第 62 卷　第 2 期
2026 年 3 月

南京大学学报(自然科学)
JOURNAL OF NANJING UNIVERSITY
(NATURAL SCIENCE)

Vol. 62, No. 2
Mar., 2026

DOI:10.13232/j.cnki.jnju.2026.02.011

基于可解释多任务学习模型揭示糖尿病联合并发症的关键特征及预测建模

王志轩,罗冬梅*

(安徽工业大学微电子与数据科学学院,马鞍山,243002)

摘　要:糖尿病并发症是引起糖尿病患者死亡的重要因素,揭示并发症的关键特征能有效地帮助医生制定针对性干预策略,从而降低糖尿病患者并发症状况下的死亡风险.然而,既往研究大多集中在识别糖尿病单一并发症的风险因素上,忽略了并发症之间的潜在关联,因此,基于国家人口健康科学数据中心提供的糖尿病并发症预警数据集,采用皮尔逊相关系数和卡方检验筛选出显著相关的糖尿病并发症,并将其纳入多任务学习模型中进行联合建模.接着使用SHAP(SHapley Additive exPlanations)评估各特征的重要性,筛选出SHAP的值高于75%分位数的11个特征作为糖尿病联合并发症的重要风险因素.基于随机森林、逻辑回归、梯度提升模型、极限梯度提升模型、自适应增强算法以及类别特征梯度提升模型构建糖尿病联合并发症预测模型,输入变量为SHAP的值高于25%分位数的特征,结合网格搜索选择最优参数组合,并通过准确率、精确率、$F1$-score、AUC等指标评估模型的预测性能.结果表明,采用可解释的多任务学习模型筛选出来的特征是关键特征,六种预测模型的AUC均接近0.90.最后引入LIME(Local Interpretable Model-Agnostic Explanations)对模型进行解释,进一步验证所构建的可解释多任务学习模型筛选关键特征的有效性与可靠性.可解释多任务学习模型充分考虑了并发症之间的潜在关系,能够准确地识别糖尿病联合并发症的关键风险因素,辅助医生制定针对性干预策略,有助于减少患者因并发症导致的死亡.

关键词:糖尿病,可解释技术,多任务学习,联合并发症

中图分类号:R318,R587.1　　**文献标志码**:A

Interpretable multitask learning-based model reveals key features and predictive modelling of joint complications in diabetes mellitus

Wang Zhixuan, Luo Dongmei*

(School of Microelectronics and Data Science, Anhui University of Technology, Maanshan, 243002, China)

Abstract: Complications of diabetes mellitus are important factors in patient mortality, and revealing their key features can effectively help physicians develop targeted intervention strategies to reduce the risk of death in comorbid conditions. However, most previous studies have focused on identifying risk factors for a single complication of diabetes, ignoring potential associations between complications. Therefore, based on the Diabetes Complications Early Warning Dataset provided by the National Population Health Sciences Data Centre, we used Pearson′s correlation coefficient and the chi-square test to screen out significantly associated diabetic complications and incorporated them into a multi-task learning model for joint modeling. Then the importance of each feature was assessed using SHAP (SHapley Additive exPlanations), and 11 features with SHAP values higher than the 75% quartile were screened as significant risk factors for diabetes co-morbidities. A predictive model for diabetes-related complications was constructed using random forest, logistic regression, gradient

基金项目:安徽省教育教学改革研究项目(2024sx047),安徽省高校自然科学基金重点研究项目(2022AH050328)

收稿日期:2026－01－26

* 通信联系人,E-mail:luodmahut@126.com

boosting, extreme gradient boosting, adaptive boosting, and categorical feature gradient boosting. Input variables comprised features with SHAP values exceeding the 25th percentile. Optimal parameter combinations were selected via grid search, with model predictive performance evaluated using metrics including accuracy, precision, $F1$-score, and AUC. Results indicated that features selected through the interpretable multi-task learning model constituted key predictors, with all six predictive models achieving AUC values approaching 0.90. Finally, LIME (Local Interpretable Model-Agnostic Explanations) was introduced to interpret the model outcomes, thereby further validating the effectiveness and reliability of the constructed interpretable multi-task learning model for screening key features. The interpretable multi-task learning model comprehensively accounts for the underlying relationships between complications, enabling the precise identification of key risk factors for concurrent diabetic complications. This assists clinicians in formulating targeted intervention strategies, thereby helping to reduce patient mortality attributable to complications.

Keywords: diabetes mellitus, interpretable technique, multi-task learning, joint complications

糖尿病是一种由于胰岛素分泌不足或胰岛素无法成功作用,或二者兼有而导致高血糖的代谢性疾病,它会潜移默化地影响患者的眼、肾等器官,常常引起视网膜病变、肾病和神经病变[1],大大增加了糖尿病患者的死亡率[2].

识别潜在风险因素是预防和控制糖尿病及其并发症的关键环节.以往研究中,多数研究人员采用机器学习算法探索糖尿病及其并发症的风险因素,并构建预测模型[3-4]. 例如,Xie et al[5]利用多种机器学习算法建立了Ⅱ型糖尿病患者的预测模型,发现睡眠不足(每日少于 6 h)和睡眠过多(每日多于 9 h)均会增加患Ⅱ型糖尿病的风险. Schallmoser et al[6]基于逻辑回归(Logistic Regression, LR)与梯度提升树(Gradient Boosting, GB)两种机器学习方法,分析了糖尿病患者微血管及大血管并发症的风险因素,结果显示血糖水平、糖化血红蛋白和血清肌酐水平的升高是微血管并发症的重要致病因素,而年龄和高血压则是大血管并发症的主要影响因素. Cui et al[7]使用支持向量机(Support Vector Machine, SVM)和随机森林(Random Forest, RF),通过分类与回归分析对肌肉减少并发症进行风险评估,发现年龄、性别和BMI等指标对模型有显著影响. Li et al[8]通过极限梯度提升模型(eXtreme Gradient Boosting, XGBoost),SVM 和 RF 等算法构建糖尿病视网膜病变(Diabetic Retinopathy, DR)的预测模型,发现肾病史、血肌酐值大于 100 $\mu mol \cdot L^{-1}$等因素与DR风险上升相关,而年龄超过 65 岁则与 DR 风险下降有关. Lian et al[9]运用朴素贝叶斯和决策树等机器学习方法识别糖尿病神经病变(Diabetic Peripheral Neuropathy, DPN)的风险因素,得出糖化血红蛋白、胰岛素抵抗指数、尿蛋白浓度等均为DPN的显著正相关因素,而总胆固醇、肌酐水平与DPN风险呈负相关.尽管上述研究在识别糖尿病并发症风险因素方面取得了重要进展,但它们普遍忽略了不同并发症之间的潜在关联,可能导致新疗法在临床试验中效果不佳.

为了考虑糖尿病并发症之间的潜在关联性,本研究首先采用皮尔逊相关系数(Pearson Correlation Coefficient, PCC)和卡方检验方法,筛选出具有显著相关性的联合并发症,将它们纳入多任务学习模型进行预测建模.随后,利用沙普利加和解释(SHapley Additive exPlanations, SHAP)可解释性技术评估模型中各特征的重要性,并选取SHAP值高于75%分位数的特征作为糖尿病联合并发症的重要风险因素.为了进一步验证所筛选特征的有效性,将数据集按7∶3的比例划分为训练集和测试集,并将SHAP值高于25%分位数的特征(为了保留特征更多的信息)输入六种常见的机器学习模型——LR、RF、XGBoost、GB、自适应增强算法(Adaptive Boosting, AdaBoost)、类别特征梯度提升(Categorical Boosting, CatBoost),进行预测建模.通过在训练集上应用网格搜索(GridSearch)算法,以五折交叉验证下的AUC最大化为优化目标,确定最优模型参数.最终,在测试集上评估模型性能,并利用LIME方法增强模型决策的可解释性.

研究结果表明,SHAP技术能有效识别糖尿

病联合并发症的关键风险因素．本研究构建的方法体系不仅为识别糖尿病重要风险因素提供了高效路径，也为糖尿病患者的早期筛查与辅助治疗提供了可行的技术支持．

1　数据和方法

1.1　数据收集和来源　采用的数据源自国家人口健康科学数据中心提供的糖尿病并发症预警数据集（http://www.ncmi.cn）．此数据集涵盖了解放军人民医院3000例Ⅱ型糖尿病患者的87项指标信息，如纤维蛋白、血肌酐、白球比等，还记录了患者是否患有肾病、神经系统疾病、高血压等其他糖尿病相关并发症，并通过标签列区分是否为糖尿病视网膜病变（DR）患者．

1.2　方法　首先进行数据清洗，利用均匀加权K近邻（KNN）方法填充数据缺失值，然后通过*PCC*和卡方检验筛选出显著相关的糖尿病并发症，并将其纳入多任务学习模型．接着，使用可解释性技术SHAP对特征重要性进行评估，筛选出SHAP值高于75%分位数的特征作为联合并发症的重要风险因素．

为了保留数据的更多信息，选择SHAP值高于25%分位数的特征作为输入变量，用于构建六种机器学习模型（RF，XGBoost，LR，GB，AdaBoost和CatBoost）．以五折交叉验证下的最大*AUC*为目标函数，结合网格搜索（GridSearch）进行参数调优，最终得到糖尿病联合并发症的最优预测模型．然后，通过准确率、精确率、*F*1-score、*AUC*等指标来评估模型的预测性能．同时，引入局部可解释性技术LIME对模型进行解释，进一步验证基于可解释性方法筛选关键特征的有效性与可靠性．

1.2.1　数据预处理　首先对原始数据集进行缺失值分析，发现有15个指标出现大量缺失（缺失率大于45%），有34个指标出现部分缺失（缺失率小于35%），其余指标无缺失．

选择剔除缺失值较多（缺失率大于45%）的指标，对于部分缺失（缺失率小于35%）的指标，采用KNN方法进行填补．该方法考虑了数据的整体分布特征，充分利用多个特征的信息，并具有较强的抗异常值能力[10]，其计算式如下：

$$\begin{cases} d(X_i,X_j)=\sqrt{\sum_{k\in S}\left(X_{i,k}-X_{j,k}\right)^2} \\ X_{\text{missing}}=\dfrac{1}{K}\sum_{j=1}^{K}X_j \end{cases} \tag{1}$$

其中，$d(X_i,X_j)$表示第i个样本与第j个样本的欧式距离，$X_{i,k}$为第i个样本在第k个特征上的取值，$X_{j,k}$同理，S为特征集合，K是最近邻数量．

对于样本量适中、缺失率不高的医学数据集，设置$K=5$比较稳健[11-12]，本文也取$K=5$，即对于缺失特征值X_{missing}，用该特征上的五个最近邻样本的均值进行填充．此外，还对输入特征进行标准化处理，消除不同特征之间的量纲差异．随后，通过标准随机抽样将数据集划分为训练集（70%）和测试集（30%），利用训练集寻找模型的最佳参数，再基于测试集评估最佳模型的性能．

1.2.2　统计学方法　使用Python软件进行统计分析与建模．针对分类变量，采用*PCC*与卡方检验进行比较分析．模型评估方面，借助Scikit-learn库中的Metrics模块，选用准确率、精确率、*F*1-score和*AUC*作为主要评估指标．所有统计分析中，$p_value<0.05$被视为具有统计学意义．全部建模与分析过程均在Python 3.8.5中完成，部分图形绘制使用R 4.3.2软件实现．

1.2.3　重要风险因素的识别、验证以及解释

1.2.3.1　显著相关的联合并发症筛选　*PCC*能衡量两个变量之间的线性相关程度，已被广泛应用于医学研究领域[13-14]．卡方检验常用于评估分类变量之间关系的统计显著性，其p值可用于判断变量是否相关[15]．Jiang et al[16]采用*PCC*来分析预测与实际的*IC*50之间的线性相关性．DeGroat et al[17]结合*PCC*与卡方检验方法，评估健康个体与心血管疾病（Cardiovascular Disease，CVD）患者在转录组表达和临床特征方面的差异．首先，从数据中筛选出糖尿病相关的各类并发症，如糖尿病视网膜病变（Diabetic Retinopathy，DR）、高血压（HYPERTENSION）、动脉粥样硬化（Atherosclerosis，A_S）等34种病类，并定义并发症变量为二元变量，若患者患有某种并发症，标记为1，否则标记为0．随后，计算各并发症的平衡度，将平衡度小于0.2或大于5的情况视为高度不平衡类别，予以剔除．接着，采用*PCC*和卡方检验

量化各并发症之间的相关性,筛选出与糖尿病显著相关的联合并发症.其中,平衡度为不患病样本与患病样本的数量比,计算式如下:

$$IR=\frac{Counts[0]}{Counts[1]} \tag{2}$$

1.2.3.2 多任务学习模型构建 基于筛选出的显著相关联合并发症,构建一个基于多层感知机(Multilayer Perceptron,MLP)框架的多任务学习模型(Multi-Task Learning,MTL),采用硬参数共享策略[18],即在不同任务中共享同一组隐藏层参数.该模型的核心在于通过共享特征表示并引入跨任务一致性机制,实现多个任务的联合学习,从而提升模型在各项任务上的整体性能[19].

具体地,本研究构建的模型结构如下.

(1)静态结构.

输入层:传入矩阵 $X\in R^{n\times m}$,其中,n 是所有样本数,m 是除联合并发症以外的特征数.

共享隐藏层:定义两层共享隐藏层,设置神经元数量分别为 64,32. 基于 *Rule* 激活函数引入非线性,提高模型学习复杂关系的能力,并在第一层与第二层之间设置正则化技术(Dropout),在训练中随机丢弃部分神经元,抛弃比例设置为 0.5,增强模型的泛化能力.

输出层:最终的分类任务分别通过两个独立的全连接层输出预测值.

(2)动态学习.

损失函数:损失函数是一个数学函数,用于衡量模型预测结果与真实值之间的差异.其基本原理是模型在训练数据上的表现越优,对应的损失值就越小[20]. 针对二分类任务,采用二元交叉熵(Binary Cross-Entropy)作为损失函数,以衡量模型预测概率与真实标签之间的偏差.该方法旨在最小化预测值与真实值之间的误差,在多样本情况下,二元交叉熵可定义为[21-22]:

$$L=-\frac{1}{N}\sum_{i=1}^{N}\left[y_{i_{\text{true}}}\lg\left(\sigma\left(y_{i_{\text{pred}}}\right)\right)+\left(1-y_{i_{\text{true}}}\right)\lg\left(1-\sigma\left(y_{i_{\text{pred}}}\right)\right)\right] \tag{3}$$

其中,N 为样本总个数,$y_{i_{\text{true}}}$ 是第 i 个样本的真实标签,$y_{i_{\text{pred}}}$ 是第 i 个样本的预测值,$\sigma(x)$ 为 *Sigmoid* 激活函数,确保输出值位于 0~1. 基于联合优化的思想,采用联合加权优化总损失:

$$L_{\text{Totle}}=L_1+L_2+\cdots+L_n \tag{4}$$

其中,L_{Totle} 是总损失,L_n 是第 n 个任务的损失.

(3)优化算法.针对神经网络,最常用的训练方法之一是反向传播算法,该算法基于梯度下降法并需要采用自适应步长策略[23-24],其中,Adam-Optimizer 是使用最广泛的自适应步长优化方法之一[25]. 该方法 2015 年由 Kingma and Ba[26] 提出,具有实现简便、计算效率高以及内存需求小等优点.本研究中模型共训练 50 轮,并设置学习率为 0.001,以进行梯度更新并提升优化过程的稳定性,后称该模型为 MLP-MTL.

1.2.3.3 SHAP 值筛选重要风险因素 采用 SHAP 结合 MLP-MTL 模型进行特征重要性分析,识别对多种并发症均具有影响的关键因素. 2017 年 Lundberg and Lee[27] 提出 SHAP 方法,其核心在于通过计算某一特征在所有可能特征组合中存在与缺失时模型输出的平均差异得到该特征的 SHAP 值,进而量化其对模型预测的整体贡献.该方法已被广泛应用于各类"黑箱"模型的可解释性分析[27-29]. SHAP 值的计算式如下:

$$\varnothing_j=\sum_{S\subseteq f/\{j\}}\frac{|S|!(|F|-|S|-1)}{|F|!}\left[f\left(S\cup\{j\}\right)-f(S)\right] \tag{5}$$

其中,$\varnothing_j$ 为特征 j 的 SHAP 值,F 是所有输入特征的集合,$|F|$ 是特征集合 F 中包含的特征数量,S 为不包含 j 特征的子集,$|S|$ 是特征集合 S 中包含的特征数量,$f(S)$ 是分类模型在特征子集上的输出概率,$f\left(S\cup\{j\}\right)$ 是向子集 S 中添加特征 j 后的输出概率.分别计算各特征在不同分类任务中的 SHAP 值,并取其平均值以评估该特征在整体任务中的贡献程度.随后,利用 *PCC* 分析糖尿病联合并发症之间特征重要性的相关性,相关性较高表明各任务模型在决策过程中依赖相似的特征集合.此外,将 SHAP 贡献值高于 75% 的特征定义为影响多种并发症的重要特征.图 1 展示了识别糖尿病联合并发症关键风险因素的整体流程.

1.2.4 重要风险因素的验证

1.2.4.1 糖尿病联合并发症预测模型构建 机器学习能从大量复杂的数据集中提取有效信息并提

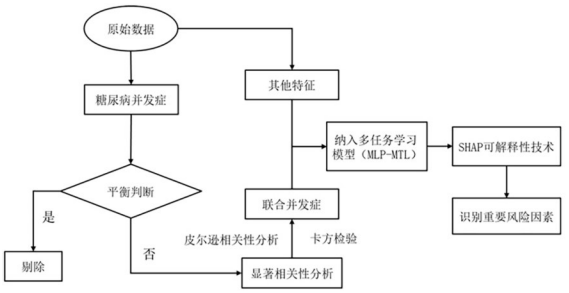

图1　识别联合并发症的重要风险因素流程

Fig. 1　Process for identifying significant risk factors for co-morbidities

升模型性能，已被广泛应用于疾病预测领域，显著提升了医疗保健的潜力[30−33]. 本研究采用当前主流的机器学习算法，如RF[34]，XGBoost[35]，LR[36]，GB[37]，AdaBoost[38]和CatBoost[39]，来构建糖尿病联合并发症预测模型，用于预测糖尿病患者是否同时患有多种显著相关的并发症. 数据被处理为二分类变量，若一名患者同时患有M种显著相关并发症（等于M），标记为1，若无并发症或仅患有一种或部分非显著并发症（小于M），标记为0. 随后，通过模型预测结果证明SHAP技术可用于识别联合并发症的重要风险因素. 为了更充分地挖掘数据中的信息，将SHAP值高于25%分位数的特征纳入模型，并通过GridSearch结合五折交叉验证，以最大化平均AUC为优化目标，在训练集中确定每个模型的最佳参数配置.

1.2.4.2　模型的评估　对于二分类问题，采用准确率（*Accuracy*）、精确率（*Precision*）、$F1$-score、AUC作为模型预测性能的评估指标[1,40]，计算式如下：

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN} \tag{6}$$

$$Precision=\frac{TP}{TP+FP} \tag{7}$$

$$F1\text{-score}=2\times\frac{Precision\times Recall}{Precision+Recall} \tag{8}$$

$$Recall=\frac{TP}{TP+FN} \tag{9}$$

$$\begin{cases} AUC=\int_0^1 TPR\mathrm{d}FPR \\ FPR=\dfrac{FP}{FP+TN} \\ TPR=\dfrac{TP}{TP+FN} \end{cases} \tag{10}$$

其中，TP（模型正确预测为正类的样本数）、FP（模型错误预测为正类的样本数）、TN（模型正确预测为负类的样本数）和FN（模型错误预测为负类的样本数）是评估分类模型的基本指标. FPR（假阳性率）表示负类样本中被错误预测为正类的比例，即在未患糖尿病及其并发症的个体中被误判为患病的比例；TPR（真正率）表示正类样本中被正确预测为正类的比例，即在实际患有糖尿病及其并发症的患者中被准确识别的比例.

1.2.4.3　模型的局部解释　LIME是一种局部可解释技术，广泛用来解释已训练完成的黑箱模型[41]. 从技术原理上看，LIME对选定样本生成新的扰动数据集并基于该数据集训练一个线性可解释模型，从而在局部范围内实现对原始模型预测结果的良好近似，并展示每个类别中各个特征对预测结果的具体贡献[33−43]. 本研究采用LIME技术对基于机器学习算法构建的预测模型进行解释，以揭示各特征在不同模型中的重要性. 图2展示了重要风险因素验证和解释的流程.

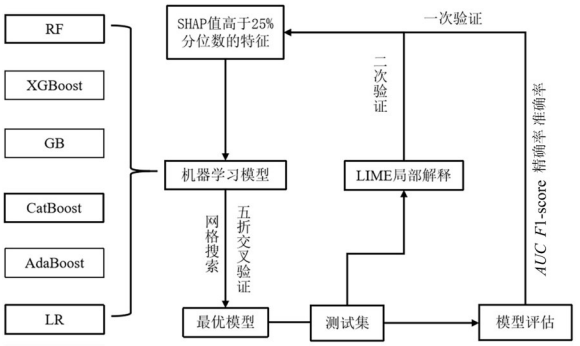

图2　重要风险因素的验证解释流程

Fig. 2　Validated interpretation process for important risk factors

2　结果和分析

2.1　识别糖尿病联合并发症重要风险因素

2.1.1　显著相关性分析　共纳入3000例样本，涵盖33种疾病类别. 由于研究重点为糖尿病并发症之间的关联性，首先筛选出所有与糖尿病相关的并发症并剔除不符合条件的疾病类别，最终确定三种主要并发症，即DR、肾病（NEPHROPATHY）和冠心病（Coronary Heart Disease，CHD）. 随后，计算这三种并发症之间的PCC（如图3）及

卡方检验的p值,筛选出具有显著相关性的疾病组合.结果显示,DR和NEPHROPATHY是一组显著相关的病组$(PCC>0.3, p_value<0.05)$.

2.1.2 多任务学习模型的训练 将除DR和NEPHROPATHY外的70个指标纳入MLP-MTL模型,制定了预测是否患有DR以及是否患有NEPHROPATHY的两个不同的分类任务,并基于训练集数据将模型循环训练50次.

2.1.3 重要风险因素识别 采用SHAP方法,结合已经训练好的多任务学习模型,分别计算各特征对DR和NEPHROPATHY的贡献程度(如图4所示),同时利用PCC来评估不同任务间特征重要性的排序一致性,结果为$PCC=0.80$,说明两个分类任务在特征重要性排序上具有高度的一致性.

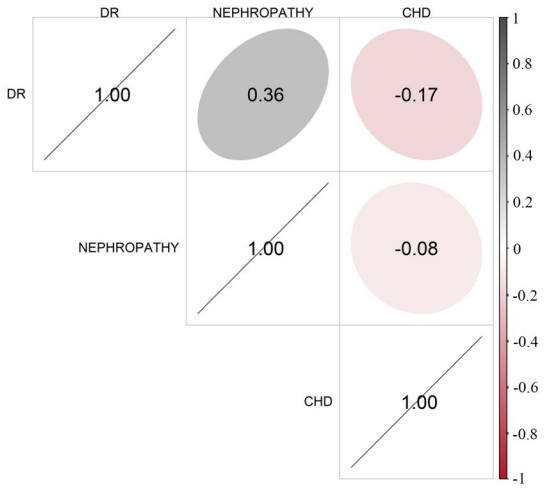

图3 并发症的相关系数图

Fig. 3 Plot of correlation coefficients for complications

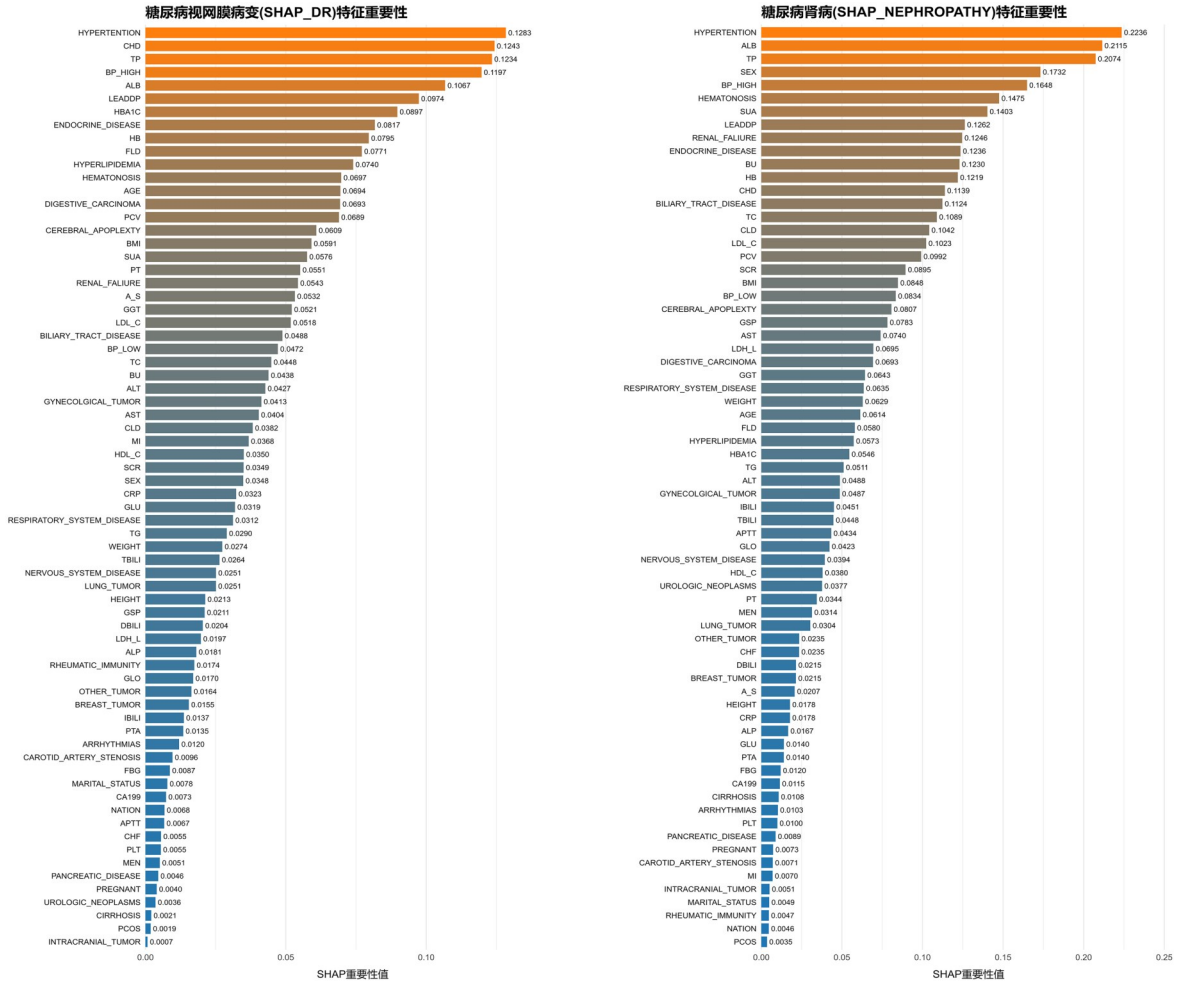

图4 DR (左)和NEPHROPATHY (右)的特征重要性分析

Fig.4 Analysis of the importance of DR (left) and NEPHROPATHY (right) characteristics

通过特征重要性分析以及对两种疾病的全局解释,发现部分特征是糖尿病视网膜病变(DR)和肾病(NEPHROPATHY)的共同关键因素,如图5所示. 由图可见,高血压(HYPERTENSION)、白蛋白(Albumin,ALB)、总蛋白(Total Protein,TP)和收缩压(High Blood Pressure,BP_HIGH)的重要性均位列前四,也有部分特征对两疾病的影响较小,如多囊卵巢综合征(Polycystic Ovary Syndrome,PCOS)、颅内肿瘤(INTRACRANIAL_TUMOR)和颈动脉狭窄(CAROTID_ARTERY_STENOSIS). 依据SHAP重要性值的75%分位数,筛选出对两种并发症均具有显著影响的特征,如BP_HIGH、HYPERTENSION、CHD、糖尿病下肢动脉病变(Lower Extremity Arterial Disease due to Diabetes,LEADDP)、血色素沉着症(HEMATONOSIS)、内分泌疾病(ENDOCRINE_DISEASE)、血清尿酸(Serum Uric Acid,SUA)、血红蛋白(Hemoglobin,HB)、红细胞压(Packed Cell Volume,PCV)、TP、ALB.

2.2　重要风险因素验证

2.2.1　预测模型的构建与评估　为了更充分地提取数据中的信息,选取SHAP值高于25%分位数的特征,并将其作为输入变量纳入六种机器学习模型(RF,XGBoost,LR,GB,AdaBoost和CatBoost). 以训练集中五折交叉验证的*AUC*为优化目标,采用网格搜索法对各模型参数进行调优,最终确定各模型的最佳参数组合,如表1所示.

模型参数确定后,在测试集上分别计算六种模型的准确率、精确率、*F*1-score以及*AUC*,评估模型的性能,验证特征选择方式的合理性,结果如表2和图6所示.

从表2可以看出,各模型的*Accuracy*基本一致,但LR模型的*Precision*仅为0.80,低于其他模型(均在0.85左右),说明LR模型在识别正类样本时存在较多误判. 相比之下,XGBoost和CatBoost模型的*F*1-Score均为0.75,高于其余四种

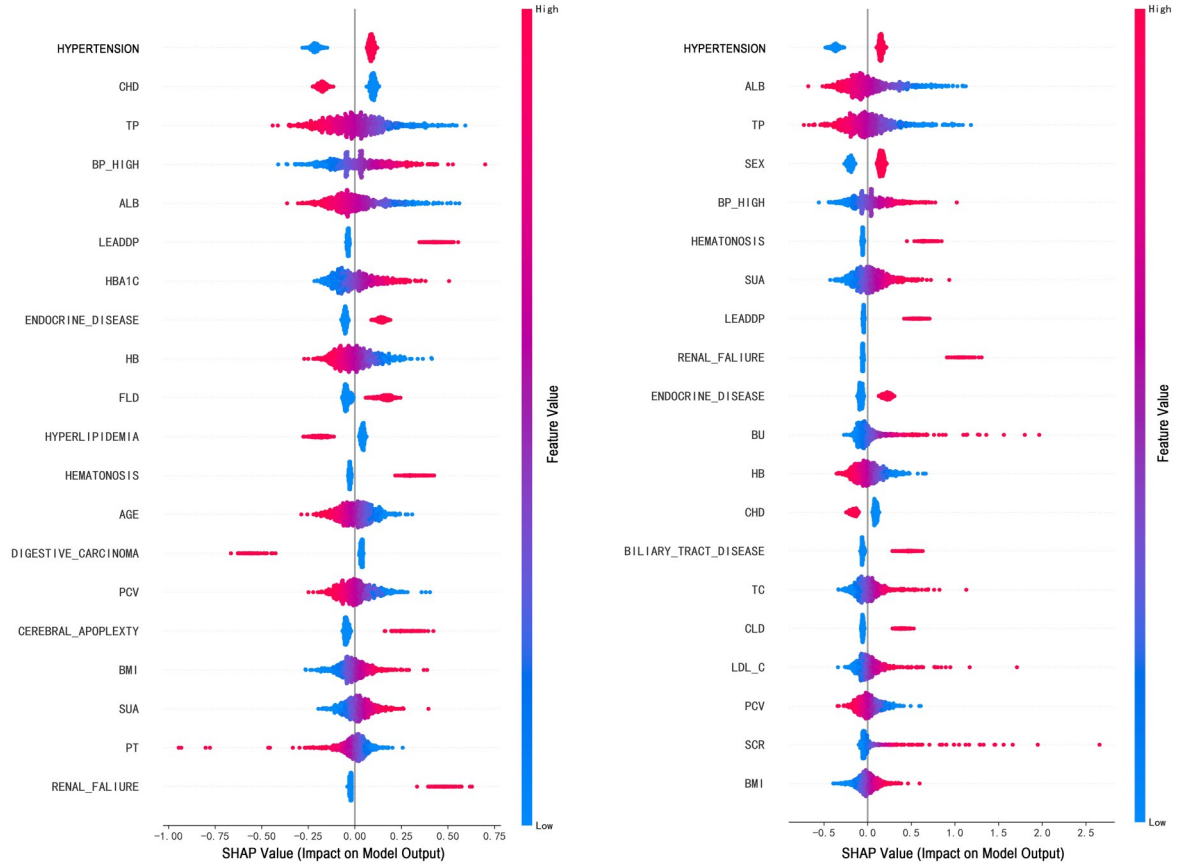

图5　DR和NEPHROPATHY的全局解释

Fig.5　Global explanation of DR and NEPHROPATHY

表 1 各模型的最佳参数

Table 1 Optimal parameters for each model

模型	参数	值	模型	参数	值
RF	max_depth	9	XGBoost	colsample_bytree	0.8
	max_features	log2		learning_rate	0.1
	min_samples_split	5		max_depth	3
	n_estimators	125		n_estimators	75
LR	C	1	GB	learning_rate	0.1
	penalty	L1		max_depth	3
	solver	liblinear		n_estimators	50
AdaBoost	learning_rate	0.1	CatBoost	depth	3
	n_estimators	200		learning_rate	0.1
				n_estimators	150

表 2 六种模型的预测性能的评估结果

Table 2 Evaluation for the predictive performance of six models

模型	准确率	精确率	*F*1-score	*AUC*
RF	0.84	0.85	0.70	0.90
XGBoost	0.86	0.84	0.75	0.92
LR	0.84	0.80	0.71	0.89
GB	0.85	0.85	0.73	0.91
AdaBoost	0.85	0.86	0.72	0.92
CatBoost	0.86	0.86	0.75	0.92

模型,表明其综合性能更强,具有更高的鲁棒性. 此外,CatBoost,XGBoost 和 AdaBoost 模型的 *AUC* 达到 0.92,为所有模型中最高,而 LR 模型的 *AUC* 最低,仅为 0.89,进一步说明这三种模型在分类能力上显著优于 LR 模型. 综上,所选特征包含了对预测任务的关键信息,验证了基于 SHAP 结合 MLP-MTL 模型进行 DR 关键特征筛选的有效性. 其中,CatBoost 模型在各项指标综合评估下表现最优.

图 6 展示了六个模型的 ROC 曲线及其对应的 *AUC*. 由图可见,除了 LR 模型以外,其余五种模型的 *AUC* 均不低于 0.90,整体分类性能较强,进一步证明所选特征是联合并发症的共同重要因素.

2.2.2 预测模型的局部解释 LIME 局部可解释技术有助于医生更好地理解基于机器学习构建的黑盒模型. 随机选取一名患者,计算其各特征

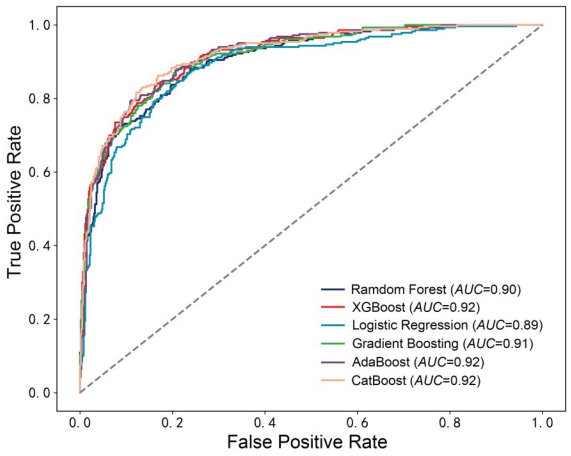

图 6 各模型在测试集上进行评估的 ROC 曲线

Fig. 6 ROC curves of each model evaluated on the test set

的权重,并对 Top15 特征的权重进行可视化,结果如图 7 所示. 由图可见,ALB、直接胆红素(Direct Bilirubin, DBILI)、ENDOCRINE_DISEASE、HYPERTENSION、LEADDP 和肾功能衰竭(RENAL_FALIURE)这六个特征均出现在各模型的 Top15 特征中,表明它们是影响模型预测的重要变量. 结合图 5 中的 SHAP 全局解释结果,二者在特征重要性分析上具有一致性,进一步验证了特征选择方法的合理性和可靠性. 此外,尽管各特征在模型中的权重有所不同,但其对预测结果的正负影响方向保持一致,说明模型的解释具有逻辑性和合理性.

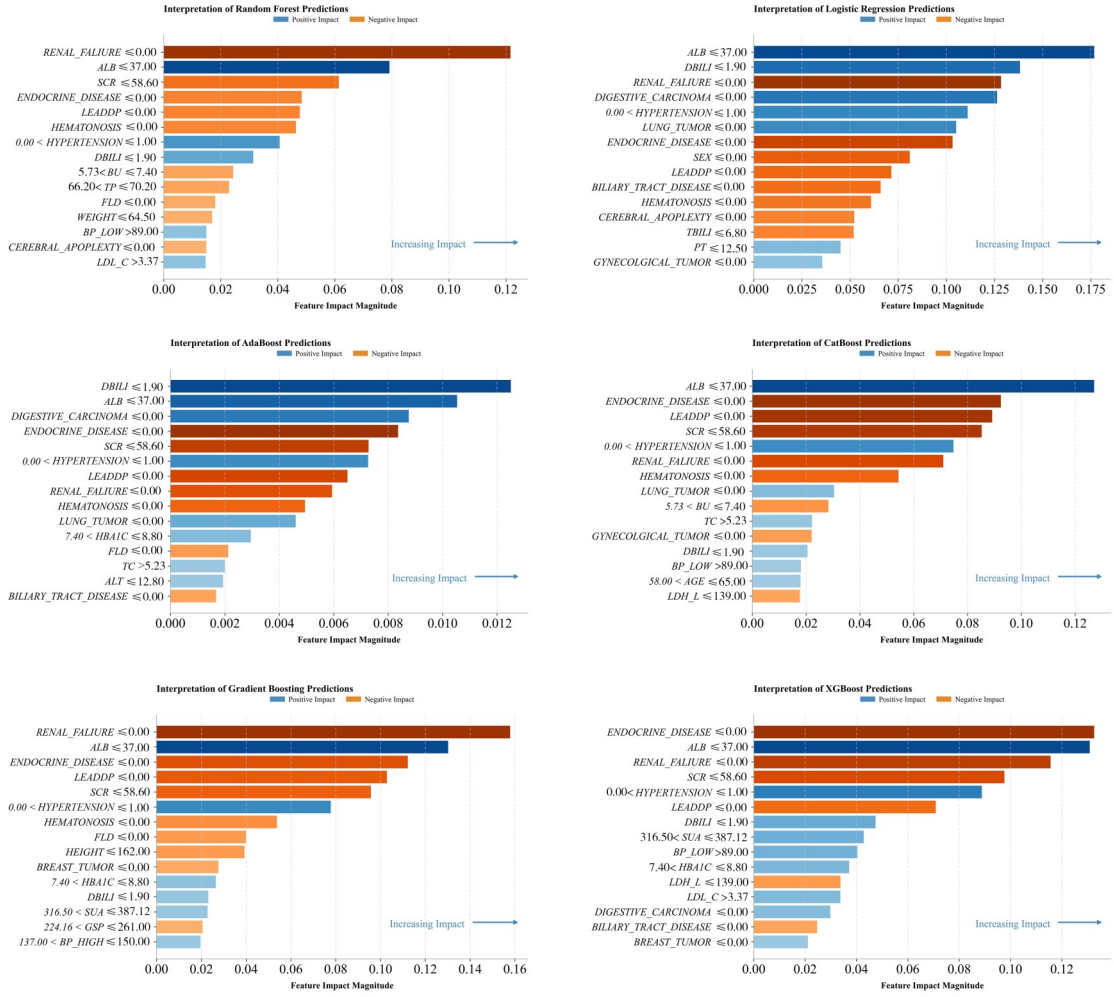

图7　各模型Top15重要特征LIME单一样本分析

Fig.7　Single-sample analysis of the Top15 important features of each model by LIME

3　讨论

尽管科技进步已催生出多种降血糖药物,糖尿病及其并发症仍对数百万患者构成沉重的健康和经济负担.糖尿病并发症是导致患者死亡的主要原因之一,因此糖尿病的管理与预防重点在于对各类并发症的干预[44].

以往研究多集中于糖尿病本身或单一并发症的风险因素分析,往往忽略不同并发症之间可能存在的潜在关联.为了更全面地考虑并发症之间的相互关系,本研究首先进行糖尿病并发症间的显著相关性分析.具体地,采用*PCC*筛选出具有相关性的并发症组合,并通过卡方检验识别具有统计学意义的联合并发症组合.

为了进一步识别这些联合并发症的重要风险因素,本研究构建了一个基于多层感知机的多任务学习模型(MLP-MTL),并在训练完成后应用SHAP可解释性技术对关键风险因素进行排序和筛选.结果显示,BP_HIGH,HYPERTENSION,CHD, LEADDP, HEMATONOSIS, ENDOCRINE_DISEASE,SUA,HB,PCV,TP,ALB等特征在DR与NEPHROPATHY中具有重要影响,这一发现与既往的研究结果一致[45-46].此外,还发现收缩压(BP_HIGH)、血清尿酸(SUA)及其他内分泌疾病等因素在联合特征重要性分析中排名靠前,提示其可能是糖尿病视网膜病变与肾病共存的重要高风险因素.

为了验证所识别的关键风险因素的有效性,

本研究将SHAP值高于25%分位数的特征纳入六种主流机器学习算法(包括随机森林、Logistic回归、GBM、CatBoost、AdaBoost、XGBoost)构建的联合并发症预测模型中.实验表明,各模型均表现出良好的预测性能,其中,以Catboost模型表现最优($Accuracy=0.86$, $Precision=0.86$, $F1\text{-score}=0.75$, $AUC=0.92$).研究表明,CatBoost在分类预测任务中优于其他机器学习方法[47],本研究结果与其一致.此外,本研究还采用局部可解释技术LIME结合所采用的六种机器学习模型实现单一患者的特征分析,发现结果与识别出的关键风险因素具有一致性,进一步验证所识别风险因素的可靠性与合理性.

本研究在考虑糖尿病并发症间潜在关联的基础上,提出一种将SHAP与MLP-MTL模型相结合的方法,高效识别了联合并发症的重要风险因素.该方法有助于医生针对可能有联合并发症的糖尿病患者采取有效的管理与预防措施,降低患者的死亡率.此外,基于筛选出的关键特征,结合机器学习算法构建的联合并发症预测模型,能实现对糖尿病人群中潜在联合并发症患者的快速识别.但本研究仍存在一定局限性.首先,由于数据样本量相对较少而特征维度较高,模型仅设置了两层隐藏层,可能无法充分捕捉所有特征的复杂关系;其次,受限于数据集中并发症类型的数量,仅聚焦于DR(糖尿病视网膜病变)和NEPHROPATHY(肾病)的共同关键特征分析.

综上,本研究通过引入SHAP与MLP-MTL模型相结合的方法,在识别联合并发症重要风险因素方面取得了良好效果,为医生提供依据以制定针对性干预策略,有助于减少患者因并发症导致的死亡.研究成果不仅提高了联合并发症风险因素的识别效率,还表明结合机器学习的预测方法在糖尿病联合并发症的早期筛查中具有良好的应用前景,并具备向其他疾病联合并发症研究推广的潜力.

参考文献

[1] American Diabetes Association. Diagnosis and classification of diabetes mellitus. Diabetes Care, 2014,37(Suppl 1):S81-S90.

[2] Sheetz M J, King G L. Molecular understanding of hyperglycemia's adverse effects for diabetic complications. JAMA,2002,288(20):2579-2588.

[3] Milanović M, Milošević N, Milić N, et al. Food contaminants and potential risk of diabetes development: A narrative review. World Journal of Diabetes,2023,14(6):705-723.

[4] Tan K R, Seng J J B, Kwan Y H, et al. Evaluation of machine learning methods developed for prediction of diabetes complications: A systematic review. Journal of Diabetes Science and Technology, 2023, 17(2): 474-489.

[5] Xie Z D, Nikolayeva O, Luo J B, et al. Building risk prediction models for type 2 diabetes using machine learning techniques. Preventing Chronic Disease, 2019,16:E130.

[6] Schallmoser S, Zueger T, Kraus M, et al. Machine learning for predicting micro- and macrovascular complications in individuals with prediabetes or diabetes: Retrospective cohort study. Journal of Medical Internet Research,2023,25:e42181.

[7] Cui M Z, Gang X K, Gao F, et al. Risk assessment of sarcopenia in patients with type 2 diabetes mellitus using data mining methods. Frontiers in Endocrinology,2020,11:123.

[8] Li W Y, Song Y N, Chen K, et al. Predictive model and risk analysis for diabetic retinopathy using machine learning: A retrospective cohort study in China. BMJ Open,2021,11(11):e050989.

[9] Lian X Y, Qi J Z, Yuan M Q, et al. Study on risk factors of diabetic peripheral neuropathy and establishment of a prediction model by machine learning. BMC Medical Informatics and Decision Making,2023,23(1):146.

[10] Emmanuel T, Maupong T, Mpoeleng D, et al. A survey on missing data in machine learning. Journal of Big Data,2021,8(1):140.

[11] Zhang S C. Nearest neighbor selection for iteratively kNN imputation. Journal of Systems and Software, 2012,85(11):2541-2552.

[12] Troyanskaya O, Cantor M, Sherlock G, et al. Missing value estimation methods for DNA microarrays. Bioinformatics,2001,17(6):520-525.

[13] Schober P, Mascha E J, Vetter T R. Statistics from A (agreement) to Z (z score): A guide to interpreting

common measures of association, agreement, diagnostic accuracy, effect size, heterogeneity, and reliability in medical research. Anesthesia and Analgesia, 2021, 133(6): 1633－1641.

[14] Schober P, Boer C, Schwarte L A. Correlation coefficients: Appropriate use and interpretation. Anesthesia and Analgesia, 2018, 126(5): 1763－1768.

[15] Ugoni A, Walker B F. The Chi square test: An introduction.COMSIG Review, 1995, 4(3): 61－64.

[16] Jiang L K, Jiang C Z, Yu X Y, et al. DeepTTA: A transformer-based model for predicting cancer drug response. Briefings in Bioinformatics, 2022, 23(3): bbac100.

[17] DeGroat W, Abdelhalim H, Patel K, et al. Discovering biomarkers associated and predicting cardiovascular disease with high accuracy using a novel nexus of machine learning techniques for precision medicine. Scientific Reports, 2024, 14(1): 1.

[18] Sun T X, Shao Y F, Li X N, et al. Learning sparse sharing architectures for multiple tasks// Proceedings of the AAAI Conference on Artificial Intelligence. Menlo Park, CA, USA: AAAI, 2020: 8936－8943.

[19] Klingner M, Fingscheidt T. Online performance prediction of perception DNNs by multi-task learning with depth estimation. IEEE Transactions on Intelligent Transportation Systems, 2021, 22(7): 4670－4683.

[20] Ruby U, Yendapalli V. Binary cross entropy with deep learning technique for image classification. International Journal of Advanced Trends in Computer Science and Engineering, 2020, 9(10): 9603－9608.

[21] Mao A Q, Mohri M, Zhong Y T. Cross-entropy loss functions: Theoretical analysis and applications// Proceedings of the 40th International Conference on Machine Learning. New York, NY, USA: PMLR, 2023: 23803－23828.

[22] Hurtik P, Tomasiello S, Hula J, et al. Binary cross-entropy with dynamical clipping. Neural Computing and Applications, 2022, 34(14): 12029－12041.

[23] Werbos P J. Backpropagation through time: What it does and how to do it. Proceedings of the IEEE, 1990, 78(10): 1550－1560.

[24] Wright L G, Onodera T, Stein M M, et al. Deep physical neural networks trained with backpropagation. Nature, 2022, 601(7894): 549－555.

[25] Bock S, Goppold J, Weiß M. An improvement of the convergence proof of the ADAM-optimizer. 2018, arXiv: 1804.10587.

[26] Kingma D P, Ba J. Adam: A method for stochastic optimization. 2017, arXiv: 1412.6980.

[27] Lundberg S M, Lee S I. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 2017, 30: 4765－4774.

[28] 黎子豪,蒋恕. 基于机器学习和SHAP算法的声波测井曲线重构及可解释性分析. 地质科技通报, 2025, 44(1): 321－331.

[29] 李佳思. 基于机器学习的糖尿病预测及SHAP特征分析. 智能计算机与应用, 2023, 13(1): 153－157.

[30] Lee A, Taylor P, Kalpathy-Cramer J, et al. Machine learning has arrived!. Ophthalmology, 2017, 124(12): 1726－1728.

[31] Polyzotis N, Zinkevich M, Roy S, et al. Data validation for machine learning. Proceedings of Machine Learning and Systems, 2019, 1: 334－347.

[32] Lee S, Mohr N M, Street W N, et al. Machine learning in relation to emergency medicine clinical and operational scenarios: An overview. Western Journal of Emergency Medicine, 2019, 20(2): 219－227.

[33] Wiens J, Shenoy E S. Machine learning for healthcare: On the verge of a major shift in healthcare epidemiology. Clinical Infectious Diseases, 2018, 66(1): 149－153.

[34] Schonlau M, Zou R Y. The random forest algorithm for statistical learning. The Stata Journa, 2020, 20(1): 3－29.

[35] Li W, Yin Y B, Quan X W, et al. Gene expression value prediction based on XGBoost algorithm. Frontiers in Genetics, 2019, 10: 1077.

[36] Schober P, Vetter T R. Logistic regression in medical research. Anesthesia and Analgesia, 2021, 132(2): 365－366.

[37] Konstantinov A V, Utkin L V. Interpretable machine learning with an ensemble of gradient boosting machines. Knowledge-Based Systems, 2021, 222: 106993.

[38] Ochs R A, Goldin J G, Abtin F, et al. Automated classification of lung bronchovascular anatomy in CT using AdaBoost. Medical Image Analysis, 2007, 11(3):315－324.

[39] Hancock J T, Khoshgoftaar T M. CatBoost for big data: An interdisciplinary review. Journal of Big Data, 2020, 7(1):94.

[40] 徐良辰,郭崇慧. 基于集成学习的胃癌生存预测模型研究. 数据分析与知识发现,2021,5(8):86－99.

[41] Ribeiro M T, Singh S, Guestrin C. "Why should I trust you?": Explaining the predictions of any classifier//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: Association for Computing Machinery, 2016:1135－1144.

[42] Alabi R O, Elmusrati M, Leivo I, et al. Machine learning explainability in nasopharyngeal cancer survival using LIME and SHAP. Scientific Reports, 2023, 13(1):8984.

[43] Salih A M, Raisi-Estabragh Z, Galazzo I B, et al. A perspective on explainable artificial intelligence methods: SHAP and LIME. Advanced Intelligent Systems, 2025, 7(1):2400304.

[44] Kong M J, Xie K, LÜ M H, et al. Anti-inflammatory phytochemicals for the treatment of diabetes and its complications: Lessons learned and future promise. Biomedicine & Pharmacotherapy, 2021, 133:110975.

[45] Wang G, Ouyang J, Li S, et al. The analysis of risk factors for diabetic nephropathy progression and the construction of a prognostic database for chronic kidney diseases. Journal of Translational Medicine, 2019, 17(1):264.

[46] 宋亚男,武惠韬,应俊,等. 基于机器学习算法探讨糖尿病视网膜病变的风险因素. 解放军医学院学报,2021,42(9):906－912,992.

[47] Ibrahim A A, Ridwan R L, Muhammed M M, et al. Comparison of the CatBoost classifier with other machine learning methods. International Journal of Advanced Computer Science and Applications, 2020, 11(11):738－748.

(责任编辑 杨可盛)