0	0.421 8	0.500 4	0.837 6	1.669 6
	CWT	0.909 1	0.421 6	0.527 8	0.814 3	1.717 4
LSTM	RAW	0.967 8	0.251 0	0.592 4	0.756 5	1.848 6
	CWT	0.745 3	0.705 9	0.525 3	0.895 4	1.575 0
CNN	RAW	0.620 4	0.861 6	0.059 1	1.149 4	1.048 6
	MSC	0.752 7	0.695 5	0.330 6	0.969 5	1.243 2

2.4 特征波长选择结果

基于预处理后的 XRD 数据, 本研究进一步探究了 CARS、SPA、VIP 三种特征选择方法对 5 种预测模型性能的影响。不同特征选择算法对模型性能的优化程度与适配性存在显著差异。整体而言, CARS 算法的特征筛选与性能优化效果最为突出, 在多数模型中实现了预测精度与泛化性能的跃升。如图 6 所示, CARS 算法对应曲线在 R^2 和 RPD 轴上明显外延, 在 RMSE 轴上明显内敛, 表明该算法能剔除谱图冗余并保留关键预测信息。相比之下, VIP 和 SPA 算法逊于 CARS, 在雷达图中呈现收缩趋势。进一步结合表 4 数据可见, 经 CARS 特征筛选后, PLSR、SVR、RF、LSTM、CNN 模型的验证集 R_v^2 和 RMSE 均得到显著改善, 模型综合性能全面优于基于全谱数据构建的对应模型, 其中经过 SNV+CARS 处理后的 PLSR、SVR、LSTM 模型校正集的 R_c^2 均达到 0.90 以上, 验证集 R_v^2 均在 0.81 以上, 校正集与验证集性能指标差异小, 模型具有良好的泛化性能, 未出现明显过拟合现象。

该研究结果证明, CARS 算法可高效锁定与待测属性高度相关的特征波长, 剔除无关变量和冗余信息, 在简化模型结构的同时, 显著增强模型的拟合精度与泛化稳健性。

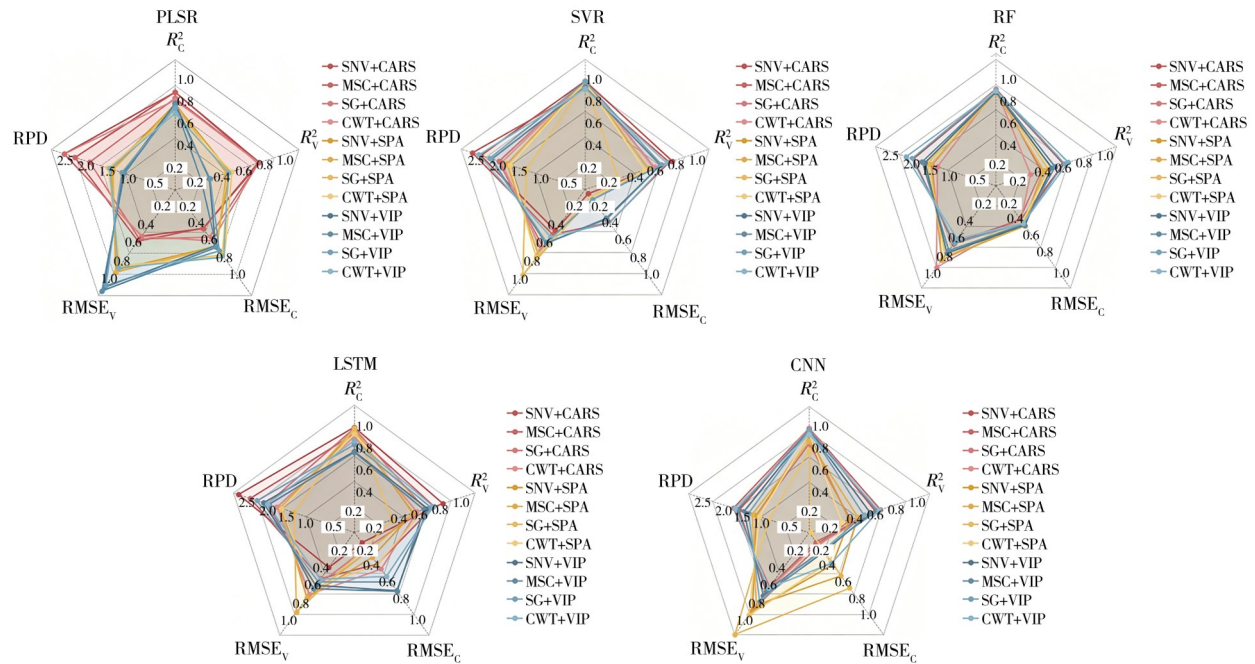


图 6 各模型在不同预处理与特征选择方法组合下的综合性能评价

Fig. 6 Comprehensive performance evaluation of each model under different combinations of preprocessing and feature selection methods

表 4 各模型特征选择最优预测结果

Table 4 Optimal prediction results of feature selection for each model

Model	Preprocessing	Feature selection	Calibration set		Validation set		RPD
			R_C^2	RMSE _C	R_V^2	RMSE _V	
PLSR	SNV	CARS	0.901 4	0.439 2	0.816 4	0.525 2	2.685 4
	SG	SPA	0.720 0	0.740 0	0.521 1	0.866 5	1.627 7
	CWT	VIP	0.703 6	0.761 4	0.515 4	0.887 9	1.588 4
SVR	SNV	CARS	0.998 8	0.048 5	0.843 6	0.468 6	2.528 7
	SNV	SPA	0.995 3	0.092 3	0.612 7	0.737 4	1.896 4
	CWT	VIP	0.924 6	0.384 0	0.790 3	0.542 7	2.577 1
RF	MSC	CARS	0.919 0	0.398 1	0.681 4	0.668 9	2.090 7
	CWT	SPA	0.907 2	0.421 6	0.546 6	0.797 9	1.752 8
	CWT	VIP	0.902 3	0.437 1	0.721 6	0.625 3	2.236 7
LSTM	SNV	CARS	0.992 6	0.120 5	0.879 4	0.411 5	2.879 7
	CWT	SPA	0.928 6	0.373 8	0.695 4	0.654	1.811 8
	CWT	VIP	0.863 9	0.516 0	0.785 9	0.548 3	2.161 4
CNN	MSC	CARS	0.994 1	0.107 4	0.709 9	0.638 3	1.888 4
	MSC	SPA	0.868 5	0.507 2	0.484 0	0.851 2	1.415 9
	CWT	VIP	0.933 6	0.360 4	0.699 7	0.649 4	1.856 0

2.5 不同定量预测模型的功能对比及评价

基于表4所示的最优预处理与特征选择方案，5种建模架构的内在泛化能力呈现显著差异。PLSR作为线性基准，虽训练拟合能力精度最低，但校正集与验证集性能相差较小，验证集 R_V^2 达0.816 4，RPD为2.685 4，具备一定的定量预测能力，可作为简化建模的备选。相比之下，RF与CNN模型的泛化性能均不理想。RF模型的验证集 R_V^2 为0.721 6，RPD为2.236 7；CNN模型虽训练拟合能力优异，但验证集 R_V^2 仅为0.709 9，RPD仅1.888 4，二者均未达到可靠定量分析的基本要求，无法用于未知样本的定量预测。CNN泛化能力受限的关键原因在于其网络特性：作为高容量深度架构，CNN高度依赖大规模样本进行特征抽象与权重优化，而本研究总样本量仅90例，模型易过度记忆训练特征，导致训练-验证性能出现显著断层。SVR模型在校正集上拟合精度极高，但验证集 R_V^2 较校正集下降超0.15，存在一定的过拟合风险。LSTM模型在所有模型中表现最优，验证集 R_V^2 和RPD均为最高，分别达到0.879 4和2.879 7，RMSE_V低至0.411 5，该优势主要来源于LSTM对序列数据结构的适配性，可有效捕捉特征维度的长程依赖关系。

各模型在最优参数组合下的预测值-真实值散点分布如图 7 所示。整体来看, RF 与 CNN 模型的散点呈现明显发散趋势, 偏离对角线较远, 印证了其泛化能力的不足。相比之下, LSTM 模型的校正集与验证集样本均紧密分布于对角线两侧, 未见显著系统性偏差或异常离散, 在实现高精度拟合的同时有效抑制了过拟合, 展现出最优的未知样本预测能力与模型稳定性。与任小英等^[26]在大青盐 XRD 定量研究中报道的同架构最优模型(验证集 $R^2_v=0.725$)相比, 本文 LSTM 模型的预测精度提升显著, 且同时开展了更全面的主流建模架构对比分析, 为矿物药 XRD 快速定量方法的算法选型提供了更充分的参考依据。

需指出的是, SVR 模型的参数采用手动调试, 难以获得最优参数组合, 因此, 模型精度仍有较大提升空间。为进一步优化模型性能, 下文将采用 PSO(粒子群优化)算法对 SVR 模型的关键参数进行自适应寻优, 构建性能更优的 SNV+CARS+PSO-SVR 定量预测模型。

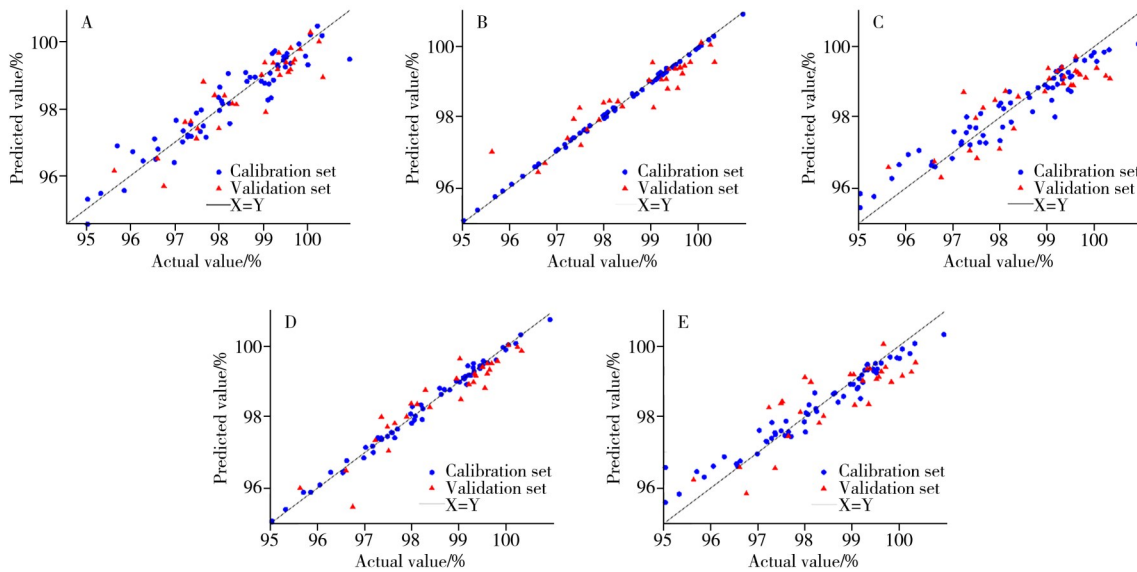


图 7 5 种模型预测值与真值的散点图

Fig. 7 Scatter plots of predicted versus reference values for five models

A. PLSR model; B. SVR model; C. RF model; D. LSTM model; E. CNN model

2.6 粒子群算法优化 SVR 超参数

支持向量回归在处理小样本、高维数据方面具有独特优势, 适用于本研究的数据特点。为增强模型泛化能力, 并进一步挖掘 SVR 在回归任务中的性能优势, 本研究引入粒子群优化算法对 SVR 的惩罚参数 C 、核函数参数 γ 及不敏感损失系数 ε 进行自动寻优, 以实现模型参数的最优配置。优化后的 PSO-SVR 模型性能指标见表 5。与优化前相比, 验证集 R^2_v 从 0.843 6 提升至 0.865 0, $RMSE_v$ 从 0.468 6 降至 0.435 4; 虽然校正集指标较之前相比变差, 但这并非是模型拟合性能的下降, 而是通过 PSO 全局寻优合理约束了模型复杂度, 有效避免了对校正集局部特征与随机噪声的过度学习, 从而使得校正集和验证集的差距减小, 模型泛化能力增强, 同时 RPD 也得到提升。从散点图 8A 中可以观察到, 优化后的校正集和验证集数据点整体分布更集中, 尤其是验证集样本偏离理想预测线的幅度明显降低。

表 5 PSO-SVR 模型预测结果

Table 5 Prediction results of the PSO-SVR model

Model	Calibration set		Validation set		RPD	SVR parameter		
	R^2_c	$RMSE_c$	R^2_v	$RMSE_v$		C	γ	ε
PSO-SVR	0.961 1	0.275 8	0.865 0	0.435 4	2.721 4	9.99	0.002 9	0.065

图 8B 展示了 PSO 算法优化 SVR 超参数时的收敛过程。由收敛曲线可知, 前 15 次迭代内算法收敛迅速, 交叉验证 RMSE 从初始的 0.678 降至 0.618 左右, 初步定位到较优参数区域; 第 15~20 次迭代期间, 曲线下降趋势明显放缓, 搜索过程由全局探索逐步过渡至局部开发; 20 次迭代后曲线稳定, 直至 100 次迭代结束时稳定在 0.616 左右。该收敛特性表明, 对数空间编码与动态惯性权重策略有效调节了粒子的搜索步长, 使算法在前期保持较大探索范围, 后期自动收缩以进行精细寻优, 从而兼顾了全局

与局部搜索能力。在此最优参数组合下, PSO-SVR 模型的预测误差明显降低, 其泛化能力与预测精度达到了与 LSTM 模型相当的水平。

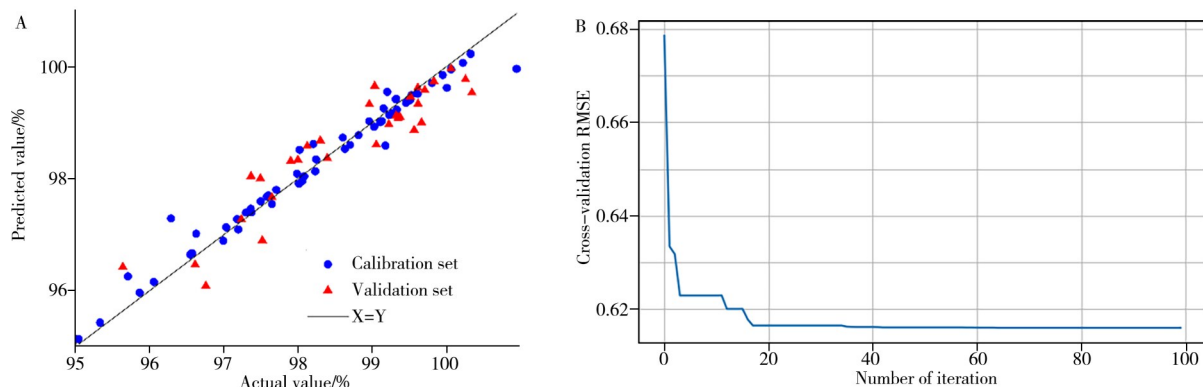


图8 PSO-SVR 校正集和验证集散点图(A); PSO收敛曲线图(B)

Fig. 8 Scatter plots of calibration and validation sets for PSO-SVR(A); convergence curve of PSO(B)

2.7 模型外部验证

为了评估 PSO-SVR 和 LSTM 模型的泛化能力, 本研究选取独立的 10 批朱砂样本作为预测集进行外部验证。如图 9 所示, 两个模型的含量预测值与真值绝大多数高度吻合, 仅个别样本存在微小偏差。表 6 列出了模型在预测集上的评价指标。结果显示, PSO-SVR 与 LSTM 模型的判定系数(R^2)均高于 0.80, 均方根误差(RMSE)均低于 1.0, 表明两模型在独立样本上具有较高的预测精度。结合前述验证结果, 所构建的定量模型表现出良好的泛化能力与预测稳定性, 可为朱砂含量的快速、准确测定提供有效的方法学支持。

相较于近红外光谱、拉曼光谱等常用快速检测技术, 本研究采用的 XRD 技术基于晶体结构特征进行定量分析, 对朱砂这类晶型单一、成分明确的矿物药具有更高的特异性, 且不受样品颜色、基质荧光效应及水分含量的干扰, 检测结果的重现性更好。

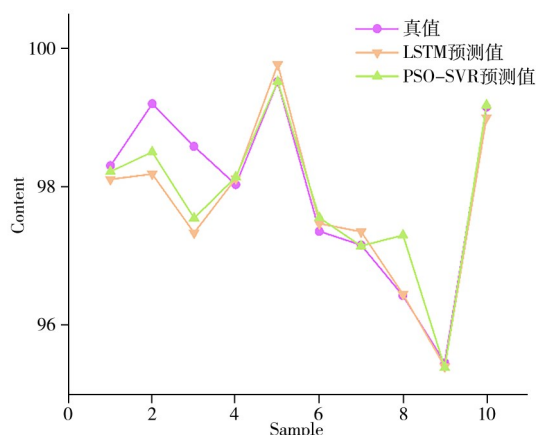


图9 预测集实测值与预测值相关性图

Fig. 9 Correlation between measured and predicted values of the prediction set

表6 两个最优模型的预测集结果

Table 6 Prediction results of the prediction set for the two optimal models

Model	Optimal combination	R^2_p	RMSE _p	RPD
PSO-SVR	SNV+CARS	0.845 9	0.488 3	2.547 6
LSTM	SNV+CARS	0.820 6	0.526 9	2.361 0

2.8 模型可解释性分析结果

为阐明模型预测机制并识别对 HgS 含量预测贡献最为显著的 XRD 衍射特征, 本研究采用 SHAP 方法对 SNV+CARS+LSTM 和 SNV+CARS+PSO-SVR 两个最优模型进行可解释性分析, 量化各特征的贡献度并排序, 从而揭示关键特征变量与目标含量之间的内在关联^[27]。

2.8.1 LSTM 模型可解释分析结果 基于 SHAP 方法计算得到各衍射特征对 LSTM 模型 HgS 含量预测的平均绝对 SHAP 值(图 10A), 该指标可量化特征对模型输出的平均贡献程度, 数值越大代表特征对预测结果的重要性越高。结果显示, $2\theta=52.85^\circ$ 处的衍射特征贡献度最高, 平均绝对 SHAP 值接近 0.10, 是 LSTM 模型预测 HgS 含量的核心主导特征; 其次为 27.92° 、 30.84° 、 79.93° 处的衍射特征, 其平均绝对 SHAP 值均高于 0.09, 共同构成模型的高贡献特征集。从特征的衍射角分布来看, 模型筛选出的高贡献特征主要集中于 3 个区间: $2\theta=24^\circ\sim35^\circ$ 区间, 包含 27.92° 、 30.84° 、 34.29° 、 29.41° 、 24.78° 、 29.49° 、 29.47° 、 32.91° 、 26.10° 共 9 个核心特征, 该区间为朱砂 HgS 的强特征衍射峰集中区域; $2\theta=$

50°~60°区间, 包含 52.85°、59.18°、52.06°、51.87°共 4 个特征, 对应 HgS 的次级特征衍射峰; 其余高贡献特征如 45.59°、64.26°、79.93° 则分别分布于 40°~50°、60°~80° 区间。

SHAP 蜂群图 (图 10B) 揭示了单一样本的特征值对模型输出的影响方向与影响程度, 27.92°、45.59°、34.29°、52.06°、24.78°、64.26°、51.87° 处的衍射特征与模型预测值呈正相关, 其高衍射强度集中在正 SHAP 区, 低衍射强度集中在负 SHAP 区, 表明该类特征的衍射强度越高, 模型对 HgS 含量的预测输出值越大。

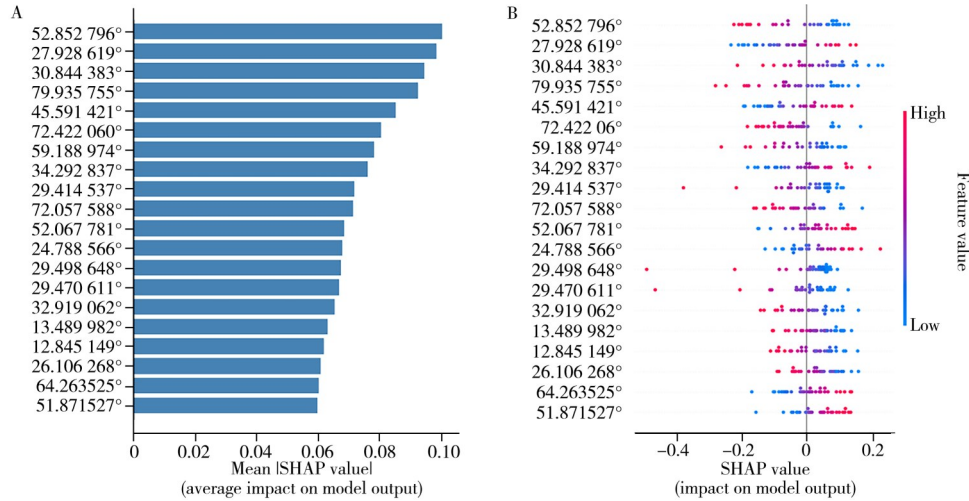


图 10 朱砂硫化汞含量 LSTM 预测模型的 SHAP 可解释性分析

Fig. 10 SHAP interpretability analysis of the LSTM prediction model for mercury sulfide content in cinnabar

A. SHAP bar plot; B. SHAP beeswarm plot (A. SHAP 柱状图; B. SHAP 蜂群图)

2.8.2 PSO-SVR 模型可解释分析结果 针对 PSO-SVR 模型的 SHAP 分析结果显示, HgS 核心贡献的衍射特征集中于 $2\theta=24^\circ\sim35^\circ$ 区间, 包含 32.91°、29.44°、24.78°、30.87°、27.92°、26.02°、29.41°、32.52°、29.47°、31.06°、26.13°、30.90° 共 12 个特征, 与朱砂 HgS 的主特征衍射峰范围完全匹配; 其余核心特征分别分布于 $2\theta=40^\circ\sim50^\circ$ 区间 (44.55°、45.59°)、 $50^\circ\sim60^\circ$ 区间 (52.85°、59.18°、52.06°), 覆盖了 HgS 的次级特征衍射峰范围。从特征贡献度的层级差异来看 (图 11A), 与 LSTM 模型分析结果一致的是, $2\theta=52.85^\circ$ 处的衍射特征仍为贡献度最高的核心变量, 其平均绝对 SHAP 值超过 0.14, 远高于其余所有特征, 是决定 PSO-SVR 模型预测结果的最核心变量; 32.91°、59.18°、29.44° 处的特征为次级高贡献特征。SHAP 蜂群图 (图 11B) 中可以进一步看出, 52.06°、24.78°、44.55°、15.28°、27.92°、45.59° 处的衍射特征与 HgS 含量预测值呈现稳定正相关。

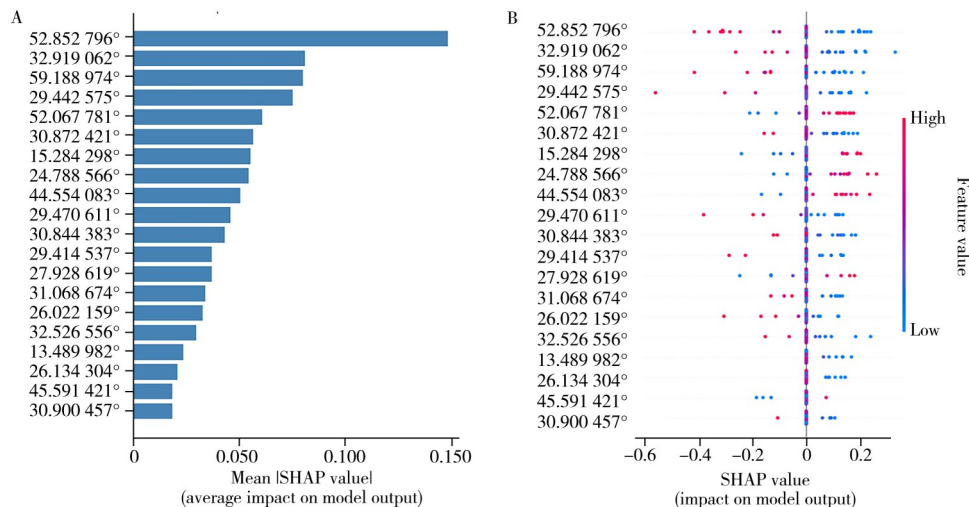


图 11 朱砂硫化汞含量 PSO-SVR 预测模型的 SHAP 可解释性分析

Fig. 11 SHAP interpretability analysis of the PSO-SVR prediction model for mercury sulfide content in cinnabar

A. SHAP bar plot; B. SHAP beeswarm plot (A. SHAP 柱状图; B. SHAP 蜂群图)

综合两个最优模型的SHAP分析结果可知, LSTM模型与PSO-SVR模型的核心贡献特征具有高度一致性, 二者均将 $2\theta=24^{\circ}\sim 35^{\circ}$ 、 $50^{\circ}\sim 60^{\circ}$ 区间的HgS特征衍射峰作为预测的核心依据, 且 27.92° 、 24.78° 、 45.59° 、 52.06° 等特征在两个模型中均与HgS含量呈正相关。该分布与HgS标准晶体学衍射图谱吻合, 表明模型有效捕捉到了目标成分的物理衍射规律, 预测机制具有明确的物相依据。同时, 两个模型的特征学习机制存在差异: LSTM模型对高角度衍射特征(如: 64.26° 、 79.93°)的利用率更高, 特征贡献度的梯度变化较缓, 体现了循环神经网络在处理长序列光谱时全局上下文关联的优势; 而PSO-SVR模型的单特征贡献集中度较高, 对中低角度的特征衍射峰更为敏感, 反映了经粒子群优化的支持向量机模型对局部非线性特征的强拟合能力。

3 结 论

针对传统朱砂(HgS)检测方法前处理繁琐、耗时长、具有破坏性等局限, 本研究构建了基于XRD与可解释机器学习的快速定量预测模型。结果表明, 单一光谱预处理对模型性能的提升有限, 引入特征选择后预测精度显著改善。其中, CARS算法在剔除衍射图谱冗余信息、保留关键预测特征方面优于SPA与VIP方法, 能有效降低模型复杂度并抑制过拟合。模型对比表明, CNN与RF的预测精度与泛化能力相对有限, LSTM与PSO-SVR表现最优, 外部验证进一步证实二者对未知样本均具有较高预测精度。SHAP可解释性分析揭示, 模型的核心贡献特征高度集中于 $2\theta=24^{\circ}\sim 35^{\circ}$ (HgS主衍射峰区)与 $50^{\circ}\sim 60^{\circ}$ (次级峰区)。其中, LSTM更善于挖掘高角度特征的全局上下文关联, 而PSO-SVR对中低角度局部非线性特征更为敏感。

本研究所建立的SNV+CARS+LSTM与SNV+CARS+PSO-SVR两条路径融合了光谱预处理、特征筛选、序列建模、智能寻优与可解释性, 为朱砂主成分的无损、快速定量分析提供了可靠的方法学参考。构建的标准化、可解释的定量分析流程, 有效解决了矿物药光谱分析中基线漂移、多重共线性及高维冗余等共性难题, 为同类复杂基质样品的成分定量提供了可复制的技术路径与方法学框架。未来, 该建模框架可进一步应用于中药材产地快检、生产线在线质控及市场监管抽检等实际场景, 并有望拓展至其他矿物药及含矿物复方制剂的成分分析, 对推动中药质量评价向标准化、快速化与智能化发展具有积极的推广价值。

参考文献:

- [1] Chai S W, Fan B S, Song X J, Zhuang P W. *Tianjin J. Tradit. Chin. Med.* (柴士伟, 范栢爽, 宋学皎, 庄朋伟. 天津中医药), **2024**, 41(3): 403-408.
- [2] Ren Y H, Song F Y, Li Y W, Li H, Liu Y L, Du M B, Peng L H. *Mater. Today Bio.*, **2025**, 34: 102277.
- [3] Liu J, Wei L X, Wang Q, Lu Y F, Zhang F, Shi J Z, Cherian M G. *J. Ethnopharmacol.*, **2018**, 210: 340-350.
- [4] Wang Y, Zhou S Y, Ma H H, Shi J S, Lu Y F. *Environ. Toxicol. Pharmacol.*, **2019**, 66: 83-90.
- [5] Chinese Pharmacopoeia Commission. *Pharmacopoeia of the People's Republic of China*. 2025 ed. Beijing: China Medical Science Press(国家药典委员会. 中华人民共和国药典. 2025年版. 北京: 中国医药科技出版社), **2025**.
- [6] Qian X L, Zhang Q Q, Zhu X Y, Liu Z, Huang Z X, Zhao J, Liu S J. *Mod. Chin. Med.* (钱喜龙, 章芊芊, 朱星宇, 刘政, 黄梓轩, 赵劲, 刘圣金. 中国现代中药), **2025**, 27(4): 779-784.
- [7] Shen X Y, Huang Q Y, Wang C J, Wang M T, Shan Q. *Mine. Eng.*, **2025**, 230: 109397.
- [8] Yang H, Liu S J, Wu D K, Xu C X, Lin R C, Tian J G, Wu L T. *Chin. J. Pharm. Anal.* (杨欢, 刘圣金, 吴德康, 徐春祥, 林瑞超, 田金改, 吴露婷. 药物分析杂志), **2014**, 34(12): 2171-2180.
- [9] Wu S C, Ma Y L, Yang W G, Fang F, Wang Y, Yang W, Liu S J. *Chin. J. Exp. Tradit. Med. Formulae* (吴思澄, 马瑜璐, 杨文国, 房方, 王迎, 杨薇, 刘圣金. 中国实验方剂学杂志), **2023**, 29(13): 166-172.
- [10] Wang Y F, Li Q, Liu Q, Chi G C, Wu Y, Hu J F. *Rock Miner. Anal.* (汪寅夫, 李清, 刘琦, 迟广成, 伍月, 胡建飞. 岩矿测试), **2014**, 33(5): 706-713.
- [11] Yang D, Liu S J, Yan K, Ma H Y, Li Q, Wang H B, Di L Q. *Chin. Med. Mater.* (杨丹, 刘圣金, 燕珂, 马宏跃, 李全, 王恒斌, 狄留庆. 中药材), **2018**, 41(12): 2767-2773.
- [12] Liu Y, Zhang L W, Zhang X Z, Bian X H, Tian W L. *Microchem. J.*, **2025**, 213: 113605.
- [13] Xie W X, Zhang Z X, Liu Y, Qiu F, Ding L Q. *Chin. Tradit. Herb. Drugs* (解文欣, 张紫萱, 刘越, 邱峰, 丁丽琴. 中草药), **2025**, 56(15): 5616-5631.
- [14] Luo Y Z, Jin Q T, Lu H Z, Li P, Qiu G J, Qi H J, Li B, Zhou X X. *Agriculture*, **2025**, 15(3): 281.
- [15] Ponce-Bobadilla A V, Schmitt V, Maier C S, Mensing S, Stodtman S. *Clin. Trans. Sci.*, **2024**, 17(11): e70056.
- [16] Hashem M A, Ahmed A, Hasan M M. *LWT*, **2025**, 236: 118659.

- [17] Záhonyi P, Fekete D, Szabó E, Nagy Z K, Nagy B. *Eur. J. Pharm. Sci.*, **2025**, 212: 107173.
- [18] Xiao X, Yuan J, Li J, Wang W N, Nie Y, Jiao F, Yang J. *Food Ferment. Ind.* (肖徐, 袁进, 李静, 王薇娜, 聂叶, 焦富, 杨洁. 食品与发酵工业), **2025**, 51(24): 368–374.
- [19] Cai X, Li L, Zong X Y, Peng H B, Cao X Y. *Chin. J. Inorg. Anal. Chem.* (蔡旭, 李丽, 宗绪岩, 彭厚博, 曹欣悦. 中国无机分析化学), **2026**, 16(2): 259–269.
- [20] Ezenarro J. *Anal. Chim. Acta*, **2025**, 1364: 344229.
- [21] Zhang X Z, Saimaiti P, Zhang L W, Yang Y F, Bian X H. *Spectrosc. Spectral Anal.* (张馨之, 帕提姑丽·赛买提, 张露文, 杨亚非, 卞希慧. 光谱学与光谱分析), **2025**, 45(9): 2410–2416.
- [22] Chen W, Chen Z G, Liu S, Liu J M, Wang H. *J. Instrum. Anal.* (陈雯, 陈争光, 刘烁, 刘金明, 王河. 分析测试学报), **2025**, 44(10): 2079–2086.
- [23] Shi Y B, He T Y, Zhong J J, Mei X, Li Y, Li M X, Zhang W, Ji D, Su L L, Lu T L, Zhao X L. *Talanta*, **2024**, 268(1): 125266.
- [24] Zhang M T, Zhang J, He C F. *Comput. Eng. Appl.* (张梦婷, 张军, 何承烽. 计算机工程与应用), **2026**, 62(2): 40–53.
- [25] Li Z H, Liu Y T, Wang S T, Li B. *Food Sci.* (李子豪, 刘羽炯, 王树桐, 李波. 食品科学), **2026**, 47(5): 324–334.
- [26] Ren X Y, Li J W, Dong Y N, Sang M J, Qin M T, Lin L, Lin Y Q. *Chin. Pharm. Stand.* (任小英, 李佳薇, 董钰宁, 桑梦娇, 秦梦廷, 林林, 林永强. 中国药品标准), **2025**, 26(3): 304–311.
- [27] Lu Z P, Dong Y Y, Du K, Shan J J, Xie T. *J. Instrum. Anal.* (卢志鹏, 董莹莹, 杜柯, 单进军, 谢彤. 分析测试学报), **2026**, 45(6): 1161–1173.

(责任编辑: 盛文彦)