Computer Science and Application 计算机科学与应用, 2025, 15(8), 21-33
Published Online August 2025 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2025.158194

Hans汉斯

融合多尺度结构与通道注意力机制的多模态情感识别研究

徐晓婧

河北地质大学信息工程学院，河北 石家庄

收稿日期：2025年6月30日；录用日期：2025年7月28日；发布日期：2025年8月5日

摘　要

基于现有生理信号情感识别研究中存在的特征表达能力不足、模态干扰严重及情感分析精度较低等问题。对此，本文提出一种基于脑电图(EEG)、心电图(ECG)与皮肤电(EDA)信号融合的多模态情感识别方法。针对生理信号特征表达能力弱的问题，设计了融合标准卷积与1D-Inception结构的多尺度特征提取模块(MSI-Block)，在提取丰富特征的同时控制参数复杂度；引入通道交互注意力机制提升关键模态响应，减少冗余干扰；并采用双向长短期记忆网络(BiLSTM)建模融合特征的时序信息。实验在DEAP数据集上验证了该方法的有效性，在Valence、Arousal和Valence-Arousal四分类任务中准确率分别达到90.72%、89.48%和83.62%，显著优于传统单模态与双模态方法，表明所提方法具有良好的情感识别性能与稳定性。

关键词

情感识别，多模态，多尺度特征提取，特征融合，双向长短期记忆网络

Multimodal Emotion Recognition Based on Multi-Scale Structures and Channel Attention Mechanism

Xiaojing Xu

School of Information Engineering, Hebei GEO University, Shijiazhuang Hebei

Received: Jun. 30th, 2025; accepted: Jul. 28th, 2025; published: Aug. 5th, 2025

Abstract

To address the limitations in existing physiological signal-based emotion recognition studies—such

文章引用: 徐晓婧. 融合多尺度结构与通道注意力机制的多模态情感识别研究[J]. 计算机科学与应用, 2025, 15(8): 21-33. DOI: 10.12677/csa.2025.158194

as insufficient feature representation, severe modality interference, and low emotion classification accuracy—this paper proposes a multimodal emotion recognition method that fuses electroencephalogram (EEG), electrocardiogram (ECG), and electrodermal activity (EDA) signals. To enhance the feature extraction capability for physiological signals, a Multi-Scale Inception Block (MSI-Block) is designed by integrating standard convolution with a 1D-Inception structure, enabling rich feature extraction while controlling model complexity. A channel-wise interactive attention mechanism is introduced to enhance the response of key modalities and suppress redundant interference. Furthermore, a bidirectional long short-term memory (BiLSTM) network is employed to model the temporal dependencies of the fused features. Experiments conducted on the DEAP dataset demonstrate the effectiveness of the proposed approach, achieving classification accuracies of 90.72%, 89.48%, and 83.62% in valence, arousal, and valence-arousal tasks, respectively. The results significantly outperform traditional unimodal and bimodal approaches, indicating that the proposed method provides robust and reliable emotion recognition performance.

Keywords

Emotion Recognition, Multimodal, Multi-Scale Feature Extraction, Feature Fusion, Bidirectional Long Short-Term Memory Network

Copyright © 2025 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

随着人工智能和情感计算的不断发展，如何使计算系统具备理解与响应人类情绪的能力，已成为人机交互研究的重要方向之一。情感不仅影响人的认知与行为决策，也是实现智能系统自然交互的关键因素。同时，情绪状态不仅直接影响个体的心理健康，还与多种生理疾病密切相关。因此准确识别和分类这些情绪维度具有重大的理论和实践意义。传统的情绪识别方法大多依赖于面部表情、语音和文本分析，但这些方法往往受到个体主观感知和环境因素的影响，难以准确反映个体的真实情绪状态[1]。相比之下，脑电(EEG)、心电(ECG)和皮肤电(EDA)等相关生理信号提供了一种更客观、实时的情绪监测手段。

然而，生理信号本身通常呈现非线性、非平稳及个体差异性强等特点，使得传统的时域或频域特征提取方法在建模情绪变化方面存在一定局限，表现为识别准确率不高、泛化能力弱等问题。近年来，深度学习方法被广泛应用于生理信号的特征建模任务，显著提升了情绪识别性能。但同时，深层神经网络通常存在模型参数庞大、训练成本高等问题，限制了其实用性。

为提升情感识别的鲁棒性和泛化能力，多数研究引入多模态融合策略，将脑电、心电、皮肤电等不同类型的生理信号进行联合建模，以挖掘各模态间的互补性特征[2]。Li 等人[3]研究发现，融合多模态生理信号能够显著提升跨主体情感识别的效果。然而，决策层融合方法通常需要为每个模态单独设计分类器，导致模型复杂度增加。相比之下，特征层融合方法在统一框架中实现多模态数据的联合建模，有助于减少参数冗余并提高模型训练效率。

然而不同模态生理信号之间存在信息互补性，如何高效提取并融合多模态的情绪判别特征仍面临诸多挑战[3] [4]。

1) 如何从众多信号通道中识别出对情绪具有显著表征能力的关键通道；

2) 如何高效提取能准确表征情感强度的判别性特征；

3) 如何设计有效的特征融合策略以提升模态间协同表达能力，并增强模型的泛化性能与鲁棒性[5] [6]。

针对上述挑战，本文提出一种融合多通道注意力与多尺度特征提取机制的多模态情感识别方法，主要工作包括：

1) 使用设计的多尺度时序特征提取模块对生理信号进行特征学习，相比较传统的 Inception 结构，在丰富特征表达的同时降低参数复杂度和计算量，更适合处理非平稳的生理信号。

2) 引入通道交互注意力机制和双向长短期记忆网络(Bi LSTM)，在特征层对不同模态通道间的信息进行自适应建模融合特征；对全局时序特征学习并通过全连接层实现情绪状态分类。

2. 相关工作

EEG、ECG 以及眼动信号均呈现明显的非稳态特性，缺乏稳定的波形结构。为有效挖掘其潜在信息，已有研究尝试从时频域和统计特征角度进行分析与建模。Jerritta 等人[7]指出高阶统计量(HOS)是一种有效的特征提取手段。他们通过情绪诱发实验收集面部肌电(fEMG)数据，并提取偏度、峰度等 HOS 特征及多个统计参数。使用 KNN 构建分类模型，对六种不同情绪的总识别率达到 69.5%。Li 等[8]基于心率变异性(HRV)信号提取多维特征，采用逻辑回归算法进行情绪识别，取得了 84.3%的准确率。对于皮肤电信号，Zhang 等[9]基于皮肤电信号，采用小波去噪和协方差特征选择，结合支持向量机(SVM)进行情绪识别，取得了一定识别效果。

然而，生理信号的个体差异性使得传统时频域及统计学特征提取方法难以全面反映，因而识别率较低且跨个体分类的泛化能力有限。深度学习方法在脑电(EEG)及外围生理信号的特征提取和情绪识别领域逐渐成为研究热点。Lopez 等[10]提出超复数多模态网络，有效融合 EEG 与外围生理信号，实现了优异的情绪识别性能。Zhang 等[11]结合图卷积网络和广义学习系统对脑电及外围生理信号进行多模态融合，显著提升了情绪识别准确率。Wu 等[12]则利用基于 Transformer 的自监督学习方法，实现了在数据稀缺条件下对多模态生理信号的高效联合表示学习。但是，深度学习在处理高维特征时常采用多层堆叠策略，虽然有效，但同时引发了模型复杂度和训练开销的显著增加，降低了模型的应用可行性。

Inception 结构通过在同一层中并行引入不同尺度的卷积核，如 1×1、3×3、5×5 等，实现对局部与全局信息的联合建模，既增强了模型对多尺度特征的表达能力，又显著降低了网络深度和参数冗余。该结构已被成功引入生理信号分析任务中。例如，Zhang 等人[13]提出了 EEG-Inception 网络，该模型基于 Inception-Time 架构设计，能有效提取 EEG 时间序列中的多尺度特征，在 BCI Competition IV 2a 与 2b 数据集上分别取得 88.4%和 88.6%的平均准确率，展现出优异的泛化性能。此外，Salami 等人[14]设计了 EEG-ITNet 模型，将 Inception 模块与因果卷积相结合，提升了模型对频谱与时序信息的提取能力，在多个数据集上分类准确率提升最多达 5.9%，同时保持较低的模型复杂度。这些研究表明 Inception 结构在处理复杂的多通道非稳态生理信号时具有良好的适应性与有效性。在多模态信号中采用多维卷积核会导致不同生理信号在卷积中相互干扰，因此，本文拟将 1D Inception 模块引入多模态生理信号情绪识别中，结合标准卷积结构以构建多尺度特征提取框架，以提升特征的表达能力，避免因模态间干扰而导致的特征可识别性下降。

在多模态情绪识别任务中，特征融合策略对于模型性能的提升起着至关重要的作用。当前主流的融合方法大致可分为三类：特征拼接融合、决策级融合以及特征层融合。其中，特征拼接融合通常将不同模态的原始数据或初级特征直接拼接输入统一模型，但该方法容易受到模态间尺度差异和冗余干扰的影响，导致融合特征的表达能力有限。决策级融合则在各模态独立建模后再融合决策结果，Li 等[15]采用了基于决策级融合的多模态情绪识别方法，将 EEG 和面部表情的分类结果进行加权融合，缓解了模态间干

扰问题，然而缺乏模态间深层次语义交互，难以充分发挥多模态数据协同建模的优势。特征层融合方法可在中间层实现不同模态特征的联合表达，Tripathi 等[16]提出基于深度神经网络的特征层融合方法，学习多模态联合表示提升情绪识别性能，在 MAHNOB-HCI 数据集上准确率达到 84.9%，但仍面临模态异构性强、情绪动态变化大等挑战，且静态融合策略难以捕获模态间通道依赖，易导致无关信息干扰关键特征表达，影响最终识别效果。引入注意力机制的特征融合方法被广泛认为是有效的，特别是通道交互注意力机制(Channel-wise Interactive Attention)。Du 等[17]设计 SCA-Net 模型，在融合阶段引入通道注意力模块，以增强多模态特征中更具判别力的通道响应。在 MAHNOB-HCI 数据集上的实验显示，SCA-Net 准确率达到 87.3%，F1 分数为 86.8%，均优于无注意力融合模型，验证了通道交互注意力机制在多模态融合中的有效性。

3. 方法

本文基于 EEG、ECG 和 EDA 信号，设计了一种多通道多尺度特征提取模型，整体结构由三部分组成：多尺度时序特征提取模块、通道交互注意力模块和分类决策模块。模型对三种生理信号进行有效的特征提取与融合，并最终通过全连接层输出情绪预测概率。该多通道多尺度特征提取模型的结构如图 1 所示，主要由三个功能模块构成。其中，多尺度时序特征提取模块(MSI-Block)用于从预处理后的信号中学习多尺度时序特征；通道交互注意力模块建模各生理信号中不同通道之间的依赖关系并分配权重；最后，使用 BiLSTM 网络进一步挖掘时序特征，并通过全连接层完成最终的分类任务。

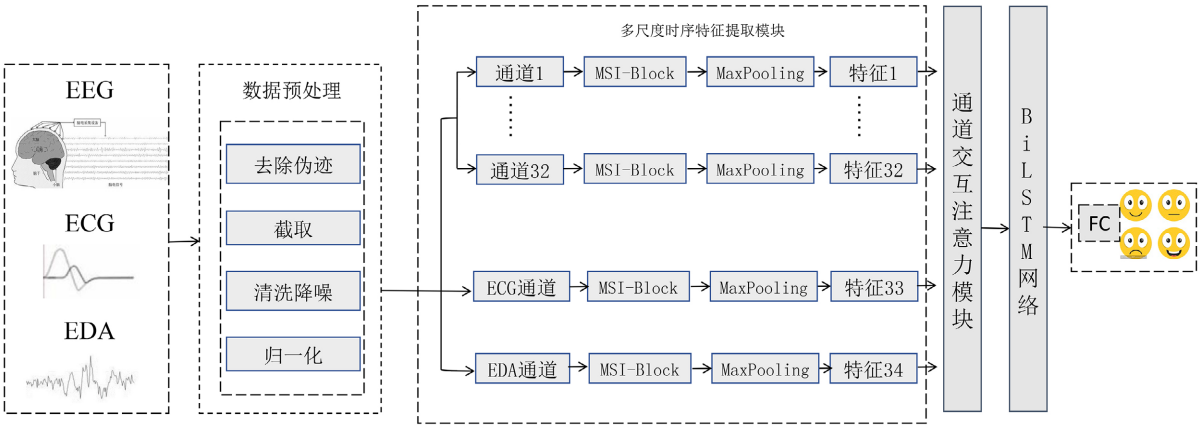

Figure 1. Overall architecture diagram of multimodal physiological signal emotion recognition model
图 1. 多模态生理信号情感识别模型整体架构图

3.1. 多尺度时序特征提取模块

在生理信号情感计算中，卷积核的卷积运算是基于 CNN 自动学习特征的核心且该方法被证明是有效的。然而，由于单层卷积的感受野有限，常通过增加卷积层数来扩展其覆盖范围，但这会导致参数量和训练开销提升，同时增加过拟合的风险。然而，小尺度卷积核的感受野也比较小，在网络加深的过程中容易过分“提取”局部特征，所以通常会在卷积层的后边加上池化层，从而减少输入的大小，增加卷积核的感受区域。但是上述方法中单尺度卷积核感受野提取的特征相对比较稀少，而信号的特征通常是多类型多维度的，标准的单尺度卷积无法提取多模态信号的多个尺度的特征信息，同时池化过程中容易丢失信号的部分特征信息。而传统 Inception 结构通过在单层卷积中并行采用多种尺度的卷积核，实现了多尺度感受野的覆盖，能够有效提取丰富的特征信息，同时保持较低的参数复杂度。因此，本文借鉴

Inception 结构的设计思想，构建了一种并行多分支的多尺度特征提取模块(Multi-Scale Inception Block, MSI-Block)。该模块由三个并行分支组成，其中前两个分支分别采用宽度为 d 和 2d 的卷积核，以获得不同感受野范围内的时序特征。第三个分支则通过采样宽度为 1.5d 的池化操作对原始数据进行降采样和通道映射，既有效保留了原始信息，又增强了特征表达的多样性。最终，三个分支的输出在通道维度上进行拼接，融合了多尺度卷积特征与原始数据特征，实现了比传统单尺度卷积更为丰富且具有更强表征能力的特征提取。

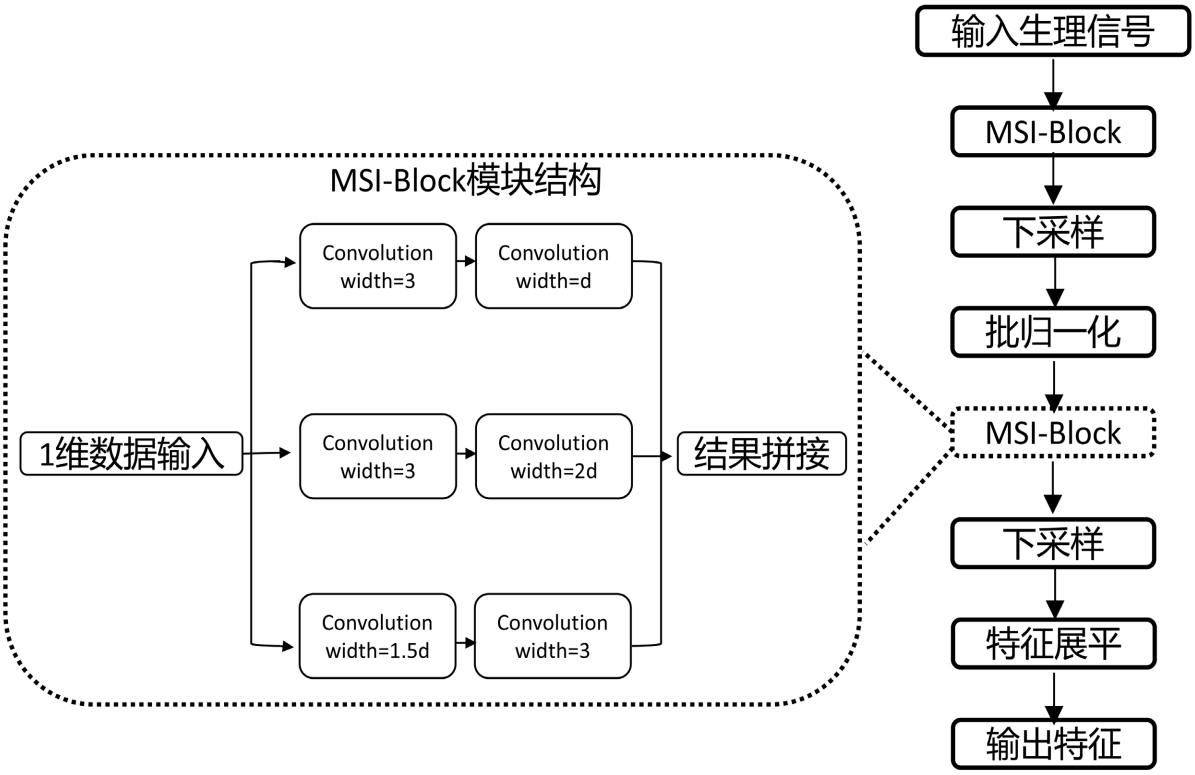

Figure 2. Overall architecture of MSI Block feature learning module
图 2. MSI-Block 特征学习模块整体架构

利用池化层和批标准化处理 MSI-Block 的卷积结果，以进一步减少参数的规模和降低训练所产生的代价大小。采用平均池化将计算结果采样后，利用批量标准化使损失函数变得更加平坦，加速学习过程，将批次内的特征调整为标准正态分布。由于单模块的参数规模和训练成本都不高，本文将两个 MSI-Block 模块堆叠起来，增加模块学习能力，中间用池化和批量标准化连接起来，组成本文使用的 MSI-Block 特色学习模块，如图 2 所示，通过特征展平最终输出一维特征，模块最终输出的特征是维度是时间和通道的二维特征。在使用 MSI-Block 特征学习模块时，单独向模块发送 EEG 的每个通道，ECG 信号和 EDA 信号的分布，以达到提取特征的目的，避免数据在不同的模态间相互干扰。

为进一步降低模型参数规模与训练计算开销，本文在多尺度特征提取过程中引入了池化操作与批量归一化机制，对经 MSI-Block (Multi-Scale Inception Block)提取的卷积特征进行优化处理。具体而言，首先通过平均池化对卷积结果进行下采样，以压缩特征空间、增强特征的鲁棒性；随后采用批量归一化将每一小批次中的特征分布标准化为近似正态分布，从而有效缓解内部协变量偏移问题，提升模型收敛速度并优化训练稳定性。

考虑到单个 MSI-Block 的参数量较小，计算开销较低，本文进一步将两个 MSI-Block 模块进行堆叠以增强网络的特征提取能力。两个模块之间通过池化和批量归一化连接，构成本文所提出的 MSI-Block 特征学习模块(如图 2 所示)。该模块最终输出的特征为时间维度与通道维度组成的二维表示，并在后续处理中通过展平操作转化为一维向量特征，便于与其他模态特征进行融合。在使用时，为避免不同模态数据之间的相互干扰，将 EEG 各通道、ECG 信号以及 EDA 信号分别独立输入 MSI-Block 特征学习模块，以实现各模态特征在其原始语义空间中的高效建模与表征。

3.2. 通道交互注意力模块

完成时序特征学习后的各种生理信号，需要有效地从空间上整合特征。如果只是简单拼接不同模态的特征，由于不同通道的生理信号特征存在差异，可能会出现特征间相互干扰，降低识别准确率的情况，所以本文引入了多模态生理特征自适应加权融合的通道交互注意力机制。通过引入通道交互注意力机制，模型可以自适应地分配不同通道的关注权重，从而突出关键模态特征，抑制冗余信息，提升特征融合的判别性。

为减少参数量，本文以 1 × 1 卷积替代全连接层，对通道描述向量进行权重编码。实验中共使用 34 个通道，包括 32 个 EEG 通道、1 个 ECG 通道和 1 个 EDA 通道，首先在通道维度对卷积特征矩阵执行最大池化与平均池化，以编码各通道的整体空间特征，从而获得两个维度为 $1 \times C$ 的通道描述向量(其中 C 表示通道数)。随后，这两个通道描述向量依次通过两层全连接网络，第一层用于降维以提取紧凑特征，第二层用于恢复维度并输出每个通道的权重表示，用于建模通道间的依赖关系。然后将两组权重向量进行相加得到通道特征权重向量 Z^*。最终，使用将相加后的权重信息向量与原始输入特征矩阵相乘，为每个特征通道赋予不同权重信息。

具体而言，$X \in R^{C \times T}$ 为输入的特征矩阵，其中 $C = 34$ 为多模态特征通道总数，T 表示每个通道对应的特征维度。首先针对输入特征 X 在时间维度上进行全局平均池化(Global Average Pooling, GAP)和全局最大池化(Global Max Pooling, GMP)，获得两个通道级描述向量：

$$F_{\text{avg}} = \text{GAP}\left(X\right) \in R^{C \times T} \tag{1}$$

$$F_{\max} = \text{GMP}\left(X\right) \in R^{C \times T} \tag{2}$$

将两个通道描述向量分别输入到由两层 1 × 1 卷积构成的权重编码网络，f_1 和 f_2 分别代表 1 × 1 的卷积和对权重信息的编码过程：

$$W_{\max} = f_1\left(\text{ReLU}\left(f_2\left(F_{\max}\right)\right)\right) \tag{3}$$

$$W_{\text{avg}} = f_2\left(\text{ReLU}\left(f_1\left(F_{\text{avg}}\right)\right)\right) \tag{4}$$

将两组权重向量相加并通过沿通道维度的 softmax 归一化，得到最终的通道特征向量 Z^*

$$Z^* = \text{softmax}\left(W_{\max} + W_{\text{avg}}\right) \in R^{C \times T} \tag{5}$$

注意力权重计算完成后，将权重向量 Z^*作用于特征矩阵 X 的通道维度，采用逐通道的逐元素乘法赋予各通道不同的重要性权重。由此，每个通道特征在赋权过程中综合考虑了其他通道的相关性。

3.3. 分类决策模块

生理信号属于连续的生物电信号，其特征片段之间存在显著的时间相关性，具有天然的时序特征。在时序信号建模中，当前状态往往受到历史状态的显著影响，不能将各个时刻的特征孤立对待。卷积神

经网络受其结构影响，对时间维度上的序列关系处理能力有限。在时间序列建模中，尽管门控循环单元(GRU)在参数量与训练速度方面具有一定优势，但考虑到情绪相关的生理信号往往存在复杂的非线性动态与长时依赖，本文引入长短期记忆网络(Long Short-Term Memory, LSTM)对特征进行时序建模，LSTM 通过引入遗忘门(forget gate)增强了长期记忆的保留能力，在特征提取的精度和表示能力方面表现更优。此外，相较于参数冗余的 Transformer 结构，LSTM 在中等规模数据集上具备更强的收敛稳定性和较低的过拟合风险。考虑到生理信号的时序方向具有不确定性，即时间依赖可能同时存在于前向和后向两个方向，传统的单向 LSTM 模型存在一定局限性。因此，本文选用 BiLSTM 作为本模型的时序特征建模单元，以兼顾表达能力与计算效率通过对融合后的多模态特征进行时序建模，以同时捕捉正向与反向的时间依赖关系，使每个时间步的特征向量都能融合来自前后两个方向的上下文信息，与多模态生理信号的特点相匹配。

本文设置的 BiLSTM 网络结构如图 3 所示，为了增强模型的非线性建模能力并缓解梯度消失问题，该网络设置为两层堆叠的 BiLSTM 结构，每层的隐藏单元数设置为 256，分别对应正向与反向的两个 LSTM 子网络，网络整体输出维度为 512。

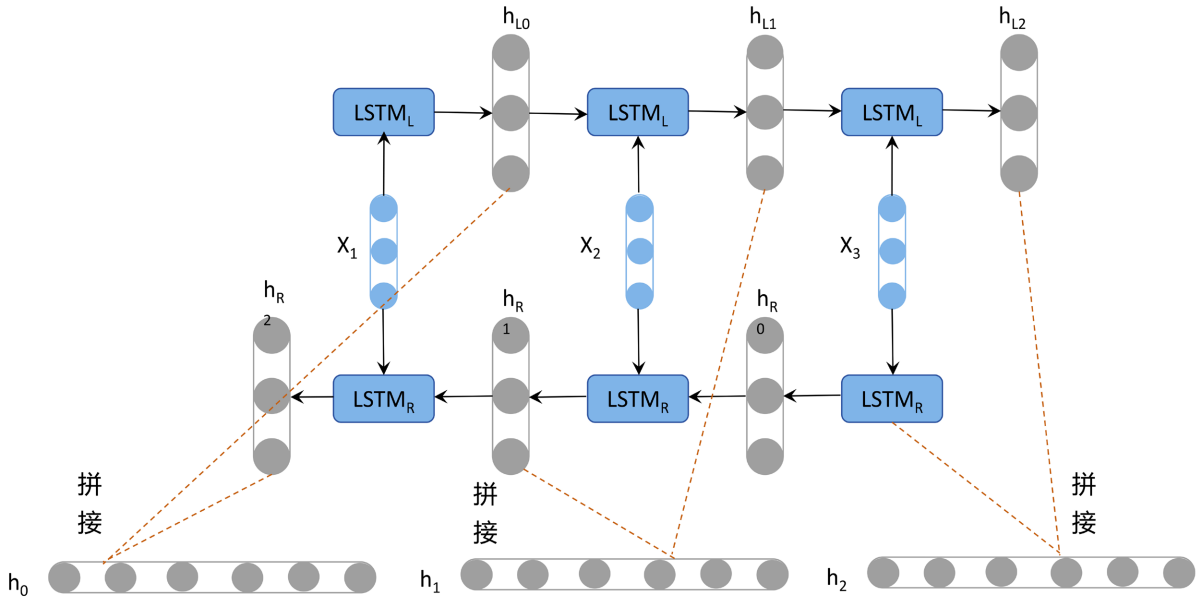

Figure 3. Schematic diagram of BiLSTM structure
图 3. BiLSTM 结构示意图

每个 LSTM 单元的前向计算公式如下：

$$\begin{cases} f_t = \sigma\left(W_f x_t + U_f h_{t-1} + b_f\right) \\ i_t = \sigma\left(W_i x_t + U_i h_{t-1} + b_i\right) \\ \tilde{c}_t = \tanh\left(W_c x_t + U_c h_{t-1} + b_c\right) \\ c_t = f \odot c_{t-1} + i_t \odot \tilde{c}_t \\ o_t = \sigma\left(W_o x_t + U_o h_{t-1} + b_o\right) \\ h_t = o_t \odot \tanh\left(c_t\right) \end{cases} \tag{6}$$

其中，f_t、i_t、$o_t \in R^d$ 分别表示遗忘门、输入门和输出门的激活值，$c_t \in R^d$ 表示单元记忆状态，$c_t \in R^d$，

$h_t \in R^d$ 为隐藏状态，符号 $\odot$ 表示 Hadamard 逐元素乘法，$\sigma(\cdot)$ 为 Sigmoid 激活函数。

双向输出经展平拼接后输入至全连接层(fully connected layer)，进一步整合时序特征。为缓解过拟合问题，在 BiLSTM 输出与全连接层之间引入 Dropout 层，丢弃概率设为 0.5。同时，为提高训练稳定性并加快收敛速度。最终，输出层采用 Softmax 函数将神经网络的输出转换为表示不同类别概率的向量。如公式(7)所示：

$$\hat{y}_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_i}}, i = 1, 2, \cdots, K \tag{7}$$

其中，$\hat{y}_i$ 为样本属于第 i 类情绪的预测概率，z_i 为第 i 类对应的网络输出，K 为情绪分类的类别数。

4. 实验

4.1. 数据集介绍

本研究主要选用 DEAP (Database for Emotion Analysis using Physiological Signals)和 AMIGOS (A Dataset for Affect, Personality and Mood Research on Individuals and Groups)两个公开的多模态情感识别数据库作为实验数据源。

DEAP 数据集，采集自伦敦玛丽女王大学 Koelstra 等研究人员，现已被广泛用于多模态情感识别的相关研究。该数据集通过播放音乐视频的方式诱发被试的情绪反应，采集了 32 名健康参与者的脑电(EEG)和多种外围生理信号(Peripheral Physiological Signals, PPS)，包括皮肤电(EDA)、呼吸(Respiration)、脉搏(BVP)、体温(Temperature)等。其中脑电信号来自 32 个通道，外围信号来自 8 个通道；参与者中括 16 名男性和 16 名女性，年龄范围为 19~37 岁，平均年龄为 26.9 岁。实验过程中，每位被试观看 40 段时长为 1 分钟的视频，并在自我情绪评估量表(Self-Assessment Manikins, SAM)上对效价(Valence)、唤醒度(Arousal)、支配度(Dominance)和喜好度(Liking)四个情绪维度进行 1~9 分的评分。数据库还包含前 22 名被试者的视频面部表情数据，可用于后续视觉模态研究。

本研究使用 DEAP 数据集提供的预处理版本，其原始信号已进行如下处理：将采样率从 512 Hz 降采样至 128 Hz，应用 4~45 Hz 带通滤波器以去除工频和低频干扰，并采用独立成分分析(ICA)去除眼电(EOG)伪迹。最终选取包含 32 个 EEG 通道、1 个 EDA 通道及 1 个心电(ECG)通道在内的 34 个通道数据。每条样本数据时长为 63 秒，其中包含 3 秒静息基线和 60 秒刺激段，用于构建多模态情绪特征。

为进一步验证所提出方法的泛化能力，本研究引入 AMIGOS 数据集进行验证迁移实验。AMIGOS 数据集由 Correa 等人发布，旨在支持在个体和群体层面上研究情绪、性格及情绪状态。该数据集收集了 40 名受试者在观看 16 段情绪诱导视频过程中产生的 EEG、EDA、ECG 等多模态生理信号，包含静息状态与观看状态两个阶段的记录。每位参与者在观看视频后填写 SAM 量表，对其当前的效价与唤醒状态进行主观评价。

4.2. 评估指标

为了全面且准确地评估模型性能，本文选用准确率(Accuracy)、精确率(Precision)、召回率(Recall)、F1 分数(F1-score)以及混淆矩阵五种常用的情感分析评估指标。各指标的具体定义如下：

1) 准确率(Accuracy)指所有样本中预测正确的样本数占总样本数的比例，计算公式如式(8)所示。

$$\text{Accuracy} = \frac{T_p + T_N}{T_p + F_p + T_N + F_N} \tag{8}$$

2) 精确率(Precision)指的是在所有被预测为正类的样本中，实际正确预测为正类的比例，公式如式(9)

所示。

$$\text{Precison} = \frac{T_p}{T_p + F_p} \tag{9}$$

3) 召回率(Recall)表示在所有真实为正类的样本中，被正确预测为正类的比例，具体定义见公式(10)。

$$\text{Recall} = \frac{T_p}{T_p + F_N} \tag{10}$$

4) F1-score 为 Precision 和 Recall 的调和值，见公式(11)。

$$\text{F1} = \frac{2T_p}{2T_p + F_N + F_p} \tag{11}$$

其中，T_P 是真正例，T_N 是真负例，F_P 是假正例，F_N 是假负例。

5) 混淆矩阵(Confusion matrix)，混淆矩阵也是一种有效的模型评估指标，能够更直观地展示数据集中的分类准确性。混淆矩阵以概率值和样本数量进行可视化。

4.3. 多尺度特征提取模块的有效性分析

尽管 MSI-Block 借鉴了 Inception 结构的多尺度并行卷积思想，但本文并未将传统 1DInception 作为对比对象，是考虑到其原始设计面向图像任务，直接迁移至一维信号需大幅调整结构，可能导致特征提取不稳定或计算资源浪费。因此，为了更清晰地验证 MSI-Block 相较于单尺度与浅层模型的优势，本文设计了消融实验，将优先选择传统特征方法与单尺度一维卷积网络进行对比。

本研究设计了三组对照实验。首先，选取差分熵(Differential Entropy, DE)作为手工提取的典型特征，结合高斯核支持向量机(SVM)进行分类，作为传统浅层方法的参考；其次，构建了一种由三层一维卷积网络组成的标准 1D-CNN 结构，直接对原始信号进行建模与分类；最后，MSI-Block 模块单独作为特征提取结构，通过其输出的特征接入全连接层与 Softmax 分类器进行分类，以验证该模块的特征建模能力。

实验在相同的数据集和评价标准下进行，结果如表 1 所示。与传统手工特征提取方法相比，MSI-Block 模块在效价和唤醒度两维度上的平均准确率分别提升了 18.98 个百分点和 20.05 个百分点；相较于固定尺度的标准 1D-CNN，平均准确率也分别提升了 5.72 个百分点和 7.03 个百分点。同时，MSI-Block 在跨被试实验中的 Macro-F1 更高，分别达到 73.01% (效价)和 72.94% (唤醒度)，表明该模块在不同主体间具有更强的泛化能力。尽管 MSI-Block 采用多分支卷积结构，其整体参数规模仍低于传统单尺度 CNN 模型，显示出更优的结构轻量性与计算效率。

Table 1. Comparison of accuracy between MSI block and other feature extraction methods (Unit: %)
表 1. MSI-Block 与其他特征提取方法的准确率对比(单位：%)

提取方法	Valence		Arousal	
	Acc	Macro-F1	Acc	Macro-F1
单尺度 1D-CNN	66.73	68.8	65.18	65.57
SVM	53.47	55.24	52.16	51.21
MSI-Block	72.45	73.01	72.21	72.94

4.4. 特征融合方法有效性分析

为验证本文所提出的基于通道交互注意力机制的特征融合方法在多模态情感识别任务中的有效性，

设计了三种融合策略的对比实验：特征拼接融合、决策级融合以及本文提出的注意力融合方法。考虑到生理信号在短时尺度内情绪相关特征具有一定波动性，结合前期实验经验与相关研究成果，本文将BiLSTM 模块的输入序列长度固定为 6。该设置在保证时序依赖提取能力的同时，有效控制了模型复杂度与计算成本，适用于本任务中的多模态融合与分类需求。特征拼接融合方法直接将各模态特征在特征维度拼接后送入统一的 BiLSTM 网络进行分类；决策级融合方法则将各模态特征分别输入独立的 BiLSTM 网络，分类结果通过最大值策略进行决策层融合；本文方法则引入通道交互注意力机制，在融合前动态建模各模态间的相关性，以增强融合特征的判别能力。实验结果如表 2 所示，在效价、唤醒度两个分类任务中，本文方法分别获得了 89.99%和 90.78%的平均识别准确率，显著优于其余两种方法。同时，该方法对应的 F1 分数分别为 80.37%与 89.21%，表明其在个体差异较大的情境下具有更强的稳定性与泛化能力。

Table 2. Comparison of results between channel interaction attention and other fusion methods (Unit: %)
表 2. 通道交互注意力和其他融合方法的结果对比(单位：%)

融合方法	Valence		Arousal	
	Acc	Macro-F1	Acc	Macro-F1
特征拼接融合	84.62	75.21	65.18	78.47
决策级融合	86.24	75.00	52.16	77.48
通道交互注意力融合	89.99	80.37	90.78	89.21

4.5. 多模态融合方法的有效性分析

为验证所提出多模态融合方法的有效性，本文在 DEAP 数据集上分别对三种生理信号(EEG、ECG、EDA)的单模态，EEG + ECG 双模态、EEG + EDA 双模态，以及 EEG + ECG + EDA 三模态融合方案进行了系统性对比实验。实验结果如表 3 所示。在实验设置方面，统一采用交叉熵作为损失函数，优化器选用 Adam 以加速收敛并提升稳定性。训练过程中采用 10 折交叉验证策略以增强模型的泛化能力，BatchSize 设置为 64，最大训练迭代次数为 125。

单模态实验结果显示，EEG 模态在四分类任务中的整体表现优于 ECG 与 EDA，尤其在 Valence-Arousal 四分类任务中取得了 79.27%的准确率，显著高于 ECG 的 56.39%与 EDA 的 35.73%，同时 EEG 单模态在各分类任务中的 Macro-F1 值亦优于其他单模态。这表明脑电信号在情感识别中的判别力更强，具备更高的信息表达能力，适合作为主模态用于情绪建模任务。

在模态融合实验中，双模态及三模态融合策略均表现出优于单模态的分类性能。具体而言，EEG + ECG 双模态在 Valence、Arousal 和 Valence-Arousal 四分类任务中的准确率分别提升至 89.47%、87.23%和 82.79%，Macro-F1 和 AUC 指标均高于 EEG 单模态，显示出更佳的跨主体识别鲁棒性。进一步引入 EDA 构建的三模态融合模型，在四项任务中分别取得了 90.72%、89.48%和 83.62%的准确率，并在各类别上获得更高的 Macro-F1 分数，充分体现出融合策略在多模态特征互补上的优势。

与 EEG 单模态相比，三模态融合在准确率上提升了 1.26%~4.35%，在 Macro-F1 指标上提升了 2.04%~5.12%；相较 EEG + ECG 双模态则进一步提高了 0.83%~2.25%。更为重要的是，三模态融合模型在不同被试间的表现更为稳定，Macro-F1 和 AUC 在 10 折交叉验证中的标准差显著低于其他模态组合，反映出其更优的跨被试泛化能力与鲁棒性。综合结果表明，引入 ECG 与 EDA 信号可有效弥补 EEG 在部分情绪状态下的感知不足，从而提升整体识别精度与跨主体的泛化能力。

Table 3. Comparison of experimental results between single modal, dual modal, and multimodal methods (Unit: %)
表 3. 单、双模态与多模态方法的实验结果对比(单位：%)

模态	Valence			Arousal			Valence-Arousal		
	Acc	F1	AUC	Acc	F1	AUC	Acc	F1	AUC
EEG	79.27	77.63	85.41	78.52	76.82	84.12	79.27	76.48	83.55
ECG	56.39	53.84	66.91	54.76	52.47	65.23	56.39	52.11	64.04
EDA	35.73	33.58	51.37	38.24	35.71	54.62	35.73	33.80	51.02
EEG + ECG	89.47	88.16	92.54	87.23	86.30	91.33	82.79	81.10	89.10
EEG + ECG + EDA	90.72	89.67	93.41	89.48	88.51	92.79	83.62	82.41	90.55

4.6. 实验结果与分析

为确保训练集、验证集与测试集之间划分的合理性与客观性，在上述实验中本文采用随机分割策略对 DEAP 数据集进行样本划分，并在划分过程中充分考虑了数据的代表性与类别平衡性。从完整数据集中随机选取 80%的样本作为训练集，其余 20%用于测试集。同时，为进一步缓解模型过拟合风险，训练过程中引入了交叉验证机制。通过对训练集进行多轮划分与验证，不仅增强了模型对不同数据子集的适应能力，也有效提升了模型评估结果的稳健性。此外，本文在样本划分过程中特别保证了情绪标签在各个子集中的分布尽可能均衡，从而保持数据集中不同情绪类别在训练集和测试集中的一致性。这一策略有助于确保模型在学习过程中不偏向某一类别，提高整体分类性能的公平性和泛化能力。

为了进一步验证模型在不同数据环境下的适应性与鲁棒性，本文引入 AMIGOS 数据集作为外部验证数据进行跨数据集测试。该实验设计有效提升了研究结果的可信度和方法的实用性。图 4 中展示了模型在 AMIGOS 数据集训练过程中的性能变化趋势。其中，图 4(a)与图 4(b)分别对应于效价(Valence)与唤醒度(Arousal)两个情感维度的训练过程。图中，蓝色虚线与实线分别表示训练集与验证集的识别准确率，红色虚线与实线则分别表示训练集与验证集的损失值。从图中可以观察到，模型在训练初期即展现出良好的收敛特性，准确率稳步提升，损失逐渐下降，且训练与验证曲线变化趋势一致，未出现明显的过拟合现象。实验结果表明，本文所提出的方法在情感识别任务中不仅具备良好的训练稳定性与收敛效率，同时也在多个评价指标上取得了较高的识别精度与鲁棒性。

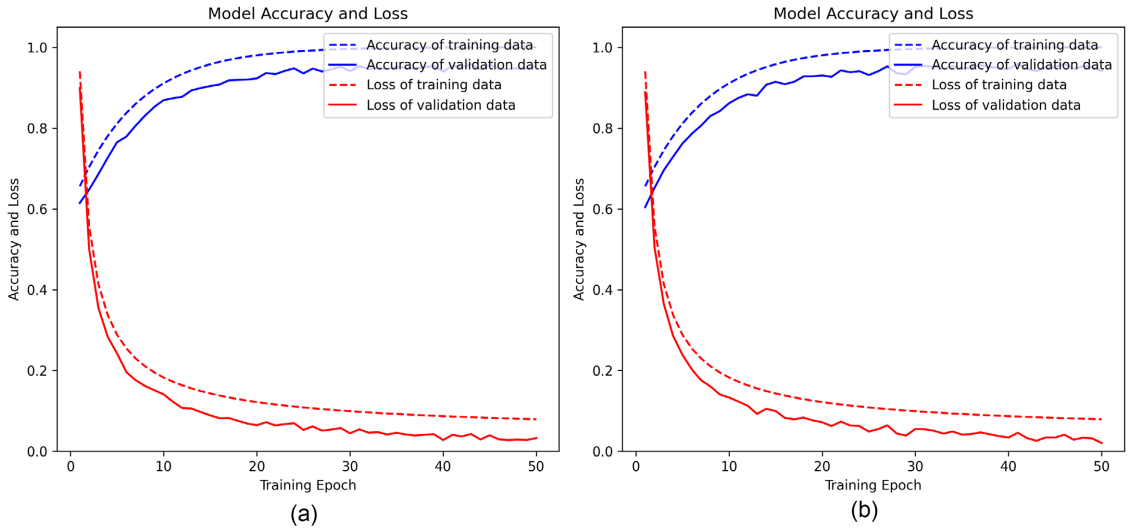

Figure 4. Training and validation accuracy and loss curves: (a) Valence dimension, (b) Arousal dimension
图 4. 训练和验证精度和损失曲线：(a) Valence 维度，(b) Arousal 维度

图 5(a)和图 5(b)分别通过混淆矩阵显示了模型在 AMIGOS 测试集上对 Valence 和 Arousal 两个分数的分类结果。从图 5 可以看出，该模型对 Valence 的分类效果最好，可达 92.03%；其次是对 Arousal 的分类效果，可达 91.37%。

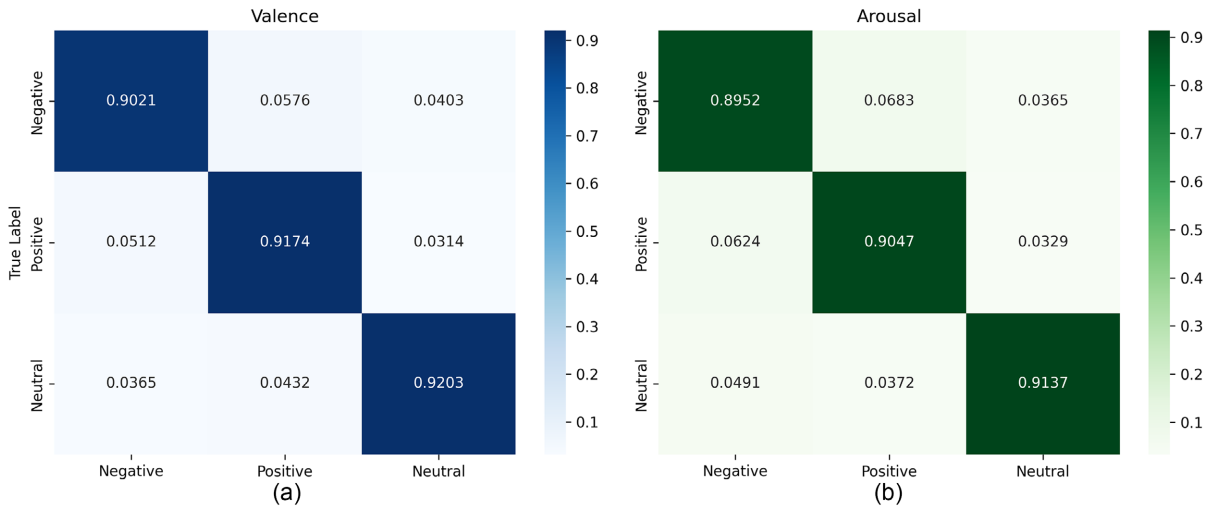

Figure 5. Confusion matrix for emotion classification on AMIGOS test set: (a) Valence dimension, (b) Arousal dimension
图 5. AMIGOS 测试集上情绪分类的混淆矩阵：(a) Valence 维度，(b) Arousal 维度

5. 总结

本文提出了一种基于多模态生理信号的情绪识别方法，融合了脑电图(EEG)、心电图(ECG)和皮肤电(EDA)三种模态信号，充分挖掘不同模态生理信号所蕴含的情绪相关信息。本文设计了改进的多尺度特征提取模块(MSI-Block)，结合通道交互注意力机制，在特征融合过程中动态建模模态间的相关性，有效提升了融合特征的判别能力。实验结果表明，该方法在 DEAP 数据集上的效价(Valence)、唤醒度(Arousal)和效价－唤醒度四分类任务中分别达到了 90.72%、89.48%和 83.62%的平均识别准确率，显著优于传统融合方法，验证了所提方法的有效性和先进性。

首先，本文构建了适用于生理信号特征提取的 MSI-Block 模块，该模块结合传统卷积与 Inception 结构，能够在不同尺度下提取丰富的局部时序特征，增强了模型对情绪变化特征的敏感性。其次，本文在多模态融合阶段引入通道交互注意力机制，从通道层级建模不同模态间的耦合关系。该机制通过显著提升关键通道的空间特征响应并抑制冗余或噪声通道的干扰，在融合过程中自适应分配各模态对情绪识别任务的贡献权重，从而实现更优的分类性能。再次，本文采用 BiLSTM 网络作为情绪识别的分类器，充分利用生理信号的时序信息，进一步提升了模型对情绪状态的建模能力。在不同情绪维度分类任务中的性能验证结果表明，该结构具有良好的泛化性能与学习能力。

除了对 DEAP 单个数据集进行实验比较外，本文还将公开的 AMIGOS 数据集作为外部验证集引入实验流程中。实验结果显示，本文方法在 AMIGOS 数据集上同样取得了较好的分类准确率，表明所提出的多模态融合方法具有良好的跨被试与跨数据集推广能力。最后，尽管本文在特征提取与融合机制方面进行了较为系统的研究，但仍存在一定的研究局限性。当前模型尚未深入探讨 EEG、ECG 与 EDA 信号中各类特征的生理学基础及其与情绪维度之间的具体关联。未来研究可进一步结合认知神经科学与心理生理理论，从特征解释性角度出发，探索情绪状态与生理响应之间的因果联系，并在模型设计中融合更多可解释性机制，以提升情感识别系统在实际场景下的可靠性和可推广性。

参考文献

[1] 权学良, 曾志刚, 蒋建华, 等. 基于生理信号的情感计算研究综述[J]. 自动化学报, 2021, 47(8): 1769-1784.

[2] 中文信息处理发展报告(2021)第十五章情感计算研究进展、现状及趋势[C]//中国中文信息学会. 中文信息处理发展报告(2021). 2021: 13.

[3] 章蕴晗. 多生理信号驱动的情绪识别关键技术研究[D]: [硕士学位论文]. 武汉: 华中师范大学, 2020.

[4] 焦蕊. 基于深度学习的情绪识别技术研究[D]: [硕士学位论文]. 北京: 中央民族大学, 2022.

[5] Wang, Z. and Wang, Y. (2025) Emotion Recognition Based on Multimodal Physiological Electrical Signals. *Frontiers in Neuroscience*, **19**, Article 1512799. https://doi.org/10.3389/fnins.2025.1512799

[6] Keelawat, P., Thammasan, N., Numao, M. and Kijsirikul, B. (2021) A Comparative Study of Window Size and Channel Arrangement on EEG-Emotion Recognition Using Deep CNN. *Sensors*, **21**, Article 1678. https://doi.org/10.3390/s21051678

[7] Lawhern, V.J., Solon, A.J., Waytowich, N.R., Gordon, S.M., Hung, C.P. and Lance, B.J. (2018) EEGNet: A Compact Convolutional Neural Network for EEG-Based Brain-Computer Interfaces. *Journal of Neural Engineering*, **15**, Article ID: 056013. https://doi.org/10.1088/1741-2552/aace8c

[8] Wang, L., Hao, J. and Zhou, T.H. (2023) ECG Multi-Emotion Recognition Based on Heart Rate Variability Signal Features Mining. *Sensors*, **23**, Article 8636. https://doi.org/10.3390/s23208636

[9] Lopez, E., Chiarantano, E., Grassucci, E. and Comminiello, D. (2023) Hypercomplex Multimodal Emotion Recognition from EEG and Peripheral Physiological Signals. 2023 *IEEE International Conference on Acoustics*, *Speech*, *and Signal Processing Workshops* (*ICASSPW*), Rhodes Island, 4-10 June 2023, 1-5. https://doi.org/10.1109/icasspw59220.2023.10193329

[10] Kumar, S.P., Selvaraj, J., Krishnakumar, R. and Sahayadhas, A. (2020) Detecting Distraction in Drivers Using Electroencephalogram (EEG) Signals. 2020 *Fourth International Conference on Computing Methodologies and Communication* (*ICCMC*), Erode, 11-13 March 2020, 635-639. https://doi.org/10.1109/iccmc48092.2020.iccmc-000118

[11] Fan, D., Liu, M., Zhang, X. and Gong, X. (2023) Human Emotion Recognition Based on Galvanic Skin Response Signal Feature Selection and SVM. arXiv: 2307.05383. https://doi.org/10.48550/arXiv.2307.05383

[12] Wu, W., Chen, X., Wang, Z., *et al.* (2020) Self-Supervised Representation Learning for Multimodal Physiological Signals. *Proceedings of the* 28*th ACM International Conference on Multimedia*, 12-16 October 2020, 2289-2297.

[13] Zhang, J., Wang, Y., Lin, Y., *et al.* (2021) EEG-Inception: A Deep Learning Framework for EEG-Based Emotion Recognition. *Frontiers in Neuroscience*, **15**, Article 674647.

[14] Salami, M. and Subasi, A. (2022) A Novel Deep Learning Model for EEG-Based Emotion Recognition Using Inception and Causal Convolution. *Computers in Biology and Medicine*, **140**, Article ID: 105045.

[15] Li, X., Chen, H., Zhang, J., *et al.* (2018) Multimodal Emotion Recognition Using Facial Expression, EEG and Eye Tracking Data. *Proceedings of the* 2018 *International Conference on Multimodal Interaction*, Boulder, 16-20 October 2018, 598-602.

[16] Tripathi, S., Acharya, S., Sharma, R., Mittal, S. and Bhattacharya, S. (2017) Using Deep and Convolutional Neural Networks for Accurate Emotion Classification on DEAP Data. *Proceedings of the AAAI Conference on Artificial Intelligence*, **31**, 4746-4752. https://doi.org/10.1609/aaai.v31i2.19105

[17] Du, J., Li, H., Zeng, J., *et al.* (2024) SCA-Net: A Self-Attention Based Channel-Aware Network for Multimodal Emotion Recognition. *IEEE Transactions on Affective Computing*, 9, 160926-160937.