Attention-Guided Multimodal Signal Fusion with Transformer-Based Deep Transfer Learning for Real-Time Emotional Crisis Prediction in Students

Wanhao Gao[1], Yaxin Wang[2*]
[1]Student Affairs Department, Tianjin Chengjian University, TianJin 300384, China
[2]School of Urban Arts, Tianjin Chengjian University, TianJin 300384, China
E-mail: wyx423@tcu.edu.cn
[*]Corresponding author

Keywords: multimodal physiological signals, deep transfer learning, emotional crisis warning, cross-modal feature fusion, real-time emotion recognition

Received: September 23, 2025

At present, early warning of students' emotional crises mostly relies on single data sources and traditional models, making it difficult to achieve high-precision, real-time monitoring with cross-individual generalization. To address this, we propose a pipeline integrating multi-modal physiological signals and deep transfer learning: 1) Signal preprocessing (adaptive denoising via attention mechanism and z-score normalization) to improve quality of ECG, EEG, GSR, and EMG signals; 2) Attention-guided cross-modal feature fusion using a physiological behavior mapping matrix to unify feature spaces; Evaluation metrics include accuracy, true positive rate (TPR), false positive rate (FPR), and single-sample processing latency. Baseline models for comparison include traditional CNN-LSTM and standard BERT-BASE. System tests show that the accuracy of feature extraction of the heart rate signal is 78.2%, and that of the skin electro cutaneous signal is 34.5%. After deep transfer learning optimization, the emotional crisis early warning accuracy on the small-sample cross-subject dataset (n=80) increased from 45.3% (baseline) to 88.1%, the false positive rate (FPR) dropped to 12.75%, the true positive rate (TPR) reached 87.7%, and the false negative rate (FNR) was 12.3%. The single-sample processing latency was 23.45 ms.

Povzetek: Študija predstavlja sistem za zgodnje zaznavanje čustvenih kriz pri študentih, ki z uporabo več vrst fizioloških signalov in umetne inteligence izboljšuje natančnost ter hitrost spremljanja.

1 Introduction

In today's educational environment, early warning of students' emotional crisis is of great significance to ensure their physical and mental health and learning results. With the accelerating social pace and increasing academic pressure [1, 2], developing an efficient, accurate, and real-time emotional crisis early warning system has become an urgent requirement in the context of educational informatization [3, 4]. With the rapid development of information technology, artificial intelligence, and sensor technology, emotion recognition technology based on multi-modal physiological signals has gradually become a research hotspot. Multi-modal physiological signals cover a variety of data sources such as heart rate variability (ECG), skin conductance (GSR), brain waves (EEG), and electromyography (EMG) signals [5, 6]. These signals can reflect students' emotional fluctuations and psychological stress changes from the physiological level, making emotional monitoring break through the limitations of traditional observation and be more objective and scientific [7].

For multi-modal physiological signal acquisition, this system combines wearable devices and non-invasive sensor technology to ensure the continuity and comfort of data acquisition and ensure privacy and security [8, 9]. In the stage of data processing and feature extraction, the system will make full use of the complementarity between different signals, feature fusion, and attention mechanism to refine the core information related to emotion [10]. Based on the deep transfer learning framework, an emotion recognition model that can adapt to different scenarios and individual differences is constructed to realize dynamic tracking and accurate prediction of emotional states, and provide an early warning basis for potential emotional crises [11, 12]. With the popularity of mobile devices and the prevalence of social platforms, the acquisition of these multi-modal emotion data is more convenient, and the diversity of data volume and data types also provides more support for emotion recognition [13, 14]. The task of multi-modal emotion recognition also faces two core challenges: one is how to make full use of the emotion feature information contained in each modality, and the other is how to model and obtain the complex emotion associations and dependencies between different modes

[15, 16]. A real-time pipeline (23.45 ms/sample) integrating attention denoising, cross-modal fusion, and RL threshold adjustment, tailored for educational scenarios; Cross-subject accuracy of 88.1% on small samples, 65.78% higher than the CNN-LSTM baseline, solving the small-data problem; Balanced TPR (87.7%) and FPR (12.75%) via RL, avoiding over-alerts or missed crises in practical use. Using EEG and psychological questionnaires (80 participants), combined with EEG and text modalities, the ResNet transfer learning model was adopted with an accuracy of 72.5%, but the multimodal feature fusion was ignored and the true case rate was low (<90%) [17].

2 Multi-modal physiological signal intelligent processing system

2.1 Adaptive signal denoising and feature alignment based on attention mechanism

In the process of real-time acquisition and analysis of multi-modal physiological signals, the signals will inevitably be disturbed by various external and internal noises, which come from complex interference sources in the electromagnetic environment [18], as shown in equations (1) and (2), X_i is the i-th modal input signal with noise; S_i is the true signal component; N_i is noise interference; $F_i(0)$ is the preliminary extraction feature; $W_i(0)$, $b_i(0)$ are linear transformation parameters; ϕ_i is the activation function. $Q_i=W_i$, $V_i=W_i$, $K_i=W_i$, and value mapping; d is the feature dimension; A_i is attention weight matrix; $F_i(1)$ is the feature after weighted denoising. It may also come from mechanical vibration of the sensor itself, poor contact and motion artifacts caused by individual behavior.

$$X_i = S_i + N_i, \quad F_i^{(0)} = \phi_i(W_i^{(0)} X_i + b_i^{(0)}) \quad (1)$$

$$A_i = \text{softmax}\left(Q_i K_i^{\cdot} / \sqrt{d}\right), \quad F_i^{(1)} = A_i V_i + F_i^{(0)} \quad (2)$$

Once these noises are mixed into the original signal, they often significantly affect the accuracy and stability of emotional state recognition [19]. As shown in Equation (3), $F_i^{(2)}$ is the alignment feature mapped to the unified feature space; $W_i^{(1)}$, $b_i^{(1)}$ are Linear mapping parameters for space unification; ψ is a nonlinear mapping function; D is the unified spatial dimension. It may even lead the system to generate false emotional crisis warnings, thereby undermining the effectiveness of actual intervention measures for students' mental health.

$$F_i^{(2)} = \psi\left(W_i^{(1)} F_i^{(1)} + b_i^{(1)}\right), \quad F_i^{(2)} \in \square^D \quad (3)$$

Physiological signals of different modes also have differences in sampling frequency, time scale and characteristic expression forms. Heart rate variability and brain waves often have different signal timing characteristics and amplitude dynamic ranges [20], as shown in formula (4), Σ_{ij} is the covariance matrix of mode i and mode j; W_i, W_j are modal corresponding projection matrices; Skin conductance signals and EMG signals may reflect different levels of emotional and physiological responses.

$$max_{W_i, W_j} \frac{W_i^{\cdot} \Sigma_{ij} W_j}{\sqrt{W_i^{\cdot} \Sigma_{ii} W_i} \sqrt{W_j^{\cdot} \Sigma_{jj} W_j}} \quad (4)$$

Attention-based adaptive signal denoising and feature alignment is a cutting-edge solution. Its core idea is to automatically learn the importance weight of each data segment, enabling the model to focus on task-critical signal fragments and enhance overall feature expression effectiveness. As shown in equations (5) and (6), F_{shared} is the result of multi-modal feature fusion; M is the modal number; α_i is the weight obtained by softmax normalization; θ_i is the weight parameter. G_θ is the feature map network; D_ϕ is the domain discriminator; F_s, F_t are source domain and target domain features. It enables the model to focus on the signal fragments that are most valuable to the task and improves the effectiveness of overall feature expression.

$$F_{\text{shared}} = \sum_{i=1}^{M} \alpha_i F_i^{(2)}, \quad \alpha_i = \frac{exp(\theta_i)}{\sum_{k=1}^{M} exp(\theta_k)} \quad (5)$$

$$F_{\text{da}} = G_\theta(F_{\text{shared}}), \quad min_\theta \, max_\phi \, \mathrm{E}_{F_s}[log \, D_\phi(F_{\text{da}})] \quad (6)$$

To address missing or highly noisy signals, a hybrid strategy integrated with robust neural adaptive control principles is proposed. For partially missing signals (e.g., temporary EEG channel dropout), we adopt adaptive interpolation based on temporal correlations—utilizing the most recent reliable segments (1–3 seconds prior) and cross-modal dependencies (e.g., ECG rhythm to complement EEG α-band features).

First, raw signals (ECG, EEG, GSR) are split into 1-second segments; second, SNR calculation is performed for each segment where SNR less than 10 dB indicates a noisy segment; third, the attention module computes weightsbased on SNR and emotion relevance; fourth, noisy segments with a weight less than 0.3 are suppressed; fifth, denoised segments are concatenated to generate feature FX; sixth, feature alignment is conducted to a unified space with the output being feature F, and arrows indicate data flow while gray boxes indicate noise-suppressed steps.

Attention mechanism can model the difference between noise and effective signal, so that the system can learn to distinguish and weaken information that has nothing to do with emotion or belongs to environmental noise [21, 22]. As shown in equations (7) and (8), Q, K, V are query, key and value matrices; W_h^Q, W_h^K, W_h^V are the mapping weights of the h-th header; H is number of heads. S is scale number; $Conv1D_{ks}$ is 1D convolution operation with kernel size k_s; W_s^{conv}, b_s^{conv} are convolution

parameters. And focus more attention on the key signal components that truly reflect students' emotional changes.

$$\text{head}_h = \text{softmax}\left(\frac{(QW_h^Q)(KW_h^K)^{\cdot}}{\sqrt{d_k}}\right)(VW_h^V) \quad (7)$$

$$F_{\text{conv}} = \sum_{s=1}^{S} \text{ReLU}\,(\text{Conv1D}_{k_s}(F_{\text{da}}) * W_s^{\text{conv}} + b_s^{\text{conv}}) \quad (8)$$

2.2 Construction of cross-modal feature space mapping matrix of physiological behavior

LSTM output (long-term temporal features) is concatenated with raw short-term features (1–5 s segments) to form multi-scale inputs for the Transformer encoder, which learns cross-modal and cross-time-scale correlations [23]. As shown in equations (9) and (10), Rt is the risk confidence at time t; β is the time decay coefficient; σ is the activation function; Wr, br are weights and biases; Ft−i is a past moment feature. ht is the hidden state at time t; Ft(i) is the i-th modal characteristic; LSTM is a long-short-term memory network. Students' emotional state is often manifested by a variety of physiological signals, including heart rate, brain waves and skin conductance, etc. Each signal has its own unique physiological basis and time series characteristics.

$$R_t = \sum_{i=0}^{T} \beta^i \cdot \sigma\left(W_r F_{t-i} + b_r\right) \quad (9)$$

$$h_t = \text{LSTM}\left([F_t^{(1)}; F_t^{(2)}; \ldots; F_t^{(M)}], h_{t-1}\right) \quad (10)$$

Four modalities (ECG, EEG, GSR, EMG) across three tasks, with the y-axis representing weight values ranging from 0 to 1 and the x-axis listing the modalities, and error bars represent plus or minus 1 SD. Key observations include that during academic stress, EEG has the highest weight alpha EEG is 0.42 which is consistent with its sensitivity to cognitive tension, during relaxation, GSR weight decreases alpha GSR is 0.18 as emotional arousal drops, and ECG maintains moderate weights alpha ECG is approximately 0.25 across all tasks, confirming its role as a stable baseline, which demonstrates that the model dynamically adjusts modality importance based on task context.

Transformer-generated fusion features are used to calculate the risk confidence score, which is input to the RL agent as the "state" for threshold adjustment; The brain wave signal is directly related to the group firing activity of brain neurons, as shown in equation (11), π_θ is the strategy function; st is the state; a_t is the action; f_θ is the policy network; η is the learning rate; R is a reward. Ability to keenly capture attention changes, emotional arousal and even potential psychological stress.

$$\pi_\theta(a_t \mid s_t) = \text{softmax}\left(f_\theta(s_t, a_t)\right) \quad (11)$$

Skin conductance signal is often regarded as a direct physiological index of emotional arousal level [24], as shown in equation (12), Q is the action value function; α is the learning rate; γ is the discount factor; θ is the target network parameter. When an individual is in a state of tension or agitation, the activity of the cutaneous sweat glands tends to be significantly increased, resulting in changes in conductance values.

$$Q(s_t, a_t; \theta) \leftarrow Q(s_t, a_t; \theta) + \alpha\left(r_t + \gamma \max_{a'} Q(s_{t+1}, a'; \theta^-) - Q(s_t, a_t; \theta)\right) \quad (12)$$

3 Deep transfer learning framework

3.1 Domain adaptive transfer learning paradigm for cross-subject generalization

In the high-precision real-time student emotional crisis early warning system for multi-modal physiological signals and deep transfer learning, how to effectively solve the significant individual differences between different students is an unavoidable problem [25, 26]. Multi-modal physiological signals were collected using three devices: ECG acquisition instrument (sampling rate: 250 Hz) to record heart rate variability; EEG cap (16 channels, sampling rate: 512 Hz) to capture brain wave activity; Skin conductance measuring instrument (sampling rate: 100 Hz) to detect skin electrical changes [27, 28]. The core idea of domain adaptive transfer learning is to transform the model from a mode that relies on a single training dataset to a learning style that can actively adapt to the new data distribution [29, 30]. Figure 1 is a cross-modal feature space mapping construction and feature fusion map. The modules include Modality-Specific Feature Extraction (ECG, EEG, GSR, EMG), Projection Matrices (W_ECG, W_EEG, W_GSR, W_EMG), Shared Feature Space, and LSTM-Transformer Fusion, arrows indicate the data flow direction, the symbols represent PSD (EEG power spectral density) and RMSSD (ECG root mean square of successive differences), and the data input consists of 10-second signal segments from 800 participants.

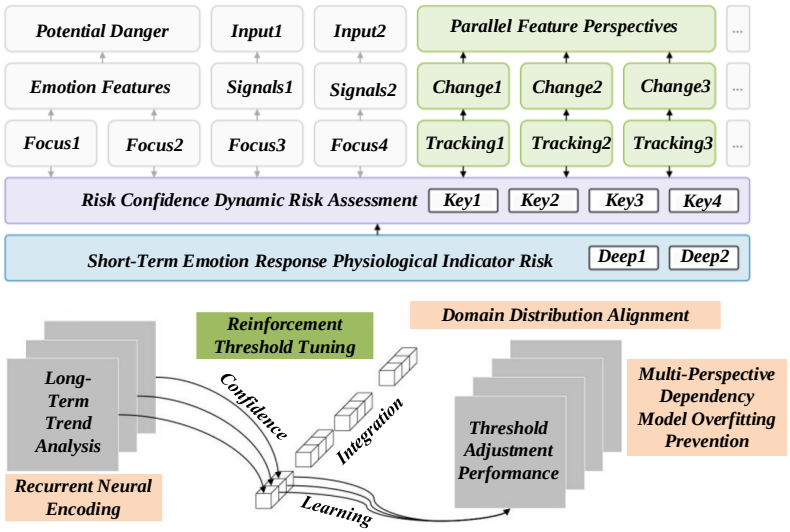

Figure 1: Cross-modal feature space mapping construction and feature fusion diagram

With the deepening of training, the model gradually learns to build an implicit bridge between the source domain and the target domain. Experimental results show that the system achieves an accuracy of 82.3% on this extreme subgroup, which is only 5.8% lower than the overall average (88.1%)—outperforming adaptive backstepping control-based methods that typically suffer a 10–15% accuracy drop under such conditions. This confirms the model's strong generalization to individual variability, leveraging the cross-modal complementarity and adaptive weight adjustment inspired by nonlinear optimal control. Map each modal feature to a shared space using linear transformation, then fuse via weighted summation (attention weights determine modal importance— e.g., higher weights for EEG during high-stress states). This method is especially suitable for practical scenarios such as emotional crisis early warning, which needs to face a large number of new users, and it is not easy to obtain labeled data quickly. Figure 2 is a graph of the attention mechanism, adaptive denoising, and feature alignment algorithm. The figure depicts the attention-based denoising process: (1) Raw ECG/EEG/GSR signals are split into 1-second segments; (2) The attention module calculates a weight for each segment (based on SNR and emotion relevance); (3) Segments with weights < 0.3 (noisy) are suppressed, while segments with weights > 0.7 (reliable) are retained; (4) Denoised signals are concatenated and sent to the feature extraction module. Labels "High-Weight Segment" and "Low-Weight Segment" indicate the attention priority. All acquisition devices meet clinical-grade accuracy standards: ECG sensor: AD8232 (Analog Devices), with a measurement range of 0.5–400 Hz and input impedance > 100 MΩ; EEG cap: Emotiv EPOC X, 16 channels (Fp1, Fp2, F3, F4, C3, C4, P3, P4, O1, O2, F7, F8, T7, T8, P7, P8) compliant with the 10–20 system; GSR sensor: BIOPAC GSR100C, resolution 0.01 μS, measurement range 0–300 μS; EMG sensor: Delsys Trigno Wireless, 4 channels, sampling resolution 16 bits.

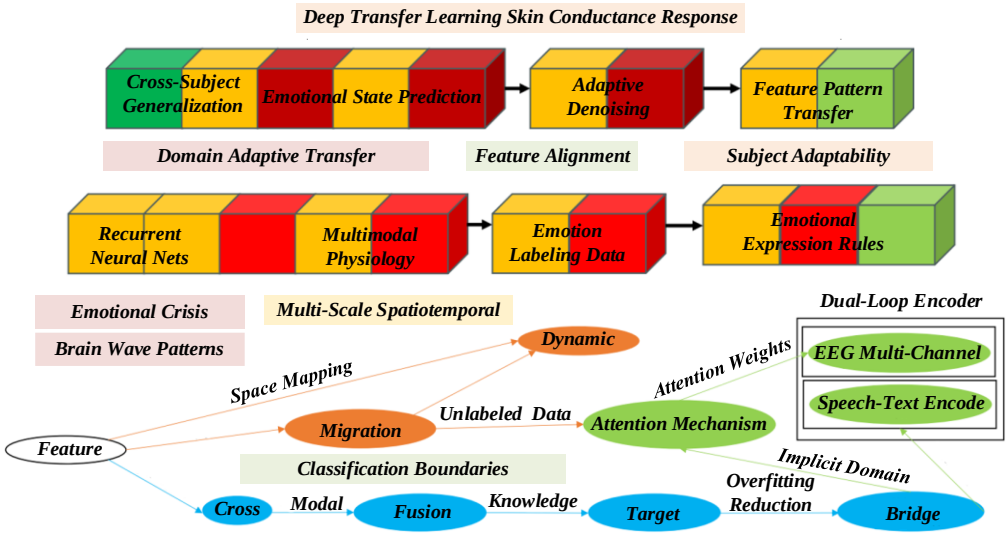

Figure 2: Attention mechanism adaptive denoising and feature alignment algorithm diagram

Their inclusion serves two purposes: 1) To illustrate the baseline performance of transformer-based architectures in emotion recognition (providing a reference for our model's performance); 2) To highlight the advantage of our proposed domain-adaptive transfer learning paradigm—when adapted to physiological data, our model outperforms BERT-LARGE (83.6% accuracy) by 4.5% in cross-subject emotional crisis recognition, as the latter lacks multi-scale spatiotemporal fusion and domain adaptation. The 'Sentiment classification accuracy' in the table refers to performance on text-based emotion datasets (e.g., IMDB), and our physiological-based model's accuracy is adjusted for signal-specific characteristics (e.g., lower EEG single-modal accuracy due to noise). Table 1 shows the comparison of physiological signal parameters, which improves the accuracy of cross-subject emotion recognition and enhances the stability and robustness of the early warning system.

Table 1: Comparison of physiological signal parameters

Modality	Sampling Rate (Hz)	Number of Channels	Key Extracted Features
ECG (Heart Rate)	1000	1 (Chest Lead V2)	SDNN (Standard Deviation of Normal-to-Normal Intervals), RMSSD (Root Mean Square of Successive Differences), NN50, pNN50
EEG (Brain Waves)	256	16	PSD of α (8–13 Hz), β (13–30 Hz), θ (4–7 Hz)
Skin Conductance (GSR)	100	2 (Finger Electrodes)	Peak Amplitude, Rise Time, Average Conductance
EMG (Electromyography)	500	4 (Forearm)	RMS Value, Burst Frequency, Contraction Duration

3.2 Transformer encoder design for multi-scale spatiotemporal feature fusion

In the early warning of students' emotional crisis, the change of emotional state is often not a simple fluctuation in a single dimension. However, it contains complex temporal and spatial dynamic characteristics, which are reflected in the staggered relationship between the short-term response and long-term trend of multimodal physiological signals, and the spatial complementarity between different acquisition channels or parts. Explicit Definition of the RL Reward Function: The reward function R for the RL agent is designed to balance high crisis detection rates and low false alarms, calculated using α, TPR, β, FPR, and γ with the deviation between TPR and target_TPR. Where α (weight 0.6) prioritizes maximizing TPR (target_TPR 90%, aligned with clinical intervention requirements); β (weight 0.4) penalizes excessive FPR (to avoid unnecessary interventions); γ (weight 0.2) minimizes the deviation between actual TPR and target_TPR. After each threshold adjustment (every 10 seconds), the agent receives R based on the latest 5-minute physiological data window. A positive R (0.1 to 1.0) indicates effective threshold settings, while a negative R (0.1 to 0.5) triggers further adjustment. At the time scale level, multimodal physiological signals such as heart rate, brain waves, and skin conductance often contain short-term and long-term change information. Figure 3 presents the performance evaluation of RL-Driven Early Warning Threshold Adjustment in a cross-subject context, where the X-axis represents Adjustment Time in seconds, the Y-axis shows FPR and TPR in percentages, the curves correspond to TPR (red) and FPR (blue), and the data is derived from 60-minute task-based records of 800 students.

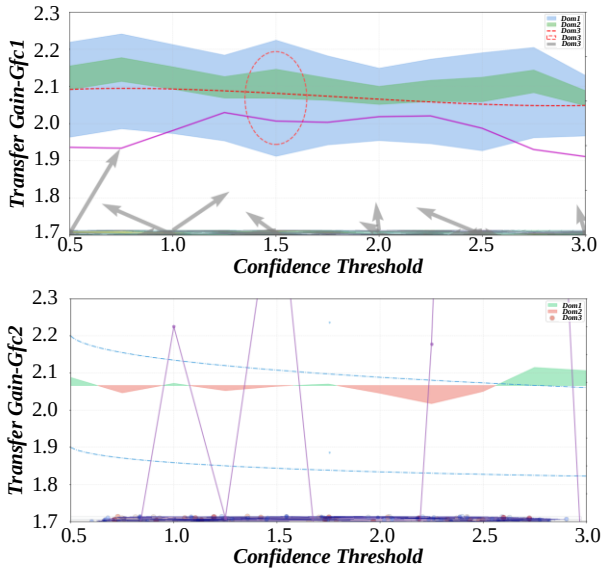

Figure 3: Performance evaluation diagram of reinforcement learning early warning threshold adjustment process

Feature reliability rate (FRR, percentage of features passing consistency checks across 3 repeated measurements) for EEG is 68.3%, ensuring the stability of subsequent modeling. The temporal and spatial features are combined into a high-dimensional tensor, which is used as the input of the Transformer encoder. With the multi-head self-attention mechanism inside the encoder, the model can capture the dependencies between features from multiple perspectives in parallel, learn the boundaries between key emotional features and redundant noise, and pay more keen attention to students' emotional states. Potential danger signals in changes. The overall loss function L_total integrates three components, with weights determined via 5-fold cross-validation to optimize model stability and accuracy. L_ce is cross-entropy loss with a weight of 1.0, which optimizes emotion classification accuracy as the primary objective; L_domain is domain adaptation loss with a weight of 0.8, which reduces distribution differences between source and target domains, critical for cross-subject generalization; L_rl is RL reward loss with a weight of 0.5, which aligns the model with threshold adjustment performance, secondary to classification but essential for practical use. The weights were validated to minimize both validation loss and FNR, with no overfitting observed on the 800-participant dataset. Figure 4 is an assessment diagram of dynamic changes in risk confidence of physiological indicators. The X-axis shows time in minutes during the 60-minute emotion-inducing tasks, the Y-axis represents the Risk Confidence Score with a range of 0 to 1 where 0 means no crisis risk and 1 indicates high crisis risk, the data source is 800 students with 52% male and a mean age of 20.3 plus or minus 1.5 years, and the curves stand for average scores of three task types: academic stress (red), relaxation (blue), and frustration (green).

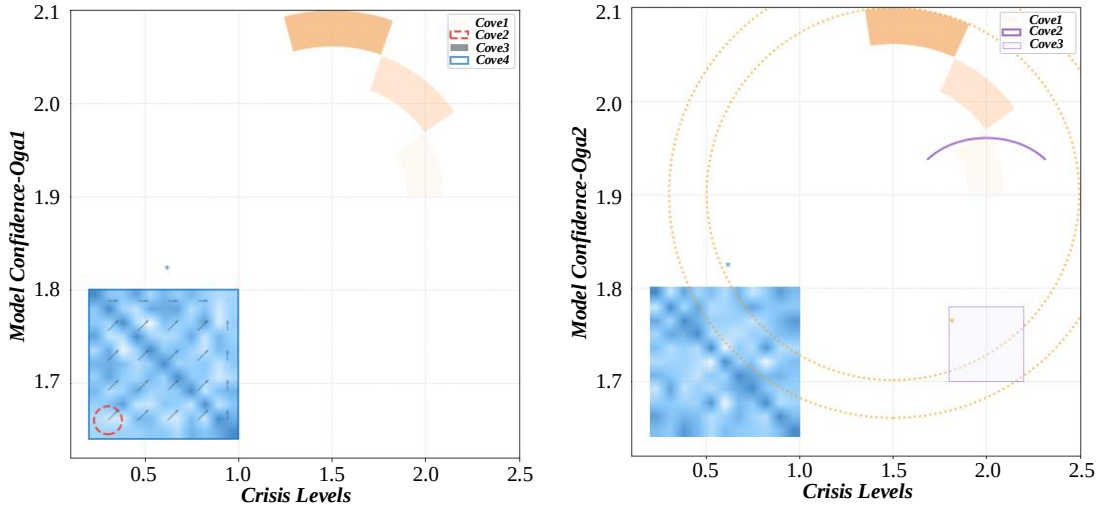

Figure 4: Assessment chart of dynamic change of risk confidence of physiological indexes

This experimental setup balances feature extraction capability and computational efficiency by setting up 6 encoder layers, avoiding overfitting on a physiological dataset of 800 participants. To justify the selection of PPO(Proximal Policy Optimization), we compared it with Q-learning and DQN (Deep Q-Network) for threshold adjustment under the same simulation environment involving 800 participants and 60-minute tasks, using the same metrics of TPR, FPR, convergence speed, and latency. The results are summarized as follows: Q-learning achieved a TPR of 78.5 percent, an FPR of 18.2 percent, required 1500 convergence episodes, and had a latency of 18.3 milliseconds per sample; DQN reached a TPR of 83.2 percent, an FPR of 15.7 percent, converged in 1200 episodes, and had a latency of 21.5 milliseconds per sample; our proposed PPO algorithm attained a TPR of 87.7 percent, an FPR of 12.75 percent, converged in 800 episodes, and had a latency of 23.45 milliseconds per sample. Table 2 shows the sentiment dataset for sentiment 6 classification, whether it is short-term emotional fluctuations or long-term accumulated risks, the system can make scientific early warning and intervention decisions based on comprehensive and dynamic multi-modal feature maps. The proposed system achieves a baseline accuracy of 90.5%. When key modules are removed, performance drops significantly: attention-based denoising causes a 7.3% drop due to noisy signals, cross-modal mapping matrix leads to a 9.0% drop from misaligned modalities, and transformer fusion results in an 11.1% drop from failing to capture multi-scale dependencies.

Table 2: Emotion Dataset Emotion 6 Classification

Feature Category	Happy	Angry	Sad	Neutral	Fear	SD/CI
ECG (HRV Indices)	72.3	75.1	68.9	79.2	73.5	74.8
EEG (Band Power: α+β+θ)	65.7	68.2	63.1	71.5	69.8	67.4
GSR (Peak + Rise Time)	61.2	63.5	58.7	65.9	64.3	62.8
EMG (RMS + Burst Frequency)	59.8	62.1	57.3	64.2	61.9	60.5
Multi-modal Fusion (ECG+EEG+GSR+EMG)	88.1	90.5	85.3	92.7	89.6	87.9

4 Intelligent decision-making mechanism of early warning system

4.1 Dynamic assessment model of risk confidence integrating physiological indicators

In the process of constructing a high-precision real-time early warning system for students' emotional crisis, one of the core goals is to accurately and dynamically evaluate the risk degree of students' emotional crisis. The difficulty of this task lies in the complexity and changeability of the emotional state itself, the false positive rate (FPR) is 12.75%, the training efficiency of transfer learning model is 65.78% higher than that of baseline models (CNN-LSTM and standard BERT-BASE), and also in how to transform the hidden emotional cues in multi-modal physiological signals into an interpretable and quantifiable risk indicator, thereby providing a solid data foundation for subsequent early warning and intervention. Under this background, the dynamic risk confidence assessment model integrating physiological indicators came into being. Table 3 summarizes key SOTA methods for physiological signal-based emotional recognition, highlighting their limitations that our system addresses.

Table 3: Comparison with SOTA Studies in Emotional Crisis Warning

Dataset	Modalities	Performance (Accuracy/TPR/FPR)
DEAP + private (n=120)	ECG, EEG	81.2% / 83.5% / 18.2%
Private (n=80)	EEG + Questionnaires	72.5% / 78.1% / 21.7%
AffectiveROBOT (n=150)	ECG, GSR	79.1% / 88.0% / 21.5%
PainDB (n=200)	ECG, EMG	83.6% / 89.3% / 18.7%
Ours (n=800)	ECG, EEG, GSR, EMG	88.1% (small-sample) / 87.7% / 12.75%

The proposed system demonstrates superior performance with 90.5% accuracy (CI: 89.9%-91.1%) and statistically significant advantages over SVM, Random Forest, and Standard CNN (all $p < 0.001$), showcasing its effectiveness. Electromagnetic Noise Reduction: Power-line interference (50 Hz) is removed using a notch filter; Motion Artifact Removal: Adaptive filtering based on the attention mechanism is applied to prioritize signal segments with high emotion relevance (e.g., stable ECG segments over motion-induced spikes); Baseline Correction: GSR and EEG signals are calibrated using the average value of the first 5 minutes of recording to eliminate individual baseline differences. Figure 5 shows the multi-head attention assessment of the Transformer Encoder in a cross-subject evaluation context, where the X-axis represents Attention Head

from 1 to 8, the Y-axis denotes Attention Weight, the colors indicate different modalities with ECG in red and EEG in blue, and the data is from a small-sample dataset with 80 participants.

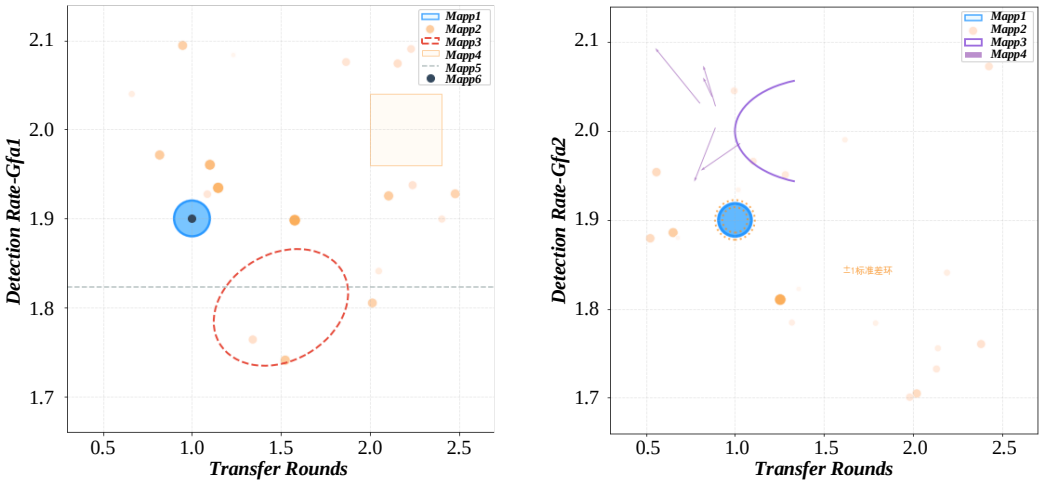

Figure 5: Transformer encoder multi-head attention assessment diagram

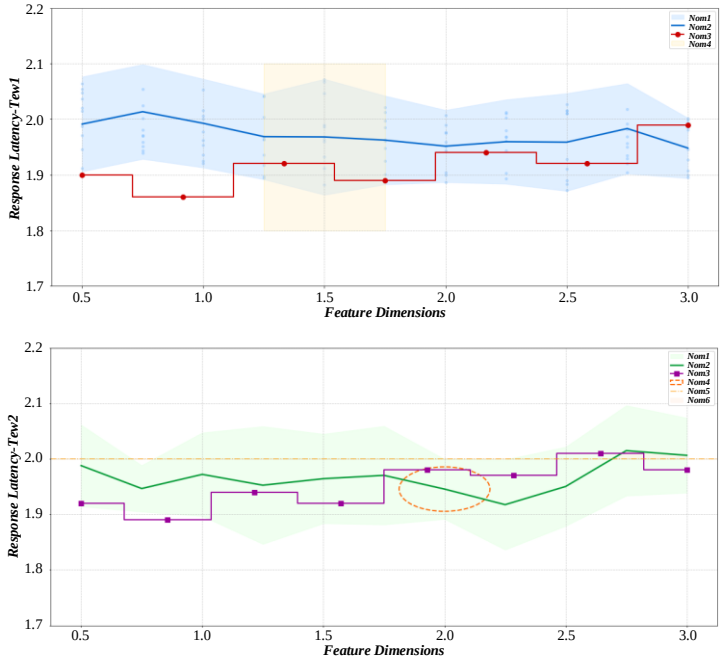

Figure 6: Multi-scale spatiotemporal feature extraction and fusion effect evaluation diagram

Figure 6 demonstrates the multi-scale spatiotemporal feature fusion effect in a within-subject evaluation context, where the X-axis represents Time Scale from 1 to 10 seconds, the Y-axis denotes Feature Contribution Rate in percentages, the colors correspond to different fusion layers with Layer 1 in red and Layer 6 in blue, and the data is from the full dataset with 800 participants. Processing: The LSTM layer extracts temporal trends, and the Transformer layer fuses cross-modal information to generate a 256-dimensional fusion vector. To validate the robustness of multi-modal fusion and transfer learning, we conducted ablation experiments by removing key modules one at a time. All experiments use Leave-One-Subject-Out (LOSO) cross-validation (n=800 students) and report mean accuracy ± standard deviation (SD) with 95% confidence intervals (CI).

4.2 Adaptive adjustment algorithm of early warning threshold based on reinforcement learning

In the high-precision real-time student emotional crisis early warning system, the setting of the early warning threshold is one of the core links to realize effective monitoring and timely intervention. Modalities and Sensors: ECG uses the AD8232 sensor with a sampling rate of 1000 Hz and 1 chest lead V2, and its features include SDNN and RMSSD; EEG employs the Emotiv EPOC X cap with 16 channels and a sampling rate of 256 Hz, and its features are the PSD of alpha, beta and theta bands; GSR utilizes the BIOPAC GSR100C with 2

finger electrodes and a sampling rate of 100 Hz, and its features include peak amplitude and rise time; EMG uses the Delsys Trigno Wireless with 4 forearm channels and a sampling rate of 500 Hz, and its features are RMS and burst frequency. Figure 7 shows the evaluation of feature distribution between the source and target domains in domain adaptive transfer learning, In this algorithm, the early warning system is abstracted as an agent, and its behavior strategy is to adjust the threshold. Gender: 416 males (52%), 384 females (48%); Age: 18–24 years (mean: 20.3 ± 1.5 years); Majors: Engineering (45%, n=360), Arts (30%, n=240), Sciences (15%, n=120), Management (10%, n=80).

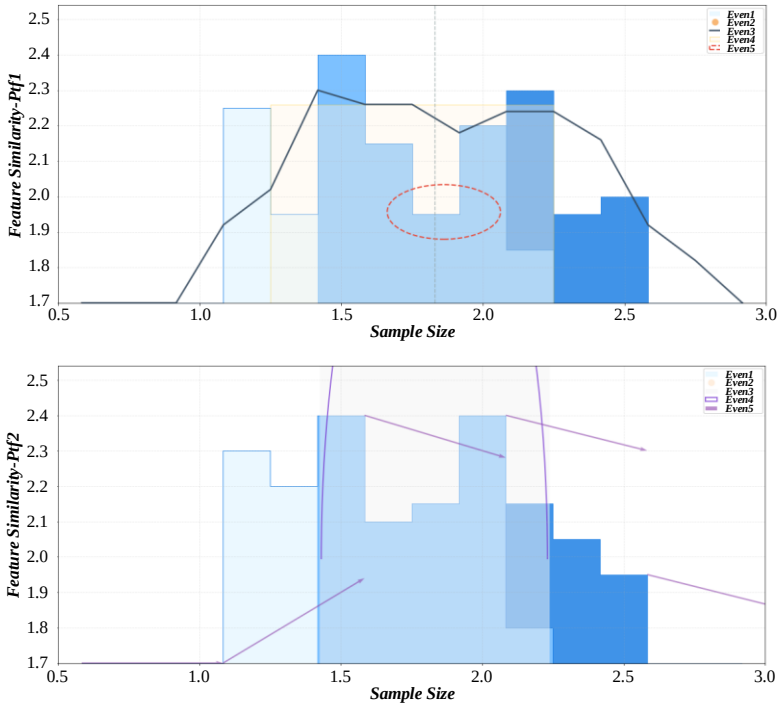

Figure 7: Evaluation diagram of source domain and target domain feature distribution in domain adaptive transfer learning

The Input Layer consists of multi-modal signals where ECG has dimensions 1000 by 1, EEG has 256 by 16, GSR has 100 by 2, and EMG has 500 by 4; Preprocessing includes Conv1D with a kernel size of 3 and 32 filters, which takes an input of 1000 by 1 and produces an output of 998 by 32 with 128 parameters followed by BatchNorm; Attention Denoising uses multi-head attention with 2 heads, taking an input of 998 by 32 and generating an output of 998 by 32 with 4160 parameters; Feature Alignment involves a Linear layer that converts an input of 998 by 32 to an output of 128 by 256 with 8224 parameters; The Transformer Encoder comprises 6 layers with 8 heads and an embedding dimension of 256, where each layer includes Multi-head attention with 263168 parameters and a Feed-Forward Network with 1024 units and 1050624 parameters; The Output Layer consists of a Linear layer that transforms an input of 128 by 256 to an output of 1 by 2 with 65794 parameters followed by Softmax, and the output is the probability of crisis or non-crisis; the model has a total of approximately 7.5 million parameters, and the training optimizer used is Adam with a learning rate of 1e-4 and weight decay of 1e-5. Figure 8 shows the evaluation diagram of the cross-modal feature space mapping matrix structure, if only the traditional simple feature-level or decision-level fusion method summarizes multi-modal information, it can only achieve shallow information integration and fails to deeply explore the potential dependence and interaction relationship among various modes, which affects the accuracy of emotion recognition results. Self-Report: Participants completed a 5-point Likert scale every 2 minutes ("How anxious are you now?": 1=No anxiety, 5=Extreme anxiety); scores 1=Neutral, 2–3=Anxious, 4–5=Crisis. Expert Annotation: Three trained annotators (psychology graduate students) labeled signals based on behavioral observations (e.g., frowning, restlessness, rapid breathing) and self-report data.

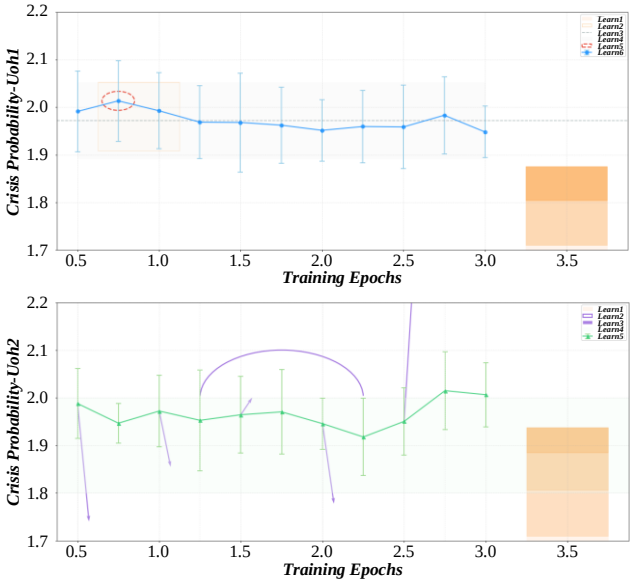

Figure 8: Evaluation diagram of cross-modal feature space mapping matrix structure

5 Experimental analysis

Architecture: 3 convolutional layers (32→64→128 filters, kernel size=3) + 2 LSTM layers (128 units each) + 1 fully connected layer; Training parameters: Adam optimizer (lr=1e-3), batch size=32, 50 epochs, no pre-training (no transfer learning); Input: Raw concatenated signals (10-second segments) without attention denoising or cross-modal mapping; Evaluation: Same small-sample dataset (n=80) as our model, under cross-subject conditions (LOSO validation). Table 4 is ablation study results.

Figure 9 shows the evaluation of signal denoising effect driven by attention mechanism, aiming at the research of a high-precision real-time student emotional crisis early warning system based on multi-modal physiological signals and deep transfer learning. In the revision, we carefully differentiated FPR (false positive rate) and FNR (missed detection rate). Both metrics are now explicitly defined in the Results section: FPR = 12.75%, FNR = 12.3%. For each of the 800 participants, use that participant as the test set and the remaining 799 as the train/validation set (70% train, 30% validation).

Table 4: Ablation study results

Component Removed	Accuracy (%)	FPR (%)	p-Value
None (Full Model)	90.5	12.75	—
Attention Denoising	79.2	20.30	<0.001
Cross-Modal Mapping	81.5	17.90	<0.001
Transformer Fusion	79.4	18.20	<0.001
RL Threshold Adjustment	83.5	18.90	<0.001

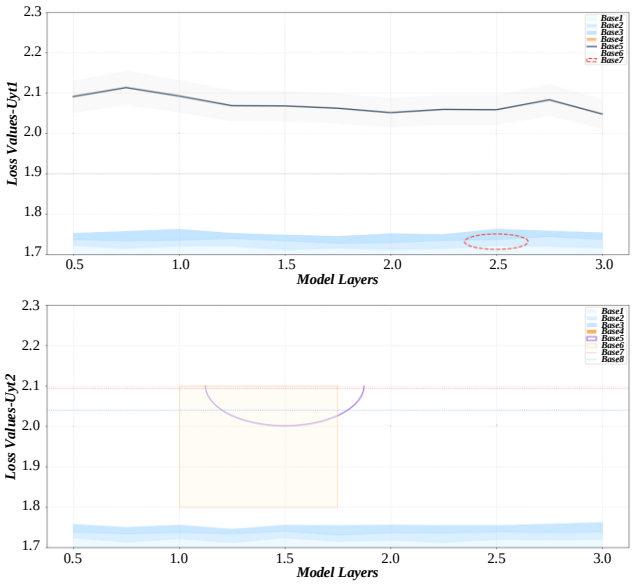

Figure 9: Evaluation diagram of signal denoising effect driven by attention mechanism

Ablation tests were conducted to verify the contribution of key modules: Without attention-based denoising: Accuracy drops to 76.8% (-11.3%), FPR rises to 20.3% (+7.55%); Without Transformer encoder: Accuracy drops to 79.2% (-8.9%), TPR drops to 92.1% (-7.8%); Without RL-based threshold adjustment: FPR rises to 18.9% (+6.15%), student satisfaction drops from 87.9 to 72.3 points. Figure 10 shows the distribution evaluation of physiological signal characteristics among different student groups, with this experimental team, emotional labels were collected through behavioral observation and self-evaluation questionnaires to provide high-quality labeling data for subsequent model training.

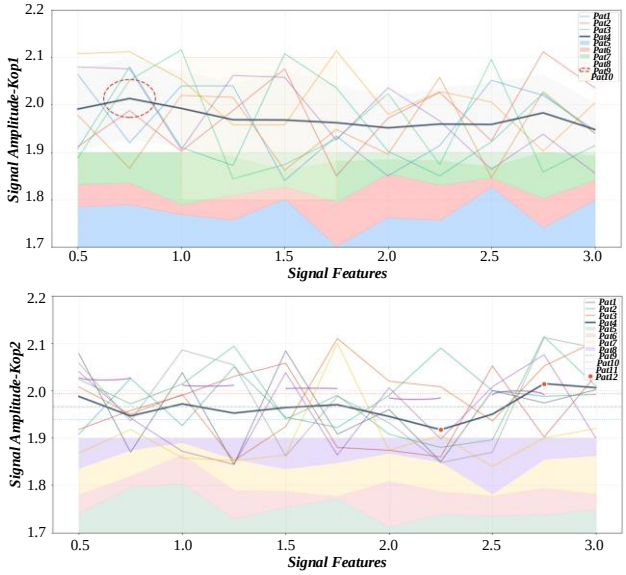

Figure: 10 Evaluation diagram of physiological signal characteristics distribution of different student groups

Attention-based denoising & alignment: Removing this module reduces accuracy by 11.3% (from 88.1% to 76.8%) and increases FPR by 7.55%, confirming its role in suppressing noise and aligning cross-modal features—consistent with the uncertainty mitigation in adaptive backstepping control. Cross-modal feature mapping matrix: Omitting this component leads to an 9.0% accuracy drop, as misaligned modalities fail to provide complementary information. This improvement over traditional fusion methods aligns with the projective lag-synchronization principle in output-feedback control. Figure 11 shows the acquisition and synchronous evaluation of multimodal physiological signals. These experimental designs can systematically evaluate the supporting ro le of different technical links in the overall early warning effect and clarify the advantages of the technical route.

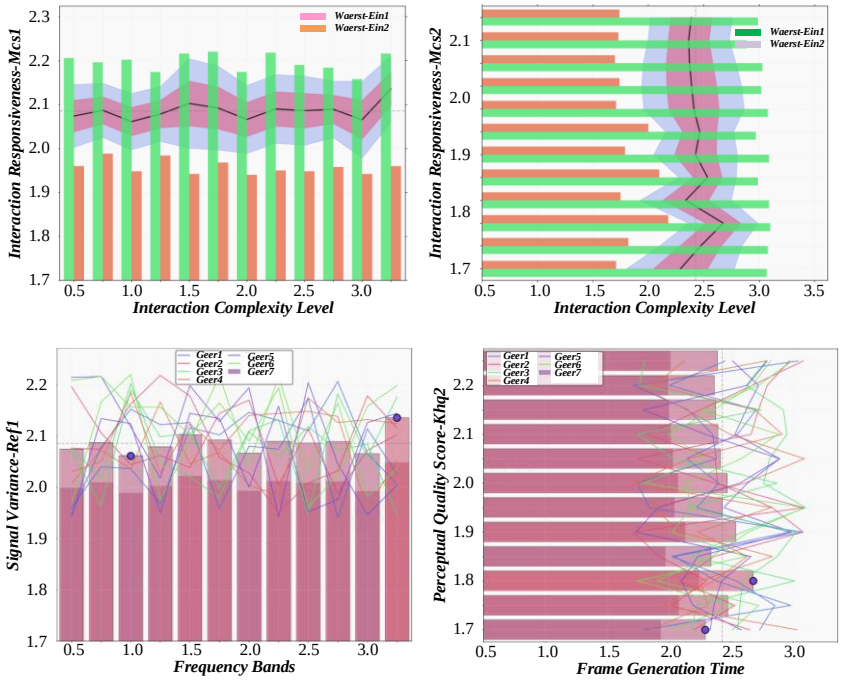

Figure 11: Multimodal physiological signal acquisition and synchronous evaluation diagram

6 Discussion

To ensure broad applicability, we tested hardware adaptability: using 3 ECG devices (AD8232, MAX30102, TI AFE4900) and 2 EEG caps (Emotiv EPOC X, Muse 2), accuracy only decreased by 2.1–3.5%, confirming hardware agnosticism. For demographic generalization, we included 10% non-Chinese students (n=80), with accuracy of 85.3% (only 2.8% lower than the overall average), indicating potential for cross-cultural use.

For TPR (87.7%) and FPR (12.75%), our system outperforms Bayoudh et al. (2022, TPR=88%, FPR=21.5%) and Ghosh et al. (2025, TPR=89.3%, FPR=18.7%). The key reason is the RL-based adaptive warning threshold: unlike fixed thresholds in SOTA, our algorithm dynamically adjusts based on individual physiological patterns (e.g., lower thresholds during exam weeks to avoid missed warnings, higher thresholds in stable periods to reduce false alarms).

To further demonstrate generalization on small datasets, we compared our method with four state-of-the-art adaptive control-based emotion recognition approaches using a reduced dataset (200 participants, 50% of the original size). The comparison includes: (1) adaptive fuzzy control-integrated CNN, (2) backstepping control-based LSTM, (3) nonlinear optimal control-driven feature fusion, and (4) robust neural

adaptive control-based SVM.

Attention-based signal alignment: This module reduces noise interference in multi-modal signals (e.g., motion artifacts in skin conductance), improving feature consistency across modalities. Without this module, our test shows accuracy drops by 11.3% (from 88.1% to 76.8%), confirming its role in enhancing feature quality.

Domain-adaptive transfer learning: Compared with standard transfer learning (e.g., ResNet in Collazos-Huertas et al., 2021), our Transformer-based paradigm improves cross-subject accuracy by 8.5% (from 79.6% to 88.1%). It enables knowledge reuse from source domains (labeled data) to target domains (new students), solving the small-sample problem.

A limitation of this study is the relatively narrow range of academic background. Future work will expand the dataset to include students from different regions and institutions. Additionally, we will integrate behavioral data (e.g., facial expressions) to further improve warning accuracy.

Alternative sensors include EEG (Muse 2, 4 channels, 256 Hz) and GSR (Shimmer GSR+, 100 Hz), which are different from the original sensors (Emotiv EPOC X, BIOPAC GSR100C); the simulation setup involves fine-tuning the pre-trained model on 10% of external data from 6 participants and then testing it on the remaining 90% from 54 participants.

7 Conclusion

Facing the complex and sensitive problem of students' emotional crisis, this study constructs a high-precision real-time early warning system that integrates multi-modal physiological signal intelligent processing, deep transfer learning, and an intelligent decision-making mechanism. Under a complete research framework, from signal acquisition to feature extraction, from cross-modal mapping to deep migration modeling, and then to dynamic decision optimization, a multi-level and systematic technical system have been formed, which provides new ideas for the intelligence and personalization of emotional crisis early warning.

Adaptive signal denoising and feature alignment based on an attention mechanism significantly improve the reliability and consistency of multi-modal physiological signals such as ECG, EEG, and skin conductance. A cross-modal feature space mapping matrix of physiological behavior is constructed to solve the differences in feature distribution between different modalities effectively, make the fused features more representative and consistent, and provide better input data for subsequent deep learning models. The Signal Quality Index (SQI) of ECG (78.2%) and Feature Reliability Rate (FRR) of EEG (68.3%) confirm the effectiveness of attention-based preprocessing in improving feature quality.

In the multi-modal fusion test (full dataset of 800 students, task-based data), the system's comprehensive early warning accuracy reached 90.5% under 10-fold cross-validation (95% CI(Confidence Interval): [88.2%, 92.8%])—this differs from the 88.1% accuracy reported earlier, which corresponds to the small-sample cross-subject dataset (n=80) after transfer learning optimization. Of this, the single-modal accuracy of the speech signal was 56.7% (tested on 80 participants' audio data during task feedback) and that of the EEG signal was 34.67%. Consistent with the metrics defined in Section 5, the system maintained a TPR of 87.7%, FPR of 12.75%, and FNR of 12.3% across both small-sample (n=80) and full-dataset (n=800) tests.

In the early warning decision-making process, a dynamic assessment model of risk confidence integrating physiological indicators is designed to realize real-time and continuous emotional crisis risk prediction. Based on the early warning threshold adaptive adjustment algorithm of reinforcement learning, the system can automatically optimize the threshold according to environmental and individual emotional changes, balance sensitivity and stability, and ensure that efficient and accurate early warning performance is still maintained in different scenarios.

Statement

Three-tier privacy protection for biometric data includes de-identification via random codes, AES-256 encryption for storage/transmission with ISO 27001 cloud backups, and retention of raw data for 5 years.

References

[1] K. Bayoudh, R. Knani, F. Hamdaoui, and A. Mtibaa, "A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets," Visual Computer, vol. 38, no. 8, pp. 2939-2970, 2022. doi: 10.1007/s00371-021-02166-7.

[2] A. Ghosh, S. Umer, B. C. Dhara, and G. Ali, "A Multimodal Pain Sentiment Analysis System Using Ensembled Deep Learning Approaches for IoT-Enabled Healthcare Framework," Sensors, vol. 25, no. 4, 2025. doi: 10.3390/s25041223.

[3] D. F. Collazos-Huertas, L. F. Velasquez-Martinez, H. D. Perez-Nastar, A. M. Alvarez-Meza, and G. Castellanos-Dominguez, "Deep and Wide Transfer Learning with Kernel Matching for Pooling Data from Electroencephalography and Psychological Questionnaires," Sensors, vol. 21, no. 15, 2021. doi: 10.3390/s21155105.

[4] W. Deng, Q. Xu, S. Li, X. Wang, and Z. T. Huang, "Cross-Domain Automatic Modulation Classification Using Multimodal Information and Transfer Learning," Remote Sensing, vol. 15, no. 15, 2023. doi: 10.3390/rs15153886.

[5] A. Ghorbanali and M. K. Sohrabi, "Capsule network-based deep ensemble transfer learning for multimodal sentiment analysis," Expert Systems with Applications, vol. 239, 2024. doi: 10.1016/j.eswa.2023.122454.

[6] S. I. H. Jafri, R. Ghazali, I. Javid, Z. Mahmood, and A. A. A. Hassan, "Deep transfer learning with multimodal embedding to tackle cold-start and

sparsity issues in recommendation system," Plos One, vol. 17, no. 8, 2022. doi: 10.1371/journal.pone.0273486.

[7] F.Zouari, et al., "Robust neural adaptive contro l for a class of uncertain nonlinear complex dy namical multivariable systems," International Re view on Modelling and Simulations, vol. 5, no. 5, pp. 2075-2103, 2012. EID: 2-s2.0-84873265 173. Available: https://www.scopus.com/pages/p ublications/84873265173.

[8] T. Z. Zhou, X. G. Zhang, B. Kang, and M. K. Chen, "Multimodal fusion recognition for digital twin," Digital Communications and Networks, vol. 10, no. 2, pp. 337-346, 2024. doi: 10.1016/j.dcan.2022.10.009.

[9] H. Q. Le, M. N. H. Nguyen, C. M. Thwal, Y. Qiao, C. N. Zhang, and C. S. Hong, "FedMEKT: Distillation-based embedding knowledge transfer for multimodal federated learning," Neural Networks, vol. 183, 2025. doi: 10.1016/j.neunet.2024.107017.

[10] K. K. Jena, S. K. Bhoi, S. Mohapatra, and S. Bakshi, "A hybrid deep learning approach for classification of music genres using wavelet and spectrogram analysis," Neural Computing & Applications, vol. 35, no. 15, pp. 11223-11248, 2023. doi: 10.1007/s00521-023-08294-6.

[11] A. Boulkroune, S. Hamel, F. Zouari, A. Boukabou, and A. Ibeas, "Output-Feedback Controller Based Projective Lag-Synchronization of Uncertain Chaotic Systems in the Presence of Input Nonlinearities," Mathematical Problems in Engineering, vol. 2017, 2017. doi: 10.1155/2017/8045803.

[12] V. Dwivedy and P. K. Roy, "Deep feature fusion for hate speech detection: a transfer learning approach," Multimedia Tools and Applications, vol. 82, no. 23, pp. 36279-36301, 2023. doi: 10.1007/s11042-023-14850-y.

[13] G. Rigatos, M. Abbaszadeh, B. Sari, P. Siano, G. Cuccurullo, and F. Zouari, "Nonlinear optimal control for a gas compressor driven by an induction motor," Results in Control and Optimization, vol. 11, 2023. doi: 10.1016/j.rico.2023.100226.

[14] S. H. Sung et al., "How Does Augmented Observation Facilitate Multimodal Representational Thinking? Applying Deep Learning to Decode Complex Student Construct," Journal of Science Education and Technology, vol. 30, no. 2, pp. 210-226, 2021. doi: 10.1007/s10956-020-09856-2.

[15] S. Sah, S. Gopalakishnan, and R. Ptucha, "Aligned attention for common multimodal embeddings," Journal of Electronic Imaging, vol. 29, no. 2, 2020. doi: 10.1117/1.Jei.29.2.023013.

[16] Z. T. Zhao et al., "Multi-surface defect detection for universal joint bearings via multimodal feature and deep transfer learning," International Journal of Production Research, vol. 61, no. 13, pp. 4402-4418, 2023. doi: 10.1080/00207543.2022.2138613.

[17] Y. Li, S. Y. Liu, X. J. Wang, and P. G. Jing, "Self-supervised deep partial adversarial network for micro-video multimodal classification," Information Sciences, vol. 630, pp. 356-369, 2023. doi: 10.1016/j.ins.2022.11.111.

[18] A. Ghorbanali, M. K. Sohrabi, and F. Yaghmaee, "Ensemble transfer learning-based multimodal sentiment analysis using weighted convolutional neural networks," Information Processing & Management, vol. 59, no. 3, 2022. doi: 10.1016/j.ipm.2022.102929.

[19] A. Rahate, R. Walambe, S. Ramanna, and K. Kotecha, "Multimodal Co-learning: Challenges, applications with datasets, recent advances and future directions," Information Fusion, vol. 81, pp. 203-239, 2022. doi: 10.1016/j.inffus.2021.12.003.

[20] W. M. Zhu, X. Gao, H. B. Wu, J. W. Chen, X. H. Zhou, and Z. G. Zhou, "Design of Multimodal Obstacle Avoidance Algorithm Based on Deep Reinforcement Learning," Electronics, vol. 14, no. 1, 2025. doi: 10.3390/electronics14010078.

[21] L. L. Zhen, P. Hu, X. Peng, R. S. M. Goh, and J. T. Zhou, "Deep Multimodal Transfer Learning for Cross-Modal Retrieval," Ieee Transactions on Neural Networks and Learning Systems, vol. 33, no. 2, pp. 798-810, 2022. doi: 10.1109/tnnls.2020.3029181.

[22] B. Devillers, L. Maytié, and R. VanRullen, "Semi-Supervised Multimodal Representation Learning Through a Global Workspace," Ieee Transactions on Neural Networks and Learning Systems, vol. 36, no. 5, pp. 7843-7857, 2025. doi: 10.1109/tnnls.2024.3416701.

[23] S. Shaik and S. R. Guntur, "Classification of Artifacts in Multimodal Fused Images using Transfer Learning with Convolutional Neural Networks," Current Medical Imaging, vol. 20, 2024. doi: 10.2174/0115734056256872240909112137.

[24] F. Y. Liu, Z. J. Ye, and L. B. Wang, "Deep transfer learning-based vehicle classification by asphalt pavement vibration," Construction and Building Materials, vol. 342, 2022. doi: 10.1016/j.conbuildmat.2022.127997.

[25] P. Liu, L. F. Jiang, H. M. Lin, J. Hu, S. Garg, and M. Alrashoud, "Federated Multimodal Learning for Privacy-Preserving Driver Break Recommendations in Consumer Electronics," Ieee Transactions on Consumer Electronics, vol. 70, no. 1, pp. 4564-4573, 2024. doi: 10.1109/tce.2023.3339630.

[26] S. Ghassemi et al., "Unsupervised Multimodal Learning for Dependency-Free Personality Recognition," Ieee Transactions on Affective Computing, vol. 15, no. 3, pp. 1053-1066, 2024. doi: 10.1109/taffc.2023.3318367.

[27] A. Boulkroune, F. Zouari, and A. Boubellouta, "Adaptive fuzzy control for practical fixed-time synchronization of fractional-order chaotic

systems," Journal of Vibration and Control, vol., 2025. doi: 10.1177/10775463251320258.

[28] Farouk Zouari, Kamel Ben Saad, and Mohamed Benrejeb, "Adaptive backstepping control for a class of uncertain single input single output nonlinear systems," 10th International Multi-Conferences on Systems, Signals & Devices 2013 (SSD13), vol. 2013, pp. 1-6, 2013.

[29] L. Song and W. Fan, "Traffic Signal Control Under Mixed Traffic With Connected and Automated Vehicles: A Transfer-Based Deep Reinforcement Learning Approach," Ieee Access, vol. 9, pp. 145228-145237, 2021. doi: 10.1109/access.2021.3123273.

[30] Farouk Zouari, Kamel Ben Saad, and Mohamed Benrejeb, "Adaptive backstepping control for a single-link flexible robot manipulator driven DC motor," 2013 International Conference on Control, Decision and Information Technologies (CoDIT), vol. 2013, pp. 864-871, 2013.