Computer Science and Application 计算机科学与应用, 2026, 16(2), 366-380
Published Online February 2026 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2026.162066

Hans汉斯

基于交叉注意力融合与模糊C均值聚类的多模态抑郁识别

王亚腾，张金珠，王姝童

河北工业大学理学院，天津

收稿日期：2026年1月6日；录用日期：2026年2月6日；发布日期：2026年2月14日

摘　要

抑郁症的早期识别对有效干预至关重要，捕获深层次的音频特征和具有长距离依赖关系的文本特征并进行特征融合提升模态耦合能力是当前主要挑战。基于深度学习和多模态数据建立端到端的抑郁症识别模型：采用VGGish-NetVLAD-GRU模型提取音频深层时序特征；RoBERTa-BiLSTM模型捕捉文本长程语义依赖；通过交叉注意力融合实现语音－文本特征的动态权重分配与跨模态语义对齐；引入模糊C均值聚类算法(Fuzzy C-means, FCM)，基于概率隶属度对情感相近的样本进行软划分，实现抑郁症分类。实验结果表明，在EATD-Corpus和CMDC中文数据集上该模型准确率分别达97.0%和94.0%，F1值分别达97.0%和94.0%。在EATD-Corpus数据集上设计对照实验，交叉注意力缺失准确率下降13%，FCM缺失准确率降低7%。

关键词

抑郁检测，交叉注意力机制，模糊c均值聚类，多模态，深度学习

The Multimodal Depression Recognition Based on Cross-Attention Fusion and Fuzzy C-Means Clustering

Yateng Wang, Jinzhu Zhang, Shutong Wang

School of Science, Hebei University of Technology, Tianjin

Received: January 6, 2026; accepted: February 6, 2026; published: February 14, 2026

Abstract

Early identification of depression is crucial for effective intervention, with current primary

文章引用: 王亚腾, 张金珠, 王姝童. 基于交叉注意力融合与模糊 C 均值聚类的多模态抑郁识别[J]. 计算机科学与应用, 2026, 16(2): 366-380. DOI: 10.12677/csa.2026.162066

challenges being the extraction of deep audio features and long-range dependent text features, along with enhancing modal coupling capability through feature fusion. An end-to-end depression recognition model was established based on deep learning and multimodal data: the VGGish-NetVLAD-GRU model was employed to extract deep temporal audio features; the RoB-ERTa-BiLSTM model captured long-range semantic dependencies in text; dynamic weight allocation and cross-modal semantic alignment of speech-text features were achieved through parametric cross-attention fusion; the Fuzzy C-means (FCM) clustering algorithm was introduced to perform soft partitioning of samples with similar emotional characteristics based on probabilistic membership, thereby enabling depression classification. Experimental results demonstrated that the model achieved accuracies of 97.0% and 94.0% on the EATD-Corpus and CMDC Chinese datasets, respectively, with corresponding F1 scores of 97.0% and 94.0%. Ablation studies on the EATD-Corpus dataset showed that the absence of cross-attention led to a 13% decrease in accuracy, while the absence of FCM resulted in a 7% reduction in accuracy.

Keywords

Depression Recognition, Cross-Attention Fusion, Fuzzy C-Means Clustering, Multimodal, Deep Learning

1. 引言

1.1. 研究背景

抑郁症作为全球性公共健康问题，具有高危害性和高复发性，临床表现为持续性情绪低落、兴趣减退及认知功能损害等[1]。据《2022 年国民抑郁症蓝皮书》数据显示，我国抑郁症患者约 9500 万，每年因抑郁症自杀身亡人数逾 70 万。早期检测对及早干预和降低患抑郁症人数至关重要，但全球范围内仅不足 30%的患者获得规范化治疗，抑郁症检测系统效能不足已成为制约抑郁症防治的关键问题。传统的抑郁诊断手段依赖于自我报告工具和临床访谈，其中，自我报告工具虽具有良好信效度，但其封闭式问题容易诱发社会期望偏差，冗长的测评流程可能导致受访者疲劳效应，影响评估效度；临床访谈则受限于筛查工作量、疾病羞耻及患者病识力不足或症状隐匿化倾向，导致主诉信息与真实病理状态产生系统性偏差。这种双重局限性制约了诊断准确率，因此，抑郁症的早期检测、定量精准化诊疗已成为当今亟需解决的社会热点问题。

1.2. 研究现状

在现有的抑郁症自动诊断研究中，生理信号脑电、声学特征、面部表情、行为举止都为抑郁症的早期诊断提供了有效信息，捕捉抑郁症患者的真实反应和情感变化。Pandey 等人[2]指出使用多模态融合技术整合异构资源，为抑郁症早期检测提供了新的可能。Shen 等人[3]结合音频和文本数据，提出基于 GRU 和带有注意力层的 BiLSTM 模型的抑郁症检测方法，在 DAIC-WOZ 和 EATD-Corpus 两个数据集上进行实证研究，F1 得分分别达到 0.85 和 0.71。张亚洲等人[4]使用 RoBERTa-BiLSTM 模型从抑郁文本序列中充分提取和利用上下文特征，捕获长距离上下文语义，在 DAIC-WOZ 和 EATD-Corpus 两个数据集上进行实证研究，准确率分别达到 0.74 和 0.93 以上。赵小明等人[5]从文本引导重构的音频特征中解离共享和

私有特征，提出了基于音频－文本数据的跨模态特征重构与解耦网络的多模态抑郁症检测方法。Chen 等人[6]提出基于文本引导的跨模态特征重构与分解框架，结合双向交叉注意力模块增强模态交互，使用 Transformer 融合共享与私有特征，显著提升了多模态抑郁症检测的性能。Li 等[7]使用交叉注意力机制进行多模态特征融合，在测试集上实现 0.95 的准确率，证明了交叉注意力在提升多模态抑郁检测精度方面的有效性。

1.3. 研究内容

尽管学界对多模态数据融合方面已经有一定成就，但是当前抑郁症检测研究仍面临公开数据集类别分布失衡和现有模态融合方法难以充分挖掘模态间的深层交互两个挑战。为应对上述问题，本文旨在构建端到端的多模态抑郁检测模型，实现对抑郁程度的四分类(非抑郁、轻度、中度、重度)。

具体研究框架如图 1 所示，包含以下四个环节：

1) 构建 VGGish-NetVLAD-GRU 模型进行音频时频特征提取与序列建模。
2) 构建 RoBERTa-BiLSTM 模型捕获文本深层语义表征。
3) 引入交叉注意力融合机制，实现语音－文本模态间的深度交互、动态权重分配与跨模态语义对齐。
4) 使用 FCM 算法对样本特征进行分类。

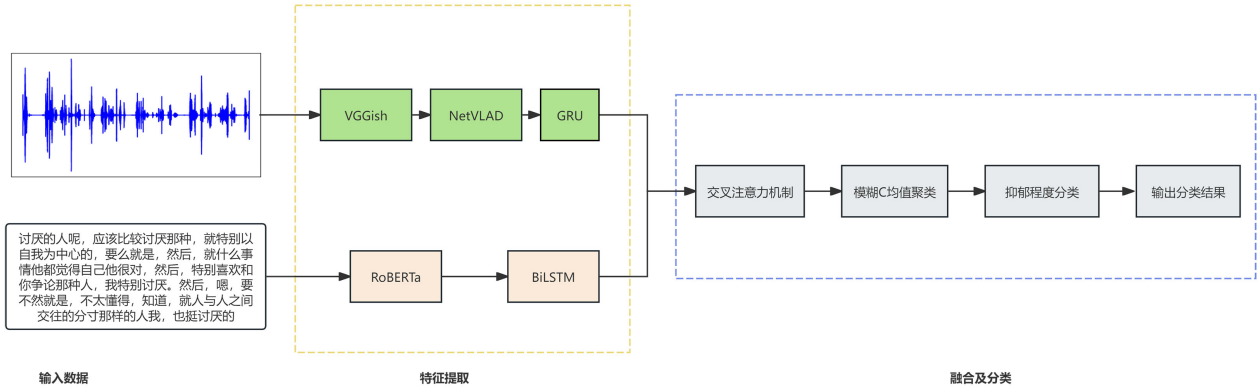

Figure 1. Multimodal depression detection framework diagram
图 1. 多模态抑郁检测框架图

2. 研究方法

2.1. VGGish-NetVLAD-GRU 模型

为充分表征音频中与抑郁状态相关的局部声学结构与全局时序依赖，本文构建 VGGish-NetVLAD-GRU 模型，将谱图级深度特征、聚类式全局表示与循环网络建模有机结合。对原始访谈录音统一重采样至 16 kHz，按照 0.96 s 分段，每段信号再以 25 ms 窗长、10 ms 帧移进行加窗与分帧，采用 Hamming 窗抑制边缘效应，通过短时傅里叶变换(STFT)得到线性功率谱，经过 64 个 Mel 滤波器组映射为 Mel 频谱。鉴于人耳对声音强度的感知呈对数特性，将 Mel 频谱进行自然对数变换得到对数 Mel 频谱图，模拟听觉感知并突显情感相关的声学细微差异。将对数 Mel 频谱图进行 z-score 标准化，以减轻量纲差异对后续卷积与聚类的影响。对标准化后的第 i 个音频片段的对数 Mel 频谱图输入到预训练 VGGish 网络得到从时频图到情感相关嵌入的非线性映射，将整段音频获取的嵌入向量输入 NetVLAD 模型[8]，通过可学习的聚类中心将局部特征的分布模式聚合为固定长度的全局特征。为刻画跨句段的情绪动态，NetVLAD 全局特征在时间维度上进一步通过 GRU 建模，记第 t 个时间步的输入为记第 t 个时间步的输入为 v_t，GRU 的

更新过程为：

$$z_t = \sigma\left(W_z v_t + U_z h_{t-1} + b_z\right) \tag{1}$$

$$r_t = \sigma\left(W_r v_t + U_r h_{t-1} + b_r\right) \tag{2}$$

$$\tilde{h}_t = \tanh\left(W_h v_t + U_h\left(r_t \odot h_{t-1}\right) + b_h\right) \tag{3}$$

$$h_t = \left(1 - z_t\right) \odot h_{t-1} + z_t \odot \tilde{h}_t \tag{4}$$

其中 z_t 为更新门，r_t 为重置门，控制在生成候选隐状态时保留历史记忆，$\sigma(\cdot)$ 为 sigmoid 函数，$\tilde{h}_t$ 为候选隐状态，$\odot$ 表示逐元素乘，$\tanh(\cdot)$ 为非线性变换，刻画当前输入及被重置后的历史信息所形成的新表示，W_z，U_z，b_z，W_r，U_r，b_r，W_h，U_h，b_h 为可学习的参数矩阵和偏置项，h_t 为最终隐状态，将旧状态与候选状态按更新门进行加权融合，实现对长程依赖的选择性记忆与遗忘。将最后时间步的隐状态向量作为整段语音的全局音频表征 h_{audio}。

2.2. RoBERTa-BiLSTM 模型

RoBERTa-BiLSTM 模型用于从访谈转录文本中抽取上下文相关的深层语义特征及双向时序依赖，使用将文本序列编码为词元序列 $S = \{w_1, w_2, \cdots, w_N\}$ 输入至 RoBERTa 模型得到上下文相关的词向量序列：

$$e_t = \text{RoBERTa}\left(w_t\right), t = 1, \cdots, N \tag{5}$$

其中 w_t 为第 t 个词元，N 为序列长度，$e_t \in \mathbb{R}^{768}$ 为融合前后语义信息的词元嵌入。

将所有词元嵌入组成文本模态的上下文语义序列表示 E，输入到 BiLSTM 模型中，生成前向隐藏状态序列 $\{\vec{s}_1, \vec{s}_2, \cdots, \vec{s}_m\}$ 和后向隐藏状态序列 $\{\overleftarrow{s}_1, \overleftarrow{s}_2, \cdots, \overleftarrow{s}_m\}$：

$$E = \left[e_1, e_2, \cdots, e_N\right] \in \mathbb{R}^{N \times 768} \tag{6}$$

$$\vec{s}_i = \text{LSTM}_f\left(E_i, \vec{s}_{i-1}\right) \tag{7}$$

$$\overleftarrow{s}_i = \text{LSTM}_b\left(E_i, \overleftarrow{s}_{i+1}\right) \tag{8}$$

最终得到同时编码过去和未来的上下文信息的文本特征表示：

$$h_{text} = \left[\vec{s}_m; \overleftarrow{s}_1\right] \tag{9}$$

2.3. 交叉注意力融合

为显式建模音频与文本模态间的互补信息与语义对齐，本文采用交叉注意力机制进行融合：以音频为查询增强语音表征，以文本为查询增强文本表征，并在统一潜在空间中实现跨模态对齐。在音频特征增强部分，将音频特征映射为查询 Q_{audio}，将文本特征映射为键 K_{text} 与值 V_{text}：

$$Q_{audio} = h_{audio} W_{Q_a} \tag{10}$$

$$K_{text} = h_{text} W_{K_t} \tag{11}$$

$$V_{text} = h_{text} W_{V_t} \tag{12}$$

其中 W_{Q_a}，W_{K_t}，W_{V_t} 为可学习投影矩阵，将两模态嵌入映射到相同维度 d 的潜在语义空间，实现跨模态“投影对齐”。

使用缩放点积计算音频对文本的注意力权重 $A_{a \to t}$：

$$A_{a\to t}=\mathrm{softmax}\left(\frac{Q_{audio}K_{text}^{\mathrm{T}}}{\sqrt{d}}\right) \tag{13}$$

其中 $A_{a\to t}$ 为注意力矩阵，每一行给出某一音频时间步在所有文本位置上的对齐权重。$\sqrt{d}$ 为缩放因子，用于避免点积值过大导致梯度不稳定。

增强后的音频特征 h'_{audio} 计算公式如下，在增强特征中加入原始向量，避免信息丢失并提高网络可训练性：

$$h'_{audio}=A_{a\to t}V_{text}+h_{audio} \tag{14}$$

在文本特征增强部分，使用同样的处理方式，文本特征作为查询，探寻音频特征中能够为其提供韵律和情感色彩的声学线索，获得增强后的文本特征 h'_{text}，将双向增强后的特征在特征维度上拼接，生成交叉注意力融合特征：

$$h_{fused}=\left[h'_{audio},h'_{text}\right] \tag{15}$$

2.4. 模糊 C 均值聚类算法 FCM

与“硬”聚类只给出单一簇标签不同，本文使用 FCM [9]对样本特征进行无监督学习，引入隶属度柔性地描述样本与簇之间的关系，目标函数为最小化样本数据点到簇族中心的加权距离和：

$$D=\sum_{l=1}^{L}\sum_{j=1}^{k}\mu_{jl}^{m}\left\|x_l-y_j\right\|^2 \tag{16}$$

其中 μ_{jl}^{m} 表示第 l 个样本属于第 j 个簇族中心的隶属度，隶属度之和为 1 且 $\mu_{jl}^{m}\in[0,1]$，x_l 为第 l 个样本的融合特征 h_{fused}，y_j 为第 j 个簇族的中心点。k 和 m 均是可调节的参数，其中 k 为预先设定的簇族数量，决定了样本数据分类的精细化程度；m 为模糊系数，控制分类结果的模糊程度，模糊系数越大，簇族间的重叠程度就越大，隶属度的分布越平滑。

第 j 个簇族的中心位置 y_j 计算公式如下：

$$y_j=\frac{\sum_{l=1}^{L}\left(\mu_{jl}^{m}\right)x_l}{\sum_{l=1}^{L}\mu_{jl}^{m}} \tag{17}$$

第 l 个样本属于第 j 个簇族中心的隶属度 μ_{jl} 计算公式如下：

$$\mu_{jl}=\frac{v_{jl}}{Z_l} \tag{18}$$

$$Z_l=\left(\sum_{j=1}^{c}v_{jl}^{\frac{1}{m-1}}\right)^{m} \tag{19}$$

$$v_{jl}=\mathrm{e}^{-\beta_j d_{jl}} \tag{20}$$

$$\beta_j=-\frac{\ln t}{\min\left\|y_h-y_j\right\|^2} \tag{21}$$

其中 β_j 表示第 j 个簇的宽度控制参数，控制簇间最大重叠程度。在模型训练阶段，使用隶属度公式计算每个样本到所有簇族的隶属度矩阵 μ，使用聚类中心公式更新所有的聚类中心 y_j，更新隶属度和聚类中心，直到样本数据点到簇族中心的加权距离和最小。

获得每个样本到各簇族的隶属度矩阵之后，将样本分配到隶属度最高的簇中，对于第 j 个簇族，统计该簇族内各样本的真实四分类标签 $j\in\{0,1,2,3\}$，将第 j 簇映射为出现次数最多的类别，即每个簇族均获得类别标签，并对该簇内所有样本点赋予这一类别。

本文使用 F1 得分、召回率和精确度作为指标来评估模型的性能，具体公式如下：

$$\text{Precision}=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{22}$$

$$\text{Recall}=\frac{\text{TP}}{\text{FP}+\text{FN}} \tag{23}$$

$$\text{F1}=2\cdot\frac{\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}} \tag{24}$$

3. 数据集及数据预处理

本文选用两个中文数据集 EATD-Corpus 和 CMDC 进行抑郁症检测分类，两个数据集的样本构成如表 1 所示。EATD-Corpus 是首个中文多模态抑郁症数据集，包含 162 名参与者的三个问题的临床访谈数据及 SDS 得分。CMDC 数据集[10]是专门用于抑郁症识别研究的中文多模态语料库，包含 78 名参与者的十二个问题的临床访谈数据，该数据库使用 PHQ-9 问卷得分量化参与者的抑郁诊断结果。以上两个数据集的类别分布不均衡，可能导致模型在训练过程中对“多数类”预测良好但是“少数类”分类效果较差，造成高准确率低精确度、召回率的情况。本文使用 SMOTE 过采样算法[11]均衡各类样本，通过在少数类样本之间合成新样本来缓解类别不均衡问题。

Table 1. Four-class sample composition of the dataset
表 1. 数据集四分类样本构成

数据集	非抑郁	轻度	中度	重度	总样本数
EATD-Corpus	132	17	7	6	162
CMDC	48	5	9	16	78

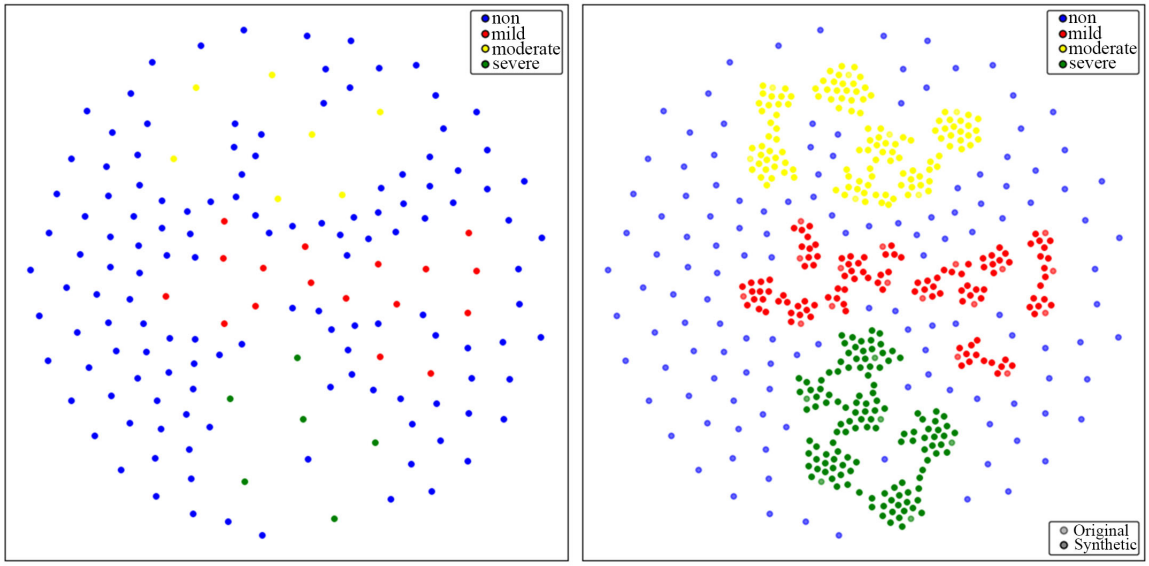

Figure 2. The feature distribution before and after applying SMOTE on the EATD-Corpus dataset
图 2. EATD-Corpus 数据集使用 SMOTE 技术前后特征分布

如图 2、图 3 所示，左图展示了一个代表性交叉验证折的训练集在未采用 SMOTE 时各类别样本在特征空间中的原始分布，右图为仅对该交叉验证折训练集进行 SMOTE 后样本分布情况。实心点表示通过 SMOTE 合成的样本，可以观察到，新增样本有效填补了原少数类样本点之间的空白区域，使少数类在特征空间中的分布更加稠密，并在一定程度上扩展了其特征支撑域。在数据预处理阶段，对音频数据，裁剪所有录音的开头与结尾的静音片段；对文本数据，针对错别字及同音词混淆进行修正，以提升文本数据质量。

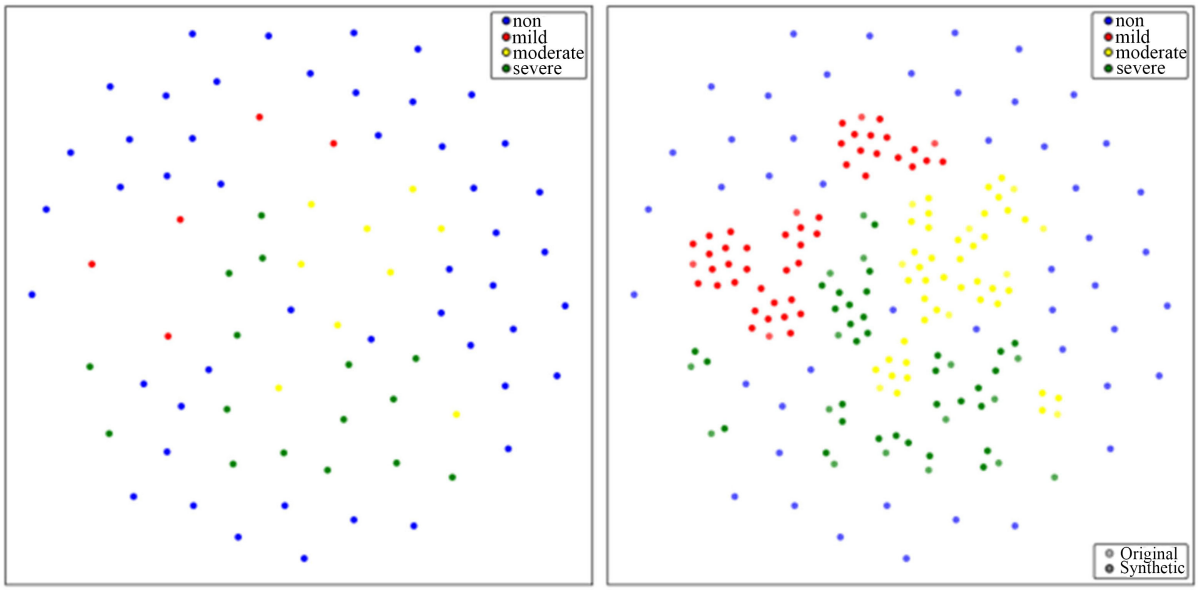

Figure 3. The feature distribution before and after applying SMOTE on the CMDC dataset
图 3. CMDC 数据集使用 SMOTE 技术前后特征分布

4. 实证研究

4.1. 实验设置和超参数调优

经过细致调参，VGGish-NetVLAD-GRU 模型、BERT-BiLSTM 模型的最优参数配置如表 2 所示。在 EATD-Corpus 数据集和 CMDC 数据集中，分别将每段音频对应的音频和文本向量进行交叉注意力融合，获得 1024 维的交叉注意力融合向量，即 EATD-Corpus 数据集的每个样本获得形状为(3, 1024)的融合向量，CMDC 数据集的每个样本获得形状为(12, 1024)的融合向量，将该融合向量作为后续 FCM 的输入。

Table 2. Hybrid model parameter configuration
表 2. 混合模型参数配置

模型	参数设置
VGGish-NetVLAD-GRU	num_clusters: 10 GRU hidden: 512 Layers: 1 Dropout: 0.2
RoBERTa-BiLSTM	Input:768 Hidden: 256 Layers: 1 Dropout: 0.3 Bidirectional: True

为量化模型在 EATD 和 CMDC 数据集上的过拟合风险，并验证模型分类性能的统计显著性，本文采用了 5 折交叉验证。每一折训练中，采用 SMOTE 对训练集进行过采样，平衡各类别的样本数量。在每一折中，分别记录了训练集和验证集的准确率。以 EATD 数据集为例，如图 4 所示，在验证集上准确率稳定在 97.1%到 97.8%之间，训练集和验证集的准确率平均差值为 2.54%，该差值较小，过拟合风险可控。为了进一步评估模型性能的统计显著性，本文对每一折验证集准确率进行了单样本 t 检验，结果表明，验证集平均准确率为 97.3%，95%置信区间为[97.1%, 97.7%]，$p = 2.21\times10^{-10}$，显著高于随机分类水平，表明所提出方法在 EATD 数据集上具有显著优于随机分类的判别能力和较高的稳定性。图 5 为处理音频数据时，NetVLAD 使用不同的聚类中心数量 num_clusters 进行特征提取后，通过 MLP 进行四分类的结果展示，最优聚类中心数为 10。

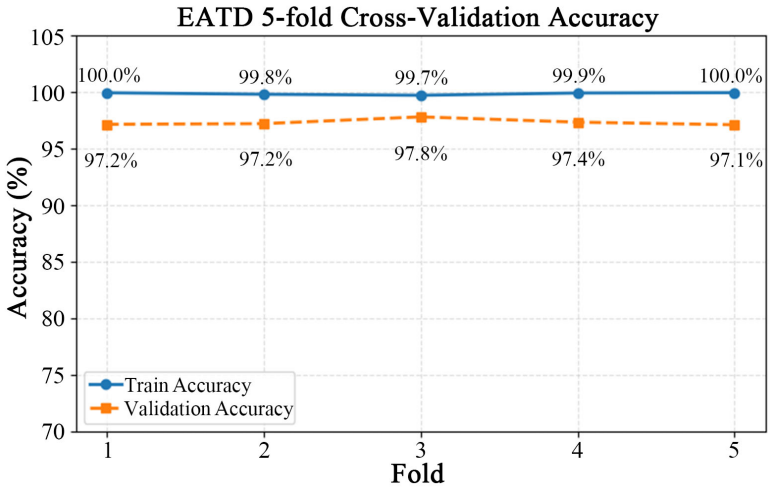

Figure 4. Five-fold cross-validation on the EATD dataset
图 4. EATD 数据集五折交叉验证

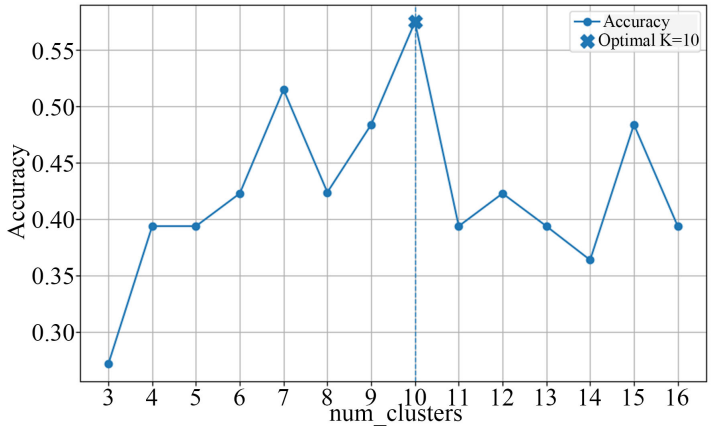

Figure 5. Optimal parameter selection for NetVLAD
图 5. NetVLAD 最优参数选取

4.2. 交叉注意力融合可视化

为更好地呈现交叉注意力融合前后样本的分布情况，以 EATD-Corpus 数据集为例，使用 t-分布随机近邻嵌入(t-SNE)技术进行可视化，将高维数据投影到二维空间中。图 6 为文本模态下样本点分布，图 7

为音频模态下样本点分布，在单模态表征下，非抑郁类别样本呈现广泛而分散的分布，而轻度、中度及重度抑郁样本则零星分布于空间各处，类间边界模糊且重叠严重。图 8 为使用交叉注意力融合机制后样本的投影分布，在多模态融合空间中非抑郁样本点分布更加密集，轻度与中度抑郁样本在投影中心区域出现较明显的空间过渡结构，体现出类别间的渐进差异。由此可见，跨模态互补信息的深度融合，不仅提升了各类样本的可分性，而且实证了多模态特征融合在抑郁症检测识别中的必要性。

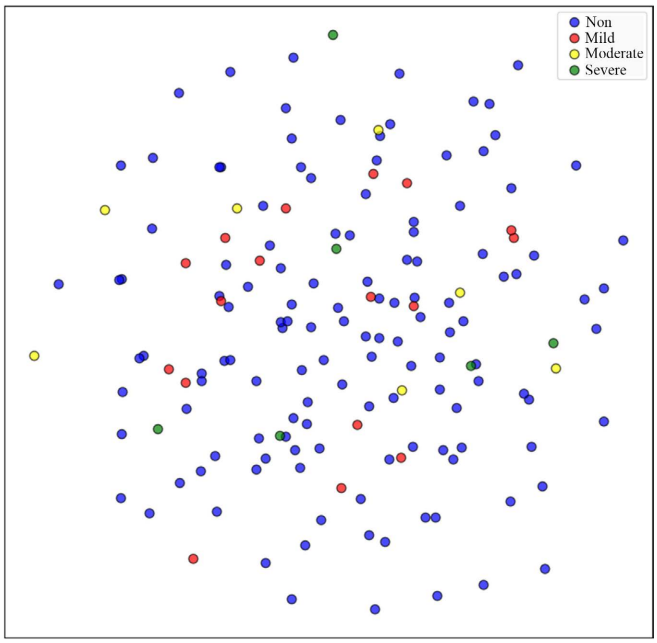

Figure 6. Sample distribution plot based on text modality
图 6. 基于文本模态的样本分布图

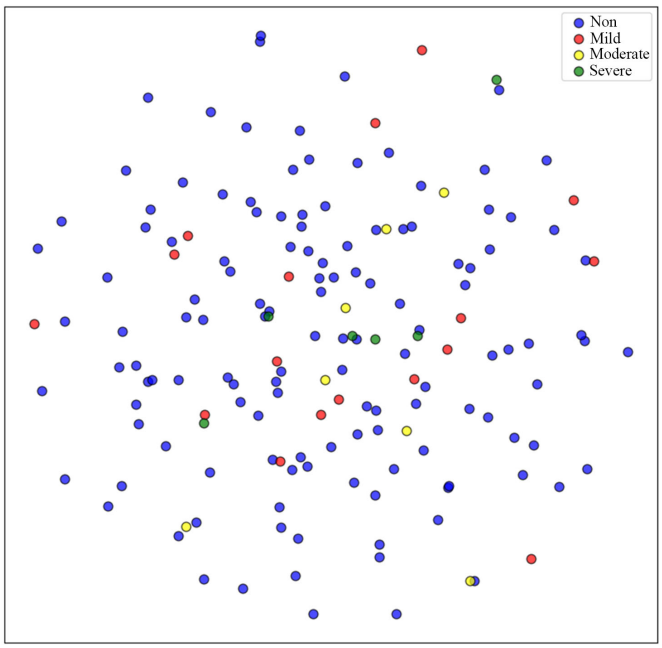

Figure 7. Sample distribution plot based on audio modality
图 7. 基于音频模态的样本分布图

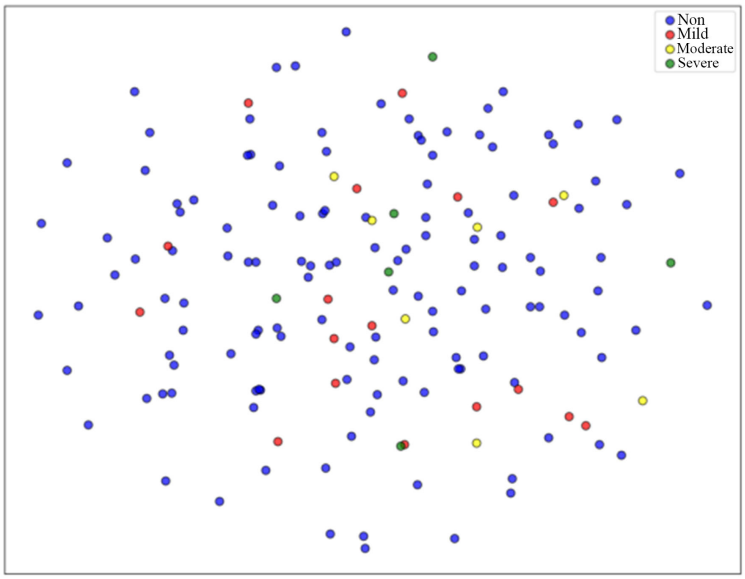

Figure 8. Sample distribution plot based on audio-text bimodal
图 8. 基于音频和文本模态的样本分布图

4.3. FCM 性能评估

簇族数量 k 和模糊化系数 m 的最佳参数组合如表 3 所示，EATD-Corpus 数据集融合特征维度较低、样本分布相对集中，25 个簇能够细致地划分数据结构。CMDC 数据集样本数据更加丰富，42 个簇有助于更好建模数据结构。将模糊化系数设置为 1.1 有助于将聚类标签映射为最终分类结果，更清晰地区分非抑郁、轻度、中度、重度四类。

Table 3. Fuzzy c-means clustering algorithm parameter configuration
表 3. 模糊 c 均值聚类算法参数配置

数据集	簇族数量 k	模糊化系数 m
EATD-Corpus	25	1.1
CMDC	42	1.1

为可视化 FCM 最优参数选取策略，以 EATD-Corpus 数据集为例，图 9 展示了 FCM 算法的超参数优化结果，以 F1 分数作为性能评价指标，对不同模糊系数 和簇数 组合下的分类效果进行可视化，在 $m=1.1$，$k=25$ 时取得了全局最优的 $F1=0.97$。在抑郁症检测识别中，轻度与中度症状表现出交叉特征，即既不完全健康也未达重度阈值，选择略微倾向“硬聚类”但仍保留少量交叠的 $m=1.1$，能够既清晰地区分各簇，又适度保留类别边界的模糊性，有助于对轻度、中度边缘样本进行正确分类。簇族数量设置为 25 可将样本空间划分为更多细粒度子区域，捕捉抑郁症状在音频与语言表达上的微细变异。

图 10 展示了对融合特征使用 FCM 后，各簇族分布情况及第一簇族样本点的分类情况。左侧子图使用不同颜色标识了 25 个簇族在投影空间中的聚集结构，可发现簇内同质性、簇间可分性以及局部连续性均得到较好体现，为后续的簇族－类别标签映射与分类提供了可靠的可视化支撑。图中用红色虚线圈出的为第一簇族，右侧子图对该簇族样本点进行局部放大，并按照真实四分类标签对样本点着色，同时统计簇内各标签的出现频次以实现“簇族－类别”映射。如图所示，第一簇族由 25 个样本点构成，且全部样本真实标签均为“重度抑郁”，因此该簇族被映射为类别 3，验证了所采用多数投票映射策略的有效性

和簇内标签一致性的高度聚集特征。

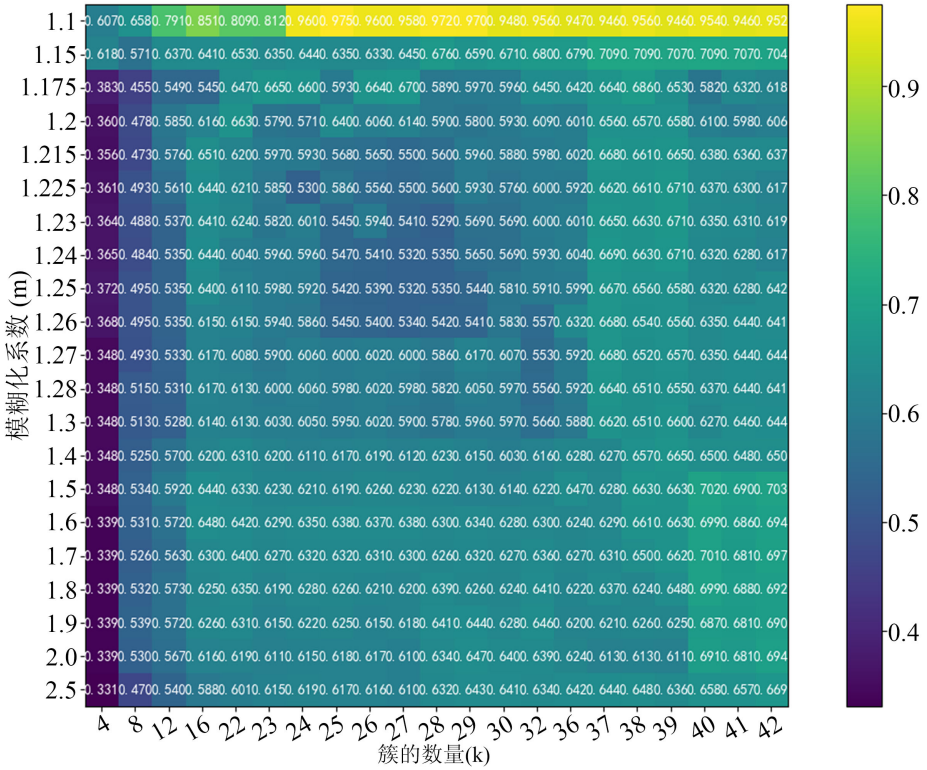

Figure 9. EATD-Corpus dataset: FCM algorithm hyperparameter optimization
图 9. EATD-Corpus 数据集使用 FCM 算法的超参数优化

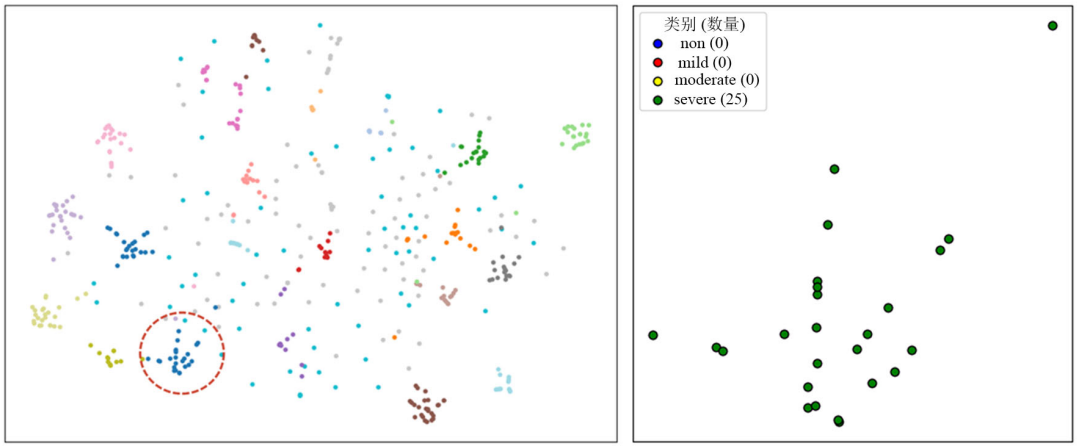

Figure 10. Distribution of clusters and classification of sample points in the first cluster
图 10. 各簇族分布情况及第一簇样本点的分类情况

4.4. 实验结果及分析

本文在 EATD-Corpus 和 CMDC 数据集上对本文提出的多模态抑郁检测模型进行实证研究，并与传统机器学习方法如支持向量机 SVM、随机森林 RF、决策树 DT 和深度学习模型 BiLSTM、BERT 进行比较。在模型性能评估中，本文采用准确率、召回率和 F1 得分作为主要指标。鉴于 EATD-Corpus 和 CMDC

数据集的四分类样本分布高度不均衡，单纯依赖准确率易造成模型偏向非抑郁类。召回率用于衡量抑郁类别的漏诊风险，F1 得分在漏诊与误诊之间取得综合平衡，更能反映模型对抑郁少数类的识别能力及其在不平衡场景下的稳健性。因此，本文评估性能主要关注 F1 得分。

Table 4. Experimental results of different modalities on the EATD-Corpus dataset
表 4. EATD-Corpus 数据集的不同模态实验结果

模态	模型	准确率	召回率	F1 得分
A	SVM	0.54	0.41	0.46
	RF	0.48	0.53	0.50
	DT	0.47	0.44	0.45
	BiLSTM	0.44	0.56	0.49
	VGGish-NetVLAD-GRU	**0.94**	**0.94**	**0.94**
T	SVM	0.48	**1.00**	0.64
	RF	0.61	0.53	0.57
	BiLSTM	0.53	0.63	0.57
	BERT	0.78	0.51	0.61
	RoBERTa	**0.96**	0.51	0.66
	RoBERTa-BiLSTM	**0.96**	**0.96**	**0.96**
A + T	GRU/BiLSTM	0.85	0.84	0.71
	本文多模态融合模型	**0.97**	**0.97**	**0.97**

如表 4 所示，在 EATD-Corpus 数据集上，传统机器学习模型在音频与文本单模态下整体性能有限，文本下 SVM 虽然召回率达 1.00，但准确度仅 0.48，表明模型通过过度预测阳性样本牺牲了分类特异性。本文构建的 VGGish-NetVLAD-GRU 音频模型和 RoBERTa-BiLSTM 模型分别获得 0.94 和 0.96 的 F1 得分，在单模态水平上优于各类基线模型，说明深度时序建模与预训练语言模型能够更充分地挖掘语音韵律变化和上下文语义依赖。本文提出的多模态融合模型进一步将 F1 得分提升至 0.97，表明音频与文本模态之间存在互补性，跨模态交互有助于增强抑郁相关特征的判别性并有效缓解类别不平衡带来的评估偏差。

Table 5. Experimental results of different modalities on the CMDC dataset
表 5. CMDC 数据集的不同模态实验结果

模态	模型	准确率	召回率	F1 得分
A	SVM	0.58	0.42	0.46
	RF	0.58	0.89	0.50
	BiLSTM	0.57	0.80	0.49
	VGGish-NetVLAD-GRU	**0.76**	**0.76**	**0.76**
T	SVM	0.48	1.00	0.53
	BiLSTM	0.53	0.63	0.67
	BERT	0.78	0.61	0.71
	RoBERTa	0.76	0.61	0.66
	RoBERTa-BiLSTM	**0.80**	**0.80**	**0.80**

续表

A + T	BiLSTM [10]	0.91	0.89	0.91
	本文多模态融合模型	**0.94**	**0.94**	**0.94**

表 5 为各模型在 CMDC 数据集上的实验结果，在音频模态下，VGGish-NetVLAD-GRU 模型的 F1 得分达到 0.76，优于 SVM 和 RF 等传统方法；在文本模态下，RoBERTa-BiLSTM 的 F1 得分为 0.80，优于 BERT 与 BiLSTM 等基线模型。在音频－文本双模态下，文献[10]提出的模型在 CMDC 上的 F1 得分为 0.91，本文提出的融合框架将准确率、召回率和 F1 得分提升至 0.94，说明在样本量更少且类别分布更不均衡的 CMDC 数据集上，本文方法依然能够稳定提升四分类任务的整体性能。

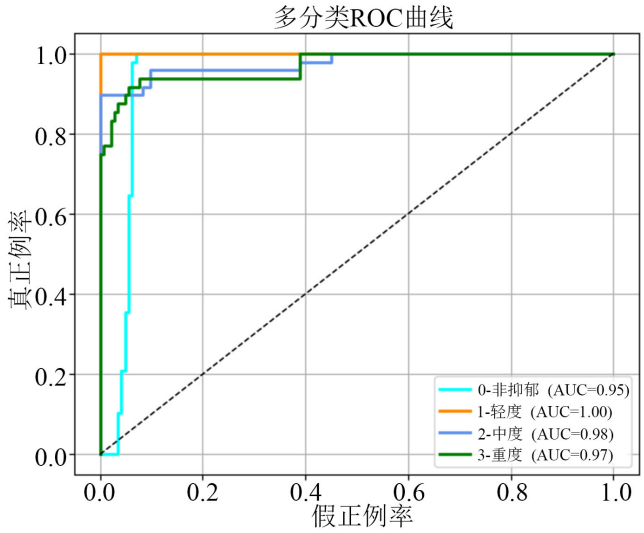

Figure 11. Multi-class ROC curve of the CMDC dataset
图 11. CMDC 数据集多分类 ROC 曲线

图 11 为 CMDC 数据集四分类任务的多分类 ROC 结果，非抑郁、轻度、中度和重度四个类别的 AUC 分别为 0.95、1.00、0.98 和 0.97，均保持在 0.95 以上，高 AUC 值表明模型在各抑郁等级与非抑郁类别之间具有较强的区分能力和稳定的判别性能。对比两个数据集的实验结果，EATD-Corpus 上准确率和 F1 得分均为 0.97，而在样本量更小、类别分布更不均衡的 CMDC 数据集上整体准确率降至 0.94，主要由 CMDC 中轻度抑郁样本极度稀缺、类别边界更为模糊所致，该结论与图 3 所示的 CMDC 数据集使用 SMOTE 技术前后特征分布变化相对应。由此可见，本文在训练阶段结合 SMOTE 进行少数类重采样，通过交叉注意力机制对齐音频与文本的互补信息，利用 FCM 实现模糊边界的软划分，使得模型在小样本、不平衡场景下能够保持对不同抑郁程度的高判别力和良好鲁棒性。

为检验模型是否真正学习到抑郁特征而非过拟合数据集背景噪声，验证本文模型在小样本场景下的跨数据集泛化能力，本文增加跨库实验。以 EATD-Corpus 作为训练集，在该训练集内部进行数据预处理并训练本文提出的多模态抑郁症识别模型，将训练完成的模型直接迁移至 CMDC 数据集进行测试，结果显示，在音频－文本双模态下准确率 0.89，召回率 0.90，F1 得分 0.88。相较于同库交叉验证结果，跨库指标出现一定幅度下降，说明不同语料在采集条件、受试者分布及文本转写风格等方面存在域偏移，但是跨库实验的准确率、召回率、F1 得分仍表现出较好的分类结果，表明本模型提出的交叉注意力融合机制与 FCM 算法能够在一定程度上学习到可迁移的抑郁相关判别特征。

4.5. 消融实验

Table 6. Results of the ablation experiment on the EATD-Corpus dataset
表 6. EATD-Corpus 数据集消融实验结果

交叉注意力	FCM	准确率(↑越大越优)	F1 得分(↑越大越优)	召回率(↑越大越优)	精确度(↑越大越优)
+	+	0.97 (基准)	0.97 (基准)	0.97 (基准)	0.97 (基准)
+	−	**↓0.07**	**↓0.07**	**↓0.07**	**↓0.07**
−	+	↓0.13	↓0.13	↓0.13	↓0.13
−	−	↓0.16	↓0.17	↓0.16	↓0.16
+	Softmax	**↓0.09**	**↓0.08**	**↓0.08**	**↓0.09**

为研究交叉注意力融合机制和 FCM 算法分别对本文提出的模型的贡献程度，本文在 EATD-Corpus 数据集上进行消融实验，实验结果如表 6 所示，其中“+”表示保留对应模块，“−”表示移除。以本文使用交叉注意力及 FCM 方法作为基线，将交叉注意力融合更改为简单拼接融合方式，对比说明交叉注意力融合的贡献。将 FCM 算法更改为 MLP 分类器，对比说明 FCM 算法的贡献。结果显示，仅使用交叉注意力融合机制时，准确率和其他指标下降 7 个百分点，表明没有 FCM 进行分类，模型难以对相似样本进行细粒度区分，导致整体性能下降。仅使用 FCM 进行分类时，虽然通过隶属度降低决策边界的硬度，但缺乏音频和文本数据的深度语义对齐，多模态信息融合不充分，聚类与分类效果仍受限。使用简单拼接融合方式和 MLP 分类器构成对照模型时，模型性能效果最差。该消融实验证明了交叉注意力融合机制与 FCM 算法在模型中的互补作用，交叉注意力负责多模态特征的深度对齐与增强，FCM 算法负责分类边界的模糊建模与细化，两者结合时可提高抑郁检测模型的准确性和 F1 得分。

为严格验证 FCM 相比与传统深度学习分类头的有效性，本文进一步补充“仅替换分类头”的对照实验，保持前端特征提取器(RoBERTa-BiLSTM + VGGish-NetVLAD + Cross-Attention)不变，仅将 FCM 层替换为标准的全连接 Softmax 层，以交叉注意力融合后的特征向量作为输入，通过全连接层映射至四分类，输出类别概率。实验结果表明，将 FCM 替换为 Softmax 层分类后，准确率下降 9 个百分点，F1 得分下降 8 个百分点，说明 FCM 算法能够在小样本且类别边界模糊的场景下形成更平滑的决策机制，降低硬判别分类头对多数类的偏置，提高对边缘样本的区分能力。

5. 结束语

本文针对多模态抑郁症检测领域中存在的单一模态表征不充分与模态间互补信息未能充分挖掘的关键问题，提出端到端深度融合框架：采用 VGGish-NetVLAD-GRU 模型捕获患者语调、语速等与情绪状态相关的长期依赖关系，增强了语音模态的表征完备性；采用 RoBERTa-BiLSTM 模型深入挖掘访谈转录中隐含的语义层级与上下文关联，增强模型对抑郁症状描述的敏感度。在多模态融合方面，使用交叉注意力融合机制融合不同模态间的互补信息，提升多模态特征的表征能力和模型整体性能。最后，在分类阶段引入 SMOTE 过采样技术，缓解少数类别样本匮乏所导致的过拟合风险，并结合 FCM 算法，刻画不同抑郁程度下患者特征的模糊边界，从而在抑郁症四分类任务中实现了漏诊率与误诊率的双重优化，提高了分类精度与 F1 得分。

参考文献

[1] 熊俊. 如何辨别抑郁症的表现[J]. 特别健康, 2019(21): 46.

[2] Pandey, A. and Vishwakarma, D.K. (2024) Progress, Achievements, and Challenges in Multimodal Sentiment Analysis Using Deep Learning: A Survey. *Applied Soft Computing*, **152**, Article 111206. https://doi.org/10.1016/j.asoc.2023.111206

[3] Shen, Y., Yang, H. and Lin, L. (2022) Automatic Depression Detection: An Emotional Audio-Textual Corpus and a Gru/Bilstm-Based Model. *ICASSP* 2022-2022 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Singapore, 23-27 May 2022, 6247-6251. https://doi.org/10.1109/icassp43922.2022.9746569

[4] 张亚洲, 和玉, 戎璐, 等. 基于上下文知识增强型 Transformer 网络的抑郁检测[J]. 计算机工程, 2024, 50(8): 75-85.

[5] 赵小明, 谌自强, 张石清. 基于跨模态特征重构与解耦网络的多模态抑郁症检测方法[J]. 计算机应用研究, 2025, 42(1): 236-241.

[6] Chen, Z., Wang, D., Lou, L., Zhang, S., Zhao, X., Jiang, S., *et al*. (2025) Text-Guided Multimodal Depression Detection via Cross-Modal Feature Reconstruction and Decomposition. *Information Fusion*, **117**, Article 102861. https://doi.org/10.1016/j.inffus.2024.102861

[7] Li, S., Xiao, Y. and Hu, S. (2025) A Depression Detection Method Based on Multi-Modal Feature Fusion Using Cross-attention. 2025 8*th International Conference on Advanced Algorithms and Control Engineering* (*ICAACE*), Shanghai, 21-23 March 2025, 1825-1831. https://doi.org/10.1109/icaace65325.2025.11019096

[8] Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T. and Sivic, J. (2016) NetVLAD: CNN Architecture for Weakly Supervised Place Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, 27-30 June 2016, 5297-5307. https://doi.org/10.1109/cvpr.2016.572

[9] Rovetta, S., Mnasri, Z., Masulli, F., *et al.* (2020) Emotion Recognition from Speech: An Unsupervised Learning Approach. *International Journal of Computational Intelligence Systems*, **14**, 23-35. https://doi.org/10.2991/ijcis.d.201019.002

[10] Zou, B., Han, J., Wang, Y., Liu, R., Zhao, S., Feng, L., *et al.* (2023) Semi-Structural Interview-Based Chinese Multimodal Depression Corpus towards Automatic Preliminary Screening of Depressive Disorders. *IEEE Transactions on Affective Computing*, **14**, 2823-2838. https://doi.org/10.1109/taffc.2022.3181210

[11] Wang, Y., Chen, S., Liu, J., *et al*. (2025) Unveiling Sex Difference in Factors Associated with Suicide Attempt among Chinese Adolescents with Depression: A Machine Learning-Based Study. *Journal of Mental Health*, **34**, 409-419.