公共人工智能平台在膝关节骨性关节炎分期中的应用

赵晓阳, 许树林, 潘为领, 唐慧勇, 张守波
(中国人民解放军第 960 医院淄博医疗区, 山东 淄博, 255300)

摘 要: 目的 探讨利用 X 线片在公共人工智能平台上训练模型对膝关节骨性关节炎(KOA)严重程度自动分期的可行性。**方法** 选取按照 Kellgren-Lawrence (KL)分期系统进行分期的 X 线片,在公共人工智能平台上训练模型。最终使用了 1 445幅图像进行自动训练及测试评估。使用 50 幅图像的测试集对模型和放射科医师进行测试,计算放射科医师的准确率和 F1-score, 并与人工智能平台中模型返回的结果进行比较。**结果** 模型对人工智能平台自动训练集的准确率为 0.73, F1-score 为 0.72; 模型对 50 幅图像的测试子集的准确率为 0.70, F1-score 为 0.69。放射科医师测试的准确率为 0.64, F1-score 为 0.63。模型效能达到甚至超过了高年资放射科医师测试水平。**结论** 基于公共人工智能平台进行模型训练,利用 X 线图像进行 KOA 的自动 KL 分期,具有可行性和一定的优越性。

关键词: 人工智能; X 线片; Kellgren-Lawrence 分期系统; 膝关节骨性关节炎; 骨关节炎倡议
中图分类号: R 816.8; R 684.3 **文献标志码:** A **文章编号:** 1672-2353(2022)08-022-05 DOI: 10.7619/jcmp.20212888

Application of public artificial intelligence platform in staging of knee osteoarthritis

ZHAO Xiaoyang, XU Shulin, PAN Weiling, TANG Huiyong, ZHANG Shoubo
(*Zibo Medical District of the* 960*th Hospital of Chinese People's Liberation Army, Zibo, Shandong*, 255300)

Abstract: Objective To explore the feasibility of automatic grading of knee osteoarthritis (KOA) severity by using X-ray film training model on public artificial intelligence platform. **Methods** The Kellgren-Lawrence (KL) staging system was selected to determine stages of the X-ray films, and the model was trained on a public artificial intelligence platform. Finally, 1 445 images were used for automatic training and test evaluation. A test set of 50 images was used to test the model and the radiologists, and accuracy and F1-score of the radiologists were calculated and compared with the results returned by the model in the artificial intelligence platform. **Results** The accuracy of the model to the automatic training set of artificial intelligence platform was 0.73 and F1-score was 0.72; the accuracy of the model was 0.70 and F1-score was 0.69 for the test subset of 50 images; the accuracy of the radiologists test was 0.64 and F1-score was 0.63. Model performance matched or even exceeded that of senior radiologists. **Conclusion** It is feasible and advantageous to train the model based on public artificial intelligence platform and use X-ray image to perform automatic staging of KOA by KL.

Key words: artificial intelligence; X-ray image; Kellgren-Lawrence staging system; knee osteoarthritis; osteoarthritis initiative

膝关节骨性关节炎(KOA)是常见的慢性退行性骨关节病,以疼痛和功能障碍为特征[1-2]。KOA 严重影响患者生活质量,给其家庭和社会造成严重的经济负担[3]。KOA 准确分期可避免患者病程的快速进展[4]。目前,骨关节炎(OA)常用的 3 种分期方法[5-6]中, Kellgren-Lawrence (KL)分期使用最广泛,其结果与疼痛及功能障碍相关,且术前 KL 分期能预测手术成功率。但人工分期不仅耗时,还会因个人主观偏倚存在差异。传统人工智能模型的建立需要大量的专家和资源,普通医师难以获得。近年来,谷歌、百度等集团提供了公共人工智能云平台,使普通医师能够在没有人工智能经验的情况下建立人工智能模型。但关于公共人工智能平台对 KOA 自动分期效果的研究较少。故本研究探讨利用公共人工智能平台对 KOA 严重程度自动分期的可行性。

收稿日期: 2021 - 07 - 19
通信作者: 张守波, E - mail: 2983790749@qq.com

1　资料和方法

1.1　一般资料

骨关节炎倡议(OAI)是有关 KOA 研究的公共数据库,其可供公众调阅使用。拍摄 X 线片：由 2 名训练有素的肌骨放射学医师使用 KL 系统[7-9]对每张 X 线片的每个关节进行分期。如有分歧,则由第 3 位医师协议解决,最终公布的为共识结果。

本研究下载了一组按照 KL 分期完成分组的数据。训练模型最多时使用了其中 5 777 个关节 X 线片,并在训练过程中进行了不同数据集的多次训练,以求获得最佳的模型方案,模型训练在百度公共人工智能平台 EasyDL(https：//ai. baidu. com/easydl/)中进行。该平台提供免费的人工智能培训、评估和基于图像的预测、分类。该平台计算能力出色,每个模型都可在 20 min 内完成训练并进行自我评估[11-14]。平台分别随机选择图像进行训练,并使用约为上传数据集 30% 的数据进行自我评估。之后,平台返回本研究模型整体的准确率、F1-score、精确率、召回率以及按照每一个 KL 分期的 F1-score，用以评价模型价值。训练过程中,本研究通过调整图像数据集,训练了多个迭代版本,最终得到效果最好的模型。其中不同的训练集以 OAtrain 加不同下标命名,例如 OAtrain 5.0，各版本所用训练及测试数据集详情见表 1。

1.2　医师评价

为了判断 OAI(作为标准)、人工智能模型和单个医师之间的评分一致性,本研究选取了 2 名高年资放射科副主任医师进行 KL 评分(在应用 KL 评分系统方面具有多年经验)。然后按照 KL 分期,从每个 KL 等级分别随机选择 10 张 X 线照片,构成总量为 50 张图像的测试集(命名为 50-test)，以供上述医生评分。

表 1　训练及测试数据集详细信息

用途及命名		图像数量					
		0	1	2	3	4	合计
5 级(KL 0 ~ 4 期)	OAtrain5.0	2 285	1 046	1 516	757	173	5 777
	OAtrain5.1	290	290	290	290	285	1 445
2 级(发病与否)	OAtrain2.0	580	865	—	—	—	1 445
测试	50-test	10	10	10	10	10	50

50-test：图像与训练数据集均无重复。

1.3　数据预处理及模型训练

训练使用的图像格式为 PNG，分辨率为 299 像素 ×299 像素。利用多个不同的数据集和多个训练参数进行组合训练,得到多个迭代版本。

1.4　效能评估

使用 F1-score 和加权 *Kappa* 系数进行效能评估,其可反映整体及每个 KL 分类的效能,而且其他关于 KL 分期的研究中也使用了该参数,使得本研究能够与之进行比较。F1-score 对某类别而言为精确率和召回率的调和平均数,范围为 0 ~ 1，其中 1 表示完全一致。对于多类分类,平台及本研究分别计算每个分类的 F1-score，并对结果进行平均。*Kappa* 系数的大小用来衡量 2 种方法的一致程度，*Kappa* 系数越大说明 2 种结果越一致,若 *Kappa* ≥0. 75，说明结果一致性较好,若 *Kappa* <0.40，说明缺乏一致性[14]。

1.5　统计学分析

使用 SPSS 26.0 及 Python 3.8 软件进行数据分析,计算加权 *Kappa* 系数、准确率、召回率和 F1-score，并对结果进行直接比较。

2　结　果

2.1　模型训练

以 OAI 的分期结果为标准,在百度公共人工智能平台,经过多次迭代训练,各版本效能结果显示，5 级 V4 版本效能最好,其中 F1-score 为 0.72，准确率为 0.73，见图 1、表 2。

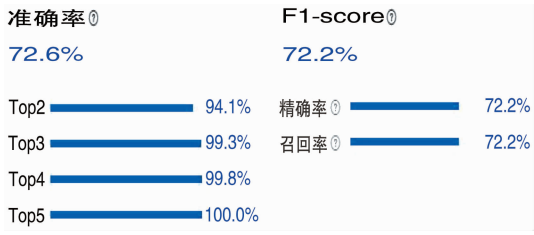

图 1　5 级 V4 版本平台整体评估结果截图

2.2　KL 分期系统效能比较

对于 50-test 测试子集,本研究 2 位医师的 F1-score 和准确率分别为 0.63 和 0.64。模型对该测试子集的F1-score为0.69,准确率为0.70。

表 2 各版本效能统计

版本	数据集	算法	结果			
			准确率	F1-score	精确率	召回率
5 级 V1	OAtrain5.0	高精度	0.63	0.65	0.65	0.65
5 级 V2	OAtrain5.0	AutoDL Transfer	0.63	0.64	0.66	0.63
5 级 V3	OAtrain5.1	高精度	0.69	0.67	0.68	0.69
5 级 V4	OAtrain5.1	AutoDL Transfer	0.73	0.72	0.72	0.72
发病率模型的训练一代	OAtrain2.0	AutoDL Transfer	0.91	0.91	0.91	0.92

模型对单个 KL 分期 0 期、3 期和 4 期的 F1-score 超过了医师，而医师的 KL 分期为 2 期的 F1-score 更高，另外对 KL 分期为 1 期的 F1-score 两者相等。这些结果可与 THOMAS K A 等[10]报告的 F1-score 进行直接比较。同时，因为子集包含来自每个 KL 分类的相等数量的图像，所以这些得分结果可以直接与 ANTONY J 等[12]研究中报告的加权 F1-score 进行比较。见表 3。

2.3 发病率效能比较

KL 分期系统中 2 期特别重要，因为在使用 KL 分期系统进行队列选择时，其经常被用作确定 OA 发病率的阈值[10]。为了评估模型，确定 OA 发病率模型的效能，本研究将 0 期和 1 期的 KL 评分合并到一个类别中，并将 2 期、3 期和 4 期的 KL 评分合并到另一个类别中。本研究对此在百度人工智能平台中重新训练了相应的模型，该模型的总体 F1-score 和准确率均达到了 0.91，而针对 50-test 测试子集，F1-score 为 0.89，准确率为 0.90。2 位医师对此效能的得分为 F1-score 为 0.87，准确率为 0.88。见表 4。

表 3 在 KL 分期系统中医师及各模型效能比较

评估来源	测试集	KL 分期	精确率	召回率	F1-score	准确率
医师	50-test	0	0.72	0.80	0.76	0.64
		1	0.38	0.30	0.33	
		2	0.60	0.60	0.60	
		3	0.67	0.60	0.63	
		4	0.75	0.90	0.82	
		平均	0.62	0.64	0.63	
模型	50-test	0	0.75	0.90	0.82	0.70
		1	0.38	0.30	0.33	
		2	0.55	0.60	0.57	
		3	0.89	0.80	0.84	
		4	0.90	0.90	0.90	
		平均	0.69	0.70	0.69	
	平台测试集	0	0.83	0.89	0.86	0.73
		1	0.62	0.58	0.60	
		2	0.57	0.57	0.54	
		3	0.71	0.73	0.72	
		4	0.89	0.85	0.87	
		平均	0.73	0.72	0.72	
THOMAS K A 等[10]报告	完整测试集	0	0.73	0.87	0.79	0.71
		1	0.38	0.27	0.31	
		2	0.71	0.67	0.69	
		3	0.82	0.81	0.81	
		4	0.87	0.86	0.87	
		平均	0.70	0.69	0.70	
ANTONY J 等[12]报告	—	0	0.57	0.92	0.71	0.60
		1	0.32	0.14	0.20	
		2	0.71	0.46	0.56	
		3	0.78	0.73	0.76	
		4	0.89	0.73	0.80	
		平均	0.61*	0.62*	0.59*	

*ANTONY J 等[12]报告的平均精确率、召回率和 F1-score 是根据其样本中每个 KL 分期的频率进行加权计算得出。

平台测试集准确率原始数据为 308/424，完整测试集准确率原始数据为 2 890/4 090。

表 4　在发病率判断中医师及各模型效能比较

评估来源	测试集	精确率	召回率	F1-score	准确率
医师	50-test	0.89	0.87	0.87	0.88
模型	50-test	0.90	0.89	0.89	0.90
	平台测试集	0.91	0.91	0.91	0.91
THOMAS K A 等[10]报告	完整测试集	0.89	0.85	0.87	0.87

平台测试集准确率原始数据为 385/424，完整测试集准确率原始数据为 3 568/4 090。

2.4　一致性评估

在以 OAI 为标准一致性评估时，在 50-test 测试子集中医师的加权 *Kappa* 系数为 0.76，此测试子集的模型获得的加权 *Kappa* 系数为 0.82，模型对完整测试集的加权 *Kappa* 系数为 0.82，与 THOMAS K A 等[10]报告的 0.86 及 TIULPIN A 等[13]研究模型的最佳 *Kappa* 系数 0.83 相近，见表 5。

表 5　评价者与金标准一致性比较

数据	医师-OAI	模型-OAI		THOMAS K A 等[10]的模型	TIULPIN A 等[13]的模型
数据集	50-test	50-test	平台测试集	—	—
加权 *Kappa* 系数	0.76(0.66～086)	0.82(0.73～0.90)	0.82(0.79～0.85)	0.86(0.86～0.86)	0.83(0.83～0.83)

括号中的数据为 95% *CI*。

在评估评价者之间一致性时，医师与模型之间的加权 *Kappa* 系数分别为 0.75 和 0.74。医师之间的加权 *Kappa* 系数为 0.76，略低于 THOMAS K A 等[10]报告中的医师间加权 *Kappa* 系数 0.79，高于 RIDDLE D 等[15]报告中最一致的 2 个评价者之间的 *Kappa* 系数 0.65，见表 6。

表 6　不同研究医师间一致性比较

模型来源	加权 *Kappa* 系数
本研究	0.76(0.64～0.88)
THOMAS K A 等[10]研究	0.79(0.84～0.94)
RIDDLE D 等[15]研究	0.65(0.38～0.92)

括号中的数据为 95% *CI*。

3　讨　论

目前，KOA 的发病率日益增高，而其诊断和分期依据主要为影像学检查结果，因此进行准确的影像学分期，对 KOA 的治疗和预后有重要意义。本研究利用公共人工智能平台建立模型，实现对 KOA 的自动分期，并取得了良好的效果。

从本研究训练的模型的表现来看，无论是对 KOA 按照 KL 分期系统进行 5 期分期，还是在 KOA 发病率的判断上，本研究模型均取得较好的效能，许多表现达到甚至超过了本研究的高年资医师。本研究针对 KL 分期系统的整体效能达到 F1-score 为 0.72，准确率为 0.73，与之前研究中 THOMAS K A 等[10]模型的 F1-score(0.70)、准确率(0.71)相近。在发病率模型测试中，本研究模型 F1-score 为 0.91，优于 THOMAS K A 等[10]报道的 0.87，说明本模型在发病率判断中的表现较优。在各项一致性评估中，本研究的模型加权 *Kappa* 系数为 0.82，略低于 THOMAS K A 等[10]报告的 0.86 及 TIULPIN A 等[13]研究模型的最佳 *Kappa* 系数 0.83，但仍可表明其具有较好的一致性，与之前的研究差异较小。

本研究提出的临床医师利用公共人工智能平台训练模型和利用 X 线片对 KOA 进行自动 KL 分期具有可行性和一定的优越性。首先，由于模型是在云平台上自动、迅速地进行训练，因此其在普通个人的计算机上便可运行，不需要专门的、价格高昂的计算机设备及人工智能专业知识储备。本研究在百度公共人工智能平台所建立的模型的效能可以达到甚至超出经验丰富的医师的评估效能。其次，既往研究往往依赖于手动标注，对图像进行标注可能会增加噪声和错误的发生，并且需要额外的时间和人力成本。而本研究模型只需上传图片数据即可，其操作简单、便捷，即使毫无人工智能经验的医师也可进行操作。既往研究往往需要大量的原始图片数据，比如 THOMAS K A 等[10]研究总共使用了 40 280 张图像，而本研究最终使用 1 445 张图像进行模型训练，且取得了与其模型相当的效能结果。本研究还发现，提高模型训练效果的重要因素除增大数据量，还需每个子分类的数据量相当，这一点百度公共人工智能平台在训练时也进行了相应提示。在本研究模型训练过程中，子分类数据量比例失调的数据集得到的结果更好，且选择 AutoDL Transfer 算法，在

训练时间及效果上均具有良好表现，推荐在训练模型时选择此算法。

本研究使用的公共人工智能模型本质上是一个分类器，平台可以根据本研究提供的不同类别的图像进行模型训练。因此，公共人工智能平台的潜能不仅限于本研究范围内，在其他医疗领域范围同样具有巨大潜能。随着越来越多的公共人工智能平台出现，更多的基层普通医师可以获得人工智能服务。本研究认为，公共人工智能平台将促进医学和人工智能的共同发展。本研究仍具有一定局限性。首先，本研究使用相对较小的训练数据集来训练模型，随着训练数据集的增加，模型的性能可能会被影响。其次，本研究将模型性能与仅使用 50 张图像进行测试的医师的评估结果进行比较，医师测试样本相对较小，结果可能存在偏差。此外，本研究模型是针对标准的膝关节 X 线片设定，对一些特殊体位或不标准位置的图像的分类效果无法判断。

综上所述，本研究使用公共人工智能平台进行模型训练，利用 X 线图像进行 KOA 的自动 KL 分期，具有可行性和优越性，为利用人工智能平台进行临床研究与工作提供了良好依据。

参考文献：

[1] HERMAN A, CHECHIK O, SEGAL G, *et al*. The correlation between radiographic knee OA and clinical symptoms: do we know everything[J]. Clin Rheumatol, 2015, 34(11): 1955 – 1960.

[2] HUNTER D J, SCHOFIELD D, CALLANDER E. The individual and socioeconomic impact of osteoarthritis[J]. Nat Rev Rheumatol, 2014, 10(7): 437 – 441.

[3] 中华医学会骨科学分会. 骨关节炎诊治指南(2007 年版)[J]. 中华骨科杂志, 2007, 27(10): 793 – 796.

[4] 张伟强, 李波, 李帆冰. 膝骨性关节炎分期与疼痛部位的相关性研究[J]. 云南中医学院学报, 2016, 39(5): 78 – 81.

[5] KELLGREN J H, LAWRENCE J S. Radiological assessment of osteo-arthrosis[J]. Ann Rheum Dis, 1957, 16(4): 494 – 502.

[6] 王弘德, 李升, 陈伟, 等.《骨关节炎诊疗指南(2018 年版)》膝关节骨关节炎部分的更新与解读[J]. 河北医科大学学报, 2019, 40(9): 993 – 995, 1000.

[7] HAYES B, KITTELSON A, LOYD B, *et al*. Assessing radiographic knee osteoarthritis: an online training tutorial for the kellgren-Lawrence grading scale[J]. MedEdPORTAL, 2016, 12: 10503.

[8] ÖZDEN F, NADIYE K Ö, TUGAY N, *et al*. The relationship of radiographic findings with pain, function, and quality of life in patients with knee osteoarthritis[J]. J Clin Orthop Trauma, 2020, 11(Suppl 4): S512 – S517.

[9] MAZZUCA S A, BRANDT K D, SCHAUWECKER D S, *et al*. Severity of joint pain and Kellgren-Lawrence grade at baseline are better predictors of joint space narrowing than bone scintigraphy in obese women with knee osteoarthritis[J]. J Rheumatol, 2005, 32(8): 1540 – 1546.

[10] THOMAS K A, KIDZIŃSKI Ł, HALILAJ E, *et al*. Automated classification of radiographic knee osteoarthritis severity using deep neural networks[J]. Radiol Artif Intell, 2020, 2(2): e190065.

[11] SHAMIR L, RAHIMI S, ORLOV N, *et al*. Progression analysis and stage discovery in continuous physiological processes using image computing[J]. EURASIP J Bioinform Syst Biol, 2010, 2010(1): 107036.

[12] ANTONY J, MCGUINNESS K, O′CONNOR N E, *et al*. Quantifying radiographic knee osteoarthritis severity using deep convolutional neural networks[J]. 2016 23rd Int Conf Pattern Recognit ICPR, 2016: 1195 – 1200.

[13] TIULPIN A, THEVENOT J, RAHTU E, *et al*. Automatic knee osteoarthritis diagnosis from plain radiographs: a deep learning-based approach[J]. Sci Rep, 2018, 8(1): 1727.

[14] 郭轶斌, 郭威, 秦宇辰, 等. 基于 Kappa 系数的一致性检验及其软件实现[J]. 中国卫生统计, 2016, 33(1): 169 – 170, 174.

[15] RIDDLE D L, JIRANEK W A, HULL J R. Validity and reliability of radiographic knee osteoarthritis measures by arthroplasty surgeons[J]. Orthopedics, 2013, 36(1): e25 – e32.

（本文编辑：周娟）

（上接第 21 面）

[10] HEMPHILL J C, BONOVICH D C, BESMERTUS L, *et al*. The ICH score: a simple, reliable grading scale for intracerebral hemorrhage[J]. Stroke, 2001, 32(4): 891 – 897.

[11] CAO D, LI Q, FU P, *et al*. Early Hematoma Enlargement in Primary Intracerebral Hemorrhage[J]. Curr Drug Targets, 2017, 18(12): 1345 – 1348.

[12] 中华医学会神经外科学分会, 中国医师协会急诊医师分会, 中华医学会神经病学分会脑血管病学组, 等. 高血压性脑出血中国多学科诊治指南[J]. 中华神经外科杂志, 2020, 36(8): 757 – 770.

[13] 杨俊, 侯自明, 王浩, 等. 影像组学模型对高血压脑出血早期血肿扩大的预测作用研究[J]. 中华神经医学杂志, 2019, 18(1): 49 – 54.

[14] ROMERO J M, HITO R, DEJAM A, *et al*. Negative spot sign in primary intracerebral hemorrhage: potential impact in reducing imaging[J]. Emerg Radiol, 2017, 24(1): 1 – 6.

[15] 曹勇, 张谦, 于洮, 等. 中国脑血管病临床管理指南(节选版)——脑出血临床管理[J]. 中国卒中杂志, 2019, 14(8): 72 – 76.

[16] CORDONNIER C, DEMCHUK A, ZIAI W, *et al*. Intracerebral haemorrhage: current approaches to acute management[J]. Lancet, 2018, 392(10154): 1257 – 1268.

[17] LI Y G, LIP G Y H. Anticoagulation Resumption After Intracerebral Hemorrhage[J]. Curr Atheroscler Rep, 2018, 20(7): 32.

[18] KURAMATSU J B, SEMBILL J A, HUTTNER H B. Reversal of oral anticoagulation in patients with acute intracerebral hemorrhage[J]. Crit Care, 2019, 23(1): 206.

[19] GE C, ZhAO W, GUO H, *et al*. Comparison of the clinical efficacy of craniotomy and craniopuncture therapy for the early stage of moderate volume spontaneous intracerebral haemorrhage in basal ganglia: Using the CTA spot sign as an entry criterion[J]. Clin Neurol Neurosurg, 2018, 169(1): 41 – 48.

[20] HUSSEIN O, SAWALHA K, HAMED M, *et al*. The intraventricular-spot sign: prevalence, significance, and relation to hematoma expansion and outcomes[J]. J Neurol, 2018, 265(10): 2201 – 2210.

（本文编辑：梁琥）