DOI: 10.13382/j.jemi.B2407820

全血光谱融合评估心血管风险方法研究*

何 洋 李志刚 杨蕊歌 王睿鑫 杨子龙
(东北大学信息科学与工程学院 沈阳 110819)

摘 要:心血管疾病是世界人口发病和死亡的最大原因之一。及时且可靠的心血管疾病风险评估是减轻患病风险,保障生命安全的关键。提出了一种高效、便捷的心血管疾病风险评估方法。采集了 108 个全血样本的傅里叶变换红外衰减全反射光谱和拉曼光谱进行风险评估模型的构建与评价。针对基于传统最小二乘法(PLS)、联合区间偏最小二乘法(siPLS)等算法进行特征提取而建立的风险评估模型效能低下的问题,提出了化学键驱动的区间联合偏最小二乘算法(CBDsiPLS)用于特征提取,并结合机器学习构建了单一数据的风险评估模型,测试结果表明该方法优于传统的特征提取算法。此外,利用中红外与拉曼光谱的信息互补性,进行特征级信息融合后结合机器学习方法建立融合数据的风险评估模型。最终的融合数据风险评估模型的准确率均超过 90%,灵敏度均超过 80%,特异性均达到 95%。实验结果表明,所提出的方法可以实现对心血管疾病风险的有效评估。

关键词: 心血管疾病;风险评估;全血光谱;特征提取;信息融合
中图分类号: TN214;TH741 **文献标识码**: A **国家标准学科分类代码**: 150.25

Research on cardiovascular disease risk assessment method based on whole blood spectral information fusion

He Yang Li Zhigang Yang Ruige Wang Ruixin Yang Zilong
(School of Information Science and Engineering, Northeastern University, Shenyang 110819, China)

Abstract: Cardiovascular disease is one of the leading causes of morbidity and mortality worldwide. Timely and reliable risk assessment is crucial for reducing disease risk and ensuring safety. The aim of this research is to propose an efficient and convenient risk assessment method for cardiovascular disease. In this research, Fourier transform infrared attenuated total reflectance spectra and Raman spectra of 108 whole blood samples were collected for the construction and evaluation of risk assessment models. To address the issue of low efficiency in risk assessment models based on traditional PLS, siPLS, and other feature extraction algorithms, a chemical bond-driven synergy interval partial least squares algorithm (CBDsiPLS) is proposed for feature extraction, and combined with machine learning to construct a risk assessment model using single data sets. The test results show that the proposed method outperforms traditional feature extraction algorithms. In addition, by utilizing the complementary information from mid-infrared and Raman spectroscopy, a risk assessment model for fused data was established through feature-level information fusion combined with machine learning methods. The final fused data risk assessment model achieves an accuracy of more than 90%, a sensitivity of more than 80%, and a specificity of 95%. The experimental results show that the proposed method can effectively assess the risk of cardiovascular disease.

Keywords: cardiovascular disease; risk assessment; whole blood spectrum; feature extraction; information fusion

0 引 言

心血管疾病是全球的头号死因,每年死于心血管疾病的人数多于任何其他死因。据世界卫生组织发布,心血管疾病每年约夺走 1 790 万人的生命。中国的情况同样令人担忧。据估计,我国目前有约 3.3 亿心血管疾病患者,并且随着人口老龄化和生活方式的改变,心血管疾

收稿日期: 2024-09-11 Received Date: 2024-09-11
*基金项目:天津市卫健委科技项目(ZC20121)、国家自然科学基金(61601104)、河北省自然科学基金(F2017501052)、中央高校基本科研经费(N2023021)项目资助

病的发病率还在不断上升[1]。此外,心血管疾病往往会对患者造成不可逆的伤害[2]。常见的病变包括心脏肌肉损害、血管硬化、血栓形成等[3],这些病变一旦发生,会对心脏和血管造成永久性的损伤,使其无法恢复到健康状态。尽管医学技术和治疗手段不断进步,但患者只能依靠合理的管理和治疗来延缓疾病的进展,并维持生命功能。因此,对心血管疾病的风险评估进行研究具有重要的理论和实际意义,有助于早期发现和干预,提高患者的生活质量,降低疾病的危害性。

目前,血液检测是判断心血管疾病的普遍方法[4-5]。但这种侵入性检测方法存在耗费血样量大、感染风险高、耗时长、需要不同化学试剂分析不同物质等缺点。相比之下,振动光谱具有低成本、侵入性低、易操作等优点,可实现对微量物质的无创、快速检测,在医学筛查与检测方面展现出了巨大的优势[6]。

近年来,由于分子光谱技术,尤其是拉曼(Raman)光谱和红外光谱在分子检测方面展现的高灵敏度,使其逐渐成为一种新型高效的疾病辅助诊断技术。Chen 等[7]通过对采集的血液样本进行拉曼光谱分析,结合适当的数据处理算法和机器学习模型,实现了对癌症、囊肿和正常样本的三元分类,其识别灵敏度和特异性分别达到了 81.0%和 97.3%,63.6%和 91.5%,以及 100%和 90.6%。Song 等[8]利用拉曼光谱技术成功区分了甲状腺乳头状癌和甲状腺微小乳头状癌,采用了主成分分析(PCA)降维方法结合 Adaboost 模型建模,准确率达到了 84.61%。Yue 等[9]利用傅里叶变换红外光谱结合深度学习算法对甲状腺功能异常患者和对照组样本构建了分类模型,其准确率高达 95.1% 。许多研究表明分子光谱技术在临床诊断领域具有很大的应用潜力[10-12]。

尽管现有研究在应用拉曼光谱和红外光谱技术进行疾病检测方面取得了一定的进展,但大多数研究主要集中在单一光谱技术的应用上,缺乏对不同光谱信息的融合与综合分析。由于红外光谱和拉曼光谱在检测原理和应用场景上具有明显的互补性,即红外光谱常用于研究极性基团的不对称振动,而拉曼光谱常用于研究非极性基团和骨架的对称振动[13],基于此,本文旨在探索结合两种全血光谱信息,通过数据融合方法构建心血管疾病风险评估模型,以提高风险评估的精度和可靠性。

1　理论论证

1.1　傅里叶变换红外衰减全反射光谱(FTIR-ATR)和 Raman 光谱理论基础

FTIR-ATR 是一种用于分析物质组成和结构的红外光谱技术,光谱能够通过测量红外光被分子吸收后的振动模式来提供分子结构信息。它适用于检测极性分子的振动,例如羰基(C=O)和羟基(O-H)。

拉曼光谱通过测量散射光的频率偏移(拉曼效应)来分析分子的振动、旋转,常用于分子结构的研究。它适用于检测非极性分子的振动,例如甲基(C-H)和亚甲基。

心血管疾病与全血中物质的变化密切相关[14-15]。高水平的甘油三酯(triglyceride,TG,)被认为是心血管疾病的重要因素,常伴随动脉粥样硬化等相关疾病。高密度脂蛋白胆固醇(high density lipoprotein cholesterol,HDL-C)具有清除血液胆固醇,降低心血管疾病风险的作用,HDL-C 水平过低也被认为是心血管疾病风险因素。光谱技术能够通过识别官能团的峰值反映这些化学物质的含量,从而有效评估和筛查心血管疾病的风险。

1.2　化学键驱动的区间来联合偏最小二乘(CBDsiPLS)算法的理论依据

由于光谱数据量巨大,为了剔除无关波段的影响,减少噪声,在进行定性分析之前,需要筛选与分类目标相关的特征谱线,以确保所选取的波段与研究目标密切相关。

现有的偏最小二乘算法[16](partial least squares,PLS)是在光谱领域经常使用的一种降维方法,通过建立预测变量和响应变量的线性关系降低数据维度,但该方法在特征提取前未进行波段筛选,这可能会导致提取的特征仍包含冗余信息。而联合区间偏最小二乘算法[17-18](synergy interval partial least square,siPLS)在 PLS 的基础上加入了波段筛选,但是其波段划分是随机的平均划分,而且在波段选取时也需要进行大量的实验去筛选有效波段,所以,这种波段划分方法具有主观性强的缺点,并且该算法计算复杂度高。

因此,本文在 siPLS 算法的基础上,对波段选择部分进行改进,提出了 CBDsiPLS 算法。首先在波段筛选时,选择与心血管疾病相关物质关系密切的波段,基于 TG 和 HDL-C 的分子特征,选择与羰基、羟基、甲基等振动相关的波段,这样可以在波段选择时去除无关波段的干扰,强化了对目标特征的关注,并且减少了计算量。波段筛选后,再使用 PLS 算法进行进一步的特征提取。

1.3　信息融合的理论分析

在某些场景中,单一光谱进行物质检测难以达到所需精度要求。针对这一问题,基于信息融合的光谱检测技术开始受到众多学者的关注和研究[19]。FTIR-ATR 和 Raman 光谱在信息检测上的侧重不同,并且存在一定的互补性。FTIR-ATR 对于极性较强的化学键能够产生比较强的信号,缺点是对非极性分子不敏感,相反,Raman 光谱适用于极性分子的检测。TG 和 HDL-C 作为心血管疾病的关键指标,都含有极性和非极性分子特征。使用信息融合手段,将 FTIR-ATR 和 Raman 光谱结合,可以全面捕捉这些分子的特征波段,提高模型的分类性能。

2　实验材料与方法

2.1　实验材料

将108份全血标本收集入EDTA抗凝管中。在测定前,将样本保存在-80 ℃,在测定时在室温下解冻。该实验已获得海港医院人体伦理委员会的批准签署。同时,采用检验科的贝克曼AU5800全自动分析仪,对全血中的TG和HDL-C等指标进行测定。由于血液中高水平的TG或TG与HDL-C比值与发生心血管疾病的风险增加有关[20],医生根据所测指标将病人归为有患心血管疾病风险和无患心血管疾病风险两类,分别有41和67个。

2.2　光谱数据采集

1)全血FTIR-ATR光谱

采用配有单反应ATR采样附件(ZnSe池)的Bruker Alpha FTIR光谱仪进行FTIR-ATR的采集,分辨率设定为6 cm^{-1},扫描范围为4 000~650 cm^{-1}。每次全血光谱测量之前,需对背景进行重新测量,消除背景影响。光谱重复测量3次,并从每个样品的平均光谱中减去在相同实验条件下获得的水光谱。全血光谱测量时,先使用注射器将0.6 μL解冻后的样本注入液体池,然后进行光谱测量。每次测量之后,使用5 mL的去离子水对液体池进行清洗,再进行干燥。这些全血样品的FTIR-ATR谱图如图1所示。

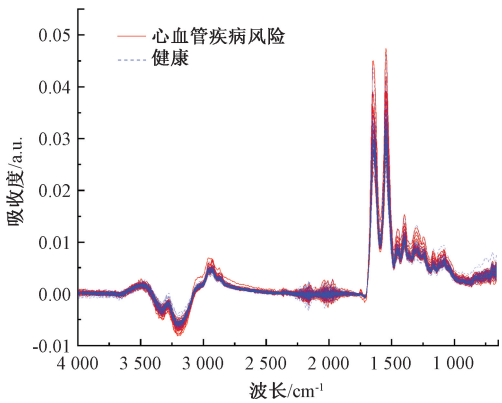


图1　全血FTIR-ATR光谱数据集

Fig. 1　FTIR-ATR spectral data set of whole blood

2)全血Raman光谱

采用配备了1 064 nm激光线和锗探测器Bruker MultiRAM(位于德国不来梅)傅里叶变换(FT)拉曼光谱仪进行全血样本的Raman光谱采集。锗探测器采用液氮冷却,确保高效的性能。在室温下进行测量,光谱范围涵盖4 000~650 cm^{-1},具有6 cm^{-1}的测量分辨率。为了防止样品在分析过程中发生热降解,激光功率被控制在150 mW。进行背景测量后开始全血光谱测量,先使用注射器将0.6 μL样本注入液体池,干燥5 min后进行光谱测量。测量之后,使用5 mL的去离子水对液体池进行清洗,之后进行液体池干燥,重新测量背景等一系列操作。这一系列的配置和措施旨在确保准确而可靠的全血光谱Raman光谱采集。全血的原始Raman谱图如图2所示。

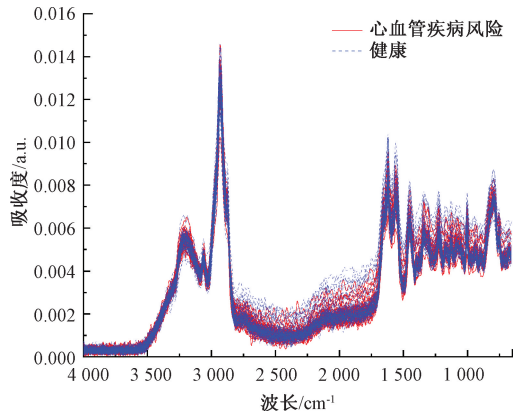


图2　全血Raman光谱数据

Fig. 2　Raman spectral data set of whole blood

2.3　数据扩充

本文采集的样本个数是108个,按照7∶3的比例分为训练集和测试集。划分后,训练集的样本有47例阴性样本和29例阳性样本,为增加样本空间多样性,选取线性插值法进行样本扩充[21]。首先进行数据平衡,将阳性样本扩充到47例。在此基础上,为避免过度扩充导致的数据冗余和模型过拟合,进一步将训练集样本扩充至原来的2倍。最终使训练集中阳性样本和阴性样本均达到94例。

使用线性插值方法生成的新样本位于已有的样本连线上,在原始数据分布空间内,在增加数据多样性的同时能够有效的避免数据偏离真实分布的问题。每个插值样本都是基于两个同类别样本生成,类别标签继承原始样本,保证了新样本的类别一致性。相比于其他数据生成方法,线性插值法具有简单高效,且不会引入过多额外噪声的优点,适合本文样本规模有限的场景。

2.4　特征提取

使用所提出的CBDsiPLS算法进行特征提取,该方法具体可分为波段筛选和特征提取两部分。

1)波段筛选

心血管疾病的发展与血液中TG、HDL-C等指标的含量密切相关。因此,在特征谱线的筛选过程中,需要优先考虑与这些生物标志物相关的光谱区域,以确保所获得的数据能够更有效地反映心血管疾病的相关信息。

对于FTIR-ATR光谱,与TG相关的波段有两段,分别是1 500~1 100和1 800~1 600 cm^{-1},与HDL-C相关

的波段有 3 段,分别是 1 800～1 700、3 500～2 800 和 1 500～900 cm^{-1}[22]。1 800～1 700 cm^{-1} 波段包括了羰基的振动,特别是酯键的振动。HDL-C 与脂质分子中的羰基振动有关,例如脂质中的酯键。这一区域的变化可能反映了与脂质组分相关的结构或含量变化,3 500～2 800 cm^{-1} 这个范围涵盖了羟基和甲基的振动。HDL-C 中可能包含有羟基官能团,因此这一区域的变化可能与胆固醇醚链中的羟基有关。甲基振动则涉及到脂肪酸链的存在。1 500～900 cm^{-1} 这个较宽的范围可以涵盖多种振动模式,包括脂肪酸的弯曲和拉伸振动,以及其他可能与 HDL-C 相关的化学键振动[23]。这个范围提供了更广泛的信息,可以用于分析复杂的分子结构。

对于 Raman 光谱,TG 的 Raman 光谱与相应脂肪酸的光谱相似,主要表现在 1 800～1 000 和 3 100～2 800 cm^{-1},1 800～1 000 cm^{-1} 范围通常与脂肪酸链中的甲基和亚甲基振动有关[24],而三酰甘油中的甲基基团在这个区域可能表现出特定的光谱特征,3 100～2 800 cm^{-1} 这个范围涉及到羰基的振动,特别是脂肪酸酯键的振动。与 HDL-C 相关的光谱信息主要在 3 100～2 800、1 500～1 400 和 1 200～1 050 cm^{-1}。分别对应于脂肪酸链中的甲基和亚甲基振动,以及 C-C 和 C-H 的伸缩模式[25]。

将上述提到的波段组合起来,可以得到一个全面且信息丰富的初步波段筛选,可用于心血管疾病的风险评估。这些选择的波段组合不仅聚焦于脂肪酸链和羰基的振动,还充分考虑了与蛋白质和其他生物分子有关的振动。这种全面性的初步筛选使得算法能够捕捉多个生物分子之间的相互作用,为光谱提供更为综合的信息,进一步揭示了与心血管健康状况密切相关的信号。

2)特征提取

经过初步筛选,已经大大降低了光谱数量。然而,光谱融合会使光谱的数量进一步提升,为了降低融合数据模型的计算复杂度,剔除无关变量影响,需要对光谱数据进行进一步的特征选择。本文采用 PLS 算法对光谱数据进行降维。PLS 的主要目标是通过建立新的变量来捕捉预测变量和响应变量之间的关系,从而实现降维,已广泛应用于医疗、食品等各个领域[26]。

2.5　信息融合

信息融合是将多个信息源进行融合,生成更具体、更全面的样本数据集。本文采用了中层融合(又称特征融合)方法[27]对数据特征进行融合,中层融合能够在信息融合的同时尽可能的减少冗余信息,降低计算成本,获得更好的效果,其原理图如图 3 所示。分别从两种信息源提取大小为 $n \times p$ 和 $n \times q$ 的特征矩阵 $\boldsymbol{a}$ 和特征矩阵 $\boldsymbol{b}$,n 代表样本个数,p 和 q 分别代表从不同信息源中提取的特征变量的个数。然后将提取的特征矩阵 $\boldsymbol{a}$、$\boldsymbol{b}$ 进行拼接,获得大小为 $n \times (p+q)$ 的融合特征矩阵 $\boldsymbol{c}$,作为分类模型的最终输入数据。

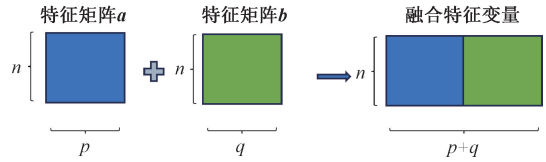


图 3　特征融合原理

Fig. 3　Feature fusion principle

2.6　模型选择

为验证算法的普适性,选取 3 个机器学习模型作为分类模型。

随机森林算法[28]是一种强大的集成学习方法,其基础构建模块为多个决策树,通过联合分类训练和趋势预测,提供了对样本数据更为稳健的分类器。在随机森林中,每颗决策树的训练数据都是通过独立的随机抽样得到的,使得每棵树都具有差异性。这样的多样性有助于提高整体模型的泛化能力,减小过拟合的风险。对于新数据的分类结果,随机森林采用投票机制,即每颗决策树都对新数据进行分类,最终的分类结果由各个树的投票多少来决定。随机森林具有精度高、对小规模数据性能好、泛化能力强、并行计算能力强等诸多优点。在生物医疗和食品检测等领域取得很好的分类效果。在生物医疗和食品检测等领域,随机森林取得了显著的分类效果。

极致梯度提升[29](eXtreme gradient boosting,XGBoost)是一种基于梯度提升框架的机器学习算法,通过迭代地构建多个决策树模型,每一步都针对前一步模型的残差进行优化。XGBoost 在训练过程中引入了正则化以防止过拟合,通过特征重要性评估和并行计算等优势,使其在性能和效率上具有显著优势,适用于分类、回归等多种任务。XGBoost 以其高性能、鲁棒性和灵活性被广泛应用于疾病检测和辅助诊断等方面,因此选择 XGBoost 作为第 2 个分类器。

支持向量机[30](support vector machine,SVM)是振动谱中最常用的二分类模型之一。SVM 的原理是通过找到能够有效区分不同类别的超平面,使得样本点到该超平面的距离最大化,从而实现分类任务。该算法具有泛化能力强,鲁棒性强等诸多优点,被广泛应用于光谱数据的疾病分类任务,例如癌症、甲状腺疾病等。因此选择 SVM 作为第 3 个分类器。

2.7　模型评价指标

模型评估是衡量不同数据或模型对分类任务有效性的关键步骤。在单一数据模型评价中,本文采用了准确率、灵敏度和特异性对分类结果进行评价。准确率直观地反映了整体分类结果的精确性,较高的灵敏度意味着漏诊率较低,而较高的特异性则表示误诊率较低。这 3

个指标越趋近于 1,代表分类效果越为出色。

$$准确率 = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (1)$$

$$灵敏度 = \frac{TP}{TP + FN} \times 100\% \quad (2)$$

$$特异性 = \frac{FP}{FP + TN} \times 100\% \quad (3)$$

式中:*TP* 为测试集中阳性样本被分类正确的样本数量;*FN* 为测试集中阳性样本被分类错误的样本数量;*FP* 为测试集中阴性样本被分类错误的样本数量;*TN* 测试集中阴性样本被分类正确的样本数量。

在数据融合模型实验中,为了使单一数据模型和数据融合模型的分类效果对比更加直观,在以上 3 个评价指标的基础上增加 ROC 曲线[31]评价指标。ROC 曲线的横坐标代表 1 减去特异性,纵坐标代表灵敏度。曲线下面积用 AUC 表示,AUC 越大,标志着模型有更为出色的分类效果。

3 实验结果与分析

3.1 单一数据模型

首先,采用基线矫正和归一化对实验数据进行预处理,用来去除基线漂移等干扰因素,增强了数据的可靠性。将两种数据集按照相同的方式划分为训练集和测试集,采用 7∶3 的比例划分。然后,分别对 FTIR-ATR 光谱数据和 Raman 光谱数据使用 CBDsipls、sipls 和 pls 方法进行特征提取。特征维数的选择基于五折交叉验证的方法,根据交叉验证的准确率大小选取最优特征维数。最后,针对不同方法提取的特征,分别使用随机森林、XGBoost 和 SVM 建立分类模型,并进行参数优化。实验结果如表 1~3 所示。

表 1 单一数据随机森林模型分类结果

Table 1 Classification results of random forest model with single data (%)

数据类型	特征提取方法	准确率	灵敏度	特异性
FTIR-ATR	CBDsiPLS	81.25	75.00	85.00
	siPLS	75.00	66.67	80.00
	PLS	75.00	66.67	80.00
Raman	CBDsiPLS	62.50	58.33	65.00
	siPLS	62.50	58.33	65.00
	PLS	59.38	50.00	65.00

从表 1~3 可以明显看出,CBDsiPLS 方法在不同模型上的的表现显著优于其他方法,无论是准确率、灵敏度还是特异性均处于领先地位。例如,在随机森林模型中,FTIR-ATR 数据集的 CBDsiPLS 方法准确率为 81.25%,明显高于 siPLS(75.00%)和 PLS(75.00%)。在 Raman 数据集上,CBDsiPLS 的准确率也高于 siPLS 和 PLS,但整体表现低于 FTIR-ATR 数据集,这可能是由于不同光谱技术在化学信息上的差异所致。由此可以看出,CBDsiPLS 方法展现出了其相对于传统方法的优势。

表 2 单一数据 XGBoost 模型分类结果

Table 2 Classification results of XGBoost with single data (%)

数据类型	特征提取方法	准确率	灵敏度	特异性
FTIR-ATR	CBDsiPLS	84.38	83.33	85.00
	siPLS	78.13	75.00	80.00
	PLS	78.13	75.00	80.00
Raman	CBDsiPLS	62.50	58.34	65.00
	siPLS	68.75	58.33	75.00
	PLS	62.50	66.67	60.00

表 3 单一数据 SVM 模型分类结果

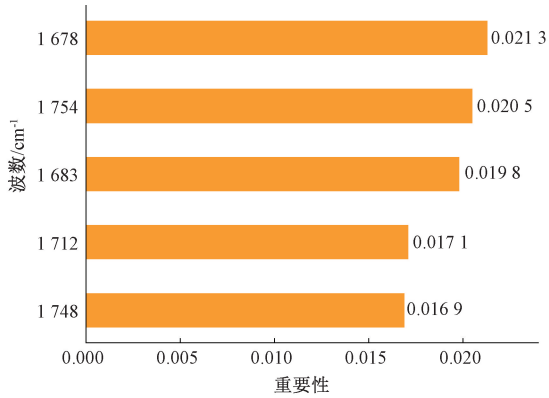
Table 3 Classification results of SVM with single data (%)

数据类型	特征提取方法	准确率	灵敏度	特异性
FTIR-ATR	CBDsiPLS	78.13	66.67	75.00
	siPLS	71.88	50.00	85.00
	PLS	71.88	50.00	85.00
Raman	CBDsiPLS	62.50	66.67	60.00
	siPLS	59.38	41.67	70.00
	PLS	50.00	41.67	55.00

CBDsiPLS 方法之所以能够在光谱数据集上取得如此出色的表现,主要得益于算法的设计。该方法基于化学键的振动特性,在光谱数据中选择具有特定化学意义的区间进行联合分析。这种策略不仅减少了冗余数据对模型性能的干扰,还使得模型更加专注于与心血管疾病风险相关的特征。因此,CBDsiPLS 方法能够更有效地提取出与心血管疾病相关的光谱信息,从而提高模型的预测准确性。

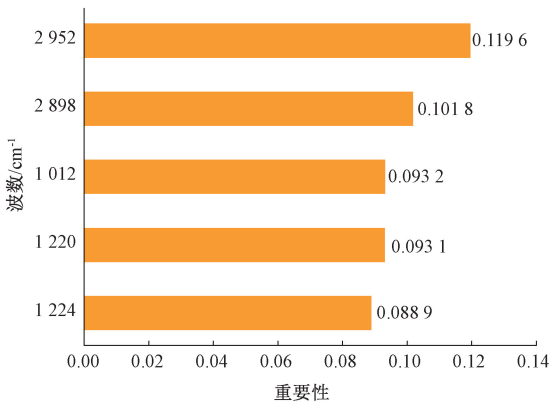
为了进一步验证 CBDsiPLS 方法的优势,对特征重要性排序如图 4 所示,可以看出特征重要性排名前五的特征波数以及其对应的重要性程度。从图 4(a)可以看出,FTIR-ATR 的特征波数主要集中在 1 750~1 650 cm^{-1} 区间内,这一区间对应于酰胺Ⅰ带的波峰。酰胺Ⅰ带包含了丰富的蛋白质二级结构信息,且与心血管疾病相关的某些化学物质,如 TG、HDL-C 等,有着密切的关联。此外,图 4(b)中 Raman 光谱特征波数集中在 3 000~2 800 和 1 250~1 000 cm^{-1},与血液中脂肪酸,甲基基团等物质相关性很高。据此分析,CBDsiPLS 方法选出的贡献度较高的特征主要集中在与心血管疾病标志物相关的波段上。这些特征不仅与心血管疾病的风险密切相关,而且在光谱数据中具有较高的辨识度。这一发现进一步证明了 CBDsiPLS 方法的有效性和针对性。因此决定将

CBDsiPLS 提取的特征作为特征融合的基础。



(a) FTIR-ATR光谱特征重要性排序

(a) FTIR-ATR spectral feature importance ranking



(b) Raman光谱特征重要性排序

(b) Raman spectral feature importance ranking

图 4　特征重要性排序

Fig. 4　Feature importance ranking

3.2　融合数据模型

在单一数据模型的基础上,将进一步采用 CBDsiPLS 算法提取的 FTIR-ATR 特征和 Raman 特征进行融合,形成融合数据矩阵,并建立融合数据分类模型。分别使用随机森林,XGBoost 和 SVM 建立了基于融合数据的分类模型,融合后的特征维度分别是 14 维,10 维和 15 维。为显示数据融合的作用,将融合模型的实验结果与最优单一模型(使用 CBDsiPLS 进行特征提取而建立的分类模型)的实验结果进行了对比,实验结果对比如表 4~6 所示,ROC 曲线结果对比如图 5 所示。

表 4　单一数据和融合数据的随机森林模型分类结果对比

Table 4　Comparison of classification results of random forest models with single data and fusion data　(%)

数据类型	准确率	灵敏度	特异性
FTIR-ATR	81. 25	75. 00	85. 00
Raman	62. 50	58. 33	65. 00
融合数据	90. 63	83. 33	95. 00

表 5　单一数据和融合数据的 XGBoost 模型分类结果对比

Table 5　Comparison of classification results of XGBoost models with single data and fusion data　(%)

数据类型	准确率	灵敏度	特异性
FTIR-ATR	84. 38	83. 33	85. 00
Raman	62. 50	58. 34	65. 00
融合数据	90. 63	83. 33	95. 00

表 6　单一数据和融合数据的 SVM 模型分类结果对比

Table 6　Comparison of classification results of SVM models with single data and fusion data　(%)

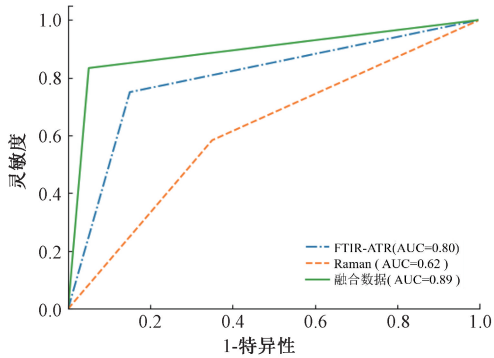
数据类型	准确率	灵敏度	特异性
FTIR-ATR	78. 13	66. 67	75. 00
Raman	62. 50	66. 67	60. 00
融合数据	90. 63	83. 33	95. 00

从表 4~6 可以看出,通过将 FTIR-ATR 和 Raman 数据进行融合,融合数据的模型分类性能得到了显著提升,表现出更高的分类准确率、灵敏度和特异性。例如,随机森林模型的融合数据准确率提升至 90. 63%,灵敏度提升至 83. 33%,特异性保持在 95. 00%。这表明融合数据能够更有效地捕捉到心血管疾病的相关特征。融合模型在不同算法下(随机森林、XGBoost 和 SVM)均显示出性能提升,表明该数据融合方法具有较好的稳定性和普适性。这对于实际应用中的模型选择提供了更大的灵活性。此外,观察图 5(a)~(c)可知,ROC 曲线下的积分面积表示 AUC 值,AUC 值越大,模型越可靠,可以看出融合数据模型的曲线下面积是最大的,均高于单一数据模型,这进一步支持了数据融合的有效性。

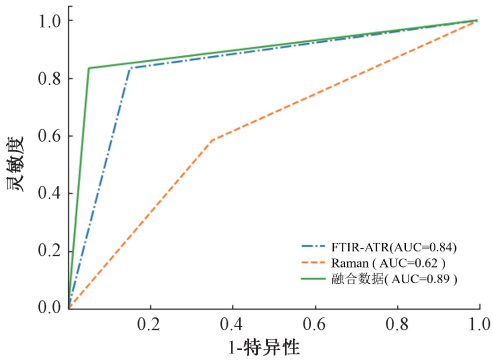
从信息互补性的角度分析,FTIR-ATR 和 Raman 光谱提供了不同维度的分子信息,FTIR-ATR 捕捉的是分子中的化学键振动,Raman 则对分子中对称振动更敏感,提取的融合特征包括 FTIR-ATR 中蛋白质的酰胺 I 带和 Raman 中与 CH2 等相关区域,这些特征分别代表了样本中与疾病相关的蛋白质和脂质的生化信息,融合后能够更全面地反映疾病的风险特征。数据融合通过结合两者的优势,显著提升了模型的分类性能。

4　结　论

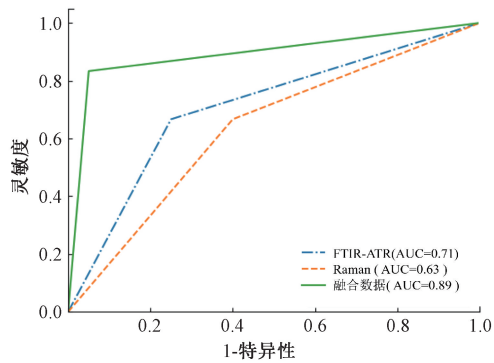
针对使用全血光谱数据建立的心血管疾病风险评估模型性能指标低下的问题,提出了一种 CBDsiPLS 用于特征提取。CBDsiPLS 算法通过依据分类目标的相关化学键进行波段选取,并结合 PLS 算法提取特征。相比传统的 siPLS 算法,CBDsiPLS 在波段划分时不再依赖于传统的平均划分区域的方法,能更有效地从复杂的光谱信息中提取出与心血管疾病风险相关的特征,在减少随机性



(a) 单一数据和融合数据的随机森林模型ROC曲线
(a) ROC curve of random forest model with single data and fusion data



(b) 单一数据和融合数据的XGBoost模型ROC曲线
(b) ROC curve of XGBoost model with single data and fused data



(c) 单一数据和融合数据的SVM模型ROC曲线
(c) ROC curve of SVM model with single data and fused data

图 5　单一数据和融合数据各模型的 ROC 曲线

Fig. 5　ROC curves of single data and fused data models

的同时,特征选择的针对性更强,同时减低了计算复杂度。进一步,本研究采用信息融合的方式,将 FTIR-ATR 和 Raman 光谱数据进行特征融合,构建了融合数据模型。与单一光谱数据模型相比,融合数据模型在各项评价指标上均表现出显著提升,充分发挥了光谱数据信息互补的优势。综上所述,本文提出的 CBDsiPLS 算法结合信息融合策略有效突破了单一光谱的局限性,为心血管疾病的风险评估提供了更为精确和可靠的解决方案。未来,随着临床数据的不断积累和算法的进一步优化,考虑结合其他生物标志物或多模态信息,以拓展该方法在更广泛的疾病筛查领域的应用。

参考文献

[1] IZZO C, CARRIZZO A, ALFANO A, et al. The impact of aging on cardio and cerebrovascular diseases[J]. International Journal of Molecular Sciences, 2018, 19(2): 481.

[2] BAYS H E, TAUB P R, EPSTEIN E, et al. Ten things to know about ten cardiovascular disease risk factors[J]. American Journal of Preventive Cardiology, 2021, 5: 100149.

[3] TOWNSEND N, KAZAKIEWICZ D, LUCY W F, et al. Epidemiology of cardiovascular disease in europe[J]. Nature Reviews Cardiology, 2022, 19(2): 133-143.

[4] SANI M H, KHOSROABADI S. A novel design and analysis of high-sensitivity biosensor based on nano-cavity for detection of blood component, diabetes, cancer and glucose concentration[J]. IEEE Sensors Journal, 2020, 20(13): 7161-7168.

[5] WANG Q, SONG S, LI L, et al. An extreme learning machine optimized by differential evolution and artificial bee colony for predicting the concentration of whole blood with Fourier transform raman spectroscopy[J]. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2023, 292: 122423.

[6] CAMERON J M, BUTLER H J, ANDERSON J D, et al. Exploring pre-analytical factors for the optimization of serum diagnostics: Progressing the clinical utility of ATR-FTIR spectroscopy[J]. Vibrational Spectroscopy, 2020,109:103092-103104.

[7] CHEN F Y, SUN C, YUE Z, et al. Screening ovarian cancers with Raman spectroscopy of blood plasma coupled with machine learning data processing[J]. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2022, 265: 120355.

[8] SONG H T, DONG C, ZHANG X, et al. Rapid identification of papillary thyroid carcinoma and papillary microcarcinoma based on serum Raman spectroscopy combined with machine learning models[J]. Photodiagnosis and Photodynamic Therapy, 2022, 37: 102647.

[9] YUE F, CHEN C, YAN Z, et al. Fourier transform infrared spectroscopy combined with deep learning and data enhancement for quick diagnosis of abnormal thyroid function[J]. Photodiagnosis and Photodynamic Therapy, 2020, 32: 101923.

[10] MORDECHAI S, SHUFAN E, KATZ B P, et al. Early diagnosis of Alzheimer's disease using infrared spectroscopy of isolated blood samples followed by

multivariate analyses[J]. Analyst, 2017,142(8): 1276-1284.

[11] LI Q, ZHANG Z, MA Z. Raman spectral pattern recognition of breast cancer: A machine learning strategy based on feature fusion and adaptive hyperparameter optimization[J]. Heliyon, 2023, 9(7):e18148.

[12] WU X, CHEN C, CHEN X, et al. Raman spectroscopy combined with machine learning algorithms for rapid detection primary Sjögren′s syndrome associated with interstitial lung disease [J]. Photodiagnosis and Photodynamic Therapy, 2022, 40: 103057.

[13] GAO F, BEN-AMOTZ D, YANG Z, et al. Complementarity of FT-IR and raman spectroscopies for the species discrimination of meat and bone meals related to lipid molecular profiles[J]. Food Chemistry, 2021, 345: 128754.

[14] LISAK M, DEMARIN V, TRKANJEC Z, et al. Hypertriglyceridemia as a possible independent risk factor for stroke[J]. Acta Clin Croat, 2013, 52(4): 458-463.

[15] PETERS S A E, SINGHATEH Y, MACKAY D, et al. Total cholesterol as a risk factor for coronary heart disease and stroke in women compared with men: A systematic review and meta-analysis [J]. Atherosclerosis, 2016, 248: 123-131.

[16] 魏新园,钱牧云,冯旭刚,等. 基于偏最小二乘的数控机床热误差稳健建模算法[J]. 仪器仪表学报,2021, 42(5):34-41.
WEI X Y, QIAN M Y, FENG X G, et al. Robust modeling algorithm for thermal error of numerical control machine tool based on partial least squares[J]. Journal of Instrumentation and Measurement, 2021, 42 (5): 34-41.

[17] 武小红, 孙俊, 武斌,等. 基于联合区间偏最小二乘判别分析的猪肉近红外光谱定性建模分析[J]. 激光与光电子学进展, 2015, 52(4): 043003.
WU X H, SUN J, WU B, et al. Qualitative modeling analysis of pork near-infrared spectroscopy based on joint interval partial least squares discriminant analysis [J]. Progress in Laser and Optoelectronics, 2015, 52(4):043003.

[18] YANG Z, XIAO H, ZHANG L, et al. Fast determination of oxides content in cement raw meal using NIR spectroscopy combined with synergy interval partial least square and different preprocessing methods [J]. Measurement, 2020, 149: 106990.

[19] XU Z, LI X, CHENG W, et al. Rapid and accurate determination methods based on data fusion of laser-induced breakdown spectra and near-infrared spectra for main elemental contents in compound fertilizers [J]. Talanta, 2024, 266: 125004.

[20] BARAL S, SHYAM KUMAR B K, KSHETRI R. Study of triglyceride glucose index and total cholesterol/HDLc for assessment of cardiovascular outcomes in patients with diabetes and hypertension[J]. American Heart Journal, 2021, 242:148.

[21] DING Z, MEI G, CUOMO S, et al. Comparison of estimating missing values in iot time series data using different interpolation algorithms [J]. International Journal of Parallel Programming, 2020, 48: 534-548.

[22] LI Z, LAN X, JIANG X, et al. Triglyceride and high density lipoprotein cholesterol concentrations quantitative analysis in whole blood by FTIR-ATR spectroscopy and FT-Raman spectroscopy[J]. Analytical Methods, 2018, 10(46): 5493-5498.

[23] PEREZ-GUAITA D, GARRIGUES S. Infrared-based quantification of clinical parameters[J]. TrAC Trends in Analytical Chemistry, 2014, 62: 93-105.

[24] PARASKEVAIDI M, MATTHEW B J, HOLLY B J, et al. Clinical applications of infrared and Raman spectroscopy in the fields of cancer and infectious diseases[J]. Applied Spectroscopy Reviews, 2021, 56(8-10): 804-868.

[25] CZAMARA K, MAJZNER K, PACIA M Z, et al. Raman spectroscopy of lipids: A review[J]. Journal of Raman spectroscopy, 2015, 46(1): 4-20.

[26] HAIR J F, HOWARD M C, NITZL C. Assessing measurement model quality in PLS-SEM using confirmatory composite analysis[J]. Journal of Business Research, 2020, 109: 101-110.

[27] 戴嘉伟,王海朋,陈瀑,等. 多光谱数据融合分析技术的研究和应用进展[J]. 分析化学, 2022, 50(6): 839-849.
DAI J W, WANG H P, CHEN P, et al. Research and application progress of multispectral data fusion analysis technology[J]. Analytical Chemistry, 2022,50(6):839-849.

[28] 梁海波,王怡. 基于深度学习的天然气钢制管道缺陷检测方法研究[J]. 电子测量与仪器学报, 2022, 36(9):148-158.
LIANG H B, WANG Y. Research on defect detection method of natural gas steel pipeline based on deep learning[J]. Journal of Electronic Measurement and Instrumentation, 2022,36(9):148-158.

[29] CHEN T Q, HE T, BENESTY M, et al. Xgboost: Extreme gradient boosting[J]. R Package Version 0.4-2,2015,1(4):1-4.

[30] HUANG S, CAI N, PACHECO P P, et al. Applications

of support vector machine (SVM) learning in cancer genomics[J]. Cancer Genomics & Proteomics, 2018, 15(1): 41-51.

作者简介

何洋,2022 年于东北大学获得学士学位,现为东北大学在读硕士研究生,主要研究方向为光谱分析与信号处理。
E-mail: 18841092611@163.com

He Yang received his B. Sc. degree from Northeastern University in 2022. Now he is a M. Sc. candidate at Northeastern University. His main research interests include spectral analysis and signal processing.

李志刚(通信作者),1999 年于燕山大学获得学士学位,2002 年于燕山大学获得硕士学位,2005 年于天津大学获得博士学位,现为东北大学副教授,主要研究方向为光谱分析与信号处理。
E-mail: lizhigang@neuq.edu.cn

Li Zhigang (Corresponding author) received his B. Sc. degree from Yanshan University in 1999, M. Sc. degree from Yanshan University in 2002, Ph. D. degree from Tianjin University in 2005. Now he is an associate professor in Northeastern University. His main research interests include spectral analysis and signal processing.