2017年2月13日

检测视网膜眼底照片中的糖尿病视网膜病变的深度学习算法开发和验证

重要事项：

深度学习是一系列计算方法，允许算法通过从大量示例中学习来标示自己的程序，这些示例演示了在无需明确指定规则的情况下可以达到理想的效果。这些方法应用于医学成像需要进一步评估和验证。

目的：应用于糖尿病视网膜病变和糖尿病性视网膜黄斑水肿的自动检测算法。

设计和设置：

我们训练了一种针对图像分类做过优化的、被称为深卷积神经网络的特定类型的神经网络。 训练使用了有128175张视网膜图像的回顾性开发数据集，所述视网膜图像按照糖尿病性视网膜病变，糖尿病性黄斑水肿做了3 - 7次的分级。整个分级由54名美国持牌眼科医生和眼科高级专家在2015年5月至12月完成的。所得的算法在2016年1月和2月使用2个独立的数据集进行验证，这两个数据集由至少7个具有高内标一致性的美国董事会认证的眼科医生分级。

曝光：深层学习训练算法。

主要成果和方法：

基于眼科医师小组的多数决定的参考标准，产生定义为中度和严重的糖尿病性视网膜病变，可疑的糖尿病性黄斑水肿或两者都有的用于检测可疑的糖尿病性视网膜病变（RDR）的算法的灵敏度和特异性。用选自开发数据集的2个操作点来评价算法，一个选择高特异性，另一个选择高灵敏度。

结果：

EyePACS-1数据集由来自4997名患者（平均年龄，54.4岁; 62.2％女性; RDR的患病率，683/8878；完全可分级图像[7.8％]）的9963张图像组成; Messidor-2数据集具有来自874名患者（平均年龄，57.6岁; 42.6％女性; RDR的患病率，254/1745；完全可分级图像[14.6％]）的1748张图像。为了检测RDR，算法在接受者操作曲线下对于EyePACS-1具有0.991（95％CI，0.988-0.993）的面积，对于Messidor-2具有0.990（95％CI，0.986-0.995）的面积。使用具有高特异性的第一操作切割点，对于EyePACS-1，灵敏度为

90.3%（95%CI，87.5%-92.7%），特异性为98.1%（95%CI，97.8%-98.5%）。对于Messidor-2，灵敏度为87.0%（95%CI，81.1%-91.0%），特异性为98.5%（95%CI，97.7%-99.1%）。在开发集中使用具有高灵敏度的第二操作点，对于EyePACS-1，灵敏度为97.5%，特异性为93.4%，对于Messidor-2，灵敏度为96.1%，特异性为93.9%。

结论和相关性：

在对来自糖尿病成人的视网膜眼底照片的评价中，基于深度机器学习的算法对于检测可疑糖尿病视网膜病变具有高灵敏度和特异性。为了确定此算法在临床应用的可行性，并确定该算法与目前的眼科评估相比是否可以导致护理和结果的改善，还需要进一步的研究。

关键点

问题：

自动深度学习算法的性能如何与眼科医生用于在视网膜眼底照片中识别糖尿病性视网膜病变的手动分级进行比较？

发现：

用于检测定义为中度或严重的糖尿病性视网膜病变和可疑糖尿病性视网膜病变，由至少7名美国董事会认证的眼科医生组成的小组的大多数决定可引起黄斑水肿。在2个验证集的9963图像和1748图像中，在选择用于高特异性的操作点，该算法具有90.3%和87.0%的灵敏度和98.1%和98.5%的特异性。在选择用于高灵敏度的操作点，该算法在2个验证集中具有97.5%和96.1%的灵敏度和93.4%和93.9%的特异性。

意义：

深度学习算法对于检测视网膜眼底照片中的糖尿病视网膜病变和黄斑水肿具有高灵敏度和特异性。

在糖尿病患者中，美国的透析性视网膜病变的发病率约为28.5%，印度为18%。大多数指南建议对无视网膜病变或轻度糖尿病性视网膜病变的患者进行年度筛查，6个月内对中度糖尿病性视网膜病变进行重复检查，对于严重、更严重的糖尿病性视网膜病变或存在可疑的糖尿病性黄斑水肿（称为临床显着的黄斑水肿），在几周至几个月内进行眼科医师转诊治疗评估。推荐的管理方法从年度筛查变为针对更严重病变的更近的筛查。由人来解读的视网膜摄影是糖尿病视网膜病变的广泛接受的筛选工具，其性能可超过无创扩张的眼科检查。

糖尿病性视网膜病变的自动分级具有潜在的价值，例如提高效率，可重复性和筛选程序的覆盖;减少获取障碍;并通过提供早期检测和治疗来改善患者结果。为了最大化自动分级的临床效用，需要一种检测可疑糖尿病视网膜病变的算法。机器学习（计算机科学中专注于教导机器以检测数据中的模式的学科）已被用于各种分类任务，包括糖尿病性视网膜病变的自动分类。然而，大部分工作集中于“特征工程”，其涉及计算由专家指定的显式特征，从而设计用于检测特定病变或预测是否存在任何水平的糖尿病性视网膜病变的算法。深度学习是一种机器学习技术，通过直接从给定大量数据集的标记示例的图像学习最多的预测特征来避免这样的工程。该技术使用称为反向传播的优化算法来指示机器如何改变其内部参数以最好地预测图像的期望输出。

在这项研究中，深度学习被用来训练一个算法来检测可疑糖尿病视网膜病变并在2个临床验证集中评估算法的性能。

方法

数据集

对于算法开发，在糖尿病性视网膜病变筛查的患者中，从美国的EyePACS和印度的3家眼科医院（Aravind Eye Hositalia，Sankara Nethralaya和Narayana Nethralaya）回顾性地获得以黄斑为中心的视网膜基底图像。所有图像在转移到研究调查员之前根据健康保险便携性和责任法案安全港被识别。使用Quorum Review IRB获得伦理审查和机构审查委员会豁免。

另两个数据集用于临床验证。第一个未识别的数据集由2015年5月至2015年10月期间在EyePACS筛选站点拍摄的以黄斑为中心的图像的随机样本组成。使用多种照相机，包括Centervue DRS，Optovue iCam，Canon CR1 / DGi / CR2和Topcon NW使用45°视野。获取EyePACS图像作为用于糖尿病性视网膜病变筛选的常规临床护理的一部分，并且约40％的图像是用瞳孔扩张获取。该数据集与开发中使用的EyePACS数据不重叠。第二组数据是公开可用的Messidor-2数据集，已被其他组用于糖尿病视网膜病变自动检测算法的基准性能。这些图像是2005年1月至2010年12月在法国3家医院获得的，使用Topcon TRC NW6非照相机和45°视野，以中心凹为中心。大约44％的图像是瞳孔扩张获得的。

表：基线数据特征

特征	开发数据集	EyePACS-1 验证数据集	Messidor-2验证数据集
图像数量	128175	9963	1748
眼科医师人数	54	8	7
每个图像的等级	3-7	8	7

特征	开发数据集	EyePACS-1 验证数据集	Messidor-2验证数据集
每个眼科医生的分级数，中位数（四分位数范围）	2021（304-8366）	8906（8744-9360）	1745（1742-1748）
患者人口统计			
唯一个体数	69573	4997	874
年龄，平均（SD），y	55.1	54.4	57.6
女性，No./total（%）基于知道性别的图片	50769/84734 59.9%	5463/8784 62.2%	743/1745 42.6%
图像质量分布			
可评价图像质量的图像中的完全可分级，总数（%）	52311/69598 75.1%	8788/9946 88.4%	1745/1748 99.8%
由眼科医生的多数决定分类的疾病严重程度分布（参考标准）			
评估糖尿病性视网膜病变和糖尿病性黄斑水肿的总图像，编号（%）	118419 100%	8788 100%	1745 100%
无糖尿病性视网膜病变	53759	7252	1217
轻度糖尿病性视网膜病变	30637	842	264
中度糖尿病视网膜病变	24366	545	211
严重糖尿病性视网膜病变	5298	54	28
增殖性糖尿病性视网膜病变	4359	95	25
潜在的糖尿病性黄斑水肿	18224	272	125
可参考糖尿病视网膜病变RDR	33246	683	254

分级

在开发和临床验证集中的所有图像由眼科医生使用注释工具分级为（补充图1和2）：存在糖尿病视网膜病变，糖尿病性黄斑水肿和图像质量。糖尿病性视网膜病变严重程度（无，轻度，中度，重度或增殖性）根据国际临床糖尿病视网膜病变评分进行分级[14]。可疑的黄斑水肿定义为具有硬质渗出物的。图像质量由评分者使用补充中“评分说明”部分中的量规评估。优良，良好和足够质量的图像被认为是可分级的。

推导

该研究组的54个分级者是他们的美国执照眼科医生或眼科学员学习的最后一年（研究生四年）。每个个体分级了20到62508图像（平均值，9774;中位数，2021）。虽然只有3名学员分级超过1000图像，但实习生的表现并不逊色于许可的眼科医生较通过取得分级者与另一分级者在成对比较的总数中一致的次数来测量每个医生的分级员可靠性。大约10％的推导集（128175图像）被随机选择为被同一个分级员重复分级以确定医生的内在的一致性和可靠性。所有的分级者的工作都支付了报酬。

验证

被邀请分级临床验证集的是具有最高的自我一致性的美国董事会认证的眼科医生。EyePACS-1（n = 8）和Messidor-2（n = 7）。一个简单的多数决定（如果≥50％的眼科医生将其分类为可疑的，则将图像分类为可疑的）作为可引用性和可分级性的参考标准。分级被其他分级者盲评。（有关详细信息，请参阅“补充”中的“分级质量控制”部分。）

算法的发展

深度学习是训练神经网络（具有数百万个参数的大数学函数）以执行给定任务的过程。该函数从眼底图像中的像素的强度计算糖尿病视网膜病变严重性。创建或“训练”此功能需要糖尿病视网膜病变严重程度已知的大量图像（训练集）。在训练过程中，神经网络的参数（数学函数）最初被设置为随机值。然后，对于每个图像，将由函数给出的严重等级与来自训练集合的已知等级进行比较，然后稍微修改函数的参数以减小该图像上的误差。对训练集中的每个图像重复该过程多次，并且函数“学习”如何根据训练集中的所有图像的图像的像素强度准确地计算糖尿病视网膜病变严重性。使用正确的训练数据，结果是一个足以在新图像上计算糖尿病视网膜病变严重程度的函数。本研究中使用的网络是卷积神经网络，其函数首先将附近像素组合成局部特征，然后将它们聚合成全局特征的函数。虽然算法没有明确地检测病变（例如，出血，微动脉瘤），但是它可能学习使用局部特征来识别它们。在这项工作中使用的特定神经网络是Szegedy等人提出的Inception-v3架构。

数据根据“附录”中描述的协议进行预处理。用于训练网络权重的优化算法是Dean等人的分布随机梯度下降函数。为了加快训练，批量归一化以及使用来自同一网络的权重的预初始化，训练已分类对象的网络使用ImageNet数据集。预初始化也提高了性能。训练单个网络进行多个二元预测，包括图像是（1）中度或严重的糖尿病性视网膜病变（即中度，重度或增殖性），（2）严重或更严重的视网膜病变，（3）糖尿病性黄斑水肿，或（4）完全可分级。可参考的糖尿病性视网膜病变定义为满足标准1，标准3或两者的任何图像。

通过绘制灵敏度对-特异性产生的接受者操作曲线下面积（AUC）测量算法的性能。因为本研究中的网络具有大量参数（2千2百万），所以使用早期停止标准（当在单独的调谐集合上达到峰值AUC时停止训练）在收敛之前终止训练。开发集分为两部分：（1）训练：80％的数据用于优化网络权重;（2）调整：20％的数据用于优化超参数（例如，图像预处理选项）。使用对相同数据训练的10个网络的集合，并且通过对集合预测的线性平均值来计算最终预测。

评估算法

对可疑的糖尿病视网膜病变和其他糖尿病视网膜病变分类训练的神经网络产生0和1之间的连续数，对应于图像中存在该病症的概率。通过改变操作阈值来绘制接收器操作曲线，并且从开发集中选择用于算法的2个操作点。第一个操作点近似于眼科医生在用于检测可疑糖尿病视网膜病变（约98％）的推导集中的特异性，并且允许在算法性能与分级验证集的7或8个眼科医生的性能之间更好地进行比较。第二个操作点对应于97％的灵敏度，用于检测可疑的糖尿病视网膜病变，因为高灵敏度是潜在筛选工具的先决条件。

临床验证集统计分析和性能比较

基于2个操作点，产生2×2个表，以表征算法相对于参考标准的灵敏度和特异性，其被定义为基于所有可用级别的眼科医师读数的多数决定。算法在2个操作点的灵敏度和特异性的95％置信区间被计算为“精确”Clopper-Pearson区间，其对应于单独的双侧置信区间，单独的覆盖概率为sqrt（0.95）≈0.975。内在和内部可靠性的95％置信区间是z置信区间。使用StatsModels版本0.6.1和SciPy版本0.15.1 python包计算统计显着性和同时的双侧置信区间。

子采样实验

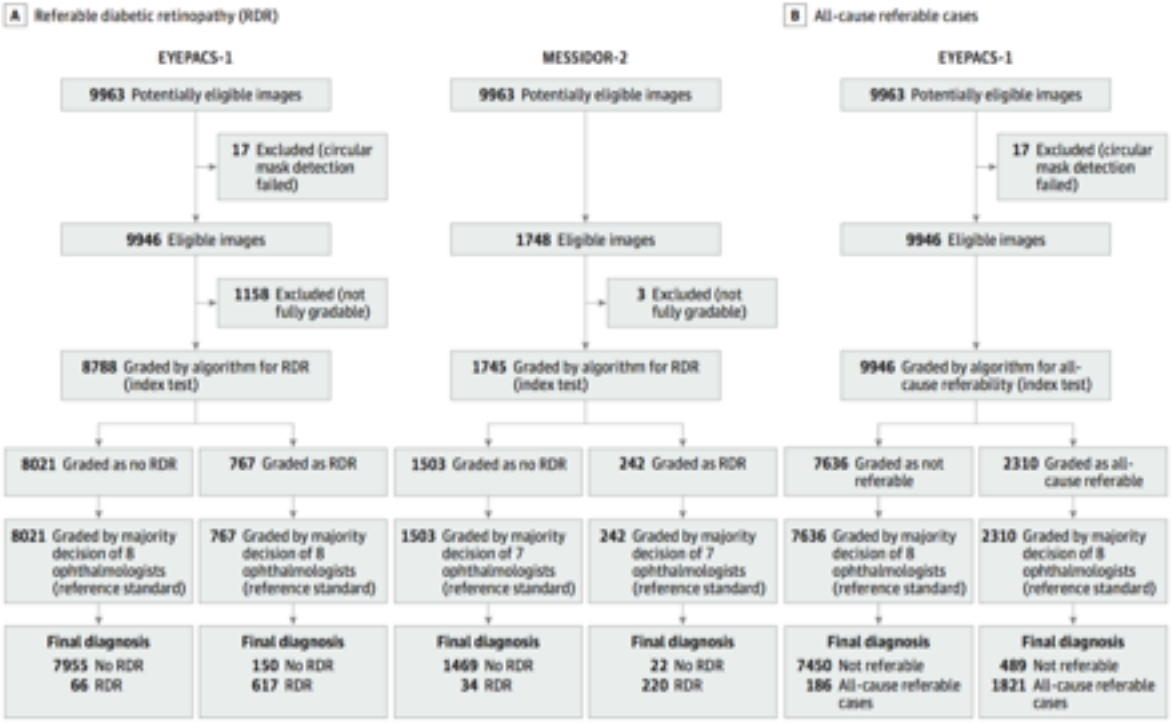
还进行了实验来理解开发数据量与生成算法的性能之间的关系。为了理解减少训练集中图像数量的效果，以0.2％，2％和N×10％的比率采样图像，其中N范围从1到10;针对每个数据集训练新算法;并在固定调优组上测量其性能。为了理解减少每个图像的等级数的效果，运行2个实验：（1）训练：训练集中的等级以N×20％的速率进行子采样，其中N范围从0到5，每个图像采样的最小数量级为1.为每个N训练一个新的算法，其性能在所有可用等级的固定调谐集上测量，以及（2）调整：等级在调优集（用于测量性能）采样与培训实验中相同的过程。训练集和算法是固定的，并使用所有可用的等级。

结果

患者人口统计和图像特征总结在表中。发展集包括128 175图像，其中118 419评估可疑糖尿病视网膜病变和33 246（28.1%）有可引起的糖尿病性视网膜病变。每个图像由眼科医生进行3 - 7次分级。 EyePACS-1和Messidor-2临床验证集分别由9963个图像（8788个完全可分级，683个[7.8%]可参考）和1748个图像（1745个完全可分级，254 [14.6%。仅对开发集的子集评估图像质量，开发集完全可分级图像的范围为75.1%（针对图像质量评估的69 598个图像中的52 311个），Messidor -2验证集为99.8%（1748年的1745）（表和图1）。

图1、EyePACS-1和Messidor-2临床验证集，用于检测糖尿病性视网膜病变和全因可能的糖尿病性视网膜病变

A，可指的糖尿病性视网膜病变，定义为中度或更差的糖尿病性视网膜病变或可疑的糖尿病性黄斑水肿。 B，所有原因的可疑病例，定义为中度或更差的糖尿病性视网膜病变，可疑的糖尿病性黄斑水肿或不可分级的图像质量。



在开发集中，可以在16个评级者中评估眼科医师之间的内标可靠性，这16个评级者已经分级了足够的重复图像。这些分级的可引起糖尿病视网膜病变的平均角膜内可靠性为94.0%（95%CI，91.2%-96.8%）。可靠性可以通过26个分级人员来评估。这些分级人员的平均分级可靠性为95.5%（95%CI，94.0%-96.9%）。

在验证集中，对于EyePACS-1数据集，每个图像总共获得8个等级，对于Messidor-2，每个图像获得7个等级。对于EyePACS-1的可引起糖尿病视网膜病变的平均角膜内可靠性为95.8%（95%CI，92.8%-98.7%）。没有评估Messidor-2的内在可靠性。 EyePACS-

1的平均分级可靠性为95.9%（95%CI，94.0%-97.8%），Messidor-2的平均分级可靠性为94.6%（95%CI，93.0%-96.1%）。

在EyePACS-1上，眼科医师对可引起的糖尿病视网膜病变图像的平均一致性为77.7%（SD，16.3%），完全一致性为19.6%。在不可逆的图像上，平均一致性为97.4%（SD，7.3%），完全一致性为85.6%。在Messidor-2上，眼科医生对可疑糖尿病视网膜病变的平均一致性为82.4%（SD，16.9%），完全一致为37.8%。在不可逆的图像上，平均一致性为96.3%（SD，9.9%），完全一致为85.1%。

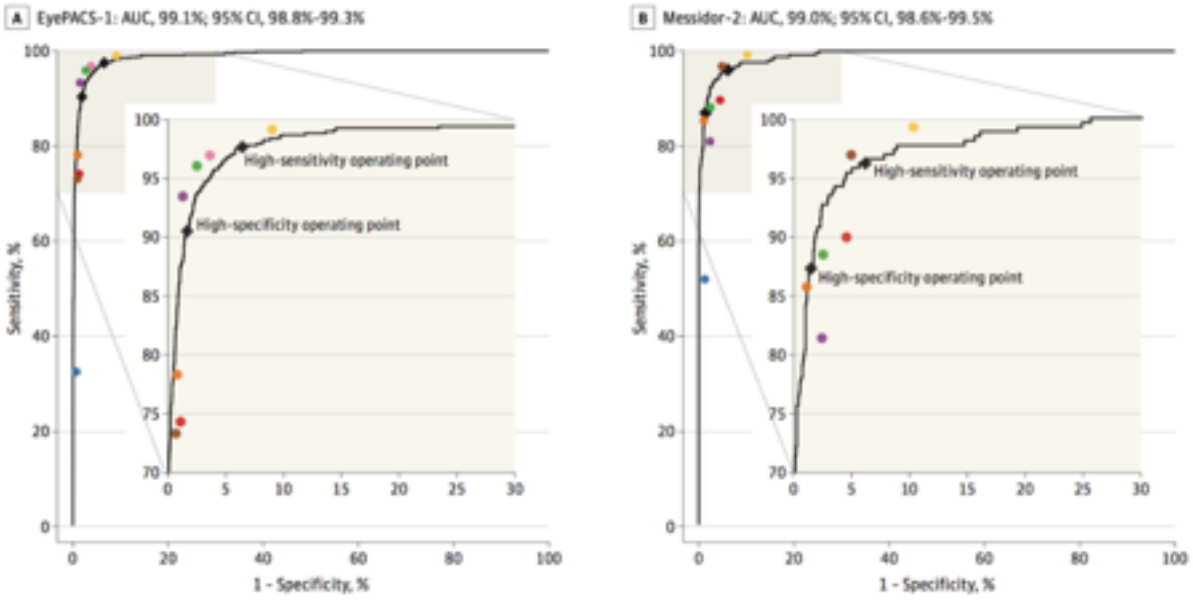


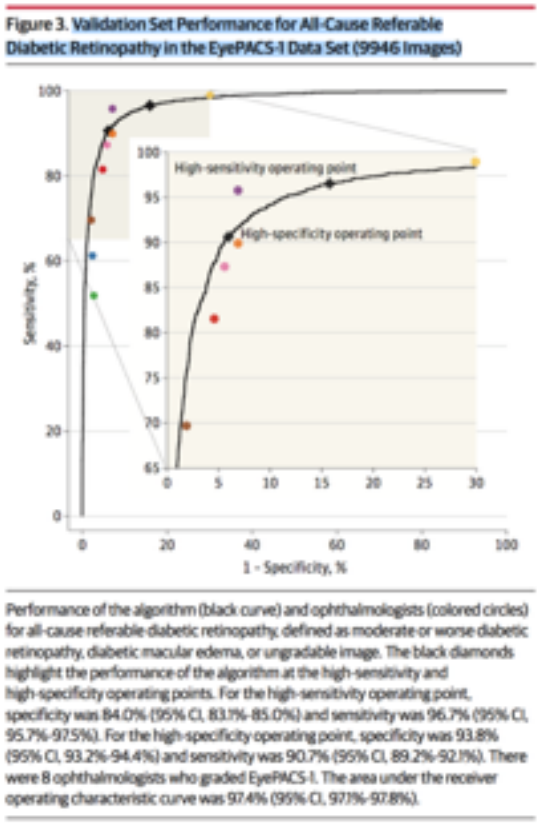
图2 可行性糖尿病性视网膜病变的验证集性能

对于A，EyePACS-1（8788完全可分级图像）和B，Messidor-1和B，存在可引起的糖尿病性视网膜病变（中度或更差的糖尿病性视网膜病变或可疑的糖尿病性黄斑水肿）的算法（黑色曲线）和眼科医生2（1745个完全可分级图像）。图上的黑色菱形对应于算法在高灵敏度和高特异性操作点的灵敏度和特异性。在A中，对于高灵敏度操作点，特异性为93.4%（95%CI，92.8%-94.0%），灵敏度为97.5%（95%CI，95.8%-98.7%）;对于高特异性操作点，特异性为98.1%（95%CI，97.8%-98.5%），灵敏度为90.3%（95%CI，87.5%-92.7%）。在B中，对于高灵敏度操作点，特异性为93.9%（95%CI，92.4%-95.3%），灵敏度为96.1%（95%CI，92.4%-98.3%）。对于高特异性操作点，特异性为98.5%（95%CI，97.7%-99.1%），灵敏度为87.0%（95%CI，81.1%-91.0%）。有8名眼科医生对EyePACS-1评分，7名眼科医生对Messidor-2评分。 AUC表示受试者工作特征曲线下的面积。

图2总结了该算法在用于完全可分级图像的EyePACS-1和Messidor-2验证数据集中检测可引起的糖尿病性视网膜病变中的性能。对于可参考的糖尿病性视网膜病变，算法在EyePACS-1上实现0.991（95%CI，0.988-0.993）的AUC，在Messidor-2上实现AUC为0.990（95%CI，0.986-0.995）。使用具有高度特异性的第一操作切割点，近似在开发

集中的眼科医生在EyePACS-1上的特异性，该算法的灵敏度为90.3%，特异性为98.1%。在Messidor-2中，灵敏度为87.0%，特异性为98.5%。

评估了算法的第二操作点，其对开发集具有高灵敏度，反映将用于筛选工具的输出。使用此操作点，在EyePACS-1上，算法具有97.5%（95%CI，95.8%-98.7%）的灵敏度和93.4%（95%CI，92.8%-94.0%）的特异性。在Messidor-2中，灵敏度为96.1%（95%CI，92.4%-98.3%），特异性为93.9%（95%CI，92.4%-95.3%）。鉴于约8%的可引起糖尿病视网膜病变的患病率（基于每个图像[表]），这些发现对应于EyePACS-1的99.8%和Messidor-2的99.6%的阴性预测值。



还使用EyePACS-1数据集来评估定义为中度或更差的糖尿病性视网膜，可参考的糖尿病性黄斑水肿或不可分级图像（图3）的所有原因可预测的预测的算法性能。 Messidor-2数据集仅具有3个不可分级的图像，因此在该分析中省略。对于该任务，算法实现了AUC为0.974（95%CI，0.971-0.978）。在第一个（高特异性）操作点，算法的灵敏度为90.7%（95%CI，89.2%-92.1%），特异性为93.8%（95%CI，93.2%-94.4%）。在第二个（高灵敏度）操作点，该算法具有96.7%（95%CI，95.7%-97.5%）的灵敏度和84.0%（95%CI，83.1%-85.0%）的特异性。

图3 EyePACS-1数据集中所有原因的RDR的验证集性能

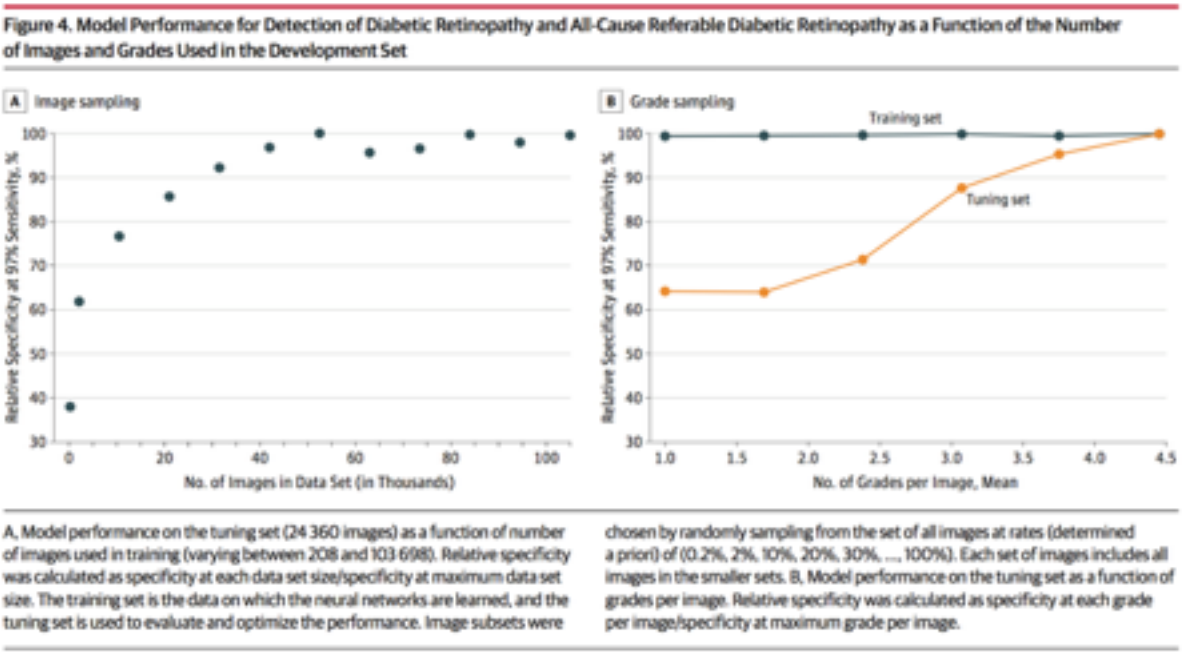
对几个亚类进行了额外的灵敏度分析：（1）仅检测中度或更差的糖尿病视网膜病变;（2）仅检测严重或更差的糖尿病视网膜;（3）仅检测可疑的糖尿病性黄斑水肿;（4）图像质量;和（5）可引起的糖尿病性视网膜病变2个数据集，每个分别限于散瞳和非散瞳图像。对于每个子类别1至4，该算法实现了高灵敏度和特异性（参见“关于个体糖尿病视网膜病变亚型，图像质量的性能”，“eTable 1”和“补充图4”中的部分）。例如，对于EyePACS-1数据集，在中度或更差的糖尿病视网膜病变的第一个操作点，该算法具有90.1%（95%CI，87.2%-92.6%）的灵敏度和98.2%的特异性95%CI，97.8%-98.5%）。对于严重或更严重的糖尿病视网膜病变，只有在第一个操作点，算法的敏感性为84.0%（95%CI，75.3%-90.6%）和特异性98.8%（95%CI，98.5%-99.0%）。

仅对于糖尿病性大鼠水肿，算法的灵敏度为90.8%（95%CI，86.1%-94.3%），特异性为98.7%（95%CI，98.4%-99.0%）。该算法在散瞳图像上的性能非常接近其在非散瞳图像上的性能（并且两者都与整体算法性能相似;参见补充中的eTable 2）。

训练具有不同数量的图像和等级的每个图像的多个网络，以确定更小的训练数据集与训练的算法的性能相关。在第一个子采样实验（图4A）中，考察了数据集大小对算法性能的影响，并显示在大约60 000个图像（或大约17 000个可参考图像）处达到稳定。在二次抽样等级的第二个实验（图4B）中，出现了2个趋势：（1）增加训练集上的每个图像的等级数目不会导致相对性能（31.6%绝对差）和2）在调音集上每幅图像只使用1级，与使用调音集上的所有可用音级（平均为4.5级）相比，性能下降了36%，性能稳定地随着更多的级别而增加可用于调谐集合。这意味着额外的分级资源应用于对调整集（对其进行评估）进行分级，这提高了参考标准的质量和算法性能。

图4 用于检测糖尿病性视网膜病变和全因可能的糖尿病性视网膜病变的模型性能作为开发集中使用的图像和成绩的数量的函数

A，作为训练中使用的图像数量（在208和103 698之间变化）的函数的调谐集合（24个360图像）上的模型性能。 将相对特异性计算为每个数据集大小/特异性在最大数据集大小下的特异性。 训练集是在其上学习神经网络的数据，并且调整集用于评估和优化性能。 通过以（0.2%，2%，10%，20%，30%，...，100%）的速率（先验确定）从所有图像集随机抽样来选择图像子集。 每组图像包括较小集合中的所有图像。 B，模型性能作为每图像等级的函数的调谐集合。 将相对特异性计算为每个图像的每个等级的特异性/每个图像的最大等级的特异性。



讨论

这些结果表明，只使用大数据集，而不必指定基于病变的特征，深层神经网络可以被训练以识别糖尿病性视网膜病变或视网膜基底图像具有的糖尿病性黄斑水肿，这个识别具备高敏感性和高特异性。这种用于检测糖尿病性视网膜病变的自动化系统提供了几个优点，包括解释的一致性（因为机器将每次对特定图像进行相同的预测），高灵敏度和高特异性，以及几乎实时的结果报告。此外，由于算法可以具有多个操作点，其灵敏度和特异性可以被调整以匹配特定临床的需求，例如筛选设置的高灵敏度。在本研究中，实现了97.5％和96.1％的高敏感性。

自动化和半自动糖尿病视网膜病变评价以前已被其他组研究。 Abràmoff等人报道了在公开可用的Messidor-2数据集上检测可疑糖尿病视网膜病变，特异性为59.4％时灵敏度为96.8％. Solanki等人报道的基于相同数据集的敏感性为93.8％，特异性为72.2％。菲利普等人的一项研究报道，在14 406幅图像的数据集上，76.8％的特异性预测疾病与无疾病的灵敏度为86.2％。在最近的Kaggle机器学习竞赛中，使用深度学习来预测糖尿病视网膜病等级（无糖尿病黄斑水肿预测）。获奖作品的表现与眼科医生对同一患者的目视分级的表现相当。尽管数据集和参考标准与以前的数据集和参考标准相比存在差异研究，本研究通过使用深卷积神经网络和每个图像具有多个等级的大数据集来扩展这一工作体系，以产生具有97.5％灵敏度和93.4％特异性的算法（在筛选操作点，其被选择为具有高灵敏度）。当筛选具有实质性疾病的人群时，实现高灵敏度和高特异性对于使假阳性和假阴性结果最小化是至关重要的。

在未来，基于本研究的观察，使用深度学习的类似的高性能算法的开发具有2个前提条件。首先，必须收集一个大的发展集，有数以万计的异常病例。虽然在调优集上的性能在60000幅图像时达到饱和，但通过增加训练数据（即来自新诊所的数据）的多样性可以获得额外的收益。第二，用于测量最终性能的数据集（调优和临床验证数据集）应该每个图像有多次分级。这提供了对模型最终预测能力的更可靠的测量。虽然在许多环境中，人类的解释被用作参考标准（而不是“硬”结果，如死亡率），例如在放射学或病理学中，疾病具有明确的解释，可能不需要每个图像额外的分级。

这个系统有限制。用于本研究的参考标准是所有眼科学家评分者的主要决定。这意味着该算法可能不适用于大多数眼科医生不能识别的具有微小发现的图像。另一个基本限制来自深层网络的本质，其中神经网络仅提供有图像和相关联的等级，而没有特征（例如，微动脉瘤，渗出物）的显式定义。因为网络“隐含地”学习了对于最具预测性的特征，所以该算法可能使用先前未知或被人忽略的特征。虽然本研究使用来自各种临床设置（数百个临床站点：印度3个，美国数百个，法国3个）的图像，以减少算法使用数据中的异常的风险采集进行预测，使用的确切特征仍然是未知的。在大型机器学习社区使用深度神经网络做出预测是一个非常活跃的研究领域。另一个开放的问题是用户界面的设计和眼科医生使用的分级的在线设置是否对其相对于临床设置的性能

具有任何影响。这需要进一步的实验来解决。该算法已被训练为仅鉴定糖尿病性视网膜病变和糖尿病性黄斑水肿。它可能错过未被训练识别的非糖尿病性视网膜病变。因此，该算法不能代替全面的眼部检查，其具有许多部件，例如视敏度，折射，裂隙灯检查和眼压测量。此外，该算法在数据集中进一步的验证是必要的，其中金标准不是参与算法推导的专家的共识。

结论

在对来自糖尿病成人的视网膜眼底照片的评价中，基于深度机器学习的算法对于检测可疑的视网膜病变具有高灵敏度和高特异性。进一步的研究是必要的，以确定在临床中应用此算法的可行性，并确定与目前的眼科评估相比是否使用该算法可以导致护理和结果的改善。