

[专题研究]

孤独症儿童的肠道菌群特征分析和筛查模型构建

庞许颖¹, 张强², 王玥珠³, 赵红洋⁴, 郑华军³, 蒋泓¹

(1. 复旦大学公共卫生学院妇幼与儿少卫生教研室, 国家卫生健康委员会卫生技术评估重点实验室, 上海 200032;
2. 遵义医科大学附属医院妇产科, 贵州 遵义 563000; 3. 上海市生物医药技术研究院, 上海市疾病与健康基因组学重点实验室, 国家卫生健康委生育调节药械重点实验室, 上海 200237;
4. 山东第一医科大学附属中心医院儿科, 山东 济南 250013)

[摘要]目的 探讨孤独症谱系障碍(ASD)患儿和健康儿童之间肠道菌群组成的差异, 并使用机器学习算法构建疾病筛查模型, 提供基于肠道菌群生物标志物的非侵入性孤独症筛查手段。方法 本研究于 2019 年 12 月至 2023 年 4 月, 在济南市、遵义市、香港特别行政区以及上海市招募 149 名 2.5~4.5 岁孤独症儿童为孤独症组, 按年龄、性别 1:1 匹配的 149 名健康儿童为对照组, 采集粪便样本, 通过 16S rRNA 基因 V3-V4 区测序收集两组儿童肠道菌群相关指标。在属水平, 使用随机森林、支持向量机、K 近邻算法、朴素贝叶斯分类器 4 种机器学习算法在模型开发数据集中构建孤独症分类模型, 识别最具判别性的菌属组合, 并评估模型在两个独立外部测试数据集中的泛化能力。结果 ①孤独症组的菌群多样性显著高于对照组(Chao 指数=118.00、105.00, Shannon 指数=3.46、3.00, $P=0.023, 0.001$)。②孤独症儿童和对照儿童肠道菌群结构存在显著差异($F=5.198, R^2=0.052, P<0.001$)。③共筛选出 14 个特征菌属。其中孤独症组中丰度较高的菌属为 *Phocaeicola*、*Anaerobutyricum*、*Faecalibacterium*、*Blautia*、*Oscillibacter*、*Lachnospira*、*Parabacteroides*、*Flintibacter* 和 *Anthropogastromicrobium*, 对照组中丰度较高的菌属为 *Ruthenibacterium*、*Flavonifractor*、*Bifidobacterium*、*Anaerostipes* 和 *Eisenbergiella*。④基于 14 个菌属组合的随机森林模型在模型开发数据上具有最优分类性能, 训练集中曲线下面积(AUC)为 100%(95%CI: 100%~100%), 验证集中 AUC 为 93.94%(95%CI: 88.13%~99.74%)。在两个独立的外部测试集中, 朴素贝叶斯模型则展现出最佳的泛化性能, AUC 分别为 63.83%(95%CI: 51.99%~75.67%)和 60.19%(95%CI: 47.83%~72.55%)。结论 孤独症和健康儿童肠道微生物群落存在显著差异, 且特定肠道生物标志物对孤独症疾病状态具有分类能力, 提示肠道微生物具有作为儿童早期孤独症无创筛查标志物的潜在作用。

[关键词] 肠道菌群; 孤独症; 儿童; 生物标志物; 机器学习; 16S rRNA

Doi: 10.3969/j.issn.1673-5293.2024.07.001

[中图分类号] R174; R179

[文献标识码] A

[文章编号] 1673-5293(2024)07-0001-11

Characteristic analysis of gut microbiota and screening model construction in children with autism spectrum disorder

PANG Xuying¹, ZHANG Qiang², WANG Yuezhu³, ZHAO Hongyang⁴, ZHENG Huajun³, JIANG Hong¹

(1. Department of Maternal and Child Health, School of Public Health, Fudan University, NHC Key Laboratory of Health Technology Assessment, Shanghai 200032, China; 2. Department of Obstetrics and Gynecology, Affiliated Hospital of Zunyi Medical University, Guizhou Zunyi 563000, China; 3. Shanghai Institute for Biomedical and Pharmaceutical Technologies, Shanghai-MOST Key Laboratory of Health and Disease Genomics, NHC Key Lab of Reproduction Regulation, Shanghai 200237, China; 4. Department of Pediatrics, Central Hospital Affiliated to Shandong First Medical University, Shandong Jinan 250013, China)

[Abstract] **Objective** To investigate the differences of gut microbiota composition between children with autism spectrum disorder (ASD) and health children, and to construct a disease screening model using machine learning algorithm to provide a non-invasive method for autism screening based on biomarkers of gut microbiota. **Methods** From December 2019 to April 2023, this study recruited 149 ASD children aged 2.5 to 4.5 years from Jinan, Zunyi, Hong Kong and Shanghai, as the autism group. Additionally, 149 healthy children matched 1:1 by age and gender were recruited as the control group. Fecal samples were collected, and gut

[收稿日期] 2024-05-09

[基金项目] 国家自然科学基金(82181220077, 8237120466); 复旦大学公共卫生学院嘉定区卫生健康委公共卫生高质量发展重点学科、重点专项(GWGZLXK-2023-04); 上海市加强公共卫生体系建设三年行动计划(2023—2025 年)重点学科(GWVI-11.1-32)

[作者简介] 庞许颖(1999—), 女, 在读硕士研究生, 主要从事妇幼与社区保健研究。

[通讯作者] 蒋泓, 教授; 郑华军, 研究员。

microbiota-related indices were gathered through 16S rRNA gene V3-V4 region sequencing for both groups. At the genus level, four machine learning algorithms, random forest, support vector machine, K-nearest neighbors, and naive bayes classifier, were used to construct an autism classification model in the model development dataset, identifying the most discriminative bacterial genus combinations, and the generalization ability of the models was evaluated in two independent external test datasets. **Results** ①The gut microbiota diversity of the autism group was significantly higher than that in the control group (Chao index = 118.00, 105.00; Shannon index = 3.46, 3.00; $P = 0.023, 0.001$). ②There were significant differences in gut microbiota structure between autism children and control children ($F = 5.198, R^2 = 0.052, P < 0.001$). ③A total of 14 characteristic genera were identified. The genera with higher abundance in the autism group were *Phocaeicola*, *Anaerobutyricum*, *Faecalibacterium*, *Blautia*, *Oscillibacter*, *Lachnospira*, *Parabacteroides*, *Flintibacter*, and *Anthropogastromicrobium*, and the genera with higher abundance in the control group were *Ruthenibacterium*, *Flavonifractor*, *Bifidobacterium*, *Anaerostipes*, and *Eisenbergiella*. ④The random forest model based on the combination of 14 genera showed the best classification performance in the model development dataset, with the training set AUC of 100% (95% CI: 100%-100%) and validation set AUC of 93.94% (95% CI: 88.13%-99.74%). In two independent external test datasets, the Naive Bayes model showed the best generalization performance, with AUC of 63.83% (95% CI: 51.99%-75.67%) and 60.19% (95% CI: 47.83%-72.55%), respectively. **Conclusion** There are significant differences in the gut microbiota communities between autism children and control children, and specific gut microbiota biomarkers have the capability to classify autism disease states, suggesting that gut microbiota has potential significance as a non-invasive screening biomarker for early autism detection in children.

[Key words] gut microbiota; autism; children; biomarkers; machine learning; 16S rRNA

孤独症谱系障碍 (autism spectrum disorder, ASD), 以下简称孤独症或 ASD, 是一种复杂的发生在生命早期的神经发育障碍性疾病, 以社交障碍、兴趣狭隘、刻板行为及不寻常的感官反应等为特征^[1]。美国疾病控制与预防中心 (Centers for Disease Control and Prevention, CDC) 的调查数据显示, 2020 年 8 岁儿童 ASD 患病率为 2.76%, 较 2018 年增长近 22%^[2]。而中国 ASD 发病率从 2000 年的 2.80/10 000 增长至 2015 年的 63/10 000^[3]。而当前孤独症尚无标准化且严格的诊断标准, 缺乏明确的生物标志物来辅助诊断^[4]。既往研究表明, 孤独症儿童相较于对照组, 其肠道菌群生物量、丰富度和多样性存在显著差异^[5]。为了进一步探索孤独症生物标志物, 目前已有研究利用机器学习算法中的随机森林模型进行探索, 然而该领域仍处于不断发展与完善的阶段^[6]。因此, 本研究旨在分析孤独症与健康儿童肠道菌群的特征差异, 并借助四种机器学习算法, 构建孤独症的早期筛查模型, 以在生命早期及时识别孤独症, 为无创生物筛查工具提供参考。

1 材料与方法

1.1 研究对象选择

本研究于 2019 年 12 月至 2023 年 4 月共纳入 298 例年龄 2.5~4.5 岁的受试者进行筛查模型开发

构建。孤独症组共 149 例, 其中 101 例于山东省济南市中心医院儿保科就诊 ASD 患儿中招募, 48 例于贵州省遵义市孤独症康复中心的孤独症患儿中招募。按性别和年龄 1:1 匹配的健康儿童, 即神经典型 (neurotypical, NT) 发育儿童共 149 例为对照组, 其中 117 名来自上海市闵行区出生队列, 另有 32 名儿童来自香港特别行政区 (原始数据下载自 NCBI, BioProject ID: 912769, 网站: <https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA912769>)。孤独症的诊断由儿科医生根据第五版《精神障碍诊断与统计手册》的标准进行^[3]。纳入标准: 参加研究前至少 1 个月没有服用抗生素、益生菌或其他可能影响肠道微生物群的药物。排除标准: 在过去一年中有精神分裂症、其他精神疾病、情绪障碍和癫痫发作的患者。

选择两个独立的外部测试队列来评估模型的泛化能力。一个队列包括来自韩国的 85 名受试者 (38 名对照, 47 名孤独症), 而另一个队列包括来自厄瓜多尔的 50 名受试者 (27 名对照, 23 名孤独症) 和来自意大利的 23 名受试者 (13 名对照, 10 名孤独症)。外部测试数据从 NCBI 下载 (BioProject ID: PRJEB45948、PRJEB27306、PRJEB29421, 网站: <https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJEB45948>, <https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJEB27306>, <https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJEB29421>)。

= PRJEB27306、<https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJEB29421>。

本研究的实施方案已通过上海市生物医药技术研究院伦理委员会的审查,并获得批准(PJ2019-17)。在充分详尽地解释研究内容后,所有参与的研究对象及其监护人均自愿签署知情同意书。

1.2 粪便样本采集

将带有保护液的分体式粪便采集器(江苏一米科技,中国)交予对象及其监护人,保证采样当天身体状态良好,与平时相比无特殊不适症状。采集的样本在回收后,立即存放于 -80°C 的超低温冰箱中。

1.3 肠道菌群测定

采用 QIAamp PowerFecal Pro DNA Kit 粪便抽提试剂盒(QIAGEN,德国)对样本的基因组 DNA 进行提取。取稀释后的 DNA $1\mu\text{L}$,使用 TransStart Fastpfu DNA 聚合酶(北京全式金生物,中国),和带有不同 Barcode(样品标签)的特异性引物 338F($5'-\text{ACTCTACGGGAGGCAGCAG}-3'$)和 806R($5'-\text{GGACTACHVGGGTWTCTAAT}-3'$),对样本的 16S rRNA 基因的 V3-V4 区进行聚合酶链式反应(polymerase chain reaction,PCR)扩增。PCR 产物用 QIAquick PCR 纯化试剂盒(QIAGEN,德国)进行纯化,然后 Qubit 定量,所有样本的 PCR 产物等量混合。用 TransNGS[®] DNA Library Prep Kit for Illumina(北京全式金生物,中国)对 PCR 产物进行文库构建,构建好的文库经定量和文库检测合格后,采用 Illumina MiSeq 测序仪(Illumina,美国)进行 $2\times 300\text{bp}$ 的双端测序。

1.4 生物信息学分析

1.4.1 上游分析

测序完成后,在去除样本中 Barcode 特异性引物的基础上,对得到的原始序列数据执行生物信息学的上游分析。首先对双端测序得到的成对末端(paired end, PE)序列数进行拼接。使用 Quantitative Insights Into Microbial Ecology 2(QIIME,美国)软件处理,通过插件 dada2 对序列进行过滤和剪切处理,移除上下游引物,并去除冗余的序列,保留一组唯一的代表序列集合,构建参数误差模型来推测真实的序列并进行去噪,从而构建扩增子序列变异体(amplicon sequence variant, ASV)表格。接着利用

软件 RDP classifier 确定 ASV 序列的分类信息(门、纲、目、科、属),阈值设定为 0.8,菌种注释采用生物大分子序列比对搜索工具(basic local alignment search tool, BLAST)中的 BLASTn 与 SILVA 参考数据库(网站:<https://www.arb-silva.de/>,版本 138.1),和美国国家生物技术信息中心(National Center for Biotechnology Information, NCBI)数据库进行比对,置信阈值设定:identity $>97\%$ 、alignment $>97\%$,以具有最高匹配分值的菌种作为该代表序列的物种名称。

1.4.2 下游分析

在 R 4.2.2 中主要使用了“amplicon”和“vegan”包来进行生物信息学下游分析,包括 α 多样性指标(衡量单个样本内部的多样性水平)和 β 多样性指标(衡量组间的微生物群落结构差异)。本研究主要采用的 α 多样性指标为 Chao 指数(估计群落中 ASV 的数目,值越大表示丰富度越高)和 Shannon 指数(同时考虑丰富度和均匀度,值越大表示多样性越高)。两组间 α 多样性指数的统计学差异采用 Mann-Whitney U 检验。采用基于 Bray-Curtis 距离矩阵的主坐标分析(principal coordinates analysis, PCoA)可视化肠道菌群组成差异,并使用置换多元方差分析(permutational multivariate analysis of variance, PerMANOVA)确定孤独症组与对照组细菌结构差异的显著性。在置换 999 次的基础上,进行组间差异的显著性分析, $P<0.05$ 表示差异有统计学意义。

1.5 基于机器学习构建筛查模型

1.5.1 特征选择

使用机器学习构建筛查模型之前,有必要采用多种特征选择技术的组合来筛选出有重要贡献的菌属。第一种方法涉及使用 STAMP 软件基于错误发现率(false discovery rate, FDR)的 Welch's 双侧 t 检验分析, $\text{FDR}<0.05$ 的菌属具有统计学显著差异。然后利用主成分分析(principal component analysis, PCA)筛选出方差贡献率(每个主成分占整个数据集方差的比例)超过 1% 的菌属。最后,采用线性判别分析(linear discriminant analysis, LDA)在线工具(网站:<http://huttenhower.sph.harvard.edu/galaxy/>)

筛选组间有差异的物种。该方法首先应用 Kruskal-Wallis H 检验筛选显著差异物种,再通过 Wilcoxon 检验组间差异一致性,最后用 LDA 评估物种丰度对差异的贡献。本研究中设置的 $\text{Log}(\text{LDA})$ 的阈值为 3,超过该阈值即认定该菌在两组之间存在显著差异。采用 FDR 对 P 值进行调整,以校正后的 $q < 0.05$ 作为相对丰度在两组间差异有统计学意义的判断标准。将上述组合方法确定的属纳入模型。

为了识别特征噪声,避免模型过拟合,反复调整特征子集,使模型性能最大化。使用 R 4.2.2 中的“Boruta”包创建随机阴影变量,并反复比较特征变量的重要性得分即 Z -score,来识别重要变量。将每个菌属的 Z -score 与最大阴影属性(maximum Z -score among shadow attributes, MZSA)的 Z -score 进行比较,认为各特征菌属的 Z -score 显著高于 MZSA 的属为重要属。此外,采用随机森林交叉验证方法进行递归特征消除,识别出模型错误率最低的最优变量组合,并得到各特征的重要度排序,利用最后筛选出的差异特征构建模型。

1.5.2 筛查模型构建

在模型开发数据集中,298 例样本(孤独症为 149 例,对照为 149 例)被随机划分为 80% 的训练集和 20% 的验证集。在训练集中采用内部 5 次十折交叉验证法(fold=10),即将训练数据分成 10 个子集,每次使用其中 1 个子集作为验证集,其余作为训练集,重复 5 次以获得稳定的性能评估结果,最终取平均性能作为评估结果,进行超参数选择,并对留存的验证集进行模型评价,从而构建最优模型。将构建好的最优模型在外部测试数据集中进行测试,以评估模型泛化能力。

使用 R 4.2.2 中的“randomForest”和“e1071”包,使用随机森林(random forest, RF)、支持向量机(support vector machine, SVM)、K 近邻算法(K-nearest neighbor, KNN)和朴素贝叶斯分类器(naive bayes classifier, NBC)四种机器学习筛查模型来区分孤独症与对照,并明确特定的生物标志物。完整的机器学习筛查模型构建和外部测试流程(见图 1)。

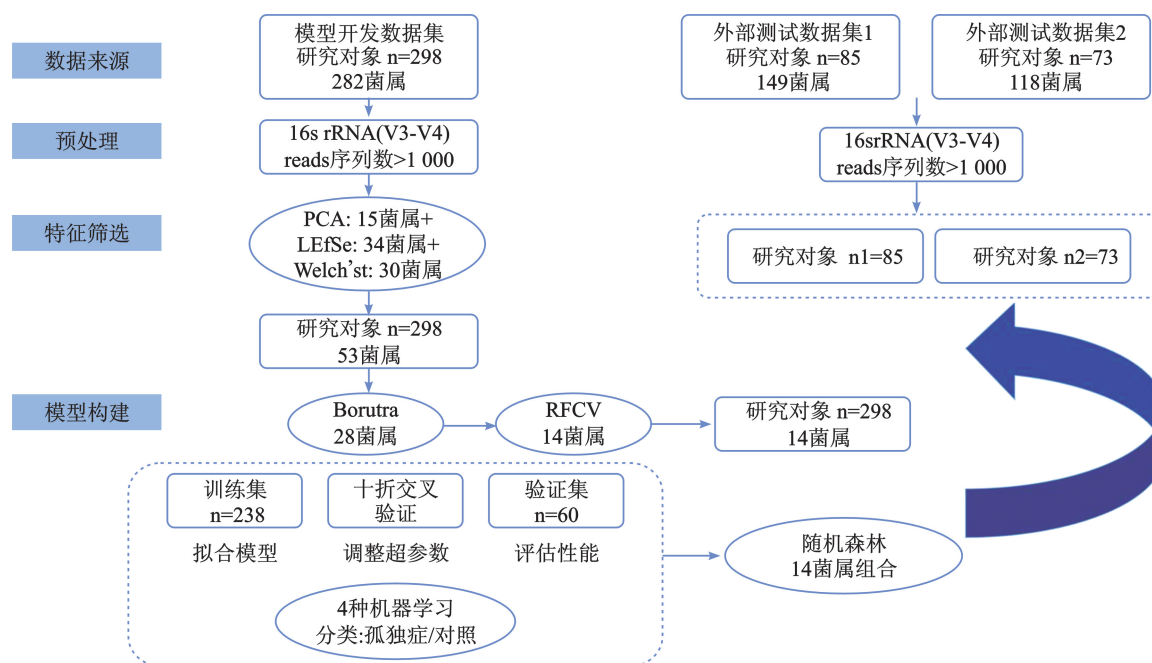


图 1 筛查模型构建和测试流程

Fig.1 Flow chart of screening model construction and testing

1.5.3 模型结果评价和可视化

利用 R 4.2.2 中“pROC”和“caret”包,进行受试者工作特征(receiver operating characteristic, ROC)和曲线下面积(area under the curve, AUC)的计算,

并绘制 ROC 曲线。AUC 值介于 0.5 与 1 之间,越接近 1,说明模型的分性能越好,反应模型的总体表现。使用机器学习评估指标衡量模型的性能,包括准确率、召回率、精度、马修斯相关系数(Matthews

correlation coefficient, MCC)。

2 结果

2.1 孤独症组与对照组儿童基本信息

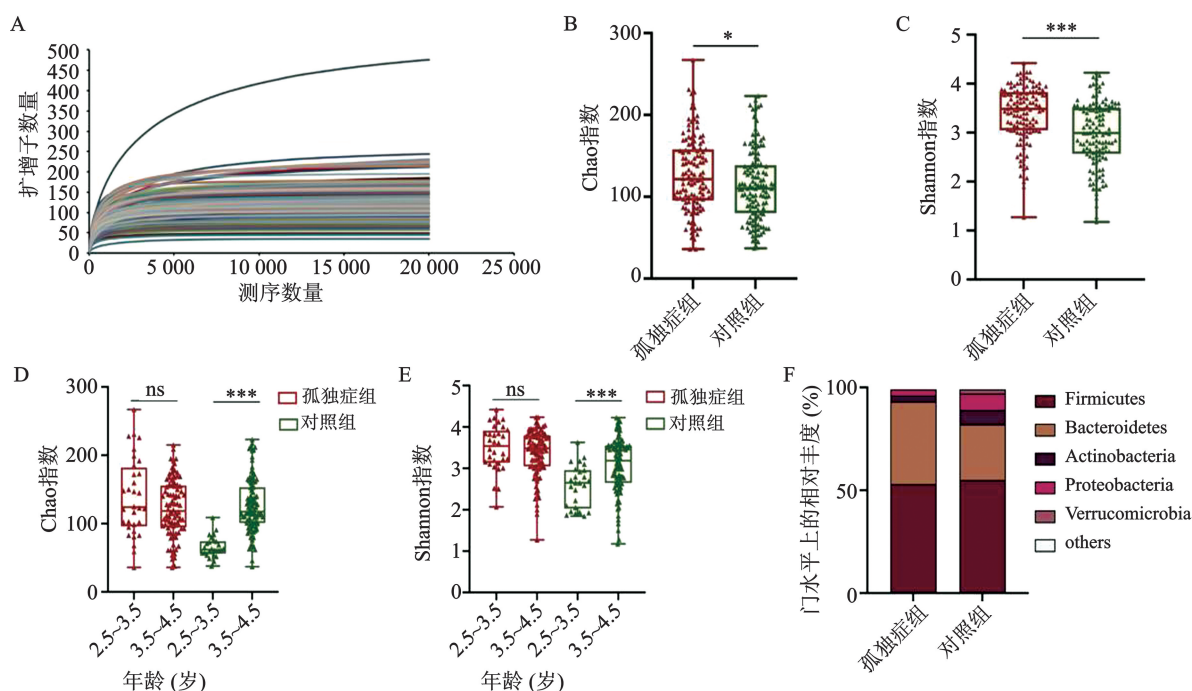
孤独症组 149 例患儿平均年龄 3.9 岁(年龄范围在 2.5~4.5 岁),对照组 149 例健康儿童平均年龄 3.8 岁(年龄范围在 3.0~4.0 岁)。两组男女比例为 1:1,均为男性 115 例,女性 34 例。孤独症组 101 例(67.8%)来自华北地区的济南,48 例(32.2%)来自西南地区的遵义。在对照组中,117 例儿童(78.5%)来自华东地区的上海,32 名儿童(21.5%)来自中国华南地区的香港特别行政区。

2.2 孤独症组与对照组儿童肠道菌群组成

完成质量控制和去除冗余序列后,298 份儿童粪便样本中总共获得了 12 374 919 条高质量的序列,最低为 10 647 条,最高为 316 537 条,均一化分析用数据量为 10 647 条。本次共聚类 4 995 个 ASV,分属于 338 个属。在 338 个属中,有 67 个属在至少一个组中相对丰度大于 0.1%,占孤独症组总丰度的 97.5%,占对照组总丰度的 98%。根据不同

的测序深度构建稀释曲线(见图 2A),稀释曲线呈现明显的平台期,所有样本的采样覆盖率都很高,说明本次测序深度已经足够覆盖样本中大部分的物种,测序深度足以进行后续的粪便微生物群研究。

两组的 α 多样性分析显示,孤独症组的 Chao 指数[118.00 (93.50, 154.00) vs. 105.00 (75.00, 135.00), $P=0.023$]和 Shannon 指数[3.46 (3.04, 3.80) vs. 3.00 (2.56, 3.48), $P=0.001$]显著高于对照组,说明孤独症组具有更高的丰度和多样性(见图 2B、图 2C)。进一步基于年龄分层进行 α 多样性分析。与 2.5~3.5 岁亚组相比,对照组 3.5~4.5 岁亚组的 Chao 指数显著增加($P<0.05$),而孤独症组无显著变化($P>0.05$)(见图 2D),对照组 3.5~4.5 岁亚组的 Shannon 指数较 2.5~3.5 岁亚组显著升高($P<0.05$),而孤独症组的变化仍无统计学意义($P>0.05$)(见图 2E)。通过比较两组在门水平上的相对丰度(见图 2F),孤独症组厚壁菌门(*Bacillota*)和拟杆菌门(*Bacteroidota*)的比值显著低于对照组($P<0.05$)。



注: * 为 $P<0.05$, *** 为 $P<0.001$, ns 为 $P>0.05$ 。

图 2 孤独症和对照组儿童肠道菌群组成分析

Fig.2 Composition analysis of gut microbiota between ASD and control children

2.3 孤独症和对照组儿童肠道菌群结构差异

基于 Bray-Curtis 距离进行的 PCoA 可视化结果(见图 3),孤独症组与对照组的粪便微生物群落

展现一定程度的重叠。在 PCoA 中,能够最大程度区分所有样本的两个主要坐标轴解释了 7.18% 和 11.34% 的差异。进一步通过 PerMANOVA 评估 β 多

样性,结果表明两组间的微生物群落在整体分布上呈现出显著性差异($F=5.198, R^2=0.052, P<0.001$)。

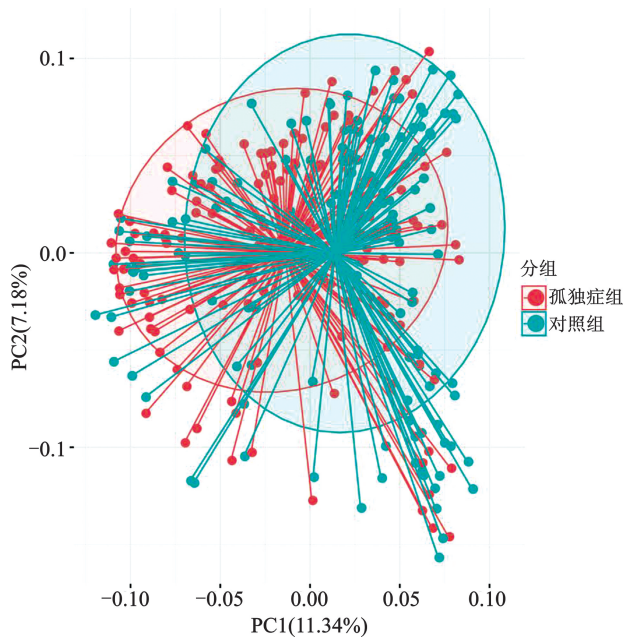


图3 孤独症组与对照组肠道菌群组间结构差异

Fig.3 The structural differences of gut microbiota between ASD and control group

2.4 差异物种选择

首先通过 Welch's t 检验(双侧),识别出 30 个 $FDR<0.05$ 的属作为差异菌属。另外,PCA 识别出 15 个方差贡献率超过 1% 的菌属,将这 15 个重要菌属从高到低排序,占总方差的 98.3%。使用线性判

别分析(linear discriminant analysis effect size, LEfSe)识别孤独症组与对照组中有差异的物种,设定 LDA 分值为 3.0,筛选 $LDA>3.0$ 且 FDR (调整后的 P 值) <0.05 的物种作为组间差异物种,获得 34 个差异菌属。结合这三种方法筛选的变量,共有 53 个差异菌属被纳入筛查模型。

考虑到 53 个属可能存在的冗余性,使用 R 语言中的“Boruta”包算法进一步进行特征筛选,最后共有 28 个菌属的重要性得分大于最大阴影属性,在重要性图中被标记为绿色(见图 4A)。最后,使用 R 语言中的“rfcv”函数进行 5 次十折交叉验证,发现当筛查模型中包含 14 个菌属时,模型错误率最低(见图 4B)。根据“MeanDecreaseGini”重要性指数,分类能力最高的菌属是 *Ruthenibacterium*。其中,发现 9 个菌属在孤独症儿童中富集,这些属包括 *Phocaeicola*、*Anaerobutyricum*、粪杆菌属(*Faecalibacterium*)、经黏液真杆菌属(*Blautia*)、颤螺菌属(*Oscillibacter*)、毛螺菌属(*Lachnospira*)、*Parabacteroides*、*Flintibacter* 和 *Anthropogastromicrobium*。而在对照组中,*Ruthenibacterium*、*Flavonifractor*、双歧杆菌(*Bifidobacterium*)、*Anaerostipes* 和 *Eisenbergiella* 的丰度较高(见图 4C)。最后筛选出的 14 个菌属在孤独症组中的总丰度达 49.26%,在对照组中为 38.65%(见表 1)。

表 1 孤独症和对照组中 14 种生物标志物的平均相对丰度

Table 1 Average relative abundance of 14 biomarkers in ASD and control groups

菌属	孤独症平均丰度(%)	孤独症丰度标准差(%)	对照平均丰度(%)	对照丰度标准差(%)	FDR	丰度变化(%)
<i>Ruthenibacterium</i>	0.07	0.25	1.02	2.06	<0.001	0.95
<i>Phocaeicola</i>	19.85	15.08	10.73	15.82	<0.001	9.12
<i>Anaerobutyricum</i>	0.56	0.62	0.21	0.41	<0.000	0.35
<i>Faecalibacterium</i>	15.84	10.55	12.07	15.41	0.104	3.78
<i>Blautia</i>	3.62	3.10	3.06	4.47	0.490	0.56
<i>Oscillibacter</i>	0.20	0.31	0.05	0.21	<0.001	0.15
<i>Lachnospira</i>	2.68	3.67	1.11	3.37	0.006	1.56
<i>Parabacteroides</i>	1.52	2.29	0.89	1.68	0.066	0.63
<i>Flavonifractor</i>	0.19	0.31	0.58	1.03	0.001	0.39
<i>Flintibacter</i>	0.18	0.38	0.06	0.16	0.011	0.12
<i>Anthropogastromicrobium</i>	0.25	0.60	0.06	0.19	0.009	0.19
<i>Bifidobacterium</i>	3.45	4.00	6.90	13.36	0.035	3.45
<i>Anaerostipes</i>	0.81	1.03	1.39	6.10	0.532	0.58
<i>Eisenbergiella</i>	0.03	0.16	0.52	1.68	0.013	0.49

注:计算两组平均相对丰度的平均值(%)和标准差(%);FDR 采用 STAMP 计算;丰度变化代表两组间相对丰度变化的绝对值。

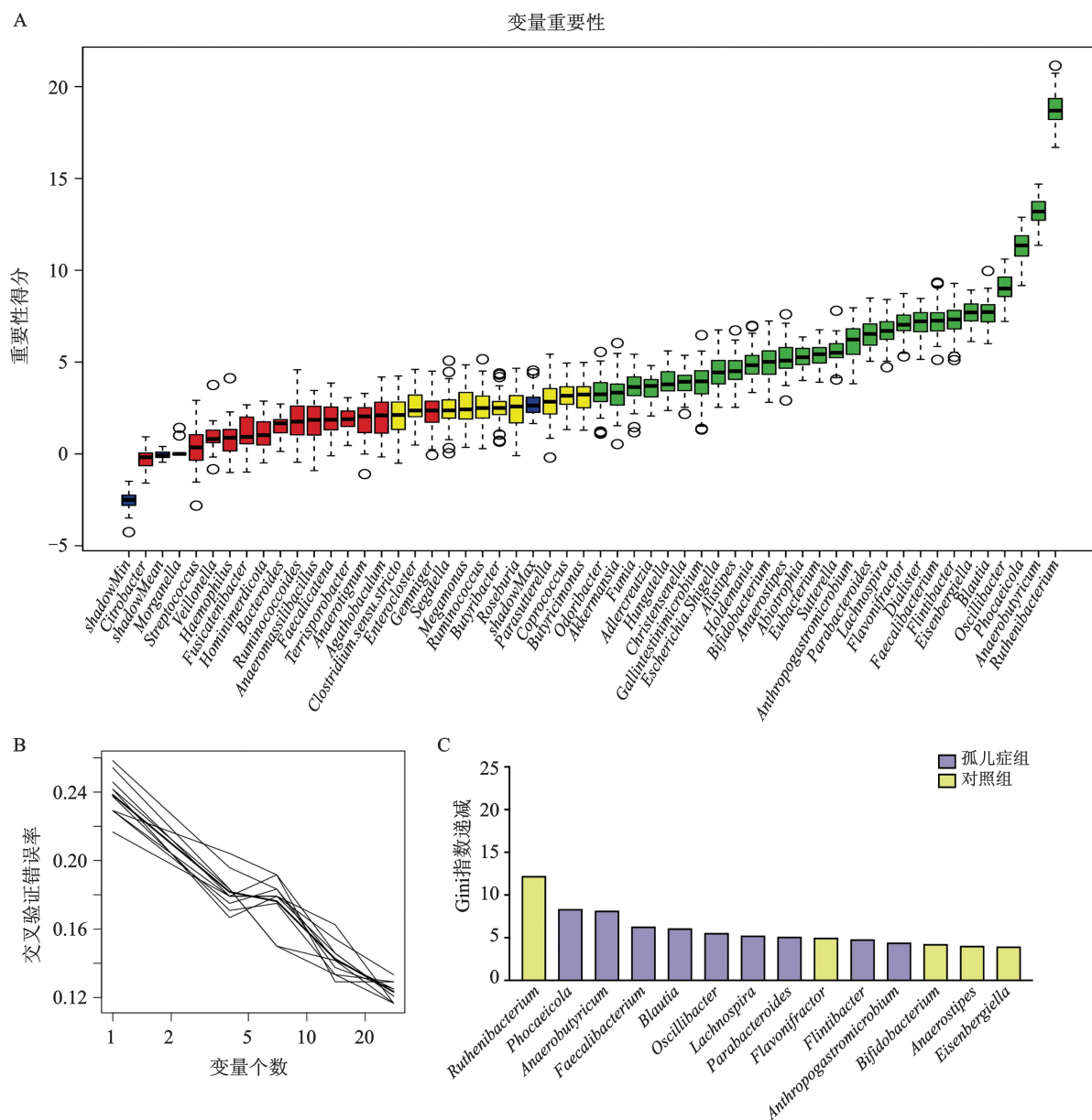


图 4 基于 Boruta 和 rfcv 算法的特征菌属筛选

Fig.4 Screening of characteristic genera based on Boruta and rfcv algorithms

2.5 机器学习模型

为了测试潜在的生物标志物是否可以用于孤独症患病状态的分类,使用机器学习算法基于 14 个菌属建立了孤独症筛查模型。计算 RF、SVM、KNN、NBC 这四种分类算法在开发模型数据集中训练集和验证集上的正确率、召回率、精度和 MCC 指标。从可以看出,随机森林分类器在训练数据集(准确率=1,召回率=1,精度=1,MCC=1)和验证数据集(准确率=0.84,召回率=0.79,精度=0.88,MCC=0.69)上都表现出较好的预测性能,AUC 分别为 100%(95%CI:100%~100%)和 93.94%(95%CI:88.13%~99.74%)(见表 2、图 5A、图

5B)。

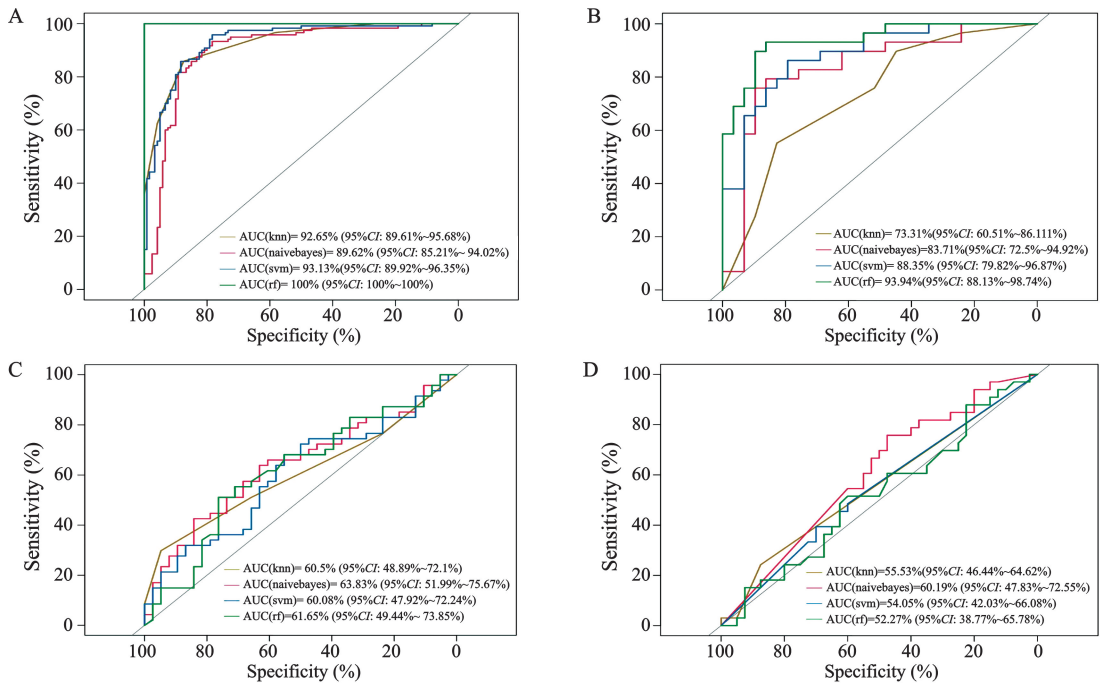
运用两个独立外部测试队列对上述构建的四种分类模型进行评估。表 2 的外部测试结果表明,朴素贝叶斯模型在外部测试数据集 1(准确率=0.56,召回率=0.83,精度=0.57,MCC=0.08)和外部测试数据集 2(准确率=0.56,召回率=0.61,精度=0.51,MCC=0.13)中均具有更好的泛化能力,AUC 值分别为 63.83%(95%CI:51.99%~75.67%)和 60.19%(95%CI:47.83%~72.55%)(见表 2、图 5C、图 5D)。虽然这 4 个模型均能区分孤独症和对照,但与开发模型数据分类结果相比,对外部测试数据的分类性能有所下降。

表 2 模型开发队列和外部测试队列机器学习评价指标

Table 2 Machine learning evaluation indicators for model development cohort and external test cohort

数据集	机器学习模型	准确率	召回率	精度	MCC	AUC(%)
训练集	RF	1	1	1	1	100
	KNN	0.87	0.86	0.87	0.73	92.65
	SVM	0.87	0.86	0.87	0.73	93.13
	NBC	0.77	0.96	0.69	0.58	89.62
验证集	RF	0.84	0.79	0.88	0.69	93.94
	KNN	0.64	0.76	0.61	0.28	73.31
	SVM	0.81	0.76	0.85	0.62	88.35
	NBC	0.72	0.9	0.67	0.48	83.71
外部测试 1	RF	0.59	0.78	0.6	0.14	61.65
	KNN	0.41	0.7	0.48	-0.31	60.50
	SVM	0.55	0.81	0.57	0.06	60.08
	NBC	0.56	0.83	0.57	0.08	63.83
外部测试 2	RF	0.56	0.18	0.55	0.08	52.27
	KNN	0.56	0.03	1	0.13	55.53
	SVM	0.45	0.61	0.43	-0.07	54.05
	NBC	0.56	0.61	0.51	0.13	60.19

注:准确率为正确分类的样本占样本总数的比例;召回率为 ASD 样本中被正确识别为自闭症的百分比;精度为 ASD 样本在模型预测的所有样本中所占的比例;AUC 越大,模型性能越好,范围在 0~1;MCC 即马修斯相关系数,是用来评价二元分类模型性能的指标,它考虑真、假阳性、真阴性和假阴性情况的比例,适用于不平衡数据集。MCC 的取值范围从-1 到+1,+1 表示完美预测,0 表示随机预测,-1 表示完全不一致的预测。



注:A和B为在开发数据的训练和验证集中使用14个特征菌属的四种不同分类模型进行对照和孤独分类的ROC曲线。C和D为在外部测试集中使用开发数据构建好的最优模型基础上,测试相同的14个菌属对孤独症的分类能力。横轴 Specificity 表示特异度,纵轴 Sensitivity 表示灵敏度,将 ROC 曲线左上角被认为是最佳临界点,在此临界点同时满足灵敏度和特异度相对最优。ROC 为受试者工作特征曲线,AUC 为曲线下面积,四个模型的 ROC 曲线用不同颜色的线表示,对应 AUC 用百分比(最高 100%)和 95%置信区间表示。

图 5 训练集、验证集、外部测试集中的模型分类性能

Fig.5 Model classification performance in training set, validation set, and external test set

3 讨论

3.1 机器学习特征筛选

本研究基于肠道菌群数据,使用不同机器学习模型进行疾病状态的预测。尽管机器学习目前在国内外研究中取得了巨大突破和应用潜力,但个体间肠道菌群组成的显著差异仍然给模型应用的有效性带来了严峻挑战。传统的孤独症预测模型通常依赖于 LEfSe 方法进行差异菌群的筛选。然而,由于 LEfSe 分析基于非参数检验,可能导致较高假阴性率^[7]。为减少假阴性率,本文采用了三种不同的方法(LEfSe、STAMP 和 PCA)联合识别预测特征变量,并筛选出 53 个菌属作为预测模型的候选特征。由于菌属间可能存在的冗余性,进一步筛选有助于降低模型的复杂性,减少泛化误差,并提升模型的预测性能和可解释性。Boruta 算法很好改善模型泛化误差,通过向系统添加随机性来降低随机波动和相关性的潜在干扰,更准确地揭示特征与疾病状态之间的内在关联。与基于单个随机森林的特征选择方法相比,Boruta 算法通常能够产生更为稳定的结果^[8]。最终构建的预测模型包含了 14 个错误率最低的属,其中五个属(*Anthropogastromicrobium*、*Bifidobacterium*、*Flavonifractor*、*Flintibacter* 和 *Oscillibacter*)先前在 LEfSe 分析中并未被识别。这 14 个菌属的丰度在孤独症患者中达到 49.26%,在对照组中为 38.65%。此外,筛选出的生物标志物在两组之间的丰度变化范围为 0.12%至 9.12%。

在传统的疾病研究中,显著性检验主要用于寻找差异特征,但这种方法在研究复杂的神经发育障碍时可能会忽略关键预测因子^[9]。有研究者^[10]指出,预测模型筛选出的变量可能包括在人类肠道微生物组成中未显示统计学显著性的预测因子。最终选定的 14 个预测因子中,有四个属在 STAMP 分析中并未显示统计学显著性,但很可能影响大脑功能。*Parabacteroides* 已被发现与一系列精神疾病的高水平相关,包括重性抑郁障碍和双相情感障碍^[11]。已有研究发现 *Faecalibacterium* 与一些参与干扰素 γ 表达的基因上调和下调有关,干扰素是胎儿发育过程中与 ASD 相关的细胞因子^[12]。炎症环境被认为与 ASD 的发生发展有关,*Blautia* 可通过其代谢

物减少炎症和代谢状态,并具有抗菌能力^[13]。*Anaerostipes* 主要与丁酸盐的产生有关^[14],可能通过短链脂肪酸的作用影响 ASD 的发生。虽然以上 4 个菌属在统计学上无显著性差异,但两组的绝对变化均较高。丰度差异最大的是 *Phocaeicola*,达到了 9.12%。动物研究表明 *Phocaeicola* 与认知有关,在社交焦虑小鼠中检测到更多的 *Phocaeicola masonii*^[15]。

3.2 机器学习模型构建

利用预测模型寻找儿童早期孤独症的生物标志物目前尚处于探索阶段。Zhao 等^[16]使用更适合小样本研究的支持向量机分类模型,筛选出对模型贡献最大的 15 个优势菌属,AUC 为 0.88。在探究疾病分类的机器学习算法方面,一项研究表明当使用来自不同队列的年龄和性别匹配的孤独症和对照样本构建预测模型时显示出良好的分类性能(AUC>0.87)。而结合所有样本进行建模时,模型预测性能下降(AUC=0.78)^[17]。可以推断,不同种群之间的微生物异质性阻碍了分类绩效的有效性。这与我们目前的研究是一致的。随机森林模型显示,来自中国四个不同地区的年龄和性别匹配的孤独症和正常儿童之间存在更好的分辨能力,突出了两组之间的微生物差异。虽然将该模型的预测扩展到两个外部跨队列数据集仍然可以预测 ASD,但其预测性能明显下降。由此可见,中国不同地区饮食差异造成的混杂因素可能人为地提高了分类性能。在外部测试和开发队列之间的受试者之间的基线不平衡可能影响概括结果的能力。预测模型的有效性因地区和人口而异。

完整的预测模型研究应该包括模型的内部和外部测试。开发队列分为 80%的训练集和 20%的验证集。在训练集通过内部十倍交叉验证选择最优超参数,以提高选择结果的可靠性,避免过拟合问题^[18]。外部效度反映了模型的一般性,需要使用非来自研究本身的数据集进行验证,这些数据集与时间无关或地理上相互独立^[19]。因此在本研究中,我们选择了两个完全独立于开发队列的外部数据集来验证模型的泛化能力。

3.3 机器学习模型评价

当处理不平衡的数据集时,准确性会被高估,因

因为它没有完全考虑最终分数计算中混淆矩阵的四分类的大小。召回率和精确度的侧重点不一样,更关注于找到尽可能多的阳性患者和测量模型对阳性患者的预测准确性^[20]。在临床实际应用中,可根据不同的目的选择不同的评价指标。AUC 高意味着模型对正样本的预测概率高于对负样本的预测概率,不受阈值选择的影响,可以综合考虑模型在各种阈值下的性能。AUC 在正负样本数量不平衡情况下,也能给出有意义的评价结果。在实际应用中,MCC 能够平衡真阳性、真阴性、假阳性和假阴性,通常为不平衡数据集提供更准确和全面的性能评估^[21]。它不仅考虑了所有可能的预测结果,而且还提供了一个直观的、范围特定的值来衡量预测与实际结果的一致性。

由于本研究采用回顾性研究设计,宿主年龄、地理、饮食、基因型、胃肠状态、疾病状态等影响因素未被纳入研究,无法证实因果关系和排除混杂因素干扰。因此,未来仍需建立纵向人群队列,探索其他宿主因素与肠道生物标志物的关联,以及它们的联合预测是否可以作为早期筛查和改善自闭症症状的更有效手段,将是我们下一步的重点。

综上所述,本研究发现了孤独症组和对照组儿童肠道微生物多样性、结构的差异,并鉴定出 14 个差异菌属。借助机器学习算法,构建孤独症儿童的早期筛查模型,分别在内部开发数据集和外部测试数据集中对构建的筛查模型进行性能评估,结果提示肠道菌群对辅助孤独症无创筛查有潜在意义。

[参考文献]

- [1] Rosen N E, Lord C, Volkmar F R. The diagnosis of autism: from Kanner to DSM-III to DSM-5 and beyond[J]. *J Autism Dev Disord*, 2021, 51(12): 4253-4270.
- [2] Zeidan J, Fombonne E, Scora J, et al. Global prevalence of autism: a systematic review update[J]. *Autism Res*, 2022, 15(5): 778-790.
- [3] Maenner M J, Warren Z, Williams A R, et al. Prevalence and characteristics of autism spectrum disorder among children aged 8 years-autism and developmental disabilities monitoring network, 11 sites, United States, 2020[J]. *MMWR Surveill Summ*, 2023, 72(2): 1-14.
- [4] 贾美香. 提升对我国孤独症谱系障碍儿童发病状况及早期干预的关注[J]. *中国妇幼健康研究*, 2023, 34(1): 1-4.
- [5] Wan Y, Zuo T, Xu Z, et al. Underdevelopment of the gut microbiota and bacteria species as non-invasive markers of prediction in children with autism spectrum disorder[J]. *Gut*, 2022, 71(5): 910-918.
- [6] Bundgaard-Nielsen C, Knudsen J, Leutscher P D C, et al. Gut microbiota profiles of autism spectrum disorder and attention deficit/hyperactivity disorder: a systematic literature review[J]. *Gut Microbes*, 2020, 11(5): 1172-1187.
- [7] Cho H, Qu Y, Liu C, et al. Comprehensive evaluation of methods for differential expression analysis of metatranscriptomics data [J]. *Brief Bioinform*, 2023, 24 (5): bbad279.
- [8] Jayasinghe W J M L P, Deo R C, Ghahramani A, et al. Deep multi-stage reference evapotranspiration forecasting model: multivariate empirical mode decomposition integrated with the boruta-random forest algorithm[J]. *Ieee Access*, 2021, 9: 166695-166708.
- [9] Monteith S, Glenn T, Geddes J, et al. Expectations for artificial intelligence (AI) in psychiatry[J]. *Curr Psychiatry Rep*, 2022, 24(11): 709-721.
- [10] Wang M, Doenyas C, Wan J, et al. Virulence factor-related gut microbiota genes and immunoglobulin a levels as novel markers for machine learning-based classification of autism spectrum disorder[J]. *Comput Struct Biotechnol J*, 2021, 19: 545-554.
- [11] Gomez-Nguyen A, Basson A R, Dark-Fleury L, et al. Parabacteroides distasonis induces depressive-like behavior in a mouse model of Crohn's disease[J]. *Brain Behav Immun*, 2021, 98: 245-250.
- [12] Inoue R, Sakaue Y, Sawai C, et al. A preliminary investigation on the relationship between gut microbiota and gene expressions in peripheral mononuclear cells of infants with autism spectrum disorders[J]. *Biosci Biotechnol Biochem*, 2016, 80(12): 2450-2458.
- [13] Li R, Li L, Hong P, et al. β -Carotene prevents weaning-induced intestinal inflammation by modulating gut microbiota in piglets[J]. *Anim Biosci*, 2021, 34 (7): 1221-1234.
- [14] Kircher B, Woltemate S, Gutzki F, et al. Predicting butyrate- and propionate-forming bacteria of gut microbiota from sequencing data [J]. *Gut Microbes*, 2022,

- 14(1):2149019.
- [15] Ritz N L, Brocka M, Butler M I, et al. Social anxiety disorder-associated gut microbiota increases social fear[J]. *Proc Natl Acad Sci U S A*, 2024, 121(1): e2308706120.
- [16] Zhao Y, Wang Y, Meng F, et al. Altered gut microbiota as potential biomarkers for autism spectrum disorder in early childhood[J]. *Neuroscience*, 2023, 523: 118-131.
- [17] Morton J T, Jin D M, Mills R H, et al. Multi-level analysis of the gut-brain axis shows autism spectrum disorder-associated molecular and microbial profiles [J]. *Nat Neurosci*, 2023, 26(7): 1208-1217.
- [18] Charilaou P, Battat R. Machine learning models and over-fitting considerations [J]. *World J Gastroenterol*, 2022, 28(5): 605-607.
- [19] Collins G S, Reitsma J B, Altman D G, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement[J]. *BMJ*, 2015, 350: g7594.
- [20] Kumar Y, Koul A, Singla R, et al. Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda[J]. *J Ambient Intell Humaniz Comput*, 2023, 14(7): 8459-8486.
- [21] Chatterjee A, Pahari N, Prinz A, et al. Machine learning and ontology in eCoaching for personalized activity level monitoring and recommendation generation[J]. *Sci Rep*, 2022, 12(1): 19825.
- [专业责任编辑: 王 丽]
[中文编辑: 向婉婉; 英文编辑: 牛 惠]