

本文引用: 林琪雄, 张柳, 李霞, 等. 面向基层医疗场景的机器学习模型在冠心病二级预防中的风险分层研究[J]. 中国动脉硬化杂志, 2026, 34(8): 743-751. DOI: 10.20039/j.cnki.1007-3949.2026.08.005.

[文章编号] 1007-3949(2026)34-08-0743-09

· 临床研究 ·

面向基层医疗场景的机器学习模型在冠心病二级预防中的风险分层研究

林琪雄, 张柳, 李霞, 史树银

新疆医科大学第五附属医院全科, 新疆乌鲁木齐市 830000

[摘要] **[目的]** 开发并验证一种基于易获取临床指标的轻量化机器学习模型, 以实现冠心病(CHD)的精准风险预测。**[方法]** 回顾性纳入 1 376 例经冠状动脉造影检查的患者数据。通过单因素 Logistic 回归(LR)、Boruta 算法和 LASSO 回归筛选出 13 个核心预测变量。利用训练集($n=963$)构建了包括 Logistic 回归、支持向量机(SVM)、随机森林(RF)等 7 种机器学习模型, 并在测试集($n=413$)上通过受试者工作特征曲线下面积(AUC)、召回率、精确率、F1 值及 Brier 评分等指标评估其性能。采用 SHAP 方法对最优模型进行可解释性分析。**[结果]** 单因素 Logistic 回归显示, 男性($OR=2.445, 95\% CI: 1.870 \sim 3.202$)、高血压($OR=1.697, 95\% CI: 1.305 \sim 2.210$)和糖尿病($OR=1.932, 95\% CI: 1.413 \sim 2.665$)与 CHD 风险升高相关, 高密度脂蛋白与 CHD 风险降低相关($OR=0.351, 95\% CI: 0.221 \sim 0.551$; 均 $P<0.001$)。测试集中, SVM 模型的 AUC 为 0.775 ($95\% CI: 0.730 \sim 0.821$), 召回率为 0.790, F1 值为 0.784, Brier 评分为 0.183 7 ($95\% CI: 0.168 3 \sim 0.202 0$)。DeLong 检验显示, SVM 与 Logistic 回归、RF、极端梯度提升(XGBoost)及轻量梯度提升机(LGBM)模型的 AUC 差异无统计学意义(均 $P>0.05$)。决策曲线分析(DCA)显示, 在预设的 10%~80% 阈值概率范围内, 依据 SVM 预测结果决定是否实施进一步检查或转诊的决策策略, 其净获益高于“全部实施进一步检查或转诊”和“全部不实施进一步检查或转诊”两种默认参照策略。SHAP 分析显示, 男性、高血压、糖尿病及炎症指标升高对模型预测呈正向影响, 而高密度脂蛋白、左心室射血分数、白蛋白水平升高呈负向影响。**[结论]** 本研究依托基层易得指标, 初步构建适配基层医疗场景、用于 CHD 二级预防风险分层的轻量化 SVM 风险预测模型, 但后续仍需多中心外部数据验证方可推广应用。

[关键词] 冠心病; 基层医疗; 机器学习; 二级预防

[中图分类号] R541

[文献标识码] A

Research on risk stratification of machine learning models in secondary prevention of coronary heart disease for primary care settings

Lin Qixiong, Zhang Liu, Li Xia, Shi Shuyin

Department of General Practice, the Fifth Affiliated Hospital of Xinjiang Medical University, Urumqi, Xinjiang 830000, China

[ABSTRACT] **Aim** To develop and validate a lightweight machine learning model based on readily available clinical indicators to achieve accurate coronary heart disease (CHD) risk prediction. **Methods** A total of 1 376 patients who underwent coronary angiography were retrospectively enrolled. Thirteen core predictive variables were identified through univariate Logistic regression, Boruta algorithm, and LASSO regression. Using the training set ($n=963$), seven machine-learning models, including Logistic regression, support vector machine (SVM), and random forest (RF), were constructed. Their performance was evaluated on the test set ($n=413$) using metrics such as the area under the curve (AUC), recall, precision, F1-score, and Brier score. The SHAP method was employed to analyze the interpretability of the optimal model. **Results** Univariate Logistic regression analysis identified male sex ($OR=2.445, 95\% CI: 1.870 \sim$

[收稿日期] 2026-03-30

[修回日期] 2026-07-30

[基金项目] 国家自然科学基金项目(81960073);新疆维吾尔自治区自然科学基金项目(2023D01C56);“天山英才”医药卫生高层次人才培养计划(TSYC202301B015)

[作者简介] 林琪雄, 硕士研究生, 研究方向为冠心病的基础与临床, E-mail: 18819797625@163.com。通信作者史树银, 主任医师, 硕士研究生导师, 研究方向为心血管内科、全科医学科、医院管理, E-mail: shy@xjmu.edu.com。

3.202), hypertension ($OR=1.697$, 95% CI : 1.305 ~ 2.210), and diabetes mellitus ($OR=1.932$, 95% CI : 1.413 ~ 2.665) as significant correlates of elevated CHD risk, whereas high density lipoprotein (HDL) was associated with a reduced likelihood of CHD ($OR=0.351$, 95% CI : 0.221 ~ 0.551; all $P<0.001$). In the independent test cohort, the SVM model achieved an AUC of 0.775 (95% CI : 0.730 ~ 0.821), with a recall of 0.790, an F1-score of 0.784, and a Brier score of 0.183 7 (95% CI : 0.168 3 ~ 0.202 0). DeLong's test revealed no statistically significant differences in AUC between the SVM model and the comparative models, including Logistic regression, RF, extreme gradient boosting (XGBoost), and light gradient boosting machine (LGBM) (all $P>0.05$). Decision curve analysis (DCA) further demonstrated that across the predefined threshold probability range of 10% to 80%, the decision strategy of guiding further diagnostic workup or referral based on SVM-derived predictions yielded superior net clinical benefit compared with the two default reference strategies: "performing universal further examinations or referrals for all participants" and "withholding further examinations or referrals from all participants". Shapley additive explanations (SHAP) analysis revealed that male sex, hypertension, diabetes mellitus, and elevated inflammatory biomarkers exerted positive contributions to the model's CHD risk prediction, while higher levels of HDL, left ventricular ejection fraction (LVEF), and albumin were associated with negative predictive effects.

Conclusion Based on routine indicators accessible at primary medical institutions, this study preliminarily constructed a lightweight SVM risk prediction model suitable for primary care settings to conduct risk stratification in the secondary prevention of CHD. The model exhibited acceptable discrimination performance and interpretability within this single-center cohort, which can provide a reference framework for early identification and risk stratification of CHD in primary care. Further multicenter external data validation is required prior to its clinical promotion and application.

[KEY WORDS] coronary heart disease; primary care; machine learning; secondary prevention

冠心病(coronary heart disease, CHD)是一种严重危害人类健康的常见病^[1]。随着国民生活水平提高,不健康生活方式较为普遍,CHD的危险因素不断加剧,其发病率和死亡率持续攀升^[2]。据《中国心血管健康与疾病报告2024》要点解读推算,我国CHD现患人数已高达1 700万,且仍在增长^[3],其疾病负担远超大多数其他疾病。当前,我国推行的国家三级诊疗制度旨在通过分级诊疗优化医疗资源配置、提升基层服务能力。然而,CHD的传统诊疗模式与该制度的推进需求存在显著矛盾:临床诊疗高度依赖传统影像学检查与有创介入治疗。前者对亚临床病变的检出敏感性有限,易导致早期筛查漏诊;后者虽能改善血流动力学,却伴随血管损伤、造影剂肾病等侵入性并发症风险。更关键的是,这两类技术存在明显的实施壁垒——高端影像设备,如计算机断层扫描(computed tomography, CT)、磁共振成像(magnetic resonance imaging, MRI)及介入器械购置成本高昂,基层医疗机构难以配备;其操作亦需专业资质认证,进一步限制了在医疗资源匮乏地区的普及。这与三级诊疗制度中“基层首诊、双向转诊”的核心要求相悖^[4],也导致了适用于基层场景、标准化且易于推广的CHD诊断标准的长期缺失。

在此背景下,现有国外风险评分工具应用于国内基层人群时区分效能有限,难以兼顾精准评估与

基层普及需求。为了应对这一挑战,人工智能(artificial intelligence, AI)技术所引领的医学领域变革提供了新的可能^[5]。机器学习作为AI的重要分支,能够通过利用大量医疗数据进行训练与优化,在临床疾病预测领域展现出巨大潜力^[6]。其独特优势在于处理高维复杂数据,并能自动分析包括自由文本、图像、波形等在内的多模态医疗信息^[7]。特别是其强大的预测与趋势识别能力,使得从海量个体患者数据中挖掘深层价值成为可能^[8]。

本研究旨在采用7种机器学习方法,构建CHD识别模型,确定最佳模型并分析影响CHD的主要因素,从而为临床早期干预提供科学依据。通过打破计算机科学与临床医学的学科壁垒,使患者能够从数字化工具中获益,实现更精确、规范、便捷的疾病诊断,有望推动疾病的早发现、早诊断和早治疗。

1 资料和方法

1.1 研究对象

本研究为回顾性临床研究分析。连续纳入2021年4月—2024年6月期间因胸痛症状就诊于新疆医科大学第五附属医院,均完善冠状动脉造影检查,且年龄 ≥ 18 岁的患者。排除存在患有心理疾病及精神分裂症等不能配合的患者、信息不全者及资料不全者。纳入研究的患者,经新疆医科大学第五附

属医院伦理委员会批准(No. XYDWFYLSk-2025-280)。

1.2 资料收集

严格按照课题组成员自行设计的冠心病电子病例报告表(coronary heart disease electronic case reported form, CHD-eCRF)收集患者住院期间的电子病例信息。内容包括以下 6 项。(1)一般资料:年龄、性别、血压、心率、呼吸频率、体重指数(body mass index, BMI)、吸烟史、饮酒史。(2)既往史:高血压病、糖尿病。(3)CHD 家族史。(4)实验室指标:血常规(白细胞计数、中性粒细胞计数等),血清电解质(血钠、血钾等),肾脏功能(估算的肾小球滤过率、血尿素氮、血肌酐、血尿酸等),胆红素三项(总胆红素、直接胆红素、间接胆红素),血脂(甘油三酯、总胆固醇等),血清白蛋白。(5)影像学检查:超声心动图中的射血分数。(6)冠状动脉造影结果:冠状动脉病变部位、狭窄程度。参照《稳定性冠心病基层诊疗指南(2020 年)》^[9], CHD 确诊需要冠状动脉造影证实心外膜下左主干、三大分支至少有一支或其主要分支直径狭窄范围 $\geq 50\%$,由 2 名以上心血管内科冠状动脉介入专科医师共同判断。

本研究相关变量统一标准化定义: BMI = 体重(kg)/身高(m)²;吸烟史定义为既往/目前每日吸烟 ≥ 1 支且持续 ≥ 1 年;饮酒史定义为每周摄入酒精 ≥ 25 g 且持续 ≥ 1 年;CHD 家族史指一级亲属男性 < 55 岁、女性 < 65 岁确诊 CHD。所有实验室指标均由本院检验科全自动生物化学分析仪、血细胞分析仪统一检测,执行室内质控与室间质评标准。

1.3 变量选择

本课题组 CHD-eCRF 收集患者住院电子病历信息,共收集 50 个变量。变量筛选遵循三步流程:首先,对所有潜在预测变量进行单因素 Logistic 回归分析,筛选出具有统计学意义的变量($P < 0.05$)进入后续分析;其次,采用 LASSO 回归对上述变量进行特征降维,通过十折交叉验证确定最优惩罚系数(λ_{1se} :在最小误差一个标准误差范围内的最大 λ),保留系数非零的变量构成最终预测变量集合;最后,应用 Boruta 算法对 LASSO 筛选出的变量进行重要性验证,确保所有纳入最终模型的变量均被确认为“确认”特征。

1.4 基线资料分析

分类变量以例数(百分比)表示,组间比较采用 χ^2 检验;符合正态分布的连续变量以 $\bar{x} \pm s$ 表示,组间比较采用独立样本 t 检验;不符合正态分布的连续变量以中位数(四分位数)表示,组间比较采用

Mann-Whitney U 检验。对于 2×2 列联表采用 Yates 连续性校正。 $P < 0.05$ 认为差异具有统计学意义。

1.5 机器学习算法与建模

符合纳排标准的 1 376 名患者,按照 7 : 3 随机划分为模型训练集($n = 963$)和测试集($n = 413$)。LASSO 回归借助十折交叉验证确定最优惩罚系数, Boruta 算法完成变量重要性显著性校验,单因素 Logistic 回归以 $P < 0.05$ 为阈值筛选候选变量。将筛选出的指标纳入多因素 Logistic 回归(Logistic regression, LR)、随机森林(random forest, RF)、轻量梯度提升机(light gradient boosting machine, LGBM)、支持向量机(support vector machine, SVM)、极端梯度提升(extreme gradient boosting, XGBoost)、K 近邻算法(k-nearest neighbors, KNN)和决策树(decision tree, DT) 7 种机器学习算法进行 CHD 预测模型构建^[10-13]。模型分类性能采用灵敏度(召回率, recall)、特异度(specificity)、阳性预测值(精确率, precision)、阴性预测值、准确率(accuracy, ACC)和 F1 值进行评价;模型判别能力采用受试者工作特征曲线下面积(area under the curve, AUC)及其 95% CI 评价,校准性能采用校准曲线和 Brier 评分评价,临床应用价值采用决策曲线分析(decision curve analysis, DCA)评价。不同模型间的 AUC 差异采用 DeLong 检验进行两两比较, $P < 0.05$ 判定模型判别性能存在统计学差异。采用特征重要性评估和沙普利可加性特征解释方法(Shapley additive explanation, SHAP)分析最优模型,识别 CHD 风险因素。本研究全部统计分析与建模操作基于 R4.2.1 软件完成。图 1 为研究整体框架流程图。

2 结 果

2.1 基本临床资料

单因素 Logistic 回归初步筛选出 18 个变量(图 2), Boruta 算法(图 3)、LASSO 回归分析(图 4)得到最终纳入机器学习模型的 13 个变量:中性粒细胞计数、白细胞计数、年龄、性别、高血压、糖尿病、心率、吸烟史、饮酒史、单核细胞计数、高密度脂蛋白(high density lipoprotein, HDL)、白蛋白和左心室射血分数(left ventricular ejection fraction, LVEF)。

本研究全面收集 1 376 名住院患者上述 13 个变量的相关信息。训练集与测试集的各项基线特征见表 1。两组在所有基线特征上均无统计学差异($P > 0.05$),证实了划分的随机性。

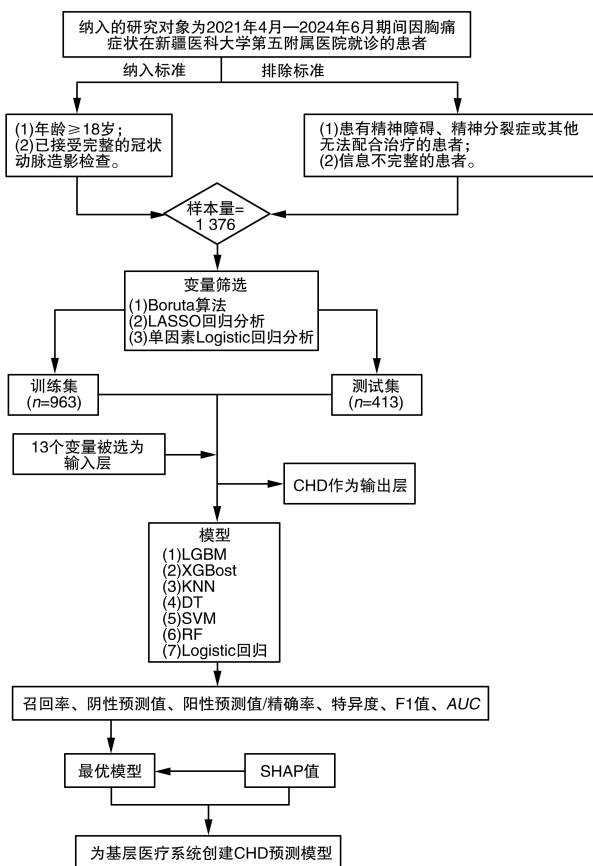


图1. 研究整体框架流程图

Figure 1. Flowchart of the overall research framework

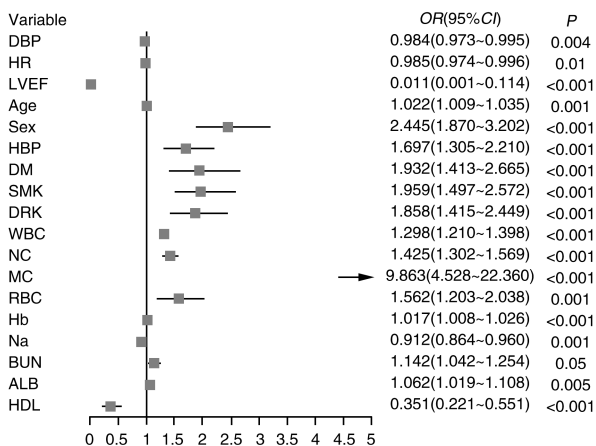


图2. 基于单因素 Logistic 回归筛选与结局事件相关的变量

DBP:舒张压(diastolic blood pressure);HR:心率(heart rate);HBP:高血压(high blood pressure);DM:糖尿病(diabetes mellitus);SMK:吸烟史(smoking);DRK:饮酒史(drinking);WBC:白细胞计数(white blood cell count);NC:中性粒细胞计数(neutrophil count);MC:单核细胞计数(monocyte count);RBC:红细胞计数(red blood cell count);Hb:血红蛋白(hemoglobin);Na:血钠;BUN:血尿素氮(blood urea nitrogen);ALB:白蛋白(albumin)。

Figure 2. Screening of variables associated with outcome events based on univariate Logistic regression

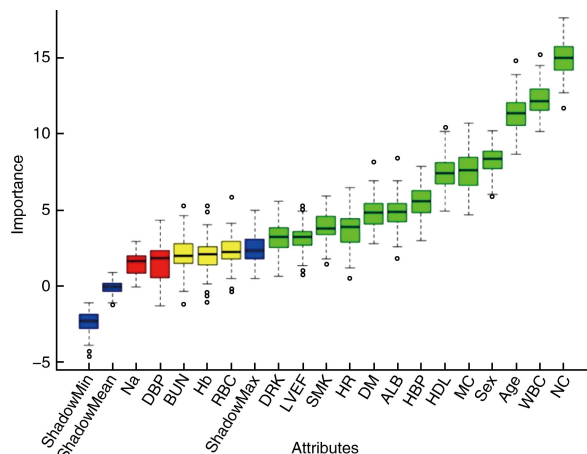


图3. 通过 Boruta 算法筛选与结局事件相关的变量

蓝色图显示了最小、平均和最大阴影分数。变量中,显示为绿色的变量代表与结局事件强相关,显示为黄色的变量的重要性未与随机基准形成稳定显著的差异,需结合领域知识进一步判断,显示为红色的变量代表与结局事件不相关。

Figure 3. Screening of variables associated with outcome events using the Boruta algorithm

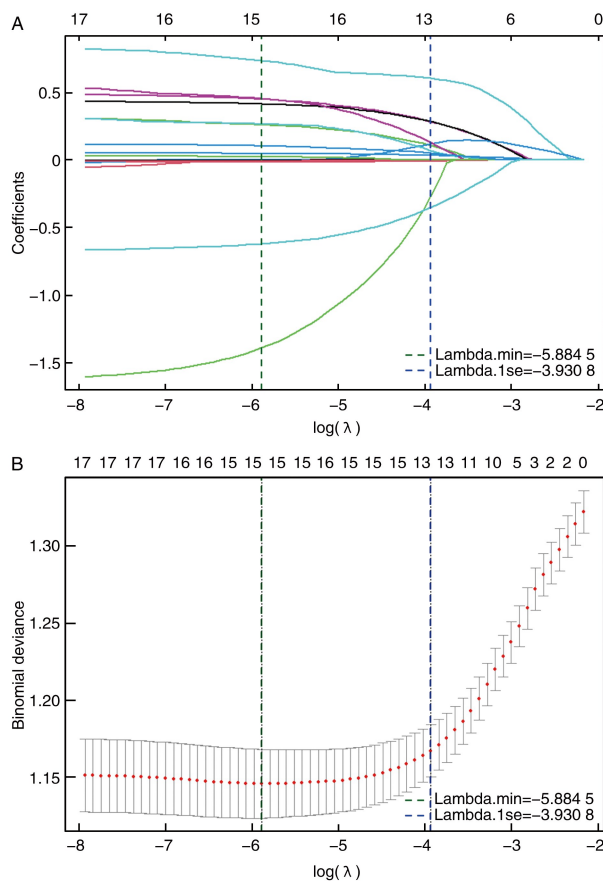


图4. 通过 LASSO 回归筛选与结局事件相关的变量

A 为 LASSO 回归中 L1 范数与不同系数之间的相关性。L1 范数是 LASSO 回归的正则化项。B 为 λ 与二项偏差(binomial deviance)之间的相关性。图中存在两条虚线。左侧虚线表示最小均方误差,右侧虚线表示与最小均方误差相差一个标准误差的位置。

Figure 4. Screening of variables associated with outcome events through LASSO regression

表 1. 研究人群的总现状况及训练集与测试集的基线特征比较
Table 1. Overall status of the study population and comparison of baseline characteristics comparison between the training set and the test set

变量	总体 (n=1 376)	测试集 (n=413)	训练集 (n=963)	P
男性/[例(%)]	824(59.9)	244(59.1)	580(60.2)	0.735
年龄/岁	59.24±10.76	59.11±10.56	59.29±10.85	0.777
吸烟/[例(%)]	614(44.6)	191(46.2)	423(43.9)	0.462
饮酒/[例(%)]	568(41.3)	179(43.3)	389(40.4)	0.338
高血压/[例(%)]	759(55.2)	232(56.2)	527(54.7)	0.663
糖尿病/[例(%)]	357(25.9)	106(25.7)	251(26.1)	0.93
心率/(次/min)	74(68,80)	73(68,80)	74(68,81)	0.224
LVEF/%	0.64(0.61,0.67)	0.65(0.61,0.67)	0.64(0.61,0.67)	0.052
白细胞计数/(×10 ⁹ L ⁻¹)	6.39(5.35,7.84)	6.37(5.29,7.70)	6.39(5.36,7.87)	0.408
中性粒细胞计数/(×10 ⁹ L ⁻¹)	3.64(2.84,4.79)	3.54(2.81,4.59)	3.69(2.86,4.84)	0.104
单核细胞计数/(×10 ⁹ L ⁻¹)	0.49(0.40,0.61)	0.49(0.39,0.60)	0.49(0.40,0.62)	0.219
白蛋白/(g/L)	41.25(39.20,43.40)	41.4(39.20,43.80)	41.20(39.20,43.25)	0.272
HDL/(mmol/L)	1.24(1.06,1.44)	1.26(1.04,1.46)	1.23(1.07,1.42)	0.427

2.2 预测模型构建

机器学习模型构建结果见表 2。模型性能采用 AUC 值评估,结果见图 5。SVM 模型(AUC=0.775)与 Logistic 回归模型(AUC=0.770)均处于较高水平。为进一步比较其性能,采用 DeLong 检验对全部模型进行了两两比较。结果表明,Logistic 回归、RF、XGBoost、SVM 和 LGBM 之间模型性能无统计学差异(P>0.05,图 6),均属于性能稳健的模型集合,在实际应用中具有较低风险。

在具体分类指标上,SVM 模型的召回率、阳性预测值和 F1 值分别为 0.790、0.778 和 0.784,整体表现相对均衡,但上述指标均非 7 种模型中的最高值。SVM 的 AUC 数值最高,且 Brier 评分数值最低;综合比较各模型的判别能力、校准性能及分类性能,SVM 整体表现较为均衡,因此选定 SVM 作为后

续解释性分析的模型。

表 2. 冠心病机器学习预测模型评价
Table 2. Evaluation of machine learning prediction models for coronary heart disease

评价指标	XGBoost	SVM	RF	Logistic 回归	LGBM	KNN	DT
F1 值	0.739	0.784	0.785	0.732	0.76	0.705	0.789
召回率/灵敏度	0.679	0.790	0.821	0.660	0.756	0.706	0.844
阳性预测值/精确率	0.809	0.778	0.752	0.820	0.764	0.703	0.742
阴性预测值	0.565	0.626	0.630	0.560	0.584	0.487	0.643
特异度	0.722	0.609	0.530	0.748	0.596	0.483	0.490

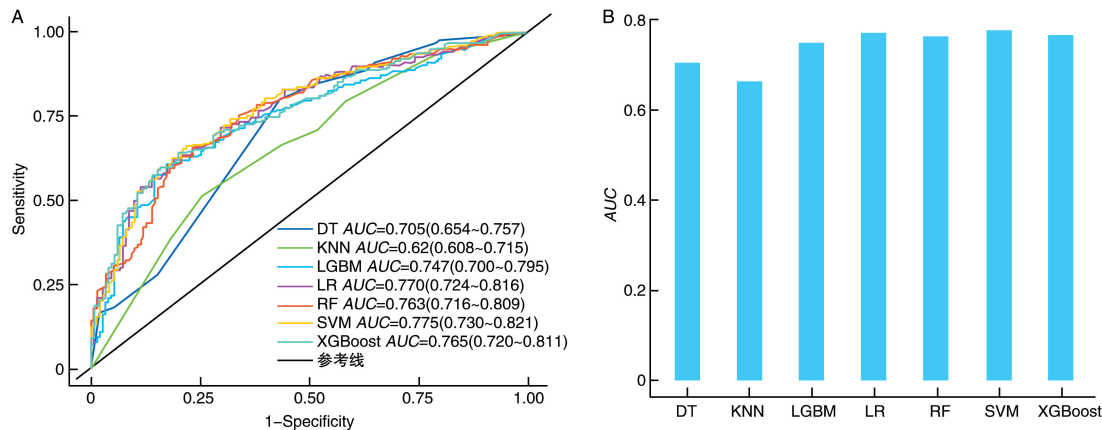


图 5. 不同机器学习模型预测性能的受试者工作特征曲线下面积比较
A 为 7 种模型的 ROC 曲线;B 为 7 种模型的 AUC 值柱状图。

Figure 5. Comparison of areas under the curve for predictive performance of different machine-learning models

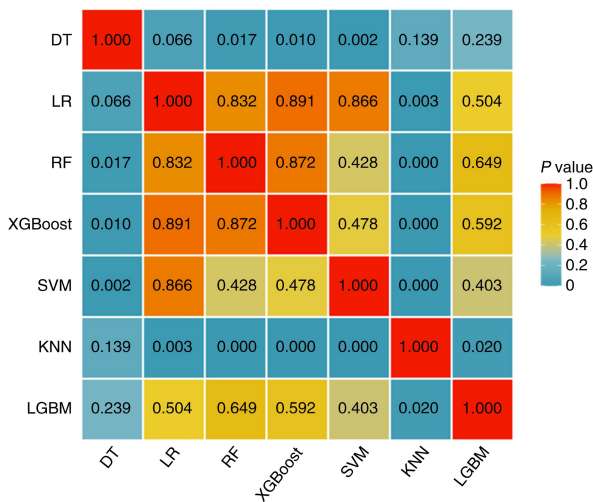


图 6. 采用 DeLong 检验对 7 种机器学习模型的 AUC 进行两两比较的 P 值矩阵

对角线为模型自身比较($P=1$)。矩阵数值反映了模型性能的差异程度, $P<0.05$ 表明两模型性能存在统计学差异。

Figure 6. P-value matrix for pairwise comparisons of AUC among seven machine-learning models using the DeLong test

2.3 特征重要性

鉴于本研究最优模型为 SVM,其具备非线性特征,传统列线图可能无法准确反映变量间的复杂交互作用,故采用 SHAP 可解释性分析替代列线图进行模型展示。结果表明,在 13 个特征的重要性排序中,中性粒细胞计数对模型影响最大。男性、白细胞计数及中性粒细胞计数升高、高血压、糖尿病、长期饮酒史属于危险因素,LVEF、血清 HDL 和白蛋白升高是保护因素,心率与 CHD 患病风险无简单线性关系,年龄与 CHD 患病风险关系复杂,受其他因素交互影响(图 7)。

2.4 模型校准性能的评估与比较

7 种模型的测试集校准曲线结果见图 8,采用 Brier 评分(表 3)衡量概率预测准确性,其取值范围在 0~1 之间,分值越低表示模型的预测概率与实际结果越接近,校准性能越好。其中,SVM 模型的 Brier 评分最低(0.183 7,95% CI:0.168 3~0.202 0),表明其校准性能最佳,预测结果与实际概率更为接近,能够更准确地识别 CHD 患者。相比之下,KNN 模型的 Brier 评分最高(0.278 6),说明其校准性能较差。此外,采用 DCA 曲线对各模型在不同风险阈值下的临床净获益进行评估,结果显示在 10%~80% 的临床决策阈值区间内,SVM 模型的净获益曲线均高于“全部实施进一步检查或转诊”和“全部不实施进一步检查或转诊”两条极端策略线,表明该模

型在基层 CHD 筛查中具有实际临床应用价值(图 9)。

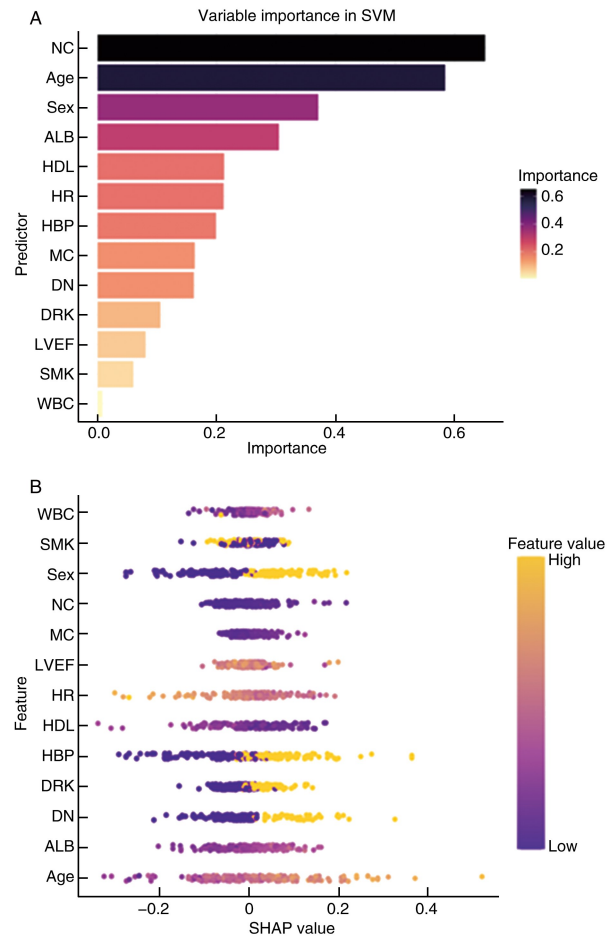


图 7. SVM 模型的预测因子解析

A 为变量重要性排序;B 为特征影响方向分析。

颜色(黄色/蓝色)和 SHAP 值的绝对值共同指明了各变量作为危险因素或保护因素的性质。

Figure 7. Analysis of predictors in the SVM model

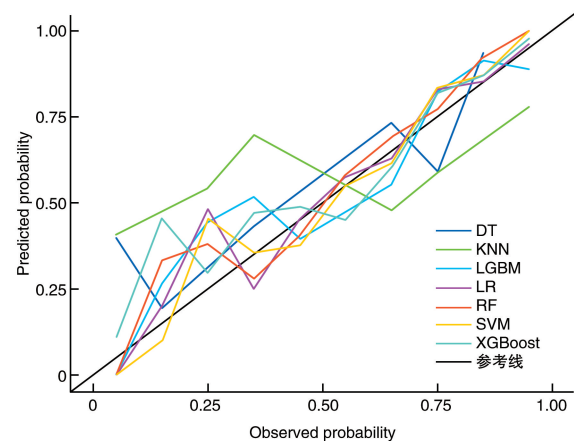


图 8. 7 种机器学习模型的测试集校准曲线对比

Figure 8. Comparison of calibration curves from the test set of seven machine-learning models

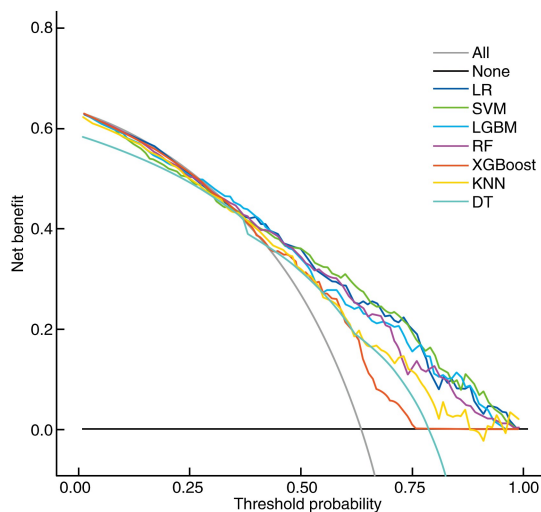


图 9. 7 种机器学习模型的 DCA 曲线
Figure 9. DCA cures of seven machine-learning models

表 3. 各测试集模型 Brier 评分
Table 3. Brier scores of models for each test set

模型	Brier 评分(95% CI)
XGBoost	0.188 4(0.170 7~0.207 6)
SVM	0.183 7(0.168 3~0.202 0)
RF	0.189 7(0.174 1~0.206 0)
Logistic 回归	0.185 7(0.168 3~0.203 6)
LGBM	0.195 5(0.178 7~0.215 6)
KNN	0.278 6(0.243 2~0.312 1)
DT	0.198 6(0.180 0~0.218 3)

3 讨论

本研究基于单中心回顾性数据,建立了 7 种机器学习模型用于预测经冠状动脉造影证实的 CHD。研究结果显示,在本组测试集内,SVM、Logistic 回归、RF、XGBoost 和 LGBM 模型性能稳健,其中 SVM 模型预测精度高($AUC=0.775$),召回率、精确率、F1 值以及校准性能等关键指标上均表现更优。这表明,在基层医疗场景下,基于易获取的临床变量构建的 SVM 模型,有望成为一种高效、可靠的 CHD 预测工具。

本研究最重要的发现是,通过严谨的特征工程(单因素 Logistic 回归、Boruta 算法和 LASSO 回归)筛选出的 13 个核心变量,足以支撑机器学习模型获得良好的预测性能^[14]。这 13 个变量涵盖了人口统计学(年龄、性别)、生命体征(心率)、生活方式(吸烟史、饮酒史)、既往史(高血压、糖尿病)、实验室检查(白细胞计数、中性粒细胞计数、单核细胞计数、

HDL、白蛋白)以及心脏超声指标(LVEF),均为基层医疗机构在常规诊疗中易于获取的数据^[15-18]。这一特点使得模型具有良好的外推性和基层应用潜力,有效规避了传统影像学和有创检查在基层普及的技术与成本壁垒。甘油三酯和低密度脂蛋白升高被认定为 CHD 的独立危险因素^[19-21]。然而,在本研究的数据体系中,这两个指标并未经三步特征筛选流程进入本研究最终核心变量集,该结果并非人为主动剔除指标所致,而是多重共线性与算法正则化筛选共同作用的客观结果。进一步探究发现,甘油三酯和低密度脂蛋白水平极有可能不仅与高密度脂蛋白水平存在关联^[22],还可能与其他代谢及生活方式变量之间存在一定程度的信息重叠。既往人群研究显示,血脂异常与吸烟、高血压和糖尿病等因素相关,且当前饮酒与高甘油三酯患病率升高相关^[23]。

在本研究 7 种机器学习模型中,SVM 各项评价指标相对均衡。SVM 的核心优势在于其通过核函数处理高维非线性数据的能力,能够有效捕捉变量与结局之间复杂的相互作用,同时对数据中的异常值不敏感^[24]。本研究所纳入的临床数据正具备高维、潜在非线性的特点,SVM 的这一特性使其能够更精细地预测 CHD 的发生。尽管 SVM 模型($AUC=0.775$)与 Logistic 回归模型($AUC=0.770$)预测能力都很强,但其在召回率和 F1 值上的优势,意味着 SVM 模型在避免漏诊方面更具临床实用性,这对于实现“早发现、早诊断、早治疗”的二级预防目标至关重要。

通过 SHAP 方法对最优的 SVM 模型进行解释,不仅量化了特征的重要性,还揭示了其影响方向,增强了模型的透明度和临床可解释性。分析结果与现有 CHD 病理生理学认知高度吻合:男性、高血压、糖尿病、吸烟、饮酒、炎症标志物(白细胞计数、中性粒细胞计数)升高被识别为重要的危险因素;而 LVEF、HDL 和白蛋白水平升高则被识别为保护因素^[25]。这一方面验证了模型构建的生物学合理性,另一方面也提示,该模型并非“黑箱”,其决策逻辑能够为临床医生提供有价值的风险因素洞察。

值得注意的是,SHAP 分析揭示了年龄与 CHD 风险之间并非简单的线性关系,心率也未显示出明确的单向影响。提示在未来的模型优化或临床应用中,需要考虑这些变量与其他因素(如合并症、用药情况)的复杂交互作用,或探索其非线性转换形式,以进一步提升模型精度^[26]。

本研究亦存在如下局限。第一,本研究为单中

心回顾性研究,数据仅来源于一家医院,可能存在选择偏倚,其普适性有待通过多中心、前瞻性,尤其是包含基层医疗机构的外部队列进行进一步验证。第二,受本研究单中心回顾数据限制,研究结论仅适用于本中心就诊人群,临床大范围应用仍需外部队列佐证。第三,尽管 SVM 模型的性能较为稳健,但其 AUC 值(0.775)仍有进一步提升的空间。未来可考虑纳入更多具有潜在价值的变量,如临床用药情况、更详细的血脂亚组分、生活方式细节或心电图特征等,并探索深度学习等更为复杂的模型架构。

综上所述,本研究基于 SVM 初步构建了一个基于基层易获取临床变量的 CHD 预测模型。该模型

在本研究中具有良好的判别性能、临床可解释性,有望为基层医疗预测 CHD 患病提供有价值的工具。未来的工作将集中于通过多中心前瞻性研究进行外部验证,并根据结果研究开发网页版风险评估工具推动该模型在基层医疗机构的集成与应用(图 10,工具网址:<https://research-dashboard.shinyapps.io/CHD-High-Risk-Screening/>),若后续多中心验证结果理想,该轻量化模型或可为基层 CHD 分级防控提供辅助评估手段,为基层心血管疾病管理提供微弱的数据支撑,以期助力“健康中国 2030”战略目标的推进^[27]。

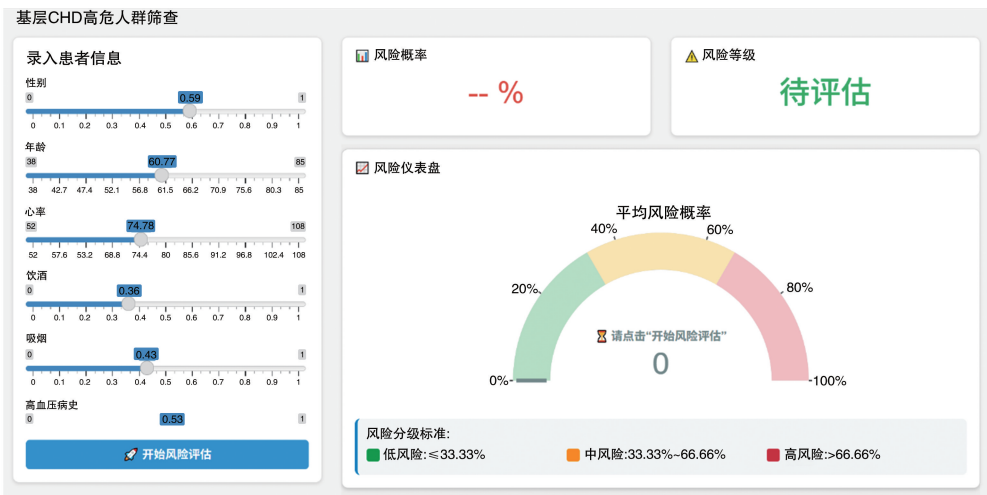


图 10. 网页版风险评估工具示意图

工具网址:<https://research-dashboard.shinyapps.io/CHD-High-Risk-Screening/>。

Figure 10. Schematic diagram of the web-based risk assessment tool

[参考文献]

[1] Heusch G. Coronary blood flow in heart failure: cause, consequence and bystander[J]. Basic Res Cardiol, 2022, 117(1): 1.

[2] 张悦, 谭驰宇, 王宇歌, 等. 机器学习在癌症患者冠心病诊疗中的典型应用[J]. 中国动脉硬化杂志, 2026, 34(1): 1-7. Zhang Y, Tan C Y, Wang Y G, et al. Typical application of machine learning in the diagnosis and treatment of coronary heart disease in cancer patients[J]. Chin J Arterioscler, 2026, 34(1): 1-7.

[3] 刘明波, 何新叶, 杨晓红, 等. 《中国心血管健康与疾病报告 2024》要点解读[J]. 中国全科医学, 2025, 28(32): 3989-4008. Liu M B, He X Y, Yang X H, et al. Interpretation of report on cardiovascular health and diseases in China 2024[J]. Chin Gen Pract, 2025, 28(32): 3989-4008.

[4] 傅利平, 王奕辉, 徐小凤, 等. 健康需求与协同治理: 分级诊疗制度建设实践逻辑——基于公共价值视角的扎根理论研究[J]. 公共管理学报, 2024, 21(3): 111-123, 173-174. Fu L P, Wang Y H, Xu X F, et al. Health demand and collaborative governance: the practical logic of hierarchical diagnosis and treatment system construction-the grounded theoretical analysis from the perspective of public value[J]. J Public Management, 2024, 21(3): 111-123, 173-174.

[5] Trayanova N A, Popescu D M, Shade J K. Machine learning in arrhythmia and electrophysiology[J]. Circ Res, 2021, 128(4): 544-566.

[6] Cai Y Q, Gong D X, Tang L Y, et al. Pitfalls in developing machine learning models for predicting cardiovascular diseases: challenge and solutions[J]. J Med Internet Res, 2024, 26: e47645.

[7] Barrios J P, Tison G H. Advancing cardiovascular medicine with machine learning: progress, potential, and perspective[J]. Cell Rep Med, 2022, 3(12): 100869.

[8] So S, Khalaf A, Yi X R, et al. Induced bioresistance via BNP detection for machine learning-based risk assessment[J]. Biosens Bioelectron, 2021, 175: 112903.

- [9] 中华医学会, 中华医学会杂志社, 中华医学会全科医学分会, 等. 稳定性冠心病基层诊疗指南(2020 年)[J]. 中华全科医师杂志, 2021, 20(3): 265-273.
Chinese Medical Association, Chinese Medical Journal Publishing House, Chinese Society of General Practice, et al. Guideline for primary care of stable coronary artery disease (2020) [J]. Chin J Gen Pract, 2021, 20(3): 265-273.
- [10] Yang X J, Meng P C, Jiang Z X, et al. Deep siamese residual support vector machine with applications to disease prediction[J]. Comput Biol Med, 2025, 196(Part A): 110693.
- [11] Redshaw J, Ting D S J, Brown A, et al. Krein support vector machine classification of antimicrobial peptides [J]. Digit Discov, 2023, 2(2): 502-511.
- [12] Mortensen M B, Dzaye O, Bøtker H E, et al. Low-density lipoprotein cholesterol is predominantly associated with atherosclerotic cardiovascular disease events in patients with evidence of coronary atherosclerosis: the western Denmark heart registry [J]. Circulation, 2023, 147(14): 1053-1063.
- [13] Allison M A, Powell J T, Golledge J. Nontraditional risk factors for peripheral artery disease: setting the scene [J]. Arterioscler Thromb Vasc Biol, 2025, 45(9): 1488-1492.
- [14] Channon K M, Newby D E, Nicol E D, et al. Cardiovascular computed tomography imaging for coronary artery disease risk: plaque, flow and fat [J]. Heart, 2022, 108(19): 1510-1515.
- [15] 刘家佳, 张璐, 张艳, 等. 二维斑点追踪分层应变参数联合白细胞衍生炎症指标诊断冠心病的临床价值[J]. 临床超声医学杂志, 2023, 25(4): 271-276.
Liu J J, Zhang L, Zhang Y, et al. Clinical value of two-dimensional speckle tracking layer-specific strain parameters combined with leukocyte-derived inflammation indicators in the diagnosis of coronary artery disease [J]. J Clin Ultrasound Med, 2023, 25(4): 271-276.
- [16] 王张羽, 曹少清, 李成思, 等. 全身免疫炎症反应指数与冠心病患者冠状动脉狭窄程度及其远期主要不良心血管事件的相关性分析[J]. 海军军医大学学报, 2024, 45(4): 477-486.
Wang Z Y, Cao S Q, Li C S, et al. Correlations between systemic immune-inflammation response index and severity of coronary artery stenosis and long-term major adverse cardiovascular events in patients with coronary heart disease [J]. Acad J Naval Med Univ, 2024, 45(4): 477-486.
- [17] 王娟, 许浩博, 张海鹏, 等. 年龄对冠心病患者冠状动脉非靶病变快速进展的影响 [J]. 中国循环杂志, 2020, 35(6): 538-545.
Wang J, Xu H B, Zhang H P, et al. Impact of age on the rapid progression of coronary non-target lesions in patients with coronary heart disease [J]. Chin Circ J, 2020, 35(6): 538-545.
- [18] 薛小琴, 陈文霞, 兰措卓玛, 等. 心率变异性对稳定性冠心病患者远期预后的预测价值 [J]. 中华老年心脑血管病杂志, 2024, 26(6): 614-618.
Xue X Q, Chen W X, Lan C Z, et al. Predictive value of heart rate variability for long-term prognosis of patients with stable coronary heart disease [J]. Chin J Geriatr Heart Brain Vessel Dis, 2024, 26(6): 614-618.
- [19] Schuitema P C E, Visseren F L J, Nordestgaard B G, et al. Elevated triglycerides are related to higher residual cardiovascular disease and mortality risk independent of lipid targets and intensity of lipid-lowering therapy in patients with established cardiovascular disease [J]. Atherosclerosis, 2025, 408: 120411.
- [20] Zhao Y M, Zhuang Z H, Li Y Y, et al. Elevated blood remnant cholesterol and triglycerides are causally related to the risks of cardiometabolic multimorbidity [J]. Nat Commun, 2024, 15(1): 2451.
- [21] Björnson E, Adiels M, Taskinen M R, et al. Triglyceride-rich lipoprotein remnants, low-density lipoproteins, and risk of coronary heart disease: a UK Biobank study [J]. Eur Heart J, 2023, 44(39): 4186-4195.
- [22] Vaura F, Kim H, Udler M, et al. Multi-trait genetic analysis reveals clinically interpretable hypertension subtypes [J]. Circ Genom Precis Med, 2022, 15(4): e003583.
- [23] Gu D X, Wang D D, Zhu Q Z, et al. Prevalence of dyslipidemia and associated factors in sedentary occupational population from Shanghai: a cross-sectional study [J]. Arch Public Health, 2024, 82(1): 21.
- [24] Balling M, Afzal S, Davey Smith G, et al. Elevated LDL triglycerides and atherosclerotic risk [J]. J Am Coll Cardiol, 2023, 81(2): 136-152.
- [25] Coleman R L, Adler A I, Clarke P M, et al. Estimating the risk of cardiovascular outcomes and all-cause mortality in individuals with type 2 diabetes: validation of the UKPDS outcomes model using TECOS and EXSCEL data [J]. Diabetes Obes Metab, 2026, 28(2): 1068-1075.
- [26] Chen J Q, Wu K X, Lin Y Y, et al. Association of triglyceride glucose index with all-cause and cardiovascular mortality in the general population [J]. Cardiovasc Diabetol, 2023, 22(1): 320.
- [27] 刘继同. 中国家庭、社区健康社会工作战略地位与健康中国、福利中国制度目标 [J]. 湖北社会科学, 2022(5): 45-56.
Liu J T. The strategic position of Chinese family-community health social work and the goals of healthy China and welfare China system [J]. Hubei Soc Sci, 2022(5): 45-56.

(此文编辑 许雪梅)