

基于深度学习模型对乳腺 X 线摄影中乳房密度分类的初步研究

马明明, 姜原, 刘义, 王祥鹏, 黄嘉豪, 张晓东, 秦乃姗, 王霄英

【摘要】 目的:探索基于深度学习对乳腺 X 线密度(MD)进行分类的可行性。**方法:**回顾性分析 2018 年 9 月—12 月本院行双侧乳腺 X 线摄影检查的 305 例连续资料,共 1220 个图像数据。由两位影像诊断医师挑选其中合格的数据分别进行乳腺密度的分类训练。由两位医师根据第 5 版乳腺影像报告和数据库(BI-RADS)中 X 线 MD 分型(a、b、c、d)标准对 1220 个图像进行分类。将上述随机分训练集(train set, n=966)、调优集(validation set, n=128)和测试集(test set, n=126),采用 AutoVGG 网络训练分类预测模型,以测试集的混淆矩阵评价模型的效能。在模型预测的四分类基础上,将 MD 进行二分类,即:非致密组(a 型和 b 型)和致密组(c 型和 d 型),评价模型二分类预测的效能。**结果:**测试集的 126 个数据的预测符合率为 0.83, a、b、c、d 各型预测符合率分别为 0.88、0.78、0.80 和 0.76。将 MD 分为致密组和非致密组时,非致密组精确度(precision)为 0.90,召回率(recall)为 0.74, F1 值为 0.81;致密组精确度为 0.86,召回率为 0.84, F1 值为 0.85。**结论:**采用深度学习方法进行乳腺 X 线摄影 MD 二分类是可行的,有望在临床工作中应用。

【关键词】 深度学习; 乳腺 X 线摄影; 乳腺密度; 结构化报告

【中图分类号】 R816.42; R-05; **【文献标识码】** A **【文章编号】** 1000-0313(2021)11-1391-05

DOI:10.13609/j.cnki.1000-0313.2021.11.011

开放科学(资源服务)标识码(OSID):



Classification of breast density in mammography based on deep learning model: a preliminary study MA Ming-ming, JIANG Yuan, LIU Yi, et al. Department of Radiology, Peking University First Hospital, Beijing 100034, China

【Abstract】 Objective: To explore the feasibility of deep learning in mammary gland X-ray density (MD) classification. **Methods:** A total of 305 continuous patients who underwent bilateral mammography in our hospital from September to December 2018 were analyzed retrospectively, including 1220 images. Two radiologists selected the qualified data for breast density classification training. The 1220 qualified images were classified into four types (a, b, c, d) based the MD classification standard in 5th edition of Breast Imaging Reporting and Data System (BI-RADS). These image data were randomly divided into train set (80%), validation set (10%) and test set (10%), and the AutoVGG network was used to train the classification prediction model. The confusion matrix of the test set was used to evaluate the performance of the classification model. Based on the four categories, MD was classified into two categories: non-dense group (type a and type b) and dense group (type c and type d), and the effectiveness of the binary prediction model was evaluated. **Results:** The accuracy (ACC Macro) in the 126-test data was 0.83, and the accuracy of the four classifications of MD (a, b, c, d) was 0.88, 0.78, 0.80 and 0.76, respectively. When MD was divided into dense group and non-dense group, the precision, recall rate, F1 value of the non-dense group were 0.90, 0.74 and 0.81 respectively; the precision, recall rate, F1 value of the dense group were 0.86, 0.84 and 0.85 respectively. **Conclusion:** It is feasible to use deep learning method to classify MD types in mammography, and it is expected to be applied in clinical work.

【Key words】 Deep learning; Mammography; Mammographic density; Structured reporting

作者单位: 100034 北京: 北京大学第一医院医学影像科 (马明明、姜原、刘义、张晓东、秦乃姗、王霄英); 100011 北京, 北京赛迈特锐医学科技有限公司 (王祥鹏、黄嘉豪)

作者简介: 马明明 (1991—), 女, 河南周口人, 博士研究生, 主要从事乳腺影像与人工智能相关研究。

通讯作者: 王霄英, E-mail: wangxiaoying@bjmu.edu.cn

乳腺 X 线密度 (mammary density, MD) 反映了乳腺中纤维腺体的相对数量, 是乳腺癌的独立危险因素^[1-2]。根据第 5 版乳腺影像报告和数据库系统

(Breast Imaging Reporting and Data System, BI-RADS)MD 分为 a、b、c、d 四型,但在临床应用时,读片专家间的一致性不高($Kappa=0.57,95\%CI\ 0.53\sim0.61$)^[3],不同级别医师之间的一致性从较差到中等($Kappa$ 值 $0.396\sim0.604$)^[4]。为了实现 MD 分类的临床意义和评价一致性,常有研究将乳腺密度进行二分类,即将 a 型和 b 型分为非致密组;将 c 型和 d 型分为致密组^[5-6]。致密组乳腺的女性患乳腺癌风险高于非致密组,患乳腺癌风险是非致密组乳腺女性的 4~6 倍^[7],因此在乳腺 X 线报告中对致密组和非致密组的分类是很重要的。近年来很多研究者使用算法测量 MD,试图对 MD 自动判断,但结果仍有提升的空间^[8]。本研究旨在训练深度学习模型,期望能用来可靠地评估 MD 密度分组,并将结果自动接入到结构化报告系统中,实现在临床工作中自动报告乳腺密度分组。

材料与方法

本研究获得了伦理审查委员会的批准(批件号: [2017(1382)]),按照本单位人工智能(AI)模型训练规范执行研究方案。

1. 用例定义

根据本单位 AI 训练管理方法,首先定义研发乳腺钼靶 X 线检查投照体位和腺体密度分类模型的用例。包括:模型的 ID、临床问题、场景描述、模型在实际工作中的调用流程、模型输入输出数据结构等。AI 模型返回结果定义为:腺体密度分型“BIRADS-a、b、c、d”。MD 分类模型预测结果返回到结构化报告中的“整体评估”模块(图 1)。

2. 研究队列建立

本研究回顾性搜集本院 PACS 系统中 2018 年 6 月 1 日—2018 年 12 月 30 日进行乳腺 X 线摄影检查的连续患者影像报告。

MD 模型训练入组标准:①常规筛查或乳腺超声检查异常而进行的诊断性乳腺 X 线检查;②行双乳的 MLO 位及 CC 位的乳腺 X 线检查。排除标准:①乳腺癌根治切除术后;②假体植入术后;③图像显示腺体不完整。根据 MD 模型训练纳入和排除标准筛选合格病例 305 例,共 1220 个数据(图 2),均为女性,年龄 29~91 岁,平均年龄 53.4 ± 11.9 岁。使用 Python 的 dicom2nii 包将 DICOM 格式转换为 NIFTI 格式。

由两位影像科医师(乳腺影像诊断工作经验分别

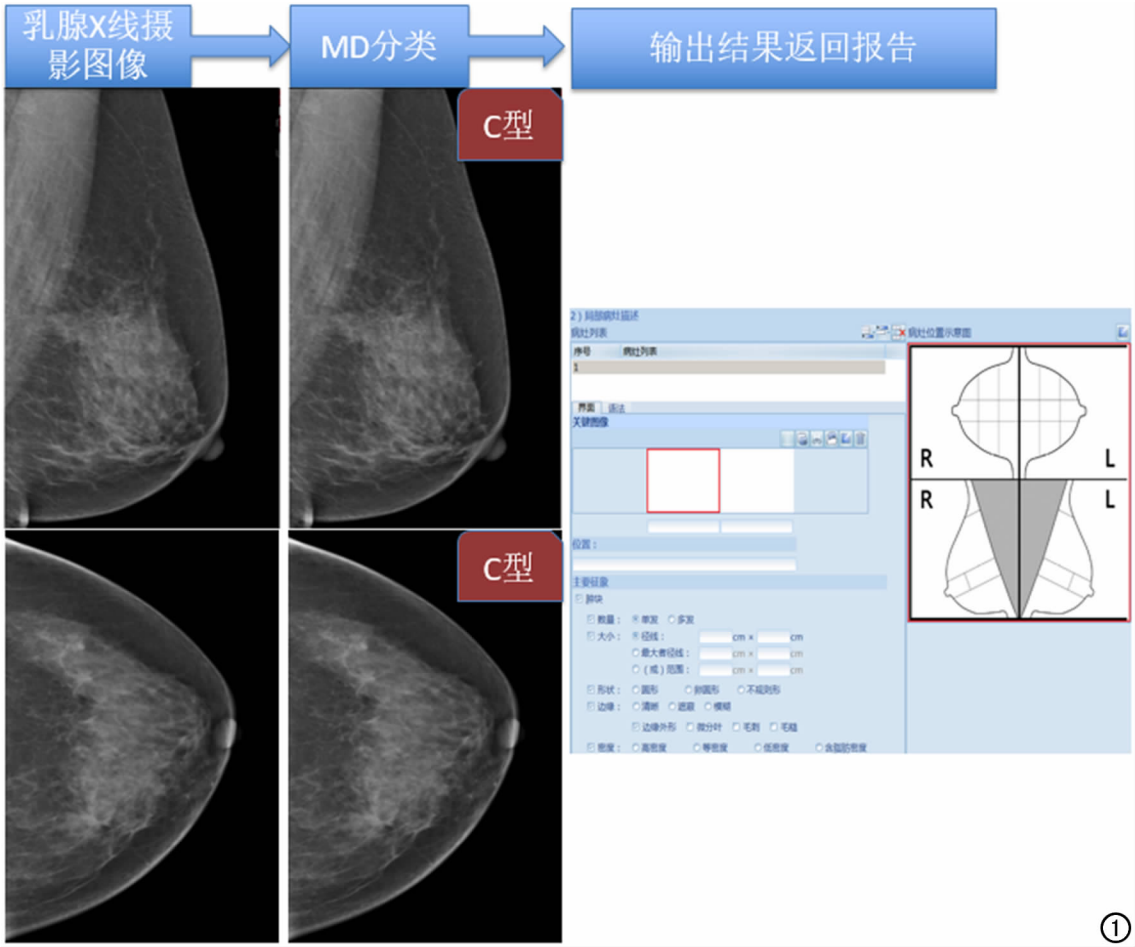


图 1 模型预测结果自动填写入结构化报告的工作流程。

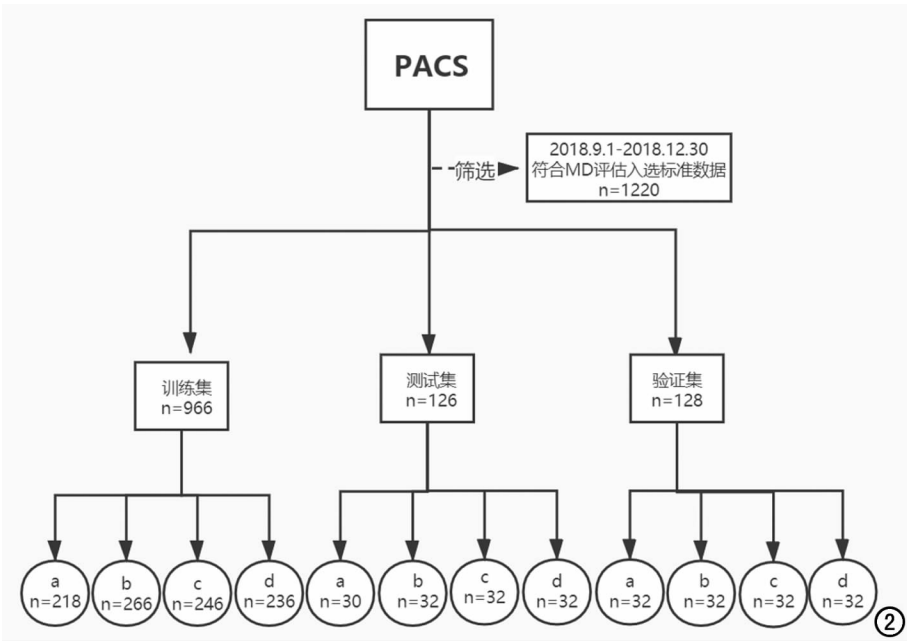


图 2 MD 模型训练纳入研究对象的流程图。

为 5 年、27 年)共同根据 BI-RADS 第 5 版的 MD 标准进行分型,标准如下:① a 型:乳腺内几乎全是脂肪组织;② b 型:乳腺内散在纤维腺体密度区域;③ c 型:乳腺组织密度不均,可能使小的肿块被遮挡;④ d 型:乳腺组织极其致密,使 X 线检查敏感度降低(图 3)。

3. 模型训练

硬件为 GPU NVIDIA Tesla P100 16G,软件包括 Python3. 6、Pytorch 0. 4. 1、Opencv、Numpy、SimpleITK。采用 Adam 作为优化器^[9]。训练腺体密度模型时,将 1220 个数据随机分训练集(train set, n= 966 个),调优集(validation set, n=128 个)和测试集(test

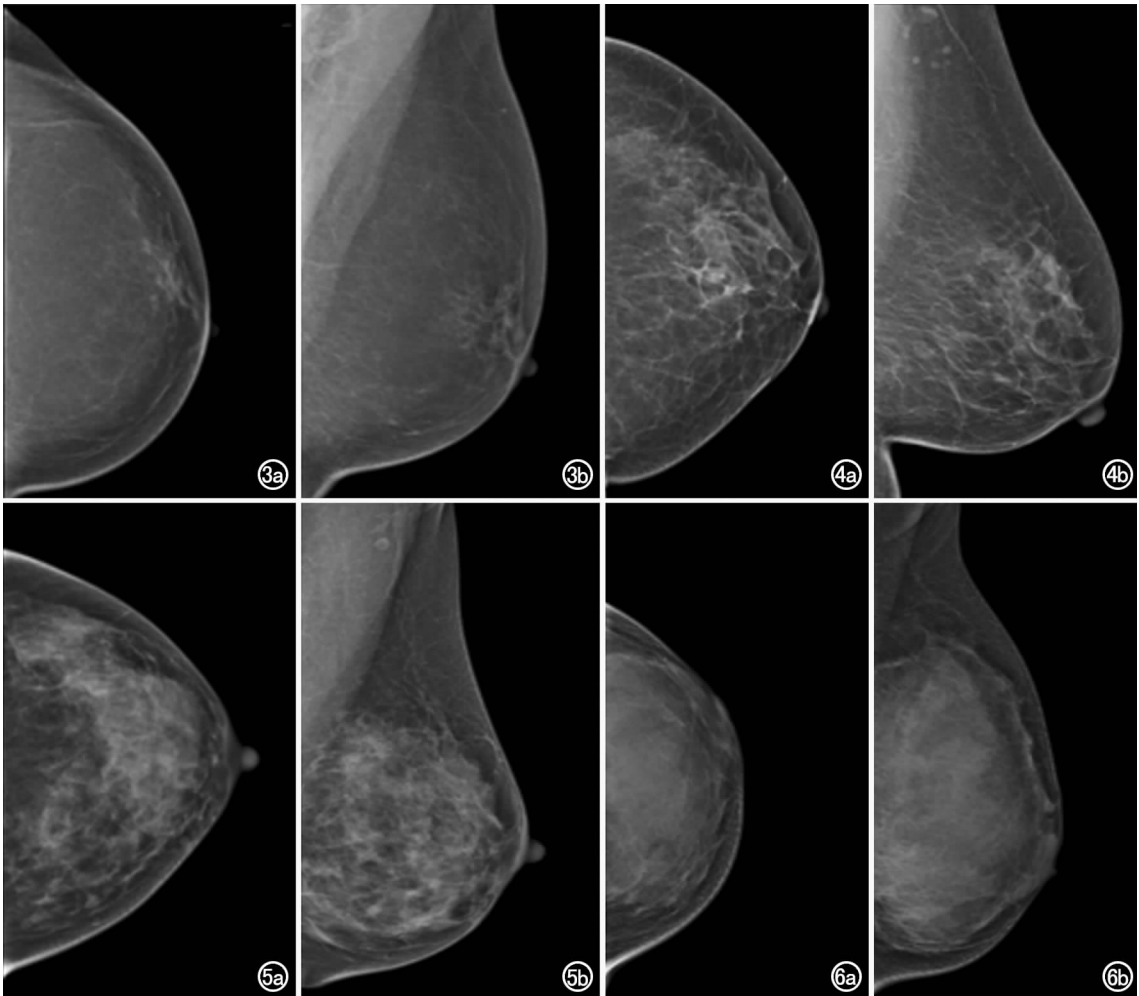


图 3 女,51 岁,a 型乳腺,乳腺内几乎全部是脂肪组织。a)CC 位;b)MLO 位。图 4 女,48 岁,b 型乳腺,乳腺内散在纤维腺体密度。a)CC 位;b)MLO 位。图 5 女,40 岁,c 型乳腺,乳腺组织密度不均,可能使小的肿块被遮挡。a)CC 位;b)MLO 位。图 6 女,27 岁,d 型乳腺,乳腺组织极其致密,降低乳腺 X 线摄影的敏感度。a)CC 位;b)MLO 位。

set,n=126 个)。输入图像设置为自动窗宽窗位,图像大小为 224×224,输出数据为模型分类预测结果。图像扩增方法包括:平移、旋转、随机噪声等。MedNet 网络主要参数如下: batch_size=150,num_epochs=300,learning_rate=0.00005。

4. 模型评价

得到测试集的混淆矩阵(confusion matrix)后,采用 PyCM 多分类混淆矩阵库进行多分类效能分析。

结 果

MD 分类模型对测试集数据预测的混淆矩阵和统计效能见表 1~2。测试集的 126 个数据的预测符合率(ACC Macro)为 0.83,a、b、c、d 各型预测符合率分别为 0.88、0.78、0.80 和 0.76。将 MD 分为致密组和非致密组时,非致密组精确度(precision)为 0.90,召回率(recall)为 0.74,F1 值为 0.81;致密组精确度为 0.86,召回率为 0.84,F1 值为 0.85。

表 1 MD 四分类模型预测结果(n=126)

分组	预测值					合计
	a	b	c	d	不确定	
致密组						
a	23	5	1	0	1	30
b	6	12	6	2	6	32
非致密组						
c	0	0	21	11	0	32
d	2	3	7	15	5	32

表 2 MD 四分类模型预测效能指标

预测效能	a	b	c	d	不确定
准确率	0.88	0.78	0.80	0.76	0.90
调整后 F 值	0.84	0.58	0.75	0.63	0.00
曲线下面积	0.84	0.64	0.75	0.67	None

讨 论

临床工作中,患者在接受乳腺 X 线摄影检查后,影像医师在书写报告时应提供 MD 评估信息,MD 不仅是乳腺癌的风险因素之一,更是对 MG 评估的准确性有影响。无论是乳腺疾病的诊断,还是乳腺癌筛查中,每年都要进行大量的乳腺 X 线摄影检查,对每次检查进行的乳腺密度评估是 ACR 影像解读的常规要求。然而,放射科医生对 MD 主观判断的重复性差,不同的放射科医生对乳腺密度的评估存在很大差异^[10-12]。深度学习已用于评估乳腺密度的初步工作^[13-14],本研究训练深度学习模型来评估乳腺 X 线摄影 MD,符合率较高。既往 Mohamed 等^[15]基于卷积神经网络(CNN)构建 MD 分类中最常见但一致性最差的两个类别(b 型、c 型)的分类模型,模型对 b 型和 c 型的判断符合率为 0.94;Ciritsis 等^[16]的研究表明,其构建的 CNN 模型根据不同体位对 MD 分类,在影

像专家共识决策方面,MLO 模型(0.92)和 CC 位模型(0.87)有很好的 consistency。本研究对 MD 四分类中,b 型和 c 型符合率接近 0.8,但二分类(致密组、非致密组)的准确率较高,在 0.9 左右,并可将结果接入到结构化报告中,不仅能减轻医师的工作强度、缩短报告时间,还能提高 MD 评估的一致性,与既往研究结果相似^[15-21]。

在 MD 四分类中,各类均有模型不能正常识别的病例(表 1),究其原因如下:①基于第 5 版的 BI-RADS MD 分类取消了关于腺体成分占比的规定,因而同一分类中图像的异质性更大,以 c 类为例,可能图像上仅有非常少的区域有高密度腺体足以产生遮蔽效应,而其他区域与 b 类相同。因此本研究训练的模型对四分类中二个相邻分类之间的区分有一定错误率,而对致密组和非致密组的分类诊断符合率较高。②本研究使用的图像数量较少,也未对图像质量进行分层,可能由于图像质量欠佳导致分类错误。③训练时未对图像中的病灶进行分层,在 a 类(5/30)和 b 类图像中肿块较大(6/32)时,标注者是以肿块之外的腺体部分为准对图像分类,而将图像输入模型训练时,肿块部分的图像特征也被模型提取特征,造成对于 a 类和 b 类的预测错误。上述原因分析为将来迭代模型提供了方向。首先,在做 MD 分类前应对图像质量进行评价,可前置一个图像甄别的模型,对投照体位、投照范围、投照条件等做出评估,将筛选过的合格图像输入到本模型中。这种流程在其他研究中也应用,前置的图像甄别模型有利于提高诊断模型的效能^[22]。另外,在预测 MD 分类时,应同时预测乳腺区域是否存在明显的肿块,将二个模型的结果集成处理,以避免大的肿块对腺体分型预测结果的影响。

本研究的局限性是非常明显的。首先,本研究是小样本回顾性研究,数据来源于同一个单位、同一台设备,常规扫描参数是相同的,所以未来需要进行前瞻性、大样本、多中心临床研究,不断迭代,才能达到较高的 MD 分类预测效能。其次,本研究虽然有 1220 个影像数据,但是只有 610 侧乳腺,而临床实践中我们一般都是综合 MLO 和 CC 位进行 MD 的判断。本研究将 CC 和 MLO 作为独立的影像数据进行分析,经过复审发现 CC 和 MLO 的分类是一致的,可能与本研究数据入组时只挑选了图像质量合格、且专家对分类达成一致的数据有关。在实际工作中,影像科医生读片时对 MLO 与 CC 的密度分类可能不一致,医生是综合两个体位做出判断的。未来在模型训练时应将体位因素增加到输入信息中,输出结果时也要先分别输出两体位图像的单独预测结果,再加以集成。第三,本研究的模型未进行临床价值的评价,在今后的模型研发过程中

应进行不同性质的临床验证试验,以了解影像医生对模型的接受情况^[23]。只有当模型在实际报告场景中应用,并确实提高了工作效率和准确率时,才能证明模型的真实价值^[24]。最后,乳腺影像诊断涉及了很多诊断任务,应将 MD 分类、病灶检出和分类等整合在一起,才能实现全流程的自动化。

总之,本研究基于深度学习训练乳腺 X 线 MD 分类模型,具有较高的分类效能,在经过更多验证和迭代后,有望用于临床工作,提高工作效率。

参考文献:

- [1] Boyd NF, Byng JW, Jong RA, et al. Quantitative classification of mammographic densities and breast cancer risk: results from the Canadian National Breast Screening Study[J]. Natl Cancer Inst, 1995, 87(9): 670-675.
- [2] Carney PA, Miglioretti DL, Yankaskas BC, et al. Individual and combined effects of age, breast density, and hormone replacement therapy use on the accuracy of screening mammography[J]. Ann Intern Med, 2003, 138(3): 168-175.
- [3] Irshad A, Leddy R, Ackerman S, et al. Effects of changes in BI-RADS density assessment guidelines (fourth versus fifth edition) on breast density assessment: intra- and interreader agreements and density distribution[J]. AJR, 2016, 207(6): 1366-1371.
- [4] 徐泽园, 曾辉, 秦耿耿, 等. 基于第 5 版乳腺影像报告和数据系统方法评价乳腺密度的一致性研究[J]. 肿瘤影像学, 2020, 29(3): 235-240.
- [5] Nazari SS, Mukherjee P. An overview of mammographic density and its association with breast cancer[J]. Breast Cancer, 2018, 25(3): 259-267.
- [6] 徐卫云, 赵洁玉, 张靖, 等. 女性 X 线乳腺密度的影响因素分析[J]. 西部医学, 2014, 26(7): 850-854.
- [7] Huo CW, Chew GL, Britt KL, et al. Mammographic density—a review on the current understanding of its association with breast cancer[J]. Breast Cancer Res Treat, 2014, 144(3): 479-502.
- [8] Youk JH, Gweon HM, Son EJ, et al. Automated volumetric breast density measurements in the era of the BI-RADS fifth edition: a comparison with visual assessment[J]. AJR, 2016, 206(5): 1056-1062.
- [9] Ruder S. An overview of gradient descent optimization algorithms [EB/OL]. ArXiv: 1609. 04747. <https://arxiv.org/abs/1609.04747>, Accessed February 25, 2021.
- [10] Nicholson BT, Lorusso AP, Smolkin M, et al. Accuracy of assigned BI-RADS breast density category definitions[J]. Academic Radiology, 2006, 13(9): 1143-1149.
- [11] van der Waal D, den Heeten GJ, Pijnappel RM, et al. Comparing visually assessed BI-RADS breast density and automated volumetric breast density software: a cross-sectional study in a breast cancer screening setting[J]. PLoS One, 2015, 10(9): e0136667.
- [12] Ciatto S, Houssami N, Apruzzese A, et al. Categorizing breast mammographic density: intra- and interobserver reproducibility of BI-RADS density categories[J]. Breast, 2005, 14(4): 269-275.
- [13] Wu N, Geras KJ, Shen Y, et al. Breast density classification with deep convolutional neural networks[EB/OL]. arXiv preprint: arXiv:1711. 03674. 2017. <https://arxiv.org/abs/1711.03674>. Accessed August 8, 2018.
- [14] Kallenberg M, Petersen K, Nielsen M, et al. Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring[J]. IEEE Trans Med Imaging, 2016, 35(5): 1322-1331.
- [15] Mohamed AA, Luo Y, Peng H, et al. Understanding clinical mammographic breast density assessment: a deep learning perspective[J]. J Digit Imaging, 2018, 31(4): 387-392.
- [16] Ciritis A, Rossi C, Vittoria De Martini I, et al. Determination of mammographic breast density using a deep convolutional neural network[J]. Br J Radiol, 2019, 92(1093): 20180691.
- [17] Dontchos BN, Meng AY, Barzilay R, et al. External validation of a deep learning model for predicting mammographic breast density in routine clinical practice[J]. Acad Radiol, 2020, 28(4). DOI: 10. 1016/j. acra. 2019. 12. 012.
- [18] Chang K, Beers AL, Brink L, et al. Multi-institutional assessment and crowdsourcing evaluation of deep learning for automated classification of breast density[J]. J Am Coll Radiol, 2020, 17(12): 1653-1662.
- [19] Pérez-Benito FJ, Signal F, Perez-Cortes JC, et al. A deep learning system to obtain the optimal parameters for a threshold-based breast and dense tissue segmentation[J]. Comput Methods Programs Biomed, 2020, 195: 105668.
- [20] 贾田菊, 马彦云, 李延涛, 等. 基于深度学习的乳腺数字化 X 线 BI-RADS 密度分类的研究[J]. 山西医科大学学报, 2019, 50(4): 506-510.
- [21] Lee J, Nishikawa RM. Automated mammographic breast density estimation using a fully convolutional network[J]. Med Phys, 2018, 45(3): 1178-1190.
- [22] 张晓东, 孙兆男, 任昕, 等. 基于深度残差网络研发辅助诊断软件用于 X 线胸片分类诊断[J]. 放射学实践, 2019, 34(9): 952-957.
- [23] 高歌, 胡娟, 王成彦, 等. 整合临床及多参数 MRI 信息的前列腺癌 CAD 系统: 诊断效能研究[J]. 放射学实践, 2016, 31(12): 1143-1145.
- [24] 朱丽娜, 高歌, 刘义, 等. CAD 整合入前列腺多参数 MRI 结构化报告: 低经验读片者诊断效能研究[J]. 放射学实践, 2020, 35(10): 1282-1287.

(收稿日期: 2020-12-16 修回日期: 2021-04-08)