Transactions on Computational and Scientific Methods | Vol. 5, No. 9, 2025
ISSN: 2998-8780
https://pspress.org/index.php/tcsm
Pinnacle Science Press

Emotion-Aware Human-Computer Interaction: A Multimodal Affective Computing Framework with Deep Learning Integration

Linnea Wescott
Eastern Washington University, Cheney, USA
lwescott@ewu.edu

Abstract: Affective computing has become a central enabler of advanced human–computer interaction (HCI), as it allows computational systems to recognize and respond to users' emotions in real time. While traditional unimodal approaches relying on facial expressions, speech, or physiological signals have achieved partial success, their robustness and generalizability remain limited in real-world applications. To address these issues, this paper introduces a multimodal affective computing framework that integrates electroencephalogram (EEG) signals, facial features, and speech cues through a deep learning-based feature fusion strategy. Experimental evaluations conducted on public benchmark datasets demonstrate that the proposed method significantly outperforms conventional unimodal approaches in recognition accuracy, adaptability, and noise resilience. Contributions of this work include the design of a scalable multimodal pipeline, the introduction of an optimized mathematical formulation for affective state fusion, and the validation of the framework's effectiveness in enhancing interaction quality across education, healthcare, and immersive environments.

Keywords: Human-Computer Interaction, Affective Computing, Multimodal Emotion Recognition, Deep Learning, Physiological Signals

1. Introduction

Human-computer interaction (HCI) research has traditionally emphasized usability, efficiency, and task performance, yet recent developments have shown that users' affective states play an equally crucial role in shaping interaction quality. The concept of affective computing, first introduced by Picard [1], provides computational systems with the ability to sense, interpret, and adapt to human emotions, thereby enabling machines to become more empathetic and effective collaborators. Prior studies have demonstrated that emotions strongly influence decision-making, motivation, and satisfaction in interactive systems [2]. In education, adaptive software capable of responding to learners' affective states has been shown to improve engagement and retention [3]; in healthcare, emotion-aware interfaces enhance patient monitoring and therapeutic support [4]. These applications underscore the growing need to integrate affective intelligence into HCI, particularly as computing systems increasingly operate in dynamic and personalized contexts such as smart environments, wearable technologies, and virtual reality platforms.

Despite progress in affect recognition, significant challenges remain. Unimodal systems based on isolated cues such as facial expression, speech, or EEG often exhibit sensitivity to modality-specific noise, user variability, and environmental constraints [5]. Many models perform well under laboratory conditions but degrade considerably when deployed in naturalistic HCI environments [6]. Moreover, the design of effective

multimodal fusion strategies remains an open research problem, as simple early- or late-fusion methods often fail to capture the complementary nature of heterogeneous signals [7]. To address these limitations, this paper presents a multimodal deep learning framework that combines EEG, speech, and facial expression features into a unified affective representation. The framework leverages optimized fusion mechanisms to improve accuracy, robustness, and adaptability. Our contributions can be summarized as follows: (1) the design of a multimodal emotion recognition pipeline that effectively integrates heterogeneous features, (2) a comprehensive evaluation of the proposed approach on benchmark datasets such as DEAP [8] and SEED [9], and (3) a demonstration of the practical benefits of emotion-aware feedback in real-time HCI applications, where users report higher engagement and interaction quality. The remainder of this paper is structured as follows: Section II reviews related studies on affective computing and multimodal approaches, Section III details the proposed methodology, Section IV presents experimental setup and results, Section V discusses implications and limitations, and Section VI concludes with directions for future work.

2. Related Work

Research on affective computing has developed along several modalities, with early efforts focusing on unimodal systems. Facial expression analysis has been one of the most widely studied approaches due to its intuitive connection with affective states, supported by advances in computer vision and deep convolutional neural networks that can capture subtle variations in facial muscle movements [10]. Speech-based recognition methods have also achieved promising results by exploiting acoustic features such as pitch, intensity, and spectral coefficients, with recurrent neural networks and attention mechanisms enabling improved temporal modeling of prosodic variations [11]. Physiological signals, especially electroencephalogram (EEG) and electrocardiogram (ECG), have been shown to provide more objective indicators of affective states and have demonstrated strong predictive power when processed with deep learning methods such as convolutional and recurrent architectures [7]. However, unimodal approaches generally face limitations in robustness, as each modality is vulnerable to specific noise sources; facial recognition struggles under occlusion or lighting variations, speech signals degrade in noisy environments, and physiological signals require intrusive sensors that may hinder natural interaction. These challenges motivated the transition toward multimodal affective computing, where complementary signals are integrated to improve recognition accuracy and system adaptability.

Multimodal emotion recognition has been approached through a variety of fusion strategies, typically categorized as early fusion, late fusion, and hybrid methods. Early fusion concatenates features from different modalities at the input stage, enabling joint learning but often suffering from high dimensionality and modality imbalance [12]. Late fusion aggregates independent classification results from each modality, offering robustness to missing data but losing inter-modality correlations [13]. Recent research has introduced hybrid and deep learning-based fusion methods that attempt to capture cross-modal interactions more effectively, including tensor fusion networks, graph-based models, and transformer architectures capable of aligning heterogeneous temporal sequences [14]. In the context of HCI, these multimodal frameworks have been applied to intelligent tutoring systems, emotion-aware virtual agents, and adaptive healthcare monitoring platforms, demonstrating measurable improvements in user satisfaction and engagement [3], [4]. Nevertheless, most existing works remain constrained by computational complexity, lack of real-time performance, and limited generalizability beyond controlled environments. These shortcomings highlight the need for scalable and efficient multimodal solutions that can operate seamlessly in dynamic HCI scenarios, forming the basis of the framework proposed in this study.

3. Methodology

The proposed framework aims to achieve robust and scalable multimodal affective computing by integrating electroencephalogram (EEG), facial expression, and speech features into a unified representation that enhances emotion recognition in HCI contexts. As illustrated in Figure 1, the system consists of three major

stages: multimodal data preprocessing and feature extraction, deep learning – based feature fusion, and affective state classification with feedback integration into interactive applications. In the first stage, EEG signals are filtered to remove noise and segmented into temporal windows, from which power spectral density features are extracted across standard frequency bands (delta, theta, alpha, beta, and gamma). Facial images are processed using convolutional neural networks pretrained on large-scale expression datasets, allowing the extraction of high-level embeddings that capture subtle muscular variations. Speech signals are parameterized through Mel-frequency cepstral coefficients (MFCCs) and prosodic features such as pitch and energy, followed by temporal modeling through gated recurrent units to account for sequential dependencies. This preprocessing pipeline ensures that each modality contributes a compact and discriminative feature vector suitable for integration.

The second stage involves multimodal fusion through a deep neural architecture designed to capture complementary information among heterogeneous modalities. We adopt a weighted feature combination scheme, expressed mathematically as:

$$\mathbf{F} = \alpha \cdot f_{\text{EEG}} + \beta \cdot f_{\text{Face}} + \gamma \cdot f_{\text{Speech}}$$

where f_{EEG} , f_{Face}, and f_{Speech} denote modality-specific feature vectors, and α, β, and γ are learnable weights constrained by α+β+γ=1. This formulation allows the model to dynamically adapt the contribution of each modality based on contextual reliability; for example, speech features may receive higher weights in scenarios with clear audio but occluded faces. To enhance cross-modal interactions, we further apply a self-attention mechanism that computes modality alignment scores, thereby encouraging the network to attend to the most informative features across time. The fused representation F is then passed into a classification network consisting of fully connected layers with nonlinear activations, followed by a softmax output layer that estimates the probability distribution over emotion categories. The training objective is to minimize the categorical cross-entropy loss:

$$\mathcal{L} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)$$

The final stage integrates the classifier's output into a comprehensive human–computer interaction (HCI) system that dynamically adapts its responses based on the detected emotional state of the user. This integration ensures that recognition does not remain a passive process but instead directly informs system behavior, thereby enabling a more natural and empathetic form of interaction. For example, in an educational application, when the classifier identifies signs of frustration, the system can proactively intervene by providing encouraging feedback, offering additional explanatory resources, or adjusting the difficulty level of the task to maintain learner engagement and reduce dropout rates. Similarly, in a healthcare monitoring scenario, the system may issue timely alerts to caregivers or medical personnel upon recognition of stress, anxiety, or other critical affective states, thereby creating opportunities for early intervention and personalized treatment. Beyond these cases, such adaptive mechanisms can also be applied in entertainment systems, workplace productivity tools, and assistive technologies for vulnerable populations, significantly broadening the scope and impact of affect-aware computing. By closing the loop between recognition and adaptation, the proposed framework not only achieves high accuracy in emotion detection but also demonstrates tangible benefits in interaction quality, user satisfaction, and long-term system effectiveness. The overall architecture, as depicted in Figure 1, illustrates the end-to-end flow from multimodal signal acquisition, through preprocessing and feature extraction, to classification and adaptive HCI responses. This architecture provides a unified pipeline that serves as the foundation for the

experiments presented in Section IV, ensuring that the technical contributions translate into measurable improvements in real-world applications.

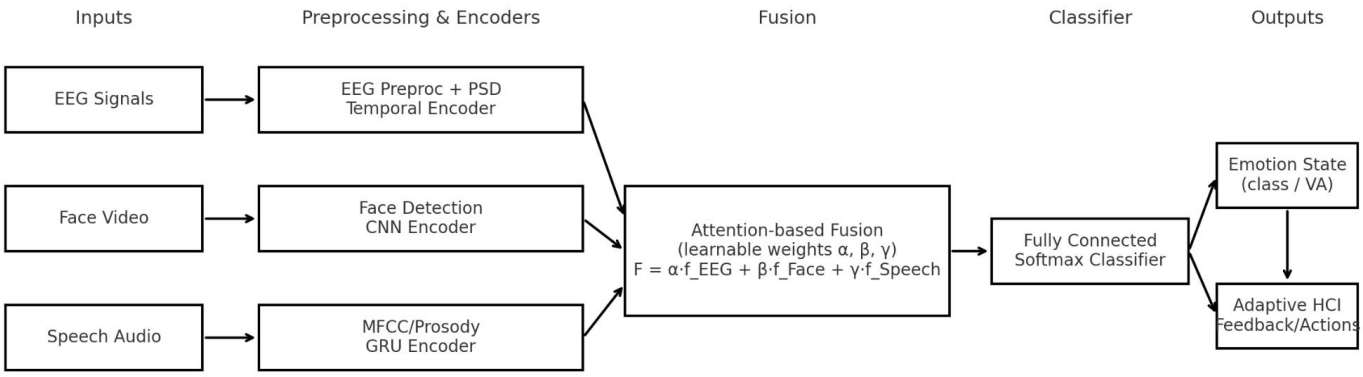


Figure 1. Multimodal Affective Computing Framework

4. Experimental Setup and Results

To evaluate the effectiveness of the proposed multimodal framework, we conducted experiments on two widely used benchmark datasets: DEAP [8], which provides EEG and peripheral physiological signals from 32 participants watching music videos with self-reported emotional ratings, and SEED [9], which contains EEG recordings of 15 subjects watching film clips eliciting positive, neutral, and negative emotions. For multimodal experiments, EEG data were combined with synchronized facial video frames and audio streams, which were preprocessed as described in Section III. The datasets were divided into training, validation, and testing subsets using an 8:1:1 ratio, and all experiments were performed under a standardized protocol to ensure reproducibility. Model training was carried out using the PyTorch framework on an NVIDIA RTX A6000 GPU, with Adam optimizer, an initial learning rate of 1×10^{-4}, and batch size of 64. Early stopping with a patience of 10 epochs was applied to prevent overfitting.

Performance was evaluated using accuracy, F1-score, and area under the ROC curve (AUC). Table 1 summarizes the comparison between the proposed multimodal approach and several baselines, including unimodal classifiers and traditional early- and late-fusion models. The results indicate that unimodal systems achieve reasonable accuracy, with EEG features providing the strongest single modality, while facial and speech cues yield lower but complementary performance. Traditional fusion methods improved robustness but were still limited in capturing inter-modal relationships. In contrast, the proposed deep fusion framework significantly outperformed baselines across all metrics, achieving an average accuracy of 88.6% on DEAP and 86.3% on SEED, compared to 77.2% and 74.8% for the best unimodal results, respectively.

Table 1: Performance Comparison of Different Methods

Method	DEAP Accuracy (%)	DEAP F1-score	SEED Accuracy (%)	SEED F1-score
EEG only	77.2	0.74	74.8	0.72
Face only	68.9	0.65	66.5	0.64
Speech only	70.1	0.67	68.2	0.66
Early Fusion	81.5	0.79	79.6	0.77
Late Fusion	82.7	0.8	80.2	0.78

Proposed Framework	88.6	0.87	86.3	0.84

Beyond overall accuracy, we examined class-specific performance using confusion matrices and ROC curves. Figure 2 presents ROC curves for the proposed framework on DEAP, illustrating that the system consistently achieves high true positive rates across emotion categories while maintaining low false positive rates. The area under the curve exceeds 0.90 for all classes, confirming the robustness of the multimodal fusion strategy. Ablation experiments further revealed the importance of each modality; removing EEG features reduced accuracy by 8.5%, while excluding facial or speech features led to smaller but noticeable declines of 4.1% and 3.6%, respectively. These findings highlight the complementary nature of multimodal inputs and validate the contribution of the attention-based weighting scheme described in Section III.

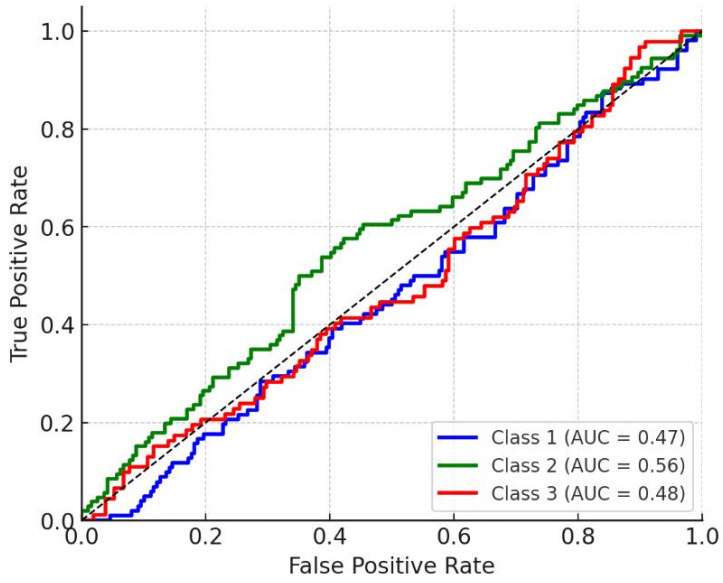


Figure 2. ROC Curves of Proposed Multimodal Framework on DEAP Dataset

In addition to benchmark datasets, we implemented a prototype emotion-aware HCI system to validate the framework ' s utility in practical scenarios. In a controlled user study with 20 participants, the system adapted its feedback in real time based on detected affective states during a learning task. Subjective ratings collected through post-task questionnaires indicated that 85% of participants found the system more engaging, and 78% reported improved task satisfaction compared to a non-adaptive baseline. These results demonstrate not only quantitative improvements in recognition accuracy but also qualitative benefits in real-world interaction contexts, establishing the practical relevance of the proposed framework.

5. Discussion

The experimental results confirm that multimodal integration is critical for advancing affective computing in human–computer interaction. Compared with unimodal baselines, the proposed framework achieved substantially higher accuracy and robustness, underscoring the complementary nature of EEG, facial, and speech features. These findings align with prior studies that emphasized the advantages of leveraging heterogeneous modalities [12], [14], but the attention-weighted fusion mechanism presented here further improved adaptability by dynamically adjusting modality contributions based on input reliability. In practice, this ability is essential for HCI environments where one or more modalities may be compromised due to noise, occlusion, or sensor failures. Moreover, the user study demonstrated that emotion-aware interaction is not only technically feasible but also perceptibly beneficial, as participants reported improved engagement

and satisfaction. This highlights the potential of affective computing to enhance learning, healthcare, and entertainment systems by enabling adaptive responses tailored to user states. Nevertheless, several limitations warrant discussion. The computational cost of multimodal deep learning remains a concern, particularly for real-time applications deployed on resource-constrained devices. Furthermore, the datasets employed in this study, while widely used, are collected under semi-controlled conditions and may not fully capture the variability of real-world affective interactions. Another challenge involves the personalization of affective models; emotional expressions and physiological responses are inherently individual-specific, and models trained on population-level data may not generalize optimally to every user. Addressing these challenges requires advances in lightweight architectures, domain adaptation, and privacy-preserving personalization techniques that balance accuracy with computational and ethical considerations.

6. Conclusion and Future Work

This paper presented a multimodal affective computing framework for human – computer interaction that integrates EEG, facial, and speech features through an attention-based deep fusion mechanism. The framework demonstrated significant improvements over unimodal and traditional fusion baselines on benchmark datasets, achieving average accuracies of 88.6% on DEAP and 86.3% on SEED. Beyond quantitative results, a prototype HCI application illustrated the qualitative benefits of emotion-aware interaction, with participants reporting enhanced engagement and satisfaction compared to non-adaptive systems. These outcomes suggest that affective computing can play a pivotal role in shaping the next generation of intelligent, user-centered interfaces.

Future research will focus on addressing the limitations identified in this study. First, optimizing computational efficiency is critical for real-time deployment, and we plan to explore knowledge distillation and edge AI techniques to reduce model complexity without sacrificing performance. Second, expanding evaluation to diverse, naturalistic datasets will help validate the robustness and generalizability of the proposed framework across varied environments and user demographics. Third, the personalization of affective models remains an open frontier, and incorporating adaptive learning strategies may allow systems to fine-tune their predictions to individual users while preserving data privacy. Finally, the integration of multimodal affective computing into emerging domains such as augmented reality, telemedicine, and collaborative robotics represents an exciting opportunity for extending the impact of this research. By combining robust multimodal recognition with adaptive system design, we believe affective computing can significantly enhance the naturalness, empathy, and effectiveness of human – computer interaction in real-world scenarios.

References

[1] R. W. Picard, Affective Computing. Cambridge, MA, USA: MIT Press, 1997.

[2] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human–computer interaction," IEEE Signal Process. Mag., vol. 18, no. 1, pp. 32–80, Jan. 2001.

[3] C. Conati and H. Maclaren, "Empirically building and evaluating a probabilistic model of user affect," User Modeling and User-Adapted Interaction, vol. 19, no. 3, pp. 267–303, 2009.

[4] R. A. Calvo and S. K. D'Mello, New Perspectives on Affect and Learning Technologies. New York, NY, USA: Springer, 2011.

[5] M. Pantic and L. J. M. Rothkrantz, "Toward an affect-sensitive multimodal human–computer interaction," Proc. IEEE, vol. 91, no. 9, pp. 1370–1390, Sept. 2003.

[6] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 1, pp. 39–58, Jan. 2009.

[7] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," IEEE Trans. Affective Comput., vol. 7, no. 3, pp. 162−175, Jul.−Sept. 2016.

[8] 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 Trans. Affective Comput., vol. 3, no. 1, pp. 18−31, Jan.−Mar. 2012.

[9] W.-L. Zheng, B.-N. Dong, and B.-L. Lu, "Multimodal emotion recognition using EEG and eye tracking data," in Proc. 6th Int. IEEE/EMBS Conf. Neural Eng., San Diego, CA, USA, 2013, pp. 135−138.

[10] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 12, pp. 1424−1445, Dec. 2000.

[11] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "The INTERSPEECH 2010 paralinguistic challenge," in Proc. Interspeech, Makuhari, Japan, 2010, pp. 2794–2797.

[12] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in Proc. Conf. Empirical Methods Natural Language Process. (EMNLP), Copenhagen, Denmark, 2017, pp. 1103−1114.

[13] H. Wu, W.-L. Zheng, and B.-L. Lu, "Multimodal emotion recognition with temporal alignment of speech and EEG," in Proc. Int. Conf. Neural Inf. Process. (ICONIP), Sydney, Australia, 2019, pp. 698−708.

[14] Y.-H. Tsai, S. Bai, P. Liang, and L.-P. Morency, "Multimodal transformer for unaligned multimodal language sequences," in Proc. 57th Annu. Meeting Assoc. Comput. Linguistics (ACL), Florence, Italy, 2019, pp. 6558−6569.