

· 脂肪性肝病 ·

DOI: 10.12449/JCH260413

代谢相关脂肪性肝病发病风险的机器学习预测模型的构建及验证

邱林杰, 任海燕, 任 燕, 李美洁, 邹姵姵, 吴梓敬, 张 晋

中国中医科学院西苑医院治未病中心/体检科, 北京 100091

通信作者: 张晋, jin_zhang2000@hotmail.com (ORCID: 0009-0001-6557-7180); 任海燕, haiyaner123@qq.com (ORCID: 0009-0002-7863-3513)

摘要: **目的** 探讨基于机器学习方法构建的预测模型对代谢相关脂肪性肝病(MAFLD)发病风险的预测价值,并分析其关键危险因素。**方法** 回顾性收集2021年1月—2024年12月于中国中医科学院西苑医院体检中心体检的2 168例体检者的人体成分、既往史及实验室检验等50个变量信息,根据是否诊断为MAFLD分为MAFLD组($n=265$)和非MAFLD组($n=1\ 903$)。计量资料两组间比较采用Mann-Whitney U 检验;计数资料两组间比较采用 χ^2 检验。将研究数据按照7:3比例随机划分为训练集和验证集。对训练集数据通过单因素分析、Lasso回归及多因素Logistic回归分析筛选预测因子,采用Logistic回归、决策树、随机森林、极端梯度提升、轻量级梯度提升机、支持向量机和人工神经网络共7种机器学习方法构建预测模型。通过绘制验证集受试者操作特征曲线,计算各模型的曲线下面积(AUC)、灵敏度、特异度和约登指数等以评价模型性能,并对最优模型使用沙普利加性解释法解析变量贡献度。**结果** 2 168例研究对象中MAFLD患病率为12.22%(265/2 168)。吸烟、舒张压、相位角、内脏脂肪面积、肌脂比、腰臀比、天冬氨酸氨基转移酶、非高密度与高密度脂蛋白胆固醇比值、甘油三酯-葡萄糖指数和胆结石均是MAFLD发生风险的独立影响因素(P 值均 <0.05)。支持向量机、极端梯度提升、决策树、轻量级梯度提升机、人工神经网络、随机森林和Logistic回归等7种预测模型在验证集中的AUC分别为0.738、0.754、0.757、0.786、0.795、0.796、0.815,其中随机森林模型区分度最佳(AUC=0.796,95%置信区间:0.754~0.839),灵敏度为81.01%,特异度为63.16%,约登指数为44.17%。沙普利加性解释法分析显示,内脏脂肪面积、腰臀比及舒张压为前3位重要预测因子。**结论** 基于人体成分与临床指标构建的随机森林模型对MAFLD发病风险具有良好的预测效能,其可解释性有助于临床早期识别高危人群。

关键词: 代谢相关脂肪性肝病;机器学习;模型;统计学**基金项目:** 国家重点研发计划资助项目(2018YFC2000600);中国中医科学院科技创新工程(CI2021A03005);北京中医药“薪火传承3+3工程”(2023-SZ-A-51)

Construction and validation of machine learning predictive models for the risk of metabolic associated fatty liver disease

QIU Linjie, REN Haiyan, REN Yan, LI Meijie, ZOU Chacha, WU Zijiang, ZHANG Jin

Center for Disease Prevention/Department of Health Assessment, Xiyuan Hospital, China Academy of Chinese Medical Sciences, Beijing 100091, China

Corresponding authors: ZHANG Jin, jin_zhang2000@hotmail.com (ORCID: 0009-0001-6557-7180); REN Haiyan, haiyaner123@qq.com (ORCID: 0009-0002-7863-3513)

Abstract: Objective To investigate the value of predictive models established based on machine learning methods in predicting the risk of metabolic associated fatty liver disease (MAFLD), and to analyze its key risk factors. **Methods** A retrospective analysis was performed for the 50 variables of 2 168 healthy individuals who underwent physical examination in Department of Health Assessment, Xiyuan Hospital, China Academy of Chinese Medical Sciences, from January 2021 to December 2024, including body composition, past history, and laboratory tests, and according to whether they were diagnosed with MAFLD or not,

they were divided into MAFLD group with 265 individuals and non-MAFLD group with 1 903 individuals. The Mann-Whitney *U* test was used for comparison of continuous data between two groups, and the chi-square test was used for comparison of categorical data between two groups. Randomly split the research data into a training set and a validation set in a 70% to 30% ratio. Predictive factors were screened from the training set data using univariate analysis, LASSO regression, and multivariate Logistic regression analysis. Predictive models were then constructed using seven machine learning methods: Logistic regression, decision tree, random forest (RF), eXtreme gradient boosting, light gradient boosting machine, support vector machine, and artificial neural network. Model performance was evaluated by plotting receiver operating characteristic curve for the validation set and calculating the area under the curve (AUC), sensitivity, specificity, and Youden index for each model. Furthermore, the SHapley Additive exPlanation (SHAP) method was used to analyze the contribution of variables in the optimal model. **Results** The prevalence rate of MAFLD among the 2 168 subjects was 12.22% (265/2 168). Smoking, diastolic blood pressure, phase angle, visceral fat area, muscle fat ratio, waist-to-hip ratio, aspartate aminotransferase, non-HDL-C/HDL-C ratio, triglyceride-glucose index, and gallstones were independent risk factors for MAFLD (all $P<0.05$). The seven predictive models of support vector machine, eXtreme gradient boosting, decision tree, light gradient boosting machine, artificial neural network, RF, and Logistic regression had an AUC of 0.738, 0.754, 0.757, 0.786, 0.795, 0.796, and 0.815, respectively, in the validation set, among which the RF model had the best discriminatory ability (AUC=0.796, 95% confidence interval: 0.754—0.839), with a sensitivity of 81.01%, a specificity of 63.16%, and a Youden index of 44.17%. The SHAP analysis showed that visceral fat area, waist-to-hip ratio, and diastolic blood pressure were the top three predictive factors in terms of importance. **Conclusion** The RF model, constructed based on body composition and clinical indicators, has a good performance in predicting the risk of MAFLD, and its interpretability can help to identify high-risk individuals in the early stage in clinical practice.

Key words: Metabolic Associated Fatty Liver Disease; Machine Learning; Models, Statistical

Research funding: National Key R&D Program of China (2018YFC2000600); Chinese Academy of Chinese Medical Sciences Science and Technology Innovation Project (CI2021A03005); Beijing “Mentorship and Inheritance in Chinese Medicine 3+3 Project” (2023-SZ-A-51)

代谢相关脂肪性肝病 (metabolic associated fatty liver disease, MAFLD) 是遗传易感个体由于营养过剩和胰岛素抵抗引起的慢性代谢应激性肝病^[1]。MAFLD 被认为是一种与代谢综合征及其组成成分密切相关的代谢性疾病, 包括高血压、糖尿病、血脂异常和肥胖^[2]。MAFLD 在中国城市人口中的患病率约为 28.77%^[3]。尽管 MAFLD 进展为代谢相关脂肪性肝炎的趋势较低, 但由于其高患病率, MAFLD 仍对肝脏相关健康构成重大威胁, 并可导致较高的发病率和死亡率^[4-5]。因此, 早期识别 MAFLD 高危人群, 对疾病发生发展的预防具有重要意义。本研究系统比较 7 种常用机器学习算法预测 MAFLD 的效能, 并通过可解释性分析揭示关键风险因子, 旨在为 MAFLD 的防治提供依据。

1 资料与方法

1.1 研究对象 本研究是基于中国中医科学院西苑医院体检中心开展的一项回顾性分析, 以 2021 年 1 月—2024 年 12 月于该体检中心进行健康体检的体检者为调查对象。

纳入标准: 年龄 ≥ 18 岁。排除标准: (1) 有饮酒史; (2) 患有自身免疫性肝病、病毒性肝炎、肝硬化或肝癌; (3) 妊娠或哺乳期妇女; (4) 健康体检项目中缺少腹部超声检查者; (5) 健康体检项目中缺少人体成分检查者。

1.2 MAFLD 的定义 以《代谢相关(非酒精性)脂肪性肝病防治指南(2024 年版)》^[1]为依据, 腹部超声检查诊断为脂肪肝, 并存在以下至少 1 项代谢心血管疾病的危险因素: (1) 体重指数 (body mass index, BMI) ≥ 24 kg/m², 或腰围 ≥ 90 cm (男性) 和 ≥ 85 cm (女性); (2) 动脉血压 $\geq 130/85$ mmHg, 或者既往存在高血压病史; (3) 空腹血糖 (fasting plasma glucose, FPG) ≥ 6.1 mmol/L, 或糖化血红蛋白 $\geq 5.7\%$, 或者既往存在 2 型糖尿病病史; (4) 空腹血清甘油三酯 (triglycerides, TG) ≥ 1.7 mmol/L; (5) 高密度脂蛋白胆固醇 (high-density lipoprotein cholesterol, HDL-C) ≤ 1.0 mmol/L (男性) 和 1.3 mmol/L (女性)。

1.3 相关指数计算 (1) 全身免疫炎症指数 (systemic immune inflammation, SII) = 血小板计数 $\times$ 中性粒细胞计数/淋巴细胞计数; (2) 系统炎症反应指数 (systemic inflammatory response index, SIRI) = 单核细胞计数 $\times$ 中性

粒细胞计数/淋巴细胞计数;(3)全身性炎症聚集指数(aggregate index of systemic inflammation, AISI)=血小板计数 $\times$ 单核细胞计数 $\times$ 中性粒细胞计数/淋巴细胞计数;(4)甘油三酯-葡萄糖指数(triglyceride-glucose index, TyG)= $\ln[\text{TG}(\text{mg/dL}) \times \text{FPG}(\text{mg/dL})/2]$;(5)甘油三酯葡萄糖-BMI(triglyceride glucose index-body mass index, TyG-BMI)= $\ln[\text{TG}(\text{mg/dL}) \times \text{FPG}(\text{mg/dL})/2] \times \text{BMI}(\text{kg/m}^2)$;(6)血清尿酸(uric acid, UA)与HDL-C比值(serum uric acid to high-density lipoprotein cholesterol ratio, UHR);(7)非高密度脂蛋白胆固醇与HDL-C比值(non-high-density lipoprotein cholesterol to high-density lipoprotein cholesterol ratio, NHHR)=[总胆固醇(mg/dL)-HDL-C(mg/dL)]/HDL-C(mg/dL);(8)肝脂肪变性指数(hepatic steatosis index, HSI)= $8 \times \text{丙氨酸氨基转移酶}(\text{alanine aminotransferase, ALT})/\text{天冬氨酸氨基转移酶}(\text{aspartate aminotransferase, AST}) + \text{BMI}(\text{kg/m}^2) + 2$ (如果是女性)+2(如果患有糖尿病);(9)肌脂比(muscle fat ratio, MFR):肌肉与脂肪的比值;(10)脂肌比(fat muscle ratio, FMR):脂肪与肌肉的比值;(11)体脂肪、肌肉量、骨骼肌、体脂百分比、基础代谢率、内脏脂肪面积、相位角及骨骼肌指数通过Inbody 770身体成分分析仪检测获得。

1.4 缺失值处理 缺失值分为两种类型处理:(1)缺失值超过20%的变量予以删除;(2)缺失值低于20%的变量采用多重插补进行填补。

1.5 统计学方法 使用R 4.5.1和Python 3.10统计软件进行数据分析。非正态分布的计量资料以 $M(P_{25} \sim P_{75})$ 表示,两组间比较采用非参数Mann-Whitney U 检验。计数资料组间比较采用 χ^2 检验。将研究数据按照7:3比例随机划分为训练集和验证集。在单因素分析的基础上,进一步通过最小绝对收缩和选择算法(least absolute shrinkage and selection operator, Lasso)筛选MAFLD发生的特征变量,将筛选出的特征变量纳入多因素Logistic回归分析中明确独立预测因素。通过Logistic回归、决策树(decision tree, DT)、随机森林(random forest, RF)、极端梯度提升(eXtreme gradient boosting, XGBoost)、轻量级梯度提升机(light gradient boosting machine, LightGBM)、支持向量机(support vector machine, SVM)和人工神经网络(artificial neural network, ANN)7种机器学习方法构建MAFLD预测模型。通过绘制受试者操作特征曲线,计算各模型的曲线下面积(area under curve, AUC)、灵敏度、特异度和约登指数等以评价模型性能。将区分度最优模型通过沙普利加性解释(SHapley

Additive exPlanation, SHAP)法进行模型解释。使用合成少数类过采样技术处理训练集,使发病患者与未发病患者比例接近1:1,以提高模型效能,并使用五折交叉验证法和网格搜索法寻找模型的最优参数。 $P < 0.05$ 为差异有统计学意义。

2 结果

2.1 一般资料 本研究共纳入2 168例体检者,其中男853例(39.35%),女1 315例(60.65%)。根据是否发生MAFLD分为非MAFLD组和MAFLD组,其中MAFLD组265例(12.22%)。性别、吸烟、年龄、身高、体重、BMI、腰围、臀围、腰臀比(waist-to-hip ratio, WHR)、收缩压、舒张压(diastolic blood pressure, DBP)、中性粒细胞计数、血小板计数、单核细胞计数、淋巴细胞计数、SII、SIRI、AISI、ALT、AST、ALT/AST、TG、HDL-C、LDL-C、肌酐、FPG、UA、NHHR、UHR、TyG、TyG-BMI、HSI、体脂肪、肌肉量、MFR、FMR、脂肪质量指数、去脂体重指数、去脂体重、骨骼肌、体脂百分比、基础代谢率、相位角、内脏脂肪面积、骨骼肌指数、高血压、糖尿病和胆结石等风险因素在非MAFLD组与MAFLD组中,差异均有统计学意义(P 值均 < 0.05)(表1)。

2.2 MAFLD发生风险的Lasso回归分析 将单因素筛选的具有统计学意义的风险因子纳入Lasso回归分析,通过交叉验证确定最优 λ 值以筛选特征变量,最终筛选出32个特征变量,包括吸烟、身高、体重、收缩压、DBP、血小板计数、单核细胞计数、淋巴细胞计数、相位角、WHR、腰围、臀围、体脂肪、内脏脂肪面积、脂肪质量指数、骨骼肌指数、MFR、FMR、ALT、AST、ALT/AST、HDL-C、NHHR、TG、FPG、TyG、TyG-BMI、UA、UHR、高血压、糖尿病和胆结石,筛选过程见图1、2。

2.3 MAFLD发生风险的多因素Logistic回归分析 将Lasso回归筛选的32个特征变量纳入多因素Logistic回归分析中,结果表明吸烟、DBP、相位角、内脏脂肪面积、MFR、WHR、AST、NHHR、TyG和胆结石是MAFLD发生风险的独立影响因素(P 值均 < 0.05)(表2)。

2.4 模型构建与验证 基于合成少数类过采样技术处理后的训练集,将多因素Logistic回归分析筛选出的10个预测因子作为输入特征,通过Logistic回归、DT、RF、XGBoost、LightGBM、SVM和ANN共7种机器学习方法构建MAFLD预测模型,并在验证集中进行内部验证。结果显示,7种模型在验证集中的AUC取值范围为0.738~0.815,表明模型具有较好的区分度;根据AUC、灵敏度、

表1 MAFLD组和非MAFLD组基线特征的比较

Table 1 Comparison of baseline characteristics between the non-MAFLD group and the MAFLD group

指标	非 MAFLD 组(<i>n</i> =1 903)	MAFLD 组(<i>n</i> =265)	统计值	<i>P</i> 值
性别[例(%)]			$\chi^2=8.51$	0.004
女	1 176(61.80)	139(52.45)		
男	727(38.20)	126(47.55)		
吸烟[例(%)]	23(1.21)	12(4.53)	$\chi^2=14.12$	<0.001
年龄(岁)	46.00(32.00 ~ 61.00)	50.00(38.00 ~ 62.00)	$Z=-2.62$	0.009
身高(cm)	163.50(158.50 ~ 170.00)	167.00(159.00 ~ 173.50)	$Z=-3.34$	<0.001
体重(kg)	62.60(55.70 ~ 71.45)	73.70(64.40 ~ 83.30)	$Z=-11.71$	<0.001
BMI(kg/m ²)	23.50(21.40 ~ 25.80)	26.60(24.60 ~ 29.00)	$Z=-13.65$	<0.001
腰围(cm)	83.40(77.30 ~ 90.50)	93.60(87.40 ~ 100.40)	$Z=-14.38$	<0.001
臀围(cm)	94.20(90.60 ~ 98.05)	99.30(96.00 ~ 103.50)	$Z=-12.69$	<0.001
WHR	0.89(0.85 ~ 0.93)	0.95(0.92 ~ 0.99)	$Z=-15.84$	<0.001
收缩压(mmHg)	122.00(111.00 ~ 134.00)	131.00(120.00 ~ 140.00)	$Z=-7.38$	<0.001
DBP(mmHg)	70.00(64.00 ~ 78.00)	76.00(69.00 ~ 83.00)	$Z=-7.41$	<0.001
中性粒细胞计数($\times 10^9/L$)	3.23(2.61 ~ 3.99)	3.63(2.99 ~ 4.65)	$Z=-5.63$	<0.001
血小板计数($\times 10^9/L$)	240.00(205.00 ~ 279.00)	249.00(223.00 ~ 294.00)	$Z=-4.14$	<0.001
单核细胞计数($\times 10^9/L$)	0.29(0.24 ~ 0.35)	0.34(0.28 ~ 0.42)	$Z=-6.97$	<0.001
淋巴细胞计数($\times 10^9/L$)	1.83(1.50 ~ 2.21)	2.08(1.73 ~ 2.47)	$Z=-6.70$	<0.001
SII	420.38(314.13 ~ 548.78)	457.96(327.28 ~ 598.05)	$Z=-2.39$	0.017
SIRI	0.50(0.36 ~ 0.73)	0.57(0.43 ~ 0.86)	$Z=-4.12$	<0.001
AISI	122.40(83.34 ~ 177.68)	148.09(97.73 ~ 229.27)	$Z=-5.05$	<0.001
ALT(U/L)	14.20(10.50 ~ 21.00)	22.90(16.20 ~ 35.10)	$Z=-12.71$	<0.001
AST(U/L)	18.30(15.50 ~ 22.00)	20.50(17.20 ~ 25.90)	$Z=-6.78$	<0.001
ALT/AST	0.79(0.63 ~ 1.02)	1.12(0.89 ~ 1.45)	$Z=-13.36$	<0.001
总胆红素($\mu\text{mol/L}$)	10.80(8.60 ~ 14.00)	11.30(8.30 ~ 14.20)	$Z=-0.30$	0.762
TC(mmol/L)	4.75(4.15 ~ 5.41)	4.88(4.17 ~ 5.49)	$Z=-0.95$	0.340
TG(mmol/L)	1.04(0.74 ~ 1.47)	1.37(1.03 ~ 1.89)	$Z=-8.22$	<0.001
HDL-C(mmol/L)	1.33(1.13 ~ 1.56)	1.22(1.02 ~ 1.38)	$Z=6.50$	<0.001
LDL-C(mmol/L)	2.96(2.41 ~ 3.59)	3.11(2.56 ~ 3.73)	$Z=-2.37$	0.018
肌酐(mmol/L)	68.00(59.00 ~ 79.00)	71.00(61.00 ~ 83.00)	$Z=-2.39$	0.017
FPG(mmol/L)	5.29(4.99 ~ 5.69)	5.53(5.17 ~ 6.20)	$Z=-6.40$	<0.001
UA($\mu\text{mol/L}$)	304.00(254.00 ~ 361.00)	364.00(303.00 ~ 422.00)	$Z=-9.10$	<0.001
NHHR	2.55(1.91 ~ 3.33)	3.02(2.38 ~ 3.62)	$Z=-6.37$	<0.001
UHR	0.14(0.11 ~ 0.19)	0.19(0.14 ~ 0.23)	$Z=-9.40$	<0.001
TyG	6.82(6.45 ~ 7.23)	7.12(6.86 ~ 7.50)	$Z=-8.97$	<0.001
TyG-BMI	159.95(140.95 ~ 182.54)	191.61(174.37 ~ 211.98)	$Z=-14.42$	<0.001
HSI	31.31(28.55 ~ 34.56)	36.67(34.26 ~ 40.74)	$Z=-15.96$	<0.001
体脂肪(kg)	19.10(15.40 ~ 23.00)	24.70(21.40 ~ 29.50)	$Z=-14.12$	<0.001
肌肉量(kg)	39.80(35.50 ~ 48.75)	44.80(38.30 ~ 54.20)	$Z=-6.72$	<0.001
MFR	1.24(0.98 ~ 1.62)	1.05(0.84 ~ 1.32)	$Z=-7.88$	<0.001
FMR	0.81(0.62 ~ 1.02)	0.96(0.76 ~ 1.19)	$Z=-7.88$	<0.001
脂肪质量指数	7.10(5.70 ~ 8.70)	9.10(7.70 ~ 10.90)	$Z=-12.40$	<0.001
去脂体重指数	16.10(14.80 ~ 17.90)	17.50(15.90 ~ 19.30)	$Z=-8.16$	<0.001
去脂体重(kg)	42.30(37.65 ~ 51.60)	47.60(40.50 ~ 57.50)	$Z=-6.68$	<0.001
骨骼肌(kg)	22.90(20.10 ~ 28.70)	26.10(21.80 ~ 32.40)	$Z=-6.76$	<0.001
体脂百分比(%)	30.40(25.40 ~ 35.40)	34.40(30.00 ~ 39.10)	$Z=-8.68$	<0.001

表1(续)

Table 1 (continued)

指标	非 MAFLD 组(<i>n</i> =1 903)	MAFLD 组(<i>n</i> =265)	统计值	<i>P</i> 值
基础代谢率(kJ)	5 368.07(4 949.67 ~ 6 209.06)	5 882.70(5 234.18 ~ 6 773.90)	$Z=-7.11$	<0.001
内脏脂肪面积(cm^2)	86.70(66.10 ~ 113.10)	120.00(103.90 ~ 148.10)	$Z=-15.28$	<0.001
相位角($^\circ$)	4.80(4.30 ~ 5.30)	4.60(4.10 ~ 5.20)	$Z=-1.75$	0.033
骨骼肌指数	6.60(6.00 ~ 7.60)	7.40(6.50 ~ 8.20)	$Z=-7.81$	<0.001
高血压[例(%)]	451(23.70)	115(43.40)	$\chi^2=46.78$	<0.001
糖尿病[例(%)]	186(9.77)	50(18.87)	$\chi^2=19.83$	<0.001
胆结石[例(%)]	26(1.37)	23(8.68)	$\chi^2=56.31$	<0.001

注:MAFLD,代谢相关脂肪性肝病;BMI,体重指数;WHR,腰臀比;DBP,舒张压;SII,全身免疫炎症指数;SIRI,系统炎症反应指数;AISII,全身炎症聚集指数;ALT,丙氨酸氨基转移酶;AST,天冬氨酸氨基转移酶;TC,总胆固醇;TG,甘油三酯;HDL-C,高密度脂蛋白胆固醇;LDL-C,低密度脂蛋白胆固醇;FPG,空腹血糖;UA,尿酸;NHHR,非高密度与高密度脂蛋白胆固醇比值;UHR,血清尿酸与高密度脂蛋白胆固醇比值;TyG,甘油三酯-葡萄糖指数;TyG-BMI,甘油三酯葡萄糖-体重指数;HSI,肝脂肪变性指数;MFR,肌脂比;FMR,脂肌比。

表2 MAFLD 发生风险影响因素的多因素 Logistic 回归分析

Table 2 Multivariate Logistic regression analysis of risk factors for MAFLD

变量	OR	95%CI	偏回归系数	标准误	Wald χ^2 值	<i>P</i> 值
吸烟	4.374	1.471 ~ 13.010	1.655	0.760	2.178	0.029
DBP(mmHg)	1.016	1.002 ~ 1.039	0.028	0.013	2.049	0.040
相位角($^\circ$)	0.392	0.286 ~ 0.536	-1.064	0.208	-5.110	<0.001
内脏脂肪面积(cm^2)	1.123	1.087 ~ 1.160	0.126	0.021	5.937	<0.001
MFR	0.074	0.010 ~ 0.562	-3.059	1.327	-2.306	0.021
WHR	7.382	2.964 ~ 16.337	4.132	1.275	3.367	<0.001
AST(U/L)	1.030	1.005 ~ 1.056	0.032	0.016	2.032	0.042
NHHR	0.672	0.517 ~ 0.875	-0.528	0.164	-3.218	0.001
TyG	8.670	4.964 ~ 15.495	6.615	1.497	4.418	<0.001
胆结石	5.974	2.700 ~ 13.218	1.805	0.479	3.767	<0.001

注:MAFLD,代谢相关脂肪性肝病;DBP,舒张压;MFR,肌脂比;WHR,腰臀比;AST,天冬氨酸氨基转移酶;NHHR,非高密度与高密度脂蛋白胆固醇比值;TyG,甘油三酯-葡萄糖指数;OR,比值比;95%CI,95%置信区间。

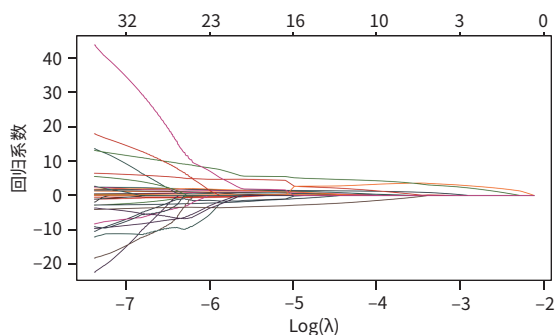


图1 基于 Lasso 回归分析的代谢相关脂肪性肝病发生风险变量选择路径图

Figure 1 Path diagram of risk factor selection for MAFLD based on Lasso regression analysis

特异度及约登指数判别,基于 RF 构建的模型区分度最佳(AUC=0.796,95% 置信区间:0.754 ~ 0.839),灵敏度为 81.01%,特异度为 63.16%,约登指数为 44.17%(图3,表3)。

2.5 SHAP 模型解释 对模型区分度最好的 RF 模型进

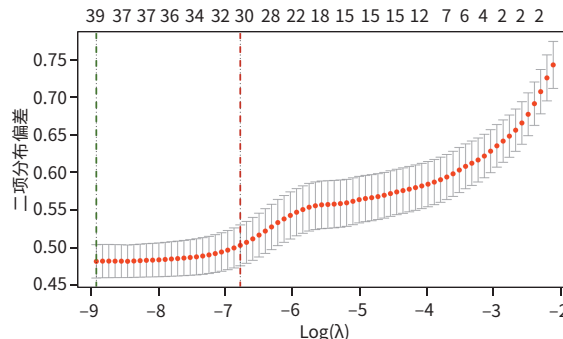
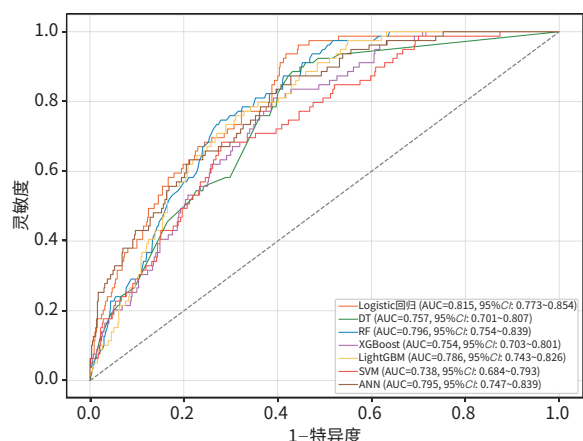


图2 Lasso 回归分析筛选影响代谢相关脂肪性肝病发生的关键特征的交叉验证结果

Figure 2 Cross-validation results of Lasso regression analysis screening key characteristics influencing the occurrence of MAFLD

行 SHAP 解释。SHAP 特征重要性排序结果显示,内脏脂肪面积、WHR 及 DBP 是 MAFLD 发生风险的前 3 个重要预测因子(图4)。SHAP 蜂群图直观地展示了各特征值与模型预测的关系,在前 3 个重要预测因子中,内脏脂



注: Logistic 回归, 逻辑斯谛回归; DT, 决策树; RF, 随机森林; XGBoost, 极端梯度提升; LightGBM, 轻量级梯度提升机; SVM, 支持向量机; ANN, 人工神经网络; AUC, 曲线下面积; 95%CI, 95% 置信区间。

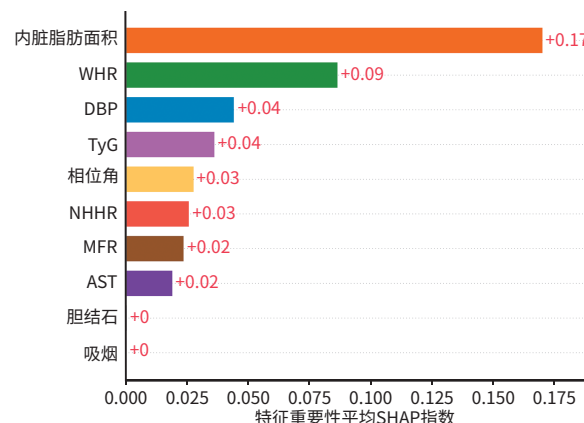
图3 基于机器学习模型的验证集受试者操作特征曲线
Figure 3 Validation set operating characteristic curve based on machine learning model

肪面积、WHR 及 DBP 数值越大, 其 SHAP 值越高, 表明其对模型有正向影响, 即会增加 MAFLD 发生风险 (图 5)。

3 讨论

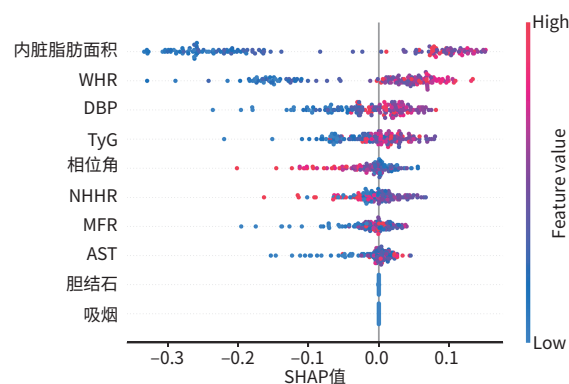
MAFLD 是一种以非酒精因素导致的肝脏脂质过度蓄积为主要特征的慢性代谢性疾病, 其可发展为代谢相关脂肪性肝炎、肝硬化甚至肝细胞癌。MAFLD 通常起病隐匿, 进展缓慢, 多数患者没有表现出明显的临床症状或不适, 早期易被忽视^[6-7], 因此早期发现和干预至关重要。

本研究构建的 7 种机器学习模型均具有良好的预测效能, 其中 RF 模型区分度最佳。进一步对 RF 模型进行 SHAP 解释, SHAP 特征重要性排序结果显示, 内脏脂肪面积、WHR 及 DBP 是 MAFLD 发病风险的三大重要预测因子。流行病学调查研究发现, MAFLD 多合并肥胖, 尤其是腹型肥胖, 其比例高达 70% ~ 80%^[8-9]。内脏脂肪面



注: SHAP, 沙普利加性解释; WHR, 腰臀比; DBP, 舒张压; TyG, 甘油三酯-葡萄糖指数; NHHR, 非高密度与高密度脂蛋白胆固醇比值; MFR, 肌脂比; AST, 天冬氨酸氨基转移酶。

图4 特征重要性排序图
Figure 4 Feature importance ranking chart



注: SHAP, 沙普利加性解释; WHR, 腰臀比; DBP, 舒张压; TyG, 甘油三酯-葡萄糖指数; NHHR, 非高密度与高密度脂蛋白胆固醇比值; MFR, 肌脂比; AST, 天冬氨酸氨基转移酶。

图5 SHAP 蜂群图
Figure 5 SHAP swarm diagram

积和 WHR 均是反映腹部脂肪堆积程度的体脂指标, 是腹型肥胖判定的重要指标。多项基于不同人群的预测模型研究均证实, 内脏脂肪面积是 MAFLD 的独立预测因素, 并确定了其有效的预测临界值, 其预测效能 (AUC

表3 7种机器模型验证集筛检结果

Table 3 Screening results for seven machine model validation sets

机器学习方法	AUC(95%CI)	灵敏度(%)	特异度(%)	约登指数(%)
Logistic 回归	0.815(0.773 ~ 0.854)	18.99	97.54	16.53
DT	0.757(0.701 ~ 0.807)	83.54	59.47	43.02
RF	0.796(0.754 ~ 0.839)	81.01	63.16	44.17
XGBoost	0.754(0.703 ~ 0.801)	81.01	59.47	40.49
LightGBM	0.786(0.743 ~ 0.826)	81.01	59.47	40.49
SVM	0.738(0.684 ~ 0.793)	64.56	72.81	37.36
ANN	0.795(0.747 ~ 0.839)	60.76	79.47	40.23

注: Logistic 回归, 逻辑斯谛回归; DT, 决策树; RF, 随机森林; XGBoost, 极端梯度提升; LightGBM, 轻量级梯度提升机; SVM, 支持向量机; ANN, 人工神经网络; AUC, 曲线下面积; 95%CI, 95% 置信区间。

可达0.826)已得到验证^[10]。此外,无论是国际大规模的横断面调查^[11],还是针对中国人群开发的列线图模型^[12-13],均一致发现WHR和内脏脂肪面积与MAFLD的发生风险呈显著正相关。这提示腹部脂肪过度堆积是MAFLD发病的核心环节,与本研究结果具有一致性。这可能是因为内脏脂肪堆积过多会增加肝脏中的脂肪沉积,促进TG的合成,将更多的游离脂肪酸释放到血液中,干扰糖代谢,导致脂质代谢紊乱和胰岛素抵抗,从而促进脂肪肝的形成^[14]。此外,内脏脂肪的积累还会降低脂联素水平并释放其他炎症因子,引发胰岛素抵抗、炎症和脂肪肝等一系列代谢紊乱的发生^[15]。本研究还凸显了DBP在MAFLD风险评估中的重要性,这与高血压人群MAFLD患病率显著升高(约49.5%)的流行病学发现相一致^[16]。DBP升高是机体处于胰岛素抵抗、交感神经兴奋和慢性炎症等不利代谢状态的标志,这些状态不仅导致DBP升高,更直接作用于肝脏,通过增加脂质输入、促进脂质合成、抑制脂质氧化以及触发炎症和纤维化反应,显著增加了MAFLD的发病风险^[17-18]。

本研究构建的RF模型在验证集中取得了较好的区分度(AUC=0.796),AUC值虽略低于部分文献报道(如康艳红等^[19]基于类别梯度提升算法构建的模型在外部验证中AUC达0.807),但应结合模型的应用场景、特征体系与临床可解释性进行综合评价。首先,本研究模型的优势在于其构建基础为常规体检数据,所有预测变量(如人体成分指标、基础代谢率等)均易于在基层医疗机构或体检中心获取,无须依赖昂贵的特殊检测(如肝脏瞬时弹性成像)或复杂实验室指标,这使得模型具备良好的普适性与推广潜力,尤其适用于大规模人群的早期初筛。其次,本研究创新性地融合了人体成分分析指标与常规临床指标,构建了一个更为全面的评估体系。更重要的是,本研究通过SHAP方法赋予了模型良好的可解释性,也为后续针对性的个体化干预提供了明确的临床路径。

本研究也存在一定局限性:(1)本研究为单中心研究,缺乏外部数据对模型进行验证,未来需通过大样本多中心的队列数据进一步拟合并验证模型;(2)由于体检数据的局限性,本研究未对参与者的饮食、运动情况等进行调查,可能导致影响因素的遗漏;(3)研究人群主要为北方人群,未来需要进一步收集各地区的数据以证实结果的稳健性。

综上所述,本研究基于常规体检指标建立的RF模型对MAFLD发病风险具有良好的预测效能,能为临床

医生提供有价值的决策支持,有利于实现MAFLD的早期干预与防治。

伦理学声明:本研究于2025年3月12日经由中国中医科学院西苑医院伦理委员会审批通过,批号:2025XLA031。

利益冲突声明:本文不存在任何利益冲突。

作者贡献声明:邱林杰负责研究的设计与构思,实施研究,撰写论文,筛选整理数据并进行统计分析,制作图表;任海燕、任燕、李美洁、邹婉婉、吴梓敬审阅论文并提出修改建议;张晋负责文章的质量把关与审核。

参考文献:

- [1] Chinese Society of Hepatology, Chinese Medical Association. Guidelines for prevention and treatment of metabolic dysfunction-associated (non-alcoholic) fatty liver disease (version 2024) [J]. J Pract Hepatol, 2024, 27(4): 494-510. DOI: 10.3760/cma.j.cn501113-20240327-00163.
中华医学会肝病学分会. 代谢相关(非酒精性)脂肪性肝病防治指南(2024年版)[J]. 实用肝脏病杂志, 2024, 27(4): 494-510. DOI: 10.3760/cma.j.cn501113-20240327-00163.
- [2] TANASE DM, GOSAV EM, COSTEA CF, et al. The intricate relationship between type 2 diabetes mellitus (T2DM), insulin resistance (IR), and nonalcoholic fatty liver disease (NAFLD) [J]. J Diabetes Res, 2020, 2020: 3920196. DOI: 10.1155/2020/3920196.
- [3] HOU MM, GU Q, CUI JW, et al. Proportion and clinical characteristics of metabolic-associated fatty liver disease and associated liver fibrosis in an urban Chinese population [J]. Chin Med J, 2025, 138(7): 829-837. DOI: 10.1097/CM9.0000000000003141.
- [4] KRISHNAN A, MUKHERJEE D. Association of cardiovascular health metrics and metabolic associated fatty liver disease: Methodological limitations, and future directions [J]. World J Hepatol, 2025, 17(3): 105635. DOI: 10.4254/wjh.v17.i3.105635.
- [5] RIAZI K, AZHARI H, CHARETTE JH, et al. The prevalence and incidence of NAFLD worldwide: A systematic review and meta-analysis [J]. Lancet Gastroenterol Hepatol, 2022, 7(9): 851-861. DOI: 10.1016/S2468-1253(22)00165-0.
- [6] ZHOU M, BO T, FAN XD, et al. Metabolic dysfunction-associated fatty liver disease: A central hub in systemic metabolic dysregulation [J]. J Clin Hepatol, 2025, 41(9): 1725-1728. DOI: 10.12449/JCH250902.
周蒙, 薄涛, 范修德, 等. 代谢相关脂肪性肝病: 全身代谢性紊乱的核心枢纽之一 [J]. 临床肝胆病杂志, 2025, 41(9): 1725-1728. DOI: 10.12449/JCH250902.
- [7] YANG B, ZHANG R. Progress on the treatment of metabolic associated fatty liver disease [J/CD]. Chin J Liver Dis (Electronic Version), 2024, 16(4): 25-30. DOI: 10.3969/j.issn.1674-7380.2024.04.004.
杨彬, 张瑞. 代谢相关脂肪性肝病治疗进展 [J/CD]. 中国肝脏病杂志(电子版), 2024, 16(4): 25-30. DOI: 10.3969/j.issn.1674-7380.2024.04.004.
- [8] TENG ML, NG CH, HUANG DQ, et al. Global incidence and prevalence of nonalcoholic fatty liver disease [J]. Clin Mol Hepatol, 2023, 29(Suppl): S32-S42. DOI: 10.3350/cmh.2022.0365.
- [9] QUEK J, CHAN KE, WONG ZY, et al. Global prevalence of non-alcoholic fatty liver disease and non-alcoholic steatohepatitis in the overweight and obese population: A systematic review and meta-analysis [J]. Lancet Gastroenterol Hepatol, 2023, 8(1): 20-30. DOI: 10.1016/S2468-1253(22)00317-X.

- [10] YU PP, YANG HC, QI XY, et al. Gender differences in the ideal cut-offs of visceral fat area for predicting MAFLD in China[J]. *Lipids Health Dis*, 2022, 21(1): 148. DOI: 10.1186/s12944-022-01763-2.
- [11] LI HJ, ZHANG Y, LUO HC, et al. The lipid accumulation product is a powerful tool to diagnose metabolic dysfunction-associated fatty liver disease in the United States adults[J]. *Front Endocrinol*, 2022, 13: 977625. DOI: 10.3389/fendo.2022.977625.
- [12] ZHOU BQ, GONG N, HUANG XJ, et al. Development and validation of a nomogram for predicting metabolic-associated fatty liver disease in the Chinese physical examination population[J]. *Lipids Health Dis*, 2023, 22(1): 85. DOI: 10.1186/s12944-023-01850-y.
- [13] YUAN Y, XU MY, ZHANG XF, et al. Development and validation of a nomogram model for predicting the risk of MAFLD in the young population[J]. *Sci Rep*, 2024, 14(1): 9376. DOI: 10.1038/s41598-024-60100-y.
- [14] ANTONIO-VILLA NE, BELLO-CHAVOLLA OY, VARGAS-VÁZQUEZ A, et al. Increased visceral fat accumulation modifies the effect of insulin resistance on arterial stiffness and hypertension risk[J]. *Nutr Metab Cardiovasc Dis*, 2021, 31(2): 506-517. DOI: 10.1016/j.numecd.2020.09.031.
- [15] MAVILIA MG, WU GY. Liver and serum adiponectin levels in non-alcoholic fatty liver disease[J]. *J Dig Dis*, 2021, 22(4): 214-221. DOI: 10.1111/1751-2980.12980.
- [16] KYHL LK, NORDESTGAARD BG, TYBJÆRG-HANSEN A, et al. High fat in blood and body and increased risk of clinically diagnosed non-alcoholic fatty liver disease in 105, 981 individuals[J]. *Atherosclerosis*, 2023, 376: 1-10. DOI: 10.1016/j.atherosclerosis.2023.05.015.
- [17] FAHED G, AOUN L, BOU ZERDAN M, et al. Metabolic syndrome: Updates on pathophysiology and management in 2021[J]. *Int J Mol Sci*, 2022, 23(2): 786. DOI: 10.3390/ijms23020786.
- [18] KATSIKI N, MIKHAILIDIS DP, MANTZOROS CS. Non-alcoholic fatty liver disease and dyslipidemia: An update[J]. *Metabolism*, 2016, 65(8): 1109-1123. DOI: 10.1016/j.metabol.2016.05.003.
- [19] KANG YH, KUANG YM, WEI JF, et al. Establishment and validation of a machine learning-based model for predicting metabolic dysfunction-associated fatty liver disease[J]. *J Chin Pract Diagn Ther*, 2025, 39(7): 611-618. DOI: 10.13507/j.issn.1674-3474.2025.07.005.
- 康艳红, 邝亚梅, 魏君锋, 等. 基于机器学习算法的代谢功能障碍相关脂肪性肝病预测模型构建及验证[J]. *中华实用诊断与治疗杂志*, 2025, 39(7): 611-618. DOI: 10.13507/j.issn.1674-3474.2025.07.005.

收稿日期: 2025-10-06; 录用日期: 2025-11-18

本文编辑: 林姣

引证本文: QIU LJ, REN HY, REN Y, et al. Construction and validation of machine learning predictive models for the risk of metabolic associated fatty liver disease[J]. *J Clin Hepatol*, 2026, 42(4): 848-855.

邱林杰, 任海燕, 任燕, 等. 代谢相关脂肪性肝病发病风险的机器学习预测模型的构建及验证[J]. *临床肝胆病杂志*, 2026, 42(4): 848-855.

· 消息 ·

《临床肝胆病杂志》入驻科学数据银行 (ScienceDB)

为提升肝胆病研究成果的学术透明度与全球影响力,《临床肝胆病杂志》于2026年4月正式入驻科学数据银行 (ScienceDB),启用论文关联数据汇交机制,完善对论文原始数据材料的开放共享策略,实现论文与数据的双向互链与长效引用,进一步提升研究成果的可验证性与学术公信力。

ScienceDB是由中国科学院计算机网络信息中心建设运营的通用型科学数据存储与共享平台。作为支撑我国开放科学发展的重要国家级基础设施,ScienceDB致力于促进科学数据的可发现性、可访问性、互操作性和可重用性 (FAIR Principles),推动数据共享文化氛围在科研共同体中的培育及良性发展,是目前国内最具权威性与影响力的科学数据共享枢纽。

《临床肝胆病杂志》编辑部

2026年4月25日