RESEARCH ARTICLE

A quantitative evaluation system for music emotion regulation based on multi-physiological signal feature fusion

Minhui Li*

School of Cruise and Art Design, Jiangsu Martime Institute, Nanjing, Jiangsu, China.

Received: August 28, 2025; **accepted:** December 26, 2025.

Music intervention has been widely recognized as an effective means of emotional regulation for mental health. However, existing evaluation methods mainly rely on subjective self-reports, lacking objectivity and real-time tracking capabilities. To address this limitation, this study proposed a quantitative evaluation system for music-induced emotional regulation effects based on the fusion of multimodal physiological signal features, which addressed the limitations of traditional subjective questionnaires by constructing an objective assessment framework using multimodal physiological responses. The system synchronously collected four types of physiological signals including electroencephalogram (EEG), electrocardiogram (ECG), electrodermal activity (EDA), and electromyography (EMG). The study also applied an improved deep forest algorithm for feature selection and dimensionality reduction. A temporal convolutional network (TCN) was employed to extract spatiotemporal features from EEG signals, while phase locking value (PLV) was used to quantify functional connectivity between brain regions. For ECG, an adaptive heartbeat segmentation algorithm was developed to enhance the robustness of heart rate variability (HRV) features. A novel multi-source attentional feature fusion (MAFF) mechanism was introduced to learn dynamic cross-modal feature weights using a gated recurrent unit (GRU), enabling optimized multimodal feature integration. A regression estimator based on a gradient boosting decision tree (GBDT) was constructed within a valence-arousal dimensional emotion model and evaluated using transfer learning on the database for emotion analysis using physiological signals (DEAP) and PMEmo datasets. The results showed that the proposed system achieved mean squared error (MSE) of 0.0410 (valence) and 0.0380 (arousal) with R^2 values of 0.81 and 0.83 respectively on the DEAP dataset, significantly outperforming unimodal approaches. The MAFF mechanism reduced the arousal MSE from 0.0462 to 0.0380, representing a 17.8% improvement. After fine-tuning on the PMEmo dataset *via* transfer learning, the model achieved further MSE reductions to 0.0380 (valence) and 0.0361 (arousal), demonstrating strong generalization and robustness across datasets. By bridging the gap between physiological signals and emotional states, this study provided a reliable objective quantitative benchmark for music therapy and offered a promising technical reference for future affective computing research in mental health care.

Keywords: multimodal physiological signal; music emotion regulation; feature fusion mechanism; deep forest algorithm; emotion evaluation system.

***Corresponding author:** Minhui Li, School of Cruise and Art Design, Jiangsu Martime Institute, Nanjing, Jiangsu 210000, China. Email: 18635679547@163.com.

Introduction

In today's rapidly evolving society, emotional problems have become a major factor affecting mental health and quality of life. Effectively assessing and regulating individual emotional states has emerged as a key research focus across psychology, neuroscience, and artificial intelligence (AI). Music as a unique form of art has been widely recognized for its role in emotion induction and regulation [1, 2]. In recent years, music-based interventions have been increasingly applied in psychotherapy, emotion management, and health promotion, demonstrating significant potential as a non-pharmacological approach [3]. However, current evaluation methods for music-induced emotional regulation rely heavily on subjective questionnaires or interviews. These methods are often biased and inconsistent, making it difficult to achieve objective, continuous, and personalized tracking of emotional changes [4].

With the continuous integration of AI and affective computing, multimodal emotion recognition has emerged as a prominent research focus. Udahemuka *et al*. highlighted that visual cues, acoustic features, and physiological signals each possessed distinct advantages in emotional expression. The integration of these factors effectively overcame the limitations of single-modal approaches in emotion recognition [5]. Kim *et al*. proposed a dual-function music classification system based on physiological signal features by analyzing parameters of electroencephalogram (EEG) and electrodermal activity (EDA). The system enabled automatic recognition and recommendation of music emotion types [6]. However, current studies still mainly focus on static emotion classification and lack continuous modeling and evaluation of the emotion regulation process. In the area of music and physiological signal integration, Yin *et al*. proposed a large-scale emotion recognition framework that combined music content with EDA signals. The research employed deep neural networks to jointly model musical audio features and physiological response patterns, exploring the mapping between emotional labels and multidimensional feature spaces [7]. However, the framework lacked comprehensive modeling of other physiological channels such as EEG and electrocardiogram (ECG). Focusing on the key technologies underpinning multimodal emotion recognition, Zhu *et al*. provided a systematic review of current deep learning models, feature selection mechanisms, and fusion strategies used in emotion analysis and emphasized that the major challenges in multimodal fusion included inter-modal inconsistency, synchronization discrepancies, and redundancy in high-dimensional feature spaces [8]. In the field of multi-source physiological signal fusion, Zhu *et al*. proposed the multi-language font generation network (MF-Net) model, which integrated EEG, ECG, and electromyography (EMG) signals using residual structures and attention mechanisms to achieve efficient fusion of emotional features [9]. However, the model still has room for improvement in temporal dependency modeling and feature selection. Additionally, Du *et al*. conducted an empirical study on emotional responses induced by traditional Chinese-style music and developed a hybrid model combining one-dimensional convolutional neural network (CNN) and bidirectional long short-term memory network (Bi-LSTM) to analyze EEG data from university students. The model achieved dual classification in the valence and arousal dimensions, demonstrating the influence of musical cultural context on the transferability of emotion recognition models [10]. In the context of attention-based fusion, Ghaleb *et al*. explored joint modeling strategies for audio and visual cues and introduced attention mechanisms to enhance the model’s sensitivity to emotionally salient segments [11]. Although their study focused on audiovisual modalities, its proposal of temporal selective modeling offered valuable insights for dynamically capturing salient features in physiological signals. Similarly, Yang *et al*. developed a mobile-based emotion recognition method that integrated behavioral data with physiological signals. Their work demonstrated the potential of lightweight deep models for use in wearable devices, highlighting the need to balance computational efficiency with model performance in future emotion recognition systems [12]. In terms of cross-modal coordination strategies for multimodal emotion recognition, Vamsidhar *et al*. proposed a hierarchical cross-modal attention mechanism. Through dual-channel audio pathway modeling, the approach achieved fine-grained emotional semantic extraction and improved semantic alignment across modalities [13]. Although the work was primarily applied to affective computing tasks, the proposed mechanism provided theoretical basis and feasible technical approach for addressing the heterogeneity of multi-source physiological signals. Although traditional emotion recognition methods have achieved certain progress using unimodal signals, they still face limitations due to the complexity of emotional responses and individual variability. The accuracy and robustness of these approaches remain to be significantly improved [14].

Compared to unimodal signals, multimodal physiological signals offer stronger representational capacity in terms of information richness and response specificity. EEG captures real-time cortical activity, making it suitable for tracking the spatiotemporal dynamics of emotional states. Heart rate variability (HRV) derived from ECG reflects the autonomic nervous system’s response to emotional stimuli. EDA as a direct indicator of sympathetic nervous activity is commonly used to assess emotional arousal levels [15]. Surface EMG measures facial or bodily muscle activity, which is closely linked to emotional experience [16]. Therefore, integrating features from multiple physiological sources enables multi-perspective modeling and more accurate evaluation of music-induced emotional regulation. However, one of the key challenges in multimodal fusion lies in the heterogeneity across signal channels including differences in sampling rates, signal patterns, response latencies, and noise distributions [17]. Traditional fusion strategies often fail to effectively capture inter-modal relationships, leading to feature redundancy or information loss, which limits both the expressiveness and generalization capacity of the model [18].

This study aimed to address the limitations of traditional emotional assessment methods by developing a quantitative evaluation system for music-induced emotional regulation effects based on multimodal physiological signal fusion to provide an objective, accurate, and personalized description of emotional state changes. The study adopted a temporal convolutional network (TCN) to extract spatiotemporal features, which introduced a novel multi-source attentional feature fusion (MAFF) mechanism, leveraged gated recurrent units (GRUs) to dynamically learn cross-modal weights, and constructed a regression model combined with a gradient boosting decision tree (GBDT). This study provided a reliable objective quantitative benchmark for the evaluation of music therapy effects and offered a robust methodological framework for solving the problem of multi-source heterogeneous data fusion in affective computing, thereby promoting its practical application in the field of mental health monitoring.

Materials and methods

Data resources and preprocessing

The public accessible Database for Emotion Analysis using Physiological signals (DEAP) (https://www.eecs.qmul.ac.uk/mmv/datasets/deap/) and PMEmo (http://huisblog.cn/PMEmo/) were employed in this research as the data resources. DEAP dataset was used for proposed model construction [19], while PMEmo dataset was used to verify the generalization ability of the proposed model when dealing with single-modal and cross-dataset scenarios [20]. All data were uniformly processed using a fifth-order zero-phase bandpass filter. The EEG, ECG, EMG signals were filtered within 1 - 45 Hz, 0.5 - 40 Hz, 20 - 150 Hz, respectively, while the EDA signals underwent high-pass filtering at 0.05 Hz to remove baseline drift [21]. For noise reduction, independent component analysis (ICA) was applied to EEG channels to eliminate artifacts caused by eye movements, blinks, and muscle activity. ECG signals were processed using an adaptive threshold detection algorithm to extract the R-peak sequence and remove motion artifacts, providing a foundation for subsequent heart rate variability analysis. For time synchronization, the system employed a trigger pulse-based marking mechanism. All modal signals were aligned at event trigger points and segmented accordingly. Each segment corresponded to a 30-second music excerpt, forming a data window for unified feature extraction. Additionally, all signal samples were normalized using the z-score method to ensure comparability of amplitude features across different modalities as follows [22].

$$Z_i = \frac{x_i - \mu}{\sigma} \tag{1}$$

where x_i was the sample value of the original signal. μ was the mean value of the signal channel. σ was the standard deviation. Z_i was the normalized value. To minimize the impact of invalid samples on model performance, the system set artifact detection thresholds. If more than 25% of data points in a segment exceeded physiologically plausible limits such as EEG amplitude beyond ± 100 μV or abnormal ECG rhythm, the segment was automatically marked as invalid and excluded. During EDA preprocessing, transient changes in skin conductance were extracted using a peak detection algorithm, and their rates of change were calculated *via* a sliding window approach. For EMG signals, after bandpass filtering, rectification and moving average smoothing were applied to enhance the distinguishability of muscle contraction activity. To facilitate unified feature extraction, multimodal signals within each time window were converted into a standardized data frame structure that contained four types of synchronized time-series data with the sampling frequency of all signals unified to 256 Hz. For EEG signals, the data structure included raw signals from 32 channels and was formatted as $[n_{samples}, 32]$. ECG signals were recorded with a single lead and formatted as $[n_{samples}, 1]$. EDA was also recorded with a single channel and formatted as $[n_{samples}, 1]$. EMG signals were collected from 8 channels on the face and upper limbs to capture muscle activity with a data format of $[n_{samples}, 8]$. This unified matrix representation ensured that the subsequent feature extraction module could efficiently process multi-source heterogeneous data.

Feature extraction

To comprehensively evaluate the regulatory effects of music stimuli on emotions, this study designed a customized multimodal physiological feature extraction method, which encompassed time-domain, frequency-domain, and spatiotemporal dynamic features combined with neural network architectures to enhance the representation of higher-order features. The proposed method constructed separate feature extraction pipelines for EEG, ECG, EDA, and EMG signals, which were ultimately integrated into a unified multimodal fusion module, serving as the input for subsequent modeling and evaluation. In EEG signal processing, temporal convolutional network (TCN) was used to extract its temporal and spatial dynamic characteristics using causal convolution structure to maintain sequence order and improving deep feature transmission ability through residual connection. Letting the EEG signal segment be the matrix $X_{EEG} \in R^{T \times \mathrm{C}}$, where T was the number of time steps and C was the number of channels, the characteristic expression of TCN extraction was shown below.

$$H^{(l)} = ReLU(W^{(l)} * H^{(l-1)} + b^{(l)}) \tag{2}$$

where $H^{(0)} = X_{EEG}$. $W^{(l)}$ was the convolution kernel of the l-th layer. $*$ was one-dimensional convolution operation. $ReLU$ was the activation function. $b^{(l)}$ was the bias term. TCN structure effectively captured the time delay correlation and, at the same time, had a longer receptive field to describe the dynamic adjustment process of music to EEG [23]. Meanwhile, to model the functional connection between EEG multi-brain regions, this study calculated the degree of phase synchronization between channels and quantified the coupling relationship between channels by using phase locking value (PLV). Letting the instantaneous phases of two EEG channel signals obtained by Hilbert transform be $\phi_1(t)$ and $\phi_2(t)$, respectively, the definition of PLV was then as follows.

$$PLV = \left|\frac{1}{N}\sum_{t=1}^{N} e^{j(\phi_1(t)-\phi_2(t))}\right| \tag{3}$$

where j was imaginary unit. N was the length of time window. The closer the PLV value was to 1, the stronger the phase synchronization between the two brain regions, reflecting the neural coupling characteristics related to emotional processing. For ECG signal, to extract the characteristics of heart rate variability (HRV), an adaptive heartbeat detection algorithm was designed to identify the peak value of R wave based on the change of waveform slope. Letting R-R interval sequence be $\{RR_1, RR_2, \cdots, RR_n\}$, the basic time-domain characteristics of HRV including average RR interval $\overline{RR}$ and standard deviation $SDNN$ were shown as follows.

$$\overline{RR} = \frac{1}{n}\sum_{i=1}^{n} RR_i \tag{4}$$

$$SDNN = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(RR_i - \overline{RR})^2} \tag{5}$$

In addition, frequency domain features such as low-frequency power and high-frequency power were extracted, and the RR interval sequence was transformed into frequency domain by fast Fourier transform (FFT) to calculate its power spectral density. In EDA signal processing, the study focused on two indicators including rapid skin conductance response (SCR) and slowly changing skin conductance level (SCL) [24]. SCR was composed of discrete peak response of skin electricity, which usually reflected the instantaneous activation state of autonomic nervous system. Letting the peak response sequence be $\{p_1, p_2, \cdots, p_m\}$, the SCR frequency per unit time was then defined below.

$$SCR_{freq} = \frac{m}{\Delta t} \tag{6}$$

where Δt was the length of observation time. SCL was the average value of the signal in the window, which was used to measure the degree of baseline activation. EMG signal characteristics were mainly based on the amplitude and power characteristics of muscle potential, which was full wave rectified and smoothed by moving average. Letting the processed EMG signal be $s(t)$, the root mean square (RMS) was defined below.

$$RMS = \sqrt{\frac{1}{T}\sum_{t=1}^{T} s(t)^2} \tag{7}$$

Meanwhile, to enhance the time-frequency characterization of muscle activity patterns, short-time Fourier transform (STFT) was applied to extract the temporal evolution of energy distribution [25].

The Feature Dimension Reduction and Selection

The layered structure of deep forest was introduced to perform multi-layer nonlinear mapping and screening on the original feature matrix. Letting the initial input feature be $X \in R^{n\times d}$, where n was the number of samples and d was the feature dimension, the output feature representation of the l-th layer was shown as follows.

$$H^{(l)} = F^{(l)}(H^{(l-1)}) \tag{8}$$

where $F^{(l)}$ was the l -layer deep forest transformation operation. $H^{(0)} = X$. Each layer of deep forest included several decision tree sub-models, whose structure dynamically determined the number of layers and forest depth through cross-validation, learned nonlinear feature combinations layer by layer and output class confidence vectors. On the candidate feature set of deep forest output, mutual information was further introduced as a feature selection criterion to measure the correlation between individual features and emotional tags. Letting the characteristic variable be X_i and the label variable be Y, the mutual information between them was then defined below.

$$I(X_i; Y) = \sum_{x\in X_i}\sum_{y\in Y} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) \tag{9}$$

where $p(x, y)$ was a joint probability distribution. $p(x)$ and $p(y)$ were edge distribution functions, respectively. The greater the mutual information, the stronger the predictive power of features on labels. If the threshold was set as θ, only the values that satisfied $I(X_i; Y) > \theta$ features were kept and used for subsequent modeling. In addition to the static selection mechanism, to enhance the cooperative expression ability of cross-modal features, this study further introduced the neural attention mechanism and dynamically learned the importance weight of each modal feature channel through the time memory ability of the gated cycle unit [26]. Letting the input be the multi-modal fusion feature sequence $Z = \{z_1, z_2, \cdots, z_T\}$, the GRU unit state update of each step was shown as follows.

$$r_t = \sigma(W_r z_t + U_r h_{t-1}) \tag{10}$$

$$z_t = \sigma(W_z z_t + U_z h_{t-1}) \tag{11}$$

$$\tilde{h}_t = tanh(W_h z_t + U_h(r_t \odot h_{t-1})) \tag{12}$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \tag{13}$$

where r_t and z_t were reset gate and update gate. $\odot$ was Hadamard product. W and U were trainable weight matrices. By finally hiding the attention weight vector α of state h_T, the dynamic importance of each channel could be calculated as follows.

$$\alpha_i = \frac{\exp(w^T h_i)}{\sum_{j=1}^{T} \exp(w^T h_j)} \tag{14}$$

where w was a trainable parameter vector. α_i was the contribution degree of the i-th channel in the whole sequence.

Regression Model Construction

After feature dimensionality reduction *via* deep forest and dynamic fusion through the MAFF mechanism, a regression estimator based on gradient boosting decision tree (GBDT) was constructed to map the optimized fused features to the two-dimensional valence-arousal emotional space. As a powerful ensemble learning algorithm, GBDT was employed as the final regression predictor in this framework to achieve quantitative output of music-induced emotional regulation effects. According to Russell's circumplex model of affect [27], the valence dimension described the pleasantness of emotions, ranging from negative such as sadness and stress to positive such as happiness and relaxation. The arousal dimension reflected the intensity of physiological and psychological activation, ranging from low arousal such as calmness and drowsiness to high arousal such as excitement and tension. These two orthogonal dimensions formed a continuous space, which could effectively map and quantify the complex emotional states induced by music. Letting the training sample set be $\{(x_i, y_i)\}_{i=1}^{N}$, where $x_i \in R^d$ was the multimodal fusion feature vector of the i-th sample, and $y_i \in R^2$ was the corresponding emotion tag vector including two dimensions of potency and arousal, the GBDT model constructed the final prediction function by superimposing M-tree learning tree f_m as follows.

$$\hat{y}_i = \sum_{m=1}^{M} f_m(x_i), f_m \in F \tag{15}$$

where F was the function space of the regression tree. The goal of the model was to minimize the loss function L, and the square error loss was usually selected as follows.

$$L = \sum_{i=1}^{N} \|y_i - \hat{y}_i\|^2 \tag{16}$$

During the training process, each iteration updated the model by fitting the current residual $r_i^{(m)}$ as follows.

$$r_i^{(m)} = y_i - \hat{y}_i^{(m-1)} \tag{17}$$

$$f_m = arg \min_{f \in F} \sum_{i=1}^{N} (r_i^{(m)} - f(x_i))^2 \tag{18}$$

$$\hat{y}_i^{(m)} = \hat{y}_i^{(m-1)} + \eta f_m(x_i) \tag{19}$$

where η was the learning rate, which was used to control the contribution of each tree to the overall model and prevent over-fitting. Aiming at multi-dimensional emotional output, the model adopted multi-task learning structure, optimized titer and arousal as two regression tasks, and captured the potential correlation between them by using shared tree structure. The loss function was extended to the weighted sum of two-dimensional outputs as below.

$$L = \sum_{i=1}^{N}(\alpha\left\|y_i^{valence} - \hat{y}_i^{valence}\right\|^2 + (1 - \alpha)\left\|y_i^{arousal} - \hat{y}_i^{arousal}\right\|^2) \quad (20)$$

where $\alpha \in [0,1]$ was a hyperparameter for adjusting the weights of two tasks. At the input end of the model, the multi-modal feature vectors were extracted and screened, and the features were normalized by the system to ensure that all input variables were at similar numerical scales to avoid unstable training caused by differences in feature scales. The minimum-maximum normalization method was adopted in the normalization process as follows.

$$x_i^{norm} = \frac{x_i - \min(X)}{\max(X) - \min(X)} \quad (21)$$

where $\min(X)$ and $\max(X)$ were the minimum and maximum values of the characteristic column, respectively. The training of GBDT model adopted greedy segmentation strategy layer by layer. Aiming at the training samples of current nodes, the decline of loss function was maximized by selecting the optimal features and segmentation points. The goal of node division was to maximize information gain as defined below.

$$\Delta L = L_{parent} - (L_{left} + L_{right}) \quad (22)$$

where L_{parent}, L_{left} and L_{right} were the square error losses of the parent node and the left and right child nodes, respectively. By traversing all possible segmentation points and features, the partition scheme that maximized ΔL was selected. To prevent over-fitting, regularization terms were introduced into the model including the maximum depth limit of the tree, the minimum sample size limit of leaf nodes, and L2 regularization of leaf weights as shown below.

$$\Omega(f) = \gamma \mathrm{T} + \frac{1}{2}\lambda\sum_{j=1}^{T} w_j^2 \quad (23)$$

where T was the number of leaf nodes. w_j was the weight of the j-th leaf. γ and λ were regularization hyperparameters to adjust the complexity of the model respectively.

Experimental setup and implementation protocol

All experiments in this study were conducted in a unified computing environment to ensure the fairness of results. The hardware platform was equipped with an Intel Core i9-10900K CPU and an NVIDIA GeForce RTX 3090 GPU (24 GB video memory). The software environment was based on the 64-bit Windows 10 operating system using Python 3.8 programming language. The deep learning model was built on the PyTorch 1.10 framework, and GBDT was implemented using the Scikit-learn 1.0 library (https://scikit-learn.org). To evaluate the model performance, a 10-fold cross-validation strategy was adopted. The dataset was randomly divided into a training set (80%), a validation set (10%), and a test set (10%). To verify the superiority of the proposed method, traditional support vector machine (SVM), shallow multimodal fusion models, and a deep forest model without an attention mechanism were selected as comparative baselines in the study. All comparative experiments used the same preprocessed data and evaluation metrics.

Experimental comparison model and evaluation index

The baseline models and performance evaluation metrics for comparative verification were first clarified. To verify the advancement of the proposed method, it was compared with traditional single-modal model, shallow multimodal fusion model, and deep forest without attention mechanism. Traditional single-

Table 1. Paired t-test statistical significance analysis results of proposed method and other baseline models on DEAP dataset.

Comparison	Valence		Awakening degree	
	t-value	*P* value	t-value	*P* value
Proposed method vs. Single modality EEG	-12.45	< 0.001	-14.21	< 0.001
Proposed method vs. Single modality ECG	-15.82	< 0.001	-16.03	< 0.001
Proposed method vs. SVM baseline	-18.33	< 0.001	-19.45	< 0.001
Proposed method vs. Multimodal (w/o attention)	-4.12	< 0.01	-5.67	< 0.001

Note: a negative t-value indicated that the error of the proposed method was significantly lower than that of the comparative model.

modal model used support vector regression (SVR) for EEG and ECG data, respectively. This model is implemented based on the Scikit-learn machine learning library. Shallow multimodal fusion model directly concatenated the extracted multimodal features and input them into a fully connected neural network without involving deep feature extraction or attention mechanisms. Deep forest without attention mechanism only used the cascaded forest structure for regression, removing the proposed MAFF module. To comprehensively quantify the model performance, this study adopted four metrics. Mean squared error (MSE) was applied to measure the average of the squared differences between predicted values and true values, reflecting the overall error level of the model. Root mean square error (RMSE) was the square root of MSE, which was more sensitive to outliers and had the same unit as the original data. The mean absolute error (MAE) was the average of the absolute differences between predicted values and true values, reflecting the actual magnitude of prediction deviations. Coefficient of determination (R^2) was used to evaluate the model’s ability to explain the variability of data with the values closer to 1 indicating better fitting performance.

Results and discussion

The prediction performance of valence and arousal across different datasets demonstrated that on the DEAP dataset, the proposed method achieved the best results for both emotional dimensions. Specifically, the MSE for valence was 0.0410, significantly lower than those of single-modality EEG (0.0653) and ECG (0.0721). For arousal, the MSE was 0.0380, also outperforming EEG (0.0687) and ECG (0.0745). Furthermore, the R^2 scores reached 0.81 and 0.83 for valence and arousal, respectively, indicating a high level of predictive accuracy. Compared with multimodal fusion methods without attention mechanisms, the proposed multi-source attention mechanism substantially enhanced the model's ability to integrate multimodal information, resulting in more accurate emotion prediction. On the PMEmo dataset, the proposed approach also demonstrated superior performance. The MSE for valence was 0.0380 with an R^2 of 0.84, while, for arousal, the MSE was 0.0361 with an R^2 of 0.85. In contrast, single-modality approaches performed significantly worse with ECG alone yielding an MSE of 0.0783 in arousal prediction, which was nearly twice that of the proposed method (Figure 1).

To further verify the statistical significance of the above performance improvements, this study conducted paired-samples t-tests on the results of 10-fold cross-validation. The results showed that the proposed method demonstrated significant performance improvement compared to single-modal methods of EEG, ECG, and the traditional SVM model (P < 0.001). More importantly, in the comparison verifying the effectiveness of the MAFF mechanism, the proposed method was significantly better than multimodal fusion without an attention mechanism in both valence dimension (P < 0.01) and the arousal dimension (P < 0.001) (Table 1). The result statistically confirmed that introducing the GRU-based dynamic attention mechanism could significantly reduce prediction errors,

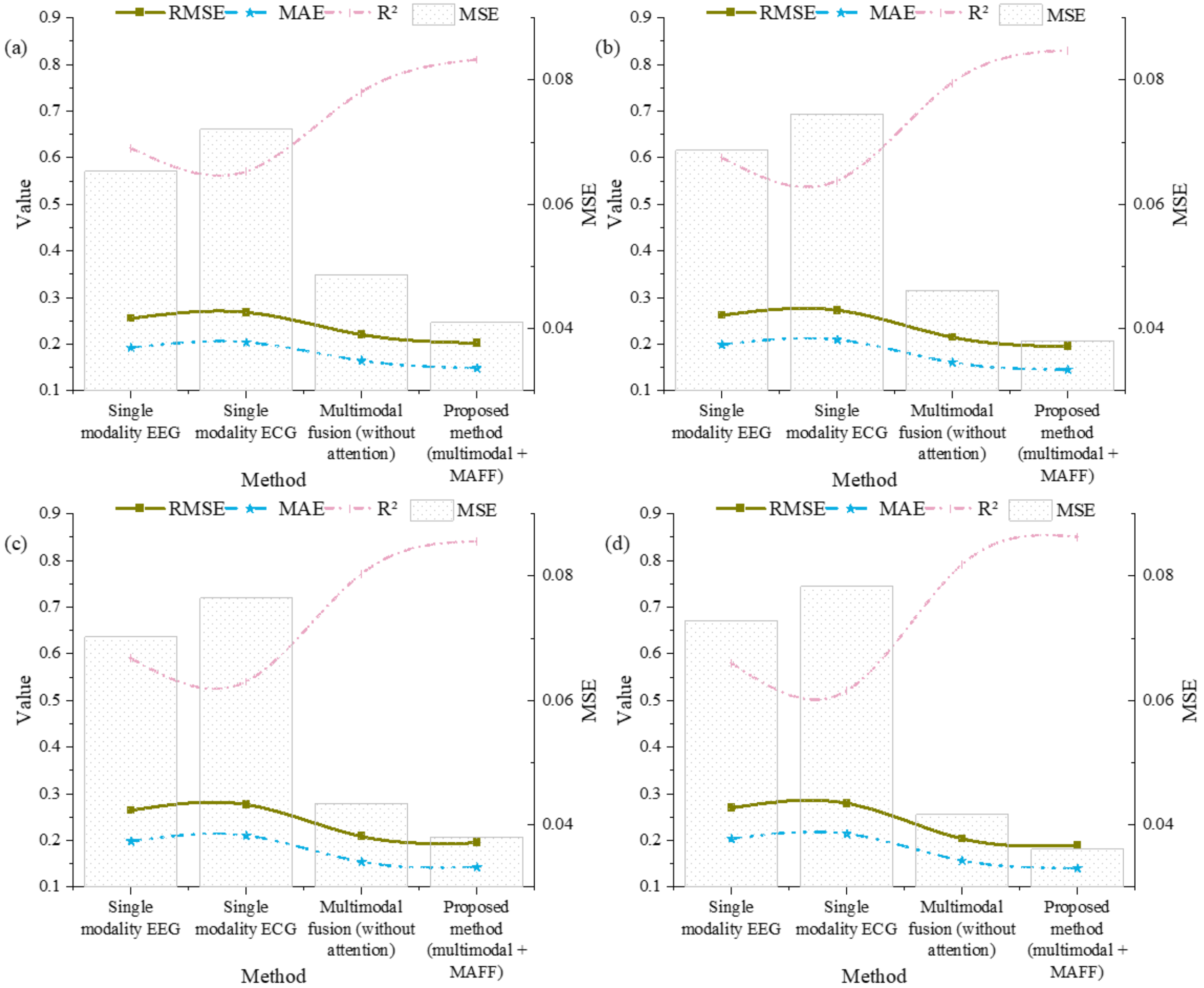

Figure 1. The potency and wake-up prediction performance of different datasets. (**a**) the potency prediction of DEAP dataset. (**b**) wake-up prediction of DEAP dataset. (**c**) titer prediction of PMEmo dataset. (**d**) PMEmo dataset wake-up prediction. (**Note**: the values of the left vertical axis corresponded to the values of RMSE, MAE, and R^2).

rather than merely random fluctuations, thereby demonstrating the core contribution of the MAFF module in capturing cross-modal dynamic correlations.

A comparison of the individual prediction performance across different modalities showed that there were clear differences in the modeling capabilities of each modality when used independently in the DEAP dataset. The EEG modality achieved the lowest MSE in both valence and arousal dimensions with values of 0.0653 and 0.0687, respectively, resulting in an average MSE of 0.0670. The results indicated that the spatiotemporal features extracted from EEG signals provided strong discriminative power for emotion estimation. In contrast, the EDA modality showed the weakest performance with an average MSE of 0.0798, which might be attributed to its relatively low feature dimensionality and limited ability to capture the complexity of emotional states. On the PMEmo dataset, a similar trend was observed, where EEG again outperformed the other modalities, achieving the lowest average MSE of 0.0715, which further confirmed the robustness and generalizability of EEG features across datasets (Figure 2). Other modalities such as ECG and EMG

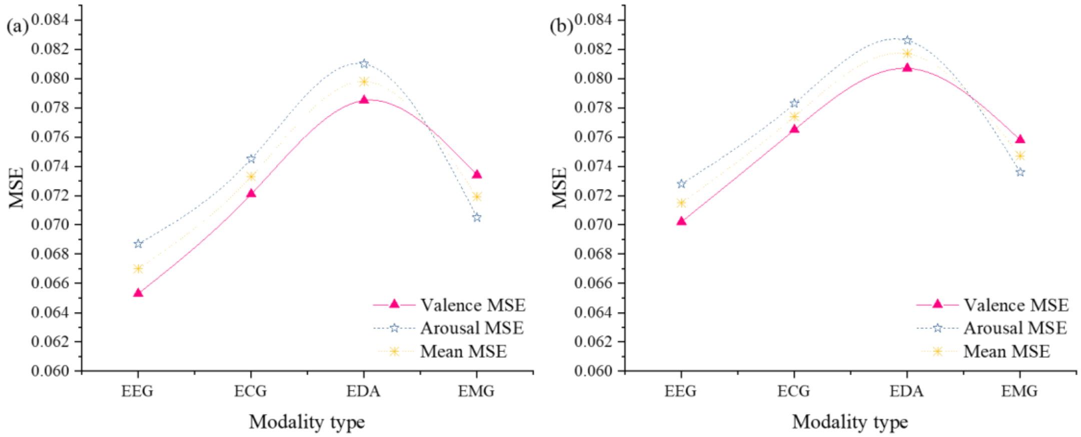

Figure 2. Comparison of individual prediction performance of different modes. (**a**) DEAP dataset. (**b**) PMEmo dataset.

showed comparable performance in both datasets, indicating their potential in predicting emotion regulation states, though still falling short of EEG in predictive accuracy.

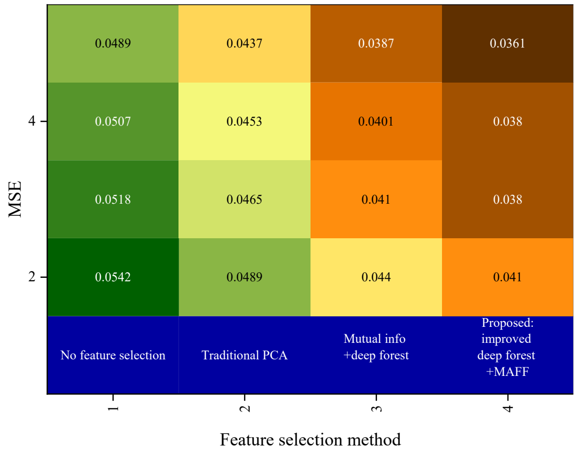

Figure 3. Comparison of different feature selection strategies for multimodal fusion.

A comparison of different feature selection strategies for multimodal fusion showed that the proposed method that combined an enhanced deep forest architecture with MAFF achieved significant improvements in feature selection performance. On the DEAP dataset, the MSE for valence and arousal dimensions reached 0.0410 and 0.0380, respectively, substantially outperforming traditional approaches such as principal component analysis (PCA) and mutual information-based selection. Compared to the baseline model without any feature selection strategy, the proposed method reduced MSE by approximately 24% in the valence dimension and 26% in the arousal dimension (Figure 3). These results demonstrated the effectiveness of the joint deep forest-MAFF approach in enhancing feature relevance and model generalization in multimodal emotional state prediction. The MAFF played a crucial role in feature fusion. After integrating MAFF into the system, the MSE for both valence and arousal significantly decreased on the DEAP and PMEmo datasets (Figure 4).

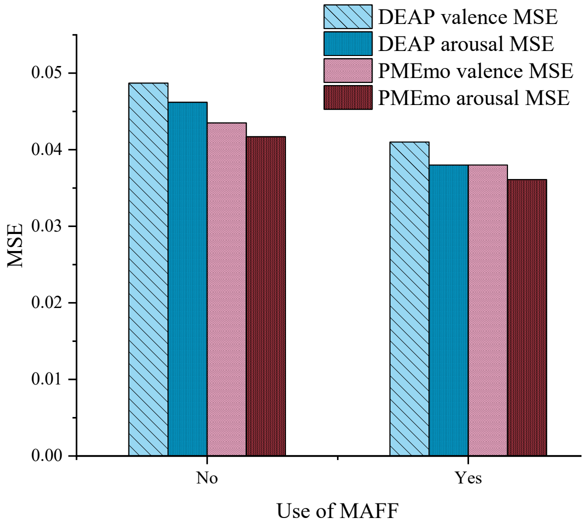

Figure 4. MAFF contribution analysis.

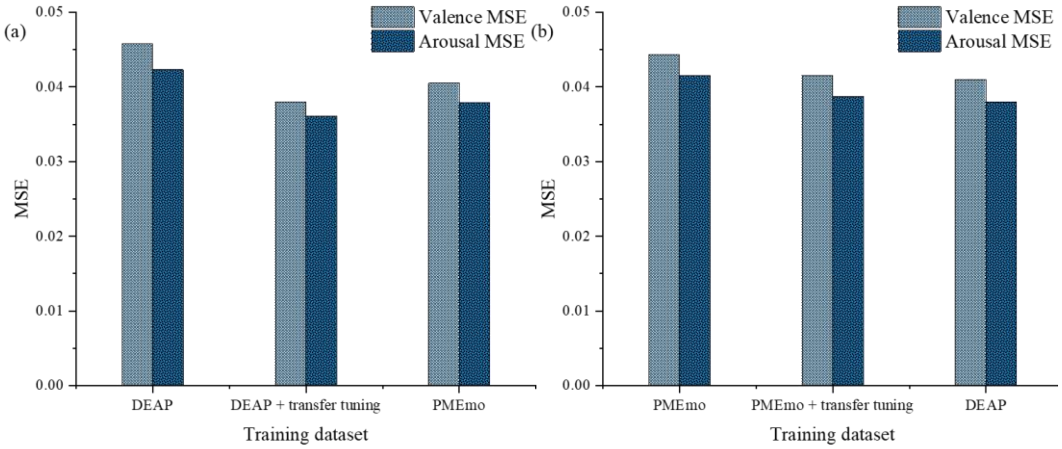

Figure 5. The influence of transfer learning strategy on the generalization performance of the model. (**a**) PMEmo dataset. (**b**) DEAP dataset.

The impact of transfer learning on model generalization showed that, when using PMEmo as the test dataset, the model being trained solely on DEAP without transfer learning yielded MSE of 0.0458 for valence and 0.0423 for arousal, significantly higher than the baseline performance achieved through direct training on the target dataset. After applying transfer learning with fine-tuning, the MSE dropped to 0.0380 and 0.0361, outperforming the models trained directly on PMEmo. The results indicated that the transfer strategy effectively enhanced the model's adaptability to new data domains. Similarly, in evaluations using DEAP as the test set, models transfer from PMEmo and fine-tuned showed improved performance. The valence and arousal MSE reduced from 0.0443 and 0.0415 to 0.0415 and 0.0387, respectively, approaching or even matching the performance of models trained directly on DEAP (Figure 5).

Model training time and complexity comparison demonstrated that the proposed model combined deep forest and MAFF as a complete system including feature extraction and fusion modules with the parameter quantity covering the entire network, while GBDT ensemble model played as the independent baseline model using only GBDT. Although the proposed method involved a larger number of parameters and longer training time compared to traditional models with 3.8 million parameters and 3.6 minutes of training, it maintained a reasonable inference time of 15.8 milliseconds per sample (Figure 6). While ensuring high prediction accuracy, the system also demonstrated balanced computational efficiency, highlighting its potential for real-time applications.

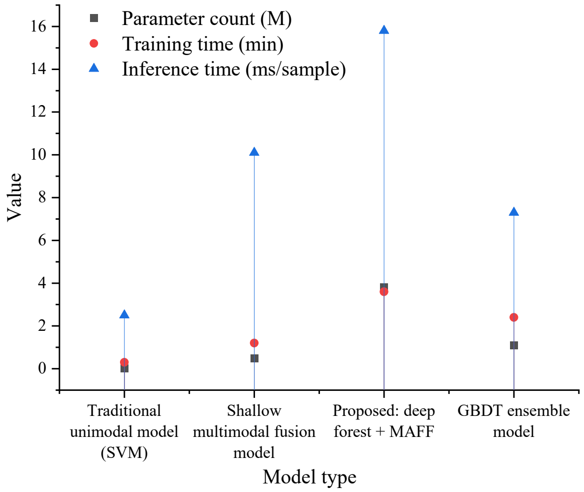

Figure 6. Comparison of model training time and complexity. (**Note:** The ordinate "Value" corresponded to the different measurement units with the parameter quantity unit as millions (m), the training time unit as minutes (min), and the reasoning time unit as milliseconds per sample (ms/sample).

Conclusion

This study proposed a multimodal physiological signal fusion-based system for evaluating the effects of music-induced emotion regulation. The framework integrated EEG, ECG, EDA, and EMG data to provide an objective assessment. A combination of temporal convolutional networks, phase locking values, adaptive heart rate algorithms, and multi-source attention mechanisms was employed for efficient feature extraction and fusion. The results showed that the proposed method significantly outperformed traditional approaches in predicting valence and arousal, demonstrating strong generalization capability. Despite limitations such as dataset discrepancies and reduced performance in extreme emotion recognition, future work might incorporate behavioral data and explore lightweight deployment strategies to extend its application potential in mental health field.

Acknowledgements

This work was supported by Jiangsu Maritime Institute Doctoral Research Startup Project (Grant No. 2024BSKY01).

References

1. Wang S, Xu C, Ding AS, Tang Z. 2021. A novel emotion-aware hybrid music recommendation method using deep neural network. Electron. 10(15):1769.
2. Garg A, Chaturvedi V, Kaur AB, Varshney V, Parashar A. 2022. Machine learning model for mapping of music mood and human emotion based on physiological signals. Multimed Tools Appl. 81(4):5137-5177.
3. Lin W, Li C. 2023. Review of studies on emotion recognition and judgment based on physiological signals. Appl Sci. 13(4):2573.
4. Wang L, Hao J, Zhou TH. 2023. ECG multi-emotion recognition based on heart rate variability signal features mining. Sensors. 23(20):8636.
5. Udahemuka G, Djouani K, Kurien AM. 2024. Multimodal emotion recognition using visual, vocal and physiological signals: A review. Appl Sci. 14(17):8071.
6. Kim HG, Lee GY, Kim MS. 2021. Dual-function integrated emotion-based music classification system using features from physiological signals. IEEE Trans Consum Electron. 67(4):341-349.
7. Yin G, Sun S, Yu D, Li D, Zhang K. 2022. A multimodal framework for large-scale emotion recognition by fusing music and electrodermal activity signals. ACM Trans Multimed Comput Commun Appl. 18(3):1-23.
8. Zhu X, Guo C, Feng H, Huang Y, Feng Y, Wang X, *et al*. 2024. A review of key technologies for emotion analysis using multimodal information. Cogn Comput. 16(4):1504-1530.
9. Zhu L, Ding Y, Huang A, Tan X, Zhang J. 2025. MF-Net: A multimodal fusion network for emotion recognition based on multiple physiological signals. Signal Image Video Process. 19(1):58.
10. Du R, Zhu S, Ni H, Mao T, Li J, Wei R. 2023. Valence-arousal classification of emotion evoked by Chinese ancient-style music using 1D-CNN-BiLSTM model on EEG signals for college students. Multimed Tools Appl. 82(10):15439-15456.
11. Ghaleb E, Niehues J, Asteriadis S. 2023. Joint modelling of audio-visual cues using attention mechanisms for emotion recognition. Multimed Tools Appl. 82(8):11239-11264.
12. Yang K, Wang C, Gu Y, Sarsenbayeva Z, Tag B, Dingler T, *et al*. 2021. Behavioral and physiological signals-based deep multimodal approach for mobile emotion recognition. IEEE Trans Affect Comput. 14(2):1082-1097.
13. Vamsidhar D, Desai P, Shahade AK, Patil S, Deshmukh PV. 2025. Hierarchical cross-modal attention and dual audio pathways for enhanced multimodal sentiment analysis. Sci Rep. 15(1):25440.
14. Zhang J, Chen W. 2025. A decade of music emotion computing: A bibliometric analysis of trends, interdisciplinary collaboration, and applications. Educ Inf. 41(3):227-255.
15. Cai Y, Li X, Zhang Y, Li J, Zhu F, Rao L. 2025. Multimodal sentiment analysis based on multi-layer feature fusion and multi-task learning. Sci Rep. 15(1):2126.
16. Chou YC, Chien SK, Chao PC, Lin YJ, Chen CY, Yeh KK, *et al*. 2025. SEM-Net: A social–emotional music classification model for emotion regulation and music literacy in individuals with special needs. Appl Sci. 15(8):4191.
17. Li Z, Zhang G, Wang L, Wei J, Dang J. 2023. Emotion recognition using spatial-temporal EEG features through convolutional graph attention network. J Neural Eng. 20(1):016046.
18. Zhao Y, Guo M, Chen X, Sun J, Qiu J. 2023. Attention-based CNN fusion model for emotion recognition during walking using discrete wavelet transform on EEG and inertial signals. Big Data Min Anal. 7(1):188-204.
19. Koelstra S, Muhl C, Soleymani M, Lee JS, Yazdani A, Ebrahimi T, *et al*. 2011. DEAP: A database for emotion analysis using physiological signals. IEEE Trans Affect Comput. 3(1):18-31.
20. Zhang K, Zhang H, Li S, Yang C, Sun L. 2018. The PMEmo dataset for music emotion recognition. In: Proc ACM Int Conf Multimed Retrieval. p. 135-142.
21. Khan M, Tran PN, Pham NT, El Saddik A, Othmani A. 2025. MemoCMT: Multimodal emotion recognition using cross-modal transformer-based feature fusion. Sci Rep. 15(1):5473.
22. Li Y, Guo W, Wang Y. 2024. Emotion recognition with attention mechanism-guided dual-feature multi-path interaction network. Signal Image Video Process. 18(Suppl 1):617-626.
23. Hassan CAU, Ehatisham-ul-Haq M, Murtaza F, Yasin AU, Ullah SS. 2025. EmoTrans attention-based emotion recognition using EEG signals and facial analysis with expert validation. Sci Rep. 15(1):22004.
24. García-Hernández RA, Luna-García H, Celaya-Padilla JM, García-Hernández A, Reveles-Gómez LC, Flores-Chaires LA, *et al*. 2024. A systematic literature review of modalities, trends, and

limitations in emotion recognition, affective computing, and sentiment analysis. Appl Sci. 14(16):7165.

25. Bao Y, Xue M, Gohumpu J, Cao Y, Weng S, Fang P, *et al*. 2024. Prenatal anxiety recognition model integrating multimodal physiological signal. Sci Rep. 14(1):21767.
26. Ma Y, Huang Z, Yang Y, Zhang S, Dong Q, Wang R, *et al*. 2024. Emotion recognition model of EEG signals based on double attention mechanism. Brain Sci. 14(12):1289.
27. Russell JA. 1980. A circumplex model of affect. J Pers Soc Psychol. 39(6):1161.