

## 基于药性理论的中药抗辐射预测模型的构建

杨 淇<sup>1</sup>, 郝二伟<sup>2,3,4</sup>, 侯小涛<sup>2,3,4</sup>, 邓家刚<sup>2,3,4</sup>, 杜正彩<sup>2,3,4,5</sup>, 范丽丽<sup>1\*</sup>, 夏中尚<sup>2,3,4\*</sup>

1. 广西中医药大学药学院, 广西 南宁 530000

2. 广西中医药大学 广西中药药效研究重点实验室, 广西 南宁 530000

3. 广西中医药大学 广西农作物废弃物功能成分研究协同创新中心, 广西 南宁 530000

4. 广西中医药大学 广西中医湿病方药理论与转化重点实验室, 广西 南宁 530000

5. 广西中医药大学 广西中医药科学实验中心, 广西 南宁 530000

**摘要:** **目的** 以中药药性作为特征描述符构建机器学习抗辐射作用预测模型, 并解释抗辐射中药药性的重要药性特征以指导临床和日常辐射防治。**方法** 通过药智网、中国知网、PubMed 等数据库获取报道具有抗辐射作用的中药研究文献, 通过 SymMap 数据库获得《中国药典》等权威著作记载的中药性味归经等药性作为特征描述符构建数据库, 并将数据处理为适合机器学习的格式。使用随机森林、支持向量机、梯度提升、逻辑回归、全连接神经网络 5 种机器学习模型对数据集进行五折交叉训练并进行性能评估, 再使用 10 个未参与训练的报道具抗辐射作用的中药和 10 个报道无抗辐射作用的中药作为外部验证集测试模型, 最后利用 SHapley 加性解释 (SHapley additive explanations, SHAP) 解释器对决定抗辐射作用有无的重要药性特征进行可视化。**结果** 收集到涉及单味药研究、保健食品注册和复方研究共 136 味报道具抗辐射作用的中药, 总频次为 447 次, 其中中药使用频次及频率排名前 3 的为灵芝、红景天、枸杞子, 频次 $\geq 10$  的中药共 10 味。在 5 个机器学习的性能评估中, 随机森林性能最佳, 其准确率、平衡 F 分数 (F1) 和曲线下面积 (area under curve, AUC) 分别为 0.804 4、0.773 2 和 0.879 8。在外部中药抗辐射的验证中, 随机森林模型能较好地预测已报道具有抗辐射作用的中药。性能最佳的随机森林 SHAP 解释器认为“补虚、清热”功效, “心、肝、脾、肺”归经, “酸、甘、苦”味, “寒”药性特征对抗辐射作用贡献最大。**结论** 首次将中药药性作为特征描述符应用到机器学习当中并取得了较好的预测模型性能, 并指导了放射病的中医药防治, 即当以扶正祛邪, 滋阴降火, 主治心、肝、脾、肺为治疗原则。此外, 机器学习模型较好的预测结果反映了中医药理论对疾病的防治具有良好的可解释性与可重现性。

**关键词:** 机器学习; 中药; 药性; 四气五味; 抗辐射; 灵芝; 红景天; 枸杞子

**中图分类号:** G350; R285.1 **文献标志码:** A **文章编号:** 0253-2670(2024)08-2684-10

**DOI:** 10.7501/j.issn.0253-2670.2024.08.018

## Construction of antiradiation prediction model of traditional Chinese medicine based on medicinal property theory

YANG Qi<sup>1</sup>, HAO Erwei<sup>2,3,4</sup>, HOU Xiaotao<sup>2,3,4</sup>, DENG Jiagang<sup>2,3,4</sup>, DU Zhengcai<sup>2,3,4,5</sup>, FAN Lili<sup>1</sup>, XIA Zhongshang<sup>2,3,4</sup>

1. School of Pharmacy, Guangxi University of Chinese Medicine, Nanning 530000, China

2. Key Laboratory of Pharmacodynamic Research of Chinese Medicine in Guangxi, Guangxi University of Chinese Medicine, Nanning 530000, China

3. Collaborative Innovation Center for Functional Component Research of Crop Waste in Guangxi, Guangxi University of Chinese Medicine, Nanning 530000, China

4. Key Laboratory of Theory and Transformation of Chinese Medicine for Dampness and Wetness Diseases in Guangxi, Guangxi University of Chinese Medicine, Nanning 530000, China

5. Guangxi Experimental Center of Traditional Chinese Medicine Science, Guangxi University of Chinese Medicine, Nanning 530000, China

收稿日期: 2023-12-16

基金项目: 广西中医药大学“桂派中医传承创新团队”资助项目 (2022A005)

作者简介: 杨 淇 (1997—), 男, 硕士研究生, 研究方向为中药药效筛选与应用研究。E-mail: yangqi2023@stu.gxtcmu.edu.cn

\*通信作者: 夏中尚, 男, 博士, 助理研究员, 研究方向为中药学。E-mail: 2671691281@qq.com

范丽丽, 女, 副教授, 研究方向为中药药效筛选与应用研究。E-mail: fanli228@163.com

**Abstract: Objective** To construct a machine learning model for predicting anti-radiation effects using the medicinal properties of traditional Chinese medicine (TCM) as feature descriptors, and elucidate the significant medicinal properties of anti-radiation TCM to guide clinical practice and daily radiation prevention and control. **Methods** Research literature on TCM with reported anti-radiation effects was collected from databases such as the Medicinal Herb Database, CNKI, and PubMed. The SymMap database was utilized to obtain the medicinal properties of TCM, such as taste, attributed meridians, etc., as recorded in authoritative texts like the Chinese Pharmacopoeia. These properties were used as feature descriptors to construct a database, which was then processed into a format suitable for machine learning. Five machine learning models-random forest, support vector machine, gradient elevator, logistic regression, and fully connected neural network-were employed to perform five-fold cross-validation on the dataset and assess performance. An external validation set, comprising 10 TCMs with reported anti-radiation effects and 10 TCMs without such reports, was used to test the models. Finally, the SHapley Additive exPlanations (SHAP) interpreter was used to visualize and interpret the significant medicinal properties that contribute to the anti-radiation effects. **Results** A total of 136 TCMs with reported anti-radiation effects were collected, including single herbs, health food registrations, and compound formulations, with a total frequency of 447 occurrences. The top three most frequently used TCMs were Lingzhi (*Ganoderma*), Hongjingtian (*Rhodiola Crenulatae Radix et Rhizoma*), and Gouqizi (*Lycii Fructus*), with each being used more than 10 times. Among the five machine learning models, the random forest exhibited the best performance, with an accuracy rate, F1 score, and area under curve (AUC) of 0.804 4, 0.773 2, and 0.879 8, respectively. The random forest model also demonstrated a good ability to correctly predict TCMs with reported anti-radiation effects in the external validation. The SHAP interpreter for the best-performing random forest model identified the functions of “tonifying deficiency, clearing heat”, the attributed meridians of “heart, liver, spleen, lung”, the tastes of “sour, sweet, bitter”, and the properties of “cold” as the most contributive to the anti-radiation effects. **Conclusions** This study is the first to apply the medicinal properties of TCM as feature descriptors in machine learning, achieving good predictive model performance. It provides guidance for the prevention and treatment of radiation sickness with TCM, suggesting that principles such as reinforcing healthy *qi* and eliminating pathogen, nourishing yin and reducing fire, and focusing on the heart, liver, spleen, and lung should be followed. Furthermore, the satisfactory performance of the machine learning models reflects the good interpretability and reproducibility of TCM theory in disease prevention and treatment.

**Key words:** machine learning; traditional Chinese medicine; drug properties; four natures and five flavors; antiradiation; *Ganoderma*; *Rhodiola Crenulatae Radix et Rhizoma*; *Lycii Fructus*

2023 年 8 月 24 日日本开始首轮排放福岛储存的核污水,使得抗辐射成为了热点话题<sup>[1]</sup>。辐射分为电离辐射与非电离辐射,其中电离辐射对人体的伤害最大,其作用于人体主要是诱导机体产生自由基导致损伤,从而导致机体所有器官不同程度的伤害,另还可致癌和引起胎儿死亡或畸形<sup>[2]</sup>。放射病目前主流的临床治疗方法为生物治疗、基因转染等,其中防护效果好的手段往往毒副作用太明显,严重限制了其应用<sup>[3]</sup>,因此,寻求安全有效的抗辐射药物对人体健康有重要意义,对此祖国医学提供了许多抗辐射新思路<sup>[4-5]</sup>。中医药治未病,防大于治,除物理防护辐射手段和化学药之外,中药或为安全且有效防治放射病的选择。

现有的研究资料显示,中药关于抗辐射作用的研究相对于其他功效的研究较少<sup>[5]</sup>,所以如何在当前的形势下进一步理解中药的抗辐射作用从而预测出潜在具有抗辐射作用的中药来指导放射病防治与药学研究则显得意义重大。但目前国内外大多关于药物药效的预测研究均是局限于分子水平的物质基础<sup>[6-9]</sup>,在药效预测常用的机器学习方法中,特征的

描述符也都以描述药物分子的理化性质、分子结构等为主<sup>[10-11]</sup>。为此本研究首次尝试使用中药药性作为特征描述符构建抗辐射作用预测模型,观察模型的性能与外部验证来判断其可靠性,同时分析中药的四气五味、归经等特征在抗辐射作用中的重要性,为后续的中药抗辐射研究提供创新方法,指导放射病的中药防治。

## 1 抗辐射中药数据库的构建

### 1.1 数据来源

本研究从公开的药智网数据库 (<https://db.yaozh.com/>) 收集具有抗辐射作用的药食同源中药,从中国知网与 PubMed 数据库收集具有抗辐射作用的单味中药和中药复方。

### 1.2 纳入标准

药食同源中药:①来源于包含“抗辐射”保健功能的产品,且有明确的“主要原料”记载;②“主要原料”至少包含 1 个中药品种;③中药品种符合《按照传统既是食品又是中药材的物质目录管理规定》<sup>[12]</sup>;④制剂辅料或添加剂可能为主要原料时,经评定其在保健食品配方中的作用,考虑删除或保

留,如白酒、醋等。

单味中药和中药复方:在中国知网与 PubMed 数据库中检索“抗辐射”“中药”等相关主题词或关键词,得到具有抗辐射作用的中药研究。纳入标准如下:①动物模型由辐射诱导;②干预措施为中药或复方;③结论明确表示干预的中药或复方有抗辐射作用。对于具有抗辐射作用报道的中药复方,根据纳入标准将复方中的所有中药品种都纳入到数据库当中并计算频次。本研究认为同一种中药若存在相同的抗辐射研究报道(即也被复方包含在内的品种),应当进行报道次数的加权作为其抗辐射的“证据力度”,所以保留了研究的报道次数用作后续机器学习的特征加权。

### 1.3 排除标准

①没有确切标注保健功能为“抗辐射”的保健食品配方;②同一单位注册且主要原料完全相同的保健食品配方;③“主要原料”未有记录的保健食品配方;④中药单体成分的抗辐射研究。

### 1.4 数据库的规范化和构建

使用 Excel 提取符合纳入标准的中药信息,提取内容包括名称、性味归经、功效、保健功能、主要原料等。随后对中药进行规范化处理,参考《中国药典》2020 年版、《中药大辞典》等统一中药的名称,如“高山红景天”规范为“红景天”、“怀山药”规范为“山药”等。最后检索《中国药典》2020 年版、《全国中草药汇编》《中药大辞典》《中华本草》及地方标准,以及结合中医药证候关联数据库 SymMap<sup>[13]</sup>得到中药对应的药性、药味、归经、功效分类等,其中微寒、微苦等带有“微”的均删除,统一为寒、苦等,规范化完成后构建出具抗辐射作用的中药数据库。

## 2 机器学习模型的训练与性能评估

### 2.1 数据集的构建

本研究的机器学习问题属于二元分类问题,需要准备正负标签的数据集。对“1.4”项得到的具抗辐射作用的中药数据库进行处理,提取其中不同中药的“药性”“药味”“归经”“功效”4 个特征类别,并以收集到的“具抗辐射作用”的中药作为正标签。对于负标签中药的收集(即无抗辐射作用报道的中药),通过以下方法在 SymMap 数据库的 704 种中药中获取:①参考正标签“具抗辐射作用”中药的功效归属,选择功效关系较远的中药,如具有抗辐射作用中药中“补虚”功效出现的次数较多,则选

择与“补虚”功效关系较远的“活血化瘀”或“涌吐”等类型的中药;②收集到的负标签中药的数量超过正标签 2 倍后,首先返回正标签数据集中进行检查,确保负标签中药不会与正标签重复出现,随后对待确认的负标签数据集进行随机抽取至数量与正标签一致。通过此方法保证了正、负标签数据集的差异,同时正、负标签的数量均衡使数据集更具有代表性且能涵盖更广泛的中药药性,进而提升模型的泛化性<sup>[14]</sup>。

### 2.2 数据集的格式转化

为了使中药药性特征成为适应机器学习的格式,本研究使用“热编码”技术对药性特征进行处理。独热编码(one-hot encoding)是一种处理分类变量的方法<sup>[15]</sup>,假设有一个“药味”变量,其包含多种可能的类别:“辛”“甘”“酸”等,在独热编码中会为每种药味创建 1 个新的变量:“是辛”“是甘”“是酸”等。每个新变量都赋予二进制 0 或 1 的值。如人参的药味是“甘、微苦”,那么在热编码后,“是甘”和“是苦”的值就会是 1,而其他的变量“是辛”“是酸”等则都是 0,这就表示人参的药味是甘、苦,而不是辛、酸或咸等,其他特征同理。最后对“1.4”项得到的数据中存在不同中药出现频次不同的情况,在进行机器学习之前将其进行频次加权,如灵芝有 36 次抗辐射作用的研究报道,那么在训练的数据集中会出现 36 行。

### 2.3 模型的训练

本研究使用随机森林、支持向量机、梯度提升、逻辑回归和全神经网络 5 种机器学习模型对数据集进行训练,涵盖了传统机器学习与深度学习,使用的编程语言为 Python。本研究的机器学习逻辑:①读取“2.2”项得到的适合学习格式的数据集并使用 5 折交叉进行训练(随机分配的训练集:测试集=4:1,5 次);②提取数据集的特征与标签;③对特征所属的中药频次进行加权;④定义模型并训练,参数设置为默认,主体代码为 models = {'Random Forest': RandomForestClassifier(), 'Support Vector Machine': SVC(probability=True), 'Gradient Boosting': Gradient BoostingClassifier(), 'Logistic Regression': Logistic Regression(max\_iter = 1 000), 'Neural Network': MLP Classifier(max\_iter = 1 000)};⑤对随机森林、支持向量机、梯度提升和逻辑回归进行特征重要性的 SHapley 加性解释(SHapley additive explanations, SHAP)<sup>[16]</sup>,而全神经网络由于复杂的非线性关系和

多层结构使得直接解释其工作原理和做出的特定预测变得极其困难，在本研究中不作 SHAP 解释；⑥性能评估与外部预测验证等。

由于本研究机器学习使用的多个 Python 库无法识别中文，所以将“2.1.1”项所构建的数据集表头以拼音表示，并将标签中“有抗辐射作用”的中药设置为“1”，“无抗辐射作用”的中药设置为“0”，需要注意的是“辛(xin)”与“心(xin)”等会被认为是同一事物，所以本研究将“辛(xin)”转化成了“辛味(xinwei)”。

### 3 模型性能评估与特征评价

#### 3.1 模型性能的评估

本研究首次探究使用中药药性作为特征描述符来构建机器学习的预测模型，因此能够较好预测抗辐射作用能力的模型不仅体现在预测的准确性，还能反映出中医药理论的可重现性。本研究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)、曲线下面积(area under curve, AUC)、受试者工作特征(receiver operator characteristic, ROC)曲线以及混淆矩阵<sup>[17]</sup>，在计算模型准确率前采用五折交叉法取平均值来评价模型性能并可视觉化为箱线图<sup>[18]</sup>。

准确率 = 正确分类的样本数 / 所有样本数

F1 分数 =  $2 \times (\text{精确率} \times \text{召回率}) / (\text{精确率} + \text{召回率})$

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 是评估二元分类器性能的指标之一，简单地说，AUC 测量了模型在不同阈值下的分类能力；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。混淆矩阵是一个表格，用来描述分类模型的性能，包括真正例、假负例、真负例和假正例，简单来说可以看到正负标签的预测正误情况。需要注意的是混淆矩阵将会体现出模型对正标签（即具抗辐射作用）和负标签（即不具抗辐射作用）各自的预测能力，将有力地指导预测的方向。五折交叉验证是将数据集分成 5 个部分，轮流将其中 4 个部分作为训练数据，剩下的一部分作为测试数据，进行 5 次训练和测试，最后取 5 次测试结果的平均值。

真正例率 = 真正例 / (真正例 + 假负例)

假正例率 = 假正例 / (假正例 + 真负例)

#### 3.2 抗辐射作用药性特征重要性评估

本研究选择 SHAP 解释器对随机森林、支持向

量机、梯度提升和逻辑回归 4 种机器学习模型的训练结果进行特征重要性可视化<sup>[19]</sup>，而全神经网络学习模型无法使用 SHAP 可视化特征的重要性。本研究抗辐射作用预测模型中使用的药性特征包括辛、甘、酸、苦、咸、寒、热、平和凉等，通过 SHAP 可视化可以了解到哪些药性特征对抗辐射效果的预测最为重要，以及这些特征是如何影响预测结果的。例如，SHAP 值显示“辛”特征对预测结果的贡献较大，那么就可以推断出具有“辛”特征的中药可能对放射病有较好的防治效果；相反，如果“寒”特征的 SHAP 值为负，便可以推断出“寒”特征可能会降低这味中药的抗辐射效果。

#### 3.3 抗辐射中药外部验证

为了验证模型的预测可靠性，本研究从中国知网、PubMed 等数据库获取了 10 个报道具有抗辐射作用的中药，并与数据集进行了遍历对照保证这 10 个中药不在先前构建的抗辐射中药数据库当中；同时，本研究使用随机法在 SymMap 数据库中收集了不重复的 10 种报道不具有抗辐射作用的中药，再到中国知网等数据库中确认，确认后作为预测共组成 20 味中药的外部验证数据集，如表 1 所示，并使用 5 种机器学习模型中综合性能最佳的模型进行预测。

### 4 结果

#### 4.1 抗辐射中药数据库

本研究获得 136 味具有抗辐射作用的中药，总频次为 447 次，其中中药使用频次及频率排名前 3 的为灵芝(36 次, 8.05%)、红景天(30 次, 6.71%)、枸杞子(24 次, 5.37%)，频次  $\geq 10$  的中药共 10 味，见表 2。涉及补虚药、清热药、活血化瘀药等 20 种功效，心、肝、肾等全部经络，以及全部药味和寒热属性。

由于本研究是初次使用中药药性作为机器学习特征，所以展示训练集构建的详细过程。表 3 展示了部分中药药性特征初步处理的结果，是从数据库初步收集到中药药性后整理得到的。而表 4 则是使用热编码方法，将表 3 的数据进行向量化，以符合机器学习的训练格式要求。需要注意的是 RDKit 库识别不了中文，所以从表 3 开始主要特征需要使用英文字母来进行描述，本研究采用拼音表示。通过“2.2”项的方法本研究最终得到了 44 个特征用以机器学习。

#### 4.2 模型测试集性能评估结果

5 个机器学习模型的准确率平均值经五折交叉

表 1 待预测抗辐射作用的 20 种中药药性

| Table 1 Properties of predicted 20 kinds of traditional Chinese medicines with radiation resistance effect |      |     |    |        |         |
|------------------------------------------------------------------------------------------------------------|------|-----|----|--------|---------|
| 序号                                                                                                         | 中药名称 | 药味  | 药性 | 归经     | 功效归属    |
| 1                                                                                                          | 木香   | 辛、苦 | 温  | 脾、胃、大肠 | 理气药     |
| 2                                                                                                          | 天山雪莲 | 苦   | 温  | 肝、脾、肾  | 补虚药     |
| 3                                                                                                          | 牛大力  | 甘、苦 | 平  | 肺、肾    | 补虚药     |
| 4                                                                                                          | 五指毛桃 | 甘   | 平  | 脾、肺、肝  | 补虚药     |
| 5                                                                                                          | 淡豆豉  | 辛、苦 | 凉  | 肺、胃    | 解表药     |
| 6                                                                                                          | 沙枣   | 酸、甘 | 凉  | 肺、胃、肝  | 补虚药     |
| 7                                                                                                          | 肿节风  | 辛、苦 | 平  | 心、肝    | 清热药     |
| 8                                                                                                          | 沙苑子  | 甘   | 温  | 肝、肾    | 补虚药     |
| 9                                                                                                          | 秦皮   | 苦、涩 | 寒  | 肝、胆、大肠 | 清热药     |
| 10                                                                                                         | 徐长卿  | 辛   | 温  | 肝、胃    | 祛风湿药    |
| 11                                                                                                         | 鬼箭羽  | 苦、辛 | 寒  | 肝、脾    | 活血化瘀药   |
| 12                                                                                                         | 红粉   | 辛   | 热  | 肺、脾    | 拔毒化腐生肌药 |
| 13                                                                                                         | 葫芦巴  | 苦   | 温  | 肾      | 补虚药     |
| 14                                                                                                         | 鸡骨草  | 甘、苦 | 凉  | 肝、胃    | 利水渗湿药   |
| 15                                                                                                         | 胡黄连  | 苦   | 寒  | 肝、胃、大肠 | 清热药     |
| 16                                                                                                         | 京大戟  | 苦   | 寒  | 脾、肺、肾  | 峻下逐水药   |
| 17                                                                                                         | 橘核   | 苦   | 平  | 肝、肾    | 理气药     |
| 18                                                                                                         | 莱菔子  | 辛、甘 | 平  | 脾、肺、胃  | 消食药     |
| 19                                                                                                         | 老鹳草  | 辛、苦 | 平  | 肝、肾、脾  | 祛风湿药    |
| 20                                                                                                         | 连钱草  | 辛、苦 | 寒  | 肝、肾、膀胱 | 利水渗湿药   |

表 2 抗辐射中药频次分布 (频次≥10)

| Table 2 Frequency distribution of traditional Chinese medicines with radiation resistance effect (frequency ≥ 10) |     |    |      |    |     |    |      |
|-------------------------------------------------------------------------------------------------------------------|-----|----|------|----|-----|----|------|
| 序号                                                                                                                | 名称  | 频次 | 频率/% | 序号 | 名称  | 频次 | 频率/% |
| 1                                                                                                                 | 灵芝  | 36 | 8.05 | 6  | 蜂胶  | 11 | 2.46 |
| 2                                                                                                                 | 红景天 | 30 | 6.71 | 7  | 葡萄籽 | 11 | 2.46 |
| 3                                                                                                                 | 枸杞子 | 24 | 5.37 | 8  | 当归  | 16 | 3.58 |
| 4                                                                                                                 | 黄芪  | 24 | 5.37 | 9  | 人参  | 15 | 3.36 |
| 5                                                                                                                 | 刺五加 | 12 | 2.68 | 10 | 螺旋藻 | 10 | 2.24 |

表 3 抗辐射中药药性特征的初步处理 (示例)

| Table 3 Preliminary treatment of property characters of traditional Chinese medicines with radiation resistance effect (example) |           |           |           |           |             |       |
|----------------------------------------------------------------------------------------------------------------------------------|-----------|-----------|-----------|-----------|-------------|-------|
| Name                                                                                                                             | Frequency | Taste1... | Character | Nature... | Attribution | label |
| 灵芝                                                                                                                               | 36        | 甘         | 平         | 心         | 补虚药         | 1     |
| 红景天                                                                                                                              | 30        | 甘         | 寒         | 肺         | 补虚药         | 1     |
| 枸杞子                                                                                                                              | 24        | 甘         | 平         | 肝         | 补虚药         | 1     |
| 胆矾                                                                                                                               | 1         | 酸         | 寒         | 肝         | 涌吐药         | 0     |
| 常山                                                                                                                               | 1         | 苦         | 寒         | 肺         | 涌吐药         | 0     |
| 白薇                                                                                                                               | 1         | 苦         | 寒         | 肝         | 清热药         | 0     |

验证后如图 1 所示，可以看到随机森林、梯度提升和支持向量机模型性能相对较好，特别是随机森林预测结果稳定且准确率高。逻辑回归与全连接神经网络的五折交叉准确率较低，中位数处于 0.5 附近，说明模型的预测能力几乎与随机预测没有区别。从表 5 和图 2 的 ROC 曲线可以看出，随机森林

的综合性能仍然是最佳的，其准确率高达 0.8 以上，F1 分数为 0.773 2，AUC 为 0.879 8，说明随机森林模型可以较好地预测待输入的数据集；梯度提升模型次之，准确率和 F1 分数在 0.76 左右，AUC 为 0.824 4；其余 3 个模型的综合性能均较差。而不同模型的 ROC 曲线越接近左上角其在不同阈值下的

表 4 最终用于机器学习的中药药性特征数据集 (示例)

| Name  | Frequency | xinwei... | han... | xin... | buxuyao... | label |
|-------|-----------|-----------|--------|--------|------------|-------|
| 灵芝    | 0.060 2   | 0         | 0      | 1      | 1          | 1     |
| 红景天   | 0.050 2   | 0         | 1      | 1      | 1          | 1     |
| 枸杞子   | 0.040 1   | 0         | 0      | 0      | 1          | 1     |
| 瓜蒂    | 0.001 7   | 0         | 0      | 0      | 0          | 0     |
| 贯叶金丝桃 | 0.001 7   | 1         | 0      | 0      | 0          | 0     |
| 核桃仁   | 0.001 7   | 0         | 0      | 0      | 0          | 0     |

表 5 5 种机器学习模型的性能评估

Table 5 Performance evaluation of five machine learning models

| 模型    | 准确率     | F1 分数   | AUC     |
|-------|---------|---------|---------|
| 随机森林  | 0.804 4 | 0.773 2 | 0.879 8 |
| 支持向量机 | 0.685 7 | 0.447 8 | 0.871 4 |
| 梯度提升  | 0.766 4 | 0.760 7 | 0.860 5 |
| 逻辑回归  | 0.526 1 | 0.327 4 | 0.824 4 |
| 全神经网络 | 0.561 2 | 0.119 5 | 0.820 2 |

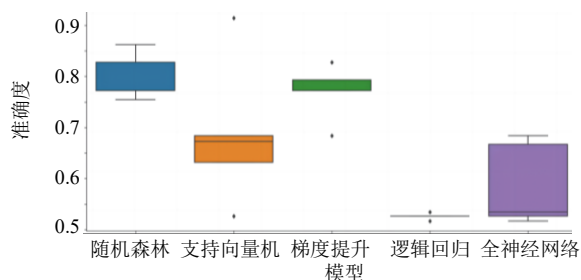


图 1 5 种机器学习模型的准确率交叉验证

Fig. 1 Cross-validation of accuracy of five machine learning models

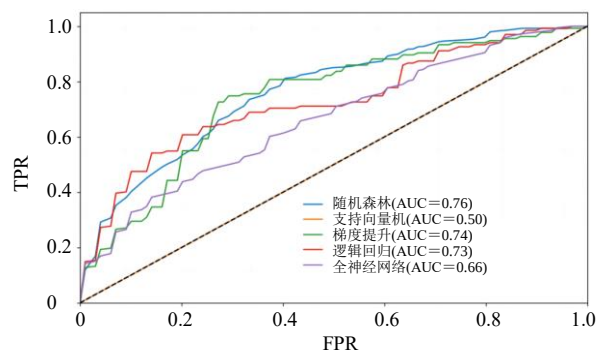


图 2 5 种机器学习模型的 ROC 曲线

Fig. 2 ROC curves for five machine learning models

预测能力越好, 可以看到随机森林模型的 ROC 曲线 AUC 最高。而支持向量机在不同阈值的预测能力都一样为 0.5, 可以认为其没有根据特征分类的能力。图 3 反映了 5 种机器学习模型区分预测测试集正、负标签的能力, 可以认为是左上与左下图的颜色越浅, 其对正、负标签的预测能力越强; 左下与左上的颜色越深, 其误判率越低。可以看到随机森林模型最符合上述“答对率”最高与“误判率”最低的描述。

#### 4.3 抗辐射作用药性特征重要性评估结果

由于全神经网络模型无法进行 SHAP 解释, 所以仅计算随机森林、支持向量机、梯度提升和逻辑回归模型的 SHAP 值并可视化。如图 4 所示, 红色代表“特征重要性”, 蓝色代表“特征不重要性”, 横坐标的左侧代表着“负标签(无抗辐射作用)”的贡献程度, 右侧代表着“正标签(有抗辐射作用)”的贡献程度, 简单来说, 红色散点在正坐标轴(右侧)越高越远越多, 那么这个特征在抗辐射作用中重要性越大, 反之红色散点在负坐标轴(左侧)越高越远越多, 则这个特征在抗辐射作用中重要性越小。如图 4 所示, 性能较好的随机森林模型给出的结果可信度相对高, 被放大以加强显示。可以看到 4 个机器学习模型虽性能差异较大, 但给出的药性特征重要性基本相同, 以随机森林模型认为“补虚”“清热”“心经”是具抗辐射作用中药的前 3 重要特征, 而“祛风湿”“辛味”“热性”是使中药远离抗辐射作用的前 3 重要的特征。

#### 4.4 抗辐射作用中药的外部验证结果

20 味中药的抗辐射作用预测与验证结果见表 6, 其中编号 1~10 的中药为已报道有抗辐射作用, 可以看到随机森林对已经具有确切抗辐射作用的木香、天山雪莲等 10 味中药给出了具抗辐射作用可

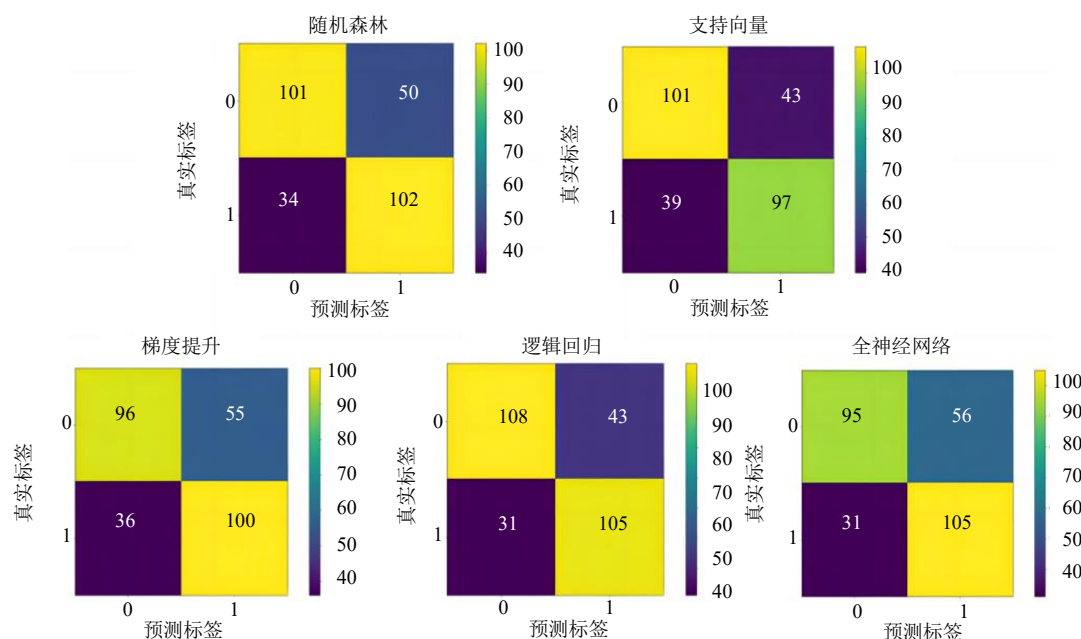


图 3 5 种机器学习模型的测试集预测结果

Fig. 3 Predict outcomes of five machine learning models for test sets

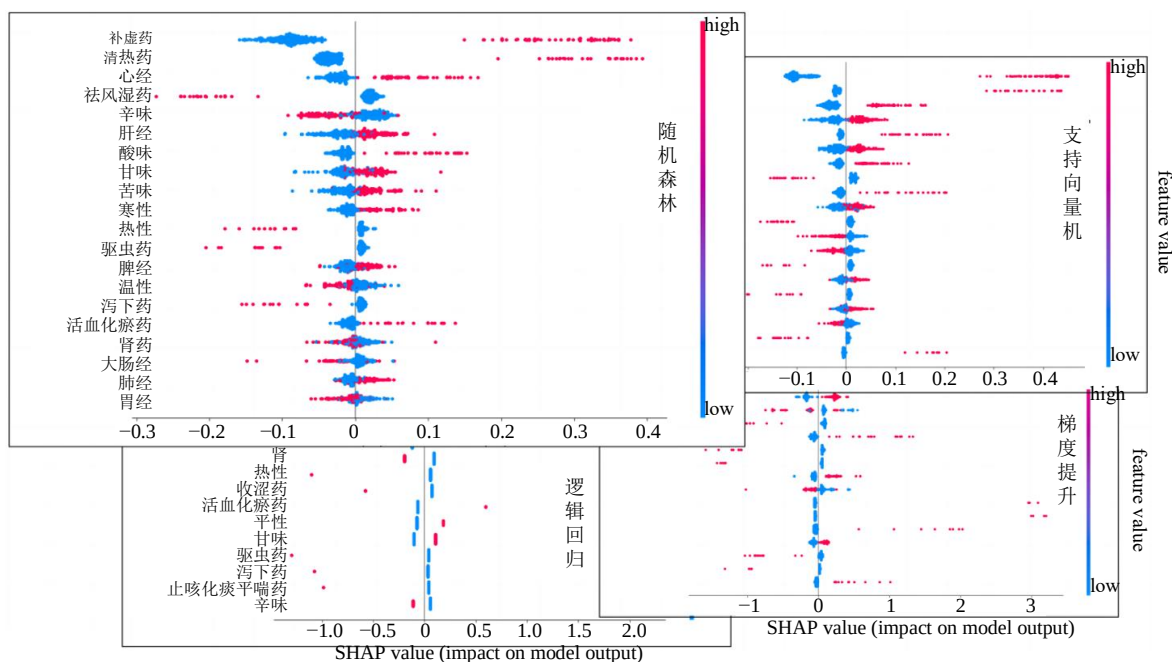


图 4 基于 4 种机器学习模型的抗辐射作用重要特征评估

Fig. 4 Evaluation of important features of anti-radiation effects based on four machine learning models

能性均在 0.7 以上的结果, 而对于暂未报道有抗辐射作用的鬼箭羽、鸡骨草、京大戟、莱菔子和连钱草给出了 0.9 以上的具抗辐射作用可能性, 红粉与葫芦巴可能性较低。

## 5 讨论

本研究使用中药药性作为机器学习的特征描述

符成功构建了抗辐射作用的中药预测模型, 并使用 SHAP 解释器对其中 4 种机器学习训练的结果进行了药性特征重要性的描述, 指导了放射病的中医药防治研究方向, 同时本研究是首次将中药药性作为特征应用到机器学习中并取得了良好的模型性能。但在探索的过程存在着许多需要讨论和注意的问题。



表 6 20 味中药的抗辐射作用预测与验证

Table 6 Prediction and verification of with radiation resistance effect of 20 traditional Chinese medicines

| 序号 | 中药名称 | 抗辐射与否 | 参考文献 | 预测可能性   | 序号 | 中药名称 | 抗辐射与否 | 参考文献 | 预测可能性   |
|----|------|-------|------|---------|----|------|-------|------|---------|
| 1  | 木香   | 是     | 20   | 0.940 0 | 11 | 鬼箭羽  | 不明    | 未报道  | 0.950 0 |
| 2  | 天山雪莲 | 是     | 21   | 0.950 0 | 12 | 红粉   | 不明    | 未报道  | 0.650 0 |
| 3  | 牛大力  | 是     | 22   | 0.770 0 | 13 | 葫芦巴  | 不明    | 未报道  | 0.640 0 |
| 4  | 五指毛桃 | 是     | 23   | 1.000 0 | 14 | 鸡骨草  | 不明    | 未报道  | 0.950 0 |
| 5  | 淡豆豉  | 是     | 24   | 0.920 0 | 15 | 胡黄连  | 不明    | 未报道  | 0.780 0 |
| 6  | 沙枣   | 是     | 25   | 0.983 3 | 16 | 京大戟  | 不明    | 未报道  | 0.930 0 |
| 7  | 肿节风  | 是     | 26   | 0.980 0 | 17 | 橘核   | 不明    | 未报道  | 0.870 0 |
| 8  | 沙苑子  | 是     | 27   | 0.987 5 | 18 | 莱菔子  | 不明    | 未报道  | 0.990 0 |
| 9  | 秦皮   | 是     | 28   | 0.980 0 | 19 | 老鹳草  | 不明    | 未报道  | 0.780 0 |
| 10 | 徐长卿  | 是     | 29   | 0.700 0 | 20 | 连钱草  | 不明    | 未报道  | 0.960 0 |

### 5.1 使用中药药性作为特征描述符的合理性

中药药性基于中医药理论，是多成分作用于多靶点协同调控多条信号通路从而发挥整体作用<sup>[30]</sup>。本研究使用的“中药药性”属于一种表征的特征，或者说是复杂的因素（如多成分-多靶点）交织融合而表现出来的现象<sup>[31]</sup>，从而被记录下来形成“四气五味”“归经”等药性。这些药性实际上已经体现了中药的药理效应、药物动力、药物代谢等，与中医药理论的“整体观念”“对症治疗”理论高度吻合<sup>[32]</sup>。目前已有研究使用机器学习来解释中医理论，如秘红英等<sup>[33]</sup>采用贝叶斯算法研究吴以岭教授的用药规律，总结了其治疗冠心病的“病、证、症、法、方、药”特点。但关于中药的机器学习研究中，目前鲜有使用“中医药理论-疾病治疗”的方法进行中中药的机器学习研究<sup>[34]</sup>，而都是围绕着微观分子的特征来展开<sup>[35-36]</sup>。但本研究所训练出来的机器学习模型特别是随机森林性能良好，可以较好地应对不同药性特征的输入，从而给出较为准确的预测结果，这点体现在模型性能评估和外部验证以及 SHAP 解释的结果当中。

目前关于机器学习的研究国内外大多使用“分子描述符”，如相对分子质量（ $M_w$ ）、对数分配系数（ $\lg P$ ）、拓扑极性表面积（TPSA）、电荷分布（FASA）、拓扑指纹等<sup>[37]</sup>，但都是基于分子理化性质与“原子论”学说<sup>[38]</sup>，存在着化学物质基础收集不全面与仅有定性描述而可能造成认识片面的情况发生，如基于单纯理化性质（如柔性）的配体-靶点的分子对接就会出现结果与实验验证时常不一致的情况发生。

### 5.2 使用中药药性作为特征描述符的局限性

自然辩证法阐释了矛盾的对立统一<sup>[39]</sup>，虽然相比于化学药常使用的分子拓扑结构描述符的“黑箱操作”（即难以解释特征）<sup>[40]</sup>，使用表征的中药药性来描述特征具有较好的解释性，但也更容易出现“不确定”的情况，其中最容易发生的情况就是“质量稳定”与“认可度”问题。如大黄可能会因为基原、炮制、产地、历史等差异而有“泻下攻积、清热泻火、凉血解毒、逐瘀通经、利湿退黄”多种功效，在“功效归类”中可以是泻下药也可以是清热药、活血化瘀药、利水渗湿药等<sup>[41]</sup>；同时一种中药甚至会包含寒热、浮沉、补泻等相反药性的情况<sup>[42]</sup>，根据阴阳学说这是一种必然的现象<sup>[43]</sup>。

除此之外，由于描述的表征性使得本研究训练得到的机器学习预测模型可能存在“过度泛化”的现象，即只要归心经或寒凉药，机器学习模型可能会认为其或多或少有抗辐射的作用，因为从机制上来说抗氧化、抗炎等也属于抗辐射的范畴<sup>[44]</sup>，这使得机器学习模型在指导抗辐射治疗研究的过程中可能会出现特异性较差的情况，弱化了“同病异治”“辨证论治”等中医独特治法的应用。

### 5.3 关于机器学习过程的讨论

本研究采用了随机森林、支持向量机、梯度提升、逻辑回归和全神经网络 5 种机器学习模型，是取决于特征的特点。根据中药的四气五味、归经和功效，最终标准化特征得到了中药 44 个特征描述，相比于摩根指纹的 1 024 位特征<sup>[45]</sup>或 166 位的分子接入系统键（molecular access system key, MACCS）键型指纹<sup>[46]</sup>等来说特征数量较少，而随机森林等传



统机器学习模型是比较适用于特征数量较少的情况,并且解释能力强<sup>[47]</sup>,与本研究的初衷一致,即希望通过机器学习获得抗辐射作用的重要特征以指导放射病的中医药防治方法。除此之外,本研究也使用了深度学习中的全神经网络,目的是用以比较传统机器学习和深度学习在学习中药药性特征中的性能,结果显示传统机器学习特别是随机森林模型的综合性能最佳。

需要注意的是,本研究通过了详尽的检索,获得了 136 味不重复的有实验研究报道的抗辐射作用中药,频次在 400 以上,在常用中药中占比较高,《中国药典》2020 年版收载 616 种中药,以单味中药为单位的数据库有别于以亿为数量级的单体化合物数据库,所以不论是抗辐射作用还是其他功效,使用药性作为特征描述符进行机器学习的过程中,数据库的样本数量就不可避免地存在着较少的情况。药性特征本研究仅采纳了四气五味、归经与功效归类 4 种,其他如升降沉浮、发散收敛、有毒无毒等药性<sup>[48]</sup>暂未被纳入,在后续的研究中可以丰富药性特征以更全面地描述中药,提高预测模型的性能和泛化能力。

## 6 结论

从随机森林模型的 SHAP 解释器结果可以看出,以补虚、清热功效,归心、肝、脾、肺经,酸、甘、苦味,寒性特征对抗辐射正向作用贡献最大,在临床用药中一般应优先考虑使用此类药性的中药;而祛风湿、驱虫、泻下功效,归肾、大肠、胃经,辛味,热、温性特征对抗辐射起到反作用,在临床用药中一般应考虑避免此类药性的中药。辐射导致的放射病在中医学上被认为是火热温毒致病<sup>[49]</sup>,辐射应为六淫之中的“火邪”,且有经脉传化、脏腑传化、卫气营血传化而致全身患病的特点<sup>[50]</sup>。发病之初,热伤皮毛,卫先受之,病邪由此而入。肺主皮毛,肺卫受损。热邪日久,传及他脏<sup>[51]</sup>,最终“火毒伤阴”,导致气血两虚等证。本研究的机器学习结果与中医药理论、放射病的临床表现均吻合,总结出对于辐射火毒所致损伤,当以扶正祛邪、滋阴降火、主治心、肝、脾、肺为治疗原则的结论。

但使用中药药性作为特征描述符来进行机器学习,若要取得精准的预测结果仍是一个挑战,学习的结果在一定程度上可以指导抗辐射的中医药防治或加以解释,也可以在研究的开始阶段指导中药的抗辐射或其他替换功效的药效研究,但将分子性质

与表观药性“中西结合”,从小到大、由浅入深地描述物质可能会得到更好的结果。总之,本研究结果表明这将是有效的传承创新的必然过程。

**利益冲突** 所有作者均声明不存在利益冲突

## 参考文献

- [1] 刘国信. 日本核污染水排海带来的影响不可小觑 [J]. 防灾博览, 2023(5): 48-53.
- [2] 李秀婷, 胡晓江, 董秋, 等. 低剂量电离辐射长期暴露对人体健康影响研究进展 [J]. 职业卫生与应急救援, 2023, 41(5): 652-656.
- [3] 付伟, 杜光. 核辐射防治药物的基础研究进展 [J]. 世界临床药物, 2022, 43(8): 1012-1018.
- [4] 辛佳芸, 魏艳平, 丁倩倩, 等. 中药抗辐射活性成分及其作用机制研究进展 [J]. 中华中医药学刊, 2023, 41(7): 166-172.
- [5] 韦唯唯, 杜光. 抗辐射中药的研究进展 [J]. 湖北科技学院学报: 医学版, 2023, 37(1): 84-89.
- [6] Juul R V, Rasmussen M H, Agersø H, et al. Pharmacokinetics and pharmacodynamics of once-weekly somapacitan in children and adults: Supporting dosing rationales with a model-based analysis of three phase I trials [J]. *Clin Pharmacokinet*, 2019, 58(1): 63-75.
- [7] Ebata S, Oba K, Kashiwabara K, et al. Predictors of rituximab effect on modified Rodnan skin score in systemic sclerosis: A machine-learning analysis of the DesiReS trial [J]. *Rheumatology*, 2022, 61(11): 4364-4373.
- [8] Alves V M, Bobrowski T, Melo-Filho C C, et al. QSAR modeling of SARS-CoV m<sup>pro</sup> inhibitors identifies sufugolix, cenicriviroc, proglumetacin, and other drugs as candidates for repurposing against SARS-CoV-2 [J]. *Mol Inform*, 2021, 40(1): e2000113.
- [9] Yang X, Feng Y, Liu Y, et al. Fuzheng Jiedu Xiaoji formulation inhibits hepatocellular carcinoma progression in patients by targeting the AKT/CyclinD1/p21/p27 pathway [J]. *Phytomedicine*, 2021, 87: 153575.
- [10] 刘皓淼, 杨志伟, 王力卓, 等. 基于机器学习的药物-靶标相互作用预测 [J]. 中国生物工程杂志, 2022, 42(4): 40-48.
- [11] 王振娟, 韩乐, 李雪婵, 等. 机器学习在药物筛选中的研究进展 [J]. 中国医院药学杂志, 2023, 43(22): 2584-2588.
- [12] 国家卫生健康委关于印发《按照传统既是食品又是中药材的物质目录管理规定》的通知 [J]. 中华人民共和国国家卫生健康委员会公报, 2021(11): 7-9.
- [13] Wu Y, Zhang F L, Yang K, et al. SymMap: An integrative database of traditional Chinese medicine enhanced by symptom mapping [J]. *Nucleic Acids Res*, 2019, 47(D1): D1110-D1117.
- [14] Wu Y F, Fang Y. Stroke prediction with machine learning methods among older Chinese [J]. *Int J Environ Res Public Health*, 2020, 17(6): 1828.
- [15] Okada S, Ohzeki M, Taguchi S. Efficient partition of

- integer optimization problems with one-hot encoding [J]. *Sci Rep*, 2019, 9(1): 13036.
- [16] Ma Y B, Zhang D, Xu J, *et al*. Explainable machine learning model reveals its decision-making process in identifying patients with paroxysmal atrial fibrillation at high risk for recurrence after catheter ablation [J]. *BMC Cardiovasc Disord*, 2023, 23(1): 91.
- [17] Cagirici H B, Budak H, Sen T Z. G4Boost: A machine learning-based tool for quadruplex identification and stability prediction [J]. *BMC Bioinformatics*, 2022, 23(1): 240.
- [18] Gossert A D. Assessing molecular interactions with biophysical methods using the validation cross [J]. *Biochem Soc Trans*, 2019, 47(1): 63-76.
- [19] El-Sappagh S, Alonso J M, Islam S M R, *et al*. A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease [J]. *Sci Rep*, 2021, 11(1): 2660.
- [20] 李硕, 王志允, 常华杰, 等. 木香 HPLC 指纹图谱的建立及其抗氧化、抗辐射作用谱效关系研究 [J]. 中草药, 2023, 54(8): 2580-2587.
- [21] 周湘洁, 周晓春, 张丽芬, 等. 雪莲培养细胞提取物对  $^{60}\text{Co}$   $\gamma$  射线照射后小鼠外周血细胞的保护作用 [J]. 航天医学与医学工程, 2015, 28(1): 16-21.
- [22] 陈晓白, 王晓平, 赵仕花, 等. 牛大力对  $^{60}\text{Co}$   $\gamma$  射线致小鼠造血系统损伤的保护作用 [J]. 中成药, 2013, 35(9): 1852-1856.
- [23] 黄明欣, 郑思新, 王培仁, 等. 若干香豆素衍生物对小鼠抗辐射作用的研究 [J]. 军事医学科学院院刊, 1986, 10(6): 409-412.
- [24] 毛峻琴, 徐超斗, 陈大贵, 等. 淡豆豉提取物对急性辐射损伤小鼠的保护作用 [J]. 解放军药学报, 2014, 30(2): 110-113.
- [25] 杨建华, 闫瑶, 胡君萍, 等. 沙枣总酚抗辐射活性研究 [J]. 时珍国医国药, 2013, 24(9): 2120-2122.
- [26] 张海东, 王仁生, 马姗姗, 等. 肿节风颗粒对小型猪腮腺放射损伤所致活性氧簇的清除作用 [J]. 南方医科大学学报, 2011, 31(1): 93-95.
- [27] 齐琳, 吴文倩, 陈韶华, 等. 沙苑子黄酮对辐射损伤的抗氧化保护作用 [J]. 上海中医药杂志, 2011, 45(5): 73-77.
- [28] 王波. 秦皮提取物的抗辐射研究 [J]. 陕西农业科学, 2010, 56(4): 40-42.
- [29] 朱世权, 蔡文秀, 薛玲, 等. 徐长卿多糖的分离纯化及其抗辐射和升高白细胞的作用 [J]. 中草药, 2010, 41(1): 103-106.
- [30] 翟兴英, 李洋, 涂珺, 等. 基于多靶点协同作用中药药效物质评价指数构建及其在葛根治疗糖尿病药效物质基础分析中应用 [J]. 中国药理学与毒理学杂志, 2023, 37(S1): 33.
- [31] 秦林, 彭欣. 现代中药法象理论研究 (I): 论药象与药性的辩证关系及意义 [J]. 山东中医药大学学报, 2023, 47(5): 548-553.
- [32] 彭亚倩, 唐辉, 韩彦琪, 等. 中药五味药性理论的现代研究进展 [J]. 药物评价研究, 2023, 46(9): 2014-2023.
- [33] 秘红英, 魏聪, 李红蓉, 等. 基于贝叶斯算法的吴以岭教授治疗冠心病心绞痛的病案分析研究 [J]. 中国中西医结合杂志, 2018, 38(5): 534-538.
- [34] 姜皓, 张冰, 张晓滕, 等. 基于 4 种机器学习算法的妊娠期中药“禁忌慎”判别 [J]. 中草药, 2021, 52(24): 7596-7605.
- [35] 刘滇, 王海. 基于加权基因共表达网络分析和机器学习筛选肾母细胞瘤特征基因及其靶向中药预测 [J]. 中医临床研究, 2023, 15(20): 43-49.
- [36] 刘洋, 曹赛雅, 冯月娇, 等. 应用机器学习和神经网络模型识别结直肠癌“炎癌转化”过程的关键基因及防治中药预测 [J]. 中草药, 2023, 54(19): 6386-6399.
- [37] 庞国庆, 严沛鑫, 周康乔. 基于机器学习的抗乳腺癌候选药物筛选模型优化 [J]. 南通职业大学学报, 2023, 37(2): 67-72.
- [38] 佟艺辰. 在原子论、元素论和经验主义之间: 若弗鲁瓦的化学研究 [J]. 自然辩证法通讯, 2023, 45(6): 53-64.
- [39] 王鹏伟. 马克思恩格斯共同体思想的本质与发展——基于自然辩证法的维度 [J]. 系统科学学报, 2024(4): 1-6.
- [40] Hedderich D M, Eickhoff S B. Machine learning for psychiatry: Getting doctors at the black box? [J]. *Mol Psychiatry*, 2021, 26(1): 23-25.
- [41] 段展辉, 刘桂荣. 从量效关系探讨张志远运用大黄经验 [J]. 中医杂志, 2022, 63(13): 1219-1222.
- [42] 王洋洋, 匡海学, 苏发智, 等. 中药四性的临床应用价值及与五味的关系 [J]. 中草药, 2023, 54(4): 1329-1341.
- [43] 姚春鹏, 姚丹. 论李杲医学的易理基础 [J]. 周易研究, 2021(4): 85-91.
- [44] 杨露勋, 杨新瑞, 刘晶, 等. EGCG 在抗辐射方面的研究进展 [J]. 中国辐射卫生, 2023, 32(6): 700-704.
- [45] Zhong S F, Guan X H. Count-based morgan fingerprint: A more efficient and interpretable molecular representation in developing machine learning-based predictive regression models for water contaminants' activities and properties [J]. *Environ Sci Technol*, 2023, 57(46): 18193-18202.
- [46] Xia Z H, Yan A X. Computational models for the classification of mPGES-1 inhibitors with fingerprint descriptors [J]. *Mol Divers*, 2017, 21(3): 661-675.
- [47] Rigatti S J. Random forest [J]. *J Insur Med*, 2017, 47(1): 31-39.
- [48] 杨曙光, 王鹏, 开菲. 桔梗升降浮沉药性考辨 [J]. 中国中医基础医学杂志, 2022, 28(8): 1323-1325.
- [49] 高明泽, 徐文慧, 王天琪, 等. 急性辐射损伤的中医病因学探讨 [J]. 中医杂志, 2016, 57(6): 454-457.
- [50] 张悦, 梁建庆, 李金田, 等. 辐射毒与中医火毒相关性探讨 [J]. 中华中医药杂志, 2022, 37(10): 6084-6087.
- [51] 王子骁, 梁小兰, 李淦, 等. 电磁辐射损伤的中医研究进展 [J]. 职业与健康, 2021, 37(5): 707-709.

[责任编辑 潘明佳]