基于自然观看时眼球运动的ADHD检测

注意力缺陷/多动障碍（ADHD）是一种高度流行的神经发育障碍，需要临床专家进行诊断。众所周知，一个人的观看行为，反映在他们的眼球运动中，与注意力机制和高阶认知过程直接相关。因此，我们探索在自由观看任务中，是否可以根据记录的眼球运动以及视频刺激的信息来检测ADHD。为此，我们开发了一个基于端到端深度学习的序列模型，我们对相关任务进行了预训练，其中有更多数据可用。我们发现，该方法实际上能够检测ADHD，并且优于相关基线。我们调查了消融研究中输入特征的相关性。有趣的是，我们发现模型的性能与视频内容密切相关，这为未来的实验设计提供了见解。

注意力缺陷/多动障碍（ADHD）是儿童期最常见的神经发育障碍之一，根据所使用的诊断程序，该年龄组约有5%至13%的儿童受到影响[37，42，26]。ADHD的特点是持续注意力不集中、高度多动和冲动[2]。

ADHD的诊断需要专家进行临床评估，通常需要通过临床访谈和使用评分表进行自我报告和线人报告。告密者报告可从亲密的家庭成员处获得，教师或合作伙伴，具体取决于应聘者的年龄。由于临床评估受到主观报告和评级的严重影响，因此也会导致反映社会或认知偏见的风险。ADHD症状和正常行为的优势和劣势（SWAN）评分表[36]是一种成熟的筛查工具，基于父母或教师必须填写的问卷。SWAN量表记录了注意力不集中、多动和冲动的症状，得出所谓的SWAN评分。具体而言，SWAN评分量表根据症状严重程度的全谱来调查行为，从功能性到功能障碍[36,5]

缺乏全面、客观的评估工具、症状表现的发展变化[4]和高并发率[2]是ADHD评估的主要挑战，并最终增加了诊断不足或过度诊断的风险。虽然假阴性可能导致拒绝治疗，但假阳性可能导致不适当的治疗，这两种情况都可能对个人在学校、职业和社会方面的能力以及他们的整体福祉产生不利影响。这推动了全自动筛查工具的开发，该工具可广泛应用于有ADHD风险或怀疑患有ADHD的人群，从而增加了ADHD筛查机会的可及性，并提高了专家评估前筛查方法的客观性。

眼球运动可分为所谓的眼球运动事件。这些包括固定(≈ 200–300毫秒），在此期间眼睛相对静止并获得视觉信息，以及扫视，这是眼睛注视在任意两个注视之间的快速重新定位运动(≈ 30–80毫秒）[12]。一系列固定被称为扫描路径。众所周知，眼球运动反映了包括注意力机制在内的认知过程[17，10]，因此它们被视为心灵和大脑的窗口[40]。几十年来，它们一直被用作认知心理学的黄金标准衡量标准[27]。认知心理学领域的研究人员通常将眼球运动视为因变量，以研究实验操纵刺激的效果，从而将其建模为目标变量。相比之下，最近的研究表明，将眼球运动视为独立变量（即模型输入）来推断观看者的属性的潜力。例如，已经表明，眼动追踪数据可以用于区分不同的认知状态[11]、个人特征[13]或认知负荷[31]。使用眼球运动来推断观看者的主要挑战是眼球跟踪信号中的高度个体可变性。眼动追踪数据中个体特征的优势解释了为什么用于观众识别的机器学习方法表现得非常好[22，24]，而用于其他推理任务的模型通常表现得最好是概念验证水平或略高于机会水平。开发用于眼睛跟踪数据分析的机器学习方法的另一个主要挑战是数据稀缺。由于高质量眼动追踪数据的收集是资源密集型的，因此只有极少数大型数据集存在。

使用眼动追踪任务，研究发现患有和不患有ADHD的个体在观看行为上存在差异，在眼动追踪任务中，参与者被要求进行朝向或远离刺激物的自愿眼动（所谓的支持或反扫视任务）[25，18]。这些发现激发了我们开发筛选工具的方法，该工具处理每个人的眼球运动，同时考虑到视觉刺激的信息。

这篇论文的贡献是四倍。首先，我们提供了一种新的最先进的模型，以从自然自由观看任务中的眼球运动中检测ADHD，并在真实世界数据集上评估该模型和相关参考方法的性能。其次，我们通过计算特征重要性，对i）消融研究和ii）中不同输入特征的相关性进行了广泛调查。第三，我们证明转移学习有潜力克服眼动追踪研究中的数据稀缺问题。最后但并非最不重要的是，我们发布了一个用于检测ADHD的预处理的免费观看眼动追踪数据集。

本文的其余部分结构如下。第2节讨论了相关工作，第3节阐述了问题设置。我们在第4节中开发了ADHD检测的模型架构，并在第5节中介绍了数据集。在第6节中，我们介绍了实验结果，而在第7节中，讨论了结果。第8节结束。

**2相关工作**

机器学习方法已应用于不同类型的诊断数据的ADHD检测；e、 g.康纳斯成人ADHD评分量表[6]的数据、EEG信号[38]和静息状态下记录的功能磁共振成像（fMRI）数据[8]。廉价眼动追踪硬件的快速发展为无创、快速甚至隐式筛查提供了新的可能性，无需依赖自我、家长或老师的报告。在下一节中，我们简要回顾了与使用机器学习方法识别ADHD患者相关的工作，特别关注眼动数据。

ADHD检测是基于在不同类型的任务中收集的眼球运动进行的，例如阅读[7]、阅读广度任务[14]或连续性能测试[21]。这些任务对参与者提出了某些要求，以确保测量的有效性；e、 例如，参与者需要已经获得一定水平的阅读技能，或者必须理解并遵守复杂的任务说明。此外，已经表明，在指示的条件下，与自然观看相比，显示的内容类型（例如，情感内容）对眼睛运动的影响较小[19]。为了减少这种局限性，人们首次尝试在无任务观看中根据他们的眼球运动来检测ADHD。与以前的方法不同，这种方法有可能已经应用于很小的儿童，从而使他们从很小的时候就可以获得治疗。早期识别和治疗对于缓解ADHD的发展及其对个人功能和整体健康的长期负面影响至关重要[15，28]。

Galgani等人[9]提出了三种通过图像观看任务检测ADHD的方法，他们对年龄范围相对较宽（9-59岁）的参与者进行了评估。在这些方法中，表现最好的方法是基于Levenshtein距离。该方法使用基于感兴趣区域（ROI）的字母编码，通过将符号分配给不同ROI，将固定序列转换为符号序列。为了对一个新的实例进行分类，他们计算了相应符号序列与ADHD组和对照组实例的Levenshtein距离。与组的平均距离越小，表示与该组的相似性越大，因此相应的组标签被分配给实例。虽然这种方法考虑了注视序列的空间信息，但它没有考虑注视的时间信息；i、 即固定持续时间。

Tseng等人[39]提出了一种三类分类器，根据观看15分钟视频片段时记录的眼球运动，来区分ADHD儿童、胎儿酒精谱障碍儿童和对照儿童，而不是仅使用二元分类器检测ADHD。他们将注视特征与用显著性模型计算的刺激的视觉显著性信息相结合。然而，它们依赖于工程特征，该工程特征以眼睛注视信号中的顺序信息未被使用为代价，在时间上聚合眼睛注视事件（例如，中等扫视持续时间或扫视峰值速度）。

更多的研究集中在使用机器学习来检测其他神经发育障碍[16，41]。例如，Jiang等人[16]提出从观看图像时收集的眼动数据中检测自闭症谱系障碍（ASD），其中他们使用神经网络明确地模拟了两组之间眼球运动模式的差异。该方法的主要局限性是，对于每个图像，只分析固定数量的注视点，这可能会导致信息丢失。

3问题设置

我们研究ADHD检测问题。观看视频时，第j个人的眼睛注视被记录为一系列注视，表示为Pj＝{（x1，y1，t1），…，（xM，yM，tM）}，其中xM，yM是第m个注视位置，tM是注视持续时间，m是记录的注视总数。提供固定的视频帧速率，我们可以使用时间信息将注视映射到对应的视频帧V，从而语义信息可以与眼睛注视相关联。训练集由D={（P1，V，c1），…，（PJ，V，cJ）}组成，其中PJ和V表示第j个个体的对齐固定序列和视频帧，cJ是个体是否患有ADHD的标签。目的是训练一个识别ADHD个体的分类器，这是一个二元分类问题。

通过改变学习模型的决策阈值，我们可以绘制真阳性率与假阳性率的接收器操作特性（ROC）曲线，并最终计算曲线下面积（AUC），即ROC曲线下面积，用作分类性能的定量指标。我们使用AUC作为评估度量，它对类的不均匀分布不敏感。

4方法

在本节中，我们介绍了我们的模型和用于初始化ADHD分类最终任务权重的预训练任务。

4.1型号

我们提出了一种端到端训练的神经序列模型，将注视序列分类为属于患有或不患有ADHD的个体。图1显示了我们提出的方法的概述。我们使用离散阈值识别算法[29]对原始眼睛跟踪进行预处理，该跟踪包括以60或120 Hz的采样率记录到注视序列中的水平和垂直屏幕坐标。该模型将眼睛注视序列（扫描路径）和生成该扫描路径的视频剪辑作为输入。

根据我们对文献的回顾，我们假设ADHD患者的眼睛注视与视觉刺激的交互作用不同于通常发育中的对照。因此，我们使用显著性图来突出显示场景中可能的感兴趣区域。我们使用最先进的显著性模型DeepGaze II[20]来计算视频刺激的显著性图。DeepGaze II使用在对象识别任务中训练的VGG-19特征[34]，并将其输入第二个网络，该网络被训练以预测给定图像上的注视位置的概率分布。

对于大小（W，H）的每个视频帧i，预训练的DeepGaze II模型生成显著性图S（i）∈ 然后，我们应用最小-最大归一化将S（i）变换到[0，1]的范围。为了提取每个固定位置的归一化显著性值，我们创建了一个提取掩码E（i）m∈ R H×W，用于第i个视频帧上的第m个固定。更具体地，E（i）m通过将固定位置设置为一个而将所有其他单元设置为零来生成。然后，我们用高斯核（标准偏差σ=1.5◦) 并将其标准化。高斯核被用于解释注视中心周围的副中心信息摄入[12]。最后，第m个固定的显著性值由下式给出：

其中是Hadamard乘积，1d是维度d的全1行向量。如果固定跨越多个帧，我们使用中心帧进行显著性计算。提取的显著性值序列与注视位置（以视角角度表示）和注视持续时间（见图1）相连接。最后，我们将z分数归一化应用于这些特征通道中的每一个。