

DOI:10.3969/j.issn.1005-202X.2024.01.017

医学生物信息

## 基于机器学习的胃癌关键基因筛选及预测模型构建

王泽朋, 李坤鹏, 周玉, 李四海

甘肃中医药大学信息工程学院, 甘肃 兰州 730100

**【摘要】目的:**为了验证与胃癌相关的遗传特征,提出一种混合式特征选择方法确定靶基因,进一步分析其意义并建立新的诊断预测模型。**方法:**对原始胃癌数据进行生物信息学方差分析,使用随机森林、支持向量机的递归特征消除、套索算法等机器学习方法筛选胃癌相关基因,对结果取交集,获得关键基因集。进行富集分析,确定关键基因并验证;依据关键基因构建基于多层感知器(MLP)、逻辑回归、决策树等8种机器学习分类算法的诊断预测模型。**结果:**混合式的特征选择方法筛选出的关键基因与肿瘤发生和发展的生物学过程密切相关;8个关键基因(TXNDC5、BMP8A、ONECUT2、COL10A1、JCHAIN、INHBA、LCTL和TRIM59)被确定为诊断效果较好的胃癌潜在标志物;根据8种分类模型的ROC曲线和准确率结果可知,MLP为最佳胃癌预测模型,其准确率高达97.77%,比他人构建的Xgboost胃癌预测模型准确率高出3.83%。**结论:**本研究获得了诊断和预防胃癌的8个关键基因,并建立了最佳预后模型。

**【关键词】**胃癌;基因筛选;关键基因;生物信息学;机器学习

**【中图分类号】**R318;R735.2

**【文献标志码】**A

**【文章编号】**1005-202X(2024)01-0115-10

## Key gene screening and prediction model construction of gastric cancer based on machine learning

WANG Zepeng, LI Kunpeng, ZHOU Yu, LI Sihai

School of Information Engineering, Gansu University of Chinese Medicine, Lanzhou 730100, China

**Abstract: Objective** To verify the genetic characteristics associated with gastric cancer, and to propose a hybrid feature selection method for identifying target genes, further analyzing their significance and establishing a new diagnostic prediction model. **Methods** Analysis of variance in bioinformatics was performed on the original gastric cancer data, and then machine learning methods such as random forest, recursive feature elimination of support vector machine, and LASSO algorithm were used to screen gastric cancer associated genes, and the intersection of results was taken as the key gene set. The key genes were identified and verified through enrichment analysis. The diagnosis and prediction models based on 8 kinds of machine learning classification algorithms such as multi-layer perceptron, logistic regression and decision tree, were constructed using the key genes. **Results** The key genes selected by the hybrid feature selection method were closely related to the tumorigenesis and development. Eight key genes (TXNDC5, BMP8A, ONECUT2, COL10A1, JCHAIN, INHBA, LCTL and TRIM59) were identified as potential markers of good diagnostic efficacy in gastric cancer. The ROC curve and accuracy results demonstrated that among the 8 classification models, MLP is the best gastric cancer prediction model, with an accuracy of 97.77%, which was 3.83% higher than that of Xgboost gastric cancer prediction model. **Conclusion** The study identifies 8 key genes for the diagnosis and prevention of gastric cancer, and establishes the optimal prognosis model.

**Keywords:** gastric cancer; gene screening; key gene; bioinformatics; machine learning

**【收稿日期】**2023-10-12

**【基金项目】**甘肃省科技计划项目(21JR1RA272);甘肃省教育厅高校教师创新基金项目(2023B-105)

**【作者简介】**王泽朋,硕士,研究方向:生物信息学、机器学习, E-mail: 19193137887@163.com

**【通信作者】**李四海,副教授,研究方向:数据挖掘、机器学习、光谱分析, E-mail: lisihai@gszy.edu.cn

## 前言

胃癌是一种严重危害人类健康的恶性肿瘤,其病死率在世界各地都很高<sup>[1]</sup>。引起胃癌的主要原因有以下几点:幽门螺旋杆菌、饮食习惯不规律及烟酒过度摄入。由于早期发病症状不明显,难以引起人们的重视,使得很多患者错过最佳的治疗时机。因此,如何实现胃癌早期预测诊断,成为攻克现代胃癌的重要难关。近年来国内外学者在分子基因层面对癌症

基因筛选做了大量研究,其中靶向治疗工程最为引人注目,利用基因检测手段对胃癌进行诊断,随后再基于诊断出的致病基因进行靶向治疗<sup>[2]</sup>。因此,关键基因的筛选检测对于胃癌早期诊断、预后分析及靶向治疗具有重要意义。

在癌症基因筛选问题上,国内外科学家主要分为两个方向,一个是利用生物信息学研究方法,另一个是在基因筛选过程中加入机器学习模型。赵博璇等<sup>[3]</sup>对胃癌基因表达进行详细的研究,并通过芯片分析建立胃癌早期预测和分类模型,模型准确率为96.7%,为建立胃癌诊断和预后预测模型提供思路和启示。在基因筛选应用问题上,刘辉等<sup>[4]</sup>应用癌症基因组图谱(TCGA)胃癌数据构建基于加权基因共表达网络和套索算法(LASSO)预测模型,并找到靶基因HKR1,侧面验证了机器学习对于基因筛选结果的重要性和准确性。本文为了进一步提高基因筛选的准确性,主要通过基因筛选过程中加入最先进的机器学习方法,将随机森林(RF)、支持向量机特征递归消除(SVM-RFE)和LASSO与生物信息学相结合,建立一种混合型的新算法<sup>[5]</sup>。在建立预测模型的问题上,主要运用了以极限梯度提升、轻量级梯度提升、支持向量机、多层感知器等8种机器学习分类算法为基础,和其他研究构建的分类模型进行比较,本文算法效果更好,为胃癌早期预防提供有力的科学依据,也为病理生理过程的分子机制提供新的见解。

## 1 数据来源与预处理

### 1.1 数据来源

本文使用的数据来自公开的TCGA<sup>[6]</sup>和基因型-组织表达(GTEX)<sup>[7]</sup>数据库。首先下载TCGA数据库中446例胃癌标本数据,从GTEX数据库网站下载359个正常组织样本,以弥补TCGA数据库中正常组织样本数量较少的不足,并与TCGA表达矩阵数据合并,为建立分类模型增加样本平衡,提高模型的说服力。

### 1.2 数据预处理

合并TCGA和GTEX的基因表达数据,对数据进行预处理。再利用R语言对基因表达数据进行分类,判定01A为胃癌组织样本,11A为正常组织样本。合并后的数据集包含805个组织样本,其中胃癌组织样本有410个,正常组织样本有395个。

## 2 方法

### 2.1 差异表达分析

由于TCGA与GTEX两种基因序列来自于不同的测序平台,因此两种基因序列之间可能会有批次

间的差异,因此应先进行批次效应处理<sup>[8]</sup>。利用R语言中的Deseq2软件包实现对数据的批次效益去除和差异表达基因的筛选<sup>[9]</sup>。差异分析设置 $|\log FC| > 2, P < 0.05$ 为差异有统计学意义,筛选出的胃癌相关的差异基因通过火山图可视化展出。具体原理如下:将数据进行log2转换,得到 $[-1, 1]$ 作为阈值,识别差异基因,具体计算公式为:

$$\text{Fold change} = \frac{\bar{x}_{\text{tumor}}}{\bar{x}_{\text{normal}}} \quad (1)$$

### 2.2 富集分析

富集分析也叫通路分析,是分析基因信息中常用的方法之一。基因本体论(Gene Ontology, GO)<sup>[10]</sup>分析和京都基因与基因组百科全书(Kyoto Encyclopedia of Genes and Genomes, KEGG)<sup>[11]</sup>的通路分析是目前最常用的两种富集分析方法。其中,GO分析分为3部分:分子功能(MF)、细胞组分(CC)、生物学过程(BP);KEGG通路包含基因组、化学以及系统的功能信息。对于关键基因集,常用DO疾病富集分析,用来分析关键基因集富集的具体疾病通路。本研究中,利用R语言中的clusterProfiler软件包来实现差异基因的DO疾病分析、GO分析和KEGG通路分析,通过富集显著性(P-value)的超几何检验来衡量判定富集结果是否显著,计算公式如下:

$$P = 1 - \sum_{i=0}^{m-1} \frac{\binom{M}{i} \binom{N-M}{n-i}}{\binom{N}{n}} \quad (2)$$

当评估基因富集的程度时,通常会使用 $P < 0.05$ 作为标准,如果该标准被满足,则表明该基因在差异表达基因中具有显著的富集特征,最后以柱状图和圈图的形式可视化富集分析结果。

### 2.3 基于机器学习算法的基因筛选

在创建基因筛选模型时,首先使用差异表达分析进行筛选,之后使用机器学习中的3种特征选择算法:RF、SVM-RFE、LASSO算法,通过不断训练和调整模型参数进一步完善待筛选基因的范围,从而达到每个模型的最佳状态。

**2.3.1 RF基因筛选** RF的核心思想是通过构建多颗决策树,并将它们的预测结果进行集成处理,提高模型的性能,在分类问题上,每颗决策树都有一个分类标记,最后的结果就是每颗决策树的投票结果。在基于RF思想的基因筛选中,由于面对的基因变量较大,本文采用逐步筛选的思想,以通过模型的多次训练,提高筛选结果的精度,进而更精准地找到关键基因。

**2.3.2 SVM-RFE基因筛选** Guyon等<sup>[12]</sup>在SVM基础

上首次提出改进的SVM-RFE方法,采用一种基于回归特征剔除(RFE)的方法,将一个又一个的基因剔除,结果令人满意,是一种经典的基因选择算法。其核心思想是通过多次迭代,每次迭代过程中使用SVM模型对特征排序和选择,其工作流程图如图1所示。本文根据数据的特点,进行交叉交叉验证,每次使用2、4、6、8、30步长的基因进行训练。本文设定最优特征子集的最大容量为30,选出每次实验排序前30名基因作为筛选集,根据特征出现的次数和排名位置进行加权求和再次排序,最终构成筛选出的关键基因子集。

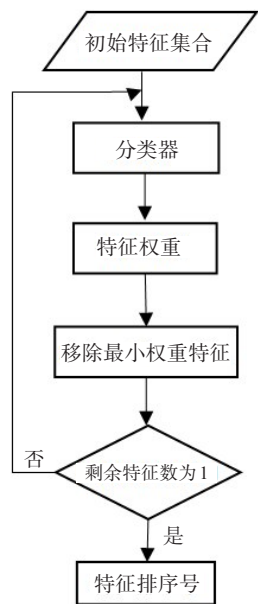


图1 SVM-RFE流程图  
Figure 1 SVM-RFE flowchart

**2.3.3 LASSO基因筛选** 在本研究中,采用LASSO算法对基因集进行筛选,LASSO的主要优点是能够自动选择对目标变量有预测能力的特征,并将不重要的特征系数压缩为零,降低模型的复杂性,提高通用性,并减少过拟合的风险。其数学公式如下所示:

$$L(\beta) = \frac{1}{2n} \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j)^2 \tag{3}$$

其中, $L(\beta)$ 代表损失函数,表示为拟合误差; $n$ 代表样本数量; $p$ 代表特征数量,本文代表患癌样本; $y_i$ 表示第*i*个样本实际观测值; $x_{ij}$ 表示第*i*个样本的第*j*个特征的取值; $\beta_0$ 表示截距项; $\beta_j$ 表示第*j*个特征的系数。

在LASSO算法中,正则化项是最关键项,它是由系数的绝对值之和构成,公式如下所示:

$$R(\beta) = \lambda \sum_{j=1}^p |\beta_j| \tag{4}$$

其中, $R(\beta)$ 表示正则化项, $\lambda$ 表示正则化参数,它控制了正则化的强度, $\lambda$ 越大会导致更多的系数被压缩为0,从而实现特征选择。

LASSO的目标是最小化损失函数和正则化项的和,由此可以得出LASSO的计算公式为:

$$\hat{\theta}^{lasso} = \min_{\beta} \left[ L(\beta) + \lambda \sum_{j=1}^p |\beta_j| \right] \tag{5}$$

2.4 诊断效能分析

采用ROC曲线法,曲线下面积(AUC)对候选关键基因进行诊断评估<sup>[13]</sup>。设置AUC值大于0.9的候选基因确定为胃癌早期诊断的关键基因。

2.5 诊断预测模型构建

使用Python机器学习扩展包scikit-learn开发基于极限梯度提升(Xgboost)<sup>[14]</sup>、轻量级梯度提升(LightGBM)<sup>[15]</sup>、支持向量机(SVM)<sup>[16]</sup>、多层感知器(MLP)<sup>[17]</sup>、逻辑回归(Logistic)<sup>[18]</sup>、决策树(DecisionTree)<sup>[19]</sup>、高斯朴素贝叶斯(GaussianNB)<sup>[20]</sup>、自适应提升(Adaboost)<sup>[21]</sup>等8种方法的胃癌早期诊断预测模型。

2.6 模型验证与评估

由于每种分类算法在不同的训练数据集上有不同的训练效果,为了减少每种模型的过拟合问题,采用常用的准确度、精确度、召回率、F1分数、ROC曲线和AUC值等方法来估算指标<sup>[22]</sup>。其中ROC曲线的下限被定义为AUC,作为一个数值,AUC越高代表分类准确率越高。此外,混淆矩阵是评估二元分类模型的另一个常用指标<sup>[23]</sup>。

3 结果

3.1 胃癌差异表达基因

使用Deseq2软件包对TCGA和GTEx联合数据集中的差异表达基因进行批量收益去除和筛选,结果发现了908个DEGs,包括301个上调基因和607个下调基因,并得到火山图如图2所示。

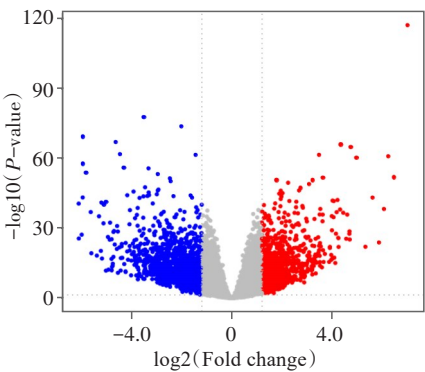


图2 胃癌组织与正常组织间DEGs火山图  
Figure 2 DEGs volcano diagram between gastric cancer and normal tissues

3.2 胃癌差异表达基因 GO 注释和 KEGG 分析结果

对与胃癌相关的关键基因展开富集分析,筛选出实际概率  $P<0.05$  的富集途径,其中,GO 富集分析结果包括 296 个条目:生物过程(Biological Process, BP)条目 220 条,分子功能(Molecular Function, MF)条目 53 条,细胞组分(Cellular Component, CC)2 条目 3 条。

条目  $P$  值按特定顺序排列,3 个过程中每个过程的前 8 条记录被选中并显示在图 3 中。富集结果显示,BP 与细胞外基质的组织、细胞外基质结构的组织和外囊结构的组织、胶原纤维排列等密切相关,提示胃癌的发生多与膜结构联系密切。主要结果见表 1。

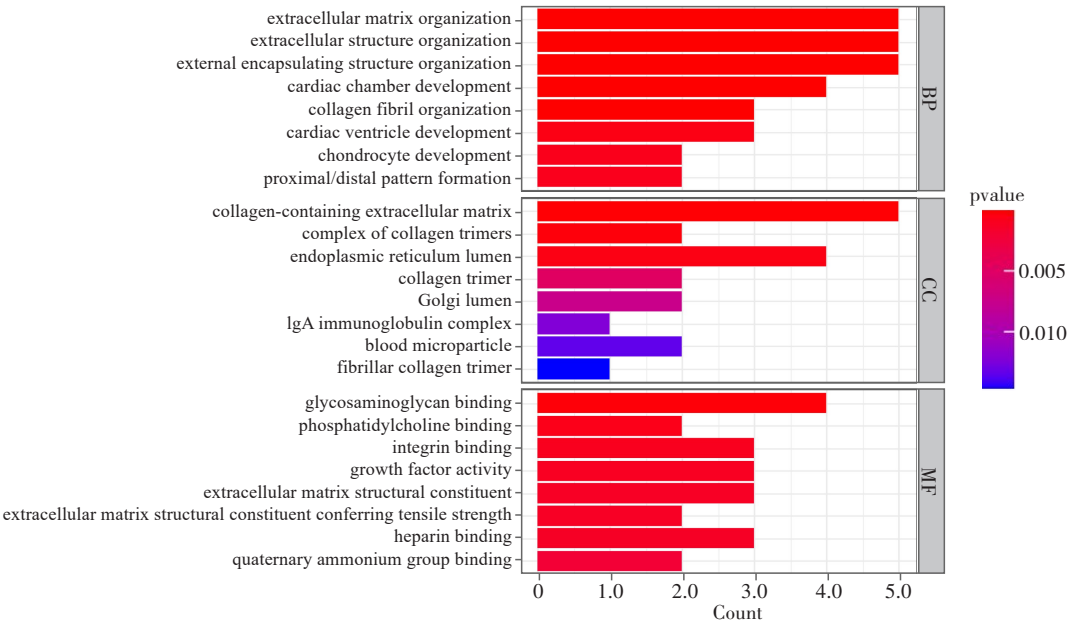


图3 GO疾病分析柱状图  
Figure 3 Histogram of GO disease analysis

表1 GO功能富集分析部分结果  
Table 1 Partial results of GO functional enrichment analysis

类别	编号	描述	$P$ 值	调整后的 $P$ 值	数量
BP	GO:0030198	extracellular matrix organization	$4.32\times10^{-5}$	$9.71\times10^{-3}$	5
	GO:0043062	extracellular structure organization	$4.32\times10^{-5}$	$9.71\times10^{-3}$	5
	GO:0045229	external encapsulating structure organization	$4.32\times10^{-5}$	$9.71\times10^{-3}$	5
CC	GO:0062023	collagen-containing extracellular matrix	$1.29\times10^{-4}$	$8.78\times10^{-3}$	5
	GO:0098644	complex of collagen trimers	$3.00\times10^{-4}$	$1.02\times10^{-2}$	2
	GO:0005788	endoplasmic reticulum lumen	$5.29\times10^{-4}$	$1.20\times10^{-2}$	4
MF	GO:0005539	glycosaminoglycan binding	$2.16\times10^{-4}$	$1.79\times10^{-2}$	4
	GO:0031210	phosphatidylcholine binding	$6.96\times10^{-4}$	$1.79\times10^{-2}$	2
	GO:0005178	integrin binding	$9.71\times10^{-4}$	$1.79\times10^{-2}$	3

KEGG 途径结果包含 21 个条目,关键基因主要在蛋白质消化与吸收、细胞外基质与受体分子之间的相互作用、氨基酸的生物合成、癌症中的蛋白聚糖、胃酸分泌等方面富集明显。通过 Benjamini-Hochberg 校正后减少错误发现率,将  $P$  值按升序排列,以圈形图的方式展示前 21 个通路(图 4)。表 2 展

示部分通路包含的基因数量等条目结果。

3.3 基因特征选择得到基因标志物

在差异分析的基础上,使用 RF、SVM-RFE 和 LASSO 算法重新研究了 908 个与胃癌相关的差异基因。RF 通过生成决策树的数量筛选,本文选择  $n_{tree}=1\ 000$ ,最终提取前 25 个特征基因作为目标结果;SVM-RFE 通过

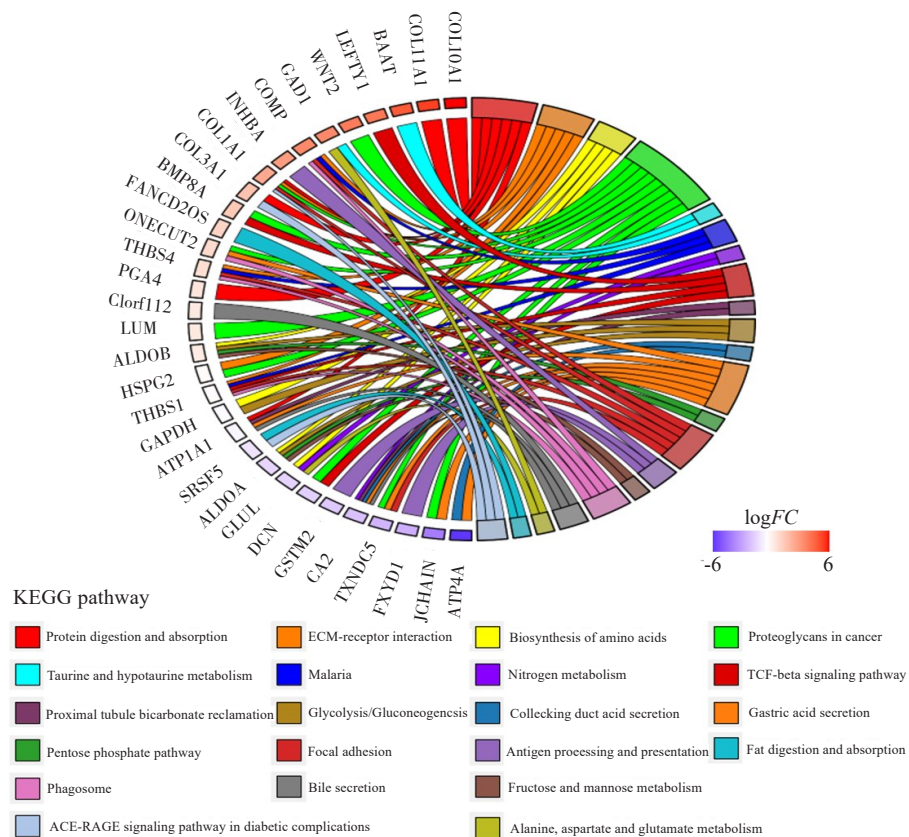


图4 KEGG 富集分析的圈图

Figure 4 Circle diagram of KEGG enrichment analysis

表2 KEGG通路富集分析部分结果

Table 2 Partial results of enrichment analysis of KEGG pathway

编号	描述	<i>P</i> 值	调整后的 <i>P</i> 值	数量
hsa04974	Protein digestion and absorption	1.96×10 <sup>-4</sup>	3.00×10 <sup>-2</sup>	6
hsa04512	ECM-receptor interaction	8.12×10 <sup>-4</sup>	6.22×10 <sup>-2</sup>	5
hsa01230	Biosynthesis of amino acids	3.35×10 <sup>-3</sup>	1.71×10 <sup>-1</sup>	4
hsa05205	Proteoglycans in cancer	6.80×10 <sup>-3</sup>	1.85×10 <sup>-1</sup>	9
hsa00430	Taurine and hypotaurine metabolism	7.51×10 <sup>-3</sup>	1.85×10 <sup>-1</sup>	2
hsa04971	Gastric acid secretion	2.48×10 <sup>-2</sup>	2.70×10 <sup>-1</sup>	6
hsa04510	Focal adhesion	2.60×10 <sup>-2</sup>	2.70×10 <sup>-1</sup>	5

十折交叉检验后,保存前30个特征基因结果;LASSO选择 lambda.min 参数后得到30个特征基因。将三者两两取交集得到:RF和LASSO取交集而没有SVM-RFE时有4个最佳标志基因:INHBA、COL10A1、ONECUT2、JCHAIN;RF和SVM-RFE取交集而没有LASSO时有8个最佳标志基因:LCTL、TRIM59、MYZAP、SRSF5、FANCD2OS、C1orf112、FXYD1、GSTM2;RF和LASSO和SVM-RFE同时取交集得到2个共同基因:TXNDC5、BMP8A,见图5。RF前20个特征重要性见图6。LASSO筛选见图7。SVM-

RFE筛选的前20个特征重要性见图8。

3.4 标志物的验证

对筛选出来的特征基因进行DO疾病功能分析,发现基因主要富集在胃癌、胃腺癌、屈光不正、退行性椎间盘病变等疾病,见图9。ROC曲线表明, TXNDC5的AUC值为0.95, BMP8A的AUC值为0.92;ONECUT2、COL10A1、JCHAIN和INHBA在数据集AUC值分别为0.95、0.90、0.91、0.92; LCTL、TRIM59、MYZAP、SRSF5、FANCD2OS、C1orf112、FXYD1和GSTM2在数据集AUC值分别为0.91、

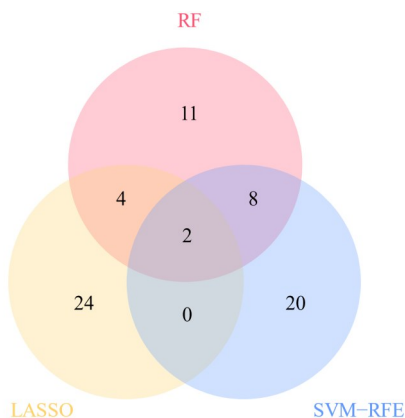


图5 重叠基因韦恩图  
Figure 5 Wayne diagram of overlapping genes

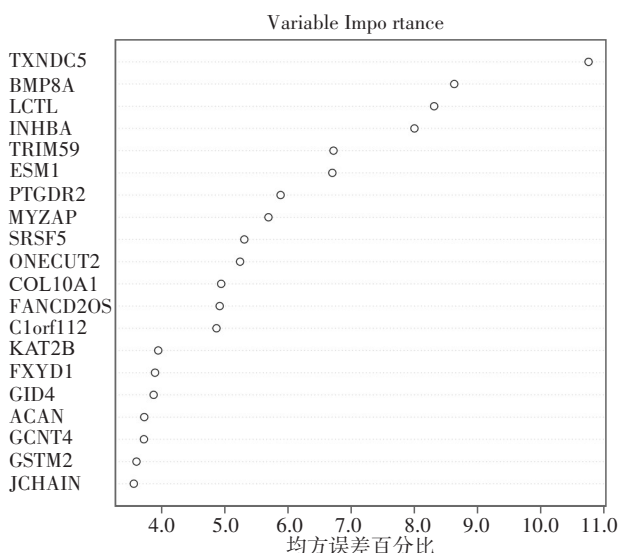


图6 RF算法筛选的前20个基因的特征重要性图  
Figure 6 Feature importance of the top 20 genes screened by RF algorithm

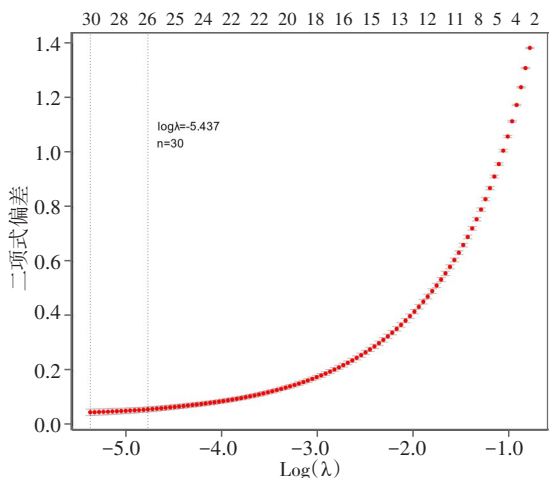


图7 LASSO算法筛选的特征基因  
Figure 7 Feature genes screened by LASSO algorithm

0.92、0.80、0.84、0.85、0.86、0.89、0.85, 结果见图10。当AUC值在0.9以上时,提取出的基因显示出较高的诊断价值,故最终关键基因集为TXNDC5、BMP8A、ONECUT2、COL10A1、JCHAIN、INHBA、LCTL、TRIM59。

### 3.5 胃癌诊断预测模型的构建

利用筛选出的8个胃癌早期关键基因构建胃癌的早期诊断预测模型,其步骤如下:(1)提取出8个关键基因在联合数据集中的表达值从而形成新的表达矩阵;(2)从TCGA和GTEx联合数据集中,410个早期胃癌组织样本和395个正常组织样本被随机分成训练集,前5轮训练集由724个样本组成,后5轮训练集包含725个样本,前5轮测试集中包含81个样本,后5轮测试集中包含80个样本。

采取十折交叉验证法构建基于Xgboost、LightGBM、SVM、MLP、Logistic、DecisionTree、GaussianNB、Adaboost 8种算法的诊断预测模型。在测试集中8种分类方法表现效果优秀,各指标均高于0.9,见表3。根据图11可知,各模型具有较高的AUC值。在含有81个样本的独立测试集中进行模型效能验证。8种分类模型算法中,MLP模型表现较为优异,准确率高达97.77%。本文采用的MLP分类模型相比于文献[3]构建的Xgboost模型准确度提高3.83%(文献[3]预测模型的结果如表4所示)。结果表明,基于MLP构建的胃癌诊断预测模型性能较好,鲁棒性较强。

## 4 讨论

近年来,许多研究表明,分子标记物在疾病的诊断、预后和靶向治疗中发挥着重要作用。随着诊断和治疗水平的不断进步,胃癌的病因和治疗因素不断被发现,但胃癌的具体病因仍不清楚,尤其是幽门螺杆菌感染、不良饮食习惯、不卫生的环境和吸烟等常见危险因素<sup>[24]</sup>。此外,误诊误治和转移也是导致胃癌死亡的主要原因。本研究使用生物信息学方法初步筛选了胃癌基因表达数据的差异基因,结果显示,共有908个差异基因,其中包括301个上调基因,607个下调基因,并绘制了火山图,以供差异基因的可视化;在差异分析的基础上,采用机器学习中的特征选择方法,利用3种特征选择方法(RF、LASSO、SVM-RFE)分别进行筛选,最后两两取交集,确定关键基因集,并对关键基因集做富集分析,GO富集结果发现在生物过程中其与胞外基质组织、细胞外结构组织、外部包膜结构组织、胶原纤维排列等密切相

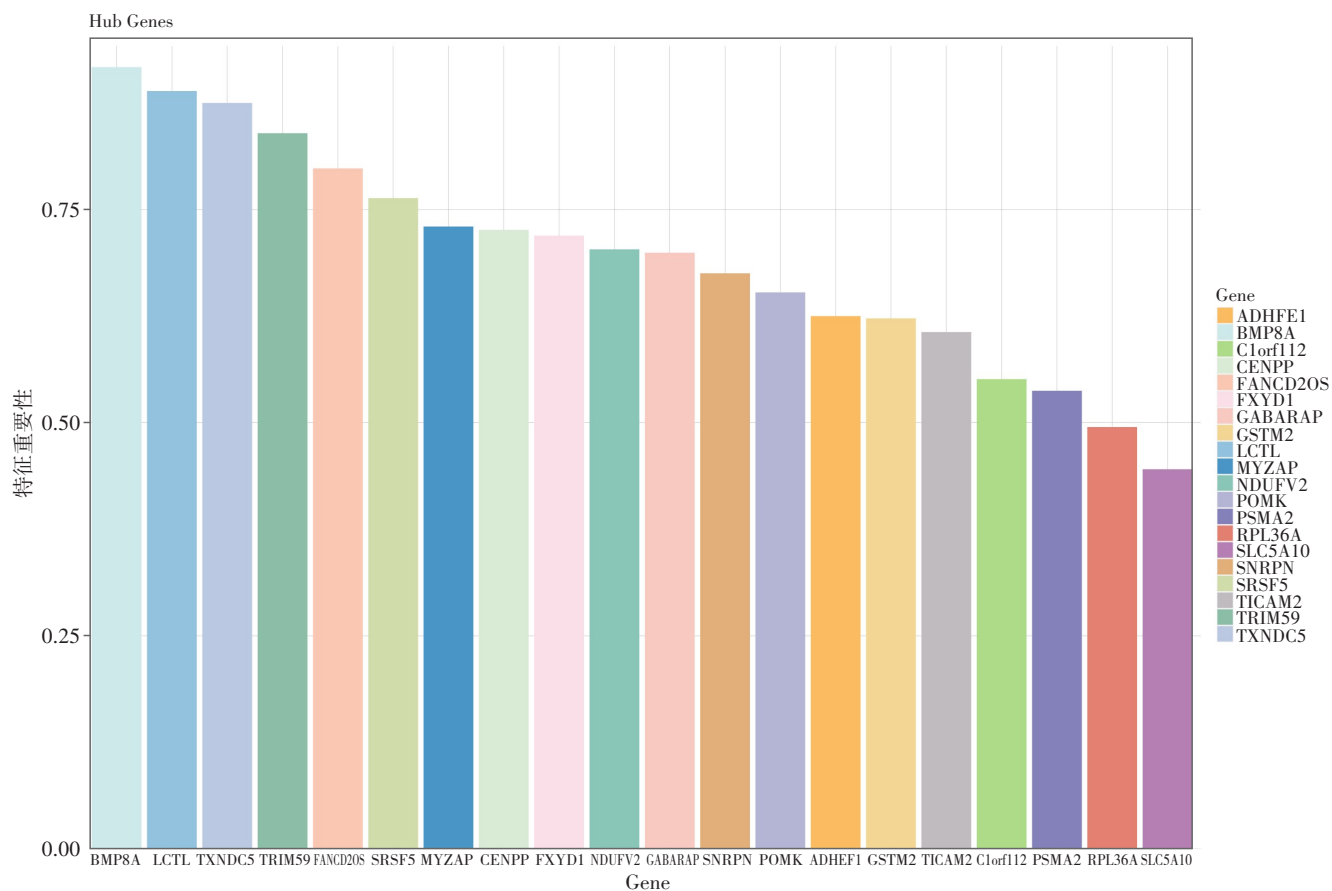


图8 SVM-RFE 算法筛选的特征基因重要性图

Figure 8 Importance of feature genes screened by SVM-RFE algorithm

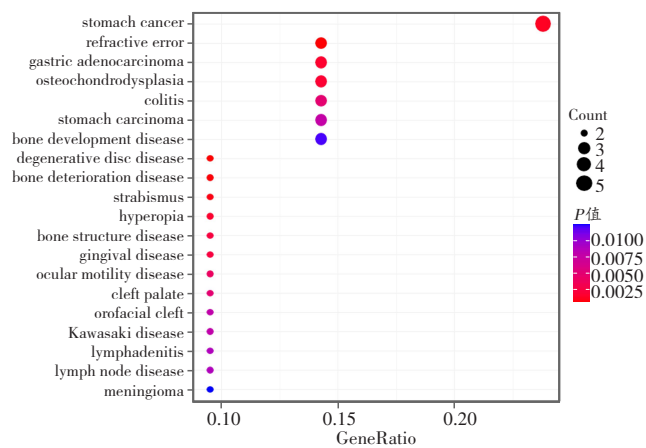


图9 DO 疾病分析气泡图

Figure 9 Bubble diagram of DO disease analysis

关,KEGG 通路分析表明,在蛋白质消化和吸收、受体分子与细胞外基质的相互作用以及癌症中的蛋白多糖等方面都有明显的富集。在疾病富集分析中主要富集在胃癌疾病问题上。基于基因表达数据,对 14 个候选关键基因(TXNDC5、BMP8A、ONECUT2、COL10A1、JCHAIN、INHBA、LCTL、TRIM59、

MYZAP、SRSF5、FANCD2OS、C1orf112、FXYD1、GSTM2)进行诊断效能分析,通过 AUC 值大于 0.9 进行判断选取最终的关键基因,最终发现 TXNDC5、BMP8A、ONECUT2、COL10A1、JCHAIN、INHBA、LCTL、TRIM59 为胃癌最佳关键基因。

张林等<sup>[25]</sup>通过 Co-IP 方法证实 TXNDC5 在胃癌细胞和组织中的高表达,证明其在胃癌发生过程中有着促进作用。在对胃癌的模型研究中,李相辉等<sup>[26]</sup>通过胃癌种植瘤裸鼠模型治疗实验显示, TXNDC5 siRNA 靶向纳米微粒在体内条件下对胃癌治疗作用较为显著,表明 TXNDC5 可能会成为治疗胃癌的重要靶点。BMP8A 在肿瘤的产生过程中起到促进作用,在甲状腺乳头状癌研究中,曾学宇等<sup>[27]</sup>发现 BMP8A 在甲状腺乳头状癌病理发生过程中高表达,但其分子功能机制在胃癌的产生过程中还尚未研究清楚,有待成为未来的重要研究靶点之一。为了探究基因 ONECUT2 在人胃癌中的表达意义,丁鹏等<sup>[28]</sup>利用生物信息学方法探究 ONECUT2 的蛋白相互作用,最终发现在胃癌组织中现高表达,并与胃癌的产生有着重要的关系。在胃癌预后研究问题上,

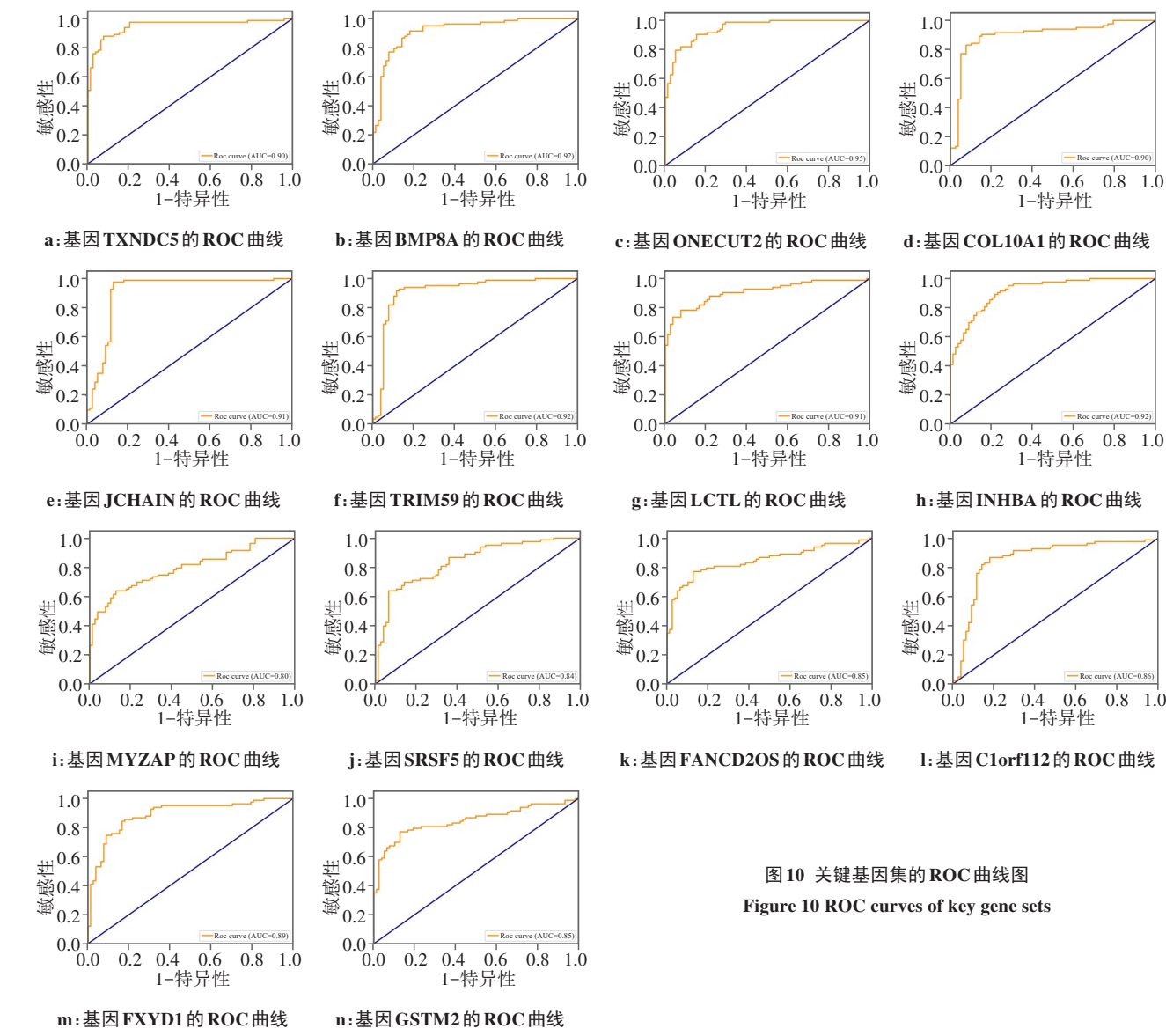


图 10 关键基因集的 ROC 曲线图  
Figure 10 ROC curves of key gene sets

表 3 8 种模型在独立测试集上的评价指标  
Table 3 Evaluation indicators of 8 models on independent test set

模型	准确度	精确度	召回率	F1 分数
MLP	0.977 7	0.978 4	0.977 5	0.977 6
LightGBM	0.975 1	0.975 6	0.975 1	0.975 1
SVM	0.972 7	0.974 6	0.972 2	0.972 6
Xgboost	0.966 5	0.967 1	0.966 4	0.966 4
Logistic	0.957 8	0.959 2	0.957 6	0.957 7
GaussianNB	0.955 2	0.957 5	0.954 9	0.955 1
Adaboost	0.949 1	0.949 8	0.949 0	0.949 0
DecisionTree	0.949 1	0.949 9	0.949 0	0.949 0

牛刚等<sup>[29]</sup>发现 COL10A1 血清与胃癌复发转移及患者预后有关。在胃癌分类模型问题上, Pan 等<sup>[30]</sup>通过构建免疫细胞浸润程度对胃癌患者进行分类的评估模型, 利用 Cox 和 LASSO 回归分析确定关键基因

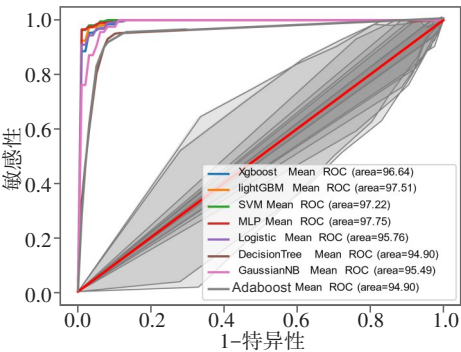


图 11 训练集 ROC 曲线  
Figure 11 ROC curves on the training set

JCHAIN, 结果表明在胃癌肿瘤组织中表达低水平 JCHAIN 的患者表现出更好的预后。基于生物信息学方法, 张亮等<sup>[31]</sup>发现 INHBA 在胃癌中显著高表达, 并与胃癌的临床分期和 T 期密切相关, INHBA 高表达的患者预后较差。LCTL 也称为“乳糖酶样蛋白”,

表4 文献[3]预测模型数据  
Table 4 Prediction model data from reference [3]

模型	准确度	精确度	召回率	F1 分数
SVM	0.939 4	0.882 4	1.000 0	0.937 5
RF	0.939 4	0.882 4	1.000 0	0.937 5
NBM	0.909 1	0.833 3	1.000 0	0.909 1
KNN	0.909 1	0.875 0	0.933 3	0.933 2
Xgboost	0.939 4	0.933 3	0.933 3	0.933 3
Adaboost	0.939 4	0.882 4	1.000 0	0.937 5

其具体功能和生理意义仍在研究中,Su等<sup>[32]</sup>研究发现LCTL在神经胶质瘤中高表达,对于肿瘤的发生发展有着促进作用,有一定的研究价值。研究胃癌中一种新的致癌驱动因子TRIM59的多态性,Luo等<sup>[33]</sup>利用HapMap标签 snp 方法,筛选出3个标签TRIM59单核苷酸多态性进行基因分型,结果表明变异rs1141023A等位基因的携带显著增加了胃癌发生的风险( $P=0.006$ )。

通过以上研究分析发现,利用机器学习方法筛选胃癌关键基因是可行的,机器学习筛选方法在大量冗余数据中表现优异,筛选出的关键基因有一定的研究价值,为后续相关学者进行不同肿瘤疾病关键基因检测问题上提供一定的结论支撑。此外本研究结果还需要在临床随访中进一步证实。

5 结 论

利用生物信息学和不同的机器学习算法相结合的联合方法研究发现TXNDC5、BMP8A、ONECUT2、COL10A1、JCHAIN、INHBA、LCTL、TRIM59在胃癌病理过程中呈高表达,可能作为胃癌的潜在标志物,利用这8个关键基因进行胃癌诊断预测,结果显示对胃癌样本和正常样本分类识别能力较为准确,表明这些基因与胃癌的产生和发展有着密切的联系,可作为胃癌早期诊断及研究的潜在靶点。

综上所述,本文提出一种生物信息学与机器学习联合的方法,对胃癌中的关键基因进行筛查,并建立多个机器学习分类模型,实现对胃癌的早期诊断与预测,最终选择在测试集上表现优异的MLP为最佳的诊断预测模型,准确率为97.77%,与其他人构建的Xgboost模型相比,准确率提高3.83%。因此,MLP分类模型可用于早期胃癌诊断预测,为肿瘤基因筛选提供新思路。本研究不足之处有以下几点:(1)目前对胃癌发病机理的认识还不够深刻,转录组数据无法全面阐明机体的总体变化;(2)目前的研究主要

集中在基因筛选和预测模型的建立上,缺乏体内或体外实验的支持。在下一步的研究中,将会加强与生物学实验相结合,对筛选出的基因进行更精确、更可靠的验证。

【参考文献】

[1] 胡珊,黄奔,姜扬,等. 基于lncRNA-miRNA-mRNA调节网络胃癌关键基因筛选与分析[J]. 中华肿瘤防治杂志, 2020, 27(7): 511-518.  
Hu S, Huang B, Jiang Y, et al. Screening and analysis of key genes in gastric cancer based on lncRNA-miRNA-mRNA regulatory network [J]. Chinese Journal of Cancer Prevention and Treatment, 2020, 27(7): 511-518.

[2] 吴思滢,向丽娟,包楚阳,等. 胃癌嘌呤代谢通路差异基因筛选及其与预后关系的研究[J]. 安徽医科大学学报, 2021, 56(11): 1802-1806.  
Wu SH, Xiang LJ, Bao CY, et al. Differential gene screening of purine metabolic pathway in gastric cancer and its relationship with prognosis [J]. Acta Universitatis Medicinalis Anhui, 2021, 56(11): 1802-1806.

[3] 赵博璇,刘明,李建伟. 基于生物信息学的胃癌早期诊断预测模型研究[J]. 生物信息学, 2022, 20(4): 274-283.  
Zhao BX, Liu M, Li JW. Study on early diagnosis and prediction model of gastric cancer based on bioinformatics [J]. Bioinformatics, 2022, 20(4): 274-283.

[4] 刘辉,张超,景丽伟,等. 基于WGCNA联合LASSO算法胃癌预后lncRNA分子标志物筛选[J]. 临床肿瘤学杂志, 2021, 26(10): 891-897.  
Liu H, Zhang C, Jing LW, et al. Screening of lncRNA molecular markers for gastric cancer prognosis based on WGCNA combined with LASSO algorithm [J]. Journal of Clinical Oncology, 2021, 26(10): 891-897.

[5] 文宏伟,陆菁菁,何晖光. 机器学习在神经精神疾病诊断及预测中的应用[J]. 协和医学杂志, 2018, 9(1): 19-24.  
Wen HW, Lu JJ, He HG. Application of machine learning in diagnosis and prediction of neuropsychiatric diseases [J]. Union Medical Journal, 2018, 9(1): 19-24.

[6] Liñares-Blanco J, Pazos Sierra A, Fernandez-Lozano C. Machine learning analysis of TCGA cancer data [J]. Peer J Comput Sci, 2021, 7: e584.

[7] Ardlie KG, Dermitzakis ET. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans [J]. Science, 2015, 348(6235): 648-660.

[8] Randall DW, Kieswich J, Swann J, et al. Batch effect exerts a bigger influence on the rat urinary metabolome and gut microbiota than uraemia: a cautionary tale [J]. Microbiome, 2019, 7(1): 127.

[9] Liu S, Wang Z, Zhu R, et al. Three differential expression analysis methods for RNA sequencing: limma, EdgeR, DESeq2 [J]. JoVE, 2021, 2021(175): e62528.

[10] Klopfenstein DV, Zhang L, Pedersen BS, et al. GOATOOLS: a python library for gene ontology analyses [J]. Sci Rep, 2018, 8(1): 10872.

[11] Chen H, Zhang Y, Awasthi SK, et al. Effect of red kaolin on the diversity of functional genes based on Kyoto Encyclopedia of Genes and Genomes pathways during chicken manure composting [J]. Bioresour Technol, 2020, 311(0): 123584.

[12] Guyon I, Weston J, Barnhill S, et al. Gene selection for cancer classification using support vector machines [J]. Mach Learn, 2002, 46(1-3): 389-422.

[13] Janssens A, Martens FK. Reflection on modern methods: revisiting the area under the ROC Curve [J]. Int J Epidemiol, 2020, 49(4): 1397-1403.

[14] Hou N, Li M, He L, et al. Predicting 30-days mortality for MIMIC-III patients with sepsis-3: a machine learning approach using XGboost [J]. J Transl Med, 2020, 18(1): 462.

[15] Yan J, Xu Y, Cheng Q, et al. LightGBM: accelerated genomically designed crop breeding through ensemble learning [J]. Genome Biol, 2021, 22(1): 271.

[16] Zhou S. Sparse SVM for sufficient data reduction [J]. IEEE Trans Pattern Anal Mach Intell, 2022, 44(9): 5560-5571.

[17] Qin Y, Li C, Shi X, et al. MLP-based regression prediction model for compound bioactivity [J]. Front Bioeng Biotech, 2022, 10: 946329.

[18] Pelánek R. Bayesian knowledge tracing, logistic models, and beyond:

- an overview of learner modeling techniques[J]. User Model User-Adap, 2017, 27(3-5): 1-38.
- [19] Sharma H, Kumar S. A survey on decision tree algorithms of classification in data mining[J]. Int J Sci Res, 2016, 5(4): 2094-2097.
- [20] Kaviarasi R, Gandhi Raj R. Accuracy enhanced lung cancer prognosis for improving patient survivability using proposed gaussian classifier system[J]. J Med Syst, 2019, 43(7): 201.
- [21] Pham BT, Nguyen MD, Nguyen-Thoi T, et al. A novel approach for classification of soils based on laboratory tests using adaboost, tree and ANN modeling[J]. Transp Geotech, 2021, 27(2): 100508.
- [22] Muschelli J. ROC and AUC with a binary predictor: a potentially misleading metric[J]. J Classif, 2020, 37(3): 696-708.
- [23] 张开放, 苏华友, 窦勇. 一种基于混淆矩阵的多分类任务准确率评估新方法[J]. 计算机工程与科学, 2021, 43(11): 1910-1919.
- Zhang KF, Su HY, Dou Y. A new method for accuracy evaluation of multiple classification tasks based on confusion matrix[J]. Computer Engineering and Science, 2021, 43(11): 1910-1919.
- [24] Smyth EC, Nilsson M, Grabsch HI, et al. Gastric cancer[J]. Lancet, 2020, 396(10251): 635-648.
- [25] 张林, 侯艳红, 李湘辉, 等. TXNDC5 蛋白在胃癌细胞及组织中相互作用蛋白的免疫共沉淀验证[J]. 胃肠病学和肝病学杂志, 2022, 31(8): 861-866.
- Zhang L, Hou YH, Li XH, et al. Verification of TXNDC5 protein interaction protein in gastric cancer cells and tissues by immunoprecipitation[J]. Journal of Gastroenterology and Hepatology, 2022, 31(8): 861-866.
- [26] 李相辉, 侯艳红, 吴凯, 等. TXNDC5 siRNA 靶向纳米微粒对胃癌模型动物抑瘤作用的实验研究[J]. 实用医学杂志, 2023, 39(13): 1634-1640.
- Li XH, Hou YH, Wu K, et al. Experimental study on the inhibitory effect of TXNDC5 siRNA targeting nanoparticles on gastric cancer model animals[J]. Journal of Practical Medicine, 2023, 39(13): 1634-1640.
- [27] 曾学宇, 陈柱, 毛敏, 等. 甲状腺乳头状癌的 BMP8A 表达及其与颈部淋巴结转移的相关性[J]. 西安交通大学学报(医学版), 2021, 42(3): 443-447.
- Ceng XY, Chen Z, Mao M, et al. Expression of BMP8A in papillary thyroid carcinoma and its correlation with cervical lymph node metastasis[J]. Journal of Xi'an Jiaotong University (Medical Edition), 2021, 42(3): 443-447.
- [28] 丁鹏, 闫洁, 秦艳茹. ONECUT2 基因在人胃癌中的表达及其临床意义[J]. 中国肿瘤生物治疗杂志, 2020, 27(4): 433-439.
- Ding P, Yan J, Qin YR. Expression of ONECUT2 gene in human gastric cancer and its clinical significance[J]. Chinese Journal of Tumor Biotherapy, 2020, 27(4): 433-439.
- [29] 牛刚, 王建锋, 裴瑜, 等. 胃癌根治术后 COL10A1 血清表达与其复发转移及远期预后的相关性[J]. 实用癌症杂志, 2022, 37(12): 2032-2035.
- Niu G, Wang JF, Gong Y, et al. Correlation of COL10A1 serum expression with recurrence and metastasis and long-term prognosis after radical gastrectomy of gastric cancer [J]. Practical Cancer Journal, 2022, 37(12): 2032-2035.
- [30] Pan S, Gao Q, Chen Q, et al. Integrative analysis-based identification and validation of a prognostic immune cell infiltration-based model for patients with advanced gastric cancer[J]. Int Immunopharmacol, 2021, 101(Part B): 108258.
- [31] 张亮, 王斌, 陈东风. 基于数据库分析 INHBA 基因在胃癌中的表达及临床意义[J]. 胃肠病学和肝病学杂志, 2021, 30(2): 133-139.
- Zhang L, Wang B, Chen DF. The expression and clinical significance of INHBA gene in gastric cancer were analyzed based on database[J]. Journal of Gastroenterology and Hepatology, 2021, 30(2): 133-139.
- [32] Su J, Ma Q, Long W, et al. LCTL is a prognostic biomarker and correlates with stromal and immune infiltration in gliomas[J]. Front Oncol, 2019, 9: 1083.
- [33] Luo D, Wang Y, Huan X, et al. Identification of a synonymous variant in TRIM59 gene for gastric cancer risk in a Chinese population[J]. Oncotarget, 2017, 8(7): 11507-11516.

(编辑:陈丽霞)