文章编号:1003-6199(2026)01-0151-07 DOI:10.16339/j.cnki.jsjsyzdh.202601024

基于改进残差网络和SHAP的糖尿病预测及可解释性分析

魏国政[1],魏丽丽[2],宋廷强[1†],渠蓉蓉[1],孙媛媛[3],董凡琦[1]

(1.青岛科技大学信息科学技术学院,山东 青岛 266061;2.青岛大学附属医院院长办公室,山东 青岛 266000;
3.青岛科技大学数据科学学院,山东 青岛 266061)

摘 要:针对糖尿病预测领域中可靠性与可解释性不足问题,提出了基于改进深度残差网络的预测算法。该算法嵌入了根据数据集特性设计的特征自注意力机制,并辅以SHAP模型以增强可解释性。SHAP能够精准定位并可视化影响糖尿病预测的关键因素,提升预测逻辑的透明度与实用价值。实验在Pima公开数据集及青岛某三甲综合医院私有数据集上展开,RAC模型与朴素贝叶斯、逻辑回归、支持向量机等模型进行了对比。结果显示,RAC的分类准确率、灵敏度、特异性、F_1分数值均优于其他模型,验证了其在临床实践中早期预警或辅助诊断的潜力。

关键词:糖尿病预测;可解释性;改进深度残差网络;特征自注意力机制;SHAP模型

中图分类号:TP301 **文献标识码**:A

Diabetes Prediction and Interpretability Analysis Based on Improved Residual Network and SHAP

WEI Guozheng[1], WEI Lili[2], SONG Tingqiang[1†], QU Rongrong[1], SUN Yuanyuan[3], Dong Fanqi[1]

(1. College of Information Science and Technology, Qingdao University of Science and Technology, Qingdao, Shandong 266061, China;
2. Office of the Dean, The Affiliated Hospital of Qingdao University, Qingdao, Shandong 266000, China;
3. College of Data Science, Qingdao University of Science and Technology, Qingdao, Shandong 266100, China)

Abstract: To address the lack of reliability and interpretability in the field of diabetes prediction, a prediction algorithm based on improved deep residual network is proposed. The algorithm embeds a feature self-attention mechanism designed according to the characteristics of the dataset, and is complemented by the SHAP model to enhance the interpretability, which can pinpoint and visualise the key factors affecting the prediction of diabetes mellitus, and enhance the transparency and practical value of the prediction logic. The experiments were carried out on the public dataset of Pima and the private dataset of a tertiary general hospital in Qingdao, and the RAC model was compared with the plain Bayes, logistic regression, and support vector machine models. The results show that the classification accuracy, sensitivity, specificity, and F1 score values of RAC are better than those of other models, validating its potential for early warning or assisted diagnosis in clinical practice.

Key words: diabetes prediction; interpretability; improved deep residual network; feature self-attention mechanism; SHAP model

收稿日期:2024-08-09
基金项目:国家自然科学基金青年项目(32301702);中华护理学会科研项目(ZHKY202118);山东省护理学会科研项目(SDHLKT202209)
作者简介:魏国政(1999—),男,山东济宁人,硕士,研究方向:人工智能与医学交叉应用领域。
† **通信联系人**,E-mail: songtq@qust.edu.cn

糖尿病作为一种全球性的代谢性疾病，其预测和监测对于早期干预至关重要。建立高准确率的预测模型[1-4]能够提前识别高危人群，有效控制糖尿病的发病率，能为大众早期糖尿病风险预警提供有效工具，对于辅助医生进行诊断具有重要现实意义。

目前，利用机器学习算法实现糖尿病预测已有大量研究。Iyer 等[5]分别采用了决策树和朴素贝叶斯模型在 Pima 糖尿病数据集上对糖尿病进行预测，分类准确率分别为 74.8% 和 79.5%。Zheng 等[6]设计了 Multivariate Bayesian logistic regression 预测模型，Specificity 和 Sensitivity 分别达到 75% 和 66%。Yan 等[7]使用 Logistic regression 算法设计预测模型，Sensitivity 达到了 70.6%。Kumari 团队[8]开发了依托 SVM 算法的糖尿病预测模型，准确率达到 78%。Xiong 等[9]则采取 LightGBM 技术对临床数据训练预测模型，Specificity 和 Sensitivity 分别达到 99.5% 和 88.3%。尽管以上研究取得了进展，但现有的传统模型在诸如准确率等关键性能指标上大多未超过 90%。鉴于以上研究，许多研究者开始聚焦于开发既高效又具有可解释性的糖尿病预测模型。例如，王鑫等[10]利用 LightGBM 算法构建了糖尿病预测模型，并使用 SHAP 技术对影响糖尿病的关键因素进行了分析，实现了模型的可解释性增强。

尽管糖尿病预测领域持续涌现新进展，但现有预测模型仍面临多重局限性。首先，预测模型结构简单及数据不平衡[11-13]导致预测精度不高；其次，多数预测方法为黑箱[14]机器学习模型，阻碍了临床医生对结果的理解与信任[15-16]。为解决上述问题，本文的主要贡献如下。

(1)鉴于糖尿病数据集的特性，针对性地改进了深度残差网络架构，将残差网络原始结构中的卷积层改进为更适应数据集特性的全连接模块。

(2)自主设计并嵌入了符合数据集特性的特征自注意力模块于改进残差网络内，强化模型对糖尿病关键特征的捕捉与调整，实现模型内部动态的特征处理。

(3)融合 SHAP 模型增强了糖尿病风险预测的可解释性[17]，同时实现了对糖尿病影响因子的可视化分析。

1 模构型建

1.1 融合特征自注意力机制的改进深度残差网络

本文引入了残差网络思想[18]和针对糖尿病数据集设计的特征自注意力模块来解决糖尿病预测中的退化和性能问题。在残差网络中，目标底层映射被表述为 $H(x)$。基于多层非线性变换能够逼近复杂函数的假设，这意味着它们能逼近一个特定的残差函数：

$$F(x) = H(x) - x \tag{1}$$

图 1 是本文针对糖尿病数据集设计的特征自注意力模块。特征自注意力机制旨在让模型学会如何在给定输入特征集合中分配注意力，从而强调对预测任务重要的特征并抑制冗余的信息。本文在每个残差块后添加了该特征自注意力模块。

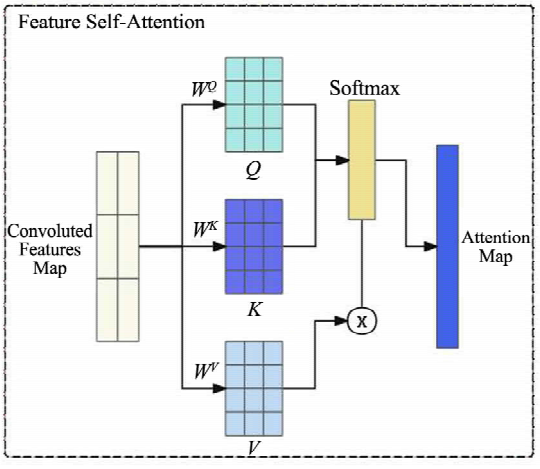


图 1 糖尿病数据集上的特征自注意力模块

假设输入特征向量为 $\boldsymbol{X} \in \boldsymbol{R}^{d \times N}$ 其中 d 是特征维度，N 是样本数量。自注意力机制首先通过三个可学习的权重矩阵 $\boldsymbol{W}^Q$、$\boldsymbol{W}^K$、$\boldsymbol{W}^V$ 将输入特征映射到查询、键和值空间，这三个空间的维度分别为 d_q、d_k、d_v。具体来说：

$$\boldsymbol{Q} = \boldsymbol{X}\boldsymbol{W}^Q \tag{2}$$

$$\boldsymbol{K} = \boldsymbol{X}\boldsymbol{W}^K \tag{3}$$

$$\boldsymbol{V} = \boldsymbol{X}\boldsymbol{W}^V \tag{4}$$

然后使用点积注意力函数来计算特征之间的相关性得分，注意力得分矩阵 A 可以通过下面的公式计算得到：

$$\boldsymbol{A} = \mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right) \tag{5}$$

这里的 softmax 函数用于归一化注意力得分，使其成为概率分布。接下来将注意力权重矩阵 $\boldsymbol{A}$ 与值矩阵 $\boldsymbol{V}$ 相乘来得到加权后的特征表示 $\boldsymbol{Z}$：

$$\boldsymbol{Z} = \boldsymbol{A}\boldsymbol{V} \tag{6}$$

最终，将加权后的特征表示 Z 与原始输入特征 X 进行融合，以保留原始信息的同时增强重要

特征的表示能力。这个过程可以通过简单的元素级相加或通过额外的可学习参数矩阵 $\boldsymbol{W}^O$ 进行线性变换完成：

$$\boldsymbol{X}' = \boldsymbol{X}\boldsymbol{W}^O + \boldsymbol{Z} \tag{7}$$

结合以上步骤，最终定义特征自注意力机制的完整计算流程为：

$$\boldsymbol{X}' = \boldsymbol{X} + \mathrm{softmax}\left(\frac{(\boldsymbol{X}\boldsymbol{W}^Q)(\boldsymbol{X}\boldsymbol{W}^K)^T}{\sqrt{d_k}}\right)(\boldsymbol{X}\boldsymbol{W}^V) \tag{8}$$

在这个过程中，$\boldsymbol{W}^Q$、$\boldsymbol{W}^K$、$\boldsymbol{W}^V$、$\boldsymbol{W}^O$ 都是需要学习的参数，它们通过反向传播算法进行优化，以最小化网络的预测误差。

图 2 所示为 RAC 网络结构图。由于糖尿病数据集通常为 csv 格式，和图像数据集中各个特征承载的含义不同，在实验反复验证的过程中发现应用卷积运算会对性能产生负优化。而全连接层能将输入数据映射到输出空间，并通过学习权重参数来实现模式识别和特征提取。残差块中执行全连接层的可行性得到了文献[19]的支撑。因此本文用全连接层改进原结构中的卷积层以优化针对糖尿病数据集的预测性能。融合特征自注意力机制能使模型诊断更符合现实的医疗诊断经验，使得各个体检指标都能得到全局信息，增强了模型对数据内在联系的捕捉能力。改进后的 RAC 网络公式如下：

$$y = F'(x_i, W_i') + \alpha \cdot x \tag{9}$$

$$z_1 = W_2' \cdot \sigma(\mathrm{BN}(W_1' \cdot x_1 + b_1')) + b_2' \tag{10}$$

$$F' = \sigma(\mathrm{BN}(z_1)) + \mathrm{SelfAttention}(x) \tag{11}$$

式中，α 是一个可学习的标量，用于调整残差连接的强度；σ 是激活函数；b_1' 和 b_2' 为全连接层的偏执；BN 表示批量归一化操作；$\mathrm{SelfAttention}(x_i)$ 即本文的特征自注意力机制。

1.2　引入 SHAP 可解释性模型

研究所提出的 RAC 模型在训练阶段表现出了优异的预测性能。然而，若未能提供预测结果的解释依据，可能会引发使用者的信任壁垒。为此，本研究在 RAC 模型的基础上整合了 SHAP 模型，后者以其高度可解释性而著称。

SHAP 由 Lundberg 等[20]所创，引入了 SHAP 值这一通用性强的解释技术，用以解构模型特别是难以理解的“黑箱”模型。其核心在于评估特征对模型输出的边际贡献。考虑 x_i 作为第 i 个样本，其第 j 个特征为 x_{ij}，mc_{ij} 定义为特征边际贡献，边的权重为 w_i，SHAP 值 $f(x_{ij})$ 为 x_{ij} 对预测的贡献，以第 i 个样本的首个特征的 SHAP 值计算为例。

$$f(x_{i1}) = \mathrm{mc}_{i1}\ w_{l_1} + \cdots + \mathrm{mc}_{i1}\ w_n \tag{12}$$

模型对样本 i 的预测值为 y_i，整个模型的均值基线设为 y_{base}，与各特征的 SHAP 值相关，具体体现在下式中。

$$y_i = y_{\mathrm{base}} + f(x_{i1}) + f(x_{i2}) + \cdots + f(x_{is}) \tag{13}$$

式中，$f(x_{i1})$ 即第 i 个样本中首个特征对预测结果 y_i 的贡献值。SHAP 模型能够为预测结果提供特征权重的可视化，为医护人员提供直观的决策支持和更可信赖的工具。

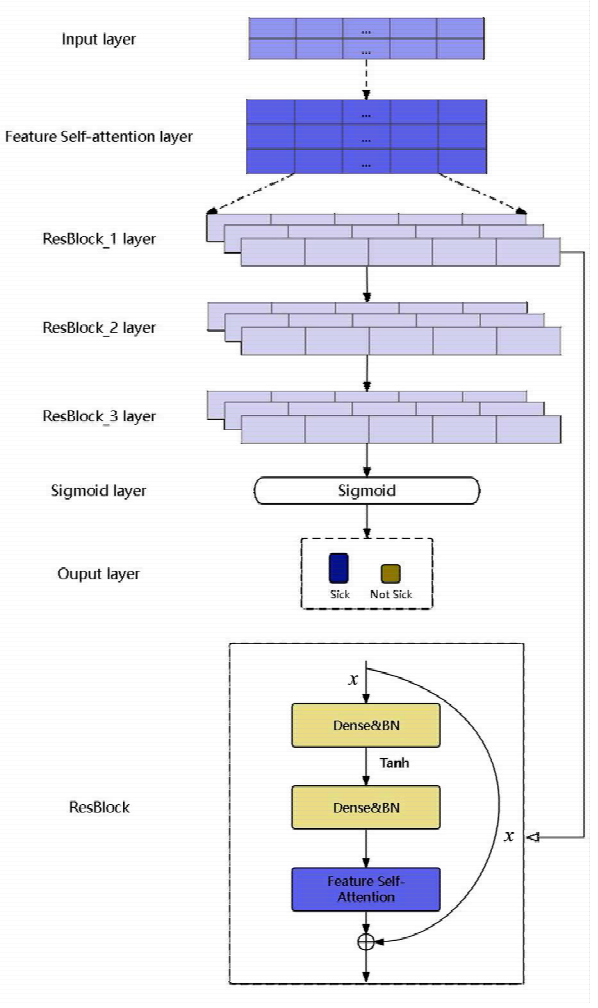


图 2　RAC 网络结构图

2　数据选择与处理

2.1　数据来源

实验在 Pima 公开数据集和青岛某三甲综合医院收集的私有数据集上分别验证。Pima 数据集

共计 768 个样本，分为 268 个糖尿病患者和 500 个非患者，覆盖 8 个特征和 1 个标签。青岛某三甲综合医院提供的私有数据集中包含 4807 个样本，包含 3006 个糖尿病患者和 1801 个非患者，涵盖了 60 余项特征和 1 个标签。该数据集综合性较强，从基本的身体指标到复杂的环境与行为互动，再到医疗与遗传背景，全方位覆盖了可能影响妊娠期糖尿病发病风险的因素。

2.2 数据预处理

2.2.1 缺失数据填充及数据标准化

Pima 数据集的缺失值可视化结果如图 3 所示。通过观察发现，特征变量 Glucose、Blood Pressure、BMI、Skin Thickness 和 Insulin 均存在缺失值。处理缺失特征时，将数据集分为患者组与非患者组两大类别，并根据不同类别中位数进行分别填充操作。私有数据集中无缺失情况出现。鉴于两类数据集中各属性特征单位不一致，为消除单位差异并统一属性间比较基准，同时规避异常值对模型训练的噪声影响，选择使用 RobustScaler 方法对数据进行标准化处理。

2.2.2 特征选择

面对私有数据集中繁复的特征构成，本文采用皮尔逊相关系数统计方法，来量化并解析特征间的相互依存关系。相关系数绝对值趋向于 1 时，明确指示了特征间的高度相关性，反之则揭示独立性。去除冗余特征有助于减少模型对噪声和异常值的敏感度，增强模型的鲁棒性。最终选取了权重排名前 24 项作为核心特征集(图 4)用于后续的仿真实验。

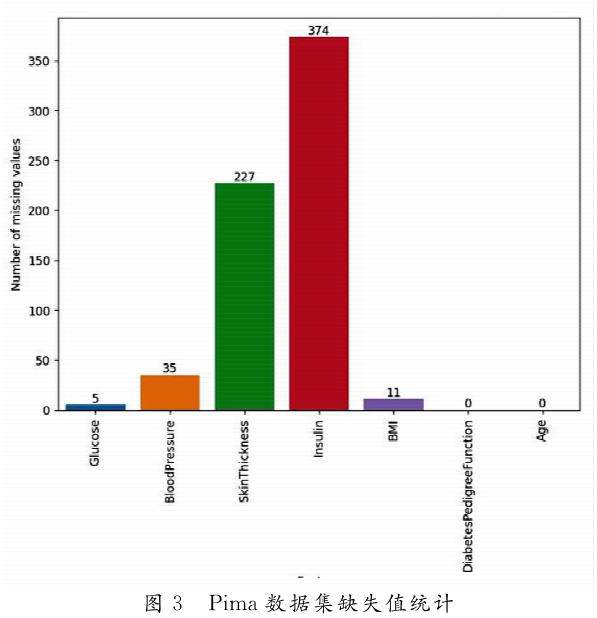


图 3 Pima 数据集缺失值统计

3 仿真实验与结果分析

3.1 评价指标

在模型构建过程中选择特异度(Specificity)、灵敏度(Sensitivity)、F_1-score 和准确率(Accuracy)作为评价指标，旨在全面考察模型在识别糖尿病和非糖尿病病例方面的能力。特异度反映的是模型正确识别非糖尿病病例的能力，即避免误诊为糖尿病的能力；而灵敏度反映的是模型正确识别糖尿病病例的能力；F_1-score 在数据集类别不平衡的情况下综合考虑了模型的性能；准确率能够直观全面的了解模型整体性能。同时关注这四项评价指标以确保模型在实际应用中能够做出真实有效的判断。

3.2 消融实验

本节构建了 RAC 模型的消融实验，旨在比较有无残差结构和特征自注意力模块情况下模型的各项指标表现。实验在相同预处理后的 Pima 数据集上进行，分别构建了针对残差结构和特征自注意力机制的消融实验。结果表明，带有残差结构和嵌入特征自注意力模块的模型性能指标均得到了显著提升。具体结果如表 1、表 2 所示。

表 1 有无残差结构的性能比较

标签	Res	Non-Res
准确率	92.87%	83.52%
F_1-score	93.37%	84.73%
特异度	99.23%	92.34%
灵敏度	93.45%	78.82%

表 2 有无特征自注意力结构的性能比较

标签	FSAM	Non-FSAM
准确率	92.87%	90.61%
F_1-score	93.37%	89.67%
特异度	99.23%	97.28%
灵敏度	93.45%	86.17%

3.3 多模型对比实验

本节构造了 RAC 与 Random Forest、SVM、Logistic Regression、Naive Bayes 及 Decision Tree

算法于糖尿病预测任务中性能表现的对比实验。实验均采用一致预处理流程的数据集进行训练与评估。通过 Specificity、Sensitivity、F_1-score 与 Accuracy 四项指标，表 3、表 4 分别展示了各算法模型在公开数据集、私有数据集对比下的性能表现。上述实验结果表明，RAC 模型在各项关键指标上展现出显著的性能提升，充分证实了其在糖尿病预测领域的优越潜力与实际应用价值。

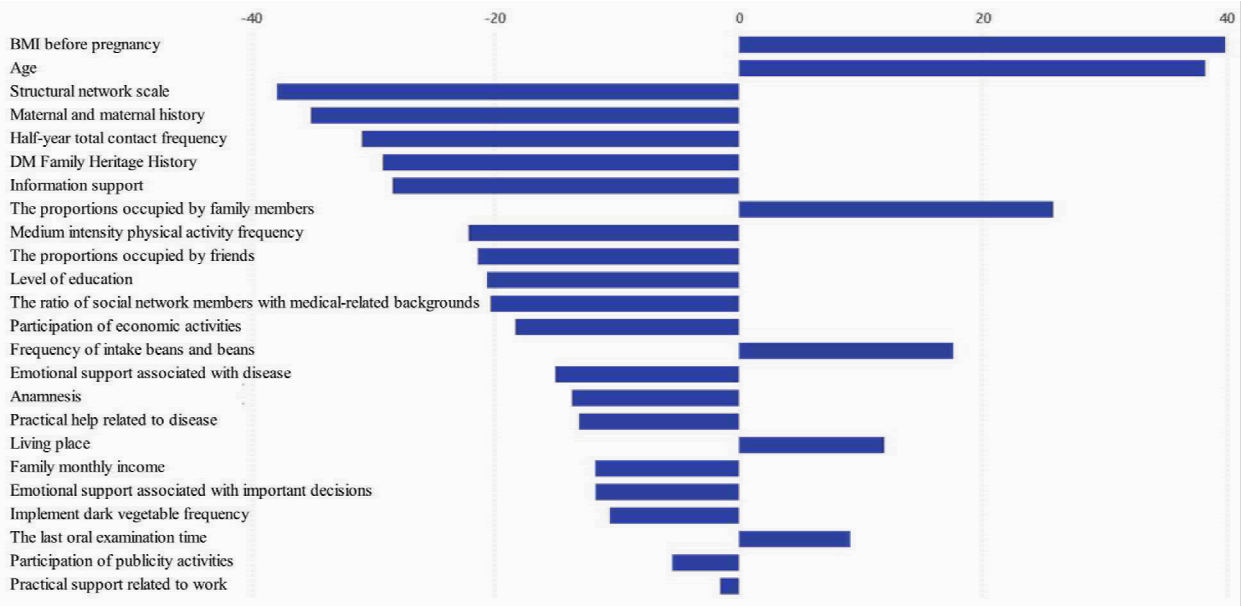


图 4　特征权重可视化排序

表 3　基于公开数据集的模型性能比较

模型	特异度	灵敏度	F_1-score	准确率
RAC	99.23%	93.45%	93.37%	92.87%
Random forest	80.70%	67.27%	73.14%	75.97%
SVM	87.86%	56.36%	66.28%	76.62%
Logistic regression	78.69%	67.27%	74.36%	74.67%
Naive Bayes	79.68%	70.90%	76.11%	76.62%
Decision tree	75.70%	65.45%	70.54%	72.08%

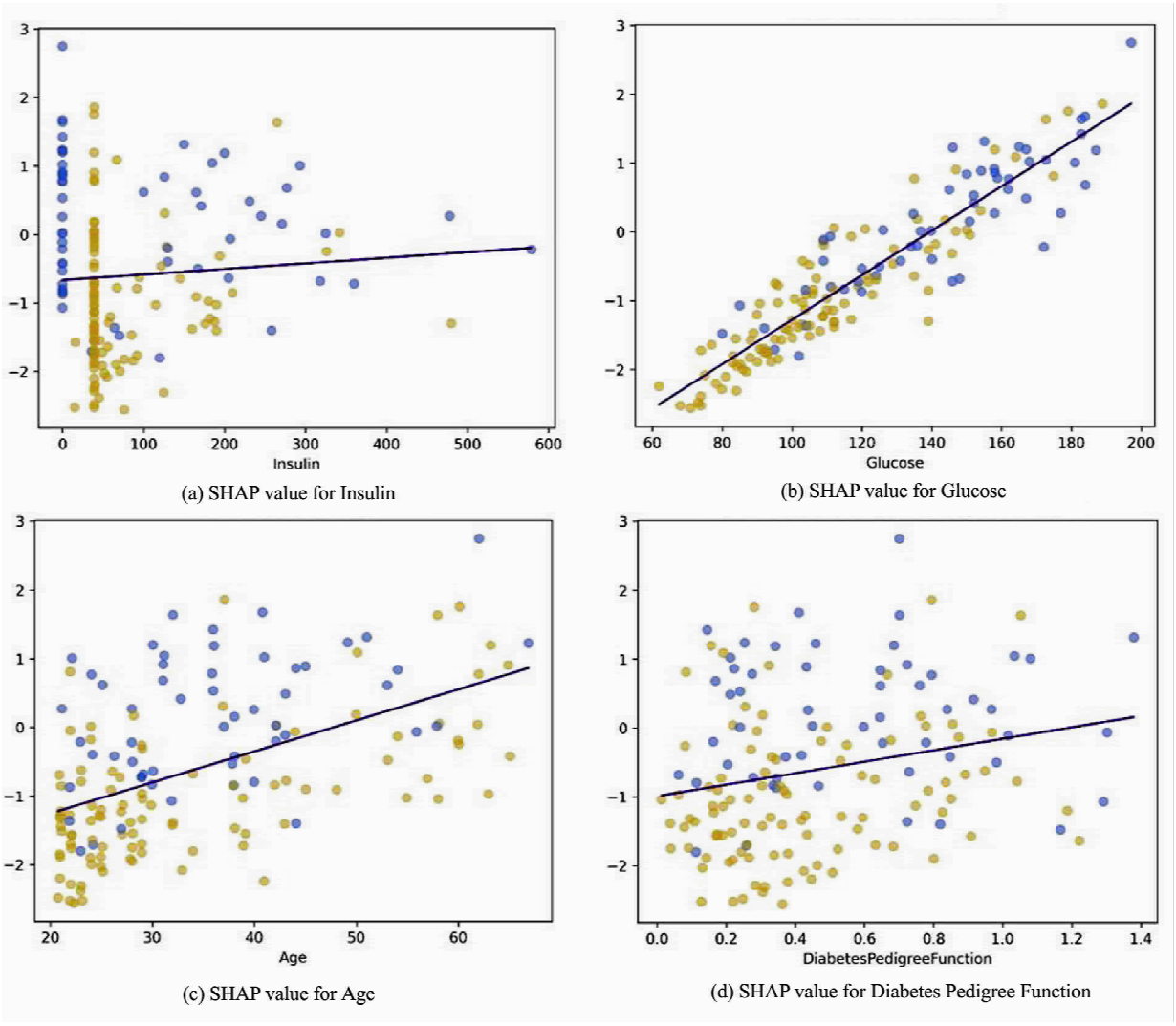
表 4　基于私有数据集的模型性能比较

模型	特异度	灵敏度	F_1-score	准确率
RAC	98.44%	89.58%	91.23%	91.74%
Random forest	94.22%	61.40%	72.27%	83.36%
SVM	88.92%	69.12%	76.17%	81.10%
Logistic regression	92.16%	73.14%	81.56%	81.53%
Naive Bayes	75.42%	69.90%	74.23%	75.86%
Decision tree	75.91%	65.97%	72.13%	73.11%

3.4　可解释性实现

图 5 展示了 SHAP 特征依赖分析后的结果，选取权重最大的四个特征进行展示。图中的深色直线表示“最佳拟合线”，反映特征值与糖尿病风险间强烈的正相关性，即特征值的升高对应着更高的患病风险，这一趋势符合临床诊断的预期[21]。

图 6 为 SHAP 可视化摘要图，它具体呈现了 Pima 数据集中某一确诊糖尿病病例各项指标对疾病风险预测的局部贡献度的 SHAP 值。该图按照各特征指标对糖尿病风险影响的重要性进行了排序，清晰揭示了 Glucose、Age、Pregancies、BMI 等指标对模型决策具有显著影响，明确这些因素为判定该对象罹患糖尿病较高风险的决定性条件。结果显示，SHAP 可视化不仅增强了模型的可信度，也能为医疗工作者在临床诊断中加速决策过程提供有力支持。



(a) SHAP value for Insulin

(b) SHAP value for Glucose

(c) SHAP value for Age

(d) SHAP value for Diabetes Pedigree Function

图 5 SHAP 特征依赖分析

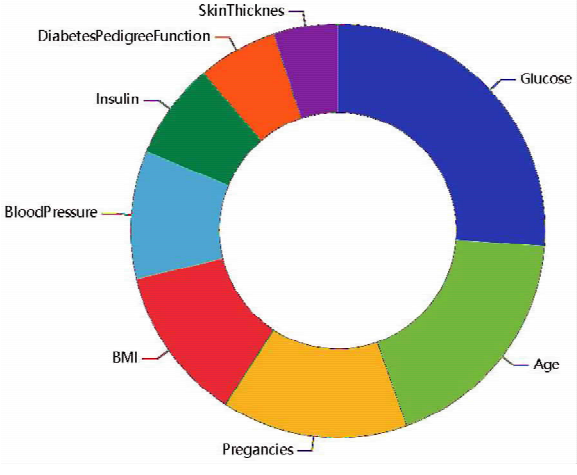


图 6 SHAP 可视化摘要图

4 结 论

提出了一个创新的预测模型，融合特征自注意力机制的改进深度残差网络 RAC。借助 SHAP 的加持，RAC 能够为医疗专家提供翔实的诊断见解，使糖尿病风险判断变得既精准又透明。为了验证 RAC 模型的有效性和优越性，本文将其在多个数据集上与一系列机器学习模型以及领域内其他研究者提出的先进方法进行了全面比较。实验结果显示，在多个评估指标上，RAC 模型均展现出显著的优势。展望未来，我们的研究重心将转向 RAC 模型在真实临床环境中的应用与优化，确保其能够无缝融入并服务于实际医疗决策。

参考文献

[1] CHAKI J, GANESH S T, CIDHAM S K, et al. Machine learning and artificial intelligence based diabetes mellitus detection and self-management: a systematic review [J]. Journal of King Saud University-Computer and Information Sciences, 2022, 34(6): 3204—3225.

[2] CAHN A, SHOSHAN A, SAGIV T, et al. Use of a machine learning algorithm improves prediction of progression to diabetes [J]. Diabetes, 2018, 67(Supplement 1).

[3] KHANWALKAR A, SONI R. A survey on prediction of diabetes using classification algorithms [J]. Journal of Achievements in Materials and Manufacturing Engineering, 2021, 104(2): 77—84.

[4] PRADHAN N, RANI G, DHAKA V S, et al. Diabetes prediction using artificial neural network [C]// Deep Learning Techniques for Biomedical and Health Informatics. Academic Press, 2020: 327—339.

[5] IYER A, JEYALATHA S, SUMBALY R. Diagnosis of diabetes using classification mining techniques [J]. ArXiv Preprint ArXiv: 1502.03774, 2015.

[6] ZHENG T, YE W P, WANG X P, et al. A simple model to predict risk of gestational diabetes mellitus from 8 to 20 weeks of gestation in Chinese women [J]. BMC Pregnancy and Childbirth, 2019, 19: 1—10.

[7] YAN J Z, GENG Y A, XU H X, et al. A prediction model of gestational diabetes mellitus based on first pregnancy test index [C]// Health Information Science: 9th International Conference, HIS 2020, Amsterdam, The Netherlands, October 20—23, 2020, Proceedings 9. Springer International Publishing, 2020: 121—132.

[8] KUMARI V A, CHITRA R. Classification of diabetes disease using support vector machine [J]. International Journal of Engineering Research and Applications, 2013, 3(2): 1797—1801.

[9] XIONG Y, LIN L, CHEN Y, et al. Prediction of gestational diabetes mellitus in the first 19 weeks of pregnancy using machine learning techniques [J]. The Journal of Maternal-Fetal & Neonatal Medicine, 2022, 35(13): 2457—2463.

[10] 王鑫，廖彬，李敏，等. 融合 LightGBM 与 SHAP 的糖尿病预测及其特征分析方法 [J]. 小型微型计算机系统，2022，43(9)：1877—1885.

[11] KHAKZAR A, LI Y W, ZHANG Y, et al. Analyzing the effects of handling data imbalance on learned features from medical images by looking into the models [J]. ArXiv Preprint ArXiv, 2022: 2204.01729.

[12] WEIDER C. L, KIKIN-GIL R, NORI H P. Method and system of correcting data imbalance in a dataset used in machine-learning: 20200380309[P]. 2020—12—03.

[13] HUANG C X, HUANG X, FANG Y, et al. Sample imbalance disease classification model based on association rule feature selection [J]. Pattern Recognition Letters, 2020, 133: 280—286.

[14] CASTELVECCHI D. Can we open the black box of AI? [J]. Nature News, 2016, 538(7623): 20.

[15] TONEKABONI S, JOSHI S, MCCRADDEN M D, et al. What clinicians want: contextualizing explainable machine learning for clinical end use [C]// Machine Learning for Healthcare Conference. PMLR, 2019.

[16] 陈小昆，左航旭，廖彬，等. 融合 XGBoost 与 SHAP 的冠心病预测及其特征分析模型 [J]. 计算机应用研究，2022，39(6)：9.

[17] CAKIROGLU C, DEMIR S, OZDEMIR M. H, et al. Data-driven interpretable ensemble learning methods for the prediction of wind turbine power incorporating SHAP analysis [J]. Expert Systems with Applications, 2024: 237.

[18] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[19] WU Y T, ZHANG C J, MOL B. W, et al. Early prediction of gestational diabetes mellitus in the Chinese population via advanced machine learning [J]. The Journal of Clinical Endocrinology & Metabolism, 2021, 106(3): e1191—e1205.

[20] LUNDBERG S M, LEE S I. A unified approach to interpreting model predictions [C]// Advances in Neural Information Processing Systems, 2017, 30.

[21] 中华医学会糖尿病学分会. 中国 2 型糖尿病防治指南(2017 年版)[J]. 中国实用内科杂志，2018，38(4)：53.