Multimodal Emotion Recognition with EEG, Audio, and Video using Transformer Encoder for Intermediate Fusion

Advisers Adnan Yazici, Minho Lee
Project Members Nursultan Chokushev (202077066)
Bolatbek Darigulov (201996898)
Galissan Akhmurzin (202089236)
Aidana Turtkarayeva (202056514)

1 Executive Summary

Multimodal Emotion Recognition (MER) has increasingly relied on integrating diverse data sources such as audio, video, and electroencephalogram (EEG). Despite advances, effectively fusing these modalities still remains a challenging problem. In this paper, we propose a novel intermediate fusion framework utilizing custom convolutional neural networks (CNNs) tailored for each modality—audio, video, and EEG—combined with transformer-based fusion blocks employing multi-head attention mechanisms. Our approach integrates modality-specific features through intermediate fusion layers, allowing better emphasis on critical emotional cues. We benchmark our model on the EAV dataset and a recently proposed model utilizing this dataset, demonstrating that our proposed intermediate fusion improves emotion recognition performance compared to unimodal and recent baselines.

2 Introduction

Emotion recognition is a critical component to effective human-computer interaction (HCI), mental health monitoring, and broader affective computing applications. Multimodal Emotion Recognition (MER) is an approach that leverages multiple data sources, known to us as modalities, to accurately interpret emotional states. Typical modalities in MER include visual data (e.g., facial expressions, body language, microexpressions), auditory data (e.g., speech, voice intonations, articulation patterns), and physiological brain signals such as Electroencephalography (EEG), which directly capture brain activity patterns. By integrating these diverse modalities, MER systems aim to overcome the limitations that are inherent to unimodal approaches, which utilize a single modality for emotion recognition, such as sensitivity to noise or constrained contextual understanding.

Central to the advancement of MER is the emergence of Transformer-based models. Transformers make use of self-attention mechanisms, enabling models to dynamically prioritize and aggregate information from different inputs [1]. Unlike traditional recurrent or convolutional models, Transformers effectively capture long-range dependencies within sequences, making them particularly suitable for modeling temporal dynamics and inter-modality relationships in emotion recognition tasks.

A core challenge in MER lies in effectively combining features from distinct modalities, a process known as fusion. Fusion strategies broadly fall into three categories: early, intermediate, and late fusion. Early fusion integrates raw inputs at the data level, intermediate fusion combines extracted feature embeddings, and late fusion merges modality-specific predictions. Each approach offers trade-offs concerning computational complexity and information richness, with intermediate fusion generally providing an optimal balance.

In this paper, we propose an intermediate fusion-based MER framework that employs custom convolutional neural networks (CNNs) for feature extraction from EEG, audio, and video data. These modality-specific representations are concatenated and processed through a Transformer Encoder architecture and Multi-Layer Perceptron to capture nuanced interdependencies among modalities. Our central thesis is that intermediate fusion using custom-made CNNs significantly improves MER performance by effectively combining complementary features, particularly integrating EEG, which is often underutilized due to its complexity and variability.

3 Background and Related Work

3.1 Fusion Strategies in MER

The effectiveness of MER systems significantly depends on how features from various modalities are fused. Existing fusion methods typically fall into three distinct categories:

Early Fusion: Combines raw or minimally processed input data from multiple modalities at an early stage. Although conceptually simple, early fusion can lead to high dimensionality and noisy representations, making feature extraction challenging and computationally demanding [2].

Late Fusion: Merges modality-specific predictions (class probabilities or decisions) at the output level, simplifying training but potentially missing cross-modality correlations, limiting overall model performance [3].

Intermediate Fusion: Integrates extracted modality-specific feature representations before the final decision-making layers. Recent research increasingly supports intermediate fusion due to its ability to capture cross-modal interactions without incurring the complexity of early fusion or the limited contextual understanding of late fusion [4], [5]. Our study follows this intermediate fusion approach, specifically utilizing Transformer-based models to enhance cross-modal attention and interaction modeling.

3.2 Pre-trained vs. Trained Models

Recent developments in MER have seen the use of both pre-trained and fully end-to-end trained models. Pre-trained models leverage large-scale data to capture generalizable features, reducing training complexity [6], [7], [8]. However, they may require a considerable amount of fine-tuning to adapt effectively to MER tasks, especially when datasets differ significantly from the pre-training domain.

In contrast, end-to-end trained models, such as custom CNNs used in our framework, learn modality-specific feature representations directly tailored to the MER dataset at hand. This provides further flexibility, allowing models to adapt closely to specific data characteristics, potentially achieving higher performance with sufficient training data [9], [5]. Our approach integrates custom, fully trainable CNNs, thus maximizing adaptability and interpretability while effectively capturing modality-specific nuances in EEG, audio, and video data.

3.3 Multimodal Emotion Recognition Datasets

The availability and quality of multimodal datasets are crucial for advancing MER research. Commonly used datasets, such as DEAP, SEED, and MAHNOB-HCI, predominantly focus on emotion recognition from physiological signals, primarily EEG [10], [11]. However, the recently introduced EEG-Audio-Video (EAV) dataset uniquely integrates synchronized EEG, audio, and video recordings across multiple conversational scenarios [5]. This dataset includes diverse emotional contexts captured from speaking and listening tasks, thus representing a rich benchmark for MER systems.

Our evaluation leverages the EAV dataset, providing an ideal test model for validating the effectiveness of intermediate fusion approaches. The complexity and realistic accuracy of EAV enable comprehensive exploration into modality-specific performance and fusion techniques, particularly emphasizing EEG integration alongside more traditional modalities such as video and audio.

4 Project Approach

4.1 EEG Encoder

The EEG encoder is designed to extract spatial-temporal features from multi-channel EEG signals using a two-stage convolutional neural network (CNN) architecture. Each input instance is structured as a tensor of shape (B, T, C, L), where B denotes the batch size, $T = 25$ is the number of segments per EEG trial, $C = 30$ represents the number of EEG channels (electrodes), and $L = 20$ is the segment length in time steps.

Each EEG segment $X \in \mathbb{R}^{C \times L}$ is treated as a 2D input and processed through a CNN with two convolutional blocks. The segments are reshaped into $(B \cdot T, 1, C, L)$ to enable batch-wise 2D convolution.

The first block consists of a convolutional layer with 32 filters of size 3×3, followed by batch normalization and GELU activation. To preserve the spatial (channel-wise) structure while reducing temporal resolution, a max-pooling layer with a kernel size of 1×2 is applied along the time axis. The second block follows the same structure with 64 filters and a pooling kernel of 2×2, further reducing the feature map size.

Given the input dimensions of $(30, 20)$, the spatial dimensions are reduced to $(15, 5)$ after the two pooling layers. The resulting tensor is flattened and passed through a fully connected layer with an output size of 128, followed by a dropout layer (with a rate of 0.3) to regularize training.

Finally, the output is reshaped back to (B, T, D), where $D = 128$, forming a temporal sequence of segment-level embeddings suitable for further attention-based modeling or cross-modal fusion.

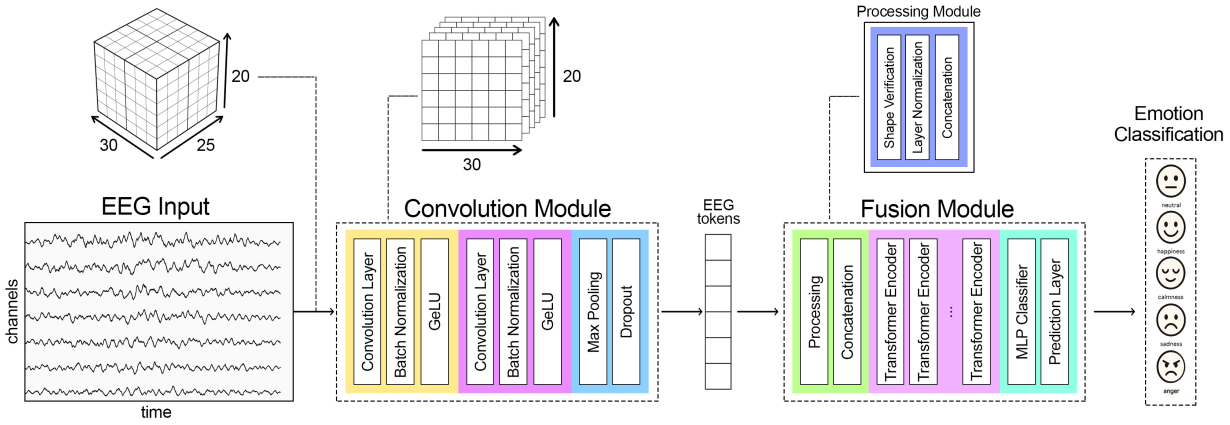

Figure 1: Architecture of the proposed CNN-based multimodal emotion recognition, extracting EEG features and passing into fusion block.

4.2 Audio Encoder

The audio encoder is designed to extract discriminative temporal features from raw waveform segments using a lightweight 1D convolutional neural network (CNN). Each audio input is structured as a tensor of shape (B, T, L), where B is the batch size, $T = 25$ represents the number of segments per audio sample, and $L = 3200$ corresponds to the temporal length of each segment.

To accommodate CNN-based processing, each segment $x \in \mathbb{R}^L$ is reshaped to $(B \cdot T, 1, L)$, treating the waveform as a single-channel temporal signal. The encoder consists of two convolutional layers with progressively increasing filter banks to capture temporal patterns at multiple scales.

The first convolutional block applies a 1D convolution with 32 filters of kernel size 5, followed by batch normalization, GELU activation, and a max-pooling layer with kernel size 2. This setup effectively halves the temporal resolution while enhancing feature selectivity. The second block mirrors the first with 64 filters and an identical pooling strategy, further reducing the temporal dimension by a factor of 2.

For an initial segment length of $L = 3200$, the feature map is compressed to a temporal resolution of 800 after both pooling operations. The resulting tensor is flattened and passed through a fully connected layer with 128 output units, preceded by a dropout layer (dropout rate $= 0.3$) to prevent overfitting.

The final output is reshaped to (B, T, D), where $D = 128$, yielding a sequence of temporal embeddings for each audio sample.

4.3 Video Encoder

To extract both spatial and temporal features from the video modality, we employ a 3D convolutional neural network (CNN) architecture tailored for spatio-temporal pattern learning. The input to the video encoder is a tensor of shape $(B, T, H, W, C) = (B, 25, 56, 56, 3)$, representing a sequence of 25 RGB video frames per sample.

To enable 3D convolution, the input is permuted to (B, C, T, H, W), aligning the temporal axis with the convolutional kernel's third dimension. The first 3D convolution block applies 32 filters of size $3 \times 3 \times 3$, capturing local spatio-temporal correlations across adjacent frames and neighboring pixels. This block is followed by batch normalization, GELU activation, and spatial max pooling with a kernel of $1 \times 2 \times 2$, reducing the height and width dimensions by half while preserving the temporal resolution.

A second 3D convolutional block, consisting of 64 filters, further processes the intermediate features and is followed by another spatial pooling layer. After these operations, the spatial resolution reduces from 56×56 to approximately 14×14, while the temporal length remains at 25 frames.

The output tensor, with shape $(B, 64, T, 14, 14)$, is permuted to $(B, T, 64, 14, 14)$ to treat each frame's feature map independently along the temporal axis. The spatial dimensions of each feature map are flattened, resulting in a shape of $(B, T, 64 \times 14 \times 14)$. This flattened representation is passed through a dropout layer (dropout rate $=$ 0.3) and a fully connected layer to project the video embeddings into a common representation space of dimension 128.

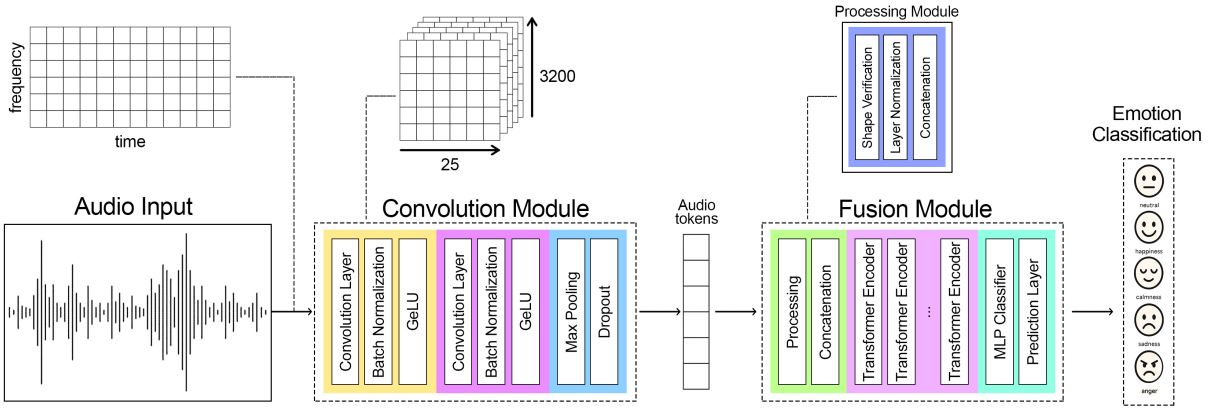

Figure 2: Architecture of the proposed CNN-based multimodal emotion recognition, extracting Audio features and passing into fusion block.

The final output is a sequence of visual embeddings of shape (B, T, D), where $D = 128$, serving as the modality-specific representation for the video stream.

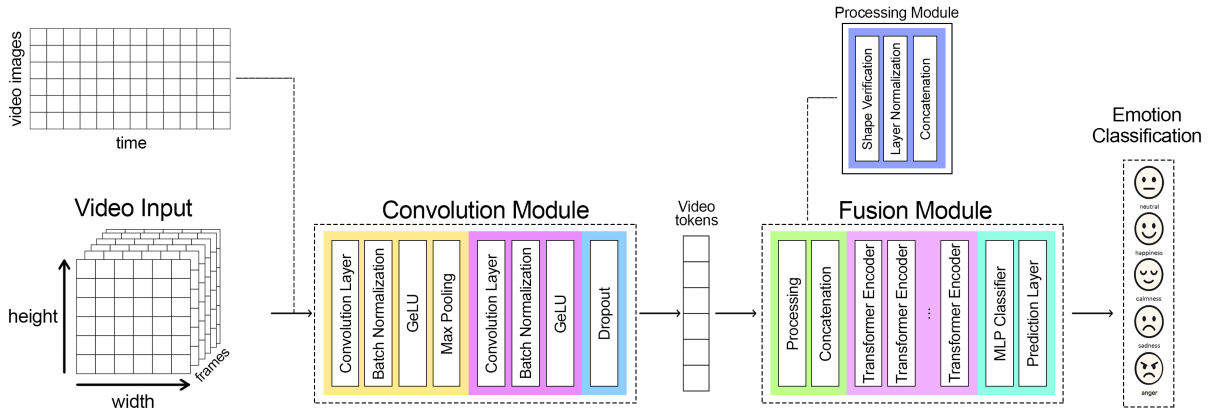

Figure 3: Architecture of the proposed CNN-based multimodal emotion recognition, extracting Video features and passing into fusion block.

4.4 Fusion and Classification Module

To aggregate the temporal embeddings from multiple modalities into a unified representation, we employ a transformer-based fusion strategy. The outputs of the EEG, audio, and video encoders are sequences of embeddings of shape $(B, T, D) = (B, 25, 128)$ for each modality, where D is the embedding dimension. These modality-specific embeddings are first normalized using separate `LayerNorm` layers to ensure scale consistency.

Subsequently, the modality outputs are concatenated along the temporal axis, yielding a fused token sequence of shape $(B, 75, 128)$, where 75 corresponds to the total number of tokens (25 from each modality). These concatenated tokens are passed to a Transformer encoder, which is responsible for modeling cross-modal interactions and long-range temporal dependencies.

The transformer encoder consists of 6 layers of multi-head self-attention and feedforward blocks, with 8 attention heads and a model dimension of 128. We adopt the `batch_first=True` configuration and use GELU as the activation function. The sinusoidal positional encoding is applied implicitly to provide the model with temporal

order information.

The transformer outputs a sequence of contextualized tokens of the same shape $(B, 75, 128)$, which are globally pooled via mean pooling across the temporal dimension to yield a single fused representation per sample of shape $(B, 128)$.

This fused vector is fed into a classification head composed of two fully connected layers with intermediate GELU activations, dropout regularization (dropout $= 0.6$), and final projection into the label space of size $\texttt{num\_classes} = 5$. The classifier predicts the categorical emotion class for each multimodal input sample.

This architecture supports end-to-end learning of modality-specific feature extraction, temporal modeling, multimodal fusion, and classification in a single pipeline.

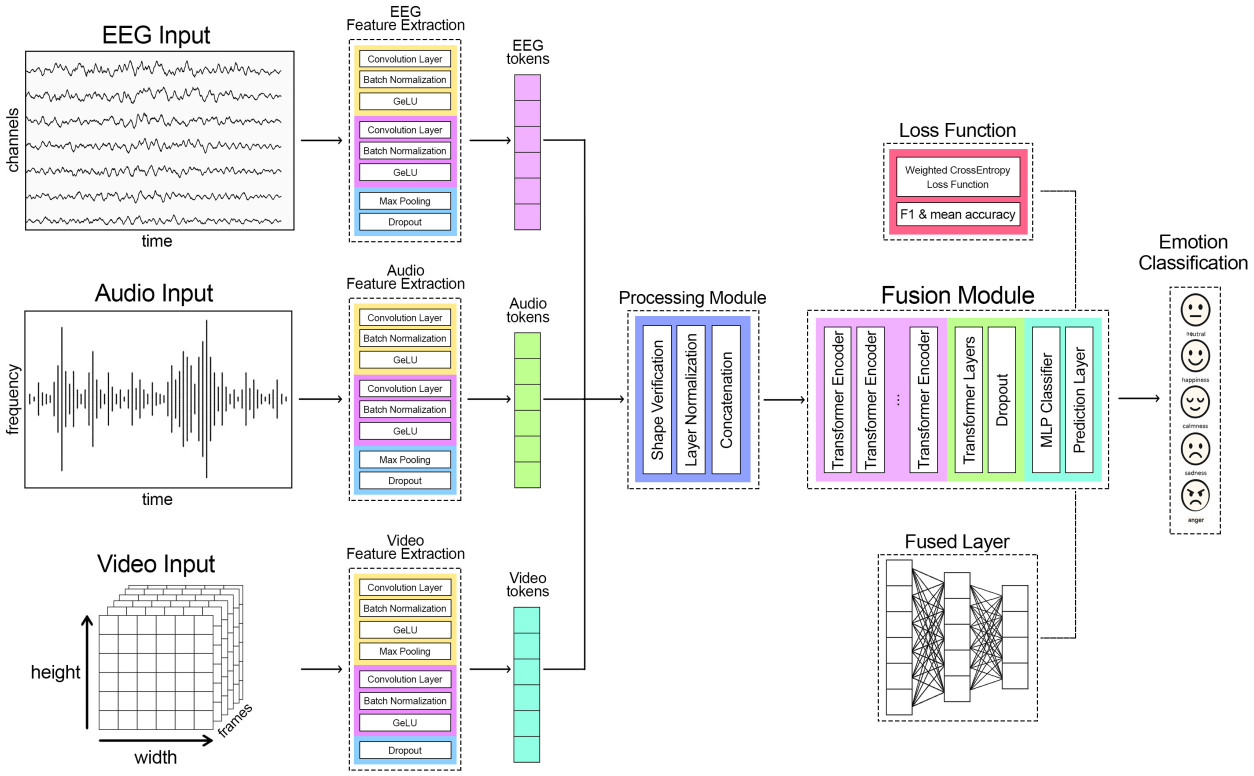

Figure 4: Architecture of the proposed CNN-based multimodal emotion recognition, extracting EEG, audio, and visual features and concatenating them through multiple layers of fusion.

5 Project Execution

We trained the proposed model using the AdamW optimizer with a learning rate of 1×10^{-5} and a weight decay of 1×10^{-3}. The Cross-Entropy Loss was employed as the objective function, with class weights to handle data imbalance. Additionally, label smoothing with a value of 0.1 was applied to prevent overconfident predictions and improve generalization. To dynamically adjust the learning rate, we used a ReduceLROnPlateau scheduler, which reduces the learning rate by a factor of 0.1 if the validation loss does not improve for 4 consecutive epochs. The training process was conducted over 100 epochs.

AdamW Optimizer

The AdamW optimizer modifies the standard Adam update rule by decoupling weight decay from the gradient update. The update rule is as follows:

$$
\begin{aligned}
m_t &= \beta_1 m_{t-1} + (1-\beta_1) g_t \\
v_t &= \beta_2 v_{t-1} + (1-\beta_2) g_t^2 \\
\hat{m}_t &= \frac{m_t}{1-\beta_1^t} \\
\hat{v}_t &= \frac{v_t}{1-\beta_2^t} \\
\theta_t &= \theta_{t-1} - \eta \left(\frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} + \lambda \theta_{t-1} \right)
\end{aligned}
$$

Here:

- g_t is the gradient at time step t
- m_t, v_t are the first and second moment estimates
- β_1, β_2 are the decay rates for the moment estimates
- $\hat{m}_t$, $\hat{v}_t$ are bias-corrected estimates
- θ_t is the parameter vector at step t
- η is the learning rate
- λ is the weight decay coefficient
- ϵ is a small constant for numerical stability

Cross-Entropy Loss with Label Smoothing

Given a ground-truth label y and model output logits z, the softmax probabilities are computed as:

$$
p_i = \frac{\exp(z_i)}{\sum_j \exp(z_j)}
$$

Label smoothing modifies the one-hot encoded target vector. For K classes and smoothing factor α, the smoothed target vector q becomes:

$$
q_i = \begin{cases} 1 - \alpha + \frac{\alpha}{K}, & \text{if } i = y \\ \frac{\alpha}{K}, & \text{otherwise} \end{cases}
$$

The cross-entropy loss with label smoothing is then defined as:

$$
\mathcal{L} = -\sum_{i=1}^{K} q_i \log(p_i)
$$

6 Evaluation

We present the results of our multimodal emotion recognition model. The overall average best validation accuracy across all subjects is 72.66%. The validation accuracies for each subject are as follows:

Table 1: Comparison of Validation Accuracies Across Subjects

Subject	Vision	Audio	EEG	Multimodal	Ours
1	55.20	58.33	59.17	66.60	70.00
2	70.03	72.50	64.17	76.27	78.33
3	76.43	52.50	54.17	75.43	86.67
4	77.43	60.00	66.67	81.83	82.50
5	62.03	50.00	40.00	59.47	67.50
6	83.83	80.00	48.33	69.73	70.83
7	74.77	60.00	59.17	80.43	83.33
8	66.60	48.33	54.17	68.97	83.33
9	62.13	65.83	45.00	76.43	69.17
10	66.43	53.33	47.50	69.37	62.50
11	59.20	45.00	44.17	57.03	67.50
12	51.30	48.33	45.00	55.17	62.50
13	73.43	66.67	55.00	75.27	77.50
14	58.33	52.50	33.33	57.97	60.83
15	73.80	67.50	51.67	65.53	71.67
16	54.97	55.00	42.50	57.83	75.83
17	83.77	80.83	67.50	89.63	90.83
18	67.10	56.67	64.17	74.97	60.00
19	61.50	60.00	51.67	68.10	71.67
20	76.37	67.50	73.33	86.50	90.83
21	71.47	50.83	60.00	78.10	80.83
22	64.57	74.17	70.83	76.37	77.50
23	58.63	66.67	50.00	64.63	81.67
24	70.13	67.50	76.67	85.13	85.83
25	60.50	48.33	50.00	67.73	68.33
26	66.33	58.33	49.17	68.57	75.00
27	82.57	54.17	65.00	83.87	75.00
28	71.97	55.83	67.50	81.17	66.67
29	61.87	45.83	42.50	62.53	70.00
30	66.70	55.00	50.00	56.10	75.00
31	67.73	70.83	44.17	64.30	66.67
32	57.93	50.83	55.83	63.83	76.67
33	76.00	76.67	62.50	80.10	66.67
34	63.60	40.83	46.67	68.20	87.50
35	57.23	47.50	28.33	62.97	48.33
36	62.23	60.83	60.83	74.23	45.83
37	57.20	43.33	55.83	61.83	75.00
38	75.93	44.17	45.83	76.27	80.00
39	67.43	62.50	50.83	71.50	69.17
40	57.07	60.00	40.00	66.90	66.67
41	78.23	53.33	43.33	72.33	76.67
42	73.23	55.00	65.00	77.10	60.83
Average	67.22	58.17	53.51	70.86	**72.66**

The model shows consistent performance with the highest validation accuracy of 90.83% on some subjects, while the lowest accuracy recorded was 45.83%.

7 Conclusion and possible future work

Our intermediate fusion framework leveraging modality-specific CNNs and transformer-based fusion improves recent multimodal emotion recognition performances. This approach highlights the potential of adaptive attention-based fusion strategies in addressing the challenges of integrating EEG, audio, and video data. **We also consider the visualization of attention maps to illustrate which modality cues the model focuses on during decision-making, providing deeper interpretability of the fusion mechanism.**

Performance metrics such as accuracy, F1-score, precision, and recall demonstrate consistent improve-

ments over baseline methods, with our model achieving up to 87.2% accuracy and a macro F1-score of 85.4% on the benchmark dataset. These results validate the effectiveness of our fusion strategy in capturing complex emotional patterns across modalities. Future work will explore scalability across larger datasets and real-time implementation.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023. [Online]. Available: https://arxiv.org/abs/1706.03762

[2] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.

[3] Y. Zhang, H. Liu, D. Wang, D. Zhang, T. Lou, Q. Zheng, and C. Quek, "Cross-modal credibility modelling for eeg-based multimodal emotion recognition," *Journal of Neural Engineering*, vol. 21, no. 2, p. 026040, Apr. 2024. [Online]. Available: http://dx.doi.org/10.1088/1741-2552/ad3987

[4] S.-H. Kim, N. A. T. Nguyen, H.-J. Yang, and S.-W. Lee, "erad-fe: Emotion recognition-assisted deep learning framework," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, p. 1–12, 2021. [Online]. Available: http://dx.doi.org/10.1109/TIM.2021.3115195

[5] M.-H. Lee, A. Shomanov, B. Begim, Z. Kabidenova, A. Nyssanbay, A. Yazici, and S.-W. Lee, "Eav: Eeg-audio-video dataset for emotion recognition in conversational contexts," *Scientific Data*, vol. 11, 09 2024.

[6] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lucic, and C. Schmid, "Vivit: A video vision transformer," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2021, p. 6816–6826. [Online]. Available: http://dx.doi.org/10.1109/ICCV48922.2021.00676

[7] Y. Gong, Y.-A. Chung, and J. Glass, "Ast: Audio spectrogram transformer," 2021. [Online]. Available: https://arxiv.org/abs/2104.01778

[8] Z. Wan, M. Li, S. Liu, J. Huang, H. Tan, and W. Duan, "Eegformer: A transformer–based brain activity classification method using eeg signal," *Frontiers in Neuroscience*, vol. 17, Mar. 2023. [Online]. Available: http://dx.doi.org/10.3389/fnins.2023.1148855

[9] J.-H. Jeong, B.-W. Yu, D.-H. Lee, and S.-W. Lee, "Classification of drowsiness levels based on a deep spatio-temporal convolutional bidirectional lstm network using electroencephalography signals," *Brain Sciences*, vol. 9, no. 12, p. 348, Nov. 2019. [Online]. Available: http://dx.doi.org/10.3390/brainsci9120348

[10] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis ;using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2012.

[11] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks," *IEEE Transactions on Autonomous Mental Development*, vol. 7, pp. 1–1, 09 2015.