Multimodal Sentiment Analysis and Improved Transformer-Based Emotional Music Generation Method for Elderly People in Nursing Homes

Shijia Fan

Tonbridge School, Tonbridge, United Kingdom

Abstract

Aiming at the predicaments of loneliness, depression, and cognitive decline commonly faced by elderly people in nursing homes-where existing manual emotional support struggles to achieve accurate responses due to limited nursing resources. This study conducts research on multimodal sentiment analysis (integrating facial images, EEG signals, and EOG signals) and improved Transformer-based emotional music generation for this group. The research holds significant theoretical supplementary value and practical application significance: it not only provides a new paradigm of multimodal fusion for the field of elderly affective computing but also offers a precise emotional and cognitive intervention solution for the elderly in nursing homes, alleviating their negative emotions and reducing the burden on caregivers. The core innovations of the study are reflected in two aspects: First, a multimodal sentiment analysis model is constructed with optimized design in the feature extraction stage. An improved PFLD model incorporating a Feature Module is adopted to enhance the accuracy of facial landmark detection through multi-scale feature fusion; the Linear Dynamic System (LDS) is used to smooth EEG signal features for reducing noise interference; continuous wavelet transform and peak detection are combined to extract key features (e.g., blinking, fixation) from EOG signals; and a 1D Convolutional Neural Network (1DCCNN) is designed to adapt to the temporal characteristics of physiological signals. Finally, multimodal features are fused via a convolutional encoder and input into a Recurrent Neural Network (RNN) for sentiment classification, breaking the limitations of single-modal recognition. Second, the Discrete-Emo Transformer emotional music generation model is proposed: the Flash attention mechanism is introduced to alleviate memory bottlenecks and improve inference speed; the segment recurrence mechanism of Transformer-XL is integrated to enhance the model's ability to learn long-range dependencies in long music sequences; and the REMI-EMO representation method is used to simultaneously model note features and emotional information, enabling high-quality symbolic music generation under emotional conditions. Experimental results show that the multimodal sentiment analysis model achieves an accuracy of 97.5%, outperforming mainstream models such as FDMER and MISA. The Discrete-Emo Transformer model reaches an emotional accuracy of 78.4% and performs optimally in the Perplexity (PPL) index (1.71), generating music that is more in line with target emotions and human creative characteristics. This study provides key technical support for "technology-empowered elderly care" and effectively improves the accuracy and efficiency of emotional intervention in nursing homes.

Keywords

Multimodal sentiment analysis, emotional music generation, improved transformer.

1. Introduction

Currently, loneliness has become a core issue prevalent among the elderly in nursing homes, and its chain reactions pose a serious threat to their physical and mental health. Social isolation is not only a lack of emotional connection but also directly induces a series of mental health problems, such as high incidence of negative emotions like depression and anxiety. Some elderly people even experience persistent cognitive decline due to long-term lack of emotional interaction and social support. Although nursing homes have now recognized the importance of emotional and cognitive support for the elderly and attempted to alleviate related problems through companionship, emotional communication, and cognitive activities, some institutions have even introduced multimodal stimulation (e.g., intervention combining visual, auditory, and tactile stimuli) to enhance the elderly's cognitive engagement, existing support methods still have obvious limitations. On one hand, traditional manual companionship and emotional support are restricted by the quantity and professional level of caregivers, making it difficult to achieve real-time monitoring and personalized responses to the emotional state of each elderly person. They often fail to timely capture subtle emotional changes of the elderly, resulting in inaccurate interventions. On the other hand, existing multimodal stimulation mostly adopts fixed content for batch application, failing to fully consider individual differences among the elderly-such as differences in preferences for music and images, and differences in cognitive needs of dementia patients at different disease stages-greatly reducing the intervention effect. The development of artificial intelligence technology provides a new idea for solving this dilemma: its applications in natural scenery video generation and soothing music customization have shown potential in alleviating the anxiety of the elderly.

Conducting research on multimodal sentiment analysis (integrating facial images, EEG signals, and EOG signals) and emotional music generation for the elderly in nursing homes has important theoretical significance and practical value. Theoretically, this study will break the limitation of single-modal sentiment recognition. By integrating multi-dimensional data of vision (facial expressions) and physiology (EEG, EOG), it will construct a more accurate and robust elderly sentiment recognition model, fill the research gap in the field of affective computing for special elderly groups (especially dementia patients), and provide empirical support for the improvement of the theoretical system of elderly affective computing. Practically, the research results will provide a "personalized, real-time, and precise" emotional and cognitive intervention solution for the elderly in nursing homes: through multimodal sentiment analysis technology, real-time capture of the elderly's emotional fluctuations and cognitive state changes (e.g., depressive tendencies, anxiety levels, and attention concentration) can be realized; based on the recognition results, AI technology is used to generate customized emotional music, which not only adjusts the content according to the elderly's personal preferences (e.g., music style, rhythm) but also matches their current emotional state (e.g., generating soothing melodies for anxious elderly people and lively rhythms for elderly people with low mood), realizing personalized intervention of "music tailored to emotion." This intervention method can not only effectively alleviate the loneliness and negative emotions of the elderly but also enhance cognitive engagement through music stimulation, helping to slow down the rate of cognitive decline and improve their quality of life. At the same time, this technology can reduce the workload of caregivers: through automated emotional monitoring and intervention assistance, it lowers the emotional and physical pressure of caregivers, reduces care burnout, further optimizes the service efficiency of nursing homes, and provides key technical support for building a new model of "technology-empowered elderly care."

2. Related Work

2.1. Multimodal Information-Based Sentiment Analysis

As one of the external manifestations of human emotions, facial expressions carry rich emotional information and are an important part that cannot be ignored in daily communication. With the continuous development of deep learning technologies, innovative models such as Transformer have also been widely used in facial expression recognition, improving recognition accuracy [1]. Zhao Z et al. proposed a cross-modal Transformer model [2], which integrates spatiotemporal features of video sequences and multimodal information, solving the shortcomings of traditional methods in modeling temporal dependencies and occlusion issues in dynamic expression sequences.

In the field of affective computing research, scholars mostly select nine types of physiological indicators as biomarkers for emotional state recognition, including EEG, ECG, GSR, electrooculogram (EOG), respiratory rate (RR), photoplethysmographic (PPG), electromyogram (EMG), blood volume pulse (BVP), and skin temperature (SKT). Atefeh G et al. conducted research on emotion recognition by collecting multiple physiological signals including EMG, ECG, GSR, and respiratory rate [3]. Kim K H et al. collected three types of physiological data (GSR, ECG, and skin temperature) [4], extracted multiple emotion-related feature parameters, and used a support vector machine (SVM) to classify and recognize four different emotional states.

Currently, multimodal sentiment recognition has gradually become a research hotspot. The fusion analysis of physiological signals and facial expressions can more comprehensively reflect an individual's emotional state [5]. Among them, physiological signals show advantages in sentiment recognition due to their high stability and strong anti-interference ability, and EEG signals have great application potential [6]. Although there are still technical challenges such as equipment wearing and signal noise, with the continuous development of signal processing methods and deep learning algorithms, multimodal fusion will improve the accuracy of sentiment recognition and provide strong support for fields such as intelligent interaction [7]. The core idea of feature-level fusion is to integrate features from different modalities into a unified feature vector, which contains information from all modalities and can be used as input for subsequent classification tasks. Generally, concatenation operation is used to splice multiple features into a higher-dimensional feature. S. Wang proposed a deep learning model for multimodal sentiment recognition based on EEG signals and facial expressions [8]. The basic idea of decision-level fusion is that after the features of different modalities are independently analyzed and classified, a specific mathematical strategy is used to integrate the discrimination results of each modality to obtain the final multimodal sentiment recognition result. Common fusion methods include maximum strategy, minimum strategy, arithmetic average method, and vector dot product method, among which the weighted summation method is most frequently used. Lian Y et al. proposed a bimodal emotion recognition method based on EEG and facial images [9]. The hybrid fusion method combines the information fusion strategies of feature-level and decision-level, which not only exerts the advantages of each modality but also effectively makes up for their shortcomings. With the hybrid fusion method, in the feature extraction stage, some modalities are subjected to feature-level fusion to obtain cross-modal features, which are then further extracted with the unimodal features that have not been fused. Finally, the classification predictions of cross-modal and unimodal features are fused through a decision fusion method to output the final prediction result. Cimtay et al. [10] performed feature-level fusion on GSR and EEG signals to obtain cross-modal physiological signals, output the classification result of physiological signals through a neural network, and finally fused this result with the classification result of facial expressions at the decision level.

2.2. Transformer-Based and Emotion-Driven Music Generation

The Transformer model is a time-series model that relies entirely on the self-attention mechanism [11]. It consists of two modules: encoder and decoder. It processes sequence data through multi-head attention mechanism and relative positional encoding, solving the long-range dependency problem of sequence models such as RNN and LSTM caused by excessively long sequences [12], and improving the training efficiency of the model. Currently, it has been widely used in natural language processing. Since the self-attention mechanism can effectively capture the dependency between different positions in the sequence and ensure the long-term consistency of the generated sequence structure, research on Transformer-based music generation is the most extensive in the field of symbolic music generation. In 2018, Huang et al. [13] first applied Transformer to the field of symbolic music generation and proposed the symbolic music generation model Music Transformer.

In 2023, Sarmento et al. [14] proposed a Transformer-based music generation model GTR-CTRL, which is specifically designed for guitar music creation. GTR-CTRL uses information in guitar scores, including not only rhythm and pitch but also performance techniques and chord instrument fingerings. During the generation process, it controls the generation process by adjusting specific instrument configurations (inst-CTRL) and music genres (genre-CTRL). In addition, a BERT model for music genre classification is proposed to evaluate whether the generated music conforms to the expected genre style. Experimental results show that compared with unconditional generation models, the GTR-CTRL model is more flexible and controllable in providing guitar-centered symbolic music generation.With the diversification of music generation tasks, the quality of generated music has gradually improved. Therefore, the current research direction of music generation should not be limited to improving music quality, but should focus on human needs to conduct research on controllable music generation. As an important branch of the field of music information retrieval, emotional music generation has received extensive research attention from scholars at home and abroad in recent years. In 2018, Madhok et al. [15] studied the generation of corresponding music based on emotions predicted from facial expressions and proposed a framework named SentiMozart. This framework consists of a CNN-based image classification model and an LSTM-based music generation model. In 2021, Zheng et al. [16] proposed the EmotionBox emotional music generation system, which uses RNN as the core model and does not require additional datasets with specific emotional labels. It uses pitch histogram and note density as features (representing mode and rhythm respectively) to control music emotion. In 2022, Bao et al. [17] proposed an automatic and annotation-free method to construct a lyrics-melody dataset with emotional labels, where each sample includes lyrics, melody, and emotional labels.

3. Modeling Algorithms

3.1. Facial Feature and Physiological Signal Feature Extraction Models

In the research on human emotion recognition, facial expressions, as important external manifestations of emotional expression, have attracted much attention for a long time. Facial landmark detection involves inputting a facial image into a landmark detection model, which outputs a set of landmark coordinates with specific semantics. Guo et al. proposed the PFLD (A Practical Facial Landmark Detector) facial landmark detection model. To address the problem of unbalanced data of different categories in the dataset, a new loss function is designed to increase the penalty for small samples, and multi-scale fusion is used to expand the receptive field and accurately locate facial landmarks. This study selects a feature fusion-improved PFLD model for facial landmark detection, and its structure is shown in Figure 1.

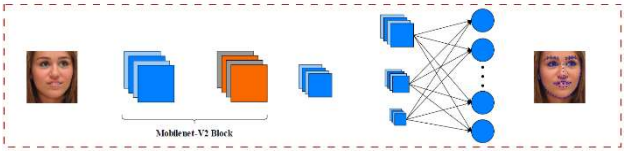

Figure 1. Structure of the Feature Fusion-Improved PFLD Facial Landmark Detection Model

The fusion of large-scale and small-scale features using multi-scale feature fusion technology can improve the accuracy of facial landmark detection. Therefore, we propose a Feature Module to replace the ordinary convolution operation. The Feature Module consists of three parts: the first part is an ordinary convolution operation, which generates condensed features of the input feature map; the second part is a grouped convolution operation, which generates multi-scale feature maps; the third part splices the condensed feature map generated in the first part with the multi-scale feature maps generated in the second part. Replacing the ordinary convolution operation with the Feature Module and integrating it into the existing neural network structure can further extract and fuse multi-scale features, thereby improving detection accuracy. The detailed principle of the Feature Module is shown in Figure 2.

Since EEG electrodes conduct electrical signals in contact with the subject's scalp, and EEG signals are extremely weak and easily interfered by other factors, the extracted EEG features will contain some outliers. Although some noise is removed through band-pass filtering and other methods in the data preprocessing stage, the EEG features after feature extraction still contain a lot of information irrelevant to the fatigue state. Therefore, feature smoothing can not only reduce the impact of irrelevant features but also make the performance of features more stable. In this study, the Linear Dynamic System (LDS) averaging method is used to process the extracted EEG signal features.

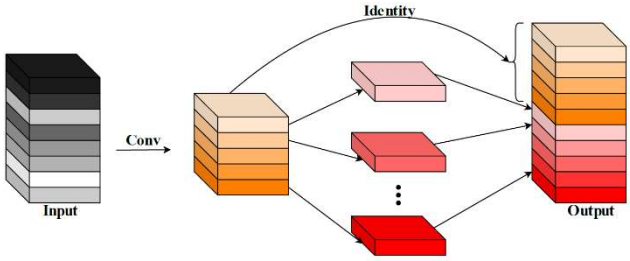

Figure 2. Structure of the Feature Module

A resting potential exists between the retinal nerve tissues of the human eye, with the cornea being positively charged and the retinal pigment epithelium being negatively charged. This resting potential forms an electric field around the eye, which is called the electrooculogram (EOG) signal. The EOG signal is the potential difference between the cornea and the retina, and corresponding electrical signals are generated when the eyes move. This signal has the characteristics of simple waveform, high signal-to-noise ratio, and easy collection. Usually, only a few electrodes are needed to collect various EOG signals, and it has been widely used in the field of medical health. In this study, continuous wavelet transform and peak detection algorithm are used to extract features related to blinking, fixation, and saccade from the separated EOG signals.

Different from image data, EEG signals and EOG signals are one-dimensional time-series data, so it is more reasonable to use a 1D Convolutional Neural Network (1DCCNN) for fatigue detection. The detailed structure of the 1DCCNN detection model is shown in Figure 3, where m1 is a one-dimensional feature vector, m2, m3, and m4 are feature maps after convolution, m5 is a fully connected output feature vector, and m6 is the category of fatigue state output by

softmax. Unlike traditional CNN models, it does not contain any pooling layers, so as to fully extract features from the input signals. The convolutional layers and fully connected layers learn hierarchical features from low to high from the input signal features. The high-level features with semantic expression are used as input to the classification layer, which completes the recognition and classification of features.

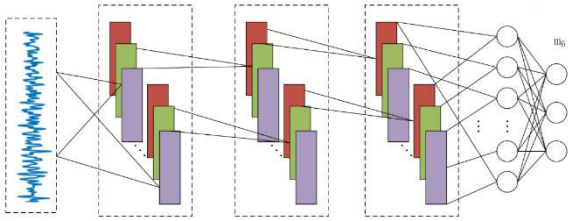

Figure 3. Structure of the 1DCCNN Detection Model

This study proposes a multimodal feature fusion-based sentiment analysis model, which combines a convolutional encoder network and a recurrent neural network. The overall system structure is shown in Figure 4.

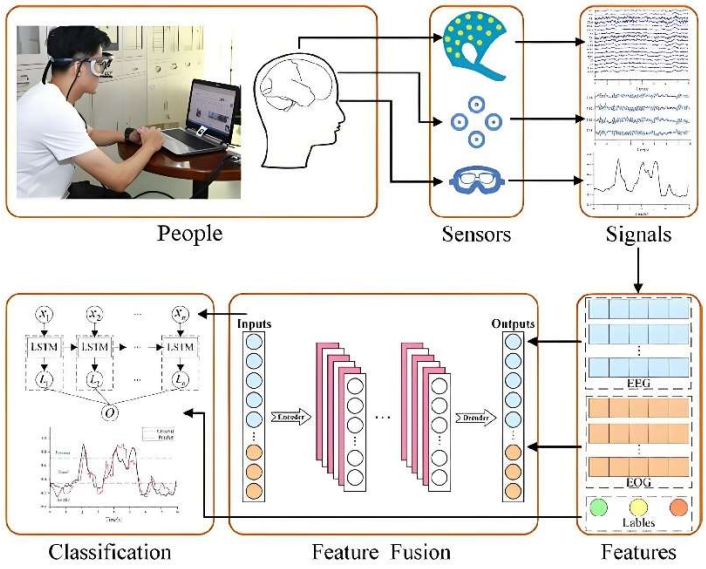

Figure 4. Structure of the Multimodal Feature Fusion-Based Sentiment Analysis Model

After extracting the aforementioned facial visual features, EEG signal features, and EOG signal features, the multimodal features are fused. The multimodal fatigue feature X is input into the constructed network model. The encoder part fuses and reduces the dimension of the features, and the encoder output features are subjected to dimension transformation to obtain deep fatigue features. Then, the decoder part realizes the reconstruction of the input vector Y. The network is trained until convergence using the input feature vector and the reconstructed feature vector. Finally, LSTM regression classification is performed. The input features of the LSTM layer are the hidden layer feature sequence in the CAE model. A batch normalization layer and a LeakyReLU activation layer are set after the LSTM layer, followed by a fully connected network and an activation function for classification, and the output result is the predicted label.

3.2. Symbolic Music Generation Model: Discrete-Emo Transformer

Symbolic music representation methods can convert a piece of MIDI music into a character sequence based on note events. Neural network models process these note sequences to

complete the symbolic music generation task. However, due to the limitations of computing resources and defects in the design of the self-attention mechanism, Transformer-based symbolic music generation models cannot process long note sequences and struggle to learn long-range dependencies between sequences, resulting in low quality of generated music that cannot meet human creative needs.

To generate high-quality emotional music, this study proposes the Discrete-Emo Transformer model for emotional music generation. It uses the segment recurrence mechanism and relative positional encoding in Transformer-XL to enhance the model's ability to learn potential features of music and improve its capacity to learn long-range dependencies. At the same time, the Flash attention mechanism is used to reduce the memory overhead during model training and increase the model's inference speed, enabling the model to process longer music sequences. In terms of music emotion representation, the REMI-EMO emotional music representation method is used to model note features and emotional features. The model architecture is shown in Figure 5, and each module of the model will be introduced in detail below.

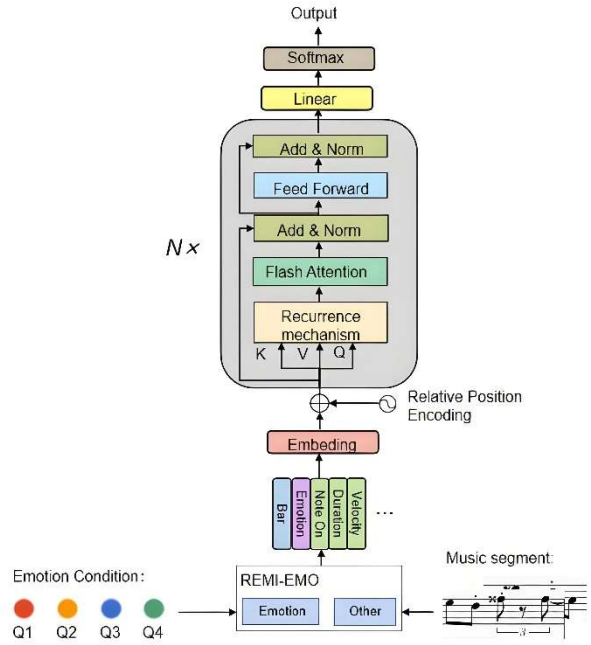

Figure 5. Architecture of the Discrete-Emo Transformer Model

The self-attention mechanism used in traditional Transformer models can capture the dependency between different positions in the sequence to a certain extent and improve the computing speed through parallel computing. However, the time and space complexity of the self-attention mechanism during operation is the square of the sequence length, which means that the model requires a large amount of computing resources to process long sequences. Therefore, researchers often choose to reduce the length of the input sequence, which makes it difficult for the model to capture long-range dependencies between sequences. For the symbolic music generation task, this leads to incoherent generated music, thus affecting the overall quality of the music.

The development of accelerator hardware currently focuses on enhancing computing capabilities rather than data transmission between memory and hardware, which leads to memory bottlenecks in attention operations and limits the ability of Transformer models to process long-sequence data. The Flash attention mechanism is a new type of attention algorithm that considers the read-write operations between different levels of GPU memory,

enabling attention computing to have IO awareness. It alleviates the problem of reduced model computing capabilities caused by insufficient memory, improves the model's ability to learn long-range dependencies, and thus achieves faster training and inference. Therefore, in the Discrete-Emo Transformer model, this study uses the Flash attention mechanism to replace the original self-attention mechanism, helping the model process long note sequences more efficiently.

The Flash attention mechanism loads the query (Q), key (K), and value (V) into SRAM once, adopts the fused attention mechanism operation, and finally writes them back to HBM once. Considering the small memory capacity of SRAM, the matrix in HBM needs to be partitioned to facilitate computing in SRAM. The entire computing process is shown in Figure 6. When computing attention, the Flash attention mechanism partitions the vectors Q, K, and V and performs multiple passes on the input blocks to execute the softmax function incrementally. At the same time, the Flash attention mechanism proposes not to use the intermediate attention matrix and reduces the HBM memory consumption by storing normalization factors.

To better process long music sequences, the Discrete-Emo Transformer replaces the traditional self-attention mechanism with the Flash attention mechanism. Compared with the Transformer-XL model, this model can use less memory to process longer music sequences, increase the coherence between generated sequences, and enable the model to better learn the potential features of music sequences.

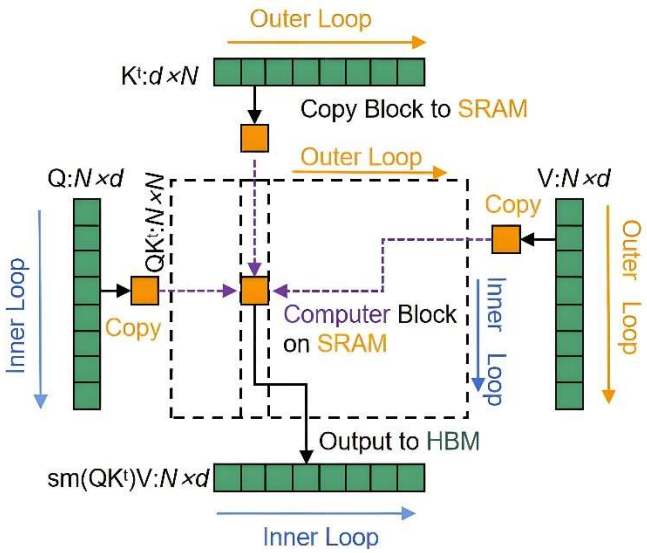

Figure 6. Attention Computing Process of the Flash Attention Mechanism

4. Experimentation and Analysis

4.1. Experimental Setup

This study first constructed a multimodal sentiment analysis dataset. The stimulus source for the experiment was 30 pieces of music of different types, each with a duration of 90 seconds. The subjects listened to these 30 pieces of music with different emotional types, and their facial videos, EEG signals, and ECG signals were collected simultaneously during the listening process. For the emotional music generation model, this study first pre-trained the model on the AILabs dataset, which contains 1748 pieces of popular piano solo music but no emotional labels. Compared with the EMOPIA dataset, the music segments in the AILabs dataset are longer. Therefore, to control the length of the input sequence and balance the model training efficiency and training resources, the length of the input sequence was set to 1024, and the REMI-EMO symbolic music representation method was adopted to represent the music sequence. However,

the EMOTION event was set to 0 to indicate unconditional generation. The pre-training was terminated when the negative log-likelihood loss reached 0.3. Then, fine-tuning was performed on the EMOPIA emotional music dataset, and emotional events were added. The value of the EMOTION event was set to the emotional category (Q1-Q4) of the corresponding sample. The training strategy combining pre-training and fine-tuning can solve the problem of small data volume of the emotional dataset, improve the generalization ability of the model, and help the model generate high-quality emotional music.

The experiment built a deep learning framework based on PyTorch 1.12.1, with Windows 10 as the operating system. The following hyperparameters were set during the simulation to train the algorithm model: batch size = 128, number of training iterations = 100, loss function = MSE, and Adam optimizer was used for gradient update. The hardware environment for the entire simulation included an AMD Ryzen 7-4800H processor, 16G memory, and two NVIDIA RTX 4090 Ti GPUs to accelerate the operation speed. The convolutional neural network was fine-tuned, with a learning rate of 1E-4 for the convolutional neural network and 1E-4 for the recurrent neural network during training. If the evaluation index did not improve for 5 consecutive epochs, the learning rate was reduced to 0.8 times the original value, and the training was terminated after 25 epochs, which could effectively avoid gradient explosion.

4.2. Experimental Results

The proposed multimodal sentiment analysis model was compared with existing mainstream models: FDMER, MISA, and MuIT. The results are shown in Table 1. The results indicate that the detection accuracy of the proposed multimodal sentiment analysis model is higher than that of other existing models. The proposed model fuses facial images, EEG, and EOG features through a convolutional encoder and then inputs them into the RNN network for recognition and classification, which proves its effectiveness compared with existing algorithm models.

Table 1. Comparison of Experimental Results

Model	**Accuracy（%）**
FDMER	95.3
MISA	96.1
MuIT	96.8
Ours	**97.5**

Researchers have proposed a large number of objective evaluation indicators to judge the quality of generated music. To comprehensively evaluate the generated music, this study objectively evaluated the generated music from four aspects: (1) Perplexity (PPL): PPL is a common indicator for evaluating the performance of language models, which reflects the degree of matching between the content generated by the model and the training set. An excellent language model can generate content similar to that in the dataset. (2) Unique Pitch Classes (UPC): UPC represents the number of different pitches used in one measure, which can reflect the pitch diversity of the generated samples. (3) Grooving Consistence (GC): GC refers to the similarity in rhythm between adjacent music segments. A higher GC value means the music sounds more smooth and powerful, which can be used to evaluate the rhythm consistency of the music. (4) Empty-Beat Rate (EBR): EBR represents the proportion of empty beats (i.e., beats with no notes played) in the total number of beats of a piece of music. A higher EBR means looser rhythm of the music, and vice versa. EBR can reflect the rhythm changes of the music to a certain extent. The comparison results of the music generated by different models in terms of objective indicators are shown in Table 2.

Table 2. Evaluation Results of Music Quality

Model	PPL	UPC	GC	EBR
mLSTM	2.24	9.31	0.64	0.22
Pop Music Transformer	1.75	9.25	0.84	0.17
Compound Word Transformer	1.76	9.34	0.82	0.17
Discrete-Emo Transformer (Ours)	**1.71**	**8.79**	**0.81**	**0.14**

As can be seen from the results in the table, in terms of the PPL indicator, the scores of Discrete-Emo Transformer, Pop Music Transformer, and Compound Word Transformer are all higher than that of mLSTM, indicating that Transformer-based music generation models can better memorize the long-range dependencies between sequences when processing long music sequences. Moreover, the Discrete-Emo Transformer has the lowest score, indicating that the music generated by this model is most consistent with the music distribution in the real dataset and has higher authenticity. In addition, in terms of the three indicators (UPC, GC, and EBR), the scores of the Discrete-Emo Transformer are close to those of the Pop Music Transformer and Compound Word Transformer but are all higher than those of the mLSTM model, which indicates the rationality of the proposed model architecture. Furthermore, the scores of the proposed model in the UPC and EBR indicators are closest to those of the dataset and have a large gap with other models, which shows that the music generated by the model has good pitch diversity, rhythm consistency, and structural stability, further proving that the music generated by the model is more similar to human creation.

This study mainly focuses on generating emotional music using discrete emotional values. Therefore, the music generated under the model's conditional generation needs to correctly reflect the given music emotion. Therefore, a classifier was used to quantify the emotional accuracy of the music generated by the model. To facilitate the calculation of the emotional accuracy of the music generated by the model, each model generated 100 pieces of music for different emotional categories, and the proportion of correctly predicted music samples in the total samples was calculated to obtain the emotional accuracy of the model. The emotional accuracy results of different models are shown in Table 3. The emotional accuracy of the music generated by the Discrete-Emo Transformer model is more than 10 percentage points higher than that of the other two models, and the emotional expression accuracy is greatly improved. This indicates that the emotional music generated by this model can better reflect the target emotion compared with the baseline models, and it has a stronger ability to learn emotional conditions and can learn more potential features related to music emotion.

Table 3. Evaluation Results of Music Emotional Accuracy

Model	Emotional Accuracy
mLSTM	41.5%
Pop Music Transformer	53.8%
Compound Word Transformer	61.7%
Discrete-Emo Transformer (Ours)	**78.4%**

Piano roll diagrams were used to visualize the music generated by the model, intuitively showing the duration, pitch, position, and other information of the notes. Figure 7 shows the piano roll diagrams of the music generated by the model under different emotional conditions. The horizontal axis represents time, the vertical axis represents pitch, and each note is represented by a rectangle in the diagram, where the width of the rectangle represents the duration of the note. Analysis of the figure shows that as time steps progress, the model can

generate repeated music segments (as shown in the boxes in the figure), which indicates that the model can learn the repetitive structure of music and express the emotion of the music by repeatedly emphasizing specific segments of the music. The generated music has structural stability, which is similar to real music creation- the emotion of music must be expressed by repeatedly repeating music segments. The experimental results further confirm the authenticity of the generated music.

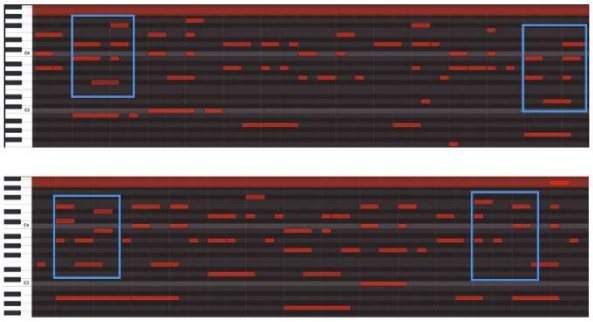

Figure 7. Visualization results

5. Conclusion

To solve the problems of loneliness, depression, and cognitive decline in elderly people in nursing homes, and to break through the limitations of insufficient accuracy of traditional manual intervention and the failure of single AI modality to form an intervention closed loop, this study focuses on multimodal sentiment analysis and personalized emotional music generation technologies, forming a solution with both technical innovation and practical applicability. The core innovations of the study are reflected in two aspects: First, a sentiment analysis model with multi-dimensional fusion is constructed. The accuracy of facial landmark detection is improved by an improved PFLD model; the LDS is used to smooth EEG signal noise; the 1DCCNN is used to adapt to the temporal features of EOG signals; then, multimodal data is fused through a convolutional encoder and input into the RNN for classification, solving the problem of poor robustness of single-modal recognition. Second, the Discrete-Emo Transformer music generation model is proposed. The Flash attention mechanism is introduced to optimize memory and speed; the Transformer-XL is integrated to enhance the learning of long-sequence dependencies; the REMI-EMO representation method is used to realize the simultaneous modeling of emotional and note features, breaking the bottleneck of insufficient personalization in traditional AI music generation. Experimental results verify the effectiveness of the solution: the multimodal sentiment analysis model achieves an accuracy of 97.5%, significantly outperforming mainstream models such as FDMER and MISA; the Discrete-Emo Transformer model has an emotional accuracy of 78.4% and a low Perplexity of 1.71, generating music that is more in line with the emotional needs of the elderly and human creative logic. This study provides a "emotion perception-personalized intervention" technical closed loop for nursing homes, helping to improve the mental health level of the elderly, while reducing the care burden and providing strong support for technology-empowered elderly care.

References

[1] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.

[2] Zhao Z, Liu Q. Former-dfer: Dynamic facial expression recognition transformer[C]//Proceedings of the 29th ACM International Conference on Multimedia. 2021: 1553-1561.

[3] Goshvarpour A, Abbasi A, Goshvarpour A. An accurate emotion recognition system using ECG and GSR signals and matching pursuit method[J]. Biomedical Journal, 2017, 40(6): 355-368.

[4] Kim K H, Bang S W, Kim S R. Emotion recognition system using short-term monitoring of physiological signals[J]. Medical and Biological Engineering and Computing, 2004, 42: 419-427.

[5] Wang Z, Wang Y. Emotion recognition based on multimodal physiological electrical signals[J]. Frontiers in Neuroscience, 2025, 19: 1512799.

[6] 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.

[7] 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.

[8] Wang S, Qu J, Zhang Y, et al. Multimodal emotion recognition from EEG signals and facial expressions[J]. IEEE Access, 2023, 11: 33061-33068.

[9] Lian Y, Zhu M, Sun Z, et al. Emotion recognition based on EEG signals and face images[J]. Biomedical Signal Processing and Control, 2025, 103: 107462.

[10] Pan H, Pang Z, Wang Y, et al. A new image recognition and classification method combining transfer learning algorithm and mobilenet model for welding defects[J]. IEEE Access, 2020, 8: 119951-119960.

[11] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30.

[12] Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult[J]. IEEE Transactions on Neural Networks, 1994, 5(2): 157-166.

[13] Huang C Z A, Vaswani A, Uszkoreit J, et al. Music transformer: Generating music with long-term structure[C]//Proceedings of the International Conference on Learning Representations. 2018: 1-14.

[14] Sarmento P, Kumar A, Chen Y H, et al. GTR-CTRL: Instrument and genre conditioning for guitar-focused music generation with transformers[C]//International Conference on Computational Intelligence in Music, Sound, Art and Design (Part of EvoStar). Cham: Springer Nature Switzerland, 2023: 260-275.

[15] Madhok R, Goel S, Garg S. SentiMozart: Music Generation based on Emotions[C]//Proceedings of the International Conference on Agents and Artificial Intelligence (ICAART). 2018, 2: 501-506.

[16] Zheng K, Meng R, Zheng C, et al. EmotionBox: A music-element-driven emotional music generation system using Recurrent Neural Network[J]. arXiv preprint arXiv:2112.08561, 2021.

[17] Bao C, Sun Q. Generating music with emotions[J]. IEEE Transactions on Multimedia, 2022.