Comparing Fusion Strategies for Multimodal Emotion Prediction Using Deep Physiological Features

Fatemeh Rahimi[1], Christian Tamantini[2,*], Andrea Orlandini[2], Francesca Fracasso[2] and Roberta Siciliano[1]

[1]Federico II University of Naples, 80138 Naples, Italy
[2]Institute of Cognitive Sciences and Technologies, National Research Council of Italy, 00196 Rome, Italy

Abstract

Recognizing affective states from physiological signals is essential for enabling emotion-aware systems, particularly in human–robot interaction. This paper presents a hybrid deep learning framework for multimodal emotion recognition that integrates deep feature extraction with handcrafted physiological descriptors. The system processes electrocardiogram, photoplethysmogram, and galvanic skin response signals to predict arousal and valence in a continuous regression setting. To our aim, we evaluate two fusion strategies — feature-level and decision-level fusion — using two public affective datasets (AMIGOS and DEAP). Features extracted from each modality via a shared one-dimensional convolutional neural network and signal-specific physiological metrics are either concatenated (feature-level fusion) or separately modeled and combined at the prediction level (decision-level fusion). A broad set of machine learning regressors, including boosting methods and tree ensembles, is explored. Experiments were conducted with a leave-one-subject-out cross-validation protocol to assess generalization across users. Results show that feature-level fusion generally outperforms decision-level fusion, achieving the best root mean square error of 0.089 for arousal and 0.053 for valence. Statistical analyses confirm the significance of these differences, particularly favoring adaptive boosting and random forest under feature fusion. The proposed architecture offers a robust and interpretable solution for physiological emotion recognition and provides a solid foundation for real-time applications in emotion-aware social robotics and human-centered adaptive systems.

Keywords

Emotion Recognition, Affective Computing, Deep Feature Extraction, Physiological Signals.

1. Introduction

Among key areas of modern human–computer interaction, affective computing aims to enable machines to recognize, interpret, plan, and respond to human emotions. A central task in this field is emotion recognition, which involves automatically analyzing emotional cues from multiple sources such as facial expressions, speech, and physiological signals, mimicking human perception of emotions [1]. Foundational scientific models guide emotion classification: Ekman's model defines six basic universal emotions—joy, sadness, anger, fear, surprise, and disgust—plus a neutral state [2], while Russell's circumplex model represents emotions along continuous valence and arousal dimensions [3]. These frameworks underpin emotion recognition system design, which has been extensively studied across diverse modalities.

Recent advances have broadened affective computing applications into mental health monitoring, personalized interfaces and intelligent decision-making systems, highlighting the need for robust Multimodal Emotion Recognition (MER) [4]. A particularly promising application is in socially assistive robots, which leverage emotion recognition to enable adaptive, user-centered interactions in, e.g., caregiving, rehabilitation, and education [5]. Integrating continuous valence-arousal estimation into robot control architectures supports these applications by allowing real-time perception of user affective

Workshop on Social Robotics for Human-Centered Assistive and Rehabilitation AI (a Fit4MedRob event) - ICSR 2025

[*] Corresponding author.

f.rahimi@studenti.unina.it (F. Rahimi); christian.tamantini@cnr.it (C. Tamantini); andrea.orlandini@cnr.it (A. Orlandini); francesca.fracasso@cnr.it (F. Fracasso); roberta@unina.it (R. Siciliano)
0009-0002-4234-3055 (F. Rahimi); 0000-0001-6238-2241 (C. Tamantini); 0000-0001-6458-5202 (A. Orlandini); 0000-0001-9442-870X (F. Fracasso); 0000-0002-8012-0461 (R. Siciliano)

states. A modular design, where an independent *emotion recognition component* interacts with a robot's deliberative and reactive layers, is effective [6, 7, 8]. This component can process physiological signals through feature extraction and fusion pipelines to produce continuous affective state estimates. These estimates can inform behavior planners, enabling dynamic adjustment of robot responses, e.g., detecting distress in eldercare and modulating speech or assistance accordingly [9]. Embedding this framework fosters socio-emotional intelligence in assistive robots, improving user acceptance, care adherence, and natural interaction [10, 11]. However, real-world deployment introduces practical challenges such as signal degradation from motion artifacts, variable sensor placement, and latency constraints that can impact real-time responsiveness. Addressing these challenges requires efficient, noise-tolerant models and lightweight architectures suitable for on-device inference, as highlighted in applications involving robot-assisted rehabilitation systems [12].

A critical component of effective MER systems lies in the feature extraction process, which transforms raw multimodal data into informative and discriminative representations for emotion classification [13]. Traditional methods [14] often rely on time-domain features, including statistical and signal-based measures such as peak count, peak amplitude, variability, signal power, mean, standard deviation, minimum, maximum, and mean differences. Additional frequency-domain and time-frequency features have also been widely explored to capture complex temporal and spectral patterns [15]. However, with the rise of deep learning, there has been a paradigm shift toward automated feature learning, which is especially beneficial for processing complex, heterogeneous data sources [16]. Convolutional Neural Network (CNN) has become essential for extracting deep features in physiological and non-physiological modalities due to its ability to learn hierarchical, discriminative representations.

CNNs capture complex spatiotemporal patterns in Electroencephalogram (EEG) [17], with models like ScalingNet achieving strong results on DEAP and AMIGOS [18, 19, 20]. Extensions incorporate global-local receptive fields [21] and hierarchical fusion [22]. For Electrocardiogram (ECG) and Photoplethysmogram (PPG), CNN autoencoders and multimodal CNNs effectively classify emotions [23, 24]. Galvanic Skin Response (GSR) signals, indicative of arousal, benefit from CNN-long short-term memory models for end-to-end learning [25]. Our work applies CNN-based extraction on GSR (DEAP, AMIGOS), ECG (AMIGOS), and PPG (DEAP), leveraging their proven efficacy [26]. In vision, CNNs extract facial expression features from images and videos [27]. For audio, CNNs analyze spectrograms and are often paired with recurrent layers to capture temporal dynamics [28]. Textual emotion recognition also uses CNNs to capture semantic and syntactic cues [29, 30]. CNNs additionally facilitate cross-modal fusion, such as EEG-text [21] and audio-visual [31] integrations. Fusion strategies are critical in MER to integrate signals from physiological, visual, and audio modalities, enhancing accuracy and robustness. Three main strategies exist:

Feature-Level Fusion combines features from multiple modalities before classification. For physiological data, Hassan et al.[32] extracted features from Electrodermal Activity (EDA), Electromyography (EMG), and PPG using deep belief networks, achieving high accuracy on DEAP. Zhang et al.[33] fused EEG, EMG, GSR, and respiration signals with a deep regularized framework, improving valence and arousal prediction. Similarly, CNN-based fusion of facial and vocal features showed superior performance [34]. Decision-Level Fusion combines outputs from modality-specific classifiers. Zhao et al.[35] fused CNN-based EEG, electrooculography, and GSR decisions, outperforming unimodal models. Xu et al.[36] blended manual and deep features from audio-visual inputs using ensemble classifiers. This preserves modality specificity but may miss inter-modality correlations. Hybrid Fusion integrates both feature and decision levels. Yan et al. [37] combined facial, texture, and audio features early, then fused decisions later, improving recognition in unconstrained environments. These studies collectively suggest that while decision-level fusion preserves the uniqueness of each modality, feature-level and hybrid strategies can better exploit inter-modal relationships, especially relevant for physiological signals, which offer involuntary and robust indicators for affective state recognition. Therefore, identifying the optimal fusion strategy is crucial for developing personalized emotion-aware robotic systems.

To this end, our work proposes a hybrid deep learning framework that combines deep features with handcrafted physiological descriptors to enhance affect recognition from ECG, PPG, and GSR signals. This work builds on our recently proposed DeepPhysioNet, a deep physiological feature extraction method for affective state recognition from wearable sensing. Here, we focus on evaluating its effectiveness under different fusion strategies, highlighting its potential for real-world affective computing. By evaluating both feature-level and decision-level fusion strategies on the AMIGOS and DEAP datasets, we demonstrate that integrating modalities at feature level consistently leads to superior performance. Through rigorous experimentation using Leave-One-Subject-Out Cross-Validation (LOSO-CV) and a diverse set of regression models, we achieve state-of-the-art results, particularly in predicting valence and arousal. Unlike existing works that focus either on handcrafted descriptors or fixed fusion architectures, our contribution lies in validating the discriminative power of DeepPhysioNet's features and providing a systematic comparison of fusion strategies tailored for physiological signals. These findings not only validate the effectiveness of feature-level fusion but also underscore the potential of our architecture as a foundation for real-time, personalized, and emotion-aware robotic systems.

2. Proposed Framework

To effectively capture both low-level temporal patterns and high-level physiological descriptors from multimodal biosignals, we propose a hybrid deep learning framework that integrates data-driven feature learning with domain-specific physiological knowledge. As shown in Figure 1, the system is designed to process signals such as ECG and GSR, which are acquired independently per subject and trial, and to flexibly support multimodal affective state estimation.

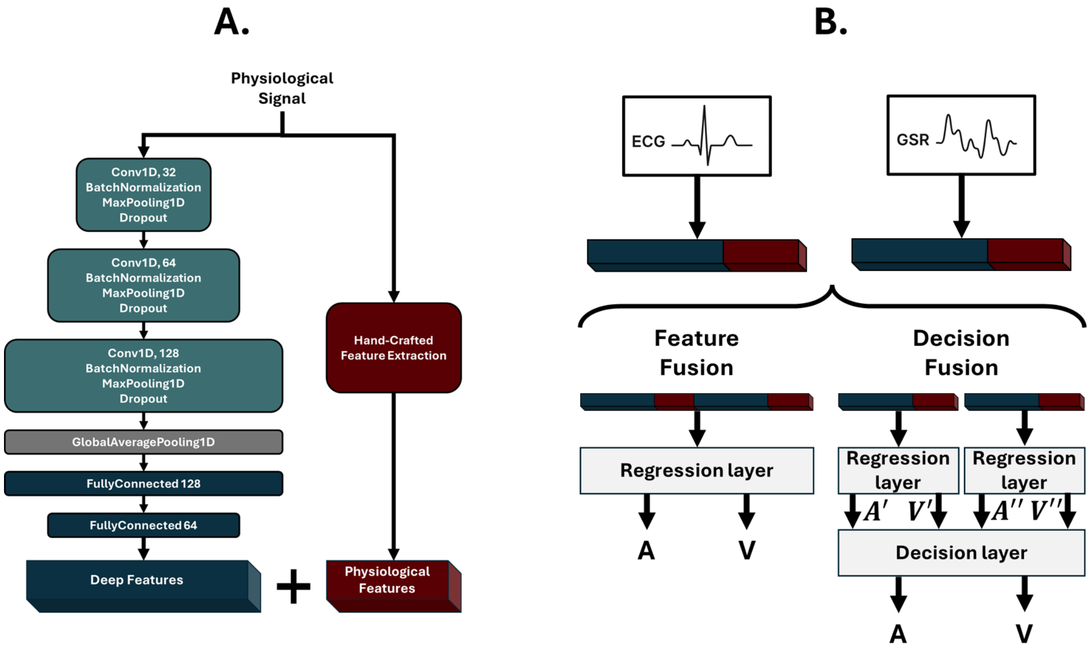

Figure 1: A. Overview of the proposed deep learning framework for multimodal affective state prediction from physiological signals. Raw physiological signals (e.g., ECG, GSR) are independently processed by a shared One-Dimensional (1D) CNN architecture to extract deep features. These are concatenated with handcrafted features to form hybrid representations. **B.** Two fusion strategies are supported: feature-level fusion, where combined features from multiple modalities are fed to a single regression model; and decision-level fusion, where separate regressors are trained per modality and their predictions are integrated via a late fusion layer.

At its core, the architecture relies on a shared 1D CNN, composed of stacked convolutional layers with increasing filter widths, each followed by batch normalization, max pooling, and dropout. This sequence enables the extraction of deep, hierarchical features from raw signals. A global average pooling layer compresses the temporal dimension, followed by fully connected layers that generate compact deep embeddings of each modality. To improve both interpretability and physiological robustness, these deep features are concatenated with handcrafted features computed per modality. These can include classic time- and frequency-domain metrics, such as Heart Rate Variability (HRV) parameters and skin conductance indices, depending on the physiological signal given in input to the network. The result is a joint feature vector that integrates learned representations and expert-designed descriptors. Moreover, the proposed deep feature extraction framework is particularly suitable for managing deep multimodal learning. It supports two complementary fusion strategies:

- **Feature-Level Fusion**, where the feature vectors from multiple modalities are concatenated into a single representation and passed to a regression layer. This approach enables end-to-end learning across modalities and supports direct exploitation of multimodal dependencies.
- **Decision-Level Fusion**, in which independent regressors are trained for each modality. Their outputs, i.e., predicted arousal or valence values, are later combined through a late fusion ensemble, introducing robustness to sensor-specific noise and missing data. This strategy preserves the individual modality characteristics and has been shown to outperform feature-level fusion in scenarios with degraded or noisy signals. For instance, in [38], decision-level fusion achieved a significantly higher accuracy by separately learning from audio and visual features and combining results via an ensemble method, while [39] highlights its advantage in low-quality multimodal data environments.

Thanks to its modular and modality-agnostic design, the proposed framework can be easily adapted to different physiological channels and experimental settings. Its ability to combine physiological insight with deep learning makes it particularly suited for affective computing applications in real-world human–robot interaction scenarios, such as emotion-aware social robots deployed in domestic environments for personalized monitoring and adaptive interaction. Moreover, its low computational footprint and flexible signal handling make it appropriate for embedded use, where resources are limited and robustness against sensor noise or dropout is essential for sustained user engagement.

3. Experiments

In order to validate the proposed framework, we present experiments carried out to cope with the problem of affective state recognition based on multimodal physiological signals. The task is formulated as a continuous regression problem, targeting the prediction of arousal and valence dimensions. Both feature-level and decision-level fusion strategies are evaluated, and their comparative performance is analyzed in the following section. To this end, two widely adopted affective computing datasets were used to develop and evaluate the proposed framework. Both datasets provide multimodal physiological recordings collected during exposure to emotionally evocative stimuli and include subjectively annotated arousal and valence values for each trial. The AMIGOS dataset [20] contains recordings from 40 participants in both individual and group settings. In this study, only the individual sessions were considered, focusing on ECG and GSR signals acquired while participants watched short video clips. After each clip, participants reported their perceived arousal and valence using continuous self-assessment scales, and external annotations of arousal and valence were provided by independent observers, enabling evaluation against both subjective self-assessments and externally judged emotional states. The DEAP dataset [19] includes data from 32 participants who each watched 40 one-minute music videos in a controlled laboratory environment. For our experiments, PPG and GSR signals were used. Participants rated their affective responses on a 9-point Likert scale for both arousal and valence dimensions.

3.1. Feature Extraction

Each physiological signal (ECG, PPG, GSR) was processed using the hybrid pipeline introduced in this work, which combines deep feature extraction through a shared 1D CNN architecture with domain-informed handcrafted features. This design enables the generation of compact, modality-independent representations that capture both latent signal dynamics and physiologically meaningful descriptors.

For cardiovascular signals such as ECG and PPG, a shared set of HRV metrics was extracted based on Inter-Beat Interval (IBI), which was derived through peak detection algorithms [40]. The specific features computed in this study are listed in Table 1. For GSR, which is a well-established indicator of sympathetic nervous system activity, the extracted features are summarized in Table 2. All handcrafted features were normalized and concatenated with CNN-based embeddings to form a 68-dimensional vector for ECG/PPG and a 67-dimensional vector for GSR per trial.

Table 1
Description and mathematical formulation of HRV-related features extracted from ECG and PPG signals.

Feature	Description	Equation
IBI	Average time between consecutive heartbeats, computed from R-peaks (ECG) or pulse peaks (PPG). t_i denotes the timestamp of the i-th detected peak.	$\text{IBI} = \frac{1}{N-1}\sum_{i=1}^{N-1}(t_{i+1} - t_i)$
Root Mean Square of Successive Differences (RMSSD)	Time-domain HRV metric that quantifies short-term variability between successive IBIs.	$\text{RMSSD} = \sqrt{\frac{1}{N-2}\sum_{i=1}^{N-2}(IBI_{i+1} - IBI_i)^2}$
Standard Deviation of NN Intervals (SDNN)	Standard deviation of all inter-beat intervals, reflecting overall HRV. $\overline{IBI}$ is the mean IBI.	$\text{SDNN} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(IBI_i - \overline{IBI})^2}$
Frequency Signal Quality Index (FSQI)	Spectral-based quality index indicating the proportion of total power contained in the LF (0.04–0.15 Hz) and HF (0.15–0.4 Hz) bands. Computed using Welch's periodogram.	$\text{FSQI} = \frac{P_{\text{LF}} + P_{\text{HF}}}{P_{\text{Total}}}$

Table 2
Description and mathematical formulation of GSR-based physiological features extracted per trial.

Feature	Description	Equation
Skin Conductance Responses (SCR) Peak Count	Number of SCR detected during the trial using a threshold-based peak detection algorithm (amplitude $> 0.01\ \mu S$). Detected via NeuroKit2's `eda_process()` method [41].	
Mean SCR Amplitude	Average amplitude of the detected SCR responses. M is the number of SCRs, A_j is the amplitude of the j-th SCR.	$\text{Mean SCR} = \frac{1}{M}\sum_{j=1}^{M} A_j$
Mean GSR Level	Average value of the raw GSR signal over the trial duration. In discrete terms, the mean over all N samples.	$\text{Mean GSR} = \frac{1}{N}\sum_{i=1}^{N} GSR_i$

3.2. Supervised Model for Emotion Recognition

A supervised regression pipeline was implemented to predict arousal and valence from the extracted multimodal features, exploring two complementary fusion strategies: feature-level fusion and decision-level fusion. In both configurations, model hyperparameters were optimized via grid search within each fold of the LOSO-CV protocol, ensuring consistent and unbiased evaluation.

In the feature-level fusion approach, the deep and handcrafted features extracted from different modalities (ECG and GSR) were concatenated into a single feature vector per trial. This unified representation allows regression models to capture inter-modality correlations and learn joint patterns across signals [42]. In contrast, the decision-level fusion strategy involves training separate models for each modality and subsequently combining their predictions using ensemble methods [43], a technique shown to enhance robustness in multimodal affective computing scenarios [44].

To evaluate the proposed framework across both fusion strategies, a diverse set of machine learning regressors was employed, encompassing both simple and advanced models. This variety ensures broad coverage of learning paradigms and robustness to overfitting, nonlinearity, and noise. Specifically, we included:

- Linear Regression (LR), for its interpretability and as a baseline linear model [45];
- Support Vector Regressor (SVR), effective for capturing nonlinear relations with good generalization [46];
- Random Forest (RF), an ensemble of decision trees that reduces variance via bagging [47];
- Gradient Boosting (GB) and Adaptive Boosting (AdaBoost), which sequentially build additive models to minimize error, offering strong performance in many real-world regression tasks [48], [49];
- XGBoost (XGB) and LightGBM (LGBM), highly efficient GB implementations that support regularization and scalability, particularly suited for tabular data [50], [51];
- CatBoost, which natively handles categorical features and stabilizes training through ordered boosting [52].

All models were applied both in the feature-fusion pipeline and in the decision-fusion meta-regressor framework, enabling a comprehensive comparison of their capacity to model multimodal physiological data in the context of affective state estimation.

3.3. Experimental Evaluation.

The proposed framework was validated through a LOSO-CV scheme. In each iteration, data from one subject in the AMIGOS dataset were excluded from training and used solely for testing, while the remaining AMIGOS subjects, together with all participants from the DEAP dataset, were used for training. This evaluation protocol reflects a realistic deployment scenario in which models must generalize to new users whose physiological patterns are not seen during training.

Model performance was quantified using the Root Mean Square Error (RMSE), a commonly adopted metric in affective computing tasks involving continuous affect prediction [14]. RMSE penalizes larger deviations more heavily and is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2} \tag{1}$$

where $\hat{y}_i$ and y_i represent the predicted and true values for the i-th trial, respectively, and N is the total number of predictions. Lower RMSE values indicate more accurate predictions of arousal and valence dimensions.

3.4. Statistical Analysis.

To evaluate the impact of fusion strategies on predictive performance, we first assessed, for each regression model individually, whether feature-level fusion or decision-level fusion yielded significantly lower RMSE values. This comparison was conducted separately for arousal and valence using the Wilcoxon signed-rank test, applied to paired RMSE scores computed across subjects. This approach allowed us to determine which fusion strategy was more effective on a per-model basis, accounting for subject-level variability.

Subsequently, we investigated the relative performance of all models within each fusion strategy and emotion dimension. A Friedman test [53] was employed to assess whether statistically significant differences existed in model performance, considering the repeated-measures design. The null hypothesis (H_0) stated that all models performed equally (i.e., no difference in median RMSE), while the alternative hypothesis (H_1) assumed that at least one model differed significantly from the others.

To identify which specific model pairs contributed to any significant effects found by the Friedman test, we performed pairwise comparisons using the Wilcoxon signed-rank test with Bonferroni correction to control for multiple comparisons. This procedure was repeated for both arousal and valence, and for both fusion types.

Only results with adjusted p-values ≤ 0.05 were considered statistically significant and annotated in the corresponding visualizations. This two-level statistical analysis framework enabled us to draw robust conclusions on the effectiveness of fusion strategies and the relative merits of different models in multimodal affective state prediction.

4. Results and Discussion

To provide a comparative overview of the predictive performance across models and fusion strategies, Figure 2 reports the RMSE distributions obtained for each model under both feature-level and decision-level fusion. Results are shown separately for the arousal and valence prediction tasks. For each pair (model, task), statistical significance was assessed using Wilcoxon signed-rank tests, comparing feature- vs. decision-level fusion. Statistically significant differences are annotated with standard asterisk notation, while non-significant results are labeled as "ns".

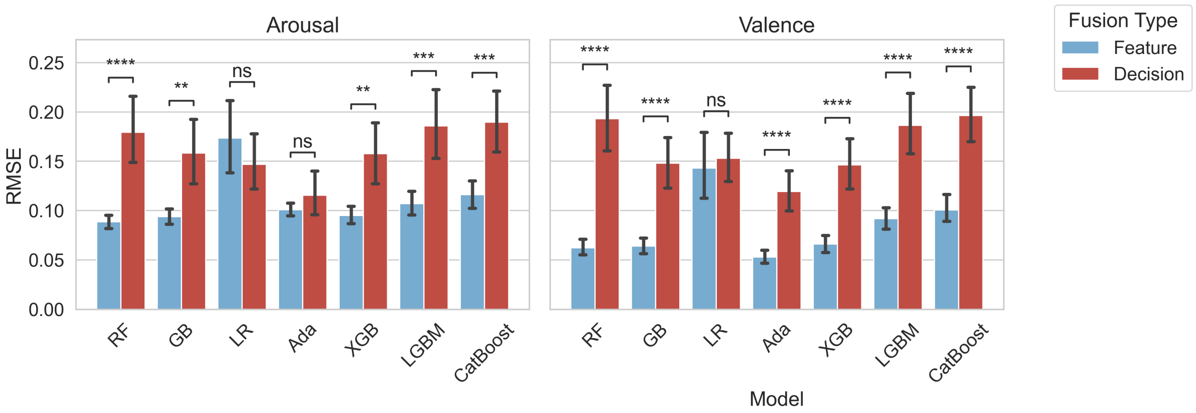

Figure 2: Comparison of model performance across fusion strategies for arousal and valence prediction. Bars show mean RMSE ± standard deviation per model, grouped by fusion type. Significance markers denote Wilcoxon test results comparing fusion strategies within each model: ns = not significant, $^{*}\ p < .05$, $^{**}\ p < .01$, $^{***}\ p < .001$, $^{****}\ p < .0001$.

The results shown in Figure 2 confirm that, in general, feature-level fusion leads to better performance than decision-level fusion across most of the tested models. This outcome aligns with established findings in multimodal learning, where early fusion strategies often benefit from the ability to capture joint dependencies and correlations across modalities at the feature level [42]. By integrating the complementary characteristics of physiological signals (e.g., GSR and ECG) into a shared representation before model training, feature fusion allows the regressors to exploit richer contextual information. However, this trend is not consistent across all models. For instance, in the case of LR and AdaBoost during arousal prediction, decision-level fusion shows slightly better or equivalent performance. This inversion may be attributed to the limited capacity of linear or shallow models to exploit the high-dimensional fused feature space effectively. In such cases, learning separate unimodal models and aggregating their outputs can act as a form of regularization, reducing the risk of overfitting and improving robustness to noisy features or modality-specific artifacts. Moreover, ensemble-based methods like RF and GB generally benefit more from feature-level fusion, likely due to their capacity to handle heterogeneous features and non-linear interactions. In contrast, models with more rigid assumptions or sensitivity to feature scaling (e.g., LR) may struggle when exposed to the increased complexity introduced by early fusion. These findings underscore the importance of selecting appropriate fusion strategies in relation to the model architecture and the characteristics of the input modalities. While feature-level fusion appears generally preferable, decision-level fusion can still provide competitive performance in scenarios where model simplicity or modularity is required.

To further investigate the comparative performance of the different regression models across fusion strategies and emotion dimensions, we conducted a non-parametric Friedman test. This test assesses whether there are statistically significant differences in model performance when evaluated on the same subjects, based on RMSE rankings. The results of this analysis are reported in Table 3. For both arousal and valence prediction tasks, and under both feature-level and decision-level fusion strategies, the Friedman test returned extremely low p-values (all $< 10^{-19}$), clearly rejecting the null hypothesis that all models perform equally. This confirms that the choice of model has a significant impact on performance, regardless of the fusion strategy adopted. Interestingly, the highest Friedman test statistic was observed in the valence prediction task under feature-level fusion (186.86), indicating particularly large differences in model performance in this configuration. This may be due to the nature of valence representation in physiological signals, which could benefit more from richer multimodal embeddings learned during feature fusion. Overall, the analysis highlights that both the fusion strategy and the emotion dimension being predicted play a crucial role in shaping the relative effectiveness of the regression models.

Table 3
Friedman Test Results

Fusion Strategy	Emotion	Friedman Test Statistic	p-value
Feature-level Fusion	Arousal	123.85	1.21×10^{-23}
	Valence	186.86	6.93×10^{-37}
Decision-level Fusion	Arousal	111.45	1.92×10^{-20}
	Valence	134.69	3.01×10^{-25}

A natural continuation of the statistical analysis following the Friedman test is provided by the pairwise Wilcoxon comparisons summarized in Figure 3. This set of heatmaps displays the adjusted p-values resulting from multiple pairwise tests between models within each combination of fusion strategy and emotion dimension. The Bonferroni correction was applied to account for multiple comparisons, and significance levels are indicated using a standard asterisk notation.

The results confirm and extend the Friedman test findings, revealing several significant pairwise differences in model performance. In particular, models such as AdaBoost and XGB consistently outperform others under feature-level fusion, especially for the valence dimension. Conversely, the performance gaps under decision-level fusion appear slightly narrower, although significant differences still emerge.

4.1. Comparison with Related Studies.

To further contextualize the effectiveness of the proposed framework, Table 4 presents a comparison with the benchmark results reported by [14]. That study employed traditional hand-crafted features—such as Hjorth parameters, spectral entropy, wavelet-based energy and entropy, and empirical mode decomposition descriptors—combined with conventional classifiers like k-Nearest Neighbors (KNN) and RF for affective state prediction using the AMIGOS dataset.

In contrast, our approach integrates deep feature learning with handcrafted physiological metrics and explores both feature-level and decision-level fusion strategies. The results indicate a consistent

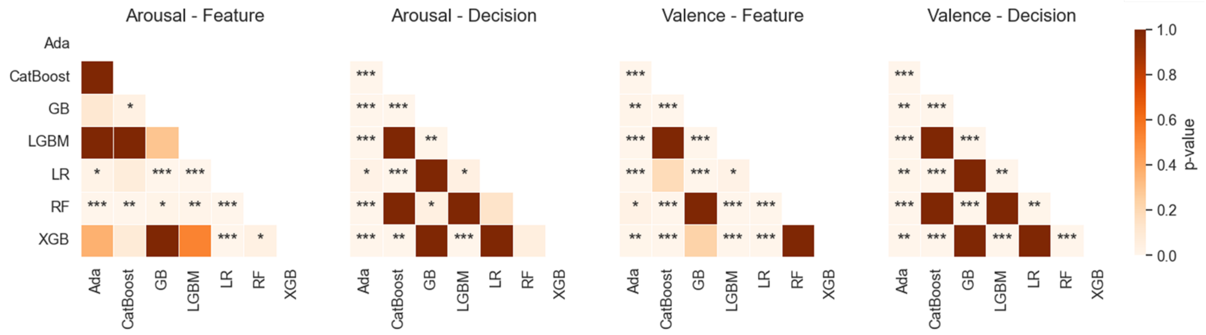

Figure 3: Adjusted p-value heatmaps from pairwise Wilcoxon signed-rank tests between predictive models, shown separately for each combination of fusion strategy (feature-level vs. decision-level) and emotion dimension (arousal vs. valence). The tests were computed on subject-level RMSE values, and Bonferroni correction was applied. Asterisks indicate significance levels (* $p < .05$, ** $p < .01$, *** $p < .001$, **** $p < .0001$.).

Table 4
RMSE comparison of prior studies and our proposed methods for Arousal and Valence prediction using physiological signals. The comparison emphasizes the impact of feature extraction techniques and fusion strategies. Best-performing models in each configuration are highlighted in bold.

Ref.	Dataset	Feature Extraction	Fusion Type	Arousal RMSE	Valence RMSE
[14]	AMIGOS	Traditional	Feature-level (KNN, RF)	KNN:0.129, RF:0.148	KNN:0.175, RF:0.195
Ours	AMIGOS+DEAP	Deep Feature + Handcrafted	Decision-level (AdaBoost)	0.116	0.119
Ours	AMIGOS+DEAP	Deep Feature + Handcrafted	Feature-level (RF, AdaBoost)	**RF:0.089**	**Ada:0.053**

improvement in RMSE performance for both arousal and valence dimensions. Specifically, decision-level fusion using AdaBoost achieves RMSE scores of 0.116 for arousal and 0.119 for valence, outperforming the baseline models in [14]. Even greater performance gains are observed with feature-level fusion, where RF and AdaBoost models attain RMSE values as low as 0.089 and 0.053, respectively.

These findings confirm the advantages of combining deep representations with early fusion mechanisms, especially when dealing with complex, multimodal physiological data. They also demonstrate the superiority of the proposed pipeline in comparison to existing handcrafted approaches, supporting its suitability for real-world affective computing applications.

5. Conclusion

This study presented a neural framework for multimodal affective state recognition based on physiological signals. The proposed architecture integrates deep feature extraction via a shared 1D CNN with signal-specific handcrafted physiological metrics, enabling a robust representation of autonomic responses related to emotional arousal and valence. In addition to designing an effective feature extraction pipeline, we systematically investigated two widely adopted fusion strategies, i.e, feature-level and decision-level fusion, within a supervised regression setting. The framework was validated on two benchmark datasets, AMIGOS and DEAP, using a LOSO-CV protocol to simulate real-world generalization to unseen subjects. A wide range of machine learning regressors was tested to assess the flexibility and robustness of the extracted features under different fusion paradigms. Our results demonstrate that feature-level fusion consistently outperforms decision-level fusion in most scenarios, particularly when coupled with ensemble-based models such as RF and AdaBoost. Statistical analyses using Wilcoxon and Friedman tests confirmed the significance of these findings, highlighting the impact of both model selection and fusion strategy on performance. When compared with previous work based on traditional feature engineering and classical classifiers, our approach achieved lower RMSE values for both arousal and valence prediction tasks, confirming the value of combining deep representations with physiological insights. In future work, we plan to systematically evaluate fusion strategies under non-optimal conditions (e.g., simulated noise or missing modalities), to better understand their robustness and suitability for real-world settings. This analysis will help determine whether feature-level fusion remains effective or if decision-level fusion offers greater resilience in such scenarios. In addition, we plan to conduct ablation studies by removing specific modalities to assess the contribution of each physiological signal. We will also compare our approach with classical machine learning models trained on handcrafted features tailored to each modality. These insights will inform the deployment of our proposed framework in real-time applications, using biosignals collected during human–robot interaction to support emotionally adaptive behavior. In particular, we aim to integrate the model into social robotic platforms, enabling them to continuously estimate users' affective states and adapt their communicative strategies accordingly, paving the way toward more empathic and responsive assistive technologies.

Acknowledgments

This work was partially supported by the Italian Ministry of Research, under the complementary actions to the NRRP "Fit4MedRob - Fit for Medical Robotics" Grant PNC0000007, (CUP: B53C22006990001) and partially by Next Generation EU – "Age-It – Ageing Well in an Ageing Society" project (PE0000015), National Recovery and Resilience Plan (NRRP).

Declaration on Generative AI

During the preparation of this work, the authors used generative AI tools (specifically, OpenAI's GPT-4) to assist with grammar and spelling checks.

References

[1] B. Pan, K. Hirota, Z. Jia, Y. Dai, A review of multimodal emotion recognition from datasets, preprocessing, features, and fusion methods, Neurocomputing 561 (2023) 126866.

[2] Y. Wang, W. Song, W. Tao, A. Liotta, D. Yang, X. Li, S. Gao, Y. Sun, W. Ge, W. Zhang, et al., A systematic review on affective computing: Emotion models, databases, and recent advances, Information Fusion 83 (2022) 19–52.

[3] R. Cittadini, C. Tamantini, F. Scotto di Luzio, C. Lauretti, L. Zollo, F. Cordella, Affective state estimation based on russell's model and physiological measurements, Scientific Reports 13 (2023) 9786.

[4] A. Geetha, T. Mala, D. Priyanka, E. Uma, Multimodal emotion recognition with deep learning: advancements, challenges, and future directions, Information Fusion 105 (2024) 102218.

[5] C. Tamantini, A. Umbrico, A. Orlandini, Automated planning and scheduling in robot-aided rehabilitation: a review, Journal of NeuroEngineering and Rehabilitation 22 (2025) 180.

[6] G. Beraldo, C. Tamantini, A. Umbrico, A. Orlandini, Fostering behavior change through cognitive social robotics, in: International Conference on Social Robotics, Springer, 2024, pp. 279–291.

[7] R. D. Benedictis, A. Umbrico, F. Fracasso, G. Cortellessa, A. Orlandini, A. Cesta, A dichotomic approach to adaptive interaction for socially assistive robots, User Modeling and User-Adapted Interaction 33 (2023) 293–331.

[8] C. Tamantini, A. Umbrico, A. Orlandini, Repair platform: Robot-aided personalized rehabilitation, in: International Conference of the Italian Association for Artificial Intelligence, Springer, 2024, pp. 301–314.

[9] A. Umbrico, R. De Benedictis, F. Fracasso, A. Cesta, A. Orlandini, G. Cortellessa, A mind-inspired architecture for adaptive hri, International Journal of Social Robotics 15 (2023) 371–391.

[10] M. Spezialetti, G. Placidi, S. Rossi, Emotion recognition for human-robot interaction: Recent advances and future perspectives, Frontiers in Robotics and AI Volume 7 - 2020 (2020). doi:10.3389/frobt.2020.532279.

[11] S. Rossi, F. Ferland, A. Tapus, User profiling and behavioral adaptation for hri: A survey, Pattern Recognition Letters 99 (2017) 3–12.

[12] e. a. Zhang, Emotion recognition using eeg and physiological data for robot-assisted rehabilitation systems, in: Companion Publication of the International Conference on Multimodal Interaction (ICMI), 2020.

[13] N. Kim, S. Cho, B. Bae, Smate: A segment-level feature mixing and temporal encoding framework for facial expression recognition, Sensors 22 (2022) 5753.

[14] F. Galvão, S. M. Alarcão, M. J. Fonseca, Predicting exact valence and arousal values from eeg, Sensors 21 (2021) 3414.

[15] Q. Li, A. Zhang, Z. Li, Y. Wu, Improvement of emg pattern recognition model performance in repeated uses by combining feature selection and incremental transfer learning, Frontiers in Neurorobotics 15 (2021) 699174.

[16] N. Jia, C. Zheng, W. Sun, A multimodal emotion recognition model integrating speech, video and mocap, Multimedia Tools and Applications 81 (2022) 32265–32286.

[17] S. K. Khare, V. Bajaj, Time–frequency representation and convolutional neural network-based emotion recognition, IEEE transactions on neural networks and learning systems 32 (2020) 2901–2909.

[18] J. Hu, C. Wang, Q. Jia, Q. Bu, R. Sutcliffe, J. Feng, Scalingnet: Extracting features from raw eeg data for emotion recognition, Neurocomputing 463 (2021) 177–184.

[19] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, I. Patras, Deap: A database for emotion analysis; using physiological signals, IEEE transactions on affective computing 3 (2011) 18–31.

[20] J. A. Miranda-Correa, M. K. Abadi, N. Sebe, I. Patras, Amigos: A dataset for affect, personality and mood research on individuals and groups, IEEE transactions on affective computing 12 (2018) 479–493.

[21] S. Wang, J. Qu, Y. Zhang, Y. Zhang, Multimodal emotion recognition from eeg signals and facial expressions, IEEE Access 11 (2023) 33061–33068.

[22] Y. Zhang, C. Cheng, Y. Zhang, Multimodal emotion recognition using a hierarchical fusion convolutional neural network, IEEE access 9 (2021) 7943–7951.

[23] N. Hajarolasvadi, E. Bashirov, H. Demirel, Video-based person-dependent and person-independent facial emotion recognition, Signal, Image and Video Processing 15 (2021) 1049–1056.

[24] G. Yin, S. Sun, D. Yu, D. Li, K. Zhang, A multimodal framework for large-scale emotion recognition by fusing music and electrodermal activity signals, ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 18 (2022) 1–23.

[25] T.-P. Jung, T. J. Sejnowski, et al., Utilizing deep learning towards multi-modal bio-sensing and vision-based affective computing, IEEE Transactions on Affective Computing 13 (2019) 96–107.

[26] F. J. Ordóñez, D. Roggen, Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition, Sensors 16 (2016) 115.

[27] Q. Wei, X. Huang, Y. Zhang, Fv2es: A fully end2end multimodal system for fast yet effective video emotion recognition inference, IEEE Transactions on Broadcasting 69 (2022) 10–20.

[28] M. Sharafi, M. Yazdchi, R. Rasti, F. Nasimi, A novel spatio-temporal convolutional neural framework for multimodal emotion recognition, Biomedical Signal Processing and Control 78 (2022) 103970.

[29] F. Chen, J. Shao, A. Zhu, D. Ouyang, X. Liu, H. T. Shen, Modeling hierarchical uncertainty for multimodal emotion recognition in conversation, IEEE Transactions on Cybernetics 54 (2022) 187–198.

[30] S. Liu, P. Gao, Y. Li, W. Fu, W. Ding, Multi-modal fusion network with complementarity and importance for emotion recognition, Information Sciences 619 (2023) 679–694.

[31] Z. Farhoudi, S. Setayeshi, Fusion of deep learning features with mixture of brain emotional learning for audio-visual emotion recognition, Speech Communication 127 (2021) 92–103.

[32] M. M. Hassan, M. G. R. Alam, M. Z. Uddin, S. Huda, A. Almogren, G. Fortino, Human emotion recognition using deep belief network architecture, Information Fusion 51 (2019) 10–18.

[33] J. Zhang, Z. Yin, P. Chen, S. Nichele, Emotion recognition using multi-modal data and machine

learning techniques: A tutorial and review, Information Fusion 59 (2020) 103–126.

[34] J. Xu, H. Li, Y. Wang, Cnn-based multimodal emotion recognition using facial and vocal features, IEEE Access 8 (2020) 36774–36785.

[35] Y. Zhao, X. Cao, J. Lin, D. Yu, X. Cao, Multimodal affective states recognition based on multiscale cnns and biologically inspired decision fusion model, IEEE Transactions on Affective Computing (2021).

[36] J. Xu, L. Wang, Y. Zhang, H. Li, An ensemble learning framework for multimodal emotion recognition using audio and visual features, Neurocomputing 412 (2020) 251–259.

[37] J. Yan, W. Zheng, Z. Cui, C. Tang, T. Zhang, Y. Zong, Multi-cue fusion for emotion recognition in the wild, Neurocomputing 309 (2018) 27–35.

[38] M. Hao, W.-H. Cao, Z.-T. Liu, M. Wu, P. Xiao, Visual-audio emotion recognition based on multi-task and ensemble learning with multiple features, Neurocomputing 391 (2020) 42–51.

[39] Q. Zhang, Y. Wei, Z. Han, H. Fu, X. Peng, C. Deng, Q. Hu, C. Xu, J. Wen, D. Hu, et al., Multimodal fusion on low-quality data: A comprehensive survey, arXiv preprint arXiv:2404.18947 (2024).

[40] C. Tamantini, M. L. Cristofanelli, F. Fracasso, A. Umbrico, G. Cortellessa, A. Orlandini, F. Cordella, Physiological sensor technologies in workload estimation: A review, IEEE Sensors Journal (2025).

[41] M. Benedek, C. Kaernbach, A continuous measure of phasic electrodermal activity, Journal of neuroscience methods 190 (2010) 80–91.

[42] P. K. Atrey, M. A. Hossain, A. El Saddik, M. S. Kankanhalli, Multimodal fusion for multimedia analysis: a survey, Multimedia systems 16 (2010) 345–379.

[43] J. Kittler, M. Hatef, R. P. Duin, J. Matas, On combining classifiers, IEEE transactions on pattern analysis and machine intelligence 20 (1998) 226–239.

[44] R. A. Calvo, S. D'Mello, J. M. Gratch, A. Kappas, The Oxford handbook of affective computing, Oxford University Press, 2015.

[45] G. A. Seber, A. J. Lee, Linear regression analysis, John Wiley & Sons, 2003.

[46] H. Drucker, C. J. Burges, L. Kaufman, A. Smola, V. Vapnik, Support vector regression machines, Advances in neural information processing systems 9 (1996).

[47] L. Breiman, Random forests, Machine learning 45 (2001) 5–32.

[48] J. H. Friedman, Greedy function approximation: a gradient boosting machine, Annals of statistics (2001) 1189–1232.

[49] Y. Freund, R. E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, Journal of computer and system sciences 55 (1997) 119–139.

[50] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785–794.

[51] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, Lightgbm: A highly efficient gradient boosting decision tree, Advances in neural information processing systems 30 (2017).

[52] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, A. Gulin, Catboost: unbiased boosting with categorical features, Advances in neural information processing systems 31 (2018).

[53] M. Friedman, The use of ranks to avoid the assumption of normality implicit in the analysis of variance, Journal of the american statistical association 32 (1937) 675–701.