Vol. 28, No. 3 May 2022

295

文章编号 1006-8147(2022)03-0295-06

论著

# 三维 ResNet 网络预测肺腺癌结节亚型的效能及其 稳定性

骆源1,徐启飞2,吕泽政1,蔡娜1,郭丽1

(1.天津医科大学医学技术学院医学图像处理教研室,天津 300203;2.山东省临沂市人民医院影像科,临沂 276000)

摘要 目的:探究 ResNet 模型对肺腺癌不同亚型结节的分类表现及稳定性。方法:回顾性收集 2014 年 2 月—2020 年 10 月期 间的 364 例肺腺癌结节 CT 影像数据,以 7:3 的比例分为训练集和内部测试集,将 2020 年 4 月到 2020 年 11 月的 58 例结节数据作为外部测试集。使用基于 ResNet 的三维卷积神经网络在训练集中进行训练以及调参,并使用内部测试集和外部测试集对模型的准确性及泛化性进行评估。使用随机中心移动和掩膜处理的方式分别以内部测试集和外部测试集为基础构造新的测试集,新数据集对模型进行测试验证模型的稳定性。结果:模型在内部测试集 AUC 为 0.949 1(95%CI:0.910 8~0.987 4),模型在随机中心移动以及掩膜处理之后的数据集的 AUC 值分别为 0.940 4 和 0.918 1,与其差异无统计学意义(P值分别为 0.425 3 和 0.239 3)。在外部测试集中模型 AUC 为 0.959 6(95%CI:0.901 2~1.000 0),在用于稳定性测试的随机中心移动以及掩膜处理之后的数据集中,模型所得 AUC 分别为 0.948 5 和 0.947 3,与其同样差异无统计学意义(均 P>0.05)。结论:ResNet 模型对肺腺癌结节亚型有优异的鉴别能力,并且具有一定稳定性。

关键词 人工智能:深度学习:卷积神经网络:肺腺癌:诊断

中图分类号 R734.2; R730.44

文献标志码 A

### Efficacy and stability of 3D ResNet for predicting nodule subtypes in lung adenocarcinoma

LUO Yuan<sup>1</sup>, XU Qi-fei<sup>2</sup>, LYU Ze-zheng<sup>1</sup>, CAI Na<sup>1</sup>, GUO Li<sup>1</sup>

(1.Department of Medical Image Processing, School of Medical Technology, Tianjin Medical University, Tianjin 300203, China; 2.Department of Imaging, Linyi People's Hospital, Linyi 276000, China)

**Abstract Objective**: To investigate the classification performance and model stability of ResNet models for different subtypes of nodules in lung adenocarcinoma. **Methods:** The CT image of 364 lung adenocarcinoma nodules collected retrospectively between February 2014 and October 2020 were divided into a training set and an internal test set in a ratio of 7:3, and data of 58 nodules from April 2020 to November 2020 were used as the external test set. The ResNet-based 3D convolutional neural network was trained and tuned in the training set, and the accuracy and generalization of the model was evaluated using both internal and external test sets. To verify the stability of the model, two new test sets were constructed using random center shifts and masking process in both internal and external test set, and the model was tested using the new test set. **Results:** The model obtained an AUC of 0.949 1(95% CI:0.910 8–0.987 4) on the internal test set, and the AUC values of the model were not statistically different (P=0.425 3 and 0.239 3, respectively) from those measured on the data sets with the random center shift and after the masking process. The model AUC in the external test set was 0.959 6(95% CI:0.901 2–1.0000). In the dataset after the random center shift and mask processing used for stability testing, the AUC obtained for the model (0.948 5 and 0.947 3, respectively) was again not statistically different from it (all P>0.05). **Conclusion:**ResNet model has excellent ability to discriminate subtypes of lung adenocarcinoma, and the model has considerable stability.

Key words artificial intelligence; deep learning; convolutional neural network; lung adenocarcinoma; diagnosis

在癌症相关致死率排名第一的肺癌中,肺腺癌 占比最大<sup>[1]</sup>。2021 年 5 月出版的《WHO 胸部肿瘤分 类,第 5 版》已将其目录中的原位腺癌(adenocarci-

基金项目 国家自然科学基金(8197070539); 天津市自然科学基金(18JCYBJC95600)

作者简介 骆源(1997-),男,硕士在读,研究方向:医学影像技术;通信作者:郭丽,E-mail:yxgl@tmu.edu.cn;蔡娜,E-mail:caina302@aliyun.com。

noma in situ, AIS)归类为前驱病变<sup>[2]</sup>,其他常见亚型如:微浸润性腺癌(minimally invasive adenocarcinoma, MIA)和浸润性腺癌(invasive adenocarcinoma, IA)仍为腺癌类别。并且先前的研究表明,在肺腺癌的各种亚型中 IA 的预后最差,但其他亚型都有着几乎 100%的生存率<sup>[3-4]</sup>。不同的预后决定了不同的临床诊疗策略。由此,术前如何鉴别 IA 与 non-IA (AIS+MIA)非常重要,有助于辅助并指导临床诊疗

策略的构建。

与从原始图像中人工计算提取特征并利用机器学习模型识别疾病特有特征模式的模型建立方式相比,深度学习技术能以数据驱动的方式自动从原始高维度图像中寻找低维特征来实现端到端的学习。前者因需要人工确定感兴趣区,不可避免的不稳定性将会被引入提取的特征之中,这在一定程度上影响了模型的性能及其泛化能力。相应的,深度学习可以减少人工标注的成本,提高数据分析的效率和客观性。在先前的研究中深度学习模型能较好完成肺腺癌的精准诊断任务,但其模型稳定性仍有待验证[5-6]。在本研究中,笔者在两个批次的数据中进行建模和测试,旨在探究二元分类的三维卷积神经网络对肺腺癌亚型影像数据的鉴别效能、泛化能力以及模型的稳定性。

## 1 材料与方法

1.1 材料 回顾性收集山东省临沂人民医院 2014 年2月-2020年10月和2020年4月-2020年11月 两个批次的经病理证实为肺腺癌患者的 CT 影像资 料。纳入标准:经病理证实为肺腺癌;有完整的 CT 影像资料;CT图像上表现为孤立的结节。排除标准: 图像质量差,有明显伪影;结节直径不在 3~30 mm 内。最终收集 311 例患者的 CT 图像数据,共计 422 枚结节。从第1批数据随机分层抽样70%的结节数 据(n=256,IA=123 例,non-IA=133 例)作为模型训 练数据集,剩余 30%的结节数据(n=108,IA=51 例, non-IA=57 例)作为内部测试集,将第2批次的数据 (n=58, IA=24 例, non-IA=34 例)作为外部测试集。 1.2 结节区域分割 使用 3D-slicer 软件(https://www. slicer.org/)进行结节边界的勾画,每个结节的勾画至 少有 2 名医师的参与。在相同的窗(1 400 HU)窗位 (-500 HU) 下,1 名医师对肺腺癌患者 CT 影像的 DICOM (medical digital imaging and communication) 格式数据逐层手动分割,在标注过程中沿结节与正 常组织交界处进行分割且尽量避免纳入气管和血

管,最终得到结节 3D-VOI(volume of interest)的掩膜。在结节区域勾画完毕之后,由另外一名高年资的医师进行核查,对于有异议的分割区域,在两名医师商议之后重新进行勾画。

素值, $\bar{x}$ 和 $\sigma_x$ 分别为数据块均值和标准差。

1.4 网络结构 使用了一个基于 ResNet<sup>□</sup>的 3D 卷 积神经网络,见图 1。为避免过拟合,不对原文章中 的基础模块进行堆叠,最后得到一个10层的网络 (ResNet10)。网络包含一个 64 个 7×7×7 卷积核的 3D 卷积层以及 4 个基础模块,每个基础模块中都有 两个卷积核大小为 3×3×3 的卷积层,并在两次卷积 之后添加捷径链接(shortcut connection),4个基础 模块中卷积核的个数分别为64、128、256、512。一个 核大小为 3×3×3 的最大汇聚层(max pooling layer) 在第1个基础模块之前,另外一个平均汇聚层 (average pooling layer)在第4个基础模块之后,输出 大小为 1×1×1。在全连接层(fully connected layer)将 卷积得到的特征映射到输出之后,使用 sigmoid 函数 输出网络的预测概率。网络中使用的激活函数都为 整流线性单元(rectified linear units, ReLUs)。为了避 免内部协变量偏移图的问题,每层卷积层之后都会 使用批量规范化(batch normalization)。

1.5 网络训练与测试 将 25%的训练集数据作为 验证数据集,在使用网格搜索以及早停(early stop-

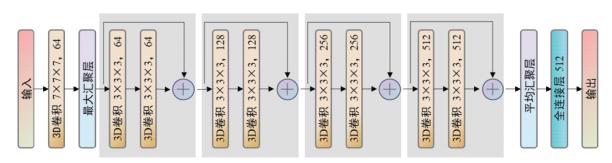


图 1 卷积神经网络结构

Fig 1 Structure of the convolutional neural network

ping)<sup>®</sup>技术搜寻验证损失最佳的批大小以及学习率 组合。使用基于梯度的随机优化器 Adam[10]来对网络 权重进行优化,以及周期性学习率(cyclical learning rate)叫技术动态修改每轮训练过程中的学习率,将 网络 10 次迭代作为 1 周期, 动态学习率的最大值 为基础学习率的 5 倍。在训练过程中对数据进行概 率为 50%的随机水平镜像翻转和随机 90°的旋转来 进行数据增广。选择对应验证损失最小的模型分别 对内部测试集以及外部测试集进行预测,并使用受 试者工作特性曲线(receiver operator characteristic curve, ROC curve)的曲线下面积(AUC)、敏感度、特 异度对模型性能进行评估。作为对比,原文献中的 ResNet-18的3D版本模型也按上述步骤,在相同的 数据集上进行了训练、调试以及测试。最后两个模 型所得参数均为:批大小为 16,基础学习率为 3.6× 10<sup>-4</sup>。最大训练轮数设置为 200。本研究基于 Pytorch<sup>[12]</sup> 深度学习框架,设备硬件配置为 Inteli9-10900X 中 央处理器、64 GB 内存、Nvidia GeForce RTX3080 图 形处理器,系统版本为 Ubuntu 18.04.6LTS。模型训 练及测试时间见表 1。

表 1 模型训练以及测试用时

Tab 1 Time elapse of model training and test

		训练月	用时	测试用时			
	周期	总耗时	平均每周期	样本	总耗	平均每样本	
	数	(s)	耗时(s)	数	时 $(s)$	耗时(s)	
ResNet-10	18	29.12	1.62	166	0.32	1.91×10 <sup>-5</sup>	
ResNet-18	25	33.59	1.34	166	0.33	2.00×10 <sup>-5</sup>	

#### 1.6 ResNet10 模型的稳定性测试

1.6.1 标注敏感测试 结节中心由掩膜边界计算得到。为测试深度学习模型是否对不稳定人工标注敏感,对测试集数据,对计算得到的结节中心随机增加一个各维度上的偏移,并根据此中心提取用于测试的数据块。比较模型对增加随机偏移前后的数据的预测效能。各维度的可移动量为4个像素单位。1.6.2 掩膜测试 从原始图像数据中提取固定大小

的数据块时部分数据块会纳入结节周围信息,仅对每个测试集使用数据对应结节掩膜将三维数据块内结节周围的信息剔除,并设置为-1 000 Hu 的 CT 值。不重新训练模型,使用处理后的两组测试集对已经在训练集中训练完成的模型进行测试。使用 ROC 及其 AUC、敏感度、特异度评估模型分类表现。用混淆矩阵展示模型在各测试集中的具体分类情况。

1.7 统计学处理 正态分布计量数据使用  $\bar{x}$ ±s 表示,非正态分布计量数据使用  $M(P_{25}, P_{75})$ 表示。使用 Delong 等[13]检验来比较模型稳定性测试中 ROC 的 差异。P<c0.05 为差异有统计学意义。

## 2 结果

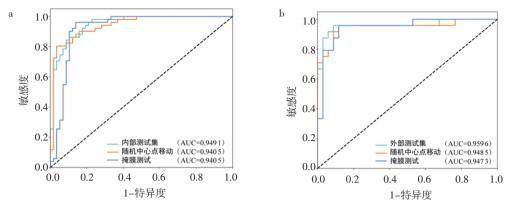
2.1 数据基本统计信息 第一批次收集的数据共364个结节。将每个结节作为一个样本,男性有232人,女性有132人。平均年龄为(55.1±10.8)岁。结节直径10.3(6.9,16.0)mm,。第二批次收集的数据共58个结节。将每个结节作为一个样本,男性有44人,女性有14人。年龄均值为(53.8±11.9)岁。结节直径10.2(6.9,15.2)mm。

2.2 模型测试结果 表 2 列出了训练完成的 ResNet10 模型在内部测试集和外部测试集以及相应的用于模型稳定性测试的数据集中的分类表现。在原始的内部测试集以及外部测试集中 AUC 值分别为0.949 1 和 0.959 6。ResNet18 模型在内部测试集和外部测试集中的分类性能与 ResNet10 相似,在内部测试集以及外部测试集中 AUC 值分别为0.945 3 和 0.957 1。在内部测试集测试的敏感度和特异度分别为92.16%和87.72%,在外部测试集中敏感度和特异度分别为79.17%和91.18%,较低于ResNet10模型所得结果。18 层结构的模型在本数据集上出现了过拟合的现象。内部测试集和外部测试集中模型分类的ROC 曲线见图 2,分类的混淆矩阵见图 3,在各种测试情景下,模型均能较好的对各类结节进行分类,体现在混淆矩阵较高的主对角线值。在随

表 2 不同测试集中 ResNet10 模型的分类性能

Tab 2 ResNet10 classification performance in different test dataset

		AUC[95% <i>CI</i> ]		敏感度[95%CI]		特异度 [95%CI]	统计值	P
内部测试集	原数据	0.949 1	[0.910 8~0.987 4]	88.24%	[76.62%~94.49%]	82.46% [70.63%~90.18%]		参考
	随机中心移动	0.940 5	[0.896 5~0.984 5]	82.35%	[69.75%~90.43%]	89.47% [78.88%~95.09%]	0.797 3	0.425 3
	掩膜测试	0.918 1	[0.858 1~0.978 2]	98.04%	[89.70%~99.65%]	66.67% [53.72%~77.51%]	1.176 7	0.239 3
外部测试集	原数据	0.959 6	[0.901 2~1.000 0]	87.50%	[69.00%~95.66%]	94.12% [80.91%~98.37%]		参考
	随机中心移动	0.948 5	[0.882 1~1.000 0]	79.17%	[59.53%~90.76%]	91.18% [77.04%~96.95%]	1.087 4	0.276 9
	掩膜测试	0.947 3	[0.888 3~1.000 0]	95.83%	[79.76%~99.26%]	52.94% [36.74%~68.55%]	0.313 6	0.753 9



注:a:模型在内部测试集中各数据集中分类的 ROC 曲线;b:模型在外部测试集中各数据集中分类的 ROC 曲线;ROC:受试者工作特征

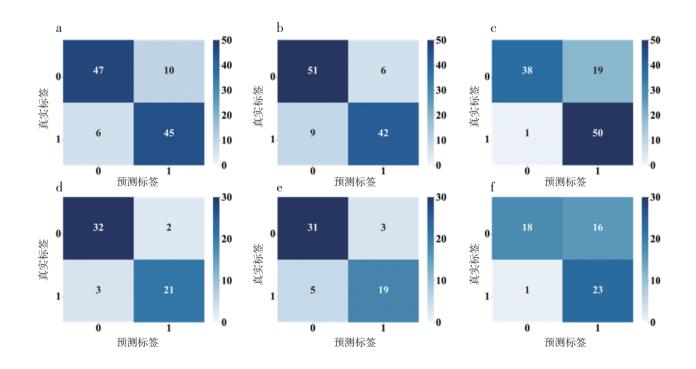
图 2 内部测试集和外部测试集中 ResNet10 模型分类表现的 ROC 曲线

Fig 2 ROC curve of ResNet10 classification performance in internal test set and external test set

机结节中心移动的模型稳定性测试中,模型在内部测试集和外部测试集的分类混淆矩阵和其在原数据中的混淆矩阵相近,模型 AUC 值与在原数据测得的相比差异均无统计学意义(P值分别为 0.425 3和 0.276 9)。在使用掩膜处理原数据的模型稳定性测试中,混淆矩阵右上角值增大,模型分类的假阳性样本增多。相应的,模型特异度在内部测试集和外部测试集中与在原数据测试的结果(分别为 82.46%和 94.12%)相比都有不同程度的下降,分别为 66.67%

和 52.94%, 统计比较的 P 值分别为 0.085 4 和 0.000 4。但在两测试集中模型 AUC 值与在原数据 测得的相比差异均无统计学意义(P 值分别为 0.239 3 和 0.753 9)。

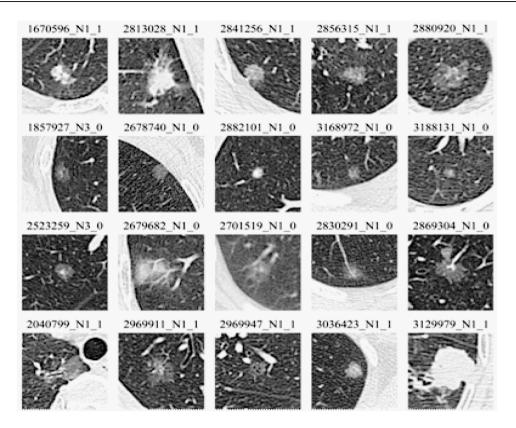
2.3 模型分类模式 部分结节中心轴位面的影像数据见图 4。对于第 3 行的 non-IA 结节图像,模型做出了误判,诊断为 IA 类结节。第 4 行假阴性样本的图像中多数结节与周围组织没有很强的对比,与阴性样本相似,导致了模型的误判。



注: $a \ b \ c$  分别为模型在内部测试集原数据、内部测试集随机中心点移动处理后的数据,内部测试集掩膜处理后数据上分类的混淆矩阵;  $d \ c \ f$  分别为模型在外部测试集原数据、外部测试集随机中心点移动处理后的数据,外部测试集掩膜处理后数据上分类的混淆矩阵

## 图 3 ResNet10 模型在各测试数据上分类的混淆矩阵

Fig 3 Confusion matrices of ResNet10 classification on each test data



注:部分结节的轴位中心层面的图像;第1行为IA类结节,第2行为non-IA类结节,第3行为假阳性结节图像,第4行为假阴性结节图像

#### 图 4 部分结节图像数据

Fig 4 Images data of some nodules

# 3 讨论

本研究试图通过使用 CT 影像数据训练三维 ResNet 网络对不同侵袭性的肺腺癌进行分类,并使 用内部测试集和外部测试集评估模型分类的准确 性和泛化性。通过随机中心点移动和掩膜处理的方 式构建新的测试集以评估神经网络的稳定性。

随着神经网络权重层的增多,模型过拟合的风险也会增大。仅通过与相同的调参以及训练步骤的 18 层三维 ResNet 网络进行对比,可以看出在本数据集中适当减少含有权重参数的层,能在不降低模型分类性能的情况下增加模型的泛化性能。且在使用早停方法训练模型的前提下,ResNet10模型仅在 18 个训练周期内就达到了 ResNet18 训练 25 个周期的效果,由此可见 ResNet10 模型更适合于当前数据集。

借助深度学习模型来对图像数据进行诊断,同样也能通过模型对于图像的行为分析来对模型进行诊断。在 2.3 中,展示了部分结节中心轴位面的影像数据。希望通过图像分析得到模型的分类规律以及模型的偏倚。从图像的角度,真阳性结节在图像上一般为较大的内部密度不均的区域,且在密度上与瘤周正常组织形成了强的对比。真阴性结节则一

般较小,内部密度较均匀,与周围组织对比不强。对于图 4 中第 3 行的 non-IA 结节图像,模型做出了误判,诊断为 IA 类结节。一种可能的解释是:结节区域大多都出现了高密度的噪声,原本均质的结节内部密度变得不一致,从而导致了模型的误判。图 4 中第 4 行假阴性的样本与周围组织没有很强的对比,这与大多数 IA 类结节和周围组织形成高对比的特征不一致,而编号为 3129979\_N1\_1 的图像则可能是由于结节本身较大,在数据截取过程中仅截取到了其内部部分区域,因此失去了与周围组织密度高对比度的特征,从而导致了模型的误判。

在前人的研究中<sup>[6.14-15]</sup>,深度学习在肺腺癌结节的分类任务中同样都有较优异的表现,证实了深度学习在肺腺癌侵袭性早期分类的应用前景。笔者的试验中,深度学习模型在训练之后对一定范围内随机中心点移动后提取的数据仍具有优异的分类表现,且与在原始数据得到的结果无统计学差异。说明深度学习模型对人工标注的不稳定性有一定的容受能力。使用经过掩膜处理之后的数据对模型进行测试,无论是在内部测试集还是在外部测试集中,其 AUC 与在原始数据得到的 AUC 同样无统计学差异。由此可以看出,在提取用于训练深度学习

模型的数据时,只需确定大致的中心点并按一定大小提取数据块,深度学习模型就能自动聚焦于差异区域并提取特征用于分类。

本研究仍然存在一定的局限性:深度学习模型是数据驱动的,422个肺结节的数据集对于深度学习模型仍然不足。结果 2.3 可看出,当数据中掺入噪声或者数据存在偏倚,模型将不能准确做出判断。需要采集更多的数据或使用更多的数据增广方式用以训练模型,使用在较大 3D 医学图像数据集中训练的神经网络的迁移学习也可能缓解这一问题。虽然体素重采样能部分解决获取的图像在采集过程中由于成像协议不同带来的数据差异,但仍需要尝试更加有效的方法来增加数据一致性。笔者虽然分两个批次收集了数据用以评估模型的泛化性,但仍需要在多中心的数据中进行测试。

在本研究中深度学习模型展示了肺腺癌结节 侵袭性分类的优异性能表现。同时它相较于手工提取 特征的建模方式更加高效且具有一定的稳定性。深 度学习模型在肺腺癌的诊断任务中是有应用前景 的,能辅助临床进行精确诊断并做出治疗决策。

#### 参考文献:

- MEZA R, MEERNIK C, JEON J, et al. Lung cancer incidence trends by gender, race and histology in the United States, 1973-2010 [J]. PLoS One, 2015, 10(3): e0121323.
- [2] 李媛,谢惠康,武春燕. WHO 胸部肿瘤分类(第 5 版)中肺肿瘤部分解读[J]. 中国癌症杂志,2021,31(7):574-580.
- [3] YOSHIZAWA A, MOTOI N, RIELY G J, et al. Impact of proposed IASLC/ATS/ERS classification of lung adenocarcinoma: prognostic subgroups and implications for further revision of staging based on analysis of 514 stage I cases [J]. Mod Pathol, 2011, 24(5):653–664.
- [4] YANAGAWA N, SHIONO S, ABIKO M, et al. New IASLC/ATS/ERS classification and invasive tumor size are predictive of disease recurrence in stage I lung adenocarcinoma[J]. J Thorac Oncol, 2013,8(5):

- 612-618
- [5] 陶雪敏,方瑞,吴重重,等. 深度学习模型对纯磨玻璃结节肺腺癌病理亚型的预测分析[J]. 中国医学科学院学报,2020,42(4): 477-484
- [6] XIA X,GONG J,HAO W,et al. Comparison and fusion of deep learning and radiomics features of ground-glass nodules to predict the invasiveness risk of stage-I lung adenocarcinomas in CT scan [J]. Front Oncol, 2020, 10:418.
- [7] HE K,ZHANG X,REN S,et al. Deep residual learning for image recognition; proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, F, 2016 [C].
- [8] IOFFE S,SZEGEDY C. Batch normalization; accelerating deep net work training by reducing internal covariate shift[J]. arXiv preprint arXiv:1502.03167,2015.
- [9] PRECHELT L. Early stopping -but when? [M]. Neural Networks: Tricks of the trade. Springer, 1998:55-69.
- [10] KINGMA D P, BA J. Adam: a method for stochastic optimization [J]. arXiv preprint arXiv:1412.6980,2014.
- [11] SMITH L N. Cyclical learning rates for training neural networks; proceedings of the 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), F, 2017 [C]. IEEE.
- [12] PASZKE A, GROSS S, MASSA F, et al. Pytorch; an imperative style, high-performance deep learning library[J]. Adv Neural Inf Process Syst, 2019, 32; 8026–8037.
- [13] DELONG E R, DELONG D M, CLARKE-PEARSON D L. Comparing the areas under two or more correlated receiver operating characteristic curves; a nonparametric approach[J]. Biometrics, 1988, 44(3): 837–845.
- [14] WANG X, LI Q, CAI J, et al. Predicting the invasiveness of lung adenocarcinomas appearing as ground-glass nodule on CT scan using multi-task learning and deep radiomics[J]. Transl Lung Cancer Res, 2020,9(4):1397-1406.
- [15] WANG S, WANG R, ZHANG S, et al. 3D convolutional neural network for differentiating pre-invasive lesions from invasive adenocarcinomas appearing as ground-glass nodules with diameters ≤3 cm using HRCT [J]. Quant Imag Med Surg, 2018, 8(5):491-499.

(2021-11-05 收稿)

#### (上接第294页)

- [22] Vincent J L, Backer D de, Wiedermann C J. Fluid management in sepsis: the potential beneficial effects of albumin[J]. J Crit Care, 2016, 35:161–167.
- [23] Mendes R d S, Pelosi P, Schultz M J, et al. Fluids in ARDS; more pros than cons[J]. Intensive Care Med Exp, 2020, 8 (Suppl 1): 32.
- [24] 李真玉,张璐,宗晓龙. 肺血管内皮钙黏蛋白在急性肺损伤血管 内皮屏障中的作用及机制[J]. 中华急诊医学杂志,2019,28(3):
- [25] 齐颖,李真玉,王斌,等. Wnt5a 与 Syndecan-1 在小鼠脓毒症致 急性肺损伤中的表达研究[J]. 中国急救医学,2017,1(37):43-48.
- [26] Delaney A P, Dan A, McCaffrey J, et al. The role of albumin as a resuscitation fluid for patients with sepsis: a systematic review and meta-analysis[J]. Crit Care Med, 2011, 39(2):386-391.
- [27] Vincent J L, Russell J A, Jacob M, et al. Albumin administration in the acutely ill: what is new and where next?[J]. Crit Care, 2014, 18 (4):231.
- [28] Pan J,Zhu Q,Zhang X,et al. Factors Influencing the prognosis of patients with intra-abdominal infection and its value in assessing prognosis[J]. Infect Drug Resist, 2021, 14:3425–3432.

(2021-10-26 收稿)