(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 111436215 A
(43)申请公布日 2020.07.21

(21)申请号 201980000264.8

(22)申请日 2019.01.15

(30)优先权数据
201821042894 2018.11.14 IN

(85)PCT国际申请进入国家阶段日
2019.03.07

(86)PCT国际申请的申请数据
PCT/IB2019/050315 2019.01.15

(87)PCT国际申请的公布数据
WO2020/099940 EN 2020.05.22

(71)申请人 库雷人工智能科技私人有限公司
地址 印度孟买

(72)发明人 萨桑克·奇拉姆库尔希
罗希特·高希 斯威萨·塔纳马拉
普贾·拉奥 普拉桑特·瓦瑞尔

(74)专利代理机构 北京安信方达知识产权代理有限公司 11262
代理人 陆建萍 杨明钊

(51)Int.Cl.
G06T 7/00(2017.01)
G06K 9/62(2006.01)

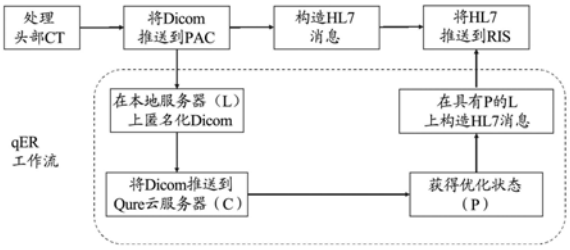
权利要求书2页 说明书16页 附图7页

(54)发明名称

用于医学成像评估的深度学习的应用

(57)摘要

本公开总体上涉及用于处理从成像或其他诊断和评估医疗流程中获得的电子数据的方法和系统。某些实施例涉及用于深度学习算法的开发的方法，该算法对成像和其他医学数据中的特定特征和状况执行机器识别。另一实施例提供了被配置为通过深度学习算法来检测并定位医学成像扫描上的医学异常的系统。



CN 111436215 A

1.一种用于开发深度学习系统以检测并定位非对比头部CT扫描上的医学异常的方法，包括：

选择医学成像扫描并使用自然语言处理(NLP)算法来提取医学异常，其中，每种类型的所述医学异常都在扫描级、切片级和像素级处被注释；

用所选择的医学成像扫描来训练切片式深度学习算法，以分割像素级注释的扫描；

用所述选择的医学成像扫描来训练所述深度学习算法，以生成切片级置信度；

预测对于每种类型的医学异常的存在的置信度；以及

生成对应于所述医学异常的识别水平的分数，并且输出表示所述医学异常的精确位置和程度的掩码。

2.根据权利要求1所述的方法，其中，所述深度学习算法包括卷积神经网络架构。

3.根据权利要求2所述的方法，其中，通过使用多个并行的全连接层来修改所述架构。

4.根据权利要求1所述的方法，其中，使用全连接层跨切片组合切片级处的置信度，以预测对于所述医学异常的存在及其类型的扫描级置信度。

5.根据权利要求1所述的方法，其中，所述医学异常包括颅内出血，并且5种类型的颅内出血中的每一种包括脑实质出血(IPH)、脑室内出血(IVH)、硬膜下颅内出血(SDH)、硬膜外出血(EDH)以及蛛网膜下出血(SAH)；中线移位；肿块效应；头颅骨折以及颅骨骨折。

6.根据权利要求1所述的方法，其中，通过向三个单独的窗口开窗来预处理所述扫描。

7.根据权利要求6所述的方法，其中，所述三个单独的窗口包括脑窗口、骨窗口和硬膜下窗口。

8.根据权利要求1所述的方法，其中，通过与放射科医师报告进行比较来验证用于检测所述医学异常的所述深度学习算法的准确性。

9.一种被配置为使用深度学习算法来检测并定位头部CT扫描上的医学异常的系统，其中，通过以下步骤来开发所述深度学习算法：

选择医学成像扫描并使用自然语言处理(NLP)算法来提取医学异常，其中，每种类型的所述医学异常都在扫描级、切片级和像素级处被注释；

用所选择的医学成像扫描来训练切片式深度学习算法，以分割像素级注释的扫描；

用所述选择的医学成像扫描来训练所述深度学习算法，以生成切片级置信度；

预测对于每种类型的医学异常的存在的置信度；以及

生成对应于所述医学异常的识别水平的分数，并且输出表示所述医学异常的精确位置和程度的掩码。

10.根据权利要求9所述的系统，其中，所述深度学习算法包括卷积神经网络架构。

11.根据权利要求10所述的系统，其中，通过使用多个并行的全连接层来修改所述架构。

12.根据权利要求9所述的系统，其中，使用全连接层跨切片组合切片级处的置信度，以预测对于所述医学异常的存在及其类型的扫描级置信度。

13.根据权利要求9所述的系统，其中，所述医学异常包括颅内出血，并且5种类型的颅内出血中的每一种包括脑实质出血(IPH)、脑室内出血(IVH)、硬膜下颅内出血(SDH)、硬膜外出血(EDH)以及蛛网膜下出血(SAH)；中线移位；肿块效应；头颅骨折以及颅骨骨折。

14.根据权利要求9所述的系统，其中，通过向三个单独的窗口开窗来预处理所述扫描。

15.根据权利要求14所述的系统，其中，所述三个单独的窗口包括脑窗口、骨窗口和硬膜下窗口。

16.根据权利要求9所述的系统，其中，通过与放射科医师报告进行比较来验证用于检测所述医学异常的所述深度学习算法的准确性。

17.根据权利要求9所述的系统，其中，所述算法针对检测ICH、IPH、IVH、SDH、EDH和SAH分别实现了0.94±0.02、0.96±0.03、0.93±0.07、0.95±0.04、0.97±0.06和0.96±0.04的AUC(ROC曲线下的面积)。

用于医学成像评估的深度学习的应用

[0001] 相关申请

[0002] 本申请要求享有于2018年11月14日提交的第201821042894号印度专利申请的优先权权益,通过引用将其全部并入本文用于所有目的。

技术领域

[0003] 本公开总体上涉及用于处理从成像或其他诊断和评估医疗程序中获得的电子数据的方法和系统。一些实施例涉及用于深度学习算法的开发的方法,该算法对成像和其他医学数据中的特定特征和状况执行机器识别。

背景技术

[0004] 诸如计算机断层摄影(CT)和X射线成像的医学成像技术广泛用于诊断、临床研究和治疗计划。存在对提高医学成像评估的效率、准确性和成本效益的自动化方法的新兴需求。

[0005] 非对比(non-contrast)头部CT扫描在最常用的急诊室诊断工具当中,用于头部受伤患者或者暗示有中风(stroke)或颅内压升高症状的患者。它们的广泛可用性和相对较低的获取时间使它们成为常用的一线诊断方式(modality)。过去几十年来,美国急诊每年进行CT扫描的比例一直在增加,并且使用头部CT来排除对神经外科介入(intervention)的需求也在增加。

[0006] CT扫描上可以容易地检测到的最关键的、时间敏感的异常包括颅内出血、颅内压升高和颅骨(cranial)骨折。中风患者的关键评估目标是排除颅内出血。这取决于CT成像及其迅速(swift)解释。类似地,对于疑似急性(acute)颅内出血的患者,即时(immediate)CT扫描解释对于评估神经外科治疗的需求至关重要。颅骨骨折如果是开放性的或凹陷的,通常需要紧急的神经外科介入。颅骨骨折也是头部CT扫描上最常遗漏的主要异常,尤其是如果在轴向平面上运行(course)的情况下。

[0007] 虽然仅在一小部分CT扫描上发现这些异常,但通过自动化初始筛查和分诊(triage)过程来简化头部CT扫描解释工作流(workflow),将大大减少诊断时间并加速治疗。这反过来会减少由中风和头部受伤产生的发病率和死亡率。自动头部CT扫描筛查和分诊系统对于繁忙的创伤护理中的队列管理是很有价值的,或者在没有放射科医师即时可用的情况下有助于远程位置的决策。

[0008] 过去一年,深度学习在医学成像解释任务中的应用取得了许多进展,有力的证据表明,深度学习可以执行特定的医学成像任务,包括识别糖尿病视网膜病变(diabetic retinopathy)并对其进行分级,以及以相当于专家医师(specialist physician)的准确度将皮肤病变(lesion)分类为良性(benign)或恶性(malignant)。深度学习算法也被训练,以通过“分类”算法来检测放射线图像(诸如胸片、胸部CT和头部CT)上的异常;以及通过“分割”算法来定位并量化疾病模式或解剖体积。

[0009] 用于放射学的准确的深度学习算法的开发,除了合适的模型架构外,还需要使用

大量准确标记的扫描来训练算法。当训练数据集很大并且包括来自不同源的扫描时，算法很好地推广(generalize)到新设置的可能性增加了。

[0010] 有几项关于低体积的头部CT扫描的计算机辅助诊断(CAD)算法的开发和验证的研究。早期，深度学习被用来检测颅内出血。传统的计算机视觉技术更常用于检测骨折和中线移位(midline shift)。对于大多数研究的训练和验证数据集具有<200次头部CT扫描，这引起了对这些算法的鲁棒性的担忧。此外，没有标准的公共头部CT数据集来直接比较算法的性能。

[0011] 发明概述

[0012] 本公开描述了全自动深度学习系统的开发和临床验证，该系统被训练以从医学成像扫描检测并定位异常。

[0013] 某些实施例涉及深度学习系统的开发和验证，以检测并定位头部CT扫描异常。经训练的算法检测五种颅内出血(ICH)(即，脑实质出血(IPH)、脑室内出血(IVH)、硬膜下颅内出血(SDH)、硬膜外出血(EDH)和蛛网膜下出血(SAH))，以及具有最大化AUC(ROC曲线下面积)的头颅(skull)/颅骨(calvarial)/颅顶(cranial vault)骨折。经训练的算法还检测肿块效应(mass effect)和中线移位，这两者都被用作脑受伤严重性的指标。

[0014] 具体地，实施例提供了用于开发深度学习系统以检测并定位头部CT扫描上的医学异常的方法，包括：

[0015] 选择医学成像扫描并使用自然语言处理(NLP)算法来提取医学异常，其中，每种类型的医学异常都在扫描级、切片级和像素级处被注释；

[0016] 用选择的医学成像扫描来训练包括卷积神经网络架构的切片式深度学习算法，以分割像素级注释的扫描；

[0017] 用选择的医学成像扫描来训练包括卷积神经网络架构的深度学习算法，其中，通过使用多个并行的全连接层来修改架构以生成切片级置信度；

[0018] 预测对于每种类型的医学异常的存在的置信度，其中，使用全连接层跨切片组合在切片级处的置信度，以预测对于医疗异常的存在及其类型的扫描级置信度；

[0019] 生成对应于医学异常的识别水平的分数，并且输出表示医学异常的精确位置和程度的掩码(mask)；以及

[0020] 通过与放射科医师报告的比较来验证用于检测医学异常的深度学习算法的准确性。

[0021] 根据实施例，所述医学成像扫描包括但不限于CT、X射线、核磁共振成像(MRI)和超声过程。对于头部CT扫描，所述医学异常包括但不限于包括脑实质出血(IPH)、脑室内出血(IVH)、硬膜下颅内出血(SDH)、硬膜外出血(EDH)以及蛛网膜下出血(SAH)的5种类型的出血中的每一种；中线移位；肿块效应；以及头颅/颅骨骨折。

[0022] 此外，对于给定的头部CT扫描，通过向三个单独的窗口(包括脑窗口、骨窗口和硬膜下窗口)开窗(window)并且将这些窗口作为通道堆叠来预处理扫描。

[0023] 另一实施例提供了被配置为通过深度学习算法来检测并定位头部CT扫描上的医学异常的系统，其中，通过以下步骤来开发深度学习算法：

[0024] 选择医学成像扫描并使用自然语言处理(NLP)算法来提取医学异常，其中，每种类型的医学异常都在扫描、切片和像素级处被注释；

[0025] 用选择的医学成像扫描来训练包括卷积神经网络架构的切片式深度学习算法,以分割像素级注释的扫描;

[0026] 用选择的医学成像扫描来训练包括卷积神经网络架构的深度学习算法,其中,通过使用多个并行的全连接层来修改架构以生成切片级置信度;

[0027] 预测对于每种类型的医学异常的存在的置信度,其中,使用全连接层跨切片组合切片级处的置信度,以预测对于医学异常的存在及其类型的扫描级置信度;

[0028] 生成对应于医学异常的识别水平的分数,并且输出表示医学异常的精确位置和程度的掩码;以及

[0029] 通过与放射科医师报告的比较来验证用于检测医学异常的深度学习算法的准确性。

[0030] 此外,系统被配置为通过深度学习算法来检测并定位头部CT扫描上的医学异常,其中,算法针对检测ICH、IPH、IVH、SDH、EDH和SAH分别实现了0.94±0.02、0.96±0.03、0.93±0.07、0.95±0.04、0.97±0.06和0.96±0.04的AUC。

[0031] 发明的有益效果

[0032] 本发明提供了深度学习算法,以从头部CT扫描中单独地检测多达九个关键调查结果(finding)。与临床放射学报告相比,所述算法已经在大数据集上得到了验证。与三名放射科医师对从与开发数据集完全不同的源获取的数据集的一致看法(consensus)相比,所述算法也得到了验证。此外,到目前为止,很少的文献描述检测颅骨骨折的深度学习算法的准确使用。

[0033] 本发明提供了能够以高准确度执行该任务的深度学习算法。对这种大量患者检测肿块效应和中线移位(两者都用于估计颅内状况变化的严重性和对紧急介入的需要)的算法的临床验证也是独一无二的。同样重要的是,一旦获得头部CT扫描,所述算法可以用于自动分流或通知具有关键调查结果的患者。

[0034] 附图简述

[0035] 图1由算法产生的定位。这些可以提供结果的视觉显示。

[0036] 图2数据集部分流程开发,Qure25k和CQ500数据集。

[0037] 图3对于Qure25k和CQ500数据集上的算法的受试者操作特性(ROC,receiver operating characteristic)曲线。蓝色线用于Qure25k数据集,并且红色线用于CQ500数据集。个人评分员根据他们对CQ500数据集的一致看法测量的真阳性率和假阳性率也与ROC一起被绘制,用于比较。

[0038] 图4使用qER进行头部CT扫描的推荐工作流。

[0039] 详细描述

[0040] 应当理解,本发明不限于本文描述的特定的方法论、协议和系统等,且因此可以变化。本文使用的术语仅为了描述特定实施例的目的,而不旨在限制本发明的范围,本发明的范围仅由权利要求限定。

[0041] 如说明书和所附权利要求中所使用的,除非有相反的指定,否则以下术语具有以下所指示的含义。

[0042] “架构”是指描述计算机系统的功能、组织和实现的一组规则和方法。

[0043] “卷积神经网络(CNN)”是指一类深度、前馈人工神经网络,最常用于分析视觉影

像。CNN使用多层感知器的变体，这些感知器被设计为需要最少的预处理。CNN由输入和输出层以及多个隐藏层组成。CNN的隐藏层通常由卷积层、池化(pooling)层、全连接层和标准化层组成。卷积层对输入应用卷积运算，将结果传递给下一层。局部或全局池化层将一层处的神经元簇(cluster)的输出合并到下一层中的单个神经元中。全连接层将一层中的每个神经元连接到另一层中的每个神经元。与其他图像分类算法相比，CNN使用相对较少的预处理。这意味着网络学习传统算法中手工设计(hand-engineered)的过滤器。在特征设计中，独立于先前知识和人类努力是主要的优势。

[0044] “启发法”是指被设计用于在经典方法太慢时更快地解决问题，或者用于在经典方法找不到任何精确解(exact solution)时找到近似解的技术。这是通过交易最优性、完整性、准确性或速度精度来实现的。从某种意义上来说，这可以被认为是捷径。启发式函数(也简称为启发法)是基于可用信息在每个分支步骤中对搜索算法中的备选方案进行排序以决定跟随哪个分支的函数。启发法的目标是在合理的时间范围内产生足以解决手头问题的解。该解可能不是对该问题的所有解中最好的，或者它可以仅仅近似精确解。

[0045] “自然语言处理(NLP)”是指计算机以智能且有用的方式来分析、理解人类语言并从人类语言获得含义的方法。通过利用NLP，开发人员可以组织并构造知识来执行任务，诸如自动摘要、翻译命名实体识别、关系提取、情感分析、语音识别和主题分割。

[0046] 本公开示出了能够在数据驱动的图像评估工作流中集成并使用机器学习分析的各种技术和配置。例如，可以对被产生作为医学成像研究的一部分的医学成像过程数据执行机器学习分析(诸如某些医学状况的图像检测的经训练的模型)。医学成像过程数据可以包括由成像模态捕获的图像数据，以及命令(order)数据(诸如指示对放射线图像读取的请求的数据)，每个都是为了便于医学成像评估而产生的(诸如由放射科医师执行的放射学读取或者由另一合格的医学专业人员执行的诊断评估)。

[0047] 例如，机器学习分析可以接收并处理来自医学成像过程数据的图像，以识别经训练的结构、状况以及特定研究的图像内的状况。机器学习分析可以导致自动检测、指示或确认图像内的某些医疗状况，例如紧急或生命攸关的医疗状况、临床严重异常和其他关键调查结果。基于机器学习分析的结果，对于图像的医学评估和相关联的成像过程可以被设置优先级，或者以其他方式被改变或被修改。此外，医学状况的检测可以用于帮助将医学成像数据分配给特定评估者、对于医学成像数据的评估过程，或在医学成像评估之前或与医学成像评估同时执行其他动作(或生成诸如来自这样的医学成像评估的报告的数据项)。

[0048] 如本文进一步讨论的，机器学习分析可以代表任何数量的机器学习算法和经训练的模型来被提供，包括但不限于已经被训练来执行图像识别任务的深度学习模型(也称为深度机器学习或分层模型)，特别是对于人体解剖和解剖表示的医学图像上的某些类型的医学状况。如本文所使用的，术语“机器学习”用于指各类人工智能算法和能够执行经训练的结构的机器驱动(例如，计算机辅助)识别的算法驱动方法，其中术语“深度学习”是指使用多层表示和抽象的这种机器学习算法的多层操作。然而，明显的是，在当前描述的医学成像评估中应用、使用和配置的机器学习算法的角色可以由任何数量的其他基于算法的方法(包括人工神经网络的变体、有学习能力的算法、可训练的对象分类和其他人工智能处理技术)来补充或替代。

[0049] 在下面的一些示例中，参考放射学医学成像过程(例如，计算机断层摄影(CT)、核

磁共振成像(MRI)、超声波和X射线过程等)和由这样的成像过程产生的图像的诊断评估,这样的成像过程由经许可并认证的放射科医师通过图像评估(例如,放射学读取)来执行。应当理解,当前描述的技术和系统的适用性将扩展到由各种医学过程和专业(包括不涉及传统放射学成像模态的那些)产生的各种各样的成像数据(和其他数据表示)。这样的专业包括但不限于病理学、医学摄影、诸如脑电图学(EEG)和心电图学(EKG)过程的医学数据测量、心脏病学数据、神经科学数据、临床前成像以及与远程医疗、远程病理学(telepathology)、远程诊断相关的其他数据收集过程,以及医学过程和医学科学的其他应用。因此,本文描述的数据识别和工作流修改技术的性能可以应用于各种医学图像数据类型、设置和用例(use case),包括捕获的静态图像和多图像(例如视频)表示。

[0050] 以下描述和附图充分说明了具体实施例,以使本领域技术人员能够实施它们。其他实施例可以结合结构、逻辑、电气、过程和其他变化。一些实施例的部分及特征可以被包括在其它实施例中或是代替其他实施例的部分及特征。

[0051] 示例

[0052] 示例1.用于检测头部CT扫描中的关键调查结果的深度学习算法

[0053] 1.1数据集

[0054] 从印度的几个中心回顾性地(retrospectively)收集了313,318份匿名头部CT扫描。这些中心包括住院和门诊放射中心,其采用各种CT扫描仪型号(model)(表1),其中每次旋转的切片的范围从2到128。扫描的每一次都有与之相关联的电子临床报告,我们在算法开发过程期间将其用作黄金标准(gold standard)。

[0055] 表1.用于每个数据集的CT扫描仪的型号。

[0056]

数据集	CT 扫描仪型号
Qure25k 和开发	GE BrightSpeed、GE Brivo CT315、GE Brivo CT385、GE HiSpeed、GE LightSpeed、GE ProSpeed、GE Revolution ACT、Philips Brilliance、Siemens Definition、Siemens Emotion、Siemens Sensation、Siemens SOMATOM、Siemens Spirit
CQ500	GE BrightSpeed、GE Discovery CT750 HD、GE LightSpeed、GE Optima CT660、Philips MX 16-slice、Philips Access-32 CT

[0057] 在这些扫描中,23,263名随机选取的患者的扫描(Qure25k数据集)被选择用于验证,并且其余患者的扫描(开发数据集)用于训练/开发算法。从Qure25k数据集移除术后(Post-operative)扫描和年龄小于7岁的患者的扫描。算法开发过程期间未使用该数据集。

[0058] 印度新德里的Centre for Advanced Research in Imaging,Neurosciences and Genomics(CARING)提供了临床验证数据集(称为CQ500数据集)。该数据集是在新德里各个放射中心进行的头部CT扫描的子集。大约一半的中心是独立的门诊中心,并且另一半是大型医院内含的放射科。这些中心和从中获取开发数据集的中心之间没有重叠。在这些中心处使用的CT扫描仪每次旋转的切片从16变化到128。CT扫描仪的型号在表1中列出。这些数据是从本地PACS服务器提取(pull)的,并按照内部定义的HIPAA指南匿名化。由于这两个数据集都是回顾性获得的并且完全匿名化,因此该研究免于IRB批准。

[0059] 与开发和Qure25k数据集类似,CQ500数据集中与扫描相关联的临床放射学报告也是可用的。如下所述,临床放射学报告用于数据集选择。

[0060] 分两个批次(B1和B2)收集CQ500数据集。通过选择在上述中心处进行从2017年11月20日开始的30天的所有头部CT扫描来收集批次B1。批次B2是通过以下方式从其余扫描中选择的:

[0061] 1.自然语言处理(NLP)算法被用于检测来自临床放射学报告的IPH、SDH、EDH、SAH、IVH、颅骨骨折。

[0062] 2.然后随机选择报告,使得IPH、SDH、EDH、SAH和颅骨骨折中的

[0063] 每一个具有大约80次扫描。

[0064] 然后,针对以下排除标准来筛选选择的扫描中的每一个:

[0065] • 无术后缺陷

[0066] • 没有覆盖整个大脑的非对比轴系(axial series)。

[0067] • 年龄<7岁(如果数据不可用,从颅骨缝估计)。

[0068] 1.2读取扫描

[0069] 三名高级放射科医师担任对于CQ500数据集中的CT扫描的独立评分员。他们在颅骨CT解释有相应的8年、12年和20年的经验。三名评分员中没有一人参与对登记患者的临床护理或评估,他们也没有查询任何患者的临床病史。放射科医师的每一位独立评估CQ500数据集中的扫描,并给出对于记录调查结果和查询分辨率的说明。扫描的呈现的顺序是随机化的,以尽量最小化患者的后续扫描的回忆(recall)。

[0070] 评分员的每一位记录了对于每次扫描的以下调查结果:

[0071] • 颅内出血的存在或不存在、以及如果存在,颅内出血的类型(脑实质、脑室内、硬膜外、硬膜下和蛛网膜下)。

[0072] • 中线移位和肿块效应的存在或不存在。

[0073] • 骨折的存在或不存在。如果存在,是否为(部分)颅骨骨折。

[0074] 由于诸如出血性挫伤、肿瘤/出血性脑梗死等任何病因引起的轴内血液的存在也被包括在脑实质出血的定义中。在该研究中,慢性出血被认为是阳性的。肿块效应被定义为以下的任何一种:局部肿块效应、心室消失(effacement)、中线移位和疝形成(herniation)。如果移位量大于5mm,则中线移位被认为是阳性的。如果至少有一处骨折延伸到头颅盖(skullcap)内,则扫描被认为有颅骨骨折。

[0075] 如果三名评分员未能就调查结果的每一个达成一致意见,则多数评分员的解释被用作最终诊断。

[0076] 在开发和Qure25k数据集上,放射科医师撰写的临床报告被视为黄金标准。然而,这些都是以自由文本而不是结构化格式撰写的。因此,基于规则的自然语言处理(NLP)算法被应用于放射科医师的临床报告,以自动推断上面记录的调查结果。在来自Qure25k数据集的报告的子集上验证该算法,以确保推断的信息是准确的,并且可以用作黄金标准。

[0077] 1.3.开发深度学习算法

[0078] 深度学习是机器学习的形式,其中使用的模型是具有大量(通常是卷积)层的神经网络。训练该模型需要大量的数据,对于这些数据,真相(truth)是已知的。通常通过被称为反向传播(back propagation)的算法来执行训练。在该算法中,模型被反复修改,以最小化

模型的预测和对于每个数据点的已知实际真相(ground truth)之间的误差。

[0079] 算法的开发中的主要挑战之一是CT扫描的三维(3D)性质。这主要是由于被称为“维数灾难”的问题,在这个问题上,训练机器学习算法所需的数据随着数据的维数呈指数级增长。深度学习技术已经被广泛研究用于二维图像的分割和分类的任务。虽然3D图像的分割是在多个上下文中被研究的,但是它们的分类没有得到很好的研究。一个密切相关的问题是从短视频剪辑中识别人类行为(因为视频是三维的,其中时间作为第三维度)。尽管该问题在文献中得到很好的探讨,但并没有出现用于该项任务的领先的架构。分类的方法与Simonyan和Zisserman的分类方法(Advances in neural information processing systems,第568-576页,2014)密切相关,并且涉及大量扫描的切片级和像素级注释。

[0080] 在该研究中,针对子任务的每一个训练了单独的深度学习模型,即颅内出血、中线移位/肿块效应和颅骨骨折,我们将在下面描述。

[0081] 1.3.1颅内出血

[0082] 使用自然语言处理(NLP)算法搜索开发数据集,以选择一些非对比头部CT扫描,这些扫描报告了脑实质出血(IPH)、脑室内出血(IVH)、硬膜下颅内出血(SDH)、硬膜外出血(EDH)、蛛网膜下出血(SAH)中的任一个,以及都不是这些的那些问题。这些扫描中的每个切片都被手动地标记有在该切片中可见的出血。总共对4304次扫描(165809个切片)进行了注释,其中具有IPH、IVH、SDH、EDH、SAH以及都不是这些的扫描(切片)的次数分别为1787次(12857个)、299次(3147个)、938次(11709个)、623次(5424个)、888次(6861个)和944次(133897个)。

[0083] ResNet18是一种流行的卷积神经网络架构,稍加修改就可以用于预测对于切片中每种出血类型的存在的基于SoftMax的置信度。通过使用五个并行的全连接(FC)层代替单个FC层来对架构进行修改。该设计基于这样的假设,即用于检测出血的图像特征对于所有出血类型都是类似的。使用随机森林(random forest)将切片级处的置信度进行组合,以预测对于颅内出血的存在及其类型的扫描级置信度。

[0084] 独立的模型被进一步训练来定位以下类型的出血:IPH、SDH、EDH。定位要求对扫描中的每个像素进行出血存在或不存在的密集预测。为了训练用于密集预测的模型,针对上面切片注释的图像的子集注释了对应于每个出血的像素,以提供对于模型的实际真相。该集合包含1706幅图像,其中具有IPH、SDH、EDH以及都不是这些的图像的数量分别为506、243、277和750。基于切片式2D UNet7的架构用于每种类型的出血的分割。

[0085] 由于对出血检测算法的分割网络进行了训练,因此除了检测出血的存在之外,还输出了表示出血(蛛网膜下出血除外)的精确位置和范围的掩码。参见图1a。

[0086] 1.3.2中线移位和肿块效应

[0087] 用于检测中线移位和肿块效应的算法与用于检测颅内出血的算法非常类似。来自选择扫描的每个切片被标记为在该切片中的中线移位和肿块效应的存在或不存在。总体上,699次扫描(26135个切片)被注释,其中具有肿块效应的扫描(切片)的次数为320次(3143个),并且中线移位为249次(2074个)。

[0088] 具有两个并行的全连接层的修改的ResNet18分别用于预测针对肿块效应和中线移位的存在的切片式置信度。因此,使用随机森林来组合这些切片级置信度,以预测对于两种异常的扫描级置信度。

[0089] 1.3.3颅骨骨折

[0090] 使用NLP算法搜索开发数据集,以选择一些具有颅骨骨折的扫描。通过在骨折周围标记严格的边界框(tight bounding box)来注释这些扫描中的每个切片。注释的扫描的次数为1119次(42942个切片),其中9938个切片显示颅骨骨折。

[0091] 切片连同目标边界框一起被送入基于DeepLab的架构中,以预测对于骨折的像素式热图(heatmap)(如图1b所示)。头颅骨折在该表示中极其稀疏(sparse)。对于这种稀疏信号,反向传播算法中的梯度流往往会受到阻碍(hindered)。因此,采用难分负样本挖掘(hard negative mining)损失来抵消注释的稀疏性。

[0092] 从整个扫描的生成的热图中,设计了代表局部骨折损伤及其体积的特征。这些特征被用来训练随机森林,以预测颅骨骨折的存在的扫描式置信度。

[0093] 1.3.4预处理

[0094] 对于给定的CT扫描,使用并重新采样了使用软重构内核的非对比轴系,使得切片厚度约为5mm。在传递到我们的深度学习模型之前,该系的所有切片都被调整到224×224像素的大小。不是作为单个通道来通过CT密度的整个动态范围,而是密度通过使用三个单独的窗口进行开窗,并被作为通道堆叠。使用的窗口是脑窗口(l=40,w=80),骨窗口(l=500,w=3000)和硬膜下窗口(l=175,w=50)。这是因为骨窗口中可见的骨折能够指示脑窗口中存在额外的轴向出血,并且相反地,脑窗口中头皮血肿的存在可能与骨折相关。硬膜下窗口有助于区分头颅和额外的轴向出血,这在正常的脑窗口中可能是无法区分的。

[0095] 1.3.5训练细节

[0096] 注释被分成对患者随机分层的train集合和val集合。train集合用于训练网络,而val集合用于选择超参数。网络架构的所有权重都是随机初始化的。在训练期间,使用的数据增强策略与He等人的策略相同(Proceedings of the IEEE conference on computer vision and pattern recognition,第770-778页,2016。):任意大小的裁剪(crop)、水平翻转和像素强度增强。在网络的最后两个下采样块之后,使用0.5的丢弃(dropout)。SGD以32的批次大小来使用。学习率从0.01开始,并且每20个时期(epoch)下降2个。对于阳性类的权重为20的加权交叉熵被用作损失标准。在运行200个时期(即提前停止)之后,基于对val集合的性能选择最终模型。在Nvidia Titan X GPU上用PyTorch框架进行训练。训练每个模型大约需要15个小时。

[0097] 1.3.6算法与放射科医师的比较

[0098] 双边(two-sided)费希尔精确检验(Fisher's exact tests)用于比较在高敏感性(sensitivity)操作点上的算法的性能和在CQ500数据集上的单个放射科医师的性能。零假设是放射科医师和算法分别在阳性扫描和阴性扫描中表现同样好。请注意,关于阳性扫描的假设表明对于放射科医师和算法的敏感性是一样的。类似地,对于阴性扫描,这意味着特异性(specificity)是相同的。选择费希尔精确检验是因为卡方检验(chi-squared test)中使用的近似法在数据分布非常不均匀(即,敏感性/特异性≈1)时无效。表2中将检验的P值制成表格。

[0099] 表2.对于双边检验具有零假设的p值,零假设中算法和评分员对阳性扫描和阴性扫描表现同样好。

[0100]

调查结果	阳性扫描			阴性扫描		
	评分员 1	评分员 2	评分员 3	评分员 1	评分员 2	评分员 3
颅内出血	0·54	0·24	0·11	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$
脑实质	1·00	0·17	1·00	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$
脑室内	1·00	0·42	1·00	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$
硬膜下	0·01	0·07	1·00	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$
硬膜外	1·00	1·00	0·16	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$
蛛网膜下	0·11	1·00	1·00	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$
颅盖骨折	1·00	$<10^{-4}$	1·00	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$
肿块效应	0·03	1·00	0·13	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$
中线移位	1·00	0·03	0·36	0·73	$<10^{-4}$	$<10^{-4}$

[0101] 从表2可以看出，对于几乎所有的调查结果，不能排除算法的敏感性与评分员的敏感性不可区分的零假设。对于那些具有显著不同的调查结果-评分员对(表中粗体)，进一步的单边费希尔检验发现算法具有更好的敏感性($p<0.05$)。从表2中得出的另一推论在于，算法和评分员的特异性明显不同(除了一对：肿块效应&评分员1)。单边检验表明，在该操作点处评分员的特异性更好($p<10^{-4}$)。

[0102] 总之，在高敏感性点处，算法的敏感性与评分员的敏感性并不可区分，但是特异性明显更低。

[0103] 1.4评估算法

[0104] 当在扫描上运行时，算法会产生在[0,1]范围内的9个实值置信度分数的列表，其指示存在以下调查结果：颅内出血和5种类型出血中的每一种、中线移位、肿块效应和颅骨骨折。如前所述，使用对于CQ500数据集的多数投票和通过对于Qure25k数据集的报告的NLP算法来获得相应的黄金标准。表3示出了当单个评分员被认为是黄金标准时，以及当多数投票被认为是黄金标准时，CQ500数据集上算法的AUC，并且表4示出了评分员的敏感性和特异性相对于CQ500数据集上他们的多数投票。

[0105] 表3.当单个评分员被认为是黄金标准时，以及当多数票被认为是黄金标准时，CQ500数据集上算法的AUC。

[0106]

黄金标准	评分员 1	评分员 2	评分员 3	多数投票
调查结果				

[0107]

颅内出血	0·9080	0·9413	0·9356	0·9419
脑实质	0·9236	0·9430	0·9275	0·9544
脑室内	0·9416	0·9211	0·9364	0·9310
硬膜下	0·8847	0·8957	0·9374	0·9521
硬膜外	0·9049	0·8343	0·9769	0·9731
蛛网膜下	0·9270	0·9507	0·8975	0·9574
颅骨骨折	0·9198	0·8653	0·9523	0·9624
肿块效应	0·8617	0·9310	0·8850	0·9216
中线移位	0·9545	0·9386	0·9461	0·9697

[0108] 表4.评分员的敏感性和特异性相对于CQ500数据集上他们的多数投票。

[0109]

调查结果	**评分员 1**		**评分员 2**		**评分员 3**	
	敏感性	特异性	敏感性	特异性	敏感性	特异性
颅内出血	0·9805	0·8986	0·9268	0·9790	0·9122	0·9790
脑实质	0·9552	0·9580	0·9403	0·9608	0·8955	0·9720
脑室内	0·9643	0·9741	0·9286	0·9914	0·8214	0·9935
硬膜下	0·9245	0·9247	0·7547	0·9795	0·8113	0·9954
硬膜外	0·6154	0·9958	0·9231	0·9895	0·9231	0·9979
蛛网膜下	0·9167	0·9722	0·8000	0·9884	0·9167	0·9466
颅骨骨折	1·0000	0·9519	0·9519	0·9628	0·4118	0·9978
肿块效应	0·9606	0·9148	0·8031	0·9835	0·9055	0·9560
中线移位	0·8769	0·9883	0·9385	0·9038	0·8000	0·9883

[0110] 对于CQ500和Qure25k数据集，通过改变阈值并在每个阈值处绘制真阳性率(即，敏感性)和假阳性率(即，1-特异性)来获得对于上述每个数据集的受试者操作特性(ROC)曲线。在ROC曲线上选择了两个操作点，分别使得敏感性≈0.9(高敏感性点)和特异性≈0.9(高特异性141点)。选择具有敏感性最接近0.95的高敏感性操作点。如果在该操作点处特异性>0.7，则使用该操作点。否则，选择其敏感性略高于0.90(如果可用)的操作点，否则选择最接近0.90的操作点。选择其特异性最接近0.95的高特异性操作点。如果在该操作点处敏感性>0.70，则使用该操作点。否则，使用其特异性略高于0.90(如果可用)的操作点，否则使用最接近0.90的操作点。ROC曲线下的面积(AUC)和这两个操作点处的敏感性和特异性被用于评估算法。

[0111] 1.5统计分析

[0112] 分别使用正态逼近(normal approximation)和由Hanley和McNeil概述的方法146(Radiology，143(1)：29-36，1982)来计算对于比例的样本大小和AUC。在随机选择的CT扫描的样本中，我们的目标异常的患病率(prevalence)往往较低。这意味着，在未浓缩(un-enriched)的数据集上建立具有相当高置信度的算法的敏感性需要非常大的样本大小。例如，为了在一半长度(half-length)为0.10的95％置信区间内建立期望值为0.7的敏感性，要读取的阳性扫描的次数≈80。类似地，对于患病率为1％的调查结果，为了在一半长度为0.05的95％置信区间内建立AUC，要读取的扫描的次数≈20000。

[0113] 该研究中使用的Qure25k数据集是从人口分布中随机采样的，并且在上述样本大小计算之后具有的扫描次数>20000次。然而，对155放射科医师时间的限制需要在对于CQ500数据集的章节2.1中概述的浓缩策略。扫描的手动管理(curation)(通过参考扫描本身)会有偏向于更显著的阳性扫描的选择。该问题通过随机选择得到缓解，其中从临床报告确定阳性扫描。

[0114] 在选择的操作点处，为每个调查结果生成混淆矩阵。参见表5。使用基于Beta分布的“精确”Clopper-Pearson方法从这些矩阵计算对于敏感性和特异性的95％置信区间。按照由Hanley和McNeil描述的“基于分布的”方法(Radiology，143(1)：29-36，1982)来计算AUC的置信区间。在CQ500数据集上，使用一致(agreement)百分比和Cohen’s kappa(κ)统计(Viera等人，Fam Med，37(5)：360-363，2005)来测量每个调查结果上的成对评分员之间的一致性。此外，使用Fleiss’kappa(κ)统计(Fleiss等人，Statistical methods for rates and proportions.364John Wiley&Sons，2013)来测量每个调查结果上所有三名评分员之间的一致性。

[0115] 表5.类别之间的混淆：这些表格中的每一行都表示在扫描的子集上计算的AUC，对于该子集，调查结果是阳性的。例如，SDH行表示具有SDH的扫描上的不同调查结果的AUC。SDH行的EDH列中的低值意味着如果扫描中存在SDH，算法就不能很好地检测EDH。(ICH-SAH是颅内出血及其子类型、Frac是颅骨骨折、ME和MLS分别是肿块效应和中线移位。)

[0116]

	ICH	IPH	IVH	SDH	EDH	SAH	Frac	ME	MLS
ICH	–	0·80	0·92	0·87	0·90	0·80	0·88	0·82	0·87
IPH	–	–	0·92	0·88	0·91	0·81	0·89	0·83	0·86
IVH	–	0·79	–	0·94	0·94	0·80	0·97	0·79	0·83
SDH	–	0·78	0·81	–	0·73	0·78	0·83	0·80	0·85
EDH	–	0·79	0·83	0·78	–	0·77	0·72	0·71	0·77
SAH	–	0·74	0·92	0·82	0·85	–	0·85	0·77	0·84
Frac	0·87	0·81	0·90	0·81	0·86	0·80	–	0·78	0·84
ME	0·92	0·90	0·92	0·93	0·91	0·84	0·94	–	0·79
MLS	0·88	0·87	0·92	0·95	0·90	0·80	0·93	–	–

[0117] (a)Qure25k数据集

[0118]

	ICH	IPH	IVH	SDH	EDH	SAH	Frac	ME	MLS
ICH	–	0·92	0·90	0·94	0·97	0·93	0·95	0·86	0·92
IPH	–	–	0·88	0·95	0·96	0·95	0·93	0·83	0·88
IVH	–	0·99	–	1·00	1·00	0·98	1·00	0·69	0·69
SDH	–	0·99	0·84	–	0·95	0·98	0·97	0·89	0·98
EDH	–	0·79	0·88	0·83	–	0·71	1·00	0·92	1·00
SAH	–	0·87	0·87	0·89	0·94	–	0·91	0·94	0·89
Frac	0·91	0·81	0·86	0·89	0·90	0·88	–	0·84	0·97
ME	0·75	0·94	0·88	0·97	0·98	0·94	0·96	–	0·86
MLS	0·84	0·98	0·90	0·98	0·96	0·94	0·98	–	–

[0119] (b)Qure25k数据集

[0120] 1.6结果

[0121] 表6汇总了对于每个调查结果的患者人口统计和患病率。Qure25k数据集包含21095次扫描，其中报告对于颅内出血和颅骨骨折为阳性的扫描的次数分别为2495次和992次。CQ500数据集包括491次扫描，其中批次B1有214次扫描，并且批次B2有277次扫描。B1包含35次和6次扫描，分别报告颅内176出血和颅骨骨折。同样对于B2分别为170次和28次。

[0122] 表6.对于CQ500和Qure25K数据集的数据集特性。

[0123]

特性	Qure25K 数据集	CQ500	CQ500
		数据集批次 B1	数据集批次 B2
扫描次数	21095	214	277
每次扫描的评分员的编号	1	3	3
患者人口统计			
年龄			
已知年龄的扫描的次数	21095	189	251
平均值	43·31	43·40	51·70
标准偏差	22·39	22·43	20·31
范围	7－99	7－95	10－96
女性的数量/已知性别的扫描的次数（百分比）	9030/21064 (42·87%)	94/214 (43·92%)	84/277 (30·31%)

[0124] 患病率

颅内出血的扫描次数（百分比）	2495 (11·82%)	35 (16·36%)	170 (61·37%)
脑实质	2013 (9·54%)	29 (13·55%)	105 (37·91%)
脑室内	436 (2·07%)	7 (3·27%)	21 (7·58%)
硬膜下	554 (2·63%)	9 (4·21%)	44 (15·88%)
硬膜外	290 (1·37%)	2 (0·93%)	11 (3·97%)
蛛网膜下	611(2·90%)	9 (4·21%)	51 (18·41%)
骨折	1653 (7·84%)	8 (3·74%)	31 (11·19%)
颅骨骨折	992 (4·70%)	6 (2·80%)	28 (10·11%)
中线移位	666 (3·16%)	18 (8·41%)	47 (16·97%)
肿块效应	1517 (7·19%)	28 (13·08%)	99 (35·74%)

[0125] 期望Qure25k数据集和CQ500数据集的批次B1代表头部CT扫描的人口分布。这是因为Qure25k数据集是从头部CT扫描的大型数据库中随机采样的，而批次B1包括一个月内在选择的中心180处获取的所有头部CT扫描。对于这两个数据集的年龄、性别和患病率统计类似的事实进一步支持了该假设。然而，CQ500数据集整体上并不代表人群，因为批次B2被选择用于更高的出血发生率。尽管如此，性能度量(即，AUC、敏感性和特异性)应当代表人群的性能，因为这些度量与患病率无关。

[0126] 在CQ500数据集的选择过程中分析的临床报告的数量为4462份。其中，对于批次B1和B2的选择的扫描的次数分别为285次和440次。排除的次数分别为71次和163次，结果总共扫描了491次。排除的原因是图像不可用(113)，术后扫描(67)，扫描没有非对比轴系(32)，

以及患者年龄小于7岁(22)。图2a-2b呈现了CQ500数据集的数据集选择过程的示意图。

[0127] 1.7 Qure25K数据集

[0128] 对总共1779份报告评估用于从Qure25k数据集中的临床报告推断调查结果的自然语言处理(NLP)算法。NLP算法的敏感性和特异性相当高;性能最差的调查结果是硬膜下出血,其敏感性为0·9318(95%CI 0·8134-0·9857),并且特异性为0·9965(95%CI 0·9925-0·9987),而骨折被完美地推断,其敏感性为1(95%CI 0·9745-1·000),并且特异性为1(95%CI 0·9977-1·000)。表7a示出了对于评估的1779份报告上所有目标调查结果的敏感性和特异性。

[0129] 表8a和图2总结了深度学习算法在Qure25k集合上的性能。算法实现了颅内出血的AUC为0·9194(95%CI 0·9119-0·9269)、颅骨骨折的AUC为0·9244(95%CI 0·9130-0·9359)、以及中线移位的AUC为0·9276(95%CI 0·9139-0·9413)。

[0130] 表7.对于Qure25k和CQ500数据集的黄金标准的可靠性。在Qure25k上,我们使用NLP算法从放射科医师的报告推断出调查结果。三名放射科医生审查了CQ500数据集上的491个病例中的每一个,并且评分员的多数投票被用作黄金标准。表7a示出了使用的NLP算法的准确性的估计,而表7b示出了放射科医师的读数的可靠性和一致性。

[0131]

调查结果	敏感性（85% CI）		敏感性（95% CI）	
颅内出血	203/207	0·9807 (0·9513-0·9947)	1552/1572	0·9873 (0·9804-0·9922)
脑实质	154/157	0·9809 (0·9452-0·9960)	1603/1622	0·9883 (0·9818-0·9929)
脑室内	44/44	1·0000 (0·9196-1·0000)	1735/1735	1·0000 (0·9979-1·0000)
硬膜下	41/44	0·9318 (0·8134-0·9857)	1729/1735	0·9965 (0·9925-0·9987)
硬膜外	27/27	1·0000 (0·8723-1·0000)	1749/1752	0·9983 (0·9950-0·9996)
蛛网膜下	51/51	1·0000 (0·9302-1·0000)	1723/1728	0·9971 (0·993-0·9991)
骨折	143/143	1·0000 (0·9745-1·0000)	1636/1636	1·0000 (0·9977-0·0000)
颅盖骨折	88/89	0·9888 (0·9390-0·9997)	1681/1690	0·9947 (0·9899-0·9976)
中线移位	53/54	0·9815 (0·9011-0·9995)	1725/1725	1·0000 (0·9979-1·0000)
肿块效应	129/132	0·9773 (0·9350-0·9953)	1636/1647	0·9933 (0·9881-0·9967)

[0132]

[0133] (a)Qure25k数据集:NLP算法在从报告推断调查结果方面的性能。这是在来自Qure25k数据集的1779份报告上测量的。

[0134]

调查结果	评分员 1 和 2		评分员 2 和 3		评分员 3 和 1		全部
	一致 %	Cohen's κ	一致 %	Cohen's κ	一致 %	Cohen's κ	Fleiss' κ
颅内出血	89·00% (437/491)	0·7772	90·84% (446/491)	0·8084	88·39% (434/491)	0·7646	0·7827
脑实质	91·24% (448/491)	0·7865	90·63% (445/491)	0·7651	90·84% (446/491)	0·7719	0·7746
脑室内	96·13% (472/491)	0·7042	97·15% (477/491)	0·7350	95·72% (470/491)	0·6550	0·6962
硬膜下	87·98% (432/491)	0·4853	93·08% (457/491)	0·6001	90·02% (442/491)	0·5624	0·5418
硬膜外	97·35% (437/491)	0·5058	98·37% (437/491)	0·7251	98·17% (437/491)	0·5995	0·6145
蛛网膜下	93·08% (478/491)	0·6778	90·84% (483/491)	0·6058	90·84% (482/491)	0·6363	0·6382
颅骨骨折	91·85% (451/491)	0·5771	92·06% (452/491)	0·3704	91·24% (448/491)	0·3637	0·4507
中线移位	88·19% (433/491)	0·5804	87·17% (428/491)	0·5344	93·69% 460/491)	0·7036	0·5954
肿块效应	86·35% (423/491)	0·6541	86·35% (432/491)	0·6747	86·97% (427/491)	0·6837	0·6698

[0135] (b)CQ500数据集:评分员之间的一致性。Fleiss等人[24]的指南将κ值>0.75描述为极好的一致,将0.40-0.75描述为相当于良好的一致,并且将<0.40描述为不太可能的差的一致。

[0136] 表8.算法在Qure25k和CQ500数据集上的性能。在训练过程期间,两个数据集都没有使用。示出了对于这两个数据集上9个关键CT调查结果的AUC。在ROC上选择了两个操作点,分别用于高敏感性和高特异性。

[0137]

调查结果	AUC (95% CI)	高敏感性操作点 敏感性 (95% CI)	高敏感性操作点 特异性 (95% CI)	高特异性操作点 敏感性 (95% CI)	高特异性操作点 特异性 (95% CI)
颅内出血	0·9194 (0·9119-0·9269)	0·9006 (0·8882-0·9121)	0·7295 (0·7230-0·7358)	0·8349 (0·8197-0·8492)	0·9004 (0·8960-0·9047)
脑实质	0·8977 (0·8884-0·9069)	0·9031 (0·8894-0·9157)	0·6046 (0·5976-0·6115)	0·7670 (0·7479-0·7853)	0·9046 (0·9003-0·9087)
脑室内	0·9559 (0·9424-0·9694)	0·9358 (0·9085-0·9569)	0·8343 (0·8291-0·8393)	0·9220 (0·8927-0·9454)	0·9267 (0·9231-0·9302)
硬膜下	0·9161 (0·9001-0·9321)	0·9152 (0·8888-0·9370)	0·6542 (0·6476-0·6607)	0·7960 (0·7600-0·8288)	0·9041 (0·9000-0·9081)
硬膜外	0·9288 (0·9083-0·9494)	0·9034 (0·8635-0·9349)	0·7936 (0·7880-0·7991)	0·8207 (0·7716-0·8631)	0·9068 (0·9027-0·9107)
蛛网膜下	0·9044 (0·8882-0·9205)	0·9100 (0·8844-0·9315)	0·6678 (0·6613-0·6742)	0·7758 (0·7406-0·8083)	0·9012 (0·8971-0·9053)
颅骨骨折	0·9244 (0·9130-0·9359)	0·9002 (0·8798-0·9181)	0·7749 (0·7691-0·7807)	0·8115 (0·7857-0·8354)	0·9020 (0·8978-0·9061)
中线移位	0·9276 (0·9139-0·9413)	0·9114 (0·8872-0·9319)	0·8373 (0·8322-0·8424)	0·8754 (0·8479-0·8995)	0·9006 (0·8964-0·9047)
肿块效应	0·8583 (0·8462-0·8703)	0·8622 (0·8439-0·8792)	0·6157 (0·6089-0·6226)	0·7086 (0·6851-0·7314)	0·9068 (0·9026-0·9108)

[0138] (a)Qure25k数据集:算法的性能

[0139]

调查结果	AUC (95% CI)	高敏感性操作点 敏感性 (95% CI)	高敏感性操作点 特异性 (95% CI)	高特异性操作点 敏感性 (95% CI)	高特异性操作点 特异性 (95% CI)
颅内出血	0·9419 (0·9187-0·9651)	0·9463 (0·9060-0·9729)	0·7098 (0·6535-0·7617)	0·8195 (0·7599-0·8696)	0·9021 (0·8616-0·9340)
脑实质	0·9544 (0·9293-0·9795)	0·9478 (0·8953-0·9787)	0·8123 (0·7679-0·8515)	0·8433 (0·7705-0·9003)	0·9076 (0·8726-0·9355)
脑室内	0·9310 (0·8654-0·9965)	0·9286 (0·7650-0·9912)	0·6652 (0·6202-0·7081)	0·8929 (0·7177-0·9773)	0·9028 (0·8721-0·9282)
硬膜下	0·9521 (0·9117-0·9925)	0·9434 (0·8434-0·9882)	0·7215 (0·6769-0·7630)	0·8868 (0·7697-0·9573)	0·9041 (0·8726-0·9300)
硬膜外	0·9731 (0·9113-1·0000)	0·9231 (0·6397-0·9981)	0·8828 (0·8506-0·9103)	0·8462 (0·5455-0·9808)	0·9477 (0·9238-0·9659)
蛛网膜下	0·9574 (0·9214-0·9934)	0·9167 (0·8161-0·9724)	0·8654 (0·8295-0·8962)	0·8667 (0·7541-0·9406)	0·9049 (0·8732-0·9309)
颅骨骨折	0·9624 (0·9204-1·0000)	0·9487 (0·8268-0·9937)	0·8606 (0·8252-0·8912)	0·8718 (0·7257-0·9570)	0·9027 (0·8715-0·9284)
中线移位	0·9697 (0·9403-0·9991)	0·9385 (0·8499-0·9830)	0·8944 (0·8612-0·9219)	0·9077 (0·8098-0·9654)	0·9108 (0·8796-0·9361)

[0140]

肿块效应	0·9216 (0·8883-0·9548)	0·9055 (0·8408-0·9502)	0·7335 (0·6849-0·7782)	0·8189 (0·7408-0·8816)	0·9038 (0·8688-0·9321)

[0141] (b)CQ500数据集:算法的性能

[0142] 1.8 CQ500数据集

[0143] CQ500数据集上三个评分员之间的一致性被观察到是对于颅内出血(Fleiss'κ=0·7827)和脑室内出血(Fleiss'κ=0·7746)的最高值,表示对于这些调查结果的极好的

一致。颅骨骨折和硬膜下出血分别具有0.4507的Fleiss'κ和0.5418的Fleiss'κ的最低一致性,这表明相当于中等的一致。对于目标调查结果的每一个,表8b示出了一对评价员之间的Cohen's kappa和对于所有评价员的Fleiss'kappa的百分比一致。

[0144] 算法通常在CQ500数据集上比在Qure25k数据集上执行的更好。表7b示出了AUC、敏感性和特异性,并且图3示出了ROC。对于颅内出血的AUC为0·9419(95%CI 0·9187-0·9651),对于颅骨骨折的AUC为0·9624(95%CI 0·9204-1·0000),以及对于中线移位的AUC为0·9697(95%CI 0·9403-0·9991)。

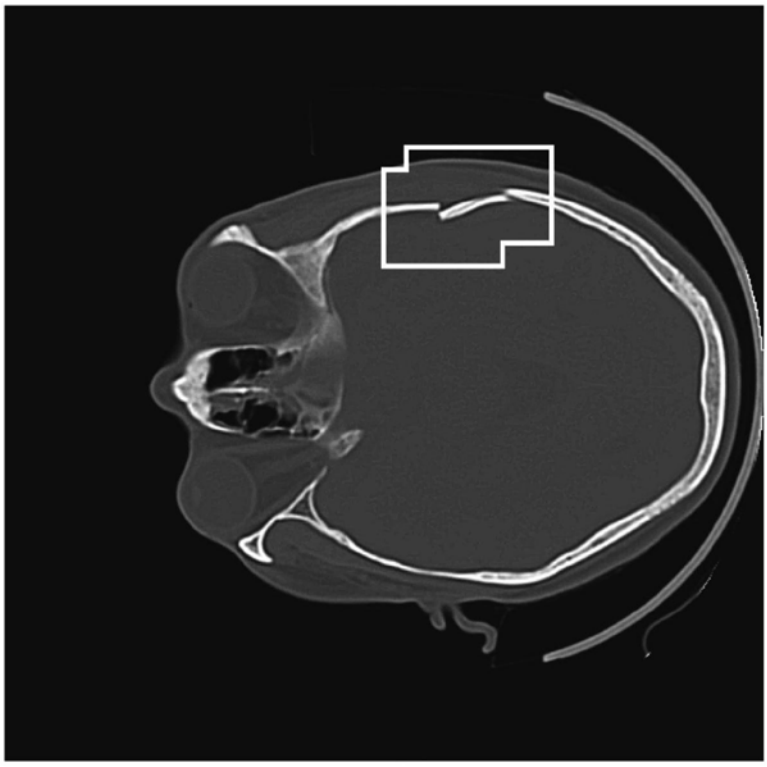
[0145] 1.9头部CT扫描工作流

[0146] 图4示出了使用qER进行头部CT扫描的推荐工作流。

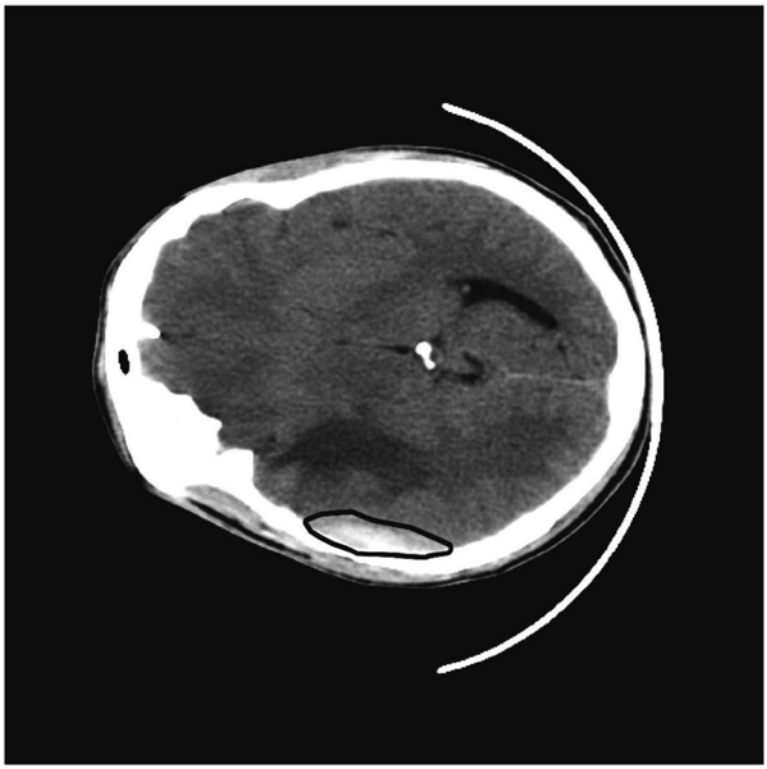
[0147] 1.10结论

[0148] 在Qure25k数据集上,算法针对检测ICH、IPH、IVH、SDH、EDH和SAH分别实现了0.92±0.01、0.90±0.01、0.96±0.01、0.92±0.02、0.93±0.02和0.90±0.02的AUC。同样CQ500数据集上的AUC分别为0·94±0·02、0·96±0·03、0·93±0·07、0·95±0·04、0·97±0·06和0·96±0·04。对于检测颅骨骨折、中线移位和肿块效应,Qure25k数据集上的AUC分别为0·92±0·01、0·93±0·01和0·86±0·01,而CQ500数据集上的AUC分别为0·96±0·04、0·97±0·03和0·92±0·03。

[0149] 示例表明,深度学习算法可以准确地识别需要紧急关注的头部CT扫描异常。这为使用这些算法来自动化分诊过程开辟了可能性。



(b) 由颅骨骨折检测算法产生的定位输出。



(a) 由出血分割算法产生的输出。

图1

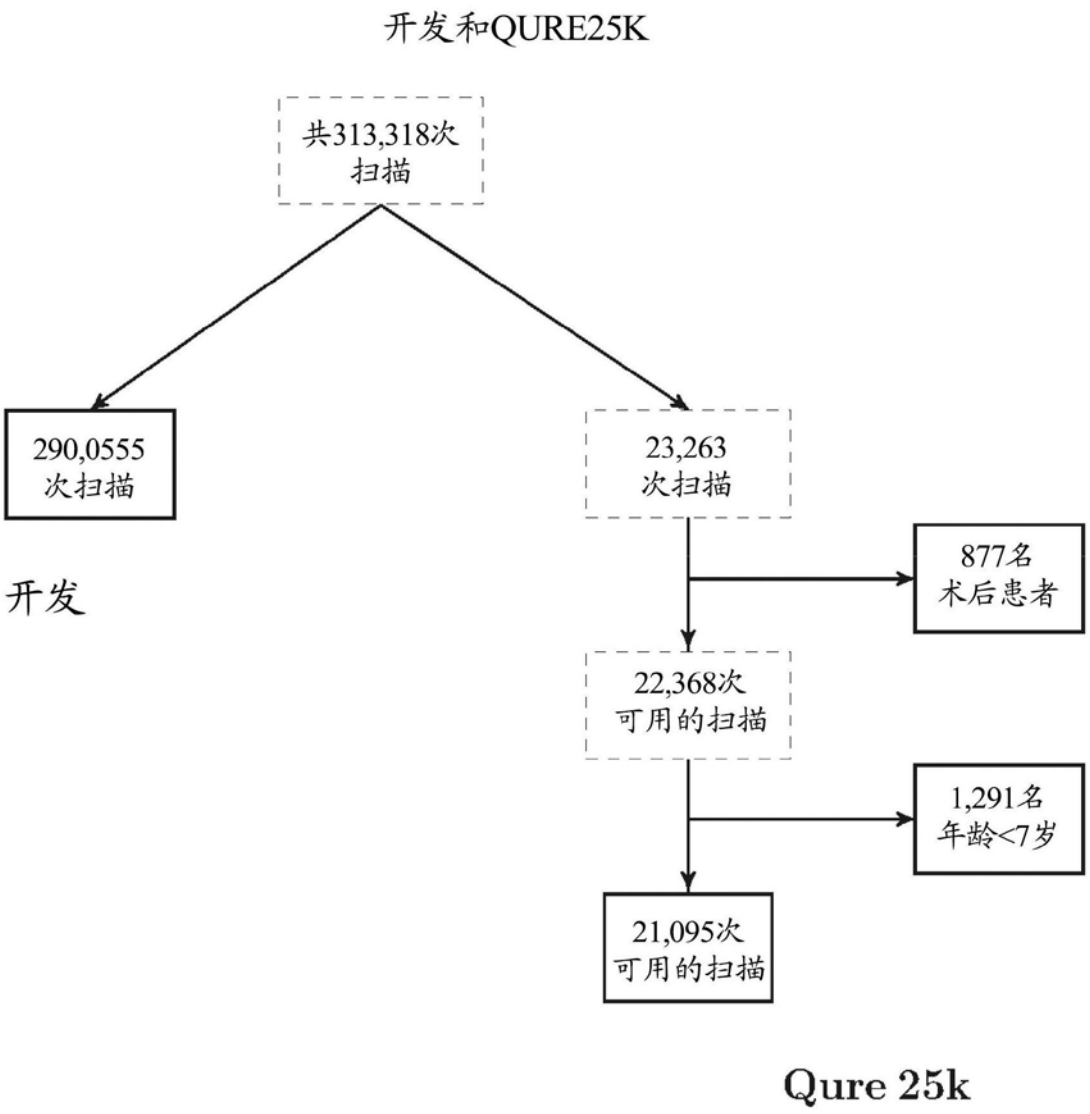

开发和QURE25K
共313,318次
扫描
290,0555
次扫描
开发
23,263
次扫描
877名
术后患者
22,368次
可用的扫描
1,291名
年龄<7岁
21,095次
可用的扫描
Qure 25k

图2a

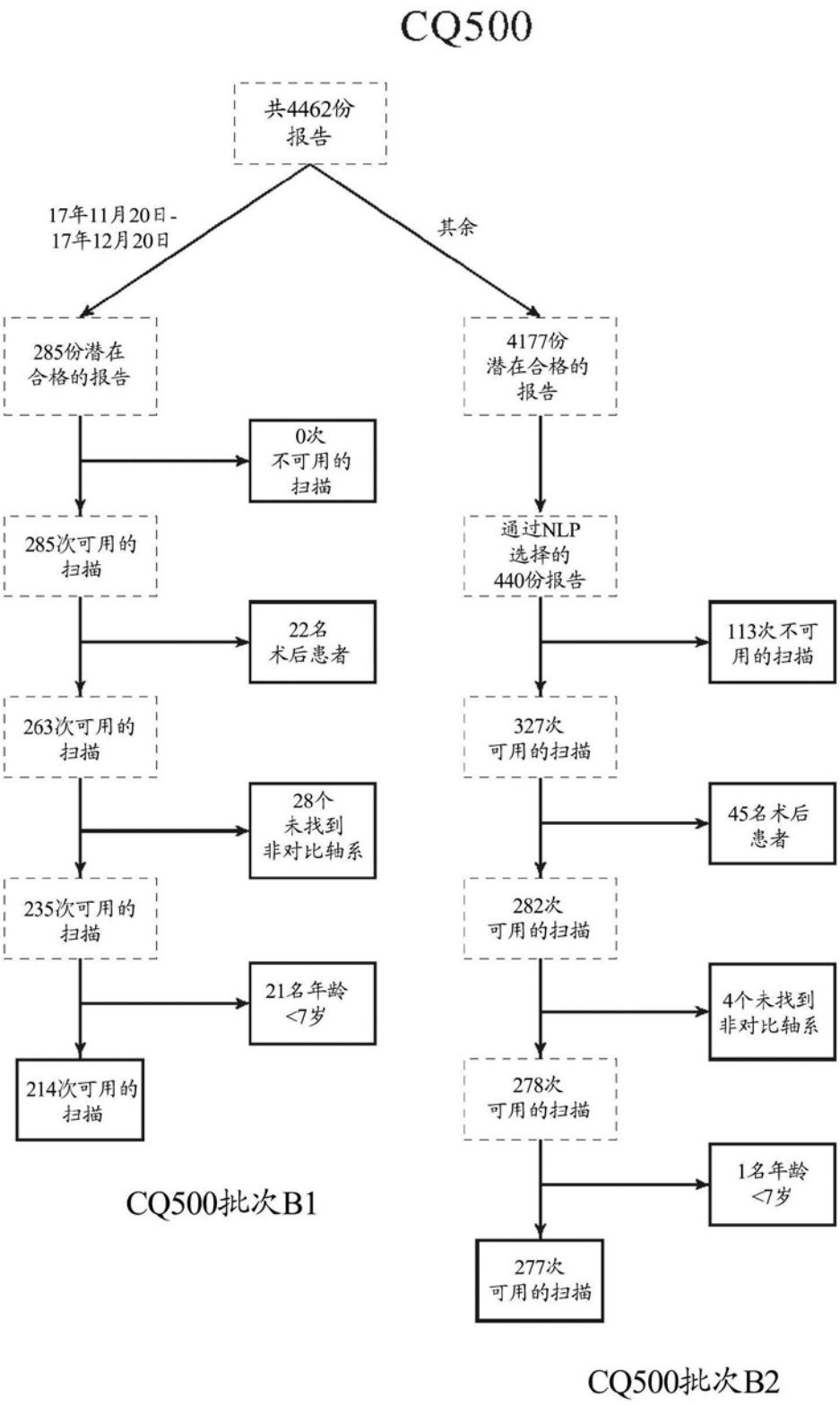
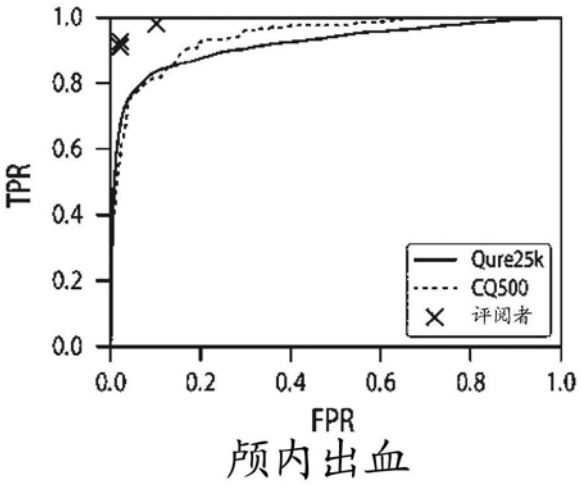
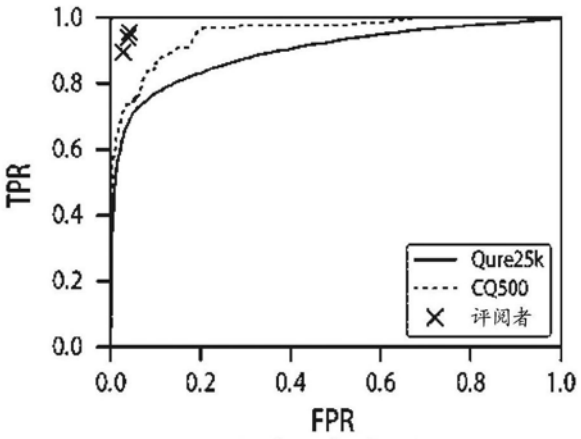

CQ500
共4462份报告
17年11月20日-17年12月20日
其余
285份潜在合格的报告
0次不可用的扫描
285次可用的扫描
22名术后患者
263次可用的扫描
28个未找到非对比轴系
235次可用的扫描
21名年龄<7岁
214次可用的扫描
CQ500批次B1
4177份潜在合格的报告
通过NLP选择的440份报告
113次不可用的扫描
327次可用的扫描
45名术后患者
282次可用的扫描
4个未找到非对比轴系
278次可用的扫描
1名年龄<7岁
277次可用的扫描
CQ500批次B2

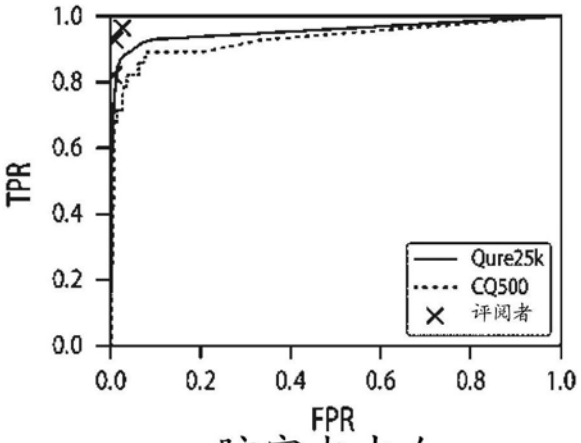
图2b



颅内出血

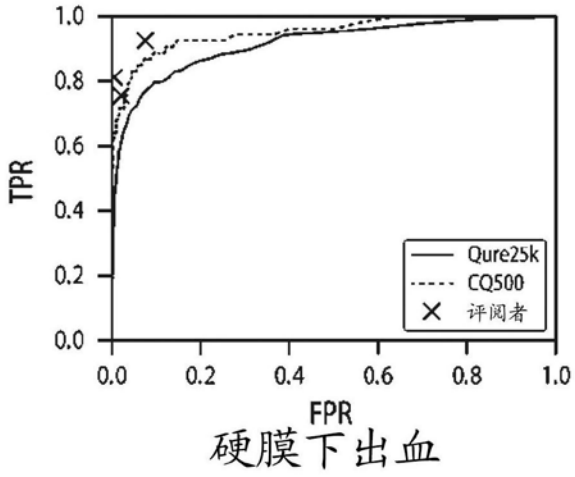


脑实质出血

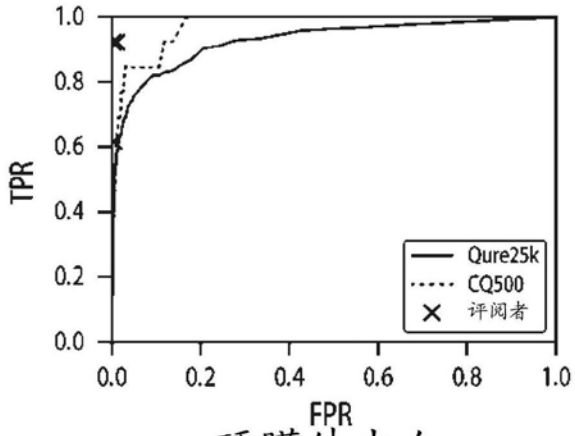


脑室内出血

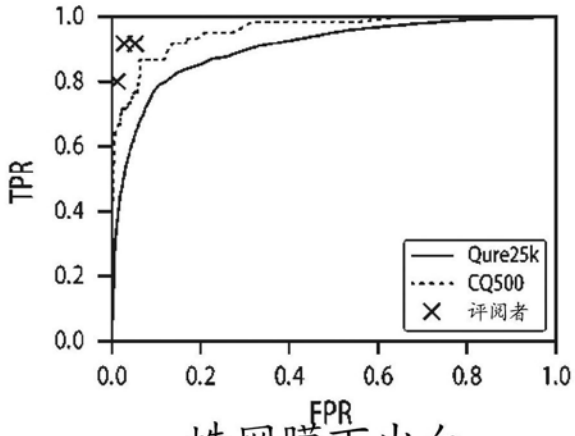
图3



硬膜下出血

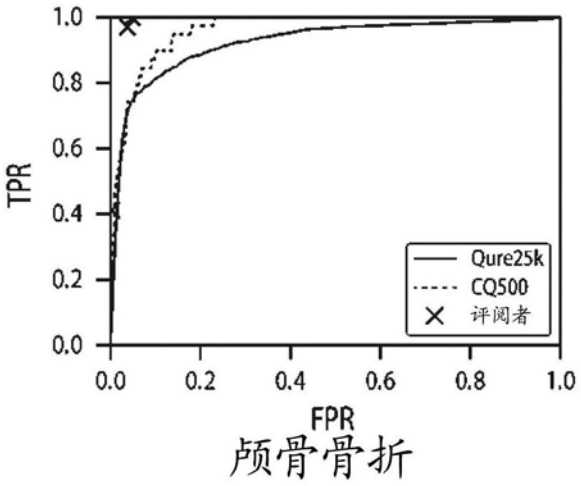


硬膜外出血

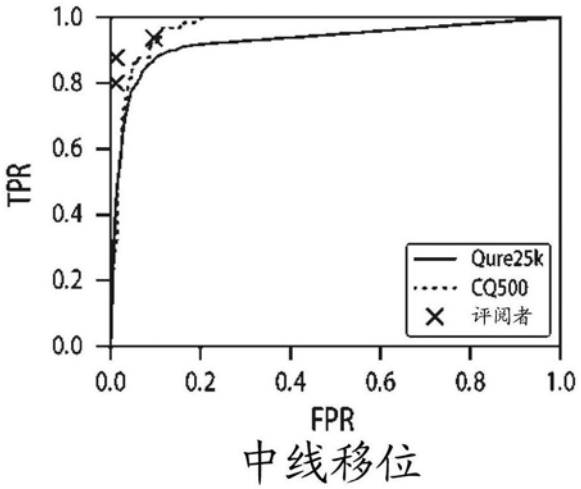


蛛网膜下出血

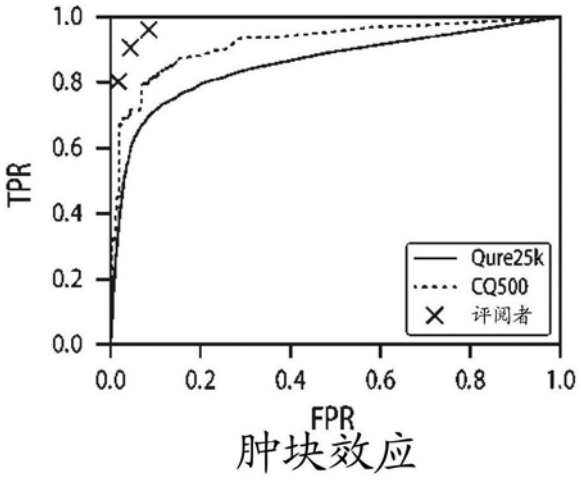
图3继续



颅骨骨折



中线移位



肿块效应

图3继续

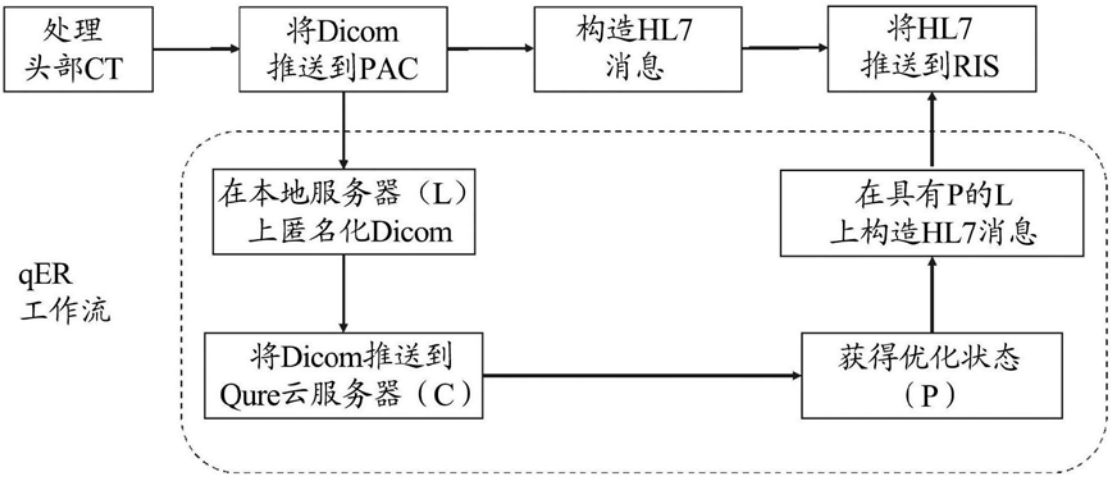

处理
头部CT
将Dicom
推送到PAC
构造HL7
消息
将HL7
推送到RIS
在本地服务器（L）
上匿名化Dicom
在具有P的L
上构造HL7消息
qER
工作流
将Dicom推送到
Qure云服务器（C）
获得优化状态
（P）

图4