

文章编号:1671-5896(2026)04-1015-12

抑郁症识别中的动态加权多模态融合框架

王天乐, 陈文实

(辽宁工业大学 电子与信息工程学院, 辽宁 锦州 121001)

摘要: 针对现有抑郁症单模态识别方法(如 EEG(Electroencephalogram)、面部表情、文本分析)易受环境干扰且识别率有限的问题,提出了 MFF-DR-DWA(Multimodal Fusion Framework for Depression Recognition with Dynamic Weight Adjustment)多模态融合框架。其通过动态整合脑电信号、语言特征和面部表情3种异质数据源,并采用自适应加权算法优化各模态贡献度。实验结果表明,在标准测试数据集上,该模型的准确率为99.09%、 F_1 值为99.12%和AUC(Area Under Curve)值为99.97%,其性能显著优于传统单模态分析方法和固定权重融合方案。该多模态融合策略有效提升了抑郁症识别的准确性和鲁棒性。

关键词: 多模态; 动态加权; 多模态融合; 抑郁症识别; 注意力机制

中图分类号: TP181

文献标志码: A

Dynamic Weighted Multimodal Fusion Framework in Depression Identification

WANG Tianle, CHEN Wenshi

(School of Electronics and Information Engineering, Liaoning University of Technology, Jinzhou 121001, China)

Abstract: To address the limitations of existing unimodal recognition methods for depression (e. g., EEG (Electroencephalogram), facial expressions, text analysis), which are susceptible to environmental noise and exhibit limited accuracy, an innovative MFF-DR-DWA (Multimodal Fusion Framework is proposed for Depression Recognition with Dynamic Weight Adjustment). The framework dynamically integrates three heterogeneous data modalities-electroencephalography (EEG), linguistic features, and facial expressions-while employing an adaptive weighting algorithm to optimize the contribution of each modality. Experimental results on benchmark datasets demonstrate that the proposed model achieves an accuracy of 99.09%, an F_1 -score of 99.12%, and an AUC (Area Under Curve) of 99.97%, significantly outperforming conventional unimodal approaches and static fusion methods. This multimodal fusion strategy effectively enhances the accuracy and robustness of depression recognition.

Key words: multimodal; dynamic weighting; multimodal fusion; depression recognition; attention mechanism

0 引言

抑郁症是一种常见的精神疾病,影响着全球数亿人口。早期识别对有效干预和治疗至关重要。然而,由于个体症状的多样性和主观报告的不稳定性,传统的抑郁症诊断方法存在一定的局限性。近年来,借助人工智能与多模态数据分析技术^[1],使自动化抑郁症检测成为研究热点。

当前的研究主要集中于单一模态的抑郁症检测,例如基于文本分析的自然语言处理(NLP: Natural Language Processing)方法^[2],以及基于人脸表情分析的计算机视觉方法^[3]。虽然其在一定程度上提高了

收稿日期: 2025-06-11

作者简介: 王天乐(2001—),男,山东济宁人,辽宁工业大学硕士研究生,主要从事情感分析研究,(Tel)86-18654776808 (E-mail) 2818258247@qq.com; **通讯作者:** 陈文实(1971—),男,辽宁工业大学副教授,硕士生导师,主要从事自然语言处理研究,(Tel)86-13804169488 (E-mail) chenwenshi@lnut.edu.cn。

检测准确率,但由于抑郁症的复杂性,仅依赖单一模态容易受到个体差异、环境及数据质量等因素的影响,导致鲁棒性不足。因此,多模态融合方法逐渐受到人们关注。

研究表明,文本、面部表情和脑电信号在抑郁症检测中具有重要作用。文本模态可以捕捉患者语言中的情感信息,如负面词汇使用频率、句法结构等^[4]。人脸模态能揭示面部表情及微表情变化,与抑郁症状的外在表现密切相关^[5]。而脑电信号作为反映大脑活动的生理指标,可提供客观的神经生理学特征^[6],有助于弥补主观行为数据的不足。

基于上述,笔者提出了一种融合文本、人脸和脑电3种模态的抑郁症检测框架,并在决策级融合阶段采用注意力机制对不同模态的权重进行动态分配,以提高检测精度和鲁棒性。

1 相关研究

1.1 单模态抑郁症识别研究现状

近年来,基于文本单模态的抑郁症检测研究取得了显著进展,人们通过不同的模型和方法从社交媒体文本中提取抑郁症相关特征。Farruque 等^[7]提出的 TUD(Temporal User-Level Clinical Depression)模型,其使用了 BiLSTM^[8](Bidirectional Long Short-Term Memory)和注意力机制对用户 in 社交媒体上的时间序列数据进行深度建模,同时整合临床抑郁症状特征和用户行为模式,通过抑郁症症状检测(DSD: Depressive Symptom Detection)模型计算抑郁症评分,最终在多个数据集上验证了其有效性。Wu 等^[9]针对传统方法忽视个体差异问题,提出了个性化信息嵌入(PIE: Personalized Information Embedded)模型,其通过深度学习技术从用户历史文本中提取个性化特征表示,有效缩小了通用临床特征与个体真实体验之间的语义鸿沟。上述研究为基于文本的抑郁症检测提供了丰富的理论支持和实践参考。

近年来,基于人脸表情的抑郁症识别研究取得了显著进展,人们通过改进深度学习模型和特征提取方法,不断提升识别性能。Ding 等^[10]提出了一种基于视频时序面部关键点角度特征方法,并结合轻量级 GhostNet^[11]进行建模,在 DAIC-WOZ(Distress Analysis Interview Corpus)数据集上 F_1 值为 0.80,显著提升了单模态抑郁症检测效果。Li 等^[12]开发了 SFTNet(Spatial-Feature Temporal Network)框架,该网络包含 STNet(Spatial-Temporal Network)和 FTNet(Feature-Temporal Network),分别用于局部时空特征提取和全局时间建模,通过在情绪刺激实验和真实医患对话数据集上的验证,证明了其在抑郁症识别任务中的优越性能。

对 EEG(Electroencephalogram)抑郁症识别,Ahmed 等^[13]提出堆叠集成学习模型,并结合多分类器处理时域统计特征,取得 98.01% 的准确率,显著优于传统方法。Xu 等^[14]采用 EEG 多尺度熵特征的实验表明,KNN(K Nearest Neighbor)在尺度 9 时准确率达 96.42% ($F_1 = 0.956$),RBF-SVM(Radial Basis Function-Support Vector Machine)在尺度 6~8 时敏感性最高(94.07%),尺度 2 时特异性最佳(98.64%),验证了多尺度熵在 EEG 信号分析中的有效性,并显著提升了抑郁症识别的准确性。

虽然部分单一模态在特定任务中已取得较高的识别准确率,但受限于其信息维度的单一性,难以全面反映抑郁症的复杂病理特征与个体差异。为此,当前越来越多的研究开始探索多模态融合方法,并尝试将脑电(EEG)、文本语言特征以及面部表情等异构信息相结合,以更全面刻画抑郁症的多维表现,进而提升模型的识别能力与鲁棒性。

1.2 多模态抑郁症识别研究现状

对多模态抑郁症识别的研究,人们通过结合多种数据源,如文本、语音、面部表情和脑电图(EEG)等,构建了多种模型以提高抑郁症检测的准确性和可靠性。

He 等^[15]提出了 DPDNet(De Pression Detection Network)的深度神经网络模型,用于多模态抑郁症检测。由于其融合了文本、音频和视觉特征,适用于临床和社交媒体两种应用场景。DPDNet 由单模态、多模态编码器和检测模块组成。同时还提出了 DPD-ENet 模型,将脑电图信号和语音数据相结合,用于抑郁症检测。Cheng 等^[16]提出了一种基于多模态融合的抑郁症检测模型,分别使用 BERT-TextCNN(Bidirectional Encoder Representations from Transformers-Text Convolutional Neural Network)和 CNN-LSTM(Convolutional Neural Network-Long Short-Term Memory Network)处理文本和 EEG 数据,并通过加权融合策

略结合两种模态, 显著提升了检测性能。Mohammad 等^[17]提出了一种基于多模态深度学习的抑郁症诊断方法 MDD (Multimodal Depression Diagnosis), 其结合文本和语音数据, 并使用 T5 (Text-To-Text Transfer Transformer) 模型^[18]提取文本特征, WaveNet 模型提取语音特征, 通过典型相关分析 (CCA; Canonical Correlation Analysis) 将两种模态的特征进行融合, 最后通过神经网络进行抑郁症检测。该模型在 DAIC-WOZ 数据集上达到了 92.75% 的准确率、92.05% 的精确率和 92.22% 的召回率。赵小明等^[19]提出了一种基于多模态特征增强网络的抑郁症检测方法, 通过融合视频、音频和 rPPG (Remote Photoplethysmography) 信号, 利用模态间 Transformer、模态内 Transformer 和多头自注意力机制增强特征表示, 最终在 AVEC2013 数据集上的平均绝对误差为 7.07, 显著优于传统单模态方法。

由于目前抑郁症识别研究主要基于文本和人脸表情等外部特征, 虽能提供情感线索, 但易受主观因素干扰。而相比较, 脑电信号能更直接反映神经活动状态。当前多模态融合方法存在 3 个关键局限: 首先, 文本、表情和脑电信号在特征空间分布和时序特性上存在显著差异, 现有方法对特征异构性处理不足; 其次, 模态融合多采用静态权重分配, 未能根据样本特性动态调整各模态贡献度; 最后, 现有方法对跨模态的深层关联 (如文本情感与脑电节律的对应关系) 挖掘不足, 导致潜在互补信息未被充分利用。这些局限制约了模型在复杂场景下的识别性能。基于上述问题, 笔者提出 MFF-DR-DWA (Multimodal Fusion Framework for Depression Recognition with Dynamic Weight Adjustment) 多模态融合框架。

2 研究方法

该研究框架结构如图 1 所示。针对特征异构性问题, 采用 3 级特征对齐策略: 首先通过可学习投影矩阵将 EEG (CNN-GRU (Gated Recurrent Unit) 网络提取)、文本 (RoBERTa (Robustly Optimized BERT Approach) 编码) 和面部表情 (CBAM-ResNet (Convolutional Block Attention Module-Residual Network) 提取) 3 种异构特征映射到统一语义空间; 其次引入双重标准化处理解决特征尺度差异; 最后基于对比学习优化语义一致性。同时针对静态融合局限, 设计了动态门控权重机制, 通过样本自适应的注意力网络实时调整各模态贡献度。通过模态间交叉注意力层建立跨模态关联, 深度挖掘文本语义、面部微表情与脑电节律之间的潜在互补信息。

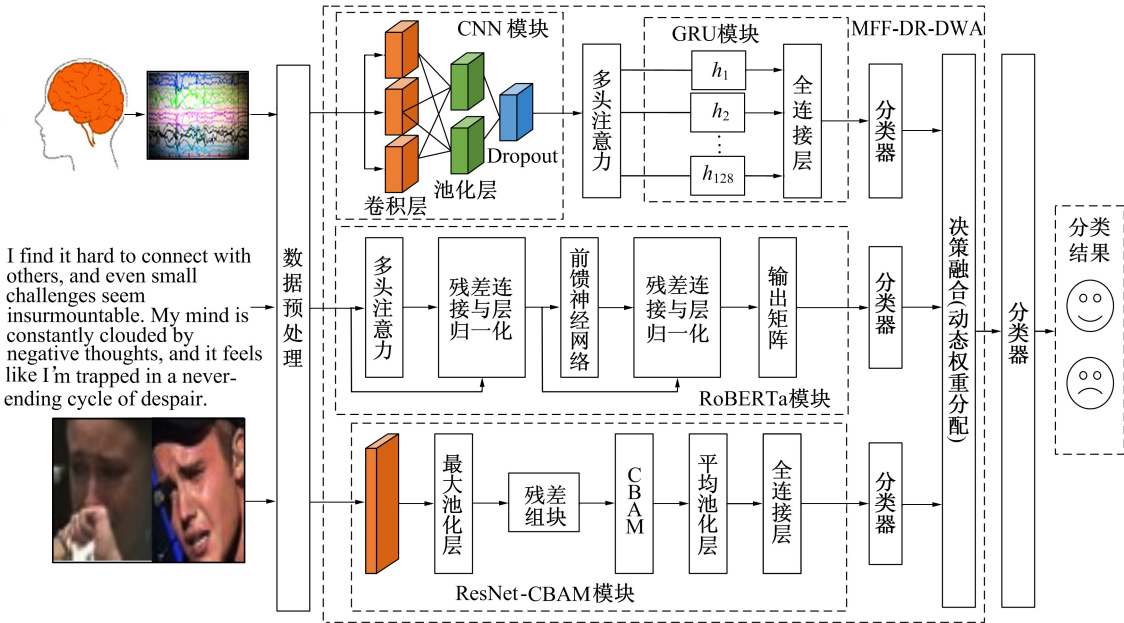


图 1 框架结构

Fig. 1 Framework structure diagram

2.1 基于脑电 (EEG) 的抑郁症检测方法

笔者使用 Mumtaz 等^[20]在 2017 年提供的公开数据集,其中包括重度抑郁患者 (MDD) 30 名 (平均年龄 40.3 ± 12.9 岁) 和健康对照组 (HC: Healthy Control) 28 名 (平均年龄 38.3 ± 15.6 岁)。所有参与者全部由马来西亚理科大学招募,并在医院门诊进行数据采集,患者和健康参与者均签署了参与同意书。根据 10-20 系统^[21],使用不同脑区的 19 个通道采集 EEG 信号,采样频率 256 Hz,电极在头皮上放置位置可分为 5 个区域,如图 2 所示,分别是额叶 (Fp1、Fp2、Fz、F3、F4、F7、F8)、颞叶 (T3、T4、T5、T6)、顶叶 (P3、P4、Pz)、枕叶 (O1、O2)、中央区 (C3、C4、Cz)。

首先对信号进行 1~40 Hz 的带通滤波处理,去除 EEG 信号中的高频噪声和低频噪声。然后,采用滑动窗口技术将信号划分为 1 s (256 采样点) 无重叠的子序列,以保留其时序特性。脑电信号具有时空动态特征,且存在复杂的非线性关系。为此,笔者采用了结合 CNN 与 RNN (Recurrent Neural Network) 的混合结构,其中 CNN 用于提取局部空间特征,RNN 用于捕捉时间序列依赖,进一步辅以注意力机制聚焦于关键时间段,从而实现对脑电信号的高效建模。

脑电数据的特征学习由 3 部分组成,模型结构如图 3 所示。

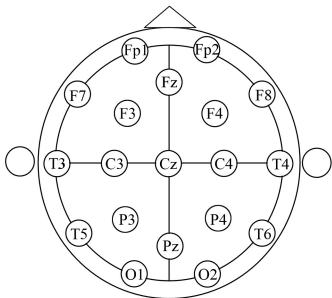


图 2 10-20 系统中电极位置

Fig. 2 Electrode positions in the 10-20 system

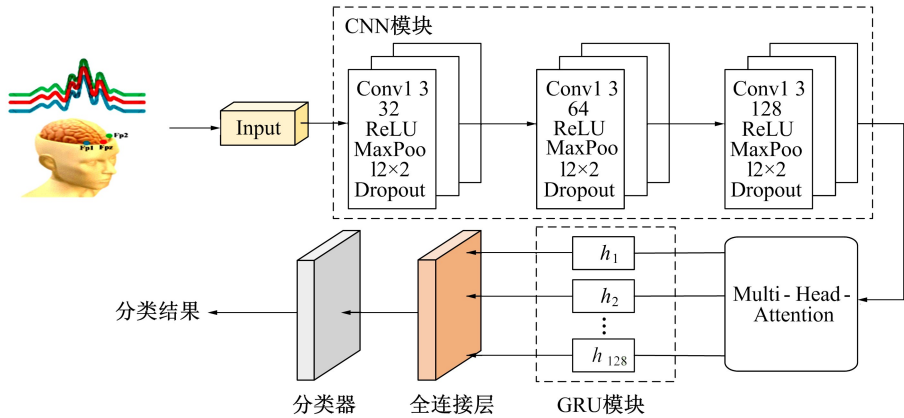


图 3 脑电结构

Fig. 3 EEG structure diagram

模型的第 1 部分是一个串行卷积结构,采用 3 层 1D 卷积结构 (通道数 32/64/128,核大小 3, padding=‘same’),每层后接 ReLU、最大池化和 Dropout。通过局部卷积核捕获时序特征,池化层压缩维度,Dropout 防止过拟合。第 2 部分为多头注意力模块,在卷积特征基础上引入多头注意力机制,通过并行计算多个注意力头,建立全局特征关联并动态加权关键信息,增强模型对长程依赖的建模能力。第 3 部分是 GRU 部分,在卷积和注意力机制提取特征后,特征被送入 GRU 网络进行时序建模。

对抑郁症识别任务,模型的最后部分是通过全连接层将 GRU 的输出转换为分类结果,最终通过分类器将输出转化为抑郁症类别的概率值,包括“健康”与“抑郁症”两个类别。分类器会将每个类别的输出映射到一个概率分布,保证各类别的概率之和为 1,用于确定输入样本是否属于抑郁症类别。

2.2 基于文本的抑郁症检测方法

文本模态涉及自然语言的上下文语义理解,笔者选用预训练语言模型 RoBERTa,其在多个自然语言处理任务中表现卓越,能捕获丰富的语义信息,有助于提取抑郁症相关的语言线索。

RoBERTa^[22]是在 2019 年基于 BERT^[23]的改进模型。与 BERT 模型相比,RoBERTa 模型采用了更大的数据训练量,其表明 RoBERTa 对文本上下文的理解比 BERT 更加全面,同时 RoBERTa 采用了动态

Mask 机制, 每次向模型输入一个新的序列, 都会生成一种新的遮盖方式, 从而增强了模型的泛化能力和鲁棒性。因此将使用 RoBERTa 作为文本数据提取和特征学习的模型。

2.3 基于人脸的抑郁症识别方法

残差网络(ResNet)^[24]旨在解决深层神经网络中梯度消失与性能退化问题。其核心结构为残差块, 包含卷积层、批归一化、ReLU 激活函数以及跳跃连接。传统网络直接学习输入到输出的映射, 而 ResNet 通过引入跳跃连接, 使输入可直接与输出相加, 有效缓解梯度问题并提升深层网络的训练稳定性与效率。

虽然 ResNet 通过残差连接缓解了梯度消失问题, 并提升了训练效率与稳定性, 但其在特征提取方面仍存在局限。传统 ResNet 未充分关注通道与空间维度特征的重要性, 处理面部图像时易引入冗余信息, 削弱对关键区域的聚焦能力, 并且其人脸表情分析方法在抑郁症识别任务中表现有限。这是由于其依赖于 HOG(Histogram of Oriented Gradients)、LBP(Local Binary Pattern)等手工特征, 难以有效表征与抑郁相关的微妙表情变化; 并且缺乏注意力机制, 无法实现对关键面部区域的动态加权, 降低了判别能力。同时, 浅层模型在建模复杂非线性关系方面能力不足, 限制了其泛化性。

针对上述问题, 笔者提出将 ResNet 与 CBAM^[25]相结合, 以提升面部特征的代表能力。ResNet 作为一种深层卷积神经网络, 能自动学习多层次的高阶语义特征, 尤其适用于识别与抑郁症相关的细微面部表情。通过残差结构, ResNet 不仅加深了网络层次, 还保持了良好的训练稳定性与收敛效率。

为进一步增强模型的注意力感知能力, 笔者在 ResNet-50 主干网络中引入了 CBAM 模块, 其是一种轻量级的卷积注意力机制, 包含通道和空间注意力模块两部分。通道注意力模块通过全局池化提取通道信息, 自适应评估各通道的重要性, 增强与抑郁表情高度相关的特征通道, 同时抑制冗余或噪声信息; 空间注意力模块则进一步分析特征图中的空间位置信息, 聚焦图像中的关键区域(如嘴角、眼神、眉间等), 提升模型对微表情变化的敏感性。

CBAM 同时具备样本自适应的动态注意力分配能力, 能根据不同输入自动调整注意力分布, 实现更细粒度的特征筛选与增强, 可有效抑制背景干扰, 增强模型的鲁棒性与泛化能力。其工作机制为, 首先对输入特征图 F 应用通道注意力, 获得增强后的特征图 F_1 ; 随后再施加空间注意力, 输出最终特征图 F_2 (见图 4)。集成 CBAM 后的整体网络结构如图 5 所示。

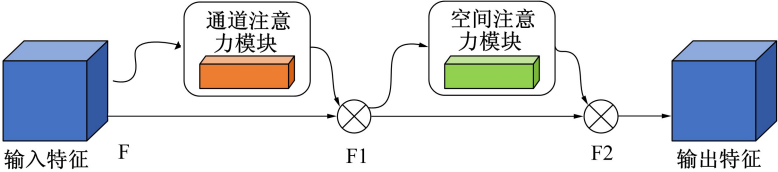


图 4 CBAM 原理
Fig. 4 CBAM schematic diagram

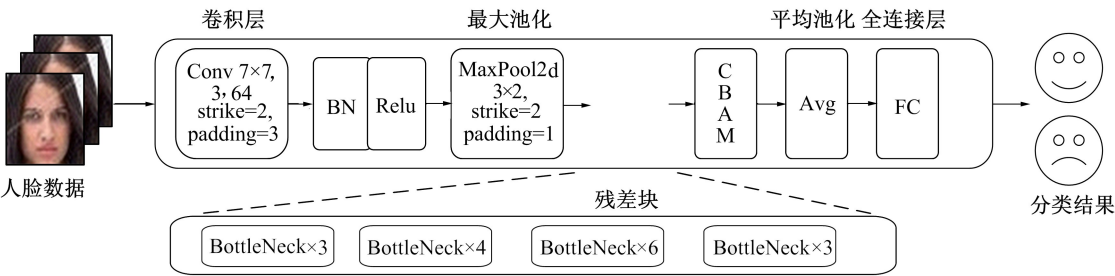


图 5 CBAM-ResNet-50 模型结构
Fig. 5 Structure diagram of CBAM ResNet-50 mode

2.4 基于决策融合的抑郁症识别方法

2.4.1 决策融合方法

笔者针对 EEG (Mumtaz)、文本 (Reddit) 与人脸 (MMAFEDB: Micro-Expression and Macro-Expression

Database) 3 个数据集异构模态的融合问题, 提出了一套多模态特征对齐与融合方案。特征对齐采用 3 级处理框架: 首先在维度对齐方面, 利用可学习的共享投影矩阵将各模态映射至统一的 128 维语义空间, 并施加正交正则化以保持结构稳定性; 其次对于统计对齐, 结合滑动窗与层归一化, 统一动态范围与特征尺度, 配合自适应池化对齐时序长度; 最后在语义对齐阶段, 引入基于对比学习的优化策略, 采用改进的 InfoNCE 损失 ($\tau=0.1$) 强化多模态语义一致性。

在决策层融合方面, 笔者采用注意力机制和动态权重分配算法整合基于不同特征提取网络 (CNN-GRU、RoBERTa 和 CBAM-ResNet) 的模型输出。该动态权重分配算法能根据各模型在验证集上的表现自动调整权重, 使性能更优的模型在最终决策中获得更大权重, 从而显著提升抑郁症检测的准确率。

2.4.2 动态权重分配算法

传统动态权重分配方法在多模态融合中存在粒度较粗、模态交互单向、缺乏样本级适应性问题, 主要集中于模态级静态权重分配, 忽略了模态内特征重要性差异及跨模态语义对齐需求。为解决上述问题, 笔者提出一种多层次动态权重分配策略, 构建由子特征注意力、模态间交叉注意力与样动态门控组成的 3 阶段融合机制, 分别针对模态内建模精度、模态间交互深度以及融合权重的个体适应性进行优化, 有效提升多模态系统的语义表达能力与判别性能。算法具体如下。

算法: 多层次动态权重分配算法

输入: 多模态预测分数 $X, X, \dots, X$

输出: 最终预测概率 Y

1) // Step 1: 子模态特征注意力 (Sub-Attention) 层

2) for each 模态 i in $\{1, 2, \dots, M\}$ do

3) // 通过两个串联的一维卷积提取时序上下文信息 4) $H'_i = \text{ReLU}(\text{Conv1D}(X_i, W_1)) \in \mathbb{R}^{T \times d}$

5) $H_i = \text{ReLU}(\text{Conv1D}(H'_i, W_2)) \in \mathbb{R}^{T \times d}$

6) // 基于特征图 H_i 计算注意力权重

7) $A_i = \text{softmax}(H_i \cdot W_i) \in \mathbb{R}^{T \times d}$

8) // 应用注意力权重, 增强关键通道响应

9) $F_i = H_i \odot A_i \in \mathbb{R}^{T \times d}$

10) end for

11) // Step 2: 模态间交叉注意力 (Cross-Modal Attention) 层

12) for each 模态 i in $\{1, 2, \dots, M\}$ do

13) // 初始化交叉表示为当前模态自身的表示

14) $F_i^{\text{cross}} = F_i$

15) // 与其他模态进行交叉注意力计算

16) for each 其他模态 $j \neq i$ do

17) $Q_i = F_i \cdot W_Q^{(i)}$ // 查询向量

18) $K_j = F_j \cdot W_K^{(j)}$ // 键向量

19) $V_j = F_j \cdot W_V^{(j)}$ // 值向量

20) // 基于缩放点积计算注意力分数

21) $\alpha_{ij} = \text{softmax}\left(\frac{Q_i K_j^T}{\sqrt{d}}\right)$

22) // 得到加权的交叉注意力表示

23) $F_i^{(j)} = \alpha_{ij} \cdot V_j$

24) // 将交叉注意力结果累加到 F_i^{cross}

25) $F_i^{\text{cross}} = F_i^{\text{cross}} + F_i^{(j)}$

26) end for

```

27) //加入残差连接与归一化, 增强稳定性
28)  $\tilde{\mathbf{F}}_i = \text{LayerNorm}(\mathbf{F}_i^{\text{cross}})$ 
29) end for
30) //Step 3: 动态门控融合(Dynamic Gating)层
31) //拼接所有交互增强后的模态特征
32)  $\mathbf{F} = [\tilde{\mathbf{F}}_1; \tilde{\mathbf{F}}_2; \dots; \tilde{\mathbf{F}}_M] \in \mathbb{R}^{M \cdot T \times d}$ 
33) //进行平均池化以获得全局表示
34)  $\mathbf{F}_{\text{mean}} = \text{MeanPool}(\mathbf{F})$ 
35) //生成每个模态的融合权重
36)  $\mathbf{G} = \text{softmax}(\text{MLP}_g(\mathbf{F}_{\text{mean}})) \in \mathbb{R}^M$ 
37) //按模态权重加权融合最终表示
38)  $\mathbf{F}_{\text{fused}} = \mathbf{0}$ 
39) for  $i$  in  $\{1, 2, \dots, M\}$  do
40)  $\mathbf{F}_{\text{fused}} = \mathbf{F}_{\text{fused}} + \mathbf{G}_i \cdot \text{MeanPool}(\tilde{\mathbf{F}}_i)$ 
41) end for
42) //Step 4: 最终预测
43)  $\mathbf{Y} = \sigma(\mathbf{W}_f \cdot \mathbf{F}_{\text{fused}} + \mathbf{b}_f) \in [0, 1]$ 

```

3 实验与结果

3.1 实验环境与数据集

所有模型都由 Python 语言以及 Pytorch 深度学习框架实现。同时在 Windows 11 操作系统下的 NVIDIA GTX4090D GPU 上进行训练。实验采用的数据集有脑电、文本和人脸数据集。脑电数据集是由 Mumtaz 在 2017 年提供的公开数据集; 文本数据集是由 InFamousCoder 在 kaggle 平台发布的一个开源的数据集 Reddit; 人脸数据集是 MahmoudiMA 于 2020 年公开的一个人脸表情数据集 MMAFEDB。整个数据集共包含 119 240 张图片。

3.2 评估指标

为评估所提出的抑郁症识别框架的性能, 选择了 6 个指标进行评估, 分别是准确率(Accuracy, A)、精确率(Precision, P)、召回率(Recall, R)、特异性(Specificity, S)、 F_1 值和 AUC(Area Under Curve)。AUC 表示的是真正例率(TPR: True Positive Rate)与假正例率(FPR: False Positive Rate)的关系。TP(True Positive, T_p)表示真阳性、TN(True Negative, T_N)表示真阴性、FP(False Positive, F_p)表示假阳性、FN(False Negative, F_N)表示假阴性, 如下:

$$A = \frac{T_p + T_N}{T_p + T_N + F_p + F_N}, \quad (1)$$

$$P = \frac{T_p}{T_p + F_p}, \quad (2)$$

$$R = \frac{T_p}{F_N + T_p}, \quad (3)$$

$$S = \frac{T_N}{T_N + F_p}, \quad (4)$$

$$F_1 = \frac{2T_p}{2T_p + F_p + F_N}, \quad (5)$$

$$T_{\text{TPR}} = \frac{T_p}{T_p + F_N}, \quad (6)$$

$$F_{FPR} = \frac{F_P}{F_P + T_N}, \tag{7}$$

$$A_{AUC} = \int_0^1 T_{TPR} F_{FPR} dF_{FPR} \circ \tag{8}$$

3.3 实验对比结果和分析

为能对实验做出更全面和深入的分析，笔者设计了一系列的对比实验和消融实验。通过对不同模态间模型的比较，能有效评估各个因素对抑郁症识别性能的影响。

3.3.1 脑电模态实验结果对比

笔者在脑电模态方面与近几年使用同一数据集的深度学习模型进行了对比，并在 CNN 和 LSTM 模型中进行对比实验。Rafiei 等^[26]在 2022 年提出了一种自定义的 Inception time 模型，其最高分类准确率为 91.67%，Xia 等^[27]在 2023 年提出了一种基于脑连通性特征的抑郁症分类方法，其平均分类准确率达到 91.06%。比较结果如表 1 所示。

表 1 脑电模态对比试验
Tab.1 Comparison test of EEG modes

模型	准确率/%	精确率/%	召回率/%	特异性/%	F_1	AUC
CNN	93.17	94.97	87.40	95.63	94	90.65
LSTM	87.2	87.54	80.29	86.36	88.75	89.83
Inception ^[26]	91.67		94.9	88.2		
DSNet ^[27]	91.69		92.11	90.54		
MFF-DR-DWA	99.09	98.88	99.36	98.80	99.12	99.97

实验结果表明，笔者提出的 MFF-DR-DWA 框架表现最佳，准确率达 99.09%，精确率、召回率、 F_1 值和分别为 98.88%、99.36%、99.12%，显著优于其他模型。相比较，Inception time 模型和 DSNet(Dual Stream Network)模型的召回率较高，但准确率和特异性稍低。传统 CNN 和 LSTM 模型在整体性能上表现较为一般，尤其是召回率偏低。

3.3.2 文本模态实验结果对比

在文本模态方面分别使用了传统的机器学习方法(朴素贝叶斯)，深度学习方法 BERT, XLNet(eXtra Large Network), DistilBERT 以及 TinyLlama 与本文的 MFF-DR-DWA 模型在同一数据集上进行对比试验，结果如表 2 所示。

表 2 文本模态对比实验
Tab.2 Text modal comparison experiment

模型	准确率/%	精确率/%	召回率/%	特异性/%	F_1	AUC
朴素贝叶斯	88.88	88.94	88.95	88.95	89.68	88.95
BERT	95.51	95.52	95.51	95.23	96.36	95.37
XLNet ^[28]	94.70	93.17	96.16	93.32	94.64	95.74
DistilBERT ^[29]	95.29	94.29	96.13	94.50	95.20	95.31
TinyLlama ^[30]	90.83	91.02	90.03	91.59	90.52	91.59
MFF-DR-DWA	99.09	98.88	99.36	98.80	99.12	99.97

实验结果表明，笔者 MFF-DR-DWA 模型在抑郁症识别任务中表现出全面优势，其准确率 (99.09%)、 F_1 值(99.12)和 AUC(99.97)等核心指标均显著优于所有对比模型。与传统机器学习方法相比，MFF-DR-DWA 较朴素贝叶斯(准确率 88.88%)性能提升超过 10%。与主流预训练语言模型相比，该模型较 BERT(95.51%)、DistilBERT(95.29%)等模型在准确率上提升 3.5%；与 TinyLlama 相比模型准确率提升达 8%。上述结果充分验证了多模态动态加权融合策略的有效性。

3.3.3 人脸模态实验结果对比

在人脸模态上分别使用了 5 种深度学习的模型进行对比实验，分别是 DenseNet, GoogleNet, MobileNetV2, SqueezeNet, ConvNeXtV2 以及本文的 MFF-DR-DWA，结果如表 3 所示。

表 3 人脸模态对比实验

Tab. 3 Facial modal comparison experiment

模型	准确率/%	精确率/%	召回率/%	特异性/%	F_1 /%	AUC/%
DenseNet ^[31]	83.52	81.30	85.86	81.31	83.59	89.59
GoogleNet ^[32]	81.87	79.63	84.87	79.60	84.36	88.24
MobileNet ^[33]	81.02	78.20	84.53	77.70	81.24	89.68
SqueezeNet ^[34]	78.56	72.81	89.23	68.46	80.19	89.12
ConvNeXt ^[35]	76.57	73.08	82.02	71.41	80.51	86.84
MFF-DR-DWA	99.09	98.88	99.36	98.80	99.12	99.97

实验结果表明, MFF-DR-DWA 在抑郁症识别任务中表现最佳, 准确率为 99.09%, 精确率为 98.88%, 召回率为 99.36%, F_1 值为 99.12%, AUC 值为 99.97%, 远超其他模型。相比较, DenseNet、GoogleNet、MobileNet、SqueezeNet 和 ConvNeXtV2 的准确率分别为 83.52%、81.87%、81.02%、78.56% 和 76.57%, 在精确率、召回率和特异性等指标上均逊色于 MFF-DR-DWA。

3.3.4 模态消融实验对比

为评估模型中不同模态对模型的影响, 笔者设计了不同模态相互组合以及模块间的消融实验, 结果如表 4 所示。

表 4 消融对比实验

Tab. 4 Comparison experiment of ablation

模型	准确率/%	精确率/%	召回率/%	特异性/%	F_1 /%	AUC/%
EEG+文本	92.04	91.00	93.82	90.14	96.17	96.59
EEG+人脸	91.53	87.09	98.11	84.54	95.80	95.99
文本+人脸	91.64	87.42	97.87	85.03	95.31	96.54
普通融合	95.42	95.17	96.44	95.16	96.35	96.98
无特征对齐	98.04	97.87	98.39	97.67	98.25	98.84
固定权重	98.50	98.03	97.41	97.93	98.63	99.00
无交叉注意力	98.86	98.05	98.72	98.00	98.35	98.92
MFF-DR-DWA	99.09	98.88	99.36	98.80	99.12	99.97

实验结果表明, 笔者提出的各关键结构模块(特征对齐、动态门控、交叉注意力)在提升多模态融合性能方面均发挥了积极作用。相较于普通融合方式(AUC 为 96.98), 逐步引入上述模块后, 模型性能持续提升, 最终完整结构(MFF-DR-DWA)取得了最高的 AUC(99.97), 显示出结构设计的显著增益效应。

移除特征对齐模块后, 模型 AUC 降至 98.84, 说明该模块在缓解模态间特征分布差异、增强融合协同性方面起到了关键作用。当将动态门控机制替换为固定模态权重后, AUC 降至 99.00, 表明虽然固定融合方式在一定程度上可行, 但动态机制能更有效地捕捉样本间模态重要性的差异。进一步地, 在移除交叉注意力模块后, AUC 降为 98.92, 表明模态间的信息引导机制对于充分挖掘互补特征仍具有重要意义。

为进一步排除性能提升仅来源于模态数量增加的可能性, 笔者设计了 3 组双模态组合实验(EEG+文本、EEG+人脸、文本+人脸), 对应的 AUC 分别为 96.17、95.80 和 95.31, 均低于普通融合和结构增强后的结果。这种现象进一步印证了最终性能的提升不仅归功于多模态信息本身, 更依赖于对特征协同机制的精细建模与结构设计的合理性。

4 结 论

针对抑郁症检测中单一模态判别能力有限的问题, 笔者提出了一种多模态动态融合框架 MFF-DR-DWA, 通过自适应整合脑电、文本和人脸模态信息, 显著提升了检测性能。主要贡献包括以下 4 个方面:

- 1) 针对脑电(EEG)信号, 提出了一种融合卷积神经网络(CNN)、多头自注意力机制与门控循环单

元(GRU)的混合模型,有效提取了脑电信号中的时序依赖与空间特征;

2) 针对文本模态,采用 RoBERTa 预训练模型深入挖掘语言表达中蕴含的语义层面的抑郁特征;

3) 在面部表情分析中,引入 ResNet 主干网络与卷积注意力模块 CBAM,增强了对细微情绪表征的提取能力;

4) 提出了包括特征对齐模块、动态门控机制与交叉注意力机制在内的多模态融合关键结构,显著提升了多模态特征融合的效果。

实验结果表明,本方法在多项评价指标上均优于单一模态基准模型和常规融合策略,充分证明了多模态融合在抑郁症识别中的有效性。

然而,本研究仍存在一定局限性。一方面,实验所采用数据集的规模和多样性仍显不足,可能对模型的泛化能力造成影响;另一方面,当前研究主要集中于脑电、文本和人脸3种模态,尚未涵盖语音、心率等其他可能具有判别价值的生理与行为信号。未来的研究工作可从以下2个方向展开:1) 扩充数据集规模,纳入更广泛人群样本和更丰富模态数据;2) 探索更多模态的有效融合机制,进一步提升模型的鲁棒性与实用价值。

参考文献:

- [1] 马思梦, 曹勇, 王培琳, 等. 深度学习算法在面向行为分析的抑郁症辅助诊断中的研究进展 [J]. 中华精神科杂志, 2020, 53(5): 460-463.
MA S M, CAO Y, WANG P L, et al. Research Progress of Deep Learning Algorithms in the Auxiliary Diagnosis of Depression Based on Behavior Analysis [J]. Chinese Journal of Psychiatry, 2020, 53(5): 460-463.
- [2] 张鑫衍. 基于文本挖掘的社交网络用户精神疾病筛查 [J]. 现代信息科技, 2023, 7(15): 157-161.
ZHANG X Y. Social Network User Mental Disease Screening Based on Text Mining [J]. Modern Information Technology, 2023, 7(15): 157-161.
- [3] 姜月武, 路东生, 党良慧, 等. 人脸表情识别研究进展 [J]. 智能计算机与应用, 2021, 11(6): 43-50.
JIANG Y W, LU D S, DANG L H, et al. Research Progress on Facial Expression Recognition [J]. Intelligent Computer and Applications, 2021, 11(6): 43-50.
- [4] 刘峻杉. 基于多模态数据的青少年抑郁症诊断模型关键技术研究 [D]. 成都: 四川大学, 2021.
LIU J S. Research on Key Technologies of Adolescent Depression Diagnosis Model Based on Multimodal Data [D]. Chengdu: Sichuan University, 2021.
- [5] 朱永新. “五育并举”构建一体化青少年心理健康教育体系——关于青少年心理健康教育问题的探讨 [J]. 苏州大学学报(教育科学版), 2025, 13(1): 1-12.
ZHU Y X. Constructing an Integrated Adolescent Mental Health Education System through the “Five Education” Approach: A Discussion on Adolescent Mental Health Education Issues [J]. Journal of Soochow University (Education Science Edition), 2025, 13(1): 1-12.
- [6] 李宸芮, 吴毅明, 陈新旺, 等. 脑电静息微状态在广泛性焦虑症中的应用与研究进展 [J]. 中国比较医学杂志, 2025, 35(2): 158-164.
LI C R, WU Y M, CHEN X W, et al. Application and Research Progress of EEG Resting Microstate in Generalized Anxiety Disorder [J]. Chinese Journal of Comparative Medicine, 2025, 35(2): 158-164.
- [7] FARRUQUE N, GOEBEL R, SIVAPALAN S, et al. Deep Temporal Modelling of Clinical Depression through Social Media Text [J/OL]. Natural Language Processing Journal, 2024, 6: 100052 (2024-03-01) [2025-03-23]. <https://www.sciencedirect.com/science/article/pii/S2949719123000493>.
- [8] WANG S, WANG X, WANG S, et al. Bi-Directional Long Short-Term Memory Method Based on Attention Mechanism and Rolling Update for Short-Term Load Forecasting [J]. International Journal of Electrical Power & Energy Systems, 2019, 109: 470-479.
- [9] WU Y, LIU Z, YUAN J, et al. PIE: A Personalized Information Embedded Model for Text-Based Depression Detection [J/OL]. Information Processing & Management, 2024, 61(6): 103830 (2024-11-01) [2025-08-23]. <https://www.sciencedirect.com/science/article/pii/S0306457324001894>.
- [10] DING Z, HU Y, JING R, et al. A Depression Recognition Method Based on the Alteration of Video Temporal Angle Features

- [J/OL]. *Applied Sciences*, 2023, 13(16): 9230 (2023-08-14) [2025-03-23]. <https://www.mdpi.com/2076-3417/13/16/9230>.
- [11] HAN K, WANG Y H, TIAN Q, et al. GhostNet: More Features from Cheap Operations [C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle: IEEE/CVF, 2020: 1580-1589.
- [12] LI X Y, YI X Y, YE J Y, et al. SFTNet: A Microexpression-Based Method for Depression Detection [J/OL]. *Computer Methods and Programs in Biomedicine*, 2023, 243: 107923 [2025-05-10]. <https://doi.org/10.1016/j.cmpb.2023.107923>.
- [13] AHMED N N, BHAT T K, POWAR O S. Stacked Ensemble Machine Learning Approach for Electroencephalography Based Major Depressive Disorder Classification Using Temporal Statistics [J/OL]. *Systems Science & Control Engineering*, 2024, 12(1): 2427028 (2024-07-04) [2025-03-23]. <https://www.tandfonline.com/doi/full/10.1080/21642583.2024.2427028>.
- [14] XU X, XU J, DU R, et al. Research on Depression Recognition Model and Its Temporal Characteristics Based on Multiscale Entropy of EEG Signals [J/OL]. *Entropy*, 2025, 27(2): 142 [2025-03-23]. <https://pubmed.ncbi.nlm.nih.gov/40003139/>.
- [15] HE M, BAKKER M E, LEW S M. DPD (DePression Detection) Net: A Deep Neural Network for Multimodal Depression Detection [J/OL]. *Health Information Science and Systems*, 2024, 12(1): 53 (2024-11-12) [2025-03-23]. <https://link.springer.com/article/10.1007/s13755-024-00311-9>.
- [16] CHENG Z, HUANG X, DING Y. An Intelligent Depression Detection Model Based on Multimodal Fusion Technology [J/OL]. *Journal of Mechanics in Medicine and Biology*, 2024, 24(8) (2024-08-22) [2025-03-23]. <https://www.worldscientific.com/doi/full/10.1142/S0219519424400463>.
- [17] MOHAMMAD F, ANSOOR A M K. MDD: A Unified Multimodal Deep Learning Approach for Depression Diagnosis Based on Text and Audio Speech [J]. *Computers, Materials & Continua*, 2024, 81(3): 4125-4147.
- [18] RAFFEL C, SHAZEER N, ROBERTS A, et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer [J]. *Journal of Machine Learning Research*, 2020, 21: 5485-5551.
- [19] 赵小明, 范慧婷, 张石清. 一种基于多模态特征增强网络的抑郁症检测方法 [J]. *软件工程*, 2024, 27(10): 68-73.
ZHAO X M, FAN H T, ZHANG S Q. A Depression Detection Method Based on Multimodal Feature Enhancement Network [J]. *Software Engineering*, 2024, 27(10): 68-73.
- [20] MUMTAZ W, XIA L, YASIN M A M, et al. A Wavelet-Based Technique to Predict Treatment Outcome for Major Depressive Disorder [J/OL]. *PLoS ONE*, 2017, 12(2): e0171409 (2017-02-02) [2025-03-23]. <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0171409>.
- [21] KUMAR S J, BHUVANESWARI P. Analysis of Electroencephalography (EEG) Signals and Its Categorization: A Study [J]. *Procedia Engineering*, 2012, 38: 2525-2536.
- [22] LIU Y, OTT M, GOYAL N, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach [EB/OL]. (2019-07-26) [2025-03-23]. Available: <https://arxiv.org/abs/1907.11692>.
- [23] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding [C]//*Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019: 4171-4186.
- [24] HE K, ZHANG X, REN S, et al. Deep Residual Learning for Image Recognition [C]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway, New Jersey, USA: IEEE, 2016: 770-778.
- [25] WOO S, PARK J, LEE J Y, et al. CBAM: Convolutional Block Attention Module [C]//*Proceedings of the European Conference on Computer Vision (ECCV)*. Cham: Springer, 2018: 3-19.
- [26] RAFIEI A, ZAHEDIFAR R, SITAULA C, et al. Automated Detection of Major Depressive Disorder with EEG Signals: A Time Series Classification Using Deep Learning [J]. *IEEE Access*, 2022, 10: 73804-73817.
- [27] XIA M, ZHANG Y S, WU Y H, et al. An End-to-End Deep Learning Model for EEG-Based Major Depressive Disorder Classification [J]. *IEEE Access*, 2023, 11: 41337-41347.
- [28] YANG Z, DAI Z, YANG Y, et al. XLNet: Generalized Autoregressive Pretraining for Language Understanding [C]//*Advances in Neural Information Processing Systems*. Red Hook: Curran Associates Inc, 2020: 5730-5740.
- [29] SANH V, DEBUT L, CHAUMOND J, et al. DistilBERT: A Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter [DB/OL]. (2019-10-02) [2025-03-23]. <https://arxiv.org/abs/1910.01108>.
- [30] ZHANG P, ZENG G, WANG T, et al. TinyLlama: An Open-Source Small Language Model [DB/OL]. (2024-01-04) [2025-

- 06-05]. <https://arxiv.org/abs/2401.02385>.
- [31]HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely Connected Convolutional Networks [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Piscataway, New Jersey, USA: IEEE, 2017: 2261-2269.
- [32]SZEGEDY C, LIU W, JIA Y, et al. Going Deeper with Convolutions [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Piscataway, New Jersey, USA: IEEE Computer Society, 2015: 1-9.
- [33]SANDLER M, HOWARD A, ZHU M, et al. MobileNetV2: Inverted Residuals and Linear Bottlenecks [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, New Jersey, USA: IEEE, 2018: 4510-4520.
- [34]QIANG B H, ZHAI Y J, ZHOU M L, et al. SqueezeNet and Fusion Network-Based Accurate Fast Fully Convolutional Network for Hand Detection and Gesture Recognition [J]. IEEE Access, 2021, 9: 77661-77674.
- [35]WOO S, DEBNATH S, HU R, et al. ConvNeXtV2: Co-Designing and Scaling ConvNets with Masked Autoencoders [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2023). Piscataway, New Jersey, USA: IEEE Computer Society, 2023: 16133-16142.

(责任编辑:刘俏亮)