Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Research
Artificial Intelligence—Article

基于深度序列特征学习的临床感染性角膜炎图像分类

许叶圣[a,#]，孔鸣[b,#]，谢文加[a,#]，段润平[a]，方钲清[b]，林宇萧[b]，朱强[b]，汤斯亮[b]，吴飞[b,*]，姚玉峰[a,*]

[a] Department of Ophthalmology, Sir Run Run Shaw Hospital, School of Medicine, Zhejiang University, Hangzhou 310016, China
[b] College of Computer Science and Technology, Zhejiang University, Hangzhou 31002, China

ARTICLE INFO

Article history:
Received 17 January 2020
Revised 18 March 2020
Accepted 20 April 2020
Available online 15 July 2020

关键词
深度学习
角膜病
序列特征
机器学习
长短时记忆

摘要

感染性角膜炎是最常见的角膜疾病之一，病原体在角膜中生长引发炎症反应并损伤角膜组织。感染性角膜炎作为一种临床急症，需要快速、精准的诊断，确保患者能够得到及时、准确的治疗，从而遏制疾病的发展，并将其对角膜的损伤降到最低。否则患者会有失明的风险，严重者甚至会失去眼球。本文提出了一种深度序列特征学习模型，该模型能够通过对临床图像的分类高效地鉴别不同的感染性角膜炎。我们针对感染性角膜炎的特点设计了一种能够解耦临床图像中最具区别性的特征并保持其空间结构的机制。通过比较，我们提出的深度序列特征学习模型在120张图像的测试集上的准确率能够达到80%，远高于421位眼科医生所能达到的平均水平[(49.27 ± 11.5)]%。

1. 引言

通常来说，角膜疾病的诊断是由医生通过观察，结合个人的经验和知识得出的。近年来，使用深度卷积神经网络（CNN）的深度学习算法在医疗图像处理领域大放异彩。疾病分类和诊断算法已在各医学成像技术领域得到了应用与验证，包括计算机断层扫描（CT）、磁共振成像（MRI）、眼底照相、光学相干断层扫描（OCT）和病理图像等[1]。这得益于医学成像技术可以非常自然地导出丰富的图像数据，商业化的医学成像技术对这些医学图像设定了统一的标准，可以在短时间内从单个或多个医学中心收集这些图像。

然而大多数临床疾病的诊断并不需要商业化的医学成像技术，因此许多医疗机构在临床诊疗中不进行图像记录保存，导致这类疾病图像数据的收集主要靠各医疗中心的历史积累，但针对这些临床疾病开发机器学习诊断系统也同等重要。针对皮肤病变良恶性分类的研究[2]是在非传统医疗图像领域的一次成功尝试。角膜疾病分类也可以效仿这种方式。角膜病是世界上最常见的致盲原因之一[3,4]。保守估计，世界上约有450万人因罹患角膜病导致角膜混浊，从而蒙受中到重度的视力损伤[4]。感染性角膜炎是最常见的角膜病[5]。正常的角膜拥有独有的透明性，感染性角膜炎最显著的特点是病原体在角膜内生长，导致角膜局灶性团块混浊和粗糙，不

* Corresponding authors.
E-mail addresses: wufei@zju.edu.cn (F. Wu), yaoyf@zju.edu.cn (Y.-F. Yao).
# These authors contributed equally to this work.

英文原文：Engineering 2021, 7(7): 1002–1010
引用本文：Yesheng Xu, Ming Kong, Wenjia Xie, Runping Duan, Zhengqing Fang, Yuxiao Lin, Qiang Zhu, Siliang Tang, Fei Wu, Yu-Feng Yao. Deep Sequential Feature Learning in Clinical Image Classification of Infectious Keratitis. *Engineering*, https://doi.org/10.1016/j.eng.2020.04.012

可避免地使每种病原微生物在组织内生长并呈现独特的特点[6]。感染性角膜炎的诊断主要取决于眼科医生对角膜感染性病变视觉特征的鉴别。临床上，眼科医生通常依靠裂隙灯显微镜来观察角膜是否异常。除了用作观察工具外，裂隙灯显微镜还可以用于拍摄照片并同时记录每位患者的角膜状态，从而有助于开发出具有优质标注的人工智能（AI）数据集，为感染性角膜炎的识别和分析打下基础。

从1998年开始，我们积累了一个大型且拥有正确标注的裂隙灯显微图像数据集，该数据集来自10 609名角膜疾病患者，总共115 408张图像。基于这个数据集我们设计出一种基于深度学习的方法，以端到端的方式进行感染性角膜炎的诊断。为了直观地模拟眼科医生诊断感染性角膜炎的方式，我们提出了一种特征学习机制，通过学习序列特征来识别不同的视觉模式，其中蕴含着丰富的视觉信息。对于一张临床图像，从感染病灶区域的中心到边缘的样本子块被重新排列，组成一个序列有序集（SOS），并输入神经网络进行特征学习。我们提出的序列特征学习机制可以利用感染性病变区域子块之间的空间关系，并且可以分离出数据样本中蕴含的值得探索的差异性因素。此外，该数据集提供了一个潜在的策略以实现更可靠、有效、准确的诊断。使用该数据集对我们的模型进行评估，获得了比400名眼科医生更高的诊断准确率。

2. 相关工作

2.1. 医疗数据挖掘

多年来，电子病历（EMR）积累了大量的医学数据，这使研究人员从中发现了许多隐藏知识。 数据挖掘方法被广泛应用于医学数据，以发现隐藏的知识，并利用提取的知识来辅助各种有害疾病的预测、诊断和治疗。

疾病预测对于预防疾病的发生和减少疾病的损伤具有重要意义。Yang等[7]使用患者的健康记录来预测糖尿病潜在并发症，同时发现了并发症与实验室类型之间的隐含关系。He等[8]使用EMR数据集预测肺癌术后并发症，并同时从数据集中提取关键变量。

具有预测诊断标签和药物信息的EMR可帮助自动助理预测疾病诊断并为医生提供快速诊断参考。Nee等[9]使用大型EMR文本数据集对每种疾病的EMR上下文进行建模，并在EMR中执行准确的疾病诊断预测。Wright等[10]使用数据挖掘方法从医学数据集中获取有用的关系和规则集，以预测接下来要开哪些药。

2.2. 传统模型在医疗图像中的应用

传统的医学图像使用手工构造特征，通过浅模型进行分类与分割。Scott等[11]在2003年使用梯度取向、拐角和边缘强度检测双能量X射线图像中的椎骨。此外，区域划分和合并也是基于区域的方法中的较为知名的技术。Manousakas等[12]应用区域划分和合并技术，试图克服在MRI中使用均匀性测量时遇到的困难。Zhao等[13]介绍了形态学的基本数学理论和操作，并提出了一种新颖的形态学边缘检测方法，以区分带有椒盐噪声的CT图像中的肺部边缘。实验结果表明，与2006年最佳边缘检测方法相比，该方法在医学图像降噪和边缘检测方面均更有效。除此之外，*K*-means聚类也曾被Kaus等[14]应用在心脏MRI中进行自动左心室的分割。Cordes等[15]通过使用分层聚类来衡量MRI的连通性，该方法可以检测到低频波动的相似性，结果表明，类似于已知神经元连接的层次聚类能够检测心脏连通模式。2006年，Pohl等[16]提出了一种将有符号距离图嵌入线性对数优势空间的方法，可以解决建模问题。上述方法专注于区域、边缘和聚类等简单手段，所以它们在现实世界数据上往往性能有限[17]。

2.3. 深度学习在医疗图像中的应用

在计算机辅助诊断中，深度学习现已广泛用于医学图像识别[18,19]。深度学习的基本结构是CNN，它的结构有三层，即卷积层、池化层和全连接层。为了开发基于CNN的强大AI算法，我们通常需要大量带标注的数据。

医学图像的标准化收集并不像收集一般自然图像那样容易。但是，如今几个公共医学图像数据库和多中心数据收集可以帮助解决该问题。某些类型的医学图像数据可以被大量收集，如X射线图像、CT、心电图和病理图像。通过使用这些大数据，基于CNN的AI算法可以对CT图像进行解剖结构分割[20]，对胸部X射线图像正常或异常结果进行分类[21]，对肺癌或乳腺癌进行筛查[22,23]，检测出颅脑CT扫描中的危险情况[24]，使用基于生成对抗网络（GAN）的模型对肝脏病变进行分类[25]，进行心脏病筛查[26,27]，以及在病理图像中探测淋巴结转移等[27,28]。

在眼科领域，由于眼底照相和OCT图像易于收集，因此基于CNN的AI算法主要应用领域是探查视网膜疾病，如糖尿病性视网膜病变、年龄相关性黄斑变性和青光眼[29–31]。

当前，人工智能辅助的医学诊断系统主要应用于医学成像领域。如果疾病的诊断需要依赖自然观察，则其主要取决于医生的个人经验。皮肤病损是一个例子，当前的AI算法可以将数码皮肤照片中的恶性黑色素瘤与良性病变区分开[2]。角膜病也是一个例子，眼科医生可以使用裂隙灯显微镜来获得正确的诊断，但迄今为止，还没有有关利用AI来提高角膜病的诊断准确性的研究。

3. 方法

3.1. 图像数据集构建

经机构伦理委员会批准，此研究的图像数据集包括了1998年5月至2018年在浙江大学医学院附属邵逸夫医院眼科通过裂隙灯显微镜获得的115 408张临床数字图像，这些图像来自10 609例89种角膜病患者。临床图像由两种类型的裂隙灯显微镜拍摄，即蔡司裂隙灯显微镜SL 130（德国卡尔·蔡司公司），集成了SL Cam for Imaging模块，每个图像的分辨率为1024×768像素；以及附带数码相机Unit DC-1的Topcon裂隙灯显微镜（日本拓普康公司），其图像分辨率为1740×1536像素或2048×1536像素。

在数据集中，拍摄的活跃期角膜感染的图像，包括细菌性角膜炎（BK）、真菌性角膜炎（FK）和单纯疱疹病毒基质性角膜炎（HSK）的图像，被选择用于算法分类的训练或测试集。来自角膜感染患者的所有图像均带有明确的临床诊断标注，至少有以下证据中的两种：①角膜感染的临床表现如图1（a）所示。②通过相关的诊断性单药或联合药物治疗，影响和终止了角膜感染的进展，从而最终治愈。③感染部位样本的病原体鉴定：在细菌和真菌感染中，通过显微镜检查涂片或生物培养确认病原诊断；在病毒感染中，通过聚合酶链反应（PCR）评估泪液或角膜刮片组织样本进行病原诊断。除了角膜感染的类别外，患有具有相似视觉特征的其他角膜病患者的图像被归类为“其他”类别。此类别包括各种角膜变性、泡性角结膜炎、各种角膜肿瘤、角膜乳头状瘤、角膜退行性变，甚至还包括棘阿米巴角膜炎。每个类别的代表性图像如图1（a）所示。

最终数据集包含来自867例患者的2284幅图像。训练集包括来自747例患者的387张随机选择的BK图像、519张FK图像、488张HSK图像和528张其他角膜病图像。测试集由120例患者随机选择的86张BK图像、97张FK图像、51张HSK图像和128张其他诊断图像组成。

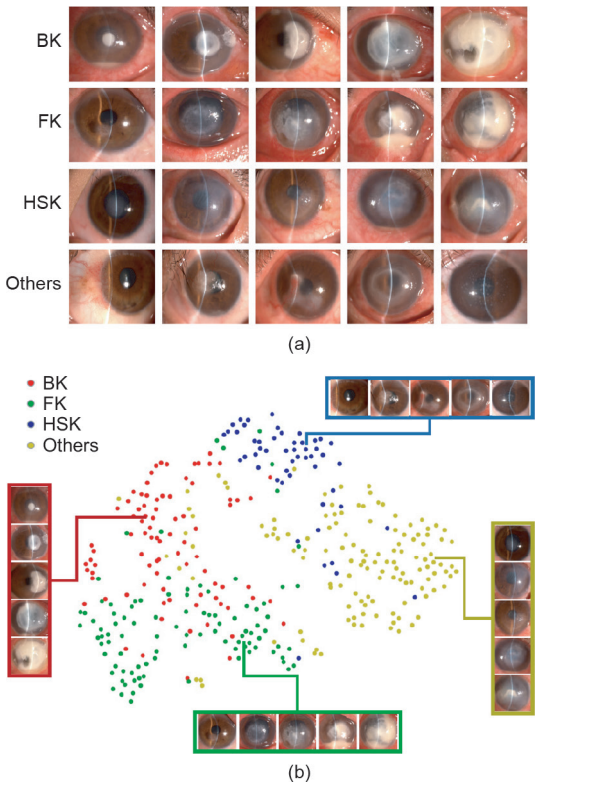


图1. 典型裂隙灯显微图像与我们提出的四类角膜疾病的SOS模型中嵌入特征的*t*-分布随机邻域嵌入（*t*-SNE）的表示可视化。（a）细菌性角膜炎（BK）、真菌性角膜炎（FK）、单纯疱疹病毒基质性角膜炎（HSK）和其他角膜炎（上述三类角膜疾病之外的角膜炎）的典型裂隙灯显微图像。不同类别的疾病或同一类别疾病的不同阶段表现出不同的视觉特征。（b）SOS模型通过*t*-SNE嵌入每种疾病类别的二维空间获得的深度序列特征。*t*-SNE用于可视化高维数据，这些数据是经过诊断验证的摄影测试集的SOS模型中的特征表示（362幅图像）。彩色的点云代表疾病的不同类别，显示了算法如何将疾病分组为不同的簇。插图显示了与各个点相对应的图像。

为了评估眼科医生的疾病分类表现，选择测试集中每位患者的首次诊断图像以构建评估眼科医生诊断水平的数据集（即总共使用了120张图像来评估眼科医生的表现）。

3.2. 基于深度序列特征的诊断模型

如前所述，我们发明了一种使用图像序列特征的感染性角膜炎分类模型训练方法。为了展示我们提出的这一方法的优势，我们将其与另两类方法——基于完整图像特征的学习方法和基于图像子块特征的方法——进行了对比。

基于完整图像特征的深度学习模型接收原始的无标注疾病图像，直接由CNN进行分类。为解决训练图像有限的问题，我们采用了迁移学习的方法[32,33]。在实验中，我们选择了如下三种经典的图像分类模型结

构：VGG-16（由英国牛津大学计算几何小组提出）[34]、GoogleLeNet-v3（由谷歌公司修改的LeNet模型的第三个版本）[35]以及稠密卷积网络（DenseNet）[36]。

在基于图像子块特征深度学习模型当中，眼前节图像由手工进行初步标注，分割出四块：角膜感染病灶区、角膜感染灶旁区、结膜充血区和前房积脓区。在该类方法中我们同样使用了三种迁移模型结构，即VGG-16、GoogleLeNet-v3和DenseNet。在对图像的每个子块完成分类后，通过多数投票的方法给出整张疾病图像的分类结果。

在基于序列的深度学习模型中，对于每张图片，如果其中存在病变区域，模型的关注点就会被设置在病变区域上。病变区的最小外接圆被进一步划分为由小到大的K个圆环，如图2所示。落在由内到外的第i个圆环上的所有图像子块组成一个子块集合S_i，所有集合组成一个由内而外的序列$\{S_1, S_2, \ldots, S_K\}$。为解决标注数据不足的问题，在训练阶段，我们通过随机失活机制来随机剔除每个集合中的一些元素，这样我们能够获得更多的序列来扩大训练集多样性，使得模型的训练更加稳定。

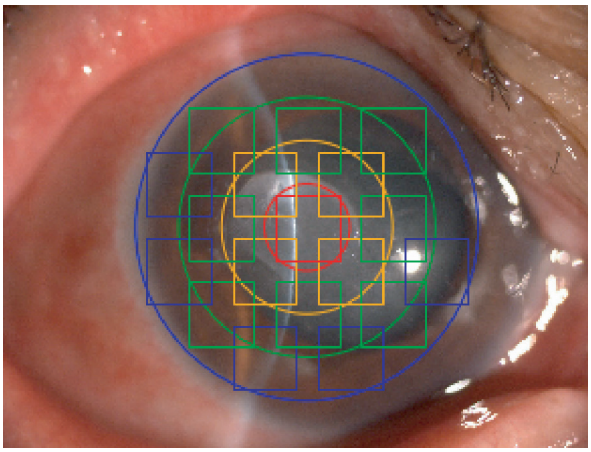


图2. 采样以及如何将病变区域分成K个集合。圆圈代表每个集合的边界，正方形代表采样区域。要注意的是为避免图片过度重叠，仅显示了一半的采样区域。

我们利用一个深层的残差CNN（如DenseNet），通过编码器-解码器框架[37–39]来提取每个集合中的每个子块的表征。卷积编码器能够将第i个集合中的第j个子块p_{ij}转化为一个能够表示其性质的表征向量f_{ij}，得到图像子块的表征的集合$\{F_1, F_2, \ldots, F_K\}$。对每个集合$F_i$，通过最大池化计算获得整个子块集合的表征$f'_i$，表示每个集合的整体性质。由于病变区由内到外的子块集合构成了一个序列，我们可以使用长短时记忆（long short-term memory, LSTM）模型[37]——一种经典的序列学习模型，来将表征序列$\{f'_1, f'_2, \ldots, f'_K\}$转化为最终用于分类的表征。最终的图像特征可以通过一个全连接层解码，通过softmax计算得到每个疾病类别的分类概率。图1（b）展示了每种疾病的图像表征在二维空间上的分布。图3展示了整个系统的结构。经由获得的概率分布与真实标签的损失函数，通过反向传播精调模型参数[40,41]。

3.3. 眼科医生基于图像的诊断能力分析

我们从全国各地招募眼科医生，测试他们基于图像进行诊断的能力，并将其与已开发的深度学习方法进行比较。呈现给眼科医生的图像和测试集中每位患者的经过诊断验证的图像都是从初次就诊时随机选择的（即总共120张图像）。所招募的眼科医生具有不同的学术头衔（从住院医生到高年资医生，甚至到医学院的临床教授）、工作机构（从大学医学院的教学医院到公立市级医院再到社区诊所）和专业经验（分为1~5年、6~10年、11~15年、16~20年以及20年以上）。我们总共招募了421名眼科医生。

眼科医生手动检视图像从而进行基于图像的诊断，

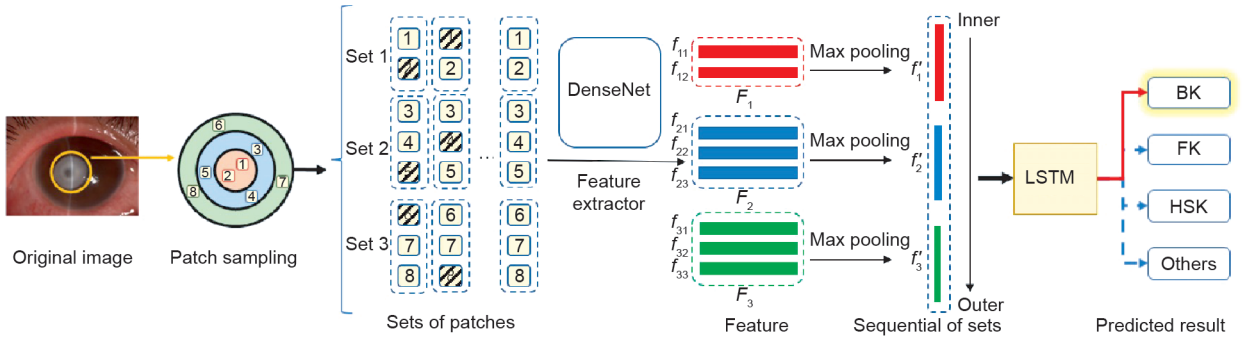


图3. 病变区域进行深度序列特征学习的过程。对于每张裂隙灯显微图像，病变区域外接圆被划分为K个圆环（直观起见，此处$K = 3$）。我们从每个圆环（从最内层到最外层）上采样，采样的区域用于生成集合序列。随后通过最大池化和长短期记忆（LSTM）模型进行序列特征学习。f：采样区域的特征；f'：集合的序列特征。

主要遵循两步流程。第一步，眼科医生仅依靠图像进行诊断，将来自测试集中每位患者的首次诊断图像中的四类角膜病图像，即BK、FK、HSK和其他角膜病的图像提供给眼科医生，眼科医生通过手动选择为每个图像做出诊断。第二步，为眼科医生提供与每个图像相关的其他标准化和结构化的医学信息，包括简短的病史、起病时间、疼痛程度和复发情况（如果有）以及药物使用史。第三步，要求眼科医生通过手动检视并考虑其他医学信息，对每个图像做出诊断决定。所有眼科医生均独立执行此程序，没有时间限制。

3.4. 统计学分析

鉴于不同的学术头衔、工作机构和专业经验导致的置信度不同，使用社会科学统计软件包（SPSS 18.0版；美国Cary公司）对眼科医生的诊断数据进行统计分析。眼科医生诊断准确度的平均水平以（均值±标准差）%表示。使用Kolmogorov-Smirnov检验验证数据的正态性。根据数据正态性，使用单因素方差分析（ANOVA）分析了不同医院级别和职称组之间的诊断准确性差异。最小显著差异法用于参数变量的事后分析。使用皮尔逊相关系数测试了诊断准确性和专业经验年资之间的相关性。采用逐步法进行多元线性回归分析，从学术头衔、医院层级和专业经验年资等方面分析了统计学因素的影响。进行了配对*t*检验（针对正态分布的变量）和Wilcoxon符号秩检验（针对非正态分布的变量），以确定在有无附加医学信息的情况下，医生在诊断准确性上是否存在显著差异。所有检验的显著性水平设置为0.05。

4. 结果

4.1. 不同深度模型的性能对比

完整图像疾病的深度模型目前在医学图像诊断任务中非常流行，此类方法直接将原始医学图像交由CNN处理。我们的工作中选择了VGG-16、GoogleLeNet-v3和DenseNet三种经典模型结构，并对比了其对BK、FK、HSK的诊断准确率，如表1所示。考虑到直接将完整图像传入CNN可能会含有一些无关的信息，我们进一步基于这三种结构设计了基于图像子块的模型[42,43]。在基于图像子块的模型中，图像包含由人工标注的角膜感染病灶区、角膜感染灶旁区、结膜充血区和前房积脓区等区域信息，我们并不传入完整的原始图像，而是使用包含这些区域的图像子块进行分类。三种基于子块的模型对每个子块的识别准确率分别能够达到49.62%、51.52%和60%。在每个图像子块分类后，通过多数投票的方法给出整张图像的诊断结果。基于图像子块的模型对图像的分类准确率分别可以达到52.50%、55.52%和66.30%，如表1所示。

最后，我们使用基于图像序列特征的模型，该方法能够保持医学图像中潜在的空间结构信息。如前所述，子块集合序列的表征是通过由内而外的顺序排列获得的（即SOS），使用SOS提取特征能达到78.73%的分类准确率。除了产生集合序列外，我们还生成了随机排序的子块（ROP）和顺序排列的子块（SOP）来获得特征序列。ROP通过随机选择产生子块序列，而SOP按照从内到外的顺序采样序列（但是没有将属于同一圆环的子块聚成一个集合）。实验结果表面，ROP方法能够达到74.23%

表1 不同深度模型在测试集上的诊断能力对比

Level	Algorithm	Test dataset (%)				
		Acc	BK	FK	HSK	Others
Image-level	VGG-16 (image)	55.24	48.84	52.57	62.74	53.90
	GoogLeNet-v3 (image)	57.73	53.49	55.67	66.67	58.59
	DenseNet (image)	61.04	60.46	56.70	80.39	57.03
Patch-level	VGG-16 (voting)	52.50	45.34	54.64	56.00	54.68
	GoogLeNet-v3 (voting)	55.52	44.19	51.55	74.51	58.59
	DenseNet (voting)	66.30	59.30	68.04	58.82	72.66
Sequence-level	Random-ordered patches (ROPs)	74.23	75.29	68.04	82.35	75.00
	Sequential-ordered patches (SOPs)	75.14	66.28	86.60	84.31	68.75
	SOSs	78.73	65.12	83.51	90.20	79.70

The image-level, patch-level, and sequence-level features are learned from the whole image, the lesion area, and the sequence of patch sets, respectively. The test dataset contains 362 images, including 86 BK, 97 FK, 51 HSK, and 128 others from 120 patients. Acc indicates the overall accuracy of each model, and columns BK, FK, HSK, and others show the recall for each corresponding category.

的准确率（BK 75.29%、FK 68.04%、HSK 82.35%）。SOP方法达到75.14%的准确率。这些实验结果表明我们基于序列的深度学习模型在基于图像的角膜病分类诊断任务上能获得最好的表现。

4.2. 与眼科医生的诊断正确率对比

我们使用数据集评估了本文中考虑的所有算法，以比较每种算法与眼科医生的表现。表2列出了所有算法的准确性以及该数据集（120幅图像）上眼科医生的平均表现。表3列出了眼科医生在临床图像诊断中的表现。从全国各地共招募421名眼科医生参加这项研究。在没有参考任何额外医疗信息的眼科医生的平均准确度为（49.27 ± 11.5）%（范围：20.00%~86.67%），远低于AI深度学习模型所获得的准确性。例如，SOS算法的诊断

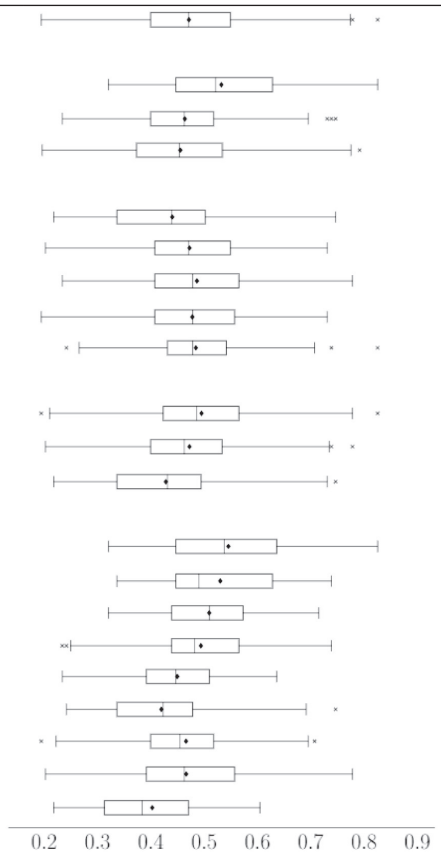
表2 在样本量为120的测试集上眼科医生与深度模型的对比

Level	Algorithm	Dataset for evaluation of ophthalmologists (%)				
		Acc	BK	FK	HSK	Others
Image-level	VGG-16 (image)	50.83	46.67	43.33	73.33	40.00
	GoogLeNet-v3 (image)	55.83	50.00	63.33	70.00	40.00
	DenseNet (image)	64.17	56.67	63.33	80.00	56.67
Patch-level	VGG-16 (voting)	51.67	23.33	43.33	76.67	63.33
	GoogLeNet-v3 (voting)	54.17	26.67	73.33	80.00	36.67
	DenseNet (voting)	71.67	46.67	86.67	73.33	80.00
Sequence-level	ROPs	77.50	66.67	70.00	93.33	80.00
	SOPs	79.17	73.33	70.00	96.67	76.67
	SOSs	80.00	53.33	83.33	93.33	90.00
Human-level	Average performance of ophthalmologists provided with image only	49.27	46.55	45.56	65.01	39.95
	Average performance of ophthalmologists provided with image together with medical history	57.16 [a]	55.55 [a]	56.28 [a]	73.25 [a]	43.56 [a]

The first-visit and diagnosis-proven images of patients with four categories of the corneal diseases were selected from the testing set to construct a dataset for evaluation and comparison of deep learning models with ophthalmologists. The dataset included 120 clinical images.
[a] $P < 0.001$ compared to the average performance of ophthalmologists provided with image only.

表3 根据医院级别、工作年限和眼科医生的专业职称的平均分类准确率

Dr. Group	Participant number	Boxplot	Mean ± STD (%)	Range (%)
Total	421		49.27 ± 11.85	[20.00, 86.67]
Hospital RK.				
Teaching	84		55.69 ± 12.19	[33.33, 86.67]
City	171		48.46 ± 10.70	[24.17, 78.33]
Community	166		46.84 ± 11.63	[20.00, 81.67]
Year of employments				
1–5	89		45.96 ± 13.22	[22.50, 78.33]
6–10	117		49.39 ± 11.80	[20.83, 76.67]
11–15	69		50.85 ± 12.03	[24.17, 81.67]
16–20	41		49.94 ± 12.30	[20.00, 76.67]
> 20	105		50.64 ± 9.66	[25.00, 86.67]
Physician RK				
Attending	173		51.76 ± 11.94	[20.00, 86.67]
Fellow	150		49.38 ± 11.18	[20.83, 81.67]
Resident	98		44.69 ± 11.66	[22.50, 78.33]
Hospital RK and physician RK				
Attending in teaching	36		57.08 ± 12.02	[33.33, 86.67]
Fellow in teaching	30		55.47 ± 12.21	[35.00, 77.50]
Resident in teaching	18		53.29 ± 12.21	[33.33, 75.00]
Attending in city	78		51.63 ± 10.45	[24.17, 77.50]
Fellow in city	59		46.96 ± 9.07	[24.17, 66.67]
Resident in city	34		43.80 ± 11.57	[25.00, 78.33]
Attending in community	59		48.69 ± 10.36	[20.00 74.17]
Fellow in community	61		48.72 ± 12.53	[20.83, 81.67]
Resident in community	46		41.99 ± 10.51	[22.50, 63.33]



RK: rank; STD: standard deviation; * $P = 0.003$; ** $P < 0.001$

准确度达到80.00%，其中BK、FK和HSK的准确度分别为53.33%、83.33%和93.33%（表2）。图4描绘了受试者工作特征（ROC）曲线、SOS模型的混淆矩阵以及眼科医生的表现。ROC曲线是分类模型的一种可视化。曲线下的面积（AUC）是性能的衡量，最大值为1。如果眼科医生的敏感度-特异性点位于分类模型的曲线下方，则该模型可达到优于眼科医生的性能。

对于工作地点对眼科医生诊断水平的影响，教学医院的眼科医生的表现比市级医院和社区诊所的要好（$P < 0.001$），而市级医院和社区诊所之间没有统计差异（$P = 0.226$）。专业级别较高的眼科医生在诊断临床图像时具有更高的准确性，如高级职称和中级职称医师的表现优于住院医师（分别为$P < 0.001$和$P = 0.003$），但高级职称和中级职称之间差异无统计学意义（$P = 0.071$）。在执业年资和诊断准确性之间未发现显著相关性（$P = 0.084$）。

综合考虑医院等级和医生等级的因素，教学医院的高级职称眼科医生的表现[准确性为(57.08 ± 12.02)%，范围：33.33%~86.67%]好于社区诊所的住院医生[准确性为(41.99 ± 10.51)%，范围：22.50%~63.33%]。逐步多元回归分析得出影响诊断准确性的三个模型。模型1（$R^2 = 0.062$）仅具有医院水平的因素（$\beta = 0.254$, $P < 0.001$）；模型2（$R^2 = 0.100$）具有医院水平（$\beta = 0.239$, $P < 0.001$）和专业职称（$\beta = 0.200$, $P < 0.001$）的因素；模型3（$R^2 = 0.109$）具有所有三个因素：医院水平（$\beta = 0.227$, $P < 0.001$）、专业职称（$\beta = 0.326$, $P < 0.001$）和工作年限（$\beta = -0.164$, $P = 0.024$）。

当进一步为眼科医生提供每张图像的附加医学信息时，包括简短的病史、发病时间、疼痛程度和复发发作（如果有）以及用药史，平均总诊断准确性从49.27%提高到57.16%，差异具有统计学意义（Wilcoxon符号秩检验，$P < 0.001$）。详细而言，附加医学信息后，BK的诊断准确性从46.55%提高到55.55%（$P < 0.001$），FK从45.56%提高到56.28%（$P < 0.001$），HSK从65.01%提高到73.25%（$P < 0.001$），404名医生的平均总准确性提高了8.28%，9名医生的准确性下降了2.13%，其他8名医生的准确性保持不变。

5. 讨论

一般而言，人类通过视觉、听觉、触觉、味觉和嗅觉对事物进行判断，并以此对事物进行适当的分类[44]。其中视觉感知发挥着最重要的作用[45]，视觉知识可以描述空间形状、大小和相关性以及颜色和纹理之间的关

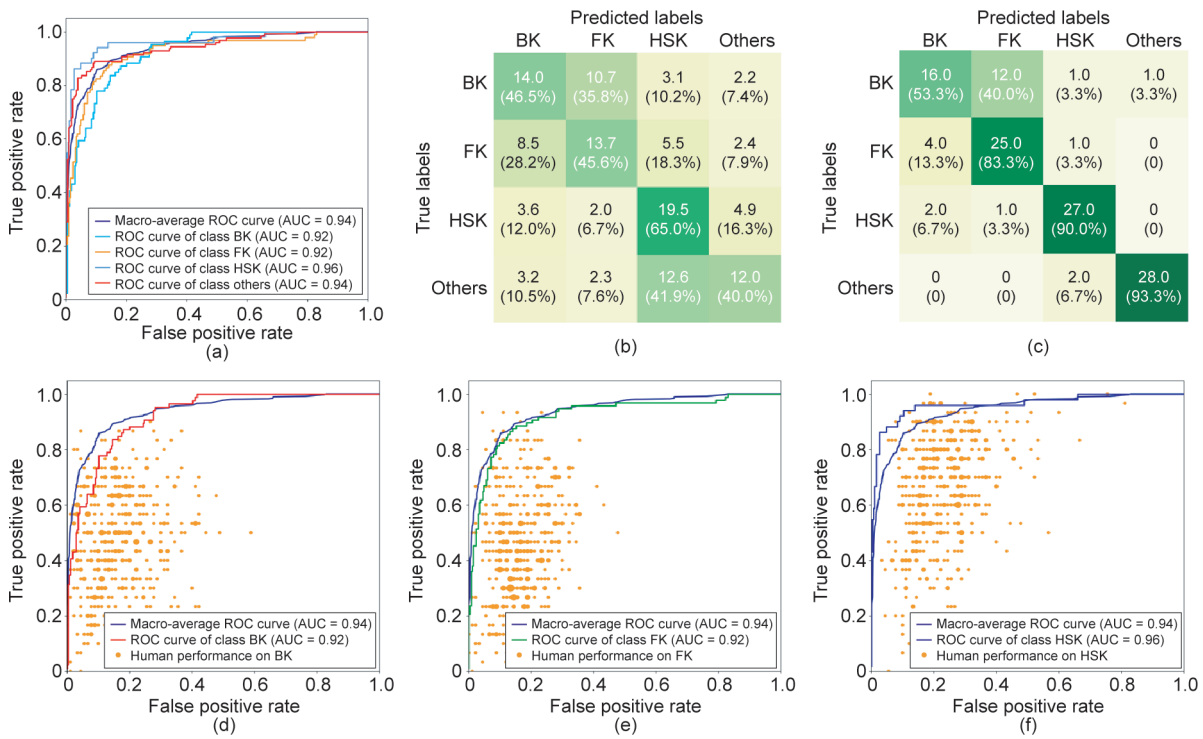


图4. SOS模型与眼科医生表现的ROC曲线及混淆矩阵对比图。（a）SOS模型的ROC曲线；（b）、（c）在用于诊断水平测试的数据集上眼科医生与SOS模型的混淆矩阵；（d）~（f）对应的BK、FK与HSK等各子类的ROC曲线。

系[46]。医生对疾病的诊断主要依靠观察和推理。在所有人类疾病中，角膜病具有最直接、最显著的视觉感知变化；因为健康角膜具有完全透明的独特特征，这与病理状况下角膜内外图像的变化形成鲜明对比。人类专家通过对图像的理解和分析来进行角膜病的诊断决策，这可能是AI辅助人类最合适的任务。

一般而言，深度学习是由大量带标注的数据驱动的[47,48]。然而，尚不清楚有多少临床图像训练数据足以被开发用于诊断临床疾病的AI系统。我们中心收集和记录具有临床图像的角膜病例已有20余年，但是，如果按照每种疾病类别对所有图像进行标注，那么在最常见的疾病类别中可以有数千张临床图像，而某些罕见疾病类别中只有少数临床图像。每个角膜病类别中带标注数据的不平衡性导致我们将注意力集中在最常见的疾病（如感染性角膜病）上，以初步开发本研究中的AI诊断系统。

在本研究中，我们证明了通过裂隙灯显微镜拍摄的临床图像，通过CNN进行的深度学习可以用于角膜感染性疾病的临床诊断。我们共评估了3组9种深度学习架构，以开发用于角膜感染性疾病的图像诊断系统。从图像级别和图像子块级别模型的结果可以看出，尽管只有4个类别，但这仍是一个困难的任务，尤其是对于VGG-16和GoogLeNet-v3。这两种结构在图像子块分类中均表现不佳，图像子块之间的投票并未显著改善其性能。相比之下，DenseNet的图像子块分类达到60%，投票后达到66.3%。研究表明，如果模型在图像子块分类中表现得足够好，那么从感染性病变区域着眼于图像子块可以产生比着眼于图片整体更高的性能。ROP方法可被视为除表决之外组合图像子块特征的另一种方法。结果表明，即使没有空间信息，采用适当的组合方法也可以进一步改善图像子块级模型。我们发现，总体而言SOS是用于角膜病中仅依赖图像进行诊断的最有前景的方法。SOS比其他方法更好的一个可能原因是在这种深度学习模型中，直接实现了对临床图像空间结构的适当利用。SOP表现不佳，是因为它没有考虑病变区域的环形结构。据我们所知，本研究首次提出了一种角膜病分类深度学习模型，该模型在仅依靠图像进行诊断时比人类眼科医生具有更高的准确性。在本研究中，一般眼科医生在仅依赖图像的角膜病诊断中的表现要比AI系统差。毫无疑问，错误的诊断可能造成长期使用不适当的药物，导致识别特征变得更为模糊，使医生诊断更加困难[6]。我们研究中的多元回归分析表明，就职称、工作机构和年资而言，这3个统计学因素对诊断性能会有影响，但3个模型中的确定系数较低。这表明上述因素可能无法如实全面地决定眼科医生对角膜病的诊断准确性，或者影响诊断性能的因素可能非常复杂，无法通过上述3个因素简单地准确总结。因此，如果AI可以帮助临床医生以更高的诊断准确性显著提高临床能力，那么这将极大地使角膜病患者受益，同时节省了医疗资源并减轻社会负担。目前全世界仍有450万人正遭受由角膜病引起角膜混浊导致的中度至重度视力障碍，特别是在发展中国家。提高诊断准确性可有两种途径，一是完善医师培训体系，加强医师专业教育和培训；另一种即开发实用的AI系统以协助诊断。我们的研究表明，通过使用临床图像来开发AI系统以提高角膜病诊断准确性是切实可行的。在测试眼科医生的表现时，我们发现当向医生提供图像和病史时，诊断准确性比单纯提供图像有所提高（从49.27%增至57.16%，$P < 0.001$）。该结果表明，其他信息可以帮助进一步提高诊断表现，AI诊断系统也可能如此。研究表明，将数据驱动的机器学习与人类知识相结合可以有效地开发出可解释、强大且通用的AI [49]；诸如病史之类的信息可能包含类似于人的常识，可以使模型利用有限的训练数据来解决许多不同的任务。为了改善我们的AI诊断系统以提高准确性，在未来的工作中可能需要设计一种多模态学习模型（即视觉和非视觉信息的有效组合），或是一个更合适的序列学习模型。

不可否认的是，在现阶段我们的AI诊断准确性只能由我们所收集的有限图像数据来证实[50]，并通过与使用相同临床图像的眼科医生进行对比。这种AI系统在辅助医生临床实践中的实际应用，需要在将来进行更大规模的深入临床评估[51]。

6. 结论

传染性角膜炎是最常见的眼科疾病，可能导致失明。眼科医生通过观察裂隙灯图像诊断疾病，利用计算机辅助图像分析算法方便诊断。在这项工作中，我们提出了一个序列水平的深部模型，端到端诊断传染性角膜炎。具体地说，利用深度卷积网络良好的特征提取性能，首先提取角膜区域的细节模式，然后将局部特征分成符合空间结构的有序集合，学习角膜图像的全局表示并进行诊断。我们收集了超过10 000名患者的110 000张图像。在此基础上，充分的实验对比结果表明，该模型是一种更为可行的结构，比传统的CNN具有更好的诊断性能。另外，通过与400多位专业眼科医生的诊断结果

进行对比发现，我们的模式可以大大超过专业人士的平均水平，达到顶级眼科医生的水平表现。据我们所知，这是第一个关于感染性角膜炎诊断的研究，我们的研究有力地证明了使用人工智能进行这些类型疾病的临床辅助诊断的潜力。

致谢

本研究得到了浙江省卫生健康委员会（WKJ-ZJ-1905、2018ZD007）、浙江省重点研究开发项目（2018C03082）和国家自然科学基金（61625107）的支持。感谢张仲非教授的意见和建议。

Compliance with ethics guidelines

Yesheng Xu, Ming Kong, Wenjia Xie, Runping Duan, Zhengqing Fang, Yuxiao Lin, Qiang Zhu, Siliang Tang, Fei Wu, and Yu-Feng Yao declare that they have no conflict of interest or financial conflicts to disclose.

References

[1] Bejnordi BE, Veta M, van Diest PJ, van Ginneken B, Karssemeijer N, Litjens G, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA 2017;318(22):2199–210.
[2] Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542(7639):115–8.
[3] Sommer A, Taylor HR, Ravilla TD, West S, Lietman TM, Keenan JD, et al. Challenges of ophthalmic care in the developing world. JAMA Ophthalmol 2014;132(5):640–4.
[4] Pascolini D, Mariotti SP. Global estimates of visual impairment: 2010. Br J Ophthalmol 2012;96(5):614–8.
[5] Clemens LE, Jaynes JM, Lim E, Kolar SS, Reins RY, Baidouri H, et al. Designed host defense peptides for the treatment of bacterial keratitis. Investig Ophthalmol Vis Sci 2017;58(14):6273–81.
[6] Gopinathan U, Garg P, Fernandes M, Sharma S, Athmanathan S, Rao GN. The epidemiological features and laboratory results of fungal keratitis: a 10-year review at a referral eye care center in South India. Cornea 2002;21(6):555–9.
[7] Yang Y, Luyten W, Liu L, Moens MF, Tang J, Li J. Forecasting potential diabetes complications. In: Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence; 2014 Jul 27–31; Quebec City, QC, Canada; 2014. p.313–9.
[8] He T, Guo J, Chen N, Xu X, Wang Z, Fu K, et al. MediMLP: using Grad-CAM to extract principal variables for lung cancer postoperative complication prediction. IEEE J Biomed Health Inf 2020;24(6):1762–71.
[9] Nee P, Li Y, Zhu J, Peng J, Dai Z, Li G, et al. Disease diagnosis prediction of EMR based on BiGRU–Att–CapsNetwork model. In: Proceedings of 2019 IEEE International Conference on Big Data (Big Data); 2019 Feb 27–Mar 2; Los Angeles, CA, USA; 2019. p. 6166–8.
[10] Wright AP, Wright AT, McCoy AB, Sittig DF. The use of sequential patternmining to predict next prescribed medications. J Biomed Inf 2015;53:73–80.
[11] Scott IM, Cootes TF, Taylor CJ. Improving appearance model matching using local image structure. In: Proceedings of Biennial International Conference on Information Processing in Medical Imaging; 2003 Jul 20–25; Ambleside, UK; 2003. p. 258–69.
[12] Manousakas IN, Undrill PE, Cameron GG, Redpath TW. Split-and-merge segmentation of magnetic resonance medical images: performance evaluation and extension to three dimensions. Comput Biomed Res 1998;31(6):393–412.
[13] Zhao Y, Gui W, Chen Z, Tang J, Li L. Medical images edge detection based on mathematical morphology. In: Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference; 2006 Jan 17–18; Shanghai, China; 2006. p. 6492–5.
[14] Kaus MR, von Berg J, Weese J, Niessen W, Pekar V. Automated segmentation of the left ventricle in cardiac MRI. Med Image Anal 2004;8(3):245–54.
[15] Cordes D, Haughton V, Carew JD, Arfanakis K, Maravilla K. Hierarchical clustering to measure connectivity in fMRI resting-state data. Magn Reson Imaging 2002;20(4):305–17.
[16] Pohl KM, Fisher J, Shenton M, McCarley RW, Grimson WEL, Kikinis R, et al. Logarithm odds maps for shape representation. In: Proceedings of International Conference on Medical Image Computing and Computerassisted Intervention; 2006 Oct 1–6; Copenhagen, Denmark; 2006. p. 955–63.
[17] Lee LK, Liew SC, Thong WJ. A review of image segmentation methodologies in medical image. In: Sulaiman HA, Othman MA, Othman MFI, Rahim YA, Pee NC， editors. Advanced computer and communication engineering technology. Cham: Springer; 2011. p. 1069–80.
[18] Shen D, Wu G, Suk HI. Deep learning in medical image analysis. Annu Rev Biomed Eng 2017;19:221–48.
[19] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42(9):60–88.
[20] Zhou X, Takayama R, Wang S, Hara T, Fujita H. Deep learning of the sectional appearances of 3D CT images for anatomical structure segmentation based on an FCN voting method. Med Phys 2017;44(10):5221–33.
[21] Dunnmon JA, Yi D, Langlotz CP, Ré C, Rubin DL, Lungren MP. Assessment of convolutional neural networks for automated classification of chest radiographs. Radiology 2019;290(2):537–44.
[22] Ardila D, Kiraly AP, Bharadwaj S, Choi B, Reicher JJ, Peng L, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat Med 2019;25(6):954–61.
[23] Wu N, Phang J, Park J, Shen Y, Huang Z, Zorin M, et al. Deep neural networks improve radiologists' performance in breast cancer screening. IEEE Trans Med Imaging 2020;39(4):1184–94.
[24] Chilamkurthy S, Ghosh R, Tanamala S, Biviji M, Campeau NG, Venugopal V, et al. Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. Lancet 2018;392(10162):2388–96.
[25] Frid-Adar M, Diamant I, Klang E, Amitai M, Goldberger J, Greenspan H. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. Neurocomputing 2018;321:321–31.
[26] Attia ZI, Kapa S, Lopez-Jimenez F, McKie PM, Ladewing DJ, Satam GP, et al. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. Nat Med 2019;25(1):70–4.
[27] Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat Med 2019;25(1):65–9.
[28] Bejnordi BE, Veta M, Van Diest PJ, van Ginneken B, Karssemeijer N, Litjen G. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA 2017;318(22):2199–210.
[29] Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016;316(22):2402–10.
[30] Ting DS, Cheung CY, Lim G, Tan GS, Quang ND, Gan A, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. JAMA 2017;318(22):2211–23.
[31] Kim SJ, Cho KJ, Oh S. Development of machine learning models for diagnosis of glaucoma. PLoS ONE 2017;12(5):e177726.
[32] Kermany DS, Goldbaum M, Cai W, Valentim CCS, Liang H, Baxter SL, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell 2018;172(5):1122–31.
[33] Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng 2009;22(10):1345–59.
[34] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014. arXiv:1409.1556.
[35] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016 Jun 26–Jul 1; Las Vegas, NV, USA; 2016. p. 2818–26.
[36] Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017 Jul 21–26; Honoululu, HI, USA; 2017. p.4700–8.
[37] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;9(8):1735–80.
[38] Cho K, van Merrienboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning phrase representations using RNN encoder–decoder for statistical machine translation. 2014. arXiv:1406.1078.
[39] Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. 2014. arXiv:1409.3215.
[40] Schmidhuber J, Wierstra D, Gagliolo M, Gomez F. Training recurrent networks by evolino. Neural Comput 2007;19(3):757–79.
[41] Greff K, Srivastava RK, Koutník J, Steunebrink BR, Schmidhuber J. LSTM: a search space odyssey. IEEE Trans Neural Netw Learn Sys 2016;28(10):2222–32.
[42] Wang D, Khosla A, Gargeya R, Irshad H, Beck AH. Deep learning for identifying metastatic breast cancer. 2016. arXiv:1606.05718.

[43] Lam C, Yu C, Huang L, Rubin D. Retinal lesion detection with deep learning using image patches. Invest Ophthalmol Vis Sci 2018;59(1):590–6.
[44] Ledley RS, Lusted LB. Reasoning foundations of medical diagnosis. Science 1959;130(3366):9–21.
[45] Claus-Christian C. Understanding human perception by human-made illusions. Front Hum Neurosci 2014;8:566.
[46] Pan Y. On visual knowledge. Front Inf Technol Electron Eng 2019;20 (8):1021–5.
[47] Lecun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436–44.
[48] Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science 2006;313(5786):504–7.
[49] Zhuang YT, Wu F, Chen C, Pan YH. Challenges and opportunities: from big data to knowledge in AI 2.0. Front Inf Technol Electron Eng 2017;18(1):3–14.
[50] Zhu Y, Gao T, Fan L, Huang S, Edmonds M, Liu H, et al. Dark, beyond deep: a paradigm shift to cognitive AI with humanlike common sense. Engineering 2020;6(3):310–45.
[51] Begoli E, Bhattacharya T, Kusnezov D. The need for uncertainty quantification in machine-assisted medical decision making. Nat Mach Intell 2019;1(1):20–3.