DOI:10.16016/j.2097-0927.202412079

基于转录组学和变分自编码器的癌症分期诊断研究

李佳芮,钱 力,沈俊杰,郭泓麟,秦茂洋,伍亚舟 陆军军医大学(第三军医大学)军事预防系医学系军队卫生统计学教研室,重庆

[摘要] **目的** 旨在对10种癌症转录组学数据开展深度分析与特征提取,进而实现对癌症样本的分期诊断。**方法** 在UCSC Xena网站收集发病率最高的10种癌症转录组学数据(包含4 938个样本和59 428个基因),以变分自编码器为基础,利用特征重要性排序思想,通过引入掩码算法和增量特征选择方法(incremental feature selection,IFS),构建了增量特征排序选优变分自编码器(incremental feature ranking and selection variational autoencoder,IFRSVAE);与随机森林(random forest,RF)、支持向量机(support vector machine,SVM)和极限梯度提升树(eXtreme gradient boosting,XGboost)结合测试了该方法的性能,并同其他方法进行比较。**结果** 提取了21个特征用于后续分类,较于传统变分自编码器、递归特征消除和Lasso回归模型,IFRSVAE模型在3种分类器上均取得了较好性能(AUC值最高,其余指标也表现良好);其中IFRSVAE-RF表现最好,AUC达到了85.49%(95%*CI*:83.24%~87.74%)。此外,还应用了沙普利加性解释(Shapley additive explanations, SHAP)模型对特征贡献进行展示。**结论** 本研究探索并验证了IFRSVAE在特征提取方面具备一定成效。基于此构建的IFRSVAE-RF模型,在癌症分期诊断任务中展现出较好的表现,为深度学习在癌症分期诊断方法的研究方向上,提供了一种可供参考的新思路。

[关键词] 癌症分期;转录组学;变分自编码器;机器学习
[中图法分类号] R195.1;R394.5;R730.49 [文献标志码] A

Cancer staging diagnosis based on transcriptomics and variational autoencoder

LI Jiarui, QIAN Li, SHEN Junjie, GUO Honglin, QIN Maoyang, WU Yazhou Department of Military Health Statistics, Faculty of Military Preventive Medicine, Army Medical University (Third Military Medical University), Chongqing, China

[**Abstract**] Objective To conduct an in-depth analysis and feature extraction of the transcriptomics data of 10 types of cancers in order to realize the staging diagnosis of cancer samples. Methods The transcriptomics data of the top 10 cancers having the highest incidence were amassed from the UCSC Xena website, which comprised 4 938 samples and 59 428 genes. With the aid of variational autoencoder, we developed an incremental feature ranking and selection variational autoencoder (IFRSVAE) based on feature importance ranking and incorporating the masking algorithm and the Incremental Feature Selection (IFS). Subsequently, the performance efficiency of our IFRSVAE model was evaluated in conjunction with Random Forest (RF), Support Vector Machine(SVM), and eXtreme Gradient Boosting (XGboost), and it was also

[基金项目] 国家自然科学基金面上项目(81872716, 82173621);2024年重庆市高等教育教学改革研究项目重点项目(242113)
[通信作者] 伍亚舟, E-mail: asiawu@tmmu.edu.cn.

2097-0927/Copyright © 2025, 陆军军医大学(第三军医大学)。这篇开放获取文章遵循 CC BY 许可协议 (https://creativecommons.org/licenses/by/4.0/)
引用格式:李佳芮,钱力,沈俊杰,等. 基于转录组学和变分自编码器的癌症分期诊断研究[J]. 陆军军医大学学报,2025,47(06):613-622. DOI:10.16016/j.2097-0927.202412079.
Citation:LI J R, QIAN L, SHEN J J, *et al*. Cancer staging diagnosis based on transcriptomics and variational autoencoder[J]. J Army Med Univ, 2025, 47(06):613-622. DOI:10.16016/j.2097-0927.202412079.

compared with other methods. **Results** Our research extracted 21 features for the ensuing classification. In comparison to the conventional variational autoencoder, recursive feature elimination, and Lasso regression models, the IFRSVAE model attained more favorable performance across all 3 classifiers (highest AUC value, and well performed other indicators). Notably, the IFRSVAE-RF exhibited the most outstanding performance, with an AUC value reaching 85.49% (95%*CI*: 83.24%~87.74%). Moreover, Shapley additive explanations (SHAP) interpretable model illustrated well contributions of the features in our model. **Conclusion** Our developed IFRSVAE shows certain effectiveness in feature extraction. The constructed IFRSVAE-RF model demonstrates relatively good performance in the task of cancer staging diagnosis, which providing a new and referable idea for research orientation of deep-learning-based diagnostic methods for cancer staging.

[**Key words**] cancer staging; transcriptomics; variational autoencoder; machine learning

Supported by the General Program of National Natural Science Foundation of China (81872716, 82173621) and the Key Research Project of Education Reform in Chongqing Colleges and Universities in 2024 (242113). Corresponding author: WU Yazhou, E-mail: asiawu@tmmu.edu.cn

当前在全球范围内癌症是过早死亡的主要原因[1-2]。除了癌症类型外,癌症分期是决定患者预后的最重要因素之一[3]。早期癌症可以通过手术切除,而中晚期癌症往往需要通过化疗或放疗手段进行治疗[4]。因此,准确有效的癌症分期可以帮助医生确定患者的最佳治疗方案,以提高患者的生存率和生活质量。近年来,随着高通量测序技术的发展,癌症基因组图谱计划(the cancer genome atlas,TCGA)使人们能够更好地了解癌症在发生、进展和耐药性方面的分子基础[5]。分子分型方案将会对传统的、有着悠久历史的解剖分期、组织学分型和分级等分型方式进行补充[6]。

如今,基于癌症组学数据的各种机器学习方法已被广泛应用于疾病的诊断和预后预测等任务。如CAI等[7]使用机器学习方法捕获无偏和紧凑的分子特征,对肺腺癌、小细胞肺癌和非小细胞肺癌进行分类。ATHIRA等[8]将机器学习方法与主成分分析和后验概率分析相结合,对乳腺癌进行分期预测。ALBARADEI等[9]提出卷积变分自编码器(convolutional variational autoencoder, CVAE)模型,用来区分原发性肿瘤和转移性肿瘤。但在利用转录组学数据和深度学习技术开展泛癌分期预测方面,相关研究鲜有报道。同时,现有模型在处理组学数据这类高维数据时,容易陷入维度灾难,难以有效提取关键特征且计算复杂度剧增,仍存在过拟合等缺陷[10]。

变分自编码器(variational autoencoder, VAE)[11]是2013年提出的一种深度学习模型,它可以将高维数据压缩到低维的潜在空间,实现数据的高效压缩,并且在这过程中学习到数据之间的潜在联系。将VAE应用到组学数据的处理上,不仅能够实现对高维数据的降维,同时也能很好地保留数据之间的相关性,从而不破坏数据的整体结构。

因此,本研究为解决癌症转录组学高维数据特征降维的问题,收集了来自10种癌症的转录组学数据,利用VAE作为基础模型,通过引入随机掩码算法和特征重要性排序思想对特征进行排序,然后应用增量特征选择方法(incremental feature selection,IFS)进行特征子集寻优。

1 材料与方法

1.1 数据收集与预处理

本研究数据来源于TCGA项目,为了使研究结果能适用于大多数癌症,研究选取了发病率最高的前10种癌症:肺癌(包括肺腺癌LUAD和肺鳞状细胞癌LUSC)、乳腺癌BRCA、结直肠癌(包括结肠癌COAD和直肠腺癌READ)、前列腺癌PRAD、胃癌(胃腺癌STAD)、肝癌(肝细胞肝癌LIHC)、胰腺癌PAAD、膀胱癌BLCA[12]。全文技术路线图如图1所示。

根据临床信息,将TNM分期为Ⅰ/Ⅱ期的样本划分为早期样本,其余划分为晚期样本[13]。在缺失值的处理上,本研究删除了缺失超过20%的特征。同样地,特征缺失超过20%的样本也被删除。经过去重和归一化处理后,将数据按照共有基因进行合并,最终有4 938个样本(其中,早期3 201例,晚期1 737例)和59 428个基因被纳入后续分析。整理后的数据如表1所示。

基于前文所述,数据整合之后每个样本包含59 428个基因信息。其中,许多基因可能在不同样本间表达差异不显著,或者是与研究目的无关。而对于深度学习模型,输入过多无关或冗余的数据可能会导致模型过拟合、训练时间延长以及模型性能下降等问题[14]。因而在将数据输入模型之前,对数据进行基因差异表达分析。差异分析使用R“limma”包进行,为了更好地适应数据的复杂性,本研究将差异基因的筛选阈值设定为动态阈值[15]。

表1　TCGA项目10种癌症样本数据

癌症名称	样本数		合计
	早期	晚期	
膀胱癌BLCA	136	290	426
乳腺癌BRCA	901	312	1 213
结肠癌COAD	288	212	500
肝细胞肝癌LIHC	289	103	392
肺腺癌LUAD	455	125	580
肺鳞状细胞癌LUSC	450	98	548
胰腺癌PAAD	172	8	180
结直肠腺癌READ	87	79	166
胃腺癌STAD	204	182	386
前列腺癌PRAD	219	328	547
合计	3 201	1 737	4 938

1.2　诊断模型构建

1.2.1　特征筛选模型　虽然传统VAE已经可以实现对高维数据的降维,但这种降维方式是对全局数据学习后生成新特征[11]。在实际运用中,仍需要对大量基因信息进行检测,这无疑加大了诊断的成本和难度,难以在临床上推行。因此,对传统VAE进行改进尤为重要。基于这种目的,本研究根据特征重要性排序的思想[16],结合随机掩码算法[17]和IFS方法,创新地提出了一个特征筛选思路,构建了增量特征排序选优变分自编码器(incremental feature ranking and selection variational autoencoder,IFRSVAE)。

本研究提出的IFRSVAE方法由特征排序和特征子集寻优两部分构成,主要框架如图2所示。其中,第一部分主要作用是根据重要性对特征进行排序,特征排序部分以VAE为主体,VAE具有强大的重构能力,通过学习数据的内在结构和特征可以得到低维的潜在表示,在关注特征信息的同时,还能够精准捕捉特征之间的关联[18];第二部分为确定最佳特征个数。通过这两部分即可筛选能代表重要基因信息的最佳特征子集。

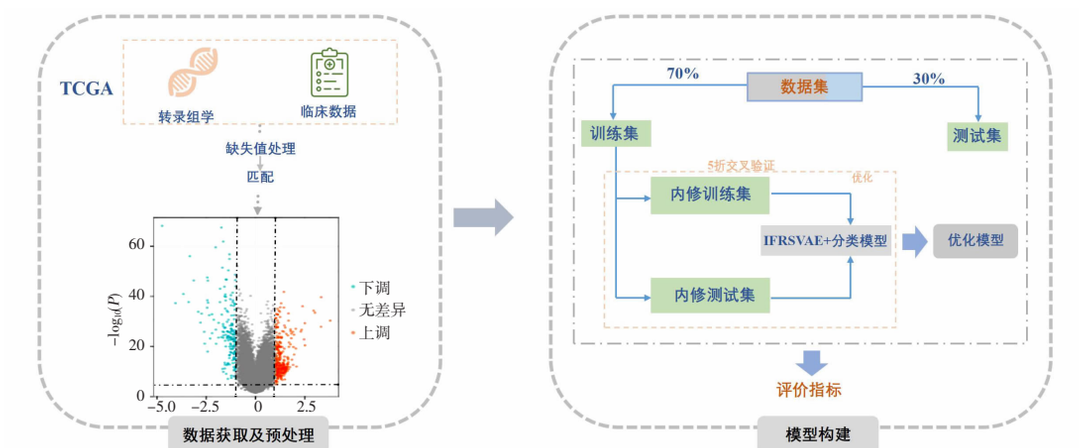


图1　基于转录组学的癌症分期模型构建研究技术路线图

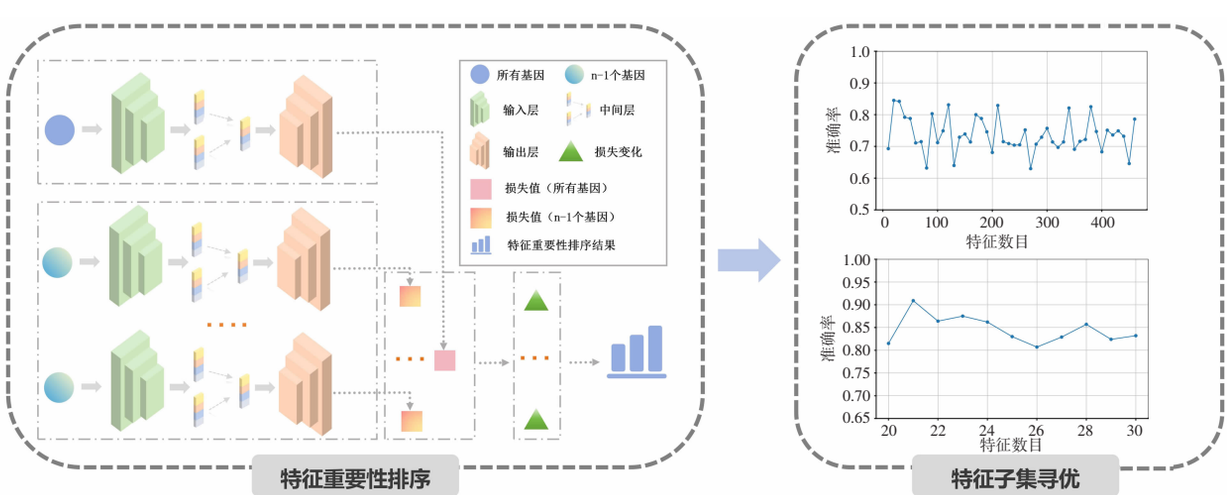


图2　IFRSVAE结构图

IFRSVAE模型特征提取步骤如下:①利用所有原始特征训练VAE模型,找到重构误差最小时的超参数,并记录此时的损失值和超参数;②在最佳超参数下多次训练VAE模型,利用随机掩码算法,在每次训练时随机遮蔽一个原始特征,记录下对应的重构误差,然后将其与第一步找到的重构误差进行比较,观察模型重构误差的变化,如果遮蔽某特征时模型重构误差变化显著,则说明这个特征对于全局特征的影响程度高,对形成潜在表示也具有更大贡献,即反复迭代地从原始特征中去除某一特征,通过潜在表示空间分布以及模型性能的变化来评估特征的重要性。通过这种方法可以为所有特征分配一个重要性评分,从而对所有特征进行排序;③将IFS方法[19]应用于特征排序列表,以提取特征子集,将特征子集应用于随机森林(random forest,RF)算法,以确定对癌症分期预测的最佳性能。

1.2.2 分类器与对比模型 分类器选用RF、支持向量机(support vector machine,SVM)和极限梯度提升树(eXtreme gradient boosting,XGboost)。作为二分类器,RF通过构建多个决策树并综合其结果,能够有效减少过拟合现象,可以处理高维数据和特征之间的复杂交互关系[20]。SVM通过寻找最大间隔超平面来实现高效分类,在小样本、高维数据情况下表现出色[21]。XGboost通过梯度提升的方式高效地组合多个弱分类器,可以自动处理缺失值,支持多种目标函数,具有较高的准确性和泛化能力[22]。

对比模型选用传统VAE、递归特征消除(recursive feature elimination,RFE)、Lasso回归模型。传统VAE能够有效学习数据的潜在分布,将高维转录组数据映射到低维空间[11]。RFE能够以一种迭代的方式评估特征的重要性,从众多基因特征中逐个剔除最不重要的特征,留下对分类或回归等任务最有价值的基因子集[23]。Lasso回归通过在损失函数中加入L1正则化项,使得部分基因特征的系数收缩为0,筛选出关键基因[24]。

1.3 模型评价指标

本研究采用准确率(accuracy)、召回率(recall)、精确率(precision)、F1分数(F1-score)、混淆矩阵(confusion matrix)作为模型评价指标[25],同时绘制ROC曲线对模型性能进行评估。模型性能评价指标计算公式如下:

$$准确率=\frac{N_{TP}+N_{TN}}{N_{TP}+N_{FN}+N_{TN}+N_{FP}}$$

$$召回率=\frac{N_{TP}}{N_{TP}+N_{FN}} \quad 精确率=\frac{N_{TP}}{N_{TP}+N_{FP}}$$

$$\text{F1 分数}=2\times\frac{精确率\times召回率}{精确率+召回率}$$

其中,N_{TP}和N_{TN}表示正确诊断为癌症晚期和癌症早期的例数,即:真正例和真负例,N_{FP}和N_{FN}表示错误诊断为癌症晚期和癌症早期的例数,即:假正例和假负例。

1.4 模型训练测试

按照7:3的比例划分训练集和测试集。在模型训练阶段,基于VAE框架对癌症转录组学数据进行建模,使用训练集全部数据对模型进行训练。将数据输入模型,通过前向传播计算预测结果,然后根据损失函数计算损失值。使用优化算法进行反向传播,更新模型参数,以减小损失值。重复这个过程,直到模型收敛。在模型优化阶段,对模型的超参数进行调整。超参数包括学习率、网络层数等。评估不同超参数组合下的模型性能,选择最佳的超参数组合。

在模型测试阶段,使用5折交叉验证,将数据集随机等分为5份,依次选取其中一份作为测试集,其余4份作为训练集,分别训练模型并评估性能,最终以5次评估结果的平均值作为模型整体性能的评价指标。

本研究采用Python软件的pytorch框架,过程中通过模拟不同的网络层数和学习率的不同组合,找出损失最小时的最优组合作为模型的参数,完成对模型的构建和训练。

2 结果

2.1 差异基因筛选结果

通过对基因的差异分析,共筛选出176个下调基因和290个上调基因,共466个相关基因。如图3所示,横坐标为差异表达倍数的对数值,反映了2个条件下基因表达水平的相对变化;纵坐标则为统计显著性的负对数,反映了基因差异表达的可信度。经过调节t检验,差异具有显著性。

2.2 特征筛选结果

将特征按重要性排序后,按照IFS方法,从特征个数=10开始,以10为间隔,形成n个(n=特征总数/10)特征子集。然后,将特征子集依次作为RF的输入,结果如图4所示。其中,横轴代表特征个数,纵轴则表示分类准确度。

由图4可知,当特征子集个数在20~30时,模型达

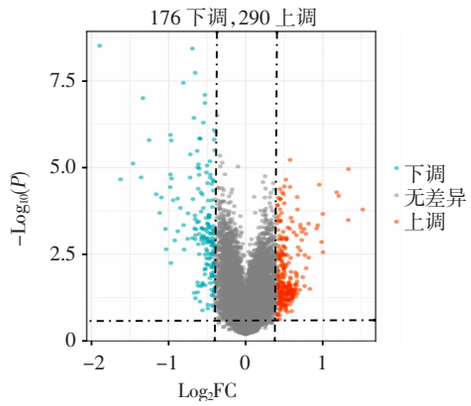


图3　基因差异分析结果

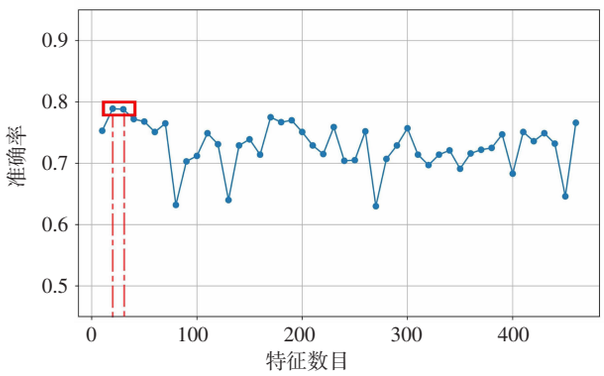


图4　第一次IFS曲线(以10为间隔)

到了最高点,且较为平缓。因此,本研究认为最优特征的可能范围在20~30之间。重复IFS方法,以1为间隔,进一步测试特征子集个数为20~30的模型性能。IFS曲线如图5所示。

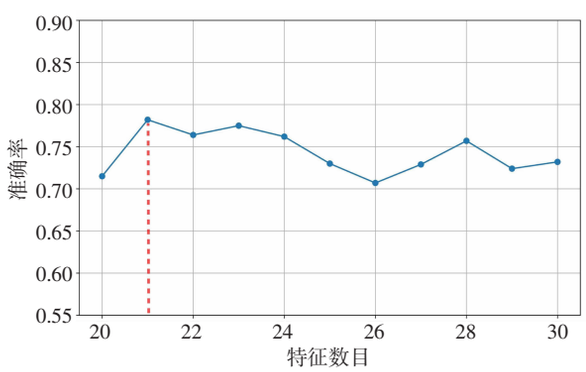


图5　第二次IFS曲线(以1为间隔)

可以看出,当使用特征列表中的前21个特征进行分类时,达到最高的准确率。因此,这些特征被认为是最优特征,构成了最优特征子集如表2所示。

2.3　分期诊断性能比较

为验证新构建的特征模型IFRSVAE对于特征的选择能力,本研究在构建的转录组学数据集(表1)上进行了比较研究,结果表明:相比于其他特征降维方法,本研究提出的IFRSVAE取得了最好的AUC值(F=12.360,P=0.002,差异具有统计学意义),模型搭配不同分类器(RF、XGboost、SVM)均能取得较为优异的诊断性能,各模型性能详见表3、图6。

表2　最优特征子集

排名	基因名称	重要性分数
1	*IGKC*	28 683.912
2	*HLA-DRB1*	26 261.019
3	*IGHA1*	24 792.539
4	*HLA-DRA*	24 048.370
5	*PPDPF*	22 929.586
6	*IGLC2*	21 232.875
7	*IGHG2*	20 493.354
8	*IGHG1*	19 861.620
9	*APOE*	19 265.340
10	*CD74*	19 146.810
11	*RPS15*	18 448.049
12	*LYZ*	17 422.336
13	*IGLC3*	17 256.354
14	*IGLV2-14*	16 704.484
15	*IGLV3-25*	16 685.056
16	*TPM2*	16 199.678
17	*NNMT*	15 906.380
18	*CD14*	15 825.326
19	*IGKV3-20*	15 602.514
20	*ATP1B1*	15 565.185
21	*CD52*	15 532.704

表3　各模型分期诊断性能展示(%)

特征选择	分类器	模型性能					F	P
		准确率	精确率	召回率	F1分数	AUC		
IFRSVAE	RF	78.54	79.59	78.79	79.19	85.49	12.360	0.002
	XGboost	76.51	76.39	75.00	75.68	83.22		
	SVM	78.33	77.57	80.29	78.90	84.33		
VAE	RF	75.16	75.68	73.53	74.57	82.33		
	XGboost	73.28	74.72	70.45	72.52	82.01		
	SVM	72.71	73.55	71.10	72.29	79.86		
RFE	RF	72.50	73.16	71.18	72.15	78.21		
	XGboost	74.17	75.69	71.28	73.41	80.21		
	SVM	70.68	71.65	68.44	69.85	77.85		
Lasso	RF	73.49	74.22	71.98	73.08	81.25		
	XGboost	73.12	74.34	70.62	72.34	81.16		
	SVM	74.11	75.49	71.49	73.44	82.36		

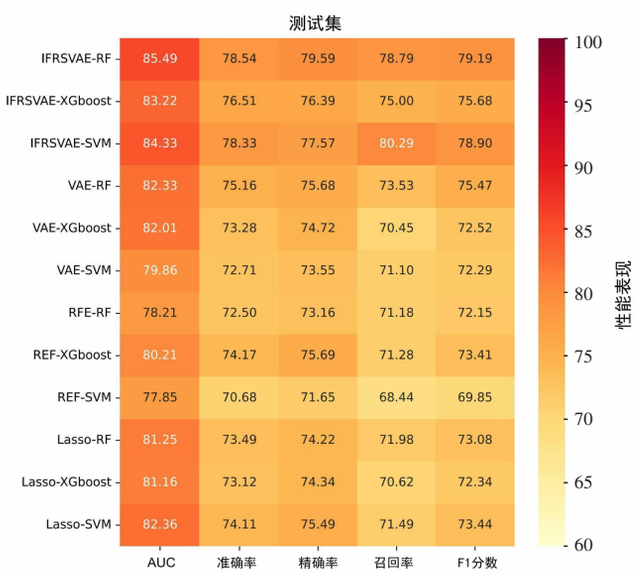


图6 各模型分期诊断性能热图

同时,为了进一步验证IFRSVAE-RF模型的分类效果优越性,本研究还绘制了ROC曲线以及混淆矩阵,4种特征选择方法在RF、XGboost、SVM上表现的混淆矩阵分别如图7、8、9所示。由混淆矩阵图可知,当IFRSVAE与RF结合时,得到了最好的效果,其正确识别晚期样本的能力较其他模型有明显的提升。ROC曲线如图10所示,ROC曲线下面积越大提示分类效果越好。

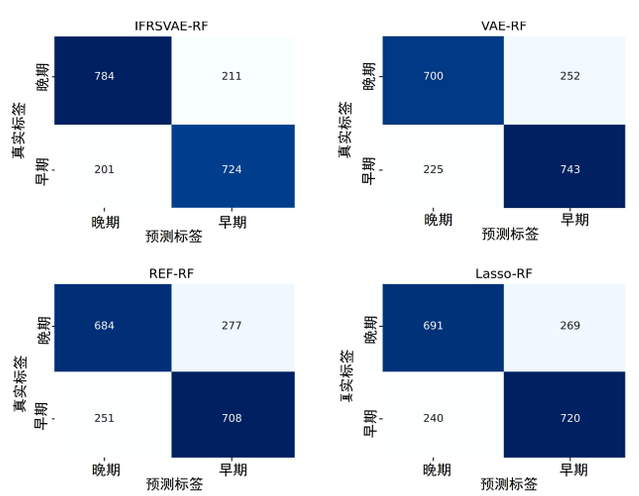


图7 4种特征选择方法在RF上的混淆矩阵

2.4 可解释性研究

前面已经证实了IFRSVAE-RF分类模型在癌症分期任务上的表现能力。但在实际应用中,充分理解模型所做的分类决策与模型的性能同样重要。因此,本研究在分类决策阶段将沙普利加性解释(Shapley additive explanations, SHAP)引入到表现最好的IFRSVAE-RF模型中,绘制的特征密度散点图如图11所示。该图展示了21个基因对模型输出分类决策的影响,直观地展示了每个特征的全局重要性。纵坐标每一行代表一个特征,横坐标为Shap值。按照Shap的平均绝对值对特征排序,一个点代表一个样本,宽的地方表示有大量的样本聚集,颜色越红说明特征本身数值越大,颜色越蓝说明特征本身数值越小。

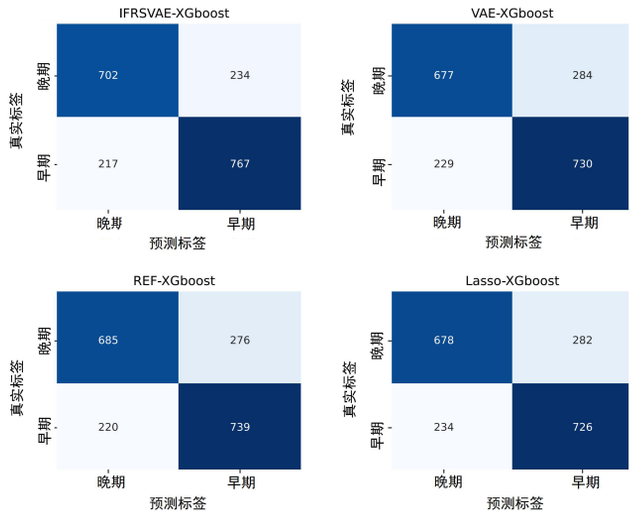


图8 4种特征选择方法在XGboost上的混淆矩阵

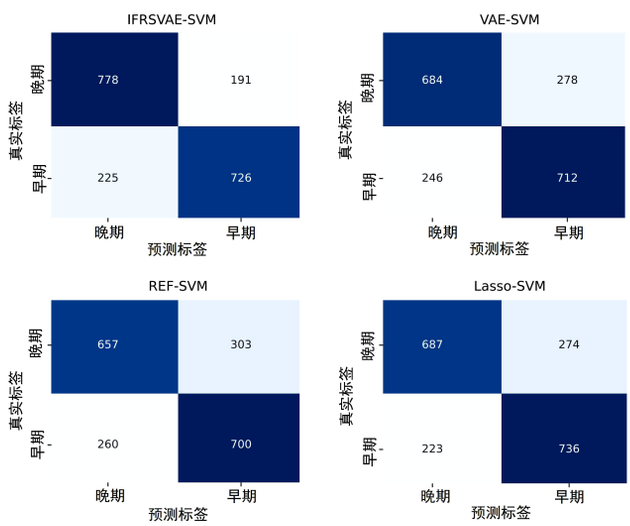


图9 4种特征选择方法在SVM上混淆矩阵

3 讨论

高维数据特征降维一直是组学数据研究的重点,在追求特征维度降低的同时,更应该关注低维特征对全局信息的代表能力。既往的深度学习等降维方法,多是利用计算机对于数据特征的学习能力来生成更

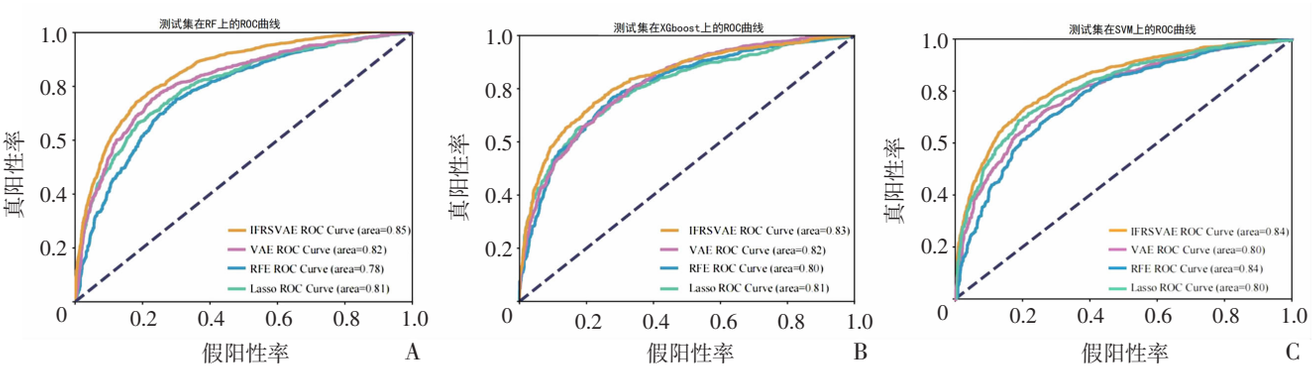


图10 4种特征选择方法在RF(A)、XGboost(B)和SVM(C)上的ROC曲线

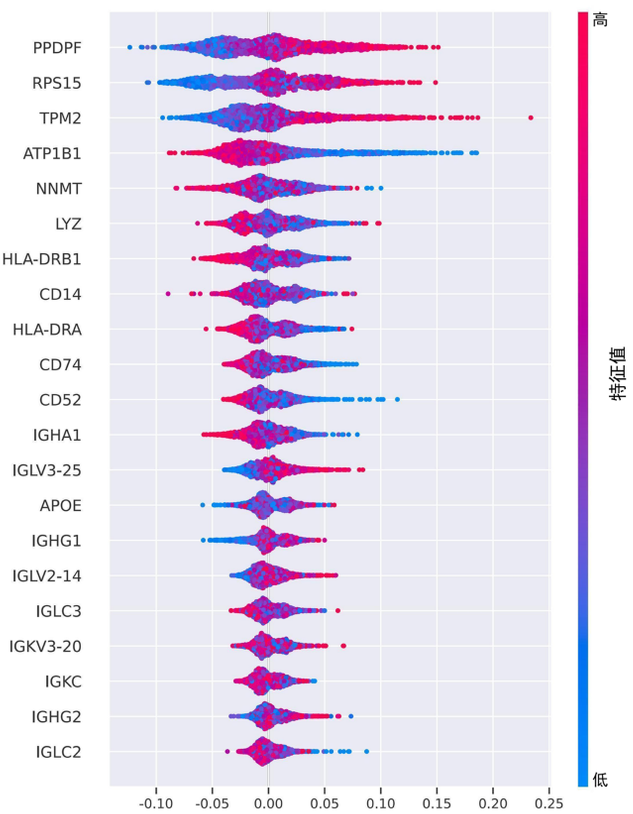


图11 SHAP值特征密度散点图

低维度的新特征[26],这些方法在应用中仍需要研究者先对所有特征进行检测,然后再输入模型进行预测,虽然能够达到较不错的模型性能,但是需要的成本大大提高,很难真正应用到实际。本文基于多种癌症转录组学数据,以VAE为主体,通过引入随机掩码算法和IFS方法,构建了IFRSVAE模型。此模型利用VAE更能够挖掘特征之间交互性这一优势,对全局信息进行分析,并根据特征对于全局信息的贡献性进行排序,IFS方法能够帮助研究者找到性能最优的特征子集。

通过IFRSVAE模型筛选出的特征在RF、XGboost、SVM上均取得了较好的性能,能够有效地识别癌症早期、晚期。在癌症治疗中,早期和晚期的治疗策略有很大差异。晚期癌症往往需要更激进的治疗方法[4],如联合化疗、靶向治疗或者免疫治疗等多种手段。IFRSVAE-RF模型,提高了对于晚期样本的发现能力。医生可以更精准、及时地制定更为合适的治疗方案,避免治疗不足以及贻误治疗时机的情况,对于临床治疗有着重要意义。

通过引入SHAP值,本研究把21个基因对分类决策影响大小进行了可视化展示,这些基因可能是癌症的危险因素或保护因素。既往研究证实,*PPDPF*、*RPS15*、*ATP1B1*、*NNMT*、*CD74*、*LYZ*、*IGHA1*、*CD14*、*IGLC3*、*IGHG2*、*IGKV3-20*[27-36]可促进肿瘤发生。而*CD52*、*IGKC*、*TPM2*、*HLA-DRA*、*HLA-DRB1*、*IGLV3-25*、*IGHG1*、*APOE*、*IGLV2-14*[36-43]等为抗肿瘤基因,其表达下调可促进肿瘤的发展,预后较差。需要说明的是,SHAP计算出的特征重要度排序与本研究模型得到的特征重要度排序不同,这是可以理解的:如上所述,IFRSVAE模型找到的特征子集可以最大程度地代表基因信息,这里的重要性是针对基因整体而言的;而SHAP是根据分类过程计算出来的,用于表示特征对于分类决策的重要性。SHAP值越高,分类器对分期决策就越重要[44]。

本模型的构建是基于多种癌症转录组学数据而开展的。由于不同癌种样本量存在差别,这一因素直接导致了所提取出的特征在稳定性以及代表性等方面可能会出现相应的差异。在后续深入研究阶段,团队将致力于单癌种数据分析与深度挖掘工作,运用更为先进的数据处理算法、优化特征筛选策略以及整合多维度的生物信息资源等一系列举措,进一步提升模型在癌症诊断、分期以及治疗指导等多方面应用场景

中的性能表现。

此外,本研究以TNM分期为基础来界定癌症的早晚期,将Ⅰ、Ⅱ期归为早期,Ⅲ、Ⅳ期划作晚期。这一划分方式与大众认知以及临床常规做法相契合。处于Ⅰ、Ⅱ期的肿瘤,通常局限在局部,患者预后相对乐观,多采用手术等局部治疗方式,整体治疗成效良好。而Ⅲ、Ⅳ期的肿瘤,往往已发生区域淋巴结的广泛转移,甚至出现远处转移,病情较为棘手,治疗方案更为复杂,预后也相对不理想。这种划分既符合临床实际状况,又契合对疾病严重程度的整体判断,方便研究成果与临床实践相互对照,促进成果转化应用。不过,这种分类方法存在一定的局限性,对于某些特定类型的癌症,可能存在更为精细或独特的分期体系,其早晚期的判定,除了TNM分期外,还可能与肿瘤的分子分型、基因特征等因素紧密相关。单纯依赖TNM分期划分早晚期,可能无法全面、精准地反映病情。在具体应用时,应结合所研究癌症的具体特性和相关临床信息,对研究结果进行综合分析与评估。在后续研究中,可以尝试运用更精细,或者结合其他因素的划分方法开展对比研究,进一步验证研究结果的可靠性与稳定性。

当然,本研究创建的分期诊断模型(IFRSVAE)也存在一定的局限性:①在数据方面,模型性能对数据质量要求较高,如果数据本身存在测序错误、样本污染或者数据缺失等情况,会影响模型对潜在特征的学习,进而干扰模型对特征交互性的挖掘和全局信息分析,导致排序和特征子集选择出现偏差;②由于模型多次迭代,计算量大,对设备要求高,运行时间相对较长;③模型需要在外部数据上进一步验证,以提高其泛化性能。

综上所述,本文基于对原始特征降维的思路构建了IFRSVAE模型,在癌症分期任务中,相较于传统方法展现出一定优势,具备较强的应用潜力。在癌症研究领域,样本的准确分类对高质量研究至关重要。IFRSVAE模型在癌症早期和晚期预测任务中表现良好,能够为后续的流行病学研究、癌症遗传学研究等提供相对准确的样本分组。这在一定程度上有助于提升研究结果的可靠性,降低因样本分类偏差而对研究造成的不良影响,为推动癌症研究水平的提升提供助力。

利益冲突 本文所有作者声明不存在利益冲突

参考文献:

[1] QI J L, LI M L, WANG L J, *et al*. National and subnational trends in cancer burden in China, 2005-20: an analysis of national mortality surveillance data [J]. Lancet Public Health, 2023, 8(12): e943-e955. DOI:10.1016/S2468-2667(23)00211-6.

[2] BRAY F, LAVERSANNE M, WEIDERPASS E, *et al*. The ever-increasing importance of cancer as a leading cause of premature death worldwide [J]. Cancer, 2021, 127(16): 3029-3030. DOI:10.1002/cncr.33587.

[3] MA B S, MENG F Y, YAN G, *et al*. Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data[J]. Comput Biol Med, 2020, 121: 103761. DOI:10.1016/j.compbiomed.2020.103761.

[4] KAUR R, BHARDWAJ A, GUPTA S. Cancer treatment therapies: traditional to modern approaches to combat cancers [J]. Mol Biol Rep, 2023, 50(11): 9663-9676. DOI: 10.1007/s11033-023-08809-3.

[5] AMIN M B, GREENE F L, EDGE S B, *et al*. The eighth edition AJCC cancer staging manual: continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging[J]. CA Cancer J Clin, 2017, 67(2): 93-99. DOI:10.3322/caac.21388.

[6] DONG Y Y, YANG W K, WANG J W, *et al*. MLW-gcForest: a multi-weighted gcForest model towards the staging of lung adenocarcinoma based on multi-modal genetic data [J]. BMC Bioinformatics, 2019, 20(1): 578. DOI: 10.1186/s12859-019-3172-z.

[7] CAI Z H, XU D, ZHANG Q, *et al*. Classification of lung cancer using ensemble-based feature selection and machine learning methods[J]. Mol Biosyst, 2015, 11(3): 791-800. DOI:10.1039/c4mb00659c.

[8] ATHIRA K, GOPAKUMAR G. Breast cancer stage prediction: a computational approach guided by transcrip-tome analysis [J]. Mol Genet Genomics, 2022, 297(6): 1467-1479. DOI:10.1007/s00438-022-01932-z.

[9] ALBARADEI S, NAPOLITANO F, THAFAR M A, *et al*. *Meta* Cancer: a deep learning-based pan-cancer metastasis prediction model developed using multi-omics data [J]. Comput Struct Biotechnol J, 2021, 19: 4404-4411. DOI: 10.1016/j.csbj.2021.08.006.

[10] MARTINEZ-GARCIA M, OLMOS P M. Handling ill-conditioned omics data with deep probabilistic models[J]. IEEE J Biomed Health Inform, 2023, 27(9): 4601-4610. DOI:10.1109/JBHI.2023.3279493.

[11] KINGMA D P, WELLING M. Auto-encoding variational Bayes [EB/OL]. 2013: 1312.6114. https://arxiv.org/abs/1312.6114v11.

[12] SUNG H, FERLAY J, SIEGEL R L, *et al*. Global cancer statistics 2020: GLOBOCAN estimates of incidence and

mortality worldwide for 36 cancers in 185 countries[J]. CA Cancer J Clin, 2021, 71(3): 209-249. DOI: 10.3322/caac.21660.

[13] MA B S, CHAI B J, DONG H, *et al*. Diagnostic classification of cancers using DNA methylation of paracancerous tissues[J]. Sci Rep, 2022, 12(1): 10646. DOI:10.1038/s41598-022-14786-7.

[14] BARBIERI M, LEE P K, BRIZI L, et al. Circumventing the curse of dimensionality in magnetic resonance fingerprinting through a deep learning approach[J]. NMR Biomed, 2022, 35(4): e4670. DOI: 10.1002/nbm.4670.

[15] LIU S Y, WANG Z T, ZHU R H, *et al*. Three differential expression analysis methods for RNA sequencing: limma, EdgeR, DESeq2[J]. J Vis Exp, 2021(175): (175). DOI: 10.3791/62528.

[16] REMESEIRO B, BOLON-CANEDO V. A review of feature selection methods in medical applications[J]. Comput Biol Med, 2019, 112: 103375. DOI: 10.1016/j.compbiomed.2019.103375.

[17] SHAHIDI L K, COLLINS L M, MAINSAH B O. Parameter tuning of time-frequency masking algorithms for reverberant artifact removal within the cochlear implant stimulus[J]. Cochlear Implants Int, 2022, 23(6): 309-316. DOI: 10.1080/14670100.2022.2096182.

[18] CAO S C, LI J J, NELSON K P, *et al*. Coupled VAE: improved accuracy and robustness of a variational autoencoder[J]. Entropy (Basel), 2022, 24(3): 423. DOI: 10.3390/e24030423.

[19] ZHANG Y H, HUANG T, CHEN L, *et al*. Identifying and analyzing different cancer subtypes using RNA-seq data of blood platelets[J]. Oncotarget, 2017, 8(50): 87494-87511. DOI:10.18632/oncotarget.20903.

[20] HU J C, SZYMCZAK S. A review on longitudinal data analysis with random forest[J]. Brief Bioinform, 2023, 24(2): bbad002. DOI:10.1093/bib/bbad002.

[21] HUANG S J, CAI N G, PACHECO P P, et al. Applications of support vector machine (SVM) learning in cancer genomics[J]. Cancer Genomics Proteomics, 2018, 15(1): 41-51. DOI: 10.21873/cgp.20063.

[22] CHEN S J, ZHOU W J, TU J H, *et al*. A novel XGBoost method to infer the primary lesion of 20 solid tumor types from gene expression data[J]. Front Genet, 2021, 12: 632761. DOI:10.3389/fgene.2021.632761.

[23] BENJAMIN K J M, KATIPALLI T, PAQUOLA A C M. dRFEtools: dynamic recursive feature elimination for omics[J]. Bioinformatics, 2023, 39(8): btad513. DOI: 10.1093/bioinformatics/btad513.

[24] DAI P P, CHANG W F, XIN Z R, *et al*. Retrospective study on the influencing factors and prediction of hospitalization expenses for chronic renal failure in China based on random forest and LASSO regression[J]. Front Public Health, 2021, 9: 678276. DOI: 10.3389/fpubh.2021.678276.

[25] ALAWAD M, ALJOUIE A, ALAMRI S, *et al*. Machine learning and deep learning techniques for optic disc and cup segmentation—A review[J]. Clin Ophthalmol, 2022, 16: 747-764. DOI:10.2147/OPTH.S348479.

[26] ZANG Z L, CHENG S H, XIA H C, *et al*. DMT-EV: an explainable deep network for dimension reduction[J]. IEEE Trans Vis Comput Graph, 2024, 30(3): 1710-1727. DOI: 10.1109/TVCG.2022.3223399.

[27] ZHENG Q W, NI Q Z, ZHU B, *et al*. PPDPF promotes lung adenocarcinoma progression *via* inhibiting apoptosis and NK cell-mediated cytotoxicity through STAT3[J]. Oncogene, 2022, 41(36): 4244-4256. DOI: 10.1038/s41388-022-02418-3.

[28] LIN B H, HUANG G, YUAN Z S, *et al*. RPS15 coordinates with CtIP to facilitate homologous recombination and confer therapeutic resistance in breast cancer[J]. Radiat Res, 2024, 202(5): 775-784. DOI:10.1667/RADE-24-00134.1.

[29] SHI J L, FU L, ANG Q, *et al*. Overexpression of ATP1B1 predicts an adverse prognosis in cytogenetically normal acute myeloid leukemia[J]. Oncotarget, 2016, 7(3): 2585-2595. DOI:10.18632/oncotarget.6226.

[30] ECKERT M A, COSCIA F, CHRYPLEWICZ A, *et al*. Proteomics reveals NNMT as a master metabolic regulator of cancer-associated fibroblasts[J]. Nature, 2019, 569(7758): 723-728. DOI:10.1038/s41586-019-1173-8.

[31] HONG W C, LEE D E, KANG H W, *et al*. CD74 promotes a pro-inflammatory tumor microenvironment by inducing S100A8 and S100A9 secretion in pancreatic cancer[J]. Int J Mol Sci, 2023, 24(16): 12993. DOI: 10.3390/ijms241612993.

[32] GU Z W, WANG L, DONG Q, *et al*. Aberrant LYZ expression in tumor cells serves as the potential biomarker and target for HCC and promotes tumor progression *via* csGRP78[J]. Proc Natl Acad Sci USA, 2023, 120(29): e2215744120. DOI:10.1073/pnas.2215744120.

[33] LI C, GUAN R, LI W M, *et al*. Single-cell RNA sequencing reveals tumor immune microenvironment in human hypopharygeal squamous cell carcinoma and lymphatic metastasis[J]. Front Immunol, 2023, 14: 1168191. DOI:10.3389/fimmu.2023.1168191.

[34] WU Z H, ZHANG Z X, LEI Z H, *et al*. CD14: biology and role in the pathogenesis of disease[J]. Cytokine Growth Factor Rev, 2019, 48: 24-31. DOI: 10.1016/j.

cytogfr.2019.06.003.

[35] WANG J P, HUANG J N, DING H, *et al*. Functional analysis of tumor-derived immunoglobulin lambda and its interacting proteins in cervical cancer[J]. BMC Cancer, 2023, 23(1): 929. DOI:10.1186/s12885-023-11426-9.

[36] FARROKHI YEKTA R, AREFI OSKOUIE A, REZAEI TAVIRANI M, *et al*. Decreased apolipoprotein A4 and increased complement component 3 as potential markers for papillary thyroid carcinoma: a proteomic study[J]. Int J Biol Markers, 2018, 33(4): 455-462. DOI: 10.1177/1724600818787752.

[37] MA Y F, CHEN Y C, FANG D L, *et al*. The immune-related gene CD52 is a favorable biomarker for breast cancer prognosis[J]. Gland Surg, 2021, 10(2): 780-798. DOI: 10.21037/gs-20-922.

[38] WU Z L, GE L Y, MA L L, *et al*. TPM2 attenuates progression of prostate cancer by blocking PDLIM7-mediated nuclear translocation of YAP1[J]. Cell Biosci, 2023, 13(1): 39. DOI:10.1186/s13578-023-00993-w.

[39] PIAO X M, KANG H W, JEONG P, *et al*. A prognostic immune predictor, HLA-DRA, plays diverse roles in non-muscle invasive and muscle invasive bladder cancer[J]. Urol Oncol, 2021, 39(4): 237.e21-237.237.e29. DOI: 10.1016/j.urolonc.2020.11.017.

[40] KAMIZA A B, KAMIZA S, MATHEW C G. HLA-DRB1 alleles and cervical cancer: a meta-analysis of 36 case-control studies[J]. Cancer Epidemiol, 2020, 67: 101748. DOI:10.1016/j.canep.2020.101748.

[41] LI Y X, WANG P, YE D M, *et al*. IGHG1 induces EMT in gastric cancer cells by regulating TGF-β/SMAD3 signaling pathway[J]. J Cancer, 2021, 12(12): 3458-3467. DOI: 10.7150/jca.56056.

[42] TAVAZOIE M F, POLLACK I, TANQUECO R, *et al*. LXR/ApoE activation restricts innate immune suppression in cancer[J]. Cell, 2018, 172(4): 825-840.e18. DOI: 10.1016/j.cell.2017.12.026.

[43] BERGHAUS N, SCHREINER S, POOS A M, *et al*. Comparison of IGLV2-14 light chain sequences of patients with AL amyloidosis or multiple myeloma[J]. FEBS J, 2023, 290(17): 4256-4267. DOI:10.1111/febs.16805.

[44] ABUZINADAH N, KUMAR POSA S, ALARFAJ A A, *et al*. Improved prediction of ovarian cancer using ensemble classifier and shaply explainable AI[J]. Cancers (Basel), 2023, 15(24): 5793. DOI: 10.3390/cancers15245793.

(**收稿**:2024-12-16;**修回**:2025-01-29)

(编辑 邓强庭)