分类号 TP391.1　　密级
U D C　　编号 10741

兰州财经大学
LANZHOU UNIVERSITY OF FINANCE AND ECONOMICS

硕士学位论文

论文题目　基于 RNA 合成指数的非小细胞肺癌患者生存分析和复发研究

研究生姓名：宋玥
指导教师姓名、职称：李兵　教授
学科、专业名称：管理科学与工程
研究方向：信息管理与信息系统

提交日期：2023 年 6 月 6 日

独创性声明

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名：宋玥　　签字日期：2023.5.20

导师签名：李乐　　签字日期：2023.5.20

关于论文使用授权的说明

本人完全了解学校关于保留、使用学位论文的各项规定，同意（选择“同意”/“不同意”）以下事项：

1.学校有权保留本论文的复印件和磁盘，允许论文被查阅和借阅，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文；

2.学校有权将本人的学位论文提交至清华大学“中国学术期刊（光盘版）电子杂志社”用于出版和编入CNKI《中国知识资源总库》或其他同类数据库，传播本学位论文的全部或部分内容。

学位论文作者签名：宋玥　　签字日期：2023.5.20

导师签名：李乐　　签字日期：2023.5.20

Survival analysis and recurrence study of non-small cell lung cancer patients based on RNA synthesis index

Candidate : Yue Song

Supervisor: Bing Li

摘 要

肺癌是全球癌症死亡的主要原因，存活率极低，其中非小细胞肺癌（NSCLC）是肺癌的最大亚群，约占病例总数的85%且总生存率很极低，5年生存率仅为24%。近年来，随着PET/CT成像技术日益成熟，高通量测序技术的发展，获得基因表达谱变得较为方便，促进了基因特征的鉴定，这也为NSCLC的二级预防提供了充足准备。本实验共纳入160例NSCLC患者的放射基因组学数据，包括PET/CT图像、临床信息和RNA-seq数据。患者按照有无RNA-seq数据被分为训练组（108例有RNA-seq数据可用，随访期间有39例死亡）和测试组（52例无RNA-seq数据可用，随访期间有15例死亡）。实验流程首先通过特征挑选从普通患者的5268条RNA序列中筛选出9条RNA序列，接着参考PET/CT容量预后指数（PVP）模型方法将9个RNA序列整合为一个一维变量，即RNA预后指数。其次，依据训练数据的PET/CT和肿瘤掩膜图像，提取放射组学特征，并使用F检验方法降低特征维数。基于RNA指数选择的放射组学特征构建支持向量回归模型（SVR，使用Scikit-learn python包）。利用训练好的SVR模型预测52例患者的RNA预后指标。最后采用Cox比例风险模型评估分析RNA合成指数的预后价值。检验的最终结果显示，在训练数据的多因素Cox比例风险模型中，RNA预后指标与患者总生存期显著相关（风险比（HR）=2.602，P<0.001）。在测试数据的单因素Cox比例风险模型中RNA预后指数预测值与患者总生存期HR达到7.155，P-value值小于0.05，并且多因素生存分析中HR=8.8038，P-value值小于0.05，此结果证明RNA合成指数也与患者生存时间显著相关。此外为了证明RNA合成指数在多模态融合实验中对NSCLC患者复发预测的潜力。本研究将预测得到的52例患者RNA合成指数作为特征纳入NSCLC复发实验研究。基于3D卷积网络模型进行二分类实验，结果证明，综合RNA合成指数，影像和临床三种模态数据，AUC可达到0.70，相较于非RNA合成指数的方法，AUC提高约10%。本研究证明RNA合成指数在生存预后和复发研究中，都具有显著性的医学意义。

关键字：非小细胞肺癌 基因序列 RNA合成指数 生存分析

Abstract

Lung cancer is the main cause of cancer death worldwide, with extremely low survival rates. Non small cell lung cancer (NSCLC) is the largest subgroup of lung cancer, accounting for approximately 85% of the total number of cases, and the overall survival rate is extremely low, with a 5-year survival rate of only 24%. In recent years, with the increasingly mature PET/CT imaging technology and the development of high-throughput sequencing technology, it has become more convenient to obtain gene expression profiles, promoting the identification of gene characteristics, which also provides sufficient preparation for the secondary prevention of NSCLC.This experiment included radiogenomic data from 160 NSCLC patients, including PET/CT images, clinical information, and RNA seq data. Patients were divided into a training group (108 cases with RNA-seq data available and 39 deaths during follow-up) and a testing group (52 cases without RNA-seq data available and 15 deaths during follow-up) based on the presence or absence of RNA-seq data. Firstly, 9 RNA-seq were screened from 5268 RNA sequences of common patients through feature selection, and then integrated into a one-dimensional variable, namely RNA prognostic index, by referring to the PET/CT volume prognostic index (PVP) model. Secondly, according to the PET/CT and tumor mask images of the training data, the radiomic features were extracted, and the F-test method

was used to reduce the feature dimension. A support vector regression model (SVR, using the Scikit-learn python package) was constructed based on the RNA index and selected radiomic features. A trained SVR model was used to predict RNA prognostic indicators in 52 patients. Finally, Cox proportional hazard model was used to evaluate the prognostic value of RNA prognostic index. Final results of the examination showed that RNA prognostic markers were significantly associated with overall survival in a multivariate Cox proportional hazard model of the training data (HR=2.602, P<0.001). In the univariate Cox proportional hazard model of the test data, the predicted value of the RNA prognotic index and the overall survival of patients reached 7.155, and the P-value was less than 0.05; in the multivariate survival analysis, the HR=8.8038, and the P-value was less than 0.05. The results showed that RNA synthesis index was also significantly correlated with survival time. In addition, to demonstrate the potential of RNA synthesis index in multimodal fusion experiments for predicting recurrence in NSCLC patients. This study included the predicted RNA synthesis index of 52 patients as a feature in the NSCLC recurrence experimental study. Based on a 3D convolutional network model, a binary classification experiment was conducted, and the results showed that by combining RNA synthesis index, imaging, and clinical modal data, the AUC can reach 0.70, which is about 10% higher than the non RNA synthesis index method. This

study demonstrates the significant medical significance of RNA synthesis index in survival prognosis and recurrence studies.

Keywords: Non-small cell lung cancer; Gene sequence; RNA synthesis index; Survival analysis

目 录

1. 绪 论

随着人工智能（AI）技术飞速发展、医学数据的持续扩增以及硬件设备的不断提升，人工智能和医疗的结合方式更加趋向多样化。目前 AI 在医疗领域中的落地应用场景主要有医学影像、智能诊疗、智能导诊、智能语音、健康管理、病例分析、医院管理、新药研发和医疗机器人等，其中在医学影像中的应用最为广泛。随着研究方法不断更新，人工智能在医学影像中的应用也逐渐深入，越来越多临床问题可以被更好的解决，使得患者能最终获益。从医生角度来看，人工智能是计算机做人类认为智能的事，其从大到小包括机器学习、神经网络、深度学习。AI 的另一重要概念即大数据，就像一个病人影像的每个断层是 512×512 的像素，200 个断层即可组成 5000 多万个体素，形成 1000~10 万个影像特征，医生看片不可能看到 1000~10 万个特征。但计算机可以通过高通量处理对信息进行降维，使得诊断更准确，从而辅助医生进行决策。医学影像现已成为人工智能在医疗领域最热门的方向，但在实际应用过程中还是存在一定挑战，例如，数据获取及数据标注问题、缺乏行业标准、注册审批缺乏指导原则、技术创新问题等等。但随着 AI 相关技术的不断发展，国家相关政策的不断完善，AI 技术在未来或将成为主流的医学图像计算方式。

1.1 研究背景及意义

1.1.1 非小细胞肺癌介绍

肺癌是最常见的癌症之一，始于肺部，并有极高的可能性扩散到身体的其他器官，它也是全世界癌症死亡的主要原因之一[1]。肺癌患者中大约 80-85%[2]的病例属于非小细胞肺癌（NSCLC）。尽管当前医疗技术在 NSCLC 的治疗方面已经取得了极大的进步，但患者的生存结局很差[3]。NSCLC 患者的生存期长短，具体要看病理分期，例如早期 NSCLC，患者在接受手术治疗以后，还能继续生存 5 年及以上的概率大于 70%，而晚期的 NSCLC 患者的 5 年生存率仅在 23.3%[4]左右。NSCLC 患者治疗关键在于可否实施手术，如果能实施手术，早期肺癌进

行手术切除后，患者的生存期可以延长，甚至可以治愈。传统意义上，肺癌不仅仅只因吸烟而引起，也可能由遗传和环境因素引起，但引起该疾病的确切原因和发病机制在医学上仍不明确。可能是吸烟、空气污染、职业因素、饮食习惯、遗传因素等。另外一些基础疾病也可能使患者感染肺癌的几率提高，比如结核病、慢性支气管肺炎等。目前，使用分期系统的肺癌预后已被研究并证明是不准确的，特别是在肺癌早期。随着精准医疗时代的到来，可以选择的治疗方法越来越多。除了癌症分期、治疗史和癌症特征外，预后方法对于复杂的多学科治疗[5]也很重要。预后预测也被证明在 NSCLC 术后患者的决策过程中起着至关重要的作用[6]。因此，延长 NSCLC 患者生存期迫切需要建立一个新的预后模型[7]。

1.1.2 PET/CT 影像介绍

随着医学研究和临床治疗水平的快速发展，许多医学成像技术被广泛应用于临床肺癌诊断[8]，例如计算机断层扫描(CT)和正电子发射断层扫描(PET)。我们的研究同样使用了 PET/CT 图像，其中 CT 图像作为一种基于高空间分辨率测量组织密度的成像方式，提供了肿瘤的解剖学描述，作为肺癌诊断有条件的常用方法，提供了良好的形态学信息，但在区分器官[9]的良恶性病变方面存在明显的局限性。PET 图像是一种功能成像技术，其特征是人体内脱氧葡萄糖(FDG)摄取的增加[10]，可能比 CT 图像更敏感，因为组织代谢的变化通常先于解剖结构的变化[11]。但 PET 的空间分辨率较差，对病变的解剖定位受到限制[12]。因此，在研究中将 PET 图像与相应的 CT 图像结合，不仅可以同时提供解剖和代谢信息[13]，还可以清晰地了解到整个身体的健康概况，并对早期发现病变提供了科学有力的帮助。与此同时，也有研究证明 PET/CT 成像可以为 NSCLC 患者的生存分析和预后提供有用的信息。例如，Roxani D. Efthymiadou 等人[14]提出的 PET/CT 成像目前广泛应用于肺病变的特征、NSCLC 的分期、复发性疾病的诊断、远端转移的检测、放疗计划和治疗监测。PET/CT 成像已被证明是一种具有成本效益的肺癌评估方法。为了改善 NSCLC 患者的总体生存预测，Amini M 等人[15]在 PET/CT 图像特征级融合和图像本身融合的基础上开发出一个多层次多模态模型。最终结果显示，在预测生存风险方面利用三维小波变换的融合策略可以实现较高的一致

性指数（C-index = 0.708）。在本文中，PET/CT 图像被用来同时成像病理生理变化和病变形态结构，显著提高了预测的准确性。

1.1.3 基因组生物标志物介绍

高通量分子技术的进步为基因组生物标记物的发展带来了巨大的希望，对特定患者的精确医疗成为可能。这些分子生物标志物提供了强大的诊断信息，以及较高的预后意义。同样，医学成像技术也提供了结构、功能和生理形态的组织的关系，通过医学图像来识别肿瘤的特性是诊断、临床分期和治疗计划的重要组成部分，同时图像解释可以让医学成像在个性化医学中发挥作用。因此，开发稳健的、标准化的图像特征用于预测分子特性、预后和治疗响应是必需的，这些标准化的特征可以是从人类观察者那里获得的语义注释的形式，也可以是放射性特征。分子技术的采用可能会随着手术的应用受到成本和投资价值的限制，此外，与分子谱相比，放射组学特征提供了更全面的肿瘤表现。虽然分子图谱仅限于活检区域，可能会导致肿瘤异质性组织的不完全表现，但另一方面，分子技术允许对组织样本中表达的基因进行图谱分析。这种互补的关系表明，结合使用分子和成像的生物标志物对改善患者护理具有潜在的意义[16]。随着 RNA 测序技术的发展，诊断和检测基因相关疾病步入了新的发展阶段，可以对难以获取和分析[17]的疾病进行更准确的分析。近年来，研究表明 RNA-seq 对 NSCLC 患者的临床治疗和风险预测具有重要意义[18]。RNA-seq 可以对疾病进行梳理和判断，尤其对于高危人群的临床治疗以及预测其患病风险具有重要意义。然而，RNA-seq 数据很难收集，因为患者并不经常进行基因检测。因此，准确预测 RNA-seq 表达作为判断 NSCLC 患者预后的标准是至关重要的。

1.1.4 研究意义

随着科技的进步，内外科治疗 NSCLC 的技术也越来越成熟。经过几十年的不断革新，形成了以微创外科手术为主的治疗体系，但总体仍然达不到预期的治疗效果。此外，初期筛查比较困难，大多病人在检查时已处于中后期。目前，临床常用 TNM 描述系统中的国际分期系统(ISS)来对 NSCLC 患者进行预后评价，

尽管对 NSCLC 分期的全球标准进行了修订和改进，但仍不足以准确预测肺癌的预后[19]。且针对由于某些因素不可切除肿瘤而接受化疗的晚期 NSCLC 患者，其体能状态、肿瘤分期等常规因素无法为患者提供预后。此外，即使使用这些明确的临床变量，也观察到预后的差异很大，这种差异很可能是由于肿瘤细胞的不同生物学特性造成的[19]。因为欠缺合理有效的初期检查技术手段，导致 NSCLC 患者的治愈率极低，预后极差。而 RNA-seq 在 NSCLC 预后中的作用日益受到重视，了解 RNA-seq 在 NSCLC 发展中的作用及机制，对于 NSCLC 的预防、诊断和预后有着关键意义。

1.2 研究现状

由于 NSCLC 个性化治疗的技术不断提升，在对患者进行预后评估分析时，其相关预后因素也变得十分重要。尤其是在临床上，与 NSCLC 预后有关的许多因素已经被发现，如 PET/CT 影像特征、临床特征、生物标志物、体积参数、RNA 信息等。下面将这些因素分为与基因相关的和不相关的两大类因素进行介绍，另外，简单介绍关于 NSCLC 患者复发预测的研究进展。

1.2.1 基于非基因标志物在 NSCLC 个体化诊疗中应用的研究进展

Shidan W 等人[25]基于深度学习卷积神经网络（CNN），建立了基于肺癌组织学病理图像的自动检测系统，用来自动检测肿瘤面积，并且将肿瘤面积、形状作为预后风险因素，进行生存分析验证。

Mukherjee P 等人[26]利用浅层 LungNet 神经网络，对来自四个医疗中心的 CT 图像进行分析，并通过此网络来预测 NSCLC 患者的生存率。最后在四个独立数据集的实验结果中，C-index 分别为 0.62、0.62、0.62 和 0.58。除此之外，实验通过生存模型实现迁移训练，并在肺部图像数据库（n=1010）中识别良性和恶性结节，AUC 达到了 0.85。

为了能更好的对 NSCLC 进行生存分析，Wu Y 等人[27]提出了一种多模态融合的深度学习方法（DeepMMSA）。该方法将 CT 影像和临床信息融合并输入到

端到端的 3D ResNet，进行肺癌全自动综合生存分析。实验可以将丰富的医学影像信息和生存数据结合起来，实现个性化的预后和决策。

Kadoya 等人[28]通过标准影像组学特征、同源影像组学特征和肿瘤大小三种类型的影像组学特征，对总生存期的预测进行了研究，实验结果的 C-index 值分别是 0.603、0.625 和 0.607，此实验结果表明基于同源影像组学特征在预测 NSCLC 患者的总生存期方面最具潜力。因此找到适合的特征和构建模型的方法就显得尤为重要。

对于晚期（IIIA-IV 期）NSCLC 患者来说，由于某种因素不能进行手术治疗，只能通过放疗和化疗的方式治疗，Yıldırım 等人[29]在这些患者中随机筛选出 110 名患者[30]进行多因素 Cox 比例风险分析研究，最终得出只有当全部病灶糖酵解（TLG）小于等于 25.7g 时，TLG 可以作为一个独立的预测因素（HR = 7.716，P < 0.05）对晚期 NSCLC 患者进行生存期预测。

Moon 等人[31]在晚期肺腺癌患者中挑选了 234 名在化疗前采取 PET/CT 检验的患者进行多因素 Cox 比例风险回归研究，结果显示，在晚期肺腺癌患者无进展生存期（HR=1.39，P < 0.05）和总生存期（HR=1.65，P < 0.05）的预测中，低指标的 TLG 可以作为一个独立的预测因素，起着决定性作用。

Sharma 等人[32]在治疗前都接受过 PET/CT 检查的 NSCLC 患者中，挑选了 60 名计划以铂金为基础进行化疗的患者，并且开展了前瞻性研究。单因素分析中，NSCLC 患者高肿瘤代谢体积 MTV 大于 120 cm^3和高全部病灶糖酵解 TLG 大于 200 g 的死亡危险比分别为 3.64 和 3.35，其差异具有统计学意义（P < 0.05）；多变量分析显示，MTV 可以作为独立预测因素对 NSCLC 患者总生存期进行分析。

Huang 等人[33]利用 LASSO 和 Cox 回归模型构建了成像组学模型，并使用 282 名第 I 期和第 II 期 NSCLC 患者。最终实验结果发现将 CE_kurtosis_0, CE_uniformity_0_0, CE_homogeneity_45_0, UE_uniformity_45_1.0 和 CE_uniformity_0_1.5 输入模型，在某种意义上是可以作为独立的生物标志物来对 GFS 进行预测，且精准度相比传统的分期系统有明显的提高，在精准医疗方面有着很深远的意义。

Timmeren 等[34]在一项回顾性研究中调查了 NSCLC 患者的总生存期和局部复发情况，该研究基于 4 次治疗后的 CT 数据，包括 1 个研究组（141 名 I-IV 期 NSCLC 患者）和 3 个外部验证组（分别为 94、61 和 41 名患者），但是对总生存期的预测价值仅在 1 次研究中就能得到验证。然而，在验证组中，或许是队列规模小的原因，导致研究者们未能开发出预测局部复发的可成功验证的预测模型。

Bousabarah 等[35]针对 110 名 I、IIA 期 NSCLC 患者的 CT 图像，分析他们的局部控制、无病生存、总生存和局部肺部纤维化损害的情况。分析结果表明，基于影像组学肿瘤体积的分析，对预测早期 NSCLC 患者立体定向放疗（SBRT）治疗后的无病生存和总生存以及局部肺部纤维化情况有很大的帮助。

1.2.2 基于基因数据在 NSCLC 患者预后中的研究进展

Gevaert 等人[36]构建了 26 名 NSCLC 患者的放射基因组关联图，以确定图像特征和元基因之间的成对关联，然后采用稀疏线性回归手段，提出了基于图像特征的元基因预测模型。同样，基于图像特征的元基因的预测模型也被开发出来。最后，当预测的图像特征被映射到一个共同的基因表达数据集时，肿瘤大小、边缘形状和清晰度在预测意义上最为重要。

Emaminejad N 等人[37]在对 8 个图像特征进行研究时，开发了一个简单的贝叶斯网络分类器，基于一个多层感知器以及 2 个基因标志物来实现早期肺癌复发的风险预测，AUC 值为 0.84。

Subramanian V 等人[38]采用具有弹性正则化的线性 Cox 比例风险模型，将 CT 影像和 RNA 测序等基因组数据进行结合，来预测 NSCLC 患者的术后复发情况，同时依靠 C-index 将风险评分的准确性进行量化，并依据 AUC 值来衡量模型分类能力。

Wang H 等人[39]提出了联合标签进行融合的方法，根据不同模态的预测结果产生的相关性，衡量不同嵌入层的全连接神经网络模型的不确定性。最终使用手术切除后 NSCLC 患者 CT 图像和基因数据来预测 1 年生存率的实验研究表明，该方法的性能良好。

Singh A 等人[40]对 85 名 NSCLC 患者的放射基因组学特征进行研究，创建了可以识别高危患者并预测生存时间的模型。通过曼-惠特尼 U 检验确定了 224 个稳定的特征子集（P>0.05），用来检测每个特征的差异（扫描片厚度、重建的核和对比度增强）。从辐射测量和基因组学特征中，选取 10 个主成分因素进行分析研究。最后，采用 5 倍交叉验证和 200 次迭代的 Cox 比例风险模型，来分析评估该模型在预测患者总生存率时的潜力，其中 C-index 为 0.62。此外，还将基于肺癌成像、基因数据的深度学习方法来预测肿瘤复发风险作为未来研究方向。

Aonpong P 等人[41]使用两步法来达到最终的复发预测，第一步先利用 CT 图像和影像组学特征对 74 维基因表达进行预测，第二步利用预测所得的基因来实现低成本高准确度的复发预测。事实证明，该方法对患者复发的预测准确率为 83.28%。但实验不足之处是对 74 个基因分别进行预测，训练了 74 个预测模型，实验较为复杂。

上述研究虽然引入了基因信息，但没有考虑到基因数据有限，获取成本较高这一问题。Aonpong P 等人提出的方法虽然对未知的 RNA 信息进行预测，缓解了数据有限和成本问题，但将 74 个基因特征作为预后研究的因素，维数过高，过程复杂。因此，本研究提出了一种低维的与 NSCLC 患者生存分析相关的预后指标，并采用基于 PET/CT 成像的机器学习方法进行预测，解决基因信息难以获取这一大难题。

1.2.3 非小细胞肺癌患者复发预测研究进展

V. Subramanian 等人[42]将 CT 图像和基因组学数据进行融合，且在线性 Cox 比例风险模型弹性正则化改进的基础上进行复发预测。研究基于近 130 名患者的 NSCLC 放射基因组学数据集进行实验，结果证明 C-index 值增加了 10%。

S. Ali Hosseini 等人[43]使用 PET 影像组学特征和机器学习算法预测肺癌患者的复发。在这项工作中，他们招募了 136 名 NSCLC 患者。手动描绘了五个亚区域或轮廓，并以不同的距离（1，2，3，4 和 5 mm）延伸。使用了三种不同的特征选择方法和多个分类器。其结果表明，具有最小冗余最大相关性（MRMR）特征选择和随机森林（RF）分类器的 contourPlus1mm，具有 MRMR 特征选择和线性判别分析（LDA）分类器的 contourPlus1mm 以及具有递归特征消除（RFE）

特征选择和逻辑回归（LR）分类器的 contourPlus4mm 具有最高的性能（AUC= 0.65）。这项研究的结果表明，手动轮廓的扩展子体积可以提高肺癌患者复发预测的性能。

Y. Ai et al 等人[44]提出了一种低成本、高精度的多层残差感知器（ResMLP）递归预测方法。首先，应用几个提出的 ResMLP 模块来构建深度回归估计模型。然后，通过该模型构建混合特征（手工特征和深层特征）和基因数据的映射函数。最后，利用回归模型获得的基因估计数据学习与复发相关的信息表示，实现复发预测任务。实验结果表明，所提出方法泛化能力强，预测准确率达到 86.38%。

1.3 本文研究内容与章节安排

1.3.1 研究内容与创新点

基于 RNA 信息引导的 NSCLC 生存分析是本文的主要任务，其目的是充分利用 RNA 数据所提供的信息用于 NSCLC 患者的预后，从而提高患者生存率。本文主要研究内容是使用机器学习方法构建预测模型，预测 RNA 信息缺失患者的 RNA 预后指标，使其预后预测更加准确。具体来说，本文首先在影像、临床、18F-FDG PET/CT 代谢体积参数和 RNA 信息等因素下对 NSCLC 患者的研究现状进行了详细综述。而后，对本文所用到的理论知识进行补充说明。最后通过各种生存分析实验以及各因素结果对比，论述 RNA 信息的预后价值，并以 NSCLC 复发预测对比实验为例，充分证明加入 RNA 合成指数后，预测准确度实现了大幅提升。

本文创新点在于：

1.在 NSCLC 患者预后研究中，不同于普遍使用的影像数据、临床信息、相关参数以及基因信息等因素，本研究使用新合成的 RNA 指数作为预后指标对 NSCLC 患者进行预后分析，从而证明 RNA 合成指数作为一种预后因素的意义。

2.鉴于大部分患者不经常进行 RNA 信息检验，基因数据有限。本研究提出 RNA 合成指数预测模型，该模型为回归任务提供了一个灵活框架，实验证明该框架基于患者影像数据可以使未接受 RNA 检测的患者得到与之对应的 RNA 信息合成指数，从而进一步提高患者的预后效果。

3.在 NSCLC 患者复发预测研究中，实验将本研究所建立与预测的 RNA 合成指数作为变量特征，分别对患者进行复发预测，实验证明综合影像、临床信息和 RNA 合成指数特征可以准确预测患者是否复发，相较于未使用 RNA 合成指数特征的预测实验是一个重大突破。

1.3.2 章节安排

本文一共划分为六章，现安排如下：

第一章：绪论。本章首先介绍了 NSCLC 这一疾病的背景以及 PET/CT 影像和基因数据在医学领域中的意义，其次介绍了 NSCLC 患者基于非基因信息与基因信息的生存分析研究现状以及 NSCLC 患者复发预测的研究进展，最后阐述了本文研究内容和章节安排。

第二章主要对传统影像组学特征类型和基本定义进行了阐述。同时，对于本文所提到的特征选择方法做了详细论述。为了让大家更加了解 RNA-seq 测序技术的原理，本文在此章节作了详细描述。此外，在深度学习，机器学习和生存分析方面，分别介绍了其技术原理和发展现状。

第三章描述了本文所使用的数据来源和类型。在数据背景资料的基础上，还对本文研究中所选择的临床数据进行了简要的统计分类。

第四章主要介绍了基于 PET/CT 影像组学特征和 RNA 合成指数的 NSCLC 回归预测模型的构成、影像组学特征提取、RNA 合成指数建立流程和生存分析实验方法等。同时也介绍了各指数预后结果的比较以及 kaplanmeier 曲线分析。实验结果中充分利用 C-index、风险比 HR 指标说明 RNA 合成指数的优越性。

第五章介绍了 NSCLC 复发预测模型。该模型使用普通 3D 深度学习网络，同时将 NSCLC 患者 PET/CT 影像以及临床等特征作为模型的输入。成功验证了 RNA 合成指数对 NSCLC 复发预测的贡献度。

第六章总结了本文中提出的相关工作，讨论了本文工作的贡献与不足，并提出了可行的解决方案。

2. 相关理论介绍

2.1 影像组学及其特征介绍

影像组学特征是一种用于可视化和分析医学图像的多模态分析方法，能够对多种类型影像信息进行复杂整合处理，可以将影像中病灶的各种特性与临床上的病史、既往症状、实验结果等信息联系起来，可以更好地表示个体病症的状况。

影像组学特征主要包括影像融合处理、影像提取和疾病分类三大类。首先、影像融合处理，是指通过将多种模态的影像，包括 CT、MR、PET、NMR 以及多种生物特征，按照设定的空间配准和分类管理，进行融合处理，获得全复合影像信息。其次、影像提取，是指从融合的全影像中获取影像特征，进行影像分割、细胞检测等处理，从而获得与病理特征适配的图像标签和具体定量病变分析结果。第三、病情分类，是根据图像特征提取、影像融合处理以及病变分析结果，对病变类型标签进行识别，并根据不同病情进行分类，从而定性地描述病情状态，以实现个体精准化的治疗以及病变检测结果的可视化。一阶强度特征、形状特征和纹理特征这些定量特征是根据某些计算结果自动计算所得的图像特征，这些特征也是影像组学特征提取[45]的核心。

一阶统计特征[46]是通过计算肿瘤或其他区域图像的灰度值得到的，通常包括一阶统计量的最大、最小、平均、中值、范围、方差、峰度、波动率和熵。一阶统计数据可以体现出肿瘤性质的差异，以及灰度强度在肿瘤内的分布形态。

形状、大小特征反映的是关于肿瘤的形状、大小和规则性的信息；肿瘤的长度、体积大小和面积反映的是关于肿瘤大小信息；椭圆度表明它是否接近球形；而紧凑性表明肿瘤是否为规则形状，其边缘是否规则。

以上介绍的一阶统计量特征和形状大小特征映射的是图像中的低维信息，而纹理特征与一阶统计量特征和形状大小特征具有明显的区别。具体来说，纹理特征是通过纹理矩阵来提取特征，矩阵类型主要包括：灰度级依赖矩阵（GLDM）、灰度级大小矩阵（GLSZM）、灰度级共生矩阵（GLCM）、以及灰度级运行长度矩阵（GLRLM）和邻域灰度差矩阵（NGTDM）。灰度级依赖矩阵量化了图像的中心像素或体素与其周围环境之间的依赖关系。灰度区域大小矩阵是一个以

元素的行和列存储灰度和大小区域（由相同灰度连接的体素）数量的矩阵。灰度共生矩阵[47]也叫二阶直方图特征，其行数和列数代表单元格中的灰度值处于角度或距离这些给定关系的次数。灰度运行长度矩阵[48]，其每个元素（i，j）描述灰度在给定方向的连续迭代次数。邻域灰度差矩阵[49]，是存储在行和列的元素中的灰度和大小区域（与同一灰度相关的体素）的数量矩阵。

虽然上述三种类型的特征从肿瘤的各个维度层中提取视觉和纹理信息，但从这些维度中提取的视觉纹理信息量非常有限。为了能够获取到来自不同频域中的信息，小波变换[50]显得尤为重要，它可以依据不同的频域分割原始肿瘤图像，然后分别从各个频域中提取上述提到的一阶统计量特征、形状大小特征和纹理类型的特征。

总体来说，通过影像组学特征的运用可以快速比较和整合不同模态的图像信息，并可以更好地对个体病情状况进行识别，有利于即时获取病情变化及治疗结果研判，从而改善临床诊疗水平，避免病情发展不良，提高治疗疗效。

2.2 特征选择方法介绍

特征选择是指根据评价标准直接选择一个合适的属性子集，或者通过以线性或非线性方式组合原始属性集，然后从新的属性集中选择一个合适的子集来生成一个新的属性集。三种常用的特征选择方法是过滤法、包裹法以及嵌入法。

2.2.1 过滤法

过滤法是一种基于特征和分类标签之间对应关系的特征选择方法[52]。基于过滤法的特征选择是最简单和最普遍的方法之一，它最大的优点是不以模型为基础，只考察特征的价值，对其进行排序和选择。事实上，基于过滤器的特征选择方案，其本质是特征排序，即通过对特征的价值排序，可以实现对任何部分或数量的特征选择或排除。显然，这里的关键环节是如何估计属性的价值，从而实现它们的排序。表 2.1 列出了各种过滤法的说明情况。

表 2.1 过滤法总结表

类	说明
方差选择法	通过设定方差阈值，选出大于阈值的特征。
卡方检验	捕捉相关性用于分类算法，追求 p 值小于 0.05 的显著性水平。
F 检验（分类/回归）	要求数据服从正态分布，只能捕捉线性相关性，追求 p 值小于 0.05 的显著性水平。
互信息（分类/回归）	捕捉任何相关性，选取互信息估计大于 0 的特征，不能用于稀疏矩阵。
皮尔逊相关系数	只能捕捉线性相关关系，追求 p 值小于 0.05 的显著性水平。

对于方差选择法，当方差作为评价特征的标准时，如果特征的价值没有明显的差异，一般认为这个特征对样本区分没有明显的贡献，因此，方差低于阈值的特征被排除在特征构建过程中。

以卡方检验统计量作挑选特征的标准，卡方检验值越高，相关性越高（卡方检验是一种估计自变量和因变量相关性关系的统计量）。

F 检验实质上是分析两组数据之间的线性关系，初始假设是 "数据之间没有显著的线性关系"。这就产生了两个统计值——F 值和 P 值。与卡方过滤类似，我们要选择 P 值小于 0.01 或 0.05 时的因素，这些因素与标签有显著的线性关系。

互信息法与卡方检验法相同，都是评价自变量对因变量的相关性。互信息用以计算两个特征或自变量与因变量之间所共有的信息，能够衡量各种相关性的特征集，计算相对复杂。互信息量计算公式包括离散随机变量 X1 和 Y1 以及连续随机变量 X2 和 Y2。公式分别为：

$$I(X1,Y1)=\sum_{y\in Y1}\sum_{x\in X1}p(x,y)log\left(\frac{p(x,y)}{p(x)p(y)}\right) \tag{2.1}$$

其中 p(x，y)是 X1 和 Y1 的联合概率分布函数，P(x)和 p(y)分别为 X2 和 Y2 的边缘概率分布函数。

$$I(X1,X2)=\int_{Y2}\int_{X2}p(x,y)log\left(\frac{p(x,y)}{p(x)p(y)}\right)dxdy \quad (2.2)$$

其中 p(x，y)是 X2 和 Y2 的联合概率密度函数，p(x)和 p(y)分别是 X2 和 Y2 的边缘概率密度函数。

皮尔逊相关系数是了解特征和反映变量之间关系的最简单和最有用的方法之一。它测量变量之间的线性相关关系，并给出一个数值范围[-1，1]，其中-1 表示完全负相关，+1 表示完全正相关，0 表示没有线性相关。

2.2.2 包裹法

包裹法是一种根据特征的预测能力来选择特征的方法，通常与给定的分类器结合使用。它可以选取对预测贡献度较高的特征，并排除预测能力最小的特征，直到提取了一定数量的特征。一个常用的反向比较方法是递归特征消除法（RFE）[51]。RFE 的稳定性很大程度上取决于迭代时，底层用的哪种模型。该方法的主要思想是反复构建模型，然后选出最好的（或者最差的）特征（根据系数来选），把选出来的特征放到一边，然后在剩余的特征上重复这个过程，直到遍历了所有的特征。在这个过程中被消除的次序就是特征的排序。

2.2.3 嵌入法

嵌入法是一种让算法自己决定使用哪些特征的方法，即特征选择和模型训练同时进行。在使用嵌入法时，我们先使用某些机器学习的算法或模型进行训练，得到各个特征的权值系数(0-1 之间)。这些权值系数往往代表了特征对模型的贡献或者说重要性。因此相比于过滤法，嵌入法的结果会更加精确到模型的效用本身，对提高模型效果有更好的作用。

2.3 RNA-seq 转录组测序技术介绍

RNA-seq 转录组测序技术出现于十年之前，自其诞生之日起，RNA-seq 就成了研究分子生物学的普遍工具，这项技术几乎构成了我们对基因组功能的认知基

础。RNA-seq 更广泛的应用已经促进了我们对生物学多方面的理解，例如通过信使 mRNA 剪接、非编码 RNAs 和增强子 RNAs 对基因表达的调控。RNA-seq 的应用和进步是由技术发展驱动的，相对于以前的基因芯片，RNA-seq 这种方法对 RNA 生物学和转录组产生更丰富并且偏见更小的信息。到目前为止，从标准的 RNA-seq 方法衍生而来的各种 RNA-seq 方法有近百种。Illumina 的短读长(short-read)测序平台能对这些由大部分不同方法的 RNA-seq 构建的文库进行测序，但是最近长读长(long-read)RNA-seq 的与直接 RNA-seq 测序(direct RNA sequencing，dRNA-seq)的进步已经能够解决以前研究人员使用短序列手段无法解决的一些问题。

RNA-seq 转录组测序技术是二代测序技术，研究特定细胞在某一功能状态下所有 RNA 的功能，主要包括 mRNA 和非编码 RNA。RNA-seq 转录组测序技术能够全而快速地获得某一物种特定组织或器官在某一状态下的几乎所有转录本序列信息，已广泛应用于基础研究、临床诊断和药物研发等领域。

RNA-seq 测序是一种可以用于研究基因组学、转录组学等领域的测序技术。它可以用来研究基因的表达模式、基因的功能和蛋白质的表达，以及物种特异性的基因组变异。RNA-seq 作为一种高通量的测序技术，可以同时测量数以万计的基因，并识别基因组中的差异，从而获得更多的信息。它可以被用来研究物种之间的差异，以及物种内部基因表达的变化。

RNA-seq 的原理是利用多种步聚的流程，即从 RNA 到 DNA 的反转录编码，然后经过克隆和序列测定，最后分析和解读。首先，将 RNA 制备好，进行反转录编码，将 RNA 转换成 DNA，从而产生 CDNA(反转录聚合酶)库。然后，将 CDNA 库进行克隆，以获得可复制和测序的 CDNA 片段。最后，将这些 CDNA 片段进行序列测定，并通过相应的软件和算法来解读序列，最终得到 RNA-seq 的结果。

RNA-seq 测序技术具有高通量，高灵敏度，低成本等优点，可以更有效地研究基因组学、转录组学等领域，可为基因组学、转录组学研究提供有价值的信息。

2.4 回归算法介绍及评价指标

回归是一个统计学研究过程，用于估计自变量和因变量之间的关系，以及研究不同自变量对于因变量影响的程度大小。回归算法的优势在于简单直接、训练速度快，而主要缺陷是要求严格的假设、需要处理异常值。常见回归算法主要包括以下几类：线性回归算法、多项式回归算法、支持向量回归算法、决策树、随机森林、LASSO、岭回归、XGBoost 等回归算法。

2.4.1 回归算法

线性回归[53]通常是人们学习机器学习和数据科学的第一个算法。线性回归算法是一种描述自变量与因变量线性相关的模型，它假设输入变量 x 和输出变量 f（x）之间存在一定的线性关系。基本公式为：

$$f(x)=\omega_1 x_1+\omega_2 x_2+\cdots+\omega_d x_d+b \tag{2.3}$$

$$f(x)=\omega^T x+b \tag{2.4}$$

多项式回归[54]样本数据属性呈非线性关系(曲线)，相较于线性回归，多项式回归打破了属性上的线性，即将属性的次幂看成独立的特征。基本公式为：

$$y(x,\omega)=\omega_0+\omega_1 x+\omega_2 x^2+\cdots+\omega_M x^M=\sum_{j=1}^{M}\omega_j x^j \tag{2.5}$$

基于支持向量机（SVM）进行回归的算法称为支持向量回归（SVR）。支持向量回归[55]是一种常见的监督学习算法，它的基本思路是找到拟合度最佳的曲线或面。决策树[56]是一种非参数型监督学习方法，主要用于分类和回归的任务，其目标是训练一个能从数据特征断定出简单决策规则的模型，来预测相对应目标变量的值。随机森林回归[57]思路上与决策树回归十分相似。可以把随机森林看做是一个元估计器，然后利用数据集的各种子样本来拟合多个决策树，并通过不断调整来控制过拟合和提高预测准确度。随机森林回归器通常在分类中表现更好，

但在回归任务中与决策树相比不确定哪一个表现更好，因为树构造的算法在本质上存在一定程度的过拟合或者欠拟合权衡。LASSO 回归[58]是使用收缩线性回归的变体。收缩是将数据值收缩到中心点作为平均值的过程。这种类型的回归非常适合显示重度多重共线性（特征相互之间高度相关）的模型。岭回归[59]与 LASSO 回归非常相似，因为这两种技术都使用了收缩。岭回归（Ridge）和 LASSO 回归都是非常适合显示重度多重共线性（特征相互之间高度相关）的模型。它们之间的主要区别在于 Ridge 使用 L2 正则化，它的系数只是接近于零，不会像 LASSO 回归那样变为零。ElasticNet 是将 LASSO 和 Ridge 的回归技术进行混合的方法。XGBoost[60]是一种高效且有效的梯度提升算法的实现。

2.4.2 回归算法评价指标

均方误差（MSE）是回归算法中最常见的损失函数。它通过计算每个样本预测值和实际值之间差的平方，以及对这些值进行求和求平均来评估数据的变化程度。该值越小，预测模型对测试数据的描述就越准确。均方根误差（RMSE）是均方误差的开平方。平均绝对误差（MAE）表示的是每个样本预测值和真实值之间差的绝对值，然后将这些值相加并取平均值。而决定系数（R^2）反映了模型拟合数据的准确性，其数值通常在 0 到 1 之间。数值越接近 1，代表方程中对应变量有着较强的解释能力，而模型对数据的拟合程度也就越强。各评价指标公示如下：

$$MSE(y,\hat{y})=\frac{1}{m}\sum_{i=1}^{m}\left(y_i-f(x_i)\right)^2 \tag{2.6}$$

$$RMSE(y,\hat{y})=\sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(y_i-f(x_i)\right)^2} \tag{2.7}$$

$$MAE(y,\hat{y})=\frac{1}{m}\sum_{i=1}^{m}\left(\left|y_i-f(x_i)\right|\right) \tag{2.8}$$

其中，y_i为真实值，$f(x_i)$和$\hat{y}$为模型的预测值。

$$R^2(y,\hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{MSE}{Var} \quad (2.9)$$

其中，$\bar{y}$是 y 的平均值。

2.5 深度学习介绍及分类模型评价指标

2.5.1 深度学习概念

深度学习是机器学习领域中的一个方向，它被引入机器学习是实现人工智能（AI）的必经路径[61]。目前，在多层神经网络模型的基础上可运用多种深度学习算法来解决多种类型的问题，例如图像、文本和语音等。同时，深度学习在搜索识别技术、数据挖掘技术、机器学习问题、机器翻译等相关领域，都在不同程度上取得了不错的结果，远远超过先前传统技术[62]。从大类来说，深度学习神经网络，其核心是掌握样本数据的内部规律和表达层次[63]，旨在利用层次网络提取层次化的特征信息，从而解决先前需要人工才能提取特征信息的重要问题。

卷积神经网络（CNN）是典型的深度学习算法之一，属于深度结构的前馈神经网络（FFN），涉及卷积计算[64-65]。它与其他神经网络非常相似：具有可以学习的权重和偏差形式的参数神经元形成。但 CNN 的一个差异特征是它们明确假设条目是图像，这允许我们对体系结构中的某些属性进行编码以识别图像中的特定元素。卷积神经网络架构通常包括卷积层、池化层、全连接层和激活函数等。下面对以上 4 种基本结构进行介绍。

1.卷积神经网络包括许多个由卷积单元组成卷积层，而任一卷积单元的参数设置都是经过反向传播算法并且在进行不断优化后得到的。在层级上，网络前段的卷积层只会提取到图像的低层次特征，例如图像的边缘、线条和图像的角落范围。而网络更深层次的卷积层才可以提取到较为复杂的图像特征。网络中卷积层的任务是从输入数据中提取不同层次的特征，它包含一些不同尺寸大小的卷积核，这在文学术语中被称为 "感受野"，与前馈神经网络中的神经元较为相似，每个元素都会有一个权重数值、一个偏置大小与之对应。卷积核的工作原理是有

规律地遍历输入特征，将接收域中的矩阵元素相乘再相加，以求得输入特征的总和，并在最后叠加偏置量的大小。

2.池化层实质上是一种形式意义的降采样。一般来说，池化层都会周期性地插入于 CNN 的卷积层之间。在多种不同形式的非线性池化函数中，“最大池化（Max pooling）”是最普遍的一种。它通过将输入的图像划分成若干个矩形区域，然后对每个子区域输出最大值。直观来讲，这种机制之所以能够有效，是因为一旦发现一个特征之后，它的精确位置就会远远不如它相对于其他特征位置的关系重要。池化层会不断的对数据空间进行缩减，所以参数的数量和计算量也会减小，这也在一定程度上控制了过度拟合。

3.位于整个网络尾部的全连接层（Fully Connected Layer），它之所以被称之为全连接层，是因为每个节点与上一层全部相连，这一特征导致全连接层的参数也是最多的。全连接层的主要作用是将上一层卷积、池化等计算得到的特征映射到样本标记空间。通俗讲就是将最终特征表示合成一个一维向量。常见的池化层包括最大池化、平均池化和全局池化[66]等。

4.激活函数在神经网络中用于引入非线性因素，凭借激活函数，神经网络就能经各种曲线拟合。当不使用激活函数时，那么每一层输出都是前一层输入的线性函数，不管神经网络的层数有多少，输出都会是输入特征的线性组合。当使用激活函数时，激活函数给网络中的神经元引入了非线性函数，这样神经网络就可以应用到更多的非线性模型中，从而提高了神经网络表达模型的能力，解决了用线性模型无法解决的问题。卷积神经网络中常用的激活函数有 Tanh 激活函数、Sigmoid 激活函数、Softmax 激活函数和 Relu 激活函数等。

2.5.2 分类模型评价指标

分类模型评价指标的重要性不亚于设计一个好的网络模型，只有通过合理的评价指标，才能衡量一个模型的好坏和选择一个合适的模型。在医学领域分类问题中，经常会用到的评价指标包括：准确度（Accuracy）、精准率（Precision）、召回率（Recall）、ROC 曲线下的面积（AUC）。

1.准确度（Accuracy）是指预测正确的样本在所有样本中比例。公式如下：

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \quad (2.10)$$

2.精准率（Precision）又叫查准率，它主要是针对预测结果来判定，即实际为正样本而被预测为正样本的概率。公式如下：

$$Precision = \frac{TP}{TP + FP} \quad (2.11)$$

3.召回率（Recall）又叫查全率，它主要是针对原样本来判定，即被预测为正样本在实际中为正样本的概率。公式如下：

$$Recall = \frac{TP}{TP + FN} \quad (2.12)$$

以上 TP 表示本身是正例预测也为正例的样本数；TN 表示本身为反例预测也为反例的样本数；FN 表示本身为正例而预测为反例的样本数；FP 表示本身为反例而预测为正例的样本数。

4.AUC 为 ROC 曲线下的面积，其反映的能力是指分类模型对正负样本的分类能力，尤其是在样本不平衡的条件下也可以进行合理且准确的评估，它已经被各种医学分类研究作为评价指标之一，获得广泛应用。AUC 值的大小位于 0 到 1 之间，如果超过 0.5 则表示模型分类的预测能力要优于随机猜测的结果，也可以说其值与 1 越接近，则分类性能越好[67]。

2.6 生存分析

2.6.1 生存分析概念

生存分析(survival analysis)是一种对生存数据(survival data)的统计分析，经常被用来探索生存因素与生存时间的发展模式以及规律。例如，分析某种药物的治疗效果、手术治疗后的存活率、医疗设备的寿命等。对生存信息和资料的分析被称为生存分析。所谓生存信息和资料是描述预期寿命或事件发生时间的数据。具体来说，人类的生存与一些因素有关，研究这些因素是否以及在多大程度上与生存有关，称为生存分析。

2.6.2 生存分析的具体方法与 Cox 回归分析评价指标

生存分析方法包括描述法、非参数法、参数法和半参数法。

1.描述性方法是以样本观察的信息为基础，用公式直接计算出每个时间点或时间间隔的生存函数、死亡率函数和风险函数，并以列表或图表的形式显示生存时间分布的规律。

2.非参数法常用到的生存分析方法是乘积极限法、寿命表法。在估计生存函数时通常不考虑生存时间的分布，就可以检验各种风险因素对患者生存时间的影响。

3.参数法从观测样本中估计假设分布模型的参数，得到生存时间的概率分布模型。

4.半参数方法不需要假设生存期的分布，但可以使用模型来分析生存期的分布和风险因素对生存期的影响，特别是 Cox 比例风险回归模型。Cox 回归模型不反映生存函数与自变量的关系，而是以风险函数为中间变量或因变量，利用生存函数与风险函数的关系间接反映生存函数与自变量的关系。公示如下：

$$h(t)=h_0(t)exp(\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p) \quad (2.13)$$

X 表示与生存时间相关的影响因素，也是自变量，在整个观察期内不随时间的变化而变化。

h(t)表示时刻 t 时的风险函数。

β 表示各自变量的偏回归系数，是需要根据实际的数据来估计的未知参数。

在 Cox 回归模型评价指标中，常见的 C-index 指标主要计算的是 Cox 模型预测值与真实值之间的区分度，与分类任务中的 AUC 指标极为相似，常用在患者预后模型的评价中，例如精度预测。一般情况下，C-index 值如果在 0.50-0.70 则表示准确度较低；当 C-index 的值在 0.71-0.90 之间表示准确度适中；而高于 0.90 则表示高准确度。同样由 Cox 风险比例模型衍生出来 HR 值，也叫风险比，用于估计因为某种因素的存在导致结局事件风险变化的倍数。回归系数的大小与正负时用来判断是正向还是负向影响。 Wald 是一个卡方值，用于对回归系数进行检

验，考察回归系数是否等于 0。而根据回归方程（常用 log-rank）得到的 p-value 值小于 0.05 的时候，这个模型才有意义。

生存分析的主要目的是研究与分析影响生存时间长短的因素，常用的方法还有 kaplanmeier 曲线法。kaplanmeier 曲线又称生存曲线，是一种生存分析的常用方法，主要分析单一因素对生存期的影响，通过绘制生存曲线来估计患者生存率和单一因素之间的关系。生存曲线横轴代表生存时间，纵轴代表患者的生存率，最后绘制而成的曲线呈连续型的阶梯形，用以说明单一因素对生存时间与生存率之间关系的影响。生存曲线在一般情况下是光滑且水平向右延伸的，当某个时间点一旦发生终点事件(如有患者死亡)，该时间点的曲线就会垂直向下降落，下降幅度代表该时间点发生终点事件的患者个数与上一个时间节点后随访的样本数量的比值。

2.7 本章总结

本章节首先对影像组学概念和构成进行简略介绍，之后从三方面介绍了特征选择算法。为了能更深层次了解 RNA-seq 测序数据，本章节详细介绍了其提取原理。此外，本章还对深度学习的概念、原理、常用分类指标及回归算法概念和评价指标以及生存分析方法进行简单介绍。

3. 实验数据介绍

3.1 数据来源介绍

本研究使用的是 NSCLC[68]放射基因组学数据集，现已在癌症影像档案[69] (TCIA) 中公开访问。该数据集是从 NSCLC 队列中收集的 162 名受试者，每个受试者的数据可以分为三种类型，包括图像数据、临床数据和基因数据 [16, 33]。

3.2 数据类型介绍

3.2.1 图像数据

在 162 名受试者空腹氟脱氧葡萄糖 18F-FDG PET/CT 数据中，斯坦福大学医学中心使用 GE Discovery D690 PET/CT 进行 PET/CT 扫描，而帕洛阿尔托退伍军人事物医疗系统 Palo Alto VA 使用 GE Discovery PET/CT 扫描仪进行 PET/CT 扫描。FDG 剂量和摄取时间分别为 138.9–572.3 MBq（平均 309.3MBq）和 23.1–128.9 分钟（平均 66.6 分钟）。图像对颅底至大腿中部覆盖，并在必要时提供额外的点视图。每个床位是 1-5 分钟采集，取决于体重。肿瘤分割的金标准由兰州大学第二医院有 5 年以上工作经验的胸放射科医师依据 CT 图像提供。由于显像剂剂量会严重影像 PET 影像灰度值，所以需要对像素值进行校正，降低其影响。本研究基于体重将 PET 代谢图像归一化为标准摄取值。一些图像数据样本和肿瘤掩模区域如图 3.1 所示。

3.2.2 基因数据

基因信息均是外科医生从手术过程中无药物志愿者的肿瘤样本中收集，他们沿着肿瘤组织的最长轴切出一个 3-5 毫米厚的切片，并在切除后 30 分钟内冷冻，然后使用 RNA-seq 测序技术进行 RNA 提取，以产生分子表型。130 个组织样本由 HiSeq 2500(Illumina)按照制造商的说明分 3 批进行了测序：16、66 和 48。测序数据由中心生物科学公司进行预处理，并以每千碱基转录本每百万(FPKM)的

片段估计基因表达。最终测序技术为 130 个患者中的每个患者提供了 22,127 个 RNA-seq。由于大多数基因数据在部分患者中表达缺失，表现为 N/A。因此从我们的工作中移除不明确的基因表达后，每个患者有 5,268 个基因可用。此外 162 个成像数据中有两个患者头文件信息不正确。因此，将剩下的 160 名患者分为训练组（108 例有 RNA 序列数据中 39 例在随访期间死亡）和测试组（52 例无 RNA 序列数据中心 15 例在随访期间死亡）。

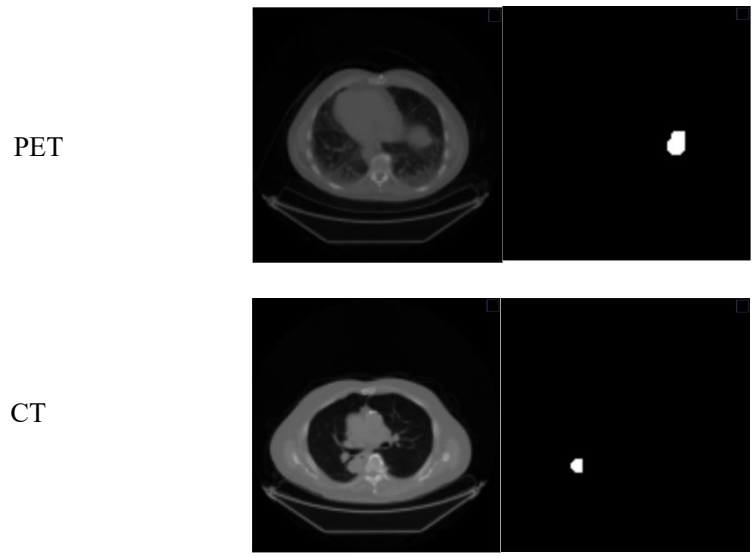


图 3.1 图像数据的两个例子(a) PET/CT 图像数据和(b)肿瘤

3.2.3 临床数据

根据 160 名患者选择具有完整且有效的临床数据类型，包括年龄、性别、吸烟状况、组织学类型、放疗、化疗和复发。数据集的详细信息如表 3.1 所示。从表中可知，患者年龄大部分集中在 61-70 岁之间，吸烟状态中显示大部分人已戒烟，腺癌与鳞癌之比大约为 5:1，复发人群只占总人数的 28.1%。

表 3.1 临床样本统计

Characteristic	Train Statistics	Test Statistics	Characteristic	Train Statistics	Test Statistics
Overall	108	52	**Chemotherapy, N(%)**		
Age, mean(SD)(year)	68.9(8.7)	67.7(11.6)	Yes	32(29.6)	9(17.0)
Age:categorical, N(%)			No	76(70.4)	43(81.1)
50 and younger	4(3.7)	4(7.7)	Not collected	-	1(1.9)
51-60	10(9.3)	3(5.7)	**Radiation, N(%)**		
61-70	49(45.4)	24(46.2)	Yes	12(11.1)	3(3.9)
71 and older	45(41.6)	21(40.4)	No	96(88.9)	49(94.2)
Gender, N(%)			Not collected	-	1(1.9)
Male	84(77.7)	28(53.8)	**Recurrence, N(%)**		
Female	24(22.3)	24(46.2)	Yes	34(31.5)	11(20.8)
Smoke status, N(%)			No	74(68.5)	41(77.3)
Former	73(67.6)	28(53.8)	Not collected	-	1(1.9)
Current	18(16.7)	8(15.4)	**Survival status, N(%)**		
Nonsmoker	17(15.7)	16(30.8)	Dead	39(36.1)	15(28.8)
Histology, N(%)			Alive	69(63.9)	37(71.2)
Adenocarcinoma	81(75.0)	49(94.2)	**Survival time(day), median:**	1282	1081
Squamous cell carcinoma	23(21.3)	1(1.9)			
NSCLC(NOS)	4(3.7)	2(3.9)			

3.3 本章总结

本章节对本文中用到的数据进行介绍，包括影像数据预处理、基因数据预处理和临床统计信息三大部分。同时对于筛选出的应用于本研究的 160 例 NSCLC 患者数据集做了临床样本统计。

4. 非小细胞肺癌患者 RNA 合成指数预后研究

4.1 基于 RNA 合成指数的 NSCLC 患者的生存分析方法

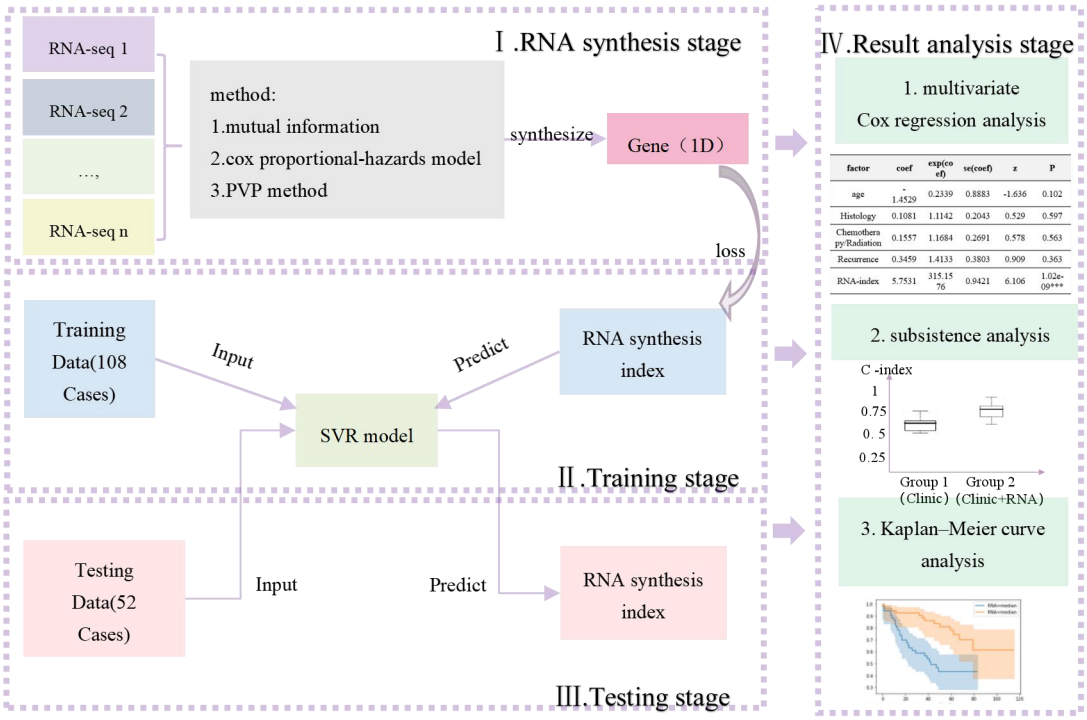
本研究首先建立 RNA 合成指数，其次建立 RNA 合成指数预测模型对患者 RNA 合成指数进行预测，最后针对建立与预测的 RNA 合成指数进行生存分析。

4.1.1 基于 RNA 合成指数的 NSCLC 患者的生存分析研究流程

本研究的工作流程如图 4.1 所示。该方法包括四个阶段：RNA 合成阶段、训练阶段、测试阶段和生存分析阶段。第一个阶段基于互信息法和 Cox 单因素法从高维 RNA-seq 中提取少量 RNA-seq（$p<0.01$）数据，并利用 PVP 合成数据方法，将 Cox 多因素分析回归系数作为权重，使得少量特征合成为 1 维，即 RNA 合成指数。训练和测试阶段基于 108 例患者的 PET/CT 图像建立了回归预测模型。模型使用 Sklearn 库中 SVR 模型包（核函数为“sigmoid”，“gamma”变量为“auto”）。实验在验证集上保存性能最佳的权重进行测试，即利用训练好的模型预测 52 名患者的 RNA 合成指数。最后一阶段包括多因素 Cox 回归分析、生存分析和 kaplanmeier 曲线分析。训练和测试阶段使用的是 NSCLC 公共放射基因组数据集[17]来进行实验。该方法的重要思想是：训练期模型是由成对的基因数据和 PET/CT 影像组学特征进行训练，这意味着训练后的 SVR 模型可以代表 PET/CT 图像与其基因表达之间的某种关系。因此，即使我们在测试阶段没有基因数据，我们也可以从 PET/CT 影像组学特征中估计一些基因信息。

4.1.2 影像组学特征提取过程

本研究从开源的 Python 软件包 "Pyradiomics"（https://pyradiomcs.readthedocs. io/en/latest/feature.html）中提取影像组学特征。该方法将 3D 医学图像(大小为 128*128*64 的肺部切片)和肿瘤 Mask 用作 PyRadiomics 的输入，对原



Data are 26-dimensional PET/CT radiomics features

图 4.1 基于 RNA 合成指标的 NSCLC 患者生存分析总体架构，包括 RNA 合成阶段、训练阶段、测试阶段和结果分析阶段四个阶段。

图和经过滤波的派生图像进行影像组学特征提取，其中图像滤波器包括“Original”和“Wavelet”。对于每个过滤器，计算了七个特征类，即一阶图像强度统计、形状特征、灰度共生矩阵、灰度偏移矩阵、灰度区域大小矩阵、相邻灰度差异矩阵和灰度相关矩阵。最终从每个病人的 PET/CT 图像的感兴趣区域提取了 825 维和 837 维的人工放射学特征。为了减少变异的可能性并提供更稳定的评估，本研究对 PET、CT 影像组学特征进行筛选，如图 4.2 所示。分别选用 F-test 方法挑出 p-value 值小于 0.05 的 7 维 CT 和 19 维 PET 特征。F 检验作为过滤式挑选特征的方法，既可应用于分类任务，也可以应用于回归任务，分类时(标签离散)使用 feature_selection 库中的 f_classif 函数；回归时(标签连续)使用 feature_selection 库中的 f_regression 函数。与卡方检验一样，这两个方法都要和 SelectKBest 连用，F 检验对于正态分布的数据表现更好，所以在进行 F 检验之前可将数据归一化或标准化。此处特征挑选的实施过程在训练数据集进行，最后将挑选出的特征应用在测试数据集中，目的是防止信息泄露。

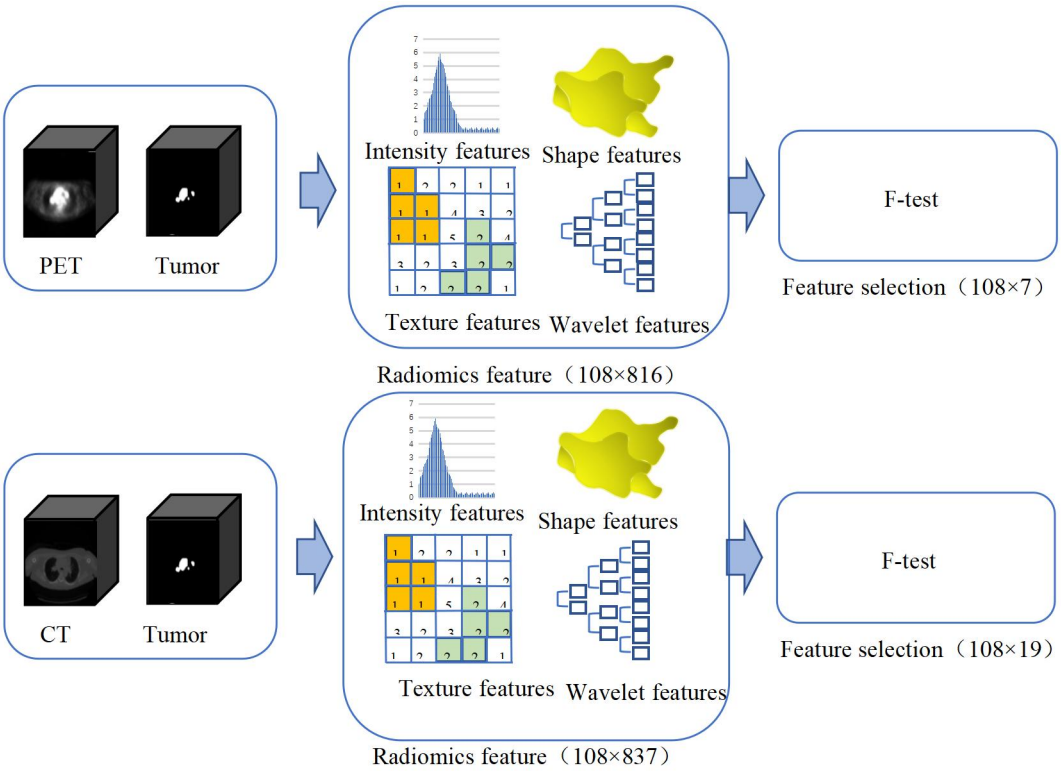


图 4.2 PET/CT 放射组学特征提取流程图

4.1.3 RNA 合成指数建立的具体方法

由于 RNA-seq 是一个相对较大的数据集，每个患者包含超过 5000 个基因数据，其中一些与生存无关，过多的基因会增加计算成本，减少递归预测结果。为了不舍弃这些特征的预后价值且方便研究，本研究引入一种策略来建立一个 RNA 预后指数（1 维），该指数整合了多个预后变量的预后能力，以实现多种 RNA 信息的降维和整合，用于 NSCLC 患者的预后评估。如图 4.1(I)所示。该方法通过特征提取，粗略地选择了与生存时间相关的基因特征，并利用 PVP 方法将多维特征合成为一维。PVP 方法是指使用产生风险比的回归系数作为权重来制定 PET/CT 体积预后（PVP）指数。Yonglin Pu[70]等人提出的 PVP 指数，为临床医生结合 MTVWB 和 TNM 分期的预后价值提供了一种实用的手段，与目前的 TNM 分期系统或代谢肿瘤负荷相比，对 NSCLC 患者的总生存提供了显著的、更好的预后准确性。本研究也利用 PVP 方法的原理来建立 RNA 合成指数，具体步骤如下：

首先利用互信息法（MI）从 5268 个 RNA-seq 中初步挑选出 3731 个具有非零系数的相关基因。近年来，MI 经常被用作特征选择方法的标准。它既能够反映两个变量的线性关系，也能够反映非线性关系。理论上，MI 是基于两个变量的熵或概率密度函数（pdfs）来表示的。在机器学习的广泛应用中，MI 估计方法已被证明可作为回归问题特征选择的标准[71]。互信息不像其他方法，返回 p-value 值或 F 值等相类似的统计值，而是基于每个特征与对应标签返回它们之间的互信息统计值，返回值的大小在 0-1 之间，完全不相关时值为 0，完全相关时值为 1，互信息也是要与 SelectKBest 函数相结合使用。

其次，使用 Cox 单因素回归分析挑出 9 个 p-value 小于 0.01 的 RNA-seq，包括 CLO4A1、HNRNPA0、IMPDH1、NES、PANK2、PYGB、SERPINE1、SETD3、SLC22A23。P-value 值是衡量特征与样本生存是否显著相关的一大标准。

最后，使用 Cox 多因素回归系数作为权重，将 9 维 RNA-seq 合成 1 维。公式如下：

$$\begin{aligned} RNA-index \;=\; & 1.671\times CLO4A1 + 1.1779\times HNRNPAO \;+ \\ & 0.404\times IMPDH1 + 3.1678\times NES + 1.3272\times \\ & PANK2 \;+ 2.1152\times PYGB + 1.9252\times SERPINE1 \\ & + 6.0416\times SEDT3 + 1.4484\times SLC22A23 \end{aligned} \tag{4.1}$$

4.1.4 SVR 回归模型训练与预测方法

支持向量回归 SVR 是 SVM 对回归问题的一种应用，SVR 的基本思路是：基于 SVR 回归模型对于实验给出的训练集进行训练时，允许模型输出值 f(x)与真实值 y 之间的偏差最多有 ε，数据在间隔带内不计算损失，当且仅当输出值与真实值之间的差距绝对值大于最大误差才对损失进行计算，可以通过最小化总损失值和最大化间隔带的宽度来优化模型[72]。公式中引入松弛变量系数 $\xi_i \geq 0$，$\xi_i^* \geq 0$，SVR 的最优化问题[73]可表示为公式(4.2)：

$$\begin{aligned} & min_{(w,b,\xi)}\ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i + \xi_i^* \\ & s.t. y_i - w^T x_i - b \leq \varepsilon + \xi_i \\ & w^T x_i + b - y_i \leq \varepsilon + \xi_i^* \\ & \xi_i \geq 0, \xi_i^* \geq 0, i = 1,2,3\cdots n \end{aligned} \tag{4.2}$$

引入拉格朗日乘数，经过一系列求解与对偶，求得线性拟合函数为公式（4.3）：

$$f(x)=w^{T}x+b=\sum_{i=1}^{n}(a_i-a_i^{*})K(x_i,x)+b \tag{4.3}$$

式中： a_i， a_i^{*} 为拉格朗日乘子；$k(x_i, x)$ 为核函数。

在 160 名患者中，本研究将 108 例有 RNA-seq 数据的患者作为训练组，52 例无 RNA-seq 数据的患者作为测试组，实验通过训练 SVR 回归预测模型来建立 PET/CT 放射组学特征和 RNA 合成指数之间的映射函数关系，如图 4.1（II）所示，在测试阶段可以直接预测 52 例无基因信息患者的 RNA 合成指数。

SVR 模型使用 Sklearn 库中的 SVR 函数实现，模型中的自由参数是正则化系数 C 和停止拟合容忍度 epsilon，其中核函数 kernel 使用的是“sigmoid”，核函数系数 gamma 变量设置为“auto”，停止拟合容忍度 tol 变量设置为 0.1，其余变量均使用默认参数。

4.1.5 生存分析实验

生存分析的基本因素包括 4 个临床特征，即年龄、组织学类型、复发、放化疗（放疗化疗合一）。本研究将关键因素 RNA 合成指数作为新的因素纳入生存分析研究。基于 Cox 比例风险模型分别报告了训练数据和测试数据中 RNA 合成指数的 C-index 和普遍的危险因素，同时绘制 kaplanmeier 生存曲线来进一步对实验结果进行解释。研究方法使用 R 软件中的“survival”和“survcomp”包 。具体实验内容包括：

1.针对训练数据的 Cox 多因素回归分析，以及有无 RNA 合成指数 Cox 回归分析所得 C-index 对比实验。

2.针对测试数据中 RNA 合成指数 Cox 单因素回归分析、基于 RNA 合成指数和治疗方式放化疗的 Cox 多因素回归分析、以及有无 RNA 合成指数 Cox 回归分析所得 C-index 对比实验。

3.基于 RNA 合成指数的 kaplanmeier 生存曲线分析。

Cox 单多因素回归分析使用 R 软件中的“survival”和“survcomp”包 。kaplanmeier 生存曲线分析使用 python3.8 环境下 lifelines 库中的 kaplanmeierfitter 函数实现。

4.2 实验结果与分析

本研究中各实验所用计算机配置为 Intel core i9-10900X CPU 处理器，系统环境为 Ubuntu 20.04，显卡为 NVIDIA RTX2080Ti GPU。

4.2.1 RNA 合成指数预测结果与分析

表 4.1 显示了从模型中得到的 RNA 合成指数估计值与真实 RNA 合成指数值之间的比较。回归预测结果 MSE 为 0.35，R2-score 为 0.21。由于数据的复杂性，一个基于影像组学特征估计的 RNA 合成指数，以及一个样本的实际值之间的比较如图 4.3 所示。在图 4.3 中，x 轴显示了训练数据中的 108 例患者，y 轴显示了 RNA 合成指数值，红色是真实值，绿色是预测值。图 4.3 显示，大多数 RNA 合成指数估计值表现出令人满意的估计性能。

表 4.1 回归预测实验结果

Method	MSE	R2-score
SVR	0.35	0.21

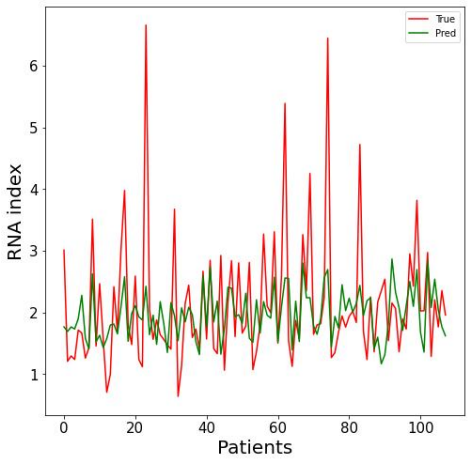


图 4.3 回归预测实验结果图

4.2.2 基于训练数据的 Cox 多因素回归分析

为了比较 108 例训练数据中 RNA 合成指数与其他临床因素对患者生存危险的影响程度，本研究进行了 Cox 多因素分析实验。实验结果如表 4.2 所示。针对患者年龄、肿瘤组织学、治疗方式和肿瘤复发进行调整的多变量 Cox 回归模型分析结果显示，RNA 合成指数（RNA-index）对生存时间的影响具有显著的统计学意义（HR=2.60208，P<0.001），此结果表明本研究所建立的 RNA 合成指数对患者预后具有极大的影响。而其余变量：年龄、组织学类型、复发以及放化疗 p-value 值均大于 0.05，无统计学意义。

表 4.2 训练数据 Cox 多因素分析实验结果

Factor	Coef	Exp(coef)	Se(coef)	Z	P
Age	0.03544	1.03607	0.02167	1.636	0.102
Histology	0.1081	1.1142	0.2043	0.529	0.597
Chemoradiotherapy	0.1557	1.1684	0.2691	0.578	0.563
Recurrence	0.3459	1.4133	0.3803	0.909	0.363
RNA-index	0.95631	2.60208	0.15661	6.106	1.02e-09***

此外，将 RNA 合成指数作为患者生存分析过程中的一个因素时，基于患者年龄、肿瘤组织学、治疗方式和肿瘤复发的 Cox 回归模型分析结果显示，模型 p-value 值在小于 0.05 的情况下 C-index 从 0.602 提升到 0.765。如表 4.3 所示，实验结果表明将 RNA 合成指数作为患者生存分析的预后因素时，C-index 提高了 0.163，这在医学分析上意义重大。

表 4.3 训练数据生存分析实验结果

Method(Cox)	C-index	S.D
Clinic	0.602	0.047
Clinic + RNA-index	0.765	0.053

Clinic: Age, Histology, chemoradiotherapy , Recurrence

Model p-value <0.05

4.2.3 基于测试数据的 Cox 单多因素回归分析

在测试数据中，对患者进行单变量 Cox 回归分析结果显示，RNA 合成指数（RNA-index）与患者总生存期（OS）显著相关（HR=7.155，P<0.05）。此外，针对患者放化疗治疗方式和 RNA 合成指数进行多变量 Cox 回归模型分析结果如表 4.4 和表 4.5 所示，RNA 合成指数（RNA-index）与患者 OS 显著相关（HR=8.8038，P<0.05）。表 4.6 显示了在治疗方式的基础上加上 RNA 合成指数预后指标对患者存活率进行预测的结果，Cox 回归模型预测性能实现提升，C-index 值从 0.483 提升到 0.700，且模型 p-value 值小于 0.05。这两个实验中治疗方式结果低于 RNA 合成指数的原因，大概率是因为本实验所用数据集中采取放化疗治疗方式的患者占少数。同时实验证明利用回归模型 SVR 预测得到 52 例患者的 RNA 合成指数，可提供与基因相关的信息，帮助医生对 52 例患者做出更好的预后。

表 4.4 测试数据 Cox 单因素分析实验结果（RNA-index）

Factor	Coef	Exp(coef)	Se(coef)	Z	P
RNA-index	1.968	7.155	0.917	2.146	0.0319

表 4.5 测试数据 Cox 多因素分析实验结果

Factor	Coef	Exp(coef)	Se(coef)	Z	P
Chemoradiotherapy	-0.3193	0.7267	0.5211	-0.613	0.5401
RNA-index	2.1752	8.8038	0.9660	2.252	0.0243

表 4.6 测试数据生存分析实验结果

Method	C-index	S.D
Chemoradiotherapy	0.483	0.142
Chemoradiotherapy +RNA-index	0.700	0.057

Model p-value <0.05

4.2.4 kaplanmeier 曲线分析

kaplanmeier 曲线分析方法其原理是按照从小到大的规则对生存时间排序。同时在出现死亡的每个时间点，计算最初的存活人数、死亡的人数、死亡的概率、存活的概率和存活率。这一想法与生命表方法大致相同，然而生命表方法中的时间段划分是人工的、等距的，而 kaplanmeier 曲线分析方法中的时段划分点是死亡发生的实际时间。这里介绍几个关键的指标：

1.失效事件：研究事件的终点。一般指患者死亡，也可以自定义为肿瘤复发、血压达标等其他感兴趣的二分类结局事件。

2.生存时间：从检测开始到事件发生所经过的时间，对于失访者，是失访前最后一次随访的时间。

3.删失（截尾）：研究对象在观察时间内没有发生事件。一种情况是研究对象在中途失访或退出；另一种情况是超过了最长的随访时间事件仍未发生。删失数据是一种不完整数据，是生存分析独有的重要组成部分。

为了测试 RNA 合成指数如何预测死亡时间，RNA 合成指数的预测被分为 2 组（RNA 合成指数<中值，RNA 合成指数>中值）。图 4.4 左侧给出了训练数据集中 RNA 合成指数值的 kaplanmeier 存活率，两曲线区别明显，证明使用 PVP 方法合成的 RNA 合成指数大小与死亡率分级关联。此外，图 4.4 的右侧给出了测试数据集中 RNA 合成指数的 kaplanmeier 生存估计值，从图中可知 RNA 合成指数估计位点与死亡率也存在分级关联。

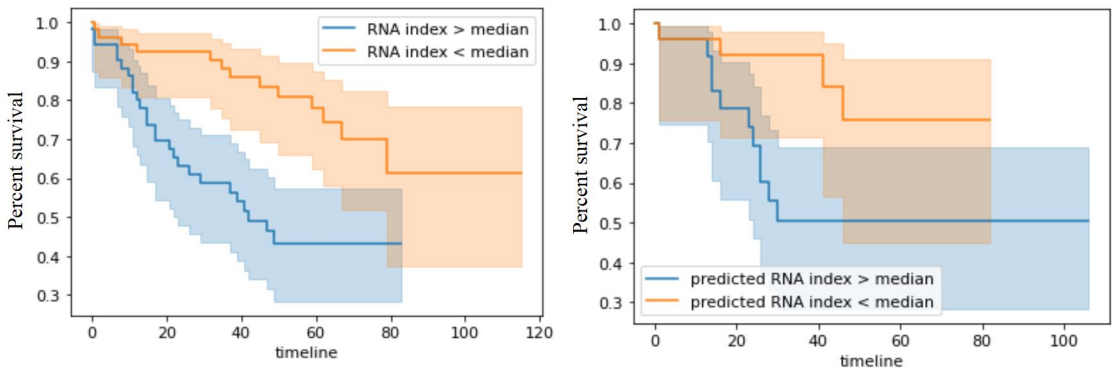


图 4.4　kaplanmeier 曲线分析

4.2.5 讨论

在本文中使用 PET/CT 放射组学特征和遗传特征来研究 NSCLC 患者的生存率，从而帮助医生和患者为可能的风险做好准备。由于传统的生存分析研究大多基于 PET/CT 图像信息、影像组学特征和临床特征，很少用到遗传信息，因此生存率预测的性能有限。同时，连续使用高维 RNA 信息来评估 NSCLC 患者的预后并不能有效估计多种 RNA 信息对患者预后的集体影响。本实验建立了一个 RNA 合成指数，它整合了多种 RNA 信息的预后能力，以便在最终治疗后立即使用。从表 4.2，表 4.3 中得知训练数据多因素 Cox 分析实验结果中 RNA 合成指数与 OS 显著相关(HR=2.60208，P<0.001)，且生存分析结果也证明 RNA 合成指数这一因素可以将 C-index 从 0.602 提升到 0.765（p<0.05)。此外，由于获取 RNA 序列的成本较高，本研究建立了预测模型，并利用 RNA 合成指数估计值来分析 NSCLC 患者的生存情况。实验结果如表 4.4、表 4.5、表 4.6 所示，结果表明测试数据单因素 Cox 分析与多因素 Cox 分析中 RNA 合成指数估计值同样与 OS 显著相关（HR=7.155，P<0.05，HR=8.8038，P<0.05），且可以使生存分析结果中的 C-index 从 0.483 提升到 0.700（p<0.05）。上述实验结果以及 kaplanmeier 曲线分析证明 RNA 合成指数对 NSCLC 患者的预后有很强的预测能力，比单独的临床信息更具预后价值。它定量地为初始治疗和预后评估提供了实用工具，有助于针对 NSCLC 患者制定个体化治疗和监测策略。

4.3 本章总结

本章节具体阐述了本文整体实验流程，从 RNA 合成指数建立到建立预测模型。具体来说，本实验在影像组学特征和 RNA 合成指数数据的基础上，运用机器学习技术训练预测模型，通过各种生存分析对比实验确定 RNA 合成指数的意义。最后还通过绘制 kaplanmeier 曲线图来进一步证明了 RNA 合成指数的有效性。

5. 基于 RNA 合成指数的非小细胞肺癌患者复发预测研究

5.1 基于 RNA 合成指数的 NSCLC 患者复发预测的方法

本研究利用上一研究中预测所得的 RNA 合成指数以及影像和临床特征，建立多模态模型对患者进行复发预测，具体方法如下。

5.1.1 基于 RNA 合成指数的 NSCLC 患者复发预测研究流程

肺癌在早期患者中复发率极高。预测肺癌患者术后复发通常是使用基因组学或放射学图像的单一模态信息。本实验研究了多模态融合在这项任务中的潜力。具体来说，首先对数据进行预处理，接着使用深度学习网络进行 NSCLC 患者复发预测。但在传统单一模态方法处理的过程中存在较大的特征缺失，诊断的性能往往较差。因此，本研究在影像数据的基础上，构建了一个多模态深度学习模型，在网络的全连接层处引入临床特征、影像组学特征和 RNA 合成指数特征，通过充分利用这三种模态的特征来对 NSCLC 患者复发进行预测。该框架包括三个步骤：影像数据预处理、深度学习特征提取、特征融合与分类。研究的具体流程如图 5.1 所示。

5.1.2 图像预处理流程

深度学习网络模型输入的图像包括两个模态，分别是 PET 和 CT 影像，如图 5.1（A）所示，列举了一例患者的 PET/CT 原始图像。不同样本分别为不同尺寸大小的全身图像。本研究所研究的内容是 NSCLC 复发预测，全身图像提供了大部分肺部以外的信息。因此，实验在数据预处理部分将全身图像肺部对应的切片截取出来，截取过程中为了提高肺部定位的准确性，实验从锁骨切片位置截取到肋骨最底端切片位置，切片数目大致在 90-110 张左右，最后将各样本图像线性插值到大小为 128*128*64 的 3D 块，输入网络前将图像进行归一化。预处理流程如图 5.2 所示。

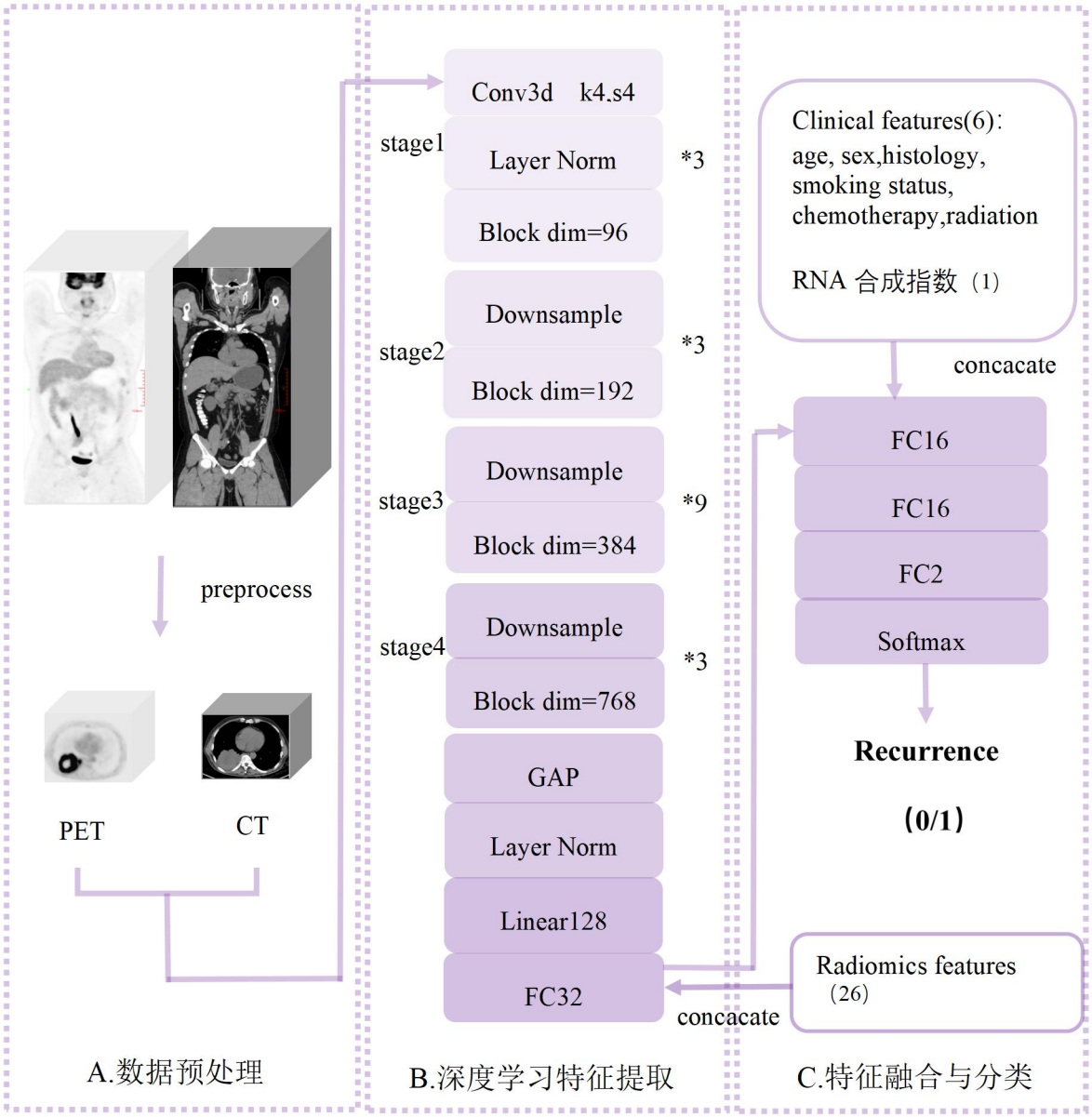


图 5.1　NSCLC 患者复发预测总体架构图，深度学习特征提取步骤中 Block 都是 ConvNeXt 模块。

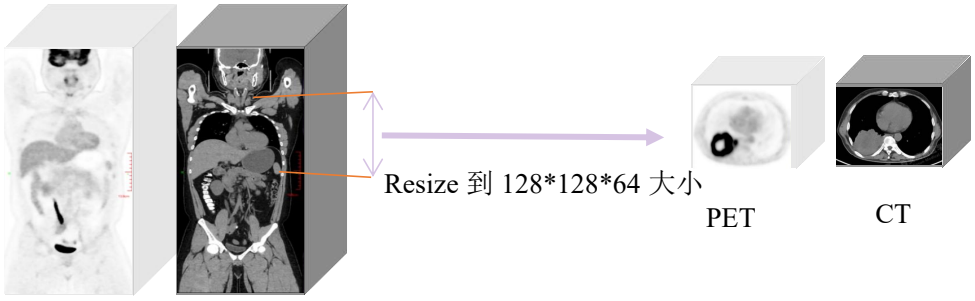


图 5.2　PET/CT 影像预处理流程图

5.1.3 影像组学特征提取方法

目前影像组学已有成熟的软件可以使用，本研究使用的是 python 环境下开源的 Pyradiomics 软件包，提取过程通过修改参数文件设置调库实现，其中图像滤波器设置为“Original”和“Wavelet”，每个滤波器的特征类别包括一阶图像强度统计量特征、形状特征、灰度共生矩阵、灰度偏移矩阵、灰度区域大小矩阵、相邻灰度差矩阵和灰度相关矩阵。最后，本研究利用该方法从每例患者的 PET/CT 图像的感兴趣区域（肿瘤对应区域）中提取出 825 维 PET 影像组学特征和 837 维 CT 影像组学特征，形成了高维的特征空间。这些高维特征是冗余的，容易出现过拟合现象，降低计算速度与效率，破坏模型分类预测的能力。所以在利用影像组学特征进行研究时，通常需要降低特征的高度相关性。本文选择 F 检验算法来达到降低特征空间维数的目的。特征选择流程如图 5.3 所示。选用 F 检验方法分别挑出 p-value 值小于 0.05 的 7 维 PET 的影像组学特征和 19 维 CT 的影像组学特征。

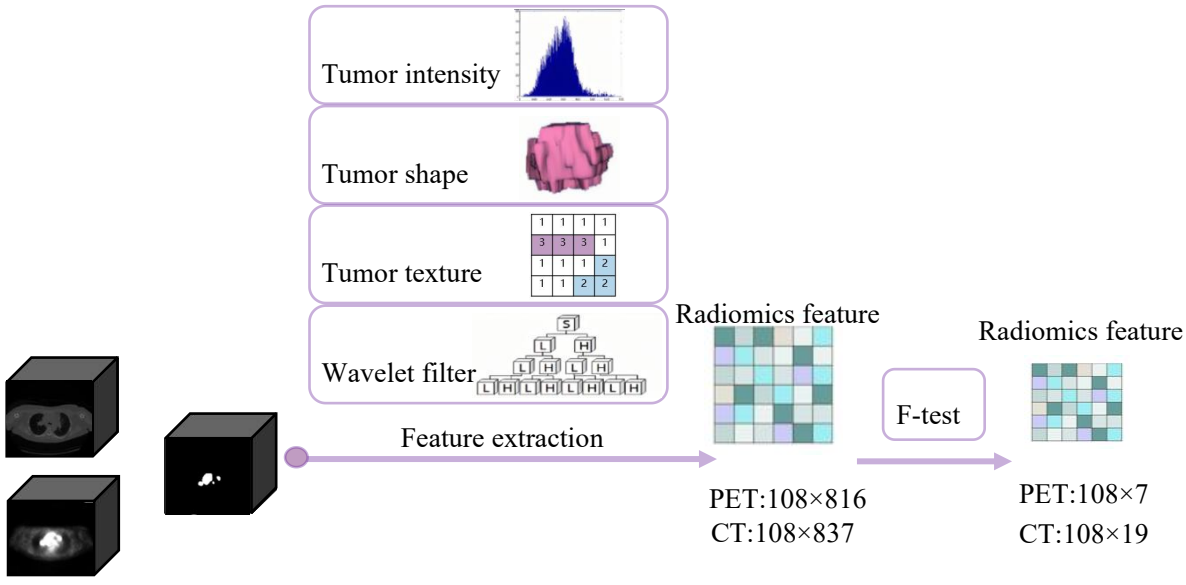


图 5.3　PET/CT 放射组学特征提取流程图

5.1.4 深度学习特征提取主干网络

深度学习主干网络使用的是 ConvNeXt[74]网络，该网络集成了 Transformer，ResNext 和 MobileNet 多个网络的优点，在多个分类任务和识别任务中均超越了 Swin Transformer 模型，达到最佳的性能表现。ConvNeXt 网络通过对现有网络的优化实现，部分细节参照 Swin Transformer 网络结构制定。ConvNeXt 使用 ResNet50 网络作为适配的骨干，网络结构较为简单。但其结果的准确率不仅比 Swin Transformer 高，更具有良好的推理速度。ConvNeXt 网络依照 Transformer 网络的一些先进思想对现有的经典 ResNet50/200 网络做一些调整改进，将 Transformer 网络的最新的思想和技术引入到 CNN 网络现有的模块中，提高 CNN 网络的性能表现。其进行的优化设计主要有以下几点：

1.ResNet 和 Swin Transformer 网络均有四个阶段，然而 Swin Transformer 各个阶段堆叠 Block 块的比例为 1:1:3:1，Swin-L 堆叠的比例为 1:1:9:3，由此可以发现 Transformer 网络的第三层的堆叠数量较多。因此 ConvNeXt 网络依照这个比例将 ResNet 各阶段的堆叠次数从(3，4，6，3)调整为(3，3，9，3)，其比例也保持在 1:1:3:1。

2.ResNeXt 网络相较于经典的 ResNet 网络而言区别在于 ResNeXt 在卷积块的中间部分采用了 group-wise convolution，使得卷积块形成了一个平行结构，ResNet 网络的卷积块是类似于瓶颈“两头粗，中间细”的结构。而 ConvNeXt 网络采用了 depth-wise convolution 构成卷积块，从而大幅度的减少网络的参数规模。

3.Swin Transformer 网络中 stem 层的输出特征通道数为 96，而 ResNet 网络 stem 层的输出只有 64 维。为了和 Swin Transformer 网络保持一致，ConvNeXt 网络加大了输出维度的数量，使其与 Swin Transformer 网络相同，大幅提升了网络的准确率，但同时也不可避免地增加了模型的参数规模。

4.在经典的 CNN 网络中一般习惯于使用 3×3 的卷积核，而 ConvNeXt 测试了各种不同尺寸的卷积核，发现当卷积核的尺寸为 7 时，网络的准确率和参数规模达到最优，其准确率的提升已经达到饱和。

5.传统的 CNN 网络中通常使用 Relu 作为网络的激活函数，而目前 Transformer 类型的网络主流上采用 Gelu 激活函数。因此 ConvNeXt 网络将 Relu 替换为 Gelu 激活函数，使网络的性能有微弱的提升。Swin Transformer 网络的每一个 Swin Transformer Block 中均只含有一个激活函数，因此受 Swin Transformer 的启发，ConvNeXt 网络减少了激活函数的使用，每个块只使用一个激活函数，部署在第二层之后。

6.与激活函数类似，ConvNeXt 网络也减少了正则化函数的使用，每个块只使用一个正则化函数，部署在第一层之后，且将正则化函数由 BN 替换成 LN。这两项操作轻微的提高了模型的准确率。参考 Swin Transformer 网络中的 Patch Merging 模块，ConvNeXt 网络单独设计了一个下采样层对特征进行单独的下采样操作。

本研究深度学习特征提取主干网络采用的是 ConvNeXt-T，主要模块是由 ConvNeXt block 和 Downsample block 组成，详细结构图如 5.4 所示。四个阶段通道数输入值分别为 96、192、384、768，每个阶段 block 块重复堆叠的次数 B 为 3、3、9、3。模型输入数据为 PET/CT 肺部区域的 3D 块（128*128*64 大小），在输入网络之前进行 Concatenate 操作。通过主干网络的四个阶段后，紧接着采用全局平均池化功能、批处理归一功能化和全连接功能得到一维的图像特征向量，用于最后一阶段的特征融合与分类任务。

5.1.5 特征融合与分类方法

实验最后一阶段将第二阶段得到的深度学习特征、原有的临床特征和 RNA 合成指数相结合，进行 NSCLC 患者复发预测，如图 5.1（C）所示。为了充分整合多变量，本研究使用 Concatenate 操作将四种类型的向量进行了合并。但是为了保证临床特征向量(包括 RNA 合成指数，长度为 7)不被图像特征和影像组学特征表示所淹没，影像组学特征首先和深度学习特征融合，当图像特征和影像组学特征经过全连接层降至 16 维时再与临床和 RNA 合成指数连接，形成长度为 23 通道的输出向量。最后通过全连接层将特征降至 2 维，并使用 Softmax 激活函数输出 NSCLC 患者复发风险概率，设置 0.5 的概率阈值确定最终的二分类结果。

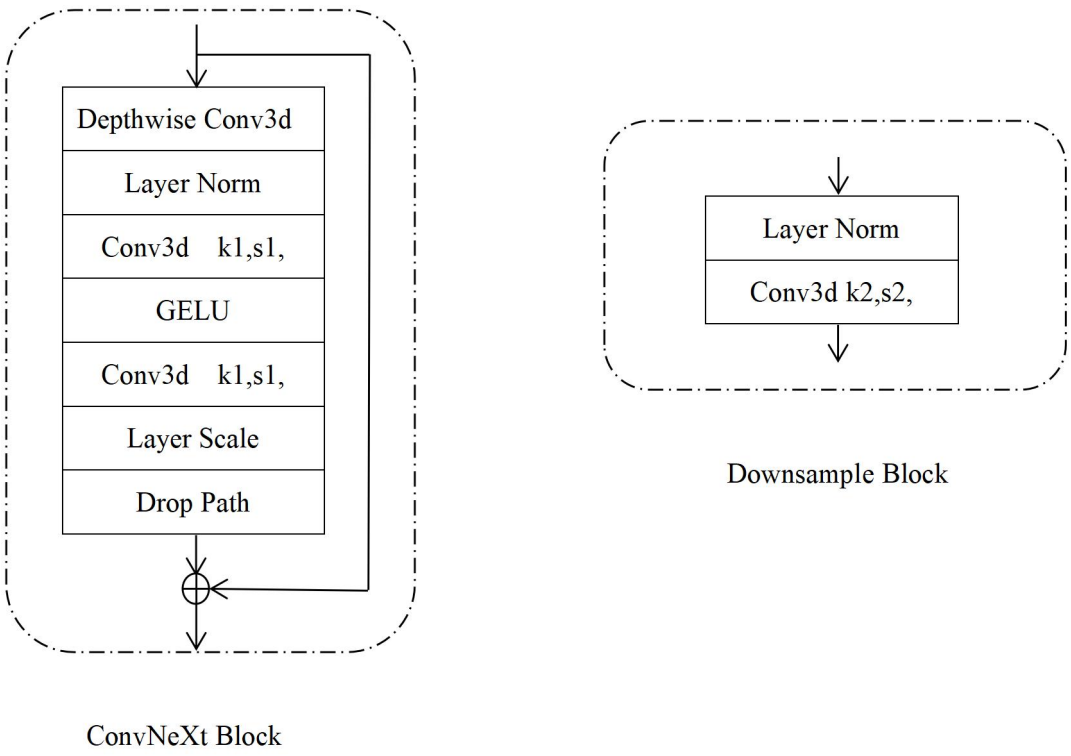


图 5.4 ConvNeXt block 和 Downsample block 详细结构图

5.1.6 实验训练过程

本研究数据共包括 160 例 NSCLC 患者，其中手术治疗之后会发生 NSCLC 复发的患者有 45 例，手术治疗后未发生 NSCLC 复发的患者有 115 例。实验目的主要是为了突出 RNA 合成指数在 NSCLC 复发实验当中的作用。因此，本研究将 RNA 合成指数作为变量进行研究，对照组实验是基于深度学习影像特征和临床特征进行实验。

本文采用二分类交叉熵作为本研究的损失函数，以此来进一步减轻数据不平衡带来的不利影响，二分类交叉熵损失函数公式可以写为：

$$L=\frac{1}{N}\sum_{i}L_i=\frac{1}{N}\sum_{i}-\left[y_i\cdot log\left(p_i\right)+\left(1-y_i\right)\cdot log\left(1-p_i\right)\right] \quad (5.1)$$

y_i表示 i 样本的标签，正类用 1 表示，负类用 0 表示，y_i表示 i 样本被预测为正类的概率大小。

交叉熵损失函数通常应用于分类任务中，特别是在使用神经网络 CNN 做分类时，也经常使用交叉熵作为损失函数。此外，交叉熵考虑到计算每个类别的概率，所以交叉熵通常都和 Sigmoid 或 Softmax 等激活函数一起出现，本研究选用的是 Softmax 函数。通过神经网络最后一层输出的情况来说明整个模型预测、获得损失和学习的流程如下：

1.神经网络最后一层得到每个类别的得分 Scores（也叫 Logits）；

2.该得分经过 Softmax 函数获得概率输出；

3.模型预测的概率输出且经过阈值操作后与真实类别的热编码形式计算交叉熵损失函数。

5.2 实验结果与分析

该模型在 Ubuntu 20.04 系统上，使用 NVIDIA RTX2080Ti GPU 进行训练，实验框架为 Python 3.8 版本的 TensorFlow2，batch size 大小设置为 6，采用随机梯度下降训练，学习率为 0.0001，且利用 Adam 优化器来训练模型。评价指标包括精准率（Precision）、准确度（Accuracy）、召回率（Recall）、AUC。

5.2.1 非小细胞肺癌患者复发实验结果

为了证明本研究提出 RNA 合成指数的有效性，进行了以下实验：基于深度学习提取的影像特征和临床特征与加入 RNA 合成指数特征的整体性能比较。

如表 5.1 所示，由于仅使用图像特征、影像组学特征和临床特征来预测 NSCLC 的复发，其预测性能非常有限，AUC 仅为 0.62，而引入 RNA 合成指数的实验结果 AUC 提高到 0.68。在这项研究中，为了提高复发预测的性能和满足所有 NSCLC 患者的需求，本研究不再使用单一输入模型直接预测复发，而是综合多模态，尤其是利用基因标签。模型性能如表 5.1 所示。在表 5.1 中，使用图像特征、组学特征和临床特征来预测 NSCLC 患者复发时 AUC 为 0.62，Precision 为 0.66，Recall 为 0.62，Accuracy 为 0.70。而综合图像特征、组学特征、临床特征和 RNA 合成指数方法预测得出的 AUC 为 0.70，Precision 为 0.76，Recall 为

0.67，Accuracy 为 0.75，AUC 比无基因数据时的预测性能有明显提高。图 5.5 是对应 ROC 曲线结果。

表 5.1 引入 RNA 合成指数特征复发预测对比实验结果

Method	Precision	Recall	Accuracy	AUC
PET/CT + clinic + radiomics	0.66	0.62	0.70	0.62
PET/CT + clinic + radiomics +RNA 合成指数（ours）	**0.76**	**0.67**	**0.75**	**0.70**

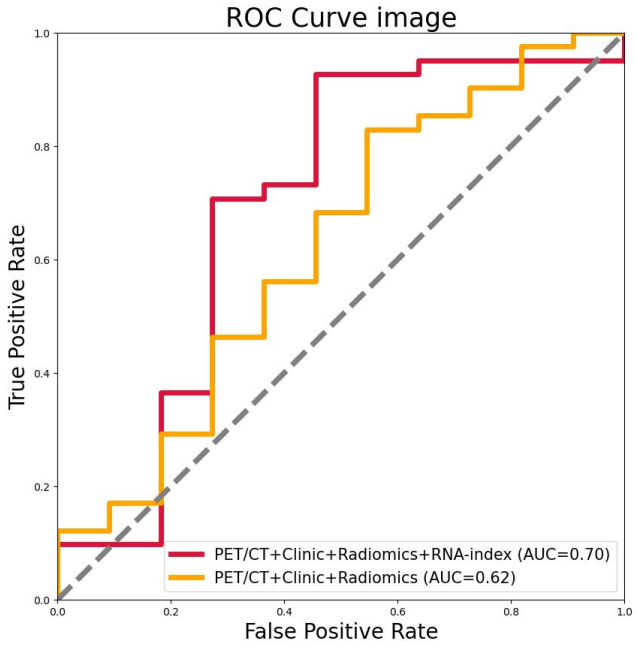


图 5.5　ROC 曲线结果

RNA 合成指数特征对于 NSCLC 患者复发判别中的作用在多模态融合模型中已经被证明，结果的大幅提升也证明这一工作引入 RNA 合成指数起到了相当重要的作用，相较于先前 NSCLC 复发研究工作是一大突破。

5.2.2 讨论

通过对 160 例患者数据集的实验验证，证明了使用 ConvNeXt 模型进行多模态融合预测复发的潜力，该结构能够综合集成多模态和多维数据，具有较强的泛

化能力。这显示了通过多模态融合可以改善模型预测效果，具有较强的学习能力、鲁棒性和高精度。该方法的预测精度从传统方法的 70%（AUC=0.62）提高到引入 RNA 合成指数方法的 75%（AUC=0.70）。在测试集的实验结果中使用四种模态方案的预测准确率要比使用三模态效果好，可以更好地实现对 NSCLC 患者复发的预测，能够满足实际应用的需要。然而，该方法的准确性依旧有待进一步提高。例如从单模态数据集进行的迁移学习可以更好地训练多模态融合网络，这是未来研究的一个有趣的方向。此外，为了进行更严格的分析和比较，还需要获得更多配对的数据集。

5.3 本章总结

在 NSCLC 患者复发预测这一章中，详细描述了基于患者 PET/CT 图像引入影像组学特征、临床特征和 RNA 合成指数进行复发预测的实验流程、具体实施方法和实验最终结果。本章使用的深度学习框架为 ConvNeXt 网络模型，该框架可以提取出与 NSCLC 患者复发相关的深度学习特征，取得了不错的性能。此外，本研究采用端到端设计，相较于低模态的方法具有更优的分类能力，能为医师提供较为准确的诊断依据，进一步改善患者的预后状态。

6. 总结与展望

6.1 总结

非小细胞肺癌是一种发病率和死亡率极高的恶性肿瘤，术后复发率也极高，直接影响到患者的生活和健康。近年来，基于计算机断层扫描 PET/CT 图像的研究被广泛应用，但图像特征不足以精准判断患者发病情况，精度较低。相比之下，利用基因表达数据预测 NSCLC 的复发具有较高的准确性。然而，基因数据的获取成本昂贵且具有侵入性，大多数患者可能因为较高成本拒绝基因检测，因此不具备完整的基因数据。本文提出了一种低成本、高精度的支持向量感知模型预测方法。首先，通过该模型构建手工特征和基因数据的映射函数。然后，利用回归模型获得基因估计数据，学习与递归相关的信息表示，实现了生存分析与复发预测任务。

在 NSCLC 的 RNA 合成指数建立与预测工作中，首先对数据进行预处理，针对图像数据提取 PET/CT 图像的传统影像组学特征。具体来说，本研究使用 Pyradiomics 库提取原发肿瘤的传统影像组学特征，包括 825 维 PET 影像组学特征和 837 维 CT 组学特征。由于特征具有高维冗余性，不利于模型训练，因此实验通过 F 检验方法进行特征降维，剔除无用特征。最后，得到 7 维 PET 影像组学特征和 19 维 CT 影像组学特征。针对基因数据，首先使用多层特征筛选的方法从 2 万多个基因中，挑出 9 个基因特征，并基于 PVP 合成方法，将 9 维特征合并成一维，大大降低了数据维度。其次，本研究通过支持向量回归模型，建立影像组学特征和基因数据之间的映射函数，由此可以实现患者基因相关信息的预测，从而帮助医生更好的制定患者治疗方案。

癌症复发的预测是针对某些疾病治疗后复发可能性的预测。临床中普遍使用类似活检的方法，但该技术成本较高。有相关研究表明，数据挖掘或机器学习的方法，能够使得该任务低成本、高精度的完成。对于 NSCLC 复发预测任务，过往研究大多采用的是图像和临床特征，即使使用了基因信息，也是高维嵌入。因此，为了进一步证明上述研究中所提出的 RNA 合成指数的关键作用，本文提出基于 RNA 合成指数特征对 NSCLC 患者复发概率进行预测。具体来说，实验基

于 ConvNeXt 网络提取 NSCLC 患者 PET/CT 影像中的高阶深度学习特征，将其与影像组学特征、临床特征和 RNA 合成指数进行多模态特征重组，并利用多个全连接层进行特征融合与降维，最后使用 Softmax 函数得到 NSCLC 患者复发的概率。该模型的预测结果显示，多模态特征的融合在 NSCLC 复发预测方面具有良好的表现，尤其是在引入 RNA 合成指数后，模型的性能得到显著提高。这一实验证明了本研究建立的 RNA 合成指数特征对 NSCLC 复发预测是可行的且有效的，为以后 NSCLC 患者复发抑或是其他疾病的研究提供了新的方向。

6.2 展望

虽然本文在基因信息预测预后和 NSCLC 复发预测方面取得了一些成果，但也存在缺陷，有待进一步探讨和研究，主要有以下几个方面：

1.在实验数据方面。目前，已有研究证明组织学图像、MRI 图像、诊断信息和医嘱等多种模态的数据已被广泛应用于医学领域的研究，而针对 NSCLC 患者生存预后的研究，尚未涉及上述模态信息，因此在之后的研究中，可对多模态信息加以利用，扩大 NSCLC 患者生存分析的研究。

2.在基因信息利用方面。本文建立的 RNA 合成指数是一种基于数学统计的方法，将高维信息合成 1 维进行研究。在之后的研究中，可以尝试更多样的降维方式，挑选出更具代表的基因信息，或是致力于多维特征的回归预测，实现点对点的基因预测，为医生提供更直接、更明确的基因信息。

3.在 NSCLC 复发预测方面。本文研究 NSCLC 复发的主要目的是证明 RNA 合成指数的有效性。在未来的研究中，可以在现有研究的基础上进行更多的理论与方法创新，例如改变特征融合方式或网络结构等，通过不断调整方法和策略来提高复发预测的准确性。

4.在 RNA 合成指数的应用方面。本文通过建立 RNA 合成指数进行了 NSCLC 生存分析研究和复发预测研究。在之后的研究中，可以将本研究建立的 RNA 合成指数应用于其他医学研究，例如肺浸润发生概率大小、鳞癌腺癌种类预测、TNM 分期预测、肿瘤危险等级预测等工作，从多方面为医生提供辅助治疗、简化工作内容，提高工作效率。

参考文献

[1] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2020. 71: 209–49.

[2] R. Qureshi, B. Zou, T. Alam, J. Wu, V. H. F. Lee and H. Yan. Computational Methods for the Analysis and Prediction of EGFR-Mutated Lung Cancer Drug Resistance: Recent Advances in Drug Design, Challenges and Future Prospects. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2023. vol. 20, no. 1, pp. 238-255, 1 Jan.-Feb. doi: 10.1109/TCBB.2022.3141697.

[3] R.L.Siegel, K.D.Miller, and A.Jemal. Cancer statistics, 2020. CA. Cancer J. Clin, 2020. vol.70, no.1, pp.7–30. doi: 10.3322/caac.21590.

[4] Noone AM, Howlader N, Krapcho M, Miller D, Brest A, Yu M, et al. SEER cancer statistics review. 1975–2015. National Cancer Institute. 2018. Accessed 27 Oct 2020. https://seer.cancer.gov/csr/1975_2015/.

[5] Jalil R, Ahmed M, Green JSA, Sevdalis N. Factors that can make an impact on decision-making and decision implementation in cancer multidisciplinary teams: An interview study of the provider perspective[J]. International Journal of Surgery, 2013, 11(5):389-394.

[6] Hu D, Zhang H, Li S, et al. An ensemble learning with active sampling to predict the prognosis of postoperative non-small cell lung cancer patients[J]. BMC Medical Informatics and Decision Making, 2022, 22(1):1-12.

[7] Wan Y W, Guo N L. Constructing gene-expression based survival prediction model for Non-Small Cell Lung Cancer (NSCLC) in all stages and early stages[C]. IEEE International Conference on Bioinformatics & Biomedicine Workshop. IEEE, 2009:338-338.

[8] Chen C. Analysis of the value of multi-slice spiral CT in the screening of high-risk lung cancer and its imaging findings[J]. Chinese J CT MRI,2016,2:42–44.

[9] H.C.Steinert. PET and PET-CT of lung cancer[J], Methods in Molecular Biology. vol.727, pp.33–51,2011.

[10] Griffeth L K. Use of PET/CT scanning in cancer patients: technical and practical considerations[J]. Proceedings, 2005, 18(4):321.

[11] Shim S. Non-small cell lung cancer: prospective comparison of integrated FDG PET/CT and CT alone for preoperative staging[J]. Radiology, 2005, 236:1011–1019.

[12] Bruzzi JF, Munden RFPET/CT imaging of lung cancer[J]. J Thorac Imaging. 2006, 21:123–136.

[13] Kim BT, Lee KS, Shim SS, Choi JY, Kwon OJ, Kim H, Kim S. Stage T1 non-small cell lung cancer: preoperative mediastinal nodal staging with integrated FDG PET/CT—a prospective study[J]. Radiology, 2006, 241:501–509.

[14] Efthymiadou, R.D. PET-CT in Lung Cancer. In: Andreou, J.A., Kosmidis, P.A., Gouliamos, A.D. (eds) Artificial Intelligence in PET/CT Oncologic Imaging[J]. Springer, Cham.23 October 2022. p39–44. https://doi.org/10.1007/978-3-031-10090-1_5.

[15] Amini M, Nazari M, Shiri I, et al. Multi-Level PET and CT Fusion Radiomics-based Survival Analysis of NSCLC Patients[C]. 2020 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC). IEEE,1-4.

[16] Bakr S, Gevaert O, Echegaray S, Ayers K, Zhou M, Shafiq M, Zheng H, Benson JA, Zhang W, Leung ANC, Kadoch M, Hoang CD, Shrager J, Quon A, Rubin DL, Plevritis SK, Napel S. A radiogenomic dataset of non-small cell lung cancer. Sci Data. 2018 Oct 16; 5: 180202. doi: 10.1038/sdata.2018.202. PubMed PMID: 30325352; PubMed Central PMCID: PMC6190740.

[17] Ergin, S., Kherad, N. & Alagoz, M. RNA sequencing and its applications in cancer and rare diseases[J]. Mol Biol Rep 49, 2325–2333 (2022). https://doi.org/10.1007/s11033-021-06963-0.

[18] X. Ye, W. Zhang and T. Sakurai, Adaptive Unsupervised Feature Learning for Gene Signature Identification in Non-Small-Cell Lung Cancer[J], in IEEE Access,

2020.vol. 8, pp.154354-154362. doi: 10.1109/ACCESS.2020.3018480.

[19] R. Harun et al., Gene expression profiles predict survival of patients with advanced non-small cell lung cancers[C], 2011 Fourth International Conference on Modeling,Simulation and Applied Optimization, 2011, pp.1-4. doi: 10.1109/ICMSAO. 2011.5775581.

[20] Fu FQ, Zhang Y, Wen ZX, et al. Distinct prognostic factors in patients with stage I non- small cell lung cancer with radiologic partsolid or solid lesions[J]. J Thorac Oncol, 2019, 14(12):2133-2142.

[21] Karp I, Sylvestre MP, Abrahamowicz M, et al. Bridging the etiologic and prognostic outlooks in individualized assessment of absolute risk of an illness: application in lung cancer[J]. Eur J Epidemiol, 2016, 31(11):1091-1099.

[22] Muller DC, Johansson M, Brennan P. Lung cancer risk prediction model incorporating lung function: development and validation in the UK biobank prospective cohort study [J]. J Clin Oncol, 2017, 35(8):861-869.

[23] 段桦, 罗楚凡, 崔慧娟等. 227 例晚期非小细胞肺癌患者的预后因素分析[J]. 癌症进展, 2020, 18(4):366-370.

[24] 王娟, 黄淼, 齐丽萍等. 周围型非小细胞肺癌 CT 影像学因素预后分析[J]. 中国介入影像与治疗学, 2016, 13(7):411-415.

[25] Wang S, Chen A, Yang L, et al. Comprehensive analysis of lung cancer pathology images to discover tumor shape and boundary features that predict survival outcome[J]. Scientific reports, 2018, 8(1): 1-9.

[26] Mukherjee P, Zhou M, Lee E, et al. A shallow convolutional neural network predicts prognosis of lung cancer patients in multi-institutional computed tomography image datasets[J]. Nature machine intelligence, 2020, 2(5): 274-282.

[27] Wu Y, Ma J, Huang X, et al. DeepMMSA: A novel multimodal deep learning method for non-small cell lung cancer survival analysis[C]. 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC).IEEE,2021:1468-1472.

[28] Kadoya N, Tanaka S, Kajikawa T, et al. Homology-based radiomic features for

prediction of the prognosis of lung cancer based on CT-based radiomics[J]. Med Phys, 2020, 47(5): 2197-2205. 33 G.

[29] Yıldırım F, Yurdakul AS, Özkaya S, et al. Total lesion glycolysis by 18F- FDG PET/CT is independent prognostic factor in patients with advanced non-small cell lung cancer [J]. Clin Respir J, 2017, 11(5): 602-611.

[30] Li XF, Yin GT, Zhang YF, et al. Predictive power of a radiomic signature based on 18F-FDG PET/CT images for EGFR mutational status in NSCLC[J]. Front Oncol,2019, 9:1062.

[31] Moon SH, Sun JM, Ahn JS, et al. Predictive and prognos- tic value of 18F-fluorodeoxyglucose uptake combined with thymidylate synthase expression in patients with advanced non-small cell lung cancer[J]. Sci Rep, 2019, 9(1): 12215.

[32] Sharma A, Mohan A, Bhalla AS, et al.Role of various metabolic parameters derived from baseline 18F- FDG PET/CT as prognostic markers in non-small cell lung cancer patients undergoing platinumbased chemotherapy[J]. Clin Nucl Med, 2018, 43(1): e8-e17.

[33] Huang Y, Liu Z, He L, et al. Radiomics Signature: A potential biomarker for the prediction of disease-free survival in early-stage (I or II) non-small cell lung cancer[J]. Radiology, 2016, 281(3): 947-957.

[34] van Timmeren JE, van Elmpt W, Leijenaar RTH, et al. Longitudinal radiomics of cone-beam CT images from non-small cell lung cancer patients: Evaluation of the added prognostic value for overall survival and locoregional recurrence[J]. Radiother Oncol, 2019, 136: 78-85.

[35] Bousabarah K, Temming S, Hoevels M, et al. Ｒadiomic analysis of planning computed tomograms for predicting radiation induced lung injury and outcome in lung cancer patients treated with robotic stereotactic body radiation therapy[J]. Strahlenther Onkol, 2019, 195(9): 830－842 .

[36] Gevaert O, Xu J, Hoang C D, et al. Non–small cell lung cancer: identifying prognostic imaging biomarkers by leveraging public gene expression microarray data—methods and preliminary results[J]. Radiology, 2012, 264(2): 387-396.

[37] Emaminejad N, Qian W, Guan Y, et al. Fusion of Quantitative Image and Genomic Biomarkers to Improve Prognosis Assessment of Early Stage Lung Cancer Patients[J]. IEEE Transactions on Biomedical Engineering, 2016, 63(5): 1034-1043.

[38] Subramanian V, Do M N, Syeda-Mahmood T. Multimodal Fusion of Imaging and Genomics for Lung Cancer Recurrence Prediction[C]. 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI). IEEE, 2020.

[39] Wang H, Subramanian V, Syeda-Mahmood T.Modeling uncertainty in multi-modal fusion for lung cancer survival analysis[C]. 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). IEEE, 2021: 1169-1172

[40] Singh A, Wang Z, Katz S, et al. Development of a radiogenomic biomarker for tumor characterization and prognosis in non-small cell lung cancer patients[C]. Medical Imaging 2021: Computer-Aided Diagnosis. International Society for Optics and Photonics, 2021, 11597: 115972W.

[41] Aonpong P, Iwamoto Y, Han X H, et al. Genotype-Guided Radiomics Signatures for Recurrence Prediction of Non-Small Cell Lung Cancer[J]. IEEE Access, 2021, 9: 90244-90254.

[42] V. Subramanian, M. N. Do and T. Syeda-Mahmood, Multimodal Fusion of Imaging and Genomics for Lung Cancer Recurrence Prediction[J], 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), Iowa City, IA, USA, 2020, pp. 804-808, doi: 10.1109/ISBI45749.2020.9098545.

[43] S. Ali Hosseini, G. Hajianfar, I. Shiri and H. Zaidi, Lung Cancer Recurrence Prediction Using Radiomics Features of PET Tumor Sub-Volumes and Multi-Machine Learning Algorithms[C], 2021 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC), Piscataway, NJ, USA, 2021, pp. 1-3, doi: 10.1109/NSS/MIC44867.2021.9875889.

[44] Y. Ai et al., Residual Multilayer Perceptrons for Genotype-Guided Recurrence Prediction of Non-Small Cell Lung Cancer[J], 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC),

Glasgow, Scotland, United Kingdom, 2022, pp. 447-450, doi: 10.1109/EMBC48229.2022.9871896.

[45] Abbasian Ardakani A, Bureau NJ, Ciaccio EJ, Acharya UR. Interpretation of radiomics features-A pictorial review. Comput Methods Programs Biomed[J]. 2022;215:106609.ISSN 0169-2607, doi:10.1016/j.cmpb.2021.106609

[46] Shahdoosti, H.R., Javaheri, N. A fast algorithm for feature extraction of hyperspectral images using the first order statistics[J]. Multimed Tools Appl 77, 23633–23650 (2018). https://doi.org/10.1007/s11042-018-5695-0

[47] A.Septiarini, H. Hamdani, E. Setyaningsih, S. Maharani, A. S. Munir and E. Winarno, Detecting Retinal Nerve Fiber Layer Using Gray Level Co-occurrence Matrix and Machine Learning Approach[C], 2022 International Conference on Information Technology Research and Innovation (ICITRI), 2022, pp. 173-178, doi: 10.1109/ICITRI56423.2022.9970211.

[48] Dash, S., Senapati, M.R. Gray level run length matrix based on various illumination normalization techniques for texture classification[J]. Evol. Intel. 14, 217–226 (2021). https://doi.org/10.1007/s12065-018-0164-2

[49] Chen, S., Harmon, S., Perk, T. et al. Using neighborhood gray tone difference matrix texture features on dual time point PET/CT images to differentiate malignant from benign FDG-avid solitary pulmonary nodules[J]. Cancer Imaging 19, 56 (2019). https://doi.org/10.1186/s40644-019-0243-3

[50] Santosh, N. Krishna and Barpanda, Soubhagya Sankar. 4. Wavelet applications in medical image processing[M]. Predictive Intelligence in Biomedical and Health Informatics, edited by Rajshree Srivastava, Nhu Gia Nguyen, Ashish Khanna and Siddhartha Bhattacharyya, Berlin, Boston: De Gruyter, 2020, pp. 63-90. https://doi.org/10.1515/9783110676129-004

[51] Kohavi R, John GH. Wrappers for feature subset selection. Artif Intell 1995; 97: 273–324.

[52] Gholami, B., Norton, I., Tannenbaum, A. R., & Agar, N. Y. Recursive feature elimination for brain tumor classification using desorption electrospray ionization

mass spectrometry imaging[J]. IEEE Engineering in Medicine and Biology Society. Annual International Conference, 2012, 5258–5261.

[53] Snee R. Development in linear regression methodology[J].2022.

[54] Chen B. Polynomial Regression[J]. Springer Texts in Statistics, 1986:235-268.

[55] AJ Smola†, BS Lkopf‡. A tutorial on support vector regression[J]. Statistics and Computing, 2004, 14(3): 199-222.

[56] Kretowski, Marek, Czajkowski, et al. The role of decision tree representation in regression problems-An evolutionary perspective[J]. Applied Soft Computing, 2016.

[57] Liaw A, Wiener M. Classification and Regression by randomForest[J]. R News, 2002, 23(23).

[58] Ranstam J, Cook J A. LASSO regression[J]. British Journal of Surgery, 2018,105.

[59] Mcdonald G C. Ridge regression[J]. Wiley Interdisciplinary Reviews Computational Statistics, 2010, 1(1):93-100.

[60] Chen T, Tong H, Benesty M. xgboost: Extreme Gradient Boosting[J].2016.

[61] 陈先昌. 基于卷积神经网络的深度学习算法与应用研究[D]. 浙江工商大学, 2014.

[62] Alzubaidi, L., Zhang, J., Humaidi, A.J. et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions[J]. J Big Data 8, 53 (2021). https://doi.org/10.1186/s40537-021-00444-8.

[63] Zhang Z, Cui P, Zhu W. Deep learning on graphs: a survey[J]. IEEE Trans Knowl Data Eng. 2020.

[64] Bengio, Y, Courville, A .Deep learning (Vol. 1) .Cambridge: MIT press, 2016: 326-366.

[65] Gu J, Wang Z, Kuen J, et al. Recent Advances in Convolutional Neural Networks[J].Pattern Recognition, 2015.1512.07108.

[66] Lin M, Chen Q, Yan S. Network In Network[J], 10.48550/arXiv.1312.4400[P]. 2013.

[67] 周志华. 机器学习[M].北京：清华大学出版社, 2000:30-35.

[68] S. Bakr, O. Gevaert, S. Echegaray, K. Ayers, M. Zhou, M. Shafiq, H. Zheng, W. Zhang, A. Leung, M. Kadoch, J. Shrager, A. Quon, Rubin, Daniel; S. Plevritis, S. Napel, Data for NSCLC Radiogenomics Collection[J]. The Cancer Imaging Archive. http://doi.org/10.7937/K9/TCIA.2017.7hs46erv, 2017.

[69] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, F. Prior. The Cancer Imaging Archive (TCIA), Maintaining and Operating a Public Information Repository[J], Journal of Digital Imaging, Dec.2013. vol. 26, no. 6, pp 1045-1057.

[70] Liu L, Zhang J, Ferguson MK, Appelbaum D, Zhang JX, Pu Y. Developing a clinical and PET/CT volumetric prognostic index for risk assessment and management of NSCLC patients after initial therapy[J]. Front Biosci (Landmark Ed). 2022 Jan 12;27(1): 16. doi: 10.31083/j.fbl2701016. PMID: 35090321.

[71] Sulaiman, M.A., & Labadin, J. (2016). Improved Feature Selection Based on Mutual Information for Regression Tasks[J]. Journal of IT in Asia, 6(1), 11-24. https://doi.org/10.33736/jita.330.2016.

[72] 张玉超. 融合用户信任关系的协同过滤推荐算法研究[D]. 兰州:西北师范大学, 2021.

[73] 程文静. 业务流程管理中的图像挖掘技术研究[J]. 现代商贸工业, 2020, 41(26): 152-153.

[74] Liu Z, Mao H, Wu CY, et al. A ConvNet for the 2020s[J]. arXiv e-prints, 2022.

致谢

凡是过往，皆为序章。岁月蹉跎，流年似水。我从一个怀抱理想的逐梦少年，一路踉踉跄跄走到现在，有些许不甘，些许留恋，些许憧憬。如今的我即将开始新的旅程，同时也对未来充满希望。回想在兰州财经大学的这三年时光，感恩从我身边出现的每一个人，正是因为有你们的陪伴和关心，才能让我对这段过往经历充满依恋与热爱，也祝福你们未来道路如阳光般夺目。回首过往，立足现在，展望未来。行文至此，唯有感恩，感谢这三年来遇到的赤诚良师，“得一良师，是吾辈之幸”，感谢益友不离不弃的陪伴，感谢我的父母给予我支持。

一朝沐杏雨，一朝念师恩。桃李不言，下自成蹊。首先我要感谢我的导师李兵老师，是他在学习中给了我极大地鼓励，让我能够在困境时努力寻找前进的方向与动力。我还要感谢我的指导老师何江萍教授，在他积极指导下，才能让我全身心投入到科学研究中；在他督促鼓励下，我才能快速前进；在他的倾囊相助下，我才能丰富阅历，提升能力；在他兢兢业业的态度下，实验室才能欣欣向荣，隆隆日上。这种氛围培养了我认真负责并且追求完美的性格，也培养了我面对困难想要退缩时的义无反顾、勇往直前。同时也感谢实验室的其他老师，你们是我前进路上坚强的后盾，是你们给予了我力量，这股无形的力量，会化作一只大手，支持着我一直前进。

山水一程，荣幸之至，与天地兮同寿，与日月兮齐光。感恩319宿舍的姐妹们，是你们让我的世界变得精彩，是你们陪伴我，包容理解我，让我看到不同女孩身上绽放的不同光彩。回望三年来的点点滴滴，我们在一起生活的时间过的真快。想念那些精力充沛的日子，一起熬夜学习，大声讨论学术的日子，那些匆忙早起，奔去实验室的美好时光。谢谢你们的陪伴，让我感受到了家一般的温暖，愿我们未来的道路能烁烁生辉。

羊有跪乳之恩，鸦有反哺之义。最后感谢我的亲人，你们是我成长的见证人，是你们教会我善良和真诚，谢谢你们一路以来对我默默无闻的陪伴，在我最艰难的时候，是你们给我肩膀依靠。前进道路上，无论是风高浪急，还是惊涛骇浪，你们永远是我们最坚实的依托，最强大的底气。有你们在，我一应俱全，也谢谢你们，把最好的都给予了我。我一定会坚守初心，勇往直前。