

可解释机器学习模型在轻度认知障碍早期筛查中的应用*

刘 刚¹, 余 强², 杜建霖², 王凯娟^{1,3△}

1. 郑州大学公共卫生学院, 河南郑州 450001; 2. 重庆医科大学附属第二医院心血管内科, 重庆 400010;
3. 河南省肿瘤流行病学重点实验室/河南省肿瘤标志物与分子影像学国际联合实验室, 河南郑州 450001

摘要:目的 基于我国社区老年人易获取的非影像学特征, 建立并验证可解释的机器学习模型, 用于轻度认知障碍(MCI)的早期筛查。方法 采用中国健康与养老追踪调查(CHARLS)2011—2022 年纵向数据, 纳入年龄 ≥ 60 岁且完成认知评估的研究对象 1 198 例, 按 7 : 3 比例随机分为训练集($n=838$)和验证集($n=360$)。在 12 项候选预测变量中以 Boruta 算法筛选预测特征, 构建 XGBoost 等多种机器学习模型。采用受试者工作特征曲线下面积(AUC)、准确度、灵敏度、特异度、Brier 分数评价模型性能, 用决策曲线分析(DCA)评估临床实用性, 采用 SHAP 方法进行可解释性分析。结果 训练集内部, 与非 MCI 组相比, MCI 组年龄更大、受教育年限更短, 农村户籍和自评健康状况差的比例更高, 抑郁症状评分更高, 即时记忆、延迟记忆和综合认知功能得分更低, 平均红细胞体积(MCV)及空腹血糖水平更高, 血红蛋白水平更低, 差异均有统计学意义(均 $P < 0.05$)。Boruta 算法从训练集 12 项候选预测变量中, 筛选出 4 项变量: 延迟记忆得分、即时记忆得分、综合认知功能得分、MCV。在验证集中, XGBoost 模型表现最佳, AUC 为 0.854 (95%CI: 0.810~0.898), 准确度为 0.819, 灵敏度为 0.831, 特异度 0.803, Brier 分数 0.122, 明显大于 AdaBoost (AUC=0.801)、MLP (AUC=0.795)、Lasso-Logistic (AUC=0.792)、递归特征消除(RFE)-Logistic (AUC=0.781)模型的 AUC, 差异均有统计学意义($P < 0.05$)。剔除 MCV 后, XGBoost 模型 AUC 由 0.854 (95%CI: 0.810~0.898)降至 0.777 (95%CI: 0.729~0.824), Δ AUC=0.077 (95%CI: 0.034~0.120), 差异有统计学意义($Z=3.52, P < 0.001$)。DCA 分析显示, XGBoost 模型在 15%~65% 阈值概率区间内净获益最高。SHAP 可解释性分析揭示, 延迟记忆得分对模型预测贡献最大, MCV 升高是发生 MCI 的独立危险因素($P < 0.05$)。结论 基于 4 项易获取指标(包括 MCV)的 XGBoost 模型具有良好的判别性能、临床实用性和可解释性, 适用于基层医疗机构开展 MCI 早期筛查。MCV 作为检测简便、低成本的血液生物标志物, 对 MCI 具有独立预测价值, 值得进一步探索其潜在机制和临床应用前景。

关键词:轻度认知障碍; 机器学习; 早期筛查; 可解释性人工智能; 平均红细胞体积

中图法分类号:R749.1; R446.1

文献标志码:A

文章编号:1672-9455(2026)16-2173-09

Application of explainable machine learning models in early screening for mild cognitive impairment*

Liu Gang¹, She Qiang², Du Jianlin², Wang Kaijuan^{1,3△}

1. School of Public Health, Zhengzhou University, Zhengzhou, Henan 450001, China; 2. Department of Cardiovascular Medicine, the Second Affiliated Hospital of Chongqing Medical University, Chongqing 400010, China; 3. Key Laboratory of Cancer Epidemiology of Henan Province/International Joint Laboratory of Cancer Biomarkers and Molecular Imaging of Henan Province, Zhengzhou, Henan 450001, China

Abstract: Objective To develop and validate an explainable machine learning model based on easily accessible non-imaging features in community-dwelling older adults in China for early screening of mild cognitive impairment (MCI). **Methods** Longitudinal data from the China Health and Retirement Longitudinal Study (CHARLS) spanning 2011—2022 were obtained. A total of 1 198 participants aged ≥ 60 years who completed cognitive assessments were included and randomly divided into a training set ($n=838$) and a validation set ($n=360$) at a 7 : 3 ratio. The Boruta algorithm was employed to select predictive features from 12 candidate

* 基金项目: 重庆市技术创新与应用发展专项重点项目(CSTB2023TIAD-KPX0048)。

作者简介: 刘刚, 男, 在读硕士研究生, 主要从事慢性病流行病学的多智能体应用、认知障碍早筛、可解释机器学习与基层医疗 AI 决策支持方向的研究。△ 通信作者, E-mail: kjwang@163.com。

中文引用格式: 刘刚, 余强, 杜建霖, 等. 可解释机器学习模型在轻度认知障碍早期筛查中的应用[J]. 检验医学与临床, 2026, 23(16): 2173-2180.

英文引用格式: Liu G, She Q, Du J L, et al. Application of explainable machine learning models in early screening for mild cognitive impairment[J]. Lab Med Clin, 2026, 23(16): 2173-2180.

variables, and multiple machine learning models including XGBoost were constructed. Model performance was evaluated using area under the receiver operating characteristic curve (AUC), accuracy, sensitivity, specificity and Brier score. Clinical utility was assessed via decision curve analysis (DCA), and SHAP was applied for model interpretability analysis. **Results** Within the training set, compared with the non-MCI group, the MCI group had older age, fewer years of education, higher proportions of rural household registration and poor self-rated health status, higher depressive symptom scores, lower scores of immediate memory, delayed memory and composite cognitive function, higher levels of mean corpuscular volume (MCV) and fasting blood glucose, and lower level of hemoglobin, with statistically significant differences ($P < 0.05$). Four variables were selected by the Boruta algorithm from the 12 candidate predictors in the training set: delayed memory, immediate memory score, composite cognitive function score and MCV. In the validation set, XGBoost demonstrated the best performance with an area under the curve (AUC) of 0.854 (95%CI: 0.810–0.898), accuracy of 0.819, sensitivity of 0.831, specificity of 0.803, and Brier score of 0.122. Its AUC was significantly higher than that of the AdaBoost (AUC=0.801), MLP (AUC=0.795), Lasso-Logistic (AUC=0.792) and recursive feature elimination (RFE)-Logistic (AUC=0.781) models, with statistically significant differences ($P < 0.05$). After excluding MCV, the AUC of the XGBoost model decreased from 0.854 (95%CI: 0.810–0.898) to 0.777 (95%CI: 0.729–0.824), with a Δ AUC of 0.077 (95%CI: 0.034–0.120), with statistically significant difference ($Z = 3.52, P < 0.001$). DCA result indicated that the XGBoost model yielded the highest net benefit within a threshold probability range of 15% to 65%. Furthermore, SHAP interpretability analysis revealed that delayed memory contributed the most to the model's predictions, and elevated MCV was identified as an independent risk factor for MCI ($P < 0.05$). **Conclusion** The XGBoost model based on four readily available indicators (including MCV) demonstrates good discrimination, clinical practicality and interpretability, making it suitable for MCI screening in primary care settings. MCV, as a convenient and low-cost blood biomarker, shows independent predictive value for MCI and warrants further investigation into its underlying mechanisms and clinical application prospects.

Key words: mild cognitive impairment; machine learning; early screening; explainable artificial intelligence; mean corpuscular volume

轻度认知障碍(MCI)是介于正常认知老化与痴呆之间的过渡状态,表现为记忆或其他认知领域的轻度损害,但日常生活能力基本保留。MCI被认为是阿尔茨海默病及其他类型痴呆的高危状态,每年有10%~15%的MCI患者进展为痴呆,而认知正常老年人的年转化率仅为1%~2%^[1]。全球痴呆患者数量预计将从2019年的5700万增至2050年的1.53亿^[2],给家庭和社会带来沉重负担。我国是全球老龄化速度最快的国家之一,≥60岁人群MCI患病率高达15.5%,约有3877万MCI患者^[3],形势尤为严峻。

早期识别MCI对于延缓认知衰退、预防痴呆具有重要意义。研究表明,在MCI阶段进行生活方式干预、认知训练和危险因素管理,可有效延缓甚至逆转认知功能下降^[4]。然而,MCI的早期症状隐匿,患者及家属常将其误认为正常老化而延误就诊。目前临床诊断MCI主要依赖神经心理学专科评估、脑脊液生物标志物检测或正电子发射断层扫描等方法,这些方法耗时长、成本高、专业性强,难以在基层医疗机构和社区大规模推广。因此,开发简便、经济、准确的MCI早期筛查工具,对于实现认知障碍防治的关口前移和分级诊疗具有重要的公共卫生意义^[5]。

近年来,机器学习方法在疾病预测领域展现出优越性能,能够自动识别复杂的非线性关系和高阶交互

作用,在医学影像分析、疾病风险预测等方面取得显著进展。在MCI预测领域,已有研究尝试利用机器学习整合多模态数据构建预测模型^[6-8],但多数研究依赖神经影像指标或昂贵的生物标志物,限制了模型在资源有限地区的推广应用。此外,传统机器学习模型常因“黑箱”特性而难以获得临床医生信任,可解释性不足成为制约其临床转化的重要瓶颈^[9]。因此,急需开发基于易获取指标、兼具良好性能和可解释性的MCI早期筛查模型。

值得关注的是,外周血生物标志物因其检测简便、成本低廉而备受关注。平均红细胞体积(MCV)是血常规检查的常规指标,近年来有研究提示其与认知功能存在关联^[10-11],但MCV对MCI的长期预测价值尚缺乏大样本前瞻性队列研究的验证。

基于上述背景,本研究利用中国健康与养老追踪调查^[12](CHARLS)2011–2022年纵向数据,旨在:(1)基于社区老年人易获取的非影像学指标,采用机器学习方法构建MCI早期筛查模型;(2)比较多种机器学习算法与传统Logistic回归模型的预测性能;(3)采用可解释性人工智能方法揭示模型决策逻辑;(4)探索MCV作为新型血液生物标志物对MCI的独立预测价值。本研究结果可为基层医疗机构开展MCI早期筛查提供简便、可靠的工具,助力实现“健康

老龄化”国家战略目标。现报道如下。

1 资料与方法

1.1 研究设计与数据来源 本研究为基于前瞻性纵向队列的二次数据分析,数据来源于 CHARLS。CHARLS 是由北京大学国家发展研究院主持的全国性代表性调查,采用多阶段分层概率比例抽样方法,覆盖全国 28 个省(自治区、直辖市)的 150 个县(区)、450 个村(居)委会,系统收集 45 岁及以上中老年人群的社会、经济、健康等多维度信息。本研究利用 CHARLS 2011 年基线调查数据作为预测变量测量时点,2022 年随访数据作为结局变量判定时点,形成平均随访时长约 11 年的前瞻性队列设计,确保了预测变量测量在前、结局发生在后的时间顺序。CHARLS 研究已通过北京大学伦理委员会审查(IRB00001052-11015),所有研究对象已签署知情同意书。

1.2 研究对象 纳入标准:(1)2011 年基线调查时年龄 ≥ 60 岁;(2)完成认知功能评估和血液生化检测;(3)参与 2022 年随访并完成相关评估。排除标准:(1)基线时已诊断为痴呆;(2)基线时已符合 MCI 操作性诊断标准,即同时满足以下 3 项,①自评记忆力“较差”或“很差”,②综合认知功能得分低于同年龄同教育水平群体均值 1.5 个标准差,③日常生活能力正常;(3)关键变量存在缺失;(4)随访期间失访或死亡。经筛选,最终纳入 1 198 例符合条件的研究对象。采用随机数字表法将样本按 7:3 比例分为训练集($n=838$)和验证集($n=360$),训练集用于特征筛选和模型训练,验证集用于独立性能验证。

1.3 结局变量定义 本研究结局变量为 2022 年随访时是否发生 MCI。鉴于 CHARLS 为大规模社区调查,无法对所有参与者进行神经心理专科医师面对面评估,故参考美国国家衰老研究所与阿尔茨海默病协会(NIA-AA)的 MCI 诊断框架,结合问卷数据制订操作性诊断标准^[13-14]。同时满足以下 3 项标准者判定为 MCI:(1)主观认知主诉,受访者对自身记忆力的主观评价为“较差”或“很差”^[15];(2)客观认知下降,综合认知功能得分(0~21 分)低于同年龄、同教育水平群体均值 1.5 个标准差;(3)日常功能正常,在日常生活活动力量表(ADL)^[16]和工具性日常生活活动力量表(IADL)所有项目中均报告无困难或可独立完成^[17]。根据 MCI 的判断结果将研究对象分为 MCI 组和非 MCI 组。

1.4 预测变量 所有候选预测变量均取自 2011 年基线调查。结合数据完整性、基层可获得性及临床解释性,本研究预先确定 12 项候选预测变量,涵盖以下 4 个维度:(1)社会人口学特征,包括年龄、性别、受教育年限和户籍类型(城镇/农村);(2)健康状态,包括自评健康状况和抑郁症状评分(简版流行病学研究中心抑郁量表 CES-D 10);(3)基线认知功能,包括即时记忆得分(0~10 分)、延迟记忆得分(0~10 分)和综合认知功能得分(0~21 分);(4)血液指标,包括

MCV、血红蛋白和空腹血糖。其中,MCV 为本研究重点考察的潜在外周血生物标志物。

1.5 统计学处理

1.5.1 数据预处理与描述性统计 对训练集和验证集进行全面描述性统计分析。符合正态分布的计量资料以 $\bar{x} \pm s$ 表示,2 组间比较采用独立样本 t 检验;偏态分布的计量资料以 $M(P_{25}, P_{75})$ 表示,2 组间比较采用 Mann-Whitney U 检验;分类资料以频数、百分率表示,组间比较采用 χ^2 检验。所有检验均为双侧,检验水准 $\alpha=0.05$ 。连续型变量采用 Z -score 标准化,分类变量采用独热编码(有序分类变量除外)。为防止数据泄露,所有预处理参数仅在训练集上拟合,再应用于验证集。

1.5.2 特征筛选 采用 Boruta 算法在训练集的 12 项候选预测变量中进行特征筛选^[18]。Boruta 算法是一种基于随机森林的全相关特征选择方法,通过比较真实特征与随机置换生成的影子特征的重要性,判断各变量是否具有稳定的预测贡献。具体流程:首先对每个真实特征进行随机置换并生成影子特征;随后合并真实特征与影子特征,训练包含 500 棵树的随机森林并计算重要性 Z 分值;再以影子特征的重要性上界为参照,通过双侧统计检验将真实特征判定为“确认”“拒绝”或“待定”;最大迭代次数设为 100 次,显著性水平设为 0.01。该方法能够同时考虑非线性关系、特征交互和冗余信息,适用于本研究的多维预测变量筛选。

1.5.3 模型构建 基于筛选出的核心特征,构建多种机器学习模型进行比较: XGBoost^[19]、LightGBM^[20]、CatBoost 模型^[21],经典集成模型 AdaBoost,深度学习模型多层感知机(MLP)。另构建传统 Logistic 回归模型作为基线对照,分别采用 Lasso 回归和递归特征消除(RFE)进行特征筛选。

为避免信息泄露导致性能高估,采用嵌套交叉验证策略^[22]:外层 10 折交叉验证评估训练集平均性能,内层 5 折交叉验证进行超参数优化。超参数优化采用网格搜索,以最大化内层交叉验证的平均受试者工作特征曲线下面积(AUC)为目标。所有模型训练在 Python 3.8 软件完成,使用 scikit-learn 1.0、xgboost 1.6、lightgbm 3.3、catboost 1.0 等软件包。

1.5.4 模型评价指标 在独立验证集上采用多维度指标评价模型性能。判别能力以 AUC 为主要指标,并报告准确度、灵敏度和特异度。校准能力采用 Brier 分数评价^[23]。Brier 分数取值为 0~1,数值越小表示预测概率与实际结局越接近。

临床实用性采用决策曲线分析(DCA)评价^[24-25],计算净获益(NB)。通过比较模型曲线与“全部干预”和“全不干预”2 种极端策略,可判断模型在不同阈值概率下是否产生额外临床净获益,并据此识别具有实际决策价值的风险阈值区间^[24]。AUC 的两两比较采用 DeLong 检验^[26]。

1.5.5 可解释性分析 对最优模型采用 SHAP 方法

进行解释。SHAP 基于合作博弈论 Shapley 值理论,为每次预测分配各特征的贡献值^[27]。全局层面,通过平均绝对 SHAP 值量化特征整体重要性;个体层面,SHAP 值展示具体患者各特征的风险贡献。绘制 SHAP 摘要图进行可视化。此方法直观展示特征重要性、作用方向及与预测结果的关系,增强模型可信度。统计分析在 R 4.2.1 和 Python 3.8 软件完成,使用 pROC、rms、shap 0.41 等软件包,检验水准 $\alpha=0.05$ (双侧)。

2 结 果

2.1 研究对象基线特征 1 198 例研究对象中共 284 例判定为 MCI(23.7%),914 例为非 MCI(76.3%)。训练集中 MCI 组 199 例(23.7%),验证集中 MCI 组 85 例(23.6%),2 个集 MCI 发生率基本一致。

训练集内部,与非 MCI 组相比,MCI 组年龄更大、受教育年限更短,农村户籍和自评健康状况差的比例更高,抑郁症状评分更高,即时记忆、延迟记忆和

综合认知功能得分更低,MCV 及空腹血糖水平更高,血红蛋白水平更低,差异均有统计学意义(均 $P<0.05$);非 MCI 组与 MCI 组性别分布比较,差异无统计学意义($P=0.635$)。训练集与验证集 12 项候选预测变量的基线特征比较,差异均无统计学意义(均 $P>0.05$)。见表 1。

2.2 特征筛选结果 在训练集 12 项候选预测变量中,Boruta 算法经迭代筛选,最终确认 4 项变量具有稳定且显著的预测重要性,按重要性 Z 分值由高到低依次为延迟记忆得分($Z=15.8$)、综合认知功能得分($Z=14.2$)、即时记忆得分($Z=12.6$)和 MCV($Z=8.9$)。图 1 显示,4 项确认特征的重要性均超过影子特征的重要性上界;其余 8 项变量(年龄、性别、受教育年限、农村户籍、自评健康状况、抑郁症状评分、血红蛋白和空腹血糖)均未超过该阈值,因此被判定为拒绝特征。Lasso 回归特征筛选见图 2,最终选取 λ_{1se} 对应的 6 个非零系数特征。

表 1 训练集与验证集 12 项候选预测变量的基线特征比较[$\bar{x}\pm s$ 或 $n(\%)$ 或 $M(P_{25},P_{75})$]

项目	训练集($n=838$)				验证集($n=360$)				2 个集整体间比较 P 值
	MCI 组 ($n=199$)	非 MCI 组 ($n=639$)	$t/\chi^2/Z$	P	MCI 组 ($n=85$)	非 MCI 组 ($n=275$)	$t/\chi^2/Z$	P	
年龄(岁)	68.5±6.3	66.0±5.4	5.260	<0.001	68.2±6.0	66.2±5.7	2.761	0.006	0.412
男性	94(47.2)	314(49.1)	0.231	0.635	42(49.4)	131(47.6)	0.081	0.775	0.845
教育年限(年)	4(2,6)	7(4,9)	-6.831	<0.001	4(2,7)	7(4,9)	-4.122	<0.001	0.528
农村户籍	136(68.3)	376(58.8)	5.890	0.015	56(65.9)	164(59.6)	1.040	0.299	0.998
自评健康差	92(46.2)	181(28.3)	21.472	<0.001	37(43.5)	81(29.5)	5.781	0.016	0.763
抑郁症状评分(分)	10(6,15)	8(5,12)	-5.321	<0.001	10(6,14)	8(5,12)	-2.580	0.010	0.601
即时记忆得分(分)	4(3,5)	5(4,7)	-10.860	<0.001	4(3,5)	5(4,6)	-5.633	<0.001	0.402
延迟记忆得分(分)	2(1,3)	4(2,5)	-11.644	<0.001	2(1,3)	3(2,5)	-6.051	<0.001	0.518
综合认知功能得分(分)	10(8,13)	14(12,16)	-13.423	<0.001	10(8,13)	14(11,16)	-7.282	<0.001	0.633
MCV(fL)	92.9±5.4	90.0±4.6	7.382	<0.001	92.5±5.1	90.3±4.8	3.580	<0.001	0.327
血红蛋白(g/L)	132.3±15.4	138.8±14.6	-5.310	<0.001	133.0±14.8	138.2±15.1	-2.781	0.006	0.456
空腹血糖(mmol/L)	5.9±1.5	5.5±1.2	3.010	0.003	5.7±1.3	5.6±1.2	0.650	0.518	0.687

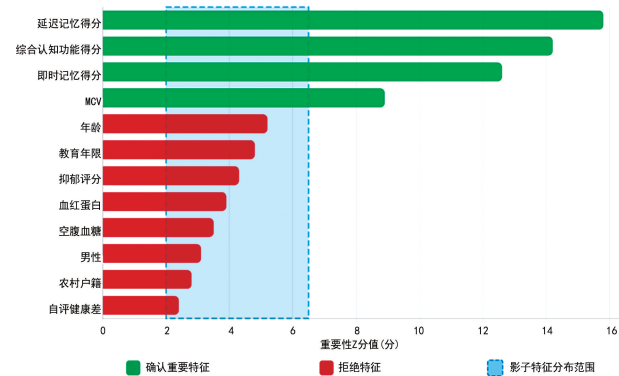


图 1 Boruta 算法特征选择重要性分布图

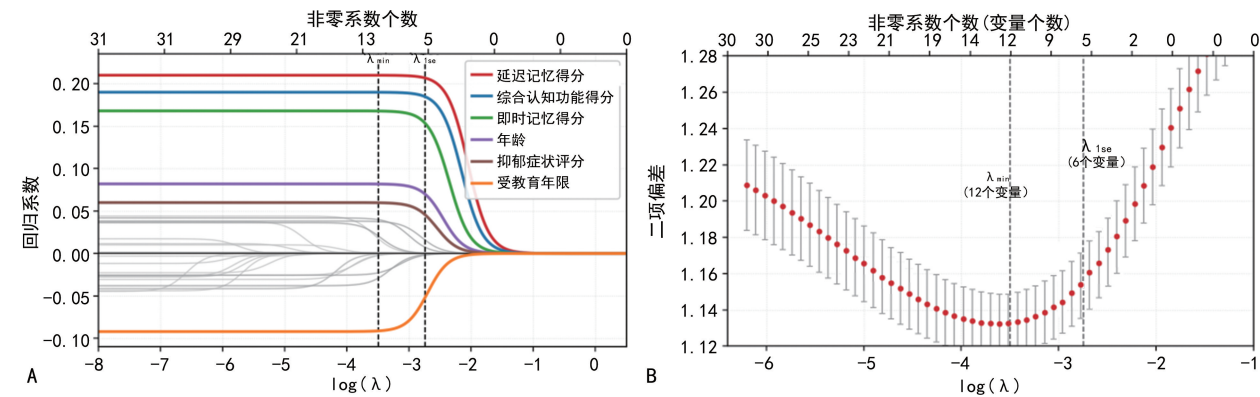
Boruta 算法、Lasso 回归和 RFE 3 种特征筛选方法的结果汇总见表 2。3 种方法均将综合认知功能得

分、即时记忆得分和延迟记忆得分识别为核心预测因子,表明基线认知储备对远期 MCI 风险具有稳定的预测价值。差异主要体现在辅助变量的选择:MCV 仅被 Boruta 算法纳入,而 Lasso 回归与 RFE 更倾向于保留年龄、受教育年限等传统人口学或健康状态变量。

2.3 模型性能比较 基于上述筛选结果分别构建模型:XGBoost、LightGBM、CatBoost、AdaBoost、MLP 模型均纳入 Boruta 算法筛选的 4 个特征(含 MCV),Lasso-Logistic 模型纳入 Lasso 回归筛选的 6 个特征(不含 MCV),RFE-Logistic 纳入 RFE 筛选的 7 个特征(不含 MCV)。各模型在验证集上的评估性能见表 3。各模型中,XGBoost 模型表现最优(AUC=0.854,

95%CI:0.810~0.898,Brier 分数 0.122),XGBoost 模型的 AUC 与 LightGBM,CatBoost 模型的 AUC 相近,差异均无统计学意义($P>0.05$);XGBoost 模型

的 AUC 大于 AdaBoost、MLP、Lasso-Logistic、RFE-Logistic 模型的 AUC,差异均有统计学意义($P<0.05$)。见表 4。



注:A 为系数路径图;B 为十折交叉验证误差曲线。

图 2 Lasso 算法回归特征筛选结果

表 2 不同特征筛选方法的结果比较

特征筛选方法	筛选特征数(个)	具体特征
Boruta 算法	4	综合认知功能得分、即时记忆得分、延迟记忆得分、MCV
Lasso 回归	6	综合认知功能得分、即时记忆得分、延迟记忆得分、年龄、受教育年限、抑郁症状评分
RFE	7	综合认知功能得分、即时记忆得分、延迟记忆得分、年龄、受教育年限、自评健康状况、居住地农村户籍

表 3 各模型在验证集上的性能比较 (n=360)

模型	AUC(95%CI)	准确度	灵敏度	特异度	Brier 分数
XGBoost	0.854(0.810~0.898)	0.819	0.831	0.803	0.122
LightGBM	0.847(0.802~0.892)	0.811	0.825	0.797	0.128
CatBoost	0.832(0.786~0.878)	0.797	0.808	0.785	0.135
AdaBoost	0.801(0.752~0.850)	0.769	0.775	0.761	0.156
MLP	0.795(0.745~0.845)	0.758	0.764	0.750	0.168
Lasso-Logistic	0.792(0.740~0.844)	0.761	0.747	0.772	0.152
RFE-Logistic	0.781(0.728~0.834)	0.750	0.739	0.758	0.164

注:Brier 分数越小表示校准度越好。

表 4 XGBoost(参照模型)与其他模型 AUC 比较的结果

比较组	XGBoost 的 AUC	对比模型的 AUC	Δ AUC	Δ AUC 的 95%CI	Z	P
XGBoost vs. LightGBM	0.854	0.847	0.007	-0.018~0.032	0.55	0.582
XGBoost vs. CatBoost	0.854	0.832	0.022	-0.011~0.055	1.32	0.187
XGBoost vs. AdaBoost	0.854	0.801	0.053	0.012~0.094	2.51	0.012
XGBoost vs. MLP	0.854	0.795	0.059	0.017~0.101	2.76	0.006
XGBoost vs. Lasso-Logistic	0.854	0.792	0.062	0.021~0.103	2.95	0.003
XGBoost vs. RFE-Logistic	0.854	0.781	0.073	0.030~0.116	3.32	<0.001

注: Δ AUC=XGBoost 的 AUC 减去对比模型的 AUC。

为进一步验证 MCV 的独立预测价值,本研究从模型比较、单变量效应及多变量校正 3 个层面进行了分析。在模型比较层面,剔除 MCV 后,XGBoost 模型 AUC 由 0.854(95%CI:0.810~0.898)降至 0.777(95%CI:0.729~0.824), Δ AUC=0.077(95%CI:0.034~0.120),差异有统计学意义($Z=3.52,P<$

0.001)。在单变量层面,仅纳入 MCV 的 Logistic 回归模型 AUC 为 0.638(95%CI:0.583~0.693,与随机水平 AUC=0.5 比较, $P<0.001$),OR(每升高 1 个标准差)为 1.67(95%CI:1.41~1.98, $P<0.001$);仅纳入 MCV 的 XGBoost 模型 AUC 为 0.645(95%CI:0.590~0.700,与随机水平比较, $P<0.001$),与

Logistic 回归模型结果基本一致,提示 MCV 本身即具有一定的独立预测能力。在多变量校正层面,纳入全部 4 个特征的 Logistic 回归模型中,MCV 仍具有独立预测效应,调整 OR(每升高 1 个标准差)为 1.52 (95%CI:1.26~1.83, $P<0.001$),表明在校正延迟记忆得分、综合认知功能得分和即时记忆得分后,MCV 对 MCI 风险的预测价值依然稳健。

2.4 最优模型的临床实用性评估 在验证集上对各模型进行 DCA(图 3)。结果显示,当阈值概率处于 15%~65% 时,XGBoost 模型的净获益曲线始终位于

其他所有模型及“全部干预”“全不干预”两种极端策略之上;例如在阈值概率为 30% 时,XGBoost 的净获益约为 0.18,而“全部干预”策略约为 0.08、“全不干预”策略为 0。LightGBM 与 CatBoost 模型的净获益曲线与 XGBoost 模型接近,在多数阈值下亦优于极端策略,但在高阈值区域(>60%)略低于 XGBoost 模型。Lasso-Logistic、RFE-Logistic 模型仅在阈值 20%~50% 优于极端策略,其余区间优势不明显;AdaBoost 与 MLP 模型在多数阈值下净获益接近或低于“全部干预”策略。

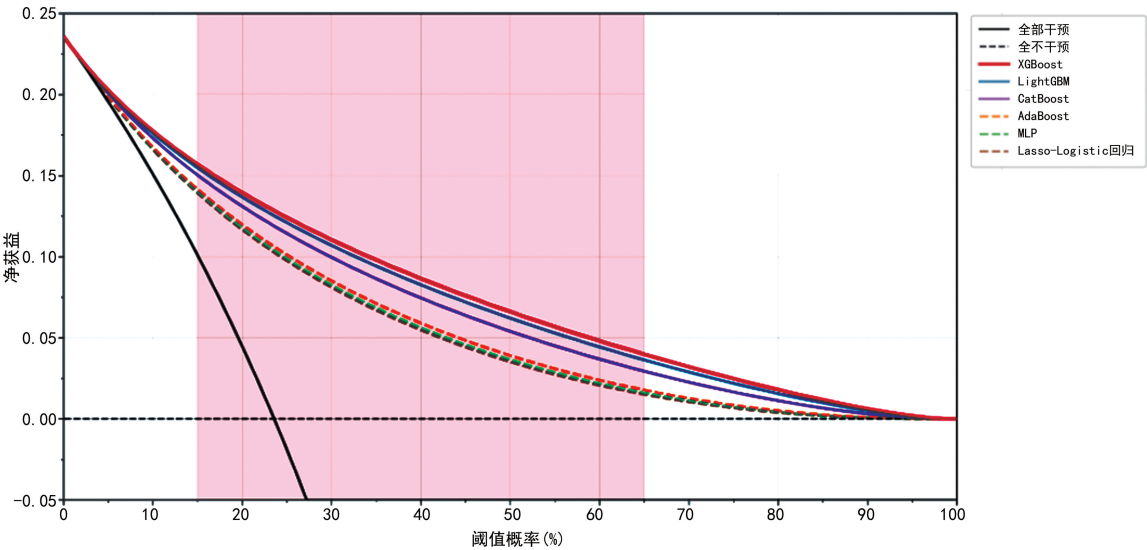
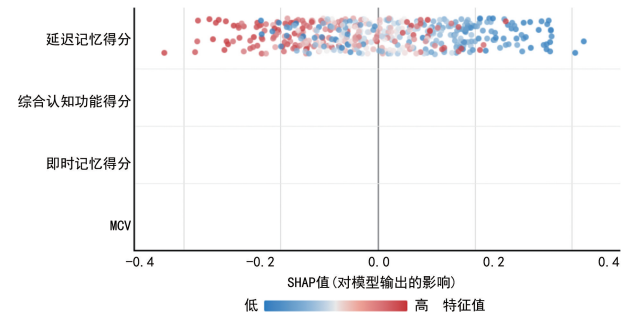


图 3 各模型的 DCA

2.5 模型可解释性分析 为揭示 XGBoost 模型的决策逻辑,采用 SHAP 方法对验证集进行解释。见图 4。结果显示,特征重要性排序与 Boruta 算法筛选结果一致:延迟记忆得分的平均绝对 SHAP 值最大,其次为综合认知功能得分、即时记忆得分和 MCV。从特征作用方向看,延迟记忆得分较低的样本主要分布在正 SHAP 值区域,提示延迟记忆表现越差,模型预测的 MCI 风险越高;综合认知功能得分和即时记忆得分呈现相同规律,即低分对应正向风险贡献、高分对应负向风险贡献。MCV 的 SHAP 分布显示,高 MCV 值主要位于正 SHAP 值区域,提示 MCV 升高与预测风险增加相关。



注:纵轴按平均绝对 SHAP 值从大到小列出 4 个特征,横轴表示 SHAP 值的分布,每个点代表一个样本在该特征上的 SHAP 值,点的颜色表示该特征的实际取值(红色为高值,蓝色为低值)。

图 4 XGBoost 模型的 SHAP 摘要图

3 讨 论

3.1 主要研究发现 本研究基于 CHARLS 纵向队列数据,构建并验证了一个仅需 4 项易获取指标的可解释机器学习模型,用于社区老年人 MCI 早期筛查。Boruta 算法从 12 项候选变量中筛选出延迟记忆得分、综合认知功能得分、即时记忆得分和 MCV 4 项核心预测特征,在精简变量数量的同时保持了良好的预测性能。XGBoost 模型在独立验证集上的 AUC 为 0.854,显著优于传统 Logistic 回归模型;DCA 显示其在 15%~65% 的阈值概率范围内具有持续净获益;SHAP 分析进一步揭示了各特征的作用方向,增强了模型的透明度与临床可信度^[28-29]。MCV 作为唯一入选的血液指标,其独立预测价值从模型剔除比较、单变量效应和多变量校正 3 个层面得到验证,为资源受限场景下的低成本风险评估提供了新的候选标志物。需要强调的是,3 项基线认知指标与随访时 MCI 操作性诊断中的认知测评虽然在构念上相关,但测量时间相隔约 11 年,且基线纳入对象均不符合 MCI 标准,因此主要反映认知储备对远期认知衰退的预测作用。

3.2 MCV 与认知功能的关系 MCV 与认知功能的关联及潜在机制值得深入探讨。本研究发现,基线 MCV 升高与 11 年后 MCI 发生风险显著相关;SHAP 分析进一步显示,在本研究预测模型中,较高 MCV 通

常对应正向风险贡献,提示 MCV 可能具有一定的长期认知风险识别价值。该结果与既往研究的发现一致。基于英国生物银行的大样本研究也报告 MCV 与认知表现呈负相关^[11,30]。Gamaldo 等^[10]基于巴尔的摩老龄化纵向研究(BLSA)对 827 名 50~96 岁参与者进行分析发现,较高 MCV 与较差的总体认知表现相关,并与总体认知、长延迟记忆及注意功能更快下降有关。这一结果从纵向认知变化的角度为本研究关于 MCV 与远期 MCI 风险的关联提供了间接支持,但 MCV 升高参与认知功能减退的具体生物学机制及其因果关系仍需进一步研究加以验证。MCV 影响认知功能的生物学机制目前仍以假说为主,主要包括以下方面:(1)微循环灌注障碍假说。红细胞体积增大导致变形能力下降,难以通过脑内直径仅 5~8 μm 的毛细血管,长期慢性缺血缺氧可损害海马和前额叶皮层等认知关键区域的神经元功能。(2)营养代谢异常假说。MCV 升高常与维生素 B₁₂ 或叶酸缺乏相关,其缺乏可导致同型半胱氨酸(Hcy)水平升高,Hcy 已被证实可通过氧化应激、血管内皮损伤和神经元凋亡等途径损害认知功能。(3)氧化应激与炎症假说。衰老过程中红细胞膜脂质过氧化可导致体积增大,氧化损伤的红细胞释放游离铁和活性氧,加剧神经炎症^[31]。(4)淀粉样蛋白毒性假说。近期研究发现 β 淀粉样蛋白可从脑组织释放入外周血,结合红细胞膜并改变其理化性质,提示 MCV 升高可能部分反映 AD 病理的早期外周表现。上述假说为 MCV 与认知下降的关联提供了理论支持,但因果链条尚未完全阐明,现有观察性证据无法排除反向因果或混杂作用。未来研究应在具备维生素 B₁₂、叶酸、Hcy 及脑影像数据的队列中进行中介效应分析和因果推断,并通过随机对照试验验证营养补充能否通过改善 MCV 延缓认知衰退。

3.3 模型的临床应用价值 本模型具有较好的临床应用潜力。首先,4 项指标均易于获得;即时记忆、延迟记忆和综合认知评估可在较短时间内完成,MCV 为血常规指标,适合基层医疗机构和社区筛查场景^[5]。其次,DCA 结果表明模型在较宽阈值范围内具有正向净获益,既可服务于强调灵敏度的早期发现,也可兼顾筛查效率与后续转诊成本。再次,SHAP 方法提供了全局和个体层面的透明解释^[26,28-29],有助于将风险预测转化为针对性干预建议,例如对 MCV 偏高者进一步评估维生素 B₁₂、叶酸及相关营养代谢指标,对延迟记忆得分偏低者加强认知训练和随访,对整体认知储备较低者鼓励参与社会活动和认知刺激训练。

在实际应用中,本模型可嵌入分级诊疗体系:基层机构完成初筛后,将高风险个体转诊至上级医院接受专科评估和诊断复核,形成“初筛—转诊—确诊—干预”的连续管理路径。该模式既能发挥基层机构覆盖面广、可及性强的优势,也有助于合理配置有限的

专科医疗资源。

3.4 研究的局限性 本研究存在以下局限性:第一,MCI 结局基于问卷和认知测评的操作性标准而非临床诊断,可能存在误分类。本研究将其定位为社区早期筛查工具,阳性者需转诊复核。未来应联合专科评估作为金标准,评估诊断效能并校正模型阈值。第二,部分预测变量与 MCI 诊断标准存在重叠,尽管测量时间相隔 11 年且基线对象均未达 MCI 标准,仍可能高估模型性能。未来可考虑以痴呆作为结局以消除循环论证疑虑。第三,MCV 分析存在方法学局限:仅使用基线单时点数据,未能捕捉随访期间的动态变化;未进行剂量-反应关系和分层分析。未来可采用联合模型或长短期记忆网络(LSTM)等方法捕捉生物标志物的时间轨迹信息。此外,随访期间的竞争风险(如死亡)可能导致风险估计偏差,后续研究可采用竞争风险模型校正。第四,候选变量有限,未纳入 APOE $\epsilon 4$ 基因型、维生素 B₁₂、叶酸、Hcy 等重要变量,可能存在残余混杂。未来应在变量更丰富的队列中重复验证。第五,模型仅完成内部验证,缺乏外部独立人群验证。本团队已拟定分层外部验证方案,计划在 CLHLS 等队列及医院门诊样本中进行多中心验证。

3.5 研究意义与展望 本研究为我国社区老年人 MCI 早期筛查提供了一个简便、经济、可解释的工具,具有重要的理论意义和实践价值。理论层面,本研究首次在大样本纵向队列中证实 MCV 作为独立预测因子对 MCI 的长期预测价值,为认知功能下降的外周血生物标志物研究提供了新的证据,拓展了 MCI 风险评估的指标谱。研究还整合了主观认知评价、客观认知测试和血液生化指标 3 个维度的信息,体现了多层次、多模态数据融合的优势,为未来更复杂模型的开发奠定了基础。实践层面,本模型仅需 4 项易得指标即可实现 AUC 为 0.85 的良好判别性能,适合在基层医疗机构大规模推广,有望填补我国 MCI 早筛工具匮乏的空白。

展望未来,本研究可从以下几个方向深化和拓展。第一,开展前瞻性临床验证研究,在独立样本中以神经心理专科医师诊断为金标准,量化本模型的灵敏度、特异度、阳性预测值和阴性预测值,并据此优化决策阈值,使筛查策略更符合不同临床场景的需求。第二,探索 MCV 与认知下降的因果机制,在具备营养代谢指标、炎症因子和脑影像数据的队列中进行中介效应分析,明确 MCV 通过何种路径影响认知功能,并开展干预试验验证补充维生素 B₁₂ 或叶酸能否通过改善 MCV 而延缓认知衰退。第三,构建动态预测模型,纳入多时点测量数据,利用联合模型或深度学习序列模型(如 LSTM)捕捉生物标志物的时间轨迹信息,提高预测精度。第四,扩展模型的多模态整合能力,在条件允许时纳入更多维度的数据,如基因型(APOE $\epsilon 4$)、脑影像标志物(海马体积、白质高信号)、数字化

行为数据(步态、睡眠模式)等,构建更全面的风险评估体系。第五,开发模型的数字化应用工具,如手机 APP 或基层医疗信息系统嵌入模块,实现自动化风险评分和智能化转诊建议,降低人工成本并提高筛查效率。通过上述努力,期待本研究成果能真正转化为可及、可用、可信的临床工具,为我国乃至全球老龄化社会的认知健康管理做出贡献。

利益冲突 所有作者均声明不存在利益冲突。

作者贡献 刘刚:研究构思与设计、数据整理与统计分析、模型构建、结果解释、论文撰写;余强:临床专业指导;杜建霖:临床专业指导;王凯娟:研究构思与设计、统计学与方法学指导、论文审阅与定稿并对全文负责。

参考文献

- [1] Mitchell A J, Shiri-Feshki M. Rate of progression of mild cognitive impairment to dementia: Meta-analysis of 41 robust inception cohort studies[J]. *Acta Psychiatr Scand*, 2009, 119(4): 252-265.
- [2] GBD 2019 Dementia Forecasting Collaborators. Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the global burden of disease study 2019[J]. *Lancet Public Health*, 2022, 7(2): e105-e125.
- [3] Jia L F, Du Y F, Chu L, et al. Prevalence, risk factors, and management of dementia and mild cognitive impairment in adults aged 60 years or older in China: a cross-sectional study[J]. *Lancet Public Health*, 2020, 5(12): e661-e671.
- [4] Campbell F C. Diet, dementia, and the hippocampus[J]. *Stem Cells Transl Med*, 2025, 14(5): e70001.
- [5] Lu Y, Liu C J, Wells Y, et al. Challenges in detecting and managing mild cognitive impairment in primary care: a focus group study in Shanghai, China[J]. *BMJ Open*, 2022, 12(9): e062240.
- [6] Arafa D A, Moustafa H E D, Ali-Eldin A M T, et al. Neuroimaging and machine learning for studying the pathways from mild cognitive impairment to Alzheimer's disease: a systematic review[J]. *BMC Neurol*, 2023, 23(1): 309.
- [7] Zhang D F, Bayrak C S, Zeng Q, et al. Prediction of mild cognitive impairment using blood multi-omics data[J]. *Front Genet*, 2025, 16: 1552063.
- [8] Fernández-Blázquez M A, Ruiz-Sánchez de León J M, Sanz-Blasco R, et al. XGBoost models based on non-imaging features for the prediction of mild cognitive impairment in older adults[J]. *Sci Rep*, 2025, 15(1): 29732.
- [9] Band S S, Yarahmadi A, Hsu C C, et al. Application of explainable artificial intelligence in medical health: a systematic review of interpretability methods[J]. *Inform Med Unlocked*, 2023, 40: 101302.
- [10] Gamaldo A A, Ferrucci L, Rifkind J, et al. Relationship between mean corpuscular volume and cognitive performance in older adults[J]. *J Am Geriatr Soc*, 2013, 61(1): 84-89.
- [11] Chen Y, Ma C N, Luo L, et al. The cross-sectional association between mean corpuscular volume and cognitive function in Chinese adults ≥ 45 years: evidence from CHARLS[J]. *PLoS One*, 2020, 15(12): e0243227.
- [12] Zhao Y, Hu Y, Smith J P, et al. Cohort profile: the China Health and Retirement Longitudinal Study (CHARLS) [J]. *Int J Epidemiol*, 2014, 43(1): 61-68.
- [13] Albert M S, DeKosky S T, Dickson D, et al. The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease[J]. *Alzheimers Dement*, 2011, 7(3): 270-279.
- [14] Silverberg N, Elliott C, Ryan L, et al. NIA commentary on the NIA-AA research framework: towards a biological definition of Alzheimer's disease[J]. *Alzheimers Dement*, 2018, 14(4): 576-578.
- [15] Parfenov V A, Zakharov V V, Kabaeva A R, et al. Subjective cognitive decline as a predictor of future cognitive decline: a systematic review [J]. *Dement Neuropsychol*, 2020, 14(3): 248-257.
- [16] Katz S, Ford A B, Moskowitz R W, et al. Studies of illness in the aged. The index of ADL: a standardized measure of biological and psychosocial function[J]. *JAMA*, 1963, 185(12): 914-919.
- [17] Lawton M P, Brody E M. Assessment of older people: self-maintaining and instrumental activities of daily living [J]. *Gerontologist*, 1969, 9(3): 179-186.
- [18] Kursa M B, Rudnicki W R. Feature selection with the Boruta package[J]. *J Stat Softw*, 2010, 36(11): 1-13.
- [19] Chen T Q, Guestrin C. XGBoost: a scalable tree boosting system[C]//Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Mining. New York: ACM, 2016: 785-794.
- [20] Hajihosseini M, Maghsoudi A, Ghezelbash R. A novel scheme for mapping of MVT-type Pb-Zn prospectivity: LightGBM, a highly efficient gradient boosting decision tree machine learning algorithm[J]. *Nat Resour Res*, 2023, 32(6): 2417-2438.
- [21] Zhang B, Xu C, Dai X G, et al. Research on mining land subsidence by intelligent hybrid model based on gradient boosting with categorical features support algorithm[J]. *J Environ Manage*, 2024, 354: 120309.
- [22] Vabalas A, Gowen E, Poliakoff E, et al. Machine learning algorithm validation with a limited sample size[J]. *PLoS One*, 2019, 14(11): e0224365.
- [23] Assel M, Sjöberg D D, Vickers A J. The Brier score does not evaluate the clinical utility of diagnostic tests or prediction models[J]. *Diagn Progn Res*, 2017, 1(1): 19.
- [24] Gonen M, Elkin E B, Cronin A M, et al. Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers [J]. *BMC Med Inform Decis Mak*, 2008, 8(1): 53.
- [25] Vickers A J, Elkin E B. Decision curve analysis: a novel method for evaluating prediction models[J]. *Med Decis Making*, 2006, 26(6): 565-574. (下转第 2186 页)

机器学习在结直肠癌相关风险预测中的研究进展*

马善源 综述, 彭 亮[△] 审校

上海交通大学医学院附属松江医院检验科, 上海 201600

摘要: 结直肠癌(CRC)是全球第 3 大常见恶性肿瘤及第 2 大癌症相关死因, 新发 CRC 患者中有 20% 在就诊时即有转移, 另外 25% 的患者会发生转移, 因此, CRC 的早期筛查与精准风险预测是降低其病死率的关键。手术和放化疗仍然是 CRC 治疗的主要手段, 但治疗效果往往与发病部位、肿瘤分期等息息相关。目前常用于 CRC 早期筛查、转移和治疗监测的手段有肠镜、直肠指检、粪便隐血、肿瘤标志物和影像学检查等, 但单一手段往往灵敏度、准确性不高。机器学习是一种通过算法和模型使计算机从多源数据中自动学习并进行预测或决策的技术, 属于人工智能的一个分支。机器学习风险预测模型正越来越多地应用于疾病的筛查、诊断、治疗及预后等各环节。该文就机器学习在 CRC 相关风险预测中的应用进行综述。

关键词: 结直肠癌; 机器学习; 风险预测; 筛查; 诊断

中图法分类号: R735.3; R319

文献标志码: A

文章编号: 1672-9455(2026)16-2181-06

Research progress of machine learning in risk prediction related to colorectal cancer*

Ma Shanyuan, Peng Liang[△]

Department of Laboratory Medicine, Songjiang Hospital Affiliated to Shanghai

Jiao Tong University School of Medicine, Shanghai 201600, China

Abstract: Colorectal cancer (CRC) is the third most common malignant tumor and the second leading cause of cancer-related death in the world, approximately 20% of newly diagnosed CRC patients present with metastasis at the time of diagnosis, and another 25% will develop metastasis subsequently. Therefore, early screening and precise risk prediction are crucial for reducing CRC mortality. Surgery, radiotherapy and chemotherapy remain the main treatment methods for CRC, but the treatment effects are often closely related to the site of onset, tumor stage, etc. Currently, colonoscopy, digital rectal examination, fecal occult blood, tumor biomarkers and imaging examination are commonly used for early screening, metastasis detection and treatment monitoring of CRC, however, a single method often has low sensitivity and accuracy. Machine learning, a branch of artificial intelligence, is a technology that enables computers to automatically learn from multi-source data through algorithms and models and make predictions or decisions. Machine learning-based risk prediction models are increasingly applied in various aspects of disease screening, diagnosis, treatment and prognosis. This article reviews the application of machine learning in CRC-related risk prediction.

Key words: colorectal cancer; machine learning; risk prediction; screening; diagnosis

国际癌症研究中心(IARC)2024 年发布的癌症负担报告显示, 2022 年全球新发结直肠癌(CRC)190 万例, 占全球所有新发肿瘤的 9.6%, CRC 相关死亡 90 万例, 占所有癌症相关死亡的 9.3%, 是全球第 3 大常见恶性肿瘤及第 2 大癌症相关死因^[1]。我国国家癌症中心 2024 年发布的最新数据显示, 2022 年我国新发恶性肿瘤中 CRC 发病率居第 2 位、病死率居第 4 位^[2], 早期筛查与精准风险预测是降低其病死率的关键。目前 CRC 筛查手段包括结肠镜、粪便免疫化学

检测、CT 结肠成像及分子标志物检测等, 但存在侵入性强、依从性低、假阳性率高、经济负担重等局限。机器学习按学习范式可分为监督学习、无监督学习和强化学习等不同类型。监督学习是一种基于标注数据训练模型的学习方式, 通过学习输入与标签之间的映射关系, 实现对新数据的预测。训练目标是预测输出与真实标签之间的误差最小化。常见算法包括线性回归、正则化、Logistic 回归、支持向量机、朴素贝叶斯、随机森林、神经网络、K 邻近(KNN)。无监督学习

* 基金项目: 上海市卫生健康系统重点学科(2024ZDXK0065); 上海市松江区科技攻关项目(2023SJKJGG058)。

[△] 通信作者, E-mail: jysjyyjyk@126.com。

中文引用格式: 马善源, 彭亮. 机器学习在结直肠癌相关风险预测中的研究进展[J]. 检验医学与临床, 2026, 23(16): 2181-2186.

英文引用格式: Ma S Y, Peng L. Research progress of machine learning in risk prediction related to colorectal cancer[J]. Lab Med Clin, 2026, 23(16): 2181-2186.