Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1474–1481, DOI 10.12694/scpe.v26i3.4274

THE INTERACTIVE SYSTEM OF MUSIC EMOTION RECOGNITION BASED ON DEEP LEARNING

JING CHEN*

Abstract. This paper proposes an interactive system based on spectrogram analysis, deep neural network and wavelet analysis aiming at the complexity and subjectivity of music emotion recognition. The system first uses a spectrogram to capture the time-frequency characteristics of music signals and then automatically extracts the deep emotion-related features through a convolutional neural network (CNN). This paper introduces the Mallat algorithm for wavelet decomposition to enhance the local details of audio signals to improve the accuracy of feature extraction. The experimental results show that the system performs well in recognizing music emotions, and the accuracy is significantly improved compared with the traditional method. In addition, the system supports real-time interaction, allowing users to personalize music experience by adjusting emotional labels, thus showing broad application prospects in music therapy, game entertainment and other fields. This study promotes the development of music emotion recognition technology and provides a new perspective for further exploration of deep learning in interdisciplinary applications.

Key words: Music emotion recognition; Deep learning; Spectrogram; Neural network; Wavelet analysis; Mallat algorithm

1. Introduction.

Music, as a universal and profound expression of emotions, has been closely linked to human emotions since ancient times. With the rapid development of artificial intelligence technology, especially the successful application of deep learning in image and speech recognition, music emotion recognition (MER) technology has gradually become the research focus of academia and industry. MER is designed to automatically identify and classify the emotional colors contained in musical works by analyzing musical signals, and this technological breakthrough will revolutionize music recommendation, psychotherapy, game design and many other fields.

In the music emotion recognition research process, scholars have tried many methods. Literature [1] proposes the MER method based on traditional machine learning, which solves the emotion classification problem in early music by manually extracting musical features, such as rhythm, melody, and harmony. However, this approach relies on expert knowledge and experience and is difficult to capture the nuances of musical emotion. Subsequently, literature [2] introduced deep learning technology, especially convolutional neural network (CNN). However, the traditional CNN has limitations in processing time-frequency domain information. Literature [3] uses a spectrogram as input to convert music signals into two-dimensional images so that CNN can better understand the time-frequency structure of music to overcome this problem. However, the spectrogram is insufficient to preserve the musical signal's details. Therefore, literature [4] combined wavelet analysis technology, used the Mallat algorithm to conduct multi-scale decomposition of music signals and extracted more abundant wavelet coefficient features. This feature shows superiority in describing the dynamic change of musical emotion. However, strategies to effectively combine wavelet analysis with deep learning still need further exploration.

This paper aims to study an interactive music emotion recognition system based on deep learning, which combines spectrogram, deep neural network, and wavelet analysis to realize more accurate music emotion recognition [5]. First, this paper will discuss how to use the spectrogram as the input of CNN to capture the time-frequency characteristics of music signals. Secondly, this paper will introduce the Mallat algorithm for wavelet decomposition to extract multi-scale features of music signals and combine them with the feature extraction layer of CNN to enhance the emotion recognition ability of the system. In addition, this paper

*Public Art Teaching Department, Zhengzhou College of Finance and Economics, Zhengzhou 450000, China (Corresponding author, cccdd1232024@163.com)

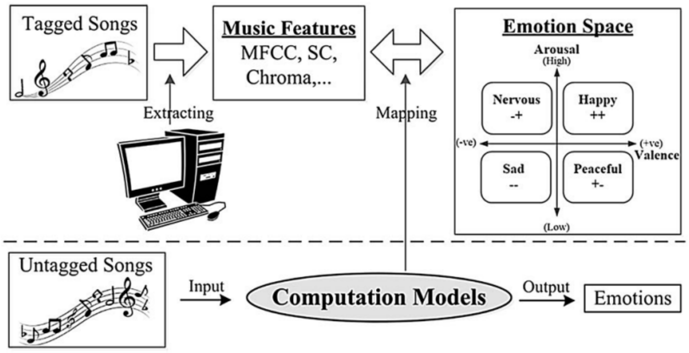

Fig. 2.1: Basic framework of music emotion recognition model.

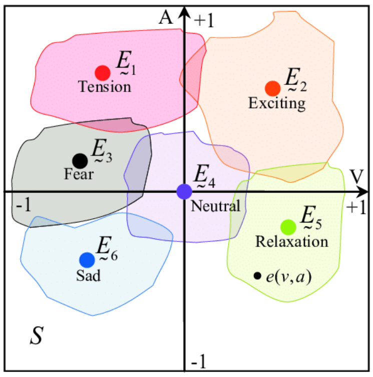

Fig. 2.2: V-A emotional space diagram.

will also study how to design a real-time interactive interface so that users can participate in the emotion recognition process and constantly optimize the system's performance through a feedback mechanism [6]. In the experimental part, this paper will collect various music data sets, including music works of different styles and emotional tendencies, to verify the generalization ability and accuracy of the system [7].

2. Deep learning neural network model.

2.1. Model Framework. This project intends to build a song emotion recognition model based on deep neural networks and machine learning technology. Figure 2.1 shows an infrastructure diagram of this pattern.

Firstly, the music library containing different emotional markers is divided into two parts: the first part is to preprocess the original score, the second part is to extract the corresponding emotional markers, and the last part is to establish the classification model with the corresponding emotional markers.

2.2. Emotional model. This paper uses Russell's Valence-Arousal model [8]. In short, effectiveness reflects two levels of emotion: positive and negative. The higher the value, the higher the positive level of emotion and the opposite negative level. The Arousal of the subject reflected the intensity of emotion. Arousal value was high, emotional intensity was high, and arousal intensity was low. The V-A emotional pattern is shown in Figure 2.2. This article will put V - A two-dimensional space-time transformation into (+ V + A), (V + A), (-v - A) (+ V - A) and so on, four different types of emotion. The corresponding results for the four types of musical emotions are given in Table 2.1.

Table 2.1: Music emotion category table.

Category	Emotion	V-A value
Emotion of the first kind	Happy	+V+A
Emotion of the second kind	Anxiety	-V+A
Emotion of the third kind	Mawkish	-V-A
Fourth emotion	Relax	+V-A

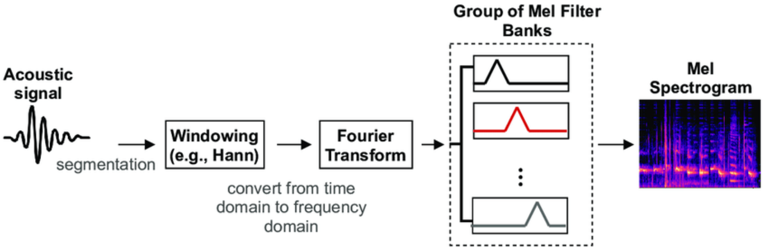

Fig. 2.3: Process of generating spectrogram.

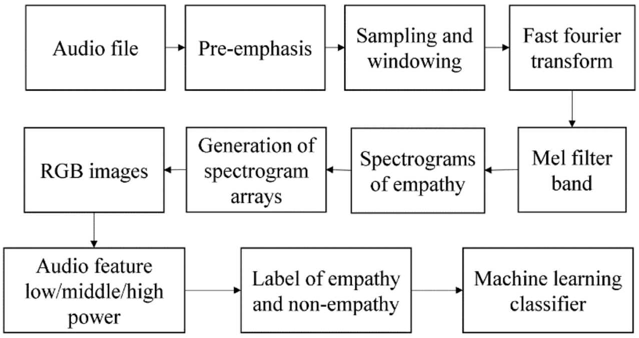
Audio file
Pre-emphasis
Sampling and windowing
Fast fourier transform
Mel filter band
Spectrograms of empathy
Generation of spectrogram arrays
RGB images
Audio feature low/middle/high power
Label of empathy and non-empathy
Machine learning classifier

Fig. 2.4: Schematic diagram of music signal generation.

2.3. Spectrogram. The spectrum is a graph obtained after Fourier analysis in the time domain. It is a two-dimensional time-frequency graph used to characterize the spectrum change horizontally and vertically. It is time horizontally and frequency vertically. The spectrum contains rich spectrum characteristics. It includes formant, energy and other frequency domain parameters and has both time and frequency domain characteristics [9]. The graph contains the entire spectrum that has not been processed, so the information about the music in the graph is not destroyed. The generation process of the spectrogram is shown in Figure 2.3.

A frame window-adding, short-time Fourier transform is performed to convert the time domain information into the frequency domain to generate the graph, and then the scale is converted into the decibel value expression of the amplitude [10]. Then, this frequency domain information is segmented and connected according to the time series to obtain the graph (Figure 2.4).

In this paper, the hearing characteristics of the human ear are taken as the primary frequency band, so the spectrum mentioned in this paper is the spectrum of the Mayer frequency band [11]. The graph takes time as the horizontal axis, Meir frequency as the vertical axis, and data energy of music signal as the coordinate. Because it is carried out in the 2D plane, its energy is represented by color, and the stronger the color, the higher the intensity of its sound.

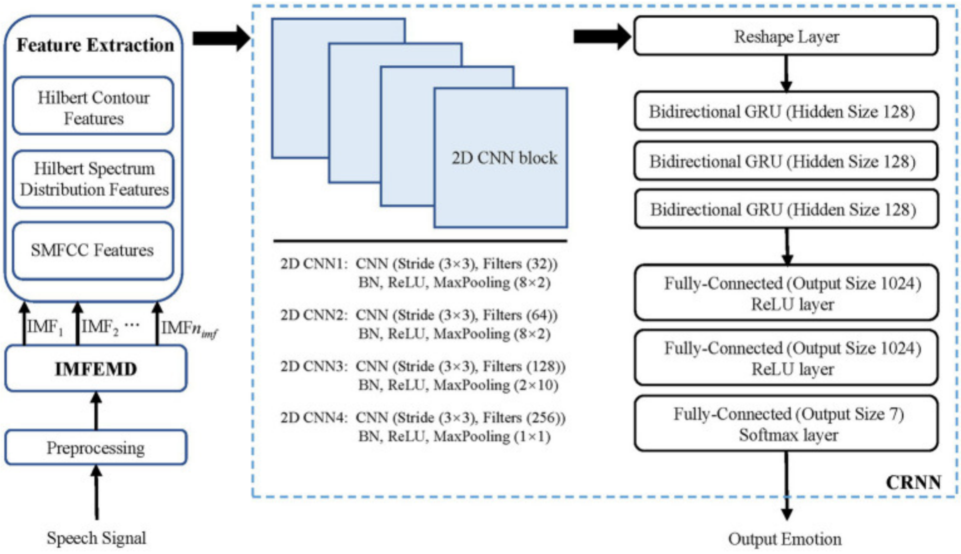
Feature Extraction
Hilbert Contour Features
Hilbert Spectrum Distribution Features
SMFCC Features
IMF$_1$
IMF$_2$ ···
IMFn_{imf}
IMFEMD
Preprocessing
Speech Signal
2D CNN block
2D CNN1: CNN (Stride (3×3), Filters (32)) BN, ReLU, MaxPooling (8×2)
2D CNN2: CNN (Stride (3×3), Filters (64)) BN, ReLU, MaxPooling (8×2)
2D CNN3: CNN (Stride (3×3), Filters (128)) BN, ReLU, MaxPooling (2×10)
2D CNN4: CNN (Stride (3×3), Filters (256)) BN, ReLU, MaxPooling (1×1)
Reshape Layer
Bidirectional GRU (Hidden Size 128)
Bidirectional GRU (Hidden Size 128)
Bidirectional GRU (Hidden Size 128)
Fully-Connected (Output Size 1024) ReLU layer
Fully-Connected (Output Size 1024) ReLU layer
Fully-Connected (Output Size 7) Softmax layer
CRNN
Output Emotion

Fig. 2.5: Structure of CRNN music emotion recognition model.

2.4. CRNN model based on deep neural network. In this project, CNN was used to obtain the time series features of the atlas, and based on preserving the time series features of the atlas, the feature maps of the time series with complete features were obtained [12]. This gives a complete time series feature. The method takes speech as the essential information and CRNN as the learning object. It was using CRNN to learn its features, and to realize the end-to-end learning of music mood. The structure diagram of the CRNN is shown in Figure 2.5(image quoted in Speech emotion recognition based on improved masking EMD and convolutional recurrent neural network).

The input to the network is the musical notation. In the substructure of the convolutional network, the advantages of the CNN network in 2D data are given full play, and the 1*15* N spectrum characteristic diagram is obtained by extracting spectrum information and maintaining the time series characteristics of the spectrum [13]. The core of this method is the convolutional pool processing of convolutional neural networks. By optimizing the convolution kernel, step size, layer number, etc., the frequency domain dimension of the obtained feature graph is reduced to 1. In this way, the signal's frequency domain and time characteristics are effectively fused [14]. It considers the feature extraction of spectrogram as an image Angle and the feature extraction of music signal time series Angle.

3. Feature information extraction and classification methods. This project takes the Simplified A-V emotional model as the research object and selects different types of emotions from four emotional modes (intense, happy, low, and soft). The intensity and speed of the music are very high in the intense areas. The music is more intense and faster in the happy zone [15]. The song is less intense in the soft area, and the tempo is slower. The music is less intense in the low zone, and the tempo is slower.

3.1. Feature Information.

3.1.1. Strength. The audience's grasp of the strength of the music is usually judged by the pitch and beat speed [16]. A physical quantity called the average energy is defined to quantify the intensity of music. The formula is:

$$E_t = \sum_{i=t*M}^{(t+1)*M} \frac{u_i}{M}; j, t = 0, 1, 2 \cdots \tag{3.1}$$

E_t is the short-term average energy of segment $t.u_i$ is the j pieces of music data collected, and M is the number of music data collected for each segment.

3.1.2. Rhythm. Strong and happy music usually has a faster speed, while low and soft music has a slower speed. The relative prosody method is used to replace the complicated prosody formula. 3.2 Classification Algorithm. Considering that the music emotion characteristic information is generally selected from the two aspects of high frequency and low frequency, this paper adopts the real-time method of wavelet analysis -The Mallat method

$$b_m[n] = \sum_t l[t-2n] b_{m+1}[t] \tag{3.2}$$

$$s_m[n] = \sum_t f[t-2n] b_{m+1}[t] \tag{3.3}$$

$l[t], f[t]$ is the signal string of the pulse response and the signal of the highpass signal. The wavelet analysis method transforms the signal by discrete Fourier transform, and the amplitude in the frequency domain is obtained. λ is used to represent the base frequency, and the following formula is obtained:

$$B(\lambda) = \sum_n b(n) \exp(-j\lambda n) \tag{3.4}$$

The wavelet analysis effectively identifies the music fragments with different simultaneous frequency characteristics. The identification of genetic information is combined with the identification of sound, which significantly improves detection efficiency. Wavelet transform is used to extract the feature of the spectrum table, extract the spectrum segment with the highest amplitude, and then the pronunciation time of the adjacent spectrum segments is timed. The duration of the large and small amplitude segments is found by comparing the adjacent spectrum segments to realize the rough spectrum recognition of the spectrum segments.

Figure 3.1 shows the contrast items' frequency-amplitude graph for each mixed tone. B_1 is its magnitude. Where y_2, y_3, y_4 is the triad tone contrast term with y_2, y_3, y_4 in each triad component, and the corresponding amplitude is B_2, B_3, B_4. y_5 is the contrasting item of tone, and its magnitude is B_5. The portion with a lower amplitude is not marked and can be excluded when setting the selection threshold.

$W_t = \{w_{t1}, w_{t2}, \cdots, w_{tn}\}$ is used to represent the defined sequence, where w_{ti} represents the i comments contained in the t filtered comment items. If it's a single tone, then $i = 1$. If it's n then $i = n$. The sequence $E_{W_t} = \{E_{w11}, E_{w+2}, \cdots, E_{w+n}\}$ may be qualified by addition, while E_{wit} represents the intensity of i comments included by the t comment items being filtered, $t = 1, 2, \cdots, i = 1, 2, \cdots, n$. Set the comparison coefficient to z_t and calculate it with the following equation:

$$z_t = E_{t+1}/E_t; t = 0, 1, 2, \cdots \tag{3.5}$$

E_t represents the average value of item t recorded. This comparison can be single or juxtaposed. Its formula goes like this:

$$E_t = \overline{E_{W_t}} = \sum_{i=1}^{n} E_{wti}/n; i = 1, 2, \cdots, n, t = 1, 2, \cdots \tag{3.6}$$

In tone contrast, it's A single tone $i = n = 1$ when the mean is $E_t = E_{W_t} = E_{w+1}$. At this time, the change of the adjacent sound contrast term can be determined by the value of z_t. If the value of z_t is in the closed interval $[0.6, 1.4]$, its change can be regarded as a slight change in the same roughness region [17]. When the value of z_t exceeds this interval, it can be regarded as a jump in different roughness regions. However, this contrast leads to a common phenomenon:

$$\begin{aligned}
&z_1, z_2, \cdots, z_{t-1} \in [0.6, 1.4] \\
&z_t, z_{t+1}, \cdots, z_{t+n} \notin [0.6, 1.4] \\
&z_{t+n+1}, \cdots \in [0.6, 1.4]
\end{aligned}$$

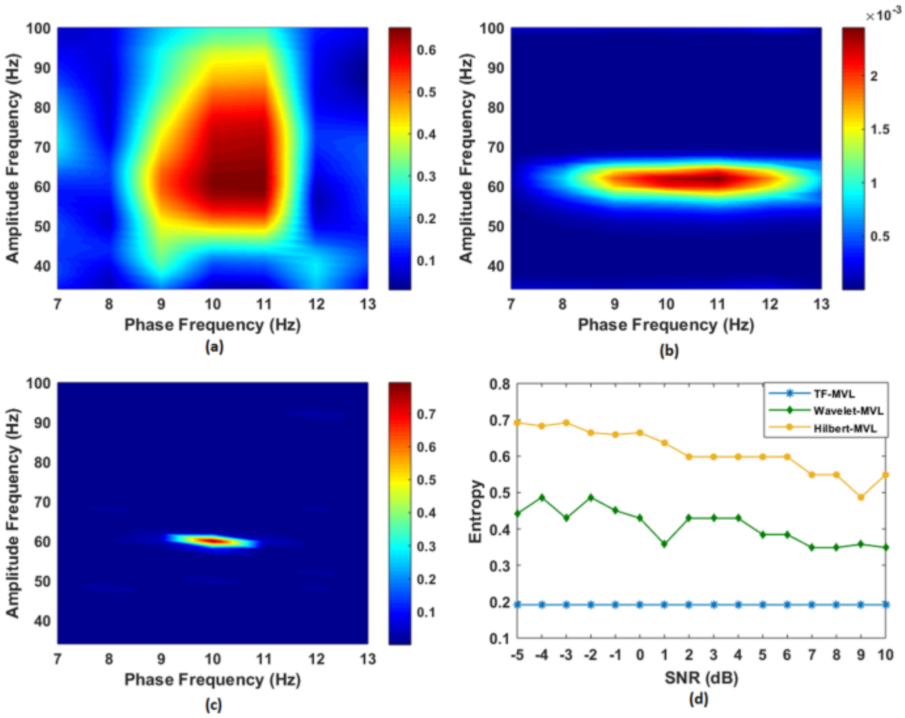

Fig. 3.1: Frequency - amplitude of the note comparison term in the mixed note bar.

The number of note comparators is M, so it only takes a simple operation to obtain a rough beat correlation value [18]. The metric-dependent value of the first paragraph is $\gamma_1 = M/t_1$. Similarly, if a song is divided into H parts, then the velocity correlation value of the segment l is

$$\gamma_l = M/t_l \tag{3.7}$$

A similar algorithm can be used for the new contrast coefficient z_t to overcome the limitation of rough partitioning based on average energy:

$$z_t = \gamma_{l+1}/\gamma_l; l = 0, 1, 2, \cdots \tag{3.8}$$

Similarly, if z_t is in a closed range [0.8,1.2], then its change can be regarded as a slight change in the same rough perception region. When z_t exceeds this interval, it can be regarded in this paper as a jump in a different roughness region.

4. Experimental results. Using the wavelet analysis software package of Matlab7.0, the rough emotional soft cutting test was carried out on the music fragments with different emotional components, which the author edited. The sampling rate is 12015 Hz. The sampling length was 50 seconds. The samples were labeled manually to determine the original emotion region [19]. In addition, a rough "soft cut" reference was made to the emotions in the test set by artificial perception in 20 researchers with good musical literacy. The results of the test are shown in Table 4.1.

Each test's maximum error time and minimum error time are 103 ms and 8 ms, respectively. The time-domain deviation of both the rough and real emotion fragments is within the acceptable range. The experiment shows that this soft-cutting technology can meet the precision requirement of the music lighting demonstration control system, and there is no apparent false connection phenomenon.

Table 4.1: Experimental results of soft cutting of music rough emotion.

Coarse affective domain	Quantity	Fall into this category	Correct number	Precision/%	Recall/%
fierce	21	23	19	85.21	94
Cheerful and cheerful	21	25	18	73.75	89
gentle	21	19	15	81.04	73
low	21	17	16	97.71	78

5. Conclusion. This paper presents an interactive system to address the challenge of music emotion recognition, which cleverly combines deep learning techniques with music signal processing. By using a spectrogram as the input of a deep neural network, the system can effectively capture the time-frequency characteristics of music, and the introduction of wavelet analysis and the Mallat algorithm further enhances the precision of feature extraction, especially in the processing of subtle changes in music emotion. The experimental results show that the designed system achieves high accuracy in the music emotion recognition task, proving the strong potential of deep learning in music emotion analysis. In addition, the interactive design of the system allows users to participate in the emotion recognition process, and the real-time feedback mechanism not only improves the user experience but also provides the possibility for continuous learning and optimization of the system.

REFERENCES

[1] Jingjing, W. A. N. G., & Ru, H. U. A. N. G. (2022). Music emotion recognition based on the broad and deep learning network. Journal of East China University of Science and Technology, 48(3), 373-380.

[2] Gómez-Cañón, J. S., Cano, E., Eerola, T., Herrera, P., Hu, X., Yang, Y. H., & Gómez, E. (2021). Music emotion recognition: Toward new, robust standards in personalized and context-sensitive applications. IEEE Signal Processing Magazine, 38(6), 106-114.

[3] Xu, L., Wen, X., Shi, J., Li, S., Xiao, Y., Wan, Q., & Qian, X. (2021). Effects of individual factors on perceived emotion and felt emotion of music: based on machine learning methods. Psychology of Music, 49(5), 1069-1087.

[4] Pandeya, Y. R., & Lee, J. (2021). Deep learning-based late fusion of multimodal information for emotion classification of music video. Multimedia Tools and Applications, 80(2), 2887-2905.

[5] Sarkar, R., Choudhury, S., Dutta, S., Roy, A., & Saha, S. K. (2020). Recognition of emotion in music based on deep convolutional neural network. Multimedia Tools and Applications, 79(1), 765-783.

[6] Abdullah, S. M. S. A., Ameen, S. Y. A., Sadeeq, M. A., & Zeebaree, S. (2021). Multimodal emotion recognition using deep learning. Journal of Applied Science and Technology Trends, 2(01), 73-79.

[7] Nawaz, R., Cheah, K. H., Nisar, H., & Yap, V. V. (2020). Comparison of different feature extraction methods for EEG-based emotion recognition. Biocybernetics and Biomedical Engineering, 40(3), 910-926.

[8] Koduru, A., Valiveti, H. B., & Budati, A. K. (2020). Feature extraction algorithms to improve the speech emotion recognition rate. International Journal of Speech Technology, 23(1), 45-55.

[9] Saxena, A., Khanna, A., & Gupta, D. (2020). Emotion recognition and detection methods: A comprehensive survey. Journal of Artificial Intelligence and Systems, 2(1), 53-79.

[10] Veltmeijer, E. A., Gerritsen, C., & Hindriks, K. V. (2021). Automatic emotion recognition for groups: a review. IEEE Transactions on Affective Computing, 14(1), 89-107.

[11] Medina, Y. O., Beltrán, J. R., & Baldassarri, S. (2022). Emotional classification of music using neural networks with the MediaEval dataset. Personal and Ubiquitous Computing, 26(4), 1237-1249.

[12] Zepf, S., Hernandez, J., Schmitt, A., Minker, W., & Picard, R. W. (2020). Driver emotion recognition for intelligent vehicles: A survey. ACM Computing Surveys (CSUR), 53(3), 1-30.

[13] Schlegel, K., Palese, T., Mast, M. S., Rammsayer, T. H., Hall, J. A., & Murphy, N. A. (2020). A meta-analysis of the relationship between emotion recognition ability and intelligence. Cognition and emotion, 34(2), 329-351.

[14] Xu, G., Guo, W., & Wang, Y. (2023). Subject-independent EEG emotion recognition with hybrid spatio-temporal GRU-Conv architecture. Medical & Biological Engineering & Computing, 61(1), 61-73.

[15] Liu, W., Qiu, J. L., Zheng, W. L., & Lu, B. L. (2021). Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. IEEE Transactions on Cognitive and Developmental Systems, 14(2), 715-729.

[16] Zhao, S., Jia, G., Yang, J., Ding, G., & Keutzer, K. (2021). Emotion recognition from multiple modalities: Fundamentals and methodologies. IEEE Signal Processing Magazine, 38(6), 59-73.

[17] Ding, Y., Robinson, N., Zhang, S., Zeng, Q., & Guan, C. (2022). Tsception: Capturing temporal dynamics and spatial asymmetry from EEG for emotion recognition. IEEE Transactions on Affective Computing, 14(3), 2238-2250.

[18] Kamble, K. S., & Sengupta, J. (2021). Ensemble machine learning-based affective computing for emotion recognition using dual-decomposed EEG signals. IEEE Sensors Journal, 22(3), 2496-2507.

[19] Panda, R., Malheiro, R., & Paiva, R. P. (2020). Audio features for music emotion recognition: a survey. IEEE Transactions on Affective Computing, 14(1), 68-88.

Edited by: Hailong Li
Special issue on: Deep Learning in Healthcare
Received: Jul 10, 2024
Accepted: Aug 17, 2024