

## 临床研究

DOI:10.13406/j.cnki.cxyb.003200

## 纯磨玻璃结节肺腺癌 CT 征象与其浸润程度的 相关性分析及预测模型构建

胡 宇,杜 铭

(重庆医科大学附属第一医院胸心外科,重庆 400016)

**【摘要】目的:**探究纯磨玻璃结节(pure ground-glass nodule, pGGN)肺腺癌计算机断层扫描(computed tomography, CT)征象与其浸润程度的相关性,建立CT征象与浸润程度的预测模型。**方法:**回顾性分析424例经手术切除、病理活检证实且胸部CT表现为pGGN的肺腺癌患者临床资料及CT征象,根据病理活检结果分为非典型腺瘤样增生、原位腺癌、微浸润腺癌和浸润性腺癌4组,对组间差异采用卡方检验或Fisher确切概率法进行统计分析。对有统计学意义的结果,使用怀卡托智能分析环境(Waikato environment for knowledge analysis, WeKa)中的6种学习算法进行预测模型构建,并验证准确性,挑选出最适用于本研究的预测模型。**结果:**4组间在结节直径、结节密度值上的差异具有统计学意义( $P<0.001$ ),对应的直径平均值分别为6.90、8.65、10.71、14.56 mm,对应的密度平均值分别为-633.16、-543.04、-401.03、-322.94 HU,随着病灶的浸润程度加重,结节的直径及密度值呈现明显的上升趋势。4组间在结节边界、分叶、毛刺、血管穿行、胸膜凹陷、空气支气管征、空泡征等的差异具有统计学意义( $P<0.05$ ),而结节的生长位置、年龄、吸烟史、直系亲属肺癌家族史等差异无统计学意义( $P>0.05$ )。随机森林算法所构建的模型预测准确率为76.42%~79.72%,Kappa系数为0.597~0.670,受试者工作特征(receiver operating characteristic, ROC)曲线下面积均大于0.9,在误差指标中表现最优,是最适合于本研究的预测模型。**结论:**pGGN的不同CT征象与其浸润程度密切相关,可以用于建立预测模型。基于随机森林算法所建模型,在有创干预前早期快速识别pGGN浸润程度的平均准确率为78.07%,准确度最高,对肺癌预测具有潜在应用价值。

**【关键词】**纯磨玻璃结节;肺腺癌;计算机断层扫描征象;浸润程度;怀卡托智能分析环境;预测模型

**【中图分类号】**R655.3

**【文献标志码】**A

**【收稿日期】**2022-10-07

## Correlation analysis and prediction model construction of CT signs and infiltration degree of pure ground-glass nodule lung adenocarcinoma

Hu Yu, Du Ming

(Department of Cardiothoracic Surgery, The First Affiliated Hospital of Chongqing Medical University)

**【Abstract】****Objective:** To explore the correlation between computed tomography(CT) signs and infiltration degree of pure ground-glass nodule(pGGN) lung adenocarcinoma, and establish a prediction model of CT signs and infiltration degree. **Methods:** The clinical data and CT signs of 424 patients with lung adenocarcinoma confirmed by surgical resection, pathological biopsy and chest CT findings of pGGN were analyzed retrospectively, and according to the results of pathological biopsy, they were divided into four groups: atypical adenomatous hyperplasia, adenocarcinoma *in situ*, minimally invasive adenocarcinoma and invasive adenocarcinoma. The Chi-square test or Fisher's exact probability test was used for statistical analysis of differences between groups. For the results with statistical significance, six learning algorithms in Waikato environment for knowledge analysis(WeKa) were used to build the prediction model, verify the accuracy, and select the prediction model most suitable for this study. **Results:** The differences in nodule diameter and nodule density between the four groups were statistically significant ( $P<0.001$ ), with the corresponding mean diameters of 6.90, 8.65, 10.71 and 14.56 mm, and the corresponding mean densities of -633.16, -543.04, -401.03 and -322.94 HU, respectively. The diameter and density of the nodules showed an obvious upward trend with the increase of the degree of invasion of the lesions. There were statistically significant differences among the four groups in nodule boundary, lobulation, burr, vascular perforation, pleural indentation, air bronchogram sign, and vacuole sign( $P<0.05$ ), but there were no statistically significant differences in nodule growth position, age, smoking history, family history of lung cancer in immediate family( $P>0.05$ ). The prediction accuracy of the model constructed by the random forest algorithm fluctuated between 76.42% and 79.72%, the Kappa coefficient fluctuated between 0.597 and 0.670, and the area under the receiver operating characteristic(ROC)

**作者介绍:**胡 宇,Email:1244403760@qq.com,

**研究方向:**胸心外科肿瘤治疗。

**通信作者:**杜 铭,Email:ljdyt1103@sina.com。

**优先出版:**<https://kns.cnki.net/kems/detail/50.1046.R.20230427.1711.006.html>

(2023-04-28)

curve was greater than 0.9, which was the best among the error indicators. Thus, it is the most suitable prediction model for this study.

**Conclusion:** The different CT signs of pGGN are closely related to the degree of infiltration, and can be used to establish a prediction model. Based on the model built by random forest algorithm, the average accuracy of quickly identifying the pGGN infiltration degree in the early stage before invasive intervention is 78.07%, which is the highest accuracy, and can potentially be used in the prediction of lung cancer in the future.

**[Key words]** pure ground-glass nodule; lung adenocarcinoma; computed tomography sign; degree of infiltration; Waikato environment for knowledge analysis; prediction model

原发性支气管肺癌是全世界范围内发病率及死亡率均较高的恶性肿瘤之一,是中国发病率及死亡率均位于第一位的恶性肿瘤<sup>[1]</sup>,严重危害人们的生命安全。早期肺癌在计算机断层扫描(computed tomography, CT)图像上常表现为肺结节。随着低剂量螺旋CT的广泛应用,肺结节检出率明显增加。肺结节可分为实性结节、混合磨玻璃结节及纯磨玻璃结节(pure ground-glass nodule, pGGN)。实性结节指CT图像上表现为明显高于正常肺组织密度的实性成分影,其中血管、支气管等结构被完全掩盖;纯磨玻璃结节指CT图像上表现为稍高于正常组织密度的磨玻璃样模糊影,其中血管、支气管能够正常显示<sup>[2]</sup>;混合磨玻璃结节则位于两者之间,指病灶同时含有实性成分及磨玻璃成分,血管、支气管等结构部分可见。当pGGN持续存在于同一部位超过3个月且无明显变化时,可认为与肺癌相关<sup>[3]</sup>。CT图像上表现为pGGN的肺腺癌,大多数为非典型腺瘤样增生(atypical adenomatous hyperplasia, AAH)、原位腺癌(adenocarcinoma *in situ*, AIS)和微浸润腺癌(minimally invasive adenocarcinoma, MIA),少部分为浸润性腺癌(invasive adenocarcinoma, IAC)<sup>[4]</sup>。国内外研究多围绕肺结节良性与恶性之间的鉴别诊断<sup>[5]</sup>,针对表现为pGGN的AAH、AIS、MIA和IAC四者间的研究较为少见,鉴别存在一定的困难。

机器学习算法是一种新型的可用于疾病预测的方法。与其他统计学方法相比,机器学习算法可给予个体化简单、直观的预测结果。怀卡托智能分析环境(Waikato environment for knowledge analysis, WeKa)是一款基于JAVA环境下集合了大量数据挖掘任务的机器学习算法,包括对数据进行预处理、分类、回归、关联等。临幊上推荐使用学习算法构建预测模型对疾病的危险因素及预后等进行量化分析<sup>[6-7]</sup>,如北大李运教授等<sup>[8]</sup>构建的孤立性肺结节良恶性判断预测模型;丁宁等<sup>[9]</sup>构建的预测肺部混合磨玻璃结节微浸润腺癌和浸润性腺癌的Nomogram模型;王成武等<sup>[10]</sup>使用WeKa建立的关于早期糖尿病风险因素的预测模型;Mansour EA<sup>[11]</sup>使用

WeKa建立的关于胸癌预后的Random Committee模型。肺癌的浸润程度预测模型,可在手术切除、支气管镜检、经皮穿刺活检等<sup>[12]</sup>一系列有创操作前早期准确地了解其发展程度,达到减少患者创伤而不过度干预的目的。由于AAH、AIS、MIA和IAC的侵袭性不同,其预后及处理方法也大不相同。因此,本研究拟通过探究pGGN肺腺癌的CT征象,通过CT征象建立预测模型,早期预测出其浸润程度,为临床早期诊断及选择合适的手术时机提供理论依据。

## 1 资料与方法

### 1.1 研究对象

回顾性分析2020年9月至2022年5月,由重庆医科大学附属第一医院胸心外科经手术切除、病理切片活检且胸部CT表现为pGGN的肺腺癌患者424例。其中男性128例,女性296例。年龄22~79岁,平均年龄(52.5±10.7)岁。

**1.1.1 纳入标准** ①胸部CT上病灶为≥5 mm且<3 cm<sup>[13]</sup>的pGGN;②术前完善胸部CT(CT重建图像≤1 mm),手术时间与CT时间间隔不超过1个月<sup>[14]</sup>;③无术前放化疗、靶向治疗、免疫治疗等肿瘤相关治疗;④术后病理类型为AAH、AIS、MIA、IAC。

**1.1.2 排除标准** ①病灶为≥3 cm的肿块;②肺实性结节、混合磨玻璃结节;③有术前放化疗、靶向治疗等肿瘤相关治疗;④术后病理类型为良性病变、鳞癌、小细胞肺癌、转移癌等。

### 1.2 病理诊断

所有肺结节标本均由专业手术医师于微创电视胸腔镜下手术或者传统开胸手术下获得。标本常规由病理科切片处理,由2~3名经验丰富的专业病理医师诊断,并出具病理学报告。肺结节的浸润程度以该病理学报告为准。

### 1.3 设备与图像

采用Discovery CT750HD或Light speed VCT螺旋CT机进行扫描。扫描前训练患者深呼吸及屏气动作,在患者深吸气未处于屏气状态时扫描,此时双肺处于膨胀状态。患者取仰卧位,双臂上举,扫描范围为双侧肺尖至双侧肋膈角消失。成像参数:管电压控制为120 kV;管电流为100~250 mA。所得图像的重组截面厚度分别为0.625 mm或1 mm。所有的患者胸部CT图像存储至放射科数据库,由2~3名专业的影像学医师对图像进行独立分析,达成统一意见后出具患者胸部

CT 报告。CT 报告上包括结节的位置、直径及其他形态学征象。

#### 1.4 临床特征及 CT 征象

患者的临床特征包括性别、年龄、吸烟史、直系亲属肺癌家族史<sup>[1]</sup>。胸部 CT 征象包括结节生长位置、结节最大直径、结节最大密度值、结节边界、分叶征、毛刺征、胸膜凹陷征、空气支气管征、血管穿行征、空泡征等<sup>[3,15-16]</sup>。将 CT 征象分为定量征象及定性征象：定量征象包括结节最大直径及结节最大密度值，其余归为定性征象。

1.4.1 定量征象 ①结节最大直径：选取重建层厚≤1 mm 的胸部 CT 图像，手动测量从结节起始层面至结束层面中每一个层面的结节直径，得到的结节直径取最大值。将手动测量的结节直径最大值与放射科出具的胸部 CT 报告结节最大直径对比，误差较大时重新测量以减小误差。②结节最大密度值：选取重建层厚≤1 mm 的胸部 CT 图像，手动测量从结节起始层面至结束层面中每一个层面上结节的密度，得到的结节密度取最大值，测量过程中避开血管、支气管、空泡，多次测量以减少误差。

1.4.2 定性征象 ①结节生长位置：包括右肺上叶、右肺中叶、右肺下叶、左肺上叶及左肺下叶。②结节边缘征象：包括边界（是否清晰）、分叶征、毛刺征、胸膜凹陷征。③结节内部征象：空气支气管征、血管穿行征、空泡征。

#### 1.5 统计学处理

采用 SPSS 25.0 统计软件进行数据处理分析。计量资料结果用均数±标准差（ $\bar{x} \pm s$ ）表示，采用单因素方差分析；计数资料用频率和百分比（n, %）表示，采取独立样本卡方检验和 Fisher 确切概率法分析。检验水准  $\alpha=0.05$ 。

#### 1.6 预测模型

采用 Weka 建立 pGGN 肺腺癌 CT 征象与其浸润程度的预测模型。采用随机原则，将 424 例肺结节按照 1:1、2:1、3:1、4:1 的比例分为 4 组，每一组中分建模组及验证组<sup>[17]</sup>，组一：建模组 212 例，验证组 212 例；组二：建模组 283 例，验证组 141 例；组三：建模组 318 例，验证组 106 例；组四：建模组 339 例，验证组 85 例。使用 Weka 对建模组数据分析，建立预测模型，用验证组数据检验该预测模型的准确率。本研究共采用 6 类不同的算法对建模组数据进行建模分析，分别是朴素贝叶斯（Naive Bayes）、支持向量机（support vector machine, SVM）、K 最近邻学习器（Instance-based Learning K-Nearest Neighbor, IBK）、随机委员会（Random Committee）、决策树 J48（Decision tree J48）和随机森林（Random Forest）。

#### 1.7 模型评价指标

评价指标包括训练集准确率（training set accuracy）、交叉验证准确率（cross validation accuracy）、测试集准确率（testing set accuracy）、受试者工作特征（receiver operating characteristic, ROC）曲线下面积（area under the ROC curve, AUC）、Kappa 系数、平均绝对误差（mean absolute error, MAE）、均方根误差（root mean squared error, RMSE）、相对绝对误差（relative absolute error, RAE）和相对平方根误差（root relative squared error, RRSE）<sup>[11,18]</sup>。

AUC：为 0~1 范围的小数。AUC 面积一般大于 0.5，这个值越接近 1，说明模型的分类效果越好。这个值在 0.5~0.7 时有较低准确度，在 0.7~0.9 时有一定准确度，在 0.9 以上时有较高的准确度。

Kappa 系数：用于衡量模型预测结果和实际分类结果是否一致。Kappa 系数取值在 -1~1 之间，实际应用中 Kappa 系数只有在 0~1 时判断一致性才有意义。Kappa 系数越大，说明一致性越好。一般而言，Kappa 系数 <0.4 时，表明一致性较差；0.40<Kappa 系数≤0.60 时，表明中度一致；0.60<Kappa 系数≤0.80 时，表明有较高度的一致性；Kappa 系数>0.80 时，表明有极好的一致性。

MAE：是绝对误差的平均值。用于评估预测结果和真实结果的一致性，结果越小表示模型越优。

RMSE：衡量变量精度的常用指标，用来衡量预测值与真实值之间的偏差，数值越小表示模型越稳定。

RAE：衡量变量精度的常用指标，预测数值与实际数值两者差的绝对值，数值越小表明模型越优。

RRSE：是预测值与实际值偏差的平方与次数比值的平方根，用来衡量预测数值同实际数值之间的偏差，数值越小表明模型越优。

## 2 结 果

### 2.1 基本情况

不同浸润程度的 pGGN 患者临床特征比较，AAH、AIS、MIA、IAC 4 组患者在性别、年龄、吸烟史、直系亲属肺癌家族史上差异均无统计学意义（ $P>0.05$ ）。结节好发于中老年人，平均发病年龄为  $(52.5 \pm 10.7)$  岁，女性较为多见，男女比约为 1:2。吸烟患者与不吸烟患者比约为 1:5，直系亲属患有肺癌与不患肺癌比约为 1:10。见表 1。

表 1 不同浸润程度的 pGGN 患者临床特征（n；n, %； $\bar{x} \pm s$ ）

项目	结节个数	性别		年龄/岁	吸烟史	直系亲属家族史
		男	女			
AAH	50	13(26.0)	37(74.0)	51.2 ± 10.2	10(20.0)	7(14.0)
AIS	232	62(26.7)	170(73.3)	52.3 ± 11.3	32(13.8)	18(7.8)
MIA	126	46(36.5)	80(63.5)	52.8 ± 9.9	30(23.8)	11(8.7)
IAC	16	7(43.8)	9(56.2)	58.2 ± 8.4	4(25.0)	2(12.5)
$\chi^2/F$ 值		5.521		1.805	6.349	2.221
P 值		0.135		0.146	0.077	0.435

## 2.2 CT 征象

不同浸润程度的 pGGN 患者胸部 CT 定量征象比较, AAH、AIS、MIA、IAC 4 组患者的胸部 CT 中, 结节最大直径与结节最大密度值差异有统计学意义 ( $P < 0.001$ )。结节的最大直径平均值总体上呈现明显上升趋势, 其中 50 例 AAH 中直径最小者为 5 mm, 最大者为 11 mm, 平均直径为 6.90 mm; 232 例 AIS 中直径最小者为 5 mm, 最大者为 23 mm, 平均直径为 8.65 mm; 126 例 MIA 中直径最小者为 5 mm, 最大者为 23 mm, 平均直径为 10.71 mm; 16 例 IAC 中直径最小者为 10 mm, 最大者为 24 mm, 平均直径为 14.56 mm。结节最大 CT 密度值总体上呈现明显上升趋势, 其中 50 例 AAH 中密度最小者为 -773 HU, 最大者为 -548 HU, 平均密度为 -633.16 HU; 232 例 AIS 中密度最小者为 -693 HU, 最大者为 -414 HU, 平均密度为 -543.04 HU; 126 例 MIA 中密度最小者为 -494 HU, 最大者为 -224 HU, 平均密度为 -401.03 HU; 16 例 IAC 中密度最小者为 -442 HU, 最大者为 -186 HU, 平均密度为 -322.94 HU。见表 2。

不同浸润程度的 pGGN 患者胸部 CT 定性征象比较, AAH、AIS、MIA、IAC 4 组患者的胸部 CT 中, 结节生长位置差异无统计学意义 ( $P > 0.05$ ), 结节边界、分叶征、毛刺征、胸膜凹陷征、空气支气管征、血管穿行征、空泡征具有统计学意义

表 2 不同浸润程度的 pGGN 患者胸部 CT 定量征象比较 ( $n; \bar{x} \pm s$ )

项目	结节数/个	结节直径/mm	结节密度/HU
AAH	50	6.90 ± 1.71	-633.16 ± 44.77
AIS	232	8.65 ± 3.12	-543.04 ± 59.62
MIA	126	10.71 ± 3.52	-401.03 ± 47.32
IAC	16	14.56 ± 4.56	-322.94 ± 69.56
$\chi^2/F$ 值		35.530	335.560
P 值		<0.001	<0.001

( $P < 0.05$ )。随着 pGGN 浸润程度的加重, 结节边界逐渐变得不规则甚至消失, 分叶、胸膜凹陷、血管穿行等征象出现的概率逐渐增加。见表 3。部分 pGGN 患者 CT 图像见图 1。

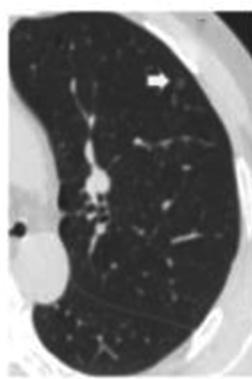
## 2.3 预测模型

分别使用 Weka 中的 6 种算法对 4 组不同建模比下的建模组数据进行分析, 建立预测模型(即训练集), 记录训练集准确率。将预测模型进行自身数据交叉验证, 记录交叉验证准确率。分别将对应的验证组数据(即测试集)代入所建模型, 记录测试集准确率、AUC、Kappa 系数、MAE、RMSE、RAE、RRSE。对比分析出最合适的预测模型。

2.3.1 模型的准确率 6 个预测模型中, 训练集准确率能达到 100% 的为 IBK, Random Committee 和 Random Forest, 最低的为建模比 1:1 时的 Naive Bayes(79.24%)。将 6 个预测模

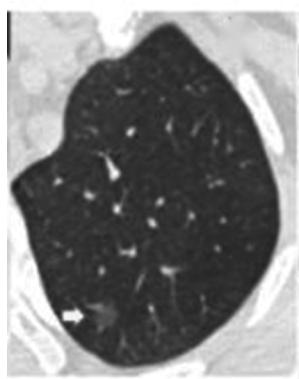
表 3 不同浸润程度的 pGGN 患者胸部 CT 定性征象比较 ( $n; n, \%$ )

项目	AAH( $n=50$ )	AIS( $n=232$ )	MIA( $n=126$ )	IAC( $n=16$ )	$\chi^2$ 值	P 值
结节位置(右肺/左肺)	30/20	151/81	79/47	5/11	7.464	0.065
边界清晰	41(0.82)	152(0.66)	65(0.52)	5(0.31)	21.933	<0.001
分叶征	0(0.00)	11(0.05)	21(0.17)	4(0.25)	25.290	<0.001
毛刺征	0(0.00)	23(0.10)	43(0.34)	3(0.19)	46.153	<0.001
胸膜凹陷	1(0.02)	11(0.05)	12(0.10)	3(0.19)	8.848	0.034
支气管征	0(0.00)	10(0.04)	6(0.05)	5(0.31)	26.323	0.001
血管穿行	24(0.48)	149(0.64)	104(0.83)	14(0.88)	32.049	<0.001
空泡征	0(0.00)	24(0.10)	23(0.18)	3(0.19)	12.954	0.001

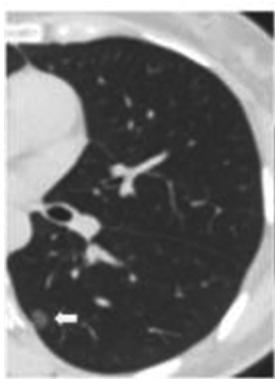


A. 男, 46岁, CT示左肺上叶 5 mm pGGN, 边界清晰。

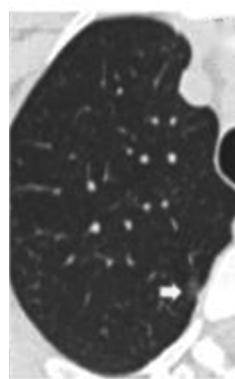
病理活检示 AAH



B. 女, 49岁, CT示左肺上叶 8 mm pGGN, 边界清晰, 边缘及内部可见小血管穿行。病理活检示 AIS



C. 女, 34岁, CT示左肺下叶 10 mm pGGN, 边界清晰, 邻近胸膜稍牵拉凹陷。病理活检示 MIA



D. 男, 40岁, CT示右肺上叶 9 mm pGGN, 边界清晰, 邻近胸膜稍牵拉凹陷。病理活检示 MIA

图 1 部分 pGGN 患者 CT 图像

型进行自身交叉验证,准确率前2位的是Naive Bayes(78.93%)和Random Forest(78.80%)。将验证组数据代入6个预测模型分别验证准确率,准确率前2位为Random Forest(79.72%)和Naive Bayes(79.25%)。就模型的准确率而言,Random Forest在建立模型、自身交叉验证以及外部验证中,都有较为突出的表现。Naive Bayes自身交叉验证准确率处于第1位,但其训练集准确率处于最低位。见表4至表7。

表4 建模比1:1时模型准确率(n,%)

模型	训练集	交叉验证	测试集
	准确率	准确率	准确率
Naive Bayes	168(79.24)	161(75.94)	168(79.25)
SVM	195(91.98)	144(67.92)	157(74.06)
IBK	212(100.00)	134(63.21)	142(66.98)
Random Committee	212(100.00)	152(71.70)	164(77.36)
Decision tree J48	187(88.21)	157(74.06)	165(77.83)
Random Forest	212(100.00)	158(74.53)	169(79.72)

表5 建模比2:1时模型准确率(n,%)

模型	训练集	交叉验证	测试集
	准确率	准确率	准确率
Naive Bayes	226(79.86)	220(77.74)	111(78.72)
SVM	262(92.58)	206(72.79)	107(75.89)
IBK	283(100.00)	196(69.26)	92(65.25)
Random Committee	283(100.00)	212(74.91)	106(75.18)
Decision tree J48	239(84.45)	218(77.03)	105(74.47)
Random Forest	293(100.00)	223(78.80)	109(77.31)

表6 建模比3:1时模型准确率(n,%)

模型	训练集	交叉验证	测试集
	准确率	准确率	准确率
Naive Bayes	257(80.82)	251(78.93)	80(75.47)
SVM	293(92.14)	233(73.27)	80(75.47)
IBK	318(100.00)	216(67.92)	75(70.75)
Random Committee	318(100.00)	237(74.53)	79(74.53)
Decision tree J48	269(84.59)	238(74.84)	79(74.53)
Random Forest	318(100.00)	249(78.30)	81(76.42)

表7 建模比4:1时模型准确率(n,%)

模型	训练集	交叉验证	测试集
	准确率	准确率	准确率
Naive Bayes	269(79.35)	262(77.29)	67(78.82)
SVM	311(91.74)	249(73.45)	65(76.47)
IBK	339(100.00)	232(68.44)	59(69.41)
Random Committee	339(100.00)	252(74.34)	62(72.94)
Decision tree J48	290(85.55)	255(75.22)	65(76.47)
Random Forest	339(100.00)	257(75.81)	67(78.82)

2.3.2 Kappa系数 6个预测模型中,Kappa系数最高的前2位为Naive Bayes(0.677)和Random Forest(0.670),提示该模

型预测值与实际值有较高的一致性。就Kappa系数而言,Naive Bayes和Random Forest在各个建模比下都有较好的表现。见表8至表11。

表8 建模比1:1测试集验证结果

模型	Kappa	AUC	MAE	RMSE	RAE	RRSE
Naive Bayes	0.677	0.917	0.151	0.293	0.501	0.759
SVM	0.509	0.737	0.130	0.360	0.433	0.932
IBK	0.465	0.741	0.167	0.398	0.556	1.031
Random Committee	0.640	0.883	0.134	0.300	0.445	0.776
Decision tree J48	0.627	0.869	0.145	0.303	0.484	0.784
Random Forest	0.670	0.926	0.139	0.273	0.463	0.708

表9 建模比2:1测试集验证结果

模型	Kappa	AUC	MAE	RMSE	RAE	RRSE
Naive Bayes	0.667	0.923	0.135	0.276	0.449	0.713
SVM	0.552	0.760	0.121	0.347	0.402	0.899
IBK	0.412	0.703	0.175	0.411	0.593	1.063
Random Committee	0.594	0.855	0.137	0.314	0.458	0.813
Decision tree J48	0.570	0.911	0.148	0.291	0.495	0.754
Random Forest	0.621	0.922	0.133	0.273	0.445	0.706

表10 建模比3:1测试集验证结果

模型	Kappa	AUC	MAE	RMSE	RAE	RRSE
Naive Bayes	0.608	0.906	0.149	0.302	0.498	0.782
SVM	0.539	0.754	0.123	0.350	0.410	0.908
IBK	0.493	0.740	0.147	0.376	0.490	0.973
Random Committee	0.567	0.854	0.136	0.322	0.454	0.834
Decision tree J48	0.569	0.904	0.147	0.293	0.491	0.758
Random Forest	0.597	0.923	0.128	0.269	0.428	0.697

表11 建模比4:1测试集验证结果

模型	Kappa	AUC	MAE	RMSE	RAE	RRSE
Naive Bayes	0.666	0.925	0.140	0.284	0.468	0.737
SVM	0.563	0.767	0.118	0.343	0.394	0.891
IBK	0.462	0.729	0.153	0.383	0.511	0.996
Random Committee	0.540	0.851	0.129	0.308	0.433	0.799
Decision tree J48	0.584	0.919	0.140	0.285	0.469	0.740
Random Forest	0.638	0.928	0.128	0.262	0.428	0.681

2.3.3 AUC 6个预测模型中,AUC最高的前2位为Random Forest(0.928)和Naive Bayes(0.925),提示该模型有较高的准确度。就AUC而言,Random Forest和Naive Bayes都有较高的准确度,4个建模比下均>0.9。见表8至表11。

2.3.4 误差指标 6个预测模型中,建模比1:1时,MAE最低的前2位是SVM和Random Committee,其余建模比下最低前2位均为SVM和Random Forest。建模比3:1时,RMSE最低的前2位为Random Forest和J48,其余建模比下最低前2位均为Random Forest和Naive Bayes。建模比1:1时,RAE最低的前2位为SVM和Random Committee,其余建模比下最

低前 2 位均为 SVM 和 Random Forest。建模比 3:1 时, RRSE 最低的前 2 位为 Random Forest 和 J48, 其余建模比下前 2 位均为 Random Forest 和 Naive Bayes。见表 8 至表 11。

综合以上实验结果, 算法 Random Forest 建立的模型最适用于本研究。4 种建模比下 Random Forest 建立模型的 Kappa 系数分别为 0.597、0.621、0.638、0.670, 提示其预测值与实际值一致性良好。AUC 均大于 0.9, 提示模型有较好的准确度。Random Forest 在 MAE、RMSE、RAE 和 RRSE 这几项误差指标中多次处于最低值, 提示该模型预测值最贴近实际值。Naive Bayes 虽然在 Kappa 系数和部分 AUC 方面略优于 Random Forest, 但其训练集准确率远远低于 Random Forest, 且为 6 个模型中最低, 不如 Random Forest 合适。

### 3 讨 论

本研究通过测量肺 pGGN 在 CT 上的不同征象, 收集对应的结节术后病理学浸润程度, 分析 CT 征象与浸润程度的相关性, 基于 Naive Bayes、SVM、IBK、Random Committee、J48 和 Random Forest 6 种算法, 在不同建模比下建立了识别浸润程度的预测模型。本研究对肺 pGGN CT 征象与浸润程度进行相关性分析, 结果表明结节直径、结节 CT 密度值、边界、分叶征、毛刺征、胸膜凹陷征等征象与浸润程度存在明显的相关性。AAH、AIS、MIA 及 IAC 4 组间的结节最大直径平均值呈现明显的上升趋势, 分别为 6.90、8.65、10.71、14.56 mm, 当 AAH 发展为 IAC 时, 结节平均直径约为原来的 2 倍。AAH、AIS、MIA 及 IAC 4 组间结节最大 CT 密度平均值分别为 -633.16、-543.04、-401.03、-322.94 HU, 表明随着病灶的浸润程度加重, 结节 CT 密度值也会上升。当由 AAH 发展为 IAC 的过程中, 结节边界逐渐由清晰变得不规则甚至消失, 分叶、毛刺、胸膜凹陷、血管穿行、空泡等征象出现的概率也逐渐增加。Cho J 等<sup>[19]</sup>分析 356 例 pGGN 得出较大结节是恶性肿瘤的独立预测因子; 柳蒋书等<sup>[20]</sup>分析 59 例 pGGN 得出肺腺癌浸润程度上的变化必然会影响 CT 值的变化; Hu HY 等<sup>[21]</sup>分析 112 例 pGGN 得出分叶征、毛刺征等与结节的恶性相关, 与本研究结果相一致。该结果说明利用肺结节的 CT 征象可以对其病理学浸润程度进行预测, 因此本研究构建了 pGGN 肺腺癌 CT 征象与其浸润程度的预测模型。6 种 Weka 算法构建的预测模型中, 以 Random Forest 算法构建的预测模型最适用于本研究。经过自身验证及交叉验证,

Random Forest 算法所建模型具备较好的分辨 AAH、AIS、MIA 和 IAC 的能力, 平均预测准确率为 78.07%, 为早期肺癌诊断提供一种非侵入性方式, 能在一定程度上减少患者的创伤, 避免过度干预。

根据美国胸科学会、欧洲呼吸学会等发布的关于肺癌分类的新标准, 将腺癌分为浸润前病变 (AAH 和 AIS)、MIA 和 IAC。多项回顾性研究表明, 在 pGGN 发展为 IAC 前实行手术切除, 其 5 年无瘤生存率接近 100%, 当发展为 IAC 后再行手术切除, 5 年生存率会显著降低, 约为 87%<sup>[22-24]</sup>。不同浸润程度的肺结节可采用不同的手术方式, AAH、AIS 和 MIA 通常行楔形切除或肺段切除即可达到目的, 而 IAC 则需行肺叶切除, 术后恢复时间及并发症会显著增加。因此若未能在 pGGN 发展为 IAC 前早期识别, 原本只需行亚肺叶手术切除的患者需被迫接受更大范围的手术切除, 增加了患者的创伤。预测模型以 pGGN 的 CT 征象为基础, 无创伤地早期区分出其浸润程度为临床医生选择合适的手术时机及手术方式提供可靠依据。预测模型可对肺结节的发展过程实施动态监控, 早期准确判断肺腺癌的浸润程度, 对保障肺癌患者生命安全具有十分重要的意义。

### 参 考 文 献

- [1] 中华医学会肿瘤临床诊疗指南(2022 版)[J]. 中华肿瘤杂志, 2022, 44(6):457-490.  
Chinese Medical Association guideline for clinical diagnosis and treatment of lung cancer (2022 edition) [J]. Chin J Oncol, 2022, 44 (6) : 457-490.
- [2] Pedersen JH, Saghiz Z, Wille MMW, et al. Ground-glass opacity lung nodules in the era of lung cancer CT screening: radiology, pathology, and clinical management[J]. Oncology (Williston Park), 2016, 30 (3):266-274.
- [3] 张 为, 李小虎, 杜丹丹, 等. 纯磨玻璃结节的 CT 特征及定量分析对肺腺癌病理分类的预测价值[J]. 中国医学影像学杂志, 2018, 26 (9):658-663.  
Zhang W, Li XH, Du DD, et al. CT features and quantitative analysis of pure ground-glass nodules in the pathological grading and prediction of lung adenocarcinoma[J]. Chin J Med Imaging, 2018, 26(9):658-663.
- [4] 张 宏, 丁必彪, 魏恒乐, 等. 高分辨率 CT 对肺纯磨玻璃结节侵袭性的预测价值[J]. 临床放射学杂志, 2019, 38(3):436-440.  
Zhang H, Ding BB, Wei HL, et al. Prediction value of the invasiveness in pure ground-glass opacity on lung high-resolution CT images[J]. J Clin Radiol, 2019, 38(3):436-440.

- [5] Gao JW, Rizzo S, Ma LH, et al. Pulmonary ground-glass opacity: computed tomography features, histopathology and molecular pathology [J]. *Transl Lung Cancer Res*, 2017, 6(1):68–75.
- [6] Gao YN, Zhou R, Lyu QW. Multiomics and machine learning in lung cancer prognosis[J]. *J Thorac Dis*, 2020, 12(8):4531–4535.
- [7] Sun T, Wang JJ, Li X, et al. Comparative evaluation of support vector machines for computer aided diagnosis of lung cancer in CT based on a multi-dimensional data set[J]. *Comput Methods Programs Biomed*, 2013, 111(2):519–524.
- [8] 李运, 陈克终, 隋锡朝, 等. 孤立性肺结节良恶性判断数学预测模型的建立[J]. 北京大学学报(医学版), 2011, 43(3):450–454.
- Li Y, Chen KZ, Sui XZ, et al. Establishment of a mathematical prediction model to evaluate the probability of malignancy or benign in patients with solitary pulmonary nodules[J]. *J Peking Univ Health Sci*, 2011, 43(3):450–454.
- [9] 丁宁, 范国华, 李梦娟. Nomogram 模型预测肺部混合磨玻璃结节微浸润腺癌和浸润性腺癌临床价值[J]. 临床放射学杂志, 2021, 40(6):1097–1101.
- Ding N, Fan GH, Li MJ. The clinical value of the nomogram model for predicting minimally invasive adenocarcinoma and invasive adenocarcinoma with mixed ground glass nodules[J]. *J Clin Radiol*, 2021, 40(6):1097–1101.
- [10] 王成武, 晏峻峰. 早期糖尿病风险预测模型的比较研究[J]. 智能计算机与应用, 2021, 11(1):64–68.
- Wang CW, Yan JF. Comparison of early diabetes risk prediction models [J]. *Intell Comput Appl*, 2021, 11(1):64–68.
- [11] Mansour EA. Prognosis of thoracic cancer using the bierman random committee machine learning[J]. *Int J Onl Eng*, 2021, 17 (12) : 135–150.
- Gould MK, Donington J, Lynch WR, et al. Evaluation of individuals with pulmonary nodules: when is it lung cancer? Diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines[J]. *Chest*, 2013, 143 (5 Suppl):e93S–e120S.
- [13] 肺结节诊治中国专家共识(2018年版)[J]. 中华结核和呼吸杂志, 2018, 41(10):763–771.
- China expert consensus on diagnosis and treatment of pulmonary nodules (2018 edition) [J]. *Chin J Tuberc Respir Dis*, 2018, 41 (10) : 763–771.
- [14] 苏大同, 王颖. 肺结节CT随访中的生长特性评估[J]. 中国肺癌杂志, 2017, 20(8):584–588.
- Su DT, Wang Y. Growth evaluation of pulmonary nodules on chest CT [J]. *Chin J Lung Cancer*, 2017, 20(8):584–588.
- [15] Miao YY, Zhang JY, Zou JW, et al. Correlation in histological subtypes with high resolution computed tomography signatures of early stage lung adenocarcinoma[J]. *Transl Lung Cancer Res*, 2017, 6 (1) : 14–22.
- [16] Yang J, Wang HL, Geng C, et al. Advances in intelligent diagnosis methods for pulmonary ground-glass opacity nodules[J]. *Biomed Eng Online*, 2018, 17(1):20.
- [17] Zheng B, Zhou XW, Chen JH, et al. A modified model for preoperatively predicting malignancy of solitary pulmonary nodules: an Asia cohort study[J]. *Ann Thorac Surg*, 2015, 100(1):288–294.
- [18] 张颖, 窦一峰. 基于 WEKA 的医学数据分类及糖尿病早期预测[J]. 医学信息, 2021, 34(6):32–35.
- Zhang Y, Dou YF. Medical data classification and early diabetes prediction based on WEKA[J]. *J Med Inf*, 2021, 34(6):32–35.
- [19] Cho J, Ko SJ, Kim SJ, et al. Surgical resection of nodular ground-glass opacities without percutaneous needle aspiration or biopsy[J]. *BMC Cancer*, 2014, 14:838.
- [20] 柳蒋书, 孟刚, 梁晓, 等. 59例纯磨玻璃结节影像与病理相关因素的定量分析[J]. 临床与实验病理学杂志, 2019, 35 (7) : 804–807.
- Liu JS, Meng G, Liang X, et al. Imaging-pathological related factors of pure ground glass nodules in lungs: quantitative analyses of 59 cases[J]. *Chin J Clin Exp Pathol*, 2019, 35(7):804–807.
- [21] Hu HY, Wang QG, Tang HM, et al. Multi-slice computed tomography characteristics of solitary pulmonary ground-glass nodules: differences between malignant and benign[J]. *Thorac Cancer*, 2016, 7 (1) : 80–87.
- [22] Wu HR, Liu CQ, Xu MQ, et al. A retrospective study of mean computed tomography value to predict the tumor invasiveness in AAH and clinical stage ia lung cancer[J]. *Zhongguo Fei Ai Za Zhi*, 2018, 21 (3):190–196.
- [23] Kim D, Kim HK, Kim SH, et al. Prognostic significance of histologic classification and tumor disappearance rate by computed tomography in lung cancer[J]. *J Thorac Dis*, 2018, 10(1):388–397.
- [24] Chen KN. The diagnosis and treatment of lung cancer presented as ground-glass nodule[J]. *Gen Thorac Cardiovasc Surg*, 2020, 68 (7) : 697–702.

(责任编辑:周一青)