DOI:10.11830/ISSN.1000-5013.202110012

糖尿病联合并发症发病风险计算与预测

郑尔昌[1],邹金串[2],薛成斌[3],张晋伟[1],陈少阳[4],陈强[4],胡国鹏[1]

(1. 华侨大学 体育与健康科学研究中心,福建 泉州,362021;
2. 华侨大学 旅游学院,福建 泉州,362021;
3. 仰恩大学 管理学院,福建 泉州,362014;
4. 福建省泉州市丰泽区华大街道社区卫生服务中心,福建 泉州,362021)

摘要: 采用国家人口与健康科学数据共享平台临床医学科学数据中心提供的 3 000 例糖尿病并发症数据作为数据集,对糖尿病联合并发症发病风险进行计算与预测. 通过关联规则查找高风险联合并发症并计算各联合并发症的关联发病率,采用随机森林算法建立高风险联合并发症发病预测模型,并查找其关键影响因素. 研究结果表明:部分联合并发症关联发病率超过 90%;在筛选出的 12 组高风险联合并发症中,高血压、动脉粥样硬化、视网膜病变、冠心病、肾病等是常见并发症;不同的联合并发症中关键影响因素(生化指标)各不相同;各联合并发症十折交叉验证法的分类平均精度均在 0.800 0 以上,曲线下面积(AUC)值均大于 0.67.

关键词: 糖尿病;并发症;关联发病率;关键因素;发病预测;关联规则;随机森林

中图分类号: R 587.1; TP 181 **文献标志码:** A **文章编号:** 1000-5013(2022)04-0498-13

Risk Calculation and Prediction of Diabetes Combined Complications Incidence

ZHENG Erchang[1], ZOU Jinchuan[2], XUE Chengbin[3],
ZHANG Jinwei[1], CHEN Shaoyang[4],
CHEN Qiang[4], HU Guopeng[1]

(1. Sports and Health Science Research Center, Huaqiao University, Quanzhou 362021, China;
2. College of Tourism, Huaqiao University, Quanzhou 362021, China;
3. College of Management, Yang'en University, Quanzhou 362014, China;
4. Community Health Service Center of Huada Street of Quanzhou Fengze District of Fujian Province, Quanzhou 362021, China)

Abstract: Using the data of 3 000 cases of diabetes complications provided by Clinical Medical Science Center of the National Population and Health Science Data Sharing Platform as a data set, the risk of diabetes combined complications was calculated and predicted. High-risk combined complications were found through association rules and the associated morbidity rate of each combined complication was calculated. The random forest algorithm was used to establish a high-risk combined complication incidence prediction model, and its key influencing factors were found. The research results show that: the related incidence rate of partial combined complications exceeds 90%. Among the slected 12 groups of high-risk combined complications, hypertension,

收稿日期: 2021-10-08

通信作者: 胡国鹏(1978-),男,教授,博士,主要从事运动与健康、运动氧科学的研究. E-mail:hugp@hqu.edu.cn.

基金项目: 国家体育总局科研基金资助项目(2017C106);福建省自然科学基金规划项目(2020J01087);福建省泉州市丰泽区科技计划项目(2020FZ34)

atherosclerosis, retinopathy, coronary heart disease, nephropathy, *et al* are common complications. The key influencing factors (biochemical indices) in different combined combinations are different. The classification average accuracy of each combined ten fold cross validation method is over 0.800 0, and the area under curve (AUC) value is all greater than 0.67.

Keywords: diabetes; complications; related incidence rate; key factor; incidence prediction; association rules; random forest

糖尿病(diabetes)作为一种慢性疾病，其发病率逐年增高. 糖尿病慢性并发症是患者致死、致残的重要原因. 根据世界卫生组织统计，糖尿病并发症目前已高达 100 多种，主要包括糖尿病肾病、糖尿病眼部并发症、糖尿病足、糖尿病心血管并发症、糖尿病性脑血管病和糖尿病神经病变等几大类. 因此，寻找糖尿病并发症发病规律并根据相关指标进行并发症预警，进而辅助医疗工作者尽早诊断及预防糖尿病并发症，是当前大健康领域的研究热点之一[1]. 文献[2-4]分别通过回归模型、机器算法模型等评估人群糖尿病患病风险.

近年来，随着医疗设备的升级与机器学习算法的应用，部分学者将研究重点转移至糖尿病并发症的诊断预测方面，主要包括关键生理生化指标预测和机器学习算法预测两类. 通过关键生理生化指标进行糖尿病并发症预测，在传统糖尿病并发症预测领域中应用较为广泛. 文献[5-8]分别通过患者血清尿酸(SUA)、尿微量蛋白(MAU)水平、皮肤无创晚期糖基化终末产物等生化指标和收缩压、心率、呼吸等生理指标对糖尿病并发症进行诊断预测. 统计学方法与机器学习的广泛应用进一步推动了糖尿病并发症预测研究的发展. 文献[9-11]通过 Cox 回归分析分别对糖尿病未来 5 年心脑血管事件和继发性功能障碍进行预测，研究结果为糖尿病社区管理提供了一定的参考. 文献[12-13]均采用了 Logistic 回归模型对糖尿病并发症进行预测，模型对糖尿病患者的并发症诊断具有较高的预测价值. 随着机器学习各类算法不断优化，模糊综合评价法[14]、神经网络模型[15-16]及其他新兴机器学习算法[17-20]均针对性地应用于糖尿病并发症的诊断预测中，并取得较好的预测效果.

目前，通过不同研究方法对糖尿病并发症进行诊断预测，取得了较丰富的研究成果. 对糖尿病并发症诊断预测的研究方法与工具也逐渐从统计学方法为主向统计学模型与机器学习算法结合使用转变，随着人工智能的发展，将会有越来越多的诊断预测工具应用于该领域的研究中. 但当前研究较多聚焦于单一糖尿病并发症或常见糖尿病联合并发症，缺乏从糖尿病联合并发症发病风险角度进行研究. 鉴于上述问题，本文对糖尿病联合并发症发病风险进行计算与预测.

1　数据与研究方法

1.1　数据来源

数据来源于国家人口健康科学数据中心《糖尿病并发症预警数据集》，包含解放军总医院 2013—2017 年的 2 型糖尿病住院患者数据共 3 000 例. 数据集 1,2 各字段内容，分别如表 1,2 所示.

33 项糖尿病并发症(表 1 中的 LABEL 与表 2 中的并发症)用于高风险联合并发症筛选. 在进行数据分析之前，对原始数据各字段进行预处理，主要包括类型转换及缺失值处理，舍弃较多缺失值的指标，其他缺失值指标通过 Python 中 sklearn 模块对缺失值进行填充处理. 完成数据预处理后，构建高风联合并发症发病预测模型. 糖尿病并发症数据处理流程图，如图 1 所示.

表 1　数据集 1 各字段内容

Tab.1　Content of each field in dataset 1

字段类型	字段	字段介绍	字段类型	字段	字段介绍
	Case_ID	患者 ID(同 1 病人 ID 对应 1 条记录)		HEIGHT	身高
	LABEL	0-糖尿病，1-糖尿病视网膜病变		WEIGHT	质量
基本信息	AGE	年龄	基本信息	BP_HIGH	收缩压
与数据	SEX	性别(0-女，1-男)	与数据	BP_LOW	舒张压
	NATION	民族(0-汉族，1-其他)		HEART_RATE	心率
	MARITAL_STATUS	婚姻状态(0-其他，1-已婚)		BMI	身体质量指数

表2 数据集2各字段内容

Tab. 2 Content of each field in dataset 2

字段类型	字段
并发症 (0-未患病,1-患病)	高血压,高脂血,动脉粥样硬化,脑卒中,颈动脉狭窄,脂肪肝,肝硬化,其他慢性肝病,胰腺外分泌疾病,胆道疾病,肾病,肾衰,神经系统疾病,冠心病,心肌梗死,心功能不全及心力衰竭,心律失常,呼吸系统疾病,下肢动脉病变,血液病,风湿免疫疾病,妊娠哺乳期,其他内分泌疾病,内分泌腺瘤,多囊卵巢综合征,消化系肿瘤,泌尿系肿瘤,妇科肿瘤,乳腺肿瘤,肺部肿瘤,颅内肿瘤,其他肿瘤
生化指标	血清白蛋白(ALB,35～50 g·L^{-1}),快速微量尿蛋白/肌酐(ALB_CR),碱性磷酸酶(ALP),谷丙转氨酶(ALT),部分活化凝血酶原时间(APTT),谷草转氨酶(AST),血尿素(BU),血尿素氮(BUN),肿瘤标志物(CA199),空腹C肽(CP),C反应蛋白(CRP),直接胆红素(DBILI,0～8.6 μmol·L^{-1}),血沉(ESR红细胞沉降率),纤维蛋白原(FRG),纤维蛋白(DIBRIN),谷氨酰胺转移酶(GGT),球蛋白(GLO),空腹血糖(GLU),餐后2 h血糖(GLU_2H),糖化血清蛋白(GSP,125～240 μmol·L^{-1}),血红蛋白(HB),糖化血红蛋白(HBAIC),高密度脂蛋白胆固醇(HDL_C,1.0～1.6 mmol·L^{-1}),间接胆红素(IBILI),空腹胰岛素(INS),乳酸脱氢酶(LDH_L,666.8～1667.5 n Kat),低密度脂蛋白胆固醇(LDL_C,0～3.4 mmol·L^{-1}),脂蛋白(LP_A),血清脂肪酶(LPS),巨噬细胞(M1_M2),红细胞压积(PCV,红细胞比积测定),磷脂(PL),血小板(PLT),凝血酶原时间(PT),凝血酶原活动度(PTA),血肌酐(SCR),血清尿酸(SUA),总胆红素(TBILI,0～21 μmol·L^{-1}),总胆固醇(TC,3.1～5.7 mmol·L^{-1}),甘油三酯(TG,0.4～1.7 mmol·L^{-1}),T辅助细胞(TH2),总蛋白(TP,55～80 g·L^{-1}),尿肌酐(UCR),24 h尿微量蛋白(UPR_24)

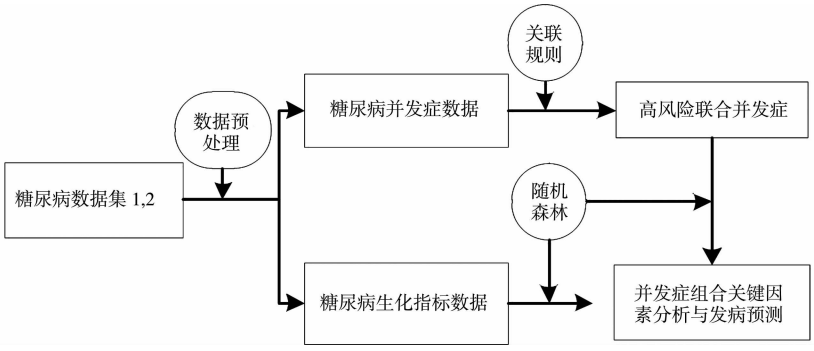


图1 糖尿病并发症数据处理流程图

Fig. 1 Data processing flow chart of diabetes complications

1.2 关联规则

关联规则由频繁项集产生,因此,每个规则都满足最小支持度(S_{min})与最小置信度(C_{min}),即支持度和置信度需满足最小阈值.其中,关联数据项支持度为几个关联的数据项在数据集中出现的次数(num(AB))占所有的样本数在数据集中出现的次数(num(Allsamples))的比例,关联数据项{A,B}的支持度$S(A,B)$计算公式为

$$S(A,B)=P(AB)=\frac{\mathrm{num}(AB)}{\mathrm{num}(\mathrm{Allsamples})}. \tag{1}$$

关联数据项{A,B}的置信度$C(A\Rightarrow B)$计算公式为

$$C(A\Rightarrow B)=P(B|A)=\frac{P(AB)}{P(A)}. \tag{2}$$

式(1),(2)中:$P(AB)$为数据项{A,B}在数据集(num(Allsamples)中出现的概率;$P(A)$为数据项{A}在数据集(num(Allsamples)中出现的概率.

通过各关联数据项的支持度与最小支持度阈值的比较,得到频繁项/项集;通过频繁项/项集中各规则的置信度与最小置信度阈值的比较,得到关联规则[19].

1.3 随机森林算法

随机森林算法是通过集成学习的思想将多棵树集成的算法,其基本单元是决策树,本质属性为机器学习领域的集成学习方法.进行分类训练时,首先,有放回地从数据集中取出数据进行训练,构建决策

树，多次训练可得到多棵决策树. 其次，通过对不同的树进行分类，得到不同的分类结果，将所有分类结果进行统计投票，即可得到最终的分类结果. 决策树主要通过信息熵和信息增益进行特征选择，信息熵的计算公式为

$$\mathrm{Info}(D) = -\sum_{i=1} P_i \times \log_2{}^{P_i}. \tag{3}$$

式(3)中：P_i 为数据集 D 中任意元组属于 C_i 的非零概率.

在数据集 D 中所有属性信息熵 $\mathrm{Info}(D_n^m)$ 计算公式为

$$\mathrm{Info}_{\mathrm{attr}}(D) = \sum_{m=1,n=1} \frac{|D_n^m|}{|D_{\mathrm{attr},n}|}\mathrm{Info}(D_n^m). \tag{4}$$

式(4)中：$|D_n^m|$ 为数据集 D 中属性 attr 第 n 个属性值属于第 m 个类别的个数；$|D_{\mathrm{attr},n}|$ 为数据集 D 中属性 attr 第 n 个属性值的个数.

完成数据集和各属性的信息熵计算后，可根据信息增益确定决策树的特征选择顺序，信息增益的计算公式为

$$\mathrm{Gain}(\mathrm{attr}) = \mathrm{Info}(D) - \mathrm{Info}_{\mathrm{attr}}(D). \tag{5}$$

将数据集中各属性的信息熵计算结果分别代入式(5)中，计算各属性的信息增益，将信息增益最大的属性作为第一特征进入决策树，并按上述步骤完成决策树剩余节点的选择. 随机森林算法实现流程，如图 2 所示.

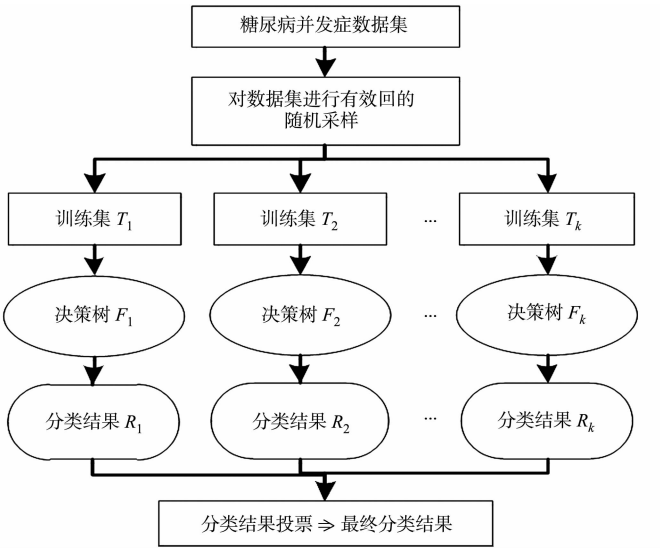


图 2　随机森林算法实现流程

Fig. 2　Algorithm implementation process of random forests

随机森林算法分为训练数据、构建模型和投票 3 个步骤：1) 训练数据，对原始数据集进行可放回随机抽样，形成 k 组训练集；2) 构建模型，对每 1 个训练集，均从样本的 n 个特征随机选取 m 个特征，构建最优学习模型(决策树)；3) 投票，输入测试数据，得到 k 个最优学习模型，给出分类结果，对 k 个分类结果进行投票，得到最终分类结果.

2　高风险联合并发症模型构建

假设数据集中存在糖尿病联合并发症$\{A,B,C\}$，该联合并发症支持度大于支持度阈值，为频繁项集. 频繁项集的关联规则，如表 3 所示. 表 3 中：$C\{A,B\Rightarrow C\}$表明若某糖尿病患者患有并发症 A 和 B，则其同时患有并发症 C 的概率为 55%，该概率为关联发病率(发病风险)，若该频繁项集中超过 1/2 的关联规则置信度大于 50%，且至少存在 1 条关联规则置信度大于置信度阈值，则认为该频繁项集中的联合并发症为高风险联合并发症. 置信度阈值可根据并发症预测实际需求设置，置信度阈值越高，表明联合并发症关联发病率越高. 根据关联规则置信度计算结果，将高风险联合并发症数量控制在 10～15 组，故置信度阈值设置为 97%.

表 3　频繁项集的关联规则

Tab. 3　Association rules for frequent itemsets

序号	关联规则	$C\{A,B\Rightarrow C\}$/%
1	$A,B\Rightarrow C$	55
2	$A,C\Rightarrow B$	60
3	$B,C\Rightarrow A$	98
4	$A\Rightarrow B,C$	30
5	$B\Rightarrow A,C$	40
6	$C\Rightarrow A,B$	62

由表 3 可知：频繁项集$\{A,B,C\}$中可产生 6 条关联规则，其中，序号为 1,2,3,6(共 4 条，占比 2/3，大于 1/2)的关联规则置信度大于 50%，且序号为 3 的关联规则置信度为 98%(大于 97%)，则糖尿病联合并发症$\{A,B,C\}$为高风险联合并发症.

通过关联规则算法对 3 000 例糖尿病患者 33 类并发症数据进行计算，查找频繁项集与关联规则，支持度阈值设置为 5%，置信度阈值设置为 50%(两个阈值均可根据研究需要进行设置，支持度阈值越大，表明该联合并发症关联发病率越高；置信度阈值越大，表明该联合并发症发病概率越高). 通过关联规则计算符合上述参数阈值的并发症依存关系，关联规则计算的关联发病率，如表 4 所示.

表4 关联规则计算的关联发病率

Tab.4 Related incidence rate computed by association rules

序号	S/%	关联规则	C/%
1	5.00	心功能不全及心力衰竭⇒动脉粥样硬化，冠心病	70.75
		心功能不全及心力衰竭，冠心病⇒动脉粥样硬化	98.04
		心功能不全及心力衰竭，动脉粥样硬化⇒冠心病	88.76
2	5.00	其他内分泌疾病，冠心病⇒视网膜病变	58.82
3	5.04	其他内分泌疾病，高脂血⇒动脉粥样硬化	70.23
4	5.10	心功能不全及心力衰竭⇒冠心病	72.17
5	5.14	呼吸系统疾病，肾病⇒动脉粥样硬化	57.68
		呼吸系统疾病，动脉粥样硬化⇒肾病	59.92
6	5.14	下肢动脉病变，冠心病⇒动脉粥样硬化	100.00
7	5.17	脑卒中⇒动脉粥样硬化	69.20
8	5.17	视网膜病变，高脂血⇒肾病	62.25
		高脂血，肾病⇒视网膜病变	63.79
9	5.17	高脂血，脂肪肝⇒动脉粥样硬化	62.00
10	5.20	下肢动脉病变，脂肪肝⇒高血压	75.73
11	5.24	肾衰⇒视网膜病变，高血压	85.79
		视网膜病变，肾衰⇒高血压	98.74
		高血压，肾衰⇒视网膜病变	88.20
12	5.24	视网膜病变，胆道疾病⇒肾病	68.26
		胆道疾病，肾病⇒视网膜病变	71.36
13	5.27	胆道疾病，脂肪肝⇒高血压	72.48
		高血压，胆道疾病⇒脂肪肝	53.56
14	5.27	其他慢性肝病，视网膜病变⇒高血压	79.40
		其他慢性肝病，高血压⇒视网膜病变	53.02
15	5.30	肾衰⇒视网膜病变	86.89
16	5.30	肾衰⇒视网膜病变，肾病	86.89
		视网膜病变，肾衰⇒肾病	100.00
		肾病，肾衰⇒视网膜病变	86.89
17	5.34	呼吸系统疾病，冠心病⇒动脉粥样硬化	96.97
		呼吸系统疾病，动脉粥样硬化⇒冠心病	62.26
18	5.37	下肢动脉病变，脂肪肝⇒视网膜病变	78.16
19	5.40	其他内分泌疾病，高脂血⇒高血压	75.35
20	5.50	其他慢性肝病，高血压⇒动脉粥样硬化	55.37
		其他慢性肝病，动脉粥样硬化⇒高血压	78.20
21	5.64	心功能不全及心力衰竭⇒动脉粥样硬化	79.72
22	5.67	高脂血，肾病⇒动脉粥样硬化	69.96
23	5.80	视网膜病变，血液病⇒动脉粥样硬化	53.21
		动脉粥样硬化，血液病⇒视网膜病变	77.33
24	5.87	心功能不全及心力衰竭⇒高血压	83.02
25	5.87	脂肪肝，冠心病⇒高血压	83.41
26	5.94	肾衰⇒高血压，肾病	97.27
		肾病，肾衰⇒高血压	97.27
		高血压，肾衰⇒肾病	100.00
27	5.94	心肌梗死⇒动脉粥样硬化	93.68
28	5.94	心肌梗死⇒动脉粥样硬化，冠心病	93.68
		心肌梗死，冠心病⇒动脉粥样硬化	98.89
		心肌梗死，动脉粥样硬化⇒冠心病	100.00
29	5.94	高血压，胆道疾病⇒视网膜病变	60.34
		视网膜病变，胆道疾病⇒高血压	77.39
30	6.00	心肌梗死⇒冠心病	94.74
31	6.00	胆道疾病，肾病⇒高血压	81.82
		高血压，胆道疾病⇒肾病	61.02
32	6.04	高脂血，脂肪肝⇒高血压	72.40
33	6.10	肾衰⇒肾病	100.00
34	6.10	视网膜病变，高脂血⇒动脉粥样硬化	73.90
35	6.13	胆道疾病，动脉粥样硬化⇒高血压	76.35
		高血压，胆道疾病⇒动脉粥样硬化	62.37
36	6.14	下肢动脉病变，脂肪肝⇒动脉粥样硬化	89.32
37	6.17	高血压，下肢动脉病变⇒其他内分泌疾病	51.82
		其他内分泌疾病，下肢动脉病变⇒高血压	73.71
38	6.20	视网膜病变，呼吸系统疾病⇒肾病	83.78
		呼吸系统疾病，肾病⇒视网膜病变	69.66
39	6.40	视网膜病变，呼吸系统疾病⇒高血压	86.49
		高血压，呼吸系统疾病⇒视网膜病变	55.81
40	6.54	视网膜病变，下肢动脉病变⇒其他内分泌疾病	54.90
		其他内分泌疾病，下肢动脉病变⇒视网膜病变	78.09

续表
Continue table

序号	S/%	关联规则	C/%
41	6.57	肾病，血液病⇒动脉粥样硬化	52.96
		动脉粥样硬化，血液病⇒肾病	87.56
42	6.70	脑卒中⇒高血压	89.73
43	6.77	呼吸系统疾病，动脉粥样硬化⇒高血压	78.99
		高血压，呼吸系统疾病⇒动脉粥样硬化	59.01
44	6.80	高脂血，肾病⇒高血压	83.95
45	6.84	视网膜病变，高脂血⇒高血压	82.33
46	6.87	高血压，血液病⇒动脉粥样硬化	54.21
		动脉粥样硬化，血液病⇒高血压	91.56
47	6.94	脂肪肝，冠心病⇒动脉粥样硬化	98.58
48	7.00	其他慢性肝病⇒高血压，肾病	51.47
		其他慢性肝病，肾病⇒高血压	87.14
		其他慢性肝病，高血压⇒肾病	70.47
49	7.04	其他慢性肝病⇒动脉粥样硬化	51.72
50	7.07	其他内分泌疾病，冠心病⇒高血压	83.14
51	7.27	胆道疾病⇒脂肪肝	51.05
52	7.34	胆道疾病⇒肾病	51.52
53	7.47	肾病，脂肪肝⇒其他内分泌疾病	54.50
54	7.50	肾衰⇒高血压	97.27
55	7.50	血液病⇒动脉粥样硬化	50.68
56	7.54	肾病，脂肪肝⇒动脉粥样硬化	54.99
57	7.57	下肢动脉病变，动脉粥样硬化⇒其他内分泌疾病	51.71
		其他内分泌疾病，下肢动脉病变⇒动脉粥样硬化	90.44
58	7.60	呼吸系统疾病，肾病⇒高血压	85.39
		高血压，呼吸系统疾病⇒肾病	66.28
59	7.60	下肢动脉病变，肾病⇒视网膜病变	83.52
		视网膜病变，下肢动脉病变⇒肾病	63.87
60	7.67	胆道疾病⇒视网膜病变	53.86
61	7.80	高血压，下肢动脉病变⇒肾病	65.55
		下肢动脉病变，肾病⇒高血压	85.71
62	7.90	视网膜病变，冠心病⇒肾病	63.37
		肾病，冠心病⇒视网膜病变	65.47
63	8.04	其他慢性肝病⇒肾病	59.07
64	8.04	胆道疾病⇒动脉粥样硬化	56.44
65	8.30	下肢动脉病变⇒肾病，动脉粥样硬化	52.31
		肾病，下肢动脉病变⇒动脉粥样硬化	91.21
		下肢动脉病变，动脉粥样硬化⇒肾病	56.72
66	8.34	其他内分泌疾病，冠心病⇒动脉粥样硬化	98.04
67	8.37	下肢动脉病变⇒其他内分泌疾病	52.73
68	8.57	呼吸系统疾病⇒动脉粥样硬化	54.45
69	8.87	下肢动脉病变⇒视网膜病变，高血压	55.88
		视网膜病变，下肢动脉病变⇒高血压	74.51
		高血压，下肢动脉病变⇒视网膜病变	74.51
70	8.90	呼吸系统疾病⇒肾病	56.57
71	9.04	视网膜病变，脂肪肝⇒其他内分泌疾病	54.20
		其他内分泌疾病，脂肪肝⇒视网膜病变	54.31
72	9.10	下肢动脉病变⇒肾病	57.35
73	9.17	其他内分泌疾病，脂肪肝⇒动脉粥样硬化	55.11
		动脉粥样硬化，脂肪肝⇒其他内分泌疾病	56.12
74	9.34	视网膜病变，脂肪肝⇒动脉粥样硬化	56.00
75	9.34	动脉粥样硬化，脂肪肝⇒视网膜病变	57.14
76	9.37	视网膜病变，脂肪肝⇒肾病	56.20
		肾病，脂肪肝⇒视网膜病变	68.37
77	9.77	高血压，高脂血⇒冠心病	59.43
		高脂血，冠心病⇒高血压	82.30
78	9.84	胆道疾病⇒高血压	69.09
79	9.90	血液病⇒视网膜病变，高血压	66.89
		视网膜病变，血液病⇒高血压	90.83
		高血压，血液病⇒视网膜病变	78.16
80	9.94	其他慢性肝病⇒高血压	73.04
81	10.30	血液病⇒视网膜病变，肾病	69.59
		视网膜病变，血液病⇒肾病	94.50
82	10.30	肾病，血液病⇒视网膜病变	83.06
83	10.44	其他内分泌疾病，肾病⇒动脉粥样硬化	58.18
		其他内分泌疾病，动脉粥样硬化⇒肾病	56.40
84	10.70	下肢动脉病变⇒视网膜病变，动脉粥样硬化	67.44
		视网膜病变，下肢动脉病变⇒动脉粥样硬化	89.92
		下肢动脉病变，动脉粥样硬化⇒视网膜病变	73.12

续表
Continue table

序号	S/%	关联规则	C/%
85	10.80	视网膜病变，冠心病⇒高血压	86.63
86	10.90	血液病⇒视网膜病变	73.65
87	10.97	肾病，冠心病⇒高血压	90.88
88	11.10	高血压，脂肪肝⇒肾病	53.28
		肾病，脂肪肝⇒高血压	81.02
89	11.10	高血压，下肢动脉病变⇒动脉粥样硬化	93.28
		下肢动脉病变⇒高血压，动脉粥样硬化	69.96
		下肢动脉病变，动脉粥样硬化⇒高血压	75.85
90	11.24	高血压，脂肪肝⇒其他内分泌疾病	53.92
		其他内分泌疾病，脂肪肝⇒高血压	67.54
91	11.47	呼吸系统疾病⇒高血压	72.88
92	11.47	血液病⇒高血压，肾病	77.48
		肾病，血液病⇒高血压	92.47
		高血压，血液病⇒肾病	90.53
93	11.57	高脂血⇒动脉粥样硬化，冠心病	52.90
		高脂血，冠心病⇒动脉粥样硬化	97.47
		高脂血，动脉粥样硬化⇒冠心病	73.21
94	11.74	其他内分泌疾病，视网膜病变⇒动脉粥样硬化	58.28
		其他内分泌疾病，动脉粥样硬化⇒视网膜病变	63.42
95	11.80	肾病，冠心病⇒动脉粥样硬化	97.79
		肾病，动脉粥样硬化⇒冠心病	52.52
96	11.87	高脂血⇒冠心病	54.27
97	11.90	下肢动脉病变⇒高血压	75.00
98	11.90	下肢动脉病变⇒视网膜病变	75.00
99	11.94	视网膜病变，脂肪肝⇒高血压	71.60
		高血压，脂肪肝⇒视网膜病变	57.28
100	12.14	视网膜病变，冠心病⇒动脉粥样硬化	97.33
101	12.20	高血压，脂肪肝⇒动脉粥样硬化	58.56
		动脉粥样硬化，脂肪肝⇒高血压	74.69
102	12.40	血液病⇒肾病	83.78
103	12.64	高脂血⇒高血压，动脉粥样硬化	57.77
		高血压，高脂血⇒动脉粥样硬化	76.88
		高脂血，动脉粥样硬化⇒高血压	79.96
104	12.67	血液病⇒高血压	85.59
105	12.97	其他内分泌疾病，视网膜病变⇒肾病	64.40
		其他内分泌疾病，肾病⇒视网膜病变	72.30
106	13.70	其他内分泌疾病，高血压⇒动脉粥样硬化	59.31
		其他内分泌疾病，动脉粥样硬化⇒高血压	74.05
107	14.64	下肢动脉病变⇒动脉粥样硬化	92.23
108	14.64	其他内分泌疾病，肾病⇒高血压	81.60
		其他内分泌疾病，高血压⇒肾病	63.35
109	14.77	其他内分泌疾病，视网膜病变⇒高血压	73.34
		其他内分泌疾病，高血压⇒视网膜病变	63.92
110	15.81	高脂血⇒动脉粥样硬化	72.26
111	16.34	脂肪肝⇒动脉粥样硬化	52.29
112	16.34	视网膜病变，肾病⇒动脉粥样硬化	54.26
		视网膜病变，动脉粥样硬化⇒肾病	65.16
		肾病，动脉粥样硬化⇒视网膜病变	72.70
113	16.44	高脂血⇒高血压	75.15
114	16.64	脂肪肝⇒其他内分泌疾病	53.26
115	16.67	脂肪肝⇒视网膜病变	53.36
116	17.94	其他内分泌疾病⇒肾病	53.69
117	18.51	其他内分泌疾病⇒动脉粥样硬化	55.39
118	19.44	肾病，动脉粥样硬化⇒高血压	86.50
		高血压，肾病⇒动脉粥样硬化	54.69
119	19.81	视网膜病变，高血压⇒动脉粥样硬化	54.35
		视网膜病变，动脉粥样硬化⇒高血压	78.99
		高血压，动脉粥样硬化⇒视网膜病变	50.73
120	20.14	其他内分泌疾病⇒视网膜病变	60.28
121	20.84	脂肪肝⇒高血压	66.70
122	22.47	肾病⇒动脉粥样硬化	52.78
123	23.11	其他内分泌疾病⇒高血压	69.16
124	25.08	视网膜病变⇒动脉粥样硬化	50.13
125	25.71	冠心病⇒高血压，动脉粥样硬化	78.27
		高血压，冠心病⇒动脉粥样硬化	97.35
		高血压，动脉粥样硬化⇒冠心病	65.84
		动脉粥样硬化，冠心病⇒高血压	80.65

续表
Continue table

序号	S/%	关联规则	C/%	序号	S/%	关联规则	C/%
126	25.98	视网膜病变⇒高血压，肾病	51.93	129	31.88	动脉粥样硬化⇒冠心病	61.96
		肾病⇒视网膜病变，高血压	61.00			冠心病⇒动脉粥样硬化	97.06
		视网膜病变，肾病⇒高血压	86.27	130	35.55	高血压⇒肾病	52.10
		视网膜病变，高血压⇒肾病	71.27			肾病⇒高血压	83.48
		高血压，肾病⇒视网膜病变	73.08	131	36.45	视网膜病变⇒高血压	72.87
127	26.41	冠心病⇒高血压	80.41			高血压⇒视网膜病变	53.42
128	30.11	视网膜病变⇒肾病	60.20	132	39.05	高血压⇒动脉粥样硬化	57.23
		肾病⇒视网膜病变	70.71			动脉粥样硬化⇒高血压	75.89

表 4 中：各关联规则置信度为该关联规则中的关联发病率，以序号 132 的关联规则为例，若糖尿病患者患有高血压，则有 57.23%的概率同时患有动脉粥样硬化；若糖尿病患者患有动脉粥样硬化，则有 75.89%的概率同时患有高血压.

算法结果符合动脉粥样硬化和高血压发病的病理学基础，因此，可参照该关联发病率计算结果，根据糖尿病患者患有并发症情况及时筛查是否同时患有其他并发症，达到尽早诊断治疗的预警目标.

根据高风险联合并发症筛选规则，结合表 4 中各关联规则置信度，筛选出 12 组符合条件的联合并发症. 高风险联合并发症，如表 5 所示.

表 5　高风险联合并发症

Tab. 5　High-risk combined complications

组号	联合并发症	组号	联合并发症
1	动脉粥样硬化＋冠心病＋心功能不全及心力衰竭	7	高血压＋动脉粥样硬化＋冠心病
2	视网膜病变＋高血压＋肾衰	8	视网膜病变＋肾病
3	视网膜病变＋肾病＋肾衰	9	动脉粥样硬化＋冠心病
4	高血压＋肾衰＋肾病	10	高血压＋肾病
5	动脉粥样硬化＋心肌梗死＋冠心病	11	动脉粥样硬化＋高血压
6	动脉粥样硬化＋冠心病＋高脂血	12	视网膜病变＋高血压＋肾病

建立上述 12 组高风险联合并发症的发病预测模型，查找各联合并发症关键影响因素（生化指标），可为联合并发症的诊断预测提供参考. 12 组高风险联合并发症和当前临床研究结论保持一致，如动脉粥样硬化是心血管疾病、慢性肾病等多种疾病的病理学基础，而和其组合的 5 组联合并发症中，也多为心血管疾病等；第 8 组高风险联合并发症也得到当前研究的不断证实.

3　发病预测模型的构建

3.1　随机森林模型的构建

以联合并发症是否发病作为类标签，构建由 100 棵决策树组成的随机森林，每棵决策树分别对分类结果进行投票，最终获得票数最多的结果为随机森林最终分类结果，并输出该联合并发症生化指标重要性排序，12 组高风险联合并发症均按照上述流程完成各自随机森林模型的构建.

将除糖尿病并发症数据外其他指标数据作为影响因素，高风险联合并发症共同发病结果作为最终分类结果纳入模型中（若同时患有该组合所有并发症，则类标号为 1，否则为 0），即可得到该糖尿病患者是否患有该类型联合并发症的预测结果.

采用十折交叉验证法判断随机森林模型的分类预测精度，即将原始糖尿病数据集随机分为 10 等份，其中，9 份作为训练集，用于训练随机森林模型；剩余 1 份作为测试集，用于测试随机森林模型的分类精度，并进行 10 次迭代，训练期间，每 1 份数据集都作为测试集对随机森林模型的分类预测精度进行测试. 随机森林模型训练过程中，根据数据集中各分类中不同数据量对不同分类赋予不同权重，确保数据均衡.

根据上述方法对数据建模，随机森林模型中重要性排名前 10 的生化指标，如表 6 所示.

若表 6 中重要性排名靠前的生化指标值异常，则对其他重要性排名靠前指标进行检测. 通过随机森

表6 随机森林模型中重要性排名前10的生化指标

Tab. 6 Top 10 important biochemical indexes in random forest model

第1组生化指标	重要性	第2组生化指标	重要性	第3组生化指标	重要性	第4组生化指标	重要性	第5组生化指标	重要性	第6组生化指标	重要性
AGE	0.085 8	SCR	0.143 1	SCR	0.158 3	SCR	0.160 2	AST	0.041 6	CP	0.056 1
GLO	0.046 9	BU	0.116 3	BU	0.116 4	BU	0.109 7	CP	0.041 2	CRP	0.040 8
HDL_C	0.034 0	UPR_24	0.100 4	UPR_24	0.101 7	PCV	0.080 6	ALB_CR	0.040 7	UPR_24	0.037 1
TC	0.033 9	PCV	0.063 8	PCV	0.057 0	HB	0.063 8	TC	0.039 7	GGT	0.035 6
LPS	0.031 9	UCR	0.042 3	UCR	0.042 9	ALB	0.046 8	LPS	0.039 3	ALB_CR	0.033 1
CP	0.031 6	ALB_CR	0.036 6	ALB_CR	0.035 8	HBA1C	0.041 3	AGE	0.039 3	LPS	0.032 5
GLU	0.029 8	ESR	0.033 7	IBILI	0.032 7	IBILI	0.035 7	LDL_C	0.037 1	ALB	0.032 3
LDL_C	0.028 7	IBILI	0.030 8	ESR	0.031 3	ALT	0.028 6	SCR	0.037 1	HB	0.031 1
INS	0.028 3	ALB	0.030 2	HB	0.029 4	ALB_CR	0.027 0	HDL_C	0.032 5	PLT	0.028 8
PT	0.027 4	HB	0.029 3	ALB	0.029 4	TBILI	0.025 6	GLU	0.031 7	PCV	0.028 3
第7组生化指标	重要性	第8组生化指标	重要性	第9组生化指标	重要性	第10组生化指标	重要性	第11组生化指标	重要性	第12组生化指标	重要性
CP	0.054 9	UPR_24	0.122 9	CP	0.058 0	HBA1C	0.058 7	CP	0.053 7	UPR_24	0.115 9
AGE	0.048 3	ALB_CR	0.108 8	ALB_CR	0.047 7	SCR	0.038 6	ALB_CR	0.044 1	ALB_CR	0.098 8
ALB_CR	0.045 0	SCR	0.081 4	UPR_24	0.045 8	GSP	0.036 4	AGE	0.036 9	SCR	0.091 7
UPR_24	0.037 2	UCR	0.047 0	AGE	0.041 6	SUA	0.032 7	UPR_24	0.035 5	BU	0.045 3
TP	0.035 3	BU	0.046 4	CRP	0.034 9	GLU	0.031 3	TP	0.032 7	UCR	0.041 7
CRP	0.033 1	ALB	0.036 7	LPS	0.032 1	BP_LOW	0.030 1	HBA1C	0.030 4	ALB	0.040 9
TC	0.028 7	PCV	0.024 9	TP	0.030 2	BU	0.030 1	ALB	0.030 1	PCV	0.029 8
INS	0.027 5	CP	0.024 3	SCR	0.029 4	BP_HIGH	0.029 0	CRP	0.027 8	BP_HIGH	0.027 9
ALB	0.026 9	FBG	0.023 9	INS	0.029 0	AGE	0.028 6	PCV	0.025 6	FBG	0.024 5
LPS	0.025 5	TP	0.020 8	ALB	0.025 6	CP	0.028 3	BP_HIGH	0.025 2	SUA	0.022 9

林模型预测该患者是否会患该联合并发症,若随机森林模型判断该患者会患该联合并发症,则需做进一步详细检查,从而确诊该联合并发症是否发病;若随机森林模型判断该患者不会患该联合并发症,则结合关联发病率计算结果,判断其患有其他并发症的风险,并加以预防. 模型预测分析过程,如图3所示.

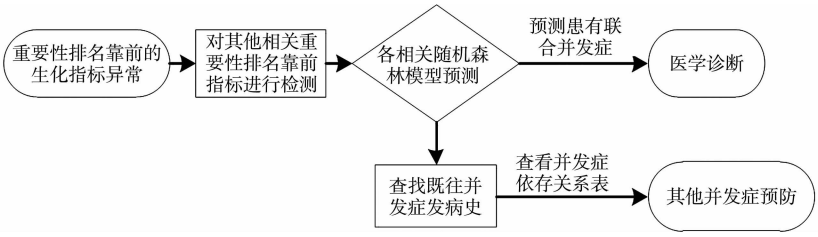


图3 随机森林模型预测分析过程

Fig. 3 Predictive analysis process of random forest model

3.2 预测效果分析

采用高风险联合并发症发病预测的精度和受试者工作(ROC)曲线,对模型的预测效果进行评估,随机森林模型对各高风险联合并发症十折交叉验证法的分类精度,如表7所示.

表7 各高风险联合并发症十折交叉验证法的分类精度

Tab. 7 Classification accuracy of high-risk combined complication in ten fold cross validation method

组号	分类精度										平均值
	1	2	3	4	5	6	7	8	9	10	
1	0.973 3	0.973 3	0.973 3	0.973 3	0.973 3	0.973 3	0.970 0	0.970 0	0.970 0	0.970 0	0.972 0
2	0.943 3	0.933 3	0.950 0	0.936 7	0.933 3	0.950 0	0.950 0	0.956 7	0.950 0	0.946 7	0.945 0
3	0.940 0	0.926 7	0.950 0	0.940 0	0.936 7	0.950 0	0.950 0	0.953 3	0.950 0	0.943 3	0.944 0
4	0.993 3	0.993 3	0.993 3	0.993 3	0.993 3	0.993 3	0.993 3	0.993 3	0.993 3	0.990 0	0.993 0
5	0.963 3	0.963 3	0.963 3	0.963 3	0.963 3	0.960 0	0.960 0	0.960 0	0.960 0	0.960 0	0.961 7

续表

Continue table

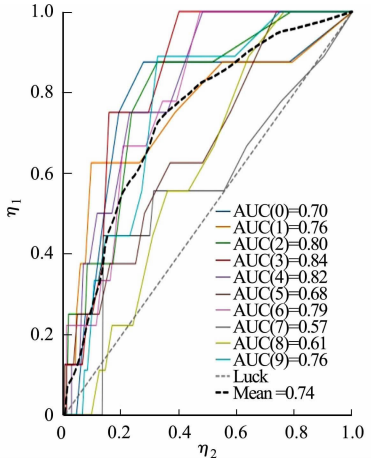
组号	分类精度										平均值
	1	2	3	4	5	6	7	8	9	10	
6	0.923 3	0.923 3	0.923 3	0.920 0	0.923 3	0.923 3	0.920 0	0.903 3	0.920 0	0.920 0	0.920 0
7	0.850 0	0.856 7	0.856 7	0.856 7	0.856 7	0.866 7	0.863 3	0.843 3	0.846 7	0.853 3	0.855 0
8	0.876 7	0.890 0	0.846 7	0.853 3	0.886 7	0.813 3	0.840 0	0.843 3	0.843 3	0.860 0	0.855 3
9	0.813 3	0.833 3	0.816 7	0.823 3	0.840 0	0.840 0	0.830 0	0.800 0	0.806 7	0.806 7	0.821 0
10	0.906 7	0.906 7	0.906 7	0.903 3	0.903 3	0.903 3	0.903 3	0.903 3	0.903 3	0.903 3	0.904 3
11	0.816 7	0.816 7	0.816 7	0.813 3	0.816 7	0.813 3	0.816 7	0.810 0	0.803 3	0.790 0	0.811 3
12	0.856 7	0.860 0	0.860 0	0.883 3	0.896 7	0.836 7	0.866 7	0.846 7	0.853 3	0.856 7	0.861 7

由表 7 可知：随机森林模型对各高风险联合并发症的分类精度大部分超过 0.900 0，对各高风险联合并发症的分类平均精度均在 0.800 0 以上.

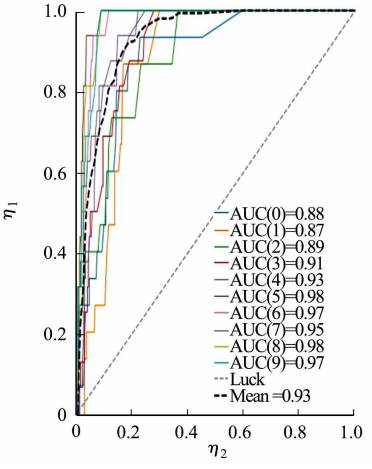
通过 ROC 曲线对模型进行评估时，ROC 曲线下面积(AUC)越接近于 1，则随机森林模型正确分类正预测的能力越强，假阳性的概率越低. 12 组高风险联合并发症的 ROC 曲线，如图 4 所示. 图 4 中：$\eta 1$ 为假阳性率；$\eta 2$ 为真阳性率；AUC(n)为 n 折交叉验证法的曲线下面积；Mean 为 AUC 的平均值；Luck 为对角线.

由图 4 可知：曲线基本位于 45°线的左上方，表明经十折交叉验证法验证后，各高风险联合并发症发病预测模型的 AUC 均大于 0.50，AUC 均值均大于 0.67，故大部分高风险联合并发症发病预测模型具有较好的发病预测效果. 联合并发症的生化指标重要性排名可为疾病的诊断和风险预测提供重要参考.

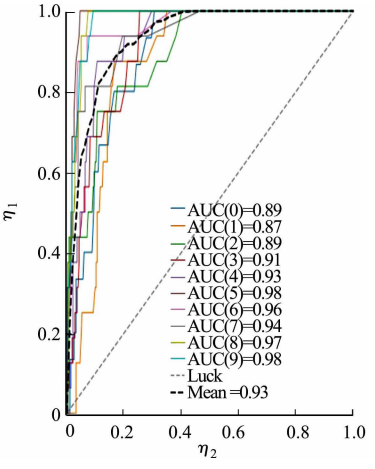
为进一步验证各高风险联合并发症发病预测模型在糖尿病患者发病预测应用的有效性，随机选取



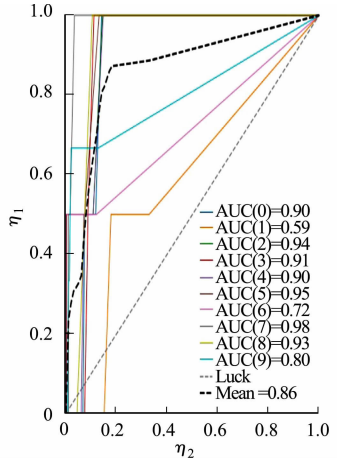
(a) 第 1 组



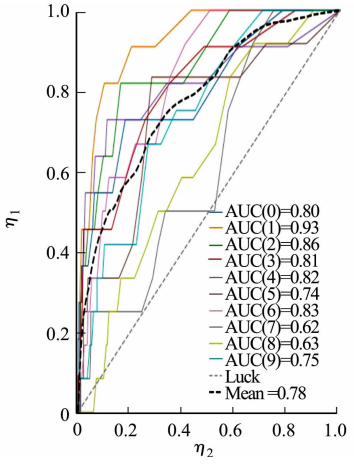
(b) 第 2 组



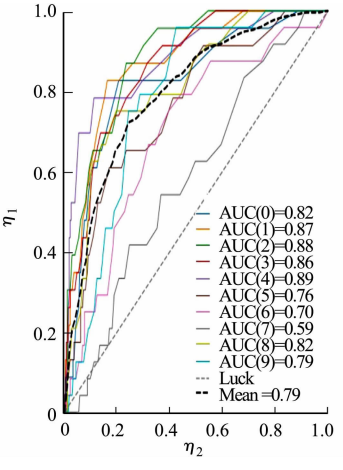
(c) 第 3 组



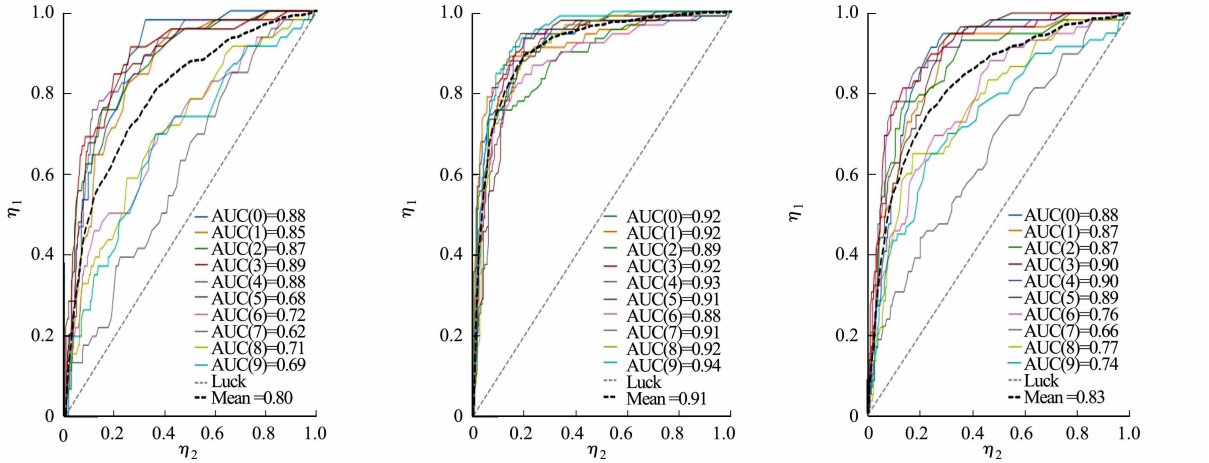
(d) 第 4 组



(e) 第 5 组



(f) 第 6 组



(g) 第7组 (h) 第8组 (i) 第9组

(j) 第10组 (k) 第11组 (l) 第12组

图4 12组高风险联合并发症的ROC曲线

Fig. 4 ROC curves of 12 groups high-risk combined complications

120例糖尿病患者数据(联合并发症患者数据在各联合并发症数据中随机选取,非联合并发症患者数据在非联合并发症数据中随机选取),按照并发症分为12组,每组包含10组数据,均为5组未患病数据与5组患病数据,将数据分别输入12组对应的高风险联合并发症发病预测模型中进行发病预测,随机森林模型对糖尿病患者并发症预测结果,如表8所示.

表8 随机森林模型对糖尿病患者并发症预测结果

Tab. 8 Prediction results of random forest model of diabetic complications

组号	结果	患者1	患者2	患者3	患者4	患者5	患者6	患者7	患者8	患者9	患者10
1	预测结果	患病	患病	患病	患病	患病	未患病	未患病	未患病	未患病	未患病
	实际情况	患病	患病	患病	患病	患病	未患病	未患病	未患病	未患病	未患病
	效果评估	√	√	√	√	√	√	√	√	√	√
2	预测结果	未患病	未患病	未患病	未患病	未患病	患病	患病	患病	患病	患病
	实际情况	未患病	未患病	未患病	未患病	未患病	患病	患病	患病	患病	患病
	效果评估	√	√	√	√	√	√	√	√	√	√
3	预测结果	患病	患病	患病	患病	患病	未患病	未患病	未患病	未患病	未患病
	实际情况	患病	患病	患病	患病	患病	未患病	未患病	未患病	未患病	未患病
	效果评估	√	√	√	√	√	√	√	√	√	√
4	预测结果	患病	患病	患病	未患病	患病	未患病	未患病	未患病	未患病	未患病
	实际情况	患病	患病	患病	患病	患病	未患病	未患病	未患病	未患病	未患病
	效果评估	√	√	√	×	√	√	√	√	√	√

续表
Continue table

组号	结果	患者 1	患者 2	患者 3	患者 4	患者 5	患者 6	患者 7	患者 8	患者 9	患者 10
5	预测结果	患病	患病	患病	患病	患病	未患病	未患病	未患病	未患病	未患病
	实际情况	患病	患病	患病	患病	患病	未患病	未患病	未患病	未患病	未患病
	效果评估	√	√	√	√	√	√	√	√	√	√
6	预测结果	患病	患病	患病	患病	患病	未患病	未患病	未患病	未患病	未患病
	实际情况	患病	患病	患病	患病	患病	未患病	未患病	未患病	未患病	未患病
	效果评估	√	√	√	√	√	√	√	√	√	√
7	预测结果	患病	患病	患病	未患病	患病	未患病	未患病	未患病	未患病	未患病
	实际情况	患病	患病	患病	患病	患病	未患病	未患病	未患病	未患病	未患病
	效果评估	√	√	√	×	√	√	√	√	√	√
8	预测结果	患病	患病	患病	患病	患病	未患病	未患病	未患病	未患病	未患病
	实际情况	患病	患病	患病	患病	患病	未患病	未患病	未患病	未患病	未患病
	效果评估	√	√	√	√	√	√	√	√	√	√
9	预测结果	患病	患病	患病	患病	患病	未患病	未患病	未患病	未患病	未患病
	实际情况	患病	患病	患病	患病	患病	未患病	未患病	未患病	未患病	未患病
	效果评估	√	√	√	√	√	√	√	√	√	√
10	预测结果	患病	患病	患病	患病	患病	未患病	未患病	未患病	未患病	未患病
	实际情况	患病	患病	患病	患病	患病	未患病	未患病	未患病	未患病	未患病
	效果评估	√	√	√	√	√	√	√	√	√	√
11	预测结果	患病	患病	患病	患病	患病	未患病	未患病	未患病	未患病	未患病
	实际情况	患病	患病	患病	患病	患病	未患病	未患病	未患病	未患病	未患病
	效果评估	√	√	√	√	√	√	√	√	√	√
12	预测结果	患病	患病	患病	患病	患病	未患病	未患病	未患病	未患病	未患病
	实际情况	患病	患病	患病	患病	患病	未患病	未患病	未患病	未患病	未患病
	效果评估	√	√	√	√	√	√	√	√	√	√

由表 8 可知：12 组高风险联合并发症发病预测模型对并发症发病预测效果较好，只有 2 例患者未成功预测发病风险，其他 118 例患者患病/未患病均正确预测，某种程度上可作为糖尿病联合并发症的预诊断参考.

4　结论

1）十折交叉验证法和 ROC 曲线对模型的评估结果表明，基于随机森林模型的高风险联合并发症发病预测模型具有较好的分类预测精度和分类效果.

2）高血压、视网膜病变、动脉粥样硬化、肾病等是糖尿病并发症中关联发病率较高且是联合发病风险最高的并发症种类，其患有某两种并发症后其他并发症关联发病率超过 97%，提示上述糖尿病患者及早诊断及预防.

3）不同高风险联合并发症发病预测模型的分类平均精度均在 0.800 0 以上，部分模型的 AUC 在 0.900 0 以上，但仍有部分模型的 AUC 未达到 0.70，需要在后续的研究中进一步探索，以提升模型的预测效果.

参考文献：

[1] 张争辉，薛爱芹，于兰. 糖尿病相关研究进展[J]. 世界最新医学信息文摘，2019，19(20)：145，149. DOI：10.19613/j.cnki.1671-3141.2019.20.067.

[2] TABAEI B P，HERMAN W H. A multivariate logistic regression equation to screen for diabetes：Development and validation[J]. Diabetes Care，2002，25(11)：1999-2003. DOI：10.2337/diacare.25.11.1999.

[3] FATIMA M，PASHA M. Survey of machine learning algorithms for disease diagnostic[J]. Journal of Intelligent Learning Systems and Applications，2017，9(1)：1-16. DOI：10.4236/jilsa.2017.91001.

[4] SOWJANYA K，SINGHAL A，CHOUDHARY C. MobDBTest：A machine learning based system for predicting di-

abetes risk using mobile devices[C]// IEEE International Advance Computing Conference. Banglore: IEEE Press, 2015:397-402. DOI:10.1109/IADCC.2015.7154738.

[5] 谭昭,李文歌.2型糖尿病患者血清尿酸及尿微量白蛋白水平与慢性血管并发症的相关性[J].中国医科大学学报,2018,47(1):67-72. DOI:10.12007/j.issn.0258-4646.2018.01.015.

[6] 李晓燕,孟凡杰,段玉龙,等.改良早期预警评分、血糖值评分及两评分结合预测糖尿病急性并发症患者预后能力的对比研究[J].实用医学杂志,2018,34(3):397-400. DOI:10.3969/j.issn.1006-5725.2018.03.014.

[7] 王雷.经颅多普勒微栓子监测在糖尿病脑血管病中的应用效果及对并发症的预测价值[J].中国医药科学,2020,10(5):201-203,214. DOI:10.3969/j.issn.2095-0616.2020.05.058.

[8] 邢美艳,姜天,夏莉,等.皮肤无创晚期糖基化终末产物测定在社区2型糖尿病血管性并发症筛查中的作用研究[J].中国全科医学,2020,23(8):913-919. DOI:10.12114/j.issn.1007-9572.2020.00.039.

[9] CEDERHOLM J, KATARINA E O, ELIASSON B, *et al*. Risk prediction of cardiovascular disease in type 2 diabetes: A risk equation from the Swedish National diabetes register[J]. Diabetes Care, 2008, 31(10): 2038-2043. DOI: 10.2337/dc08-0662.

[10] 张振堂,杨洋,韩福俊,等.基于社区2型糖尿病患者的心脑血管事件5年风险预测模型[J].山东大学学报(医学版),2017,55(6):108-113. DOI:10.6040/j.issn.1671-7554.0.2017.341.

[11] 明淑萍,刘玲,周黎,等.糖尿病急性并发症继发轻度认知功能障碍的预测模型及时间窗分析[J].中风与神经疾病杂志,2017,34(9):786-791. DOI:10.19845/j.cnki.zfysjjbzz.2017.09.004.

[12] 王洁,乔艺璇,彭岩,等.基于Logistic回归和多层神经网络的Ⅱ型糖尿病并发症预测[J].高技术通讯,2019,29(5):455-461. DOI:10.3772/j.issn.1002-0470.2019.05.006.

[13] 王冰蓉,孙阳,李益颖,等.糖化血红蛋白联合非传统血糖监测指标对妊娠期糖尿病患者急慢性并发症的预测价值[J].创伤与急诊电子杂志,2019,7(1):22-28. DOI:10.16746/j.cnki.11-9332/r.2019.01.005.

[14] 徐晓,张蕾,王珍.基于模糊综合评价法的脑中风风险预测系统[J].计算机仿真,2015,32(7):344-347,360. DOI:10.3969/j.issn.1006-9348.2015.07.076.

[15] 李攀.基于神经网络的2型糖尿病并发症预测模型的研究[D].广州:中医药大学,2016.

[16] 崔纯纯.基于神经网络的糖尿病并发症预测系统研究[D].北京:北京交通大学,2018.

[17] VIJIYAKUMAR K, LAVANYA B, NIRMALA I, *et al*. Random forest algorithm for the prediction of diabetes [C]// IEEE International Conference on System Computation, Automation and Networking. Pondicherry: IEEE Press, 2019:1-5. DOI:10.1109/ICSCAN.2019.8878802.

[18] 邱云飞,郭蕾.面向非均衡数据的糖尿病并发症预测[J].数据分析与知识发现,2021,5(2):13. DOI:10.11925/infotech.2096-3467.2020.0353.

[19] HAN Jiawei, KAMBER M, PEI Jian. Data mining: Concept and techniques[M]. Beijing: China Machine Press, 2012.

[20] 陈纪林.防治动脉粥样硬化的新动向[J].中国循环杂志,2001,16(3):163.

(**责任编辑**:陈志贤 **英文审校**:吴逢铁)