• 数据挖掘与循证医学 •

基于 Voting 集成算法的中药抗炎预测模型的构建

乔塬淏[1]，谢虹亭[2]，胡馨雨[1]，安　宸[2]，刘泽豪[2]，陈美池[3]，薛　鹏[2]，朱世杰[2*]

1. 北京中医药大学研究生院，北京　100029
2. 中国中医科学院望京医院 肿瘤科，北京　100102
3. 中山大学附属第一医院广西医院，广西 南宁　530028

摘　要：目的　以中药药性作为特征变量，构建基于 Voting 集成算法的中药抗炎作用预测模型，并通过可视化技术分析不同药性特征对于中药抗炎作用的影响。**方法**　以《中药学》与 SymMap 数据库中 1 247 味中药为研究对象，经过初筛和复筛后建立包含性味归经等特征的规范化数据库。基于决策树、支持向量机、轻量级梯度提升机等 6 种基础模型构建 Voting 集成模型，并以七折交叉验证和基于树结构的贝叶斯优化算法超参数优化提升模型性能。利用 SHAP（SHapley Additive exPlanations）解释器可视化关键药性特征。**结果**　经筛选后，共纳入 522 味抗炎中药构建数据库。Voting 集成模型综合性能最优，F1 分数为 0.797，AUC 值为 0.77，较单一模型平均提升 7.4%。SHAP 分析表明使中药发挥抗炎作用的重要特征分别是“脾经”“甘味”“补益”等，使中药不具有抗炎作用的重要特征为“性温或平”和“毒性”。**结论**　首次通过集成算法构建具有良好性能的中药抗炎作用预测模型，为中医药与机器学习结合的研究模式提供了新思路。

关键词：Voting 集成算法；中药；抗炎；机器学习；药性；四气五味

中图分类号：TP18；R285.1　　**文献标志码：**A　　**文章编号：**0253 - 2670(2025)15 - 5529 - 09

DOI: 10.7501/j.issn.0253-2670.2025.15.018

Construction of predictive model for anti-inflammatory effects of traditional Chinese medicine based on Voting ensemble algorithm

QIAO Yuanhao[1], XIE Hongting[2], HU Xinyu[1], AN Chen[2], LIU Zehao[2], CHEN Meichi[3], XUE Peng[2], ZHU Shijie[2]

1. Graduate School, Beijing University of Chinese Medicine, Beijing 100029, China
2. Oncology Department, Wangjing Hospital of China Academy of Chinese Medicine Sciences, Beijing 100102, China
3. Guangxi Hospital Division, First Affiliated Hospital of Sun Yat-sen University, Nanning 530028, China

Abstract: Objective　To develop a prediction model for the anti-inflammatory effects of traditional Chinese medicine (TCM) using medicinal properties as feature variables through a Voting ensemble algorithm, and analyzing the impact of different TCM property characteristics on anti-inflammatory activity through visualization techniques. **Methods**　We systematically analyzed 1 247 herbal medicines from the *Chinese Materia Medica* and the SymMap database. Following initial and secondary screening, we established a standardized database containing characteristic parameters including nature, flavor, and channel tropism. A Voting ensemble model was constructed by integrating six base classifiers (decision tree, support vector machine, light gradient boosting machine, etc.), with model performance enhanced through 7-fold cross-validation and Tree-structured Parzen Estimator hyperparameter optimization. The SHapley Additive exPlanations (SHAP) interpreter was employed to visualize feature importance. **Results**　The final database comprised 522 anti-inflammatory herbs. The Voting ensemble model demonstrated superior performance (F1-score: 0.797, AUC: 0.77), demonstrating a 7.4% average improvement over individual models. SHAP analysis identified “spleen meridian”, “sweet flavor”, and “tonifying properties” as critical positive predictors, while “warm/neutral nature” and “toxicity” emerged as key negative indicators. **Conclusion**　This study pioneers the application of ensemble learning in predicting TCM anti-inflammatory activity based on medicinal properties, establishing a

收稿日期：2025-04-13

基金项目：中国中医科学院望京医院高水平中医医院建设项目（WJZJ-202305）；中国中医科学院望京医院高水平中医医院建设项目（WJYY-XZKT-2023-37）。

作者简介：乔塬淏（2001—），男，硕士研究生，研究方向为中西医结合防治肿瘤。E-mail: qyh2477816173@163.com

***通信作者：**朱世杰，博士生导师，主任医师，研究方向为中西医结合防治肿瘤。E-mail: zhushij@hotmail.com

novel research paradigm that integrates traditional Chinese medicine theory with machine learning technology.

Key words: Voting ensemble algorithm; traditional Chinese medicine; anti-inflammatory; machine learning; medicinal properties; four natures and five flavors

恶性肿瘤和心血管疾病仍是威胁人类生命健康的首要难题[1-2]，而炎症反应是其共同的病理基础，相关机制研究始终是医学领域的前沿课题。中药凭借“多成分-多靶点-多通路”的协同调控特性，在干预急、慢性炎症，阻断“炎癌转化”方面展现出独特优势[3]。然而，中药抗炎作用的研究仍面临两大瓶颈：其一，传统药理学方法以实验为核心，主要使用动物模型、细胞筛选、分子对接等方法进行分析[4]，但存在靶点更新滞后、覆盖有限等问题，难以系统揭示中药“成分-药性-效应”多维关联规律；其二，多数研究聚焦于部分常见中药，仍有较多抗炎中药尚未探明。

近年来，人工智能算法的迅速发展为解决以上问题提供了新思路，目前，已有多种机器学习算法如随机森林、支持向量机等在中药数据挖掘方面得到了广泛的应用[5]。但由于中药数据具有高维度、非线性及类别不平衡等特性，导致单一机器学习模型易出现预测偏差和泛化能力不足的问题[6]。针对上述挑战，Voting 集成策略可通过概率加权融合显著提升中医复杂分类任务的准确率，为突破中药抗炎效应预测瓶颈提供了新范式[7]。本研究通过使用传统中药药性为特征变量，结合 Voting 集成算法构建中药抗炎作用预测模型，并通过可视化分析不同药性特征对中药抗炎作用的影响，为同类研究提供方法学参考。

1 抗炎药物数据库的构建

1.1 数据来源

数据来源于《中药学》[8]中记载的全部 544 味中药及 SymMap 数据库（http://www.symmap.org）中的 703 味中药。

1.2 筛选方法

①初筛：纳入中药经去重后通过查阅相关书籍确定其是否具有抗炎作用。②复筛：对初筛中未记录具有抗炎作用的中药进行复筛，复筛文献来源为中国知网（CNKI）、万方（Wangfang Data）、维普中文期刊服务平台（VIP）、中国生物医学文献数据库（CBM）、Web of Science、PubMed 等数据库，检索日期截至 2024 年 8 月 5 日，中文检索词为中药名加炎症，英文检索词为中药英文名加 inflammation。

1.3 纳入标准

初筛纳入标准：以《中药学》《中药大辞典（第 2 版）》[9]中记载的药物药理作用作为判断标准。复筛纳入标准：中药需有至少 1 篇描述其抗炎作用的文献，①文献发表期刊类型为中文核心期刊或者 SCI 源期刊；②研究类型为临床随机对照研究或队列研究；③试验组干预措施为单味中药或者中药有效成分；④结局指标有明确的抗炎指标。

1.4 排除标准

对于初筛参考书籍中未记录有明确抗炎功效且在复筛中无文献记载其有具体抗炎作用的中药予以排除。

1.5 数据库规范化构建和录入

在筛选后将中药信息录入 Excel 建立数据库，包括中药的名称、性味归经、功效毒性等，为确保数据的准确性，由双人分别录入数据后进行审核。参考《中药大辞典》[9]对中药药名信息进行规范化处理。对于中药性味中的微苦、微温等描述，将其“微”均删除，统一为苦、温等。

2 机器学习模型的构建与评估

2.1 模型数据库的构建

对建立的中药信息数据库进行处理，提取中药的“药性”“药味”“归经”“毒性”“功效”5 个类别，以收集到的“具有抗炎作用”的中药为正标签，“不具有抗炎作用”的中药为负标签。负标签中药的收集范围为《中药学》及 SymMap 数据库的所有中药，经过在上述文献数据库检索，依据以上纳排标准，若未明确其有抗炎作用，则将其作为负标签纳入。

2.2 数据信息格式的转化

由于纳入的中药性味归经等数据为分类特征，需要对其进行特征数字化以适应机器学习的格式。本研究采用独热编码的方法对数据进行处理。其方法是将分类变量转换为多个二进制特征，每个特征对应 1 个类别。举例来说，对于“药性”这一数据类别，独热编码会为其包含所有可能的类别，“寒”“热”“温”“凉”等各分别创造 1 个新的变量，每个新变量都赋予二进制 0 或 1 的值。如麻黄性温，则在独热编码后，其在“温”这一变量的值为 1，而在其他“寒”

“热”“凉”等变量的值为 0，其他特征同理。

2.3 模型训练

2.3.1 开发环境与框架 基于 anaconda3 软件，采用 Python 3.12.3 算法框架对人工智能模型进行训练。

2.3.2 数据预处理 为了确保算法能够更有效地识别出数据中的异常情况，首先对模型数据库进行孤立森林模型异常检测处理。孤立森林模型可通过计算样本异常分数以高效识别数据中的离群点[10]。通过清洗数据可减少模型过拟合的风险。对经过模型异常检测处理后的数据集使用原型选择技术进行处理，以减少可能出现的数据不平衡问题。

2.3.3 机器学习算法 使用 6 种机器学习单一模型对数据集进行训练，包括决策树（decision tree，DT）、支持向量机（support vector machine，SVM）、轻量级梯度提升机（light gradient boosting machine，LGBM）、梯度提升决策树（gradient boosting decision tree，GBDT）、梯度下降法（stochastic gradient descent，SGD）、极度随机树（extratrees，ET）。DT 模型是一种基于树结构的监督学习算法，其通过递归地选择最优特征并对数据进行分割，从而实现对未知数据的分类或连续值预测[11]。SVM 模型通过对数据进行二分类或回归分析，找到能够最大化分类间隔的超平面，同时最小化分类错误，以实现数据的精确分类或回归预测[12]。LGBM 通过采用基于直方图的算法、leaf-wise 的树生长策略和单边梯度采样等技术创新，显著提高模型的训练效率[13]；GBDT 通常采用 level-wise 的树生长策略，通过迭代地训练决策树来最小化损失函数，从而实现对复杂数据的高精度预测和强大的泛化能力[14]。SGD 模型通过在每次迭代中随机选择一个样本来近似计算损失函数的梯度，有助于跳出局部最小值，提高模型的泛化能力[15]。ExtraTrees 模型基于决策树模型，在特征选择和分割点选择上引入随机性，通过极端随机化提高计算效率，同时减少过拟合且易于实现并行化处理[16]。

采用 Voting 集成算法进行模型融合。这种方法通过集成多个算法模型的预测结果，通过投票的方式输出最终结果[17]。Voting 分硬投票（hard voting）与软投票（soft voting）2 类，本研究使用软投票作为集成分类器，考虑相较于硬投票直接选择得票数最多的类别作为最终预测结果，软投票考虑到了通过集成各基模型对类别的概率预测结果，能够综合反映模型的置信度差异，避免硬投票仅依赖多数表决导致的信息损失。本研究为多特征、非线性数据集，软投票更适用于权衡各模型的局部优势，提升分类鲁棒性。

2.3.4 模型训练 采用七折交叉验证（K＝7）的方法对 DT、SVM、LGBM、GBDT、SGD、ExtraTrees 这 6 种单一模型进行训练，七折交叉验证是将数据集分成 7 个部分，轮流使用其中一部分作为测试集，其余作为训练集，进行 7 次训练和测试，最后取 7 次测试结果的平均值为最终模型性能的评估。使用基于树结构的贝叶斯优化算法（tree-structured parzen estimator，TPE）探寻模型的最优超参数，TPE 分布通过分别建模性能优良和欠佳的超参数概率分布，指导算法高效选择更优配置[18]。基于最优超参数进行训练得到 6 种算法的最优模型，遍历所有可能的组合方式，根据 F1 分数选出最佳模型组合，对最佳组合采用软投票法进行模型融合，得到 Voting 集成模型（本研究以相同权重加权平均，基模型的最优参数经独立优化后固定，集成过程不引入额外可调超参数）。

2.4 模型性能的评估

使用机器学习研究中常见的评价指标来评价模型，包括准确率、平衡 F 分数（F1 分数）、微平均受试者工作特征（receiver operating characteristic curve，ROC）曲线、ROC 曲线下面积（Area Under Curve，AUC）。在计算模型准确率前采用七折交叉法取平均值来评价模型性能并可视化为箱线图。准确率是模型性能的基本指标，用于评估模型在整体数据集上的预测效能。ROC 曲线是以假正例率为横坐标，真正例率为纵坐标绘制的曲线[19]。

准确率＝正确分类的样本数/所有样本数

F1 分数＝2×(精确率×召回率)/(精确率＋召回率)

精确率＝真正例/(真正例＋假正例)

召回率＝真正例/(真正例＋假负例)

AUC 测量了模型在不同阈值下的分类能力，可对模型在各类样本中的预测能力进行更均衡的量化评估[20]。混淆矩阵是一种分类模型性能评估工具，其以矩阵形式直观呈现预测结果与真实标签的对应关系，使各个类别预测结果和真实结果的对比得到更直观的反映[21]。

2.5 药性特征重要性评价

使用 SHAP（SHapley Additive exPlanations，SHAP）解释器对模型的训练结果进行特征重要性可视化。SHAP 是一种基于博弈论中 Shapley 值的解释性技术，其核心思想是计算某一特征加入模型时的边际贡献，然后考虑该特征在所有特征序列条

件下的不同边际贡献，取其均值进行比选判断大小，公平量化每个特征对模型预测的影响，从而使得模型的决策透明化，增加模型可解释性[22]。

3　结果

3.1　抗炎中药数据库数据特征

本研究共筛选出抗炎中药 522 味，其中通过查阅相关书籍初筛得到 295 味中药，数据库检索复筛得到 227 味。涉及 21 种功效，其中纳入数量最多的 5 种功效依次为清热（108，20.68%）、补益（64，12.26%）、化痰止咳平喘（49，9.39%）、祛风湿（43，8.23%）、活血（38，7.28%），见图 1，中药药性特征示例见表 1。

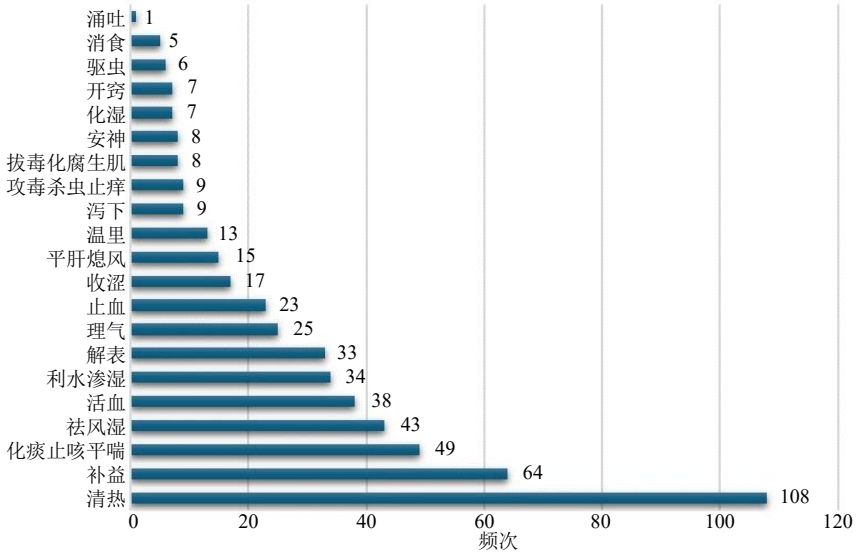


图 1　纳入中药功效频次分布

Fig. 1　Frequency distribution of efficacy of included traditional Chinese medicines

表 1　中药药性统计示例

Table 1　Statistical examples of medicinal properties of traditional Chinese medicines

序号	中药名	抗炎	药类	五味	四气	毒性	归经
1	阿胶	有	补益	甘	平	无	肺、肝、肾
2	阿魏	无	消食	苦、辛	温	无	脾、胃
3	矮地茶	有	化痰止咳平喘	苦、辛	平	无	肺、肝
4	艾片	有	开窍	辛、苦	寒	无	心、脾、肺
5	艾叶	有	止血	苦、辛	温	无	肝、脾、肾
6	安息香	有	开窍	辛、苦	平	无	心、脾
7	八角茴香	无	温里	辛	温	无	肝、肾、脾、胃
8	八角莲	无	清热	苦、辛	温	有	肺、肝
9	八月札	无	理气	苦	平	无	肝、胃
10	巴豆	有	泻下	辛	热	有	胃、大肠

3.2　数据特征信息的转化

通过独热编码的方法，对数据库中的数据进行向量化处理，使每一个中药性味都转化为 1 个单独特征变量，转化后的数据集示例如表 2 所示。数据集共纳入 47 个特征，特征包括四气、五味、归经、药类及其他功能属性。以是否具有抗炎作用为目标变量，1 表示具有该特征，0 表示无该特征，共纳入正标签 522 个，负标签 183 个，正负标签比例约为 2.85∶1。通过孤立森林模型异常检测处理剔除 10% 离群点，处理后数据集分布与原数据集相同（Kolmogorov-Smirnov 检验 $P>0.05$），对处理后的数据通过平衡采样从每类随机抽取 160 个样本，形

表 2　用于机器学习的中药药性特征数据集示例

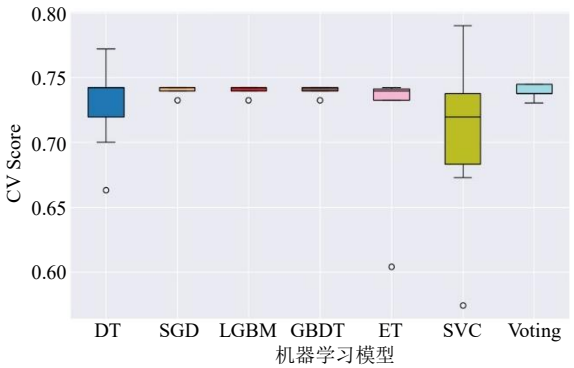
Table 2　An example of a dataset of medicinal properties of traditional Chinese medicine for machine learning

序号	中药名	抗炎	毒性	酸	热	胃	肝	药类_清热
1	阿胶	1	0	0	0	0	1	0
2	阿魏	0	0	0	0	1	0	0
3	矮地茶	1	0	0	0	0	1	0
4	艾片	1	0	0	0	0	0	0
5	艾叶	1	0	0	0	0	1	0
6	安息香	1	0	0	0	0	0	0
7	八角茴香	0	0	0	0	1	1	0
8	八角莲	0	1	0	0	0	1	1
9	八月札	0	0	0	0	1	1	0
10	巴豆	1	1	0	1	1	0	0

成平衡数据集，从平衡数据中按 8∶2 划分训练集和验证集，保证类别比例一致。

3.3 模型参数及性能评估结果

7 个机器学习模型经七折交叉验证后的性能分布情况如图 2 所示，可以看到 6 种基础模型和 Voting 集成模型性能都处于较好水平，SGD、LGBM、GBDT、ET 和 Voting 模型的表现稳定性较好，其中 Voting 模型中位数最高，且数据无异常值。表 3 为 7 种模型各自的 F1 分数、准确率及通过 TPE 算法独立优化得到的最佳超参数，图 3 为 7 种模型 ROC 曲线的对比，其中 Voting 模型的综合性能仍优于其余 6 种机器学习模型。F1 分数为 0.797，AUC 值为 0.77，说明 Voting 模型可以较好预测中药的抗炎功能。图 4 为 7 种模型的混淆矩阵图，其纵轴为标签真实值，横轴为模型预测值，当预测值与真实值匹配的数量越多，即矩阵图左上和右下的值越高时，就证明机器学习模型的预测效果越好，可信度越高。由图可见 Voting 模型对于正负标签的预测准确率是最高的。



CV Score-交叉验证得分，DT-决策树，ET-极度随机树，GBDT-梯度提升决策树，LGBM-轻量级梯度提升机，SGD-梯度下降法，SVM-支持向量机，下图同。

CV Score-cross-validation score, DT-decision tree, ET-extratrees, GBDT-gradient boosting decision tree, LGBM-lightgradient boostingmachine, SGD-stochastic gradient descent method, SVM-support vector machine, same as below figures.

图 2 7 种机器学习模型准确率交叉验证箱线图

Fig. 2 Accuracy cross-validation boxplots for seven machine learning models

表 3 7 种机器学习模型性能评估

Table 3 Performance evaluation of seven types of machine learning models

机器学习模型	准确率	F1 分数	关键超参数
DT	0.734	0.733	最大深度＝19，叶节点数＝7，分裂最小样本＝4
SGD	0.750	0.750	正则化强度 α＝0.06，L1/L2 混合比＝0.42
SVC	0.734	0.734	使用 sigmoid 核函数，惩罚系数 C＝0.13
LGBM	0.766	0.766	叶节点数＝81，最小叶样本＝51，学习率＝0.001
ET	0.719	0.718	最小分裂样本＝20，弱学习器数＝124
GBDT	0.750	0.750	学习率＝0.008，弱学习器数＝192
Voting	0.797	0.797	/

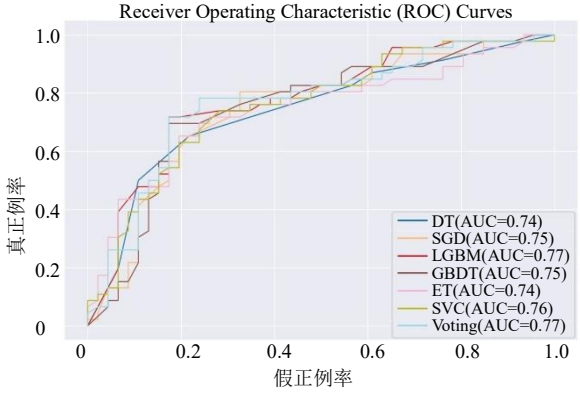


图 3 7 种机器学习模型 ROC 曲线

Fig. 3 ROC curves for seven mac hine learning models

3.4 基于 Voting 模型的抗炎药性特征重要性评估结果

对结果可信度相对高的 Voting 模型进行 SHAP 可视化分析，如图 5 所示，图中纵轴代表该特征的 SHAP 值，横轴代表特征值的大小，中线左侧代表无抗炎作用的特征值大小，右侧代表有抗炎作用的特征值的大小，其上的每一个点都代表 1 个样本，其位置根据特征值进行排列，散点排列越分散，证明这个特征对模型的影响程度越大，散点的颜色代

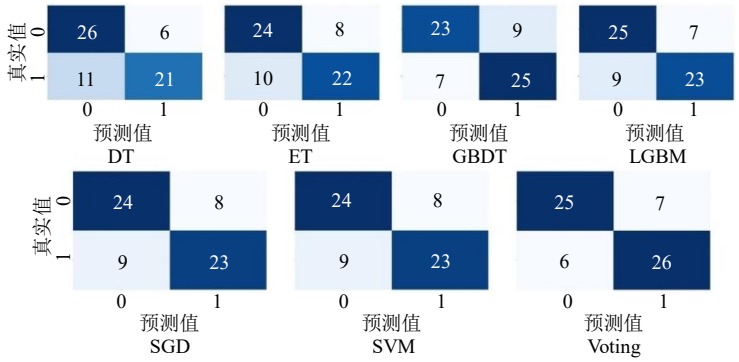


图 4　7 种机器学习模型混淆矩阵图

Fig. 4　Confusion matrix diagram of seven machine learning models

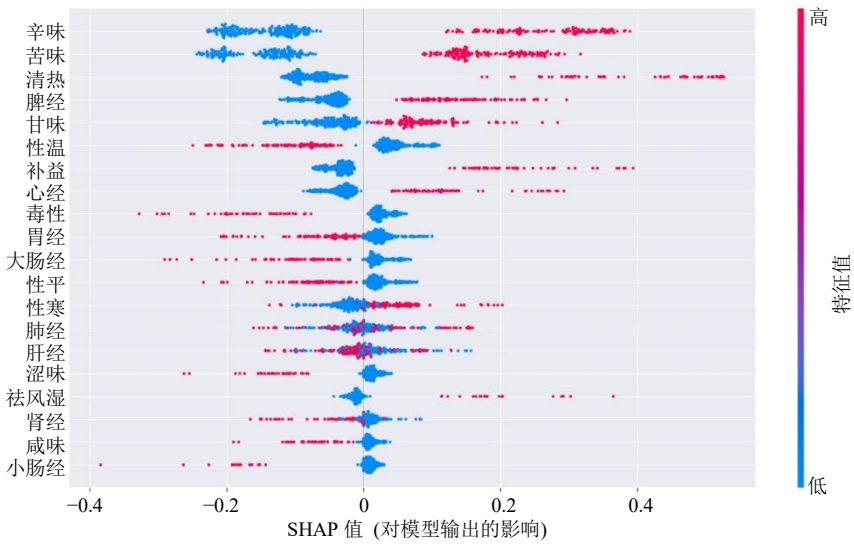


图 5　基于 7 种机器学习模型的抗炎作用特征重要特征评估

Fig. 5　Evaluation of important features of anti-inflammatory effect based on seven machine learning models

表特征值与模型预测值的正负关系，红色表示特征值较高时增加了模型的预测值，即正相关关系，蓝色表示特征值较高时减少了模型的预测值，即负相关关系。从图中可以看出，使中药具有抗炎作用的最重要的特征分别是“辛味”“苦味”“清热”“脾经”“甘味”“补益”，使中药不具有抗炎作用的最重要的特征为“性温”“毒性”“胃经”“大肠经”“性平”。

4　讨论

4.1　机器学习对比传统数据挖掘的优势

相比传统的数据挖掘方法，机器学习在自动化方面表现突出，能够自动完成特征提取和模型优化，减少了对人工干预的依赖。在数据处理方面，传统数据挖掘则需要人工设计特征，其过程易受主观经验限制，且难以捕捉高维数据中的非线性关系；而机器学习算法可以通过表征学习自动从数据中提取复杂的特征，这让它拥有更强的处理非线性和高维数据的能力，可有效解决传统方法中特征稀疏性和信息丢失问题。在模型优化层面，传统数据挖掘方法需要手动调整模型的超参数，调参结果可能受到人为因素的影响，缺乏一致性和可重复性，同时很难找到最优参数组合；而机器学习可以通过多种方式如超参数调优工具、自适应学习算法、集成学习等方法进行自动化模型优化，本研究中采用 TPE 算法自动寻找最优超参数组合，最终以集成学习的方式结合多个模型预测结果。通过对这些方法的使用，减少了人工干预的影响，使得模型性能显著提高，本研究表明集成模型的预测准确率较单一模型平均提升 7.4%。

4.2　抗炎类中药的药性、药味和功效分析

现代医学中，炎症是指机体受到致炎因子刺激

后具有血管系统的活体组织产生的防御反应，其临床表现为红肿热痛和机能障碍[23]。在中医古籍中并无对“炎症”的明确记载，“炎”字多表示火势旺盛的意思，《说文解字》曰：“炎，火光上也。从重火。”在中医八纲辨证体系下，炎症反应不能简单等同于实热证。慢性炎症过程中常伴随“虚实夹杂”的病理状态，此时单纯清热解毒可能损伤正气，需配伍补益药物形成“清补兼施”的治疗策略。本研究在对纳入中药进行特征重要性评估时发现，苦、辛、甘味对中药具有抗炎作用有重要作用，而在本研究纳入的 522 味抗炎的中药中，苦、甘、辛也是占比最多的 3 个药味，分别有 286、206、204 味，与先前文献中的研究结果相吻合[3]。同时，归脾经和补益作用也是使中药具有抗炎作用的重要特征。研究证实，卫气与机体免疫功能密切相关，可通过维护物理屏障、调节免疫细胞活性而发挥抗炎作用。《素问·痹论》曰：“卫者，水谷之悍气也。”证明了卫气的生成与脾胃运化水谷精微的功能密切相关。甘为脾之主味，药性缓柔，能顺应脾“喜缓”“恶急”的特性，正如《素问·脏气法时论》所言：“脾欲缓，急食甘以缓之”，故甘味药可通过补益脾胃、调和营卫、辅助气血运行等途径，直接或间接促进卫气的化生与功能发挥。现代药理学也证实了甘味、入脾经的中药通过调节机体免疫反应，改变抗氧化基因表达等途径发挥抗炎作用，如黄芪多糖可促进巨噬细胞向 M2 型极化，抑制促炎因子的分泌，同时促进抗炎因子的释放，具有双向免疫调节作用[24]；抗氧化层面，人参皂苷可显著减少活性氧（reactive oxygen species，ROS）的过度积累，间接抑制 ROS 依赖的核因子-κB（nuclear factor-κB，NF-κB）活化，从而减轻炎症级联反应[25]。此外，甘味中药凭借其“甘能缓、能补”的特性，通过抗炎、抗氧化、免疫调节等多个机制的协同作用，促进了炎症的恢复，减轻慢性炎症对组织的损伤，从而对炎癌转化起到了较好的防治作用，这种“扶正祛邪”的动态调节模式，既体现了中医“未病先防”的治未病理念，也与现代肿瘤预防医学强调的微环境调控策略高度契合。

目前临床上仍有医家认为“毒势愈猛则祛邪愈速”，偏好以“以毒攻毒”之法为攻邪之常法。然而，国际研究数据显示，中国 26.81%的药物性肝损伤病例与有毒中药（传统中药或草药及膳食补充剂）的使用相关[26]。本研究发现，毒性并非是中药具有抗炎作用的重要特征。现代研究也表明，中药抗炎作用源于其有效成分对炎症通路的调控，与毒性无必然关联，如附子中的有效成分去甲乌药碱通过抑制 NF-κB 和激活核因子 E2 相关因子 2（nuclear factor erythroid-2-related factor 2，Nrf2）/血红素氧合酶-1（heme oxygenase-1，HO-1）信号通路发挥抗炎和抗氧化作用[27]，但并非其主要毒性成分[28-29]。这一误区可能源于传统经验中对短期疗效的片面认知，却忽视了毒性与药效成分之间的相互独立性。临床应摒弃“以毒代效”思维，强调辨证选用低毒替代药，严格控制有毒中药的剂量、疗程及炮制配伍，并针对个体差异监测肝肾功能，通过分离药效与毒性成分，实现抗炎疗效与用药安全的平衡。

4.3 关于机器学习过程的讨论

本研究基于机器学习的方法以中医药性理论作为算法模型的特征变量成功构建了中药抗炎作用的预测模型，并通过特征重要性评估揭示关键药性参数。由于中药抗炎机制具有多成分、多靶点协同调控的特点，其药性特征空间呈现高维性与非线性关联，考虑使用传统单一模型对于多维度数据处理可能存在“维度灾难”的问题[30]，出现包括数据稀疏、过拟合风险提高、可解释性下降和特征相关性增强等诸多问题，所以本研究进行机器学习过程中使用 Voting 算法整合 6 种模型进行集成学习，利用模型多样性提升对特征空间的覆盖能力。为机器学习应用于对中医药理论特征信息可能存在的泛化能力不足、多重共线性等问题提供了可能的解决思路。

需要指出的是，本研究数据集中正负标签比例（2.85∶1）的失衡性，反映出现有文献已验证的抗炎中药数量显著多于非抗炎中药，尽管传统数据平衡技术（如欠采样）可一定程度上缓解类别不平衡问题，但其可能对药性特征的真实分布产生一定影响，导致模型对中医理论核心特征（如归经、性味）的解析出现偏差。因此，如何在保持药性数据完整性的前提下提升模型对少数类的识别能力，成为中医药智能化研究的共性挑战，本研究提出 3 种未来可尝试的研究方向，通过模型架构优化缓解类别失衡的影响：①构建中医专家知识引导的特征权重分配框架，通过邀请中医专家团队对药性特征（如四气、五味、归经等）进行系统性评分，量化其与抗炎作用的理论关联强度，将专家评分转化为各个特征的权重系数，并通过模型参数动态调节训练过程

中的特征贡献度，从而达到通过中医理论指导优化模型性能的目的。②借鉴中药复方配伍中“君、臣、佐、使”的协同理念，将不同模型按功能角色分层（如主预测、辅助校正等），通过权重分配模拟复方配伍中“主次有序”的增效机制，可能改善模型对少数类的泛化性能。③将抗炎作用预测与中医证型分类、成分靶点分析等任务结合，构建多任务学习框架，利用共享表征学习挖掘药性特征的多重关联，或可通过丰富模型的学习目标缓解单一任务中类别失衡的局限性。

5 结论

本研究通过整合 1 247 味中药数据，构建了基于 Voting 集成算法的中药抗炎预测模型，揭示了“辛味”“苦味”“清热”及“补益”等药性特征与抗炎作用的关键关联，并通过 SHAP 解释器量化其贡献，为中药抗炎作用的系统性研究提供了高效、可解释的智能化工具。未来可进一步扩展数据来源，结合深度学习方法挖掘药性-成分-靶点的多层次关联；同时探索模型在临床实践中的应用。

利益冲突 所有作者均声明不存在利益冲突

参考文献

[1] Bray F, Laversanne M, Sung H, *et al*. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries [J]. *CA Cancer J Clin*, 2024, 74(3): 229-263.

[2] 刘明波, 何新叶, 杨晓红, 等. 《中国心血管健康与疾病报告 2023》要点解读 [J]. 中国心血管杂志, 2024, 29(4): 305-24.

[3] 陈美池, 谢虹亭, 龙思丹, 等. 基于数据挖掘的抗炎抗肿瘤中药药性特征分析 [J]. 上海中医药杂志, 2023, 57(4): 44-50.

[4] 李建锋, 荀丽英, 李航, 等. 中药成分的生物学活性评价及筛选 [J]. 中草药, 2015, 46(4): 588-594.

[5] 杨淇, 郝二伟, 侯小涛, 等. 基于药性理论的中药抗辐射预测模型的构建 [J]. 中草药, 2024, 55(8): 2684-2693.

[6] 佘楷杰, 袁芳君, 马庆宇, 等. 机器学习驱动中医诊断智能化的发展现状、问题及解决路径 [J]. 中国中医基础医学杂志, 2024, 30(3): 398-406.

[7] 郭小川, 冯贞贞, 刘文瑞, 等. 基于 Stacking 集成算法的中医证候诊断模型建立: 以肺癌为例 [J]. 中医杂志, 2024, 65(17): 1775-1783.

[8] 高学敏, 钟赣生. 中药学 [M]. 第 2 版. 北京: 人民卫生出版社, 2012: 174-1988.

[9] 南京中医药大学. 中药大辞典 [M]. 第 2 版. 上海: 上海科学技术出版社, 2014: 1-3874.

[10] Liu F T, Ting K M, Zhou Z H. Isolation forest [A] // 2008 Eighth IEEE International Conference on Data Mining [C]. Pisa: IEEE, 2008: 413-422.

[11] Salzberg S L. C4.5: Programs for machine learning by J. ross quinlan. morgan kaufmann publishers, inc., 1993 [J]. *Mach Learn*, 1994, 16(3): 235-240.

[12] Cervantes J, Garcia-Lamont F, Rodríguez-Mazahua L, *et al*. A comprehensive survey on support vector machine classification: Applications, challenges and trends [J]. *Neurocomputing*, 2020, 408: 189-215.

[13] Ke G, Meng Q, Finley T, *et al*. LightGBM: A highly efficient gradient boosting decision tree [A] // Advances in Neural Information Processing Systems 30 (NIPS 2017). Long Beach: Neural Information Processing Systems Foundation, 2017: 3149-3157.

[14] Friedman J H. Greedy function approximation: A gradient boosting machine [J]. *Ann Statist*, 2001, 29(5): 1189-1232.

[15] Tian Y J, Zhang Y Q, Zhang H B. Recent advances in stochastic gradient descent in deep learning [J]. *Mathematics*, 2023, 11(3): 682.

[16] Wehenkel L, Ernst D, Geurts P. Ensembles of extremely randomized trees and some generic applications [A] // Proceedings of the 40th International Conference on Machine Learning (ICML 2023) [C]. Honolulu: Proceedings of Machine Learning Research (PMLR), 2023: 12345-12356

[17] Burka D, Puppe C, Szepesváry L, *et al*. Voting: A machine learning approach [J]. *Eur J Operational Res*, 2022, 299(3): 1003-17.

[18] Watanabe S. Tree-structured Parzen estimator: Understanding its algorithm components and their roles for better empirical performance [EB/OL]. (2023-05-26) [2025-04-11]. https://arxiv.org/abs/2304.11127.

[19] Mandrekar J N. Receiver operating characteristic curve in diagnostic test assessment [J]. *J Thorac Oncol*, 2010, 5(9): 1315-1316.

[20] Lobo J M, Jiménez-Valverde A, Real R. AUC: A misleading measure of the performance of predictive distribution models [J]. *Glob Ecol Biogeogr*, 2008, 17(2): 145-151.

[21] Heydarian M, Doyle T E, Samavi R. MLCM: Multi-label confusion matrix [J]. *IEEE Access*, 2022, 10: 19083-19095.

[22] Lundberg S M, Lee S I. A unified approach to interpreting model predictions [A] // Proceedings of the 31st International Conference on Neural Information

Processing Systems [C]. Long Beach: Curran Associates Inc., 2017: 4768-4777.

[23] Singh R, Mishra M K, Aggarwal H. Inflammation, immunity, and cancer [J]. *Mediators Inflamm*, 2017, 2017: 6027305.

[24] Zhang Z, Shan W, Wang Y F, *et al*. *Astragalus* polysaccharide improves diabetic ulcers by promoting M2-polarization of macrophages to reduce excessive inflammation via the β-catenin/NF-κB axis at the late phase of wound-healing [J]. *Heliyon*, 2024, 10(4): e24644.

[25] Chu S F, Zhang Z, Zhou X, *et al*. Ginsenoside Rg1 protects against ischemic/reperfusion-induced neuronal injury through miR-144/Nrf2/ARE pathway [J]. *Acta Pharmacol Sin*, 2019, 40(1): 13-25.

[26] Shen T, Liu Y X, Shang J, *et al*. Incidence and etiology of drug-induced liver injury in Mainland China [J]. *Gastroenterology*, 2019, 156(8): 2230-2241.

[27] Yang S W, Chu S F, Ai Q D, *et al*. Anti-inflammatory effects of higenamine (Hig) on LPS-activated mouse microglia (BV2) through NF-κB and Nrf2/HO-1 signaling pathways [J]. *Int Immunopharmacol*, 2020, 85: 106629.

[28] Qin Y, Wang J B, Zhao Y L, *et al*. Establishment of a bioassay for the toxicity evaluation and quality control of *Aconitum* herbs [J]. *J Hazard Mater*, 2012, 199/200: 350-357.

[29] Zhang D K, Li R S, Han X, *et al*. Toxic constituents index: A toxicity-calibrated quantitative evaluation approach for the precise toxicity prediction of the hypertoxic phytomedicine-aconite [J]. *Front Pharmacol*, 2016, 7: 164.

[30] 刘靖, 赵逢禹. 高维数据降维技术及研究进展 [J]. 电子科技, 2018, 31(3): 36-38.

[责任编辑　潘明佳]