Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2026, 15(2), 387-394
Published Online March 2026 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2026.152037

Hans汉斯

基于机器学习的糖尿病预测研究

陈　雪，高瑞娟，赵丽婷

河北金融学院金融科技学院，河北　保定

收稿日期：2025年11月25日；录用日期：2026年2月14日；发布日期：2026年2月26日

摘　要

随着全球糖尿病患病率的持续攀升，早期筛查与干预已成为降低疾病负担的关键手段。本文提出一种基于机器学习的预测模型，实现糖尿病的精准预测。首先对包含1006例临床样本的糖尿病数据集进行清洗、标准化及特征筛选，通过相关性分析与递归特征消除提取8个核心预测指标；随后构建逻辑回归、决策树、随机森林、支持向量机及XGBoost五种预测模型，采用网格搜索结合5折交叉验证进行参数优化；最终通过多指标综合评估，确定随机森林模型为最优预测模型(准确率84.16%, AUC = 0.9096)。该模型为糖尿病早期筛查提供了高效技术支持，可辅助基层医疗机构开展风险评估工作，具有重要的临床应用价值与社会意义。

关键词

糖尿病预测，机器学习，随机森林，模型评估

Research on Diabetes Prediction Based on Machine Learning

Xue Chen, Ruijuan Gao, Liting Zhao

School of FinTech, Hebei University of Finance, Baoding Hebei

Received: November 25, 2025; accepted: February 14, 2026; published: February 26, 2026

Abstract

As the global prevalence of diabetes continues to rise, early screening and intervention have become key measures to reduce the disease burden. This paper proposes a machine learning-based prediction model to achieve accurate diabetes prediction. First, a diabetes dataset containing 1006 clinical samples was cleaned, standardized, and subjected to feature selection: 8 core predictive indicators were extracted via correlation analysis and recursive feature elimination. Then, five predictive models (logistic regression, decision tree, random forest, support vector machine, and

文章引用: 陈雪, 高瑞娟, 赵丽婷. 基于机器学习的糖尿病预测研究[J]. 人工智能与机器人研究, 2026, 15(2): 387-394.
DOI: 10.12677/airr.2026.152037

XGBoost) were constructed, with parameter optimization performed using grid search combined with 5-fold cross-validation. Finally, through a comprehensive multi-metric evaluation, the random forest model was identified as the optimal predictive model (accuracy: 84.16%, AUC = 0.9096). This model provides efficient technical support for early diabetes screening, can assist primary medical institutions in conducting risk assessment, and holds significant clinical application value and social significance.

Keywords

Diabetes Prediction, Machine Learning, Random Forest, Model Evaluation

1. 引言

糖尿病作为一种以慢性高血糖为特征的代谢性疾病，已成为全球性公共卫生问题[1]。国际糖尿病联盟最新报告显示，全球糖尿病患者数量已突破 5 亿，且仍以每年数百万的速度增长[2]。长期高血糖状态可引发心血管疾病、肾病、视网膜病变等多种严重并发症，显著降低患者生活质量并增加医疗负担[3]。传统糖尿病诊断主要依赖血糖检测与临床症状识别，往往在疾病进展至中晚期才能确诊，错失了早期干预的最佳时机。

随着医疗信息化的快速发展，机器学习技术凭借其强大的数据分析与模式识别能力，在疾病预测领域展现出巨大潜力[4]。与传统诊断方法相比，机器学习模型能够整合生理指标、生活习惯及遗传因素等多维度数据，挖掘潜在的疾病关联模式，实现糖尿病风险的早期预警[5]。然而，现有预测模型仍存在泛化能力不足、特征冗余、临床实用性不强等问题，如何构建高精度、高稳定性的预测模型，成为当前研究的核心挑战。

2. 相关工作

近年来，机器学习在糖尿病预测领域的应用研究取得了显著进展，国内外学者从单一算法优化、多模型融合及系统开发等多个维度展开探索。Nabila [6]等人提出了一种基于机器学习算法的智能预测系统，采用基于朴素贝叶斯并结合聚类的方法，利用 532 例糖尿病患者数据评估算法性能。该研究为糖尿病早期预测提供了新的技术路径，展现了机器学习在健康管理中的潜在应用价值。Kui L [7]等人评估了机器学习模型在预测血糖水平和检测不良血糖事件中的表现，研究表明机器学习模型在糖尿病管理中的应用具有广阔前景，尤其是在血糖水平预测和不良事件预警方面。

在单一模型与系统集成研究中，苗丰顺[8]采用随机森林、XGBoost 和 CatBoost 算法训练模型，结合 IV 值分析筛选特征，最终 CatBoost 算法表现最优，被嵌入糖尿病预测系统。刘建平[9]通过 Logistic 回归、支持向量机和 XGBoost 构建糖尿病预测模型，发现 Logistic 回归模型在参数调优后表现最佳(AUC = 0.881)，并设计基于 Uniapp 和 Django 框架的预测系统，为糖尿病早期筛查和干预提供了重要技术支持。

多模型融合成为提升预测性能的重要方向，仵豪[10]提出了改进的糖尿病管理模式，通过 Stacking 方法融合逻辑回归、随机森林、支持向量机和极端梯度提升算法，构建了糖尿病风险早期识别模型，实现了对潜在患病人群的有效筛检。马吉聪[11]采用 XAR-Stacking 融合模型，结合随机森林、AdaBoost 和 XGBoost 算法，通过五折交叉验证和逻辑回归融合，实验表明该模型在精确率、召回率、F1-Score 和 AUC

值上优于单一模型，基于 Django 的 MTV 设计模式开发了糖尿病预测系统，实现模型与 Web 服务的松耦合集成。

综合来看，国内外相关研究展示出机器学习于糖尿病预测方面有巨大潜力，同时也指出多模型融合以及特征工程对于提升预测性能所起到的关键作用，通过对众多临床数据展开分析，筛选出和糖尿病发病紧密相关的关键特征，可进一步优化模型的预测能力。虽然现有研究取得了一定进展，然而依然存在一些挑战，比如，怎样提高模型的泛化能力，让其可适应不同地区、不同人群的数据，如何将临床知识与机器学习技术相结合，开发出更具临床实用性的预测工具，这些问题有待在未来研究中给予解决。

3. 模型构建与评估

3.1. 数据集描述

本研究采用的数据集来自机器学习网站和鲸社区，由某医疗机构提供并经过脱敏处理，包含 1006 例临床记录。数据集共包含 16 个变量，其中 15 个为特征变量，1 个为目标变量。特征变量涵盖人口学特征(性别、年龄)、血脂指标(高密度脂蛋白胆固醇、低密度脂蛋白胆固醇、极低密度脂蛋白胆固醇、甘油三酯、总胆固醇)、心血管相关参数(脉搏、舒张压、高血压史)及肾功能指标(尿素氮、尿酸、肌酐、体重检查结果)等；目标变量标识患者糖尿病诊断状态(0 表示非糖尿病，1 表示糖尿病)。数据集按 8:2 比例划分为训练集(804 例)与测试集(202 例)，具体分布如表 1 所示。

Table 1. Dataset partitioning
表 1. 数据集划分

	非糖尿病患者	糖尿病患者	共计
训练集	447	357	804
测试集	112	90	202
共计	559	447	1006

3.2. 数据预处理

1) 数据标准

本研究采用基于四分位距 IQR 的异常值检测方法，首先对各特征序列进行统计分析，分别计算其第一四分位数 Q1 和第三四分位数 Q3，继而求得四分位距 IQR (Q3-Q1)，最终将超出 $[Q1-1.5\times IQR, Q3+1.5\times IQR]$ 范围的观测值判定为异常数据。对于检测到的异常值，选择将其替换为该列的中位数，以减轻异常值对后续分析或建模的影响。

2) 数据标准化

此过程涉及将各个指标变量的原始值转化为以 0 为均值，1 为标准差的标准化正态分布形式。

$$X' = \frac{x-\mu}{\sigma} \tag{1}$$

其中，x 为原始数据，μ 为均值，σ 为标准差，x' 为标准化后的数据。

3) 特征选择

将各特征的线性相关程度通过计算相关系数矩阵的形式形象化成热力图，通过设置门槛值 0.8，采用上三角矩阵法来过滤删除关联度高的特征，从而为冗余特征对模型构成干扰的多重共线性问题提供了有效解决手段。从图 1 中可看出，极低密度脂蛋白胆固醇与甘油三酯的相关系数为 0.9，表明二者在脂质代

谢中高度共变。考虑到极低密度脂蛋白胆固醇在代谢综合征与胰岛素抵抗路径中更具病理生理特异性，且其与多个并发症(如心血管疾病)关联更直接，故保留极低密度脂蛋白胆固醇，剔除甘油三酯。其余特征间相关系数均低于 0.5，表明其独立性较强，全部保留进入下一步特征筛选。初步筛选后，保留特征包括：基础人口统计特征(性别、年龄)、脂肪血脂指标(高密度脂蛋白、低密度脂肪蛋白)、生命体征(脉搏、舒张压)以及肾功能指标(尿素氮、尿酸、肌酐)共 13 个特征。特征相关性热力图如图 1 所示。

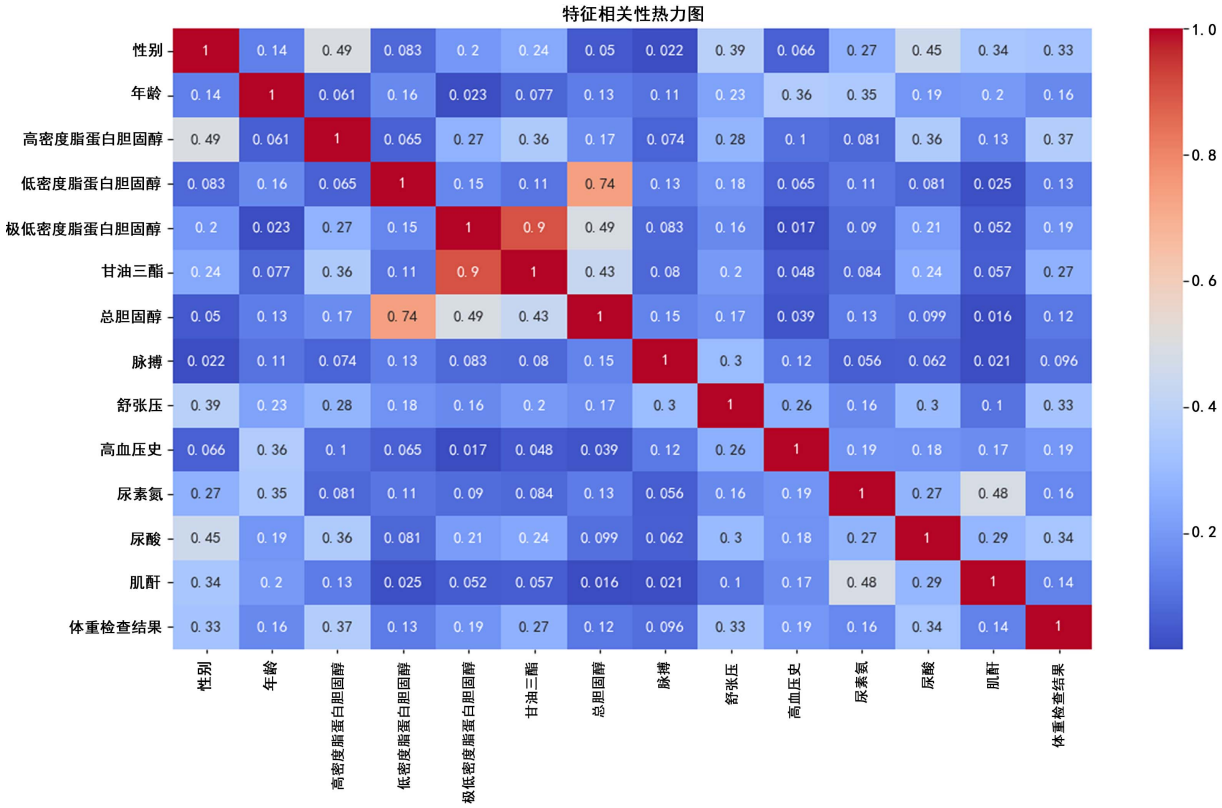

Figure 1. Feature correlation heatmap
图 1. F 特征相关性热力图

本研究在对特征进行相关性分析的基础上，采用递归特征消除对特征进行降维处理，通过实施特征选择策略，有效应对了特征冗余带来的精度衰减以及模型可解释性的挑战。具体操作中，借助递归特征消除方法逐步精简特征集，并以随机森林分类器为底层模型，以期达到更优的平衡点，对最不重要的特征进行递归移除，根据特征重要性筛选出用于糖尿病预测的八个最优特征为性别、年龄、高密度脂蛋白胆固醇、极低密度脂蛋白胆固醇、脉搏、舒张压、尿酸、肌酐。该特征组合的临床合理性分析如下：

人口学特征：年龄是糖尿病的经典危险因素，随着年龄增长，胰岛素分泌能力下降、胰岛素抵抗加重，患病风险显著升高；性别差异在糖尿病发病机制中存在明确影响，男性不良生活习惯(如吸烟、饮酒)发生率更高，女性在更年期后雌激素水平下降导致代谢紊乱，二者均为临床筛查的基础指标；

血脂指标：高密度脂蛋白胆固醇被称为“好胆固醇”，其降低会导致脂质代谢紊乱，与胰岛素抵抗密切相关，是糖尿病的重要预测因子；极低密度脂蛋白胆固醇升高会增加胰岛素抵抗风险，且与糖尿病大血管并发症的发生发展相关，二者联合可全面反映脂质代谢对糖尿病发病的影响；

生命体征：舒张压升高与胰岛素抵抗存在双向关联，高血压与糖尿病常互为并发症，共同源于代谢

综合征；脉搏异常(过快或节律不齐)可能反映自主神经功能紊乱，而糖尿病早期即可出现神经病变，二者的关联符合临床病理生理机制；

肾功能指标：尿酸升高可通过损伤胰岛 β 细胞功能、加重胰岛素抵抗诱发糖尿病，且糖尿病肾病是常见并发症，尿酸、肌酐、尿素氮作为肾功能核心指标，其异常可早期提示代谢紊乱状态，其中肌酐清除率下降与糖尿病风险升高呈正相关，尿素氮水平变化反映蛋白质代谢异常，与糖尿病患者的营养状态及肾脏灌注密切相关。

3.3. 模型构建

本研究分别使用不同的机器学习模型对糖尿病预测的参数设置进行不同选择，采用网格搜索与 5 折交叉验证进行参数调优，当机器学习模型预测参数设置不同时，通过探究机器学习模型的参数配置，可挖掘出最大化预测精度的方案，具体配置如表 2。进而，依据所得预测性能的优劣，甄选出综合表现最佳的模型。

Table 2. Basic information of model construction
表 2. 各模型构建的基础信息

模型名称	构建基础(库/类)	模型构建配置	参数探索空间
逻辑回归模型	Scikit-learn 库 LogisticRegression 类	初始化配置正则化参数 'l2'，通过优化算法迭代更新参数，最小化预测值与实际值差异	正则化强度 C：[0.1, 1.0, 10.0]；优化算法 solver：['lbfgs', 'liblinear', 'sag']
决策树模型	Scikit-learn 库 DecisionTreeClassifier 类	设置 random_state = 42 保证可重复性，通过控制树深度与节点分裂条件优化模型复杂度	最大深度 max_depth：[2, 3, 5, 7, 10]；最小样本分裂数 min_samples_split：[2, 5, 10, 15]
随机森林模型	Scikit-learn 库 Random-ForestClassifier 类	设置 random_state = 42 保证实验可重复性	决策树数量 n_estimators 考察(50、100、200)三个梯度；最大深度 max_depth 测试(None、5、10、15)四种配置(None 表示不限制深度)；最小样本分裂数 min_samples_split 评估(2、5、10)三个阈值
支持向量机模型	Scikit-learn 库 SVC 类	设置 probability = True 启用概率估计 random_state = 42 保证可重复性	正则化参数 C：[0.1, 1, 10, 100]；核函数 kernel：['linear', 'rbf', 'poly']；gamma 参数：['scale', 'auto']
XGBoost 模型	xgboost 模块 XGBClassifier 类	设置 use_label_encoder = False 避免标签编码偏差，eval_metric = 'logloss' 作为优化指标	学习率 learning_rate：[0.01, 0.1, 0.2, 0.3]；树的最大深度 max_depth：[3, 5, 7, 9]；基学习器数量 n_estimators：[50, 100, 200]

3.4. 参数调优

为提升模型性能，采用网格搜索(GridSearchCV)结合 5 折交叉验证进行参数调优。各模型的参数搜索空间如表 3 所示，以 F1-Score 作为核心优化指标(平衡精确率与召回率)。

Table 3. Optimal parameter combinations
表 3. 最优参数组合

模型名称	最优参数组合	训练集交叉验证 F1 分数
逻辑回归模型	C = 0.1, solver = 'liblinear'	0.7528 (±标准差)
决策树模型	max_depth = 5, min_samples_split = 5	0.7583 (±标准差)
随机森林模型	n_estimators = 200, max_depth = None, min_samples_split = 2	0.8030
支持向量机模型	C = 10, kernel = 'rbf', gamma = 'scale'	0.7906
XGBoost 模型	learning_rate = 0.2, max_depth = 7, n_estimators = 100	0.7963 (±标准差)

3.5. 模型评估

考虑到糖尿病预测的医疗场景需求，采用准确率(Accuracy)、召回率(Recall)、精确率(Precision)、F1-Score 和 ROC 曲线下面积(AUC)五个指标综合评估模型性能。各个模型的 ROC-AUC 值如下图 2 所示。

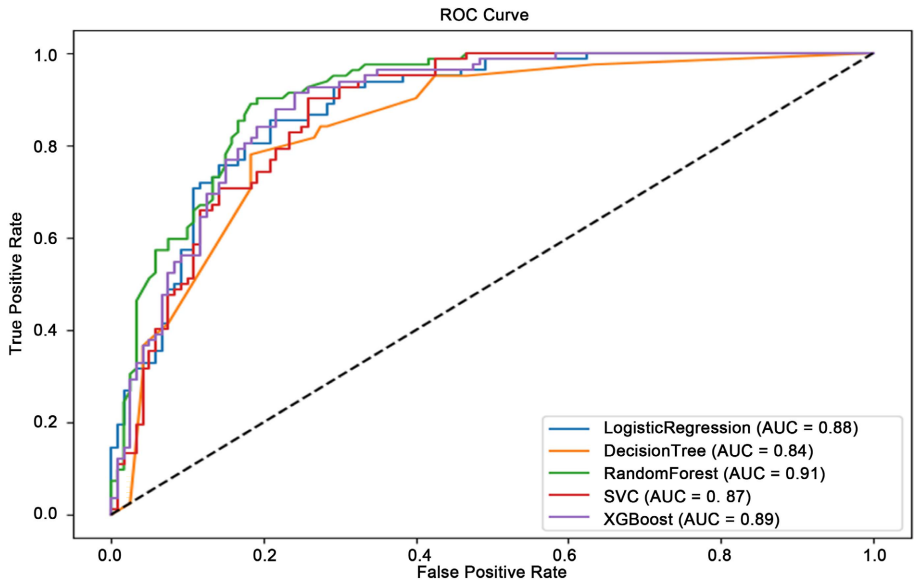

Figure 2. ROC values of different models
图 2. 各个模型 ROC 值

各个模型在测试集的指标如表 4 所示。

Table 4. Test set performance metrics of each model
表 4. 各模型测试集性能指标

模型名称	Accuracy	Recall	F1-Score	ROC-AUC
逻辑回归	0.8069	0.8293	0.7771	0.8834
决策树	0.7970	0.7683	0.7545	0.8446
随机森林	0.8416	0.8902	0.8202	0.9096
支持向量机	0.7772	0.7927	0.7429	0.8734
XGBoost	0.7772	0.8293	0.7861	0.8880

由表 4 可知，随机森林模型在各项评估指标中均表现最优：

1) 准确率：随机森林准确率为 84.16%，较次优的 XGBoost 模型的 81.68%提升 2.48 个百分点，表明其整体预测正确性最高；

2) 召回率：随机森林召回率为 89.02%，显著高于其他模型，意味着其漏诊率最低 10.98%，更符合医疗场景中“尽可能识别高危人群”的需求；

3) F1-Score：随机森林 F1-Score 为 82.02%，平衡了精确率与召回率，避免了“重漏诊轻误诊”或“重误诊轻漏诊”的问题；

4) AUC 值：随机森林 AUC 值为 0.9096，接近理想值 1，表明其区分糖尿病患者与非糖尿病患者的能力最强。

随机森林模型表现优异的原因主要有两点：一是通过集成多棵决策树，有效降低了单一模型的方差，提升了泛化能力，避免过拟合；二是其随机特征选择机制，减少了特征间的相关性干扰，更适合处理医疗数据中的多特征交互场景。因此，本研究确定随机森林模型为糖尿病风险预测的最优模型。

4. 结论

本研究围绕糖尿病风险预测展开，深入分析了五种主流机器学习算法，逻辑回归，决策树，随机森林，支持向量机和 XGBoost 在糖尿病预测任务中的表现，通过对比实验说明随机森林在学习预测精度和分类的准确度均优于其他 4 种方法，预测准确率达到 84.16%，AUC 值为 0.9096，预测效果突出。特别是通过递归特征消除和相关性分析，从原始 15 个特征中筛选出 8 个关键指标，既提高了模型效率，又增强了结果的可解释性。

虽然在某些方面取得了一定的进展，然而部分研究领域依然存在需要改进的地方，首先，所使用的数据集来源于单一公开数据库，样本量有限(1006 例)，且缺乏时间维度与外部验证集，可能导致模型在更广泛人群、不同地域或新兴风险因素(如生活方式、遗传背景)上的泛化能力不足；其次，数据中未包含血糖相关指标(如空腹血糖、糖化血红蛋白)，这虽有助于探索“非血糖指标”的预测潜力，但也限制了与现行临床诊断标准的直接对比；此外，所有特征均为静态临床指标，未能纳入动态变化数据(如连续监测数据、随访记录)，影响了模型在病程进展预测方面的适用性。未来研究可从以下几方面推进：一是整合多源异构数据，如电子健康记录、基因组学、穿戴设备监测的生活习惯数据等，构建更全面的风险画像；二是开展前瞻性队列研究或外部多中心验证，评估模型在真实世界中的预测效能与稳定性；三是开发轻量化、可解释的部署工具(如集成 SHAP、LIME 的可视化界面)，推动模型在基层医疗场景中的落地应用；四是关注模型公平性与伦理问题，确保其在不同性别、年龄、种族群体中的预测性能均衡，避免因数据偏差带来的健康不平等。此外不妨尝试运用更多的基础模型，或者探索其他更为先进的机器学习技术(如深度学习中的神经网络)，以此来提高预测的精准程度以及模型的泛化能力。

参考文献

[1] World Health Organization (2023) Global Report on Diabetes. World Health Organization.

[2] International Diabetes Federation (2025) IDF Diabetes Atlas. 11th Edition, International Diabetes Federation. https://www.diabetesatlas.org

[3] 吴晖南, 陈淑娇, 陈展峰, 等. 基于 LightGBM 模型的糖尿病预测模型的研究[J]. 中国卫生标准管理, 2023, 14(24): 64-67.

[4] Deo, R.C. (2015) Machine Learning in Medicine. *Circulation*, **132**, 1920-1930. https://doi.org/10.1161/circulationaha.115.001593

[5] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I. and Chouvarda, I. (2017) Machine Learning and Data Mining Methods in Diabetes Research. *Computational and Structural Biotechnology Journal*, **15**, 104-116.

https://doi.org/10.1016/j.csbj.2016.12.005

[6] Nabila, N., Islam, M., Hossain, M., *et al.* (2021) An Intelligent System for Diabetes Prediction: A Machine Learning Approach Using Clustering and Naive Bayes. *International Journal of Advanced Computer Science and Applications*, **12**, 78-85.

[7] Kui, L., Zhang, M., Wang, S., *et al.* (2022) Evaluation of Machine Learning Models for Predicting Blood Glucose Levels and Detecting Adverse Glycemic Events in Diabetes Management. *Journal of Medical Systems*, **46**, 1-12.

[8] 苗丰顺. 基于集成学习的糖尿病风险预测系统设计与实现[D]: [硕士学位论文]. 济南: 山东师范大学, 2023.

[9] 刘建平. 基于 Uniappand Django 的糖尿病预测系统设计[J]. 微型电脑应用, 2023, 39(8): 124-127.

[10] 仵豪. 基于 Stacking 融合算法的糖尿病风险早期识别模型研究[J]. 计算机工程与应用, 2023, 59(15): 235-242.

[11] 马吉聪. 基于 XAR-Stacking 融合模型的糖尿病预测系统开发[J]. 计算机应用与软件, 2024, 41(2): 189-196.