

尘肺胸片人工智能读片三种模型效能比较研究

汪伟^{1a}, 李宝平², 沈福海^{1b}, 孙治平², 侯博文², 高丽妮², 阎聪侠^{1a}

1. 华北理工大学 a. 公共卫生学院 b. 河北省煤矿卫生与安全实验室, 河北 唐山 063000

2. 应急总医院职业病科, 北京 100028



DOI 10.11836/JEOM21216

摘要:

[背景] 尘肺病的人工诊断受到多种因素的影响, 容易出现漏诊误诊情况。人工智能在医学影像领域的发展十分迅速, 考虑能否利用人工智能实现对尘肺病影像的读片。

[目的] 基于深度卷积神经网络的方法构建三种深度学习模型, 进行尘肺病有无的诊断。评价三种模型诊断效能并进行对比, 挑选出最优的模型。

[方法] 收集 7 家医院在 2017 年 6 月—2020 年 12 月拍摄的数字 X 线摄影(DR)胸片并对胸片进行质量控制。收集的 DR 胸片中尘肺诊断阳性为阳性组, 无尘肺胸片为阴性组。由通过阅片考核的专家对收集胸片进行标注, 标注过程中不断进行基于最大期望算法的一致性考核。标注后的数据经数据清洗、数据归档、预处理后纳入训练集和验证集。构建 TMNet、ResNet-50 和 ResNeXt-50 三种深度卷积神经网络模型, 采用十折交叉验证法对模型进行训练, 得到最优的模型。收集未纳入训练集和验证集的 500 例 DR 胸片, 经由 5 位资深专家共同鉴定作为金标准成为测试集。通过测试, 得到三种模型的准确率、灵敏度、特异度、受试者工作特征曲线下面积(AUC)等指标, 评价三种模型性能并进行对比。

[结果] 本研究共收集训练集与验证集 DR 胸片 24867 张, 其中阳性组 6978 例, 阴性组 17889 例。收集胸片中肺部异常情况如气胸、肺结核等共有 312 例。共有 9 名专家对胸片进行了标注, 尘肺异常(不分期)的标注一致性率均在 88%以上, 尘肺分期的标注一致性率从 84.68%到 93.66%不等。TMNet 的诊断准确率为 95.20%, 灵敏度达到 99.66%, 特异度为 88.61%, AUC 值为 0.987。ResNeXt 的诊断准确率为 87.00%, 灵敏度达到 89.93%, 特异度为 82.67%, AUC 值为 0.911。ResNet 的诊断准确率为 84.00%, 灵敏度达到 85.91%, 特异度为 81.19%, AUC 值为 0.912。TMNet 模型的上述指标均高于 ResNeXt-50 和 ResNet-50 模型。TMNet 与另外两个模型 AUC 差值的差异有统计学意义($P < 0.001$)。

[结论] 三种卷积神经网络模型均可有效诊断尘肺病的有无, 其中 TMNet 的效能最好。

关键词: 尘肺病; 卷积神经网络; 人工智能; 深度学习

Comparison on efficiency of three artificial intelligence-based models to read pneumoconiosis chest radiographs WANG Wei^{1a}, LI Baoping², SHEN Fuhai^{1b}, SUN Zhiping², HOU Bowen², GAO Lini², YAN Congxia^{1a} (1.a. School of Public Health b. Hebei Province Coal Mine Health and Safety Laboratory, North China University of Science and Technology, Tangshan, Hebei 063000, China; 2. Department of Occupational Disease, Emergency General Hospital, Beijing 100028, China)

Abstract:

[Background] Diagnosis of pneumoconiosis by radiologist reading chest X-ray images is affected by many factors and is prone to misdiagnosis/missed diagnosis. With the rapid development of artificial intelligence in the field of medical imaging, whether artificial intelligence can be used to read images of pneumoconiosis deserves consideration.

[Objective] Three deep learning models for identifying presence of pneumoconiosis were constructed based on deep convolutional neural network. An optimal model was selected by comparing diagnostic efficiency of the three models.

[Methods] Digital radiography (DR) chest images were collected between June 2017 and December 2020 from 7 hospitals and standard radiograph quality control protocol was also followed. The DR chest images with positive results were classified into the positive group, while those without pneumoconiosis were classified into the negative group. The collected chest radiographs were labeled by experts who had passed the assessment of reading radiographs,

作者简介

汪伟(1995—), 男, 硕士生;
E-mail: 1095507695@qq.com

通信作者

李宝平, E-mail: lbp00@sohu.com

伦理审批

已获取
利益冲突 无申报
收稿日期 2021-05-14
录用日期 2021-06-08

文章编号 2095-9982(2022)01-0041-06

中图分类号 R13

文献标志码 A

引用

汪伟, 李宝平, 沈福海, 等. 尘肺胸片人工智能读片三种模型效能比较研究[J]. 环境与职业医学, 2022, 39(1): 41-46.

本文链接

www.jeom.org/article/cn/10.11836/JEOM21216

Correspondence to

LI Baoping, E-mail: lbp00@sohu.com

Ethics approval

Obtained

Competing interests None declared

Received 2021-05-14

Accepted 2021-06-08

To cite

WANG Wei, LI Baoping, SHEN Fuhai, et al. Comparison on efficiency of three artificial intelligence-based models to read pneumoconiosis chest radiographs[J]. Journal of Environmental and Occupational Medicine, 2022, 39(1): 41-46.

Link to this article

www.jeom.org/article/en/10.11836/JEOM21216

and the experts were constantly assessed for consistency in the labeling process based on an expectation-maximization algorithm. The labeled data were cleaned, archived, and preprocessed, and then were grouped into a training set and a verification set. Three deep convolutional neural network models TMNet, ResNet-50, and ResNeXt-50 were constructed and trained by ten-fold cross-validation method to obtain an optimal model. Five hundred cases of DR chest radiographs that were not included in the training set and the validation set were collected, and identified by five senior experts as the gold standard, named the test set. The accuracy rate, sensitivity, specificity, area under curve (AUC), and other indexes of the three models were derived after testing, and the efficiency of the three models was evaluated and compared.

[Results] A total of 24 867 DR chest radiographs of the training set and the validation set were collected in this study, including 6 978 images in the positive group and 17 889 images in the negative group. There were 312 cases of pulmonary abnormalities such as pneumothorax and pulmonary tuberculosis. A total of nine experts labeled the chest radiographs, the labeling consistency rate of pneumoconiosis (non-staging) was above 88%, and the labeling consistency rate of pneumoconiosis staging ranged from 84.68% to 93.66%. The diagnostic accuracy, sensitivity, specificity, and AUC of TMNet were 95.20%, 99.66%, 88.61%, and 0.987, respectively. The indicators of ResNeXt were 87.00%, 89.93%, 82.67%, and 0.911, respectively. Those of ResNet were 84.00%, 85.91%, 81.19%, and 0.912, respectively. All these indexes of TMNet were higher than those of ResNeXt-50 and ResNet-50 models. The AUC differences between TMNet and the other two models were both statistically significant ($P < 0.001$).

[Conclusion] All the three convolutional neural network models can effectively diagnose the presence of pneumoconiosis, among which TMNet provides the best efficiency.

Keywords: pneumoconiosis; convolutional neural network; artificial intelligence; deep learning

尘肺病是全世界一种常见的职业病,是由于职业活动中长期吸入生产性矿物性粉尘并且在肺内滞留而引起的以肺组织弥漫性纤维化为主的疾病,发病率高且不可逆转。依据《2019 年我国卫生健康事业发展统计公报》,当年全国共报告 19 428 例各类职业病新病例,其中职业性尘肺病 15 947 例,全国累计报告职业病病例 99 万余例,其中尘肺 89 万余例,约占 90%,尘肺病在我国发展依旧迅猛^[1]。依据 GBZ 70—2015《职业性尘肺病的诊断》,诊断尘肺病的关键是正确判读 X 线胸片,根据小阴影密集度、肺区分布等确定尘肺分期。但人工阅片对医生水平要求较高,标准片与实际病例情况存在差距;部分地区医疗资源匮乏,缺乏足够的尘肺病诊断医师;加上疲劳看片等因素的影响,很容易出现尘肺病误诊漏诊的情况。

针对这一难题,国内外专家学者试图通过人工智能技术来解决,如 Zhu 等^[2]使用支持向量机来实现尘肺病数字 X 射线摄影(digital radiography, DR)图像的分类;Okumura 等^[3]采用人工神经网络识别 X 线胸片,判断是否存在尘肺病。相较于传统的人工诊断,医疗人工智能技术多为机器学习和深度学习的技术,利用大批量高质量数据和计算机算力,诊断结果具有较好的灵敏度和特异度^[4]。其中深度学习属于一个框架,包含多种重要算法,如卷积神经网络^[5]、限制玻耳兹曼机^[6]、循环神经网络^[7]等,目前医学图像的处理主要是依据卷积神经网络。卷积神经网络模型多为多层神经网络,即深度卷积神经网络,包括 ResNet^[7]、DenseNet^[8]、Inception^[9]等。

本研究提出一种基于深度卷积神经网络的尘肺病人工智能诊断方法,包括通用的 ResNet-50 模型、ResNeXt-50 模型^[10]和自行构建的 TMNet 模型,对标注胸片进行训练。根据测试集测试结果比较三种模型的诊断效能,选出最优模型。本研究旨在利用高质量标注数据和卷积神经网络实现对 DR 胸片有无尘肺病的诊断。

1 材料与方法

1.1 资料的选择

此次研究收集的胸片数据是应急总医院、乌海市职业病防治院、河南煤炭总医院、贵州林东总医院、同煤集团职业病防治院、赣南职业病防治医院和宁夏第五人民医院于 2017 年 6 月—2020 年 12 月期间累计上传的历史病例,数据来源为门诊病人、住院病人和职业健康体检筛查人群。7 家医院的 DR 机器来自不同的厂家,包括进口的通用电器和西门子,国产的东软,上传数据按规定对临床数据进行脱敏。纳入的数据主要包括尘肺病诊断阳性胸片和无尘肺胸片,将尘肺诊断阳性胸片作为阳性组,无尘肺胸片作为阴性组。阳性组中的尘肺病种类包括矽肺、煤工尘肺、水泥尘肺、石棉尘肺和其他种类的尘肺病。本研究获得应急总医院伦理委员会的批准,审批编号为 L21-1。

纳入标准:因为在实际的尘肺病诊断过程中许多尘肺病人有并发症,故此次研究收集的胸片数据中肺癌、肺气肿、气胸、活动性肺结核等其他肺部情况均被纳入。排除标准:①侧位胸片数据,其他部位的

DR 图像; ②年龄 < 18 岁、> 80 岁的患者的 DR 胸片;
③GBZ 70—2015《职业性尘肺病的诊断》附录 C 标准中的三级片及四级片。

1.2 胸片的质量控制

邀请 1 位有数十年影像科工作经验和尘肺病阅片经验的医师前往合作医院指导胸片摄影, 达到控制胸片质量的目的。具体措施包括: 组织合作医院影像科相关摄片技师参与学习 GBZ 70—2015《职业性尘肺病的诊断》; 调整拍摄参数, 优化图像质量, 设置拍摄模版; 对技师拍摄胸片的错误技术和操作手法予以纠正。

由 2 位副主任医师及以上职务且拥有国家尘肺病诊断资质的医师依据 GBZ 70—2015《职业性尘肺病的诊断》附录 C“胸片质量与质量评定”, 使用灰阶显示器(DG-500A, 德为, 中国)对胸片质量单独进行评定。对于有疑问的胸片质量, 两者协商讨论, 根据综合评定结果, 选择二级及以上的胸片纳入数据库中。

1.3 标注医师的选择和标注方法

因为本研究计划标注的 DR 胸片数量超过 2 万例, 需要标注专家的数量为 10 位左右, 且所有标注专家均为副主任医师及以上职称并经过考核。为提高标注的准确性和灵敏度, 防止出现记忆偏倚, 本研究采用多轮次分组交叉对数据进行标注, 并在标注过程中进行一致性检验。

参考尘肺病诊断医师读片考核方式^[11], 测试题库共 100 例胸片, 各期尘肺胸片共 50 例, 无尘肺胸片 50 例, 所有胸片均经过另外一批资深尘肺病诊断专家集体读片作为标准答案, 邀请尘肺病标注专家独立、单盲法完成读片。测试满分 100 分, 每例胸片价值 1 分。若尘肺分期判断正确记 1 分, 尘肺病有无诊断正确记 0.5 分, 其他错误情况不得分。总分达到 95 分以上者为合格, 将合格专家纳入标注专家库中。分期判断越准确, 专家对胸片的标注准确率越高, 人工智能的判断也会越准确。

为确保标注专家的一致性, 根据最大期望算法^[12]和《尘肺病数据标注规范与质量控制专家共识》^[13], 标注数量每达 1000 例时进行一次标注一致性检验, 其中有约 100 例是各位专家均标注过的共同数据, 共同数据中阴阳性分布均衡。具体流程如下:

①计算多名医师标注一例胸片的投票结果 l_0 。

②在步骤①的基础上计算每个医师的初始一致性率 P_{mi} , 即某医师的标注结果和步骤①中标注结果相同的胸片数占该医师标注胸片总数的比例。

③将每位医师对于同一例胸片的标注结果, 乘以自己的一致性率 P_{mi} , 得到标注结果系数 sl_i , 然后累加相同标注结果标签的系数, 累加结果最大额标签为胸片结果 sl_0 。

④以步骤③推算出来的每例胸片的标注结果为准, 推算每个医师标注的新的一致性率 P_{mi} 。将其标注结果与步骤②中得到的胸片标注结果对比, 结果相同的数量再除以医师总标注数量, 见公式(1)。

$$P_{mi} = \frac{\sum_{k=1}^n (sl_i^k - sl^k)}{n} \quad (1)$$

式中: P , 一致性率; i , 标注医师编号; l , 标注结果; sl , 标准结果; k , 标注影像编号; m , 一致性迭代轮次; n , DR 胸片数量。

⑤最后使用步骤④中得到的新的一致性率, 不断重复步骤②和步骤③, 直到一致性率收敛为止。

标注时参考 GBZ 70—2015《职业性尘肺病的诊断》和尘肺病标准片, 具体参数包括胸片质量、小阴影形态、肺区小阴影密集度、病变范围(肺区数)、总体密集度、小阴影聚集、大阴影、胸膜斑、心影蓬乱、尘肺分期和附加符号。标注并不在胸片图像上进行勾选, 而是给胸片一个标签(或说明), 后续人工智能根据标签和对应的胸片进行训练, 得到结果。

对于多位专家共同的标注数据, 按照各张胸片投票数的多寡进行判定, 选择投票数最多的标签为最终标注结果。若是各投票数相同, 则邀请资深专家仲裁作为最终结果。

1.4 数据库的建立和数据的预处理

标注后的数据无法直接纳入数据库, 对其进行数据清洗和数据归档^[14]。归档后图像按照数据增广^[15]、尺寸归一化^[16]和格式转换这一顺序进行预处理。

1.5 模型的训练

本研究采用三种模型: ResNet-50、ResNeXt-50 和 TMNet。其中 ResNet-50、ResNeXt-50 为通用的深度卷积神经网络模型, TMNet 则是在 ResNet 的基础上引入全卷积网络的反卷积思想^[17]形成的。研究基于惠普的小型服务器, 算力为 2 张 Nvidia Tesla V100 32G, 深度学习框架使用 Pytorch。在模型训练完成后, 将测试集胸片输入模型中, 模型根据程序会输出一个 0~1 之间的数值, 若截断值为 0.5, 则 > 0.5 的胸片为尘肺胸片, < 0.5 的胸片视为非尘肺胸片, 通过改变截断值可以调整模型的灵敏度和特异度等指标。

本研究模型训练的方式是十折交叉验证^[18]。在数据库中分层随机选取 90% 的数据作为训练集, 剩余

10%的数据作为验证集进行训练,每次训练均会得出相应准确率的模型,多次进行十折交叉验证,计算均方误差的平均值,选出误差最小的模型作为最终的模型。

1.6 测试集的构建

选择5名在尘肺病诊断方面有超过15年经验的资深专家组成专家组,新收集未纳入尘肺数据库的DR胸片共500张,所有胸片由5位专家共同阅片,对于诊断有异议的DR胸片协商确定。收集的胸片中尘肺病例占比不少于50%,尘肺病期别按照壹期数量最多,贰期次之,叁期最少分布。胸片中尘肺病种类包括煤工尘肺、矽肺和石棉肺,所有收集胸片中没有其他肺部异常情况。为了降低测试难度,所以没有纳入肺部异常情况。

1.7 统计学分析

模型的输出结果是以一个0~1之间的数值作为变量,测试集中诊断为尘肺标为1,无尘肺标为0作为分类变量。将变量和分类变量输入到MedCalc 19.7.2软件中可以得到一条受试者工作特征(receiver operating characteristic, ROC)曲线,由ROC曲线最佳临界点确定截断值,得到准确率、灵敏度、特异度、阳性似然比、阴性似然比、受试者工作特征曲线下面积(area under curve, AUC)值、F1值 [$F1=2 \times \text{阳性预测值} \times \text{灵敏度} / (\text{阳性预测值} + \text{灵敏度})$],是衡量二分类模型精确度的一种指标,被看作是模型灵敏度和阳性预测值的一种调和平均],来评价模型的诊断性能。不同模型之间的AUC值比较采用DeLong检验。除AUC比较外,所有数据采用SPSS 25.0进行分析,其中分类变量用百分比描述,连续变量用 $\bar{x} \pm s$ 描述。假设检验均为双边,检验水准 $\alpha=0.05$ 。

2 结果

2.1 一般资料和胸片情况

本研究共收集训练集与验证集DR胸片24867张,其中阳性组6978例(壹期3943例、贰期1751例、叁期1284例,分别占比56.51%、25.09%、18.40%),阴性组17889例。年龄范围为18~80岁,平均年龄为(56.0±3.5)岁;煤工尘肺+矽肺占比为90.52%,石棉肺为5.74%。DR胸片质量中一级片占比25.19%,二级片占比74.81%。

测试集胸片共500张,其中阳性组298例(59.60%),阴性组202例(40.40%)。阳性组中尘肺病壹期201例,贰期68例,叁期29例,分别占比67.45%、22.82%、

9.73%。胸片质量方面,一级片数量103例(20.60%),二级片数量397例(79.40%)。

在训练集和验证集标注的胸片中,其他肺部情况如结核、肺气肿等共有15类,总数量为312例。其中活动性肺结核数量最多,为93例,肺大泡、肺癌和胸膜间皮瘤、胸腔积液、淋巴结蛋壳样钙化、类风湿性尘肺数量均低于10例,见表1。

表1 训练集和验证集胸片中其他肺部情况分布(n=312)

Table 1 Distribution of other chest radiograph findings in the training set and the validation set (n=312)

标注符号	肺部情况	数量
bu	肺大泡	7
ca	肺癌和胸膜间皮瘤	8
cn	小阴影钙化	31
cp	肺心病	10
cv	空洞	20
ef	胸腔积液	3
em	肺气肿	15
ss	淋巴结蛋壳样钙化	2
ho	蜂窝肺	18
pc	胸膜钙化	13
pt	胸膜增厚	22
px	气胸	25
rp	类风湿性尘肺	9
tb	活动性肺结核	93
ot	除上述外的其他情况	36

2.2 标注医师的一致性

本次研究过程中共有9名专家对胸片进行了标注,分别计算他们对尘肺病分期标注的一致性和尘肺异常(不分期)标注的一致性。所有专家的尘肺异常(不分期)的标注一致性率均在88%以上,相对而言,尘肺分期的标注一致性率较低,最低为84.68%。见表2。

表2 专家标注的一致性

Table 2 Consistency of expert labeling

专家	尘肺分期的标注一致性率/%	尘肺异常的标注一致性率/%
专家1	93.66	93.66
专家2	90.38	95.19
专家3	89.43	93.29
专家4	89.12	90.33
专家5	88.27	90.39
专家6	85.94	88.13
专家7	84.75	90.85
专家8	84.68	89.06
专家9	85.62	88.53

2.3 模型训练结果对比

以测试集作为金标准,三种模型的诊断效能如表3所示。TMNet模型相较于另外两个模型,F1、诊断准确率、灵敏度、特异度、阳性似然比、阴性似然比、AUC值等指标均更高。除AUC外,ResNeXt-50模型相比ResNet-50模型其余指标均更高。

表3 三种模型尘肺病诊断效能的对比

Table 3 Comparison of diagnostic efficacy of pneumoconiosis by three models

效能指标	TMNet	ResNeXt-50	ResNet-50
约登指数	0.88	0.72	0.67
F1/%	96.11	89.18	86.48
准确率(95%CI)/%	95.20(92.90~96.80)	87.00(83.70~89.70)	84.00(80.50~86.90)
灵敏度(95%CI)/%	99.66(98.10~100.00)	89.93(85.90~93.10)	85.91(81.40~89.60)
特异度(95%CI)/%	88.61(83.40~92.60)	82.67(76.70~87.60)	81.19(75.10~86.30)
阳性似然比(95%CI)	8.75(6.00~12.90)	5.19(3.80~7.00)	4.57(3.40~6.10)
阴性似然比(95%CI)	0.01(0.01~0.03)	0.12(0.09~0.20)	0.17(0.10~0.20)
AUC(95%CI)	0.987(0.972~0.995)	0.911(0.883~0.935)	0.912(0.884~0.935)

表4显示,经DeLong检验,TMNet模型与另外两个模型的AUC差值相比差异均有统计学意义($P < 0.001$),而ResNet-50和ResNeXt-50之间差异无统计学意义。

表4 三种模型尘肺病诊断的AUC两两比较

Table 4 Pairwise comparison of AUC in the diagnosis of pneumoconiosis by three models

模型	AUC差值(95%CI)	标准差	Z	P
TMNet vs. ResNeXt-50	0.075(0.047~0.103)	0.014	5.24	<0.001
TMNet vs. ResNet-50	0.074(0.052~0.097)	0.012	6.36	<0.001
ResNeXt-50 vs. ResNet-50	0.001(-0.036~0.037)	0.019	0.04	0.969

3 讨论

本研究通过收集DR胸片训练了三种深度卷积神经网络模型,用于尘肺病的人工智能诊断。测试结果显示,三种模型中最优的模型为TMNet模型,准确率达到95.20%,AUC值为0.987,灵敏度和特异度分别为99.66%和88.61%。对尘肺病的不分期诊断,TMNet模型在保证较高特异度的同时明显提高了诊断灵敏度。北京某医院两位尘肺病医师(尘肺病诊断经验分别为<5年和>10年)诊断DR胸片的灵敏度分别为56%和76%^[19]。与尘肺病诊断医师的人工诊断结果相比,TMNet模型的诊断效能达到了尘肺病初筛的标准。TMNet模型目前可以作为计算机辅助诊断使用^[20],医生可以结合人工智能的判断结果对病人进行诊断。

因为模型诊断灵敏度特别高,代表漏诊概率较低,可以大大减轻职业病诊断医师的工作压力,提高诊断效率。

本次研究的三种模型均为深度卷积神经网络模型,相较于其他人工智能诊断尘肺病的方法,如传统机器学习方法支持向量机^[2]和深度学习方法人工神经网络,深度卷积神经网络模型诊断效果更优^[3]。支持向量机和人工神经网络均需要采用传统的手工特征提取方法且训练集可以是小样本量,但手工提取的特征常常不是最优的,而且支持向量机多用于二分类的情况,人工神经网络则容易产生过拟合情况,复杂图片分类效果比较差,且训练难度随着层数的增加而增大。深度卷积神经网络抛弃传统的手工特征提取方法,直接将图像输入模型中经过卷积、池化等过程输出结果,更为方便且分类效果更好。但是卷积神经网络必须要大量的数据进行训练,且随着数据量的增加,训练效果会更好,如果数据量不够,训练效果会较差^[21]。国内外应用卷积神经网络实现尘肺病的诊断的研究相对较少,例如王峥等^[19]利用ResNet进行尘肺病诊断,AUC值可达0.997。但是其数据均为一家医院收集的数据且测试集和训练集为同一数据来源,可能存在过拟合情况。

本研究存在一些局限性。首先是没有做尘肺病的鉴别诊断。在实际的尘肺病诊断过程中,常常有气胸、肺结核、肺气肿等肺部异常情况存在,收集含有这些肺部异常的胸片可以得到一个更贴近实际情况的数据集,提高模型的鉴别诊断能力,增强模型的泛化性。但是本次研究过程专家标注出的肺部异常仅312例。回顾研究标注过程,发现研究初始未对专家强调肺部异常这一标注内容,部分专家在标注过程中忽视了这一项而没有标注,还有部分专家无法确定阴影是何种肺部异常情况,由此导致本次研究标注的肺部异常数据相对于整个模型来说数据量太小,模型训练无法达到期望效果。故本次研究不涉及尘肺病的鉴别诊断,测试集胸片均无其他肺部异常情况。未来研究中将继续收集胸片,对标注专家强调这部分信息标注的必要性。待肺部异常数据提高到一定数量,模型经过不断训练,可以实现尘肺病的鉴别诊断。其次,黑箱效应仍是优化深度学习模型时的难点。虽然人工智能诊断已经达到了较高的准确性,并且通过类激活可视化分析能够检测模型是否学习到相关性,但是人工智能如何得到结果仍然不清楚,可解释性比较差^[22]。最后,卷积神经网络模型训练需要大量的训练数据,如果缺乏数据会影

响卷积网络模型的性能。本研究中模型的标注需要专家的大量阅片,对于尘肺分期标注的一致性相较于异常(不分期)标注的一致性更低,导致研究仅能实现DR胸片有无尘肺病的二分类诊断,无法满足实际的尘肺病诊断应用。未来需要研究如何提高标注一致性以提升模型的训练效果,使得模型能够精准识别尘肺胸片的密集度与肺区分布,实现尘肺病的分期诊断。

本研究中构建的三种卷积神经网络能够客观且稳定地诊断尘肺病的有无,其中TMNet模型的诊断效能较另两个模型更好,获得了更高的准确度、灵敏度和特异度。根据本研究结果,深度卷积神经网络模型在尘肺病的诊断中迈出了探索性的一步,为将来实现客观、高效的尘肺病人工智能分期诊断提供了基础。

参考文献

- [1] 规划发展与信息化司. 2019年我国卫生健康事业发展统计公报[EB/OL]. [2021-04-28]. <http://www.nhc.gov.cn/guihuaxxs/s10748/202006/ebfe31f24cc145b198dd730603ec4442.shtml>.
Department of Planning, Development and Information Technology. Statistical communique on the development of China's health services in 2019[EB/OL]. [2021-04-28]. <http://www.nhc.gov.cn/guihuaxxs/s10748/202006/ebfe31f24cc145b198dd730603ec4442.shtml>.
- [2] ZHU B, LUO W, LI B, et al. The development and evaluation of a computerized diagnosis scheme for pneumoconiosis on digital chest radiographs[J]. *Biomed Eng Online*, 2014, 13(1): 141.
- [3] OKUMURA E, KAWASHITA I, ISHIDA T. Computerized classification of pneumoconiosis on digital chest radiography artificial neural network with three stages[J]. *J Digit Imaging*, 2017, 30(4): 413-426.
- [4] PANAYIDES AS, AMINI A, FILIPOVIC ND, et al. AI in medical imaging informatics: Current challenges and future directions[J]. *IEEE J Biomed Health Inform*, 2020, 24(7): 1837-1857.
- [5] SCHMIDHUBER J. Deep learning in neural networks: an overview[J]. *Neural Netw*, 2015, 61: 85-117.
- [6] SRIVASTAVA N, SALAKHUTDINOV R. Multimodal learning with deep Boltzmann machines[J]. *J Mach Learn Res*, 2014, 15(1): 2949-2980.
- [7] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE, 2016: 770-778.
- [8] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE, 2017: 2261-2269.
- [9] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions[C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA: IEEE, 2015: 1-9.
- [10] XIE S, GIRSHICK R, DOLLÁR P, et al. Aggregated residual transformations for deep neural networks[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE, 2017: 5987-5995.
- [11] 邢军, 李岩, 孙大为, 等. 吉林省尘肺病诊断医师读片考核结果分析[J]. *海峡预防医学杂志*, 2015, 21(6): 77-78.
XING J, LI Y, SUN DW, et al. Analysis of examination results of reading films of doctors diagnosing pneumoconiosis in Jilin province[J]. *Strait J Prev Med*, 2015, 21(6): 77-78.
- [12] MCLACHLAN G J, KRISHNAN T. The EM algorithm and extensions[M]. 2nd ed. Hoboken: John Wiley & Sons, 2007: 1-4.
- [13] 李涛, 张建芳, 孟祥峰, 等. 尘肺病数据标注规范与质量控制专家共识(2020年版)[J]. *环境与职业医学*, 2020, 37(6): 523-529.
LI T, ZHANG JF, MENG XF, et al. Chinese expert consensus on pneumoconiosis data labeling specifications and quality control (2020 edition)[J]. *J Environ Occup Med*, 2020, 37(6): 523-529.
- [14] DZIADKOWIEC O, CALLAHAN T, OZKAYNAK M, et al. Using a data quality framework to clean data extracted from the electronic health record: a case study[J]. *eGEMs (Wash DC)*, 2016, 4(1): 11.
- [15] KRELL M M, KIM S K. Rotational data augmentation for electroencephalographic data[C]//Proceedings of the 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Jeju, Korea (South): IEEE, 2017: 471-474.
- [16] 彭思淇, 田林晓, 朱苗苗, 等. 低成本光学字符识别读表系统研制[J]. *仪器与设备*, 2017, 5(3): 55-62.
PENG SQ, TIAN LX, ZHU MM, et al. Develop of low cost meter reading system by optical character recognition[J]. *Instrum Equip*, 2017, 5(3): 55-62.
- [17] KODEWITZ A, LELANDAS S, MONTAGNE C, et al. Alzheimer's disease early detection from sparse data using brain importance maps[J]. *ELCVIA*, 2013, 12(1): 42-56.
- [18] KOHAVI R. A study of cross-validation and bootstrap for accuracy estimation and model selection[C]//Proceedings of the 14th International Joint Conference on Artificial Intelligence. Montreal, Quebec, Canada: ACM, 1995: 1137-1145.
- [19] 王峥, 贺文. 深度残差网络在尘肺病诊断中的应用初探[J]. *中国工业医学杂志*, 2019, 32(1): 31-33.
WANG Z, HE W. Primary exploration on application of ResNet in deep learning of pneumoconiosis diagnosis[J]. *Chin J Ind Med*, 2019, 32(1): 31-33.
- [20] WANG S, BURTT K, TURKBEY B, et al. Computer aided-diagnosis of prostate cancer on multiparametric MRI: a technical review of current research[J]. *J Biomed Res Int*, 2014, 2014: 789561.
- [21] YASAKA K, AKAI H, ABE O, et al. Deep learning with convolutional neural network for differentiation of liver masses at dynamic contrast-enhanced CT: a preliminary study[J]. *Radiology*, 2018, 286(3): 887-896.
- [22] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-cam: visual explanations from deep networks via gradient-based localization[C]//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017: 618-626.

(英文编辑: 汪源; 责任编辑: 汪源)