基于机器学习算法建立2型糖尿病患者冠心病辅助诊断模型

黄浩东[1] 刘小株[1] 龚 军[1] 刘 杰[2] 张祖跃[2] 向天雨[3△]

([1]重庆医科大学医学数据研究院,[2]医学信息学院 重庆 400016;[3]重庆医科大学附属大学城医院信息中心 重庆 401331)

【摘要】 目的 筛选2型糖尿病患者群合并冠心病危险因素并建立风险分类模型,为临床辅助诊断提供有价值的参考。**方法** 通过重庆医科大学大数据平台收集出院时间为2014年1月1日至2019年12月31日行冠状动脉造影术的2型糖尿病患者944例,根据造影结果分为2型糖尿病合并冠心病715例(T2DM-CAD组)和2型糖尿病非冠心病229例(T2DM组)。采用倾向得分匹配法(Propensity Score Matching,PSM)均衡组间混杂因素的影响,匹配后T2DM-CAD组389例,T2DM组221例。使用单因素分析与Logistic回归筛选冠心病发病的危险因素。采用贝叶斯优化(Bayesian Optimization,BO)算法优化支持向量机(Support Vector Machine,SVM)模型、随机森林(Random Forest,RF)模型、极限梯度上升(eXtreme Gradient Boosting,XGB)模型和Logistic回归模型,并比较4种分类模型的分类性能。**结果** 共收集缺失值<30%的指标35项,单因素分析筛选出有统计学差异的指标20项。逐步向前Logistic回归筛选出11项危险因素,包括心率、吸烟、糖尿病肾病、血肌酐、甘油三酯、脂蛋白a、白蛋白、总胆红素、谷草转氨酶、糖化血红蛋白和尿糖。基于危险因素建立的分类模型中优化后的RF模型性能在5折交叉验证(F1值=0.711,AUC=0.811)以及验证集(F1值=0.752,AUC=0.810)中表现最优。**结论** 建立了参数优化RF模型,可用于判断2型糖尿病患者是否合并冠心病,具有良好性能。

【关键词】 机器学习; 2型糖尿病(T2DM); 冠心病; 诊断

【中图分类号】 TP399,R587.1,R541.4 **【文献标志码】** A **doi**:10.3969/j.issn.1672-8467.2022.02.010

Auxiliary diagnosis model of coronary heart disease in patients with type 2 diabetes mellitus based on machine learning algorithm

HUANG Hao-dong[1], LIU Xiao-zhu[1], GONG Jun[1], LIU Jie[2], ZHANG Zu-yue[2], XIANG Tian-yu[3△]

([1]*Medical Data Science Academy*, [2] *School of Medical Informatics*, *Chongqing Medical University*, *Chongqing* 400016, *China*; [3]*Information Center*, *University Town Hospital*, *Chongqing Medical University*, *Chongqing* 401331, *China*)

【Abstract】 Objective To screen the risk factors of coronary heart disease and establish a classification model of coronary heart disease in people with type 2 diabetes, so as to provide a valuable reference for clinical auxiliary diagnosis. **Methods** A total of 944 patients with type 2 diabetes mellitus who underwent coronary angiography on the big data platform of Chongqing Medical University were collected from Jan 1, 2014 to Dec 31, 2019. According to the results of the angiography, they were divided into 715 patients with type 2 diabetes and coronary heart disease (T2DM-CAD group), 229 cases of type 2 diabetes without coronary heart disease (T2DM group). Propensity Score Matching (PSM) was used to balance the effects of confounding factors between groups. After matching, there were 389 cases in T2DM-CAD group and 221 cases in T2DM group. Univariate analysis and Logistic regression were used to screen independent risk

重庆市技术创新与应用发展专项面上项目(cstc2019jscx-msxmX0262);重庆医科大学智慧医学项目(ZHYX2019013,YJSZHYX202017)

△Corresponding author E-mail:421973525@qq.com

网络首发时间:2022-03-08 14:14:45 网络首发地址:https://kns.cnki.net/kcms/detail/31.1885.r.20220304.1750.026.html

factors of coronary heart disease. Bayesian Optimization (BO) algorithm was used to optimize Random Forest (RF) model, Support Vector Machine (SVM) model, eXtreme gradient boosting (XGB) model and Logistic regression model, and their classification performance was compared. **Results** Thirty-five indicators with missing values <30% were included, and 20 indicators with statistical differences were selected by univariate analysis. Eleven risk factors including heart rate, smoking, diabetic nephropathy, serum creatinine, triglycerides, lipoprotein a, albumin, total bilirubin, aspartate aminotransferase, glycosylated hemoglobin, and urine glucose were screened by stepwise forward Logistic regression. In the classification model established based on risk factors, the performance of the optimized RF model was the best in both the 5-fold cross validation (F1 value=0.711, AUC=0.811) and the validation set (F1 value=0.752, AUC=0.810). **Conclusion** In this study, a parameter optimized RF model with good performance was established to determine whether coronary heart disease patients with type 2 diabetes mellitus.

【Key words】 machine learning; type 2 diabetes mellitus (T2DM); coronary heart disease; diagnosis

*** This work was supported by the General Program of Technology Innovation and Application Development of Chongqing Municipality (cstc2019jscx-msxmX0262) and the Intelligent Medicine Project of Chongqing Medical University (ZHYX2019013, YJSZHYX202017).**

2型糖尿病是一种胰岛素分泌不足、胰岛素作用效果差或两者兼而有之的慢性代谢性疾病。随着我国居民生活方式的改变与人口老龄化的加剧，截至2019年我国糖尿病患者数量达到了1.16亿[1]。尽管对于糖尿病是先于冠心病发生还是在疾病早期并存的问题仍有争议，但糖尿病引起的氧化应激、晚期糖基化终末产物和慢性炎症反应对血管内皮功能有害，从而导致心血管疾病的观点已被广泛接受[2]，这表明2型糖尿病是发生微血管和大血管并发症的主要危险因素。糖尿病患者发生心血管疾病的相对风险比非糖尿病患者高2～4倍[3-4]，冠心病是其中最严重的并发症之一，且与非糖尿病的冠心病患者相比，2型糖尿病患者症状往往不典型，可能是因为2型糖尿病患者常伴有严重的自主神经功能障碍[5-6]，使得机体痛阈值增高，即使发生严重心肌缺血，患者心绞痛症状也不明显。冠状动脉造影术虽是诊断冠心病的金标准，但属于有创性检查，且价格昂贵、操作复杂、易产生不良反应，加之2型糖尿病患者痛阈值较高、患病早期无明显疼痛感，易导致疾病治疗延误。因此，本研究从数据驱动的角度，使用机器学习与统计学相关理论方法，对行冠状动脉造影术的2型糖尿病患者建立分类模型，以辅助诊断是否合并冠心病。

资料和方法

数据来源 数据来源于重庆医科大学医学大数据平台，该平台汇集了重庆7家医疗中心的电子病历数据，所有数据均已脱敏。本研究纳入2014年1月1日至2019年12月31日入院行冠状动脉造影术的2型糖尿病患者。纳入标准：(1)既往史中有明确的2型糖尿病的患病年数以及控糖史；(2)住院期间行冠状动脉造影手术且造影记录保存完整。排除标准：(1)糖尿病急性并发症、妊娠期糖尿病以及近期(半年以内)确诊2型糖尿病；(2)患风湿性心脏病、系统性红斑狼疮等自身免疫病；(3)合并癌症；(4)既往已被诊断为冠心病；(5)严重器官衰竭；(6)全身性感染。共计纳入944例2型糖尿病患者，根据冠状动脉造影情况分为2型糖尿病合并冠状动脉狭窄<50%(T2DM组，229例)和2型糖尿病合并冠状动脉狭窄≥50%(T2DM-CAD组，715例)。T2DM组中男性94例，女性135例，年龄33～87岁；T2DM-CAD组中男性422例，女性293例，年龄34～90岁。

指标选取 根据冠心病临床指南和2型糖尿病合并冠心病相关研究[7-9]收集患者行冠状动脉造影术前的35项指标，包括一般资料(如年龄、性别、合并症等)和患者入院后第一次检验的实验室指标(如尿常规、肝肾功能、血脂指标等)。

统计学处理 采用SPSS 25.0和R3.6.1进行统计分析，缺失指标使用missForest算法填补。采用Matchit包的邻近匹配(nearest neighbor matching)方法对收集的原数据按照性别、年龄和是否合并高血压进行倾向评分匹配(propensity score matching，

PSM),卡钳值设定为0.02,T2DM组与T2DM-CAD组按1∶2匹配。采用*KS*方法检验计量资料的正态性,计量资料以$\bar{x}\pm s$或M(P_{25},P_{75})表示,组间比较采用*t*检验或Mann-Whitney U检验;计数资料以例(%)表示,组间比较采用χ^2检验。将两组间有差异的指标纳入逐步向前Logistic回归($\alpha_{入}=0.05$,$\alpha_{出}=0.10$)分析2型糖尿病合并冠心病的危险因素,具体变量名与赋值如表1所示。$P<0.05$为差异有统计学意义。

表1 变量赋值

Tab 1 Variables and their assignments

Characteristic	Assignment
Disease duration of T2DM (y, ×1)	<5=1, 5-10=2, ≥10=3
Heart rate (beats/min, ×2)	<70=1, 70-80=2, ≥80=3
Smoke (×3)	No=0, Yes=1
Combined diabetic nephropathy (×4)	No=0, Yes=1
Fibrinogen (g/L, ×5)	<2.0=1, 2.0-4.0=2, ≥4.0=3
Serum creatinine (μmol/L, ×6)	<40.0=1, 40.0-132.3=2, ≥132.3=3
TC (mmol/L, ×7)	<2.8=1, 2.8-5.2=2, ≥5.2=3
LDL-C (mmol/L, ×8)	<2.1=1, 2.1-3.1=2,≥3.1=3
HDL-C (mmol/L, ×9)	<0.9=1, 0.9-2.0=2,≥2.0=3
TG (mmol/L, ×10)	<1.7=0, ≥1.7=1
Lipoprotein a (mg/L, ×11)	<300.0=0, ≥300.0=1
TP (g/L, ×12)	<65.0=1, 65.0-85.0=2, ≥85.0=3
Globulin (g/L, ×13)	<20.0=1, 20.0-40.0=2, ≥40.0=3
Albumin (g/L, ×14)	<40.0=1, 40.0-55.0=2, ≥55.0 =3
TBil (μmol/L, ×15)	<2.3=1, 2.3-20.4=2, ≥20.4=3
AST (U/L, ×16)	<15.0=1, 15.0-40.0=2, ≥40.0=3
HbA1c (%, ×17)	<6.5=1, 6.5-7.5=2, ≥7.5=3
Glucose (mmol/L, ×18)	<5.3=1, 5.3-9.5=2, ≥9.5=3
Urine protein (×19)	Negative=0, Positive=1
Urine glucose (×20)	Negative=0, Positive=1
Coronary heart disease (y)	No=0, Combined=1

TC: Total cholesterol; LDL-C: Low density lipoprotein cholesterol; HDL-C: High density lipoprotein cholesterol; TG: Triglyceride; TP: Total bilirubin; TBil: Total bilirubin; AST: Aspartate aminotransferase; HbA1c: Aspartate aminotransferase.

机器学习模型构建 分类模型构建采用python 3.8.5版本、anaconda3集成开发环境。将数据按4∶1分为训练集和测试集,训练集用于分类模型的构建。采用Scikit-learn包分别构建Logistic回归模型、随机森林(Random Forest,RF)模型、支持向量(Support Vector Machine,SVM)模型和极限梯度上升(eXtreme Gradient Boosting,XGB)模型。采用bayes_opt包中贝叶斯优化(Bayesian Optimization,BO)算法分别优化XGB算法5个主要超参数n_estimators、subsample、max_depth、learning_rate和min_chid_weight;RF算法3个主要超参数n_estimators、min_samples_split、max_features;SVM算法2个主要超参数C和gamma以及Logistic回归超参数C,设定寻找模型最大AUC对建立的4种机器学习模型进行参数优化。

模型评估 采用5折交叉验证法和验证集评估模型性能,评价指标为准确率、精确率、召回率、F1分数、ROC曲线下面积(AUC),以F1分数和AUC的最大值判断模型是否为最优模型。

结 果

匹配前后两组基线资料比较 T2DM-CAD组匹配前后,冠状动脉单支病变分别为218例(30.49%)和115例(29.56%),冠状动脉两支病变分别为199例(27.83%)和101例(25.96%),冠状动脉

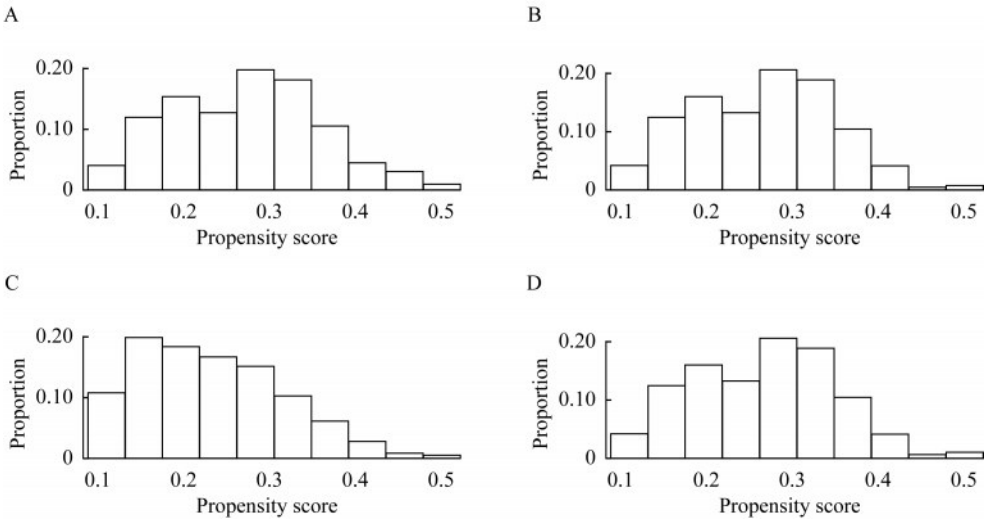
三支病变298例(41.68%)和173例(44.47%)。匹配后T2DM病程、心率、吸烟史、糖尿病肾病差异有统计学意义($P<0.05$)，其余基线资料差异无统计学意义(表2)。匹配后共筛选出610例患者，其中T2DM-CAD组389例，T2DM组221例(表2，图1)。

表2 匹配前后两组基线指标对比

Tab 2 Comparison of baseline indicators between the two groups before and after matching

Baseline indicators	Before matching		*P*	After matching		*P*
	T2DM-CAD gourp (*n*=715)	T2DM group (*n*=229)		T2DM-CAD group (*n*=389)	T2DM group (*n*=221)	
Man	422 (59.02)	94 (41.05)	**<0.001**	185 (47.56)	93 (42.08)	0.192
Age (y)	67.00 (60.00, 74.00)	65.00 (59.00,71.00)	**0.006**	66.00 (60.00, 72.00)	66.00 (60.00, 71.00)	0.933
Disease duration of T2DM (y)	8.34 (5.00, 11.00)	7.00 (4.00,10.00)	**0.001**	8.00 (5.00, 11.00)	7.00 (4.00, 10.00)	**0.014**
Heart rate (beats/min)	79.00 (70.00, 90.00)	76.00 (68.00, 86.00)	**0.003**	80.00 (70.00, 92.00)	75.00 (68.00, 85.00)	**<0.001**
Systolic blood pressure (mmHg)	136.00 (120.00, 152.00)	141.00 (127.50, 151.00)	0.138	137.00 (122.00, 151.00)	138.00 (126.00, 150.00)	0.320
Diastolic blood pressure (mmHg)	78.00 (69.00, 87.00)	80.00 (70.00, 98.50)	0.328	79.00 (70.00, 87.00)	79.00 (72.00, 88.00)	0.401
Smoke	302 (42.24)	54 (23.58)	**<0.001**	140 (35.99)	53 (23.98)	0.002
Drink	222 (31.05)	55 (24.02)	**0.042**	101 (25.96)	54 (24.43)	0.677
Family history of diabetes	63 (8.81)	29 (12.66)	0.087	38 (9.77)	29 (13.12)	0.203
Family history of coronary heart disease	24 (3.36)	14 (6.11)	0.065	14 (3.60)	12 (5.43)	0.056
High blood pressure	501 (70.07)	156 (69.43)	0.577	270 (69.41)	152 (68.78)	0.871
Hyperlipidemia	166 (23.22)	52 (22.71)	0.874	96 (24.68)	48 (21.72)	0.408
Heart block	33 (4.62)	10 (4.37)	0.875	16 (4.11)	10 (4.52)	0.809
Atrial fibrillation	29 (4.06)	12 (5.24)	0.444	13 (3.34)	12 (5.43)	0.211
Carotid atherosclerosis	161 (22.52)	61 (26.64)	0.201	83 (21.34)	59 (26.7)	0.132
Arteriosclerosis of the lower extremities	33 (4.62)	9 (3.93)	0.662	19 (4.88)	9 (4.07)	0.645
Diabetic nephropathy	63 (8.81)	8 (3.49)	**0.008**	36 (9.25)	8 (3.62)	**0.010**

The measurement data in the table is represented by M (P_{25}, P_{75}), and the counting data is represented by *n*(%).



A: T2DM group before matching; B: T2DM group after matching; C: T2DM-CAD group before matching; D: T2DM-CAD group after matching.

图1 根据PSM筛选与剔除的患者倾向评分分布图

Fig 1 Distribution of patient propensity scores screened and excluded according to PSM

单因素分析结果 共纳入22项指标,包括4项基线指标和18项检验指标。单因素分析结果显示,两组间T2DM病程、心率等20项指标差异有统计学意义($P<0.05$),谷氨酰转肽酶和谷丙转氨酶差异无统计学意义(表3)。

表3 T2DM组与T2DM-CAD组相关指标的单因素分析

Tab 3 Univariate analysis of related indexes in T2DM group and T2DM-CAD group

Characteristic	T2DM_CAD group (n=389)	T2DM group (n=221)	$\chi^2/Z/t$	P
Baseline indicators				
Disease duration of T2DM (y)	8.00 (5.00, 11.00)	7.00 (4.00, 10.00)	−2.466	**0.014**
Heart rate (beats/min)	80.00 (70.00, 92.00)	75.00 (68.00, 85.00)	−3.726	**<0.001**
Smoke	140 (35.99)	53 (23.98)	9.395	**0.002**
Diabetic nephropathy	36 (9.25)	8 (3.62)	6.685	**0.010**
Inspection indicators				
Positive urine protein	75 (19.28)	28 (12.67)	4.388	**0.036**
Positive urine glucose	171 (43.96)	46 (20.81)	32.939	**<0.001**
Fibrinogen (g/L)	3.42 (2.79, 4.17)	3.13 (2.70, 3.62)	−3.744	**<0.001**
Serum creatinine (μmol/L)	70.70 (58.00, 88.05)	61.40 (52.00, 73.97)	−6.004	**<0.001**
TC (mmol/L)	4.49 (3.77, 5.12)	4.21 (3.56, 4.79)	−3.030	**0.002**
TG (mmol/L)	1.66 (1.20, 2.29)	1.46 (1.06, 1.87)	−3.806	**<0.001**
LDL-C (mmol/L)	2.48 (1.95, 3.03)	2.24 (1.79, 2.79)	−3.374	**0.001**
HDL-C (mmol/L)	1.09 (0.92, 1.31)	1.15 (1.00, 1.33)	−2.656	**0.008**
Lipoprotein a(mg/L)	214.00 (97.90, 330.95)	138.40 (68.10, 195.90)	−6.557	**<0.001**
TP (g/L)	68.34±6.78	70.76±5.97	−4.428	**<0.001**
Albumin (g/L)	39.75 (37.00, 42.70)	41.90 (39.30, 44.10)	−5.523	**<0.001**
Globulin (g/L)	28.20 (25.71, 30.03)	29.00 (27.00, 30.48)	−3.333	**0.001**
TBil (μmol/L)	9.90 (7.00, 13.75)	10.50 (8.35, 13.60)	−2.203	**0.028**
GGT (U/L)	29.10 (20.00, 51.00)	28.00 (19.00, 41.09)	−1.392	0.164
ALT(U/L)	24.00 (15.00, 34.72)	23.00 (15.35, 31.00)	−0.805	0.421
AST (U/L)	24.20 (18.00, 48.24)	21.60 (17.00, 26.00)	−4.813	**<0.001**
HbA1c (%)	7.90 (6.90, 9.10)	7.05 (6.41, 7.60)	−7.572	**<0.001**
Glucose (mmol/L)	9.66 (7.28, 12.95)	7.74 (6.28, 10.70)	−5.203	**<0.001**

TC: Total cholesterol; TG: Triglyceride; LDL-C: Low density lipoprotein cholesterol; HDL-C: High density lipoprotein cholesterol; TP: Total bilirubin; GGT: glutamyl transpeptidase; TBil: Total bilirubin; ALT: Alanine aminotransferase; AST: Aspartate aminotransferase; HbA1c: Aspartate aminotransferase. The measurement data subject to normal distribution are represented by $\bar{x}\pm s$. The measurement data not subject to normal distribution are represented by M (P_{25}, P_{75}). The enumeration data are represented by n(%).

Logistic回归分析结果 将单因素分析有意义的20个指标进行逐步向前Logistic回归分析,其中11个变量纳入最佳回归方程(表4)。

机器学习模型结果 将表4中的11项指标纳入4种机器学习分类模型,并用BO算法优化4种分类模型,结果显示当n_estimators=2、min_samples_split=10、max_features=69时(表5),无论是5折交叉验证结果还是单独的验证集,RF算法性能最优(表6~7)。图2为4种分类模型的5折交叉验证ROC曲线图。

讨 论

本研究对行冠状动脉造影术的2型糖尿病患者就诊数据进行回顾性分析。由于存在选择偏倚,因此采用“PSM+单因素分析+多因素分析”筛选出2型糖尿病合并冠心病的危险因素,并比较了Logistic回归、SVM、RF、XGB4种分类算法性能,为2型糖尿病在慢病管理中是否发生合并症(本文为冠心病)提供了研究思路,有利于及早启动冠心病的二级预防,减少致死性心血管事件的发生。

表4 2型糖尿病合并冠心病差异性指标Logistic回归分析结果

Tab 4 Logistic regression analysis results of difference index of type 2 diabetes mellitus complicated with coronary heart disease

Indicators	β	SE	$Wald\ \chi^2$	P	OR (95%CI)
Heart rate	0.388	0.120	10.499	0.001	1.47 (1.17, 1.86)
Smoke	0.553	0.220	6.286	0.012	1.74 (1.13, 2.68)
Diabetic nephropathy	0.869	0.442	3.863	0.049	2.38 (1.00, 5.67)
Serum creatinine	1.164	0.467	6.215	0.013	3.20 (1.28, 8.00)
TG	0.405	0.201	4.049	0.044	1.50 (1.01, 2.22)
Lipoprotein a	1.393	0.277	25.378	<0.001	4.03 (2.34, 6.93)
Albumin	−0.473	0.201	5.568	0.018	0.62 (0.42, 0.92)
TBil	−1.101	0.362	9.219	0.002	0.33 (0.16, 0.68)
AST	0.920	0.206	19.993	<0.001	2.51 (1.68, 3.75)
HbA1c	0.366	0.135	7.335	0.007	1.44 (1.11, 1.88)
Urine glucose	0.599	0.236	6.416	0.011	1.82 (1.15, 2.89)

TG: Triglyceride; TBil: Total bilirubin; AST: Aspartate aminotransferase; HbA1c: Aspartate aminotransferase.

表5 参数选择与优化

Tab 5 Parameter selection and optimization

Classification model	Hyperparameter	Define Hyperparameter Ranges	Optimum value
LR	C	0−20	8.71
SVM	Gamma	0−1	0.076
	C	0−20	9.83
RF	n_estimators	0−500	69
	min_samples_split	1−11	10
	max_features	1−11	2
XGB	n_estimators	0−500	392
	subsample	0.5−1	1
	max_depth	1−11	10
	learning_rate	0−1	0.3
	min_chid_weight	1−11	10

LR: Logistic regression; SVM: Support Vector Machine; RF Random Forest; XGB: eXtreme Gradient Boosting.

表6 4种机器学习模型5折交叉验证性能评价表

Tab 6 Performance evaluation table of 4 machine learning models by 5-fold cross-validation

Classification model	Accuracy	Precision	Recall	F1 score	AUC
LR	0.680	0.707	0.692	0.697	0.763
RF	0.721	0.805	0.648	0.711	0.811
SVM	0.714	0.807	0.619	0.699	0.789
XGB	0.682	0.699	0.724	0.709	0.751

Refer to Tab 5.

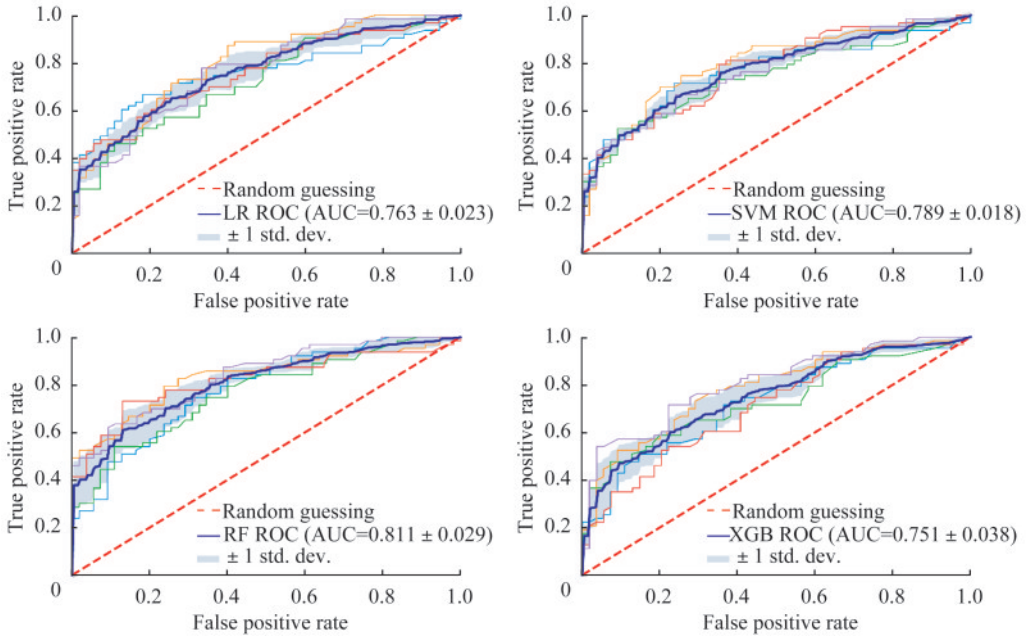
本研究筛选出的2型糖尿病合并冠心病的11项危险因素，包括心率、吸烟、糖尿病肾病、血肌酐、甘油三酯、脂蛋白a、白蛋白、总胆红素、谷草转氨酶、糖化血红蛋白和尿糖。其中血肌酐、糖尿病肾病、尿糖、谷草转氨酶在既往研究中报道较少。血肌酐在临床上常用于评估肾脏功能是否正常，糖尿病肾病是糖尿病最主要的微血管并发症之一[10]，尿糖可作为检测糖尿病患者早期肾损伤的标志物。高浓度血肌酐、合并糖尿病肾病、出现尿糖现象都表明患者肾功能下降[11]，肾功能降低可增加冠心病风险[12-13]。Salim等[14]对非糖尿病新加坡华人进行了一项病例对照研究，发现在传统危险因素中添加血肌酐可以更好地预测冠心病患病风险，这与本研究相符合。谷草转氨酶主要分布于心肌细胞中，约80%的谷草转氨酶作为一种非特异性的细胞内功能酶存在于心肌细胞的线粒体中。心肌损伤时，线粒体受损，谷草转氨酶显著升高[15-16]。因此，谷草转氨酶可以反映心肌细胞损伤的严重程度。研究表明谷草转氨酶与冠心病及其严重程度呈正相关，可以将该指标纳入各种冠心病风险预测模型[17-19]。而在本研究中谷草转氨酶为2型糖尿病合并冠心病的独立危险因素，说明对于2型糖尿病患者，谷草转氨

表7 4种机器学习模型在验证集中的性能评价表

Tab 7 Performance evaluation table of four machine learning models in validation set

Classification model	True positive	True negative	False positive	False negative	F1 score	AUC
LR	47	32	16	27	0.686	0.707
RF	50	39	9	24	0.752	0.810
SVM	49	33	15	25	0.710	0.702
XGB	48	32	16	26	0.696	0.729

Refer to Tab 5.



The thinner solid lines in the figure are the ROC curves verified for five times respectively, the solid blue line in the figure is the average ROC curve of 5-fold cross-validation. LR: Logistic regression; SVM: Support Vector Machine; RF: Random Forest; XGB: eXtreme Gradient Boosting.

图2 4种分类模型5折交叉验证ROC曲线图

Fig 2 5-fold cross-validation ROC of 4 classification models

酶依然可以作为鉴别冠心病的一种生物标志物。而其余的7项危险因素,包括心率、吸烟、甘油三酯、脂蛋白a、白蛋白、总胆红素、糖化血红蛋白,在2型糖尿病合并冠心病的研究报道中多见,与本研究的结论相似[7,20-24]。

虽然利用机器学习模型对冠心病进行疾病诊断已有较多研究[25-27],但都存在以下缺点:(1)冠心病起因不同,应分人群研究;(2)对照组与研究组同质性不高;(3)对照组缺少冠心病风险评估,而患者做过冠状动脉造影术、冠状动脉CT成像等,冠心病评估准确性较高。本研究基于冠状动脉造影术选取糖尿病患者群,根据造影结果分为两组,同质性高,在一定程度上解决了以上缺点。本研究也是国内首次从机器学习的角度判断2型糖尿病患者是否发生冠心病的综合性研究。在机器学习参数调优中,只能看到模型的输入和输出,所以很难通过求导和凸优化的方法来选择模型最佳超参数。以往通常是通过经验来选择超参数,然而这种方式往往得不到性能最优的机器学习模型。BO算法[28]可以很好地解决该问题,其思想为使用贝叶斯网格概率模型来显式反映变量之间的依赖关系及可行解的分布,具体步骤为利用先验知识逼近未知函数的后验分布从而调节超参数。XGB算法[29]是以CART回归树模型为基分类器的一种提升学习算法,是当前比较前沿的基于boosting思想的集成学习算法。SVM算法[30]的目的是寻找一个超平面对样本数据进行分割,然后转换为凸二次规划问题来求解,并且SVM算法在处理线性和非线性数据的小样本条件下具有良好的学习能力。LR算法使用Sigmoid函数作为预测函数。输入变量x通过线性函数输出变量y,然后输出变量y通过Sigmoid函数转换为带标签的结果,有着计算速度快、可解释性好、易于扩展和实现的特点。RF算法由决策树作为基分类器,是一种结合了Bagging集成学习理论和随机子空间方法的集成学习算法[31]。以上4种分类算法在目前疾病风险预测与疾病诊断中运用最多。在本研究中,优化后的RF模型(5折交叉验证:AUC=0.811,测试集:AUC=0.810)分类性能优于优化后的Logistic回归模型(5折交叉验证:AUC=0.763,测试集:AUC=0.707)、SVM模型(5折交叉验证:AUC=0.789,测试集:AUC=0.702)与XGB模型(5折交叉验证:AUC=0.751,测试集:AUC=0.709),而Logistic回归模型、SVM模型和XGB模型3者分类性能相差不大。RF算法具有分类精度高、运算速度快、鲁棒性好等优点。在一些样本量和指标数与本研究相似的研究中,RF算法的分类性能表现为

最优[32-33]，与本研究结果相似。

本研究存在一定的局限性：首先，MissForest算法对混合型缺失数据插补后为优良数据的缺失极限是缺失值＜30%[34]，因此本研究未纳入缺失值＞30%的指标（如BMI、血清C肽）。其次，本研究为回顾性临床研究，且模型缺少外部验证，结果需要进一步验证。最后，本研究建立的模型召回率较低，临床应用有一定的局限性。

综上，本研究基于2型糖尿病患者就诊数据，筛选出11项冠心病危险因素，并基于危险因素建立风险分类模型，研究结果得出贝叶斯优化后的RF算法具有较好的分类能力。可将本研究建立的模型嵌入临床决策支持系统，实现2型糖尿病患者在内分泌科就诊时收到冠心病风险提示以减少漏诊。

作者贡献声明 黄浩东 研究设计和实施，论文构思和撰写。刘小株，龚军 研究实施，数据采集和整理。刘杰，张祖跃 研究设计，论文修订。向天雨 研究选题和设计，论文终审。

利益冲突声明 所有作者均声明不存在利益冲突。

参 考 文 献

[1] IDF. IDF DIABETES ATLAS (9th edition 2019)[EB/OL]. https://www.diabetesatlas.org/data/en/country/42/cn.html.

[2] TOUSOULIS D, PAPAGEORGIOU N, ANDROULAKIS E, *et al*. Diabetes mellitus-associated vascular impairment: novel circulating biomarkers and therapeutic approaches[J]. *J Am Coll Cardiol*, 2013, 62(8): 667-676.

[3] GREGG EW, SATTAR N, ALI MK. The changing face of diabetes complications[J]. *Lancet Diabetes Endocrinol*, 2016, 4(6): 537-547.

[4] ZHENG Y, LEY SH, HU FB. Global aetiology and epidemiology of type 2 diabetes mellitus and its complications[J]. *Nat Rev Endocrinol*, 2018, 14(2): 88-98.

[5] 金学林，沈卫峰，陆林，等.2型糖尿病无症状性心肌缺血的研究进展[J].国际心血管病杂志，2008，35(3)：154-158.

[6] 杨秀颖，张莉，陈熙，等.2型糖尿病周围神经病变机制研究进展[J].中国药理学通报，2016，(5)：598-602.

[7] 《中国高血压防治指南》修订委员会.中国高血压防治指南2018年修订版[J].心脑血管病防治，2019，19(1)：1-44.

[8] 曾玲，陆泽元，余颖，等.2型糖尿病合并无症状冠心病的危险因素分析[J].中国糖尿病杂志，2018，26(5)：20-23.

[9] KOPIN L, LOWENSTEIN C. Dyslipidemia [J]. *Ann Intern Med*, 2017, 167(11): 81-96.

[10] 童国玉，朱大龙.糖尿病肾病国内外临床指南和专家共识解读[J].中国实用内科杂志，2017，37(3)：211-2116.

[11] 钟文晖.尿糖、尿微量清蛋白联合检测在糖尿病早期肾损伤诊断中的临床价值[J].国际检验医学杂志，2016，37(3)：403-404.

[12] ASTOR BC, CORESH J, HEISS G, *et al*. Kidney function and anemia as risk factors for coronary heart disease and mortality: the Atherosclerosis Risk in Communities (ARIC) Study[J]. *Am Heart J*, 2006, 151(2): 492-500.

[13] DI ANGELANTONIO E, DANESH J, EIRIKSDOTTIR G, *et al*. Renal function and risk of coronary heart disease in general populations: new prospective study and systematic review[J]. *PLoS Med*, 2007, 4(9): e270.

[14] SALIM A, TAI ES, TAN VY, *et al*. C-reactive protein and serum creatinine, but not haemoglobin A1c, are independent predictors of coronary heart disease risk in non-diabetic Chinese[J]. *Eur J Prev Cardiol*, 2016, 23(12): 1339-1349.

[15] 倪丹，张玲玲，潘洪川，等.冠心病患者血清CRP、Hcy及心肌酶与冠脉狭窄程度的相关性研究[J].标记免疫分析与临床，2019，26(12)：2048-2052.

[16] SHEN J, ZHANG J, WEN J, *et al*. Correlation of serum alanine aminotransferase and aspartate aminotransferase with coronary heart disease[J]. *Int J Clin Exp Med*, 2015, 8(3): 4399-4404.

[17] EVANS JM, OSTROW BH, POLIS GN, *et al*. Serum glutamic-oxalacetic transaminase in coronary artery disease; a review of 201 cases[J]. *Circulation*, 1956, 14(5): 790-799.

[18] MADAN SA, SINGAL D, PATEL SR, *et al*. Serum aminotransferase levels and angiographic coronary artery disease in octogenarians[J]. *Endocrine*, 2015, 50(2): 512-515.

[19] 刘志强，吴振军，杨刘顺，等.BNP联合心肌酶检测对冠心病危险分层和冠脉搭桥术疗效的预测作用[J].山东医药，2016，56(5)：57-59.

[20] MURASE T, OKUBO M, AMEMIYA-KUDO M, *et al*. Impact of elevated serum lipoprotein (a) concentrations on the risk of coronary heart disease in patients with type 2 diabetes mellitus[J]. *Metabolism*, 2008, 57(6): 791-795.

[21] ARQUES S. Human serum albumin in cardiovascular diseases[J]. *Eur J Intern Med*, 2018, 52: 8-12.

[22] WANG J, WU X, LI Y, *et al*. Serum bilirubin concentrations and incident coronary heart disease risk among patients with type 2 diabetes: the Dongfeng-Tongji cohort[J]. *Acta Diabetol*, 2017, 54(3): 257-264.

（下转第256页）