ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 1 - 2026

Research on emotion Recognition and Personalized music therapy System for the Elderly Based on deep learning

Yuan Fang[1] *and Lin Li*[2,*]

[1] School of Special Education, Changchun University, Changchun 130022, Jilin, China
[2] School of Music, Changchun University, Changchun 130022, Jilin, China

SUMMARY: *Aiming at the problems of weak emotional expression range, large modal differences and lack of dynamic adaptation of music intervention in the elderly, this paper proposes a deep learning based emotion recognition and personalized music therapy system for the elderly. This method fuses speech, face, physiological and interaction information, constructs a convolutional coding, bidirectional temporal modeling and cross-modal attention coordination framework, and realizes the stable discrimination of emotional states in the elderly. On this basis, a recommendation mechanism combining user preference, historical feedback and music content characteristics is introduced to form a closed loop of "identification-recommendation-update". The experimental results show that the Accuracy, Macro-F1 and AUC of the proposed model reach 93.84%, 92.47%and 95.12%respectively. After 4 weeks of intervention, the emotional improvement rate of the experimental group increases to 27.6%, and the average inference delay of the system is 23.6 ms. The research shows that this method has good application potential in intelligent elderly care and digital music therapy scenarios.*

KEYWORDS: *emotion recognition in the elderly; Deep learning; Personalized music therapy; Multimodal fusion*

1 Introduction

Under the dual background of population aging and the rapid development of intelligent health services, the continuous recognition and individualized intervention of emotional states of the elderly are gradually becoming an important direction in the research of computer-assisted care. Different from the emotional computing in general scenarios, the emotional changes in the elderly are often accompanied by the characteristics of voice aging, weakened facial expression amplitude, slow fluctuation of physiological signals, and complex life situations, which makes it difficult for the recognition method of a single mode and a single moment to stably depict the real mental state [1]. At the same time, depression, anxiety, loneliness and sleep disorders have a high correlation in the elderly population. If it still relies mainly on manual interview, scale evaluation or caregiver experience judgment, it not only has the problem of time lag, but also is difficult to support remote monitoring and dynamic intervention [2]. Therefore, it has clear technical value and practical significance to introduce deep learning into emotion recognition of the elderly and further couple it with personalized music therapy system.

Around elderly emotion recognition, existing research has begun to shift from feature engineering to deep representation learning. Jothimani et al. constructed a two-stage hybrid feature fusion network and combined it with Monte-Carlo dropout to improve the robustness

*lilinnuoya@163.com
https://doi.org/10.65102/is2026257 Published: 30 April, 2026

of emotional health recognition for the elderly, indicating that multi-layer feature fusion has a positive effect on complex emotion discrimination [3]. Grossi et al. released a multi-source speech emotion dataset for the elderly in Italy, which provided basic data support for the modeling research in the elderly speech scene [4]. Sreevidya et al. realized emotion classification of the elderly through intermediate layer fusion and cross-modal transfer learning, and verified that multimodal joint modeling was superior to isolated modal analysis [5]. Torcate et al. carried out exploratory research based on convolutional neural networks, indicating that deep models can provide a computable basis for individualized treatment support for the elderly [6]. In the broader research on emotion computing, context-aware modeling for digital biomarkers of stress in the elderly [7], the application of embodied emotional intelligence in companion robots [8], and a systematic review of emotion recognition techniques [9] all show that emotion sensing systems are moving from "offline recognition" to "online service".

At the same time, the development of multimodal deep learning provides a new technical path for elderly emotion recognition. Research on EEG-based emotion recognition points out that deep networks have obvious advantages in temporal dependence modeling and implicit emotion representation mining [10]. Relevant reviews further show that the collaborative modeling of multi-source information such as speech, face, text and EEG can effectively alleviate the problems of large noise, more missing and insufficient generalization of single-modal signals [11]. Pan et al. proposed a joint recognition framework of facial expression, speech and EEG, and proved that cross-modal fusion could improve the stability of emotion discrimination [12]. Islam et al. improve the performance of multimodal emotion recognition in medical analysis scenarios from the perspective of model-level fusion [13]. In the field of musical emotion computing, researchers have begun to pay attention to music-induced emotion recognition, EEG response analysis and deep learning modeling mechanism [14, 15]. However, the existing work focuses more on "whether the recognition is accurate", and the discussion on "how the recognition results can be transformed into executable and updatable music healing strategies" is still insufficient.

Music healing research suggests the feasibility of this direction from the application level. Systematic review and meta-analysis have shown that music therapy can improve depression, cognitive status and neuropsychiatric symptoms in the elderly to a certain extent [16, 17]. Recall music intervention, music listening program in nursing institutions, and remote personalized music intervention also show good adaptation potential [18-20]. The study of De Nys et al. on digital music and action intervention further suggested that digital platform could become an important carrier for the promotion of elderly well-being [21]. However, most of these studies are carried out from the perspective of clinical or nursing implementation, and the generation, recommendation and dynamic adjustment mechanism of healing content are still rarely connected with the deep learning emotion recognition model to form a closed-loop connection.

Table 1: Overview of related research and positioning of this paper

Study	Research Object/Scenario	Main Method	Main Conclusion	Implications for This Study
Jothimani et al. (2023)	Elderly emotional health recognition	Two-stage hybrid feature fusion network	Feature fusion can improve recognition robustness	Multi-level feature collaboration needs to be strengthened
Grossi et al. (2023)	Elderly speech emotion data	Construction of a multi-source speech dataset	Provides a data foundation for elderly speech emotion recognition	Highlights the importance of data specificity in elderly scenarios
Sreevidya et al. (2022)	Elderly multimodal emotion classification	Intermediate-layer fusion and cross-modal transfer learning	Multimodal methods outperform unimodal methods	Supports cross-modal joint modeling
Torcate et al. (2024)	Personalized treatment support	CNN-based emotion recognition	Deep models can support therapeutic assistance	Recognition results can be connected to intervention stages
Pan et al. (2023)	Face–speech–EEG emotion recognition	Multimodal joint framework	Cross-modal fusion improves stability	Can be used to build a unified recognition network
Islam et al. (2024)	Emotion recognition in medical analysis	Model-level fusion method	Model fusion improves overall performance	Suitable for complex elderly healthcare scenarios
Wang et al. (2023) / Hamzah et al. (2024)	EEG emotion recognition	Deep temporal modeling	Helps mine latent emotional patterns	Can provide reference for temporal feature learning
Wang et al. (2023) to De Nys et al. (2024)	Elderly music therapy and digital intervention	Music therapy, remote intervention, and digital platforms	Positive effects on emotion, cognition, and well-being	Recognition and therapy systems need to be coupled in a closed loop
This study	Elderly emotion recognition and personalized music therapy	Deep learning-based recognition + recommendation optimization + feedback updating	Constructs an integrated recognition–intervention system	A closed-loop computational framework for intelligent elderly care

Table 1 summarizes the related studies. It can be seen that although the existing achievements have made progress in the elderly emotion recognition, multi-modal fusion and music intervention, there are still several common shortcomings. First, the emotion modeling

for the elderly group often remains in single-stage recognition, lacking a unified computing link from the original signal to the intervention decision. Second, the temporal correlation and context dependence between cross-modal features have not been fully utilized. Thirdly, the music therapy system mostly relies on static rule recommendation, which is difficult to realize adaptive update according to emotional fluctuations. Based on this, this paper focuses on two questions: can deep learning models achieve more stable emotion recognition on elderly multimodal data? Can the recognition results further drive the personalized music therapy system to complete dynamic recommendation and feedback optimization?

This paper aims to construct an integrated "recognition-recommender feedback" framework for emotional intervention scenarios for the elderly. In the front-end, the emotional state discrimination is realized through multi-modal feature extraction and deep network fusion. In the back-end, the emotion tags, preference information and historical intervention responses are jointly encoded to generate a personalized music therapy plan, and the recommendation weights are continuously modified according to the feedback results. The goal of this paper is not only to improve the recognition accuracy, but also to establish a set of technical paths that balance computational performance, intervention effectiveness and system deployability, so as to provide a scalable method foundation for intelligent elderly care and digital healing services.

2 Methods and materials

2.1 Feature extraction and preprocessing framework of elderly emotional data

The quality of emotion recognition for the elderly depends to a large extent on whether the input data is preprocessed stably, meticulously and with scene adaptability. Different from ordinary emotional computing tasks, the speech signal of the elderly is often accompanied by decreased sound intensity, increased pause and decreased vocalization rate, and the change of facial expression is small. The physiological rhythm is easily affected by individual basic state and environmental disturbance. If the original multimodal data is directly fed into the network, the model often learns noise, missing segments and cross-device offset together, which weakens the discriminative ability of the emotion representation. Based on this, this paper constructs a geriatric emotional data processing framework consisting of quality screening, modal preprocessing, time alignment, feature extraction and unified coding, and its process is shown in Figure 1.

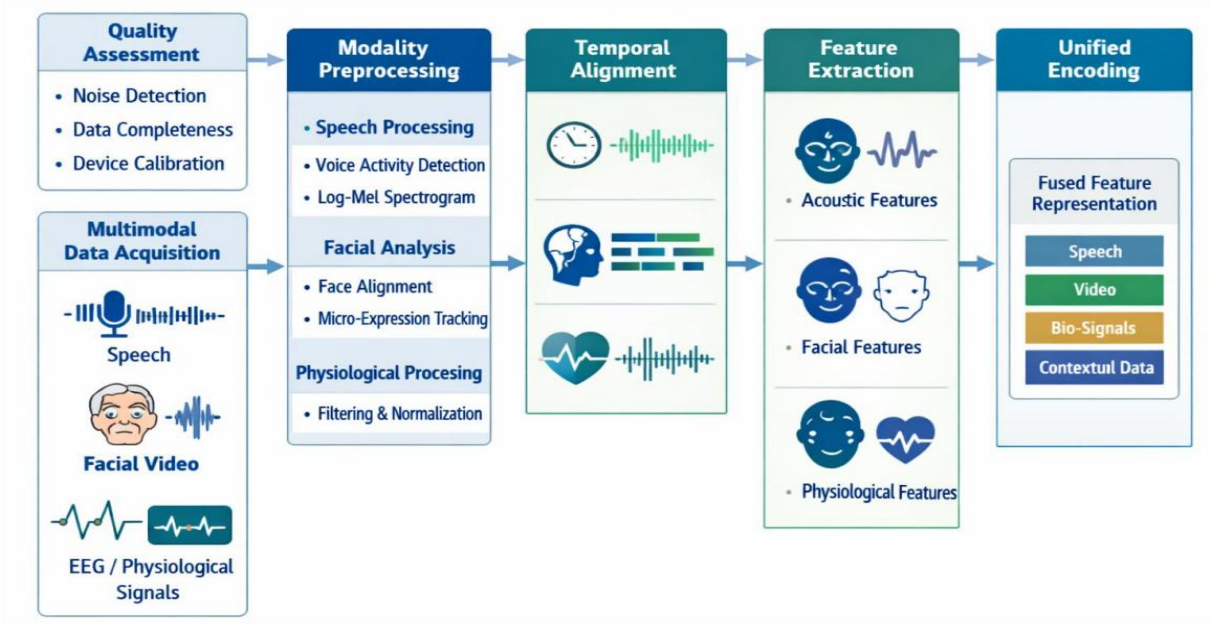

Figure 1: Framework for feature extraction and preprocessing of elderly emotional data

In the speech channel, the collected spoken speech, short sentence response and natural communication segments are subjected to endpoint detection, frame windowing and spectrum mapping, and the log-Mel spectrum is used as the main acoustic representation. It is calculated as follows.

$$S_m(t)=\log\left(\sum_{k=1}^{K}|X_t(k)|^2 H_m(k)+\varepsilon\right) \tag{1}$$

where, $X_t(k)$ represents the frequency domain amplitude of the TTH frame speech after short-time Fourier transform, $H_m(k)$ is the MTH Mel filter response, and ε is the smoothing term to prevent instability of the logarithmic calculation. This processing can compress the dynamic range of spectral domain while preserving emotion-related prosodic information, which is convenient for subsequent convolutional networks to extract local time-frequency patterns.

In the visual channel, considering the weak amplitude of facial muscle activity and the short duration of some expression changes in the elderly, we do not directly use the original image sequence. Instead, we first complete face localization, key point tracking and pose correction, and then extract the micro-expression displacement and local texture change features. In order to describe the intensity of facial motion at adjacent moments, the key point difference component is defined as follows.

$$D_t=\frac{1}{P}\sum_{i=1}^{P}\left\|p_{t,i}-p_{t-1,i}\right\|_2 \tag{2}$$

where $p_{t,i}$ is the coordinate of the ith facial keypoint at time t, P is the total number of keypoints, and D_t reflects the overall motion amplitude between the current frame and the previous frame. This index can assist the model to identify low-intensity but continuous changes in emotional expression.

For physiological signals such as EEG and heart rate, bandpass filtering and artifact removal

are used to remove EMG interference and baseline drift, and normalization processing is completed at the segment level to weaken the amplitude offset between different individuals. The normalized expression is:

$$\tilde{x}_t^{(m)} = \frac{x_t^{(m)} - \mu^{(m)}}{\sigma^{(m)} + \delta} \tag{3}$$

where, $x_t^{(m)}$ is the original observation value of mode m at time t, $\mu^{(m)}$ and $\sigma^{(m)}$ represent the sample mean and standard deviation of this mode, and δ is the smoothing constant. After this step, signals from different sources can be mapped to a relatively consistent numerical scale, which improves the stability of cross-modal joint modeling.

Since the sampling frequency of each modality is significantly different, we further use a unified time step Δt to resample and align the timestamps of speech, video, physiological and interaction logs, and construct a unified input tensor:

$$F_t = [S_t; V_t; B_t; C_t] \tag{4}$$

where, S_t, V_t, B_t and C_t represent speech features, visual features, physiological features and situational interaction features at time t, respectively. For the missing segments, linear interpolation and mask coding are jointly used to make the model retain the state information of "incomplete data" while learning the emotional patterns, and avoid the diffusion of false patterns caused by simple filling.

To facilitate subsequent network training, the original signals of each modality, core processing steps and output features are summarized in Table 2 in this paper. Overall, the framework not only serves to improve the recognition accuracy of the front-end, but also undertakes the task of providing stable emotional input for subsequent personalized music therapy recommendation. Only when the multimodal emotion data have been structured and sorted before entering the deep network, the recognition results have enough credibility to support the dynamic generation and optimization of subsequent intervention strategies.

Table 2: Preprocessing and feature composition of geriatric emotion multimodal data

Modality Source	Raw Data Format	Preprocessing Steps	Output Features
Speech signal	Reading speech, response speech, and natural conversation segments	Noise reduction, endpoint detection, framing and windowing, STFT, and Mel mapping	Log-Mel spectrograms, energy, fundamental frequency, and speech rate
Facial video	Frontal video sequences and local facial expression images	Face detection, keypoint tracking, pose correction, and inter-frame differencing	Expression texture features, keypoint displacement, and motion intensity
Physiological signals	EEG, heart rate, skin conductance, etc.	Filtering, artifact removal, segment splitting, and normalization	Band energy, rhythm statistics, and fluctuation features
Interaction logs	Clicks, dwell time, and music skip records	Timestamp alignment, outlier cleaning, and missing-value repair	Behavior frequency, response latency, and preference vectors
Unified encoding result	Multimodal heterogeneous features	Resampling, time alignment, tensor concatenation, and mask encoding	Unified input matrix for deep models

2.2 Construction of elderly emotion recognition model based on deep learning

After the feature extraction and unified coding of multi-modal emotional data for the elderly, the key to model construction is no longer whether a single feature is visible, but how to depict the collaborative changes between speech, face and physiological signals on a continuous time axis. Emotional states in the elderly usually have the characteristics of "small expression range, long duration and cross-modal asynchronization". It is easy to misjudge fatigue, pause or physiological slow change as negative emotions depending only on the static characteristics at a certain time. Based on this, this paper constructs a deep learning recognition model consisting of spatial representation extraction, temporal dependence modeling and cross-modal attention fusion, so that the preprocessed speech spectrogram, facial sequence features and physiological statistical vectors can be jointly learned in a unified framework. Its spatial feature extraction part is shown in Figure 2.

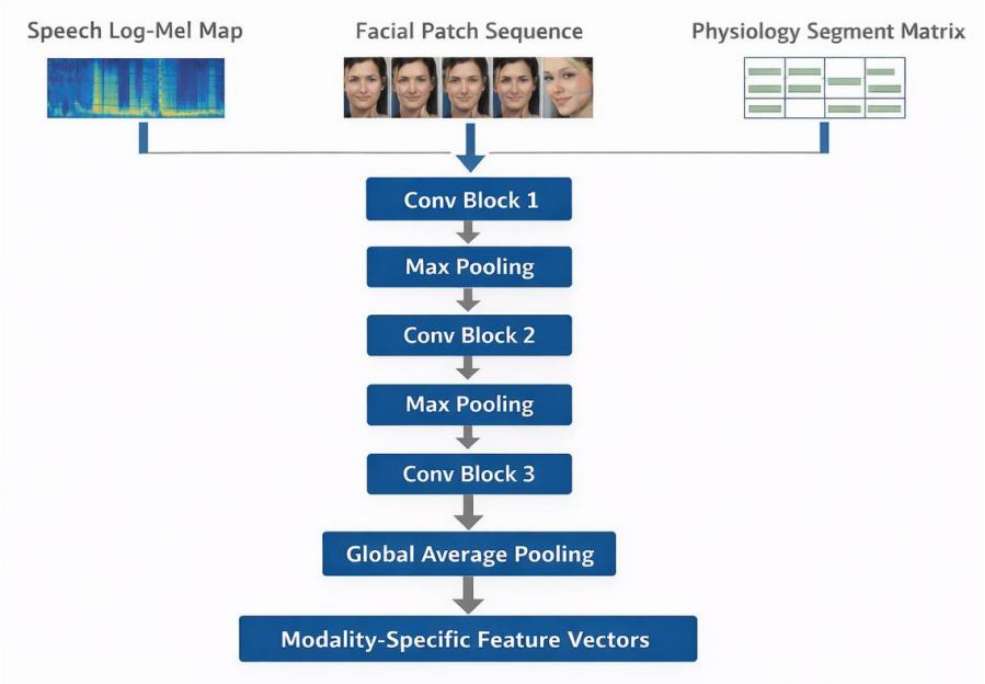

Figure 2: Spatial feature extraction module of the elderly emotion recognition model

In Figure 2, the preprocessed inputs of each modality first enter the convolutional feature extraction module. The convolutional layer extracts emotion sensitive patterns layer by layer with local receptive fields. The shallow layer focuses on preserving low-level information such as speech frequency band changes, eyebrows, eyes, and mouth texture and physiological local fluctuations, and the deep layer further combines these local responses into a discriminative higher-order emotion representation. Its convolution calculation process can be expressed as follows.

$$Y_{i,j}^{(k)}=\sigma\left(\sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{c=1}^{C}W_{m,n,c}^{(k)}X_{i+m,j+n,c}+b^{(k)}\right) \tag{5}$$

where X is the input feature map, W is the convolution kernel parameter, $b^{(k)}$ is the bias term of the KTH channel, $\sigma(\cdot)$ represents the nonlinear activation function, and $Y_{i,j}^{(k)}$ is the response

value of the output feature map at the corresponding position. The proposed structure is able to compress redundant noise under parameter sharing conditions and highlight local spatial patterns related to emotions. After the convolutional layer, the Max pooling layer was connected to reduce the dimension and retain the strong response area. At the end, the global average pooling is used to map the 3D feature map into a fixed-length vector, which provides a compact input for subsequent time series modeling.

Spatial features alone are not enough to support emotion discrimination in the elderly. The reason is that the emotional changes of elderly individuals are often reflected in the evolutionary relationship of several consecutive segments, rather than a single frame of strong stimulus response. To this end, this paper introduces a bidirectional long short-term memory network after convolutional coding to model the intra-modal temporal dependence, and then adds a cross-modal attention mechanism to realize the adaptive focus of key emotion segments. The overall process is shown in Figure 3.

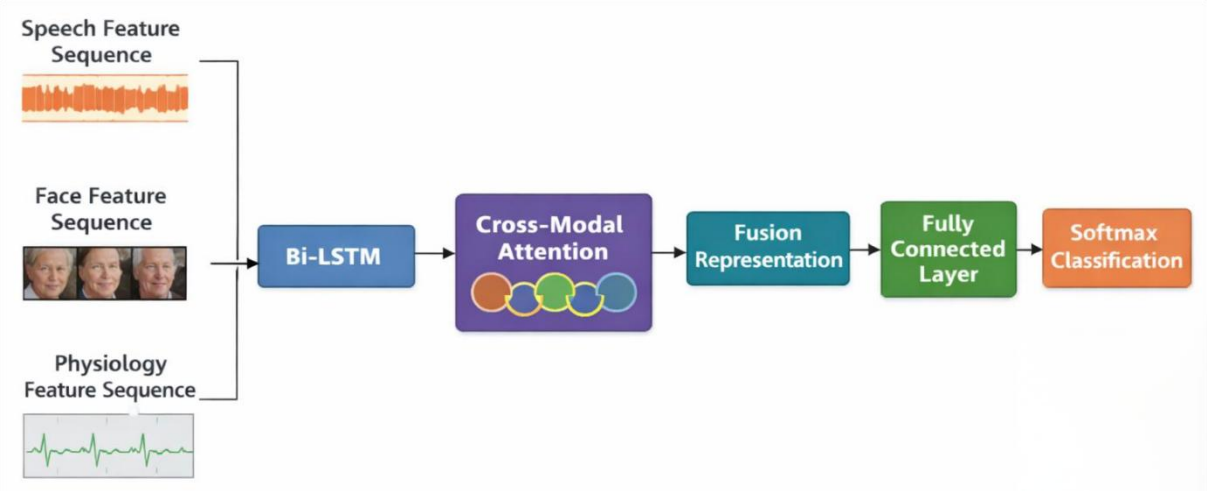

Figure 3: Process of temporal dependence modeling and cross-modal attention fusion

In Figure 3, each modal feature sequence is fed into the Bi-LSTM layer, respectively. The forward link learns the cumulative information of emotion over time, and the reverse link compensates the explanatory effect of the subsequent state on the preceding paragraph, so as to preserve the context dependence more completely. The hidden state update can be written as follows:

$$h_t=\left[\overrightarrow{h_t};\overleftarrow{h_t}\right] \tag{6}$$

where, $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$ represent the forward and backward hidden states at time t, respectively, and the concatenated h_t constitutes the bidirectional temporal representation at that time. Considering that the contribution of different modalities is not constant in the same emotional state, this paper further uses the attention mechanism to calculate the importance weight of each modality and each time segment, and forms a unified fusion vector:

$$\alpha_t^{(m)}=\frac{\exp\left(q^\top h_t^{(m)}\right)}{\sum_{m=1}^{M}\sum_{t=1}^{T}\exp\left(q^\top h_t^{(m)}\right)},\quad F=\sum_{m=1}^{M}\sum_{t=1}^{T}\alpha_t^{(m)}\,h_t^{(m)} \tag{7}$$

where, $h_t^{(m)}$ represents the temporal feature of the MTH modality at time t, q is the learnable

query vector, $\alpha_t^{(m)}$ is the corresponding attention weight, and F is the final fusion representation. This mechanism can automatically increase the proportion of key speech segments, micro-expression changes or abnormal physiological fluctuations in the overall discrimination, and reduce the interference of invalid frames on the results.

At the output end, the fusion vector is input into the fully connected layer and Softmax classifier to complete the discrimination of positive, calm, anxiety tendency and depression tendency. Compared with the single-modal convolutional model, this structure is not a simple superposition module, but a hierarchical computing link is established around the actual characteristics of the elderly's emotions: the convolutional network is responsible for extracting local spatial patterns, the Bi-LSTM is responsible for maintaining time continuity, and the cross-modal attention is responsible for screening key information with diagnostic value. The obtained emotion recognition results will be used as the core input of the subsequent personalized music therapy system for track matching, recommendation update and intervention feedback optimization.

2.3 Architecture design and recommendation optimization mechanism of personalized music therapy system

The aforementioned deep learning emotion recognition model has been able to complete the preliminary discrimination of multimodal emotional states in the elderly. However, if the recognition results are directly mapped to a fixed track list, there are still two obvious shortcomings: One is that the emotion label is instantaneous, and the same subject may have anxiety relief, attention drop and emotional rebound alternately in a short time. Static recommendation is difficult to reflect such continuous changes. The other is that older users' acceptance of music is not only determined by current mood, but also influenced by age preference, rhythm tolerance, familiarity, listening period and previous intervention feedback. In other words, emotion recognition solves "what state is in at the moment", and the personalized music therapy system must also answer "what kind of music is more likely to have a positive intervention effect in this state". Based on this, this paper further constructs a personalized music therapy system based on the emotion recognition results, and introduces the context attention and feedback update mechanism to dynamically optimize the recommendation process. Its overall architecture is shown in Figure 4.

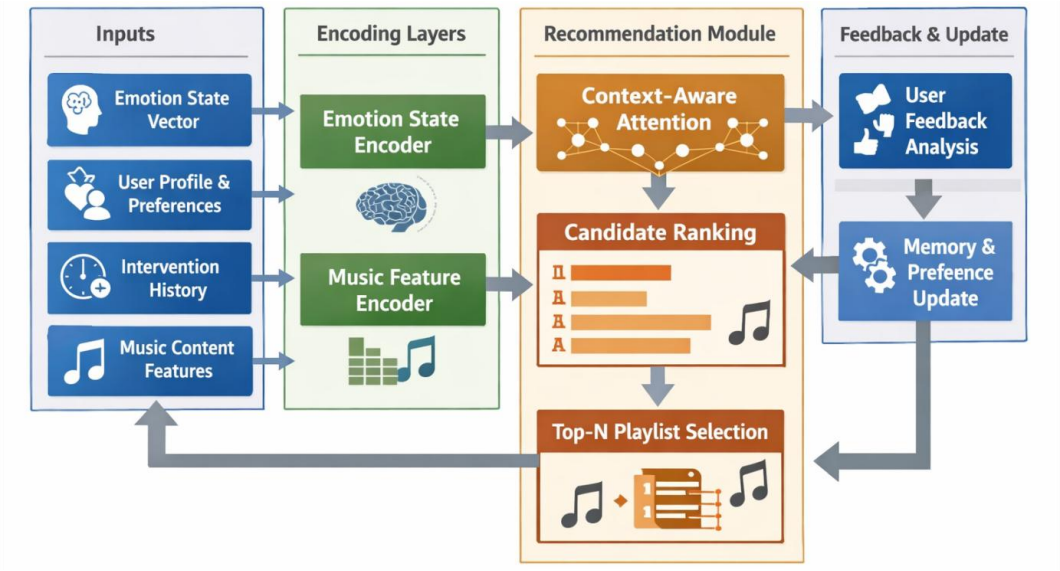

Figure 4: Overall architecture of personalized music healing system

In Figure 4, the system input consists of four parts: the state vector output by the emotion recognition model, the user static attributes and long-term preferences, the historical intervention response records, and the content features in the music library. The emotional state encoder was responsible for converting the recognition results of "anxiety tendency, depression tendency, calm state, positive state" into a computable healing need vector. The music feature encoder extracts structured representations from track rhythm, mode, energy intensity, lyric sentiment polarity, age tag and familiarity tag. In order to realize the quantitative matching between emotional needs and musical attributes, this paper defines the basic matching score of candidate track i at time t as follows.

$$r_{t,i}=\alpha\,\phi(e_t,m_i)+\beta\,\psi(p_u,m_i)+\gamma\,\eta(h_u,m_i) \tag{8}$$

where, e_t represents the emotional state vector at time t, m_i is the content representation of the ith candidate music, p_u is the user's long-term preference vector, and h_u is the historical intervention response representation. $\phi(\cdot)$ is used to describe the consistency between emotional needs and music attributes, $\psi(\cdot)$ measures the compatibility between user preferences and track features, and $\eta(\cdot)$ reflects the verified healing effectiveness in historical feedback. α, β and γ are the learnable weights. This expression does not limit the recommendation goal to a simple correspondence of "emotion tags to track tags", but incorporates the current state, individual preference and previous results into the same scoring function.

It is still difficult to characterize complex intervention situations by linear combination alone. Elderly users often need different music stimulation intensity for the same emotional state in the morning, before lunch break and at night before going to bed. In the same mood, some users were better suited to the steady comfort of a familiar melody, while others were more likely to respond to mild positive arousal. To this end, this paper adds a contextual attention fusion mechanism to the recommendation module, so that the system can dynamically weight different source features, and its process is shown in Figure 5.

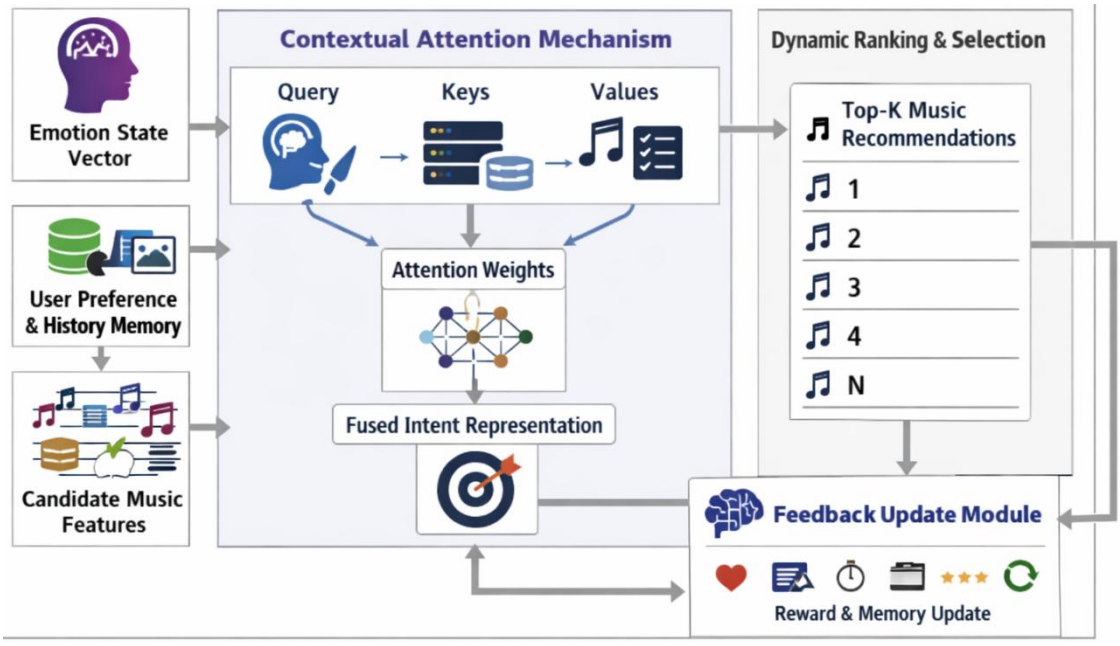

Figure 5: Recommendation optimization and feedback update mechanism based on contextual attention

In Figure 5, the sentiment vector is mapped as a query item, and the preference memory enters the attention module with the candidate track features as keys and values, respectively. The system calculates the contribution weights of different candidate tracks to the current healing task through contextual correlation and generates the fused intervention intention representation. Its attention allocation can be expressed as follows.

$$a_{t,i}=\frac{\exp\left(q_t^{\top}k_i/\sqrt{d}\right)}{\sum_{j=1}^{N}\exp\left(q_t^{\top}k_j/\sqrt{d}\right)} \tag{9}$$

where, q_t is the mapped query vector of the current emotional state, k_i is the key vector of the ith candidate track, d is the feature dimension, and $a_{t,i}$ represents the attention weight of the track in the current context. The higher the attention value, the more consistent the track is with the current user's healing needs. Then the system generates the final ranking result according to the basic score of $a_{t,i}$ and Equation (8), and outputs the Top-K track set for actual playback.

In order to make the recommendation mechanism have the ability of continuous optimization, this paper introduces the feedback update link. The system synchronously collected physiological changes, dwell time, skip rate, repeat playback, subjective rating and emotional drop amplitude after intervention during the playback process, and integrated them as reward signals. The way the user's memory state is updated is defined as follows.

$$s_u^{(t+1)}=\lambda s_u^{(t)}+(1-\lambda)\,\Delta y_t \tag{10}$$

where, $s_u^{(t)}$ represents the user's personal therapy preference state at time t, Δy_t is the comprehensive feedback increment after this round of intervention, and λ is the memory retention coefficient. This update method enables the system not only to retain long-term preferences, but also to complete partial correction according to recent responses, avoiding the recommendation strategy to stay on the initial experience setting for a long time.

The personalized music healing system formed by this is essentially a closed-loop computing framework composed of "emotion recognition, demand encoding, candidate ranking, playing and executing, and feedback writing back". Instead of treating music recommendation as an independent information retrieval task, it is embedded into the emotional intervention link for the elderly: the front-end recognition results provide the state basis, the mid-end attention fusion module completes the healing intention modeling, and the back-end feedback update continuously modifies individual preferences and recommendation weights. The system constructed in this way can maintain good adaptability in the application scenarios of slow emotional fluctuations, large individual differences and diverse intervention targets in the elderly, and also provides a unified implementation basis for the verification of intervention effects and the analysis of system efficiency in the following article.

3 Results

3.1 Analysis of experimental results of emotion recognition model for the elderly

In order to test the effectiveness of the elderly emotion recognition model constructed in this paper in real application scenarios, this paper compares the proposed model with three methods of LSTM, CNN and BiGRU under unified data division, the same input features and consistent

training environment. The experimental platform is configured with Intel Core i7-12700, 32 GB RAM, Ubuntu 22.04, and the models are implemented based on Python 3.11 and PyTorch framework. In order to reduce the occasional fluctuations caused by a single training, all experiments are repeated 5 times independently, and the results are averaged in the paper. The standard deviation of Accuracy, Macro-F1 and AUC of five independent experiments are less than 1.0%, which indicates that the model training results have good repeatability. The standard deviation of each index is controlled in a small range, which indicates that the experimental results have good stability.

The dataset consists of speech clips, facial video clips, physiological signal clips and interaction logs, and is divided into training set, validation set and test set according to the principle of subject independence, with the ratio of 8:1:1. Among them, the training set, validation set and test set are completely independent at the subject level to avoid information leakage caused by the distribution of samples from the same subject across sets. This division method can avoid the same subject sample entering the training and testing phase at the same time, so as to more truly investigate the cross-individual generalization ability of the model. In the training phase, five-fold cross validation is used, and the training set is lightly enhanced, including speech noise disturbance, video brightness jitter and random cropping of physiological signal segments, so as to reduce the dependence of the model on a single acquisition condition. In this paper, the emotional states are divided into four categories: calm, positive, anxiety tendency and depression tendency, and the evaluation indicators include Accuracy, Macro-F1, AUC and single sample average inference delay. The labels of each category were determined by the emotion scale score and manual review to improve the consistency of labeling results.

From the perspective of classification performance, the proposed model shows strong advantages on four types of elderly emotional tasks. Figure 6 shows the variation of test set accuracy for different models during training. It can be seen that CNN improves rapidly in the early stage, but tends to be flat after about 20 epochs, indicating that it is sensitive to local spatial features, but it is difficult to continue mining long-term dependencies. The rising process of LSTM is relatively stable, but the overall accuracy is limited by the ability to represent the multi-modal spatial structure. BiGRU is an improvement over the one-way time series model, although it still has fluctuations in the later stage. In contrast, the proposed model has reached a high recognition level after the 15th epoch, and remains stable at more than 90% after the 30th epoch, and the final test accuracy reaches 93.84%. This indicates that the joint introduction of convolutional coding, bidirectional temporal modeling, and cross-modal attention fusion can more effectively capture the fine-grained changes in elderly emotions.

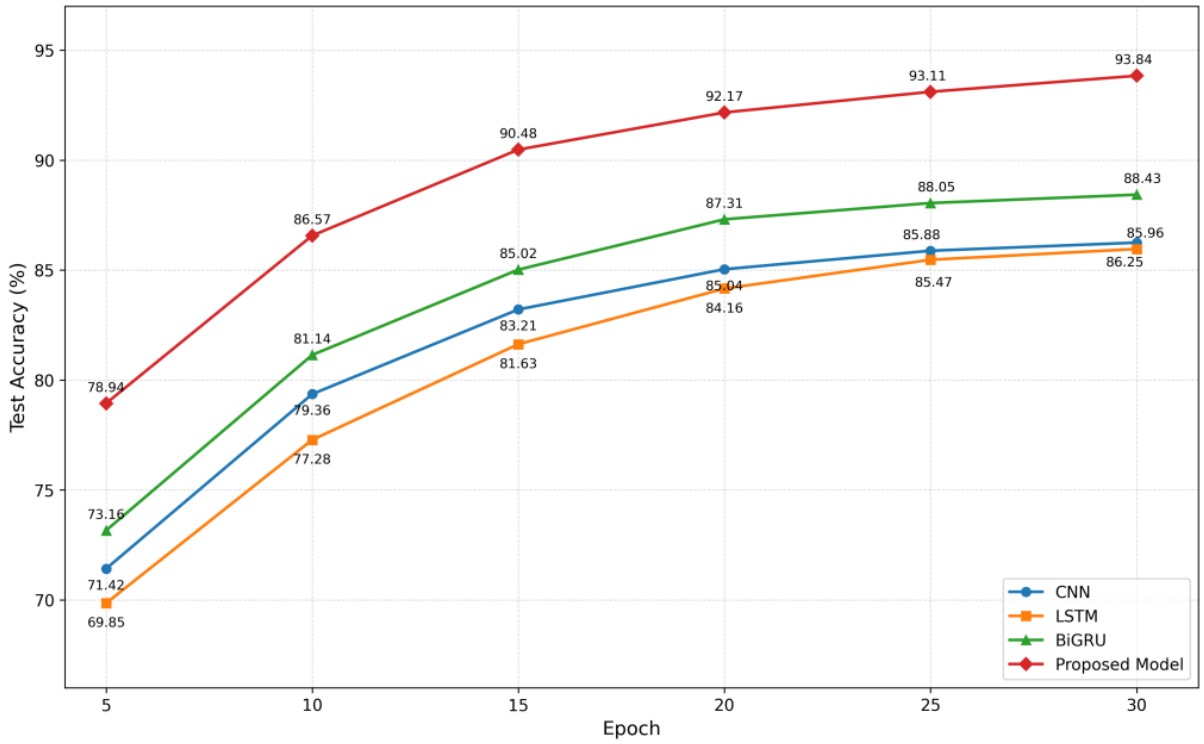

Figure 6: Variation of test set accuracy with training rounds for different models

Only the overall accuracy is not enough to show the advantages of the model, and the more critical issue of elderly emotion recognition is whether to maintain a balanced discrimination between different categories. Table 3 lists the comprehensive performance of the four models on the test set. It can be found that the proposed model is higher than the comparison methods in the three indicators of Accuracy, Macro-F1 and AUC, and the average inference delay is only 23.6 ms/ sample, without obvious computational burden due to the increase of model structure. Among them, Macro-F1 reaches 92.47%, indicating that the recognition of minority emotions by the model is not squeezed by the majority class. The AUC is 95.12%, which means that it has a good ability to rank different emotional risk boundaries. Combined with the structural analysis of the model above, it can be judged that cross-modal attention plays an important role in this. It is not simply to amplify a certain mode, but to increase the weight of these key segments in the final decision when speech pauses, expression micro-changes and physiological mild fluctuations appear at the same time.

Table 3: Comprehensive performance comparison of different models on geriatric emotion recognition task

Model	Accuracy /%	Macro-F1 /%	AUC /%	Inference Latency /ms
CNN	86.25	84.73	88.94	18.7
LSTM	85.96	84.18	88.26	21.4
BiGRU	88.43	87.05	90.77	22.1
Proposed Model	93.84	92.47	95.12	23.6

In order to further verify the stability of the model across individual conditions, this paper counted the recognition accuracy and fluctuation on different subject groups, and the results are shown in Figure 7. It can be seen from the figure that CNN and LSTM show a significant decline in some subject groups, with the lowest accuracy of 81.6% and 80.8% respectively,

indicating that they are more susceptible to the expression differences of elderly individuals. Although BiGRU is more stable as a whole, it still has a certain decline in the sample group with weak expression changes. The accuracy of the proposed model on six subject groups is maintained above 91%, and the fluctuation range is controlled within 2.4%, showing good cross-individual robustness. The reason is that the model does not establish the emotion recognition on a single salient feature, but through the collaborative modeling between multi-modal sequences to find a more stable discrimination basis.

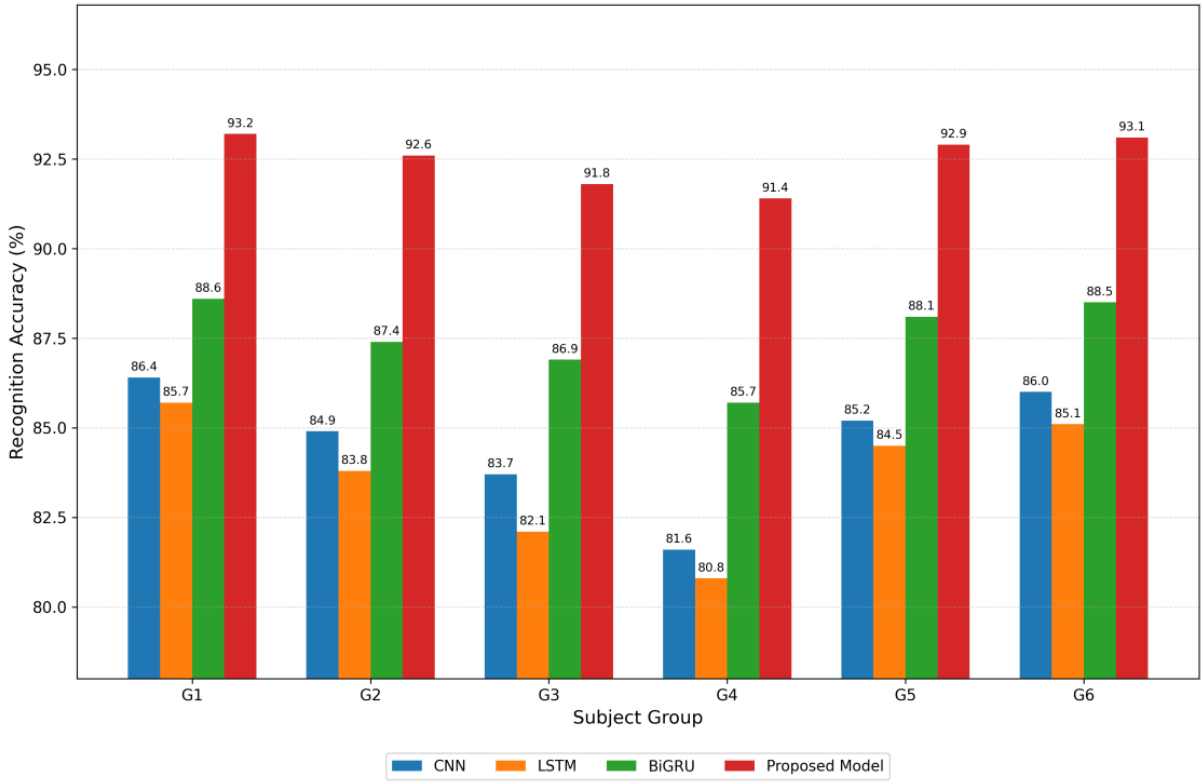

Figure 7: Cross-individual recognition accuracy comparison on different subject groups

In general, the proposed model is superior to the comparison methods in terms of overall recognition accuracy, class balance, risk ranking ability and cross-individual stability, and the inference cost is still in the acceptable range. The results show that the multimodal depth recognition for elderly emotional scenes should not stay at the level of a single sequence model or a single convolutional model, but should be completed through spatial feature extraction, temporal dependency modeling and attention fusion.

3.2 Verification of intervention effect of personalized music therapy system

After completing the performance verification of the front-end emotion recognition model for the elderly, this paper further investigates the application effect of the personalized music therapy system in the actual intervention scenario. Different from simply discussing the accuracy of emotion classification, this section focuses more on whether the system can stably output suitable tracks in different use environments after the recognition results enter the recommendation link, and have a positive impact on the emotional relief and behavior cooperation of elderly individuals. Therefore, this paper selected three application scenarios of community day care, collective intervention of elderly care institutions and remote

companishment at home for verification, and unified the mood improvement rate, recommendation acceptance rate, intervention completion rate and average response time as the core indicators. All experiments are conducted based on the same recognition model and the same music library to ensure that the results are comparable. The experimental group used the proposed system for dynamic recommendation, and the control group used a fixed playlist. The two groups were consistent in the duration of intervention and the frequency of playback.

The test results under three categories of scenarios are shown in Table 4 It can be seen that the average emotional improvement rate of the system in the community day care scene is 24.8%, and the recommended acceptance rate is 87.6%, indicating that the system can better match the mild anxiety and depression states after daytime activities. In the collective intervention environment of pension institutions, the average emotional improvement rate of the system was 22.9%, which was slightly lower than that of the community scenario, but the intervention completion rate reached 90.4%, which showed that the system had good stable execution ability in the group environment. In the home remote companiation scenario, due to the larger differences in individual work and rest and the more complex environmental interference, the emotional improvement rate is 20.7%, but the recommendation acceptance rate still remains above 82.4%, indicating that the system still has certain adaptability for elderly users under decentralized and unsupervised conditions. Overall, the average response delay in the three types of scenes is less than 160 ms, indicating that the system has good real-time performance in the process of completing emotional input, candidate track sorting and result output.

Table 4: Intervention effects of personalized music healing system in different application scenarios

Result Parameter	Community Day Care	Institutional Group Intervention	Home-Based Remote Companionship
Application task	Daytime emotional soothing	Group calming and rhythm stabilization	Bedtime companionship and low-mood relief
Average emotion improvement rate / %	24.8	22.9	20.7
Recommendation acceptance rate / %	87.6	84.3	82.4
Intervention completion rate / %	88.9	90.4	79.8
Average response latency / ms	148.6	152.1	158.4
Increase in the proportion of positive emotions after intervention / %	18.5	16.9	15.2

The differences, reflected in Table 4, are closely related to the behavior patterns of elderly users in different scenarios. Subjects in community day care usually enter the system at a fixed time, and their emotional states are more concentrated. The system is easier to output music programs with higher matching degrees based on recognition results and historical preferences. Although the collective intervention in elderly care institutions has large individual differences, the environment is relatively stable, and the system performs well in the recommendation of slow rhythm and familiar melody tracks. In the case of the lack of immediate assistance from caregivers, the acceptance rate and completion rate decreased to a certain extent, but the overall results were still within the usable range, which indicates that the closed-loop mechanism of "emotion recognition-recommendation optimization-feedback update" proposed in this paper

is not only suitable for the ideal experimental environment.

In order to further observe the continuous intervention effect of the system, this paper counted the mood improvement trend between the experimental group and the control group during the 4-week intervention period, and the results are shown in Figure 8. The control group used a fixed playlist and did not dynamically adjust the tracks according to the emotional state. The experimental group used the personalized music healing system proposed in this paper. It can be seen that the experimental group showed more significant improvement in the first week, and the mood improvement rate increased from 12.4% to 27.6% in the fourth week. The control group also showed some improvement, but only reached 16.3% by week 4. This result shows that static music playing can indeed produce a certain companiency effect, but the dynamic recommendation mechanism is easier to maintain the effect gain in long-term intervention, because the system can continuously update the track ranking according to the recent recognition results and feedback records, and reduce the reduction of intervention efficiency caused by "repeated expose-interest decay".

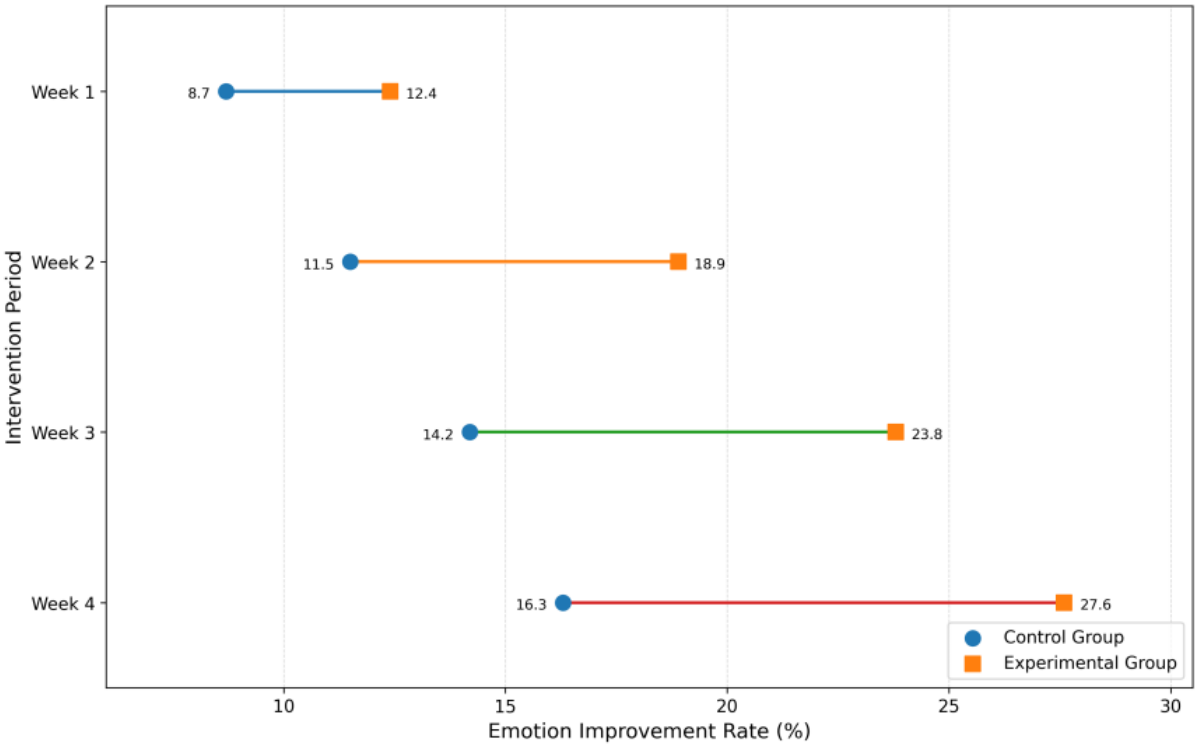

Figure 8: Changes in mood improvement rate between the experimental group and the control group during the 4-week intervention period

On the whole, the proposed system shows good intervention feasibility in different elderly application scenarios. Its advantage is not only reflected in the accuracy of the recommendation, but also in the ability to convert the emotional changes identified by the front-end into executable music healing strategies, and constantly revise the recommendation results through user feedback. In other words, the system does not simply "recognize and then play music", but gradually approaches the actual needs of the individual through continuous interaction.

3.3 Analysis of system operation efficiency and computational complexity

In addition to the recognition accuracy and intervention effect, whether the personalized music therapy system can enter the actual elderly care service scene also depends on whether its

operating efficiency and computational load are in an acceptable range. For the task of this paper, the system does not complete offline classification only once, but needs to complete multi-modal data reading, emotional state inference, candidate track ranking and feedback update in the continuous interaction process. Therefore, if the model parameter scale is too large and the inference delay is too long, even if the recognition results are better, it is difficult to meet the real-time application requirements in elderly companishment, bedtime comfort and institutional care. Based on this, this paper makes statistics on the accuracy, single-sample inference delay, parameter number and floating-point calculation of different model schemes on the unified experimental platform, and the results are shown in Table 5.

Table 5: Comparison of operational efficiency and computational complexity of different model schemes

Model Scheme	Accuracy / %	Inference Latency / ms·sample^{-1}	Number of Parameters / M	FLOPs / G	GPU Memory Usage / GB
CNN System	86.25	18.7	3.4	2.1	1.8
LSTM System	85.96	21.4	4.1	2.6	2.0
BiGRU System	88.43	22.1	4.8	2.9	2.2
Proposed System	93.84	23.6	5.6	3.2	2.4

Table 5 shows that although the proposed system is higher than the single CNN and LSTM scheme in the number of parameters and FLOPs, the growth rate remains in the controllable range, and the problem of rapid complexity expansion does not occur. In contrast, the increased computational overhead brings more obvious performance gains: the recognition accuracy reaches 93.84%, which is 7.59 percentage points higher than that of CNN system and 5.41 percentage points higher than that of BiGRU system. The average inference delay of a single sample is 23.6 ms, which is only 4.9 ms higher than that of the pure convolution scheme, which means that the system can still complete near-real-time response under regular interaction frequency, and will not cause obvious blockage to music playback and feedback update.

Structurally, the efficiency advantage of the proposed system does not come from simply compressing the network size, but from the more compact organization of the computational links. The front-end convolutional coding is responsible for extracting local emotion patterns to avoid invalid original features entering the subsequent layers in a large scale. Although the bidirectional time series modeling introduces additional overhead, it improves the long-term information utilization and reduces the repeated recommendation correction caused by misjudgment. The cross-modal attention and recommendation update module only performs weighted allocation in the fusion stage, and does not introduce too deep stacked structure, so the overall resource consumption is still maintained at a light level. In other words, the proposed system achieves a balance of "higher recognition quality and more stable intervention output at moderate complexity", rather than stacking performance at computational cost.

3.4 Comparative experimental analysis with advanced methods

In order to more accurately define the performance position of the proposed method in the task of emotion recognition and personalized music therapy for the elderly, this paper selects representative related methods in the literature as reference, and makes comparative analysis combined with the experimental results of this paper. Due to differences in the original data partition and implementation details, the comparison is mainly used to illustrate the performance position, rather than strictly reproducing the experiment under the same conditions. The comparison objects include Two-Step Hybrid Feature Fusion Network (THFN), the

emotion classification method for the elderly based on middle layer fusion and cross-modal transfer learning, the CNN method for personalized treatment support, and the model-level fusion method for medical analysis scenarios. The reason why we choose these methods is that they represent several representative technical paths of "hybrid feature fusion", "cross-modal transfer", "convolutional recognition" and "deep fusion optimization" in elderly emotion recognition, which can reflect the relative advantages of the method in this paper.

In the testing process, this paper still adopts the 8:1:1 subject independent division strategy, and uniformly reports the three indicators Accuracy, Macro-F1 and AUC. The comparison results are shown in Figure 9. It can be seen that THFN has certain advantages in multi-layer feature aggregation, but its modeling focus is still biased towards static fusion, and the slow evolution of elderly emotions on continuous time scales is insufficiently utilized. The middle layer fusion and cross-modal transfer learning methods perform well in cross-modal adaptation, but the classification boundary still has certain fluctuations when facing the elderly samples with large individual differences. The CNN method is effective in local pattern extraction, but its ability to deal with long-term dependence and cross-modal asynchronous changes is limited. Although the model-level fusion method improves the overall discrimination ability, it lacks a closer closed-loop correlation between the emotion recognition results and the subsequent healing recommendation. In contrast, the proposed method achieves the best results on the three indicators, with Accuracy, Macro-F1 and AUC reaching 93.84%, 92.47%and 95.12%respectively, which are higher than the other comparison models as a whole.

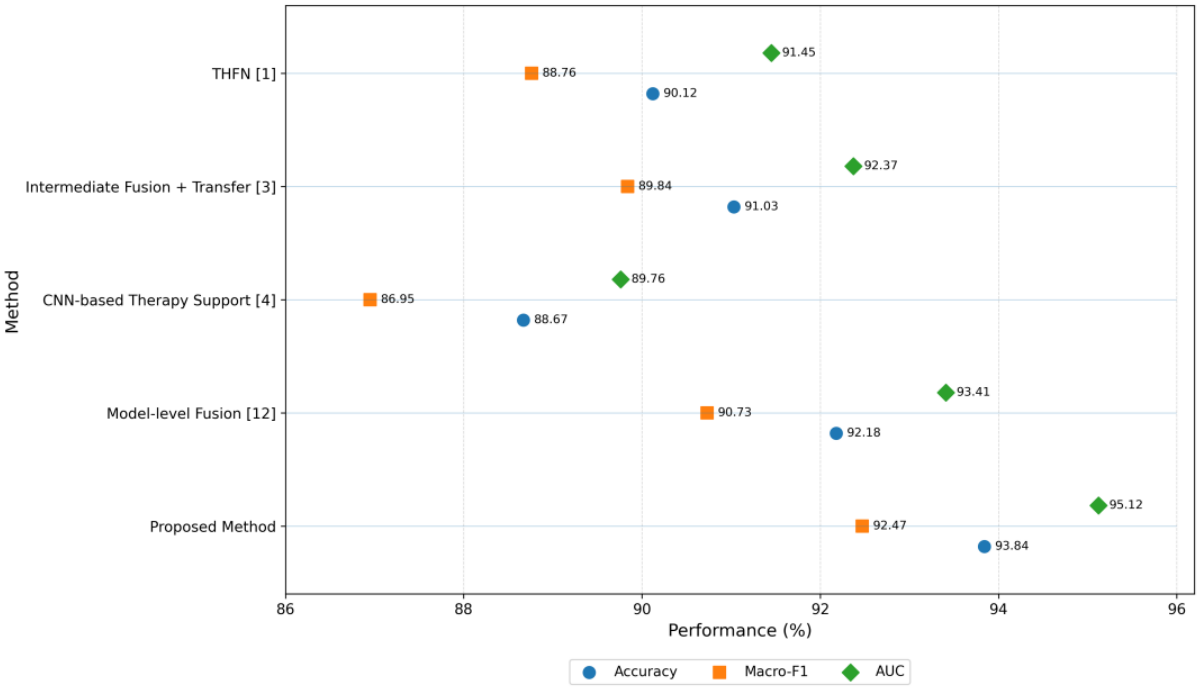

Figure 9: Comparison of comprehensive performance of different advanced methods in geriatric emotion recognition task

The results presented in Figure 9 illustrate that the advantage of the proposed method does not come from the local enhancement of a single module, but from the co-design of the overall computing link. The front-end convolutional coding module can stably extract local emotion-sensitive features in speech spectrogram, facial micro-expression and physiological rhythm. The bidirectional temporal network further makes up for the problems of strong persistence and weak instantaneous amplitude of emotional expression in the elderly. More importantly, the

proposed method does not regard emotion recognition as the end point, but directly connects the output results to the subsequent personalized music therapy system, which makes the recognition model optimized for intervention applications at the beginning of the design, rather than remaining in an isolated classification task.

4 Discussion

The emotion recognition and personalized music therapy system for the elderly constructed in this paper shows consistent advantages in three levels of recognition performance, intervention effect and operation efficiency, which shows that it is reasonable to put multimodal deep learning and individualized recommendation mechanism in the same computing framework. Experimental results show that the proposed model is superior to the comparison methods in terms of Accuracy, Macro-F1 and AUC, indicating that the joint design of convolutional coding, bidirectional temporal modeling and cross-modal attention fusion can better adapt to the actual characteristics of "low expression amplitude, slow change rhythm, and asynchronous modal response" of elderly emotions. Its performance improvement does not come from the local enhancement of a single module, but from the synergy between spatial feature extraction, temporal dependence preservation and key segment weighting.

From the perspective of system mechanism, the attention fusion module plays a similar role as an "adaptive allocator" in the proposed framework. It does not mechanically superimpose speech, expression and physiological signals, but assigns higher weights to the modal segments with more diagnostic value according to the current emotional state and context conditions. This dynamic adjustment mechanism enables the model to maintain stable recognition results in the case of large differences in elderly samples and more local noise. Furthermore, the back-end music therapy system does not stay at the static recommendation level, but constantly modifies the user preference representation through feedback updates, so that the recommendation process changes from one-time matching to continuous optimization, which is also an important reason why the intervention effect of the experimental group is better than that of the fixed playing strategy.

However, there are still some limitations in this paper. First, although the current experimental data cover multi-modal inputs, the sample size and long-term tracking time are still limited, and the generalization ability of more complex elderly groups needs to be further tested. Second, the system feedback is mainly constructed based on behavioral responses and short-term emotional changes, which has not yet been fully incorporated into the long-term healing stability indicators. Third, although the system inference delay has been controlled at a low level, there is still room to further compress the amount of parameters and optimize the resource occupation for the lightweight deployment of edge devices and community terminals. Future research can focus on cross-scenario transfer, online incremental learning, and individual long-term healing modeling, so as to improve the continuous adaptability of the system in the real elderly care service environment.

5 Conclusion

Aiming at the problems of insufficient utilization of multimodal information, insufficient modeling of temporal dependence, and lack of dynamic adaptation mechanism of music intervention in elderly emotion recognition, this paper constructs a deep learning based emotion recognition and personalized music therapy system for the elderly. Based on the front-end multi-modal feature extraction and unified coding, this method combines convolutional

network, bidirectional temporal modeling and cross-modal attention fusion to realize the stable discrimination of the emotional state of the elderly. At the same time, the recognition results, user preferences, historical feedback and music content characteristics are integrated into the recommendation link to form a healing closed loop of sustainable update. Experimental results show that the proposed model achieves good comprehensive performance in the elderly emotion recognition task, with Accuracy, Macro-F1 and AUC reaching 93.84%, 92.47%and 95.12%respectively. It also shows good emotion improvement ability and system response efficiency in the intervention scenario. It can be seen that the collaborative modeling of emotion recognition results and music recommendation mechanism is helpful to improve the adaptation ability and application value of the system in the elderly care scene. This paper still has the problem of limited sample size and long-term tracking data. Subsequent research can continue to expand the multi-scenario elderly data set, and introduce online incremental learning and lightweight deployment strategies to enhance the generalization ability and continuous application value of the system in the real elderly care service environment.

Funding

2024 Changchun University of Social Sciences and Humanities in Changchun, Jilin Province Social Science and Humanities Enterprise Entrusted Project "Background Music Production and Research for Sleep Assistance" 2024JBH20W422

References

[1] Jothimani S, Premalatha K. THFN: Emotional health recognition of elderly people using a Two-Step Hybrid feature fusion network along with Monte-Carlo dropout[J]. Biomedical Signal Processing and Control, 2023, 86: 105116.

[2] Grossi A, Gasparini F. Ser_ampel: a multi-source dataset for speech emotion recognition of italian older adults[C]//Italian Forum of Ambient Assisted Living. Cham: Springer Nature Switzerland, 2023: 70-79.

[3] Sreevidya P, Veni S, Ramana Murthy O V. Elder emotion classification through multimodal fusion of intermediate layers and cross-modal transfer learning[J]. Signal, image and video processing, 2022, 16(5): 1281-1288.

[4] Torcate A S, de Santana M A, dos Santos W P. Emotion recognition to support personalized therapy in the elderly: an exploratory study based on CNNs[J]. Research on Biomedical Engineering, 2024, 40(3): 811-824.

[5] Onim M S H, Thapliyal H, Rhodus E K. Utilizing machine learning for context-aware digital biomarker of stress in older adults[J]. Information, 2024, 15(5): 274.

[6] Abdollahi H, Mahoor M H, Zandie R, et al. Artificial emotional intelligence in socially assistive robots for older adults: a pilot study[J]. IEEE transactions on affective computing, 2022, 14(3): 2020-2032.

[7] Guo R, Guo H, Wang L, et al. Development and application of emotion recognition technology—a systematic literature review[J]. BMC psychology, 2024, 12(1): 95.

[8] Wang X, Ren Y, Luo Z, et al. Deep learning-based EEG emotion recognition: Current trends and future perspectives[J]. Frontiers in psychology, 2023, 14: 1126994.

[9] Hamzah H A, Abdalla K K. EEG-based emotion recognition systems; comprehensive study[J]. Heliyon, 2024, 10(10).

[10] Lian H, Lu C, Li S, et al. A survey of deep learning-based multimodal emotion recognition: Speech, text, and face[J]. Entropy, 2023, 25(10): 1440.

[11] Pan J, Fang W, Zhang Z, et al. Multimodal emotion recognition based on facial expressions, speech, and EEG[J]. IEEE Open Journal of Engineering in Medicine and Biology, 2023, 5: 396-403.

[12] Islam M M, Nooruddin S, Karray F, et al. Enhanced multimodal emotion recognition in healthcare analytics: A deep learning based model-level fusion approach[J]. Biomedical Signal Processing and Control, 2024, 94: 106241.

[13] Cui X, Wu Y, Wu J, et al. A review: Music-emotion recognition and analysis based on EEG signals[J]. Frontiers in neuroinformatics, 2022, 16: 997282.

[14] Su Y, Liu Y, Xiao Y, et al. A review of artificial intelligence methods enabled music-evoked EEG emotion recognition and their applications[J]. Frontiers in Neuroscience, 2024, 18: 1400444.

[15] Jiang X, Zhang Y, Lin G, et al. Music emotion recognition based on deep learning: A review[J]. IEEE Access, 2024, 12: 157716-157745.

[16] Wang M, Wu J, Yan H. Effect of music therapy on older adults with depression: A systematic review and meta-analysis[J]. Complementary therapies in clinical practice, 2023, 53: 101809.

[17] Lin T H, Liao Y C, Tam K W, et al. Effects of music therapy on cognition, quality of life, and neuropsychiatric symptoms of patients with dementia: A systematic review and meta-analysis of randomized controlled trials[J]. Psychiatry Research, 2023, 329: 115498.

[18] Tz-Han L, Wan-Ru W, Chen I H, et al. Reminiscence music intervention on cognitive, depressive, and behavioral symptoms in older adults with dementia[J]. Geriatric Nursing, 2023, 49: 127-132.

[19] Prick A E J C, Zuidema S U, van Domburg P, et al. Effects of a music therapy and music listening intervention for nursing home residents with dementia: a randomized controlled trial[J]. Frontiers in medicine, 2024, 11: 1304349.

[20] Haddad N R, Bhardwaj T, Zide B S, et al. A remotely delivered, personalized music therapy pilot intervention for lonely older adults during the COVID-19 pandemic[J]. The American Journal of Geriatric Psychiatry: Open Science, Education, and Practice, 2024, 1: 7-16.

[21] De Nys L, Oyebola E F, Connelly J, et al. Digital music and movement intervention to improve health and wellbeing in older adults in care homes: a pilot mixed methods study[J]. BMC geriatrics, 2024, 24(1): 733.