·医学工程·

基于机器学习对亚实性IA期肺腺癌气道内播散的预测价值

刘欢欢[1a] 陈业[1b] 周露[2] 齐宏亮[1b*]

（1.南方医科大学南方医院 a.放疗中心；b.医学工程科，广州 510515；2.湘潭市中心医院放射科，湘潭 411100）

【摘要】 目的 探讨IA期亚实性肺腺癌中气道内播散（Spread Through Air Spaces，STAS）的临床特征及CT影像表现，并基于机器学习算法构建预测模型。**方法** 本研究对2021年1月—2023年6月期间在湘潭市中心医院收治的IA期亚实性肺腺癌患者进行了回顾性分析。根据术后的病理结果，将患者分为STAS(+)组和STAS(–)组。使用包括Logistic回归、支持向量机（Support Vector Machine，SVM）、K近邻（K–Nearest Neighbour，KNN）、极端梯度提升（Extreme Gradient Boosting，Xgboost）、最小绝对收缩和选择算子（Least Absolute Shrinkage and Selection Operator，LASSO）、随机森林（Random Forest，RF）在内的五种机器学习算法来构建IA期亚实性肺腺癌的STAS预测模型。**结果** 在106例IA期亚实性肺腺癌患者中，有16例（15%）被诊断为STAS。在对五种机器学习算法进行比较后，Xgboost在诊断准确性方面表现最佳，曲线下面积：0.969（95% CI：0.915～0.993）。Xgboost算法确定了IA期亚实性肺腺癌中STAS的以下关键预测因子，包括实性成分比例更高、实性部分的CT值、病灶大小、周围磨玻璃影（Ground–Glass Opacity，GGO）边界的模糊程度以及实性成分体积。**结论** 通过分析患者的临床及影像学数据，Xgboost能够有效预测IA期亚实性肺腺癌的STAS，有利于临床决策。

【关键词】 IA期亚实性肺腺癌；气道内播散；机器学习；预测模型

【中图分类号】 TH77; R737.9 **【文献标识码】** A **【DOI】** 10.11876/mimt202501007

Applications of machine learning models in the prediction of tumor spread through air spaces in patients with stage IA part–solid lung adenocarcinoma

Liu Huanhuan[1a], *Chen Ye*[1b], *Zhou Lu*[2], *Qi Hongliang*[1b*]

(*1. a. Department of Radiotherapy*; *b. Department of Medical Engineering*, *Nanfang Hospital*, *Southern Medical University*, *Guangzhou 510515*, *China*; *2. Department of Radiology*, *Xiangtan Central Hospital*, *Xiangtan 411100*, *China*)

【Abstract】 Objective This study aims to explore the clinical features and CT imaging manifestations of spread through air spaces (STAS) in stage IA subsolid lung adenocarcinoma and to construct a predictive model based on machine learning algorithms. **Methods** We retrospectively analyzed patients with stage IA subsolid lung adenocarcinoma admitted to Xiangtan Central Hospital from January 2021 to June 2023. Patients were classified into STAS(+) and STAS(–) groups based on postoperative pathological results. We employed five machine learning algorithms—Logistic Regression, Support Vector Machine (SVM), K–Nearest Neighbour (KNN), Extreme Gradient Boosting (Xgboost), Least Absolute Shrinkage and Selection Operator (LASSO), and Random Forest (RF)—to develop a predictive model for STAS in stage IA subsolid lung adenocarcinoma. **Results** Among the 106 patients with stage IA subsolid lung adenocarcinoma, 16 (15%) were diagnosed with STAS. After comparing the five machine learning algorithms, Xgboost demonstrated the highest diagnostic accuracy, achieving an area under the curve (AUC) of 0.969 (95% CI: 0.915–0.993). The Xgboost algorithm identified several key predictors of STAS in stage IA subsolid lung adenocarcinoma: higher proportion of solid component, CT value of the solid component, lesion size, degree of blurring in the surrounding ground–glass opacity (GGO) border, and volume of the solid component. **Conclusion** Through the analysis of clinical and imaging data, the Xgboost algorithm can effectively predict STAS in stage IA subsolid lung adenocarcinoma, providing valuable insights for clinical decision–making.

【Key words】 Stage IA subsolid lung adenocarcinoma; Spread through air spaces; Machine learning; Predictive model

基金项目：南方医科大学南方医院院长基金项目（2022B030）

第一作者：刘欢欢，女，本科，主管护师，研究方向：放疗CT定位

通讯作者：齐宏亮，男，博士研究生，高级工程师，研究方向：影像组学

肺癌的发病率及与肿瘤相关的死亡率均位居各种恶性肿瘤之首，其预后受到多种临床因素的影响，包括肿瘤的分级、病理特征和治疗策略等。随着低剂量CT筛查模式的推广，肺癌的早期检测率显著提高。肺腺癌是最常见的病理类型，其中约有25%的病例为ⅠA期[1]。ⅠA期肺癌的主要治疗方法是手术切除，相较于根治性肺叶切除，亚肺叶切除（如肺段切除和楔形切除）能够更好地保留肺功能，并减少手术后并发症。尽管ⅠA期肺癌患者通常能够实现良好的预后，但术后局部复发或远处转移等不良结果不容忽视，特别是亚肺叶切除后，患者的复发率相对较高[2,3]。有关综述指出，气道内播散（Spread Through Air Spaces，STAS）被认为是早期肺癌不良预后的高危因素[4,5]，是肿瘤隐匿型转移及肺内播散的高危因素，并且与ⅠA期肺腺癌隐匿性淋巴结转移紧密关联，从而增加了亚肺叶切除术后复发的风险[6,7]。因此，术前有效预测STAS对评估早期肺癌患者的手术方式的选择、手术范围的确定具有重要参考价值。

已有研究发现，术前CT影像特征与肺腺癌出现STAS之间存在密切关系，这些特征包括病灶的最大直径、异常支气管的气体表现、实性成分的比例以及亚实性结节周围磨玻璃影的边界模糊等[8-10]。随着机器学习算法的发展，其逐步在肺部疾病的辅助诊断、病理评估、临床决策及预后分析等领域得到了日益广泛的应用[11]。因此，本研究致力于深度挖掘ⅠA期亚实性肺癌患者的术前临床影像特征，通过建立不同的机器学习算法来评估STAS风险，从而为临床决策提供全面而可靠的支持。

1 资料与方法

1.1 一般资料

本研究对2021年1月—2023年6月在湘潭市中心医院收治的ⅠA期亚实性肺腺癌患者进行了回顾性分析。年龄29～78岁，其中男性59例、女性47例。

纳入标准：①患者必须接受手术切除，涵盖肺叶切除和亚肺叶切除；②影像学表现为肺部亚实性结节，且CT扫描与手术时间间隔不超过15d；③术后病理结果确认为ⅠA期肺腺癌（ⅠASLC 9th）。

排除标准：①术前CT图像质量不合格，影响后续数据分析；②合并其他类型的恶性肿瘤；③有既往肺部手术史。本研究按ⅠASLC第8版TNM肺癌分期标准进行[12]。

1.2 CT扫描及图像征象获取

CT扫描采用Revolution 256 CT（GE Healthcare, Chicago, IL, USA）。扫描参数如下：管电压120kVp，管电流100～150mA，螺距1.375～1.500mm，扫描完成后以0.625～1.500mm进行薄层重建。

CT图像由两名具备10年以上肺部疾病诊断经验的高年资医师独立阅片，在未知病理结果的情况下，当影像征象存在争议时，经协商达成一致。获取CT征象包括，（1）病灶大小：在肺窗（窗位为–600HU，窗宽为1500HU）测量亚实性结节实性成份的最大径[12]。（2）实性部分体积、病灶总体积以及实性占比：采用3D slicer软件（版本号5.2.2），分别对亚实性结节的病灶轮廓、实性部分轮廓进行逐层勾画，见图1，最后计算实性占比$=\frac{\text{实性部分体积}}{\text{病灶总体积}}$。（3）实性部分CT值：在亚实性结节的实性成份上分别进行3次测量其CT值，并取其均值。（4）支气管、肺血管异常征阳性：有别于肺血管、支气管在肺门出发出逐渐变细的特点，在肺亚实性病变中穿行的血管失去渐行变细的特点，远端较近段粗，或血管出现纠集、扭曲；同理，亚实性病变内走形的支气管出现扩张、僵直等[13]。（5）周围磨玻璃边界：多维度观察亚实性结节实性成份周围磨玻璃影边界，定义为模糊或清晰[9]。最后，再纳入分析病灶位置、病灶分叶、胸膜牵拉等征象。

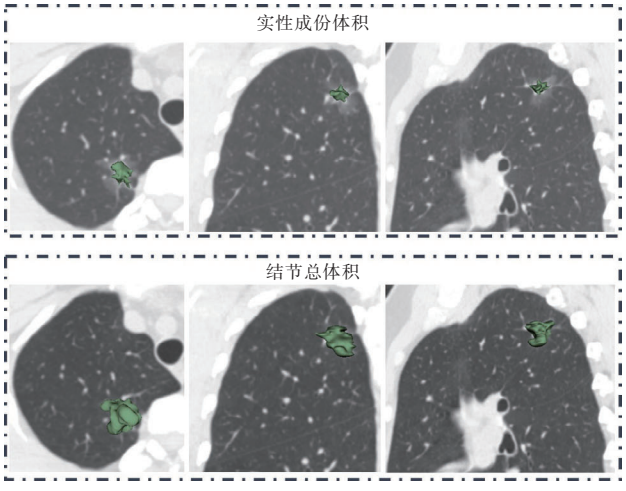


图1 使用3D slicer软件测量亚实性结节实性成份体积、结节总体积

1.3 临床病理资料的获取

临床资料包括：年龄、性别、吸烟、手术方式、病理类型、脉管侵犯、神经侵犯、胸膜侵犯，以及STAS。其中，STAS依据2015年WHO肺癌分类标准，定义为：微乳头状簇、实心巢或单个细胞散布在主要瘤体边界以外的气腔空间内[2,3]。研究设计见图2。

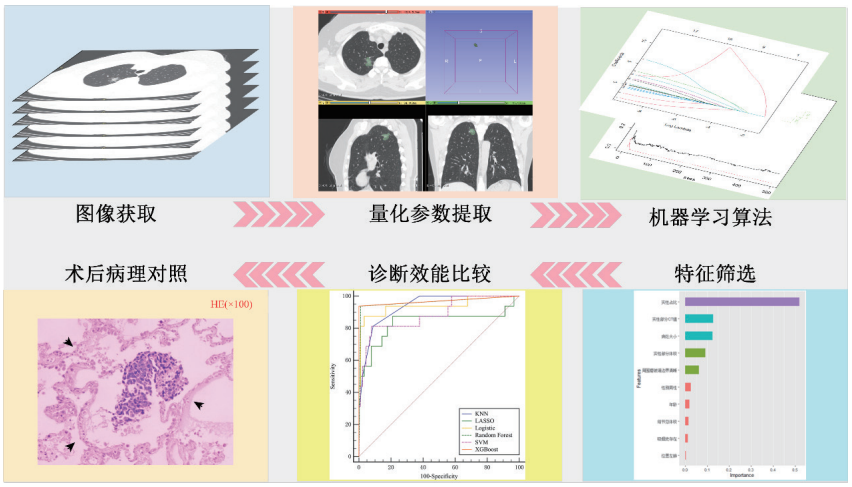


图2 设计不同机器学习算法模型对亚实性IA期肺腺癌STAS的预测评估

1.4 统计学方法

本研究使用R语言4.1.0（网址：https://www.r-project.org/）进行数据分析并构建机器学习模型。具体方法如下：（1）单因素分析及传统Logistic回归：通过“Tableone”程序包进行组间比较，利用“backward法”进行变量筛选，并建立Logistic回归模型。（2）机器学习算法：①支持向量机（Support Vector Machine，SVM），使用“e1071”软件包；②K近邻（K-Nearest Neighbour，KNN），采用“kknn”程序包；③极端梯度提升（Extreme Gradient Boosting，Xgboost），利用“xgboost”程序包；④最小绝对收缩和选择算子（Least Absolute Shrinkage and Selection Operator，LASSO），使用“glmnet”程序包；⑤随机森林（Random Forest，RF），采用“randomForest”程序包。最终，使用受试者工作特征曲线（Receiver Operating Characteristic，ROC）评估各个模型的诊断性能，选择最佳的机器学习预测模型，并筛选其特征变量。

2 结果

2.1 术后病理结果统计

本研究共计收集ⅠA期亚实性肺腺癌106例，其中16（15%）患者存在STAS。在术后病理结果中，病理类型、脉管侵犯与STAS密切相关（均P<0.05）：STAS(+)组实性成份为主型以及微乳头为主型占比更多；此外，脉管侵犯发生率更高，见表1。

表1 亚实性ⅠA期肺腺癌术后病理学特征［n（%）］

项目	STAS (–)	STAS (+)	P
手术方式			
肺叶切除	57（63.30）	10（62.50）	1.000
亚肺叶切除	33（36.70）	6（37.50）	
病理类型			
乳头为主型	0（0.00）	1（6.25）	<0.001
乳头状为主型	20（22.20）	2（12.50）	
实性为主型	3（3.33）	4（25.00）	
贴壁为主型	14（15.60）	1（6.25）	
微乳头为主型	4（4.44）	4（25.00）	
腺泡为主型	44（48.90）	3（18.80）	
其他	5（5.56）	1（6.25）	
脉管侵犯			
无	83（92.20）	10（62.50）	0.004
有	7（7.78）	6（37.50）	
神经侵犯			
无	84（93.30）	14（87.50）	0.347
有	6（6.67）	2（12.50）	
胸膜侵犯			
无	83（92.20）	13（81.20）	0.174
有	7（7.78）	3（18.80）	

2.2 STAS(+)组与STAS(–)组术前临床影像学资料的比较

性别、病灶大小、实性部分体积、结节总体

积、实性占比、实性部分CT值、胸膜牵拉以及分叶征与与STAS密切相关（均P<0.05）：STAS(+)组女性更多见，具有更大的病灶、实性部分体积、结节总体积、实性占比以及实性部分CT值；此外，胸膜牵拉、分叶征更多见，见表2。

表2 ⅠA期亚实性肺腺癌STAS的术前临床及影像学征象［$\bar{x} \pm s$，n（%）］

变量	STAS(−)	STAS(+)	P
年龄（岁）	60.4 ± 11.2	62.7 ± 13.1	0.527
性别			
男性	51（56.67）	4（25.00）	0.039
女性	39（43.33）	12（75.00）	
吸烟史			
从不	52（57.78）	8（50.00）	0.761
存在	38（42.22）	8（50.00）	
病灶大小	2.48 ± 0.40	2.72 ± 0.24	0.002
实性部分体积	782 ± 199	911 ± 228	0.046
结节总体积	2372 ± 538	1817 ± 752	0.011
实性占比	0.34 ± 0.07	0.58 ± 0.25	0.001
实性部分CT值	−58.84 ± 11.9	−50.52 ± 10.7	0.01
肺叶			
上叶	37（41.11）	9（56.25）	0.611
下叶	44（48.89）	6（37.50）	
中叶	9（10.00）	1（6.25）	
位置			
右肺	44（48.89）	6（37.50）	0.569
左肺	46（51.11）	10（62.50）	
支气管异常征阳性			
否	47（52.22）	4（25.00）	0.082
是	43（47.78）	12（75.00）	
肺血管异常征阳性			
否	38（42.22）	3（18.75）	0.134
是	52（57.78）	13（81.25）	
周围磨玻璃边界			
清晰	75（83.33）	9（56.25）	0.021
模糊	15（16.67）	7（43.75）	
胸膜牵拉征			
无	71（78.89）	7（43.75）	0.011
有	19（21.11）	9（56.25）	
分叶征			
无	75（83.33）	9（56.25）	0.021
有	15（16.67）	7（43.75）	

2.3 传统Logistic回归对ⅠA期亚实性肺腺癌STAS的多因素分析

病灶大小［OR=1.216（1.068 ~ 1.385），P=0.004］、实性占比［OR=3.924（2.749 ~ 5.601），P<0.001］、实性部分CT值［OR=1.007（1.003 ~ 1.011），P=0.002］以及周围磨玻璃边界模糊［1.213（1.065 ~ 1.382），P=0.004］是评估ⅠA期亚实性肺腺癌STAS的独立危险因素，见表3。

表3 IA期亚实性肺腺癌STAS的Logistic回归多因素分析

变量	b值	Wald值	OR_with_CI	P
病灶大小	0.196	8.913	1.216（1.068 ~ 1.385）	0.004
实性占比	1.367	58.065	3.924（2.749 ~ 5.601）	<0.001
实性部分CT值	0.007	10.367	1.007（1.003 ~ 1.011）	0.002
周围磨玻璃边界清晰	−0.193	8.700	0.824（0.724 ~ 0.939）	0.004
分叶有	0.112	2.844	1.119（0.98 ~ 1.277）	0.095

2.4 不同机器学习算法模型诊断效能比较

通过5种机器学习算法，Xgboost的诊断效能最高，且优于传统Logistic回归［曲线下面积分别为：0.969（95%CI：0.915 ~ 0.993）vs. 0.941（0.878 ~ 0.977）］，见表4、图3。基于Xgboost算法，更高的实性占比、实性部分CT值、病灶大小、周围磨玻璃边界模糊以及实性部分体积是ⅠA期亚实性肺腺癌STAS的预测因子见图4。

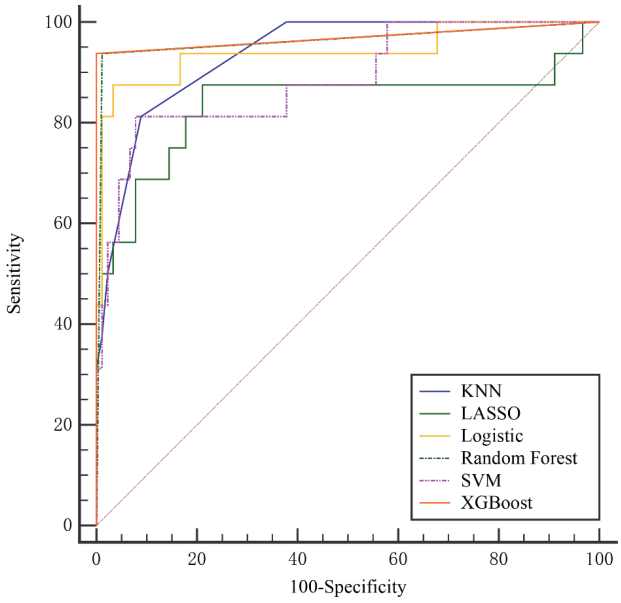


图3 不同机器学习模型的ROC比较

表 4 不同机器学习算法对 STAS 的诊断效能分析

算法模型	AUC（95%CI）	敏感度（%）	特异度（%）	阳性预测值（%）	阴性预测值（%）
Logistic	0.941（0.878 ~ 0.977）	87.50	96.67	82.40	97.80
KNN	0.936（0.872 ~ 0.975）	81.25	91.11	61.90	96.50
LASSO	0.837（0.752 ~ 0.901）	87.50	78.89	42.40	97.30
Random_Forest	0.963（0.907 ~ 0.990）	93.75	98.89	93.80	98.90
SVM	0.887（0.811 ~ 0.940）	81.25	92.22	65.00	96.50
XGBoost	0.969（0.915 ~ 0.993）	93.75	100.00	100.00	98.90

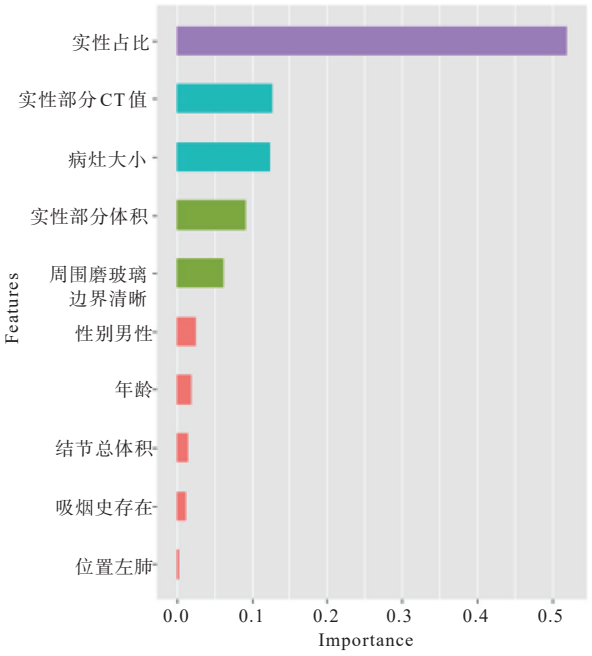


图 4 基于 Xgboost 算法筛选出亚实性Ⅰ A 期肺腺癌 STAS 的预测因子

3 讨论

STAS 被视为Ⅰ A 期肺癌亚肺叶切除术后不良预后的重要预示指标[6,7]。Eguchi 等[14]的研究揭示，在Ⅰ A 期肺腺癌且伴有 STAS 的患者中，相较于亚肺叶切除，实施肺叶切除能带来更优的预后效果。Warth 等[15]的研究则强调，无论肺腺癌患者的 STAS 是局灶性还是广泛性，其预后均劣于 STAS 阴性的患者。本研究通过对 STAS 阳性患者的术后病理特征进行深入分析，发现这类患者的脉管侵犯率显著增高，且病理亚型多以预后较差的实性或微乳头成分为主型为主，这与先前的研究结果相吻合[6,7]。由此可见，早期肺腺癌中的 STAS 会增加患者不良预后的风险，因此，对于这类患者，在手术方式的选择上应更倾向于根治性肺叶切除。

既往的一项 meta 分析指出[16]，亚实性肺腺癌具有更高的 STAS 发生风险，因此，对这一特定群体进行系统分析，探究术前临床影像学特征对 STAS 的预测价值显得尤为重要。既往研究通过术前 CT 常规影像学特征评估肺腺癌的 STAS，其主要不足在于未能充分挖掘病变的诊断信息，且存在较大的主观性，受限于评估医师的诊断水平，造成其结果的一致性较低，缺乏代表性，因此亟需量化客观的指标来评估肺腺癌术后的 STAS[8-10]。本研究运用机器学习算法，全面挖掘临床影像特征与Ⅰ A 期亚实性肺腺癌之间的关联，通过对多种机器学习算法的 ROC 曲线进行对比分析，最终确定 Xgboost 算法的诊断效能最佳。XGBoost 算法通过引入正则化项来有效控制模型的复杂度，从而有效防止过拟合；应用 XGBoost 算法，在 16 例 STAS 阳性的患者中，仅 1 例被漏诊，真阳性率（敏感度）达到 93.75%（15/16）；而在 90 例 STAS 阴性的患者中，该算法全部准确判断，真阴性率（特异度）为 100%（90/90）。此外，本研究通过特征变量降维处理，最终识别出实性成分比例更高、实性部分 CT 值、病灶尺寸、实性部分体积以及周围磨玻璃影边界不清晰是Ⅰ A 期亚实性肺腺癌 STAS 的预测因素。对于亚实性结节而言，其病理侵袭性的强弱取决于实性部分的大小及比例[17]，本研究对实性成分进行了多维度的参数化量化，并证实了其在早期肺腺癌 STAS 形成中的关键作用，这一发现与以往的研究结果相近[8-10]。Zhang 等[18]的研究表明，亚实性结节周围的磨玻璃成分，其光滑的边界往往是贴壁生长为主型肺腺癌的特异性影像学表现，此类病理亚型的恶性程度较低，预后最佳。因而可以推测，亚实

性结节周围磨玻璃影边界模糊是肺腺癌STAS的危险因素。

总而言之，本研究利用机器学习算法构建了ⅠA期亚实性肺腺癌的STAS预测模型，该模型在显著提升诊断效能的同时，也简化了特征组合，规范了思维路径，为STAS的术前诊断提供了全面且准确的参考依据，更为人工智能产品的开发奠定了研究基础。本研究存在一定的局限性：首先，作为回顾性研究，可能存在选择偏倚的风险；其次，本研究仅评估了ⅠA期亚实性肺腺癌，且样本量相对较小；最后，虽然本研究纳入了术前的各种临床影像资料，但未能包含分子生物学、基于表型等预测因子，这可能在一定程度上限制了模型的泛化能力。

[参 考 文 献]

[1] Zuo ZC, Wang LD, Peng K, et al. Development and validation of a nomogram for predicting the 1–, 3–, and 5–year survival in patients with acinar–predominant lung adenocarcinoma [J]. Curr Med Sci, 2022, 42(6):1178–1185.

[2] 张三红,左志超,叶姣,等.ⅠA期非小细胞肺癌肺叶切除术与亚肺叶切除术预后结局比较[J].肿瘤预防与治疗,2021,34(8):712–720.

[3] Bains S, Eguchi T, Warth A, et al. Procedure–specific risk prediction for recurrence in patients undergoing lobectomy or sublobar resection for small (≤2 cm) lung adenocarcinoma: an international cohort analysis [J]. J Thorac Oncol, 2019, 14(1):72–86.

[4] Kumbasar U. Spread through air spaces positivity and extent of resection in patients with Stage I non–small cell lung cancer: A contemporary review [J]. Turk Gogus Kalp Damar Cerrahisi Derg, 2022, 30(1):141–144.

[5] Jung W, Chung JH, Yum S, et al. The differential prognostic impact of spread through air spaces in early–stage lung adenocarcinoma after lobectomy according to the pT descriptor [J]. J Thorac Cardiovasc Surg, 2022, 163(1):277–284.

[6] Lv Y, Li S, Liu Z, et al. Impact of surgery and adjuvant chemotherapy on the survival of stage I lung adenocarcinoma patients with tumor spread through air spaces [J]. Lung Cancer, 2023, 177:51–58.

[7] Vaghjiani RG, Takahashi Y, Eguchi T, et al. Tumor spread through air spaces is a predictor of occult lymph node metastasis in clinical stage IA lung adenocarcinoma [J]. J Thorac Oncol, 2020, 15(5):792–802.

[8] Gu Y, Zheng B, Zhao T, Fan Y. Computed tomography features and tumor spread through air spaces in lung adenocarcinoma: A Meta–analysis [J]. J Thorac Imaging, 2023, 38(2):W19–W29.

[9] Gao Z, An P, Li R, et al. Development and validation of a clinic–radiological model to predict tumor spread through air spaces in stage I lung adenocarcinoma [J]. Cancer Imaging, 2024, 24(1):25.

[10] Zheng Y, Han X, Li H, et al. A validated model to predict spread through air space in lung adenocarcinoma [J]. J Thorac Dis, 2024, 16(4):2296–2313.

[11] Quanyang W, Yao H, Sicong W, et al. Artificial intelligence in lung cancer screening: Detection, classification, prediction, and prognosis [J]. Cancer Med, 2024, 13(7):e7140.

[12] Goldstraw P, Chansky K, Crowley J, er al. The IASLC lung cancer staging project: proposals for revision of the TNM stage groupings in the forthcoming (eighth) edition of the TNM classification for lung cancer [J]. J Thorac Oncol, 2016, 11(1):39–51.

[13] Zuo Z, Zeng W, Peng K, et al. Development of a novel combined nomogram integrating deep–learning–assisted CT texture and clinical–radiological features to predict the invasiveness of clinical stage IA part–solid lung adenocarcinoma: a multicentre study [J]. Clin Radiol, 2023, 78(10):e698–e706.

[14] Eguchi T, Kameda K, Lu S, et al. Lobectomy is associated with better outcomes than sublobar resection in spread through air spaces (stas)–positive t1 lung adenocarcinoma: A propensity score–matched analysis [J]. J Thorac Oncol, 2019, 14(1): 87–98.

[15] Warth A, Muley T, Kossakowski CA, et al. Prognostic impact of intra–alveolar tumor spread in pulmonary adenocarcinoma [J]. Am J Surg Pathol, 2015, 39(6):793–801.

[16] Yin Q, Wang H, Cui H, et al. Meta–analysis of association between CT–based features and tumor spread through air spaces in lung adenocarcinoma [J]. J Cardiothorac Surg, 2020, 15(1):243.

[17] Ahn Y, Lee SM, Noh HN, et al. Use of a commercially available deep learning algorithm to measure the solid portions of lung cancer manifesting as subsolid lesions at CT: Comparisons with radiologists and invasive component size at pathologic examination [J]. Radiology, 2021, 299(1):202–210.

[18] Zhang J, Sha J, Liu W, et al. Quantification of intratumoral heterogeneity: distinguishing histological subtypes in clinical t1 stage lung adenocarcinoma presenting as pure ground–glass nodules on computed tomography [J]. Acad Radiol, 2024, 31(10):4244–4255.