

# 重新审视 D-vlog 数据集：利用输入归因方法

黄一凡<sup>1</sup> 姚哲文<sup>1</sup>

<sup>1</sup> 北京大学 计算机学院

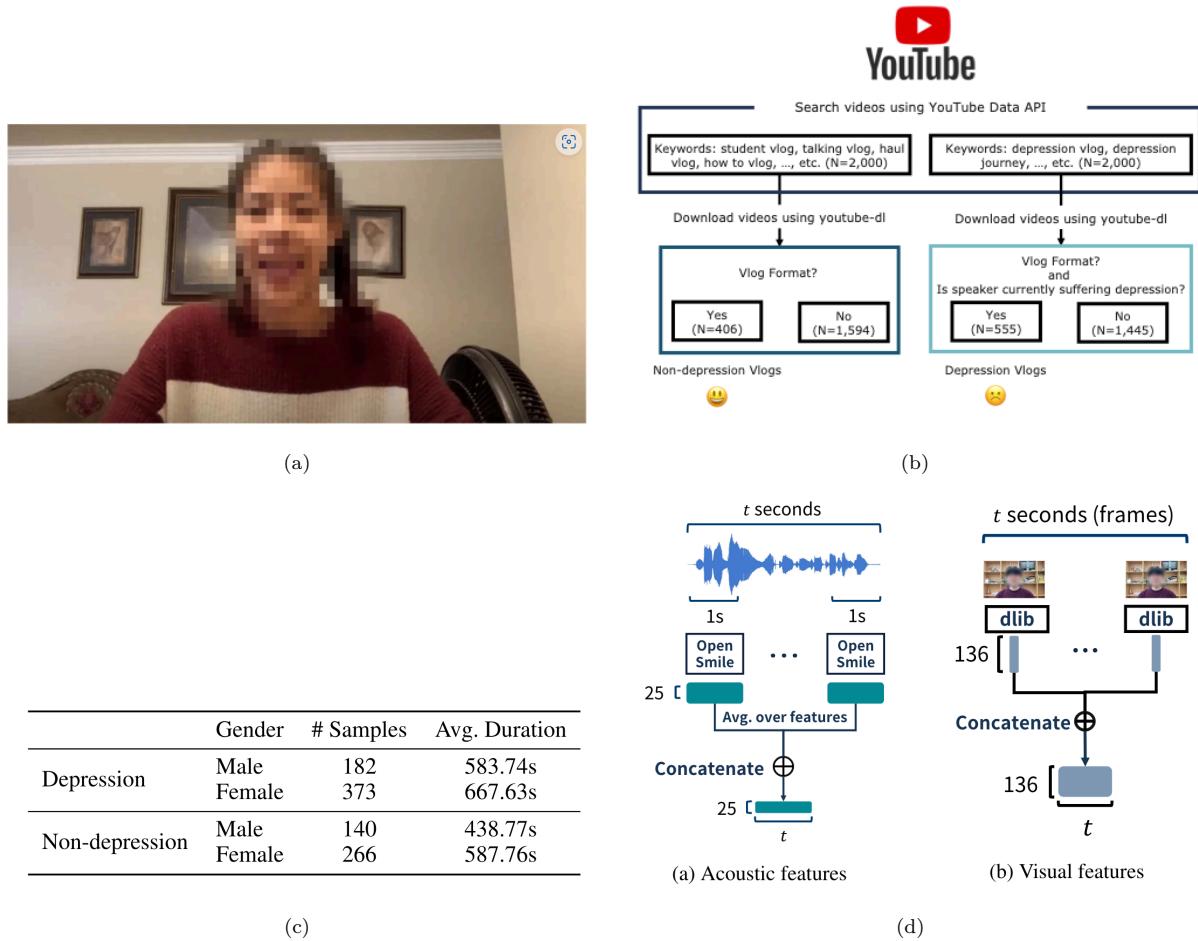
**摘要** 基于深度学习的抑郁检测能高效地对潜在抑郁人群进行筛查，对于心理健康领域具有重大意义。D-vlog 数据集作为近年公开的视听觉双模态抑郁检测数据集，因其易用性、跨数据集泛化能力和数据贴近日常生活场景的特点而受到广泛关注。本项目将重新审视 D-vlog 数据集，探究该数据集偏好什么样的模型，以及数据集中的哪些特征更加重要。我们手动实现并训练了 TMeanNet, Depression Detector 和 TAMFN 三个模型，评估它们在 D-vlog 上的性能，发现时序信息建模和早期特征融合是提升 D-vlog 抑郁检测性能的关键。进一步，我们使用积分梯度法对输入特征进行归因，发现只有少量听觉特征对模型输出有显著影响，而几乎所有视觉特征均无显著影响。这些结果指出了 D-vlog 数据集的不足之处，也为未来抑郁检测模态选择和模型设计提供了启示。

**关键词** 抑郁检测；多模态；输入归因；可解释性

## 1 背景介绍

重度抑郁障碍 (major depressive disorder, MDD)，即俗称的抑郁症，是一种常见的心理障碍和精神疾病。在 2005 年的一项统计研究指出，重度抑郁障碍在世界范围内的患病率高达 16%，在一年内的发病率也有 6% [2]。这一数据有逐年递增的趋势：截至 2017 年，全球已有约 3.5 亿人患有重度抑郁障碍 [5]。在抑郁症高度流行的背景之下，如何准确有效地对它进行诊断，成为了心理健康和精神卫生领域内的一个重要问题。当下，对重度抑郁障碍的诊断依赖于 DSM-5 [1] 和 ICD-11 [3] 上的相关条目，以基于原型 (prototype-based) 的方式进行比对和判断；诊断过程中，需要明确分辨重度抑郁障碍和其他类型的心理障碍，同时也需指出一些特殊发病模式是否存在 [1]。显然，诊断过程需要临床医师具备扎实的专业知识和丰富的临床经验；因此，对重度抑郁障碍的诊断是一项高度复杂的任务。倘若能在进行专业问诊之前先用高效而自动化的方法对潜在抑郁人群进行筛查，临床医师的工作负担将大大降低，诊断效率将会提高；而且，自动化筛查方法能够指出人们患抑郁的潜在可能性，从而提醒人们及时求助，避免状况进一步恶化。综上所述，基于人工智能技术对抑郁检测具有重要的现实意义。

近年来，基于深度学习、数据驱动的抑郁检测受到广泛关注。研究者们收集了不同模态的数据，并将其组织称标准化的数据集，以便于抑郁检测模型的训练和评估；这些模态包括但不限于文本 [6]、视频音频 [4]、步态 [11]、脑电等生理信号 [7]、以及智能手机传感器信号 [12]。其中，D-vlog [14] 是一个包含视频音频数据的多模态数据集，抽取自 Youtube 上的若干日常生活记录视频，并标记为抑郁和非抑郁组。D-vlog 任务具有优秀的跨数据集泛化性能，在多模态抑郁检测领域内具有重要的研究价值；其数据来源于真实生



**Fig. 1** D-vlog 数据集概况 [14]。(a)vlog 类型视频的示意图。(b) 视频数据集收集过程。(c)D-vlog 数据集统计信息。(d) 数据集特征提取过程。

活而非临床场景，更加符合抑郁筛查的场景需求。因此，自 2022 年 D-vlog 诞生以来，不少团队基于该数据集进行抑郁检测模型设计和开发，并在性能上取得了明显进步 [9, 15-16]。鉴于 D-vlog 数据集的以上优点以及其易用性，我们也将聚焦于这一数据集展开深入探究。

本项目中，我们试图重新审视 D-vlog 数据集。我们将回答以下两个问题：什么样的模型可以在 D-vlog 数据集上获得更好的表现？D-vlog 数据集中的那些特征对抑郁检测起到了决定性的作用？首先，我们将在 D-vlog 数据集上训练 TMeanNet, Depression Detector [14] 和 TAMFN [16] 这三个模型，并通过准确率、召回率和 F1 分数来反映它们的性能差异。此后，我们量化不同的输入特征的重要性。我们对每一个模型使用积分梯度 (integrated gradients) 算法，计算出每一个输入特征对模型输出的贡献程度。我们发现，不论在哪类模型架构上，音频特征都是最重要的特征，而视频特征对模型输出并没有显著的影响。我们的结论一方面指出了 D-vlog 数据集的不足之处，另一方面也启发了今后抑郁检测的模态选择和算法设计。

## 2 D-vlog 数据集介绍

D-vlog 数据集由 Yoon 等人于 2022 年发布 [14]，共包含 961 个 Youtube 上的日常生活记录视频 (vlog)。数据集中的所有视频均于 2020 年 1 月 1 日至 2021 年 1 月 31 日期间发布。为了保证数据质量，研究者使

用“depression daily vlog”等词汇来搜索抑郁视频，使用“talking vlog”等词汇来搜索非抑郁视频，并聘用了 4 位学生志愿者来对视频类型进行筛选（仅保留 vlog 类型视频），以及对视频进行标注（分为抑郁组或非抑郁组）。图 1(a)展示了典型的 vlog 视频的样貌，而图 1(b)展示了视频数据收集的过程。最终收集到的视频中，抑郁组视频数量略多于非抑郁组，而主人公为女性的视频数量明显多于主人公为男性的视频数量（见图 1(c))。

为了保护视频主人公隐私，研究者对视频进行了视频音频特征提取，而并没有直接公开原始视频数据。特征提取的过程如图 1(d)所示。首先，视频被分割成若干片段，每个片段的长度为 1 秒。在每个片段内，使用 OpenSmile 工具包提取出 25 个音频特征，使用 dlib 工具包提取出 68 个面部关键点作为 136 维视觉特征。于是，视频  $i$  对应的数据为  $x_i \in \mathbb{R}^{T_i \times 161}$ ，其中  $T_i$  是视频的长度。需要注意的是，这些特征均不包含任何语义信息，而仅仅是对视频音频的低层次表示；我们无法得知视频主人公话语和行为的具体内容。

我们选择 D-vlog 数据集作为探究对象，是出于以下原因：

- (1) D-vlog 中的视频来源于日常生活场景，而非临床或实验室中的场景。在这样的场景下，提取出的特征也会更加贴近主人公的真实状态，更能反应主人公的真实心理状态。这也和抑郁检测的“在日常生活中预先筛查”的目标相契合。
- (2) D-vlog 具有优秀的跨数据集泛化性能。在 D-vlog 任务上训练的模型能够在 DAIC-WOZ 任务上取得优异的表现，甚至优于直接在 DAIC-WOZ 上训练得到的模型 [14]。
- (3) D-vlog 数据集涉及较多个体。而临床数据集通常涉及的个体较少，不一定具有代表性。
- (4) D-vlog 数据集规模较小，易于使用。音频特征一共 112MB，视频特征一共 597MB。这使得我们可以在较短的时间内完成模型训练和分析。

### 3 D-vlog 抑郁检测

#### 3.1 任务定义

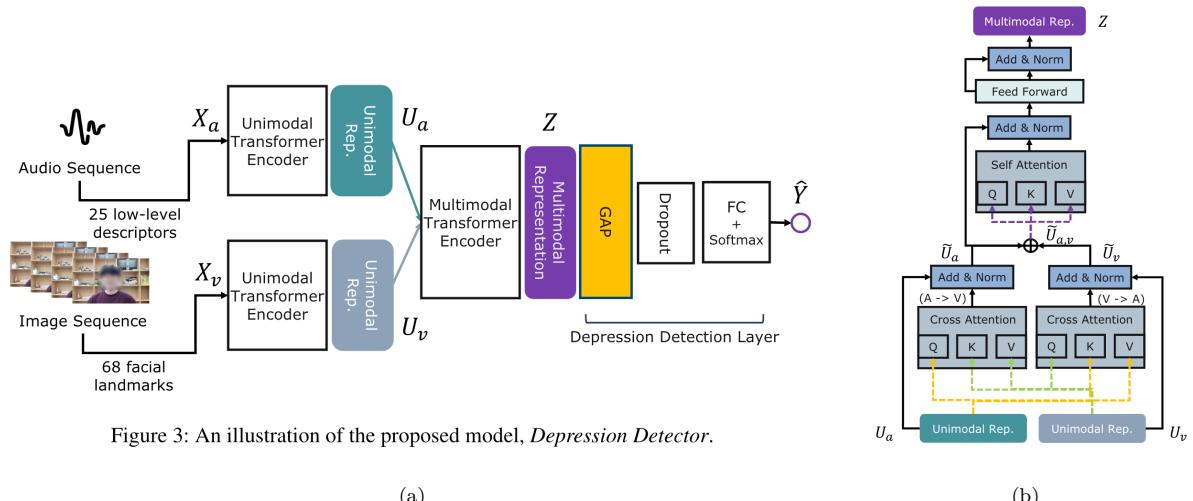
D-vlog 抑郁检测任务包含两个音频和视频模态  $\phi = \{\mathcal{A}, \mathcal{V}\}$ 。每个样本  $x_i$  是长度为  $T_i$  的序列，不同样本的长度可能不相同。样本  $x_i$  的第  $t$  帧  $x_i[t]$  是一个特征向量，维数为  $M = M_{\mathcal{A}} + M_{\mathcal{V}} = 25 + 136 = 161$ 。任务目标是训练一个二分类模型  $f$ ，使得对于任意样本  $x_i$ ，模型输出  $f(x_i) \in \{0, 1\}$  能够预测该样本属于抑郁组还是非抑郁组。我们假设 0 表示非抑郁，1 表示抑郁。

#### 3.2 模型介绍

我们采用三种模型来完成 D-vlog 抑郁检测任务：TMeanNet、Depression Detector [14] 和 TAMFN [16]。其中，TMeanNet 是我们自行提出的基线模型；而后两者虽然由其他团队在先前工作中提出，但并没有开源代码。因此，我们**自行搭建或复现**了这三个模型。

**TMeanNet** 是一种忽略局部时序信息的简单模型。首先，模型沿着时间维度对着输入样本  $x_i$  进行全局平均池化（global average pooling），得到时间无关特征向量  $z_i \in \mathbb{R}^M$ 。随后，使用含 3 个隐藏层的 MLP 对  $z_i$  进行二分类，隐藏层大小均为 512。TMeanNet 融合了不同模态，但却等效地对待每一个时间步，是我们的基线模型。

**Depression Detector** 由 D-vlog 的开发团队提出 [14]，如图 2(a)和图 2(b)所示。它使用 Transformer [13] 来处理单模态的时序信息，得到单模态特征序列  $U_{\mathcal{A}}$  和  $U_{\mathcal{V}}$ 。在交叉注意力机制作用下，两个单模态特

Figure 3: An illustration of the proposed model, *Depression Detector*.

(a)

(b)

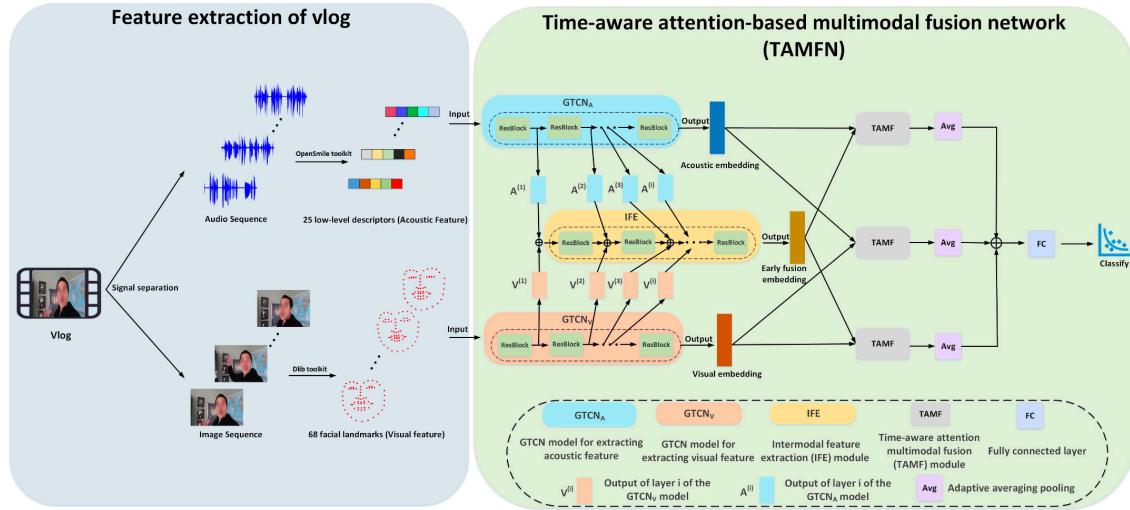
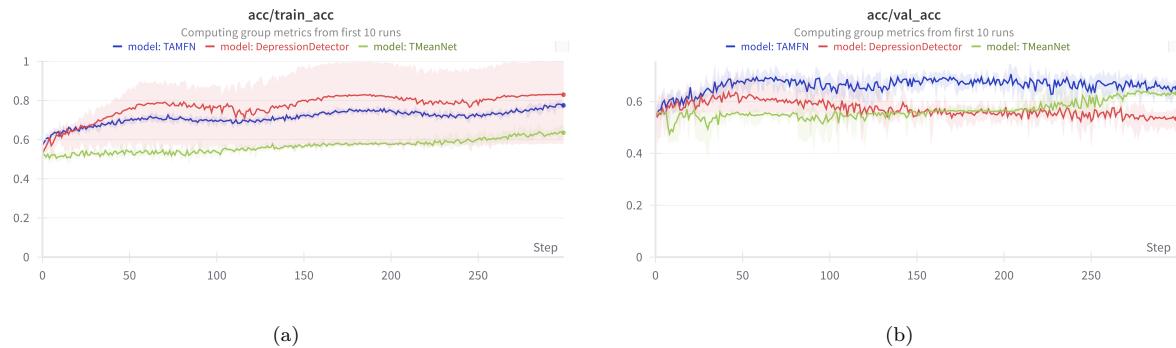


Fig. 2. The detailed structure of TAMFN.

(c)

**Fig. 2** D-vlog 抑郁检测模型。(a)Depression Detector 模型架构 [14]。(b)Depression Detector 的多模态 Transformer 结构示意图 [14]。(c)TAMFN 模型架构 [16]。



**Fig. 3** 不同模型在 D-vlog 抑郁检测任务上的训练曲线。曲线是 5 次实验的平均结果，阴影是标准差。(a) 训练正确率。(b) 验证正确率。

	Precision	Recall	F1	Accuracy
TMeanNet	0.6352	0.7902	0.7035	0.6151
Depression Detector	<b>0.6900</b>	0.7512	0.7172	0.6575
TAMFN	0.6889	<b>0.7935</b>	<b>0.7360</b>	<b>0.6708</b>

**Table 1** 不同模型在 D-vlog 任务上的测试结果。所有结果均是五次实验的平均数值。

特征序列相互融合，得到融合后的特征  $Z$ 。对  $Z$  的序列维度进行全局平均池化后，通过全连接层完成分类。显然，这一模型兼顾了局部时序信息和多模态融合。

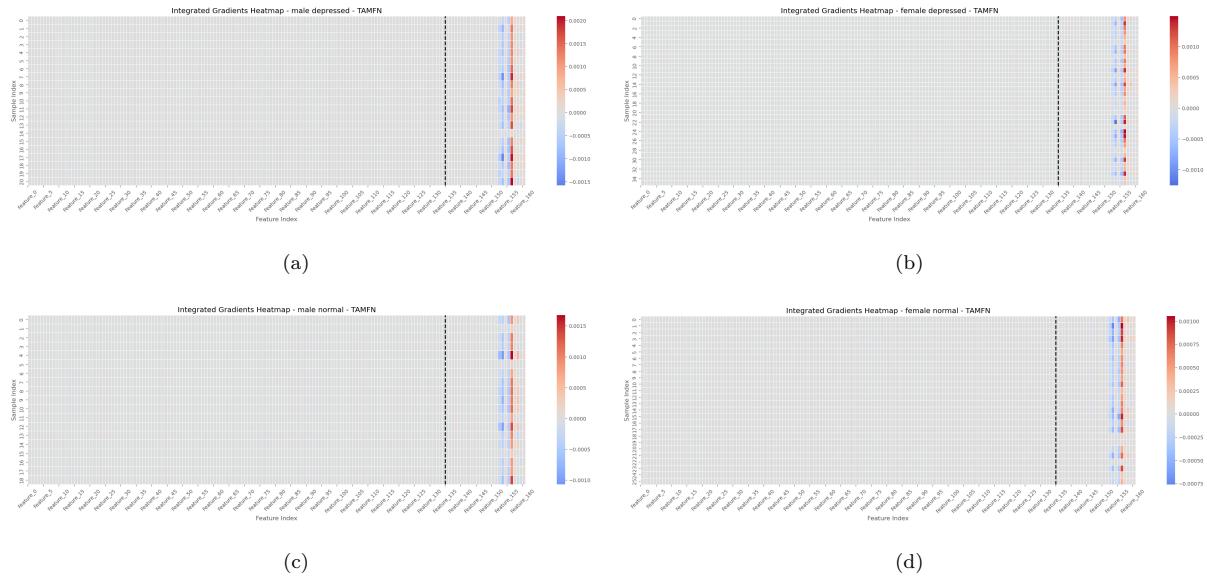
**TAMFN** 由 Zhou 等人于 2023 年提出 [16]，如图 2(c)所示。它改进了传统的残差一维卷积模块，在原先基础上添加一个新的基于时间维度平均池化的分支来加工全局时序信息，得到更加合理的单模态特征表示。它将单模态特征处理的中间结果取出，提前进行多模态融合，得到早期的多模态特征表示。随后，他将视觉特征、听觉特征和早期多模态特征两两组合，用基于时序注意力的模块来完成跨模态跨时间的融合。最后的结果被累加起来，送入全连接层完成分类。这一模型在兼顾局部和全局时序信息，并在早期和后期均进行了多模态融合，是一种较为复杂的模型。

### 3.3 实验结果

我们对模型输出的 logit 施加 sigmoid 函数得到预测概率 [8]，通过最小化交叉熵损失来训练模型。我们使用 Adam 优化器，学习率为  $10^{-4}$ ，采用 cosine 退火学习率调整策略，训练 300 个 epoch。训练过程中，在验证集上正确率最高的模型版本被选为最终模型版本，用于测试集上的评估。

三种模型的训练曲线如图 3所示。TMeanNet 作为最简单的实现方案，无法捕捉复杂的时序依赖关系，在训练集上的正确率提升极慢；而 TAMFN 和 Depression Detector 的训练正确率都能随着训练进行而提升，并最终稳定在 0.8 上下。然而，从验证集上正确率变化曲线可以看出，Depression Detector 出现了过拟合现象，后期的验证正确率甚至下降到比 TMeanNet 更低的数值；而 TAMFN 的过拟合并不明显。以上现象说明，Depression Detector 和 TAMFN 都能够在 D-vlog 数据集上学习到有效的特征表示，而 TAMFN 的泛化能力更强。

表 1展示了三种模型在测试集上的性能。TAMFN 在召回率 (recall)、F1 分数和正确率 (accuracy) 上均



**Fig. 4** TAMFN 模型对 D-vlog 数据集输入归因的热力图。每张图的横坐标是特征序号，位于黑色虚线左边的是视觉特征，右边是听觉特征；纵坐标是样本序号。红色代表正贡献度，蓝色代表负贡献度，灰色则表示贡献不明显。(a) 男性抑郁组；(b) 女性抑郁组；(c) 男性非抑郁组；(d) 女性非抑郁组。

优于 TMeanNet 和 Depression Detector，而 Depression Detector 拥有最高的精确率 (precision)。这些结果进一步反映了 TAMFN 的优越性。值得注意的是，我们得到的结果要明显优于原论文中的结果 (Depression Detector:  $F1=0.635$ ; TAMFN:  $F1=0.6582$ ) [14, 16]。这一方面是因为我们使用了更加强大的 cosine 退火学习率调整策略；另一方面，我们选用了验证集上表现最好的模型来进行测试，从而避免了过拟合现象的影响 (原文则没有报告具体的测试流程)。

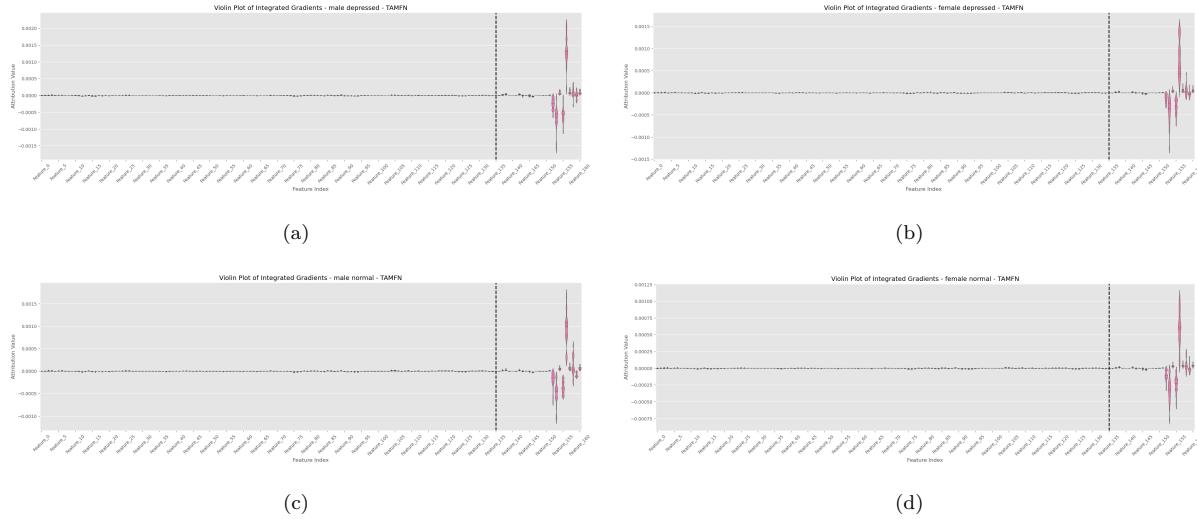
总的来看，时序依赖建模和早期特征融合是提升 D-vlog 抑郁检测性能的关键。至此，我们的第一个问题得到解决。

#### 4 D-vlog 特征重要性分析

Sun 等人 [9] 在 D-vlog 的音频特征数据集上使用预训练的情感分析模型提取情感特征，随后利用图神经网络结合情感特征来进行抑郁检测，得到了极其优异的性能：测试集上的正确率为 93.91%，准确率 91.9%，召回率 98.5%，F1 分数 95.1%，远高于其他工作中报告的结果 [14, 16]，也高于我们得到的结果 (见表 1)。这种优秀的结果表明了其所用方法的有效性；但另一方面，仅仅使用音频数据就能达到这般性能，是否说明了 D-vlog 数据集中的视频特征并不重要呢？

为了探究 D-vlog 数据集中哪些特征对于抑郁检测而言最为重要，我们建立如下的输入归因问题：给定一个训练好的模型  $f$  以及一个输入样本  $x$ ，样本的各个特征  $x_i$  对于模型输出  $f(x)$  的贡献度有多大？为解决此问题，我们采用积分梯度法 (integrated gradients)，沿着从原点  $\mathbf{0}$  到样本点  $x$  的路径对模型输出关于输入的梯度进行累积，以这个数值的分量  $G_i$  作为输入分量  $x_i$  对模型输出的贡献度 [10]。而对于 D-vlog 这样的序列数据，我们将输入的累计梯度沿着时间维度求和，即可得到某个特征对模型输出的总贡献度。积分梯度的正负表示特征对模型输出的影响是正向还是负向，而其绝对值则刻画了特征对模型输出的影响程度。

我们将样本分成男性抑郁组、女性抑郁组、男性非抑郁组、女性非抑郁组这四个类别，并以 TAMFN



**Fig. 5** TAMFN 模型对 D-vlog 数据集输入特征的归因值分布小提琴图。每张图的横坐标是特征序号，位于黑色虚线左边的是视觉特征，右边是听觉特征；纵坐标是积分梯度值。(a) 男性抑郁组；(b) 女性抑郁组；(c) 男性非抑郁组；(d) 女性非抑郁组。

模型为例，展示积分梯度分析的结果。从图 4 容易看出，只有少量的听觉特征对模型输出存在显著正向影响（红色区域），另有少量听觉特征对输出具有负向影响（蓝色区域）。大多数的听觉特征和几乎所有的视觉特征都对模型输出没有显著影响。将热力图转换成图 5 所示的小提琴图后，我们能更加直观地感受到只有少量特征在模型分类过程中起作用。这一现象在 TMeanNet 和 Depression Detector 上同样存在，见附录中的图 6、7、8、9。

自此，我们得出结论：在 D-vlog 数据集上，只有少量听觉特征对抑郁检测起到关键作用，而多数听觉特征和几乎所有的视觉特征均无法对抑郁检测起到实质性的帮助。这一结论具有跨性别、跨组别、跨模型的普适性。

## 5 结论与讨论

本项目中，我们使用三种模型（TMeanNet、Depression Detector 和 TAMFN）在 D-vlog 数据集上进行抑郁检测，并对模型的性能进行了评估。我们发现，时序依赖建模和早期特征融合是提升 D-vlog 抑郁检测性能的关键。随后，我们通过输入归因分析来衡量不同输入特征对模型输出的贡献程度，并发现不论使用哪类模型、样本属于哪种性别，都只有少量音频特征对 D-vlog 抑郁检测起关键作用，而视觉特征几乎对结果没有影响。

我们对 D-vlog 特征的重新审视揭露了该数据集的不足之处。根据我们的结果，完全抛弃 D-vlog 数据集中的视觉特征也不会造成模型表现的明显退化；这一结果和 Sun 等人的工作 [9] 相契合。我们建议数据集的开发者针对这一问题进行改进，增加视觉特征的丰富性，以提升数据集的总体质量。

另一方面，我们的结果也启发了今后抑郁检测的模态选择和算法设计。倘若我们的结果可以推广到更大的范围，那我们便可以仅基于音频特征来完成高可信度的抑郁检测，而完全不需要采集视频特征。这将大幅降低抑郁检测模型的开发难度，并让检测系统的部署应用更加方便。

## 代码与数据

本项目的代码开源于<https://github.com/AllenYolk/depression-detection>。

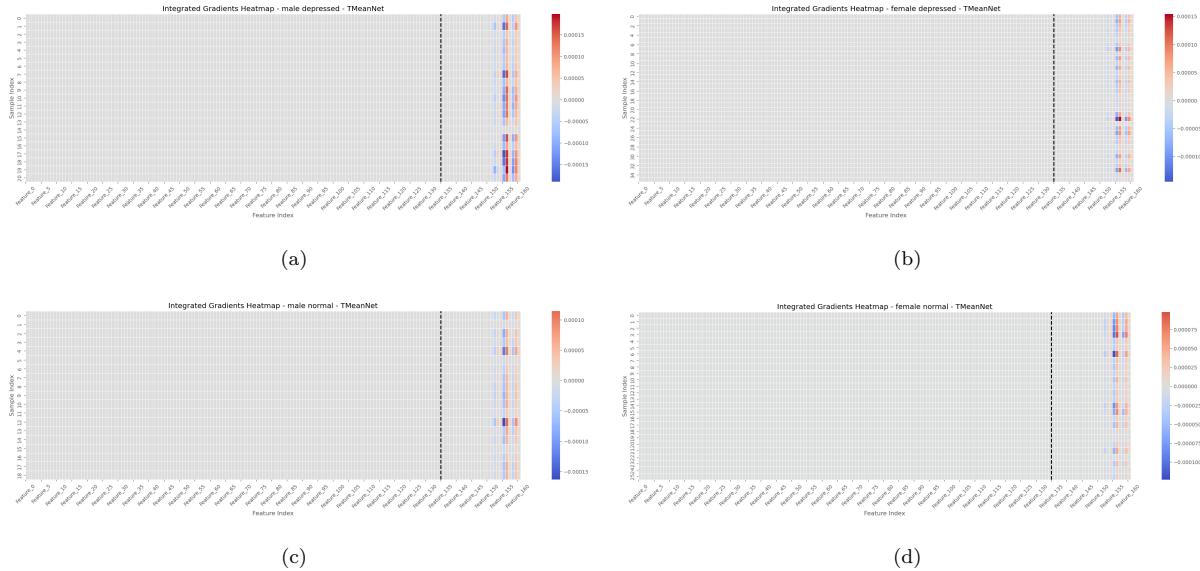
本项目使用的 D-vlog 数据集可以在<https://sites.google.com/view/jeewoo-yoon/dataset> 处获取（需向数据集开发者申请）。

## 参 考 文 献

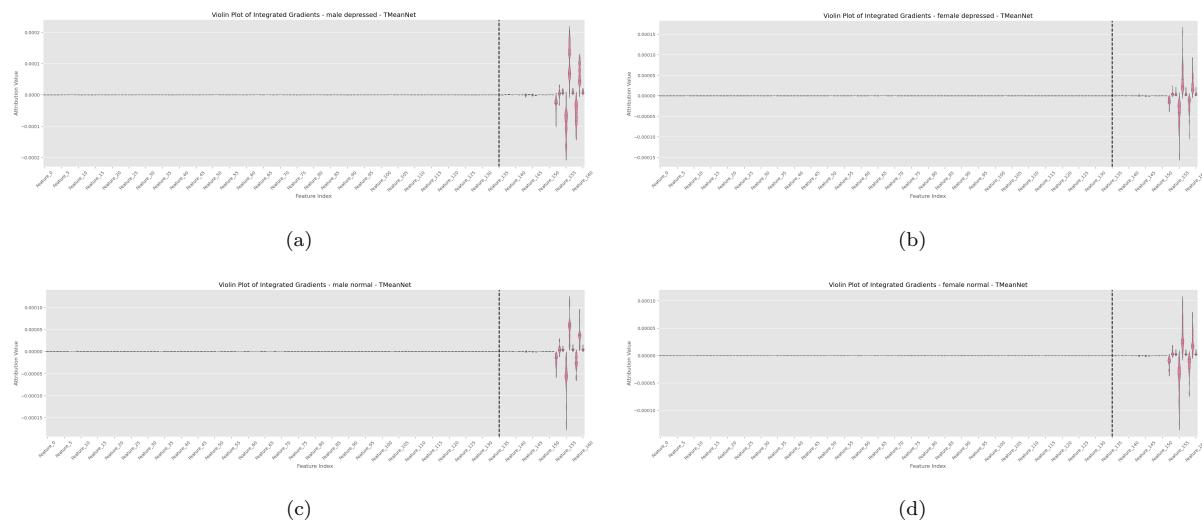
- [1] *Diagnostic and Statistical Manual of Mental Disorders: DSM-5™, 5th Ed.* Diagnostic and Statistical Manual of Mental Disorders: DSM-5™, 5th Ed. American Psychiatric Publishing, Inc., Arlington, VA, US, 2013.
- [2] Deborah S Hasin, Renee D Goodwin, Frederick S Stinson, and Bridget F Grant. Epidemiology of major depressive disorder: results from the national epidemiologic survey on alcoholism and related conditions. *Archives of general psychiatry*, 62(10):1097–1106, 2005.
- [3] World Health Organization. *International Classification of Diseases, 11th Edition (ICD-11)*. World Health Organization, 2022.
- [4] Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, Siyang Song, Shuo Liu, Ziping Zhao, Adria Mallol-Ragolta, Zhao Ren, Mohammad Soleymani, and Maja Pantic. AVEC 2019 Workshop and Challenge: State-of-Mind, Detecting Depression with AI, and Cross-Cultural Affect Recognition, July 2019.
- [5] Guangyao Shen, Jia Jia, Liqiang Nie, Fuli Feng, Cunjun Zhang, Tianrui Hu, Tat-Seng Chua, and Wenwu Zhu. Depression Detection via Harvesting Social Media: A Multimodal Dictionary Learning Solution. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 3838–3844, Melbourne, Australia, August 2017. International Joint Conferences on Artificial Intelligence Organization.
- [6] Guangyao Shen, Jia Jia, Liqiang Nie, Fuli Feng, Cunjun Zhang, Tianrui Hu, Tat-Seng Chua, and Wenwu Zhu. Depression Detection via Harvesting Social Media: A Multimodal Dictionary Learning Solution. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 3838–3844, Melbourne, Australia, August 2017. International Joint Conferences on Artificial Intelligence Organization.
- [7] Jian Shen, Xiaowei Zhang, Gang Wang, Zhijie Ding, and Bin Hu. An Improved Empirical Mode Decomposition of Electroencephalogram Signals for Depression Detection. *IEEE Transactions on Affective Computing*, 13(1):262–271, January 2022.
- [8] Leslie N. Smith. Cyclical Learning Rates for Training Neural Networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 464–472, March 2017.
- [9] Chenjian Sun, Min Jiang, Linlin Gao, Yu Xin, and Yihong Dong. A novel study for depression detecting using audio signals based on graph neural network. *Biomedical Signal Processing and Control*, 88:105675, February 2024.
- [10] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks, June 2017.
- [11] T. Wang, C. Li, C. Wu, C. Zhao, J. Sun, H. Peng, X. Hu, and B. Hu. A Gait Assessment Framework for Depression Detection Using Kinect Sensors. *IEEE Sensors Journal*, 21(3):3260–3270, 1 Feb. 2021.

- 
- [12] Ravi Prasad Thati, Abhishek Singh Dhadwal, Praveen Kumar, and Sainaba P. A novel multi-modal depression detection approach based on mobile crowd sensing and task-based mechanisms. *Multimedia Tools and Applications*, 82(4):4787–4820, February 2023.
  - [13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
  - [14] Jeewoo Yoon, Chaewon Kang, Seungbae Kim, and Jinyoung Han. D-vlog: Multimodal Vlog Dataset for Depression Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11):12226–12234, June 2022.
  - [15] Wenbo Zheng, Lan Yan, and Fei-Yue Wang. Two Birds With One Stone: Knowledge-Embedded Temporal Convolutional Transformer for Depression Detection and Emotion Recognition. *IEEE Transactions on Affective Computing*, 14(4):2595–2613, October 2023.
  - [16] Li Zhou, Zhenyu Liu, Zixuan Shangguan, Xiaoyan Yuan, Yutong Li, and Bin Hu. TAMFN: Time-Aware Attention Multimodal Fusion Network for Depression Detection. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:669–679, 2023.

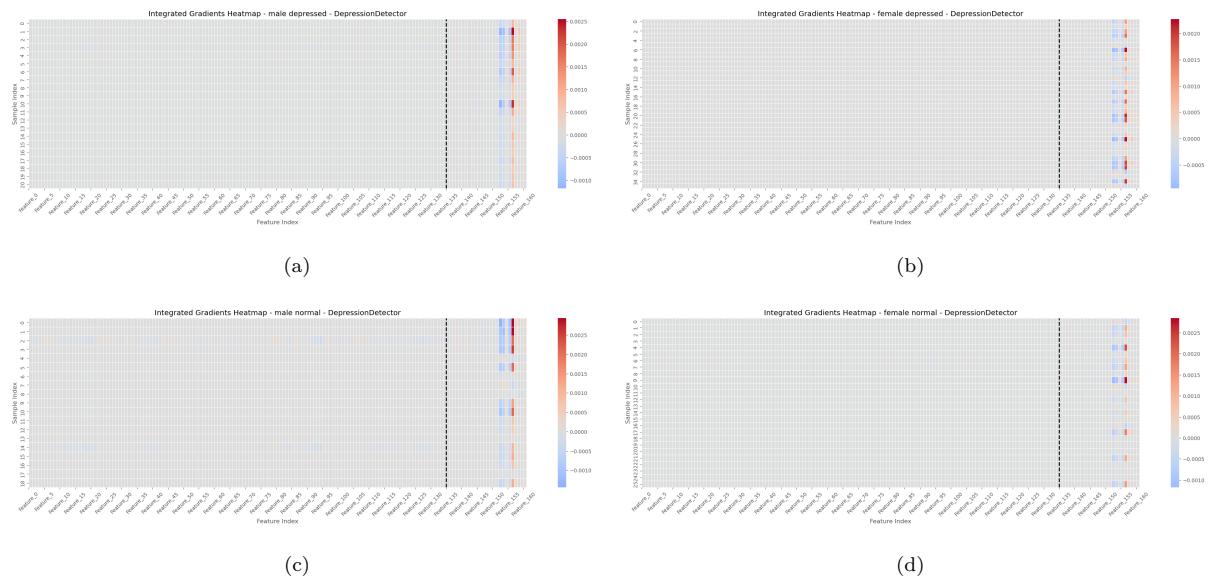
## 附录



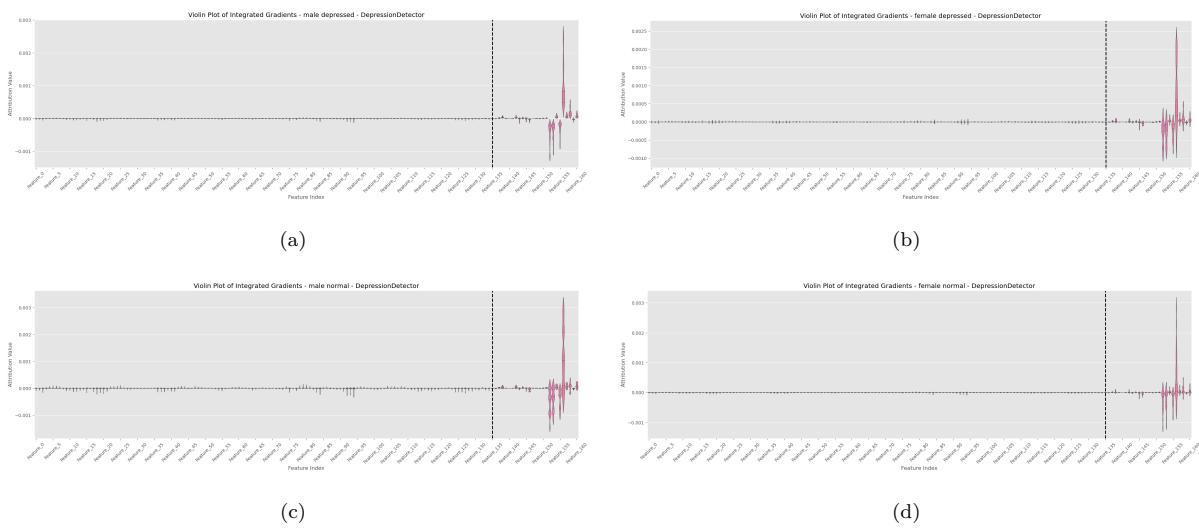
**Fig. 6** TMeanNet 模型对 D-vlog 数据集输入归因的热力图。每张图的横坐标是特征序号，位于黑色虚线左边的是视觉特征，右边是听觉特征；纵坐标是样本序号。红色代表正贡献度，蓝色代表负贡献度，灰色则表示贡献不明显。(a) 男性抑郁组；(b) 女性抑郁组；(c) 男性非抑郁组；(d) 女性非抑郁组。



**Fig. 7** TMeanNet 模型对 D-vlog 数据集输入特征的归因值分布小提琴图。每张图的横坐标是特征序号，位于黑色虚线左边的是视觉特征，右边是听觉特征；纵坐标是积分梯度值。(a) 男性抑郁组；(b) 女性抑郁组；(c) 男性非抑郁组；(d) 女性非抑郁组。



**Fig. 8** Depression Detector 模型对 D-vlog 数据集输入归因的热力图。每张图的横坐标是特征序号，位于黑色虚线左边的是视觉特征，右边是听觉特征；纵坐标是样本序号。红色代表正贡献度，蓝色代表负贡献度，灰色则表示贡献不明显。(a) 男性抑郁组；(b) 女性抑郁组；(c) 男性非抑郁组；(d) 女性非抑郁组。



**Fig. 9** Depression Detector 模型对 D-vlog 数据集输入特征的归因值分布小提琴图。每张图的横坐标是特征序号，位于黑色虚线左边的是视觉特征，右边是听觉特征；纵坐标是积分梯度值。(a) 男性抑郁组；(b) 女性抑郁组；(c) 男性非抑郁组；(d) 女性非抑郁组。