doi:10.6043/j.issn.0438-0479.202404001

机器学习指导下的椎间盘组织工程体系预测模型建立

韩照普[1],尹 正[2],叶晓健[1*]
(1.上海交通大学医学院附属同仁医院骨科,上海 200336;2.中国医学科学院阜外医院国家心血管病中心心血管代谢中心,北京 100037)

摘要:[**目的**]通过多种机器学习方法建立多特征参数对椎间盘支架生物学效应的预测模型,以降低实验试错成本,实现椎间盘退行性病变组织工程治疗的高效化.[**方法**]基于近 10 年(2014—2024 年)椎间盘支架相关材料、工艺、细胞培养数据,构建特征多维、标签多样的综合数据集,进而采用岭分类器、高斯朴素贝叶斯(Gaussian NB)、线性判别分析(LDA)、高斯过程分类器(GPC)等 21 种分类算法,评估各模型在数据集上的预测性能.[**结果**]多数分类模型在本数据集上的预测一致性较好,且运行时间较短,其中基于岭分类器算法构建的机器学习模型能最准确地预测支架上的细胞行为;进一步通过模型可解释性分析,可获知各支架特征参数生物学效应的方向和程度,从而使该模型具有较高的推广价值.[**结论**]机器学习方法对于椎间盘组织工程体系的开发具有指导意义,通过恰当的模型构建,研究人员可以有针对性地进行特征组合,从而简单快捷地设计出具有良好生物学活性的支架.

关键词:椎间盘退行性病变;机器学习;组织工程;支架
中图分类号:R681.5 **文献标志码**:A **文章编号**:0438-0479(2025)05-0886-11

Construction of a prediction model of intervertebral disc tissue engineering system guided by machine learning

HAN Zhaopu[1], YIN Zheng[2], YE Xiaojian[1*]
(1. Department of Orthopedics, Tongren Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 200336, China; 2. Cardiometabolic Center, National Center for Cardiovascular Diseases, Fuwai Hospital, Chinese Academy of Medical Sciences, Beijing 100037, China)

Abstract:[**Objective**] Intervertebral discs degeneration (IVDD) is a major contributor to back, neck, and radicular pain, which mainly involves the pathological process of dehydration and degeneration of the nucleus pulposus, weak rupture of the annulus fibrosus, and calcification of the cartilage endplate defect. These degenerative changes lead to compromised structural integrity and function of the intervertebral discs, resulting in significant pain and disability. Current treatment options for IVDD are limited and often focus on symptom management rather than addressing the underlying pathology. Intervertebral disc tissue engineering (IDTE) has emerged as a promising approach to improve outcomes of IVDD by promoting the regeneration of disc tissues and restoring their function. This study aims to construct a prediction model of the biological effects of intervertebral disc scaffolds using various machine learning (ML) methods. The goal is to reduce experimental costs associated with trial and error, and to achieve efficient treatment of tissue engineering for intervertebral disc degenerative diseases. [**Methods**] To develop a robust prediction model, data were mined from literature on intervertebral disc scaffold-related materials, technological processes, and cell culture conditions published over the past decade. This comprehensive data set included multiple features and diverse labels, capturing a wide range of

收稿日期:2024-04-01 **录用日期**:2025-03-12
基金项目:上海交通大学医工交叉重点项目(YG2021ZD34)
*通信作者:yxj4380@sjtu.edu.cn
引文格式:韩照普,尹正,叶晓健.机器学习指导下的椎间盘组织工程体系预测模型建立[J].厦门大学学报(自然科学版),2025,64(5):886-896.
Citation: HAN Z P, YIN Z, YE X J. Construction of a prediction model of intervertebral disc tissue engineering system guided by machine learning[J]. J Xiamen Univ Nat Sci, 2025, 64(5): 886-896. (in Chinese)

variables that influence scaffold performance and cell behavior. The data set was then used to train and evaluate the performance of 21 classification algorithms, including ridge classifier, Gaussian naive Bayes (Gaussian NB), linear discriminant analysis (LDA), and Gaussian process classifier (GPC). Each algorithm was assessed for predictive accuracy, generalization ability, and computational efficiency. [**Results**] Through literature retrieval, data extracting, and data preprocessing, a data set containing 10 principle scaffold materials, 5 modeling ingredients, 5 bioactive factors, 5 scaffold forms, and 10 cell types was built. Among all features, gelatin, tetrazine-norbornene (Tz-Nb), transforming growth factor β (TGF-β), hydrogel and human nucleus pulposus cells were most frequently applied. As for targets of the data set, 69 cases were matched with label 0, suggesting unsuccessful modeling of IDTE scaffolds under specific feature combination. 60 cases with label 1 were identified due to the insignificant improvement of biological effect for IVDD-related cells compared to conventional cultural environment. Finally, 82 cases of scaffold systems were proved to have positive effects on cell behavior, which were labeled with 2. After data input, ML training and predicting, a series of models indicating the cellular effects of IDTE scaffolds were established based on the data set. The results demonstrated that most of the included classification models showed relatively consistent performances, with relatively short execution time. Among them, the ML model based on ridge classifier algorithm provided the best accurate prediction for cell behavior on scaffolds. This model was found to have a high generalization value, indicating its potential for widespread application in IDTE. In addition, the shapley additive explanations (SHAP) model was applied to attribute each IDTE features to bio-target labels. [**Conclusion**] In summary, ML methods hold significant promise for guiding the design and exploitation of IDTE systems. Through appropriate model construction, researchers can selectively combine features to design scaffolds with good biological activity quickly and easily. The use of prediction model can streamline the development process, reduce the reliance on extensive empirical testing, and accelerate the translation of IDTE from the laboratory to clinical applications. By leveraging the predictive power of ML, it is possible to identify optimal scaffold designs that promote tissue regeneration and improve patient outcomes. This approach represents a major advancement in the field of tissue engineering and may offer a new avenue for the effective treatment of IVDD.

Keywords: intervertebral disc degeneration; machine learning; tissue engineering; scaffold

椎间盘退行性病变(IVDD)是一种常见的运动系统老化性改变,与盘源性腰痛、椎管狭窄、脊柱不稳等疾病密切相关[1-3]. 目前,针对 IVDD 的治疗以缓解疼痛为主要目的,而并未聚焦于退变发生发展的机制,因此往往导致疗效不确定、症状复发及邻近椎体节段出现并发症等问题[4]. 近年来,学界致力于通过激活椎间盘固有组分再生,恢复椎间盘结构与功能特性的方式治疗 IVDD 相关疾病[4-5]. 其中,椎间盘组织工程(IDTE)是颇有前景的一种方案. 一般地,IDTE 方法将具有生物相容性的材料与细胞、生物活性因子相结合,借助适当的制备技术构建辅助椎间盘修复的支架[6-7].

理想的 IDTE 支架应能模拟天然椎间盘基质,为局部组织细胞提供良好的生长、增殖、黏附、迁移及分化条件,从而创造高活性、可持续的椎间盘再生环境[7-8]. 然而,支架制备过程涉及的工艺参数过多,在没有明确理论指导的条件下,基于经验设计的 IDTE 支架并不总能如期达成良好的生物学效应. 为优选生物学效应良好的 IDTE 支架,传统上需要大量时间成本和经济成本的试错,这严重阻碍了 IDTE 支架的高效开发[9]. 近年来,机器学习(ML)在组织工程领域的引入有望降低支架开发的试错成本,精准筛选制备参数[10-11].

在优化工艺参数、提高支架的物理性能方面,ML 具有显著优势. Conev 等[12]利用随机森林(RF)分类和回归算法模拟印刷参数对 3D 打印支架质量的影响,提示树结构模型在工艺参数的选择中常具有优异的性能. Menon 等[13]提出了一种通用的分层 ML 模型,用于预测不同单体组成和链间结构下聚氨酯材料的机械行为. 在预测工程支架的生物学效应方面,ML 的应用集中于体外模型分析[11]. Echezarreta-López 等[14]利用大型数据集进行模糊神经网络分析,证实了生物活性玻璃的抗菌活性与其钙离子含量直接相关. Rafieyan 等[15]通过对多种分类模型的比较评估,较全面地建立了心脏组织工程支架相关生产参数与细胞生长作用的联系. 为获取足够的训练样本量及满足多维度特征的要求,此类研究的数据采集往往基于大规模文献汇总,具有较强的普适性.

目前在 IDTE 领域 ML 的应用尚不多见,这严重限制了 IDTE 支架的进一步开发. 为此,本研究利用 ML 建立了一个全面的分类模型来预测 IDTE 支架上的细胞行为. 首先从文献中筛选样本、规范特征参数并分配标签,建立基于支架主材、成型辅料、活性因子、支架形式和细胞种类的 IDTE 支架数据集;然后针对该数据集,应用 21 种不同的 ML 算法进行训练、测

试和评估,比较不同 ML 算法的预测性能;最后筛选出具有最佳预测性能的 ML 模型,并就该模型进行事后可解释性分析,以期阐明各特征参数对于 IDTE 体系生物学效应的贡献,推动 IDTE 支架的高效开发.

1 研究方法

1.1 数据挖掘

利用 Scopus 数据库(https://www.scopus.com/),以"intervertebral disc""scaffold""tissue engineering"为关键词检索文献. 学科领域限定为"Engineering, Multidisciplinary, Biochemistry, Genetics and Molecular Biology, Materials Science, Medicine, and Chemical Engineering",文献类型限定为"Article",发表时间限定为 2014—2024 年. 为检索更多高质量文献,避免单一数据库的遗漏,在 PubMed 数据库(https://pubmed.ncbi.nlm.nih.gov/?myncbishare=pubmedplus)和 Google Scholar 数据库(https://scholar.google.cz/schhp?hl=zh-CN)利用相同检索策略进行补充查阅. 此外,为纳入更多样本,防止因训练集过小、标签不平衡而引起的欠拟合或过拟合,还在纳入文献数据的基础上通过数据扩充技术进行数据增强.

1.2 数据预处理

根据参数在支架制备过程中的代表性及其在文献中的普遍性,选取支架主材、成型辅料、活性因子、支架形式和细胞种类 5 类特征,并提取各样本对应的数据. 在特征选择阶段,为防止样本与特征数量不平衡所引起的过拟合风险,利用最大信息系数(MIC)衡量提取的原始特征与结局指标的相关性;分别对 5 类特征进行组内 MIC 排序,筛选出 10 种支架主材、5 种成型辅料、5 种活性因子、5 种支架形式和 10 种细胞种类,共 35 个特征. 进一步经过数据约简及归一化操作,归纳一个特征维度 35、样本数量 211 的数据集来实现 ML 算法[支撑材料(https://jxmu.xmu.edu.cn/Upload/html/20250516)S1]. 该数据集的每一列代表一个具体特征,且每列中的值具有相同的单位. 对于支架主材,表格内数值表示该材料在所有参与支架制造的主材中所占质量分数;对于其余特征,只存在两种对应数值,0 表示不存在,1 表示存在. 数据集中的最后一列为标签,代表特定因素组合下支架的生物学效应. 根据细胞对支架的反应定义了 3 类标签,其具体含义如表 1 所示.

表 1 数据集标签的划分标准

Tab. 1 Classification criteria of the dataset label

标签	生物学效应
0	在给定的特征条件下,无法形成具有工程意义的支架
1	对于在支架上培养的特定细胞,其增殖、迁移、黏附、分化、分泌等行为相较于对照组(如培养板)细胞没有显著提高
2	对于在支架上培养的特定细胞,其增殖、迁移、黏附、分化、分泌等行为显著强于对照组

1.3 ML 模型

本研究共纳入 21 种 ML 模型用于多分类任务,通过多种评价指标的综合比较来确定最适合的模型. 利用 5 折交叉验证,将数据集按 8∶2 的比例随机分为训练集和测试集,在训练集上训练并通过网格搜索确定最佳超参数. 以下分别简述本研究中各类 ML 模型的定义与特点,并展示基于 scikit-learn 1.4.1.post1 及 pytorch 2.2.0 优化的超参数,详见表 2[16-31].

1.4 模型评价

参考 Rafieyan 等[15]的方法,选取 7 项指标从不同角度评估模型的预测能力,包括准确率(accuracy)、精度(precision)、召回率(recall)、F1 分数(F1 score)、科恩的 kappa 系数(CKC)、马修斯相关系数(MCC)和受试者操作特征曲线下面积(ROC_AUC)[32]. 其中,准确率是对整体正确率的综合衡量,在相关研究中应用最广泛,但当样本不平衡时该指标容易失效;精度和召回率是从两个角度对阳性预测的可靠程度进行衡量的指标,二者常常不能同时达到最大值;而 F1 分数作为精度和召回率的调和平均数,能够在保持二者平衡的情况下给出较稳健的评价;CKC 常被用来衡量两个评价体系之间的一致性程度,对模型分类过程中造成的偶然一致性进行校正;MCC 作为综合性的分类指标,对于小样本及类别不平衡的情况较适用;ROC_AUC 是综合所有可能的分类阈值而给出的一种预测效果综合评价,可以解读为模型将某个随机阳性样本排列在某个随机阴性样本之上的概率,常被视作最具代表性的评价标准. 本研究在分别测算各模型的 7 种评价指标的基础上,还求和得出一个综合指标,通过比较该指标在 21 种 ML 模型中的数值大小,选出预测 IDTE 支架上细胞行为的最佳 ML 模型,并借助混淆矩阵对该模型的预测性能进行可视化展示. 此外,模型的计算时间成本也是评价其推广价值的重要考察对象,因此本研究还利用训练时间和预测时间两个指

表 2　21 种 ML 模型的概括描述

Tab. 2　General description for 21 different kinds of ML model

模型名称	英文名称	模型特点	待优化超参数
K 最近邻[16]	K-nearest neighbors (KNN)	非参数而基于实例的学习方法,通过多数表决确定未知类别	'n_neighbors' 'weights' 'distance'
极限梯度提升[17]	extreme gradient boosting (XGBoost)	通过迭代训练多个弱学习器提升模型性能,可采用多种策略进行多分类扩展	'learning_rate' 'max_depth' 'n_estimators'
分类提升[17]	categorical boosting (CatBoost)	直接处理类别特征,支持自动特征缩放,对缺失值友好	'depth' 'iterations' 'learning_rate'
光梯度提升机分类器[17]	light gradient boosting machine (LGBM) classifier	基于直方图的决策树学习,有效减少内存占用和计算时间	'colsample_bytree' 'learning_rate' 'n_estimators' 'subsample'
决策树[18-19]	decision tree (DT)	基于树结构,通过对特征的逐步划分进行预测或分类	'criterion' 'max_depth' 'min_samples_leaf' 'min_samples_split'
随机森林[19]	random forest (RF)	通过对多个 DT 的输出进行组合,使预测更加稳健	'max_depth' 'min_samples_leaf' 'min_samples_split' 'n_estimators'
极端随机树分类器[20]	extremely randomized trees (Extra Trees) classifier	与 RF 类似,但增加了更多随机性	'max_depth' 'min_samples_leaf' 'min_samples_split' 'n_estimators'
正则化贪婪森林分类器[21]	regularized greedy forest (RGF) classifier	基于回归型树的集成模型,通过贪婪扩展和正则化损失函数最小化来构建树	'algorithm' 'l2' 'max_leaf' 'min_samples_leaf'
基于 pytorch 框架的朴素神经网络[10,22]	neural network (NN)	由多个节点组成,通过相互连接进行信息传递,可学习和推断复杂的非线性关系	'lr' 'hidden_size' 'batch size' 'activation function' 'num_epochs' 'stochastic optimization'
逻辑回归[10]	logistic regression	经典的二分类算法,通过策略扩展用于多分类问题	'C' 'penalty':'l2'
标准支持向量机分类器[23]	support vector machine (SVM)	将输入特征映射到高维空间,通过找出最适超平面来区分不同类别样本	'C' 'gamma' 'kernel'
线性支持向量机分类器[23]	linear support vector classifier (Linear SVC)	基于 SVM 的优化算法,适用于线性可分或近似线性可分的分类问题	'C' 'loss' 'penalty'
线性判别分析[24-25]	linear discriminant analysis (LDA)	通过最大化组间差异和最小化组内差异进行降维和分类	'shrinkage' 'solver'
二次判别分析[24-25]	quadratic discriminant analysis (QDA)	允许每个类别具有不同的协方差矩阵,相对于 LDA 拟合不同类别数据更准确	'reg_param'

续表

模型名称	英文名称	模型特点	待优化超参数
标签传播[26]	label propagation	基于图结构的半监督学习算法,通过已知标签节点向未知节点传播实现标签推断	‘gamma’ ‘kernel’ ‘n_neighbors’
高斯过程分类器[27]	Gaussian process classifier (GPC)	基于高斯过程,通过贝叶斯推断预测新样本类别概率分布	‘kernel’
岭分类器[28]	ridge classifier	基于 ridge 回归的分类模型,在高维数据上表现良好	‘alpha’ ‘solver’
多层感知机分类器[29]	multi-layer perceptron (MLP) classifier	基于人工神经网络,通过前向传播和反向传播训练模型进行分类预测	‘alpha’ ‘hidden_layer_sizes’
高斯朴素贝叶斯[30]	Gaussian naive Bayes (Gaussian NB)	假设特征变量服从高斯分布,以各类别下特征变量的概率密度分布进行分类	‘var_smoothing’
伯努利朴素贝叶斯[30]	Bernoulli naive Bayes (Bernoulli NB)	适用于布尔型特征变量的分类问题	‘alpha’ ‘binarize’ ‘fit_prior’
被动攻击分类器[31]	passive aggressive classifier (PAC)	基于每个实例对模型参数进行误差适应性调整,减少分类错误	‘C’ ‘fit_intercept’

标来评估 ML 模型在归纳数据集上的运行时长.

1.5 可解释性分析

多数 ML 模型属于“黑盒模型”,存在解释困难的问题. SHAP(shapley additive explanations)模型是一种较全面的模型解释方法,其核心要义是在模型预测事后,通过计算各特征对于结局指标的边际贡献,综合构建一个加性解释模型,从而在全局水平和单个样本水平反映各特征的作用方向及程度. 本研究利用 SHAP 模型对筛选出的最佳 ML 模型进行全局解释,分析多种支架特征对于 IDTE 支架上细胞行为的影响.

1.6 分析工具

使用 AutoDL 软件、Python 3.12 软件及 WPS 软件(12.1.0.16729)进行数据汇总和分析. 主要运用的 Python 包有 numpy 1.26.3、matplotlib 3.6.3、scikit-learn 1.4.1.post1、pandas 2.2.0、pytorch 2.2.0 等.

2 结果与分析

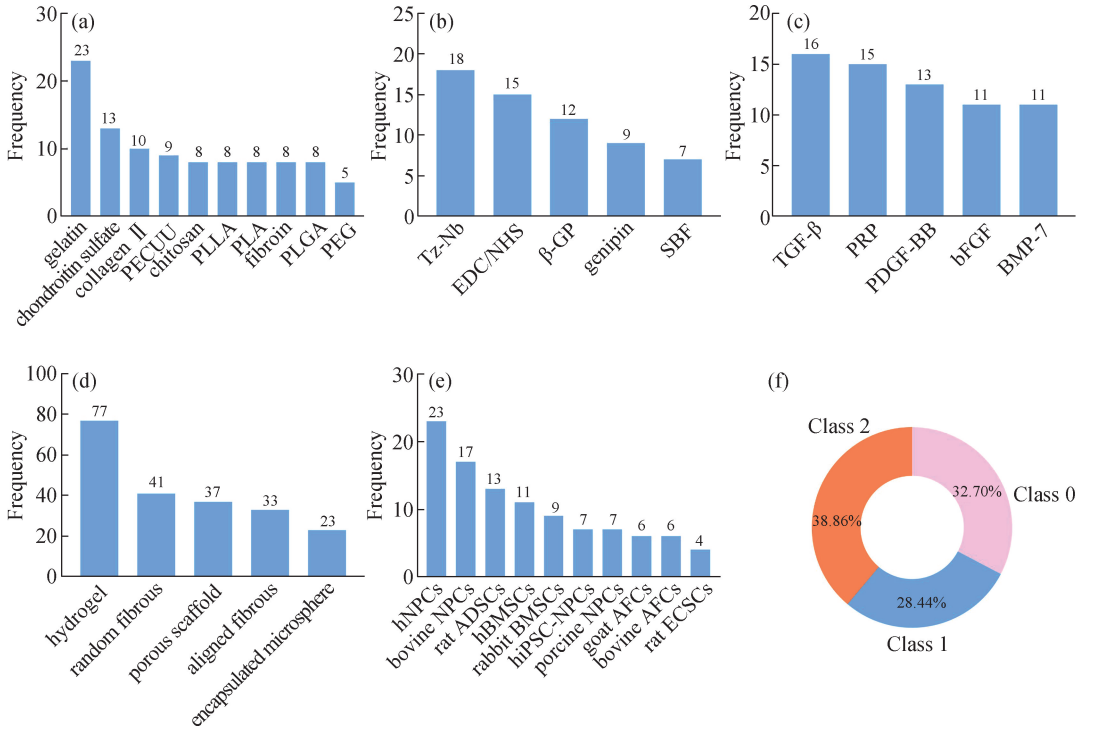
2.1 描述性统计

为理解 IDTE 中各参数的重要性,并概述该领域研究中的常用细胞,在进行正式 ML 模型训练前,先对预处理的数据集进行描述性统计(图 1). 总体而言,数据集包含 10 种支架主材、5 种成型辅料、5 种活性因子、5 种支架形式和 10 种细胞种类. 支撑材料 S2 汇总了数据集内所有特征对应的中英文名称及缩写.

如图 1(a)与(b)所示,在本数据集中制造 IDTE 最常用的支架主材和成型辅料分别为明胶(gelatin)和四嗪-降冰片烯(Tz-Nb). 图 1(c)表明,5 种筛选出的活性因子,即转化生长因子 β(TGF-β)、富血小板血浆(PRP)、血小板源性生长因子 BB(PDGF-BB)、碱性成纤维细胞生长因子(bFGF)及人骨成型蛋白 7(BMP-7)的应用分布较为平均,反映出这些活性组分在 IDTE 相关研究中皆占有重要地位. 图 1(d)显示,5 种支架形式中水凝胶(hydrogel)最常被研究,这可能与其合成工艺相对简单,机械性能适配髓核且早期研究较多有关;而封装微球(encapsulated microsphere)作为较新颖的支架形式,目前研究相对较少. 如图 1(e)所示,就细胞种类而言,经 MIC 选择后纳入数据集的 10 种细胞中,有 4 种为髓核细胞(NPCs),这凸显了髓核修复在 IVDD 组织工程治疗中的核心地位. 在细胞反应方面,69 例特征组合匹配标签 0,表明其由于缺乏支架主材或成型辅料搭配不当,无法形成 IDTE 支架;60 例特征组合匹配标签 1,说明该参数下形成的支架对于细胞的生物学效应不明显优于常规培养板的作用;82 例特征组合匹配标签 2,显示出这些样本所对应的 IDTE 支架对于细胞行为具有正向作用. 如图 1(f)所示,结局指标的类别相对平衡,降低了 ML 分类的预测偏向性风险.

2.2 模型超参数优化结果

对 KNN 模型而言,由于本研究数据集样本稀疏,



(a)支架主材频率分布;(b)成型辅料频率分布;(c)活性因子频率分布;(d)支架形式频率分布;(e)细胞种类应用频率分布;(f)细胞反应分类.

图 1　数据集描述性统计

Fig. 1　Descriptive statistics of dataset

各特征分布杂乱,故‘weights’参数选用‘distance’模式,使每个样本的近邻样本权重和距离成反比,从而降低不可信赖样本的影响.

对基于树结构的模型(XGBoost、CatBoost、LGBM classifier、DT、RF、Extra Trees classifier、RGF classifier)而言,代表子树数量的参数‘n_estimators’/‘iterations’对于模型(除 DT 外)的性能影响最大,该参数数值的增加可提高模型复杂度,使预测性能更稳健,但同时伴随过拟合风险的增加;代表树深度的参数‘depth’/‘max_depth’可控制模型(除 LGBM classifier 外)在训练过程中达到的最大层数,过度提高其数值可能导致训练集中噪声的捕获,从而引起过拟合风险的增加;学习率‘learning_rate’一般和上述两种参数呈现反相关关系,稳健的模型常伴随较低的学习率;而叶子的最小样本数量参数‘min_samples_leaf’及分支节点的最小样本数量参数‘min_samples_split’被用作限制树结构中的分枝发生,其数值越大表示模型复杂度越低,在降低过拟合风险的同时将增大模型学习的阻力. 针对本研究所归纳的数据集,几种基于树结构的模型均以适中的‘n_estimators’/‘iterations’、‘depth’/‘max_depth’、‘learning_rate’和较低的‘min_samples_leaf’、‘min_samples_split’优化超参数,反映出在处理数据量小、特征稀疏的样本时提高模型复杂度和降低过拟合风险之间的平衡. 特别地,为提高模型泛化能力,LGBM classifier 模型还使超参数‘colsample_bytree’和‘subsample’数值小于 1,在加速训练的同时降低过拟合风险;RGF classifier 模型将超参数‘algorithm’设为‘RGF’,通过‘RGF’及‘l2’正则化的单叶模型提高模型的鲁棒性.

对神经网络相关模型而言,NN 表示基于 pytorch 进行学习的朴素神经网络,而 MLP classifier 相对前者更加简化,是一类全连接的前馈神经网络. 在本研究数据集的学习中,二者均通过 ReLU 激活函数、Adam 自适应算法进行训练;但就隐藏层‘hidden_size’的规模而言,NN 优化了 64 层,而 MLP classifier 则只优化了 1 层;基于本数据集的特征推测,NN 模型可能因过于复杂而无法实现较准确的预测.

对 3 种基础线性模型(logistic regression、ridge classifier、PAC)和 2 种支持向量机分类器(SVM、linear SVC)而言,正则化是降低过拟合风险、提高泛化能力的必要手段. 本研究中,logistic regression、PAC、SVM 所设置的正则化超参数‘C’,以及 ridge

classifier 所筛选的正则化超参数‘alpha’,均采取了模型默认值,未进行调整以赋予模型更大惩罚项,从而降低欠拟合风险,印证了本数据集的小样本、高稀疏性特征.

对判别分析模型而言,模型优化过程涉及特征协方差矩阵的计算. QDA 模型中的参数‘reg_param’、LDA 模型中参数‘shrinkage’即被用来规范协方差矩阵和提升其预测性能.

对基于朴素贝叶斯假设的模型而言,Gaussian NB 的超参数优化相对简易,在本研究中超参数‘var_smoothing’以默认值代入模型即可满足计算稳定性的要求;而 Bernoulli NB 模型还要考察平滑因子‘alpha’、二值化特征阈值‘binarize’及‘fit_prior’所代表的是否需要学习类先验概率等问题,在模型训练中可能占据较多时间.

GPC 模型和 lable propagation 模型的优化过程中,超参数‘kernel’的选取至关重要. 针对本数据集,该两类模型均选用普适性较强的 RBF 内核,其可以处理特征参数与类别标签非线性相关的样例,且可以降低数值困难发生的风险.

支撑材料 S3 显示了 21 种 ML 模型的超参数优化的具体结果,反映出模型构建过程的正当性和有效性.

2.3 模型预测性能评价

经超参数优化后,利用 7 种指标多维度评价 21 种 ML 模型的预测性能. 为在各模型间横向比较,寻找性能最优算法,绘制堆叠柱状图展示 21 种算法的预测性能. 如图 2 所示,在所有模型中,3 种算法的准确率和召回率超过 75%;4 种算法的精度超过 80%;3 种算法的 F1 分数超过 75%;4 种算法的 CKC 超过 0.6;5 种算法的 MCC 超过 0.6. 由于本数据集存在样本量小、特征维度高、分布稀疏的固有缺陷,在避免过拟合、保证鲁棒性的前提下,模型在上述指标中无法达到过高的性能是可以接受的. 而作为一个综合多阈值的评价指标,ROC_AUC 在数值上更宽容,17 种模型的 ROC_AUC 大于 0.8,侧面印证了本研究所构建的 IDTE 数据集标签可靠,类别相对平衡.

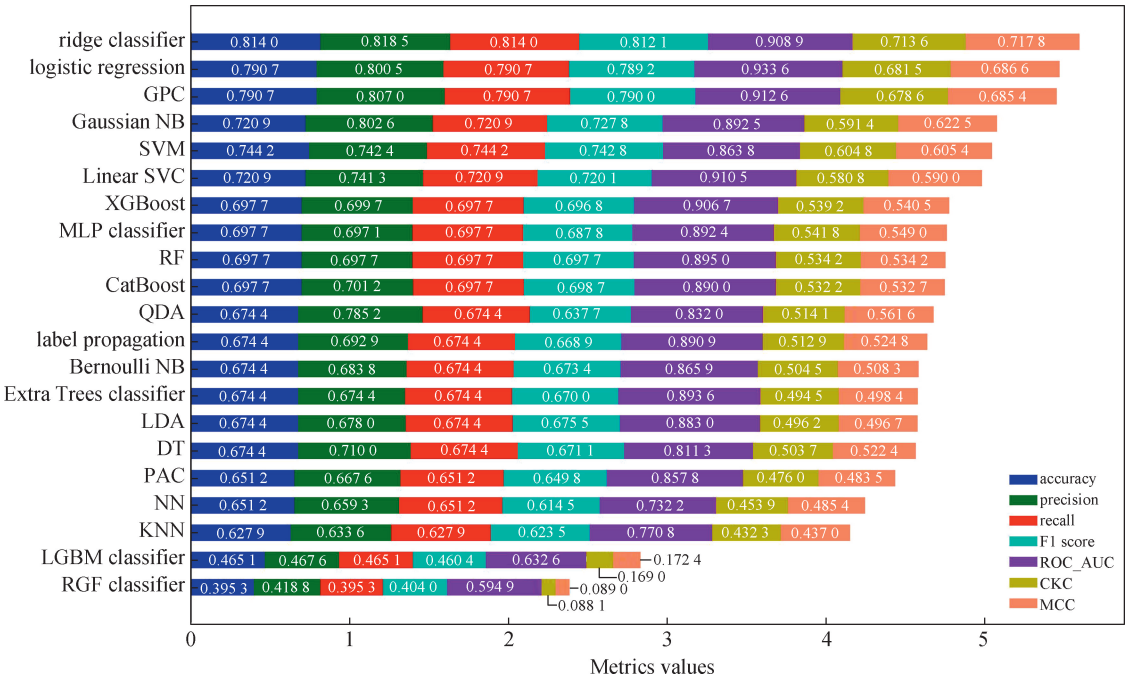


图 2 21 种 ML 模型预测性能评价

Fig. 2 Evaluation of prediction performance of 21 ML models

对于综合求和指标而言,由于 7 种独立指标的最大值均为 1,理想模型的评价指标和应尽可能接近于 7. 对 21 种算法的比较分析表明,ridge classifier 模型具有最优的综合评价,且在准确率、精度、召回率、F1 分数、CKC 和 MCC 6 项指标上单项最佳. 图 3 为该模型在本研究数据集上的混淆矩阵,可见该模型参与下,实际为 Class 0 的特征组合被准确预测的概率可达 93%,而实际为 Class 2 的特征组合被准确预测的概率为 78%. 这反映出基于本研究数据集的 ridge classifier 模型对于实验室制备前排除不具有工程意义的 IDTE 潜在参数组合,优选促进细胞增殖、迁移、黏附、分化、分泌等行为的支架具有重要意义. 该模型

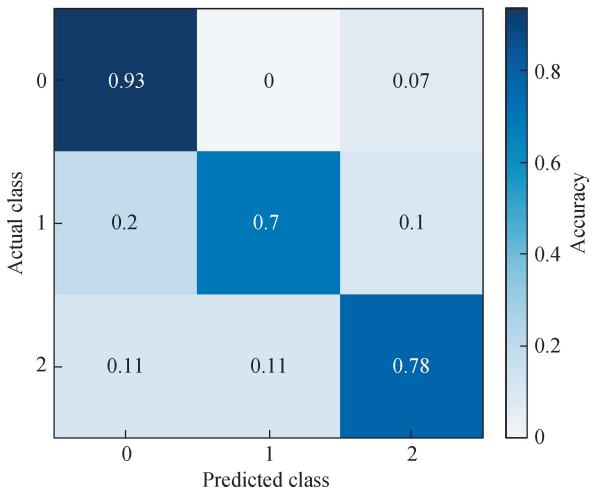


图 3　ridge classifier 模型的混淆矩阵

Fig. 3　Confusion matrix of the ridge classifier model

参数信息详见支撑材料 S4.

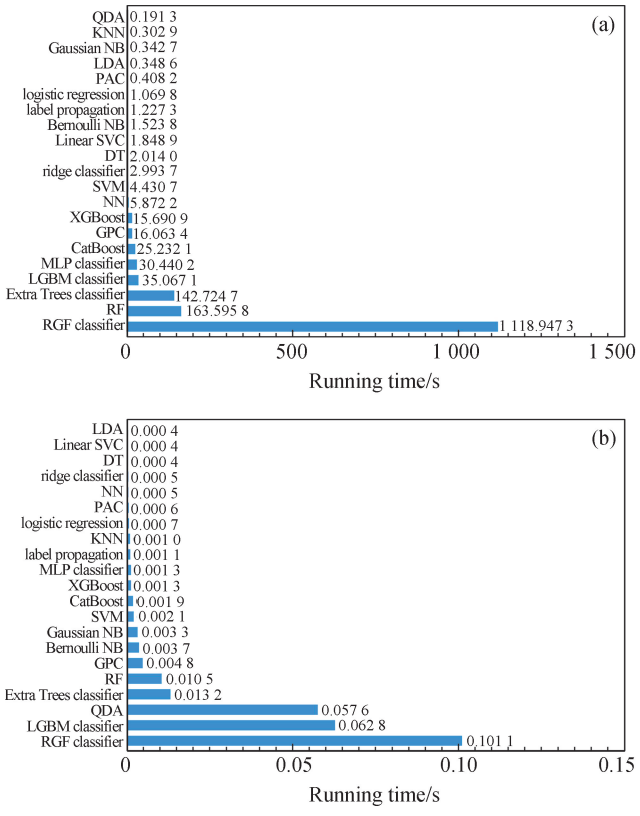
2.4　模型运行时间评价

除预测能力外,运行时间也是 ML 模型的一项重要评价指标,为此记录了本数据集上每个模型的训练及预测时间(图 4). 训练时间为超参数优化及训练集学习所耗时间总和,在所有模型中,训练时间在 1 s 以内的有 5 个,1 s～1 min 范围的有 13 个,1 min 以上的有 3 个;筛选出的预测性能最佳的 ridge classifier 模型具有较短的训练时间,在计算成本上存在优势[图 4(a)]. 预测时间是调整好的模型预测未知类别所需要的时间,对该参数的控制更具有实践价值. 理想的 ML 模型应具备尽可能少的预测时间,从而对用户输入快速响应. 21 种 ML 模型均具有低于 0.11 s 的预测时间,其中 ridge classifier 的表现也较突出,表明利用该模型指导 IDTE 支架设计具有较大可行性[图 4(b)].

综合预测性能和运行时间评价,基于 ridge classifier 算法构建的分类模型可以对特定组分支架的生物学效应进行较准确的预测,且计算成本低,预测速度快,可进一步应用于指导 IDTE 体系的开发.

2.5　模型可解释性分析

为扩展 ML 模型的应用场景,提升预测模型对 IDTE 体系开发的指导作用,对优选出的 ridge classifier 模型进行可解释性分析,解读该模型描述下各特征对细胞行为的影响. 通过 SHAP 模型计算各特征参数在每个样本上 SHAP 绝对值的平均值,该值越大表明对应的特征在 ridge classifier 模型中越重要,即该特征对结局指标的贡献越大. 图 5(a)依降序展示了重要性位次前 20 的特征. 该图直观表明,对 IDTE



(a)训练时间;(b)预测时间.

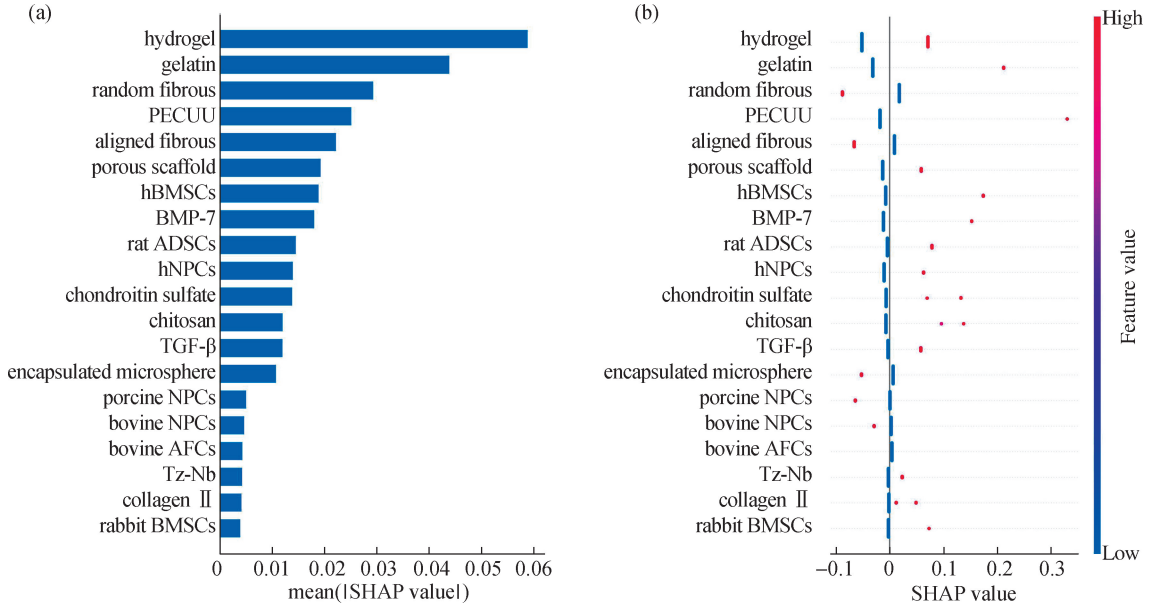
图 4　21 种 ML 模型的运行时间

Fig. 4　Running time of 21 ML models

支架上细胞行为影响最大的特征是 hydrogel. 而从特征类别的角度分析,5 种支架形式特征均有较大贡献,表明设计 IDTE 支架时需着重考虑支架形式,支架主材中 gelatin、PECUU、硫酸软骨素(chondroitin sulfate)和壳聚糖(chitosan)的贡献较大;活性因子中,BMP-7 与 TGF-β 贡献较大,该结果印证了上述材料及生物因子在已有研究中的高应用频率;5 种成型辅料对结局指标的贡献均不高,表明 IDTE 支架的制备工艺对材料的细胞学作用影响较小,提示研究人员在支架开发过程中可灵活选用合成方案;细胞种类方面,人骨髓间充质干细胞(hBMSCs)、大鼠脂肪源性干细胞(rat ADSCs)和人核心髓核细胞(hNPCs)对结局指标的贡献较大,表明这些特定种类的细胞对支架刺激的反应较明显,有较大潜力应用于 IDTE 细胞治疗.

图 5(b)是基于 SHAP 模型绘制的特征散点图,能直观反映各特征对结局指标的作用方向. 支架形式上,hydrogel 和多孔支架(porous scaffold)对结局指标的贡献是正向的,而其余 3 种支架类型对结局指标的贡献是负向的,因此研究人员在进行靶向细胞行为的支架设计时可优先考虑水凝胶及多孔形式;而对于

4 种常用的支架主材 gelatin、PECUU、chondroitin sulfate 和 chitosan,以及 2 种常用的活性因子 BMP-7 与 TGF-β,其作用方向均为正向,表明在支架合成过程中,上述主材质量比例的提高有助于增强支架对细胞行为的促进作用;在细胞种类方面,hBMSCs、rat ADSCs、hNPCs 等对结局指标的贡献为正向,而 porcine NPCs、bovine NPCs 等则表现为负向相关结局指标,提示研究人员在评价 IDTE 支架的生物相容性时,选取 hBMSCs 等正向作用细胞能更好地显示出材料组分的作用.



(a)SHAP 特征重要性排序;(b)SHAP 特征密度散点图.

图 5 ridge classifier 模型的可解释性分析

Fig. 5 Interpretability analysis of the ridge classifier model

3 讨 论

传统意义上,组织工程支架的制造具有偶然性和经验性的特点[15]. 早期的支架设计与优化往往需要研究人员进行大量的文献查阅,从而归纳出可能的改进方向. 从支架基材的构成、交联剂的使用、活性分子的添加到支架机械、化学及生物学性能的控制,大量特征变量在组织工程支架的合成中发挥作用. 因此,为找到合适的参数组合,研究人员在实验中常常需要大量试错. 从"成本-效益"的角度考虑,这种设计和制造支架的方法十分不合理,目前已经成为组织工程学科进一步发展的阻碍[10]. ML 的引入是解决工程支架设计困难的希望. 作为计算科学与统计学的拓展,ML 模型能找出既有支架工艺数据之间的内在联系,对未知参数组合下的支架性质给出合理预测,从而节约大量的经济与时间成本,减轻实验人员负担. 因此,有必要对既往积累的组织工程支架研究数据进行系统汇总,为构建 ML 指导下的预测模型奠定坚实基础.

近年来,IVDD 相关的组织工程治疗研究蓬勃发展. 已有研究利用不同的支架制备参数设计了大量 IDTE 支架. 其中部分支架具有优异的生物学效应,对于髓核细胞、纤维环细胞、间充质干细胞等椎间盘局部修复相关组分具有显著的正向作用;也有部分支架效应微弱,或难以成型应用. 简单总结文献中的成败经验难以高效地进行新型 IDTE 支架的开发;而 ML 的引入有望阐明支架制备参数和其引发的生物学效应之间的关系,从而改变 IDTE 支架开发的现状,经济、快捷、准确地制备出新型 IDTE 支架. 本研究成功建立了 IDTE 支架上的细胞行为数据集,并使用 21 种 ML 算法进行模型训练. 对于不同属性的数据集,各类 ML 算法可能表现出差异性的结果,因此通过多种指标测试了模型性能,从中挑选出综合性能最佳的模型. 在本研究的数据集上,ridge classifier 模型预测较准确,且训练时间及预测时间短,具有最优的综合评价. 通过 SHAP 算法对 ridge classifier 模型进行解释,得到构建 IDTE 体系的特征参数在细胞行为方面的作用方向与程度,可为基于预测模型的 IDTE 支架研发提供参考.

在建立 IDTE 预测模型的过程中仍有部分细节值得进一步探讨说明:

1) 在模型结局指标的选取方面,尽管数据集中纳入的文献多具有动物模型体内实验的 IVDD 相关评价,且这些评价相较于体外细胞学指标更能真实地反映椎间盘退变的进展,然而为规避不同的动物遗传、周龄、饲养条件及支架移植差异可能造成的结果偏差,本研究最终选用不同参数条件下的细胞学效应作为分类标签的依据. 在此基础上,各样本生物学效应的评估是相对客观且标准的,该体系下训练出的模型对于 IVDD 相关细胞行为具有预测价值,可间接地反映特定参数的 IDTE 支架对于 IVDD 的治疗能力.

2) 在特征参数的选取方面,本研究更多采用材料类别、工艺形式等间接要素,而非更直接的支架理化性质,这可能与研究者的科学直觉相矛盾. 理论上,选取机械强度、导电率、孔隙率、表面形貌等精确量化的理化因素更有利于建立支架性质与生物学效应间的联系;然而,本研究所使用的 5 类特征变量在实际应用中更具可操作性. 一方面,由于多数文献并未对支架的理化参数进行详细报道,使用这些特征变量容易产生大量缺失数据,不利于小样本稀疏数据集下的模型训练;另一方面,对于新支架的开发而言,基于制备组分和工艺形式的预测模型更实用,可以使研究人员从宏观可及的材料参数出发获取支架性能的预测,从而在设计和制造过程中快速排除表现不佳的特征组合,节省大量试错成本.

3) 在模型缺陷和优化方面,本研究所使用的数据集并不能涵盖 IDTE 领域所有的支架组分. 作为监督式学习模型的固有缺陷,本研究得出的 ML 模型无法预测数据集中未出现的材料或细胞制成的支架上的生物效应. 为应对这一问题,保证预测模型的活力,未来相关研究可能聚焦于两个方向:一方面随着 IDTE 相关原始研究的不断累积,势必有更多的特征变量可供加入本研究框架下的 ML 模型;新特征的不断涌现将伴随着样本扩充、数据挖掘、特征选择和标签赋值的进一步合理化,为 IDTE 预测模型的更新和完善提供源源不断的动力. 另一方面,随着分子生物学的不断发展,近年来组织工程支架的设计也逐渐倾向于为负载活性组分的靶向功能提供辅助,这提示研究人员构建更精准、小范围适用的的 ML 模型;在数据集的过程中,指向性筛除次要特征,定量化划分分类标签,从而提供更实用的 IDTE 生物学效应预测模型.

4) 在研究思路推广方面,虽然本研究关注的医学问题是 IVDD 的组织工程治疗,但是方法学部分所体现的文献检索、数据提取、特征选择、数据集构建、ML 训练与预测、模型评价的研究路线是对组织工程各亚领域普遍适用的. 在文献信息充足的基础上,进一步发掘人体各组织、器官的工程支架在特定参数下的生物学效应,以辅助组织工程研究人员进行经济而高效的支架开发,可行且值得期待.

4　结　论

本研究建立了一个基于 ridge classifier 算法的 ML 预测模型,能通过输入支架主材、成型辅料、活性因子、支架形式及细胞种类 5 类特征参数预测特定工艺条件下 IDTE 支架对 IVDD 相关细胞行为的影响. 进一步的可解释性分析识别出各支架特征对细胞行为作用的大小和方向,可用于指导基于 ridge classifier 模型的 IDTE 支架开发.

参考文献:

[1] KIRNAZ S, CAPADONA C, WONG T, et al. Fundamentals of intervertebral disc degeneration[J]. World Neurosurgery, 2022, 157: 264-273.

[2] OUYANG D C, KLECK C J, ACKERT-BICKNELL C L. Genetics of intervertebral disc degeneration[J]. Current Osteoporosis Reports, 2023, 21(1): 56-64.

[3] RISBUD M V, SHAPIRO I M. Role of cytokines in intervertebral disc degeneration: pain and disc content[J]. Nature Reviews Rheumatology, 2014, 10(1): 44-56.

[4] SAMANTA A, LUFKIN T, KRAUS P. Intervertebral disc degeneration-current therapeutic options and challenges[J]. Frontiers in Public Health, 2023, 11: 1156749.

[5] XIN J G, WANG Y J, ZHENG Z, et al. Treatment of intervertebral disc degeneration[J]. Orthopaedic Surgery, 2022, 14(7): 1271-1280.

[6] ISA I L M, MOKHTAR S A, ABBAH S A, et al. Intervertebral disc degeneration: biomaterials and tissue engineering strategies toward precision medicine[J]. Advanced Healthcare Materials, 2022, 11(13): e2102530.

[7] CHOI Y, PARK M H, LEE K. Tissue engineering strategies for intervertebral disc treatment using functional polymers[J]. Polymers, 2019, 11(5): 872.

[8] BOWLES R D, SETTON L A. Biomaterials for intervertebral disc regeneration and repair[J]. Biomaterials, 2017, 129: 54-67.

[9] YOSHIDA M, TURNER P R, CABRAL J D. Intervertebral disc tissue engineering using additive manufacturing[J]. Gels, 2022, 9(1): 25.

[10] GUO J L, JANUSZYK M, LONGAKER M T. Machine learning in tissue engineering[J]. Tissue Engineering Part A, 2023, 29(1/2): 2-19.

[11] DEO R C. Machine learning in medicine will this time be different? [J]. Circulation, 2020, 142(16): 1521-1523.

[12] CONEV A, LITSA E E, PEREZ M R, et al. Machine learning-guided three-dimensional printing of tissue engineering scaffolds[J]. Tissue Engineering Part A, 2020, 26(23/24): 1359-1368.

[13] MENON A, THOMPSON-COLÓN J A, WASHBURN N R. Hierarchical machine learning model for mechanical property predictions of polyurethane elastomers from small datasets[J]. Frontiers in Materials, 2019, 6: 87.

[14] ECHEZARRETA-LÓPEZ M M, LANDIN M. Using machine learning for improving knowledge on antibacterial effect of bioactive glass[J]. International Journal of Pharmaceutics, 2013, 453(2): 641-647.

[15] RAFIEYAN S, VASHEGHANI-FARAHANI E, BAHEIRAEI N, et al. MLATE: machine learning for predicting cell behavior on cardiac tissue engineering scaffolds[J]. Computers in Biology and Medicine, 2023, 158: 106804.

[16] ROZOS E, KOUTSOYIANNIS D, MONTANARI A. KNN vs. Bluecat: machine learning vs. classical statistics [J]. Hydrology, 2022, 9(6): 101.

[17] AHN J M, KIM J, KIM K. Ensemble machine learning of gradient boosting (XGBoost, LightGBM, CatBoost) and attention-based CNN-LSTM for harmful algal blooms forecasting[J]. Toxins, 2023, 15(10): 608.

[18] DOUBLEDAY K, ZHOU J, ZHOU H, et al. Risk controlled decision trees and random forests for precision medicine [J]. Statistics in Medicine, 2022, 41(4): 719-735.

[19] TALEKAR B, AGRAWAL S. A detailed review on decision tree and random forest[J]. Bioscience Biotechnology Research Communications, 2020, 13(14): 245-248.

[20] PENG L H, YUAN R Y, SHEN L, et al. LPI-EnEDT: an ensemble framework with extra tree and decision tree classifiers for imbalanced lncRNA-protein interaction data classification[J]. BioData Mining, 2021, 14(1): 50.

[21] JOHNSON R, ZHANG T. Learning nonlinear functions using regularized greedy forest[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36 (5): 942-954.

[22] YAN P Y, LIU Y J, JIA Y R, et al. Deep learning and machine learning applications in biomedicine[J]. Applied Sciences, 2024, 14(1): 307.

[23] CHAUHAN V K, DAHIYA K, SHARMA A. Problem formulations and solvers in linear SVM: a review[J]. Artificial Intelligence Review, 2019, 52(2): 803-855.

[24] BHATTACHARYYA S, KHASNOBISH A, CHATTERJEE S, et al. Performance analysis of LDA, QDA and KNN algorithms in left-right limb movement classification from EEG data[C]// Proceedings of the 2010 International Conference on Systems in Medicine and Biology. Piscataway: IEEE, 2010: 126-131.

[25] DHAMNETIYA D, GOEL M K, JHA R P, et al. How to perform discriminant analysis in medical research? Explained with illustrations[J]. Journal of Laboratory Physicians, 2022, 14(4): 511-520.

[26] YIN M M, LIU J X, GAO Y L, et al. NCPLP: a novel approach for predicting microbe-associated diseases with network consistency projection and label propagation [J]. IEEE Transactions on Cybernetics, 2022, 52(6): 5079-5087.

[27] CHALLIS E, HURLEY P, SERRA L, et al. Gaussian process classification of Alzheimer's disease and mild cognitive impairment from resting-state fMRI [J]. NeuroImage, 2015, 112: 232-243.

[28] PENG C, CHENG Q. Discriminative ridge machine: a classifier for high-dimensional data or imbalanced data[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(6): 2595-2609.

[29] YILDIRIM P. Chronic kidney disease prediction on imbalanced data by multilayer perceptron: chronic kidney disease prediction[C]// Proceedings of the 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC). Los Alamitos: IEEE Computer Society, 2017: 193-198.

[30] HEMACHANDRAN K, SHUBHAM T, PREETHA M G, et al. Bayesian reasoning and gaussian processes for machine learning applications[M]. Calabas: CRC Press, 2022: 1-14.

[31] KUMARASINGHE P, SURESH S, SUBBARAJU V. MiPAL: multiple-instance passive aggressive learning for identification of attention deficit hyperactive disorder from fMRI[C]// Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN). Piscataway: IEEE, 2017: 4541-4548.

[32] MAXWELL A E, WARNER T A, GUILLÉN L A. Accuracy assessment in convolutional neural network-based deep learning remote sensing studies. Part 1: literature review[J]. Remote Sensing, 2021, 13(13): 2450-2450.

(责任编辑:徐婷婷)