# OPEN: 虚拟康复学习环境中老年患者参与识别的 基准数据集和基线

Ali Abedi, Sadaf Safa, Tracey J.F. Colella<sup>†</sup>, Shehroz S. Khan<sup>†</sup>

Abstract—虚拟学习中的参与对参与者的满意度、绩效和坚持性至关重要,特别是在在线教育和虚拟康复等领域,其中交互式交流对于 成功至关重要。然而,在虚拟团体环境中准确测量参与度仍然是一个重大的挑战。越来越多的人对利用人工智能进行大规模、真实世界 和自动化的参与度识别产生了兴趣。尽管在学术背景中,关于年轻人群体的参与度已有广泛研究,但针对虚拟和远程健康学习环境中的 老年人的研究方法和数据集仍然有限。现有的方法常常忽视学习材料的情境相关性以及跨会话的参与度的纵向动态。本文介绍了 OPEN (Older adult Patient ENgagement),一个旨在支持开发用于识别参与度的人工智能模型的新颖数据集。该数据集来自参与虚拟小组学习 会话的十一名老年人,这些会话作为其心脏康复计划的一部分,持续了六周,产生了超过 35 小时的数据,代表了同类最大的此类数据 集。为了保护隐私,原始视频被保留,但公开的数据包括从视频中提取的面部、手部和身体关节位置标记,以及行为和情感特征。观察 注释包括第二级的二元参与状态、情感和行为标签以及上下文类型的细节,例如教师是针对小组还是个人发表讲话。使用长度为 5、10、 30 秒的样本,以及可变长度的片段生成了多个版本的数据集。为了证明其实用性,各种机器学习和深度学习模型在这些注释数据上进行 了训练,达到了高达 81 % 的参与识别准确率。OPEN 为推动老年群体个性化的人工智能驱动的参与识别提供了一个可扩展的基础,同 时也为更广泛的参与识别研究作出了贡献。

Index Terms—Virtual Learning, Patient Engagement, Older Adult Engagement, Virtual Rehabilitation, Engagement Recognition.

## 1 引言

**R** 康复的目标是通过锻炼、教育和咨询来改善康复过程、 减少残障并优化健康结果 [1]。然而,传统的面对面康复 面临多个出勤障碍,包括交通困难、日程冲突、经济限制和医 护人员短缺,这些都导致了高退学率 [2]。虚拟康复是远程医 疗的一个子集,指通过数字技术进行的远程康复服务。它提 供居家教育和锻炼课程,作为面对面项目的替代方案 [3],[4], [5],[6]。教育部分通常涉及由临床医生主持的团体课程,患 者在课程中获得关于慢性病管理、健康行为养成以及可持续 护理自我管理技能的指导。研究表明,虚拟康复项目可以实 现与面对面项目相当的成效,并帮助克服与访问相关的障碍。 此外,人工智能(AI)越来越多地整合到这些平台中,以加 强患者评估、监测活动,并支持健康结果的预测 [7]。

康复中的患者参与是一个随着时间发展的动态过程 [8]。 它被定义为"在康复过程中,基于并支持于患者与临床医生 之间的互动和关系的增强动机、注意力和积极参与"[9]。参 与度对于康复项目的成功至关重要,因为较高的参与度与更 好的依从性和更低的辍学率相关 [8]。在整个项目中持续监测 参与度并实施及时的干预来支持它,可以显著增强参与并带 来更好的健康结果 [8]。

在对虚拟远程医疗会议中患者参与的概念化和定义的综合 文献综述中,Liu等人[10]确定了参与的三个核心组成部分: 情感、行为和认知。这个框架与 Fredricks等人[11]在教育心 理学中的早期研究一致,他们将参与描述为学生学习背景中 参与和互动的替代指标。情感参与指的是情绪反应,如兴奋 和兴趣;行为参与涉及可观察的行为,如出勤、参与和保持 任务专注;认知参与反映了个人在学习中的投入以及愿意努 力和接受挑战。在此基础上,Sinatra等人[12]引入了粒度概 念,定义了参与被概念化和测量的水平。粒度范围从宏观层 面(例如,群体参与)到微观层面(例如,具体任务中的瞬时 个人参与)。微观层面的参与可以通过生理信号如眨眼频率、 头部姿态和心率进行评估,提供关于一个人即时参与的详尽 洞察[12]。参与也可以沿着分析水平的连续体进行观察:情 境导向、人在情境中、和人导向。宏观层面的参与对应于情 境导向分析,微观层面对应于人导向分析,而人在情境中则 介于两者之间。人在情境中的参与捕捉了个体如何与特定环 境情境进行互动,例如阅读网页或参与在线课堂。为了补充 这些概念化,Salam 等人[13] 审视了跨领域的情境感知参与 推断,识别出情境感知计算建模和参与的时间动态为关键研 究领域。

近年来,越来越多的人对使用人工智能(AI)和情感计算 在自然环境中大规模识别参与度产生兴趣[13],[14],[15],[16], [17]。目前的方法通常依赖于监督或半监督的机器学习和深 度学习方法,这些方法需要经过标注的真实数据集来进行模 型开发。然而,大多数现有的方法和数据集都集中在年轻、健 康的学生群体的参与度识别上[13],[15],[16],[17]。相比之下, 为年长患者量身定制的数据集和人工智能模型的创建和开发 进展有限[18]。

现有的参与性数据集以及利用它们开发的识别算法结合了 各种数据模态,包括视频、头部姿态、眼睛注视、面部表情、 语音、心率和心电图信号 [13], [14], [15], [16], [17]。参与的情 感、行为和认知成分在这些模态中因年龄 [19]、性别 [20]、 种族以及身体 [21] 或心理健康状况 [22] 等因素的不同而有不 同的表现。因此,在年轻、健康的学生群体数据上训练的参 与度识别算法可能不能很好地泛化到老年人或患者群体。

为了支持开发能够推广到虚拟学习环境中的老年人和患者的参与度识别算法,本文的主要贡献是引入了 OPEN (Older adult Patient ENgagement),这是一种具有以下独特特征的新公共数据集:

- 面向老年患者群体,并构成了最大型的用于人工智能驱动的互动识别的公开数据集。
- 包括由训练有素的观察者记录的注释参与状态以及情感 和行为组成部分,并提供有关学习过程的背景信息。
- 提供了多个版本的数据集,包括固定长度(例如,10秒 和 30秒)和可变长度样本,结合会话内的序列交互数据

<sup>•</sup> Ali Abedi, Sadaf Safa, Tracey J.F. Colella, and Shehroz S. Khan were with the KITE Research Institute, Toronto Rehabilitation Institute, University Health Network, Toronto, Canada.

<sup>†</sup>Shehroz S. Khan and Tracey J.F. Colella are senior authors of this work. E-mail: ali.abedj@nhn.ca. Manuscript received July 23, 2025; revised August 31, 2025.

和跨会话的纵向数据,以支持交互检测和预测。

 通过公开发布不可识别的数据,包括面部、手部和身体 关节点,以及衍生的行为和情感特征,以维护参与者的 隐私。

作为次要贡献,开发了一系列先进的机器学习和深度学习 模型,以在多个实验环境下从技术上验证 OPEN 数据集,实 现高水平的参与度识别性能,并确认了该数据集在 AI 模型开 发中的效力与适用性。此外,首次在参与度识别文献中 [23], [24], [25], [26],对两种不同的参与度推断任务,即参与度检 测和参与度预测,进行了系统研究。

本论文的结构如下。第2节回顾了现有的互动识别数据集 和方法,强调了所提出的数据集与之前工作的区别。第??节 详细介绍了数据集的收集和标注过程,以及其关键特征。第 ??节概述了互动识别的方法,描述了实验设置,并展示了使 用所提出的数据集和模型获得的结果。最后,第3节总结了 论文并讨论了未来研究的方向。

## 2 相关工作

本节回顾了用于虚拟学习环境中互动识别的现有数据集,并 以之前工作中定义的互动标注维度为指导 [16]。它还简要概 述了当前的互动识别方法。鉴于专门针对虚拟学习环境中老 年人和患者的研究有限,这一回顾包括最初为虚拟学习环境 中的学生互动开发的数据集和方法。

## 2.1 参与注释的维度

D' Mello [27] 确定了情感注解的五个维度,这后来成为 Khan 等人 [16] 通过引入两个附加维度来扩展该框架的基础。这一 扩展产生了一个由七个参与注解维度组成的综合集,强调情 感作为参与的三个核心组成部分之一的中心作用。<sup>1</sup>。这七个 维度如下 [16], [27]:

- 1) 来源:确定执行注释的人,如人类观察者或自我报告。
- 数据模态:指定用于注释的输入数据类型,例如视频录 制或屏幕捕获。
- 3) 时序:指出何时进行标注,例如,在会话期间实时进行 或事后进行。
- 4)时间分辨率(时间尺度):定义标注的频率,例如每秒、 每 10 秒或每次会话一次。
- 5)抽象水平:区分参与度是作为一个整体单一衡量标准标 注,还是作为多个组成部分(例如,情感、行为、认知) 标注。
- 6) 组合: 描述如何集成参与度组件以得出统一的参与度标签。
- 7)量化:指定如何以数值形式表示参与度,例如二进制(参与与不参与)或有序等级。

在一项关键评审中, Khan 等人检查了 31 个现有的参与度 数据集,这些数据集全部来自虚拟学习环境中的学生。通过 七个参与度注释维度分析这些数据集,揭示了在如何定义和 注释参与度方面的显著不一致性。这些不一致性对研究人员 构建既适用于又可跨数据集比较的通用性 AI 模型用于参与 度识别构成了重大挑战。值得注意的是,只有少数数据集依 赖于预先建立的心理测量学验证工具来定义和注释参与度。

Khan 等人 [16] 发现,最常见的参与注释来源是观察者 (在 31 个数据集中的 21 个) [26], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41], [42], [43], [44], [45], [46], [47],自我报告(在 31 个数据集中的 7 个) [48], [49], [50], [51], [52], [53], [54],以及观察者和自我报告相结合(在 31 个数据集中的 3 个) [55], [56], [57]。在所有涉及观察者进

1. 请注意,1节中详细描述的参与的三个组成部分与2.1小节中描述的参与注释的七个维度之间的区别。

行注释的数据集中,视频作为注释员事后观察的主要数据模态(时间安排)。时间分辨率(时间尺度)不一致,范围从1 秒到30分钟不等。五个数据集[30],[35],[36],[41],[50],[51] 使用了自适应时间尺度,其中注释不是在预定义的时间间隔 进行的。相反,只有在参与状态发生变化时才生成新的时间 戳,从而导致可变长度的数据样本。

关于抽象层次,十四个数据集将参与度定义和注释为表示参与水平 [26],[29],[31],[33],[34],[37],[42],[44],[46],[49], [52],[53],[54],[56]的简单变量。相反,其他数据集则专注于 单一 [30],[32],[35],[36],[38],[43],[45],[48],[50],[51],[57]或 参与度的多组成要素 [28],[30],[39],[40],[41],[47],其中五 个具体针对情感和行为组件 [30],[39],[40],[41],[47]。组合维 度仅适用于那些涉及参与度多个组成部分的情况。虽然有四 个数据集没有将参与度的各个组成部分结合在一起 [30],[39], [41],[47],但有两个数据集 [28],[40]采用了一套规则来组合 这些组件并得出总体参与水平。在量化方面,参与水平被定 义为顺序的 [26],[28],[32],[33],[37],[38],[39],[42],[48],[51], [54],[56],[57]、区间的 [34],[45],[52],[53],[54],以及分类或 二分的 [29],[30],[31],[35],[36],[40],[41],[43],[46],[47],[49], [55] 变量。

除了 [49] (包括年龄在 20 到 60 岁之间,平均年龄为 34 岁的健康学生)之外,所有其他在 [16] 中审查的三十个虚拟 学习数据集均来自年轻健康的学生。虚拟学习项目由离线预 录制讲座、使用互动软件、阅读和写作活动组成,只有一个 在线讲座有讲师在场 [54]。这些数据集中,有十八个是在实 地收集的,十三个是在实验室环境中收集的。所有数据集均 在单次、一时间的录制过程中从学生那里收集。数据集中的 参与者人数平均为 47.39,标准差为 40.40,范围从至少 6 人 到最多 137 人。

只有数量有限的已审查数据集可以公开获取,所有数据集 均来自印度大学就读的二十多岁的年轻人。数据是在单次会 话中收集的,参与者在各自的电脑上观看预先录制的讲座。这 些数据集包括电子环境中的情感状态数据集(DAiSEE),它包 含来自 112 名学生的 9,068 个十秒样本,总计 25.2 小时的视 频数据。"野外的情感识别"之"野外参与度预测"(EmotiW-EW)数据集包含 195 个视频样本,每个样本长 2 到 3 分钟, 收集自 78 名学生,总计 16.5 小时的视频。此外, EngageNet 包括来自 127 名学生的 11,311 个十秒样本,总计 31.4 小时的 视频数据。

#### 2.1.1 老年人参与数据集

Noceti 等人 [18] 收集了首个以老年人为对象的参与数据集。 该研究涉及 12 名年龄在 77 至 93 岁之间的健康女性参与者, 每位参与者在意大利的一家养老院进行了一次 20 分钟的虚拟 会话。该数据集包含总计 4 小时的数据, 这些数据来自虚拟 会话,在心理学家的指导下,三名老年人一组参与创造性活 动。该数据集整合了多种模态,包括视频、音频、加速度计数 据、皮电活动和心率。参与度水平被标注为从1到5的整数, 代表从最低到最高的参与度。视频数据中提取的特征包括平 均运动、头部姿势、面部动作单元(FAUs)和面部情绪,而 音频特征则包括强度、音高和梅尔频率倒谱系数。提取的特 征与生理数据结合后,进行连接,作为完全连接的前馈神经 网络的输入,用于两种时间窗口的参与度水平回归,即15秒 和2分钟窗口。出现了两个显著发现:(1)15秒窗口的效果 更好;(2)通常使用视频中的面部和身体特征可以获得最佳 结果,添加音频数据时会有轻微改善。然而,生理数据与视 频和音频特征的结合会对性能产生负面影响。该数据集尚未 公开可用。

#### 2.2 参与度识别

Dewan 等人 [15] 将参与度识别技术根据参与者的参与程度分 为三类:手动、半自动和自动。手动方法通过自我报告问卷 来评估参与度,这些问卷测量注意力、分心、兴奋和无聊等 因素 [58]。半自动方法从参与者在练习题和测试题上的表现 推断参与度 [59]。相比之下,自动方法依赖于计算机视觉和 机器学习技术,不需要参与者的主动输入。这些自动方法由 于其非侵入性、广泛适用性和在虚拟学习环境中特别有效而 受到重视。

Karimah 和 Hasegawa [23] 对虚拟学习环境中的参与度识 别方法和数据集进行了系统综述,重点强调了对参与度的情 感和行为成分的关注。这种强调源于大多数方法仅依赖视频 数据,而没有整合音频或背景信息,这限制了它们估计认知 参与度的能力。其中大多数方法使用由外部观察者 [16] 标注 的视频数据集来训练端到端、基于特征或基于标志的模型。

端到端技术使用深度神经网络直接处理原始视频帧,通 过如 3D 卷积神经网络 (CNNs)、视频 Transformers 和 2D CNNs 与长短期记忆网络 (LSTMs) 或时间卷积网络 (TCNs) 的组合等架构来学习相关特征。基于特征的方法涉及从如视 频等模态中提取参与度的情感、行为或认知指标,可以使用 领域特定的知识或预训练的模型。OpenFace 广泛用于提取面 部动作单元 (FAUs)、视线方向和头部姿态, 而来自如情感 面部对齐网络 (EmoFAN) 和用于面部视频表示学习的掩码 自动编码器(MARLIN)等模型的面部嵌入也很常见。这些 特征通常使用如词袋模型(BoW)、递归神经网络(RNNs)、 TCNs、Transformers 和集成模型等模型进行分析。基于标 志物的方法介于端到端和基于特征的策略之间,使用通过如 OpenFace 或 MediaPipe 等工具提取的面部、身体或手部标 志物的时空图。这些图通过时空图卷积网络(ST-GCNs)进 行处理,捕捉参与度中的动态模式。ST-GCNs 最初用于面部 情感分析,已经被有效应用于参与度估计;例如,Abedi和 Khan 使用 MediaPipe 从 EngageNet 数据集中提取面部标志, 并应用 ST-GCNs, 获得了比以前的方法更高的分类准确性, 且参数显著更少。

## 2.3 讨论

当前可用的公共数据集主要关注年轻学生群体,常常忽视老年人和患者群体。为了弥补这一空白,我们推出了首个从真 实家庭环境中的老年患者收集的公开虚拟学习参与度数据集。 该数据集使用验证过的参与度标注协议进行了标注,创建于 不同的环境中,例如,具有不同的时间分辨率,并包含上下 文信息。该数据集经过严格的评估,以确保其通过机器学习 和深度学习技术适合于 AI 模型的开发。此外,为了保持匿名 化并遵循研究发现强调基于标志点的算法(如 ST-GCN)优 于传统视频端到端模型,该数据集包含面部、手部和身体标 志点,以及从这些标志点派生的特征。

在经历心脏事件或手术后的急性护理出院后,符合条件的 患者通常会被推荐至门诊心脏康复计划。这些计划包括跨学 科、综合性风险减少干预,旨在提高心血管疾病患者的身体、 心理和社交功能。完成心脏康复已被证明能显著降低发病率、 心脏特异性和所有原因的死亡率。根据患者的偏好,门诊心 脏康复可以通过虚拟、混合形式或面对面来进行。然而,在 COVID-19 大流行的最后几个月期间,虚拟交付是唯一可用 的选项。

在虚拟心脏康复中,教育课程是一个关键部分,在此期间 收集了 OPEN。这些课程通常包括患者群体以虚拟方式参与, 由临床医生提供关于自我管理和采用健康心脏生活方式的指 导。常见的话题包括理解个人诊断、药物依从性、均衡营养、 体育活动、避免如吸烟等有害行为、减少久坐行为、压力管 理以及将常规锻炼纳入日常生活的策略。其主要目的是让患 者具备必要的知识和技能,以建立支持长期健康和福祉的可 持续习惯 [60] 。

虚拟心脏康复的教育课程共持续六周,每周安排在固定工作日的一小时课程,由心脏康复监督员或临床医生主持。四 到八名参与者组成的小组使用 Microsoft Teams 从他们的家 中虚拟参加这些课程。在课程期间,参与者保持坐姿,并在个 人电脑或笔记本电脑屏幕上观看内容(见图 1);然而,他们 并不总是持续专注,因为可能会出现注意力分散的时刻。课 程的主要结构是由临床医生提供信息,通常使用教育幻灯片, 并对整个小组进行讲解。临床医生还通过向小组或个别参与 者提问来促进互动。

参与者观看了临床医生的直播视频或他们共享屏幕上显示 的教育幻灯片,这些都通过 Microsoft Teams 的聚光灯功能进 行了突出显示。此功能将选定的视频或内容固定在所有与会 者的主屏幕上,确保临床医生或他们的材料在整个会议期间 始终显著可见。此外,参与者可以通过设备自带的 RGB 摄像 头或外接网络摄像头看到自己和其他与会者的实时视频。音 频通信通过设备的内置扬声器和麦克风进行,允许参与者听 到临床医生和其他与会者的声音,并在适当时发言。在会议 期间,没有参与者使用耳机或耳麦。



Fig. 1. 通过 Microsoft Teams 进行的虚拟心脏康复教育课程的截图。该 课程由一位心脏康复临床医生主持(上排从左数第二位),并有四位患者 参与。为了保护参与者的隐私,面部已被模糊处理。

#### 2.4 参与者

参与者为患有冠状动脉疾病、心脏事件或心脏手术的个体,主 要为60岁及以上的老年人;不过,也有少数参与者年龄低于 60岁。资格要求是必须注册由多伦多康复研究所心血管预防 与康复计划通过家庭进行的虚拟心脏康复计划。参与者还需 要能够访问带有互联网连接的个人电脑或平板电脑。排除标 准包括手术后并发症(例如中风)、可能妨碍参与的精神或认 知障碍(例如难以理解或遵循指令),以及英语流利度不足。

#### 2.5 伦理审查与同意

数据集收集、注释、预测建模和公共数据集发布的研究方 案已获得大学健康网络研究伦理委员会的批准(研究编号: 21-5420)。所有数据被纳入数据集的参与者均获得知情同意。

#### 2.6 数据收集

数据收集过程没有改变虚拟心脏康复教育课程的任何方面,除了记录参与者的视频外,课程遵循护理标准。在每周课程 开始时,临床医生会提醒参与者保持摄像头打开,并告知他 们课程将通过 Microsoft Teams 进行录制。

录制的视频分辨率为 1280×720 像素, 帧率为 16 帧每秒 (fps)。Microsoft Teams 不支持单独录制每位参与者的视频

流,因此每个会话都被捕获为一个综合的视频文件。这些会 话中的一些参与者不愿意参与研究。虽然他们的视频数据在 最初被录制了,但随后从数据集中被排除。

## 2.7 数据标注

数据标注包括三个任务:参与者片段的时空标注、主要参与 度标注以及上下文类型的时间标注,详细如下。

如子节 2.6 详细说明的那样,使用 Microsoft Teams 的视频录制功能,将一个会议中的所有参与者的视频录制为单个 视频文件,显示每个参与者的独立方块。两位注释者检查了 录制的会议视频,以便在参与者的位置发生变化时注释其方 块的坐标(以像素为单位)。值得注意的是,方块的坐标在会 议期间可能会发生变化,例如当临床医生开始分享其屏幕或 某位参与者暂时关闭其摄像头时。第一位注释者进行了初步 注释,第二位注释者对其进行了复查和完善。每个会议的参 与者方块的时间和坐标用于生成每个参与者对应的单独视频 文件,该文件仅包含其各自的视频内容。

#### 2.7.1 参与注释

人工专家标注过程 (HELP) [61] 是一个针对学习环境设计的 参与度标注协议,被用于标注 OPEN 中的参与度。选择该协 议是因为它与教育心理学中对参与度的定义一致 [16], [62]。 HELP 被分为三个阶段:预标注、标注和后标注。

在预注释阶段,对三名注释者进行了培训。为他们提供了 一部分数据集,其中包括一个虚拟心脏康复课程,用于练习 注释。他们的工作得到了审查,给予了反馈,解决了模糊之 处,并解答了他们的问题以确保清晰理解。

在标注阶段,被称为来源的三名标注者(见子节 2.1)通 过回顾观看录制的视频,独立地对数据集进行了标注,表示 参与度标注的时间和数据模式维度。视频是唯一的数据模式, 没有伴随音频或上下文信息。

大多数以前的数据集在标注之前将参与者的录制视频分割 成固定长度的间隔,例如,DAiSEE使用10秒段,EngageNet 也是如此。由于订阅注释是在分割之后进行的,片段中可以包 含多种参与状态;例如,参与者在片段的开头可能是参与状态,但到片段结束时变为不参与状态。这样的重叠可能给标 注人员带来挑战,并可能导致基于此类数据训练的AI模型产 生混乱。为了防止这种混淆,帮助(HELP)和开放(OPEN) 采用了适应性时间分辨率进行注释,其中数据分割由注释指 导。标注人员观看视频并标记参与者状态变化的精确时刻(精 确到秒)。这种方法确保每个数据片段代表一个单一、一致的 状态。

关于抽象层次,参与度被标注为一个多成分变量。由于视频是标注者唯一可用的数据模式,他们可以标注参与的情感和行为成分,但不能标注认知参与 [14]。借鉴 Woolf 等人的灵感 [63],HELP 识别了最优的参与成分标注类别。这些类别包括四个情感(情绪)成分:无聊、平静/满足、困惑/沮丧和有动力/兴奋,以及两个行为成分:脱离任务和参与任务。这些类别被采用用于标注 OPEN。

根据 Woolf 等人提出的框架 [63], HELP 建立了将参与度 量化为二分变量的规则:参与或未参与,这基于情感和行为 成分的组合。如果行为成分是处于离线任务状态,则状态被 量化为未参与。当行为成分处于在线任务状态时,如果情感 成分是无聊的,状态也被量化为未参与。在所有其他组合中, 状态被量化为参与。

#### 2.7.2 上下文类型的时间标注

两位注释者审阅了录制的会话视频,以识别各种情境类型的 开始和结束时间。第一位注释者进行了初步的标注,第二位 注释者对其进行了审查和修改。这些情境类型包括:(i)临床 医生对所有参与者讲话,(ii)临床医生与特定参与者交谈(标 识参与者),(iii)参与者与临床医生交谈,(iv)临床医生展示 幻灯片或视频,(v)参与者对其他所有参与者讲话,(vi)参与 者与另一参与者交谈,(vii)不活动或无互动的时期。

## 2.8 数据集特征

#### 2.8.1 参与者特征

表 1 列出了 OPEN 项目中 11 名参与者的人口统计学特征。 参与者的平均年龄为 66.5 岁(标准差为 9.6)。其中, 36.4 % (n = 4) 为女性, 90.9 %(n = 10)被认定为白人, 9.1 %(n = 1) 被认定为南亚裔。参与者经历了一系列多样的心脏事件 和手术。



Fig. 2. 数据样本在样本持续时间(以秒为单位)的不同范围内的频率固 定为(a)情感,(b)行为,以及(c)同时具有情感和行为状态。例如, 第一个数据点显示持续时间范围为1到5秒。最后一个数据点显示持续 时间范围为长于2分钟。

表1的最后一列展示了每位参与者在其虚拟程序中参加的 虚拟课程次数,以及他们在数据集中数据的总时长。该数据 集由 36 个参与者会话组成,总共记录了 35 小时 2 分钟的数 据。

#	Age	Sex	Ethnicity	Cardiac Disease / Procedure	# of Sessions (dura-
					tion)
1	62	Male	Caucasian	Aortic Valve Replacement	4 (03:56)
2	60	Male	Caucasian	Coronary Artery Bypass Graft	6 (05:26)
3	80	Female	Caucasian	Angioplasty and Stent Placement	2(02:14)
4	52	Female	South Asian	Chemotherapy-induced Cardiotoxicity	1(00:15)
5	72	Female	Caucasian	Atrial Fibrillation	1 (01:08)
6	76	Male	Caucasian	Transient Ischemic Attack	5(04:31)
7	74	Male	Caucasian	Stent Placement	5(04:33)
8	77	Female	Caucasian	Aortic Valve Replacement	3 (03:02)
9	65	Male	Caucasian	Stent Placement	2 (02:22)
10	64	Female	Caucasian	Stress Induced Cardiomyopathy	4 (04:05)
11	47	Male	Caucasian	Percutaneous Coronary Intervention	3 (03:27)

TABLE 1 参与者的人口统计数据、心脏病/手术、数据集中虚拟会话的数量(以及数据量)。

正如在子章节 2.7.1 中所描述的,数据样本是根据参与度 的情感或行为组成部分的变化自适应生成的。这种方法产生 的带注释的数据样本长度不定,每个样本有固定的情感和行 '为参与组成部分。图 2 (a)、(b) 和 (c) 分别显示了数据集中不 同持续时间范围内的数据样本比例,每个面板分别关注情感 状态、行为状态以及情感和行为状态的结合。正如图 2 所示, 对于情感、行为和结合的参与组成部分来说,数据样本最常 见的持续时间范围是1到5秒。这表明在大多数情况下,参 与者的情感和行为状态每1到5秒就会发生变化。其次最为 频繁的范围是 6 到 10 秒,其次是更长的持续时间。图 1 (a) 和 (b) 的比较表明行为状态比情感状态变化更快。结合情感 和行为状态的数据样本总数,如图 2 (c) 所示为 4,494。这些 样本的持续时间具有以下统计数据(以秒计):最小值为1, 最大值为1,133,平均值为26.29,标准差为52.28,中位数为  $9^{\circ}$ 

#### 2.8.2 情感和行为状态的动态

利用基于情感和行为状态变化的自适应数据样本创建, OPEN 是文献中第一个捕捉随时间推移情感和行为状态转变的数据 集。表2展示了这些转变的统计数据,在文献[64],[65]中被 称为情感和行为状态的动态。这些动态为开发预测未来情感 状态、行为状态和参与度水平的模型提供了有价值的见解。

例如,在表2中,当情感状态是平静/满意时,最频繁的 过渡发生在非任务状态和任务状态之间。当参与者处于任务 状态时,从平静/满意过渡到有动力/兴奋的可能性明显高于 过渡到无聊或困惑/沮丧。此外,在任务状态下,从无聊过渡 到有动力/兴奋的可能性远低于过渡到平静/满意。尽管在其 他人群中已经研究过情感和行为状态的动态,但对老年患者 人群中这些动态的分析是新颖的。值得注意的是,在任务状 态期间观察到的情感状态过渡趋势与之前在年轻健康的学生 人群中的研究结果一致[64]。

除了数据集中可变长度的数据样本外,还实现了两种方法 来生成固定长度的数据样本。在将数据集分段为长度为*1*秒 的固定长度数据样本后,如果 l 跨越超过一秒钟,某些数据样 本可能会与不同参与者情感和行为状态的过渡重叠。使用了 两种策略来确定分配给每个固定长度数据样本的参与者状态 或标签。第一个为大小为 l 的数据段标记的策略涉及多数投 票。在这种方法中,在数据样本中出现次数最多的情感和行 为状态被选为该数据样本的相应标签。第二种策略利用的原 则是,如果参与者的行为状态在数据样本中的任何时刻是离 线的,那么整个数据样本的行为状态应该标记为离线的。然 而, 情感状态仍然使用多数投票来确定。

#### 2.8.3 类别分布

表 3 展示了在不同设置下数据样本在行为、情感和参与状态 方面的分布: 原始的可变长度样本以及使用子节 ?? 中描述 的策略1和2创建的5秒、10秒和30秒的固定长度样本。 在可变长度、5秒和10秒的设置中,情感状态的分布不平衡, 明显倾向于平静/满意。然而,在这些设置中,行为和参与状 态的分布相对平衡。相反,对于 30 秒的设置,行为和参与状 态的分布变得不平衡。与策略1相比,使用策略2会增加不 在任务上的样本,这随后导致更多未参与的数据样本。数据 样本的总数在可变长度设置中为 4,494, 在 5 秒、10 秒和 30 秒固定长度设置中分别为 22,430、10,956 和 3,787。

在数据样本在情感和行为参与成分以及参与程度上的分布 方面, OPEN 数据集与现有的公众参与数据集(如 DAiSEE [32] 和 EngageNet [26]) 既有相似之处, 也有不同之处。OPEN 数据集是从居家的老年患者在连续的每周会议上收集的,而 DAiSEE 和 EngageNet 则是在受控环境下从年轻学生中收集 的,时间为单次 20 到 30 分钟的会议。在 DAiSEE 中,参与仅 被标注为一种情感状态,而忽略了行为方面。相反,EngageNet 在参与标注中考虑了情感和行为两个成分。

DAiSEE 的 10 秒数据样本与 OPEN 在 10 秒长度数据样 本配置之间的一个关键相似之处在于情感参与标注的不平衡。 在 DAiSEE 中, 95.6 % 的样本被标记为高度参与。同样地, 在 OPEN 中, 情感成分明显偏向于平静/满意状态(93.96 %) 的数据样本),这通常与参与相关联 [61],[63]。另一个相似 之处可以观察到 EngageNet 和 OPEN 之间的整体参与分布。 EngageNet 基于情感和行为线索对 10 秒的数据样本进行了标 注,其分类没有表现出明显的不平衡,30.5%低参与和69.5 % 高参与。这与 OPEN 在 10 秒数据样本的整体参与分布 (42.4 % 未参与对比 57.6 % 已参与) 相当, 该分布同样结合 了情感和行为成分,并遵循 HELP 协议。在 OPEN 的其他配 置中,如在表3中所描述的使用30秒数据样本时,参与状态 的分布与 DAiSEE 和 EngageNet 均有所不同。

如子章节 2.7.1 所概述的, 注释过程涉及根据参与的情感 或行为成分的变化创建新的数据样本,精确到秒。因此,数 据集的每一秒都包括来自三名注释者的注释。

在行为参与注释中获得了 Fleiss' Kappa 为 0.8254, Krippendorff's Alpha 为 0.8434, 以及成对的 Cohen's Kappa 分别 为 0.7376、0.8492 和 0.8891, 行为参与包括三个状态: 任务 外、任务中和不可见。

在情感参与注释中,涉及五种状态(无聊、平静/满 意、困惑/沮丧、积极/兴奋和不可见), Fleiss' Kappa 值 为 0.7124, Krippendorff's Alpha 值为 0.5746, 以及成对的 Cohen's Kappa 值为 0.6662、0.7695 和 0.7014。

具有三种状态(参与、未参与和不可见)的参与注释结果 的 Fleiss' Kappa 为 0.8108, Krippendorff's Alpha 为 0.8107, 两两 Cohen's Kappa 分别为 0.7172、0.8448 和 0.8704 [66]。

在大多数情况下,协同性指标范围在 0.61-0.80 (显著一致) 和 0.81-1.00(几乎完美一致)[67] 之间,超过了最初 HELP [61] 和先前数据集 [26] 中报告的水平。与 HELP [61] 中的趋

TABLE 2

行为(不在任务和在任务)和情感(无聊、平静/满意、困惑/沮丧和有动力/兴奋)状态的动态。第一列列出的状态代表当前状态,而后续列中的数字表示相应的下一状态出现的频率。例如,状态"不在任务,平静/满意"在 21 次中跟随"不在任务,无聊"状态。

	Off, Bored	Off, Calm	Off, Confused	Off, Motivated	On, Bored	On, Calm	On, Confused	On, Motivated
Off, Bored	_	21	0	2	0	14	0	0
Off, Calm	35	_	14	87	1	1586	0	22
Off, Confused	1	16	_	1	0	1	6	0
Off, Motivated	2	101	0	_	0	20	0	76
On, Bored	18	0	0	0	-	17	0	3
On, Calm	13	1576	2	27	3	_	5	232
On, Confused	0	0	8	0	0	4	_	0
On, Motivated	1	24	0	92	1	208	0	-

TABLE 3 在不同环境中,数据样本在行为、情感和参与状态中的分布。

Class	Variable- length	5 seconds- strategy 1	10 seconds- strategy 1	30 seconds- strategy 1	5 seconds- strategy 2	10 seconds- strategy 2	30 seconds- strategy 2
Off-Task On-Task	$\begin{array}{c} 2144 \\ 2350 \end{array}$	9544 12886	$4621 \\ 6335$	$1580 \\ 2207$	$\begin{array}{c c} 10787 \\ 11643 \end{array}$	$5747 \\ 5209$	$2543 \\ 1244$
Bored Calm/Satisfied Confused/Frustrated Motivated/Excited	$     113 \\     3822 \\     38 \\     521   $	511 20952 74 893	$248 \\ 10295 \\ 38 \\ 375$	$80 \\ 3601 \\ 11 \\ 95$	- - - -		- - - -
Engaged Not-Engaged	$2214 \\ 2280$	12828 9602	$6308 \\ 4648$	$2201 \\ 1586$	$ \begin{array}{c c} 11603 \\ 10827 \end{array} $	$5194 \\ 5762$	$1243 \\ 2544$
Total	4494	22430	10956	3787	-	_	_

势一致,情感投入注释的协同性低于行为投入注释。这归因 于情感投入注释本质上的主观性 [61] 和可能状态的更多数量。

#### 2.8.4 上下文类型分布

图 3 展示了在小节 2.7.2 中描述的七种语境类型的分布。在所 有会议中,大多数时间(64.69%)用于前两种语境类型,即临 床医生正在与所有参与者或与某个特定个体交谈。在 21.39% 的时间内,某个参与者正在与临床医生交谈,而剩下的 13.92 %则对应于所有其他语境类型。值得注意的是,每个参与者 (编号 1 到 11)都在注释中被识别出来,从而可以分析不同语 境类型如何影响个人参与者在整个会话期间的参与程度。



Fig. 3. 如小节 2.7.2 所述,汇总所有会话后,七种上下文类型的持续时 间分布。

正如前面提到的,为了保护患者隐私,原始的音频-视频数 据被保留;但是,从视频中提取了面部、手部和身体关节的 标志点,以及从这些标志点提取出的行为和情感特征。从每 个视频帧中获取了以下特征。

使用 OpenFace [68],提取了 668 个特征,包括:六个用 于眼睛凝视方向向量的世界坐标,两个用于凝视方向的弧度, 112 个用于 2D 眼睛标记 (像素),168 个用于 3D 眼睛标记 (毫米),头部位置和旋转各三个,2D 面部标记 (像素)为 68 x 2,3D 面部标记 (毫米)为 68 x 3,以及 34 个用于编号为 1、2、4、5、6、7、9、10、12、14、15、17、20、23、25、26、 45 的 17 个面部动作单元 (FAU)的出现和强度。

使用预训练的 EmoFAN [69] ,从每个视频帧的面部区域 提取了效价和唤醒度的两个值。

众所周知拥有实时和跨平台能力的 MediaPipe 深度学习 框架 [70] 被用来从视频中提取面部、手部和身体的关键点。 在 MediaPipe 中, Attention Mesh 模型 [71] 能够精确检测整 个脸部 468 个三维面部关键点及额外的 10 个虹膜关键点,总 共为面部和虹膜提供了 478×3 个三维关键点。MediaPipe 还 为手部跟踪提供了 21×3 个三维关键点并为身体关节提供了 33×3 个三维关键点。

数据集中一个数据样本的格式是一个 CSV 文件,其中每 一行对应数据样本的一个帧,列表示从帧中提取的前述特征。 接下来的内容是参与度的真实情感和行为成分,以及与帧相 关联的参与度注释。

为了对 OPEN 进行基准测试和技术验证,开发了一系列 在文献中以在先前数据集上优异表现而闻名的机器学习和深 度学习模型。鉴于 OPEN 中数据类型的多样性和数据样本长 度的变化,这些模型经过训练和评估,以确保与各自输入数 据类型的特定特征保持一致。

为了与 OPEN 中的参与度标注保持一致,参与度推理被 表述为二元分类任务。在参与度识别领域 [23], [24], [25], [26] 中,首次系统地研究了两种不同类型的参与度推理任务:参与 度检测,这是一项在现有文献中广泛涉及的任务,重点在于 使用来自当前时间戳的数据来估计当前时间戳的参与度;以 及参与度预测,这是一项新颖任务,涉及基于先前时间戳的 数据来估计未来时间戳的参与度。

此外, 在交互识别领域的首次实例中, 系统地评估了各种 模型和特征集在不同数据集设置下的性能,考虑的数据样本 长度为5秒、10秒、30秒和可变长度, 如小节 2.8.3 中所述。

此外,在参与度识别领域 [23], [24], [25], [26] 中,首次比 较了两种不同的参与度估计方法:一种是开发直接输出参与 度作为二分变量的模型,这是现有文献的重点;另一种是采 用一种新颖的方法,开发输出参与度的情感和行为组成部分 的模型,从这些部分可以随后推断参与度。

如子节?? 所述,数据集特征中的几个子集被用于识别参与度,与现有文献[24],[25],[26]一致。不同组合的下列特征(从连续的视频帧中提取)被用作能够分析序列数据的机器学习和深度学习模型的输入。这些特征包括 6 个眼动特征、6 个头部姿势(位置和旋转)特征、17 个 FAU [26]强度,以及 2 个情感特征的效价和唤醒度[72]。除了上述特征外,如子节?? 中详细描述的面部标志也用于参与度估计。所使用的模型包括 LSTM、ST-GCN [73]、Transformer 和 ROCKET [74]。

LSTM 模型由一个两层的 LSTM 架构组成,每层包含 64 个隐藏单元。接着是一个大小为 64 × 1 的全连接层,其 中输出维度 1 对应于二分类任务。Transformer 模型包含两 层 Transformer 编码器,输入维度为 64,具有四个注意力 头,以及 128 作为前馈网络模型的维度。输入特征首先通 过一个线性投影层处理,将原始输入维度映射到 64,然后 传递给 Transformer 编码器。ST-GCN 由三层 ST-GCN 层组 成,随后是一个平均池化层。平均池化层的输出进一步通过 一个卷积层处理,将其投影到单个输出维度进行分类。所有 深度学习模型都使用 Adam 优化器优化,进行了 1000 次迭 代,初始学习率为 0.0001,每 250 次迭代学习率降低 10 倍。 ROCKET 模型使用 10,000 个随机初始化的卷积核来提取特 征表示。由 ROCKET 生成的变换特征空间随后使用分类增 强 (CatBoost)模型进行分类。CatBoost 分类器进行了 1000 次训练迭代,最大树深度为 6。

作为二分类问题构建的参与度推断评估指标包括准确 率、精确率、召回率、F1-得分、接收者操作特征曲线下面积 (AUC-ROC)和精确率-召回率曲线下面积(AUC-PR)。

为了评估模型在未见参与者数据上的泛化能力,模型使用 两种不同设置的 11 折交叉验证(CV)(其中 11 表示数据集 中参与者的数量)和留一参与者法(LOPO)CV进行训练和 评估。标注过程被设计为尽可能与参与者无关,数据包括从 视频中提取的特征或面部标志,相较于原始视频,这些数据 对参与者的依赖性较低。然而,预期 LOPO CV 的性能会低 于 11 折 CV,因为后者允许同一参与者的数据在一个 CV 迭 代中的训练集和验证集中同时出现。这是因为参与感和不参 与感的情感和行为指标的表现对于每个参与者来说都是独特 的。

## 2.9 参与检测

表格 4 展示了在 OPEN 上使用 10 秒数据样本进行不同特征 集和模型的参与度检测结果,如小节 2.8.3 中所述。该时长代 表了文献中最常见的数据样本长度 [16], [26], [32]。在表格 4 中展示的所有特征集和模型组合中,11 折交叉验证的结果始 终优于 LOPO 交叉验证的结果。如小节 ?? 中所讨论的,这 种性能差距源于不同参与者独特的情感和行为状态,以及参 与度的不同表现形式,使得模型更难以泛化到 LOPO 交叉验 证测试集中未见的参与者。

如表4所示,当使用包含眼动、头部姿态和效价-唤醒特征 集时,无论是否包括 FAUs, Transformer 模型实现了最佳性能,其次是 ROCKET 和 LSTM。然而,当使用 3D 面部标志进行建模时,专门为时空图数据设计的 ST-GCN 在 LOPO CV 中优于 Transformer 和 ROCKET。值得注意的是,在11 折 CV 中,ROCKET 超越了 ST-GCN,达到了最高精度 0.8112 和 AUC-ROC 为 0.8934。

视频的原始帧率,以及从视频帧中提取的特征是 16 帧每 秒。然而,表4中呈现的结果是在每秒 8 帧的情况下获得的, 这意味着每隔一帧被使用。表5比较了在每秒 16 帧与 8 帧 情况下,10秒数据样本中 3D 面部标志和 ROCKET 的结果。 如表5所示,在 LOPO CV 中使用 16 帧每秒会降低性能,但 在 11 折 CV 中会提高结果。这是因为更高的帧率捕捉了参与 者情感和行为状态的更多细微细节。在 11 折 CV 中,同一参 与者的数据出现在训练和测试集中,这些细节增强了模型性 能。然而,在 LOPO CV 中,测试集中的参与者在训练中是 未见过的,缺乏共享的参与者特定的细节使得这些额外的细 微差别不太有益,甚至可能有害。

在表 5 中使用 10 秒数据样本在 OPEN 上获得最高结果的 模型和特征集的性能进一步在使用 5 秒、30 秒和可变长度数 据样本的 OPEN 上进行了评估,如表 5 所示。在 OPEN 上 使用 5 秒和 30 秒数据样本的结果分别优于和劣于使用 10 秒 数据样本的结果。这主要是由于可用样本的数量: 5 秒数据 的 OPEN 包含的样本数量是其两倍,而 30 秒数据的 OPEN 的样本数量是其三分之一,详见表 3 。此外,在使用 30 秒数 据样本的 OPEN 中,较长序列让模型更难有效捕获长期依赖 性。在处理可变长度序列时使用了 LSTM 模型,与眼动、头 部姿态和效价-唤醒特征一起进行处理。可变长度序列对应于 使用自适应策略标注的原始数据样本,每当情感或行为状态 发生变化时,都会创建一个新样本,如子节 ?? 中所述。如 观察到的那样,可变长度样本的结果显著低于固定长度样本, 强调了对更先进模型和替代特征集的需求。

如第??节所述,OPEN中的固定长度数据样本是通过将 1小时会话拆分为10秒段来创建的。这意味着在一次会话中, 一个10秒的数据样本可以被视为相对于下一个10秒数据样 本的前一个时间戳。对于参与度预测,给每个数据样本分配 来自同一会话的下一个数据样本的参与度标签。通过这种方 式,模型学习根据前一个时刻的数据预测未来时间戳的参与 度。表6展示了ROCKET和Transformer使用眼动、头部姿 势、效价-唤醒和FAUs在OPEN上用10秒数据样本的参与 度预测结果。在LOPO和11折交叉验证中,Transformer和 ROCKET分别表现最佳,其中ROCKET达到了0.7034的 最高准确率。

表 6 还包括使用 11 折交叉验证(CV)下不同数据样本长度的三维面部标志进行的 ROCKET 参与度预测结果。对于 30 秒的数据样本,OPEN 的结果优于 10 秒的样本,而 5 秒 的数据样本表现较差。这表明,提供更长时间的三维面部标志数据可以提高模型在未来时间戳上预测参与度的准确性。

为了评估参与度的情感和行为组成部分如何用于检测参与度,表现最佳的模型,即使用面部标志的 ROCKET,在 OPEN 上用 10 秒的数据样本进行了训练。焦点是在多任务学习环境下的参与检测。首先,ROCKET 被训练用来从输入数据中提取特征表示,而不以任何特定输出为目标。这些提取的特征随后用于训练两个独立的 CatBoost 模型以进行下游分类任务:一个是 4 类情感参与组件,另一个是 2 类行为参与组件,如子节 2.7.1 所述。接下来,检测到的情感和行为参与组件被合并以生成最终的二元参与标签,如子节 2.7.1 详细 所述。LOPO 和 11 折交叉验证的结果展示在表 7 中。如观察 所示,情感状态检测的准确性显著高于行为状态分类,并且最终的参与准确性与行为状态分类的结果高度一致。

## 3 结论与未来工作

本论文引人了一个独特的数据集,用于开发用于评估参与度 的机器学习和深度学习模型。该数据集在多个方面具有独特 TABLE 4 在 OPEN 上使用不同特征集和模型的参与度检测结果。(EG: 眼动, HP: 头部姿势, VA: 价度-唤醒度, FAU: 面部动作单元, CV: 交叉验证, 11-fold: 十一折 CV, LOPO: 留一参与者交叉验证)

CV	Features	Model	Accuracy	Precision	Recall	F1 Score	AUC-ROC	AUC-PR
LOPO	EG, HP, VA	LSTM	0.6437	0.6444	0.7486	0.6926	0.6633	0.6759
LOPO	EG, HP, VA	Transformer	0.6671	0.6888	0.6915	0.6902	0.7087	0.7178
LOPO	EG, HP, VA	ROCKET	0.6133	0.6221	0.7098	0.6631	0.6631	0.6908
11-fold	EG, HP, VA	LSTM	0.6466	0.6563	0.7155	0.6846	0.6902	0.7097
11-fold	EG, HP, VA	Transformer	0.7264	0.7731	0.6932	0.7310	0.7917	0.8120
11-fold	EG, HP, VA	ROCKET	0.7089	0.7260	0.7341	0.7301	0.7828	0.8011
LOPO	EG, HP, VA, FAU	Transformer	0.6086	0.6264	0.6685	0.6468	0.6199	0.6086
LOPO	EG, HP, VA, FAU	ROCKET	0.5925	0.5991	0.7250	0.6561	0.6313	0.6622
11-fold	EG, HP, VA, FAU	Transformer	0.7707	0.7915	0.7771	0.7842	0.8240	0.8263
11-fold	EG, HP, VA, FAU	ROCKET	0.7384	0.7419	0.7853	0.7630	0.8070	0.8110
LOPO	Facial Landmarks	Transformer	0.6345	0.6932	0.5707	0.6260	0.6839	0.7125
LOPO	Facial Landmarks	ST-GCN	0.6529	0.7050	0.5530	0.6198	0.6973	0.6815
LOPO	Facial Landmarks	ROCKET	0.6362	0.6389	0.7391	0.6854	0.7019	0.7455
11-fold	Facial Landmarks	Transformer	0.7018	0.7484	0.6685	0.7062	0.7460	0.7820
11-fold	Facial Landmarks	ST-GCN	0.7660	0.8054	0.7438	0.7734	0.8374	0.8503
11-fold	Facial Landmarks	ROCKET	0.8112	0.8313	0.8211	0.8261	0.8934	0.9125

TABLE 5 在不同数据样本长度和帧率下,使用 OPEN 的不同特征集和模型的参与度检测结果。(EG:眼动,HP:头部姿态,VA:效价-唤醒,FAU:面部动 作单元,CV:交叉验证,11-fold:11 折交叉验证,LOPO:逐个参与者留出验证,FPS:每秒帧数)

Length	FPS	CV	Features	Model	Accuracy	Precision	Recall	F1 Score	AUC-ROC	AUC-PR
10	8	LOPO	Facial Landmarks	ROCKET	0.6362	0.6389	0.7391	0.6854	0.7019	0.7455
10	16	LOPO	Facial Landmarks	ROCKET	0.6178	0.6248	0.7190	0.6686	0.6863	0.7370
5	16	LOPO	Facial Landmarks	ROCKET	0.6413	0.6604	0.8587	0.7466	0.6678	0.7768
30	16	LOPO	Facial Landmarks	ROCKET	0.5945	0.5608	0.7737	0.6503	0.7114	0.7419
10	8	11-fold	Facial Landmarks	ROCKET	0.8112	0.8313	0.8211	0.8261	0.8934	0.9125
10	16	11-fold	Facial Landmarks	ROCKET	0.8161	0.8361	0.8252	0.8306	0.8922	0.9147
5	16	11-fold	Facial Landmarks	ROCKET	0.8347	0.8384	0.8140	0.8260	0.9158	0.9131
30	16	11-fold	Facial Landmarks	ROCKET	0.7556	0.7830	0.8791	0.8283	0.8201	0.9023
Variable	16	LOPO	EG, HP, VA	LSTM	0.5207	0.5166	0.4227	0.4650	0.5039	0.4940
Variable	16	11-fold	EG, HP, VA	LSTM	0.5656	0.5577	0.4842	0.5184	0.5854	0.5614

性。它是从老年患者中收集的,并包括来自单个会话的连续 参与数据以及跨多个会话的纵向参与数据。它捕捉了会话级 别的上下文信息和参与者级别的数据。此外,还包含表现和 行为状态旁注,观察者提供的参与度标注。数据样本的特征 在于自适应时间戳,由个体的表现或行为状态变化触发,使 得能够进行参与度检测和预测。该数据集也是最大公开可用 的参与度数据集,通过仅包含非可识别的面部、手部和身体 关节特征点及从这些特征点派生的表现和行为特征来确保隐 私保护。为了进行初步的技术验证,基于数据集中不同的特 征训练了多种模型,在参与度检测和预测方面分别获得了高 达 81 % 和 70 % 的准确率。然而,该数据集存在某些限制, 包括参与者数量相对较少、人口多样性低,以及缺乏音频和 原始视频数据。此外,在本文的技术验证部分,并未完全探 讨数据集的所有独特特征。

在未来的工作中,将利用会话级数据和上下文类型信息进行面向上下文和参与者上下文的参与度估计。我们将分析虚 拟教育项目多个会话中的纵向参与度数据,以及单个会话内 的参与度,以进行纵向和会话级的参与度评估。此外,将探索 特别是能够处理可变长度序列的高级模型,用于参与度的检 测和预测。将采用先进的多任务学习和多标签学习技术,从 其组成部分来估计参与度。同时,将制定策略以解决参与度 情感成分中样本分布不平衡的问题。将研究开发更多可泛化 的模型的方法,以在未见过的参与者上实现足够的性能。将 探索多模态大语言模型,用于标志性数据的自动分析、情感 和行为状态的推断,以及学习会话上下文信息的整合,旨在 提高参与度识别管道的性能和可解释性。最后,将研究所开 发模型在不同数据集上的适用性,特别是因为以前的数据集 主要来自年轻、健康的学生,而本文引入的数据集则着重于 老年患者。这项研究由新前沿研究基金和 J.P. Bickell 基金 会医学研究资助。作者向 Christopher Ho, Azra Ince, Aliana Jamal 和 Ahmed Mokhtar 致以诚挚的感谢和欣赏,感谢他们 在数据集观察性注释中的宝贵贡献。

有兴趣获取本文中介绍的 OPEN 数据集的研究人员应联 系主要研究人员,邮箱为 shehroz.khan@uhn.ca。在签署确保 遵守伦理标准的数据共享协议后,将授予对数据集的访问权 限。

4

作者贡献声明 S.S.K. 和 T.J.F.C. 构思了这项研究。A.A. 在 S.S.K. 和 T.J.F.C. 的支持下,准备了研究伦理委员会的文 件以获得研究批准和数据集发布。S.S. 在 A.A.、S.S.K. 和 T.J.F.C. 的支持下,获取了参与者的同意并收集了数据。A.A. 设计了数据标注协议,而 S.S. 在 A.A.、S.S.K. 和 T.J.F.C. 的 支持下,管理了数据标注过程。A.A. 在手稿中描述的各种设 定下预处理、整理并创建了数据集。A.A. 进行了所有机器学 习和深度学习实验以进行数据集的技术验证。A.A. 起草了手 稿的所有部分。所有合著者都审阅了手稿。 TABLE 6 在不同数据样本长度下,不同特征集和模型在 OPEN 上的参与度预测结果。(EG: 注视点,HP: 头部姿态,VA: 价-唤醒度,FAU: 面部动作单元, CV: 交叉验证,11-fold: 11 折交叉验证,LOPO: 留一参与者交叉验证,FPS: 每秒帧数)

Length	CV	Features	Model	Accuracy	Precision	Recall	F1 Score	AUC-ROC	AUC-PR
10	LOPO	EG, HP, VA, FAU	Transformer	0.5908	0.6128	0.6533	0.6324	0.5922	0.6241
10	LOPO	EG, HP, VA, FAU	Rocket	0.5686	0.5766	0.7346	0.6461	0.5891	0.6143
10	11-fold	EG, HP, VA, FAU	Transformer	0.6854	0.7117	0.6943	0.7029	0.7347	0.7478
10	11-fold	EG, HP, VA, FAU	Rocket	0.7034	0.7070	0.7629	0.7339	0.7646	0.7791
10	11-fold	Facial Landmarks	Rocket	0.6102	0.6157	0.7258	0.6662	0.6565	0.6693
5	11-fold	Facial Landmarks	Rocket	0.5924	0.5687	0.6409	0.6029	0.6478	0.6493
30	11-fold	Facial Landmarks	Rocket	0.6457	0.6891	0.8655	0.7673	0.5933	0.7557

TABLE 7 使用在子节 ?? 中描述的多任务学习环境中训练的 3D 面部特征,通过 ROCKET 在 OPEN 上检测行为和情感参与的组件,以及整体参与检 测。(LOPO:排除一个参与者交叉验证,11 折:11 折交叉验证)

Target –Metric	LOPO	11-fold
Behavioral –Accuracy	0.6052	0.8180
Behavioral – Precision	0.5339	0.8046
Behavioral –Recall	0.8175	0.8040
Behavioral –F1 Score	0.6459	0.8043
Behavioral –AUC-ROC	0.6491	0.8946
Behavioral –AUC-PR	0.5168	0.8729
Emotional –Accuracy	0.9165	0.9462
Engagement –Accuracy	0.6059	0.8210

## References

- World Health Organization, "Rehabilitation," https://www.who.int/news-room/factsheets/detail/rehabilitation, 2023, Accessed: January 30, 2023.
- [2] S. Shanmugasegaram, L. Gagliese, P. Oh, D. E. Stewart, S. J. Brister, V. Chan, and S. L. Grace, "Psychometric validation of the cardiac rehabilitation barriers scale," *Clinical rehabilitation*, vol. 26, no. 2, pp. 152–164, 2012.
- [3] I. Nabutovsky, D. Breitner, A. Heller, Y. Levine, M. Moreno, M. Scheinowitz, C. Levin, and R. Klempfner, "Home-based cardiac rehabilitation among patients unwilling to participate in hospital-based programs," *Journal of Cardiopulmonary Rehabilitation and Prevention*, vol. 44, no. 1, pp. 33–39, 2024.
- [4] I. Boukhennoufa, X. Zhai, V. Utti, J. Jackson, and K. D. McDonald-Maier, "Wearable sensors and machine learning in post-stroke rehabilitation assessment: A systematic review," *Biomedical Signal Processing and Control*, vol. 71, p. 103197, 2022.
- [5] M. Naeemabadi, H. Fazlali, S. Najafi, B. Dinesen, J. Hansen et al., "Telerehabilitation for patients with knee osteoarthritis: a focused review of technologies and teleservices," *JMIR Biomedical Engineering*, vol. 5, no. 1, p. e16991, 2020.
- [6] S. Rahman, S. Sarker, A. N. Haque, M. M. Uttsha, M. F. Islam, and S. Deb, "Ai-driven stroke rehabilitation systems and assessment: A systematic review," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2022.
- [7] A. Abedi, T. J. Colella, M. Pakosh, and S. S. Khan, "Artificial intelligence-driven virtual rehabilitation for people living in the community: A scoping review," *NPJ Digital Medicine*, vol. 7, no. 1, p. 25, 2024.
- [8] F. A. Bright, N. M. Kayes, L. Worrall, and K. M. McPherson, "A conceptual review of engagement in healthcare and rehabilitation," *Disability and rehabilitation*, vol. 37, no. 8, pp. 643–654, 2015.
- [9] M. M. Danzl, N. M. Etter, R. D. Andreatta, and P. H. Kitzman, "Facilitating neurorehabilitation through principles of engagement," *Journal of allied health*, vol. 41, no. 1, pp. 35–41, 2012.

- [10] Z. Liu, A. Brandon-Jones, and C. Vasilakis, "Unpacking patient engagement in remote consultation," *International Journal of Operations & Production Management*, vol. 44, no. 13, pp. 157– 194, 2024.
- [11] J. A. Fredricks, P. C. Blumenfeld, and A. H. Paris, "School engagement: Potential of the concept, state of the evidence," *Review of educational research*, vol. 74, no. 1, pp. 59–109, 2004.
- [12] G. M. Sinatra, B. C. Heddy, and D. Lombardi, "The challenges of defining and measuring student engagement in science," pp. 1–13, 2015.
- [13] H. Salam, O. Celiktutan, H. Gunes, and M. Chetouani, "Automatic context-aware inference of engagement in hmi: A survey," *IEEE Transactions on Affective Computing*, vol. 15, no. 2, pp. 445–464, 2024.
- [14] B. M. Booth, N. Bosch, and S. K. D' Mello, "Engagement detection and its applications in learning: a tutorial and selective review," *Proceedings of the IEEE*, vol. 111, no. 10, pp. 1398– 1422, 2023.
- [15] M. Dewan, M. Murshed, and F. Lin, "Engagement detection in online learning: a review," *Smart Learning Environments*, vol. 6, no. 1, pp. 1–20, 2019.
- [16] S. S. Khan, A. Abedi, and T. Colella, "Inconsistencies in measuring student engagement in virtual learning–a critical review," arXiv preprint arXiv:2208.04548, 2022.
- [17] S. Mandia, R. Mitharwal, and K. Singh, "Automatic student engagement measurement using machine learning techniques: A literature study of data and methods," *Multimedia Tools and Applications*, vol. 83, no. 16, pp. 49641–49672, 2024.
- [18] N. Noceti, S. Campisi, A. Chirico, V. Cuculo, G. Grossi, M. Michelotto, F. Odone, A. Gaggioli, and R. Lanzarotti, "Predicting engagement of older people's virtual teams from video call analysis," *International Journal of Human-Computer Interaction*, pp. 1–12, 2024.
- [19] G. Guo, R. Guo, and X. Li, "Facial expression recognition influenced by human aging," *IEEE Transactions on Affective Computing*, vol. 4, no. 3, pp. 291–298, 2013.
- [20] L. Abbruzzese, N. Magnani, I. H. Robertson, and M. Mancuso, "Age and gender differences in emotion recognition," *Frontiers in psychology*, vol. 10, p. 2371, 2019.
- [21] L. Alonso-Recio, J. M. Serrano, and P. Martín, "Selective attention and facial expression recognition in patients with parkinson's disease," *Archives of clinical neuropsychology*, vol. 29, no. 4, pp. 374–384, 2014.
- [22] Y. Zhou, W. Han, X. Yao, J. Xue, Z. Li, and Y. Li, "Developing a machine learning model for detecting depression, anxiety, and apathy in older adults with mild cognitive impairment using speech and facial expressions: A cross-sectional observational study," *International journal of nursing studies*, vol. 146, p. 104562, 2023.
- [23] S. N. Karimah and S. Hasegawa, "Automatic engagement estimation in smart education/learning settings: a systematic review of engagement definitions, datasets, and methods," *Smart Learning Environments*, vol. 9, no. 1, pp. 1–48, 2022.
- [24] A. Abedi and S. S. Khan, "Engagement measurement based on facial landmarks and spatial-temporal graph convolutional networks," in *International Conference on Pattern Recognition*. Springer, 2024, pp. 321–338.
- [25] A. Vedernikov, P. Kumar, H. Chen, T. Seppänen, and X. Li, "Tcct-net: Two-stream network architecture for fast and efficient engagement estimation via behavioral feature signals," in *Pro-*

ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 4723–4732.

- [26] M. Singh, X. Hoque, D. Zeng, Y. Wang, K. Ikeda, and A. Dhall, "Do i have your attention: A large scale engagement prediction dataset and baselines," in *Proceedings of the 25th International Conference on Multimodal Interaction*, ser. ICMI '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 174–182.
- [27] S. K. D'Mello, "On the influence of an iterative affect annotation approach on inter-observer and self-observer reliability," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 136–149, 2015.
- [28] J. Whitehill, Z. Serpell, Y.-C. Lin, A. Foster, and J. R. Movellan, "The faces of engagement: Automatic recognition of student engagementfrom facial expressions," *IEEE Transactions on Affective Computing*, vol. 5, no. 1, pp. 86–98, 2014.
- [29] S. Aslan, Z. Cataltepe, I. Diner, O. Dundar, A. A. Esme, R. Ferens, G. Kamhi, E. Oktay, C. Soysal, and M. Yener, "Learner engagement measurement and classification in 1: 1 learning," in 2014 13th International Conference on Machine Learning and Applications. IEEE, 2014, pp. 545–552.
- [30] N. Bosch, "Detecting student engagement: human versus machine," in Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization, 2016, pp. 317–320.
- [31] J. Chen, N. Luo, Y. Liu, L. Liu, K. Zhang, and J. Kolodziej, "A hybrid intelligence-aided approach to affect-sensitive elearning," *Computing*, vol. 98, no. 1, pp. 215–233, 2016.
- [32] A. Gupta, A. D'Cunha, K. Awasthi, and V. Balasubramanian, "Daisee: Towards user engagement recognition in the wild," arXiv preprint arXiv:1609.01885, 2016.
- [33] A. Kamath, A. Biswas, and V. Balasubramanian, "A crowdsourced approach to student engagement recognition in elearning environments," in 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2016, pp. 1–9.
- [34] B. M. Booth, A. M. Ali, S. S. Narayanan, I. Bennett, and A. A. Farag, "Toward active and unobtrusive engagement assessment of distance learners," in 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII). IEEE, 2017, pp. 470–476.
- [35] E. Okur, N. Alyuz, S. Aslan, U. Genc, C. Tanriover, and A. Arslan Esme, "Behavioral engagement detection of students in the wild," in *International Conference on Artificial Intelligence in Education.* Springer, 2017, pp. 250–261.
- [36] N. Alyuz, E. Okur, U. Genc, S. Aslan, C. Tanriover, and A. A. Esme, "An unobtrusive and multimodal approach for behavioral engagement detection of students," in *Proceedings of the 1st* ACM SIGCHI International Workshop on Multimodal Interaction for Education, 2017, pp. 26–32.
- [37] J. Zaletelj and A. Košir, "Predicting students' attention in the classroom from kinect facial and body features," *EURASIP journal on image and video processing*, vol. 2017, no. 1, pp. 1–12, 2017.
- [38] A. Kaur, A. Mustafa, L. Mehta, and A. Dhall, "Prediction and localization of student engagement in the wild," in 2018 Digital Image Computing: Techniques and Applications (DICTA). IEEE, 2018, pp. 1–8.
- [39] I. Alkabbany, A. Ali, A. Farag, I. Bennett, M. Ghanoum, and A. Farag, "Measuring student engagement level using facial information," in 2019 IEEE International Conference on Image Processing (ICIP). IEEE, 2019, pp. 3337–3341.
- [40] O. Mohamad Nezami, M. Dras, L. Hamey, D. Richards, S. Wan, and C. Paris, "Automatic recognition of student engagement using deep learning and facial expression," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases.* Springer, 2019, pp. 273–289.
- [41] N. Alyuz, S. Aslan, S. K. D' Mello, L. Nachman, and A. A. Esme, "Annotating student engagement across grades 1–12: Associations with demographics and expressivity," in *International Conference on Artificial Intelligence in Education*. Springer, 2021, pp. 42–51.
- [42] P. Bhardwaj, P. Gupta, H. Panwar, M. K. Siddiqui, R. Morales-Menendez, and A. Bhaik, "Application of deep learning on student engagement in e-learning environments," *Computers & Electrical Engineering*, vol. 93, p. 107277, 2021.
- [43] K. Delgado, J. M. Origgi, T. Hasanpoor, H. Yu, D. Allessio, I. Arroyo, W. Lee, M. Betke, B. Woolf, and S. A. Bargal,

"Student engagement dataset," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3628–3636.

- [44] J. Ma, X. Jiang, S. Xu, and X. Qin, "Hierarchical temporal multiinstance learning for video-based student learning engagement assessment." in *IJCAI*, 2021, pp. 2782–2789.
- [45] S. Gupta, P. Kumar, and R. K. Tekchandani, "Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models," *Multimedia Tools and Applications*, pp. 1–30, 2022.
- [46] Y.-S. Jeong and N.-W. Cho, "Evaluation of e-learners' concentration using recurrent neural networks," *The Journal of Supercomputing*, pp. 1–18, 2022.
- [47] M. Verma, Y. Nakashima, N. Takemura, and H. Nagahara, "Multi-label disengagement and behavior prediction in online learning," in *International Conference on Artificial Intelligence* in Education. Springer, 2022, pp. 633–639.
- [48] Y. Chen, N. Bosch, and S. D'Mello, "Video-based affect detection in noninteractive learning environments." *International Educational Data Mining Society*, 2015.
- [49] H. Monkaresi, N. Bosch, R. A. Calvo, and S. K. D'Mello, "Automated detection of engagement using video-based estimation of facial expressions and heart rate," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 15–28, 2016.
- [50] B. De Carolis, F. D'Errico, N. Macchiarulo, and G. Palestra, " "engaged faces" : Measuring and monitoring student engagement from face and gaze behavior," in *IEEE/WIC/ACM International Conference on Web Intelligence-Companion Volume*, 2019, pp. 80–85.
- [51] S. Hutt, J. F. Grafsgaard, and S. K. D'Mello, "Time to scale: Generalizable affect detection for tens of thousands of students across an entire school year," in *Proceedings of the 2019 CHI* conference on human factors in computing systems, 2019, pp. 1–14.
- [52] P. Vanneste, J. Oramas, T. Verelst, T. Tuytelaars, A. Raes, F. Depaepe, and W. Van den Noortgate, "Computer vision and human behaviour, emotion and cognition detection: A use case on student engagement," *Mathematics*, vol. 9, no. 3, p. 287, 2021.
- [53] P. Buono, B. De Carolis, F. D' Errico, N. Macchiarulo, and G. Palestra, "Assessing student engagement from facial behavior in on-line learning," *Multimedia Tools and Applications*, pp. 1– 19, 2022.
- [54] C. Thomas, K. P. Sarma, S. S. Gajula, and D. B. Jayagopi, "Automatic prediction of presentation style and student engagement from videos," *Computers and Education: Artificial Intelligence*, p. 100079, 2022.
- [55] N. Bosch, S. K. D'mello, J. Ocumpaugh, R. S. Baker, and V. Shute, "Using video to automatically detect learner affect in computer-enabled classrooms," ACM Transactions on Interactive Intelligent Systems (TiiS), vol. 6, no. 2, pp. 1–26, 2016.
- [56] X. Zheng, S. Hasegawa, M.-T. Tran, K. Ota, and T. Unoki, "Estimation of learners' engagement using face and body features by transfer learning," in *International Conference on Human-Computer Interaction.* Springer, 2021, pp. 541–552.
- [57] K. Altuwairqi, S. K. Jarraya, A. Allinjawi, and M. Hammani, "A new emotion-based affective model to detect student's engagement," *Journal of King Saud University-Computer and Information Sciences*, vol. 33, no. 1, pp. 99–109, 2021.
- [58] H. L. O' Brien, A. T. Chen, J. Kaneshiro, and O. Zaslavsky, "User engagement in an online digital health intervention to promote problem solving," *Interacting with Computers*, vol. 36, no. 5, pp. 355–369, 2024.
- [59] J. E. Beck, "Using response times to model student disengagement," in *Proceedings of the ITS2004 Workshop on Social and Emotional Intelligence in Learning Environments*, vol. 20, no. 2004. Maceio, 2004, pp. 88–95.
- [60] Health e-University, ``Cardiaccollege patient education program cardiac rehabilitafor tion," n.d., accessed: 2025-01-04. [Online]. Available: https://www.healtheuniversity.ca/en/CardiacCollege
- [61] S. Aslan, S. E. Mete, E. Okur, E. Oktay, N. Alyuz, U. E. Genc, D. Stanhill, and A. A. Esme, "Human expert labeling process (help): towards a reliable higher-order user state labeling process and tool to assess student engagement," *Educational Technology*, pp. 53–59, 2017.
- [62] S. Khan and S. Safa, "Revisiting annotations in online student engagement," in *Proceedings of the 2024 10th International*

Conference on Computing and Data Engineering, 2024, pp. 111–117.

- [63] B. Woolf, W. Burleson, I. Arroyo, T. Dragon, D. Cooper, and R. Picard, "Affect-aware tutors: recognising and responding to student affect," *International Journal of Learning Technology*, vol. 4, no. 3/4, pp. 129–164, 2009.
- [64] S. D'Mello and A. Graesser, "Dynamics of affective states during complex learning," *Learning and Instruction*, vol. 22, no. 2, pp. 145–157, 2012.
- [65] R. S. d Baker, M. Rodrigo, T. Mercedes, and U. E. Xolocotzin, "The dynamics of affective transitions in simulation problemsolving environments," in *International Conference on Affective Computing and Intelligent Interaction*. Springer, 2007, pp. 666– 677.
- [66] K. L. Gwet, Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters. Advanced Analytics, LLC, 2014.
- [67] M. L. McHugh, "Interrater reliability: the kappa statistic," *Bio-chemia medica*, vol. 22, no. 3, pp. 276–282, 2012.
- [68] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018). IEEE, 2018, pp. 59–66.
- [69] A. Toisoul, J. Kossaifi, A. Bulat, G. Tzimiropoulos, and M. Pantic, "Estimation of continuous valence and arousal levels from faces in naturalistic conditions," *Nature Machine Intelligence*, vol. 3, no. 1, pp. 42–50, 2021.
- [70] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee *et al.*, "Mediapipe: A framework for building perception pipelines," *arXiv preprint arXiv:1906.08172*, 2019.
- [71] I. Grishchenko, A. Ablavatski, Y. Kartynnik, K. Raveendran, and M. Grundmann, "Attention mesh: High-fidelity face mesh prediction in real-time," arXiv preprint arXiv:2006.10962, 2020.
- [72] A. Abedi and S. S. Khan, "Affect-driven ordinal engagement measurement from video," *Multimedia Tools and Applications*, pp. 1–20, 2023.
- [73] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.
- [74] A. Dempster, F. Petitjean, and G. I. Webb, "Rocket: exceptionally fast and accurate time series classification using random convolutional kernels," *Data Mining and Knowledge Discovery*, vol. 34, no. 5, pp. 1454–1495, 2020.