·临床研究·

基于健康体检数据构建不同代谢综合征风险预测模型及其效能比较

张宇浪，邓艳华，周晓琦，张琼晓，龚文波，盛爱萍

【摘要】目的　基于健康体检数据构建 Logistic 回归模型（LR）、支持向量机（SVM）、决策树（DT）及随机森林（RF）代谢综合征（MS）风险预测模型，并比较其预测效能。**方法**　收集 2020 年 5 月至 2023 年 5 月来宁波市中医院健康体检者 2 538 例，其中 MS 患者 846 例（MS 组），非 MS 患者 1 692 例（非 MS 组）。收集两组体质量指数（BMI）、空腹血糖（FPG）、收缩压、三酰甘油（TG）和高密度脂蛋白胆固醇（HDL-C）作为特征变量，分别建立 LR、SVM、DT 及 RF 模型，并计算 *AUC* 值、准确率、精确度、召回率和 F_1 值等指标对模型效能进行评估。**结果**　两组 BMI、FPG、收缩压、TG 及 HDL-C 等核心代谢指标差异均有统计学意义（均 $P < 0.05$）。多因素 LR 分析显示，BMI、TG、收缩压、HDL-C 及 FPG 均为发生 MS 的独立影响因素（均 $P < 0.05$）。RF 模型在验证集中的 *AUC* 值最高达到 1.0，具备更强的泛化能力与风险预测能力。**结论**　RF 模型表现出良好的 MS 风险预测性能，是 MS 疾病风险预测的理想工具。

【关键词】 代谢综合征；风险预测模型；健康体检数据
DOI:10.3969/j.issn.1671-0800.2026.01.007
【中图分类号】 R589　**【文献标志码】** A　**【文章编号】** 1671-0800(2026)01-0028-04

代谢综合征（metabolic syndrome，MS）是一组严重影响机体健康的临床症候群[1]。随着社会经济的发展、肥胖问题加剧、居民饮食习惯和生活行为方式的转变，MS 的患病率呈上升趋势[2]。近年来，关于 MS 风险预测研究逐渐增多[3-5]，但多采用单一预测模型，准确率有待提高。本研究基于宁波市中医院健康体检数据，探讨发生 MS 的影响因素并建立多种预测模型，以期为临床提供数据，现报道如下。

1　资料与方法

1.1　一般资料　收集 2020 年 5 月至 2023 年 5 月来宁波市中医院健康体检者 2 538 例，其中 MS 患者 846 例，非 MS 患者 1692 例。纳入标准：（1）首次体检年龄 20 ～ 80 岁；（2）有≥2 次连续年份体检数据；（3）首次体检个人信息资料完整，包含体质量指数（BMI）、血压、血脂、血糖、肝肾功能及血常规等体检项目；（4）MS 患者符合《中国 2 型糖尿病防治指南（2020 年版）》中 MS 诊断标准[6]。排除标准：首次体检发现有 MS、高血压病、糖尿病、心脑血管疾病、严重肝肾功能损伤及恶性肿瘤等疾病者。本研究获得宁波市中医院医学伦理委员会批准，所有研究对象均同意参加本研究并签署书面知情同意书。。

1.2　方法

1.2.1　收集资料　（1）一般资料，包括性别、年龄、吸烟及饮酒史等；（2）体格检查，包括身高、体质量及血压，并计算 BMI；（3）实验室检查，包括空腹血糖（FPG）、总胆固醇（TC）、三酰甘油（TG）、高密度脂蛋白胆固醇（HDL-C）、低密度脂蛋白胆固醇（LDL-C）、丙氨酸氨基转移酶（ALT）、天门冬氨酸氨基转移酶（AST）、γ-谷氨酰胺转氨酶（GGT）、血尿酸（SUA）、血清肌酐（SCr）、尿素氮（BUN）、白细胞计数（WBC）、血红蛋白（Hb）及红细胞比容（HCT）。

1.2.2　数据预处理　将所有变量采用 Z-score 方法标准化处理，研究队列以计算机随机法按照 7∶3 比例分为训练集与验证集（random_state=41），即 1 776 例样本作为训练集用于构建模型，762 例样本作为验证集用于内部验证，其中训练集中 MS 患者 601 例（33.84%），验证集中 MS 患者 245 例（32.15%）。

1.3　自变量筛选及机器学习模型的建立　在训练集

基金项目： 宁波市科技计划项目（2022S074）；宁波市医疗卫生高端团队重大攻坚项目（2022030309）；浙江省优势专科建设项目（浙中医药〔2025〕12 号）

作者单位： 315010　宁波，宁波市中医院
通信作者： 邓艳华，Email：dengyanhua0912@163.com

中设计两个实验：(1)实验一。筛选训练集中所选择的特征变量所在列以及标签列，以单个变量作为输入特征，分别在不同分类器[Logistic 回归模型(logistic regression，LR)[7]、支持向量机(support vector machine，SVM)[8]、决策树(decision tree，DT)[9]及随机森林(random forest，RF)[10]]中建立二分类模型，并计算每个模型在该变量下的受试者工作特征曲线(ROC 曲线)下面积(*AUC* 值)，以此作为量化指标评估单个特征变量预测 MS 的能力。(2) 实验二。选取实验一得出的特征变量进行多因素 Logisic 回归分析以筛选 MS 的独立危险因素，并进一步综合评估四种模型整体预测 MS 的效果。方法：综合各特征变量所在列以及标签列为输入特征得出四个分类器模型的真阴性(*TP*)、真阳性(*TN*)、假阴性(*FP*)及假阳性 (*FN*) 值，以此计算各特征变量对标签变量 MS 的 *AUC*、准确率、精确率、召回率、F_1 值，并通过绘制精确率-召回率曲线 (PRC)、ROC、混淆矩阵图进行可视化分析，以直观评价各模型对各类样本的分类情况，最终对比四种机器学习模型对 MS 的预测效果。

本研究共构建四种机器学习模型进行 MS 风险预测，分别为 LR、SVM、DT 及 RF，所有模型均在不设定超参数的前提下进行训练与评估，力求反映基础分类器在临床特征预测 MS 风险中的原始判别能力，为后续模型优化提供基准参考。

1.4　统计方法　数据采用 SPSS 26.0 软件处理。正态分布的计量资料以均数±标准差表示，采用 *t* 检验；偏态分布的计量资料以 $M(P_{25}, P_{75})$ 表示，采用非参数检验。计数资料比较采用 χ^2 检验。基于 python 3.9 的标准库模块进行数据处理、机器学习模型建立及结果可视化输出，绘制 ROC 曲线，并计算 *AUC*、准确率、精确率、召回率和 F_1 值来评估各模型的预测效能。$P < 0.05$ 表示差异有统计学意义。

2　结果

2.1　一般资料比较　两组 BMI、FPG、收缩压、TG 及 HDL-C 等核心代谢指标差异均有统计学意义 (均 $P < 0.05$)，见表 1。

2.2　单变量与 MS 相关性分析　TG、收缩压、HDL-C 及 FPG 在各模型中的 *AUC* 值均＞ 0.75，其中 TG 在各模型中的 *AUC* 值普遍最高，为 MS 相关性最强的特征变量，见表 2。

2.3　多因素 Logistic 回归分析　结果显示，BMI、TG、收缩压、HDL-C 及 FPG 均为发生 MS 的独立影响因素 (均 $P < 0.05$)，见表 3。

2.4　各模型预测性能评估　结果显示，RF 模型在综合指标中表现最优，当 random_state=41 时，其 *AUC* 最高为 1.0，准确率为 0.999，精确率为 0.996，召回率为 1.0、F_1 值为 0.998，见表 4。

2.5　可视化分析　混淆矩阵图显示，RF 模型 TP 和 TN 占比最高，预测 MS 发生概率与实际 MS 发生概率结果之间具有较好的一致性，见图 1。RF 模型 *AUC* 最大，显著优于其他模型，见图 2；在样本不平衡(1∶2)情况下，RF 模型 P-R 曲线仍保持高精确率与召回率的平衡，见图 3。

3　讨论

本研究构建了 MS 风险预测模型，采用 LR、

表 1　两组一般资料比较

项目	MS 组(n=846)	非 MS 组(n=1 692)	t 值	P 值
年龄(岁)	49.67±12.83	44.14±12.03	10.45	＜0.05
身高(cm)	168.91±8.07	167.32±7.89	4.71	＜0.05
体质量(kg)	76.81±13.83	67.31±12.47	16.84	＜0.05
BMI(kg/m^2)	26.79±3.53	23.93±3.36	19.56	＜0.05
收缩压(mmHg)	151.70±16.39	128.06±17.69	33.34	＜0.05
舒张压(mmHg)	91.49±12.01	78.52±11.69	25.89	＜0.05
FPG(mmol/L)	7.04±2.37	5.25±0.98	21.09	＜0.05
TC(mmol/L)	5.32±1.17	5.12±0.97	4.27	＜0.05
TG(mmol/L)	3.41±2.28	1.53±1.11	22.70	＜0.05
HDL-C(mmol/L)	1.05±0.25	1.39±0.37	27.36	＜0.05
LDL-C(mmol/L)	3.08±0.80	3.00±0.74	2.44	＜0.05
ALT(U/L)	48.44±39.24	27.52±20.92	14.51	＜0.05
AST(U/L)	30.57±16.62	23.27±11.18	11.54	＜0.05
GGT(U/L)	64.24±82.20	35.75±33.84	9.68	＜0.05
SUA(μmol/L)	401.04±89.86	360.80±93.00	10.51	＜0.05
SCr(mg/dl)	74.32±14.66	71.03±18.07	4.93	＜0.05
BUN(μmol/L)	4.73±1.18	4.65±1.27	1.50	＞0.05
WBC(×10^9/L)	7.27±2.11	6.27±1.60	12.16	＜0.05
Hb(g/L)	157.61±14.21	148.94±16.62	13.67	＜0.05
HCT(%)	46.03±3.76	44.01±4.38	12.04	＜0.05

注：BMI 为体质量指数，FPG 为空腹血糖，TC 为总胆固醇，TG 为三酰甘油，HDL-C 为高密度脂蛋白胆固醇，LDL-C 为低密度脂蛋白胆固醇，ALT 为丙氨酸氨基转移酶，AST 为天门冬氨酸氨基转移酶，GGT 为γ-谷氨酰胺转氨酶，SUA 为血尿酸，SCr 为血清肌酐，BUN 为尿素氮，WBC 为白细胞计数，Hb 为血红蛋白，HCT 为红细胞比容；1 mmHg≈0.133 kPa

表 2　单变量在 4 种模型中的 *AUC* 值

变量	SVM	LR	DT	RF
TG	0.838	0.887	0.838	0.869
FPG	0.786	0.827	0.742	0.775
收缩压	0.781	0.816	0.813	0.818
HDL-C	0.777	0.802	0.779	0.779
舒张压	0.681	0.754	0.731	0.737
BMI	0.686	0.732	0.592	0.628
GGT	0.699	0.717	0.701	0.706
ALT	0.691	0.721	0.651	0.665
AST	0.639	0.679	0.652	0.653
SUA	0.544	0.631	0.554	0.559
SCr	0.498	0.547	0.567	0.567
BUN	0.528	0.493	0.546	0.551
WBC	0.597	0.628	0.601	0.603
Hb	0.565	0.646	0.598	0.598
HCT	0.554	0.630	0.555	0.554
TC	0.550	0.565	0.492	0.508
LDL-C	0.468	0.529	0.516	0.516

注：BMI 为体质量指数，FPG 为空腹血糖，TC 为总胆固醇，TG 为三酰甘油，HDL-C 为高密度脂蛋白胆固醇，LDL-C 为低密度脂蛋白胆固醇，ALT 为丙氨酸氨基转移酶，AST 为天门冬氨酸氨基转移酶，GGT 为γ-谷氨酰胺转氨酶，SUA 为血尿酸，SCr 为血清肌酐，BUN 为尿素氮，WBC 为白细胞计数，Hb 为血红蛋白，HCT 为红细胞比容；LR 为 Logistic 回归模型，SVM 为支持向量机，DT 为决策树，RF 为随机森林

SVM、DT 和 RF 4 种算法，对 BMI、收缩压、FPG、TG 及 HDL-C 5 项关键诊断指标进行了单变量与多变量的系统建模与评估。实验通过单变量 *AUC* 分析与多变量集成评估，从多个维度验证了各变量与 MS 的相关性及优先级排序，并明确了最优模型选择。

本研究结果显示，两组 BMI、FPG、收缩压、TG 及 HDL-C 等核心代谢指标差异均有统计学意义（均 $P < 0.05$）。提示上述指标在 MS 发病机制中具有潜在关键作用。此外，部分炎症及代谢相关指标如 WBC、GGT 等也表现出较强的关联性，这与近年来关于慢性低度炎症与 MS 关系的研究结果相一致[11]。

在实验一中，以各相关变量在每个模型的 *AUC* 值作为该变量对 MS 判别能力的量化指标。结果表明，TG 的 *AUC* 值最高，提示其在 MS 风险评估中具有最强的单因子判别能力；其次为 FPG 与收缩压，表明其均与 MS 的发生密切相关。HDL-C 与 MS 呈负相关，其 *AUC* 略低但仍具备预测价值，符合其在代谢保护机制中的角色[12]。肝肾功能指标如 GGT、SUA 及 SCr，血液分析指标如 WBC 及 Hb 也表现出

表 3　发生 MS 的独立影响因素分析

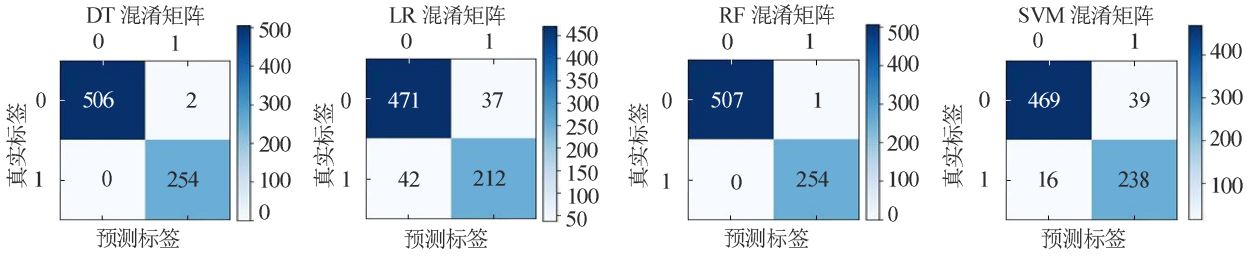
因素	*β* 值	*SE* 值	*Z* 值	*P* 值	*OR* 值（95%*CI*）
BMI	0.243	0.097	2.49	＜0.05	1.276（1.054 ～ 1.544）
TG	0.925	0.134	6.91	＜0.05	2.522（1.940 ～ 3.279）
FPG	1.633	0.151	10.84	＜0.05	5.120（3.812 ～ 6.878）
收缩压	1.798	0.127	14.10	＜0.05	6.037（4.702 ～ 7.750）
HDL-C	−2.070	0.155	12.62	＜0.05	0.126（0.091 ～ 0.174）
常数	−2.027	0.137	14.83	＜0.05	0.132（0.101 ～ 0.172）

注：BMI 为体质量指数，FPG 为空腹血糖，TG 为三酰甘油，HDL-C 为高密度脂蛋白胆固醇

表 4　4 种模型预测效能对比

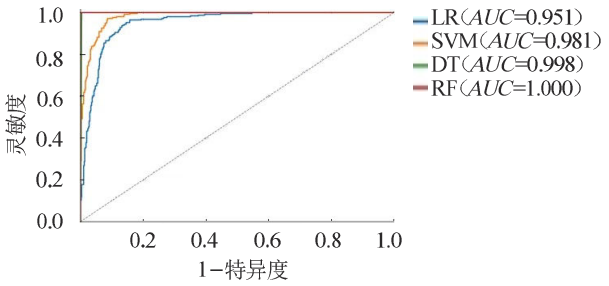
模型	*AUC*	准确率	精确率	召回率	F_1 值（95%*CI*）	*SE* 值	*P* 值
SVM	0.981	0.928	0.859	0.937	0.896（0.972 ～ 0.988）	0.004	＜0.05
LR	0.951	0.896	0.851	0.835	0.843（0.947 ～ 0.973）	0.007	＜0.05
DT	0.998	0.997	0.992	1.000	0.996（0.992 ～ 1.000）	0.002	＜0.05
RF	1.000	0.999	0.996	1.000	0.998（1.000 ～ 1.000）	0.001	＜0.05

注：LR 为 Logistic 回归模型，SVM 为支持向量机，DT 为决策树，RF 为随机森林



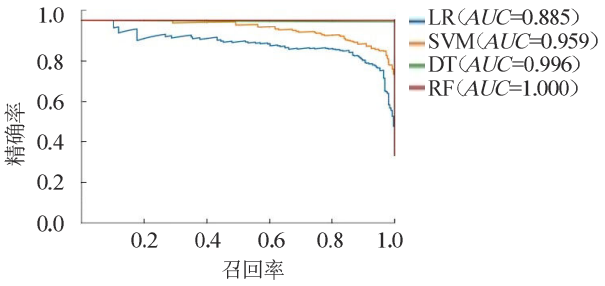
注：混淆矩阵图右下为真阳性（*TP*），左上为真阴性（*TN*），右上为误判（*FP*），左下为漏诊（*FN*）；LR 为 Logistic 回归模型，SVM 为支持向量机，DT 为决策树，RF 为随机森林

图 1　4 种模型在验证集中的二分类混淆矩阵图



注：LR 为 Logistic 回归模型，SVM 为支持向量机，DT 为决策树，RF 为随机森林

图 2　4 种模型的 ROC 曲线



注：LR 为 Logistic 回归模型，SVM 为支持向量机，DT 为决策树，RF 为随机森林

图 3　4 种模型的 PRC 曲线

一定的预测能力，提示其可能参与 MS 相关的炎症与代谢通路[13]。这些单变量模型的表现支持了传统代谢指标在 MS 判别中的重要地位，同时也提示转氨酶、SUA 等非传统变量可用于 MS 的辅助判别。

在实验二中，多因素 Logistic 回归结果显示，BMI、TG、收缩压、HDL-C 及 FPG 均为发生 MS 的独立影响因素（均 $P < 0.05$）。据此推测，高 TG 血症通常是胰岛素抵抗、高血压和肥胖等多种病理状态的共同下游后果[14]，因此承载了多种风险信号信息，使其在单变量分析中表现出的强大区分力。然而在多变量模型中控制收缩压、FPG 等其他因素后，收缩压和 FPG 则表现出与 MS 更强、更独立的关联性，提示它们在 MS 的发生发展中可能扮演着更为基础的“驱动因素”角色。同样，BMI 的 *OR* 值相对较小也提示其作为上游风险因素，其效应很大程度上是通过影响血糖、血脂等下游指标来实现的[15]。各模型预测效能对比结果表明，所有模型均可实现较好的风险预测效果，其中 RF 模型在准确率、*AUC*、精确率、召回率及 F_1 分数等各项指标上均表现最优，具有最强的泛化能力和预测精度，是 MS 疾病风险预测的理想选择。

利益冲突　所有作者声明无利益冲突

作者贡献声明　张宇浪，周晓琦：实验操作，论文撰写；张琼晓，盛爱萍：数据整理，统计学分析；邓艳华，龚文波：研究指导，论文修改，经费支持

参　考　文　献

[1] 高心宇,张昌,黄毓珊,等.代谢综合征男性体力活动水平与体成分、代谢指标、心肺耐力的相关性研究[J].中国医学创新,2021,18(6):29-34.

[2] Li Y R, Zhao L Y, Yu D M, et al. Metabolic syndrome prevalence and its risk factors among adults in China: A nationally representative cross-sectional study[J]. PLoS One, 2018, 13(6): e0199293.

[3] 张慧,陈丹丹,邵静,等.基于 KNN 算法与 logistic 回归的代谢综合征风险预测模型构建与对比研究[J].重庆医学,2023,52(13):2019-2023,2029.

[4] 孙苑漾,杨亚超,曲明苓,等.健康管理人群代谢综合征发病风险预测模型[J].山东大学学报(医学版),2017,55(6):87-92.

[5] Zhang W C, Chen Q C, Yuan Z S, et al. A routine biomarker-based risk prediction model for metabolic syndrome in urban Han Chinese population[J]. BMC Public Health, 2015, 15: 64.

[6] 中华医学会糖尿病学分会.中国 2 型糖尿病防治指南(2020 年版)(上)[J].中国实用内科杂志,2021,41(8):668-695.

[7] 孙胜男.基于 SAEM 算法的协变量缺失下 Logistic 回归模型的统计分析[D].长春:长春工业大学,2024.

[8] 孙逸尘.基于类重叠的分类和聚类方法研究[D].无锡:江南大学,2024.

[9] 孙雨晴.机器学习算法在数据挖掘中的应用[J].中国科技信息,2025(10):50-52.

[10] Carranza C, Nolet C, Pezij M, et al. Root zone soil moisture estimation with Random Forest[J]. J Hydrol, 2021, 593: 125840.

[11] Silveira R J L, Barbalho S M, Reverete de A R, et al. Metabolic syndrome and cardiovascular diseases: Going beyond traditional risk factors[J]. Diabetes Metab Res Rev, 2022, 38(3): e3502.

[12] 李艳君,郝淑梦,寇中靖,等.国内老年人代谢综合征患病危险因素研究进展[J].中国老年保健医学,2024,22(5):120-123.

[13] Naidu B T K, Santosh R K, Bhaskararao J V, et al. Gamma-glutamyl transferase as a diagnostic marker of metabolic syndrome[J]. Cureus, 2023, 15(6): e41060.

[14] Bhargava A. Dietary modifications and lipid accumulation product are associated with systolic and diastolic blood pressures in the women's health trial: Feasibility study in minority populations[J]. Curr Hypertens Rep, 2018, 20(6): 50.

[15] 田晓雪. 芳香烃受体抑制因子在糖脂代谢和肥胖中的作用及机制研究[D].广州:南方医科大学,2021.

收稿日期：2025-08-11

（本文编辑：钟美春）