Hans Journal of Computational Biology 计算生物学, 2026, 16(1), 31-39
Published Online March 2026 in Hans. https://www.hanspub.org/journal/hjcb
https://doi.org/10.12677/hjcb.2026.161003

Hans汉斯

基于多组学数据的肺癌分期预测研究

胡思亲

江西服装学院大数据学院，江西 南昌

收稿日期：2026年2月19日；录用日期：2026年3月12日；发布日期：2026年3月20日

摘　要

癌症是一类由基因变异引发的恶性疾病，其发病率和死亡率均较高，严重威胁人类健康。基因表达调控对生物体发育至关重要，在肿瘤发生与发展中，常表现为沉默基因的异常激活或活跃基因的表达抑制，这被认为是促进肿瘤发展的关键机制之一。此外，人体微生物群落参与调控多种生理过程，其结构或功能失调可提升致癌风险。本研究聚焦于肺癌，整合基因表达与微生物组数据，旨在开发一种用于肿瘤分期预测的计算模型。研究流程如下：首先对基因表达与微生物组数据进行差异分析，筛选显著变化的基因及微生物物种；其次构建融合注意力机制的深度神经网络模型；随后基于弹性网模型选出的关键特征训练模型以预测肺癌分期；最后采用五折交叉验证评估模型性能。实验结果表明，该模型在肺癌分期预测中表现优异，准确率超过80%。

关键词

多组学数据，肺癌，深度学习，分期模型

Study on Multi-Omics Data-Driven Prediction of Lung Cancer Stages

Siqin Hu

School of Mega Data, Jiangxi Institute of Fashion Technology, Nanchang Jiangxi

Received: February 19, 2026; accepted: March 12, 2026; published: March 20, 2026

Abstract

Cancer is a malignant disease driven by genetic alterations, characterized by high incidence and mortality rates, posing a severe threat to human health. Precise regulation of gene expression is essential for normal organismal development; in tumorigenesis and progression, it is frequently disrupted through aberrant activation of normally silenced genes or suppression of constitutively active genes—a mechanism widely regarded as pivotal in cancer development. Moreover, the human microbiota

文章引用: 胡思亲. 基于多组学数据的肺癌分期预测研究[J]. 计算生物学, 2026, 16(1): 31-39.
DOI: 10.12677/hjcb.2026.161003

modulates a wide array of physiological processes, and dysbiosis—either structural or functional—has been associated with an elevated risk of carcinogenesis. This study focuses on lung cancer and integrates gene expression and microbiome data to develop a computational model for tumor stage prediction. The workflow is as follows: To identify significantly dysregulated genes and microbial taxa, differential analyses were conducted on both gene expression profiles and microbiome compositions. Subsequently, a deep neural network incorporating an attention mechanism is constructed; third, key features selected by an elastic net model are used to train the network for lung cancer staging; Finally, model performance is evaluated via five-fold cross-validation. Experimental results demonstrate that the proposed model achieves superior predictive performance, with an accuracy exceeding 80%.

Keywords

Multi-Omics Data, Lung Cancer, Deep Learning, Staging Model

1. 引言

癌症源于细胞的失控性增殖，是全球致死率最高的疾病之一[1]。在我国，肺癌的发病率和死亡率均居各类恶性肿瘤首位，显著高于结肠癌、肝癌、胃癌和乳腺癌等其他常见癌种。作为一种高度异质性的原发性肿瘤，肺癌的发生发展涉及多个基因异常及多条信号通路的协同失调，并具有局部侵袭和远处转移的能力。

临床实践中，肺癌的严重程度通常依据其原发灶大小、病灶数量及扩散范围进行“分期”。目前广泛采用的 TNM 分期系统由三个核心指标构成：T (Tumor)反映原发肿瘤的体积与局部浸润程度，N (Node)表示区域淋巴结受累情况，M (Metastasis)则指示是否存在远处转移。综合 T、N、M 三要素，肿瘤被划分为 I 至 IV 期(以罗马数字表示)，其中 I 期为早期，II~IV 期归为中晚期。治疗策略依分期而定：早期患者多接受手术切除；中晚期患者则常需联合化疗、靶向治疗及放疗等多模式干预[2]。当前，肺癌分期主要依赖胸部 CT (计算机断层扫描)和 MRI (磁共振成像)等影像学手段。尽管这些技术对恶性病变具有较高敏感性[3] [4]，但其分期准确性仍有限[5]，难以全面反映肿瘤生物学行为，可能导致治疗延迟或预后不良[6]。

癌症是一种与基因表达异常密切相关的恶性疾病[7]。基因表达涉及将 DNA 信息转录为 RNA 并翻译成蛋白质，对正常发育至关重要。肿瘤发生主要受两种机制影响：一是遗传变异，如抑癌基因或原癌基因的突变及染色体结构异常[7] [8]；二是表观遗传调控，即通过化学修饰调节基因活性而不改变 DNA 序列[9]。研究表明，癌症的发展很大程度上由基因表达的动态失衡驱动。例如，在早期非小细胞肺癌(NSCLC)中，Bmi-1 基因表现出初期上调随后下调的时序性变化[10]；BRCA2 基因突变显著增加患癌风险，尤其在吸烟者中可达普通人群两倍[11]；EGFR 突变促进细胞异常增殖，于晚期肺癌尤为常见[12]。

人体健康依赖于微生物群，后者可通过诱发慢性炎症、破坏免疫稳态及产生代谢产物等方式参与肿瘤发展[13]。人体微生物组是一个包含共生菌和潜在致病菌的复杂生态系统。某些情况下，原本无害甚至有益的微生物也可能转变为促癌因素[13]-[15]。例如，肺部共生菌群可抑制炎症并维持免疫耐受，但其失衡可能导致免疫紊乱，进而促进肺癌[13]。随着肿瘤进展，微生物种类及其相对丰度亦呈现动态变化[16]。这些发现揭示了微生物组在癌症进程中的关键作用，并为诊断和治疗提供了新视角。

高通量测序技术的快速发展推动了肿瘤多组学数据(包括基因组、转录组、蛋白组和代谢组)的快速积累。有效整合这些数据有助于提升癌症分期的准确性，而精准分期对治疗决策与预后评估至关重要[17][18]，是实现个体化治疗、改善生存结局的关键基础[19]。因此，融合基因组学与微生物组学等多维信息已成为肺癌精准分期的重要方向。

为系统探究基因表达与微生物群落在肺癌进展中的作用，本文聚焦于早期(I 期)与中晚期(II~IV 期)肺癌的分类预测。通过结合统计建模、机器学习与深度学习方法，构建一个整合基因表达与微生物组特征的多组学预测框架，并引入注意力机制以增强模型对关键生物标志物的识别能力。研究旨在提升预测性能的同时，挖掘具有生物学意义的潜在标志物，揭示多组学特征与肺癌演进的内在关联，为个体化诊疗提供理论依据与技术路径。

2. 材料与方法

2.1. 构建数据集

国际癌症基因组联盟(ICGC)是专为癌症研究构建的权威数据库，其核心目标为挖掘并整合全球各类癌症中引发人类患病的基因组变异信息。本研究从 ICGC 数据库下载肺癌患者的基因表达数据与临床信息，并结合 Poore 等人[20]发表的微生物组数据，构建了一个多组学数据集。在对上述三类数据进行质量控制与统一注释后，最终筛选出 189 例临床信息完整的肺癌样本，各病程阶段的样本分布详见表 1。根据肿瘤分期所对应的治疗策略差异，将样本划分为两类：早期组包括 Stage IA 和 Stage IB 患者，中晚期组涵盖 Stage II、III 及 IV 期患者，从而构建用于肺癌分期预测的二分类数据集。

Table 1. Lung cancer sample information
表 1. 肺癌样本信息

Cancer	Stage	Size	Group	Total
LUNG	IA, IB	98	Early	98
	IIA, IIB	41	Middle-Late	91
	IIIA, IIIB	40		
	IV	10		

2.2. 特征降维

2.2.1. 基于 DESeq2 的基因表达差异分析

本研究采用 R 语言中的 DESeq2 包进行基因表达差异分析，通过局部回归刻画基因表达均值与方差的关系，并结合离散度估计与 $\log_2$ 倍数变化(Fold Change)的收缩技术，以提高结果的稳健性与可重复性[20]-[22]。DESeq2 内置的标准化机制能够有效校正中等表达基因的定量偏差，在显著控制假阳性率的同时，保持较高的检测灵敏度与特异性，因而被广泛认可为差异表达分析的可靠工具[23]。基于 DESeq2 的基因表达差异分析流程主要包括以下三个核心步骤：

1) 导入原始基因表达计数矩阵(Read Counts)，并依据临床分期将样本划分为早期组(Early，作为对照组)与中晚期组(Middle-Late，作为实验组)，构建对应的样本分组向量(Groups)。在此基础上，结合计数数据、分组信息及基因注释，生成用于统计建模的设计矩阵(Design Matrix)，最终整合上述要素构建 DESeq DataSet 对象(DDS)；

2) 调用 DESeq()函数对 DDS 对象执行完整的差异表达分析，设定筛选阈值为校正后 P 值，P.adj < 0.05 且 abs ($\log_2$ Fold Change) > 1，依据该阈值从分析结果中筛选差异表达基因；

3) 基于步骤 2)的筛选条件，剔除绝对值约束后，进一步区分出满足条件的上调基因与下调基因子集。

2.2.2. 基于 Mann-Whitney U Test 的微生物丰度差异分析

本研究采用 Mann-Whitney U 检验，比较早期与中晚期肺癌患者间微生物相对丰度的分布差异，以筛选显著差异富集的微生物物种作为候选生物标志物。

Mann-Whitney U 检验由 H. B. Mann 与 D. R. Whitney 于 1947 年提出，是一种用于比较两组独立样本的非参数统计方法。其零假设为：两组数据来自除位置参数(如中位数)外分布完全相同的总体。该检验的核心目的在于判断两个总体的分布位置是否存在显著差异，尤其适用于数据不满足正态性或方差齐性假设的情形。基于 Mann-Whitney U test 的微生物丰度差异分析流程主要包括以下三个步骤：

1) 导入原始微生物种计数数据(micro_count)，将样本划分为两个组(groups)：早期组(x)与中晚期组(y)；

2) 调用 mannwhitneyu()函数，输入 x 和 y 值，设定 alternative 为 two-sided，得到检验结果；

3) 基于步骤 2)的结果，设定筛选条件 $p < 0.05$ 且 absolute_value > 0.2，得到满足条件的差异微生物种。

2.2.3. 基于弹性网模型的特征筛选

在进行差异分析后，本研究进一步开展特征选择，以提升后续建模的效率与泛化能力[24] [25]。为此，本研究采用弹性网(Elastic Net)正则化回归方法进行特征筛选，该方法由 Zou 等人[26] [27]提出，通过同时引入 L1 (Lasso)与 L2 (Ridge)惩罚项，在“高维小样本”场景下展现出独特优势：一方面可生成稀疏解以实现自动变量选择；另一方面凭借其“分组效应”，将高度相关的特征协同纳入模型，有效克服 Lasso 在处理强相关变量时随机保留单一特征的局限性。

在具体实施中，我们分别对 mRNA 数据、微生物组数据以及二者融合的多组学数据集独立进行特征筛选。采用五折交叉验证框架：在每折训练集中拟合弹性网模型并记录所选特征。为增强结果的稳定性与可重复性，整个五折交叉验证流程独立重复 5 次，最终仅保留在所有验证中均被选中的特征，作为分期候选标志物。

2.3. 肺癌分期模型的构建

2.3.1. 基于注意力机制的深度神经网络

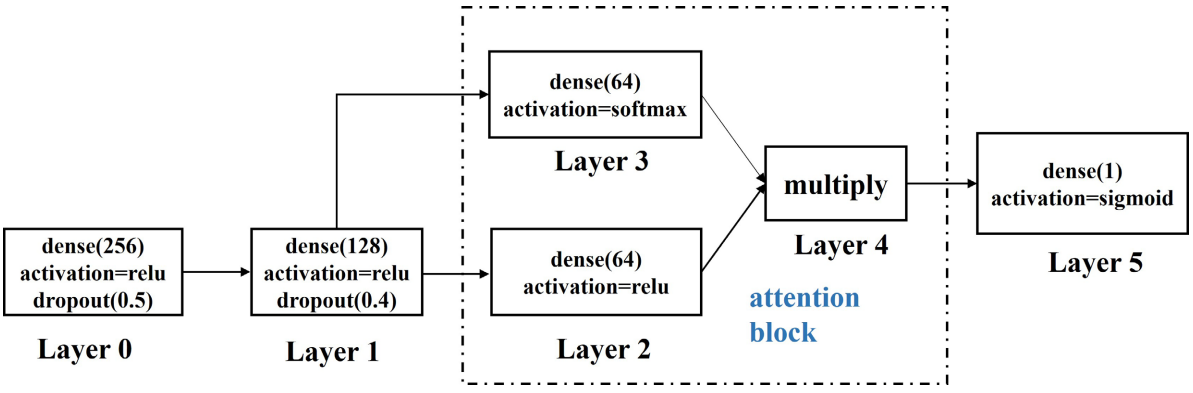

Figure 1. Diagram of a deep neural network architecture based on the attention mechanism
图 1. 基于注意力机制的深度神经网络结构图

本研究提出了一种基于深度神经网络并融合注意力机制的肺癌分期预测模型，即 ATT-DL (Attention-

Based Deep Learning Model)。该模型的架构设计包括输入层、全连接层、注意力模块以及输出层，通过这种结构堆叠形成完整的模型体系，具体结构如图 1 所示。

图中每个矩形框代表一个网络层：dense(n) 表示该全连接层包含 n 个神经元；activation 标注所采用的激活函数；dropout(p) 表示以概率 p 实施随机失活正则化策略[28]。值得注意的是，Layer 2、Layer 3 与 Layer 4 共同构成注意力模块，其中通过 multiply 操作实现特征权重与原始输入的逐元素相乘，从而动态增强关键特征的贡献。采用弹性网模型对 mRNA 数据、微生物组数据以及二者融合的多组学数据集独立进行特征筛选后分别得到的肺癌分期预测的候选生物标志物子集作为 ATT-DL 模型的输入，输出则是判断某样本所处癌症阶段。

整个网络结构基于 TensorFlow 深度学习框架实现，并完成端到端的模型训练。具体参数设置如下：训练过程共进行 100 个 epoch，以确保模型充分收敛；批量大小(Batch Size)设为 4，以平衡计算效率与梯度估计稳定性；优化器选用 Adam 算法，因其在处理非凸优化问题时具有良好的自适应学习率调节能力；初始学习率设定为 0.001，有助于在训练初期稳定地更新模型权重；损失函数采用二元交叉熵(Binary Crossentropy)，适用于二分类或多重分类任务中的概率输出建模，尤其适合将肿瘤分期视为类别标签的分类问题。

2.3.2. 评价指标

为系统评估所提出肺癌分期预测模型的有效性，本研究采用四项常用分类性能指标进行综合评价，包括：准确率(ACC)、精准率(PRE)、召回率(REC)以及受试者工作特征曲线下面积(AUC)。

3. 实验结果与分析

3.1. 基因表达差异分析结果

基于 DESeq2 的差异表达分析共鉴定出 583 个差异表达基因(DEGs)，其中 197 个上调，386 个下调。图 2 为火山图，横轴表示 $\log_2$ 倍数变化($\log_2$ Fold Change)，纵轴表示统计显著性($-\log_{10}$ P 值)。每个点代表一个基因，颜色标识其表达状态：红色为上调基因，蓝色为下调基因，灰色为无显著变化基因。从图中可以看到，显著差异表达的基因多分布于左右两侧，远离中心区域，表明其兼具较大的表达变化幅度和较高的统计显著性。

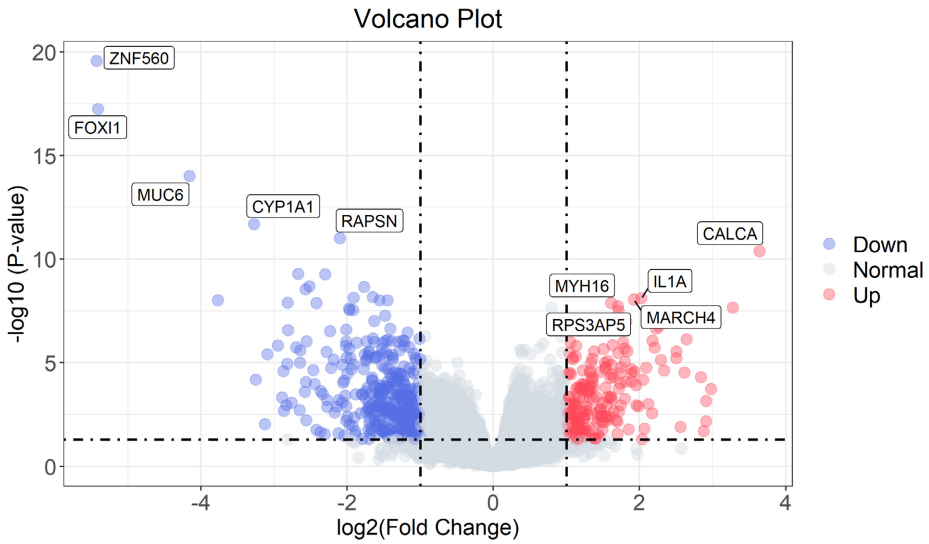

Figure 2. Volcano plot of differential analysis results
图 2. 差异分析结果的火山图

具体而言，显著上调的基因，如 CALCA、IL1A 等，在肿瘤组织中表现出较高的表达水平，提示其可能在肺癌的发生和发展过程中起到重要作用。相反，显著下调的基因，如 ZNF560、FOXI1 等，在肿瘤样本中的表达量较低，表明它们可能具有抑制肿瘤的作用。

3.2. 微生物丰度差异分析结果

本研究采用 Mann-Whitney U 检验对 1524 个微生物物种的相对丰度进行差异分析，设定显著性阈值为 $p < 0.05$，最终筛选出 15 个在早期与中晚期肺癌样本间具有显著丰度差异的微生物物种，其分布情况如图 3 所示。

其中，Pedosphaera 和 Desulfurobacterium 在早期肺癌样本中的相对丰度显著高于中晚期样本；相反，Lentimicrobium、Leptonema、Xanthomonas 和 Sediminimonas 四个物种在中晚期样本中呈现明显富集，表明其丰度随疾病进展而升高。这些差异物种可能与肺癌不同阶段的微生态特征密切相关，具有作为分期相关生物标志物的潜在价值。特别的是 Xanthomonas 已被证实会影响肺癌的发生发展，研究表明嗜麦芽黄单胞菌作用于肺腺癌 A549 细胞后，通过转录组测序发现其可显著调控 MAPK、p53、JAK-STAT、PI3K-Akt 等与肺癌发生发展密切相关的信号通路，影响肿瘤细胞基因表达[29]。

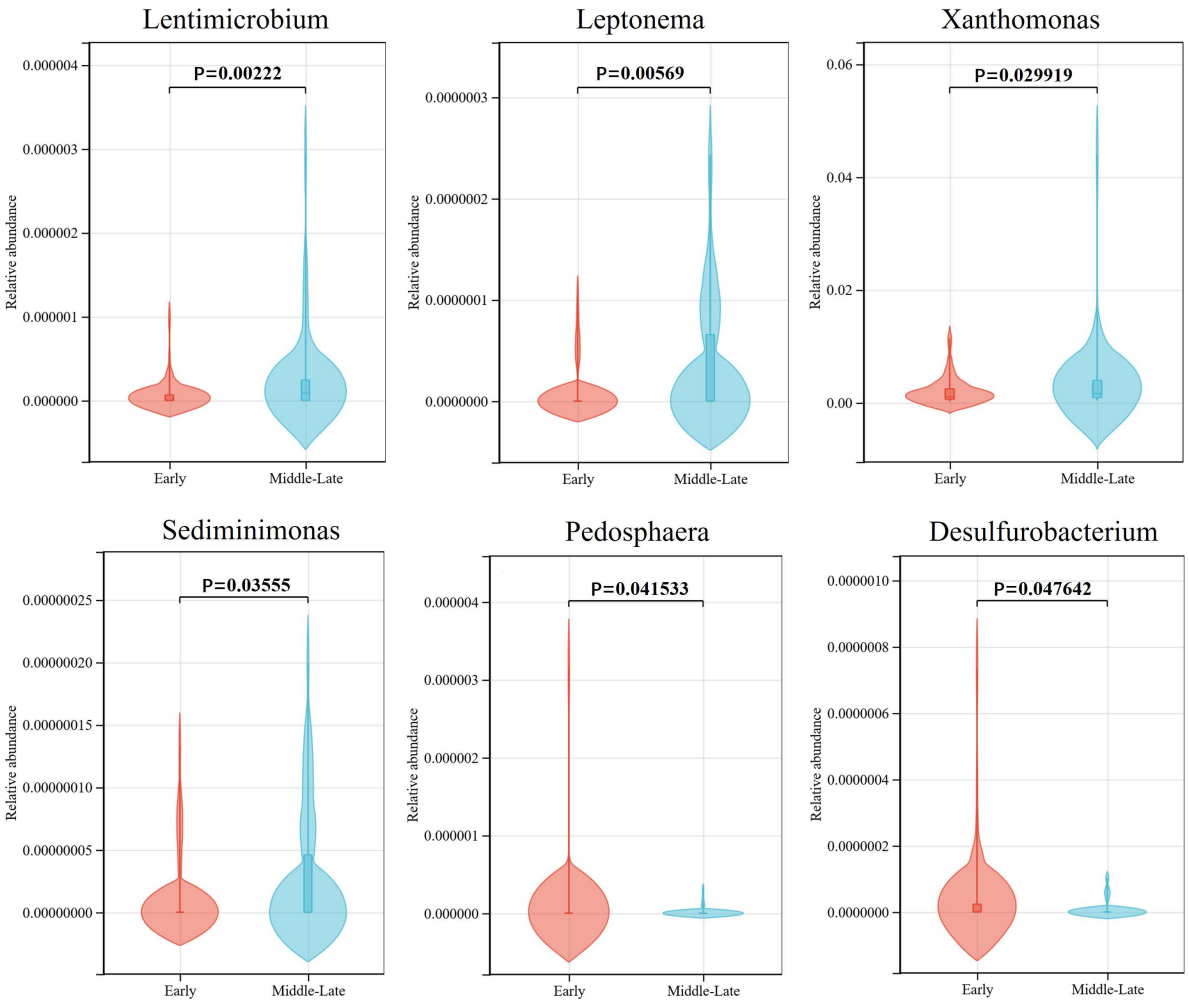

Figure 3. Expression levels of differential microbial species at different stages of cancer
图 3. 差异微生物物种在癌症不同阶段的表达水平

3.3. 特征筛选的结果

针对弹性网模型的输入数据，本研究按以下流程进行处理：首先，基于 583 个差异表达基因(DEGs)对 mRNA 数据进行特征筛选，最终保留 13 个具有判别能力的基因；其次，对微生物组数据，依据 Mann-Whitney U 检验的 p 值对 1524 个微生物物种进行升序排序，并选取前 150 个最显著的物种作为候选特征，经弹性网筛选后获得 62 个有效微生物物种；最后，将上述 583 个差异基因与前 150 个差异微生物物种合并，构建多组学融合数据集，并在此基础上执行联合特征选择，最终得到 14 个稳定入选的特征，其中包括 12 个基因和 2 个微生物物种。

在 mRNA 与融合数据集之间，共有 9 个基因被共同识别：ARMC3、EREG、CLDN8、GFRA3、SAA4 和 SLC22A9、NXPE4、TMPRSS11E、ANKRD204A。在微生物组与融合数据集之间，仅有一个微生物物种 Desulfurobacterium 被共同识别。不同组学数据间的特征重叠性，反映出融合分析不仅保留了各单一组学中的关键信号，还可能通过整合信息增强特征的生物学意义和预测效能。

3.4. 肺癌分期模型的评估结果

本研究构建的基于注意力机制的癌症分期预测模型(ATT-DL)，在三种数据集(mRNA 数据集、Microbiome 数据集及 mRNA + Microbiome 融合数据集)上开展 5 次五折交叉验证，验证结果详见表 2。结果显示，ATT-DL 模型在融合数据集上的预测性能，显著优于其在两种单一数据集上的表现，四项评估指标(ACC、REC、PRE、AUC)数值均超过 80%，体现了融合基因与微生物特征的优势。

Table 2. Prediction results of the ATT-DL model on three datasets
表 2. ATT-DL 模型在三种数据集中的预测结果

Datasets	ACC	PRE	REC	AUC
mRNA	0.7667 ± 0.0002	0.7687 ± 0.0002	0.7539 ± 0.0004	0.7821 ± 0.0002
Microbiome	0.7924 ± 0.0003	0.7835 ± 0.0003	0.8000 ± 0.0004	0.8081 ± 0.0004
mRNA + Microbiome	**0.8136 ± 0.0002**	**0.8089 ± 0.0005**	**0.8215 ± 0.0004**	**0.8275 ± 0.0003**

此外，实验最后还将本研究所提出的癌症分期预测模型(ATT-DL)模型与传统机器学习(SVM、Random Forest、XGBoost、Logistic Regression)模型在融合数据集上进行了对比，表 3 显示 ATT-DL 模型各项指标均高于四种传统机器学习模型，说明在癌症分期预测任务中使用深度学习是非常有必要的，更加体现出 ATT-DL 模型的高性能。

Table 3. Prediction results of the ATT-DL model and traditional machine learning models on the fused dataset
表 3. ATT-DL 模型与传统机器学习模型在融合数据集上的预测结果

Method	ACC	PRE	REC	AUC
ATT-DL	**0.8136 ± 0.0002**	**0.8089 ± 0.0005**	**0.8215 ± 0.0004**	**0.8275 ± 0.0003**
SVM	0.6636 ± 0.0309	0.6810 ± 0.0580	0.6154 ± 0.0487	0.6629 ± 0.0302
Random Forest	0.7030 ± 0.0169	0.7010 ± 0.0104	0.6923 ± 0.0506	0.7029 ± 0.0173
XGBoost	0.6167 ± 0.0419	0.6096 ± 0.0355	0.6092 ± 0.0956	0.6166 ± 0.0426
Logistic Regression	0.7485 ± 0.0101	0.7545 ± 0.0071	0.7262 ± 0.0429	0.7482 ± 0.0105

4. 讨论

本研究旨在探究基因组与微生物组在肿瘤演进中的协同作用，基于多组学数据构建了一种融合注意力机制的深度神经网络模型(ATT-DL)，以实现肺癌分期的精准预测。五折交叉验证结果显示，该模型在mRNA 与微生物组融合数据集上，准确率、召回率、精准率及 AUC 等核心指标均超 80%，显著优于单一组学(mRNA 或微生物组)对照模型，凸显了多组学整合对提升肺癌分期预测性能的价值。

本研究在模型构建与预测效能上取得一定成果，但仍存在以下局限：

1) 样本规模有限且癌种单一：本研究仅分析 189 例肺癌样本，未涵盖其他肿瘤类型，筛选出的差异基因与微生物特征的跨癌种普适性待验证。未来在多种癌症独立队列中评估这些特征的判别能力，可进一步揭示其在肿瘤发生发展中的共性机制。

2) 缺乏外部验证：ATT-DL 模型虽在内部交叉验证中稳定性良好，但其泛化能力仍需通过不同临床中心或测序平台的外部数据集验证。引入多源异构外部数据验证，可更全面评估模型的临床适用性，为其向实际诊疗场景转化提供支撑。

基金项目

江西省教育厅科学技术研究项目(No. GJJ2402712)。

参考文献

[1] Bray, F., Laversanne, M., Weiderpass, E. and Soerjomataram, I. (2021) The Ever-Increasing Importance of Cancer as a Leading Cause of Premature Death Worldwide. *Cancer*, **127**, 3029-3030. https://doi.org/10.1002/cncr.33587

[2] Hou, J., Aerts, J., den Hamer, B., van IJcken, W., den Bakker, M., Riegman, P., *et al*. (2010) Gene Expression-Based Classification of Non-Small Cell Lung Carcinomas and Survival Prediction. *PLOS ONE*, **5**, e10312. https://doi.org/10.1371/journal.pone.0010312

[3] Mountain, C.F. (1997) Revisions in the International System for Staging Lung Cancer. *Chest*, **111**, 1710-1717. https://doi.org/10.1378/chest.111.6.1710

[4] Ma, X., Xi, B., Zhang, Y., Zhu, L., Sui, X., Tian, G., *et al*. (2020) A Machine Learning-Based Diagnosis of Thyroid Cancer Using Thyroid Nodules Ultrasound Images. *Current Bioinformatics*, **15**, 349-358. https://doi.org/10.2174/1574893614666191017091959

[5] Mountain, C.F. and Dresler, C.M. (1997) Regional Lymph Node Classification for Lung Cancer Staging. *Chest*, **111**, 1718-1723. https://doi.org/10.1378/chest.111.6.1718

[6] Tsou, J.A., Hagen, J.A., Carpenter, C.L. and Laird-Offringa, I.A. (2002) DNA Methylation Analysis: A Powerful New Tool for Lung Cancer Diagnosis. *Oncogene*, **21**, 5450-5461. https://doi.org/10.1038/sj.onc.1205605

[7] Hanahan, D. and Weinberg, R.A. (2000) The Hallmarks of Cancer. *Cell*, **100**, 57-70. https://doi.org/10.1016/s0092-8674(00)81683-9

[8] Hahn, W.C., Counter, C.M., Lundberg, A.S., Beijersbergen, R.L., Brooks, M.W. and Weinberg, R.A. (1999) Creation of Human Tumour Cells with Defined Genetic Elements. *Nature*, **400**, 464-468. https://doi.org/10.1038/22780

[9] Jones, P.A. (2012) Functions of DNA Methylation: Islands, Start Sites, Gene Bodies and Beyond. *Nature Reviews Genetics*, **13**, 484-492. https://doi.org/10.1038/nrg3230

[10] Tan, A.C. and Gilbert, D. (2003) Ensemble Machine Learning on Gene Expression Data for Cancer Classification. *Applied Bioinformatics*, **2**, S75-S83.

[11] Wang, Y., McKay, J.D., Rafnar, T., Wang, Z., Timofeeva, M.N., Broderick, P., *et al*. (2014) Rare Variants of Large Effect in BRCA2 and CHEK2 Affect Risk of Lung Cancer. *Nature Genetics*, **46**, 736-741. https://doi.org/10.1038/ng.3002

[12] Anggaraditya, P.B., Adiputra, P.A.T. and Widiana, I.K. (2019) EGFR Nanovaccine in Lung Cancer Treatment. *Bali Medical Journal*, **8**, 844-851. https://doi.org/10.15562/bmj.v8i3.1494

[13] Guo, H., Zhao, L., Zhu, J., Chen, P., Wang, H., Jiang, M., *et al*. (2022) Microbes in Lung Cancer Initiation, Treatment, and Outcome: Boon or Bane? *Seminars in Cancer Biology*, **86**, 1190-1206. https://doi.org/10.1016/j.semcancer.2021.05.025

[14] Bhatt, A.P., Redinbo, M.R. and Bultman, S.J. (2017) The Role of the Microbiome in Cancer Development and Therapy.

CA: *A Cancer Journal for Clinicians*, **67**, 326-344. https://doi.org/10.3322/caac.21398

[15] Schwabe, R.F. and Jobin, C. (2013) The Microbiome and Cancer. *Nature Reviews Cancer*, **13**, 800-812. https://doi.org/10.1038/nrc3610

[16] Han, P., Zhou, J., Xiang, J., Liu, Q. and Sun, K. (2022) Research Progress on the Therapeutic Effect and Mechanism of Metformin for Lung Cancer (Review). *Oncology Reports*, **49**, Article 3. https://doi.org/10.3892/or.2022.8440

[17] Hu, G., Gu, J., Zheng, J., Schnöll, M. and He, F. (2019) Improved Neighborhood Covering Algorithm and Its Lung Cancer Staging Prediction. *Journal of Computational Methods in Sciences and Engineering*, **19**, 317-326. https://doi.org/10.3233/jcm-180872

[18] Qu, W., Zhao, J., Wu, Y., Xu, R. and Liu, S. (2021) Recombinant Adeno-Associated Virus 9-Mediated Expression of Kallistatin Suppresses Lung Tumor Growth in Mice. *Current Gene Therapy*, **21**, 72-80. https://doi.org/10.2174/1566523220999201111194257

[19] Xiong, D., Ye, Y., Fu, Y., Wang, J., Kuang, B., Wang, H., *et al*. (2015) Bmi-1 Expression Modulates Non-Small Cell Lung Cancer Progression. *Cancer Biology & Therapy*, **16**, 756-763. https://doi.org/10.1080/15384047.2015.1026472

[20] Robinson, M.D. and Smyth, G.K. (2007) Moderated Statistical Tests for Assessing Differences in Tag Abundance. *Bioinformatics*, **23**, 2881-2887. https://doi.org/10.1093/bioinformatics/btm453

[21] Anders, S. and Huber, W. (2010) Differential Expression Analysis for Sequence Count Data. *Nature Precedings*. https://doi.org/10.1038/npre.2010.4282.2

[22] Hardcastle, T.J. and Kelly, K.A. (2010) BaySeq: Empirical Bayesian Methods for Identifying Differential Expression in Sequence Count Data. *BMC Bioinformatics*, **11**, Article No. 422. https://doi.org/10.1186/1471-2105-11-422

[23] Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., *et al*. (2013) Comprehensive Evaluation of Differential Gene Expression Analysis Methods for RNA-Seq Data. *Genome Biology*, **14**, Article No. 3158. https://doi.org/10.1186/gb-2013-14-9-r95

[24] Chen, T. and Xie, Y. (2005) Literature Review of Feature Dimension Reduction in Text Categorization. *Journal of the China Society for Scientific and Technical Information*, **24**, 691-695.

[25] Liu, T., Liu, S., Chen, Z., *et al*. (2003) An Evaluation on Feature Selection for Text Clustering. *Proceedings of the* 20*th International Conference on Machine Learning* (*ICML*-03), Washington, 21-24 August 2003, 488-495.

[26] Zou, H. and Hastie, T. (2005) Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **67**, 301-320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

[27] Ogutu, J.O., Schulz-Streeck, T. and Piepho, H. (2012) Genomic Selection Using Regularized Linear Regression Models: Ridge Regression, Lasso, Elastic Net and Their Extensions. *BMC Proceedings*, **6**, Article No. S10. https://doi.org/10.1186/1753-6561-6-s2-s10

[28] Srivastava, N., Hinton, G., Krizhevsky, A., *et al*. (2014) Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *The Journal of Machine Learning Research*, **15**, 1929-1958.

[29] 吴仁迪, 沈吉禹, 王福栋, 等. 嗜麦芽窄食单胞菌对肺腺癌 A549 细胞系转录组基因表达的影响[J]. 中华实验外科杂志, 2023, 40(4): 682-685.