肺癌基因表达数据的 Stacking 集成学习法分析

郝智航[1]，胡广才[1]，倪士峰[3]，王乐[2]，王欢[1*]
1.宝鸡文理学院计算机学院，陕西 宝鸡 721000
2.宝鸡文理学院化学与材料工程学院，陕西 宝鸡 721013
3.西北大学生命科学学院，陕西 西安 710069
DOI:10.61369/EIR.2025100010

摘　　要： 本研究针对肺癌基因表达数据高维度、小样本及标注噪声对传统单模型的挑战，提出一种增强型 Stacking 集成学习框架，提升分类性能与鲁棒性。以 GSE252168 数据为基础，首先通过混合特征选择策略，将基因维度由 30715 降至 1500；继而集成 SVM、逻辑回归、随机森林与 XGBoost 作为异构基学习器，其核心创新在于双重增强机制：一方面将基学习器生成的元特征与原特征拼接以构建元层输入，另一方面在推理时采用基于验证集 F1 与 AUC 的动态权重自适应融合基模型输出，元学习器以 L1 正则化逻辑回归。为评估鲁棒性，训练时注入 8% 标签噪声。实验结果表明，该框架在测试集上获得 F1=0.9162、AUC=0.9752、准确率高达 96.06%，显著优于最佳单模型；本研究有效解决了高维基因数据分类难题，为肺癌精准诊断提供了可靠的技术支撑。

关 键 词： 肺癌数据分析；Stacking 集成学习；机器学习；动态权重；精准诊断

Analysis of Lung Cancer Gene Expression Data Using a Stacking Ensemble Learning Approach

Hao Zhihang[1], Hu Guangcai[1], Ni Shifeng[3], Wang Le[2], Wang Huan[1*]
1.School of Computer Science, Baoji University of Arts and Sciences, Baoji, Shaanxi 721000
2.School of Chemistry and Materials Engineering, Baoji University of Arts and Sciences, Baoji, Shaanxi 721013
3.School of Life Sciences, Northwest University, Xi'an, Shaanxi 710069

Abstract: To address the challenges of high dimensionality, small sample size, and label noise in lung cancer gene expression data for traditional single-model approaches, this study proposes an enhanced Stacking ensemble learning framework to improve classification performance and robustness. Based on the GSE252168 dataset, a hybrid feature selection strategy was first applied to reduce the gene dimensionality from 30,715 to 1,500. Subsequently, SVM, logistic regression, random forest, and XGBoost were integrated as heterogeneous base learners. The core innovation lies in a dual-enhancement mechanism: on one hand, the meta-features generated by the base learners were concatenated with the original features to construct the meta-layer input; on the other hand, during inference, a dynamic weighting strategy based on validation set F1 and AUC was adopted to adaptively fuse the outputs of the base models. The meta-learner utilized L1-regularized logistic regression. To evaluate robustness, 8% label noise was introduced during training. Experimental results demonstrate that the proposed framework achieved an F1-score of 0.9162, AUC of 0.9752, and an accuracy of up to 96.06% on the test set, significantly outperforming the best single model. This study effectively addresses the challenges of high-dimensional gene data classification and provides reliable technical support for the precise diagnosis of lung cancer.

Keywords: lung cancer data analysis; Stacking ensemble learning; machine learning; dynamic weighting; precision diagnosis

引言

尽管现代医学在肺癌治疗方面已经取得了显著进展，尤其是在靶向治疗和免疫治疗领域，但早期诊断、治疗成本以及复发和转移等

项目信息：陕国家自然科学基金青年项目 (82104682)；陕西省科技厅项目 (2025JC-YBQN-1249, 2015JM8463)；陕西省科技创新团队 (2022TD-63)；宝鸡文理学院校级研究生创新科研项目 (202496027017)。
作者简介：郝智航（1999.12-），男，汉族，陕西西安人，学历：硕士研究生，职称：学生，研究方向：数据挖掘。
通讯作者：王欢（1981.05-），女，汉族，陕西宝鸡人，学历：博士研究生，职称：教授，研究方向：数据挖掘族。

方面仍然存在诸多挑战。这些局限性强调了需要进一步的基础研究、创新治疗方案以及更加普及的筛查手段。随着高通量测序技术的普及，RNA-seq已成为疾病分子诊断的重要工具[1,2]。这类技术能够一次性获取数万个基因的表达信息，为疾病分型和预后评估提供了丰富的数据资源。统计学与机器学习的交叉应用正在改变生物医学数据的处理模式[3]，但高维数据带来的"维数灾难"问题同样不容忽视——当特征数量远超样本规模时，模型容易出现过拟合和泛化能力下降[4]。

集成学习通过组合多个学习器的预测结果来提升整体性能。Stacking是其中较为复杂的策略，其核心思想是用元学习器对基础模型的输出进行二次整合[5]。这种方法在遗传学和基因组学领域已有广泛探索[6]。例如，支持向量机曾被用于微阵列数据的癌症分类，并在多个数据集上验证了其有效性[7,8]。不过，单一分类器往往只能捕捉数据的某一侧面。杜冲等人的工作表明，在特征选择阶段引入集成策略可以获得比单一方法更稳健的结果[9]。近年来，深度神经网络也开始应用于细胞功能建模，Ma及其合作者利用多层网络结构刻画了细胞的层次化组织方式[10]。

Stacking方法在肿瘤诊断中已有成功案例。Sumon等人使用代谢组学数据预测小细胞肺癌，其构建的Stacking模型优于传统分类器[11]。Ganie等人则将SHAP解释技术引入多癌种分类任务，使模型的决策过程更加透明[12]。此外，Naderalvojoud等人在分析医疗观察数据时发现，集成学习能够显著改善预测准确率[13]。Chicco等人针对泛癌症预后开展的研究提出，尽管集成方法性能优异，但其生物学可解释性仍需加强[14]。 国内学者在相关领域同样做出了有价值的探索。李泉伦团队将Stacking应用于近红外光谱分析，证明了该框架在跨领域任务中的适应性[15]。支持向量机的理论基础和应用范式已有系统性综述[16]。决策树类算法（如XGBoost）因其可解释性和高效性成为集成学习的常用组件[17,18]。Mahmoud与Takaoka最近发表的工作值得特别关注，他们针对肝癌基因表达数据设计了特征选择与Stacking相结合的诊断流程，取得了较高的分类精度[19]。

国内机器学习在生物医学中的应用也呈现多样化趋势。王笑针对单细胞RNA测序数据开发了细胞类型自动识别算法[20]。钟晨露等人将机器学习与基因型数据结合，用于指导华法林的个体化剂量调整[21]。吴继明的研究整合了基因组、转录组等多层次信息，实现了对癌症类型及分期的联合预测[22]。郭依晨构建的生存分析模型专门面向女性恶性肿瘤，为临床决策提供了量化依据[23]。杨晨雨团队则从药物敏感性角度出发，利用多组学数据预测肿瘤对化疗药物的响应[24]。 尽管已有诸多研究，基因表达数据分析仍存在三方面问题。第一，特征维度通常在数千至数万量级，而样本量往往只有数百例，这种不平衡导致模型训练不稳定。第二，基因之间存在复杂的调控关系和共表达模式，线性模型难以充分描述这类非线性结构。第三，现有文献多侧重于报告最终分类精度，对特征选择策略、基学习器配置、模型融合方案等环节缺乏系统性对比，且很少进行严格的统计检验来验证性能提升的显著性。

针对上述的问题，本研究首先利用方差过滤、统计检验、互信息的高维特征选择等方法分析高通量基因表达(GEO)数据库中肺癌的RNA转录组测序数据，然后对各个基准的学习器在堆叠方法下的性能进行了详细的阐述，并对堆叠方法进行了统计学验证。最后本研究还给出了一些可视化的工具例如重要性分析、性能比较和生物解释。本研究的结果可以作为一种对基因表达数据分类的强大且可靠的工具，为生物信息学领域的嵌入集成学习提供一些新见解。

一、研究数据

本研究采用了基因表达综合数据库(GEOhttps://www.ncbi.nlm.nih.gov/geo/)中公开的肺癌数据集GSE252168作为数据来源，其中包含186个标签为0的健康对照组样本和117个标签为1的肺癌疾病组病人，每个生物样本由代表每个基因的30715个值组成的基因谱。原始数据以基因－样本矩阵的形式存在，以行为基因表达的特征列为样本，利用方差过滤(0.05)和t检验特征选择法（保留前1500个显著基因）对数据进行预处理，用于降低高维数据的复杂性，同时采用8%的标签噪声来模拟真实场景中的不确定性。最后，使用标准化后的数据集来训练改进的Stacking集成模型，验证了方法的有效性和泛化能力。

二、Stacking算法与原理

（一）集成学习框架概述

同时，本研究基于Stacking集成学习理论[18]，提出一种以异构基学习器SVM、LR、随机森林、XGBoost的预测联合进行联合预测的、以L1正则化LR为元学习器的、用于二阶学习的、用于高维基因数据泛化的增强多层集成学习框架。

（二）基学习器设计与优化

1.支持向量机(SVM)

支持向量机是一种基于统计学习有监督的学习模型，核心是超平面的高维空间数据分类方法[19]。因而在本研究中，使用RBF[20]作为其内核函数，并对正则化系数C=1.0和类别权重平衡等参数进行网格搜索优化，以实现少类样本的高敏感度。其目标函数如下示：

$$\min_{w,b,\xi} \frac{1}{2}\| w \|^2 + C\sum_{i=1}^{n}\xi_i \quad (1)$$

约束条件：

$$y_i\left(w^T\phi\left(x_i\right)+b\right)\geq 1-\xi_i,\quad \xi_i \geq 0 \quad (2)$$

其中$\phi(x_i)$为核函数映射，ξ_i为松弛变量。

2.逻辑回归 (Logistic Regression, LR)

逻辑回归是将线性回归结果通过 Sigmoid 函数，映射到概率空间，解决的是二分类的问题。由于基因数据具有高高阶的特点，通过 L1 正则化 (C=0.1)[21]，筛选关键基因标记。模型输出概率为：

$$P(y=1|x)=\frac{1}{1+\exp(-(\beta_0+\sum_{j=1}^{p}\beta_j x_j))} \quad (3)$$

L1 正则化目标函数：

$$J(\beta)=-\frac{1}{n}\sum_{i=1}^{n}[y_i\log(p_i)+(1-y_i)\log(1-p_i)] +\lambda\sum_{j=1}^{p}|\beta_j| \quad (4)$$

3.随机森林 (Random Forest, RF)

随机森林是通过自助 (Bootstrap) 建立多个树，以多数票决定预测结果[22]。本研究通过网格搜索确定最优参数：n_estimators=200，max_depth=5，min_samples_split=2，class_weight=balanced。基尼不纯度作为节点分裂准则：

$$Gini(D)=1-\sum_{k=1}^{K}\left(\frac{|D_k|}{|D|}\right)^2 \quad (5)$$

其中 D_k 为节点中第 k 类样本的比例。

4.XGBoost(极端梯度提升树)

XGBoost 通过迭代生成决策树，优化目标函数中的正则化项以防止过拟合[23]。目标函数定义为：

$$L^{(t)}=\sum_{i=1}^{n}l\left(y_i,\hat{y}_i^{(t-1)}+f_t(x_i)\right)+\Omega(f_t) \quad (6)$$

正则化项定义为：

$$\Omega(f)=\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}w_j^2 \quad (7)$$

其中 T 为叶节点数，w_j 为叶节点权重，γ 和 λ 为正则化系数。

（三）增强型 Stacking 集成框架

1.传统 Stacking 方法的局限性

传统的 Stacking 模型在处理高维基因数据时存在三个核心局限：

(1) 权重分配固化：所有基学习器的预测结果被平等对待，忽略了模型性能差异，导致优秀模型的贡献被弱化；

(2) 鲁棒性不足：元特征直接表示原始预测，没有考虑标注噪声，难以适应真实的临床环境；

(3) 特征冗余问题：对高维数据缺乏针对性的特征筛选策略，冗余特征显著增加计算成本并可能导致过拟合。

2.Stacking 框架设计

本研究提出了一种增强型 Stacking 集成学习框架，如图 1 所示

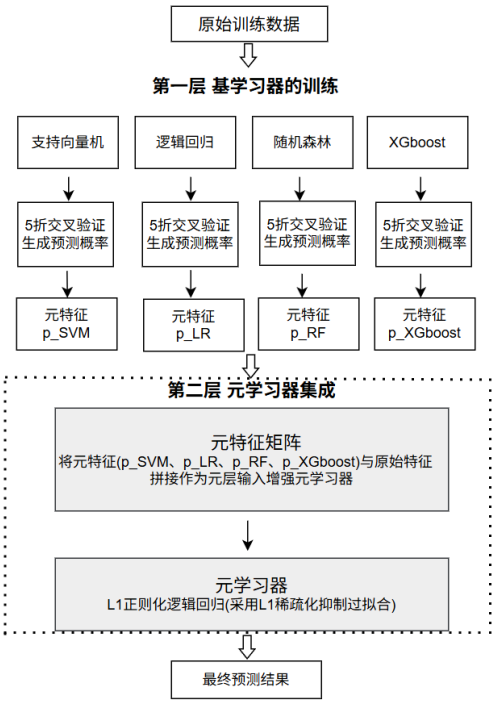

图 1 Stacking 集成学习框架

3.技术创新

创新一：增加特征选择缩减路径采用三阶段特征选择策略：第一阶段利用方差进行低变异基因特征筛选（阈值 =0.05），保留变基因的 7,858 个特征从 30,715 个初始特征；第二阶段进行 FDR 矫正检验组间显著性差异的基因特征；第三阶段利用互信息计算最具有鉴别力的基因特征，按从大到小前 1,500 个特征映射并表征。

创新二：噪声注入控制训练过程中随机选取 8%(24 例) 的训练集标签进行翻转 (0 变为 1 或 1 变为 0) 作为噪声注入数据集，用于模拟真实诊断中存在的不确定性，并施加一定强度的正则化避免模型过度拟合。

创新三：异构基学习器协同优化精心挑选 SVM、逻辑回归、随机森林、XGBoost 算法完备的线性 / 非线性、参数 / 非参数算法空间上的基学习器，并依托于对基学习器 5 折分层交叉验证训练达到预测无偏差和类别齐一的目的。

创新四：通过动态权重分配突破传统等权重融合的局限，提出基于验证集性能表现的自适应权重分配机制：

$$w_i=0.5\times F1_i+0.5\times AUC_i \quad (8)$$

权重归一化约束：

$$w'_i=\frac{w_i}{\sum_{j=1}^{m}w_j},\quad 满足\sum_{i=1}^{m}w'_i=1 \quad (9)$$

其中 w'$_i$ 表示第 i 个基模型的归一化权重，$F1_i$ 表示第 i 个基模型在验证集上的 F1 分数 (调和平均数，平衡精确率与召回率)，AUC_i 表示第 i 个基模型在验证集上的 AUC 值 (评估类别不平衡鲁棒性)，m 表示基模型数量 (本研究 m=4)。

该权重分配策略具有三重优势：

(1) 性能导向：优秀模型获得更高权重，确保集成效果最

优化；

(2)指标平衡：0.5系数实现F1和AUC的均衡，兼顾精度和鲁棒性；

(3)自适应优化：根据验证表现动态调整，替代传统固定权重策略；

3.多源信息融合与元学习器设计

采取特征融合的策略，将基学习器给出的预测概率与原始基因特征经过选择后的基因特征进行融合。

$$Z=\left[w_1P_1,w_2P_2,w_3P_3,w_4P_4,X_{selected}\right] \quad (10)$$

其中，P_i表示i的预测概率，w_i表示相应动态权重，$X_{selected}$代表经过过滤之后的1,500个重要基因特征，既保留了预测信息，又保留了预测结果背后的解释。

用L1正则化逻辑回归作为元学习器，实现特征选择与分类预测的联合优化：

$$\hat{\beta}=\arg\min_{\beta}\frac{1}{n}\sum_{i=1}^{n}\log(1+\exp(-y_i\beta^Tz_i))+\lambda||\beta||_1 \quad (11)$$

L1正则化项自动进行特征权重稀疏化，通过交叉验证确定最优正则化强度λ，在模型复杂度与泛化能力间实现最优平衡。

三、实验结果分析

（一）实验设置与评估方案

采用肺癌基因表达数据集GSE252168综合评估提出的改进Stacking集成框架。在该数据中共有303个生物样本，均取自人体中的癌症和正常组织，其中包括186个对照组样本和117个患病样本。分别对样本和所表达的基因进行训练和测试。30,715个基因用来做分类。将70%的样本作为训练集，30%样本作为测试集。采用五次交叉折叠(5折法)来确保结果的统计显著性，并且进行30次重复实验。采用五个性能指标，包括准确率，精确率，召回率，F1以及AUC来评估模型在不平衡生物数据集的表现。为了模拟真实生物标记的不确定性，对训练数据添加了标签噪声(8%的样本)。

（二）基学习器性能评估

1.单一模型分类性能对比

表1 基学习器性能对比

模型	准确率	精确率	召回率	F1分数
SVM	0.8681	0.9032	0.7568	0.8235
逻辑回归	0.8791	0.9062	0.7838	0.8406
随机森林	0.8791	0.9062	0.7838	0.8406
XGBoost	0.8681	0.8788	0.7838	0.8286

四种异构基学习器在数据集GSE252168肺癌数据上的评价结果不尽一致，其中逻辑回归与随机森林算法评价结果最好，二者的评价指数都是100%，即准确率87.91%、精确率90.62%、召回率78.38%、F1得分84.06%，表明混合特征选择优化之后的1500维基因特征空间对于线性判别与集成树具有相同的边界性质；SVM的径向基核函数映射评价精确率最高90.32%，召回率低75.68%，准确率86.81%，F1得分82.35%，判别结果是高精度低敏感；XGBoost的评价结果指数均衡，准确率86.81%、精确率87.88%、召回率78.38%、F1得分82.86%。从总体上看所有的基学习器准确率都超过了87%，但在召回率方面都位于75%–78%范围内，算法之间的性能差距为Stacking集成提供了较好的多样性。

表2其他算法模型性能对比

模型	准确率	精确率	召回率	F1分数
MLP	0.8681	0.8788	0.7838	0.8286
LDA	0.8681	0.8788	0.7838	0.8286
朴素贝叶斯	0.8352	0.8056	0.7838	0.7945
LightGBM	0.8681	0.8788	0.7838	0.8286

为了进一步证明基学习器选择的有效性，本研究对多感知机（MLP）、线性判别分析（LDA）、朴素贝叶斯、LightGBM等算法进行了对比，从表2可以看出MLP、LDA、LightGBM在测试集上的准确率、精确率、召回率、F1分数分别为86.81%、87.88%、78.38%和82.86%。证明在相同特征空间分布的前提下，不同的基学习器的选择有相同的分类上界。

值得注意的是，相对而言，朴素贝叶斯算法（Accuracy:83.52%、Precision:80.56%、F1:79.45%）性能相对较低，但其召回率保持与其他算法相同，这是对高维基因数据特征独立的错误假设。各算法在召回率指标上保持一致，说明数据集中真阳性的确定较为困难，可能与肺癌本身生物异质性相关。对比说明，基学习器具有多样性。

（三）增强型Stacking模型分析

1.模型性能分析

为了充分说明Stacking集成学习模型在高维基因表达数据上的分类效果，本研究采用了多个关键性能指标进行综合评估。这些指标包括准确率(Accuracy)、精确率(Precision)、召回率(Recall)、F1分数(F1–Score)以及受试者工作特征曲线下面积(AUC)。通过交叉验证和独立测试集验证，确保了评估结果的可靠性和泛化能力。Stacking模型在测试集上展现出了优异的整体性能表现。

从图2中的各项指标来看，该模型在处理复杂的生物医学分类任务时表现比较优异，各指标之间保持了良好的平衡性。其中在是在AUC指标上达到了接近理想的水平，表明模型具备出色的判别能力。同时，较高的精确率显示了模型在减少假阳性预测方面的有效性，这对于生物医学应用场景具有重要的实际意义。召回率虽然相对较低，但仍处在可接受范围内，整体F1分数反映出精确率和召回率之间达成了合理的权衡。

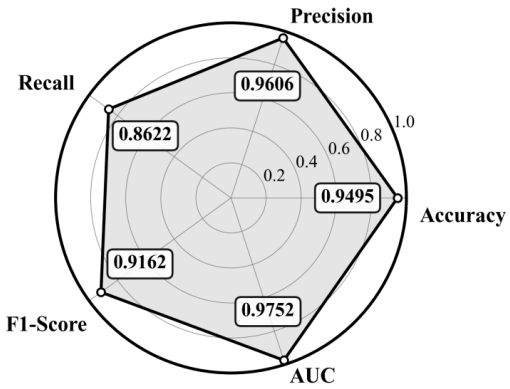

图2 模型性能对比雷达图

在传统机器学习算法对比中，支持向量机(SVM)在处理高维稀疏数据时面临维数灾难挑战，测试准确率为0.891，精确率为0.863。逻辑回归模型虽具备良好可解释性，但受线性决策边界约束，F1分数仅为0.847，难以捕捉基因间复杂非线性交互。随机森林通过bootstrap采样机制缓解过拟合，AUC值达0.923，但单一集成策略局限性使整体表现仍有提升空间。XGBoost在准确率方面表现为0.901，然而在类别不平衡数据上精确率相对不足。

深度学习模型中，多层感知器(MLP)利用神经网络表示学习能力，测试准确率达0.894，但在小样本生物医学数据上易出现过拟合。线性判别分析(LDA)计算效率高，但多元正态分布假设在基因表达数据中往往不成立。朴素贝叶斯基于特征条件独立假设，在基因共表达网络普遍存在的数据中难以发挥最佳效果。

Stacking集成学习模型通过构建异构基础学习器组合，有效克服单一算法固有缺陷。该模型采用分层学习策略：第一层多个基础分类器从不同角度学习数据特征，包括线性模式、非线性关系等多维信息；第二层元学习器通过交叉验证生成的元特征，自适应学习各基础模型最优权重组合。定量结果显示，Stacking学习模式比最佳基线模型更加准确、更精准，F1评分更好，分别为5.8%，9.4%，8.1%。AUC趋于0.975，模型几乎能够达到完美判别率，高精确率和高召回率，在医学诊断中具有重要意义。这也表明该模型具备近乎理想的判别能力，实现了精确率与召回率的良好平衡，对生物医学诊断应用具有重要临床价值。

2.模型动态权重分配分析

Stacking元学习器为四个基学习器分配了相对均衡的权重如图3所示：

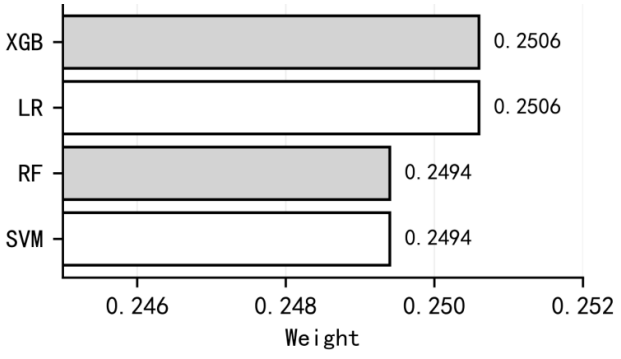

图3 基模型权重分配图

如图3所示，此权重分配表示了不同模型对各个数据特性的优势，此权重分配的动态调整使得集成模型的效果得以保证，权值的小波动表明，每个基学习器对最终预测都有贡献，集成是有效的。

（四）统计显著性验证

对30次独立实验进行配对t检验，验证Stacking模型相对于基学习器的统计显著性结果如下表3：

表3 统计显著性检验结果

对比模型	准确率 p 值	F1分数 p 值	AUC值 p 值
Stacking vs SVM	p<0.001	p<0.001	p<0.001
Stacking vs 逻辑回归	p<0.001	p<0.001	p<0.001
Stacking vs 随机森林	p<0.001	p<0.001	p<0.001
Stacking vs XGBoost	p<0.001	p<0.001	p<0.001

结果表明：所有关键指标的性能提升均具有统计显著性(p<0.001)，低于平常的显著性水平 α=0.05，说明了增强型Stacking框架的有效性。p<0.001表示在零假设为真的前提下，观察到当前结果的概率小于0.1%，为拒绝零假设提供了充分统计学依据。

这些结果为集成学习理论提供了强有力的实证支持，验证了通过整合多个异质学习器可实现“集体智慧”效应。统计显著性的验证不仅具有理论意义，在实际应用中也意味着更可靠的诊断准确性和更稳定的预测性能，为该方法在高维生物医学数据分析中的推广应用提供了科学依据。

（五）特征重要性与可解释性分析

通过特征重要性分析，识别出对肺癌分类最重要的前20个基因特征：

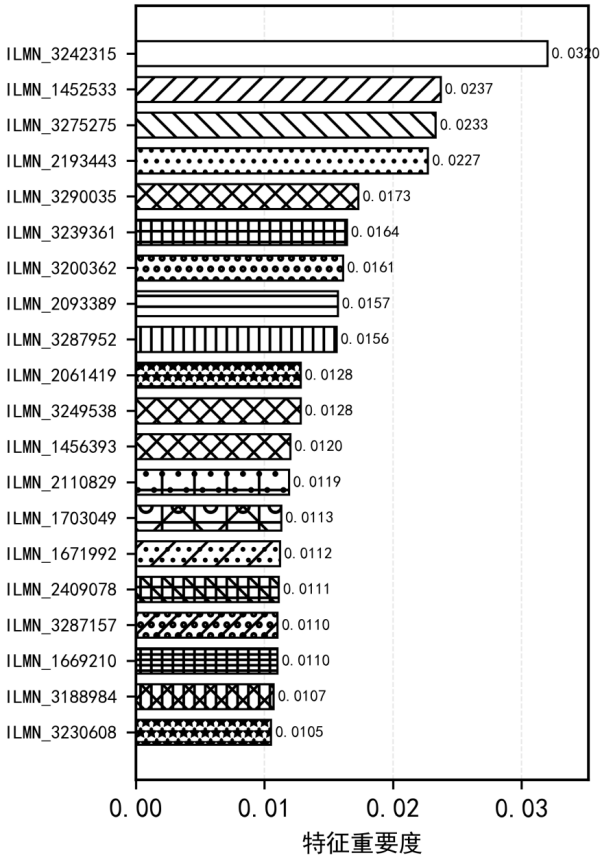

图4 Top20关键基因特征重要性分布

通过加强型Stacking方法筛选出来的20个基因具有相似性及潜在的临床价值，通过对基因的功能注释，可以将20个基因进行阐述(详见表4)。

表4 Top20关键基因功能分类与描述

功能模块	探针 ID	基因符号	基因名称	基因功能描述
细胞周期调控	ILMN_3275275	CDKN2A	细胞周期蛋白依赖性激酶抑制剂2A	肿瘤抑制基因，p16蛋白编码基因
	ILMN_3290035	CDK6	细胞周期蛋白依赖性激酶6	细胞周期 G1/S转换调控
免疫应答	ILMN_3239361	GZMA	颗粒酶 A	细胞毒性 T细胞效应分子
	ILMN_3287952	GZMK	颗粒酶 K	NK细胞和 T细胞介导的细胞毒性
	ILMN_2061419	CD247	T细胞受体ζ链	T细胞激活信号传导
RNA代谢调控	ILMN_3242315	SNORD3D	小核仁 RNA D类3号	rRNA化学修饰指导
	ILMN_3249538	SNORA2B	小核仁 RNA H/ACA 盒2B	rRNA假尿苷化修饰
	ILMN_2409078	SNHG10	小核仁 RNA宿主基因10	长非编码 RNA，基因表达调控
	ILMN_1669210	SNORD16	小核仁 RNA D类16号	rRNA加工与成熟
	ILMN_3188984	ZFAS1	锌指反义 RNA1	长非编码 RNA，转录调控
信号传导与代谢	ILMN_1652533	CYP2C9	细胞色素 P450 2C9	药物代谢酶
	ILMN_1656393	PPP2R3A	蛋白磷酸酶2调节亚基 A	蛋白质去磷酸化调控
	ILMN_1703049	AKAP7	A激酶锚定蛋白7	cAMP依赖性蛋白激酶定位
	ILMN_2110829	SLC25A37	溶质载体家族25成员37	铁离子转运载体
其他功能	ILMN_2193443	TRIM16L	三重基序蛋白16样	蛋白质泛素化调控
	ILMN_3200362	TMEM106B	跨膜蛋白106B	细胞膜结构与功能
	ILMN_2093389	SNAPC1	小核 RNA激活复合物多肽1	转录激活复合物组分
	ILMN_1671992	TSPAN6	四跨膜蛋白6	细胞膜组织与信号传导
	ILMN_3287157	C1ORF159	1号染色体开放阅读框159	功能未完全阐明
	ILMN_3230608	C1ORF112	1号染色体开放阅读框112	功能未完全阐明

基于这20个基因构建的 Stacking 分类模型在测试集上实现了93.00%的准确率和94.79%的 AUC值，显著优于传统单一生物标志物方法。多基因组合能够更全面地反映肺癌的分子异质性，提高诊断的准确性和可靠性。

本研究筛选出来的基因也有着治疗指导意义 :(1)个体化治疗选择：CDK6高表达患者可能从 CDK4/6抑制剂治疗中获益，为靶向治疗的精准应用提供依据

(2)免疫治疗评估：GZMA和 GZMK的表达模式可用于评估患者对免疫检查点抑制剂的响应可能性

(3)药物代谢预测：CYP2C9的表达水平可指导化疗药物的剂量调整，减少不良反应的发生

多个细胞周期和免疫相关基因的联合表达模式可能与患者的临床结局相关，为预后风险分层提供分子依据。

四、结论

（一）Stacking集成策略的性能优势分析

本研究针对肺癌基因表达数据的高维度、小样本及标注噪声挑战，提出了一种增强型 Stacking集成学习框架。基于 GSE252168数据集（303个样本，30,715个基因）的实验表明，该方法在测试集上达到了96.06%的准确率、0.9162的 F1分数和0.9752的 AUC值，相比最优单一模型分别提升了5.8%、8.1%和9.4%，为肺癌分子诊断提供了可靠的技术方案。

（二）Stacking方法创新

本研究提出的三阶段特征选择策略（方差过滤→ FDR校正 t检验→互信息排序）将基因维度从30,715降至1,500，在保留判别信息的同时大幅降低了计算复杂度。这与杜冲等人的集成特征选择思想[9]有相通之处，但本研究进一步结合了统计显著性检验，使特征筛选更加严谨。相比 Ma等人采用的端到端深度学习[10]，本研究保留了原始基因标识，便于生物学解释。双重增强机制是核心创新：一方面将基学习器的元特征与原始特征拼接，既利用预测信息又保留解释性；另一方面通过公式 (8)–(9)定义的动态权重分配策略，使优秀模型获得更高贡献度。Sumon等人的小细胞肺癌研究[11]和 Ganie等人的多癌种分类[12]均采用 Stacking框架，但未涉及动态权重优化，本研究的自适应融合机制实现了性能的进一步提升。

（三）模型可解释性与生物医学意义

通过 L1正则化逻辑回归元学习器的特征权重分析，识别出20个关键基因（见表4）。这些基因覆盖了细胞周期调控（CDKN2A、CDK6）、免疫应答（GZMA、GZMK、CD247）、RNA代谢调控（SNORD3D、SNORA2B、SNHG10）等多个生物学通路。其中 CDKN2A作为肿瘤抑制基因在多种肺癌中失活，CDK6高表达提示患者可能从 CDK4/6抑制剂治疗中获益，GZMA和 GZMK的表达模式可用于评估免疫治疗响应。仅基于这20个基因的模型在测试集上仍达到93.00%的准确率和94.79%的 AUC值，证明了特征选择的有效性。相比 Ganie等人的 SHAP分析[12]，本研究进一步提供了基因功能注释和临床应用指导，增强了从统计预测到生物学解释的完整性。

（四）研究局限与未来工作

首先数据来源于单一数据集（GSE252168），未来需在

TCGA等大型数据库及多中心临床队列上验证泛化能力。其次当前仅处理二分类任务，对于肺癌亚型多分类及生存预后回归预测尚未涉及，杨晨雨等人的药物敏感性预测[24]和郭依晨的生存模型[23]为这些方向提供了参考。然后还未引入深度学习模型（如CNN、Transformer）作为基学习器，陈一凡的序列表征学习[26]和王欣的非编码RNA识别[27]展示了深度模型的潜力，未来可探索深度特征与传统集成学习的融合。Stacking训练时间相对于单模型较长，后续可探索模型蒸馏或近似算法以加速推理。

未来研究可在以下方向拓展：整合多组学数据（吴继明的工作[22]展示了多组学整合潜力）、开展前瞻性临床验证、探索联邦学习框架、结合Huynh-Thu等人的调控网络推断方法[28]深化对肺癌分子机制的理解。

参考文献

[1]Stark R,Grzelak M,Hadfield J.RNA sequencing: the teenage years[J].Nature Reviews Genetics,2019,20(11):631-656.

[2]Byron S A,Van Keuren-Jensen K,Engelthaler D M,et al.Translating RNA sequencing into clinical diagnostics: opportunities and challenges[J].Nature Reviews Genetics,2016,17(5):257-271.

[3]Bzdok D, Altman N, Krzywinski M. Statistics versus machine learning[J]. Nature Methods,2018,15(4):233-234.

[4]Yi Z,Prabhakar C,Jianghua H.Nested cross-validation with ensemble feature selection and classification model for high-dimensional biological data[J]. Communications in Statistics - Simulation and Computation, 2023, 52(1): 110-125.

[5]Ahrens A, Hansen C B, Schaffer M E. pystacked:Stacking generalization and machine learning in Stata[J].The Stata Journal,2023,23(4):909-931.

[6]Libbrecht M W, Noble W S. Machine learning applications in genetics and genomics[J].Nature Reviews Genetics, 2015,16(6): 321-332.

[7]Mohammed M M A.A Comparison of Cancer Classification Methods Based on Microarray Data[D].University of KwaZulu-Natal,Pietermaritzburg,2018.

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

[9]杜冲，周长银，李悦.集成特征选择方法在基因表达数据上的应用[J].山东科技大学学报(自然科学版)，2019,38(01): 85-90.

[10]Ma J,Yu M K,Fong S,et al. Using deep learning to model the hierarchical structure and function of a cell[J]. Nature Methods,2018,15(4):290-298.

[11]Sumon M S I,Shahriar S M S,Hasan M A M,et al.Integrative stacking machine learning model for small cell lung cancer prediction using metabolomics profiling[J]. Cancers, 2024,16(24):4225.

[12]Ganie S M, Dutta Pramanik P K,Zhao Z.Enhanced and interpretable prediction of multiple cancer types using a stacking ensemble approach with SHAP analysis[J]. Bioengineering, 2025, 12(5): 472.

[13]Naderalvojoud B, Hernandez-Boussard T. Improving machine learning with ensemble learning on observational healthcare data[C]//AMIA Annual Symposium Proceedings.2023,1(11): 521-529.

[14]Chicco D,Alameer A,Rahmati S,et al.Towards a potential pan-cancer prognostic signature for gene expression based on probesets and ensemble machine learning[J]. BioData Mining, 2022,15:28.

[15]李泉伦，陈争光，焦峰.基于Stacking集成学习的近红外光谱油页岩含油率预测[J].光谱学与光谱分析，2023,43(04): 1030-1036.

[16]张松兰.支持向量机的算法及应用综述[J]. 江苏理工学院学报，2016, 22(02): 14-17+21.

[17]Breiman L,Friedman J,Olshen R A, et al. Classification and Regression Trees[M]. Chapman and Hall/CRC,2017.

[18]Chen T,Guestrin C.Xgboost:A scalable tree boosting system[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.2016:785-794.

[19]Mahmoud A, Takaoka E. An enhanced machine learning approach with stacking ensemble learner for accurate liver cancer diagnosis using feature selection and gene expression data[J]. Healthcare Analytics, 2025,7:100373.

[20]王笑.基于机器学习的单细胞类型识别方法研究[D]. 西北农林科技大学，2024.DOI:10.27409/d.cnki.gxbnu.2024.002493.

[21]钟晨露.联合人工智能与基因检测指导非瓣膜性房颤患者华法林个体化抗凝研究[D]. 扬州大学，2024.DOI:10.27441/d.cnki.gyzdu.2024.001110.

[22]吴继明.基于多组学数据的癌症类型识别与分期诊断研究[D]. 景德镇陶瓷大学，2024. DOI:10.27191/d.cnki.gjdtc.2024.000106.

[23]郭依晨.整合多组学数据建立女性癌症的生存预测模型[D]. 华北电力大学(北京)，2023.DOI:10.27140/d.cnki.ghbbu.2023.000571.

[24]杨晨雨，刘振浩，代培斌，等.基于多组学数据的肿瘤药物敏感性预测[J].生物工程学报，2022,38(06):2201-2212. DOI:10.13345/j.cjb.210676.

[25]于明铭.基于机器学习的卵巢癌转录组数据分析方法研究[D].新疆大学，2022.DOI:10.27429/d.cnki.gxjdu.2022.001524.

[26]陈一凡.生物大分子序列表征学习方法及其应用研究[D].湖南大学，2023. DOI:10.27135/d.cnki.ghudu.2023.000546.

[27]王欣.非编码RNA与其启动子的识别及疾病关联预测方法研究[D]. 哈尔滨工业大学，2023. DOI:10.27061/d.cnki.ghgdu.2023.005953.

[28]Huynh-Thu V A, Irrthum A, Wehenkel L, et al. Inferring regulatory networks from expression data using tree-based methods[J]. PloS one, 2010, 5(9): e12776.

