doi：10. 12068/j. issn. 1005－3026. 2025. 20240103

基于HRV多尺度分析的糖尿病前期检测方法

李鸿儒，李同同，石康康，杨英华
（东北大学 信息科学与工程学院， 辽宁 沈阳 110819）

摘 要：糖尿病前期（prediabetes）是糖尿病发展过程中葡萄糖代谢异常的重要阶段,及早诊断对于全球糖尿病防控至关重要. 为探索糖尿病前期的无创检测方法,基于心率变异性（heart rate variability, HRV）信号,通过引入多尺度分析策略,揭示信号的全局信息以及不同尺度内微小但重要的变化,并采用CatBoost算法完成分类任务. 结果表明,该方法在数据集上取得88. 52%的准确率、83. 40%的敏感度、91. 82%的特异度、86. 73%的精确度和87. 40%的$F1$分数. 本研究为糖尿病前期的诊断提供了新思路,尤其适用于可穿戴设备,为实现日常自我健康监测及疾病防控提供了潜在解决方案.

关 键 词：糖尿病前期；心率变异性；多尺度分析；小波散射网络；CatBoost算法
中图分类号：TP 391.5 **文献标志码**：A **文章编号**：1005-3026（2025）12-0019-10

Prediabetes Detection Method Based on Multi-scale Analysis of HRV

LI Hong-ru，*LI Tong-tong*，*SHI Kang-kang*，*YANG Ying-hua*
（School of Information Science & Engineering，Northeastern University，Shenyang 110819，China. Corresponding author：LI Hong-ru，E-mail：lihongru@ise.neu.edu.cn）

Abstract：Prediabetes is an important stage of abnormal glucose metabolism in the development of diabetes，and its early diagnosis is crucial for global diabetes prevention and control. To explore non-invasive detection methods for prediabetes，heart rate variability（HRV）signals were utilized. By introducing a multi-scale analysis strategy，the global information of the signals was revealed，as well as subtle but important changes at different scales. The CatBoost algorithm was used for classification task. The results show that this method achieves an accuracy of 88. 52%，a sensitivity of 83. 40%，a specificity of 91. 82%，a precision of 86. 73%，and an $F1$-score of 87. 40% on the dataset. This study provides a new approach for the diagnosis of prediabetes. The results are especially suitable for wearable devices，offering a potential solution for daily self-health monitoring and disease prevention.

Key words：prediabetes；heart rate variability；multi-scale analysis；wavelet scattering network；CatBoost algorithm

糖尿病是全球最为关注的、也是发展最为迅速的健康问题之一. 根据国际糖尿病联盟（IDF）[1]估计,在2021年,全球范围内20~79岁的成年人中,糖尿病患者占10.5%,即约5.36亿人. 这一趋势仍在不断上升,到2045年,患病人数预计将增至7.83亿,患病率将上升至12.2%. 全球的医疗支出也将从0.996万亿美元增加至1.054万亿美元. 由于内源性胰岛素的生产能力下降,与外周组织的胰岛素抵抗,糖尿病是根本无法治愈的,且会引发许多严重的并发症[2],包括神经病变、肾病、视网膜病变和心血管疾病等. 因此,糖尿病的管理和控制至关重要. 然而,现有的治疗主要是以预防与管理为主[3],而对于糖尿病前期的检测则成为预防糖尿病发展的重要手段. 糖尿

收稿日期：2024-05-05
基金项目：国家自然科学基金资助项目（62073062）.
作者简介：李鸿儒（1968—），男，内蒙古赤峰人，东北大学教授，博士生导师；杨英华（1970—），男，辽宁辽阳人，东北大学教授，博士生导师.

病前期是糖尿病发展的早期阶段,其特点是葡萄糖代谢异常,但尚未达到糖尿病的诊断标准[4-5].如果不及时检测和干预,将有很大风险罹患2型糖尿病[6].通过早期检测,可以采取必要的生活方式干预和药物治疗,使葡萄糖代谢恢复到正常水平,从而防止糖尿病的发生[7-8].此外,有研究表明,糖尿病前期人群患心血管疾病的风险明显增加[9-10].事实上,多达三分之一未确诊的糖尿病前期患者已被诊断出患有心血管并发症[11].因此,糖尿病前期的检测不仅可以预防糖尿病,还可以降低心血管疾病的风险[7-8].

2020年,Tobore等[12]同时采集心电图(ECG)与脑电图(EEG),手工提取了39个特征,用于检测糖尿病前期;Wang等[13]使用5 s的ECG图像并开发了IGRNet深度学习模型,进行糖尿病前期检测.研究表明,糖尿病前期与心脏自主神经系统的活动存在一定关联[14-15],HRV信号成为重要的研究方向[16].HRV信号是从ECG中提取出的生理信号,已被广泛认为是评估心脏自主神经功能的潜在生物标志物[17-18].Igbe等[19]测量了口服葡萄糖耐量试验(OGTT)前后的HRV信号,利用前后HRV信号的绝对幅值偏差来检测糖尿病前期.尽管前人的研究在揭示生理状态与疾病之间的关系方面提供了重要见解,但他们的方法仍存在一些问题和限制,所使用的生理信号需要相对复杂的实验流程和较长的时间,这限制了该方法在实际临床和日常健康监测中的应用.

为了解决这些问题并提供更加实用和精确的方法,本文提出使用5 min HRV信号来检测糖尿病前期,同时引入多尺度分析策略,以进一步提高检测效果.本文首先关注特定生理尺度的整体特征并进行分析,以捕捉HRV信号在特定生理尺度下的显著变化,这有助于理解信号的整体模式与糖尿病前期之间的关系.其次,为了更细致地揭示信号中的变化,本文利用小波散射网络将信号分解为不同尺度下的散射系数,以捕捉信号在精细化尺度内的微小变化,从而能够更准确地分析HRV信号的细节特征,揭示出信号中微弱但重要的模式.最后,本文构建了基于CatBoost算法的分类模型,将提取到的特定生理尺度的整体特征与精细化尺度特征结合起来,实现了分类任务.CatBoost算法能够有效地整合所提供的特征,进一步扩展特征维度,从而实现更精确的决策.该分类模型通过对多尺度特征学习到的信息进行综合考量,为糖尿病前期的检测提供了更准确的结果.

1 数据描述与处理

1.1 数据描述

本文使用的数据集是由大连医科大学附属第一医院和东北大学联合收集整理的私人数据库.数据来源于大连医科大学附属第一医院2020—2021年患者的电子健康记录(EHR)与单导联心电图(ECG)数据.本研究经过了医院伦理委员会的批准.在分析前,本文对患者数据进行了脱敏处理.EHR数据包含患者的生理参数和生化指标,具体包括性别、年龄、身高、体重、既往病史、腰围、身体质量指数(BMI)、肝功能生化指标、空腹血糖,以及心电参数(PR间期、QRS波群时限、心率)和病史信息.ECG数据是从深圳市博英医疗仪器科技有限公司的心电信息管理系统中获取的.患者的ECG数据以256 Hz的频率采样,时长在0~72 h不等.本文严格控制了实验条件,根据电子健康记录排除了患有心脏疾病、高血压、糖尿病、肾脏疾病等可能干扰HRV信号的样本,以及缺少ECG数据的样本.本文根据中华医学会糖尿病学分会制定的《中国2型糖尿病防治指南(2020年版)》[20]中糖尿病前期的诊断标准(空腹血糖值范围6.1~6.9 mmol/L)进行标签标注,最终获得254个样本(年龄为71.4±5.38岁),其中正常样本182例,糖尿病前期样本72例.

1.2 HRV信号的提取与预处理

心率极易受到影响,在人体进行活动、情绪产生波动时都会出现心率的变化.由于HRV信号与心率有着十分密切的联系,不同时刻、不同状态下的HRV信号存在着极大的差异,这种特性被称为心率变异性.图1为5 s的ECG信号,连续R峰之间的RR间期序列构成HRV信号.Cui等[21]针对这种波动作了深入的研究,结果表明:不同时间、不同运动甚至卧躺时不同的体位均会导致HRV产生较大波动,但在进入睡眠时期HRV会出现较长的平稳期.因此本文选取凌晨0点至6点这一时间段的HRV信号进行研究.根据欧洲心脏病协会工作组和欧洲心律协会的联合立场声明:5 min HRV信号通常用于临床检查和科研实验中,可以快速评估心脏自主神经系统的功能状态和心血管健康状况[22].经过上述处理,本研究最终得到糖尿病前期HRV信号1 240段,正常样本1 014段.

本文采用Pan-Tompkins算法[23]对5 min的ECG信号进行R波定位.图2中的圆圈为定位的R波波峰,从图中可以看出,该算法具有很高的检出率.

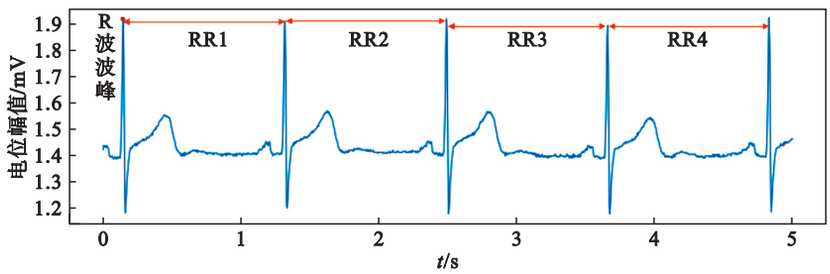

图1　5 s的ECG信号

Fig. 1　Five seconds of ECG signal

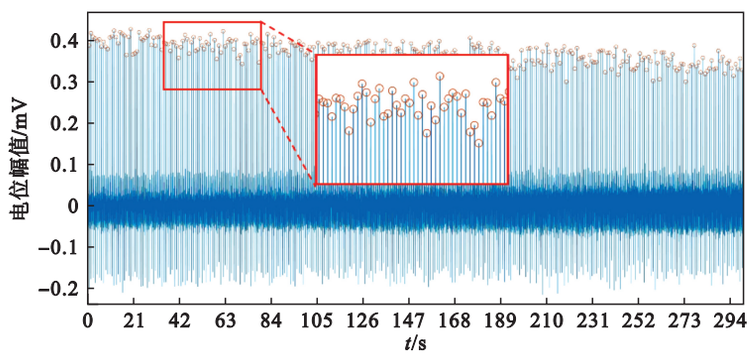

图2　R波定位

Fig. 2　R-wave localization

通过计算连续相邻R波波峰的时间差值来获取RR间隔的时间序列,即HRV信号.在人体的正常生理情况下,不可能出现RR间期大于1 700 ms或小于300 ms的情况,这些值被视为异常点.除此之外,还存在由异位搏动引起的异常,即后一个与前一个RR间期差的绝对值大于前一个RR间期的20%.运用线性插值法将上述异常RR间期值替换为前后RR间期的平均值.由于心动周期的波动,RR间期具有非均匀采样特性,对非均匀性序列进行多尺度分析时存在一定的影响,因而需要对数据进行均匀重采样.本文选择3次样条插值法对HRV信号进行2.4 Hz的重采样,得到均匀采样的HRV时间序列.图3是处理后的5 min HRV信号,可以看出,与正常人相比,该HRV信号具有复杂度较低、变化范围小的特点.

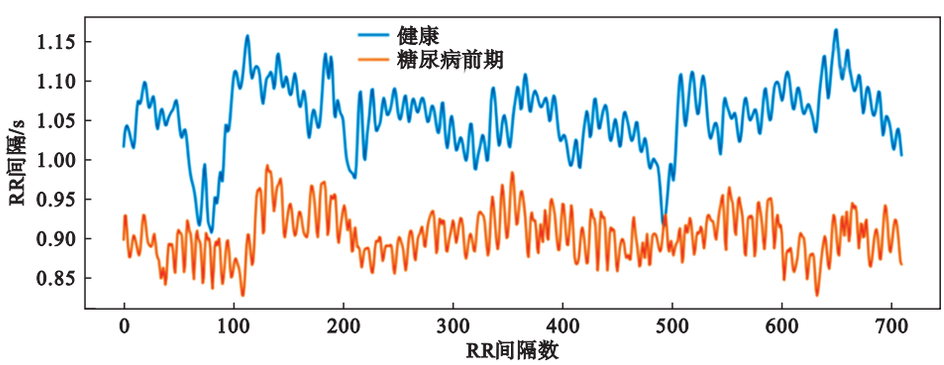

图3　健康和糖尿病前期的HRV信号

Fig. 3　HRV signals of health and prediabetes

2　方　　法

HRV信号作为一种反映心脏自主神经调节的生理指标,一直受到研究者的关注.传统频域分析法仅从特定尺度分析HRV信号,难以全面展现信号的多样特性.因此,本文利用小波散射网络将信号分解为不同的尺度成分,进而捕捉HRV信号中可能蕴含的微弱特征.本文的多尺度分析策略为糖尿病前期的检测提供了更全面的信息.这种

方法有望进一步提高糖尿病前期检测的准确性和敏感性,为早期诊断和治疗提供新的视角和工具.

2.1 频域分析法

从医学研究出发,将处理后的HRV信号利用频域分析法提取特定生理尺度特征.HRV信号的能量基本都集中在0~0.4 Hz的频谱范围,研究中计算极低频功率(VLF)、低频功率(LF)和高频功率(HF)3种成分,频段分别在0.003~0.04 Hz,0.04~0.15 Hz和0.15~0.40 Hz.在这3种成分中,高频功率能唯一反映副交感神经的调节作用,而低频功率同时反映交感神经和副交感神经的调节作用[24-25].此外,本文还计算了频段在0~0.4 Hz的总功率P_t、交感神经与副交感神经活性比R(R为LF与HF的比值)、归一化低频成分(LF_{nu})、归一化高频成分(HF_{nu}),来评价交感神经与副交感神经的平衡性能,其中:

$$LF_{nu}=\frac{LF}{P_t-VLF}\times 100\%, \tag{1}$$

$$HF_{nu}=\frac{HF}{P_t-VLF}\times 100\%. \tag{2}$$

对生理频段的VLF,LF和HF进行功率谱密度(power spectral density,PSD)估计,可以获取交感神经和副交感神经的活动状况.本文使用Welch周期图方法[26]估计HRV信号的PSD,该方法采用信号分段重叠、加窗、快速傅里叶变换(FFT)等技术来计算功率谱:首先将长度为N的时间序列$x(n)$分成有重叠的p段,每段包含M个数据点,用窗口函数对每一小段信号序列进行预处理,每一段信号序列修正的PSD计算式为

$$J_p(f)=\frac{\left|\sum_{n=0}^{M-1}x_p(n)w(n)\,\mathrm{e}^{-\mathrm{j}2\pi fn}\right|^2}{MU}. \tag{3}$$

其中:

$$U=\frac{\sum_{n=0}^{M-1}w(n)^2}{M}; \tag{4}$$

$J_p(f)$表示第p个子段的功率谱密度估计值,f为频率变量;$x_p(n)$表示第p个子段的信号样本序列;$w(n)$为加窗函数.对p段的周期图进行平均,得到整个信号的功率谱估计.

2.2 小波散射网络

频域分析法提取的是HRV信号生理尺度的整体信息.由于人体差异性,可能两类样本的整体尺度能量分布存在相同的情况,此时仅仅用整体信息进行评判无法分辨.因此,本文提出利用小波散射网络来提取精细化尺度的细节特征,旨在挖掘更具区分性的尺度特征.

小波散射网络是Mallat教授在小波变换基础上提出的一种改进的多尺度分析方法[27-28],是一种能够自动提取信号细节尺度特征的框架.网络的每一层通常由3部分组成:小波卷积、非线性操作和池化操作,类似于构成卷积神经网络的3个部分.小波散射网络使用的是预先设定好的小波滤波器,而卷积神经网络需要训练得到滤波器参数,因而小波散射网络不需要很大的数据集来训练网络参数,更适用于中小数据集.

小波散射网络以迭代的方式提取特征,将小波变换结果经过非线性取模操作和最大尺度平均化的池化操作,可以得到具有平移不变性和局部变形稳定性的特征,基本框架如图4所示.

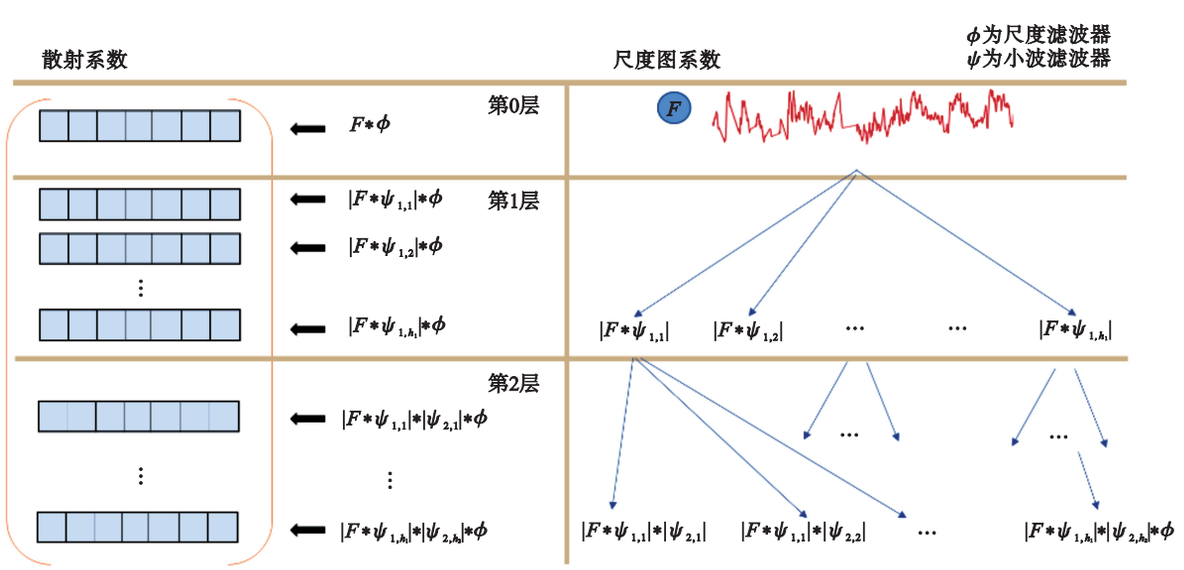

图4 小波散射网络基本框架

Fig. 4 Frame diagram of wavelet scattering network

其算法流程一般为:

①在第0层,信号F与尺度滤波器ϕ卷积,得到第0层的小波散射系数:$S_0=F*\phi$,在平均窗口内具有平移不变性;

②在第1层,信号F与复数小波算子$\psi_{1,\lambda}$卷积并通过取模运算得到第1层尺度模系数:

$$U_1=|F*\psi_{1,\lambda}|. \tag{5}$$

式中,$\lambda=1,2,\cdots,h$.取模运算后得到U_1,接着使用尺度滤波器ϕ与其进行卷积运算,得到第1层网络的散射系数输出:

$$S_1=|F*\psi_{1,\lambda}|*\phi. \tag{6}$$

③按照以上运算步骤得到第2层的尺度模系数U_2,以及小波散射系数S_2:

$$U_2=||F*\psi_{1,\lambda}|*\psi_{2,\lambda}|, \tag{7}$$

$$S_2=||F*\psi_{1,\lambda}|*\psi_{2,\lambda}|*\phi. \tag{8}$$

④通过逐步迭代复数小波算子和取模运算可以获得更多的尺度模系数,重复进行以上操作到第m层,则有

$$U_m=|\cdots||F*\psi_{1,\lambda}|*\psi_{2,\lambda}|\cdots|*\psi_{m,\lambda}|, \tag{9}$$

$$S_m=|\cdots||F*\psi_{1,\lambda}|*\psi_{2,\lambda}|\cdots|*\psi_{m,\lambda}|*\phi. \tag{10}$$

⑤最后将所有层的散射系数相组合,即得到信号的细节尺度特征:

$$\boldsymbol{S}=\begin{bmatrix}S_0\\S_1\\\vdots\\S_m\end{bmatrix}=\begin{bmatrix}F*\phi\\|F*\psi_{1,\lambda}|*\phi\\\vdots\\|\cdots||F*\psi_{1,\lambda}|*\psi_{2,\lambda}|\cdots|*\psi_{m,\lambda}|*\phi\end{bmatrix}. \tag{11}$$

2.3　CatBoost分类算法

基于树的模型能够很好地表达非线性关系,它适合用来解决本文遇到的分类问题.本研究利用CatBoost[29]分类算法进行模型训练.CatBoost是俄罗斯Yandex公司于2018年推出的基于梯度提升决策树(GBDT)[30]算法框架的改进实现,其原理图如图5所示.该算法通过构造一组由决策树组成的弱学习器,并将所有的结果累加作为最终的预测输出.每棵树都是在前一棵树的残差基础上构建的,这意味着下一棵树能在前一棵树未覆盖的区域提供更好的拟合能力.在模型训练过程中,每个决策树都从前一棵树中学习并影响下一棵树,以提高模型性能,从而构建一个强大的学习器.

CatBoost在各种任务中的表现优于其他改进的GBDT算法,其特性如下:

1)使用one-hot编码技术,可以自动处理类别特征,而无需将其转换为数值特征,从而减少数据预处理的工作量;

2)在训练过程中,CatBoost通过损失函数评估预测准确性,并依据损失函数值自动调整每棵树的权重[30];

3)采用阶增方式取代传统的梯度估计方法,减少梯度偏差和预测偏移,提高模型的准确性和泛化能力,同时降低过拟合风险[31];

4)组合现有特征生成新特征,能够利用特定生理尺度的整体信息与小波散射网络提取的精细化尺度细节特征之间的关联性,显著扩充特征维度.

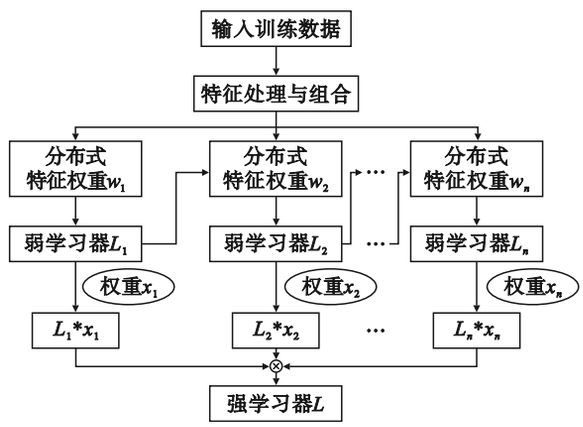

图5　CatBoost的原理图

Fig. 5　Principle of CatBoost

3　结果与分析

3.1　评价指标

本实验采用分类研究中常用的5种评价指标,包括准确率、敏感性、特异性、精确度以及$F1$分数.表1描述的是混淆矩阵,每一列代表样本的预测分类,每一行代表样本的真实分类.

表1　混淆矩阵

Table 1　Confusion matrix

类型	糖尿病前期	健康
糖尿病前期	TP	FN
健康	FP	TN

TP(真阳性):正确分类的患病样本数,即预测为患病样本且实际也为患病样本;

FP(假阳性):错误预测为正常样本的患病样本数;漏报,即实际为患病样本却被预测为正常样本;

TN(真阴性):正确分类的正常样本数,即预测为正常样本且实际也是正常样本;

FN(假阴性):错误预测为患病样本的正常样本数;误报,即实际为正常样本却被预测为患病样本.

准确率(R_{acc})是预测正确样本数量占总样本数的比值:

$$R_{acc}=\frac{TP+TN}{TP+FP+TN+FN}. \tag{12}$$

敏感度(R_{sen})是正确预测为患病样本的数量与实际患病样本总数之比:

$$R_{sen}=\frac{TP}{TP+FN}. \tag{13}$$

特异度(R_{spe})是正确预测为正常样本的数量与实际正常样本总数之比:

$$R_{spe}=\frac{TN}{TN+FP}. \tag{14}$$

精确度(R_{pre})是正确预测为患病样本的数量与预测患病样本总数之比:

$$R_{pre}=\frac{TP}{TP+FP}. \tag{15}$$

$F1$ 分数是敏感度与精确度的调和平均数:

$$F1=2\times\frac{R_{sen}\times R_{spe}}{R_{sen}+R_{spe}}. \tag{16}$$

3.2 特定生理尺度特征分析

从特定生理尺度提取到7种尺度特征(VLF, LF, HF, P_t, R, HF_{nu}, LF_{nu}),为了消除特征间单位差异的影响,并对每维特征同等看待,对每个特征使用Z-score方法进行归一化处理.特征归一化后,其在小提琴图上的表现如图6所示.小提琴图结合了概率密度图与箱线图,更清晰地显示数据的分布情况.

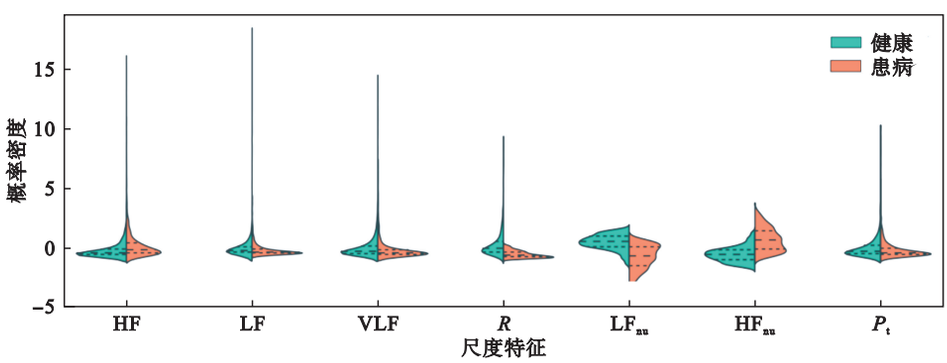

图6 特定生理尺度特征

Fig. 6 Specific physiological scale features

如图6所示,由于个体差异,不同组成部分中存在异常值,表现为概率密度图中的边缘分布.鉴于这些离群值代表了实际事件,本文并没有对其进行严格处理.LF_{nu}, HF_{nu}是将低频、高频功率经过特殊归一化后得到的(计算式如式(1)、式(2)所示),由于LF及HF等各频段的数值直接受总功率P_t的影响,特别是在短时程(5 min)分析时,不同状态下的P_t及LF, HF值各不相同,如果直接以绝对值进行比较,常可得出错误的结论.因此分别进行归一化后再比较.归一化对LF与HF起到缩放的作用,减少了离群值的干扰,使得糖尿病前期和健康人群之间的差异更加明显.与健康人相比,糖尿病前期患者的LF_{nu}成分明显降低,而HF_{nu}成分增加,对于其他成分,两者的差异不显著,这为利用HRV信号诊断糖尿病前期提供了基础.

3.3 精细化尺度特征分析

小波散射网络能够捕捉信号在精细化尺度上的细节特征,提取更丰富、更具区分性的特征,这些特征包含了传统手工提取特征所无法捕捉到的信息.当小波散射网络层数为2时,特征提取器所提取的特征能量之和已与输入信号的能量极为接近.当小波散射网络层数超过2时,特征提取器所提取的特征能量之和会出现显著损失.因此,本实验选取的网络层数为2层.领域知识和经验对于选择小波滤波器起到了重要的指导作用.本文选取的是DB12小波滤波器.分解尺度J表示信号进行尺度变换的次数,它决定了信号在时间和频率域中的分辨率.每一级的分解将信号分解成更低频的子带,同时保留一部分高频信息.品质因数Q表示每个频率带内滤波器的数量,它控制了每个频率带内子带平均分解的深度.考虑J, Q值的多种情况,并将提取出的特征用CatBoost分类器分类,得到的准确率如表2所示.

表2 不同J, Q值得到的准确率

Table 2 Accuracy obtained with different values of J and Q

J	Q			
	(2,1)	(4,1)	(6,1)	(4,2)
2	0.72	0.74	0.73	0.72
3	0.73	0.75	0.75	0.74
4	0.75	0.76	0.76	0.77
5	0.77	0.78	0.77	0.79
6	0.80	0.81	0.79	0.80
7	0.79	0.81	0.80	0.80

小波散射网络提取的精细化尺度特征可视化如图7所示,图中不同颜色代表不同能量水平,颜色越亮表示能量越高.从图7a与图7b的光强度分布可以看出,正常人在0~0.4 Hz的能量明显高于糖尿病前期患者,而0.4 Hz以上的能量大小区别不明显,表明健康个体和糖尿病前期个体的总体能量和能量分布存在显著差异.图7c与图7d之间也存在明显的差异,正常人的整体能量分布仍高于糖尿病前期患者,且在高于0.06 Hz的尺度上明显存在能量分布.无论是第一层还是第二层的小波散射特征,在糖尿病前期患者与正常人之间均存在明显差异,这进一步证明了HRV信号检测糖尿病前期的能力.

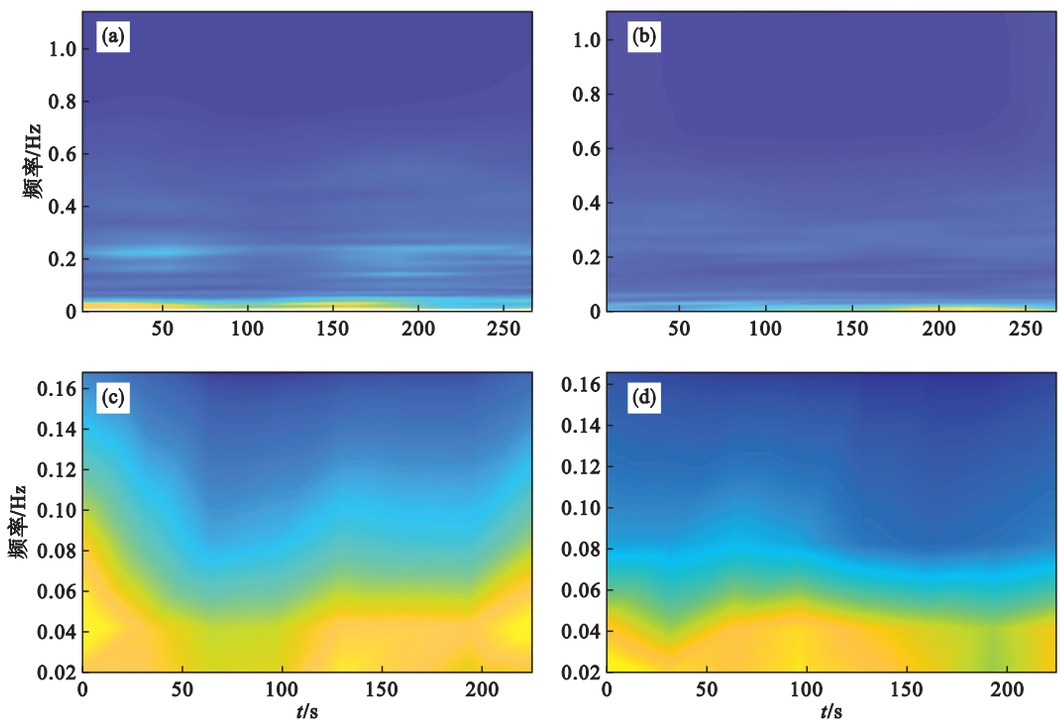

图7　精细化尺度特征的小波散射时频图

Fig. 7　Wavelet scattering-based time-frequency diagram with refined scale features

(a)—正常样本的第一层;(b)—患病样本的第一层;(c)—正常样本的第二层;(d)—患病样本的第二层.

3.4　检测结果

本文从254名受试者的心电信号中截取了1 224例糖尿病前期样本和1 780例健康对照样本.通过CatBoost分类算法,分别利用特定生理尺度特征和精细化尺度特征构建基于单尺度的糖尿病前期检测模型,并将两者进行特征融合后构建基于多尺度的糖尿病前期检测模型.首先按照4∶1的比例将原始数据集随机划分为训练集与测试集,所使用的CatBoost参数均通过网格搜索法确定为模型最优参数.在网格搜索过程采用五折交叉验证方法,将训练集均匀地随机划分为5个子集,其中4个子集用于训练模型,剩下的1个子集用于验证.这个过程将重复5次,每次选择不同的子集作为验证集,以保证模型的稳定性和结果的可靠性.将每次迭代后$F1$分数的平均值作为模型的选定标准,最终将最优模型在测试集上测试.图8是3种模型在测试集上的混淆矩阵结果图.模型经过测试后得到各模型评价指标的值如表3所示.

本文首先对特定生理尺度特征构建检测模型,准确率、敏感度、特异度、精确度、$F1$分数分别为76.71%, 65.96%, 83.61%, 72.09%, 73.74%;其次,对小波散射网络提取的精细化尺度特征构建了基于精细化尺度特征的单尺度检测模型,得到了81.03%的准确率、71.49%的敏感度、87.16%的特异度、78.14%的精确度、78.55%的$F1$分数;最后,本文将上述两种特征相融合构建了多尺度特征检测模型,得到了88.52%的准确率、83.40%的敏感度、91.82%的特异度、86.73%的精确度、87.40%的$F1$分数.在各个性能指标上,基于精细化尺度特征的单尺度检测模型的表现要比基于特定生理尺度的单尺度检测模型更好,这是由于小波散射网络提供了精细化尺度特征,能够描述隐藏在信号中的细节信息;而基于特征融合的多

尺度检测模型表现比基于精细化尺度特征的单尺度检测模型更优,这是由于多尺度模型将特定生理尺度的整体信息与小波散射网络提供的精细化尺度信息更好地融合在一起.

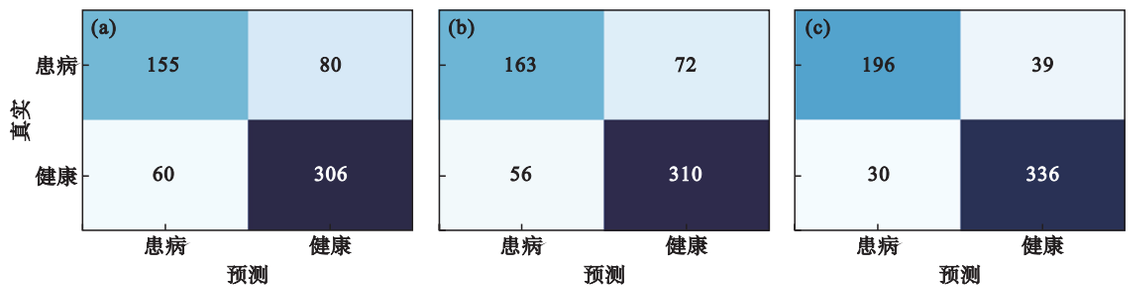

图 8 各模型的混淆矩阵

Fig. 8 Confusion matrix of each model

(a)—特定生理尺度特征模型;(b)—精细化尺度特征模型;(c)—多尺度特征模型.

表 3 单尺度与多尺度模型的效果

Table 3 Effects of single-scale and multi-scale models %

方法	准确率	敏感度	特异度	精确度	$F1$
特定生理尺度特征+CatBoost	76.71	65.96	83.61	72.09	73.74
精细化尺度特征+CatBoost	81.03	71.49	87.16	78.14	78.55
多尺度特征+CatBoost	88.52	83.40	91.82	86.73	87.40

本文绘制了 3 种模型训练后不同特征对分类过程贡献度排序的结果,如图 9 所示.

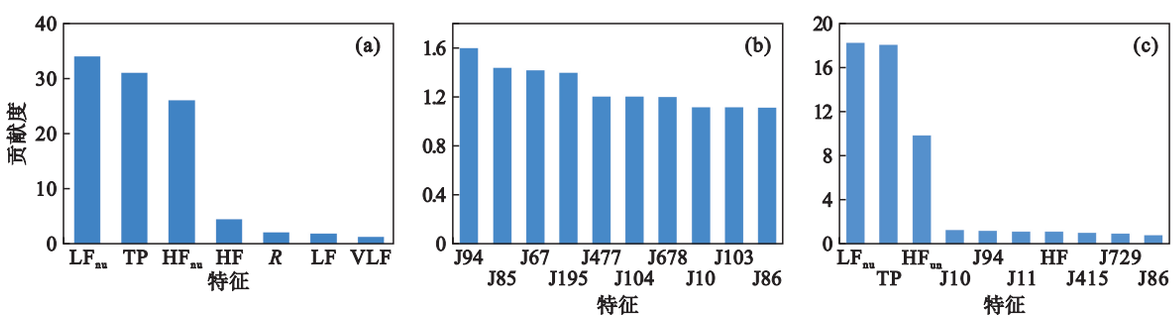

图 9 特征重要性排序

Fig. 9 Feature importance ranking

(a)—特定生理尺度特征模型;(b)—精细化尺度特征模型;(c)—多尺度特征模型.

由图 9 可知,在基于特定生理尺度特征的单尺度检测模型中,LF_{nu},P_t 和 HF_{nu} 的贡献值较大,其他特征的贡献值较小;在基于精细化尺度特征的单尺度检测模型中,列出了前 10 个贡献值最大的尺度特征,J94,J85,J67 和 J195 的贡献值较大,其他特征的贡献值表现相当;在基于多尺度特征融合的检测模型中,列出了前 10 个贡献值最大的尺度特征,LF_{nu},P_t 和 HF_{nu} 的贡献值依旧是最大的,其余特征的贡献较小.然而 J10,J94 等精细化尺度特征的贡献值仍然高于 HF 等其他特定生理尺度特征.多尺度模型的各种性能指标皆得到了大幅度提升,说明 J10,J94 等特征可以有效地辅助模型决策,小波散射网络提取的精细化尺度特征发挥了较为重要的作用.在这项工作中,所提出的多尺度特征模型在全部性能指标评估中表现最好,准确率达到 88.52%,敏感度达到 83.40%,特异度达到 91.82%,精确度达到 86.73%,$F1$ 分数达到 87.40%,比单一的特定生理尺度模型和精细化尺度模型皆有大幅提升.尤其是模型中代表检测糖尿病前期患者能力的敏感度指标,这意味着该模型可以检测到绝大多数糖尿病前期患者.此外,模型完全基于 5 min HRV 信号检测糖尿病前期,具有无创、简便的优点.

3.5 讨 论

本文使用随机森林(random forest, RF)、AdaBoost、XGBoost 和支持向量机(SVM)这几种常用的机器学习分类算法构建了糖尿病前期检测模型.表 4 比较了采用不同分类算法时模型的

性能指标结果.

表4　采用不同算法时的模型性能指标

Table 4　Performance indices of models with different algorithms　%

分类算法	准确率	敏感度	特异度	精确度	*F*1
RF	87.69	81.70	91.53	86.10	86.34
AdaBoost	85.02	79.57	88.52	81.66	83.81
CatBoost	88.52	83.40	91.82	86.73	87.40
XGBoost	88.19	83.40	91.26	85.96	87.15
SVM	73.37	75.85	71.67	64.80	69.89

从所有算法的评估结果来看,本文提出的多尺度分析方法在多种分类算法上均具有良好的分类效果.其中,SVM模型效果最差,在5项指标上的数值均最低;CatBoost模型在5项指标中的数值最大,特异度更是达到了91.82%;RF模型和XGBoost模型的表现相当,仅次于CatBoost模型;AdaBoost模型在所有指标上的数值均不如以上3种模型,但仍优于SVM模型.综合来看,基于CatBoost构建的多尺度特征糖尿病前期预测模型效果最好.这得益于CatBoost强大的特征组合能力,能够将整体特征与细节特征更好地结合在一起.

本研究采用可解释的CatBoost作为分类器,使决策过程更加透明;本文提供了特征贡献值的排序结果,增强了模型的可解释性,在糖尿病前期筛查任务中具有快速、简便的优势.然而需要指出的是,本研究存在一定局限性,实验数据具有一定的地域局限性,仅来源于大连医科大学附属第一医院的患者数据,若要进一步提升模型的性能和普适性,还需纳入更多不同地域和年龄段的数据.

4　结　　语

本文提出了基于HRV多尺度分析的糖尿病前期检测方法.首先,在领域专业知识的指导下,从HRV信号中提取了特定生理尺度的整体特征.经过特征分析发现,糖尿病前期患者的LF_{nu}成分明显减少,HF_{nu}成分明显增加;其次,通过运用小波散射网络,获取了精细化尺度的细节特征,可视化分析表明,健康人在网络第1层和第2层的能量整体上高于糖尿病前期患者.最后,采用CatBoost分类算法构建了多尺度糖尿病前期检测模型.实验结果表明,本文的方法在糖尿病前期患者检测方面具有良好的检出率.CatBoost算法在训练阶段的复杂度为$O(n\lg n)$,其中n是训练样本的数量.这部分的计算主要在训练阶段进行,不影响实时检测.在预测阶段,CatBoost算法的复杂度为$O(\lg m)$,其中m是模型树的深度.由于每次预测只需进行一次树的遍历,计算复杂度较低,适合实时应用.未来,本文的检测方法有望部署于像手环这类的可穿戴设备上,以方便人们在日常生活中实现自我健康监测,期望能够在糖尿病前期的早期诊断中实现突破,为公共卫生事业作出贡献.

参考文献:

[1] International Diabetes Federation. IDF diabetes atlas[M]. 10th ed. Brussels: International Diabetes Federation, 2021.

[2] Wan H, Wang Y Y, Fang S J, et al. Associations between the neutrophil-to-lymphocyte ratio and diabetic complications in adults with diabetes: a cross-sectional study[J]. *Journal of Diabetes Research*, 2020, 2020(1): 6219545.

[3] ElSayed N A, Aleppo G, Aroda V R, et al. 17 diabetes advocacy: standards of care in diabetes—2023[J]. *Diabetes Care*, 2023, 46(sup1): 279-280.

[4] Echouffo-Tcheugui J B, Selvin E. Prediabetes and what it means: the epidemiological evidence[J]. *Annual Review of Public Health*, 2021, 42: 59-77.

[5] Faerch K, Hulmán A, Solomon T P J. Heterogeneity of prediabetes and type 2 diabetes: implications for prediction, prevention and treatment responsiveness[J]. *Current Diabetes Reviews*, 2016, 12(1): 30-41.

[6] Liu Q, Zhou Q, He Y F, et al. Predicting the 2-year risk of progression from prediabetes to diabetes using machine learning among Chinese elderly adults[J]. *Journal of Personalized Medicine*, 2022, 12(7): 1055.

[7] Tabák A G, Herder C, Rathmann W, et al. Prediabetes: a high-risk state for diabetes development[J]. *The Lancet*, 2012, 379(9833): 2279-2290.

[8] Brannick B, Wynn A, Dagogo-Jack S. Prediabetes as a toxic environment for the initiation of microvascular and macrovascular complications[J]. *Experimental Biology and Medicine*, 2016, 241(12): 1323-1331.

[9] Cai X Y, Zhang Y L, Li M J, et al. Association between prediabetes and risk of all cause mortality and cardiovascular disease: updated meta-analysis[J]. *BMJ*, 2020, 370: m2297.

[10] Mutie P M, Pomares-Millan H, Atabaki-Pasdar N, et al. An investigation of causal relationships between prediabetes and vascular complications[J]. *Nature Communications*, 2020, 11: 4592.

[11] Cosic V, Jakab J, Pravecek M K, et al. The importance of prediabetes screening in the prevention of cardiovascular disease[J]. *Medical Archives*, 2023, 77(2): 97-104.

[12] Tobore I, Kandwal A, Li J Z, et al. Towards adequate prediction of prediabetes using spatiotemporal ECG and EEG feature analysis and weight-based multi-model approach[J]. *Knowledge-Based Systems*, 2020, 209: 106464.

[13] Wang L Y, Mu Y, Zhao J, et al. IGRNet: a deep learning model for non-invasive, real-time diagnosis of prediabetes through electrocardiograms[J]. *Sensors*, 2020, 20(9):

2556.
[14] Lin Y C, Lin C S, Chang T S, et al. Early sensory neurophysiological changes in prediabetes[J]. *Journal of Diabetes Investigation*, 2020, 11(2): 458-465.
[15] Oliveira C M, Ghezzi A C, Cambri L T. Higher blood glucose impairs cardiac autonomic modulation in fasting and after carbohydrate overload in adults [J]. *Applied Physiology, Nutrition, and Metabolism*, 2021, 46(3): 221-228.
[16] Coopmans C, Zhou T L, Henry R M A, et al. Both prediabetes and type 2 diabetes are associated with lower heart rate variability: the maastricht study [J]. *Diabetes Care*, 2020, 43(5): 1126-1133.
[17] Rajendra A U, Paul J K, Kannathal N, et al. Heart rate variability: a review [J]. *Medical and Biological Engineering and Computing*, 2006, 44(12): 1031-1051.
[18] Vijay C, Darshan M, Vishnu R. Cardiac autonomic dysfunction and ECG abnormalities in patients with type 2 diabetes mellitus—a comparative cross-sectional study[J]. *National Journal of Physiology, Pharmacy and Pharmacology*, 2016, 6(3): 178.
[19] Igbe T, Li J Z, Kandwal A, et al. An absolute magnitude deviation of HRV for the prediction of prediabetes with combined artificial neural network and regression tree methods[J]. *Artificial Intelligence Review*, 2022, 55(3): 2221-2244.
[20] 中华医学会糖尿病学分会.中国2型糖尿病防治指南(2020年版)[J].中华糖尿病杂志, 2021, 13(4): 317-411.(Chinese Diabetes Society. Type 2 diabetes prevention and treatment comprehensive guide (2020 edition)[J]. *Chinese Journal of Diabetes*,2021, 13(4): 317-411.)
[21] Cui X R, Tian L R, Li Z W, et al. On the variability of heart rate variability—evidence from prospective study of healthy young college students[J]. *Entropy*, 2020, 22(11): 1302.
[22] Sassi R, Cerutti S, Lombardi F, et al. Advances in heart rate variability signal analysis: joint position statement by the E-cardiology ESC working group and the European Heart Rhythm Association co-endorsed by the Asia Pacific Heart Rhythm Society[J]. *Europace*, 2015, 17(9): 1341-1353.
[23] Pan J, Tompkins W J. A real-time QRS detection algorithm [J]. *IEEE Transactions on Biomedical Engineering*, 1985, 32(3): 230-236.
[24] Catai A M, Pastre C M, de Godoy M F, et al. Heart rate variability: are you using it properly? standardisation checklist of procedures[J]. *Brazilian Journal of Physical Therapy*, 2020, 24(2): 91-102.
[25] Forte G, Favieri F, Casagrande M. Heart rate variability and cognitive function: a systematic review[J]. *Frontiers in Neuroscience*, 2019, 13: 710.
[26] Jwo D J, Chang W Y, Wu I H. Windowing techniques, the welch method for improvement of power spectrum estimation[J]. *Computers, Materials and Continua*, 2021, 67(3): 3983-4003.
[27] Jin Y, Duan Y L. Wavelet scattering network-based machine learning for ground penetrating radar imaging: application in pipeline identification[J]. *Remote Sensing*, 2020, 12(21): 3655.
[28] Mallat S. Group invariant scattering[J]. *Communications on Pure and Applied Mathematics*, 2012, 65(10): 1331-1398.
[29] Ostroumora L, Gusev G, Vorobev A, et al. CatBoost: unbiased boosting with categorical features[C]// Advances in Neural Information Processing Systems. Montreal, 2018: 6639-6649.
[30] Bentéjac C, Csörgő A, Martínez-Muñoz G. A comparative analysis of gradient boosting algorithms [J]. *Artificial Intelligence Review*, 2021, 54(3): 1937-1967.
[31] Zhang Y X, Zhao Z G, Zheng J H. CatBoost: a new approach for estimating daily reference crop evapotranspiration in arid and semi-arid regions of Northern China[J]. *Journal of Hydrology*, 2020, 588: 125087.