

· 人工智能影像学专题 ·

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

张晓东, 孙兆男, 任昕, 周宇, 周雯, 李建辉, 谢辉辉, 刘婧, 张虽虽, 李津书, 王霄英

【摘要】 目的:研究以深度残差网络(ResNet)为基础架构建立深度学习模型,对 X 线胸片(CXR)做出“有发现”与“无发现”鉴别诊断的可行性。**方法:**回顾性收集 2017 年 1 月 1 日至 2018 年 7 月 1 日的连续 CXR 图像及诊断报告,经过数据清洗后分为“无发现”组(无任何异常发现,诊断印象为“两肺心膈未见异常”,共 9765 例)与“有发现”组(诊断印象中提及了一种以上影像所见,共 9956 例)。使用 ResNet152(152 layers)作为二分类模型的基础架构,结合 Grad-CAM 技术生成模型激活热图,训练二分类模型。数据随机分为训练集(70%)、调优集(20%)和测试集(10%)。以测试集的预测结果检测 CXR 二分类模型的效能。**结果:**在测试集中(“有发现”者 1018 例,“无发现”者 995 例),CXR 二分类模型鉴别“有发现”与“无发现”的精确度分别为 0.885 和 0.894,召回率分别为 0.898 和 0.880,F1 一分数分别为 0.891 和 0.887,ROC 曲线下面积均为 0.96。**结论:**使用 CXR 二分类模型可对 X 线胸片做出“无发现”与“有发现”的预测。

【关键词】 深度学习;人工智能;深度残差网络;胸部 X 线片;用例;结构化报告

【中图分类号】 TP18;R814.41 **【文献标识码】** A **【文章编号】** 1000-0313(2019)09-0952-06

DOI:10.13609/j.cnki.1000-0313.2019.09.004

开放科学(资源服务)标识码(OSID):



Development of a ResNet-based CADx software for classification diagnosis in chest X-ray images

ZHANG Xiao-dong, SUN Zhao-nan, REN Xin, et al. Department of Radiology, Peking University First Hospital, Beijing 100034, China

【Abstract】 Objective: To evaluate the feasibility of residual deep neural network (ResNet) based deep learning model in the differentiation between "no significant finding" and "any significant finding" cases in chest X-ray (CXR) images. **Methods:** From January 1 2017 to July 1 2018, a consecutive cohort were retrospectively collected in this study. All the images were divided into the "no significant finding" group (defined as no significant finding was detected and reported) and the "any significant finding" group (defined as any significant findings that were detected or reported). Finally, 9765 cases were collected in "no significant finding" group and 9956 cases were collected in "any significant finding" group. The ResNet 152 (152 layers) was used as the framework of the binary classification model and the Grad-CAM (gradient class activation maps) was used to generate model activation heating maps. The cohort was randomly divided into training (70%), validation (20%), and testing dataset (10%). The results obtained in testing dataset was considered as the performance of the model. **Results:** In the testing dataset, 1018 cases were classified as "any significant finding", while 995 cases were classified as "no significant finding". The precision, recall and F1-scores were 0.885 vs 0.894, 0.898 vs 0.880 and 0.891 vs 0.887, respectively in terms of two group (any significant finding vs no significant finding). The areas under ROC (AUC) were both 0.96. **Conclusion:** The CXR binary classification model could be useful in the prediction of "no significant finding" and "any significant finding" for CXR images.

【Key words】 Deep learning; Artificial intelligence; residual deep neural network; Chest X-ray; Use case; Structured report

作者单位:100034 北京,北京大学第一医院医学影像科(张晓东,孙兆男,任昕,周宇,周雯,李建辉,谢辉辉,刘婧,李津书,王霄英);100011 北京,北京赛迈特锐医学科技有限公司(张虽虽)

作者简介:张晓东(1979-),男,山西人,博士,副主任技师,主要从事医学图像处理工作。

通讯作者:王霄英, E-mail: cjr.wangxiaoying@vip.163.com

胸部 X 线片 (chest X-ray, CXR) 是诊断呼吸系统疾病的重要影像检查方法, 在医学影像科日常工作中占很大比例。使用 CXR 进行检查的临床场景包括门诊检查、术前常规检查、健康体检、急诊检查等^[1]; 在前三种情况下, CXR 结果多数为正常, 影像检查的任务是检出少数的异常者, 再做进一步处理 (如行胸部 CT 检查等)。近年来人工智能 (artificial intelligence, AI) 领域的技术开始应用于 CXR 的质控、图像分割和诊断等工作^[2-6], 初步结果认为在适应证明确的情况下, AI 对完成影像检查任务是有帮助的。本研究拟使用深度学习方法研发影像辅助诊断工具, 对门诊、术前常规和体检人群的 CXR 进行“无发现”、“有发现”的分类 (以下简称二分类), 以提高医学影像科医师的工作效率。

材料与方法

本研究获得了本院伦理审查委员会的批准, 按照本单位 AI 项目研发规范开展研究工作。

1. 用例定义

根据本单位 AI 项目管理方法, 首先定义研发 CXR 二分类 AI 模型的用例 (Use Case)^[7], 包括以下内容: CXR 二分类 AI 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“无发现”和“有发现”两类。“无发现”的定义: CXR 上未见应写入报告的影像发现, 诊断印象为两肺心膈未见异常。“有发现”的定义: CXR 上有需要在报告中提及的任何影像

所见, 包括 (但不限于) 胸壁、肺、纵隔、横膈、上腹部的影像所见, 诊断印象中提及了一种以上的影像所见。

2. 研究队列建立

回顾性收集 CXR 数据建立二分类模型研发队列。从本院 PACS 系统中检索 2017 年 7 月 1 日至 2018 年 7 月 1 日的连续资料, 由 2 位住院医师根据登记信息、设备信息、影像报告和 CXR 图像进行筛选, 再由 1 位高年资医师再次读片, 对图像进行最终分类 (表 1)。

经过数据筛选建立研究队列, 获得“无发现”患者的 CXR 图像共 9765 幅, “有发现”患者的 CXR 图像共 9956 幅。以患者为单位随机分为训练集 (training dataset, 70%)、调优集 (validation dataset, 20%) 和测试集 (testing dataset, 10%), 各个数据集中的患者信息是不同的, 具体见表 2。

表 2 CXR 二分类模型训练过程中的数据分配 (例)

分类	训练集	调优集	测试集	合计
有发现	6903	2035	1018	9956
无发现	6795	1975	995	9765

3. 数据处理

CXR 二分类模型输入图像大小为 512×512, 模型输出为“无发现”、“有发现”二分类概率, 并可通过 Grad-CAM 方法得到该模型的激活热图。

模型训练的初始权重来源于 ImageNet 预训练 ResNet152 模型^[8]。图像预处理部分包括以下步骤: 原始图像大小缩放至 512×512, 并对图像像素值按照 ImageNet 数据集的均值和方差进行归一化处理。扩

表 1 CXR 二分类模型研发数据筛选方法

筛选依据	条件	分类*	合格值	不合格值
登记信息	检查类型	0;1	X 线	其他
	检查项目		胸部正位	胸片侧位、胸片正侧位、床旁胸片等
	年龄		>18 岁	≤17 岁
	图像数量		1 张	0 张, ≥2 张
设备信息	诊断报告	0;1	具备	不具备
	检查设备		CR; DR	其他
报告信息	影像表现	0	“肺野”、“纵隔”、“横膈”、“胸壁”、“下颈部”、“上腹部”正常	术后改变、乳腺术后、支架术后、起搏器术后; PICC、IVAP 等置入物; 膈抬高; 叉状肋、脊柱侧弯、颈 7 横突过长、胸廓萎陷; 陈旧肋骨骨折; 甲状腺增大; 间位结肠、胆囊结石等
		1	胸部疾病 (结核、结节、占位等); 胸部病变 (渗出、实变、纤维索条影、肺气肿、肺不张、钙化、陈旧病变等); 纵隔形态异常; 胸膜病变, 胸壁病变等	
	诊断印象 (报告)	0	双肺心膈未见异常、未见明显异常、未见明确异常、无异常; 两肺/心肺未见异常、未见明显异常、未见明确异常、无异常; 肺心膈未见异常; 未见## 疾病, 且影像表现中“双肺野清晰”等	/
		1	不符合 0 者均入选	/
CXR 图像	投照体位	0;1	正确	不正确
	曝光条件		适当	不适当
	体表金属		无或少量	明显
	诊断正确		是	否

注: * 0=无发现; 1=有发现

增方法:随机旋转、缩放、平移等。

4. 模型训练

在 NVIDIA GPU (P100, Nvidia Corporation, Santa Clara, CA) 上利用 Pytorch (<https://pytorch.org/>) 深度学习 Python 库,在本院收集整理的临床患者 CXR 图像数据集上进行开发和训练。

二分类模型基于 152 层的卷积神经网络模型,以 Resnet 为卷积神经网络(convolutional neural network, CNN)深度学习模型的基础架构^[9],结合梯度类激活图(gradient class activation maps, Grad-CAM)方法^[10],建立可以同时显示分类结果及激活热图的 CXR 二分类模型(图 1)。

训练和调优分别在训练集和调优集数据上进行,训练好的模型在测试集上验证。用测试集的混淆矩阵(confusion matrix)评价二分类模型的效能,得到将 CXR 分为“无发现”与“有发现”的精确度(precision)、召回率(recall)、F1-分数(F1-score)及受试者工作特征(receiver operating characteristic, ROC)曲线的曲线下面积。将 CXR 二分类模型预测结果返回到结构化报告中,以保证 AI 模型在临床工作流程中的可用性。

结 果

CXR 二分类模型对测试集数据预测的混淆矩阵(图 2)显示,分类的精确度、召回率和 F1-分数均较高

(表 3), ROC 曲线下面积均为 0.96(图 3)。在测试集 1018 个“有发现”的病例中,模型正确识别了其中 914 个,有 104 个未被识别出来;995 个“无发现”的病例中,模型正确识别了其中 876 个,其中 119 个被误认为“有发现”。以“有发现”为阳性,模型的假阴性率为 10.6%,假阳性率为 11.5%。经与影像专家的读片结果进行对照,发现假阴性者的主要“发现”种类为:少量纤维索条影、小范围钙化、心脏轮廓稍饱满、主动脉结轻度突出、肺门影稍重、肺纹理轻度增多、肋骨变异等。

表 3 CXR 二分类模型在测试集中的分类效能

分类	例数	精确度	召回率	F1-分数
有发现	1018	0.885	0.898	0.891
无发现	995	0.894	0.880	0.887

CXR 二分类模型返回了激活图(图 4~6),可以看到图像中的病灶区域与预测为“有发现”的激活区域基本吻合,而如果将该图像预测为正常的话,其激活高亮区域大部分散在分布,与解剖学上的特征区域没有相关性。

根据用例设计要求,将 CXR 二分类模型结果直接返回到结构化报告中,生成相应的结果。对分类是“无发现”的,直接生成“正常”报告,包括常规影像表现和诊断印象,不返回激活图。对分类是“有发现”的,激活“异常”控件,并返回激活图(图 7)。

讨 论

胸部 X 线检查是呼吸系统疾病重要的影像检查方法之一,是医学影像科工作中最常见的临床任务。在综合性医院,大多数常规体检和呼吸系统疾病筛查均首选 CXR^[11],且在上述诊断任务中相当多的影像诊断结果是“无发现”。使用辅助诊断软件将“有发现”与“无发现”分开,可提高医生的工作效率,有利于工作流程的优化。值得注意的是,本研究中此模型召回率和 F1 值不足 0.9,将面临着 10%左右的漏诊率和误诊率,尤其是漏诊将给临床应用价值带来较大影响。经与影像专家的读片结果进行对比,发现漏诊的情况主要集中于轻微影像改变,对临床结局影响不大,所以仍需在工作中持续改进。当前的改进方法是将在实际工作中发现的预测错误的数据持续性收集、整理、标注,用于模型

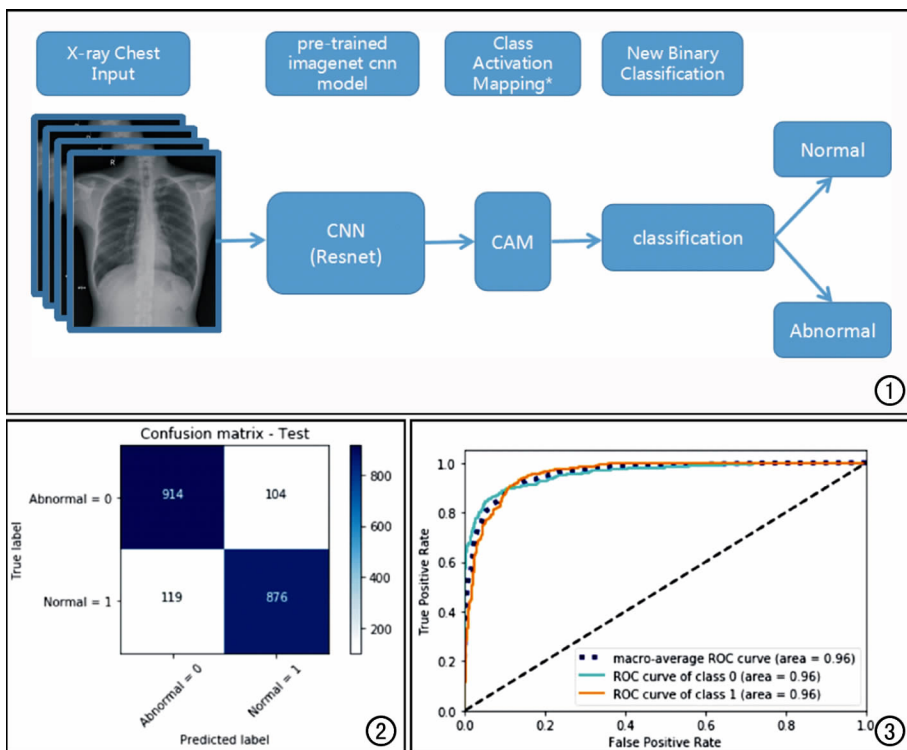


图 1 基于 ResNet152 基础架构并结合 Grad-CAM 的 CXR 二分类模型训练过程。图 2 CXR 二分类模型在测试集中的混淆矩阵。图 3 CXR 二分类模型在测试集中的 ROC 曲线。

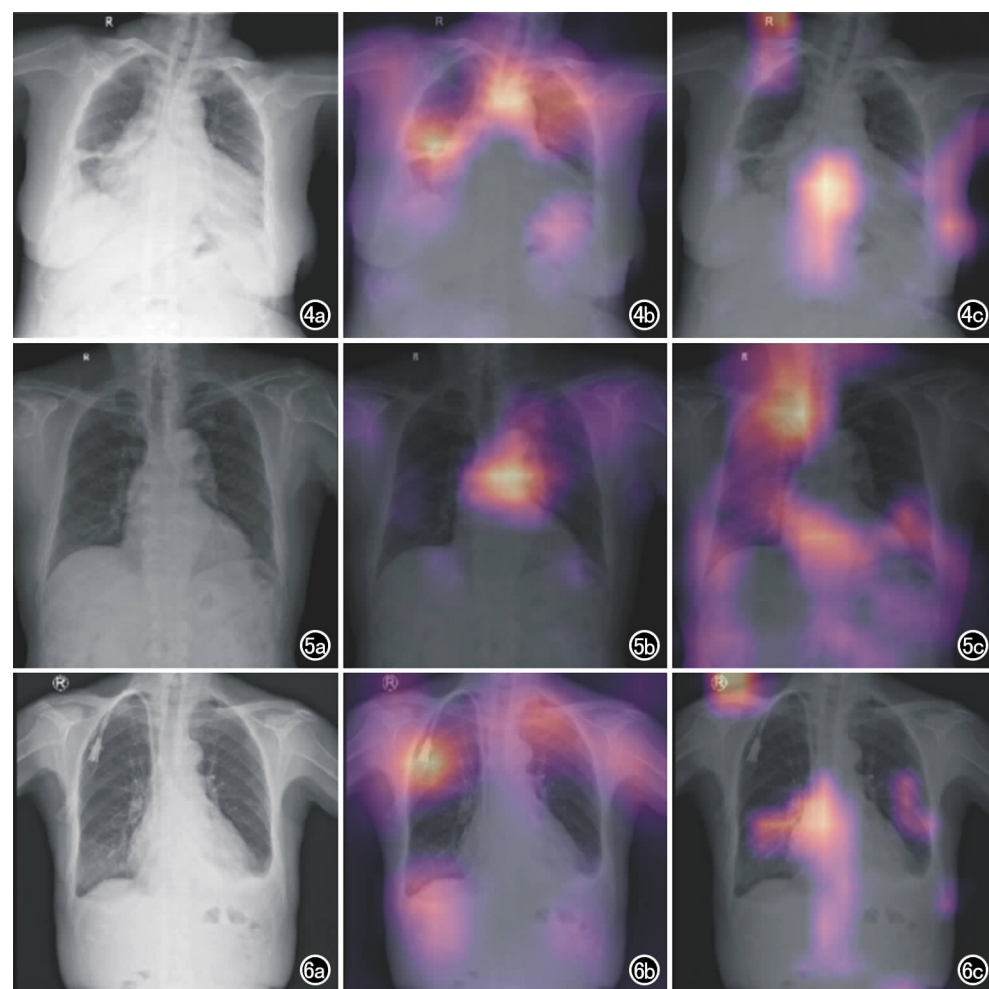


图 4 心衰合并右肺病变及胸腔积液, CXR 二分类预测结果为“有发现”。a) CXR 原图; b) “有发现”的激活图, 高亮区域为纵隔、心尖、右下肺, 与影像征象异常的区域基本吻合; c) 预测结果如果是“无发现”, 则激活区域在肺外以及脊柱区域, 与图像的关键解剖位置不相关。图 5 主动脉增宽, CXR 二分类预测结果为“有发现”。a) CXR 原图; b) “有发现”的激活图, 高亮区域为中纵隔主动脉区域, 与影像征象异常的区域基本吻合; c) 预测结果如果是“无发现”, 则激活区域散在分布于整个图像中, 与图像的关键解剖位置不相关。图 6 深静脉置管, 双侧胸腔积液, 肺内散在小片渗出, CXR 二分类预测结果为“有发现”。a) CXR 原图; b) “有发现”的激活图, 高亮区域为深静脉置管、双下胸部和左上肺, 与影像征象异常的区域基本吻合; c) 预测结果如果是“无发现”, 则激活区域散在分布于整个图像中, 无特殊规律, 与图像的关键解剖位置不相关。

的进一步训练, 以实现模型的持续迭代。另一方面, 将患者入组人群进一步细分, 统一专家诊断标准, 都有可能提高模型的准确率。

本研究结果显示, 深度学习模型可以用于 CXR 的辅助诊断, 这与既往研究结果相似^[12,13]。对于医学图像的分类问题, 模型的可解释性非常重要, 有利于对模型分类结果的准确性评估。近年来, 相关学者对于深度卷积神经网络模型的可解释性做了很多探索, 其中 CAM 方法是对不同类别的卷积层生成的特征图加权求和而得到激活热图, 通过该激活热图可以对模型分

类的结果进行解释^[14]。Grad-CAM 是 CAM 技术的扩展, 可以适用于任何 CNN 架构。本研究中 CXR 二分类模型利用 Grad-CAM 来生成分类激活热图, 在该热图上可以展示输入图像中的哪些区域对于得到该分类结果是重要的激活区域。

本研究的数据筛选过程是由医生回顾性读片完成的。对于大样本模型训练, 多是从报告中用自然语言处理 (natural language processing, NLP) 方法提取信息^[15,16], 优点是处理大量影像数据时可节省医生时间, 但报告内容常欠规范, 不完整, 因此需要大量的数据, 即便是大量数据其准确性仍欠佳^[17]。本研究由影像医生进行数据筛选, 且由高年资医生再次回顾性阅片确定图像分类, 虽然在数据清洗过程中人力成本较高, 但结果较好, 用不太多的数据训练得到的模型, 准确性仍很高^[18]。

AI 模型研发和泛化过程中应考虑技术问题^[19]。本研究 CXR 二分类模型训练过程中未对设备进行筛选, 使用了连续数据。CXR 图像来自本单位在实际临床工作中使用的多种 DR 设备, 由不同技师完成摄片工

作, 未根据设备、人员进行分组。本研究结果证明了不同 DR 设备、不同技师采集的图像均可用于模型的训练, 且无论来源于何种 DR 设备的图像, 其验证集数据的分类预测均达到较好的结果。胸部 X 线检查有明确的技术规范, 经过常规培训的技师都可按照规范完成日常工作, 操作一致性强; 且现代 DR 设备有自动曝光功能, 可自动设置最佳投照条件并调整图像的窗水平, 因此图像预处理难度不大, 可适用于多种 AI 模型的训练^[19]。由于常规 CXR 基本可以保证图像质量,

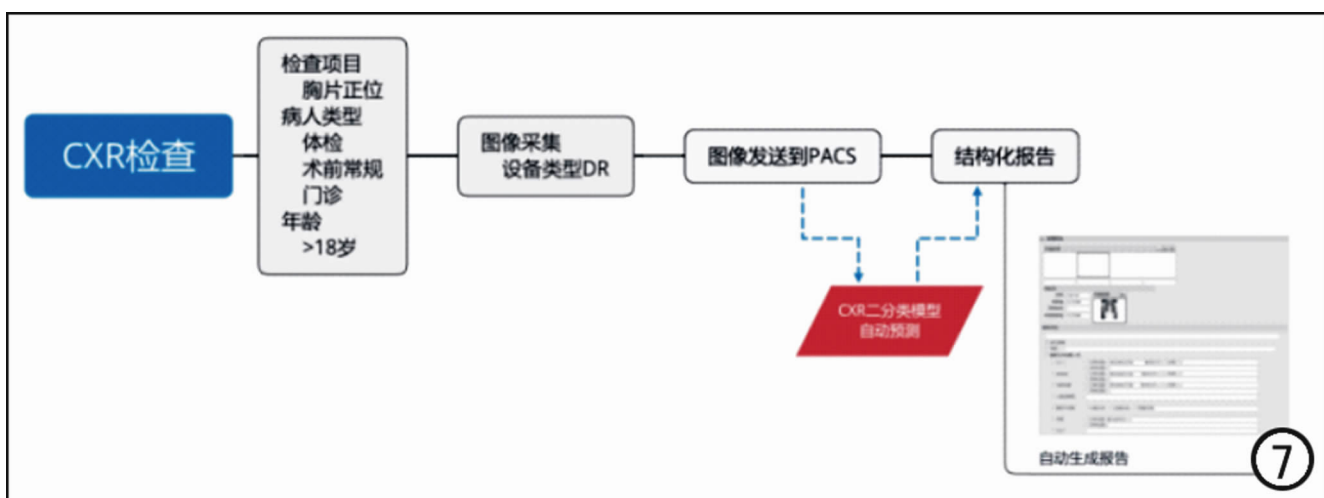


图7 CXR 二分类模型与结构化报告整合的流程。

来源于不同设备的图像性质差别不大,从这个角度考虑,在 CXR 二分类模型泛化过程中,没有图像采集技术上的风险。

从实际场景出发设定 AI 研发目标,是项目设计的关键问题,决定了模型临床应用的适用性^[20]。本研究 CXR 二分类模型训练过程中定义的分类输出结果为“无发现”和“有发现”。在最初的用例设计中,拟将返回的二分类值定义为“正常”和“异常”,但在实际建立队列过程中,发现“肺心膈正常”并不是“正常”。在“肺心膈正常”时,图像中如存在有临床意义的其他征象,也需要报告出来;如“单侧乳腺切除术后,胸壁软组织缺失,两肺心膈未见异常”,再如“PICC 置入术后,两肺心膈未见异常”等,这些术后改变、置入物等信息,都是 CXR 检查应发现并报告给临床医生的。因此,为了提高模型训练效率,本研究中将“肺心膈正常”且合并其他任何需报告的情况归为“无发现”;将有疾病影像表现,或“肺心膈正常但合并其它有临床意义的所见”情况归为“有发现”。

AI 临床应用的一个基本要求是与临床流程整合。在符合法规、伦理的框架下,国内外技术人员做过很多探索^[21],均认为把 AI 模型作为独立第三方软件使用的形式不是最优解。笔者认为将 AI 结果直接返回到临床实际使用的结构化报告,是较好的解决方案^[22]。为了提高临床工作效率,在将 CXR 二分类结果返回到报告时,如果模型预测结果为“无发现”,则激活“两肺心膈未见异常”的报告内容,此时报告是完全自动生成的。如果模型预测结果为“有发现”,则在报告中激活的是“异常”控件,需要医生进一步对异常进行描述,这时报告是半自动生成的。本单位临床实际工作中 CXR“无发现”的情况较常见,占到工作量的一半以上,因此如将 CXR 二分类模型接入报告,在分类准确

率足够高时,能明显提高临床工作效率。而且由模型返回结果自动生成的报告,语言规范、结构清晰,对未来提取报告信息做其他研究分析也有利^[23]。

本研究的局限性:①CXR 二分类模型在“有发现”情况下提供的信息不是具体疾病,临床价值受限,仍需进一步优化,训练用于多种疾病诊断的 CXR 多分类 AI 模型^[3],针对不同体位^[2]、不同年龄、不同申请目的时对多种常见疾病做出诊断^[24,25];②模型研发后应在实际临床工作中进行验证,开展前瞻性、随机对照临床试验,证明使用 AI 模型的临床获益;③CXR 二分类模型在本单位经过验证后,应尝试将其泛化^[26],应用于不同医疗机构、不同临床场景下,进一步提高其适用性。

综上所述,使用深度学习模型可对 X 线胸片做出“无发现”与“有发现”的分类诊断,在常规临床工作中使用模型分检病例,有利于提高医学影像科的工作效率。

致谢:感谢北京赛迈特锐医学科技有限公司张欣静、郭嘉欣在数据整理工作中的贡献。

参考文献:

- [1] American College of Radiology. ACR Appropriateness Criteria [EB/OL]. <https://www.acr.org/Clinical-Resources/ACR-Appropriateness-Criteria>, 2019.
- [2] Kim TK, Yi PH, Wei J, et al. Deep learning method for automated classification of anteroposterior and posteroanterior chest radiographs[J]. J Digit Imaging, 2019 Apr 10. DOI:10.1007/s10278-019-00208-0.
- [3] Novikov AA, Lenis D, Major D, et al. Fully convolutional architectures for multiclass segmentation in chest radiographs[J]. IEEE Trans Med Imaging, 2018, 37(8):1865-1876.
- [4] Wei Y, Liu YB, Lin LY, et al. Lung field segmentation in chest radiographs from boundary maps by a structured edge detector[J]. IEEE J Biomed Health Inform, 2018, 22(3):842-851.
- [5] Candemir S, Antani S. A review on lung boundary detection in

- chest X-rays[J]. *Int J Comput Assist Radiol Surg*, 2019, 14(4): 563-576.
- [6] Mao C, Yao L, Pan Y, et al. Deep generative classifiers for thoracic disease diagnosis with chest X-ray images[C]. Madrid: IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2018; 1209-1214.
- [7] American College of Radiology Data Science Institute. AI use cases 101: what radiologists need to know [EB/OL]. <https://www.acrdsi.org/Blog/AI-Use-Cases-101>, 2019.
- [8] Jia D, Wei D, Richard S, et al. Imagenet: a large-scale hier-archical image database in computer vision and pattern recognition[C]. Florida: Fontainebleau Resort, Miami Beach, 2009; 248-255.
- [9] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]. Las Vegas: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016; 770-778.
- [10] Rahimi RS, Rockey DC. Overuse of head computed tomography in cirrhosis with altered mental status[J]. *Am J Med Sci*, 2016, 351(5): 459-466.
- [11] Heo SJ, Kim Y, Yun S, et al. Deep learning algorithms with demographic information help to detect tuberculosis in chest radiographs in annual Workers' health examination data[J]. *Int J Environ Res Public Health*, 2019, 16(2): 250.
- [12] Rajpurkar P, Irvin J, Ball RL, et al. Deep learning for chest radiograph diagnosis: a retrospective comparison of the cheXNeXt Algorithm to practicing radiologists[J]. *PLoS Med*, 2018, 15(11): e1002686.
- [13] Singh R, Kalra MK, Nitiwarangkul C, et al. Deep learning in chest radiography: detection of findings and presence of change [J]. *PLoS One*, 2018, 13(10): e0204155.
- [14] Zhou B, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization[C]. Las Vegas: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016; 2921-2929.
- [15] Chen MC, Ball RL, Yang L, et al. Deep learning to classify radiology free-text reports[J]. *Radiology*, 2018, 286(3): 845-852.
- [16] 马帅, 吴静云, 刘义, 等. 自然语言处理对泌尿系结石 CT 特征的提取: 与人工诊断结果的对照研究[J]. *放射学实践*, 2016, 31(12): 1151-1154.
- [17] Wang C, Elazab A, Jia F, et al. Automated chest screening based on a hybrid model of transfer learning and convolutional sparse denoising autoencoder[J]. *Biomed Eng Online*, 2018, 17(1): 63.
- [18] Qin C, Yao D, Shi Y, et al. Computer-aided detection in chest radiography based on artificial intelligence: a survey[J]. *Biomed Eng Online*, 2018, 17(1): 113.
- [19] Lakhani P. Deep convolutional neural networks for endotracheal tube position and X-ray image classification: challenges and opportunities[J]. *J Digit Imaging*, 2017, 30(4): 460-468.
- [20] Labaki WW, Han MK. Artificial intelligence and chest imaging. Will deep learning make us smarter? [J]. *Am J Respir Crit Care Med*, 2018, 197(2): 148-150.
- [21] Parker W, Lee B, Nicolaou S. AI is bringing USB back: implementing a beta chest X-ray neural network[J]. *J Digit Imaging*, 2019 Jun 25. DOI: 10.1007/s10278-019-00247-7.
- [22] 曹敏, 郭小超, 张来运, 等. 医学影像结构式报告的现状和展望[J]. *放射学实践*, 2016, 31(12): 1130-1132.
- [23] Huesch MD, Cherian R, Labib S, et al. Evaluating report text variation and informativeness: natural language processing of CT chest imaging for pulmonary embolism[J]. *J Am Coll Radiol*, 2018, 15(3): 554-562.
- [24] Pasa F, Golkov V, Pfeiffer F, et al. Efficient deep network architectures for fast chest X-ray tuberculosis screening and visualization[J]. *Sci Rep*, 2019, 9(1): 6268.
- [25] Taylor AG, Mielke C, Mongan J. Automated detection of moderate and large pneumothorax on frontal chest X-rays using deep convolutional neural networks: A retrospective study[J]. *PLoS Med*, 2018, 15(11): e1002697.
- [26] Zech JR, Badgeley MA, Liu M, et al. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study [J]. *PLoS Med*, 2018, 15(11): e1002683.

(收稿日期: 2019-07-31 修回日期: 2019-08-12)