Cascade Multi-Modal Emotion Recognition Leveraging Audio-Video and EEG Signals

Renato Esposito[1], Vincenzo Mele[1], Stefano Verrilli[1], Stefano Minopoli[1], Lorenzo D'Errico[2], Laura De Santis[3] and Mariacarla Staffa[1,*]

[1]*University of Naples "Parthenope", Naples, Italy*
[2]*University of Naples "Federico II", Naples, Italy*
[3]*University of Salerno, Salerno, Italy*

Abstract

Emotion recognition is essential for improving human-computer interaction, but single-modality approaches often face challenges in accurately capturing the complexity of human emotions. To overcome these limitations, we introduce a novel multimodal system that combines audio, video, and electroencephalogram (EEG) data. The system employs two deep learning models: an audio-video classifier utilizing hybrid fusion for analyzing speech and facial expressions, and a Feature-Based Convolutional Neural Network (FBCCNN) designed to process EEG signals. These models are integrated through a meta-model that uses logistic regression to combine their predictions. The system is capable of classifying four emotions—happiness, sadness, anger, and neutral—and outperforms single-modality methods, particularly in handling more complex emotional states.

Keywords

Emotion recognition, multimodal classification, EEG signal processing, fusion strategies

1. Introduction

Emotion recognition has become a crucial area of study in human-computer interaction [1, 2], with significant implications for interpersonal relationships, perception, and decision-making [3]. As automated systems become increasingly integral to daily life, the precise detection and classification of emotions have grown in significance, thereby underscoring the importance of research in this domain.

Traditional single-modality approaches, such as analyzing facial expressions [4], voice patterns [5, 6], or EEG signals [7, 8], often struggle to capture the complexity of human emotions [9]. Emotions are inherently multimodal, manifesting through various physiological and behavioral channels simultaneously. Relying on one modality can lead to incomplete or inaccurate assessments [10], particularly in real-world contexts where environmental factors may degrade signal quality or introduce noise in individual channels.

Recent research has demonstrated the potential of multimodal approaches to improve recognition accuracy [11, 12, 13]. By integrating information from multiple sources, these systems can leverage complementary features and compensate for weaknesses in individual modalities. However, significant challenges persist in feature extraction, synchronization, and handling noisy or incomplete data [14, 15].

Deep learning has shown promise in addressing these challenges [16], with Convolutional Neural Networks (CNNs) and other architectures achieving success in multimodal feature extraction and classification [17, 18, 19]. These approaches can automatically learn relevant features from raw data,

EMPATH-IA: Workshop on EMpowering PAtients THrough AI, Multimedia, and Explainable HCI: Innovations in Personalized Healthcare, in conjunction with the 16th Biannual Conference of the Italian SIGCHI Chapter (CHITALY 2025), 6–10 October 2025, Salerno, Italy

[*] Corresponding author.

[†] These authors contributed equally.

renato.esposito001@studenti.uniparthenope.it (R. Esposito); vincenzo.mele001@studenti.uniparthenope.it (V. Mele); stefano.verrilli001@studenti.uniparthenope.it (S. Verrilli); stefano.minopoli@collaboratore.uniparthenope.it (S. Minopoli); lorenzo.derrico@unina.it (L. D'Errico); ldesantis@unisa.it (L. D. Santis); mariacarla.staffa@uniparthenope.it (M. Staffa)
0000-0001-8044-8224 (L. D'Errico); 0009-0008-7403-2765 (L. D. Santis); 0000-0001-7656-8370 (M. Staffa)

potentially capturing subtle emotional cues that traditional methods might miss. However, effectively integrating diverse modalities remains an active research area requiring innovative solutions.

To address these challenges, we propose a novel cascade multimodal system that integrates audio, video, and electroencephalogram (EEG) inputs for emotion recognition. Our primary contribution lies in the development of an effective meta-model integration approach that combines existing specialized models for different modalities. Specifically, for audio-video processing, we adopt the established model from Zhang et al. [6], which has demonstrated strong performance in multimodal emotion recognition through hybrid fusion strategies. For EEG signal processing, we leverage the Feature-Based CNN (FBCNN) approach described by Pan and Zheng [20].

Our approach differs from previous work in several key aspects:

1. We implement a two-stage cascade architecture that first processes individual modalities through specialized models before combining their outputs at a meta-level
2. We leverage state-of-the-art deep learning architectures already tailored to each modality unique characteristics
3. We utilize the existing hybrid fusion strategy for audio-visual processing from [6] that captures both low-level interactions and high-level semantic relationships
4. We adopt the FBCNN [20] specifically designed to effectively process EEG signals
5. Our main innovation is the development of a meta-model integration approach using logistic regression that intelligently weighs predictions from each modality to produce superior classification results

The proposed system offers key advantages: robust synchronization of modalities, effective feature extraction through specialized architectures, and an interpretable yet powerful meta-model that combines predictions. Our experimental results show significant improvements over single-modality systems, particularly in classifying complex emotional states that have traditionally been difficult to recognize.

The system is capable of classifying four primary emotions: happiness, sadness, anger, and neutral, with high accuracy across varied conditions. This approach represents an important step toward more naturalistic and robust emotion recognition systems that can function effectively in real-world human-computer interaction scenarios.

The remainder of this paper is organized as follows: Section 2 details the materials and methods, including the architecture of both the audio-video and EEG models; Section 3 describes the multimodal classifier and integration approach through our meta-model; Section 4 outlines the experimental procedure and datasets used; Section 5 presents the results and provides a detailed discussion of our findings; and Section 6 concludes with a summary of contributions and directions for future research.

2. Materials and Methods

This section presents the foundational components of our multimodal emotion recognition system. We first provide an overview of the system architecture 2.1, followed by detailed descriptions of the specialized models for each modality and their integration through the meta-model approach 2.3.

2.1. System Architecture Overview

Our proposed system follows a cascade architecture comprising three main components:

1. Audio-Video Emotion Recognition Model: A specialized deep learning model that processes both audio and video data, extracting complementary features from speech and facial expressions;
2. EEG-based Emotion Recognition Model: A Feature-Based Convolutional Neural Network (FBCNN) designed to analyze EEG signals and extract emotion-relevant patterns from brain activity;
3. Meta-Model Integration: A logistic regression-based model that takes predictions from the two specialized models as input and produces the final emotion classification.

This architecture allows each modality to be processed by models specifically designed for their unique characteristics, before combining their outputs at a higher level. The complete system is capable of classifying four primary emotions: happiness, sadness, anger, and neutral states.

2.2. Audio-Video and EEG architectures

Audio-Video Model The audio-video emotion recognition component is based on the work of Zhang et al. [6], which processes both audio and video streams through dedicated neural networks before combining them through a fusion strategy. For our implementation, we adopt a late fusion approach as described in [21], which has demonstrated effective performance in multimodal emotion recognition tasks.

As illustrated in Figure 1, the video stream is processed through EfficientFace [22], a lightweight yet powerful convolutional neural network designed specifically for facial expression recognition. Simultaneously, the audio stream undergoes processing through a series of one-dimensional convolutional blocks that extract spectral and temporal features from speech signals. The outputs from these specialized networks are then passed through transformer blocks that implement a cross-modal attention mechanism. This allows each modality to benefit from complementary information in the other stream:

- The Audio-Video transformer uses audio features as queries to attend to video features
- The Video-Audio transformer uses video features as queries to attend to audio features

The attended feature maps undergo pooling operations before being concatenated and fed into a classification head that produces emotion predictions based on the combined audiovisual information.

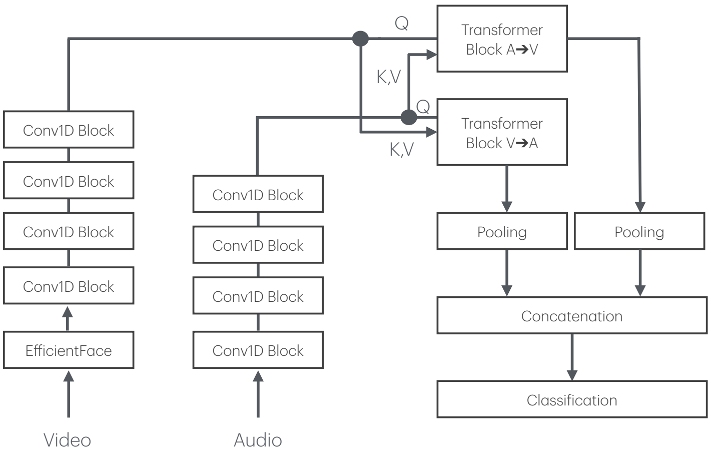

Figure 1: Diagram showing architecture used for the Audio-Video model following a late-fusion strategy.

EEG Model For the EEG modality, we implement the FBCNN (Feature-Based Convolutional Neural Network) approach described in [20]. This architecture is specifically designed to process the unique characteristics of EEG signals in the context of emotion recognition.

As shown in Figure 2, the FBCNN model first divides the EEG signals into multiple frequency bands corresponding to established brain wave patterns: alpha (8-14 Hz), beta (14-31 Hz), gamma (31-49 Hz), and theta (4-8 Hz). Each frequency band is processed separately through dedicated convolutional layers that extract spatial-temporal features from the corresponding brain activity patterns.

The architecture employs multiple convolutional layers with varying filter configurations to capture features at different scales and abstraction levels. The extracted features from all frequency bands are then concatenated and processed through a series of fully connected layers, ultimately producing emotion classification outputs.

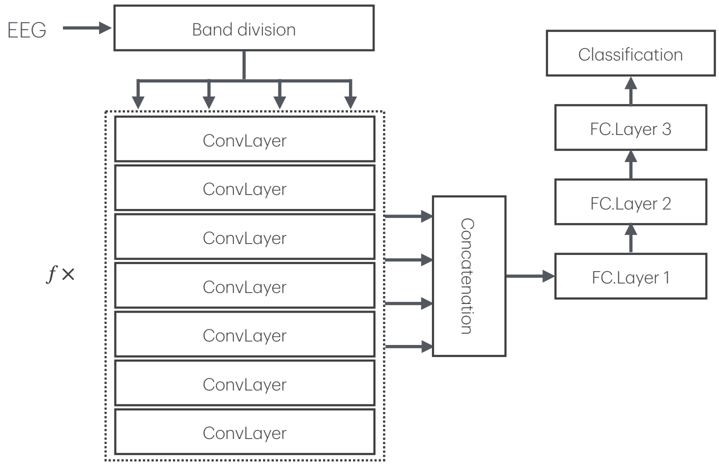

Figure 2: Main components of FBCCNN architecture.

This band-specific processing is particularly advantageous for emotion recognition, as different emotional states have been shown to manifest themselvesin distinct frequency bands of brain activity. For example, gamma activity (31-49 Hz) has been demonstrated to increase during the processing of emotionally salient stimuli, particularly in the prefrontal and temporal regions of the brain [23, 24].

For both models, we employ a reduced set of four emotion categories: neutral, happy, angry, and sad. These categories provide a balanced representation of the primary emotional dimensions according to the Russell Circumplex Model of Affect [25], covering both positive and negative valence as well as high and low arousal states.

2.3. Meta-model integration

The meta-model serves as the integrative component of our cascade architecture, combining the predictions from the audio-video and EEG models to produce a final emotion classification. This approach follows the stacking ensemble method [26], where the outputs of base classifiers become input features for a higher-level model.

As illustrated in Figure 3, the meta-model receives predictions from both specialized models and applies a logistic regression function to determine the final classification. This process involves several key steps:

Data synchronization A critical challenge in multimodal emotion recognition is ensuring temporal alignment between different data streams. We employ a batch-based synchronization strategy that creates label-matched data pairs between the audio-video and EEG modalities. This procedure ensures that both models are exposed to consistent emotional content despite coming from different datasets.

The synchronization process creates emotional bins based on the class labels, allowing samples from different modalities to be paired according to their emotional content rather than requiring strict temporal alignment (a detailed description of the synchronization strategy is reported in Sec.3.2 and Alg.1). This approach provides semantic consistency between modalities while accommodating the reality that our training data comes from separate specialized datasets.

Meta-Feature Creation The predictions from the audio-video and EEG models serve as meta-features for the final classification stage. These predictions, represented as logits (pre-softmax outputs) for each emotion category, capture the confidence levels of each specialized model regarding the emotional content of the input.

By using these prediction vectors as features, the meta-model can learn which modality tends to be more reliable for specific emotional states and weigh their contributions accordingly. This approach is more sophisticated than simple averaging or voting schemes, as it can adapt to the strengths and weaknesses of each modality.

Logistic Regression for Final Classification The final element of our system is a logistic regression classifier, which receives the meta-features as input and outputs the final emotion prediction. We chose logistic regression for this task because of its interpretability, computational efficiency, and effectiveness in combining predictive signals from multiple sources. The logistic regression model learns the optimal coefficients for the predictions for each modality, effectively determining their relative contributions to the final decision. This approach balances the strengths of deep learning models for modality-specific feature extraction with the interpretability and reliability of traditional machine learning techniques for the final integration step. The entire process results in a robust multimodal emotion recognition system that takes advantage of the complementary nature of audiovisual and neuro-physiological signals, providing more accurate and reliable emotion classification than any single modality could achieve independently.

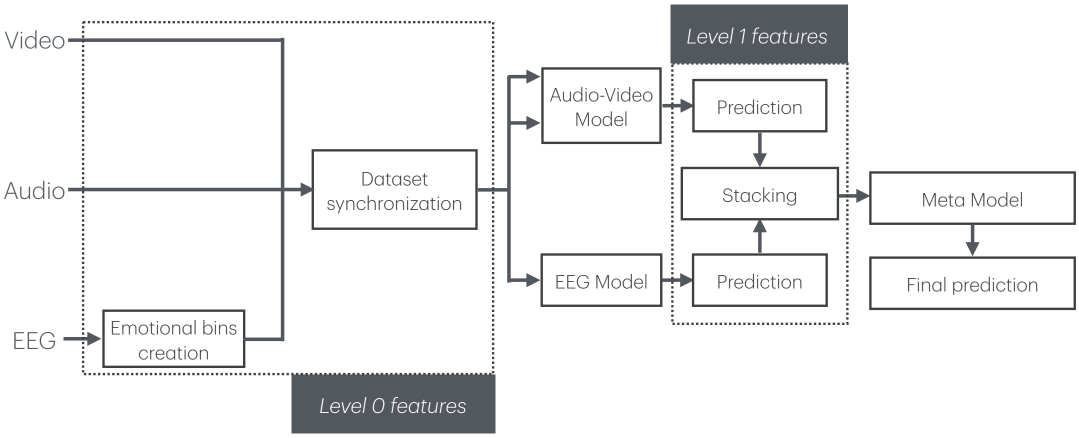

Figure 3: Final architecture responsible for the merging of classification of modality specific models.

3. Experimental Procedure

This section presents our experimental methodology and results. Following the architectures described in Figure 3, we first train individual models for audio, video, and EEG data using the datasets detailed in Section 3.1. These models' outputs serve as inputs for a meta-model that produces the final emotion prediction. Section 3.2 details the data preprocessing methodology, with particular emphasis on EEG signal processing. All experiments were conducted on a high-performance computing cluster running Linux, comprising eight interconnected computational nodes, with four nodes each equipped with four NVIDIA V100 GPUs.

3.1. The dataset

A key challenge in multimodal emotion recognition research is the scarcity of datasets that simultaneously capture audio, video, and EEG signals during emotional experiences. To address this limitation, we adopted a two-stage training approach using specialized datasets for each modality, followed by a synchronized integration procedure for the meta-model. This independent training phase ensures that each model learns modality-specific features optimally.

Audio-Video Dataset For training and evaluating the audio-video emotion recognition model, we utilized the RAVDESS dataset (Ryerson Audio-Visual Database of Emotional Speech and Song) [27].

This dataset includes recordings of 24 professional actors (12 female, 12 male) vocalizing two lexically-matched statements in a neutral North American accent. The actors express various emotions including calm, happy, sad, angry, fearful, surprise, and disgust, with each expression captured in three formats:

- Audio-only (16bit, 48kHz .wav)
- Audio-Video (720p H.264, AAC 48kHz, .mp4)
- Video-only (no sound)

For our experiments, we focused exclusively on the Audio-Video format to capture both facial expressions and vocal characteristics simultaneously. The dataset consists of 2,880 recordings, which we partitioned following a 70:15:15 split for training, validation, and testing, respectively.

To align with our four-category emotion classification scheme, we mapped the original eight emotional expressions in RAVDESS to our target categories (neutral, happy, angry, and sad) based on the Russell Circumplex Model of Affect [25]. This model arranges emotions in a two-dimensional space defined by valence (positive/negative) and arousal (high/low), providing a theoretical foundation for our emotion grouping. The specific mappings are detailed in Table 1.

EEG dataset For the EEG modality, we employed the SEED-IV dataset [28], which contains EEG recordings from 15 subjects during emotion-elicitation experiments. The dataset includes simultaneous recordings from EEG and eye-tracking devices, providing comprehensive neurophysiological data during emotional experiences.

The original dataset comprises 1,080 EEG signals, which we further segmented into sequences of 800 samples each, resulting in a total of 37,575 samples. This segmentation approach ensures uniform sample sizes and eliminates the need for padding operations. We divided the dataset using an 80:10:10 scheme for training, validation, and testing.

Similar to the approach taken with the audio-video dataset, we mapped the original emotion categories in SEED-IV (happy, sad, neutral, and fear) to match our four-category classification scheme. Specifically, we consolidated the "fear" category into the "sad" emotional bin, as detailed in Table 1. This mapping was guided by the valence-arousal coordinates in the circumplex model, where fear and sadness share negative valence characteristics.

Input Emotions	**Final Category**
Neutral, **Calm**	Neutral
Happy, **Surprised**	Happy
Angry, **(Fear)ful**, **Disgusted**	Angry
Sad	Sad

Table 1
Mapping of Emotions to Final Categories. Emotions mapped from the original datasets are presented in bold with fear/fearful belonging to both.

Meta-Model Dataset The dataset for training and evaluating the meta-model consists of the prediction outputs from both pre-trained models (audio-video and EEG). These models were individually trained on their respective datasets to extract modality-specific information, producing prediction vectors that reflect the emotional content of the inputs.

A critical aspect of our approach is the synchronization between data from separate datasets. For the meta-model, each sample is constructed by ensuring that the predictions from both modalities are based on inputs associated with the same emotional label. This process, detailed in Section 3.2 and Algorithm 1, guarantees semantic consistency between the features extracted from different modalities and preserves the integrity of the emotional information.

3.2. Feature Extraction

Effective feature extraction is crucial for capturing emotion-relevant information from heterogeneous data sources. We implemented modality-specific preprocessing pipelines optimized for audio, video, and EEG signals.

Audio and Video Data Preprocessing The audio-video model processes both audio and visual data following established methods from the literature [21]. For the audio stream, we extract Mel-frequency cepstral coefficients (MFCC) as the primary feature representation [6]. This choice was informed by comparative studies showing no significant advantages in using alternative features such as chroma or spectrograms for emotion recognition tasks in speech.

For the visual stream, we implemented a preprocessing pipeline consisting of:

1. Frame sampling at regular intervals
2. Image scaling to a standard resolution
3. Region of interest detection using multi-task cascaded convolutional networks (MTCNN) [29] to localize and extract facial regions
4. Feature extraction using EfficientFace [22], a lightweight yet powerful model specifically designed for facial expression recognition

This preprocessing strategy ensures that the visual model receives consistent and relevant facial expression data while filtering out irrelevant background information.

EEG Data Preprocessing The preprocessing of EEG signals involved several specialized steps to enhance signal quality and extract emotion-relevant features. First, we conducted channel selection to match our laboratory equipment constraints. From the 62 channels available in the SEED-IV dataset, we selected a subset of 14 channels: 'AF3', 'AF4', 'F3', 'F4', 'F7', 'F8', 'T7', 'T8', 'P7', 'P8', 'O1', 'O2', 'FC5', and 'FC6'. This selection corresponds to the channels available in the EEG headset Epoch Plus used in our laboratory, ensuring compatibility with our experimental setup for future multimodal data collection.

EEG signals were pre-processed by filtering out artifacts through a baseline noise removal procedure, eliminating parts of the signal unrelated to the emotional stimulus. The signals were then transformed into the frequency domain and divided into f bands corresponding to alpha (8–14 Hz), beta (14–31 Hz), gamma (31–49 Hz), and theta (4–8 Hz) bins. Each frequency band is known to uniquely contribute to the understanding of emotional and cognitive states. The gamma band (31–49 Hz), in particular, has been shown to play a critical role in emotion recognition tasks. Studies have highlighted that gamma activity increases during the processing of emotionally salient stimuli, particularly in the prefrontal and temporal regions of the brain [23, 24].

Subsequently, the separated channels are processed through a feature extraction mechanism employing differential entropy technique to extract meaningful information from the provided data.

The resulting EEG signals are then transformed into a spatial grid format by mapping the electrode signals onto a 2D matrix based on the spatial arrangement of the electrodes. This transformation enables spatially aware processing—such as with convolutional neural networks—and yields a final output as a 3D EEG frame of size $f \times G.width \times G.height$.

The obtained EEG signals of different channels are then projected onto a grid structure to form a 3D frame EEG signal representation with the size of [number of data points, width of the grid, height of grid] according to the electrodes position.

Label alignment A fundamental challenge in our multimodal approach is the lack of simultaneously recorded data across all three modalities. To address this, we developed a label-based synchronization strategy that creates emotionally consistent batches across modalities. The synchronization procedure, outlined in Algorithm 1, involves the following steps:

- Organizing the EEG dataset samples into emotional bins corresponding to our four emotion categories
- Using samples from the audio-video dataset as a guide for selecting corresponding EEG samples
- For each audio-video sample with a specific emotion label, randomly selecting an EEG sample from the matching emotional bin
- Creating artificial batches containing paired audio-video and EEG samples that share the same emotional label

This approach ensures semantic consistency between modalities despite the absence of temporally synchronized recordings. While not capturing the exact same emotional instances across modalities, it provides a valid basis for training the meta-model to recognize patterns in how each modality responds to similar emotional states.

The complete synchronization algorithm is presented in Algorithm 1, which describes the creation of emotional bins and the batch-based selection process that aligns data points across modalities. These synchronized batches then serve as input for generating the meta-features used in the final classification model.

Algorithm 1 Dataset Synchronization

```
Input: Dataloaders $D_1$, $D_2$; Batch size $B$
Output: Synchronized batches $S_{\text{batches}}$
procedure OrgByLabels($D$)
    Init $L \leftarrow \{\text{label} : [] \mid \text{label} \in \{0, 1, 2, 3\}\}$
    for all $(\text{data}, \text{labels}) \in D$ do
        Add data to $L[\text{label}]$
    end for
    return $L$
end procedure
procedure SyncDatasets($D_1, D_2, B$)
    $L_2 \leftarrow$ OrganizeByLabels($D_2$)
    Init empty $S_{\text{batches}}$
    $C_{\text{batch}} \leftarrow \{\text{audio:}[], \text{video:}[], \text{eeg:}[], \text{labels:}[]\}$
    for all $(\text{audio}, \text{video}, \text{labels}) \in D_1$ do
        for all $i$, label $\in$ enumerate(labels) do
            if $L_2[\text{label}] \neq \emptyset$ then
                Pop $eeg\_data$ from $L_2[\text{label}]$
                Add audio$[i]$, video$[i]$ to $C_{\text{batch}}$
                Add $eeg\_data$, label to $C_{\text{batch}}$
                if $|C_{\text{batch}}| = B$ then
                    Add to $S_{\text{batches}}$
                    Reset $C_{\text{batch}}$
                end if
            end if
        end for
    end for
    if $C_{\text{batch}} \neq \emptyset$ then
        Add remaining samples to $S_{\text{batches}}$
    end if
    return $S_{\text{batches}}$
end procedure
```

4. Results and Discussion

The integrated system's performance is expected to outperform single-modality models by combining the complementary strengths of audio, video, and EEG data. The Audio-Video Emotion Classification

Model	Dataset	Loss	Acc.
Audio-Video	RAVDESS	0.8860	0.7708
FBCCNN	SEED-IV	0.7075	0.8067
Meta model	Meta-features	0.2915	0.9145

Table 2
Average accuracy and loss values of the audio-video and EEG models at the 100th epoch.

Model enhances the system's ability to capture expressive cues from both speech and facial expressions, while the FBCNN model contributes insights from the brain activity captured in EEG data.

The meta-model further improves performance by combining the predictions from both models, ensuring that the final emotion classification is robust and accurate. Preliminary results will be analyzed in terms of accuracy and classification error rates.

We expect that the multi-modal approach will significantly improve recognition rates, particularly in distinguishing emotions such as anger and sadness, which are often more difficult to classify based on a single modality alone.

4.1. Audio-Video and EEG results

The audio-video branch of the proposed model, which integrates a Convolutional Neural Network (CNN) and a Transformer, is implemented following the architecture outlined in the corresponding reference paper. The performance metrics considered for this model are accuracy and loss, evaluated during both the training and validation processes.

Similarly, for the EEG-based model, which utilizes the FBCCNN architecture, the same performance metrics are assessed. As presented in Table 2, the audio-video model, trained on the RAVDESS dataset, achieves a loss of 0.8860 and an accuracy of 77.08%.

In contrast, the EEG model, evaluated on the SEED-IV dataset, achieves a lower loss of 0.7075 and a higher accuracy of 80.67%. These results show the better performance of the EEG-based model in terms of both accuracy and loss.

While the audio-video model demonstrates competitive results, its relatively higher loss and lower accuracy may be attributed to the inherent complexity of processing multimodal data from the RAVDESS dataset. Conversely, the EEG model benefits from the frequency-band-specific learning capabilities of the FBCNN architecture, which prove to be well-suited for the emotion recognition tasks in the SEED-IV dataset.

4.2. Meta model predictions

The integration of multimodal data through logistic regression represents the final stage of our emotion recognition pipeline. While logistic regression does involve a training phase. Unlike the complex training dynamics observed in the CNN and FBCNN networks, this last step follows a more straightforward optimization path. This simpler nature of the model involved implies that traditional training visualizations, such as loss curves or learning rate analysis, are less informative and arguably unnecessary for understanding the model's performance. In the described scenario the logistic regression serves as a meta-learner, weighing and combining the features already extracted by our networks for audio-video and EEG data.

When evaluated independently, the single-modal approaches showed varying degrees of success. The FBCNN model, which uses the DEAP dataset for EEG analysis and tested on multiple actors, achieved an average accuracy of 55.87% as reported in Table 7 in the "Experiment" section of the original paper.

The audio-video model, which leverages both audio and video signals via a 1-head dropout transformer architecture and trained on the RAVDES dataset, demonstrated better performance with an accuracy of 79.08%, as reported in Table III in the "Results and Discussion" section of the original paper.

However, the true potential of emotion recognition emerges through our multimodal integration approach, which achieves a remarkable accuracy of 91% when combining all three modalities (EEG, audio, and video signals) as shown in Table 2, demonstrating the effectiveness of multimodal emotion recognition. The effectiveness of this approach is clearly demonstrated in Figure 4 where the showed confusion matrix describes a model capable of learning and synthesizing information from multiple modalities.

Particularly robust performance is highlighted in distinguishing emotional states that are typically challenging to differentiate. The strongest performance is observed in the happy emotion category, where the model achieves high accuracy with minimal false positives. This suggests that the combination of physiological signals from EEG with audiovisual cues provides particularly strong indicators for this emotional state.

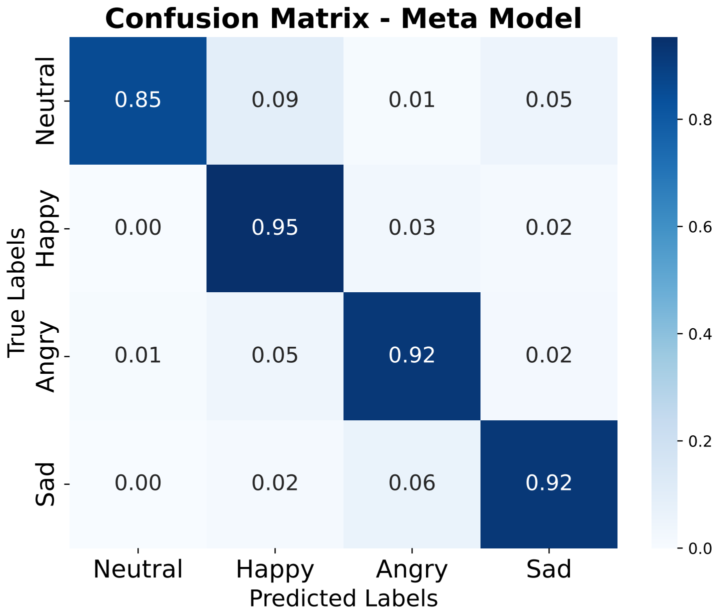

Figure 4: Confusion matrix obtained across the test set for the Meta-model employing EEG and Audio-Video models as baseline for meta-features extraction.

While the model shows strong performance across most categories, the detection of sadness presents more challenges. This aligns with existing literature in emotion recognition, where these types of emotions appear to be complex to classify correctly [30, 25].

The pattern of misclassifications we observe is notably systematic – the model typically confuses emotions that are psychologically adjacent rather than making dramatic misclassifications between opposing emotional states [31]. This suggests that our meta-learning approach has successfully captured the underlying continuous nature of emotional expressions. However, the performance on sadness detection indicates an area where the fusion strategy could benefit from further refinement. This limitation likely reflects the inherent challenge in capturing subtle emotional states across different modalities, rather than a fundamental limitation of the logistic regression approach itself.

One of the main challenges of this study, or more specifically, one of its significant limitations, was the need to align and synchronize the data. Since no existing dataset contained data acquired within the same session or set, we faced the problem of ensuring that the data from different modalities (e.g., EEG, audio, and video) were properly aligned in time.

To address this, we had to devise a solution that would enable us to effectively test our meta-classifier model. However, to overcome this limitation, we are currently conducting an experiment in which audio, video, and EEG data are being simultaneously collected. This new dataset will be more suitable for our study and is expected to be made publicly available after a thorough validation phase.

However, overall, the meta-model performance validates the choice of a simpler and more interpretable fusion strategy over more complex alternatives. In conclusion, the use of logistic regression demonstrates its effectiveness as a fusion strategy for multimodal emotion recognition by integrating predictions from audiovisual and EEG data streams. This is demonstrated by the strong diagonal dominance in the confusion matrix, indicating reliable classification across most emotional states.

5. Conclusion

This study presents a novel cascade multimodal system for emotion recognition that effectively integrates audio, video, and EEG data to achieve superior classification performance. Our approach addresses the inherent limitations of single-modality systems by leveraging the complementary strengths of each data stream through a two-stage architecture. The key contributions and findings of this work can be summarized as follows:

- First, our cascade architecture demonstrates the effectiveness of specialized modality-specific processing before high-level integration. The audio-video component, implementing the hybrid fusion approach from Zhang et al., achieved 77.08% accuracy, while the EEG component using the FBCCNN architecture from Pan and Zheng reached 80.67% accuracy. When combined through our meta-model approach, the system achieved a remarkable 91.45% accuracy, representing an improvement of approximately 11% over the best single-modality model;
- Second, the logistic regression-based meta-model proved to be an effective and interpretable integration strategy. The confusion matrix results revealed particularly strong performance in distinguishing "happy" emotions (95% accuracy) and consistently high performance for "angry" and "sad" categories (92% accuracy each). The pattern of misclassifications primarily occurring between psychologically adjacent emotions in Russell's circumplex model suggests that our system successfully captures the underlying continuous nature of emotional expressions;
- Third, our approach offers practical advantages in terms of implementation and extensibility. By leveraging pre-trained specialized models for each modality, our system can benefit from advancements in modality-specific architectures without requiring complete redesign. The meta-model integration strategy also provides flexibility in weighting the contributions of each modality based on their reliability for specific emotional states.

We acknowledge that a significant limitation of the current study is the lack of simultaneously recorded multimodal data, necessitating our bin-based synchronization strategy. To address this limitation, we are currently conducting experiments to collect a comprehensive dataset with synchronized audio, video, and EEG recordings during emotional experiences. This dataset will enable more direct evaluation of temporal dynamics across modalities and will be made publicly available following validation.

Future research directions include expanding the range of recognizable emotions beyond the four primary categories (neutral, happy, angry, sad) to include more nuanced states such as surprise, fear, and disgust. Additionally, we plan to explore more sophisticated meta-model architectures that can adapt their weighting strategies dynamically based on signal quality and context. The integration of additional physiological signals such as heart rate variability or galvanic skin response also presents promising avenues for further improving recognition accuracy, particularly for subtle emotional states.

In conclusion, our multimodal approach represents a significant step toward developing more naturalistic and robust emotion recognition systems that can function effectively in real-world human-computer interaction scenarios. By demonstrating substantial improvements over single-modality approaches, particularly for complex emotional states, this work contributes to the advancement of empathetic computing systems capable of more nuanced understanding of human emotional experiences.

Acknowledgments

The work was supported by the research project SPECTRA (Supporting schizophrenia PatiEnts Care wiTh aRtificiAl intelligencen - CUP: I53D23006050001), funded by the MIUR with DD no.861 under the PNRR and by the European Union - Next Generation EU.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] M. Spezialetti, G. Placidi, S. Rossi, Emotion recognition for human-robot interaction: Recent advances and future perspectives, Frontiers in Robotics and AI 7 (2020) 532279.

[2] D. Greco, P. Barra, L. D'Errico, M. Staffa, Multimodal interfaces for emotion recognition: Models, challenges and opportunities, in: Int. Conf. on Human-Computer Interaction, Springer, 2024, pp. 152–162.

[3] M. Zhao, The emotion recognition in psychology of human-robot interaction, Psychomachina 1 (2023) 1–11.

[4] Z. Liu, M. Wu, W. Cao, L. Chen, J. Xu, R. Zhang, M. Zhou, J. Mao, A facial expression emotion recognition based human-robot interaction system., IEEE CAA J. Autom. Sinica 4 (2017) 668–676.

[5] C. Tsiourti, A. Weiss, K. Wac, M. Vincze, Multimodal integration of emotional signals from voice, body, and context: Effects of (in) congruence on emotion recognition and attitudes towards robots, Int. Journal of Social Robotics 11 (2019) 555–573.

[6] S. Zhang, S. Zhang, T. Huang, W. Gao, Q. Tian, Learning affective features with a hybrid deep model for audio–visual emotion recognition, IEEE transactions on circuits and systems for video technology 28 (2017) 3030–3043.

[7] M. Staffa, L. D'Errico, Eeg-based machine learning models for emotion recognition in hri, in: Int. Conf. on Human-Computer Interaction, Springer, 2023, pp. 285–297.

[8] L. Galluccio, L. D'Errico, M. Giordano, M. Staffa, Advancing eeg-based emotion recognition: Unleashing the power of graph neural networks for dynamic and topology-aware models, in: 2024 Int. Joint Conf. on Neural Networks (IJCNN), IEEE, 2024, pp. 1–8.

[9] N. Ahmed, Z. Al Aghbari, S. Girija, A systematic survey on multimodal emotion recognition using learning algorithms, Intelligent Systems with Applications 17 (2023) 200171.

[10] T. Baltrušaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: A survey and taxonomy, IEEE transactions on pattern analysis and machine intelligence 41 (2018) 423–443.

[11] Y. Cimtay, E. Ekmekcioglu, S. Caglar-Ozhan, Cross-subject multimodal emotion recognition based on hybrid fusion, IEEE Access 8 (2020) 168865–168878.

[12] S. M. S. A. Abdullah, S. Y. A. Ameen, M. A. Sadeeq, S. Zeebaree, Multimodal emotion recognition using deep learning, Journal of Applied Science and Technology Trends 2 (2021) 73–79.

[13] R. Caccavale, E. Leone, L. Lucignano, S. Rossi, M. Staffa, A. Finzi, Attentional regulations in a situated human-robot dialogue, 2014, p. 844 – 849. doi:10.1109/ROMAN.2014.6926358.

[14] A. I. Middya, B. Nag, S. Roy, Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities, Knowledge-Based Systems 244 (2022) 108580.

[15] M.-I. Georgescu, R. T. Ionescu, Recognizing facial expressions of occluded faces using convolutional neural networks, in: Neural Information Processing, Springer, 2019, pp. 645–653.

[16] B. Pan, K. Hirota, Z. Jia, Y. Dai, A review of multimodal emotion recognition from datasets, preprocessing, features, and fusion methods, Neurocomputing (2023) 126866.

[17] L. D'Errico, E. Di Nardo, A. Ciaramella, M. Staffa, A 3d-cnns approach to classify users' emotion through eeg-based topographical maps in hri, in: Companion of the 2024 ACM/IEEE Int. Conf. on Human-Robot Interaction, 2024, pp. 397–401.

[18] M. Akhand, S. Roy, N. Siddique, M. A. S. Kamal, T. Shimamura, Facial emotion recognition using transfer learning in the deep cnn, Electronics 10 (2021) 1036.

[19] X. Gu, Y. Shen, J. Xu, Multimodal emotion recognition in deep learning: a survey, in: 2021 Int. Conf. on Culture-oriented Science and Technology (ICCST), IEEE, 2021, pp. 77–82.

[20] B. Pan, W. Zheng, Emotion recognition based on eeg using generative adversarial nets and

convolutional neural network, computational and Mathematical Methods in Medicine 2021 (2021) 2520394.

[21] K. Chumachenko, A. Iosifidis, M. Gabbouj, Self-attention fusion for audiovisual emotion recognition with incomplete data, in: 2022 26th Int. Conf. on Pattern Recognition (ICPR), IEEE, 2022, pp. 2822–2828.

[22] Z. Zhao, Q. Liu, F. Zhou, Robust lightweight facial expression recognition network with label distribution training, in: Proceedings of the AAAI conference on artificial intelligence, volume 35, 2021, pp. 3510–3519.

[23] M. Li, B.-L. Lu, Emotion classification based on gamma-band eeg, in: 2009 Annual Int. Conf. of the IEEE Engineering in medicine and biology society, ieee, 2009, pp. 1223–1226.

[24] K. Yang, L. Tong, J. Shu, N. Zhuang, B. Yan, Y. Zeng, High gamma band eeg closely related to emotion: evidence from functional network, Frontiers in human neuroscience 14 (2020) 89.

[25] J. A. Russell, A circumplex model of affect., Journal of personality and social psychology 39 (1980) 1161.

[26] B. Pavlyshenko, Using stacking approaches for machine learning models, in: (DSMP), 2018, pp. 255–258.

[27] S. R. Livingstone, F. A. Russo, The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english, PloS one 13 (2018) e0196391.

[28] W. Zheng, W. Liu, Y. Lu, B. Lu, A. Cichocki, Emotionmeter: A multimodal framework for recognizing human emotions, IEEE Transactions on Cybernetics (2018) 1–13.

[29] K. Zhang, Z. Zhang, Z. Li, Y. Qiao, Joint face detection and alignment using multitask cascaded convolutional networks, IEEE signal processing letters 23 (2016) 1499–1503.

[30] A. Aggarwal, S. Garg, R. Madaan, R. Kumar, Comparison of different machine learning and deep learning emotion detection models, Algorithms for Intelligent Systems (2021).

[31] C. Norman, Ai in pursuit of happiness, finding only sadness: Multi-modal facial emotion recognition challenge, ArXiv abs/1911.05187 (2019).