分类号 TP391.4　　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 基于影像组学特征的非小细胞肺癌 TNM 分期与淋巴转移预测研究

研 究 生 姓 名：吴 瑶

指导教师姓名、职称：韩金仓 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：2021 年 5 月 15 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：吴 瑶 签字日期：2021.5.25

导师签名：[illegible] 签字日期：2021.5.25

# 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：吴 瑶 签字日期：2021.5.25

导师签名：[illegible] 签字日期：2021.5.25

# Radiomics-Based Prediction for TNM stage and Lymph Node Metastasis in Non-small cell lung Cancer

**Candidate : Wu Yao**

**Supervisor: Han Jincang**

# 摘 要

肺癌是世界上发病率和死亡率最高的恶性肿瘤疾病之一。TNM 分期是国际常用的恶性肿瘤分类方法，对癌症治疗方案的制定有重要作用。影像组学从医学图像中高通量提取量化特征，定量地描述关于肿瘤的关键性病理信息。实现基于影像组学特征的 TNM 分期意味着以非侵袭的方式实现肿瘤分期。通常来说如果出现肿瘤转移，就意味着较差的生存期和预后，通过对淋巴转移的预测，可以尽早发现存在潜在转移风险的患者，提前介入治疗对于提升患者的生存期和预后有重要意义。

因此，本文基于非小细胞肺癌患者的 PET-CT 图像，利用机器学习的方法，分别对非小细胞肺癌患者的 TNM 分期和淋巴转移进行预测，分析影像组学特征在术前肺癌 TNM 分期和淋巴转移预测中的性能。

基于影像组学特征对非小细胞肺癌患者的 TNM 分期进行预测，实现肺癌 TNM 自动分期。研究内容包括特征的提取与选择、TNM 分期预测两部分。首先是影像组学特征的提取和选择。利用提取工具从患者的 PET 图像中提取影像组学特征，由于影像设备间的差异，用曼惠特尼 U 检验剔除在不同设备中不服从同一分布的特征，影像组学特征具有高维性，用方差过滤进行进一步特征选择，最后用 LASSO 选出与肺癌 TNM 分期相关的特征。其次，根据选择的特征用 XGBoost 分类器实现 TNM 分期预测。实验数据为 324 位 TNM 分期为 I 期、II 期和 III 期的非小细胞肺癌患者的 PET-CT 图像。10 折交叉验证结果显示总体精确度和 AUC 分别为 0.71 和 0.59，I 期、II 期和 III 期召回率分别为 0.95，0.38 和 0.63。结果说明利用影像组学特征可以用于非小细胞肺癌 TNM 分期，影像组学特征反映了肺癌不同分期肿瘤间的异质性。

基于影像组学特征的非小细胞肺癌淋巴转移预测，包括特征提取与选择和淋巴转移预测两部分。利用组合的特征选择策略包括曼惠特尼 U 检验、方差过滤和 LASSO 进行特征选择，然后用 XGBoost 分类器进行淋巴转移的预测。利用 SMOTE 数据不平衡策略解决实验中存在的数据不平衡问题。实验数据为 121 例非小细胞肺癌患者，其中没有淋巴转移（LNM-）102 例，出现淋巴转移（LNM+）19 例。淋巴转移预测包括两个内容，一是用不同类型的特征分别进行预测，验证不同类型特征在淋巴转移预测中的性能，二是利用 LASSO 选择的 2 个最优特征进行淋巴转移预测。单类型淋巴转移预测结果表明利用方差过滤选择后的不同类型特征在预测淋巴转移方面有良好的性能，不同类型的特征均一定程度地表达了瘤间异质性；利用 LASSO 选择后的特征进行预测时，在不加入临床特征情况下预测精确度与 AUC 分别为 0.72 和 0.7，加入临床特征后精确度和 AUC

分别为 0.78 和 0.75。通过与医生对患者淋巴转移的判断进行对比，基于影像组学的方法对于 LNM+患者的预测要优于医生的判断，对于提升存在转移风险患者的生存期和预后有重要意义。利用医学影像进行肿瘤转移预测，具有非侵袭性，实验结果说明利用 PET 影像组学特征，可以实现肺癌的淋巴转移预测。

本文通过影像组学特征和机器学习的方法，研究肿瘤临床治疗中存在的问题。实验结果表明 PET 影像组学特征与非小细胞肺癌的 TNM 分期和淋巴转移有关，可用于非小细胞肺癌的 TNM 分期和淋巴转移预测，影像组学特征与临床特征结合有更好的淋巴转移预测结果。

**关键字：**非小细胞肺癌　PET-CT 图像　影像组学　TNM 分期　淋巴转移　预测

# Abstract

Purpose: To develop and validate a radiomics-based model for preoperative prediction of TNM stage and lymph node metastasis(LNM) in Non-small cell lung cancer(NSCLC).

Methods: A total of 324 and 121 patients were enrolled respectively in our retrospective study. A radiomics evaluation of 833 high-throughput features were calculated from PET-CT images, including shape, texture, intensity and wavelet. Combined features selection strategy were used to select features related to TNM stage and LNM in NSCLC. The performance of prediction was evaluated by Confusion matrix, Accuracy and the area under the ROC curve(AUC). Synthetic Minority Oversampling Technique(SMOTE) were used to resolve data imbalance in LNM prediction.

Results: In total, 125 stage I, 58 stage II and 141 stage III were enrolled in prediction of TNM stage. After feature selection, 11 most contributing features were selected to develop XGBoost model. 10 fold cross validation showed that AUC value was 0.59, classification performance of stage I, II and III were 0.95, 0.38 and 0.63 respectively. 121 patients with 102 LNM+ and 19 LNM- were included in prediction of LNM. 2 radiomics features were reserved to develop XGBoost model finally. We both combined radiomics and clinical features (Smoker, Age and Laterality Desc) to develop XGBoost model. 5 fold cross validation showed that AUC value were 0.7 and 0.75 respectively. And then we test the performance of doctor’s prediction of LNM, AUC value was 0.66 after

the calculation.

Conclusion: This study indicated that PET-CT based radiomics combined with clinical information is valuable in prediction of TNM stage and LNM in NSCLC. Prediction of LNM with radiomics-based model is better than the judgement of doctor, especially in patients with LNM+, which have significant value in improvement of overall and prognosis.

**Keywords:** Non-small cell lung cancer; PET-CT images; radiomics; TNM stage; lymph node metastasis; prediction

# 目 录

1 绪 论 1
1.1 研究背景与意义 1
1.1.1 肺癌 1
1.1.2 PET-CT 介绍 2
1.1.3 肺癌 TNM 分期 4
1.1.4 预测 TNM 分期和淋巴转移的临床价值 5
1.2 影像组学及研究现状 6
1.2.1 影像组学介绍 6
1.2.2 影像组学研究现状 7
1.3 主要研究内容及创新 11
1.4 本文组织结构 12
2 相关理论 13
2.1 LASSO 回归分析 13
2.2 XGBoost 13
2.3 常用分类问题评价标准 15
2.4 本章小结 16
3 影像组学特征与实验数据 17
3.1 影像组学特征 17
3.2 实验数据 19
3.2.1 数据来源 19
3.2.2 实验数据选择 19
3.2.3 临床特征处理 22
3.2.4 PET-CT 影像组学特征提取 22
3.3 本章小结 25
4 非小细胞肺癌 TNM 分期预测 26
4.1 研究流程与特征选择 26

4.2 TNM 分期预测结果.....28
4.3 本章小结.....30
**5 非小细胞肺癌淋巴转移预测**.....31
5.1 研究流程与特征选择.....31
5.2 类别不平衡问题.....32
5.3 淋巴转移预测结果.....33
5.3.1 不同类型影像组学特征分别进行淋巴转移预测.....33
5.3.2 Lasso-XGBoost 淋巴转移预测.....36
5.4 本章小结.....39
**6 总结与展望**.....40
6.1 总结.....40
6.2 展望.....41
**参考文献**.....43
**致 谢**.....49
**硕士期间科研项目与成果**.....50

# 1 绪 论

医疗诊断是指在人体具有不正常状态时，找出患病部位、程度及确定病症名称的过程[1]。随着科学技术的不断进步，临床医疗诊断形式也不断发生变化。医学影像学的出现和不断发展使得现代医学更加注重非侵袭的诊断方式与传统医学的结合，在肿瘤学医疗诊断过程中，一般是以多项医学检查的综合信息诊断是否出现癌症。临床中常用的活组织检查虽然可以为肿瘤医学诊断提供非常丰富的信息，但是这种方法也存在局限性，首先活组织检查是以侵袭的方式获取部分肿瘤组织，对人体有一定伤害；其次肿瘤在时间和空间都具有异质性，随着时间和空间的变化需要多次提取肿瘤组织，这又增加了患者的风险[2]。医学影像学则避免了上述问题，医学影像不具侵袭性，且大部分的影像仪器凭借少量代谢过程信息就可以刻画出肿瘤的解剖学和形态学特征[3]，因此医学影像在肿瘤医疗诊断中可能发挥更大的作用。但是，目前医学影像在临床应用中依然很大程度依赖医生的经验对影像进行解释，这无疑提高了临床实践对医生的要求，而且查看影像大量耗费医生宝贵的时间，严重影响医生的诊疗效率。影像组学的出现为肿瘤医疗诊断带来了新方向[4,5]，影像组学具备处理并分析大量医学数据的能力，能够从中获取有价值的医学信息，如肿瘤的解剖学和形态学特征等，并且可以分析肿瘤的异质性，能够帮助开发预测、预后模型用于精准诊疗，为肿瘤学的发展助力。影像组学作为新兴领域，虽然目前在临床实践应用方面仍存在一些挑战，但展望未来，影像组学将会成为图像驱动信息集成的一个关键组成部分，在精准治疗中发挥巨大作用。

## 1.1 研究背景与意义

### 1.1.1 肺癌

随着经济和科技的不断发展，曾经对人类有强烈威胁的一些烈性疾病得到预防，人类的寿命不断增加，同时也使得癌症成为现代社会威胁人类健康的主要疾病之一。在发达国家中，癌症是发病率和死亡率最高的疾病之一，其在发展中国家的发病率和死亡率也不断增加。根据《2020 全球癌症报告》数据[6]显示，在未来 20 年中，全世界癌症病例数可能会增加 60%，在中低收入国家，增幅甚至可能高达 81%。癌症中最常见的类型包括肺癌、乳腺癌、结直肠癌、前列腺癌、胃癌和宫颈癌等，其中，肺癌是全球发病率和死亡率位于前列的癌症之一。根据中国国家癌症中心发布的数据[7]，癌症已成为中国

居民首要的死亡原因，其中肺癌居我国恶性肿瘤发病率和死亡率首位。

肺癌在临床中的分类一般有两种：小细胞肺癌（Small Cell Lung Cancer, SCLC）和非小细胞肺癌（Non-Small Cell Lung Cancer, NSCLC），其中小细胞肺癌约占全部肺癌患病人数的 25%，非小细胞肺癌占比约为 75%。在非小细胞肺癌中又包括两种最常见的组织学亚型，分别是鳞状细胞癌和腺癌，这两种类型的癌症在非小细胞肺癌中的占比分别约为 20%和 38%。鳞状细胞癌和肺腺癌具有明显不同的组织学特性，并且分布位置通常也存在明显差异，鳞状细胞癌一般与细胞间桥和单个细胞角化珠相关，大多数情况下位于肺的中间位置，而肺腺癌则是腺体结构更加显著，更多情形是分布在肺的周围[8-10]。非小细胞肺癌的组织学亚型是关乎系统治疗方案的关键影响因素[11]，目前临床实践中通常是采用活组织检查获取肿瘤的组织学信息。

癌症发生时通常形成于身体的某个部位或者器官，也称之为原发瘤，随着病情的持续发展，癌细胞可以脱离原发部位，穿过附近的淋巴结或血管壁，或是通过血液和淋巴系统在体内移动，转移到身体其他组织，并在这个组织中生长成一个小肿瘤，称其为转移瘤，一旦出现这种情况就说明患者体内肿瘤出现了转移。肺癌最常见的转移方式包括淋巴结转移和血液转移，淋巴结转移一般会出现肺门淋巴结、纵膈淋巴结以及锁骨上淋巴结的转移等多种方式。肿瘤转移与较差的预后有关，并且对患者的生存率也会产生不良影响。通过预测肿瘤的转移，就有可能提前对高危患者进行早期强化治疗，进而帮助提升患者的生存期和预后[12,13]。

### 1.1.2 PET-CT 介绍

PET 全称为正电子发射断层扫描技术（Positron Emission Tomography, PET），是一种基于检测少量正电子发射体标记物的活体生理参数定量测量的核医学成像技术。核医学中最常用的放射性示踪剂包括碳-11、氧-15、氮-13 和氟-18（11-C、15-O、13-N 和 18-F），当给病人注射极少量的放射性示踪剂时，示踪剂会逐渐在生命体内发生衰变（不同示踪剂衰变速度不同，如 15-O 的半衰期为 2 分钟，而 18-F 的半衰期为 109 分钟），衰变过程中示踪剂的放射性原子转化释放的正电子在生命体组织中移动几毫米后与电子结合，生成一对能量近乎相等但方向相反的光子（γ 射线），此时通过一对共线排列的探测器就可以检测到这对对向的光子。PET 图像采集就是基于这一对光子的同时检测，PET 扫描仪由环绕病人的许多光子探测器组成，在扫描过程中，PET 会同时收集数百万对对向的光子，环状分布使得 PET 可以在大量角度和辐射距离下沿着相关器官的路线测量放射

性，多角度的信息被用于重建断层图像的区域放射性分布。在恶性细胞中，由于己糖激酶活性的上调，恶性细胞对葡萄糖的利用增加，此外，由于人体部分组织特有的功能特性，这些组织对葡萄糖需求量相对较大或者用于存储人体代谢物（大脑、心脏、膀胱等），因此病变部位和部分人体组织在 PET 上都会显现出显著的高代谢亮信号[14,15]。

PET 有能力显示在分子水平上形态尚未显示异常的器官的异常代谢活动，对患者的诊断和随访至关重要。PET 在医学中常用于各类恶性肿瘤疾病的诊断、分期和随访过程，如孤立性肺结节、非小细胞肺癌、淋巴瘤、黑色素瘤、乳腺癌、结直肠癌等；PET 对于化疗或手术切除肿瘤的患者也很有用，因为大多数患者由于术后改变或疤痕组织往往在 CT 或 MR 上有复杂的表现。

CT 全称为电子计算机 X 射线断层扫描技术（Computed Tomography, CT），顾名思义，CT 是利用一系列的 X 射线对人体进行体层检查。X 射线管与位于病人另一侧的 X 射线探测器紧密相连，二者同时扫过病人，将狭窄的 X 射线束扫过切片。当 CT 扫描仪从不同角度向病人发射 X 射线时，扫描仪中的探测器就会测量被身体吸收的 X 射线和穿过身体传输的 X 射线之间的差异，这个过程称之为衰减。由于人体不同组织的密度不同，X 射线穿过不同组织时的吸收效果呈现差异，衰减量也就不同。人体组织密度越高，吸收的 X 射线也就越多，探测器收到的信号也就越弱，反之，组织密度越低，探测器收到的信号也越强。通过从不同角度获取投影，探测器收集到的 X 射线衰减信息会用数学算法反映到监视器上，通过图像重建，人体的三维平面成像就可以在二维监视器中显示，进而呈现出清晰的人体横断面解剖结构信息[16]。基于解剖信息的各种成像模式对肿瘤疾病的诊断和随访十分重要，但是它们也存在一个显著的缺点，即不可以检测到人体中形态显示正常但功能异常的组织。

PET-CT 则是融合了 CT 提供的横断面解剖信息和 PET 提供的代谢信息。相比于单独的 PET 图像，PET-CT 能够准确定位放射性示踪剂活性增加到特定的正常或异常解剖位置。PET-CT 提供的功能和解剖信息能够检测到形态学定义正常但是代谢功能异常的组织，弥补了基于解剖信息的成像模式的不足，对肿瘤患者的护理至关重要。PET-CT 在肿瘤疾病的管理中发挥着越来越重要的作用，在非小细胞肺癌、淋巴瘤、结直肠癌和食管癌、黑色素瘤、头颈癌、乳腺癌的诊断、分期和随访以及描述单发肺结节的特征方面得到广泛的应用[17-22]。图 1.1 为 PET-CT 示例图：

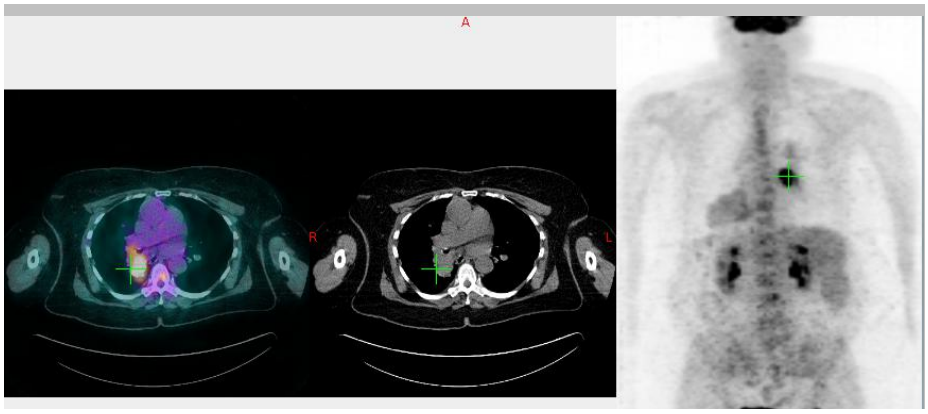


**图 1.1　PET-CT 示例图。左边为 PET 图像，中间为 CT 图像，右边为 PET 最大值投影图像。绿色十字标注区域为肿瘤**

从图 1.1 可以看出，大脑作为人体最为活跃的组织之一，葡萄糖代谢频繁，因此其亮度较人体其它组织更高，同理心脏也有较高的亮度，由于肾脏和膀胱是人体泌尿系统的重要组成部分，负责细胞代谢废物的运输，因此在肾脏和膀胱也会有较高的亮度，除部分正常人体组织显示高亮之外，恶性肿瘤的葡萄糖代谢相较人体周围组织更为频繁，因此也会显示出高亮度的示踪剂痕迹。

### 1.1.3 肺癌 TNM 分期

TNM 分期是国际抗癌协会对恶性肿瘤的分类方法，对肿瘤分期、治疗方案的制定和评估预后具有指导作用。不同的癌症的 TNM 分期系统各不相同。国际肺癌协会于 2015 年更新修订了第八版国际肺癌 TNM 分期标准，对肺癌分期系统进行了更新，详细的分期标准如图 1.2 所示[23]。

| | N0 | N1 | N2 | N3 | M1a any N | M1b any N | M1c any N |
|---|---|---|---|---|---|---|---|
| T1a | IA1 | IIB | IIIA | IIIB | IVA | IVA | IVB |
| T1b | IA2 | IIB | IIIA | IIIB | IVA | IVA | IVB |
| T1c | IA3 | IIB | IIIA | IIIB | IVA | IVA | IVB |
| T2a | IB | IIB | IIIA | IIIB | IVA | IVA | IVB |
| T2b | IIA | IIB | IIIA | IIIB | IVA | IVA | IVB |
| T3 | IIB | IIIA | IIIB | IIIC | IVA | IVA | IVB |
| T4 | IIIA | IIIA | IIIB | IIIC | IVA | IVA | IVB |

Goldstraw P et al. J Thorac Oncol 2016; 11: 39-51.

**图 1.2　第八版肺癌 TNM 分期系统**

在 TNM 分期系统中，T 期描述原发肿瘤的范围，N 期描述区域淋巴转移的存在与否及范围，M 期描述是否存在远端转移，T、N 和 M 后面的数字详细地描述肿瘤的大小和位置。根据第 8 版肺癌 TNM 分期，一共有 4 种基本类型，分别是 I 期、II 期、III 期和 IV 期，在四种基本类型下，又详细地分为 11 种具体的肿瘤发展状况描述（IA1~IVB），越往后肺癌患病情况越严重。本文将按照肺癌 TNM 分期的 I 期、II 期、III 期和 IV 期为标准，对 TNM 分期进行预测。

### 1.1.4 预测 TNM 分期和淋巴转移的临床价值

预测病人的预后一直是重要的医学实践之一，患者预后的评估受多种因素的影响，包括临床表现、功能形态、组织病理学判断，疾病发展程度以及一些生物学因素等。肿瘤大小、淋巴结状态、TNM 分期都为肿瘤评估、治疗和预后提供了基础信息[24]。TNM 分期系统是一种恶性肿瘤的分类方法，描述了肿瘤的扩散程度，作为一种解剖学定义的分类系统，TNM 系统详细描述了原发肿瘤、局部淋巴结范围以及肿瘤是否出现转移。Halsted 等人[25]认为实体瘤经过一系列的分期，从原发部位通过淋巴管一直到远端的其他组织，随着时间的推移而连续扩散，每个分期的预后也越来越差。

临床治疗过程中，肿瘤患者的 TNM 分期应该通过活组织检查确定，但是由于实际情况的约束，只有少数患者会进行活组织检查，大部分患者的 TNM 分期由医生查看患者的医学影像确定，查看医学影像不仅需要花费医生许多的时间，而且可能存在人为误差。影像组学特征在作为肿瘤潜在分子特性的替代标记物方面具有很大潜力，使其成为非侵袭方式分析癌症特性，辅助诊断和预后评估的重要手段。利用影像组学可否进行组织学分类和 TNM 分期预测，是一项非常值得研究的问题[26]。

无论是从基因组学还是从影像组学的角度，预测肿瘤的转移一直都是现代医学领域探索的热点话题。在肿瘤的发展过程中，肿瘤转移是最具风险的情形之一，例如对于乳头状甲状腺癌患者来说，一旦出现淋巴转移，就意味着病情复发率的增加和总生存期的下降[27]。对肺癌患者来说，出现淋巴转移也会对患者的预后和生存期产生不良影响。临床治疗过程中及时准确地预测肿瘤是否出现转移对临床决策过程有重要的指导意义，很大程度决定了患者是否需要进行手术治疗[28]。在外周直径小于 3cm 的肺癌患者中，约有 5.6%-20%的非小细胞肺癌患者临床病理诊断为 N1 或者 N2，其余大部分患者的病理诊断为 N0[29-31]，临床病理判断是否准确，深刻影响着非小细胞肺癌患者的预后和最佳治疗方案的选择。

近年来随着机器学习和深度学习在医学数据中的应用，也产生了大量的基于影像组学特征的机器学习方法在预测肿瘤转移方面的研究。影像组学特征很好地对肿瘤表型包括瘤内异质性等进行量化的描述，为预测肿瘤转移提供了新的思路和方向。

## 1.2 影像组学及研究现状

### 1.2.1 影像组学介绍

在肿瘤医学中有“精准肿瘤学”的概念，其本质是通过一系列组学技术和医学前沿技术，对大样本人群和特定疾病进行生物标志物分析与鉴定，以实现对某种疾病的状态和过程的精确分类，最终目的是为了对疾病和特定患者进行个性化的精准治疗。精准治疗有助于最大限度地达到预防和治疗干预的效果，并且副作用最小。目前实现精准治疗难度较大，且大都是基于基因组学和蛋白质组学等技术，因此需要获取肿瘤活组织，确定活组织的分子特性，进而制定进一步的治疗方案。尽管一些基于基因和蛋白质组学的精准治疗方法有很大的应用前景，但是由于其成本高昂、特定分组靶向药物的短缺、临床数据的缺乏和基因突变的多样性和复杂性等多种因素的影响，精准医疗的临床实践受到很大的限制，而且由于肿瘤组织在时间和空间上都存在异质性，所以基于基因组学和蛋白质组学技术的精准医疗需要多次提取肿瘤活组织，重复提取肿瘤组织又会增加患者的风险。医学影像学则避免了上述方法存在的局限性，医学影像学以非侵袭的方式获取肿瘤的关键性病理信息，包括瘤间异质性，有助于癌症的临床诊断，影像组学概念也因此被提出[32,33]。

在过去的数十年，随着医学影像数据规模的急剧增长和模式识别工具的发展，医学图像分析领域取得了长足的进步，也促进了影像组学的发展。影像组学是医学影像学和分子医学领域关注的重要话题[34]，与传统影像医学将医学图像视为图片，其作用完全依赖医生对图像的视觉理解的理念不同，影像组学是从医学图像中高通量提取定量特征，并通过对量化特征的定量分析进一步支持临床决策的过程，具体体现在影像组学通过提取和挖掘大量的医学图像特征来量化肿瘤表型的特性[35,36]。影像组学数据包含一阶、二阶和高阶统计量，通常在实践中是将影像组学数据与患者的其他数据结合，并利用复杂的生物信息学工具进行挖掘，以便更好地实现临床决策支持的目的，如潜在地提高患者的诊断、预后和预测精度。影像组学研究方法的基本流程如图 1.3 所示[35]：

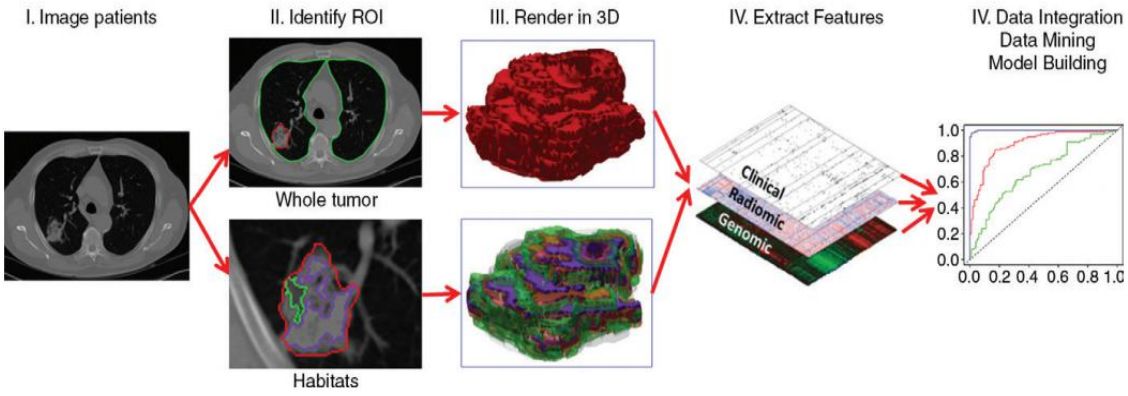



**图 1.3 影像组学研究流程示意图。**

影像组学的研究大致包括以下几个部分：首先是影像数据的获取。核医学领域有多种影像技术，包括 CT、PET、MRI、超声图像以及 PET-CT、PET- MRI 一体机等多种形式，这些医学图像都可以作为影像组学的研究内容；其次，在获得影像数据后，需要对感兴趣区域（ROI）进行分割，ROI 区域是指整个肿瘤区域或是肿瘤内的亚区域，这一部分是影像组学中更为关键的一步。目前 ROI 的提取有两个来源，一是由具有丰富临床经验的影像科医生手动勾画出肿瘤的 ROI，二是采用深度学习的方式，实现肿瘤的自动分割，近年来有很多学者也专注于各类型肿瘤的分割。这两种方式各自有优劣，具体的分割过程还是要依据研究目标而定；最后，根据得到的 ROI 高通量提取影像组学特征，并将其与临床数据、基因数据等其他的患者数据相结合，根据研究目的进行进一步的数据分析，开发诊断、预测和预后模型，为临床提供辅助决策支持。

### 1.2.2 影像组学研究现状

影像组学作为一个新兴的研究领域，还有广阔的发展空间，近年来也有大量基于影像组学的研究。目前影像组学研究主要是利用各种医学影像数据，通过机器学习或者深度学习的方式，对临床实践中存在并需要解决的医学问题进行研究。首先在实验数据的获取上，由于医学数据的隐私性和复杂性，目前大部分研究数据的获取都是研究机构自行与某些医疗机构合作，进行数据的采集，不同的影像设备之间也存在成像、参数等各方面的差异，因此影像数据呈现多样性和复杂性，而且数据的标准也不尽相同，尽管现在也有一些公开的数据集可以用来研究，但是公开数据集还存在一些问题，如公开的数据集太少、影像模式单一等，因此整体来说，数据的标准化也是目前影像组学研究的问题之一。其次在研究方法上，各个研究的选择标准也有所不同，与此同时，影像组学特征具有高维性，医学数据又难以采集，所以数据样本相较影像组学特征远远不足，所以

在研究中会涉及降维或特征选择的问题，不过，随着深度学习的发展，也有一些深度学习的方式直接对图像进行分析得出研究结果。

Aerts 等人[37]利用 1019 位肺癌或头颈癌患者的 CT 数据，提取 440 个基于肿瘤强度、形状和纹理的量化特征，并将其与患者的临床数据和基因数据结合，通过绘制放射性热图，展示了放射性热图与不同临床分析结果之间的联系。这项研究说明了影像组学特征在肺癌和头颈癌各自的独立数据集中具备预后的能力，首次证明了影像组学特征能够捕获肿瘤的瘤内异质性，与肿瘤潜在的基因表达模式相关。

Parmer 等人[38]用 878 例头颈癌和肺癌患者的治疗前 CT 图像提取 440 个影像组学特征，为了减少冗余并比较影像组学特征在两种不同癌症类型中的预后性，Parmer 等人在 4 组不同的肺癌和头颈癌患者的数据中研究指定癌症的影像组学特征簇。结果表明在肺癌和头颈癌中分别有 11 个和 13 个稳定的影像组学特征簇，实验说明无论是不同类型癌症间共同的影像组学特征还是某一类型癌症特有的影像组学特征，都与实际临床相关联，并且能够较好的预测肿瘤分期和预后（肺癌预后 CI=0.60±0.01；头颈癌预后 CI=0.68±0.01；肺癌组织学分类 AUC=0.56±0.03,；肺癌分期 AUC=0.61±0.01；头颈癌 HPV AUC=0.58±0.03；头颈癌分期 AUC=0.77±0.02），结果证明了肿瘤特有的影像组学特征簇能进一步改善放射生物标志物，为临床实践中量化和监测肿瘤表型提供一种非侵入式的方法。

准确识别根治性手术后出现预后不良的患者是结肠癌临床治疗中的关键性问题之一。为此，Dai 等人[39]利用影像组学特征，对 TNM 分期为 I 期、II 期和 III 期结肠癌患者的预后进行预测，期望利用与死亡和复发相关的特异性影像组学特征实现对结肠癌患者总生存期和无复发生存期的预测，以此实现对预后不良的结肠癌患者的识别，为临床决策提供辅助。此项研究通过 701 例结肠癌患者的 CT 图像共提取出 647 个影像组学特征，利用 Lasso 特征选择识别出分别与死亡和复发存在显著关系的相关特征，最终选择出 13 个与死亡相关的特异性特征和 26 个与复发相关的特异性特征，时间依赖性的相对工作特征曲线表明特异性特征相对于其他临床病理因素，能够更好地预测患者的总生存期（AUC 值为 0.76），生存决定曲线也证实了两组特异性影像组学特征良好的临床应用性，有助于结肠癌的临床精准治疗。

Kniep 等人[40]根据不同类型脑转移瘤患者的 MRI 影像组学特征，探讨了利用多分类的机器学习方法对诊断时原发病灶未知的脑转移瘤患者进行原发肿瘤类型预测的可行性。在这项研究中，Kniep 等人回顾性地分析了 189 位不同癌症类型癌症患者的转移病

灶，研究的癌症类型包括乳腺癌、小细胞肺癌、非小细胞肺癌、胃肠癌和黑色素瘤。利用随机森林方法对上述癌症类型患者的 MRI 影像组学特征和临床特征进行评估，结果显示在该 5 分类问题中，5 个类别的 AUC 值范围在 0.64 和 0.82 之间；除此之外，他们还将分类结果与放射科医生对同一数据集患者原发瘤类型的预测进行对比，发现分类器的性能要更好。其中，分类器对黑色素瘤的分类提升结果比两位放射科医生的判断敏感性提升了 17%。这项研究说明利用常规脑 MR 图像影像组学特征的机器学习分类器在预测脑转移瘤癌症类型方面有较好的预测性能。

Parmer 等人[33]利用 464 位肺癌患者的 CT 影像数据，共提取了 440 个影像组学特征，并利用 14 种不同的特征选择方法（Relief、Fisher Score、Gini index、Wilcoxon、T-test score、卡方检验、互信息等）和 12 种不同的分类方法（神经网络、决策树、随机森林、支持向量机、Bagging、Boosting、贝叶斯等）对肺癌患者的总生存期进行预测，以比较不同的特征选择方法和分类方法在影像组学中的性能和稳定性。最终结果显示，在该肺癌患者数据集中，基于 Wilcoxon 的特征选择方法（stability=0.84±0.05，AUC=0.66±0.02）和随机森林（RSD=3.52%，AUC=0.66±0.03）有最高的预测结果和稳定性，该项研究说明选择合适的方法是研究放射学标志物的关键因素之一。

正确的癌症分期对于最佳临床治疗方案的选择和制定极为关键，但是目前来说要做到准确的 TNM 分期依然是临床医生面临的重大挑战。并且在前列腺癌中，虽然术前临床试验结果如临床 T 分期、前列腺特异性抗原水平等因素均可被用于预测癌症的病理分期，但是并非每个患者在所有的测试中都会返回异常结果，因此预测前列腺癌分期受到很大限制。针对这一问题，Cosma 等人[41]利用一种神经模糊模型对前列腺癌患者的病理分期进行预测，具体是在患者接受治疗前，对患者癌症扩散的可能性进行估计。

Xiao 等人[42]利用 MRI 影像组学特征和机器学习的方法，对胸腺上皮性肿瘤（TETs）患者的病理分类和 TNM 分期进行预测。通过回顾性的收集 189 例 TETs 患者的 MRI 图像和临床资料，Xiao 等人提取了 2088 个影像组学特征，利用支持向量机选择与病理分类和 TNM 分期最为相关的最优特征，与病理分类和 TNM 分期最为相关的特征分别为 125 个和 69 个，通过上述若干特征和多变量 Logistic 模型的结合分别建立了病理分类和 TNM 分期预测模型，两个模型分别用于区分低危、高危胸腺瘤和胸腺癌，以及 TETs 的早期和晚期。两个模型在测试集中预测的 AUC 值分别为 0.77 和 0.91，实验证明 MRI 影像组学分析具有鉴别 TETs 病例分类和 TNM 分期的潜力。此外还有一些关于各类型影像组学特征在不同临床情形下的预测和预后性能及稳定性的研究，不同的研究分别展

示了影像组学特征对肿瘤组织学分层、肿瘤分期和临床鉴别的预测性能，还有一些研究说明了影像组学特征和潜在基因表达模式之间的联系[43,52]。

除上述研究以外，在利用影像组学特征预测肿瘤转移方面也已经有很多具体的工作。乳头状甲状腺癌是一种发病率较高，但生存率较为稳定的惰性肿瘤，但是临床治疗中往往存在过度诊断和治疗的问题，因此 Liu 等人[53]根据 450 位乳头状甲状腺癌患者的超声图像对其是否出现淋巴转移进行预测。首先从超声图像中共提取 614 个高通量影像组学特征，包括大小、形状、边界、位置、回声模式等，然后通过组合特征选择策略，选取 50 个在预测中表现最佳的特征，包括回声模式、后声学模式、钙化特征，利用支持向量机对乳头状甲状腺癌患者是否会出现淋巴转移进行预测，上述三种类型特征预测的 AUC 值分别为 0.753、0.740 和 0.743，再利用 50 个特征共同进行预测的 AUC 为 0.782，精确度为 0.712，测试集中 AUC 和精确度分别为 0.727 和 0.710。乳头状甲状腺癌回声特征复杂，后区均匀，伴有钙化或多发钙化，该研究说明影像组学特征在无创预测乳头状甲状腺癌的淋巴转移预测方面具有重要价值。

Chen 等人[28]回顾性地收集了 2011 年 1 月到 2013 年 12 月期间的 345 位 I 期肺腺癌患者的 CT 影像数据，以验证影像组学特征在预测 I 期肺腺癌出现气道播散(STAS)中的价值。共提取出 88 个影像组学特征，通过 P 值选出 5 个最佳影像组学特征，并利用朴素贝叶斯分类器构建预测模型，测试集中预测结果 AUC 为 0.69，该研究证明了影像组学在预测 I 期肺腺癌是否出现 STAS 方面的价值。

在非小细胞肺癌患者中，主要有肺腺癌和鳞状细胞癌两种组织学亚型，相比鳞状细胞癌患者，肺腺癌患者更有可能面临原发肿瘤的淋巴转移，此前也曾有人利用纵膈阴性的早期非小细胞肺癌患者的术前 CT 图像中原发肿瘤的一些特异因素来预测患者是否会出现淋巴转移，但是他们对原发肿瘤特异因素的判断来源于个体的视觉评估，这样的情形使得在研究个体和观察者之间都存在一定的误差，可能导致判断不准确。因此，Gu 等人[29]回顾性地根据 501 例 TNM 分期为 T1N0M0 的肺腺癌患者的 CT 图像、临床切除术以及系统淋巴结清扫或是淋巴结取样的记录进行研究，通过提取 CT 图像中的纹理特征，并结合临床特征包括年龄、性别、吸烟史、癌胚抗原水平（CEA）等，利用多变量 Logistic 回归模型，对 501 例患者是否会出现淋巴转移进行预测。模型预测结果显示验证集的 AUC 为 0.808，说明影像组学特征潜在地表达了肿瘤的表型、异质性等信息，并且影像组学特征与癌胚抗原水平的结合能够很好地对肺腺癌患者是否出现淋巴转移进行预测，可以帮助外科医生更好地做出后续的临床决策。

淋巴转移是胃癌中导致预后较差的主要风险因素之一，且胃癌淋巴转移发生率高达25%，内镜切除术不仅周期长而且结果慢，影响患者的及时治疗，因此，无创评估胃癌淋巴转移十分必要。Gao 等人[54]对 768 例胃癌患者进行了回顾性研究，基于 CT 影像组学特征对胃癌的淋巴转移进行了预测。首先根据是否出现转移将患者分为没有转移和有转移两组，对两组数据用曼惠特尼 U 检验进行检验，将具有统计学差异的特征纳入分析，然后用最小绝对收缩和算子（LASSO）方法进行选择，选出与肿瘤淋巴转移相关的特征，并将选定的影像组学特征与临床特征结合，用单变量和多变量 Logistic 回归模型对胃癌患者是否出现淋巴转移进行了预测，最后用决策曲线分析（DCA）评估了他们的模型在临床中的效应。模型预测总体精确度为 0.61，灵敏度为 0.5，特异度为 0.78，在验证集中模型预测淋巴转移阳性和阴性精确度分别为 0.87 和 0.52，AUC 为 0.82，决策分析曲线（Decision Curve Analysis, DCA）表明根据胃癌患者的 CT 图像，影像组学模型可以提供比淋巴结状态更多的临床净效益。Liu 等人[55]利用 62 位乳腺癌患者的动态对比增强 MRI（DCE-MRI）图像，对其是否出现前哨淋巴结转移进行了预测。除上述研究之外，还有很多利用影像组学特征对肿瘤转移预测的研究[56-59]。

## 1.3 主要研究内容及创新

本文以非小细胞肺癌患者的 PET-CT 图像为载体，以医生给定的肿瘤金标准为模板，从 PET-CT 图像中提取非小细胞肺癌患者的 PET-CT 影像组学特征，并利用机器学习的方法，对非小细胞肺癌患者的 TNM 分期和淋巴转移进行了预测。首先是特征的选择，将 TNM 分期数据集和淋巴转移数据集中分别提取的 833 个特征利用曼惠特尼 U 检验、方差过滤和 LASSO 进行特征选择，然后将选择的特征分别用于 TNM 分期和淋巴转移预测。以下是本文研究的创新点：

（1）利用影像组学特征进行 TNM 分期和淋巴转移预测，其中淋巴转移预测模型的的性能要高于医生的判断，尤其是对出现淋巴转移的患者，这样有助于癌症临床治疗中的医生对患者病情的判断、治疗方案的制定和预后评估。

（2）对医学研究中普遍存在的数据不平衡问题，加入 SMOTE 数据平衡策略，使得淋巴转移预测结果更加均衡稳定。

## 1.4 本文组织结构

本文基于非小细胞肺癌的影像组学特征，对肺癌患者的 TNM 分期和是否会出现淋巴转移进行了预测，具体的论文组织结构如下：

第一章：绪论。这一章主要阐述研究的背景和意义，包括对肺癌、PET-CT 的介绍以及预测 TNM 分期和淋巴转移的现实价值。接着对影像组学和影像组学发展和研究现状进行详细的介绍。

第二章：相关理论。简要介绍了本文用到的理论知识，包括 LASSO 特征选择算法，XGBoost 预测方法和常用的分类问题评价标准。

第三章：影像组学特征与实验数据。主要对影像组学特征、实验数据的来源、选择标准进行详细介绍。首先是影像组学特征及其类型，并对不同类型的影像组学特征进行简要描述，其次是本文具体的实验数据介绍。

第四章：非小细胞肺癌 TNM 分期预测。这一章主要对 TNM 分期预测研究流程、特征选择以及预测结果进行详细地介绍。

第五章：非小细胞肺癌淋巴转移预测。主要讲述淋巴转移预测研究流程，数据的处理过程以及实验的结果，并对不同的实验结果进行对比和分析。

第六章：总结与展望。对研究的主要工作进行总结，并指出了本文的不足以及未来进一步工作中需要注意的问题。

# 2 相关理论

对研究中用到的主要研究方法的基本理论进行介绍，包括 LASSO 特征选择方法和 XGBoost 分类方法，并且对本文用到的分类问题评价标准如精确度、精准率、召回率和 AUC 进行简要介绍。

## 2.1 LASSO 回归分析

最小绝对收缩和算子（Least Absolute Shrinkage and Selection Operator, LASSO）由 Robert Tibshirani 于 1996 年首次提出，是一种正则化和特征选择的方法。LASSO 是结合最小二乘损失和 1 范式约束或约束系数绝对值和的方法。通过向模型参数的绝对值和施加约束，就可以将数据中不重要的变量（系数接近于 0）挑选出，实现特征的选择[60]。LASSO 解决了高维数据中普遍存在的稀疏性问题，尤其适合于存在多重共线性的模型。

## 2.2 XGBoost

集成学习是机器学习中的一种重要思想，主要是通过构建并结合多个学习器来完成学习任务，通过结合多个学习器，集成学习往往能够获得比单个学习器更加优越的泛化性能[61]。集成学习主要分为两类——Bagging 和 Boosting。Bagging 是一种并行式的集成方法，多个学习器之间不存在依赖关系，可同时生成；Boosting 则恰好相反，多个个体学习器之间存在强依赖关系，因此必须串行生成。提升树（Boosting Tree）是一种非常具有影响力且被广泛应用的分类算法，是 Boosting 算法的一个分支。关于提升树，已经有很多的工作，XGBoost 作为其中的一种模型，由于其出色的性能被广泛的应用于各类机器学习竞赛中。

XGBoost 是基于梯度提升树（GBDT）算法的极端梯度提升，基本思想与 GBDT 相同，但是在许多方面进行了优化，如利用二阶泰勒公式展开优化损失函数，提升计算精确度、采用正则化项简化模型，避免出现过拟合、采用 Blocks 存储结构，使得算法可以并行计算等。其主要思想是通过持续的特征分裂来不断地添加树，以拟合上一次预测的残差，通过不断的减少残差来达到更好的分类效果。如图 2.1 所示[62]，我们以预测某一家庭不同成员对电子游戏的喜好程度为例，说明 XGBoost 中蕴含的集成树思想：首先考虑成员的年龄，年轻人相比年长的人更喜欢打游戏，而男性相较于女性更偏好电子游戏，因此就可以根据年龄和性别生成一棵分类树，然后根据每天使用电子产品的时长再进行

判断，一般较长时间使用电子产品的人更喜欢电子游戏，由此，就可以根据若干个树的预测结果，对不同目标进行打分，进而获得最终的预测结果。

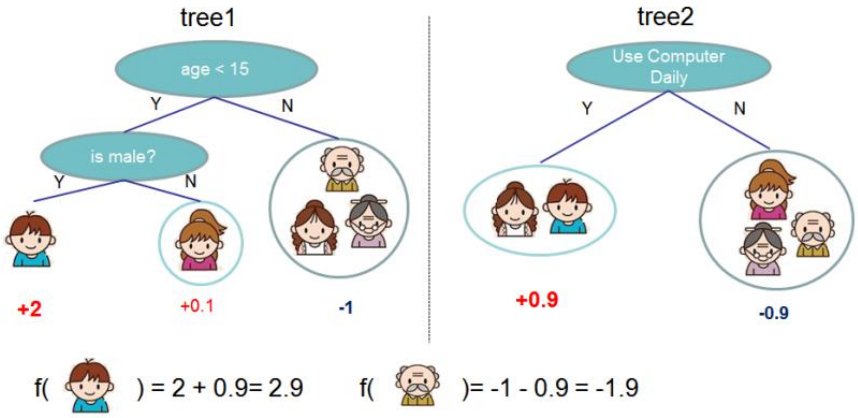



**图 2.1：集成树模型**

XGBoost 的目标函数由损失函数和正则化项两部分构成。对某已知数据集 $D=\{(x_i,y_i)\}$，其目标函数[62]为

$$L(\phi)=\sum_i l(\hat{y}_i,y_i)+\sum_k \Omega(f_k)$$
$$\text{其中}\Omega(f)=\gamma T+\frac{1}{2}\lambda\|w\|^2 \quad (2.1)$$

$l$ 为损失函数，描述的是预测结果 $\hat{y}_i$ 和目标 $y_i$ 之间的差异；$\Omega$ 降低了模型复杂度。当以加法模式训练模型时

$$\hat{y}_i=\sum_{k=1}^{k} f_k(x_i)=\hat{y}_i^{(t-1)}+f_t(x_i)$$

因此，目标函数为

$$L(\phi)=\sum_{i=1}^{n} l\left(y_i,\hat{y}_i^{(t-1)}+f_t(x_i)\right)+\Omega(f_t) \quad (2.2)$$

接着利用二阶泰勒展开和正则化项展开，对损失函数和正则化项进行优化，再通过合并同类项，得到最终的目标函数

$$L(\phi)=\sum_{j=1}^{T}\left[G_j w_j+\frac{1}{2}(H_j+\lambda)w_j^2\right]+\gamma T \quad (2.3)$$

其中，$G_j$ 为叶子结点 $j$ 所含样本的一阶偏导累加和，$H_j$ 为二阶偏导累加和，二者均为常数，因此，第 $t$ 棵树的权重向量 $w$ 为目标函数的唯一变量。

在模型实际训练过程中，通过采用分块并行、缓存访问和核外 Block 计算，XGBoost 可以用最少的资源解决大规模计算问题，具有高效、灵活和轻便的特点，在结构化或表格数据中有非常好的性能，被广泛应用于垃圾邮件分类、高能物理事件分类、恶意软件识别、商店销售预测、产品分类、灾害风险预测等多个领域，也成为各大数据挖掘挑战赛的宠儿。

## 2.3 常用分类问题评价标准

在分类问题中，精确度是最常用的评价标准之一，表示所有样本中被正确分类的样本占所有样本的比例。但是在普遍存在数据不平衡的医学研究问题中，不能仅以预测精度为评价标准，在弥漫性大 B 细胞淋巴瘤患者中，只有 16%的患者呈现骨髓受累，所以若在评价影像组学检测骨髓受累的问题时，若仅将精确度作为衡量标准，在分类方法将所有样本均预测为没有出现骨髓受累时，预测精确度为 84%，依然是一个很高的值，但是这样的结果没有任何价值。因此在医学问题中衡量结果时，必须充分考虑数据不平衡问题，多方面对结果进行综合评价，可以根据总体精确度、分类精确度、敏感性和特异性对结果进行综合评价[34]。在本次研究问题中，我们利用精确度，混淆矩阵、精准率、召回率以及受试者工作特征曲线下的面积（AUC）从多个角度对预测结果进行评价。

（1）精确度（Accuracy）

在分类任务中常用的度量是精确度与错误率，错误率即分类错误的样本数量占所有样本数的比例，若在样本数目为 m 的样本中有 a 个分类错误，那么错误率 $E = a/m$，精确度则为 1-错误率，即 $accuracy = 1 - a/m$，表示分类正确的样本占所有样本数的比例。

（2）混淆矩阵（Confusion Matrix）、精准率（precision）与召回率（Recall）

对于二分类任务，将样本真实的类别和分类模型预测的类别分别组合一共有四种情形，分别是真正例（True Positive,TP）、假正例（False Positive,FP）、真反例（True Negative,TN）和假反例（False Negative,FN），则混淆矩阵[63]如表 2.1 所示：

**表 2.1 二分类混淆矩阵**

| 预测结果 | 真 实 类 别 | |
|---|---|---|
| | 正例 | 反例 |
| 正例 | TP | FP |
| 反例 | FN | TN |

资料来源：wikipedia.confusion_matrix

混淆矩阵清晰地展示了模型在不同类别样本中的预测性能，有助于实验者更有针对性地改进模型。由混淆矩阵得到精准率和召回率，精准率表示全部预测为正例的样本中真实类别为整理的样本所占的比例，而召回率则表示所有真实类别为正例的样本中正确预测为正例的样本的比例。一般情况下，精准率高时召回率偏低，而召回率高时精准率则相对较低。

$$precison = \frac{TP}{TP + FP} \tag{2.4}$$

$$recall = \frac{TP}{TP + FN} \tag{2.5}$$

同理，对于多分类任务来说，同样也存在混淆矩阵，其基本思想与二分类任务相同。只是将其中某一类看做正样本，其余样本看作负样本，以此进行相应的计算。

（3）受试者工作特征曲线（ROC）下的面积（AUC）

根据模型对样本的预测结果对样本进行排序，并按此顺序逐个将样本作为正例进行预测，每次都可以计算出两个重要的值，分别是真正例率（True Positive Rate, TPR）和假正例率（False Positive Rate, FPR），分别以 TPR 和 FPR 为纵坐标和横坐标，就可以得到 ROC 曲线图，ROC 曲线图下的面积（AUC）可用于比较分类器之间的性能，AUC 值越大，表示分类器性能越好[61]。

$$TPR = \frac{TP}{TP + FN} \tag{2.6}$$

$$FPR = \frac{FP}{TN + FP} \tag{2.7}$$

## 2.4 本章小结

本章对本文中用到的主要的方法进行了简要介绍，包括 Lasso 特征选择算法，XGBoost 预测方法和分类问题常用的评价标准。

# 3 影像组学特征与实验数据

分子层面的基因或者蛋白质模式改变都能在医学影像中有所表达，通过深入挖掘影像组学特征，就可以从影像角度得知人体的组织、细胞和基因水平的变化。影像组学特征从强度、形状和纹理三个角度对肿瘤进行量化描述。本文利用肺癌患者的 PET-CT 影像数据，从中提取影像组学特征，进而挖掘影像组学特征与非小细胞肺癌患者的 TNM 分期和淋巴转移之间的关系。

## 3.1 影像组学特征

影像组学特征主要包括基于强度的特征、形状特征和纹理特征，可以从原始图像提取特征，也可以用小波滤波器将图像转换之后再提取特征。基于肿瘤强度的特征估计了肿瘤区域强度直方图的一阶统计量，形状特征描述了肿瘤的三维几何属性，而纹理特征量化了肿瘤的异质性。

（1）基于肿瘤强度的特征

基于强度（Intensity）的特征也称一阶特征（Firstorder），来源于图像直方图，是图像强度分布的图形表示，描述了感兴趣区域的强度特征，基于强度的特征将关于肿瘤体积的三维数据简化为单一的直方图。直方图描述了一定体素值范围内选定结构的分数值（如 CT 扫描中的亨斯菲尔德单位），通过直方图就可以计算一些常见的统计数据如均值、中位数、最值、标准差、峰度、偏度、能量、熵、均匀性和方差等，也可以计算更为复杂的数据，如绝对标准摄取值（SUV）为 5 以上的代谢量或是 CT 测量的高密度组织的比例等[64]。需要注意的是，参数的设置会影响一些特征的取值，因此在实际使用过程中，需要对特征进行归一化处理[65]。

（2）形状特征

形状（Shape）特征描述了肿瘤的几何性质，基于形状的特征相较于其他影像组学特征要简单的多，仅需要肿瘤的二维或者三维直径、坐标轴及大小比例。形状特征包括紧密度、球度和密度等，紧密度和球度描述了肿瘤的形状和标准圆（二维情形下）或者球体（三维情形下）之间的差异，而密度则依赖于包围肿瘤的最小三维体或者矩形[66]。

（3）纹理特征

在图像分析领域，纹理（Texture）是特征的定义集合之一，纹理特征最初是用来衡量二维图像中感兴趣区域的表面纹理，后来逐渐发展延伸至三维领域。纹理特征从图像

中识别肿瘤的重要特性，反映了像素强度的变化及相互之间的关系，被广泛应用于医学模式识别任务中。纹理特征作为高阶统计描述性特征，描述了具有相似(或不同)对比值的体素之间的统计相互关系，它包含像素或体素关于表面结构及其与周围环境关系的重要信息。纹理特征不仅可以从二维切片中进行提取，也可拓展至三维空间。在计算纹理特征之前，通常对图像灰度进行离散化处理[67]。

常用的纹理特征包括基于灰度游程矩阵（GLRLM）、灰度大小区域矩阵（GLSZM）、邻近灰度区域差异矩阵（NGTDM）、灰度相关矩阵（GLDM）和灰度共现矩阵（GLCM）的特征。GLRLM 特征由 Galloway 等人[68]提出，提供了具有相同灰度级的连续像素在一个或多个方向上的空间分布信息，包括分数（反映颗粒度）、长（短）游程增强、灰度和游程不均匀性等；GLSZM 计算具有相同灰度相邻像素（体素）的数量；NGTDM 由 Amadasun 等人[69]提出，主要作用是量化像素（体素）与限定距离内相邻像素（体素）平均灰度级之间的差值总和，主要特征有粒度、忙度和复杂度；GLDM 衡量中心像素（体素）与周围像素之间的灰度级关系[34]；GLCM 是由 Haralick 等人[67]提出的一种二阶灰度直方图，用以获取具有确定像素或体素之间距离的成对像素或预先确定灰度强度的体素在不同方向（水平、垂直或倾斜）上的空间关系。常用的 GLCM 特征包括熵、能量、对比度等[66]。基于 GLCM 的纹理特征是影像组学中最常用的纹理特征[70]，以 GLCM 的计算为例，说明纹理特征的计算过程。GLCM 是表示相邻像素(或三维体素)的离散灰度组合如何沿着图像的一个方向分布的矩阵。在三维纹理分析方法中，一个体素的直接邻域由 26 个直接邻域体素组成，因此，当距离为 1 时，相邻体积内有 13 个独特的矢量方向，即(0,0,1)、(0,1,0)、(1,0,0)、(0,1,1)、(0,1,-1)、(1,0,1)、(1,0,-1)、(1,1,0)、(1,-1,0)、(1,1,1)、(1,1,-1)、(1,-1,1)和(1,-1,-1)。同样地，在二维图像中，忽略切片之间的联系，一个像素的直接邻域有 8 个像素，距离为 1 时有 4 个矢量方向，即(1,0,0)、(1,1,0)、(0,1,0)和(-1,1,0)。在每个矢量方向计算 GLCM，如令 $M_{\Delta}$ 为 $N_g \times N_g$ 的 GLCM，其中，$N_g$ 为体积中出现的离散灰度数，Δ为特定方向，元素 $(i, j)$ 是离散灰度 i 和 j 的组合在方向 $\delta$ 和 $-\delta$ 的相邻体素中出现的频率，得到 $M_{\Delta} = M_{\delta} + M_{-\delta} = M_{\delta} + M_{\delta}^{T}$，根据给定的方向和距离，可以计算出若干 GLCM，再利用 GLCMs 就可以计算出若干个 GLCM 相关的纹理特征[66,67]。

（4）小波特征

小波（Wavelet）特征是影像组学中的高阶特征，影像组学中常用的高阶特征有小波特征和傅里叶特征，二者的共同原理都是在不同的频率下获取生物标志物。小波是一种

滤波变换，是影像组学中最常用的高阶特征，通过将图像与复杂的线性或径向波相乘，可以从图像中提取纹理图案较为粗糙的区域的特征。小波特征包括强度特征和纹理特征，是这两种特征的变换域表示。本次实验利用 Coiflet 滤波器对图像进行变换并进一步提取特征。

## 3.2 实验数据

### 3.2.1 数据来源

实验数据由美国芝加哥大学医学院支持提供。我们对 2004 年至 2014 年在芝加哥大学医学院接受治疗的、所有经病理证实为非小细胞肺癌（NSCLC）患者的病历进行回顾性审查，将所有符合以下标准的患者纳入研究：（1）完成 PET 阳性肿瘤的 18F-FDG（18 氟-脱氧葡萄糖）PET-CT 检查；（2）无并发医疗诊断或其他原发癌病史。所有患者都完成从头部到脚部的全身 18F-FDG PET-CT 扫描，数据共包括 935 例确诊的 NSCLC 患者。935 例非小细胞肺癌患者病例 PET-CT 图像来源于两种不同厂家的设备，部分患者于 2012 年 3 月 15 日前接受 Reveal HD PET-CT 成像仪扫描，其余患者在 2012 年 3 月 15 日及之后接受 Simens mCT 成像仪扫描。肿瘤金标准（Gold Standard）图像是由具备医师资格的有十年以上临床经验的至少 1 名以上放射科医生使用 MIMvista5.1.2 在 PET-CT 图像中分别进行肿瘤勾画。对于有争议的具体病灶，以多数医生的分割意见为标准，从而得到最终分割标准。

本次研究利用非小细胞肺癌患者的原发病灶进行 TNM 分期和淋巴转移的预测。

### 3.2.2 实验数据选择

（1）TNM 分期预测数据选择标准

在以上 935 例可得数据中，根据以下选择标准确定研究中用于肿瘤 TNM 分期预测的具体实验数据：（1）可获得临床特征包括年龄、性别、吸烟史、primary Site Desc（描述肿瘤位于肺的上叶、中叶、下叶或其他位置）和 Laterality Desc（描述原发肿瘤位于左肺还是右肺）；（2）无脑转移迹象；（3）TNM 分期为 I 期、II 期和 III 期。经过筛选，最终有 324 例患者数据列入 TNM 分期预测研究，其中 TNM 分期为 I 期、II 期和 III 期的样本数分别为 125 例、58 例和 141 例。有 83 例样本接受 Reveal HD PET-CT 成像仪

扫描，241 例接受 Simens mCT 成像仪扫描，具体的实验数据分布情况如表 3.1 所示：

**表 3.1 TNM 分期预测数据分布**

| | 总计 | TNM 分期 | | |
|---|---|---|---|---|
| | | I 期 | II 期 | III 期 |
| 数目 | 324 | 125 | 58 | 141 |
| 性别 | | | | |
| 男 | 134 | 43 | 22 | 69 |
| 女 | 190 | 96 | 36 | 72 |
| 年龄 | | | | |
| 均值 | 69 | 68 | 70.4 | 69 |
| 最小值 | 40.3 | 42.8 | 40.3 | 45 |
| 最大值 | 93 | 89 | 93 | 88 |
| 吸烟史 | | | | |
| 无烟史 | 25 | 14 | 7 | 4 |
| 既往烟史 | 117 | 50 | 19 | 48 |
| 吸烟 | 182 | 61 | 32 | 89 |
| primary Site Desc | | | | |
| Upper | 199 | 76 | 34 | 89 |
| Lower | 92 | 41 | 13 | 38 |
| Middle | 20 | 8 | 6 | 6 |
| Other | 13 | 0 | 5 | 8 |
| Laterality Desc | | | | |
| Left | 129 | 52 | 25 | 52 |
| Right | 194 | 73 | 32 | 89 |
| Other | 1 | 0 | 1 | 0 |

由上述表 3.1 可知，在非小细胞肺癌 TNM 分期预测研究数据中，有烟史或曾经有烟史的人群肺癌患病率明显高于没有烟史的人群，约占患病人数的 92%，肺癌患者年龄分布在 40~90 岁之间，约有 60%的人群原发肿瘤位于肺上叶，其余 40%位于下叶、中部或其他位置，这是由于本数据集中肺癌组织学亚型为腺癌的患者要多于鳞状细胞癌或其他亚型的患者。

（2）淋巴转移预测数据选择标准

根据以下选择标准确定本次研究中用于淋巴转移预测的具体实验数据：（1）接受外科病理学检查（活体组织检查）；（2）可获得临床特征包括年龄、性别、吸烟史、primary Site Desc（描述肿瘤位于肺的上叶、中叶、下叶或其他位置）和 Laterality Desc

（描述原发肿瘤位于左肺还是右肺）；（3）不存在远端转移；（4）无脑转移迹象。经过筛选，共有 123 例未出现远端转移的患者曾接受外科病理学检查，删除其中 2 例有错数据，最终有 121 例数据参与本次研究，其中，出现淋巴转移（LNM +）的数据 19 例，没有出现淋巴转移（LNM -）的数据 102 例。在该 121 例数据中，有 100 例在 2012 年 3 月 15 日之前接受一类 PET-CT 设备扫描，21 例在 2012 年 3 月 15 日之后接受另一类 PET-CT 设备扫描，实验数据中患者的具体的数据分布如表 3.2 所示：

**表 3.2 淋巴转移预测数据分布**

| | 总计 | LNM + | LNM - |
|---|---|---|---|
| 数目 | **121** | 19 | 102 |
| 性别 | | | |
| 男 | 82 | 11 | 71 |
| 女 | 39 | 8 | 31 |
| 年龄 | | | |
| 均值 | 66 | 65 | 67 |
| 最小值 | 42.7 | 54 | 42.7 |
| 最大值 | 87 | 85 | 87 |
| 吸烟史 | | | |
| 无烟史 | 16 | 3 | 13 |
| 曾有烟史 | 41 | 7 | 34 |
| 吸烟 | 64 | 9 | 55 |
| primary Site Desc | | | |
| Upper | 71 | 8 | 63 |
| Lower | 38 | 7 | 31 |
| Middle | 9 | 3 | 6 |
| Other | 3 | 1 | 2 |
| Laterality Desc | | | |
| Left | 56 | 6 | 50 |
| Right | 65 | 13 | 52 |

由表 3.2 可知，吸烟或是曾有吸烟史的患者多于不吸烟的患者，患病人数约为不吸烟患者的 6.5 倍。此外，由数据可知 NSCLC 患者右肺与左肺患病概率大致相同，大部分患者肺部肿瘤位于肺叶的上叶和下叶，少数肿瘤生于肺叶中部，primary Site Desc 和 Laterality Desc 两个位置特征与肺癌的组织学亚型相关。

### 3.2.3 临床特征处理

本次实验数据包含临床特征数据，实验中选择的临床特征具体包括性别（gender）、年龄（age）、是否有吸烟史（smoker）、Primary Desc 和 Laterality Desc。除年龄为数值型数据之外，其余数据均为分类数据。为便于分析对分类数据用数值进行表示，例如在性别特征中，用 0 表示女性，1 表示男性。

### 3.2.4 PET-CT 影像组学特征提取

Pyradiomics[71]是专门用于从医学图像中提取影像组学特征的开源 Python 库，本次实验使用 Pyradiomics3.0.1 版本进行影像组学特征提取。从 PET 图像中共提取强度、形状和纹理特征，也提取了经过小波变换的特征，具体是将图像通过 Coiflet 滤波器[72]分解为 8 个分量（LHH、LLH、LLL、 LHL、HHH、HHL、HLH 和 HLL，其中 L 代表低通滤波器，H 代表高通滤波器，以 LHH 为例，LHH 表示 x 轴方向的低通滤波器、y 轴方向的高通滤波器和 z 轴方向的高通滤波器），然后从每个分量中再提取图像的强度和纹理特征。具体的特征提取过程如图 3.1 所示：

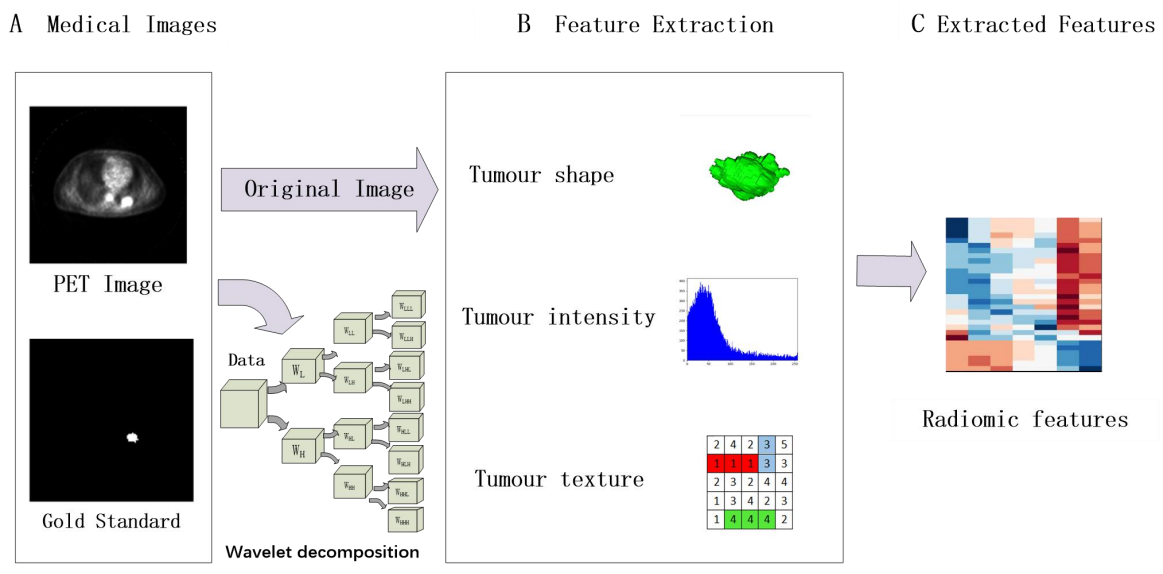



**图 3.1 影像组学特征提取过程**

特征提取过程中，金标准图像提供肿瘤的位置信息，Pyradiomics 利用金标准提供的位置从原始图像中获取肿瘤的形状、强度和纹理信息，从经小波变换后的图像图像中提

取强度和纹理信息。最终从 PET 图像中共提取出 833 个影像组学特征，具体的特征分布统计如表 3.3 所示：

**表 3.3 PET-CT 影像组学特征分布统计**

| 特征类型 | 数目 |
|---|---|
| **PET 影像组学特征** | **833** |
| 强度（Intensity） | 18 |
| 形状（SHAPE） | 14 |
| 纹理（Texture） | 73 |
| GLCM | 22 |
| GLRLM | 16 |
| GLSZM | 16 |
| GLDM | 14 |
| NGTDM | 5 |
| 小波（Wavelet） | **728** |
| Intensity | 144 |
| GLCM | 176 |
| GLRLM | 128 |
| GLSZM | 128 |
| GLDM | 112 |
| NGTDM | 40 |

表 3.4 详细地展示了全局信息、形状信息、灰度信息、GLCM、GLRLM、GLSZM、GLDM 和 NGTDM 类型下提取的特征具体名称。

**表 3.4 影像组学特征名称**

| 类型 | 特征名称 |
|---|---|
| 全局信息 | VolumeNum，VoxelNum |
| 形状信息 | Elongation，Flatness，LeastAxis，Maximum2DDiameterColumn，Maximum2DDiameterRow，Maximum2DDiameterSlice，MajorAxis，Maximum3DDiameter，MinorAxis，Sphericity，SurfaceArea，SurfaceVolumeRatio，Volume |
| 灰度信息 | 10Percentile，90Percentile，Energy，Entropy，InterquartileRange，Kurtosis，Maximum ，MeanAbsoluteDeviation，Mean， Median，Minimum， Range，RobustMeanAbsoluteDeviation，RootMeanSquared，Skewness， TotalEnergy，Uniformity，Variance |

**续表 3.4 影像组学特征名称**

| 类型 | 特征名称 |
| --- | --- |
| GLCM | Autocorrelation，JointAverage，ClusterProminence， ClusterShade，ClusterTendency，Contrast，Correlation， DifferenceAverage，DifferenceEntropy，DifferenceVariance，JointEnergy，JointEntropy，Imc1，Imc2，Idm，Idmn，Id，Idn，InverseVariance，MaximumProbability，SumEntropy，SumSquares |
| GLRLM | GrayLevelNonUniformity，GrayLevelNonUniformityNormalized，GrayLevelVariance，HighGrayLevelRunEmphasis，LongRunEmphasis，LongRunHighGrayLevelEmphasis， LongRunLowGrayLevelEmphasis，LowGrayLevelRunEmphasis，RunEntropy， RunLengthNonUniformity，RunLengthNonUniformityNormalized，RunPercentage，RunVariance，ShortRunHighGrayLevelEmphasis， ShortRunLowGrayLevelEmphasis，ShortRunEmphasis |
| GLSZM | GrayLevelNonUniformity，GrayLevelNonUniformityNormalized，GrayLevelVariance，HighGrayLevelZoneEmphasis，LargeAreaEmphasis，LargeAreaHighGrayLevelEmphasis，LargeAreaLowGrayLevelEmphasis，LowGrayLevelZoneEmphasis，SizeZoneNonUniformity，SizeZoneNonUniformityNormalized，SmallAreaEmphasis，SmallAreaHighGrayLevelEmphasis，SmallAreaLowGrayLevelEmphasis，ZoneEntropy，ZonePercentage，ZoneVariance |
| GLDM | DependenceEntropy，DependenceNonUniformity，GrayLevelVariance，DependenceNonUniformityNormalized，LargeDependenceEmphasis，DependenceVariance，LargeDependenceLowGrayLevelEmphasis，HighGrayLevelEmphasis， LargeDependenceHighGrayLevelEmphasis，LowGrayLevelEmphasis， SmallDependenceEmphasis，GrayLevelNonUniformity，SmallDependenceHighGrayLevelEmphasis，SmallDependenceLowGrayLevelEmphasis |
| NGTDM | Busyness，Coarseness，Complexity，Contrast，Strength |

影像组学特征具有高维性，因此一般在进行建模分析之前，需要对影像组学特征进行选择或降维。虽然从理论上讲，可以将提取的所有影像组学特征作为预测模型的输入，但这种情形下模型所需参数的数量将呈指数型增长，计算成本大幅增加。并且影像组学特征之间往往有较高的相关性，即数据存在冗余性，因此可以适当删除一些特征，也可以分组或采用某些方法如主成分分析或线性判别分析生成具有代表性的特征来代替若干特征[34]。Parmar 等人[33]比较了包括基于互信息等的 14 种特征选择方法和 12 个机器学习分类器方法，Leger 等人[73]也采用了类似的方法。机器学习中有很多特征选择的方法，

如 Relief、Fisher score、Wilcoxon 等，分类方法也有很多，基本的方法如决策树（Decision Tree）、支持向量机（SVM）、随机森林、集成方法（Bagging、Boosting）等。在影像组学研究中，由于数据差异性等各种实际原因，学者们所使用的方法也不尽相同，如 Liu 等人[74]在研究中对特征选择和预测方法分别选用方差膨胀因子和随机森林算法。Ren 等人[75]则是利用 LASSO 回归分析进行特征选择和预测。还有 Liu 等人[53]则是用组合的特征选择方法和支持向量机进行肿瘤转移的预测。总之，在影像组学研究中，大多数情况下需要对特征进行一定的选择和处理，根据研究目的的不同，所选择的方法也有所差异，本文对影像组学特征的处理过程将在之后的内容中详细讲述。

## 3.3 本章小结

本章首先对影像组学特征进行简要介绍，然后具体介绍了实验数据的来源、肿瘤金标准的获取以及本文 TNM 分期预测和淋巴转移预测研究中使用数据的标准和具体的选择过程，分别展示了实验数据的详细分布情况，包括对数据样本的性别、年龄、吸烟史、primary Site Desc 和 Laterality Desc 等临床特征的统计。还介绍了本文提取的基于强度、形状和纹理的影像组学特征的具体分布情况，以及不同特征类型下具体的特征名称。期间对本文 PET-CT 影像组学特征提取过程进行了详细的描述，包括使用的软件、提取的过程。

# 4 非小细胞肺癌 TNM 分期预测

在非小细胞肺癌 TNM 分期预测数据集中共包括 324 例非小细胞肺癌患者，其中 TNM 分期为 I 期患者 125 例，II 期 58 例，III 期 141 例。从患者的 PET 图像中提取影像组学特征，共提取影像组学特征 833 个，包括基于形状、强度和纹理的特征。影像组学特征数远大于样本数，因此需要对影像组学特征进行选择，本文用组合的特征选择策略对特征进行挑选，基于选择后的影像组学特征，用机器学习的方法对非小细胞肺癌患者的 TNM 分期进行预测。

## 4.1 研究流程与特征选择

本章研究内容是基于影像组学特征，利用机器学习的方法对 TNM 分期进行预测，研究分为两个部分：特征选择和 TNM 预测，具体的研究流程示意如图 4.1 所示。

首先是特征的选择。在非小细胞肺癌患者的 PET 图像中，本文共提取原发肿瘤的影像组学特征 833 个。由于影像组学特征的高维性和共线性，如果使用全部特征进行预测，不仅会大量增加计算成本，且预测性能并不会有效提升。因此在进行 TNM 分期预测之前，需要进行特征的选择，根据 TNM 分期预测数据集，影像组学特征的处理过程包括以下三个步骤：

第一步，曼惠特尼 U 检验剔除不同分布的特征。影像组学特征存在不稳定的问题，本文中的 PET 影像数据来自两种不同的 PET-CT 设备，由于不同的影像设备在尺寸和参数之间都存在差异，因此在两组不同来源的影像设备中提取的影像组学特征之间可能存在差异，所以需要选择在两种设备中分布都较为一致的影像组学特征。曼惠特尼 U 检验（也称 Wilcoxon 秩和检验）作为一种非参数检验方法，可以用于比较两组数据之间有没有差异。因此本文对两组不同影像设备中分别提取的影像组学特征进行曼惠特尼 U 检验，验证来自不同设备的同一特征是否服从相同分布，并将不同分布的高置信度特征移除，保留其余特征进行下一步的特征选择。

第二步，移除低方差特征。利用方差过滤进行特征选择基本理念是：通常来说低方差的特征由于所有样本的数值较为接近，因此该特征包含的信息量较少从而预测能力相对较差，因此方差过滤的方法会计算每个特征的方差，若特征的方差低于某个设定的阈值，则会被移除，以此达到特征选择的目的。经曼惠特尼 U 检验之后保留的特征数仍大于样本数，不适用于直接进行 TNM 分期预测，因此本文利用方差过滤进行进一步的特

征选择，将方差接近于 0 的特征剔除。

第三步，选择与 TNM 分期相关的特征。LASSO 特征选择方法能够在高维稀疏性数据中选出与研究目的相关的有效特征，尤其适合于存在多重共线性的模型，而影像组学特征之间存在多重共线性。因此，本文在方差过滤的基础上，利用 LASSO 再次进行特征选择，从保留的特征中选出与非小细胞肺癌 TNM 分期相关的影像组学特征，用于肺癌的 TNM 分期预测。

其次，基于选择的影像组学特征，利用机器学习的方法进行 TNM 分期预测。关于预测方法本文选择 XGBoost 方法。XGBoost 作为一种可扩展的提升树模型，擅长于从稀疏数据中获取知识。利用缓存访问、数据压缩和核外 Block 计算，XGBoost 可以凭借最少的资源解决现实生活中的大规模数据问题，具有高效、灵活、轻便的特点。XGBoost 在结构化或表格数据中有非常好的性能，被广泛应用于垃圾邮件分类、高能物理事件分类、恶意软件识别等多个领域。由于其良好的泛化性能和低成本的算力要求，因此我们选择用该方法进行 TNM 分期预测。

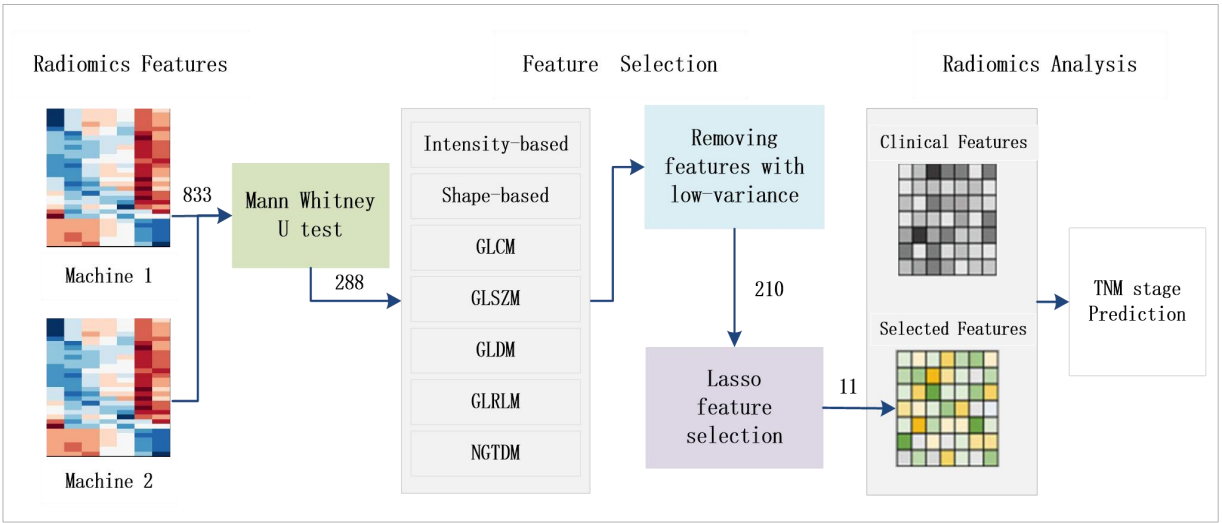



**图 4.1 非小细胞肺癌 TNM 分期预测研究流程图。**

具体的特征选择过程如图 4.1 所示，由第三章第二节实验数据介绍部分内容可知，324 例样本来源于两种不同的影像设备，分别从两组设备中提取出 833 个影像组学特征。对这 833 个影像组学特征进行选择：影像组学特征具有不稳定性，首先，用曼惠特尼 U 检验剔除在两种 PET-CT 设备中不服从同一分布的特征。在 TNM 分期预测数据集中，83 例样本使用一种 PET-CT 设备，其余 241 例使用另一种 PET-CT 设备，经过曼惠特尼

U 检验本文共剔除 545 个在两种设备中不服从同一分布的特征，保留其余 288 个影像组学特征；其次，在曼惠特尼 U 检验之后，利用方差过滤的方法进行进一步特征选择。本文将曼惠特尼 U 检验后保留的 288 个特征按照不同的特征类型进行分组，分别是基于强度、形状、GLCM、GLDM、GLSZM、GLRLM 和 NGTDM 的七种不同类型的特征，然后对不同类型特征分别计算方差并进行排序。方差过滤后 TNM 分期数据集共保留 210 个特征。接着，将 210 个特征进一步利用 LASSO 进行特征选择，选出与非小细胞肺癌 TNM 分期相关的 11 个影像组学特征，这 11 个特征分别包括形状特征、GLCM、GLSZM 和 GLRLM 四种类型，具体的特征名称及各类型中保留的特征数目统计如表 4.1 所示。最后将这 11 个特征与 XGBoost 分类器结合进行 TNM 分期的预测。

**表 4.1 用于 TNM 分期预测的影像组学特征名称**

| 类型 | 名称 | 数目 |
|---|---|---|
| 形状 | MajorAxisLength、Maximum2DDiameterSlice、Maximum3DDiameter、LeastAxisLength | 4 |
| GLCM | Wavelet LHH Idn | 1 |
| GLSZM | Wavelet (LHH SmallAreaLowGrayLevelEmphasis、HLL ZoneEntropy、HLH SmallAreaEmphasis、HLH SizeZoneNonUniformityNormalized) | 4 |
| GLRLM | Wavelet (HLH RunVariance、HHL ShortRunLowGrayLevelEmphasis) | 2 |

## 4.2 TNM 分期预测结果

按照国际癌症协会第八版肺癌 TNM 分期系统[23]的说明，本次实验数据集中的肿瘤分期主要包括肺癌 TNM 分期系统中的三种基本分期类别，分别是 I 期、II 期和 III 期。本文基于影像组学特征利用 XGBoost 分类器对这三种类别进行预测，其中，I 期样本、II 期样本和 III 期样本分别为 125 例、58 例和 141 例。在实验过程中，采用 10 折交叉验证的方法，将数据随机分成 10 份，其中 9 份作为训练集，另外 1 份作为测试集，具体得到的预测结果如图 4.2 和表 4.2 所示：

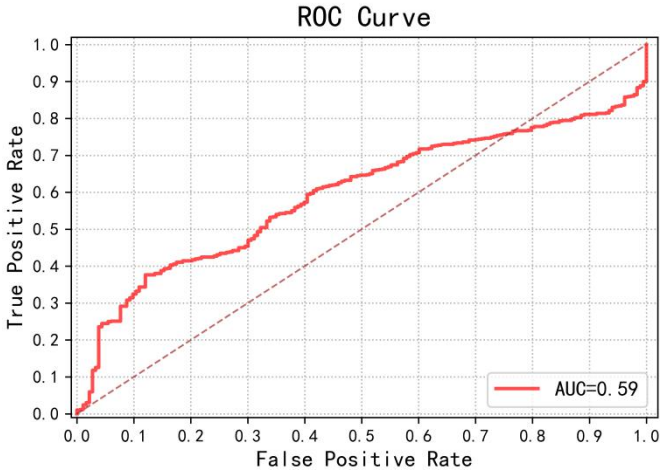



**图 4.2 TNM 分期预测 ROC 曲线图。**

**表 4.2 TNM 分期预测具体结果**

| 类别 | 精准率 (Precision) | 召回率 (Recall) | 精确度 (Accuracy) | 样本数 |
|---|---|---|---|---|
| I 期 | 0.7 | 0.95 | | 125 |
| II 期 | 0.58 | 0.38 | 0.71 | 58 |
| III 期 | 0.77 | 0.63 | | 141 |
| 合计 | | | | 324 |

由图 4.2 和表 4.2 可知，在非小细胞肺癌数据集 TNM 分期预测的三分类问题中，预测结果总的精确度和 AUC 值分别为 0.71 和 0.59。从方法对 TNM 分期为 I 期、II 期和 III 期的样本预测的具体结果来看，I 期样本的预测结果最好，召回率为 0.95，意味着在所有 I 期的 125 例样本中，有 95%的样本都被预测正确，对 III 期样本的预测结果次之，召回率为 0.63，II 期样本的预测性能相对较低，召回率为 0.38。由于 II 期样本相较于 I 期和 III 期样本来说，样本数较少，不足 I 期和 III 期样本的一半，因此方法对 II 期样本学习不足，导致预测结果较差。

单独分析方法对 III 期样本的预测结果。本次实验中虽然 III 期样本数要略多于 I 期样本数，但是在对 III 期样本的预测中，方法的预测性能同样要低于 I 期的预测结果。除了实验中参数设置的影响，通过第八版 TNM 分期系统，笔者认为可能存在的原因还包括以下内容：肺癌 I 期的具体分期只有 IA1、IA2 和 IA3 三种，三种具体分期均没有淋巴转移，且肿瘤大小≤3cm，肿瘤之间的特征较为贴近。而在 III 期的样本中，包含了

IIIA、IIIB 和 IIIC 三种不同的分期，而在 IIIA 期又包括 T4N0M0、T4N1M0 和 T4N2M0 三种情形，IIIB 期中包含 T4N2M0 和 T4N3M0 两种情形，IIIC 期也有 T3N3M0 和 T4N3M0 两种，因此 III 期的所有具体分期在原发肿瘤的大小和淋巴转移的范围等方面其实都存在差异，也就是说 III 期的 141 例样本中的特征之间存在差异，导致方法的预测性能略低，具体的结果还需要进一步的验证。整体来说，基于影像组学特征的分类器在预测非小细胞肺癌 TNM 分期中有良好的性能，影像组学特征反映了不同分期的肿瘤之间存在的异质性。在以后的 TNM 分期预测研究中还可以利用的更适合的特征选择方法和分类器，使得模型在预测不平衡 TNM 分期数据集时能有更加均衡稳定的预测性能，这在实际的应用中更具有意义和价值。

## 4.3 本章小结

在非小细胞肺癌 TNM 分期预测一章中，我们详细描述了 TNM 分期预测的研究流程和预测结果。首先是 TNM 分期预测研究中影像组学特征选择的三个具体步骤，包括曼惠特尼 U 检验、方差过滤和 LASSO 特征选择，并且将最终选择的用于 TNM 分期预测的特征名称以表格方式呈现。在本章的最后部分，详细展示了所选的 11 个影像组学特征在 TNM 分期预测中的性能，并对预测结果进行简要的分析。

# 5 非小细胞肺癌淋巴转移预测

在非小细胞肺癌的淋巴转移预测数据集中，共包括 121 例非小细胞肺癌患者，影像组学特征为 833 个，特征数远大于样本数，因此仍然需要对影像组学特征进行选择，用组合的特征选择策略对特征进行挑选，用机器学习的方法对非小细胞肺癌患者是否出现淋巴转移进行预测。

## 5.1 研究流程与特征选择

淋巴转移预测研究分为两个部分：影像组学特征的选择和淋巴转移的预测，具体的研究流程如图 5.1 所示。首先是影像组学特征的选择。在淋巴转移预测中，本文仍使用与 TNM 分期预测相同的方法，利用组合的特征选择策略实现特征选择。淋巴转移预测数据集中 121 位非小细胞肺癌患者的 PET-CT 图像从两种不同来源的影像设备获得，不同的影像设备在尺寸和参数之间的差异使得两组样本的一些影像组学特征不服从同一分布，因此采用曼惠特尼 U 检验剔除不服从同一分布的特征，然后用方差过滤的方法将低方差的特征去除，接着使用 LASSO 选出与非小细胞肺癌淋巴转移相关的有效特征。其次，将确定的影像组学特征与临床特征结合，进行非小细胞肺癌淋巴转移预测。关于预测方法的选择，由于 XGBoost 非常适合于结构化和表格数据，且成本低、灵活高效，并具有良好的泛化性能，因此在预测淋巴转移时仍采用 XGBoost 方法进行预测。

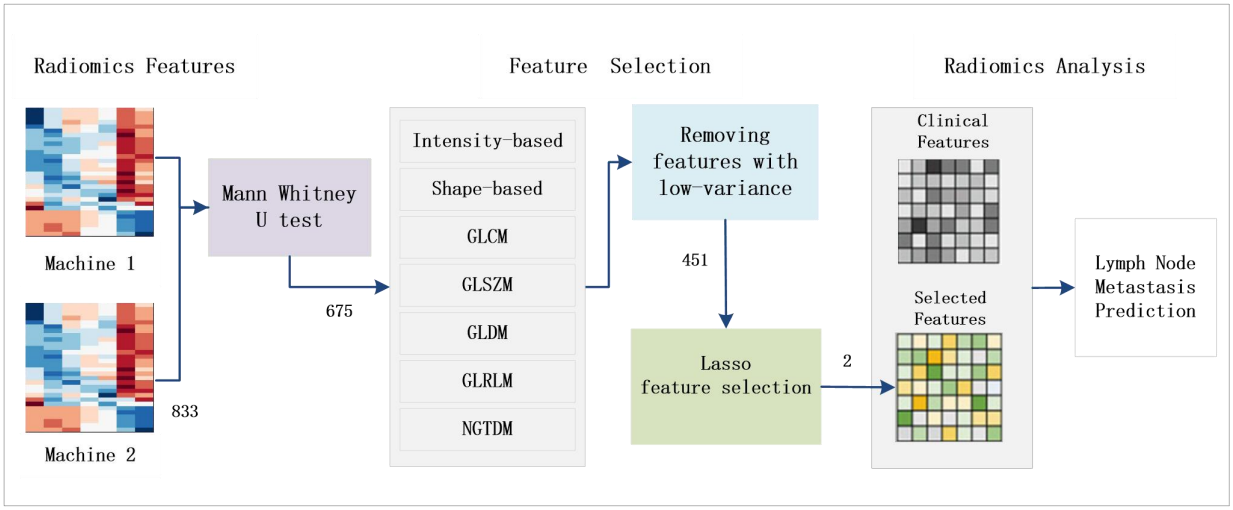



**图 5.1　非小细胞肺癌淋巴转移预测研究流程图。**

在特征选择过程中，对于提取的 833 个影像组学特征，经曼惠特尼 U 检验剔除不服从同一分布的特征，共保留 675 个影像组学特征；利用方差过滤的方法，对不同的特征类型分组进行特征选择，分别计算方差并进行排序。在这里我们分为两个步骤，一是在不同类型的特征中各自选择最优的前若干个特征，之后不再进行进一步选择，通过单类型特征进行淋巴转移的预测，以验证不同类型特征在预测淋巴转移中的表现；二是根据设定的阈值，保留每个类型中方差大于阈值的特征，在这一步中保留 451 个特征，用 LASSO 回归分析将保留的特征进行进一步的特征选择，最终选出 2 个与非小细胞肺癌淋巴转移相关的特征，将其与临床特征结合，并进行淋巴转移的预测。

## 5.2 类别不平衡问题

除了随机临床试验，医学领域中很多研究问题都存在类别不平衡的问题，特别是在使用常规临床数据进行的回顾性研究中，符合某一研究目标的样本数总是远少于不符合该研究目标的样本数。例如在弥漫性大 B 细胞淋巴瘤患者中，约有 16%的患者被发现有骨髓受累，因此在评估 18F-FDG PET 影像组学特征检测骨髓侵犯方面的表现时，必须要考虑出现骨髓受累的患者（16%）和没有出现骨髓受累的患者（84%）数据之间的不平衡[34]。在外周直径小于 3cm 的非小细胞肺癌患者的 TNM 分期临床诊断中，仅有 5.6% ~ 20%的患者病理诊断为存在淋巴转移，其余超过 80%的患者临床诊断为未出现淋巴转移[29]。而在本次实验数据中，出现淋巴转移的患者样本数仅占所有实验数据的 15.7%，具体表现在没有出现淋巴转移的患者样本为 102 例，出现淋巴转移的样本为 19 例，两类样本之间存在非常严重的数据不平衡，因此在预测非小细胞肺癌患者是否出现淋巴转移时，为了保证分类器能够充分学习到两种类别之间的差异，也必须要考虑数据不平衡的问题。

在分类问题中如果不考虑数据不平衡的影响，那么分类器很难通过样本数较少的少数类样本和相对样本数较多的多数类样本准确地获取不同类型样本之间的划分边界，从而导致分类不准确。对于数据不平衡问题，存在不同的解决方法，最简单的方式是对少数类样本进行随机重复过采样（Oversample）直至不同类型样本数达到平衡，但是这种方法不能为模型提供额外的有用信息。还有一种方法是根据现有的少数类样本合成新的样本，通过对少数类样本的数据增强达到数据平衡的目的，这种方法即合成少数类过采样技（Synthetic Minority Oversampling Technique, SMOTE）。除此之外还有一些其他的数据不平衡解决办法，如 ADASYN（自适应合成抽样），ADASYN[76]是通过对不同的

少数类样本分别赋予不同的权重，进而合成新样本的方法。本文采用 SMOTE 方法解决实验中的数据不平衡问题。

SMOTE 由 Nitesh Chawla 等人[77]提出，其主要原理是选择特征空间中贴近的样本，并在特征空间的样本之间画一条线，然后在线上某一点绘制一个新的样本。具体做法如下：对于某给定的少数类样本集合 $D$，首先从中随机选取某一样本 $x_i$，接着找到距离该样本最近的 k（通常 k 取值为 5）个近邻样本，并在特征空间中随机选择一个近邻样本 $x_{zi}$，之后在样本点和近邻之间随机选取一点生成新样本 $x_{new}$，新样本 $x_{new}$ 为两个选定样本的凸组合。

$$x_{new} = x_i + \lambda(x_{zi} - x_i) \tag{5.1}$$

其中 λ 为[0,1]之间的随机数。

## 5.3 淋巴转移预测结果

对于非小细胞肺癌患者的淋巴转移预测主要是分为两个部分：首先是将方差过滤后的不同类型的特征分别选出前若干个特征，然后利用这些特征分别进行预测，以此来观察不同类型的影像组学特征在预测肺癌淋巴转移中的表现；其次是利用经曼惠特尼 U 检验、方差过滤和 LASSO 回归分析后的保留特征进行预测，分析最后选定的特征在肺癌淋巴转移预测中的性能。本文还加入了临床特征进行预测，与仅用影像组学特征的预测结果进行对比。在实验过程中，采用 5 折交叉验证的方法，将样本数据随机分成 5 份，将其中 4 折作为训练集进行训练，另外 1 折作为测试集。在训练集中，本文利用 SMOTE 方法解决数据不平衡的问题。具体的预测结果将分为两个部分分别进行阐述：

### 5.3.1 不同类型影像组学特征分别进行淋巴转移预测

为了验证不同类型特征在淋巴转移预测中的表现，本文利用不同类型特征分别进行淋巴转移的预测，主要包括两部分内容，一是仅用不同类型影像组学特征进行淋巴转移预测，二是将影像组学与临床特征结合，再进行淋巴转移预测。

（1）不加入临床特征的预测

本文在经过曼惠特尼 U 检验和方差过滤后保留的 7 种不同类型特征（强度、形状、GLCM、GLDM、GLSZM、GLRLM、NGTDM）中，选择前若干个特征分别用于淋巴

转移的预测。由于不同类型的特征中每个具体的特征所包含的信息量各不相同，因此不同类型方差过滤后保留的特征个数也不尽相同，以保证不同类型特征有最优的淋巴转移预测结果，在这 7 种类型的特征中，本文保留的特征个数及名称如表 5.1 所示：

**表 5.1 不同类型选择的特征**

| 类型 | 名称 | 数目 |
|---|---|---|
| 强度 | firstorder_Variance、wavelet (LLL Variance、LHL Variance、LHH Median) | 4 |
| 形状 | Flatness、Elongation | 2 |
| GLCM | ClusterProminence、wavelet (HHH ClusterProminence、LHL ClusterProminence、LLL ClusterProminence、LHH ClusterProminence) | 5 |
| GLDM | LargeDependenceHighGrayLevelEmphasis、HighGrayLevelEmphasis、wavelet LHL (GrayLevelVariance、HighGrayLevelEmphasis、SmallDependenceHighGrayLevelEmphasis) | 5 |
| GLSZM | waveletLHL (HighGrayLevelZoneEmphasis、SmallAreaHighGrayLevelEmphasis、GrayLevelVariance) | 3 |
| GLRLM | ShortRunHighGrayLevelEmphasis、LongRunHighGrayLevelEmphasis、wavelet LHL (ShortRunHighGrayLevelEmphasis、HighGrayLevelRunEmphasis、 GrayLevelVariance、LongRunHighGrayLevelEmphasis) | 6 |
| NGTDM | Wavelet (LLL Contrast、LHL Contrast) | 2 |

根据表 5.1 中的选择的具体特征，分别进行不同类型影像组学特征的非小细胞肺癌淋巴转移预测。各类型特征具体的淋巴转移预测结果如图 5.2 所示，图例中展示了不同类型特征预测淋巴转移的 AUC 值。

由图 5.2 可知，方差过滤后的不同类型特征在分别预测非小细胞肺癌淋巴转移时，基于形状的特征在预测时有最好的预测结果，AUC 的值为 0.63，其次是基于 GLCM 的特征和基于 GLSZM 的特征，接下来是基于 NGTDM 的特征、基于 GLDM 的特征和基于强度（Intensity）的特征，最后是基于 GLSZM 的特征。整体来说不同类型特征在肺癌淋巴转移方面均表现出非常明显的价值。

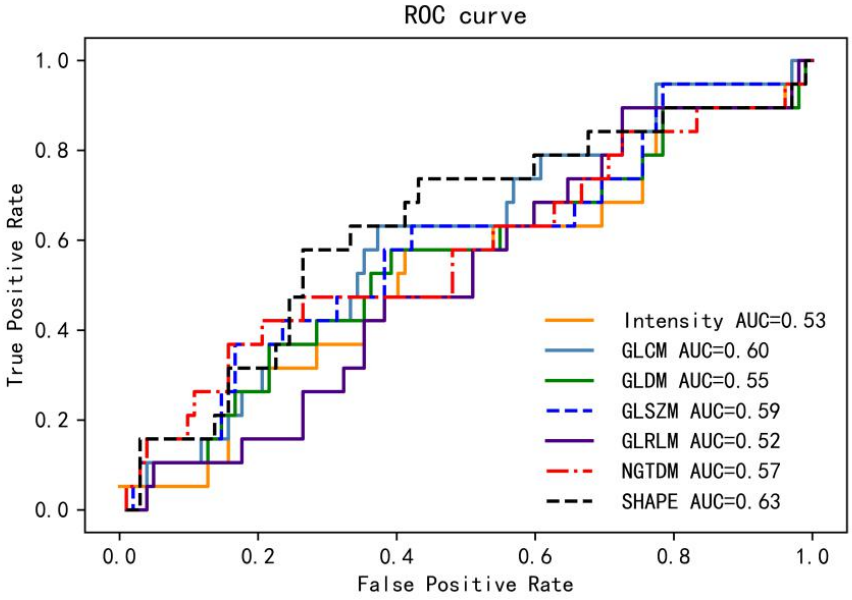



**图 5.2　不同类型特征预测 ROC 曲线。黄色实线为强度（Intensity）特征 ROC；浅蓝色实线为 GLCM 特征 ROC；绿色实线为 GLDM 特征 ROC；蓝色虚线为 GLSZM 特征 ROC；紫色为 GLRLM 特征 ROC；红色点横线为 NGTDM 特征 ROC；黑色虚线为形状（SHAPE）特征 ROC 曲线；**

（2）加入临床特征的预测

将临床特征与各类型影像组学特征结合进行预测，使用的临床特征包括吸烟史、年龄和 Laterality Desc（描述原发肿瘤位于左肺还是右肺），具体的预测结果如图 5.3 所示：

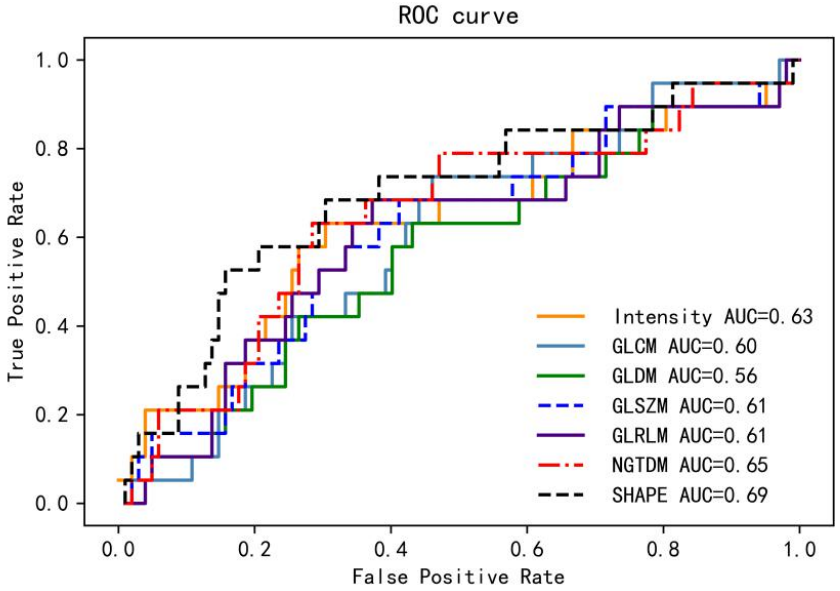



**图 5.3　加入临床特征后不同类型特征淋巴转移预测 ROC 曲线。黄色实线为强度（Intensity）特征 ROC 曲线；浅蓝色实线为 GLCM 特征 ROC 曲线；绿色实线为 GLDM 特征 ROC 曲线；蓝色虚线为 GLSZM 特征 ROC 曲线；紫色为 GLRLM 特征 ROC 曲线；红色点横线为 NGTDM 特征 ROC 曲线；黑色虚线为形状（SHAPE）特征 ROC 曲线；**

由图 5.3 和 5.3 可以看出，相较于仅用影像组学特征进行判断，在加入临床特征后，不同类型特征分别预测非小细胞肺癌淋巴转移的结果发生明显变化。整体来说基本所有类型特征在加入临床特征后预测非小细胞肺癌淋巴转移时的性能都有所提升。在加入临床特征后，基于强度的特征预测性能提升最大，AUC 值提升至 0.63，增加了 0.1，其次是基于 GLRLM 和 NGTDM 的特征，最后是基于形状、GLDM 和 GLCM 的特征，但是总的来说形状特征在预测淋巴转移时具有最好的预测性能。这个实验结果说明影像组学特征与肺癌的淋巴转移相关，影像组学特征与临床特征结合，在一定程度上能够更好地预测非小细胞肺癌患者是否出现淋巴转移。

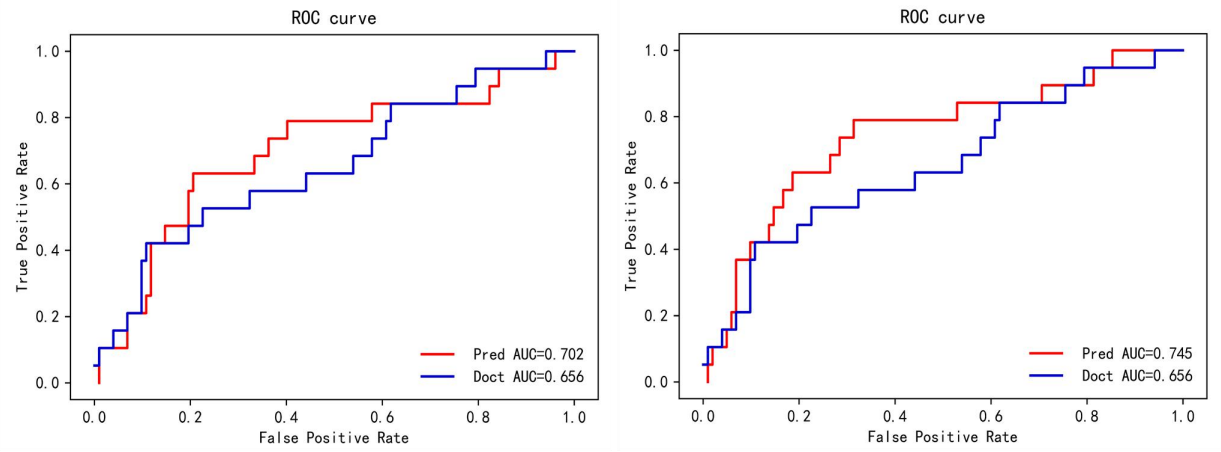
### 5.3.2 Lasso-XGBoost 淋巴转移预测

将经曼惠特尼 U 检验、方差过滤和 LASSO 特征选择后的影像组学特征用于淋巴转移预测，为便于进行描述，将方法简称为 LASSO - XGBoost 方法。最终 LASSO 回归分析选出两个与肺癌淋巴转移相关的特征用于预测，两个特征均为 GLSZM 类型特征，具体特征名称如表 5.2 所示：

**表 5.2　LASSO - XGBoost 淋巴转移预测使用特征名称**

| 类型 | 特征名称 |
|---|---|
| GLSZM | wavelet HLH SmallAreaLowGrayLevelEmphasis，<br>wavelet HHL SmallAreaEmphasis |

预测过程分为两个部分，（1）仅用影像组学特征进行预测；（2）加入临床特征（吸烟史、年龄和 Laterality Desc），再次对非小细胞肺癌淋巴转移进行预测。与此同时，我们还将医生对于非小细胞肺癌数据集中患者是否出现淋巴转移的判断与方法的预测结果进行了对比，最终的结果如图 5.4 和表 5.3 至表 5.6 所示：




1）不加临床特征　　2）加入临床特征

**图 5.4　Lasso 特征选择后的淋巴转移预测 ROC 曲线。1)为不加入临床特征的预测；2)为加入临床特征后的预测。红色为模型预测 ROC 曲线；蓝色为医生判断结果 ROC 曲线。**

**表 5.3　不加临床特征的淋巴转移预测结果**

| | 精准率 (precision) | 召回率 (Recall) | 精确度 (accuracy) | 样本数 |
|---|---|---|---|---|
| 没有转移 | 0.91 | 0.74 | 0.72 | 102 |
| 有转移 | 0.31 | 0.63 | | 19 |
| 合计 | | | | 121 |

**表 5.4　加入临床特征的淋巴转移预测结果**

| | 精准率 (precision) | 召回率 (Recall) | 精确度 (accuracy) | 样本数 |
|---|---|---|---|---|
| 没有转移 | 0.92 | 0.8 | 0.78 | 102 |
| 有转移 | 0.38 | 0.63 | | 19 |
| 合计 | | | | 121 |

**表 5.5　医生判断淋巴转移结果**

| | 精准率 (precision) | 召回率 (Recall) | 精确度 (accuracy) | 样本数 |
|---|---|---|---|---|
| 没有转移 | 0.89 | 0.87 | 0.8 | 102 |
| 有转移 | 0.38 | 0.42 | | 19 |
| 合计 | | | | 121 |

**表 5.6 混淆矩阵**

| 不加临床特征 | 加入临床特征 | 医生判断 |
|---|---|---|
| 75 27 | 82 20 | 89 13 |
| 7 12 | 7 12 | 11 8 |

由表 5.2 至 5.5 可以看出，无论是否加入临床特征，LASSO - XGBoost 方法的预测精确度均略要低于医生的判断，虽然精确度相较于医生判断略低，但是具体看方法和医生对转移和没有转移两类的具体预测结果，可以看出医生在预测没有转移的患者时分类精确度要高于方法的预测性能，相反在出现淋巴转移的患者中，无论是否加入临床特征，方法的预测性能均要高于医生的判断。不加临床特征的方法预测、加入临床特征的方法预测和医生判断的召回率分别为 0.63、0.63 和 0.42，即在 19 例 LNM+的肺癌患者中，本文方法正确预测 12 例，医生正确预测 8 例，这就意味着本文的方法比医生能更准确地对出现淋巴转移的患者进行分类。如果能够及时对存在转移风险的患者进行预警，在临床中就可以提前制定相应对策，有助于提升患者的生存期和预后，因此这一结果具有十分重要的临床价值。

为验证方法的预测性能，我们还将方法的预测结果与另外两种不同的方法进行了比较，一是特征组合的 XGBoost 预测方法，简称其为特征组合方法。这一方法主要是在特征选择方面与上述方法有所不同，具体是将方差过滤后保留的不同类型的特征进行随机组合，用组合的特征进行淋巴转移的预测。经多次实验证明，当基于 GLCM 的两个特征（Wavelet LHL gclm-ClusterProminence 和 Wavelet LHH gclm-ClusterProminence）和基于形状的两个特征（Elongation 和 Flatness）组合时，该方法的预测性能最好。二是与集成模型进行比较。该模型利用与本实验同一数据来源的非小细胞肺癌患者的 PET-CT 图像提取的影像组学特征，对非小细胞肺癌的组织学亚型（鳞状细胞癌和腺癌）进行分类。其方法主要思想是利用集成的方法，将影像组学特征充分利用。通过将所有影像组学特征随机分为 15 组，每一组影像组学特征都结合临床特征，并分别用若干不同参数设置的不同的分类器（支持向量机、Logistic 回归和多层感知机）进行分类，最后对所有分类结果的投票打分，进而得到最终的分类结果。由于本次实验的数据存在不平衡，因此在实现过程中同样加入 SMOTE 处理数据不平衡问题。上述两种方法与 LASSO-XGBoost 预测方法的具体预测结果对比如表 5.7 所示，对比结果均为加入临床特征后的预测结果：

**表 5.7 不同模型淋巴转移预测结果对比**

| | 精确度 (accuracy) | AUC |
|---|---|---|
| 集成模型 | 0.84 | 0.52 |
| 特征组合模型 | 0.81 | 0.73 |
| Lasso-XGBoost 模型 | 0.78 | 0.75 |

通过对比结果可以看到，虽然集成模型预测淋巴转移相较于其他两种方法有较高的精确度，但是 AUC 值要远低于另外两种模型，主要是因为在 LNM+和 LNM-两种类型样本的分类精确度中，集成模型对出现转移（LNM+）的患者预测较差，要么是完全预测不对此类进行预测，要么是随机猜测。个人认为出现这种情况主要有两方面的原因，一是集成模型中分类器的选择对预测性能的影响，二是所有特征的应用，即便是随机进行分组，也不能避免影像组学特征之间的相关性，导致模型预测性能受到影响。同样地，特征组合方法相较于 Lasso-XGBoost 方法虽然也有较高的预测精确度，但是 AUC 值还是相对略低，在预测 LNM+的患者淋巴转移方面略低于后者，总的来说，Lasso-XGBoost 方法在预测非小细胞肺癌淋巴转移方面具有较好的预测性能。

## 5.4 本章小结

本章对非小细胞肺癌淋巴转移预测过程进行了具体的讲述。首先是研究流程与类别不平衡问题，并对 SMOTE 数据不平衡策略简要介绍。然后对预测过程结果进行简要的分析。一是基于单个类型特征非小细胞肺癌淋巴转移预测，二是经曼惠特尼 U 检验、方差过滤和 LASSO 特征选择后的 LASSO-XGBoost 方法的淋巴转移预测；期间对加入临床特征的预测结果、不加临床特征的预测结果以及医生的判断进行对比分析，证明基于影像组学特征的预测方法的预测性能要高于医生的判断。最后，将 LASSO-XGBoost 方法的预测结果与集成模型和特征组合方法的预测结果进行了对比，从对比结果可知，Lasso-XGBoost 方法对非小细胞肺癌淋巴转移有更好的预测结果。

# 6 总结与展望

## 6.1 总结

影像组学特征通过定量的方式量化描述了肿瘤的关键性病理信息，包括瘤间异质性。充分挖掘和应用影像组学特征，可以以有别于传统医学研究的方式解决医学中存在的一些难题，同时还避免了传统医学研究的局限。医学图像分析领域已经有非常多基于影像组学特征的研究，包括肿瘤的组织学分类、患者的生存期预测和评估患者的预后等。本文根据非小细胞肺癌患者的 PET-CT 图像提取的影像组学特征，利用机器学习的方法对非小细胞肺癌患者的 TNM 分期和淋巴转移进行了预测。

首先是对影像组学特征的提取和选择。通过 Pyradiomics，共提取 833 个特征，影像组学特征具有高维性，且特征之间存在多重共线性，需要进行特征的选择。通过曼惠特尼 U 检验，剔除在两种设备中不服从同一分布的特征，再根据方差过滤的方法，过滤掉低方差的特征，之后再利用 LASSO 特征选择方法，选出分别与非小细胞肺癌 TNM 分期和淋巴转移相关的影像组学特征。其次是预测方法的选择，我们选择 XGBoost 进行预测，XGBoost 不仅泛化性能好，而且在具体运算过程中计算速度快，稳定性好，在各类竞赛中都有广泛的应用。

在 TNM 分期的三分类问题中，由于 I 期、II 期和 III 期患者的样本之间存在数据不平衡的问题，因此，在三分类问题中 XGBoost 分类器的预测性能受到限制，实验效果不够突出。但是，数据不均衡是现实环境尤其是医学问题中普遍存在的现象，如果能够在一定程度数据不均衡情况下使得多分类模型更加健全，在预测多分类问题时能够更加均衡稳定，那么将更加具有现实意义。

在对非小细胞肺癌的淋巴预测中，主要是预测没有淋巴转移和出现淋巴转移两种类型。通过单类型特征对淋巴转移的预测，我们发现 7 种不同类型的影像组学特征普遍在预测淋巴转移方面都显现出非常明显的价值，LASSO-XGBoost 淋巴转移预测方法的结果显示 PET 影像组学特征在预测非小细胞肺癌淋巴转移方面具有很好的性能，尤其对出现淋巴转移的患者，预测性能要高于医生的判断。除上述之外，我们还将加入临床特征和不加入临床特征的预测结果进行对比，实验证明影像组学特征结合临床特征，在预测非小细胞肺癌淋巴转移时具有更好的性能。

尽管本文在利用影像组学特征进行非小细胞肺癌 TNM 分期和淋巴转移预测方面取

得一些成果，但是也存在一些不足和改进之处，主要有以下几个方面：一，在医学影像利用程度方面，本文研究数据为非小细胞肺癌患者的 PET-CT 图像，但是在研究过程中仅基于 PET 图像提取的影像组学特征进行 TNM 分期和淋巴转移预测，CT 图像包含与 PET 图像互补的重要信息，也应该充分考虑加入 CT 图像进行研究；二，由于时间原因等多种因素的影响，本次研究中缺少对方法的临床效用评价，在之后的研究中还需要利用统计分析工具和医学领域中常用的评价方式如决策曲线分析对研究的临床效用进行分析和评价；三，在研究方法的选择方面，本文研究表明通过方差过滤的方法，虽然可以快速有效的筛选特征，但同时也会将一些与淋巴转移明显相关的特征也剔除掉，比如在 LASSO-XGBoost 方法中用到的对预测性能非常有效的两个特征在 GLSZM 类型特征预测淋巴转移的最佳特征中并没有出现。所以在影像组学研究中，要特别注意特征选择方法，结合研究目的对特征选择和预测方法谨慎选择。四，在医学研究中普遍存在数据不平衡问题，所以如果在数据不平衡的情况下实现较为准确地转移预测，更具有普适性。但是极度不平衡的数据对模型的训练有非常明显的影响，因此面对十分不平衡的数据，本文加入了数据不平衡策略，所以在数据允许的情况下之后的研究中还应该考虑利用验证集验证加入数据不平衡策略研究的有效性。

## 6.2 展望

医学影像学以非侵袭的方式获取肿瘤的关键性病理信息，包括瘤间异质性，有助于癌症的临床诊断。医学影像学和机器学习深度学习发展至今，已经在某些方面呈现深度的融合。在传统的机器学习与影像组学结合的研究中，应该要注意关于影像组学特征的特殊性，影像组学特征具有高维性，且特征之间存在很明显的相关，而且影像组学特征还具有不稳定性，不同的数据集得到的特征不同，同一数据集得到的特征排序也可能存在差异，因此在具体进行影像组学的研究时，要实事求是地根据研究问题制定研究方法。在特征提取过程中要特别注意参数的设置，根据实际的数据，设置合适的参数。特征处理过程同样非常重要，方差过滤虽然能够快速有效地选择特征，但是有极高概率会过滤掉一些与研究目的紧密相关的重要特征，除此之外，影像组学特征的取值范围变化很大，还要考虑应该将数据归一化。研究过程中应该关注到影像组学的特珠性，充分考虑影像组学特征之间的共线性问题。在具体的实验之后，还应该利用一些统计学的方法和医学常用的方法如决策分析曲线等对影像组学研究的临床效用进行分析和评价。

毫无疑问的是影像组学特征能够反映出有关肿瘤的一些关键病理信息，而且已经在

医工交叉的多方面研究中显现出非常重要的价值。除了机器学习的方法，现在也有很多利用深度学习方法的影像组学研究，但是深度学习在影像组学方面的研究依旧受到传统机器学习的影响，而且训练速度方面要逊于机器学习方法，随着医学影像学和机器学习深度学习的进一步发展，必然会有更多关于影像组学的研究。

# 参考文献

[1] 徐定杰,周远东,桑育黎,等.模糊神经网络控制的医疗诊断系统研究[J].中医药学刊,2005(5).

[2] DOROSHOW J H, KUMMAR S. Translational research in oncology--10 years of progress and future prospects[J]. Nat Rev Clin Oncol, 2014, 11(11):649-62.

[3] KAPOOR V, MCCook B M, TOROK F S. An introduction to PET-CT imaging[J]. Radiographics, 2004, 24(2):523-43.

[4] OIKONOMOU A, KHALVATI F, et al. Radiomics Analysis at PET/CT Contributes to Prognosis of Recurrence and Survival in Lung Cancer Treated with Stereotactic Body Radiotherapy[J]. Scientific Reports, 2018, 8(1).

[5] AERTS H J, VELAZQUEZ E R, et al. Decoding Tumour Phenotype by Noninvasive Imaging Using a Quantitative Radiomics Approach[J]. Nature Communications, 2014, 5.

[6] WILD C P, WEIDERPASS E, STEWART B W. World Cancer Report: Cancer research for cancer prevention[M]. Lyon: International Agency for Research on Cancer, 2020.

[7] 国家肿瘤质控中心. 2019 年中国癌症报告[EB/OL]. (2019-03-19)[2021-02-28].http://www.china-rt.cn/special/846.html.

[8] RISCH A, PLASS C. Lung cancer epigenetics and genetics[J]. Int J Cancer, 2008, 123:1-7.

[9] TRAVIS W D. Pathology of lung cancer[J]. Clin Chest Med, 2011, 32:669-92.

[10] REKHTMAN N, ANG D C, et al. Immunohisto-chemical algorithm for differentiation of lung adenocarcinoma and squamous cell carcinoma based on large series of whole-tissue sections with validation in small specimens[J]. Mod Pathol , 2011, 24:1348-59.

[11] LIU J, CUI J, LIU F, et al. Multi-subtype classification model for non-small cell lung cancer based on radiomics: SLS model[J]. Med Phys. 2019, 46(7):3091-3100.

[12] Predicting metastasis from primary tumor size[EB/OL].(2019-11-21)[2020-09-05]. https://medicalxpress.com/news/2019-11-metastasis-primary-tumor-size.html.

[13] ZHOU Z, WANG K, et al. Multifaceted radiomics for distant metastasis prediction in head & neck cancer[J].Physics in Medicine and Biology. 2020, 15.

[14] Sibylle l. Ziegler. Positron Emission Tomography: Principles, Technology, and Recent Developments[J]. Nuclear Physics. 2005:679-687.

[15] Principles and Practice of PET/CT [EB/OL].(2016-11-21)[2021-01-19].https://www.

eanm.org/content-eanm/uploads/2016/11/gl_Principles_and_Practice_of_PET-CT_Part_1.pdf.

[16] GOLDMAN LEE W. Principles of CT and CT Technology[J]. Society of Nuclear Medicine. 2007, 3: 115-128.

[17] GUPTA N C, FRANK A R, DEWAN N A, et al. Solitary pulmonary nodules: detection of malignancy with PET with 2-[F-18]-fluoro-2-deoxy-D-glucose[J]. Radiology. 1992, 184: 441– 444.

[18] STRAUSS L G, CLORIUS J H, SCHLAG P, et al. Recur- rence of colorectal tumors: PET evaluation[J]. Radiology. 1989, 170: 329 –332.

[19] VESSELLE H, SCHMIDT R A, PUGSLEY J M, et al. Lung cancer proliferation correlates with [F-18]fluoro- deoxyglucose uptake by positron emission tomography[J]. Clin Cancer Res. 2000, 6: 3837–3844.

[20] HOH C K, GLASPY J, ROSEN P, et al. Whole-body FDG-PET imaging for staging of Hodgkin's disease and lymphoma[J]. J Nucl Med. 1997, 38: 343– 348.

[21] ADLER L P, CROWE J P, AL-KAISI N K, et al. Evaluation of breast masses and axillary lymph nodes with [F-18]2-deoxy-2-fluoro-D-glucose PET[J]. Radiology, 1993, 187:743–750.

[22] MIRALDI F, VESSELLE H, FAULHABER P F, et al. Elimination of artifactual accumulation of FDG in PET imaging of colorectal cancer[J]. Clin Nucl Med, 1998, 23:3–7.

[23] RAMI-PORTA R. Lung Cancer Staging Changing the Clinical Practice[J]. WCLC. 2016.

[24] HUEMAN M, WANG H, HENSON D, CHEN D. Expanding the TNM for cancers of the colon and rectum using machine learning: A demonstration[J]. ESMO Open. 2019, 4.

[25] HALSTED W S. The results of radical operations for the cure of carcinoma of the breast[J]. Ann Surg. 1907, 46:1–19.

[26] JAFFE C C. Imaging and genomics: is there a synergy?[J]. Radiology. 2012, 264:329–331.

[27] LIU Z, WEN Z, LIU C, et al. Diagnostic accuracy of ultrasonographic features for lymph node metastasis in papillary thyroid microcarcinoma: a single-center retrospective study[J]. World J Surg Oncol. 2017, 15(1):32-36.

[28] CHEN D, SHE Y, WANG T, et al. Radiomics-based prediction for tumour spread through air spaces in stage I lung adenocarcinoma using machine learning[J]. Eur J Cardiothorac Surg. 2020, 58(1):51-58.

[29] GU Y, SHE Y, XIE D, et al. A texture analysis - based prediction model for lymph node

metastasis in stage IA lung adenocarcinoma[J]. Ann Thorac Surg. 2018, 106:214–20.

[30] ISHIDA T, YANO T, MAEDA K, KANEKO S, et al. Strategy for lymphadenectomy in lung cancer three centimeters or less in diameter[J]. Ann Thorac Surg. 1990, 50:708–13.

[31] TAKIZAWA T, TERASHIMA M, KOIKE T, et al. Lymph node metastasis in small peripheral adenocarcinoma of the lung[J]. Thorac Cardiovasc Surg. 1998, 116:276–80.

[32] LAMBIN P, VAN STIPHOUT R G, STARMANS M H, et al. Predicting outcomes in radiation oncology - multifactorial decision support systems[J]. Nat. Rew. Clin. Oncol. 2013, 10: 27-40.

[33] PARMAR C, GROSSMAN P, BUSSINK J, et al. Machine learning methods for quantitative radiomic biomarkers[J]. Sci Rep. 2015, 5:13087.

[34] MAYERHORFER ME, MATERKA A, LANGS G, et al. Introduction to Radiomics[J]. J Nucl Med. 2020, 61(4):488-495.

[35] Gillies R, KINAHAN P, et al. Radiomics: Images Are More than Pictures, They Are Data[J]. Radiology. 2016, 278(2):563-577.

[36] LAMBIN P, RIOS-VELAZQUEZ E, LEIJENAAR R, et al. Radiomics: extracting more information from medical images using advanced feature analysis[J]. Eur. J. of Cancer. 2012, 48: 441-446.

[37] AERTS H J, VELAZQUES E R, et al. Decoding Tumour Phenotype by Noninvasive Imaging Using a Quantitative Radiomics Approach[J]. Nature Communications. 2014, 5.

[38] PARMAR C, LEIJENAAR R T, GROSSMANN P, et al. Radiomic feature clusters and prognostic signatures specific for Lung and Head & Neck cancer[J]. Sci Rep. 2015, 5;5:11044.

[39] DAI W, MO S, HAN L, et al. Prognostic and predictive value of radiomics signatures in stage I-III colon cancer[J]. Clin. Transl. Med. 2020, 10:288-293.

[40] KNIEP H C, MADESTA F, SCHNEIDER T, et al. Radiomics of Brain MRI: Utility in Prediction of Metastatic Tumor Type[J]. Radiology. 2019, 290(2):479-487.

[41] COSMA G, ACAMPORA G, BROWN D, et al. Prediction of Pathological Stage in Patients with Prostate Cancer: A Neuro-Fuzzy Model[J]. PLoS ONE. 2016, 11(6): e0155856.

[42] XIAO G, RONG W C, HU Y C, et al. MRI Radiomics Analysis for Predicting the Pathologic Classification and TNM Staging of Thymic Epithelial Tumors: A Pilot Study[J]. AJR Am J

Roentgenol. 2020, 214(2):328-340.

[43] COROLLER T P, Grossmann P, et al. CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma[J]. Radio the Oncol. 2015.

[44] Gevaert O, MITCHELL L A, et al. Glioblastoma multiforme: exploratory radiogenomic analysis by using quantitative image features[J]. Radiology. 2014, 273: 168–174.

[45] GANESHAN B, GOH V, HENRY C, et al. Non–small cell lung cancer: histopathologic correlates for texture parameters at CT[J]. Radiology. 2013, 266:326–336.

[46] COOK G J, YIP C, SIDDIQUE M, et al. Are Pretreatment 18F-FDG PET Tumor Textural Features in Non–Small Cell Lung Cancer Associated with Response and Survival After Chemoradiotherapy?[J]. Nucl Med. 2013, 54:19–26.

[47] PARMAR C, RIOS VELAZQUEZ E, et al. Robust radiomics feature quantification using semiautomatic volumetric segmentation[J]. PLOS ONE. 2014, 9: e102107.

[48] ALIC L, NIESSEN W J, VEENLAND J F. Quantification of heterogeneity as a biomarker in tumor imaging: a systematic review[J]. PLOS ONE. 2014, 9: e110300.

[49] JAIN R, POISSON L M, GUTMAN D, et al. Outcome prediction in patients with glioblastoma by using imaging, clinical, and genomic biomarkers: focus on the nonenhancing component of the tumor[J]. Radiology. 2014, 272: 484–493.

[50] NICOLASJILWAN M, HU Y, YAN C, et al. Addition of MR imaging features and genetic biomarkers strengthens glioblastoma survival prediction in TCGA patients. Journal of Neuroradiology. 2015, 42(4):212-221.

[51] LE Q C, ARIMURA H, NINOMIYA K, et al. Radiomic features based on Hessian index for prediction of prognosis in head-and-neck cancer patients[J]. Sci Rep. 2020, 10: 21301.

[52] SAID A A, ABD-ELMEGID L A, KHOLEIF S, et al. Stage–Specific predictive models for main prognosis measures of breast cancer[J]. Future Computing and Informatics Journal. 2018, 391-397.

[53] LIU T, ZHOU S, YU J, et al. Prediction of Lymph Node Metastasis in Patients With Papillary Thyroid Carcinoma: A Radiomics Method Based on Preoperative Ultrasound Images[J]. Technol Cancer Res Treat. 2019, 1;18:1533033819831713.

[54] GAO X, MA T, CUI J, et al. A radiomics-based model for prediction of lymph node metastasis in gastric cancer[J]. Eur J Radiol. 2020, 129:109069.

[55] LIU J, SUN D, CHEN L, et al. Radiomics Analysis of Dynamic Contrast-Enhanced Magnetic Resonance Imaging for the Prediction of Sentinel Lymph Node Metastasis in Breast Cancer[J]. Front Oncol. 2019, 30;9:980.

[56] LIU Z, MENG X, ZHANG H, et al. Predicting distant metastasis and chemotherapy benefit in locally adveanced rectal cancer[J]. Nat Common. 2020, 11; 4308.

[57] CHEN LD, LIANG JY, Wu H,et al.Multiparametric radiomics improve prediction of lymph node metastasis of rectal cancer compared with conventional radiomics[J]. Life Sci. 2018, 1; 208:55-63.

[58] ZHOU L, WU X, HUANG S, et al. Lymph Node Metastasis Prediction from Primary Breast Cancer US Images Using Deep Learning[J]. Radiology. 2020, 294:19-28.

[59] GAO X, MA T, CUI J, ZHANG Y, et al. A CT-based Radiomics Model for Prediction of Lymph Node Metastasis in Early Stage Gastric Cancer[J]. Acad Radiol. 2020, 2:S1076-6332(20)30191-4.

[60] HASTIE T, TIBSHIRANI R, WAINWRIGHT M. Statistical Learning with Sparsity-The Lasso and Generalizations[M].CRC. 2016.

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

[62] CHEN T Q, GUESTRIN C. XGBoost: A Scalable Tree Boosting System[J]. In eprint arXiv:1612.07003 [cs.LG](2016).

[63] FAWCETT T. "An Introduction to ROC Analysis"[J]. Pattern Recognition Letters. 2006, 27 (8): 861–874.

[64] KUMAR V, GU Y, BASU S, et al. Radiomics: the process and the challenges[J].Magn Reso-n Imaging. 2012,30(9):1234-48.

[65] THAWANI R, MCLANE M, et al., Radiomics and Radiogenomics in Lung Cancer:A Review for the Clinician[J]. Lung cancer, 2017, 115:34-41.

[66] ZWANENBURG A, LEGER S, VALLIERES M, et al. Image biomarker standardisation initiative feature definitions[J]. In eprint arXiv:1612.07003 [cs.CV], 2016.

[67] HARALICK R, SHANMUGAN K, DINSTEIN I. Textural features for image classification[J]. IEEE Transactions on Systems, Man and Cybernetics. 1973, (3); p610-621.

[68] GALLOWAY M M. Texture classification using gray level run length[J]. Comput Graph Image Process. 1975, 4:172–179.

[69] AMADASUN M, KING R. Textural features corresponding to textural properties[J]. IEEE Trans Syst Man Cybern. 1989, 19:1264–1274.

[70] PAREKH V, JACOBS M. Radiomics: a new application from established techniques[J]. ExpertReview of Precision Medicine and Drug Development. 2016, 1(2): 207-226.

[71] GRIETHUYSEN JJ M, FEDOROV A, PARMAR C, et al. Computational Radiomics System to Decode the Radiographic Phenotype[J]. Cancer Research. 2017, 77(21): e104–e107.

[72] Á SALAZAR, H LORDUY G. Approach to wavelet multiresolution analysis using Coiflets and a two-wave mixing arrangement[J]. Optics Communications. 2008, 281:3091-3098.

[73] LEGER S, ZWANENBURG A, PILZ K, et al. A comparative study of machine learning methods for time-to-event survival data for radiomics risk modelling[J].Sci Rep. 2017.

[74] LIU Q, SUN D, LI N, et al. Predicting EGFR mutation subtypes in lung adenocarcinoma using 18F-FDG PET/CT radiomic features[J]. Transl Lung Cancer Res. 2020, 9(3):549-562.

[75] REN C, ZHANG J, QI M, et al. Machine learning based on clinico-biological features integrated 18F-FDG PET/CT radiomics for distinguishing squamous cell carcinoma from adenocarcinoma of lung[J]. Eur J Nucl Med Mol Imaging. 2020.

[76] HE H B, BAI Y, GARCIA E A, et al. ADASYN: Adaptive synthetic sampling approach for imbalanced learning[C]. 2008 IEEE International Joint Conference on Neural Networks. Hong Kong, China, 2008, pp. 1322-1328.

[77] NITESH V C, KEVIN W B, LAWRENCE O H, et al. SMOTE: Synthetic Minority Over-sampling Technique[J]. Journal of Artificial Intelligence Research. 2002, 321-357.

# 致 谢

时光如白驹过隙，转眼之间就匆匆过去。三年时光的研究生生涯，真是一点都经不起回想，似乎在一眨眼之间这一千多个日子就从脑海掠过了。但是认真思考在兰州财经大学的这三年，我的心中便有颇多感触，在这里我学习成长，受益良多。

首先要感谢我们实验室的四位老师。感谢我的导师韩金仓教授，韩老师在我的学习和生活中给予我很多的关注，不仅使我在学业上开拓了视野，增长见闻，在生活中还经常教我很多做人做事的道理和人际交往的礼仪。感谢韩老师在我研究和写作过程中给予我的支持。还要感谢何江萍老师，何老师为我打开了一扇真正走向研究的大门，在一个全新的领域我接触到了一个完全与我之前二十几年人生迥异的世界。他一丝不苟的科研精神和一心为学生着想的美好品德令我敬佩。郑重感谢以上两位老师在我的研究生生涯中给予的巨大帮助，使我在学习和工作中养成踏实严谨的良好习惯，也让我明白了好老师能够让学生在面临问题时有思考的能力和质疑的底气。同时，还要感谢李兵老师和丁晓阳老师，他们对研究和工作的责任与专注让我明白了热爱是一件多么幸福的事情。

其次，要感谢我们实验室的小伙伴，包括我的学长学姐、我的同班同学和学弟学妹，在学业上我们时常探讨问题，并且在我遇到困难时小伙伴们都会倾力帮助我，学业之余也给了我非常珍贵的友谊。还要感谢我宿舍的两位舍友，三年的时光我体会到了舍友之间纯粹真挚的美好情谊，希望我们的友谊天长地久。

郑重感谢我的父母和姐妹，对我生活以及精神的无私支持和无限帮助。

最后，要感谢莅临答辩现场的各位老师，能够在百忙之中对我的论文给予指导。

# 硕士期间科研项目与成果

**参与的科研项目：**

1 甘肃省自然科学基金：“PET/CT 图像中肿瘤组织类型自动判别方法研究”（项目批准号：20JR5RA200）