DOI:10.3969/j.issn.1005-202X.2023.08.018

医学生物信息

基于Null Importance与GS-LGBM的糖尿病视网膜病变因素分析与风险预测

曹佳悦，罗冬梅
安徽工业大学微电子与数据科学学院，安徽 马鞍山 243002

【摘要】**目的：**通过机器学习算法分析糖尿病视网膜病变（DR）关键因素，构建DR风险预测模型，为DR的预防和诊断提供参考。**方法：**采用国家人口健康科学数据中心的《糖尿病并发症预警数据集》，基于Null Importance方法去除噪声特征，筛选出与DR有关的关键因素；基于GridSearch优化LGBM模型参数，构建GS-LGBM DR风险预测模型。以准确率、精确率、召回率、F1分数、AUC值作为评价标准，与XGBoost、随机森林、Logistic以及未调优的LGBM模型进行比较。**结果：**Null Importance方法筛选出30个关键因素；与XGBoost、随机森林、Logistic以及未调优的LGBM模型相比，本研究所构建的GS-LGBM DR风险预测模型各评价指标均最优，其在测试数据上的AUC值高达0.897。**结论：**相较传统的DR预测模型，经过超参数优化后的模型具有更好的DR风险预测能力，更有助于DR的临床诊断。

【关键词】糖尿病视网膜病变；Null Importance；风险预测；GS-LGBM
【中图分类号】R318；R587.1 【文献标志码】A 【文章编号】1005-202X（2023）08-1033-06

Risk factors analysis and prediction of diabetic retinopathy based on Null Importance and GS-LGBM

CAO Jiayue, LUO Dongmei
School of Microelectronics and Data Science, Anhui University of Technology, Ma'anshan 243002, China

Abstract: Objective To analyze the risk factors of diabetic retinopathy (DR) and construct a DR risk prediction model through machine learning algorithms, thereby providing reference for DR prevention and diagnosis. **Methods** The study adopted the Diabetic Complication Early-Warning Data Set of the National Population Health Data Center. Null Importance method was used to remove noise features and screen out the key factors related to DR. LGBM model parameters were optimized with GridSearch to construct the GS-LGBM DR risk prediction model. The proposed method was compared with XGBoost, random forest, Logistic, and LGBM models in terms of accuracy, precision, recall, F1 score, and AUC values. **Results** Thirty key factors were screened out using the Null Importance method. Compared with XGBoost, random forest, Logistic and LGBM models, the GS-LGBM DR risk prediction model had the best evaluation performances, and its AUC value on the test data was as high as 0.897. **Conclusion** The hyperparameter optimized model is superior to the traditional DR prediction model, and it is more conducive to the clinical diagnosis of DR.

Keywords: diabetic retinopathy; Null Importance; risk profiling; GS-LGBM

前 言

截至2022年11月14日，根据最新发布的流行病学调查，中国糖尿病患者超1.4亿，位居世界首位[1]。糖尿病视网膜病变（Diabetic Retinopathy, DR）是糖尿病最常见最严重的微血管并发症之一，它是由于高血糖引起视网膜微血管病理性改变而发生的，是导致患者失明的主要原因，严重影响患者生活质量[2]。近些年，全球的DR患者逐年增加，预计到2030年全球将有3亿DR患者[3]。

对DR患者进行早期诊断并及时进行激光治疗可防止95%患者失明[4]。然而，现有的DR诊断方法主要基于眼底图像，需依赖眼底照相机等设备，由于

【收稿日期】2023-06-07
【基金项目】国家级创新创业训练项目（202110360094, 202210360089, 202210360086）；安徽省高校自然科学基金重点研究项目（2022AH050328）；安徽省教育教学研究项目（2020jyxm0238）
【作者简介】曹佳悦，研究方向：数据科学，E-mail: 3194889781@qq.com
【通信作者】罗冬梅，博士，讲师，研究方向：数据科学，E-mail: luodmahut@126.com

糖尿病患者人数较多，广泛诊断需耗费大量资源，且成本效益较低[5-6]。有研究表明，利用机器学习方法构建DR临床预测模型（根据预测值进行风险评分，筛选高风险人群进行眼科检查）相较于临床直接诊断具有较高的成本效益[7]。Yang等[8]通过单变量分析和LASSO回归识别DR独立因素，并基于多元逻辑回归模型开发列线图预测糖尿病患者患DR的风险。宋亚男等[9]采用递归特征消除（RFE算法）和XGBoost选取最优预测变量，应用SHAP方法对模型的风险因子进行解释分析。Houmayouni等[10]基于XGBoost的渐进消融特征选择方法预测DR的患病概率。申思源等[11]采用互信息选择DR的关键因素，并基于Stacking算法进行模型组合以提升糖尿病患者DR风险预测的算法精度。

以往的研究多采用互信息或通过机器学习算法对原始特征重要性进行排序，从而筛选出DR的关键因素。本研究采用Null Importance方法，通过比较初始特征重要性排序与打乱标签后的特征重要性排序之间的差异对DR关键因素进行筛选。然后将筛选的关键因素作为自变量，基于LGBM模型（Light Gradient Boosting Machine），采用GridSearch进行参数调优，构建GS-LGBM模型对DR患病风险进行预测，并与XGBoost[9-10]、随机森林[12-13]和Logistic回归[14-15]这3种常用预测模型以及未调优的LGBM模型进行对比。

1 材料与方法

1.1 数据来源

本研究数据来源于国家人口健康科学数据中心的《糖尿病并发症预警数据集》（http://www.ncmi.cn），数据集包含解放军人民医院3 000例糖尿病患者的87项生化指标数据，包括血尿素、脂蛋白、收缩压等以及患者的其它患病情况（如高血压、肾病、下肢动脉病变以及冠心病等）。观察对象标签为是否为DR患者，数据集中患DR和未患DR的患者各有1 500例。

1.2 方法

首先删除数据中异常值，并采用多重插补法对缺失值进行填补；然后采用Null Importance算法筛选DR关键因素[16-17]。接着将筛选出的关键因素作为自变量、用GridSearch对LGBM模型参数进行调优，构建GS-LGBM模型，并利用准确率、精确度、召回率、F1分数、AUC值等指标评估GS-LGBM模型与3种常用模型以及未调优的LGBM模型的优劣，验证GS-LGBM模型的优势。

1.2.1 数据预处理 根据3 Sigma原则发现数据没有异常值。然后，用R语言中VIM包matrixplot函数对缺失数据进行可视化（图1）。通过matrixplot函数，数据被重新转换到[0,1]区间，并用灰度表示大小：浅色表示值小，深色表示值很大，默认缺失值为红色。

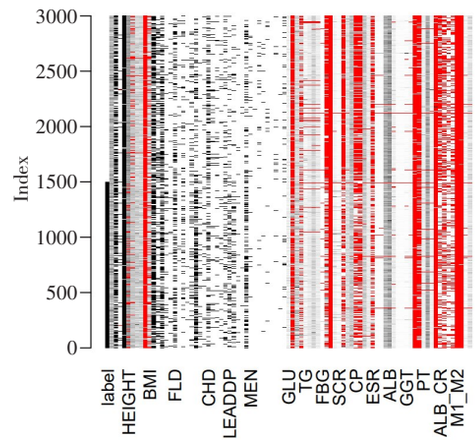


图1 缺失数据可视化

Figure 1 Visualization of missing data

纵轴表示患者编号，横轴为各项特征

从图1可以看出部分变量存在大量缺失，本研究保留缺失比例小于33%的特征，初步保留72个特征。然后利用R语言中mice包，基于多重插补法进行缺失值填补。多重插补法已广泛应用于医学研究的缺失值处理[18]。相对于单一插补法，多重插补法以每个填补数据集估计系数的平均值作为最终的关联估计，考虑了缺失值的不确定性，减小了抽样误差，从而得到更精确的点估计值，增加估计有效性。

1.2.2 筛选关键因素 本研究采用Null Importance算法筛选关键因素。已有研究表明Null Importance算法可用于各个领域且效果良好，对后续预测模型的建立、分类有重要帮助。陶世银等[19]利用Null Importance筛选特征，并将其用于闪电预报建模。刘金翰[20]基于Null Importance筛选特征，用于预测不同型号汽车在汽车测试台上完成测试所需时间。本研究首次尝试将Null Importance算法应用于DR的关键因素筛选。

Null Importanc算法的核心思想认为：如果某个特征对标签的预测非常有效，那么它在原始标签下重要性排序较高，而在打乱标签下，重要性排序很低。反之，若特征与标签无关，则在原始标签下的重要性排序小于等于打乱标签下的重要性排序。相比于单纯使用原始特征重要性进行特征筛选，Null Importance方法额外增加了一个必要条件（关键因素在打乱标签下的特征重要性排序较低），可以有效地排除一些噪声特征[19]。

Null Importance方法的实现步骤如下：首先基于树

模型(随机森林、LGBM等)的特征重要性函数[21]对原始数据进行训练得到特征原始重要性分布,其次将标签随机打乱n次,得到特征随机重要性排序,综合比较两者的偏离度是否显著,实现特征的评估和筛选。其筛选标准的score函数一般有两种设置方法:

方法1:

首先设置0-1变量a_i。

$$a_i = \begin{cases} 1, \mathrm{imp}_i < \mathrm{imp}_{\mathrm{true}}^{25\%} \\ 0, \mathrm{imp}_i \geqslant \mathrm{imp}_{\mathrm{true}}^{25\%} \end{cases} \quad (1)$$

其中,imp_i是指在第i次打乱后得到的随机特征重要性,$\mathrm{imp}_{\mathrm{true}}^{25\%}$是指原始特征重要性的25%分位数。当$\mathrm{imp}_i < \mathrm{imp}_{\mathrm{true}}^{25\%}$时,$a_i$为1,反之则为0。

其次对a_i进行求和除以打乱总次数n。

$$\mathrm{score}_1 = \frac{\sum_{i=1}^{n} a_i}{n} \quad (2)$$

score_1计算的是随机特征重要性小于原始特征重要性的次数占总次数的比重。score_1越小,说明该特征越不重要。

方法2:

$$\mathrm{score}_2 = \log\left(10^{-10} + \frac{\mathrm{imp}_{\mathrm{true}}}{1 + \mathrm{imp}_n^{75\%}}\right) \quad (3)$$

其中,$\mathrm{imp}_{\mathrm{true}}$是指特征的原始重要性;$\mathrm{imp}_n^{75\%}$是指经过$n$次打乱后,得到的$n$个随机重要性的75%分位数。若$\mathrm{score}_2$小于0则说明该特征为噪声特征但被拟合到模型中,需要进行剔除。

本研究基于LGBM模型的内置特征重要性函数,得到原始特征重要性排序与随机打乱80次($n=80$)得到的随机特征重要性,使用score_2进行综合比较,最终筛选30个关键因素。图2展示了根据score_2筛选关键因素的流程。图3展示特征的score_2在原始和随机打乱下的分布差异(score_2值的直方图)。

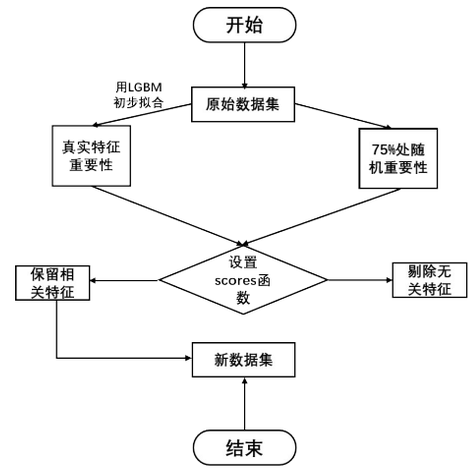
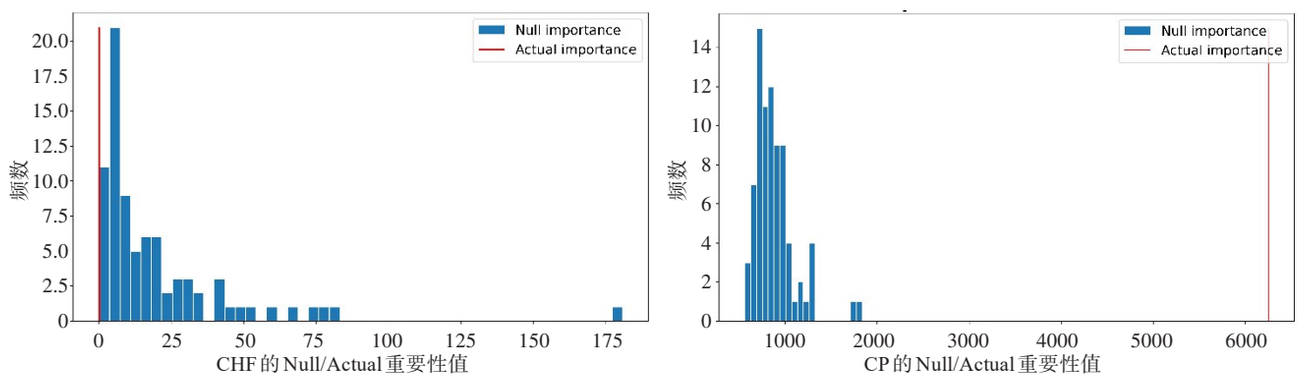


图2 Null Importance 流程图

Figure 2 Null Importance flowchart



a:CHF特征重要性分布　　**b:CP特征重要性分布**

图3 CHF及NEPHROPATHY特征重要性分布

Figure 3 Importance distributions for CHF and NEPHROPATHY

图3a展示了被剔除的因素CHF(心功能不全及心力衰竭)的原始(红色)与随机打乱下(蓝色)特征重要性分布;图3b展示了被保留下的因素CP(空腹C肽)的原始(红色)与随机打乱下(蓝色)特征重要性分布

从图3可以看出,被剔除的因素CHF(心功能不全及心力衰竭)的原始与随机打乱下特征重要性分布非常接近,而被保留下的因素CP(空腹C肽)的原始与随机打乱下特征重要性分布则差异很大。

1.2.3 模型的构建与调优　近年来,基于决策树的集成模型在机器学习领域应用广泛。LGBM、XGBoost、GBDT等均是以决策树作为基学习器的集成模型。GBDT由Freidman[22]于2001年提出,Chen等[23]基于GBDT模型提出XGBoost算法,与GBDT最大的区别是在损失函数上做二阶泰勒展开,进一步提升GBDT

模型的性能。但XGBoost和GBDT都需要遍历所有的实例来计算信息增益，从而确定最优分裂节点。当涉及到庞大的数据量和高维的数据特征时，XGBoost和GBDT等算法存在效率和可扩展性上的局限性[23]。

为了解决这些问题，微软亚洲研究院（MSRA）于2017年提出LGBM算法[24]。LGBM算法相对于XGBoost采用了单边梯度采样方法（Gradient-based One-Side Sampling，GOSS）和互斥稀疏特征绑定（Exclusive Feature Bundling，EFB）算法。其中，GOSS算法通过保留具有大梯度的实例以及随机采样后小部分具有小梯度的实例，从而在较少数据量下获得准确的信息增益估计，提高算法的效率。EFB算法则把高维稀疏数据中互斥的特征捆绑成一个特征，从而对数据进行降维。在GOSS和EFB的帮助下，LGBM在计算速度和内存消耗方面显著优于XGBoost和GBDT[24]。LGBM模型对于二分类问题的预测效果优良，已有效应用于乘用车使用寿命预测[25]、贷款违约预测[26-27]以及城市道路交通流量预测[28]等。

GridSearch的核心是将各个参数的可能取值进行排列组合形成“网格”，然后通过遍历所有可能参数组合训练模型找到最优参数。本研究采用GridSearch与LGBM模型相结合构建GS-LGBM模型，并尝试将其应用于DR的风险预测中。为了避免过拟合，本研究采用5折交叉验证确定GS-LGBM模型的最优参数。

1.2.4 模型的验证 为了验证GS-LGBM模型的预测性能，本研究设置随机种子数，将包含30个DR关键因素的3 000例糖尿病病患数据随机抽取70%作为训练集、30%作为测试集，带入XGBoost、随机森林、Logistic回归3种常用模型以及未调优的LGBM模型中，并计算它们的准确率（Accuracy）、精确率（Precision）和召回率（Recall）。各指标计算公式如下：

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{6}$$

其中，TP表示真阳性的数量；TN表示真阴性的数量；FP表示假阳性的数量；FN表示假阴性的数量。

为了更全面地评估模型，本研究还采用F1分数（F1-score）与AUC值对模型进行综合评价，F1分数与特异度（Specificity）的计算公式分别如下：

$$\text{Specificity} = \frac{TN}{FP + TN} \tag{7}$$

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

2 结 果

2.1 Null Importance筛选关键因素

本研究利用Null Importance最终筛选score$_2$大于0的30个DR关键因素。根据图4，排名前10的关键因素分别是：NEPHROPATHY（肾病）、LEADDP（下肢动脉变）、OTHER_TUMOR（其他肿瘤）、HEIGHT（高）、CHD（冠心病）、HBA1C（糖化血红蛋白）、ALB_CR（快速微量尿蛋白/肌酐测定）、FLD（脂肪肝）、ENDOCRINE_DISEASE（其他内分泌疾病）以及CP（空腹C肽）。这与申思源等[11]采用互信息得到的关键因素以及曹文哲等[29]采用随机森林算法得到的关键因素相符合，说明Null Importance算法能有效筛选关键因素。

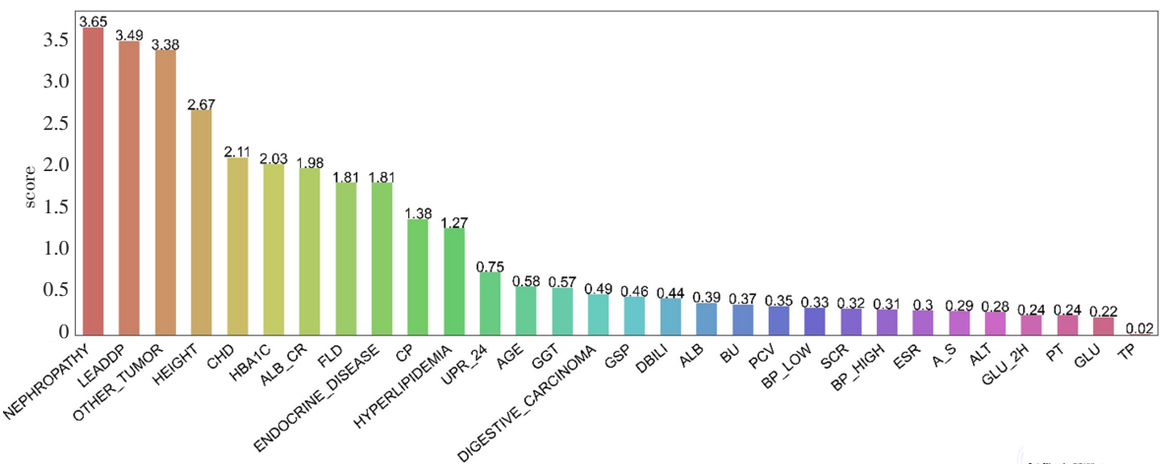


图4 重要性得分

Figure 4 Importance score

横坐标为对应的关键因素名称，纵坐标为各因素的score$_2$值

2.2 基于GS-LGBM的DR预测

本研究利用GridSearch将LGBM模型中学习率、树的最大深度、叶节点数、弱学习器个数这4个参数进行排列组合形成“网格”，并使用5折交叉验证方法选择出最佳超参数组合为：学习率为0.05，最大深度为2，叶节点数为5，弱学习器个数为500。

设计随机种子为42，随机抽取70%作为训练集、30%作为测试集分别代入GS-LGBM模型、XGBoost、随机森林、Logistic回归以及未调优的LGBM模型。ROC曲线以及AUC如图5所示，其准确率、精确率、召回率以及F1分数见表1。

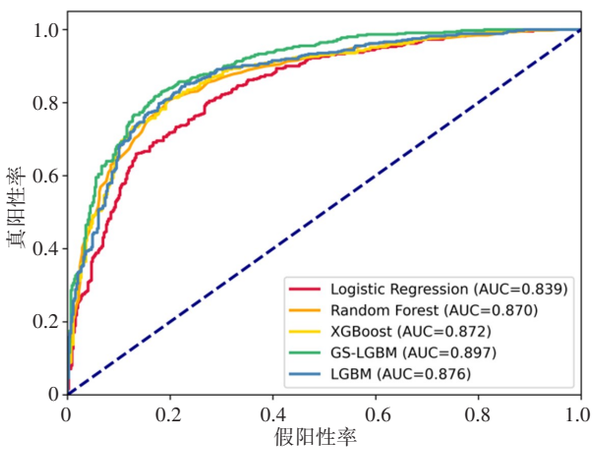


图5 各个模型ROC曲线

Figure 5 ROC curve of each model

表1 各个模型评估指标

Table 1 Evaluation metrics for each model

模型	准确率	精确率	召回率	F1分数
XGBoost	0.796	0.804	0.782	0.793
随机森林	0.799	0.816	0.771	0.793
Logistic回归	0.758	0.783	0.713	0.747
GS-LGBM	0.822	0.828	0.813	0.821
原始LGBM	0.808	0.813	0.800	0.807

从图5可以看出，所建立的GS-LGBM预测模型的ROC曲线对应的AUC值为0.897，相较于其它4种模型，AUC值为最高。

从表1可以看出，相比于其它4种模型，GS-LGBM各项评价指标（准确率、精确率、召回率、F1分数）均较优。因此，本研究所建立的GS-LGBM模型可以作为DR风险的预测模型，且预测效果良好。

3 讨 论

视网膜血管和神经的病理性改变早在眼底DR可见之前就已存在[30]。因此，筛选DR的关键因素，构建DR临床预测模型对于早期有效预防和治疗十分必要。当前的研究分为两类，一类是根据糖尿病患者视网膜图像数据库进行DR智能诊疗。Somasundaram等[31]设计了一种机器学习袋装集成分类器ML-BEC，提取包括血管、视神经、神经组织、神经视网膜边缘、视盘大小等图像特征，并使用BEC分类器建立预测模型，获得较好的分类结果。这一类方法计算量较大，对计算资源要求比较高。另一类是根据患者的临床生化指标构建DR临床预测模型。杨洪燕等[32]采用最大相关-最小冗余算法及随机森林算法筛选关键因素，并将筛选的关键因素代入Logistic回归模型，以此建立DR预测模型。这一类方法计算量不是太大，对计算资源要求不是很高，但是对数据的质量、特征工程以及模型精度的要求较高。

与以往研究不同的是，本研究首次采用Null Importance算法筛选DR关键因素，该方法通过比较原始特征重要性分布与打乱标签下的特征重要性分布之间的差异有效地排除一些噪声特征[19]。另外，本研究还基于GridSearch对LGBM模型进行参数调优，构建GS-LGBM DR风险预测模型，并通过比较发现：相较传统的预测模型，基于GS-LGBM的风险预测结果更好，精确度较高。

本研究中也存在一些需要改进的地方：所构造的模型仅仅只与单一模型进行比较，今后的研究中会考虑将组合模型也纳入比较范围；所用数据样本分布较为平衡（DR和非DR各1 500例），然而在实际医学数据中，往往存在数据分布不平衡的问题，今后的研究需更深入地优化模型，增加其泛化能力。

综上所述，基于Null Importance与GS-LGBM所构建的临床风险预测模适用于DR的早期筛检，在降低社会成本和临床应用中具有一定的价值。

【参考文献】

[1] 央视网. 我国成年人糖尿病患病率为12.8%科学控糖避开“甜蜜”陷阱[EB/OL]. (2022-11-14)[2023-5-6]. https://news.cctv.com/2022/11/14/ARTISnxsqyD9AyXPP0P8fDMj221114.shtml.
CCTV.com. The prevalence of diabetes in adults in China is 12.8%, and scientific sugar control avoids the "sweet" trap[EB/OL].(2022-11-14) [2023-5-6]. https://news. cctv. com/2022/11/14/ARTISnxsqyD9AyXPP0P8fDMj221114.shtml.

[2] Peimani M, Nasli-Esfahani E, Shakibazadeh E. Ottawa charter framework as a guide for type 2 diabetes prevention and control in Iran[J]. J Diabetes Metab Disord, 2019, 18(1): 256-261.

[3] Nakajima M, Cooney MJ, Tu AH, et al. Normalization of retinal ascular permeability in experimental diabetes with genistein[J]. Invest Ophthamol Vis Sci, 2001, 42(9): 2110-2114.

[4] Umaefulam V, Premkumar K. Diabetic retinopathy awareness and eye care behaviour of indigenous women in Saskatoon, Canada[J]. Int J Circumpol Heal, 2021, 80(1): e1878749.

[5] Jones S, Edwards RT. Diabetic retinopathy screening: a systematic

review of the economic evidence[J]. Diabet Med, 2010, 27(3): 249-256.
[6] Scanlon PH, Aldington SJ, Leal J, et al. Development of a cost-effectiveness model for optimisation of the screening interval in diabetic retinopathy screening[J]. Health Technol Assess, 2015, 19 (74): 1-116.
[7] Aspelund T, Thornórisdóttir O, Olafsdottir E, et al. Individual risk assessment and information technology to optimise screening frequency for diabetic retinopathy[J]. Diabetologia, 2011, 54(10): 2525-2532.
[8] Yang YZ, Tan JT, He YX, et al. Predictive model for diabetic retinopathy under limited medical resources: a multicenter diagnostic study[J]. Front Endocrinol, 2023, 13: e1099302.
[9] 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[J]. 解放军医学院学报, 2021, 42(9): 906-912.
Song Y, Wu HT, Ying J, et al. Risk factors analysis of diabetic retinopathy based on machine learning[J]. Academic Journal of Chinese PLA Medical School, 2021, 42(9): 906-912.
[10] Homayouni A, Tieming L, Thieu T. Diabetic retinopathy prediction using progressive ablation feature selection: a comprehensive classifier evaluation[J]. Smart Health, 2022, 26: e100343.
[11] 申思源, 罗冬梅. 糖尿病视网膜病变的风险揭示与关键因素分析[J]. 中国医学物理学杂志, 2022, 39(6): 783-787.
Shen SY, Luo DM. Risk disclosure and key factors analysis of diabetic retinopathy[J]. Chinese Journal of Medical Physics, 2022, 39(6): 783-787.
[12] 倪孝兵, 王思宏, 黄崇兵. 血清总胆固醇/甘油三酯比值对2型糖尿病视网膜病变的预测价值研究[J]. 医学理论与实践, 2023, 36(4): 655-657.
Ni XB, Wang SH, Huang CB. Predictive value of serum total cholesterol/triglyceride ratio on type 2 diabetic retinopathy[J]. The Journal of Medical Theory and Practice, 2023, 36(4): 655-657.
[13] 郝振伟. 非增生型糖尿病视网膜病变患者循环胆红素水平变化及其预测价值[J]. 中国现代医生, 2022, 60(31): 52-56.
Hao ZW. Changes of circulating bilirubin level in patients with non-proliferative diabetic retinopathy and its predictive value[J]. China Modern Doctor, 2022, 60(31): 52-56.
[14] 王夺. 基于随机森林的糖尿病足治疗策略的研究[D]. 长春: 吉林大学, 2022.
Wang D. Study on the treatment strategy of diabetic foot based on randomforest[D]. Changchun: Jilin University, 2022.
[15] 李军, 胡晓娟, 周昌乐, 等. 基于随机森林算法的糖尿病舌象特征分析和诊断模型研究[J]. 中华中医药杂志, 2022, 37(3): 1639-1643.
Li J, Hu XJ, Zhou CL, et al. Study on the feature analysis and diagnosis model of diabetic tongue based on random forest algorithm[J]. China Journal of Traditional Chinese Medicine and Pharmacy, 2022, 37(3): 1639-1643.
[16] 王威, 杨帆. 基于多重填补的广义线性模型在肾脏疾病研究中的应用[J]. 肾脏病与透析肾移植杂志, 2021, 30(5): 476-479.
Wang W, Yang F. A generalized linear model based on multiple imputation in kidney disease research[J]. Chinese Journal of Nephrology, Dialysis & Transplantation, 2021, 30(5): 476-479.
[17] Pelgrims I, Devleesschauwer B, Vandevijvere S. Using random-forest multiple imputation to address bias of self-reported anthropometric measures, hypertension and hypercholesterolemia in the Belgian health interview survey[J]. BMC Med Res Methodol, 2023, 23(1): e69.
[18] 朱荣慧, 许金芳, 王睿, 等. 多重填补技术在医学研究缺失值处理中的应用及发展[J]. 中国卫生统计, 2022, 39(2): 293-295.
Zhu RH, Xu JF, Wang R, et al. Application and development of multiple filling technology in the treatment of missing values in medical research[J]. Chinese Journal of Health Statistics, 2022, 39(2): 293-295.
[19] 陶世银, 贺敬安. 基于XGBoost与特征重要性筛选的闪电预报模型构建研究[J]. 国外电子测量技术, 2022, 41(1): 99-105.
Tao SY, He JA. Research on construction of lightning forecast model based on XGBoost and feature importance screening[J]. Foreign Electronic Measurement Technology, 2022, 41(1): 99-105.
[20] 刘金翰. 基于回归分析的汽车测试时间预测[D]. 兰州: 兰州大学, 2021.
Liu JH. The prediction of time spend on automobile test system using regression analysis[D]. Lanzhou: Lanzhou University, 2021.
[21] André A, Laura T, Oliver S, et al. Permutation importance: a corrected feature importance measure[J]. Bioinformatics, 2010, 26(10): 1340-1347.
[22] Friedman JH. Greedy function approximation: a gradient boosting machine[J]. Ann Stat, 2001, 29(5): 1189-1232.
[23] Chen TQ, Guestrin C. XGBoost: a scalable treeboosting system[C]. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016: 785-794.
[24] Guo LK, Qi M, Thomas F, et al. Lightgbm: a highly efficient gradient boosting decision tree[C]. Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017: 3149-3157.
[25] 徐国强, 徐妍, 郭德卿. 基于GS-LGBM的乘用车使用寿命预测研究[J]. 科技和产业, 2022, 22(9): 341-346.
Xu GQ, Xu Y, Guo DQ. Research on life prediction for passenger vehicles based on GS-LGBM[J]. Science Technology and Industry, 2022, 22(9): 341-346.
[26] 陈福康. 不平衡数据下信贷违约风险控制研究[D]. 济南: 山东大学, 2021.
Chen FK. Research on credit default risk control under unbalanced data [D]. Ji'nan: Shandong University, 2021.
[27] 王向鹏. 基于不平衡三分类LGBM模型的贷后风险预警研究[D]. 兰州: 兰州大学, 2019.
Wang XP. Research on post-loan risk warning based on unbalanced three-classification LGBM model[D]. Lanzhou: Lanzhou University, 2019.
[28] 何芸. 基于LGBM模型的城市道路交通流量预测研究[J]. 电子技术与软件工程, 2022, 3: 259-262.
He Y. Research on urban road traffic flow prediction based on LGBM model[J]. Electronic Technology & Software Engineering, 2022, 3: 259-262.
[29] 曹文哲, 应俊, 陈广飞, 等. 基于Logistic回归和随机森林算法的2型糖尿病并发视网膜病变风险预测及对比研究[J]. 中国医疗设备, 2016, 31(3): 33-38.
Cao WZ, Ying J, Chen GF, et al. Risk prediction and comparitive research of type 2 diabetes mellitus complicated with retinopathy based on logistic regression and random forest algorithm[J]. China Medical Devices, 2016, 31(3): 33-38.
[30] Honasoge A, Nudleman E, Smith M, et al. Emerging insights and interventions for diabetic retinopathy[J]. Curr Diabetes Rep, 2019, 19 (10): 1-16.
[31] Somasundaram SK, Alli P. A machine learning ensemble classifier for early prediction of diabetic retinopathy[J]. J Med Syst, 2017, 41(12): 201-210.
[32] 杨洪燕, 夏森, 刘赞朝, 等. 2型糖尿病视网膜病变临床预测模型的构建与评价[J]. 中国慢性病预防与控制, 2023, 31(1): 2-7.
Yang HY, Xia M, Liu ZC, et al. Establishment and evaluation of a clinic prediction model of diabetic retinopathy in patients with type 2 diabetes mellitus[J]. Chinese Journal of Prevention and Control of Chronic Diseases, 2023, 31(1): 2-7.

（编辑：谭斯允）