DOI:10.3969/j.issn.1005-202X.2019.04.011

医学影像物理

基于支持向量机的肺CT图像三维磨玻璃结节的提取和识别

徐亚楠[1],赵伟[2],李铭[2],石宏理[1]
1.首都医科大学生物医学工程学院,北京 100069;2.复旦大学附属华东医院CT室,上海 200040

【摘要】提出一种基于支持向量机的提取和识别肺CT图像三维磨玻璃结节(GGN)的算法。该算法首先根据肺实质三维图像的连通性,分割出肺实质区域,然后在肺实质区域内提取潜在GGN的孤立组织,并用三维形状特征和三维纹理特征建立识别结节的线性模型。依据临床医师标定的图像,通过支持向量机确定该线性模型参数。最后,采用该线性模型识别孤立组织中的结节。本研究采用139例临床医师标定的肺腺癌数据,其中100例作为训练集,39例作为测试集。测试结果表明,该算法可有效识别出肺CT图像的GGN,通过受试者工作特征曲线(ROC),得到ROC曲线下面积的值为0.937 2。

【关键词】肺;磨玻璃结节;支持向量机;CT;三维图像
【中图分类号】R318.13 【文献标志码】A 【文章编号】1005-202X(2019)04-0425-06

Support vector machine-based algorithm for the extraction and recognition of ground glass nodules in lung CT image

XU Ya'nan[1], ZHAO Wei[2], LI Ming[2], SHI Hongli[1]
1. School of Biomedical Engineering, Capital Medical University, Beijing 100069, China; 2. CT Room, Huadong Hospital Affiliated to Fudan University, Shanghai 200040, China

Abstract: A new approach based on support vector machine is proposed to extract and recognize the ground glass nodules in three-dimensional (3D) lung computed tomography (CT) image. Firstly, the lung parenchyma regions are segmented according to the 3D connectivity of the lung parenchyma. Then the isolated tissues which may be ground glass nodules are extracted from the lung parenchyma region. After the 3D shape features and 3D texture features are calculated, a linear model is established using these features to recognize the ground glass nodules. The coefficients of the model are determined by support vector machine based on the CT images labeled by clinicians. Finally, the linear model is used to recognize ground glass nodules from the isolated tissues. Among the labeled lung CT images of 139 patients in this study, the images of 100 patients were used as training set and the others as test set. The test results show that the proposed approach can effectively recognize the ground glass nodules in lung CT image. The area under receiver operating characteristic curve reaches 0.937 2.

Keywords: lung; ground glass nodule; support vector machine; computed tomography; three-dimensional image

前 言

国家癌症中心发布的癌症报告显示2015年肺癌的发病人数为73.3万人,死亡人数接近61万,是致死率最高的恶性肿瘤[1-2]。肺癌主要分为鳞形细胞癌、未分化癌、腺癌和肺泡细胞癌4种。其中,肺腺癌占原发肿瘤的40%,早期无特殊症状,主要分为浸润前癌变和浸润性癌变[3]。无论是浸润前癌变或浸润性癌变,其CT图像上均可表现为磨玻璃结节(Ground Glass Nodule, GGN),其形状主要呈类圆形和不规则型,早期特点为分布范围比较广范,直径较小,易漏诊[4-6]。在临床诊断中如何对早期GGN的鉴别及诊断,仍是亟待解决的难题。

GGN的计算机影像学辅助诊断(Computer Aided Diagnosis, CAD)一般可分为两种[7-9]:一种是对于是否存在GGN以及GGN的位置判别,其主要任务是对肺部的各种组织进行分割、提取,判断是否存在GGN;另一种是对于GGN是否浸润进行判别。本文

【收稿日期】2018-11-19
【基金项目】北京自然科学基金(7142022)
【作者简介】徐亚楠,硕士研究生,研究方向:医学图像处理,E-mail: xyn19940215@126.com;赵伟,博士在读,研究方向:影像数据挖掘,E-mail: zbsasd@163.com
【通信作者】李铭,副主任医师,研究方向:胸部影像诊断和医学人工智能,E-mail: minli77@163.com;石宏理,副教授,研究方向:医学图像处理,E-mail:shl@ccmu.edu.cn

主要针对GGN的提取与判别展开研究。

自动化分割肺结节，精确划分感兴趣区域一直是CAD研究的热点之一。国内外很多学者提出许多新的算法来分割肺结节。Farag等[10]利用水平集算法对肺结节进行分割，其主要是通过肺结节形状模型将分割框架与图像强度统计信息融合，该方法的优点是不依赖于结节类型或位置。Shakir等[11]提出一种半自动系统分割结节，该算法是基于水平集中平均强度阈值的模型来进行结节分割，同时通过自适应技术来估计平均强度。Santos等[12]用支持向量机算法区分结节与非结节，该算法采用高斯混合模型并结合了熵测量等纹理特征。Liu等[13]提出一种自适应模糊C均值算法。该算法根据中心像素和相邻像素之间的灰度相似性和空间相似性来计算隶属度值，并通过使用从训练样本中学习的少量先验知识来构建聚类和类别之间的概率关系矩阵。基于该矩阵，实现对未标记肺CT图像的弱监督肺结节分割。Wang等[14]和Qi等[15]利用卷积神经网络算法，分别提出中心聚焦卷积神经网络和三维卷积神经网络，对结节进行分割。

结合目前已有的肺结节分割方法，本文提出一种基于支持向量机的三维GGN自动提取和识别算法。该算法首先根据三维连通域的特性分割肺实质，然后在肺实质区域内提取可能为GGN的孤立组织，这些孤立组织直径一般不超过4 cm，可能为GGN、钙化、血管末端等组织。为了从孤立组织中识别出GGN，本方法中选取了28个三维形状特征和纹理特征参数，建立一个线性判别模型。本文共选取139例GGN数据，将数据分为训练集100例，测试集39例。其中训练集中的孤立组织已有医生标记，可直接得到GGN。在研究中，计算其形状特征和纹理特征参数作为线性模型的输入参数，训练线性模型，用支持向量机得到确定模型参数。对于测试集中的孤立组织，计算特征参数，再根据线性模型判别该组织是否为GGN。为了保证方法的实用性，本文对肺实质提取、孤立组织及其特征的提取、基于支持向量机的GGN识别和量化评估分析这4个方面对该方法进行测试。测试结果表明，该算法可以比较理想地识别出GGN。

1 算法流程

本文方法是基于三维体素数据为数字运算单元对GGN的识别和提取。算法流程主要分为3部分：首先，先对肺实质进行分割提取，分割出包含可能为GGN的孤立组织；其次，计算其三维形状特征和纹理特征，建立线性识别模型；最后，利用支持向量机确定模型参数，区分GGN和非GGN。具体方法如下。

1.1 肺实质分割

在肺CT图像中，GGN只存在于肺实质区域。为了方便提取GGN，首先分割出肺实质。文中采用的肺实质分割方法主要包括以下4个步骤：(1)归一化与二值化：首先将图像进行归一化处理，将归一化的数据根据大津阈值算法计算阈值，根据阈值进行二值化处理，得到二值化的胸部CT图像；(2)三维连通区域：根据肺部CT图像三维方向26邻域来计算连通区域，分别提取出肺实质与背景等部分；(3)肺实质提取：为了将肺实质从全部的连通区域中提取出来，本文根据肺实质区域与其他组织的局部差异性，将肺实质与其他组织进行区分，得到肺实质二值化图像；(4)掩模运算：将原始图像与计算得到的二值图像进行类似于“乘积”的掩模运算，得到包含灰度信息的肺实质图像。

1.2 提取孤立组织及其特征

孤立组织提取：对于分割肺实质后的CT图像数据，首先根据肺实质图像进行阈值计算，根据阈值将图像变为二值化图像。然后将图像做膨胀腐蚀运算，再将腐蚀后的图像根据组织的连通性，将三维方向上连通的组织分别提取，得到肺实质中孤立组织。

三维形状特征和纹理特征：将所得到的孤立性组织进行三维形状特征和纹理特征提取。其形状特征主要包括：体积、直径、区域与总边框中体素的比值、椭圆主轴长度的第二中心距、主轴长度特征值、凸体积、凸面积、曲面面积。纹理特征主要包括：能量、惯量、逆差距、熵、相关系数[16]。在计算纹理特征时，由于CT图像是三维数据，计算纹理特征采用了三维灰度共生矩阵[17]，方向(θ, φ)包含了13个不同的方向，其中θ是XY平面与正X方向之间的角度间隔，分别依次取0°、45°、90°、135°；φ是XY平面与正Z方向之间的角度间隔，分别依次取0°、45°、90°、135°，所有方向的距离值均取为1。最后，采用这些特征参数构造线性判别模型。

为了训练该判别模型，需根据医生标记的图像确定训练集。将肺实质中提取的孤立组织，根据医生所给的结节位置标记，与全部孤立组织的位置进行匹配，可得到孤立组织中的GGN，再将GGN还原成未腐蚀的实际大小，为防止图像细节丢失，将算法提取的GGN与医生标记的GGN进行结合(取并集)，得到GGN图像。

1.3 支持向量机

对于线性分类问题，支持向量机能够在特征空间中寻找一个最优超平面将数据分类，从而区分两类数据。如图1所示，H1和H2是两类数据的边缘分

类面，它们之间的距离就是两类之间的间隔，虽然能够将两类点正确分开的超平面有很多，但H为最优超平面。其中，位于H1和H2上的数据点是接近超平面的支持向量，支持向量机的最终目的即寻求一个最优超平面使两类数据之间的间隔最大，同时将数据的分类能力达到最佳[18]，即：

$$g(\boldsymbol{x})=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+\boldsymbol{w}_0=0 \tag{1}$$

其中，$\boldsymbol{x}$ 表示特征向量，$\boldsymbol{w}$ 、$\boldsymbol{w}_0$ 表示超平面参数，$\boldsymbol{w}^{\mathrm{T}}$ 表示参数 $\boldsymbol{w}$ 的转置。

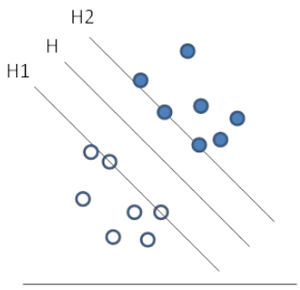


图1 线性可分情况下的最优超平面

Fig.1 Optimal hyperplane in condition of linear separability

对于一个二分类问题，为了不偏袒任何一类，选择最优超平面时应选取在每一个方向上两类数据中各自最近的点距离相同[19]，即：

$$\min J(\boldsymbol{w},\boldsymbol{w}_0)=\frac{1}{2}\|\boldsymbol{w}\|^2 \tag{2}$$

$$\text{st.}\, y_i(\boldsymbol{w}^{\mathrm{T}}x_i+\boldsymbol{w}_0)\geqslant 1,\ i=1,2,\cdots,N \tag{3}$$

其中，y_i 表示相对应类的表示器，一般用±1表示。这是一个满足一系列线性不等式条件的非线性最优化任务。由于 $J(\boldsymbol{w},\boldsymbol{w}_0)$ 是一个二次型函数，有唯一的极小点，利用拉个朗日优化方法将最优分类问题转化为其对偶形式[20]：

$$\max_{\lambda}\left(\sum_{i=1}^{N}\lambda_i-\frac{1}{2}\sum_{i,j}\lambda_i\lambda_j y_i y_j \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j\right),\ \lambda_i\geqslant 0 \tag{4}$$

$$\text{st.}\sum_{i=1}^{N}\lambda_i y_i=0,\ i=1,2,\cdots,N;\ \lambda_i\geqslant 0 \tag{5}$$

其中，λ_i 是拉个朗日乘子，$\boldsymbol{x}_i$、$\boldsymbol{x}_j$ 为核函数，$\boldsymbol{x}_i^{\mathrm{T}}$ 表示为核函数 $\boldsymbol{x}_i$ 的转置。支持向量机的基本算法可分为块算法、分解算法和序列最小优化算法。

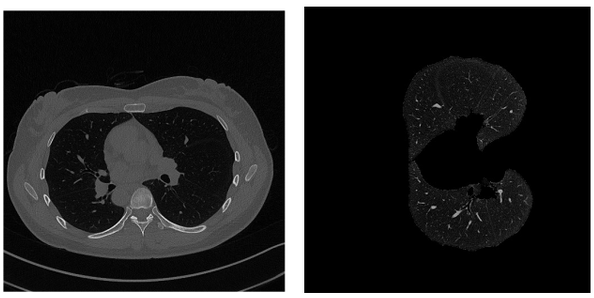
根据上述原理，本文方法中 y_i 代表结节和非结节两类类别，用±1表示；$\boldsymbol{x}_i$ 代表28个三维形状特征和纹理特征参数。采用序列最小优化算法计算线性模型参数。

2 仿真实验

本文涉及软件及硬件环境。软件环境：Windows 7操作系统，MATLAB R2017b。硬件环境：Intel(R) Xeon (R) E5-2603-1.6 GHz(CPU)，8.0 Gbyte(内存)，2.0 Gbyte显存(显卡)，500 Gbyte(硬盘)。本文采用由华东医院CT室医师标记的139例临床肺腺癌患者的CT图像作为实验样本，大小均为512×512。

2.1 肺实质分割

首先，对三维CT图像数据进行归一化和二值化处理，并根据肺CT图像不同区域的连通性不同，即肺实质区域与其他组织局部性差异，将肺实质提取，最后对提取的二值化图像进行掩模填充，获取肺实质的灰度信息。其分割结果如图2所示，图中选取的图像为139例数据中的其中一例，该CT图像一共有254层，本图为其中一层图像，其中图2a为肺CT原图，图2b为肺实质分割结果，该结果显示该方法能够分割出肺实质区域并保留图像细节。



a：肺CT原图　　b：肺实质分割结果

图2 肺实质分割结果

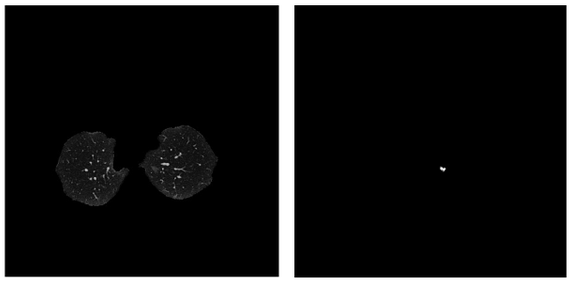
Fig.2 Segmentation of lung parenchymal

2.2 提取孤立组织及其特征

构造线性判别模型需提取肺实质中孤立组织并计算特征参数。首先，根据肺实质图像进行阈值计算，根据阈值对图像做二值化计算。然后，将图像做膨胀腐蚀运算，再将腐蚀后的图像根据组织的连通性，将三维方向上连通的组织提取，得到肺实质中孤立组织。最后，计算三维形状特征和纹理特征，建立线性判别模型。

为训练线性模型，根据CT室医师提供的GGN标记，将训练集中的GGN从孤立组织中提取。GGN标记由CT室医师使用3-D slice软件完成。软件由于是手动勾画GGN结节范围，可能会导致漏选体素点的情况出现。为了防止图像细节丢失，我们根据医师给出的GGN位置匹配出孤立组织中的GGN，再将其与医师标记的GGN结合(取并集)，构成训练集。图3为100例训练集中的某一例，图3a为带有GGN的肺实质图像，图3b为融合医师标记和自动提取的GGN区域的二值图像(一层)。GGN区域的实际大小为11×11×10，其每层结果如图4所示。该GGN的三维

形状特征和三维纹理特征根据这些图计算得到，最后构造线性判别模型。



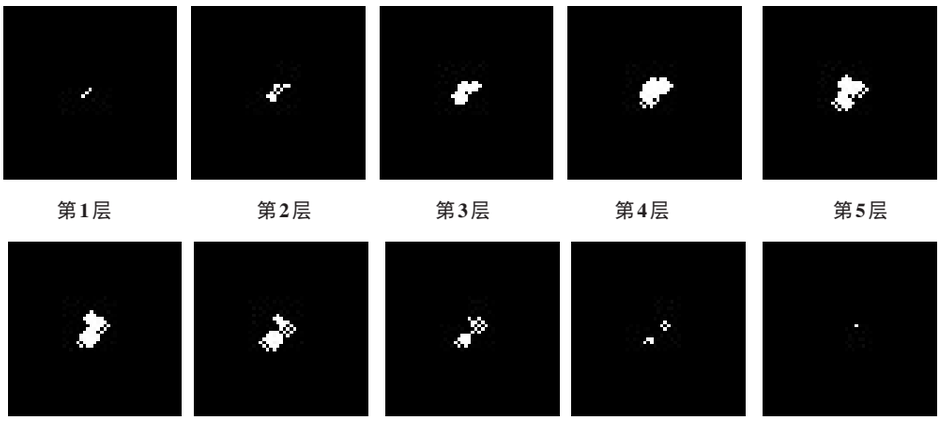
a：肺实质图像　　b：融合医师标记和自动提取的GGN

图3 训练集中GGN提取

Fig.3 Extraction of ground glass nodules (GGN) from training set

2.3 基于支持向量机的GGN识别

为了确定线性判别模型的参数，采用支持向量机训练该模型。首先，将得到的100例训练集中GGN数据的形状特征和纹理特征作为输入参数，同时在训练集中随机挑选100个其它孤立组织（非GGN）计算形状特征和纹理特征，构成训练集的输入特征参数。然后，用支持向量机训练线性模型，计算出模型参数。最后，将39例CT图像测试集数据中的孤立组织所计算的形状特征和纹理特征参数输入到线性模型中，对孤立组织进行判别，识别出孤立组织中的结节。判别与提取的结果如图5和图6所示。图5为39例测试集中某一患者CT影像的识别结果，该患者为女性，36岁，不吸烟，GGN直径为0.8 cm。



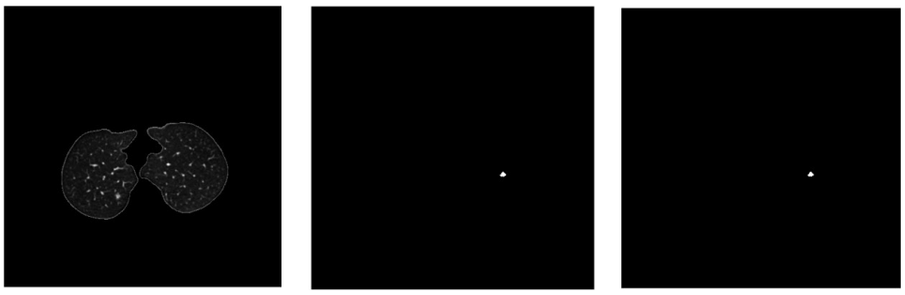
第1层　第2层　第3层　第4层　第5层

第6层　第7层　第8层　第9层　第10层

图4 训练集GGN每层结果

Fig.4 Per layer of GGN in training Set

其中图5a表示为CT图像中某一层带有GGN肺实质图像；图5b为该层医师根据观察标记的GGN图像；图5c为该层使用线性模型判别GGN的结果。该测试CT图像共192层，其中医师标记GGN在图像的149至157层，大小为13×11×9；测试结果提取的GGN在图像的150至157层，大小为9×12×8。其每层结果如图6所示，图6a显示医师标记的每层结果，图6b显示判别模型识别出GGN的每层结果，由图可得，本文模型可识别和提取GGN，其结果与医师标记非常类似。



a：肺实质图像　　b：医生标记的GGN　　c：GGN识别结果

图5 GGN的提取和识别结果

Fig.5 Extraction and recognition of GGN

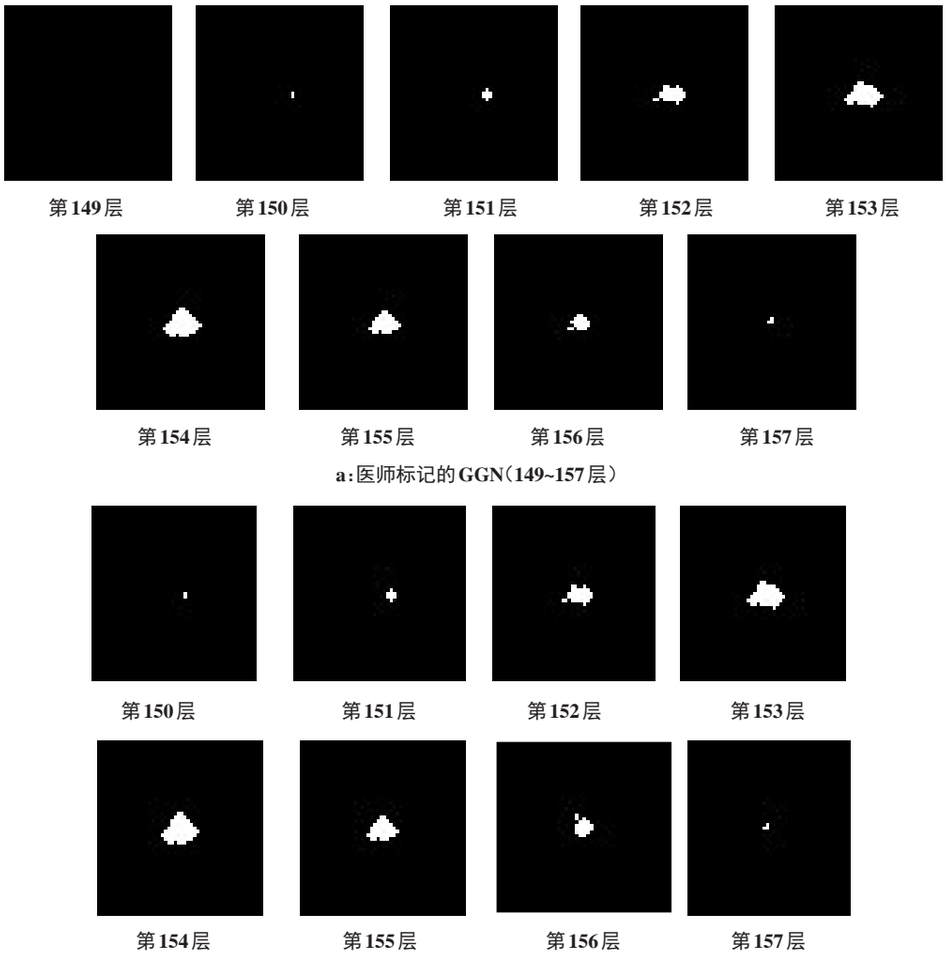


图6 提取和识别每层GGN与医师标记对比

Fig.6 Extraction and recognition of GGN compared to physician's markers

2.4 量化评估和分析

为了量化评估本算法的有效性,所以用ROC曲线对测试集结果进行评估,其ROC曲线如图7所示。根据测试集结果的ROC曲线所得的AUC的值为0.937 2。由于在提取孤立组织的过程中采用了腐蚀运算,这一过程虽然可以将与血管相连的GGN提取,但导致末端血管可能被判定为孤立组织,导致假阳性的出现。评估结果表明,本算法可比较有效地识别和提取GGN。

3 讨论与总结

本文所提出的算法,模拟了医师诊断GGN的过程,基本实现了自动化分割GGN,首先该算法可提取出肺实质区域,剔除其他组织区域,并利用GGN的孤立特性与其形状和纹理特征提取出GGN,并且本算法的实现是基于三维CT图像的GGN分割,并保留了GGN的三维结构特征,为临床提供诊断依据。

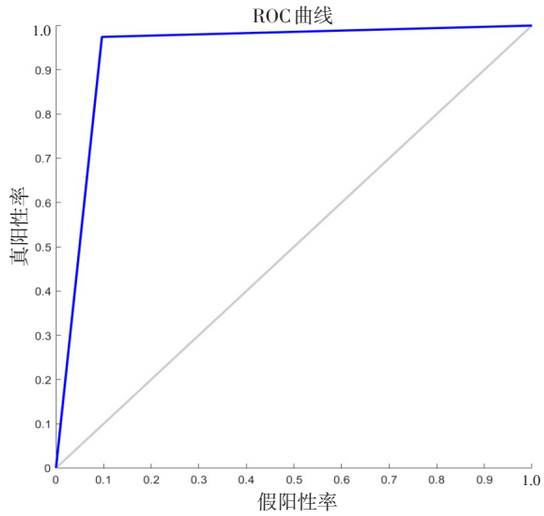


图7 测试结果的ROC曲线

Fig.7 Receiver operating characteristic curve of test set

本文中的算法自动化程度高,基本不需要人工操作,即可提取和识别肺结节;其次,本文所提到的算法

是在三维的基础上进行GGN提取，保留了GGN的三维特征；最后，本算法计算速度快，基本计算机配置即可实现本算法运算。但是，本算法还存在一些问题需要进一步解决，仍存在对一些孤立组织的误判，导致假阳性，这主要源于自动化分割肺结节算法中的误差，我们将继续改进算法，以实现临床中的应用。

【参考文献】

[1] CHEN W, ZHENG R, BAADE P D, et al. Cancer statistics in China, 2015[J]. CA Cancer J Clin, 2016, 66(2): 115-132.

[2] SIEGEL R L, MILLER K D, JEMAL A. Colorectal cancer statistics, 2017[J]. CA Cancer J Clin, 2017, 67(1): 7.

[3] TRAVIS W D, BRAMBILLA E, NOGUCHI M, et al. Prognostic significance of the International Association for the Study of Lung Cancer/American Thoracic Society/European Respiratory Society classification of stage I lung adenocarcinoma: a retrospective study based on analysis of 110 Chinese patients[J]. J Thorac Oncol, 2011, 6(2): 244-285.

[4] WU F, TIAN S P, JIN X, et al. CT and histopathologic characteristics of lung adenocarcinoma with pure ground-glass nodules 10 mm or less in diameter[J]. Eur Radiol, 2017, 27(10): 4037-4043.

[5] BANKIER A A, MACMAHON H, GOO J M, et al. Recommendations for measuring pulmonary nodules at CT: a statement, from the fleischner society[J]. Radiology, 2017, 285(2): 584-600.

[6] GILLIES R J, KINAHAN P E, HRICAK H. Radiomics: images are more than pictures, they are data[J]. Radiology, 2016, 278(2): 563-577.

[7] SLUIMER I C, VAN WAES P F, VIERGEVER M A, et al. Computer-aided diagnosis in high resolution CT of the lungs[J]. Med Phys, 2003, 30(12): 3081-3090.

[8] XU Y, VAN BEEK E J, HWANJO Y, et al. Computer-aided classification of interstitial lung diseases *via* MDCT: 3D adaptive multiple feature method (3D AMFM)[J]. Acad Radiol, 2006, 13(8): 969-978.

[9] ANTHIMOPOULOS M, CHRISTODOULIDIS S, EBNER L, et al. Lung pattern classification for interstitial lung diseases using a deep convolutional neural network[J]. IEEE Trans Med Imaging, 2016, 35 (5): 1207-1216.

[10] FARAG A A, MUNIM H E, GRAHAM J H, et al. A novel approach for lung nodules segmentation in chest CT using level sets[J]. IEEE Trans Med Imaging, 2013, 22(12): 5202-5213.

[11] SHAKIR H, RASOOL KHAN T M, RASHEED H. 3-D segmentation of lung nodules using hybrid level sets[J]. Comput Biol Med, 2018, 96: 214-226.

[12] SANTOS A M, FILHO A O, SILVA A C, et al. Automatic detection of small lung nodules in 3DCT data using Gaussian mixture models, Tsallis entropy and SVM[J]. Eng Appl Artif Intell, 2014, 36(C): 27-39.

[13] LIU H, GENG F, GUO Q, et al. A fast weak-supervised pulmonary nodule segmentation method based on modified self-adaptive FCM algorithm[J]. Soft Comput, 2017, 22(3): 1-13.

[14] WANG S, ZHOU M, LIU Z, et al. Central focused convolutional neural networks: developing a data-driven model for lung nodule segmentation[J]. Med Image Anal, 2017, 40: 172-183.

[15] QI D, HAO C, YU L, et al. Multilevel contextual 3-D CNNs for false positive reduction in pulmonary nodule detection[J]. IEEE Trans Biomed Eng, 2016, 64(7): 1558-1567.

[16] 罗述谦, 周果宏. 医学图像处理与分析[M]. 北京: 科学出版社, 2003: 33-35.
LUO S Q, ZHOU G H. Medical image processing and analysis[M]. Beijing: Science Press, 2003: 33-35.

[17] BOONSIRI O, WASHIYA K, AOKI K, et al. 3D gray level co-occurrence matrix based classification of favor benign and borderline types in follicular neoplasm images[J]. J Biosciences, 2016, 4(3): 51-56.

[18] DUDA R O, HART P E, STORK D G. 模式分类[M]. 北京: 机械工业出版社, 2004: 211-215.
DUDA R O, HART P E, STORK D G. Pattern classification[M]. Beijing: China Machine Press, 2004: 211-215.

[19] THEODORIDIS S, KOUTROUMBAS K. 模式识别[M]. 第2版. 北京: 电子工业出版社, 2004: 81-96.
THEODORIDIS S, KOUTROUMBAS K. Pattern recognition[M]. 2nd ed. Beijing: Publishing House of Electronics Industry, 2004: 81-96.

[20] 张松兰. 支持向量机的算法及应用综述[J]. 江苏理工学院学报, 2016, 22(2): 14-17.
ZHANG S L. A survey of algorithms and applications of support vector machines[J]. Journal of Jiangsu University of Technology, 2016, 22 (2): 14-17.

（编辑：陈丽霞）