

CLANet: A Hybrid Encoder for EEG Emotion Recognition via Spatiotemporal-Spectral Feature Fusion

Huiwen Chen^{1, a}, Hongyu Liu^{1, b}, and Naishi Feng^{1, c *}

¹ School of Intelligent Science and Information Engineering, Shenyang University, Shenyang 110044, China;

^a1084169481@qq.com, ^b3174833412@qq.com, ^cfengnaishi@syu.edu.cn

Abstract. Human emotion is a core link between cognition, behavior, and physiology, and its accurate recognition is crucial for advancing the development of intelligent human-computer interaction, mental health diagnosis, and other related fields. Current research mostly achieves multi-domain feature fusion through simple concatenation or weighted fusion at the algorithmic level, failing to fully reveal and exploit the feature contribution rates associated with emotional processing. To address the aforementioned issues, this study first conducts sufficient feature extraction across the temporal, frequency, and spatial domains, and then proposes a hybrid encoder model integrating 3D-CNN, LSTM, and an attention mechanism, named CLANet. This model can not only capture local dynamic patterns but also integrate global spatial configurations, thereby providing a novel approach for emotion recognition and improving recognition accuracy. Experiments conducted on the SEED IV dataset demonstrate that the proposed CLANet model achieves a test accuracy of 93.8%, outperforming state-of-the-art models such as support vector machines (SVMs), BiLSTM, Hierarchical LSTM, and EEGNet. Furthermore, the fusion of multi-domain features (temporal, frequency, and spatial) significantly enhances recognition performance, achieving a maximum accuracy of 94.0% in the θ band. This study provides a more physiologically relevant architecture for EEG-based emotion recognition and offers technical support for its practical applications in related fields.

Keywords: EEG; Emotion recognition; Machine learning; Multi-domain.

1. Introduction

Human emotion is a key driver of cognition and behavior, and its objective and accurate recognition is of profound significance in fields such as mental health assessment, neuropsychiatric disease diagnosis, intelligent human-computer interaction, and consumer neuroscience. Traditionally, emotion assessment has relied primarily on subjective scales or behavioral observations. These methods are susceptible to individual differences in expression and recall bias, making it difficult to capture the generation and changes of emotions in real time. With the development of brain imaging technologies, electroencephalogram (EEG), in particular, has become one of the core tools for exploring the neural mechanisms of emotion due to its millisecond-level high temporal resolution, non-invasiveness, and relatively low cost, and it has also emerged as an important research direction in the field of neuroscience.

EEG signals contain more complex and abundant neurodynamic information. They not only change dynamically over time but also form specific patterns in spatial distribution across the scalp; meanwhile, their rhythmic activities in different frequency bands are closely related to distinct emotional states. Therefore, the effective fusion of EEG features across the three dimensions of time, frequency, and space is the key to constructing high-precision emotion recognition models. Feature extraction is a critical step for emotion recognition based on EEG signals, and thus the selection of appropriate EEG features is essential to improving the accuracy of emotion recognition. Traditional emotion recognition methods mainly combine manual feature engineering with shallow classifiers. Manually designed features often fail to fully capture the complex and high-dimensional nonlinear patterns of EEG signals under specific emotional states, resulting in limited representation capability. Luo et al. constructed brain functional networks based on left and right hemispherical brain regions and combined sample entropy from the C3 and C4 channels as features, using a

support vector machine (SVM) model for classification [1]. Ai et al. utilized time-frequency-energy features to analyze the nonlinear dynamic characteristics of EEG signals and ultimately adopted the traditional SVM classifier for classification [2]. Chen et al. extracted features from brain functional networks based on graph theory and further classified these features using SVM, random forest, and convolutional neural network (CNN) models [3]. For existing related studies, on the one hand, most lack comprehensive multi-domain analysis of EEG signals, failing to fully exploit their inherent physiological characteristics; on the other hand, the sole reliance on traditional classifiers hinders the in-depth extraction of valid latent features in EEG signals, which restricts the improvement of recognition performance.

With the rise of deep learning, end-to-end models such as convolutional neural networks (CNNs) have been introduced into this field to enable automatic feature extraction and learning. Aldawsari et al. decomposed raw EEG signals and applied 1D-CNN for one-dimensional convolution [4]. Wang et al. proposed a self-supervised end-to-end classifier for emotion classification on public datasets, achieving better performance than traditional models [5]. Mahmoud employed 1×1 convolution to integrate multi-frequency band features, aiming to highlight the weights of high-contribution frequency bands, and stacked convolutional modules to implement end-to-end classification [6]. Beyond end-to-end models, some studies have adopted shallow networks for feature learning. For instance, Hu et al. performed spatial filtering on EEG signals using the common spatial pattern (CSP) to extract spatial information during the feature extraction phase and then combined CNN and bidirectional long short-term memory (BiLSTM) models for classification [7]; other studies have used graph convolutional networks for EEG signal classification [8]. In summary, most existing end-to-end models suffer from relatively simple architectures, failing to fully fuse the multi-dimensional information of EEG signals across time, frequency, and space, which makes it difficult to comprehensively characterize the neurodynamic mechanisms associated with emotions. Secondly, end-to-end models generally face the "black box" problem—they cannot quantify the contribution of each feature to the classification results. Although these models solve the classification task, they are unable to reveal the neurophysiological mechanisms underlying emotion generation and recognition.

To address the limitations of insufficient feature extraction and the failure of deep learning models to fully exploit features in the aforementioned methods, this paper simultaneously extracts spatiotemporal-frequency joint features and proposes a novel model framework. By leveraging the three-dimensional convolution operation of 3DCNN, the framework synchronously captures local correlations of EEG signals across the spatial, temporal, and frequency dimensions. Furthermore, we integrate LSTM and attention mechanisms to enhance the precision of temporal sequence feature extraction, making up for the shortcomings of static convolution in dynamic temporal modeling. This architecture provides a more physiologically relevant solution for EEG emotion recognition tasks and significantly improves recognition accuracy.

2. Material and Methods

2.1 Dataset

The experiments were conducted on the SEED IV dataset [9]. A total of 15 healthy subjects were recruited, including 7 males and 8 females, aged between 20 and 24 years old, all of whom were right-handed. The experimental procedure is illustrated in Fig. 1. On separate days, each participant completed 3 experimental sessions, with each session consisting of 24 individual trials.

In each trial, the participant was presented with a video clip, and a 62-channel EEG acquisition device was used to record their EEG signals at a sampling rate of 1000 Hz. The 24 emotional clips were designed to elicit four distinct emotional states, namely happiness, sadness, neutrality, and fear. The EEG recording duration for each video clip was 120 seconds, and each clip was labeled with its corresponding emotional category. All participants watched the video clips in the same fixed order.

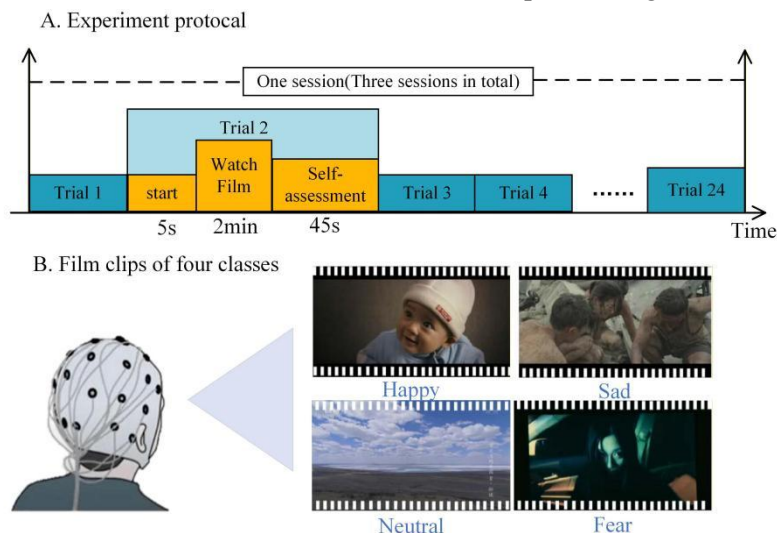


Fig. 1 Experiment Design.

The data format of the SEED IV dataset is presented in Table 1. The EEG signals in the dataset are structured in a three-dimensional array of dimensions $3 \times 24 \times 62$, corresponding to 3 experimental sessions, 24 emotion-inducing video clips, and 62 acquisition electrode channels, respectively. The duration of the EEG signal for a single video clip is 120 seconds, with an original sampling rate set at 1000 Hz, which can be downsampled to 500 Hz according to research requirements. The accompanying emotion labels are organized in a two-dimensional array of dimensions 3×24 .

Table 1. the signal format of SEED IV dataset

Name	Value	Description
EEG signal	$3 \times 24 \times 62$	session/video/channel
Duration/s	120	EEG signal length
Sampling rate/Hz	1000	Downsample to 500Hz
Label	24×4	Video/label

2.2 Data Preprocess

Relevant experiments were conducted in this study using the SEED-IV electroencephalogram dataset. First, the preset channel sequence table and the international 10-20 system standard EEG channel location table were imported to complete label mapping of the 62 EEG channels and match the corresponding three-dimensional spatial coordinates (X/Y/Z). Second, a third-order Butterworth filter was applied for multi-dimensional filtering: a 1 Hz high-pass filter was used to eliminate low-frequency baseline drift noise; a 50 Hz low-pass filter was adopted to remove high-frequency electromagnetic interference; meanwhile, a notch filter band of 48–52 Hz was set to suppress the 50 Hz power-line interference and its harmonic components. The filtered EEG signals were decomposed using extended independent component analysis (ICA) to obtain 62 mutually independent signal components. The ICLabel automatic artifact identification algorithm was combined to classify and identify each independent component [10], quantify the artifact proportion of each component, set a threshold of 50% for artifact proportion, label and discard independent components with excessive artifacts such as electrooculogram (EOG) and electromyogram (EMG), and reconstruct the pure EEG signals. Finally, a re-referencing operation was performed, with the reference electrodes set as T8 and T7.

2.3 Feature Extraction

In this study, a strategy combining multi-band signal processing and multi-domain feature fusion was adopted for feature extraction of the preprocessed EEG data. First, the preprocessed signals were filtered into five frequency bands, namely the δ band (1–4 Hz), θ band (4–7 Hz), α band (7–13 Hz), β band (13–30 Hz), and γ band (31–50 Hz), and subsequent feature extraction was performed separately for each of these five bands. The signals of each frequency band and channel were segmented using a sliding time window, with the window length set to 2 s and the overlap rate between adjacent windows set to 50%. For each extracted time window, features were extracted from three dimensions: the time domain, frequency domain, and nonlinear dynamics.

First, nonlinear dynamic features were extracted in the time domain, including Sample Entropy and Differential Entropy. In the spatial domain, this study extracted the Phase Lag Index (PLI), and five features were further derived based on the constructed network, namely the Node Clustering Coefficient, Path Distance, Node Global Efficiency, Node Betweenness, and Edge Betweenness. The seven features were fused for each frequency band, channel, and time window to obtain the input data of the model, and the specific process is illustrated in Fig. 2.

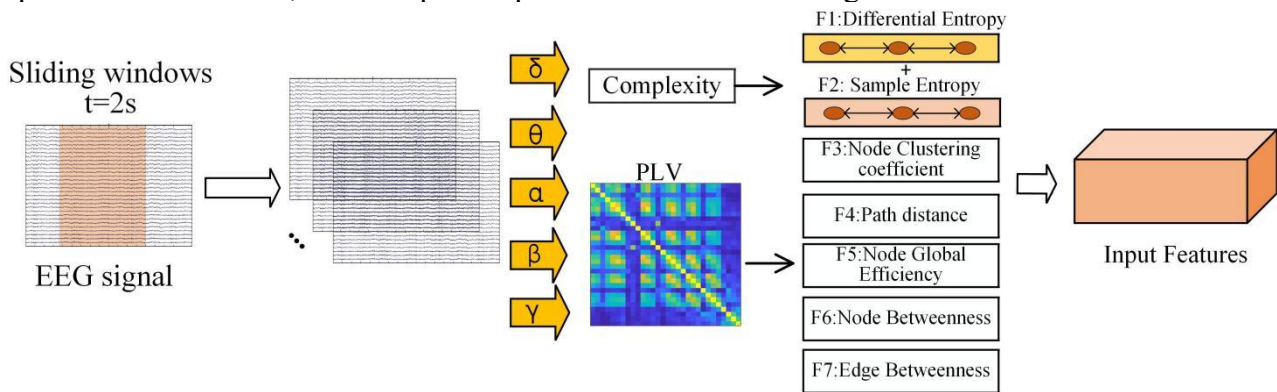


Fig.2 Process of feature extraction

2.4 Classification Model design

To fully exploit features across all domains, this study integrates 3DCNN, LSTM, and the attention mechanism to construct the CLANet (Convolution-LSTM-Attention Network). As a hybrid architecture designed specifically for electroencephalogram (EEG) signal analysis, this model enables in-depth mining of multi-dimensional features of EEG signals. It extracts local correlation features of EEG signals across the spatial, temporal, and spectral dimensions via 3DCNN, captures the long-term dynamic dependencies of signals in the time series using LSTM, and adaptively focuses on the key brain regions and time-frequency components most relevant to emotion recognition with the aid of the attention mechanism. Finally, the fully connected layer performs nonlinear mapping and classification decision-making on the fused high-level features, thereby realizing multi-dimensional joint feature extraction of EEG signals and efficient modeling of emotional states. The structure of the model is illustrated in Fig. 3.

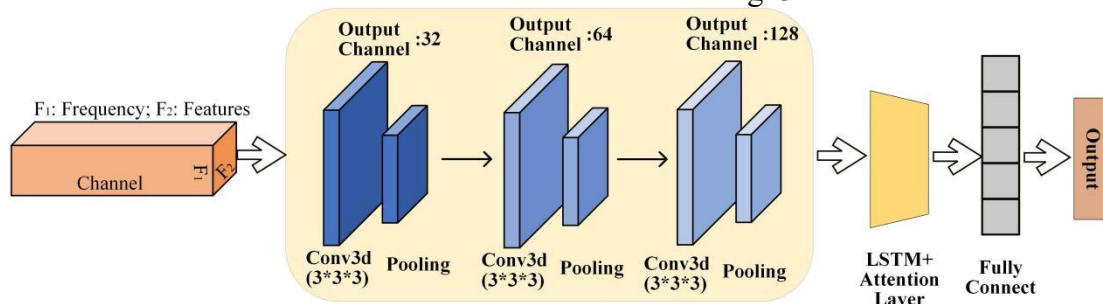


Fig.3. Model framework

First, Conv3d performs sliding convolution operations on the input features using a 3D convolutional kernel (with dimensions corresponding to spatial channels, temporal windows, and frequency sub-bands, respectively), directly fusing multi-dimensional local information. The input feature map is denoted as $X \in \mathbb{R}^{C \times T \times F}$, where C represents the number of channels, T denotes the number of time points, and F stands for the number of frequency bands. The 3D convolutional kernel is defined as $K \in \mathbb{R}^{k_c \times k_t \times k_f}$, correspond to the kernel sizes in the spatial, temporal, and frequency dimensions, respectively. The value at position (c, t, f) in the output feature map Y is calculated as shown in Equation (1).

$$Y(c, t, f) = \sigma \left(\sum_{i=0}^{k_c-1} \sum_{j=0}^{k_t-1} \sum_{k=0}^{k_f-1} K(i, j, k) \cdot X(c+i, t+j, f+k) + b \right) \quad (1)$$

Where $\sigma(\cdot)$ denotes the nonlinear activation function (e.g., ReLU), and b represents the bias term. The model performed three consecutive Conv3d operations with output channels set to 32, 64, and 128, respectively. After flattening the output features, they were fed into the LSTM and attention mechanism module to extract features targeting the temporal attributes of EEG signals. Through its gating mechanisms (input gate, forget gate, output gate), the LSTM captures long- and short-term dependencies in the temporal dimension, which can enhance the temporal dynamic information in the output features of 3DCNN and compensate for the limitation of 3DCNN in capturing local correlations over long time series. Finally, a fully connected layer was applied to yield the four-category classification results.

The training strategy employed 5-fold cross-validation, with the random seed set to 0 to ensure reproducibility. The optimizer chosen was Adam, with an initial learning rate of 0.01, combined with the cross-entropy loss function. The batch size was set to 16 for training, enabling efficient computation under memory constraints. The early stopping mechanism was designed with dual monitoring: the termination condition for training monitored validation loss, stopping when the loss exceeded the historical best by more than 0.01 with no improvement for 6 consecutive epoch, or when the learning rate dropped to the lower limit of 0.0001. The learning rate adjustment independently tracked validation loss, halving the learning rate after 5 consecutive epochs with no improvement. The maximum number of training epochs was set to 100, while the actual training duration was dynamically controlled by the early stopping mechanism. The model training for this study was conducted on an ASUS TUF Gaming F16 FX607JV laptop, with hardware specifications including: a 13th Gen Intel Core i9-13980HX processor (2.20GHz), 16GB RAM, an 8GB dedicated GPU, and 954GB storage. Training utilized the PyTorch framework, prioritizing CUDA GPU acceleration, and was configured with 4 data loading worker processes to fully leverage the multi-core performance of the processor.

3. Results

3.1 Different Feature Combinations

To comprehensively evaluate the effectiveness of different neurophysiological features and frequency bands in emotion recognition, this study systematically compared the classification accuracies of three feature sets across five classical frequency bands, as presented in Table 2. Among these, "DE+SE" denotes the time-frequency nonlinear features combining Differential Entropy and Sample Entropy, which achieved the highest accuracy of 89.1% in the θ band.

Table 2. Accuracy across different EEG features and frequency bands

Features	Accuracy(%)				
	gamma	delta	theta	alpha	Beta
DE+SE	87.2	83.5	89.1	85.8	88.4
PLV	82.6	79.3	84.5	81.2	83.9
ALL	93.8	92.1	94.0	92.7	93.3

Note: DE+SE = Differential Entropy + Sample Entropy; PLV = Phase Locking Value; ALL = Feature fusion of DE+SE and PLV.

This result indicates that the energy and complexity information of θ waves contributes prominently to emotion recognition. PLV represents the functional connectivity features based on Phase Locking Value; although its overall performance was slightly lower than that of DE+SE, it still reached an accuracy of 84.5% in the θ band, demonstrating that emotional processing is accompanied by significant inter-regional phase synchronization in the brain. The last row, ALL, corresponds to the combined feature set fusing DE+SE and PLV, whose classification accuracy was significantly improved across all frequency bands, with a peak of 94.0% in the θ band. This validates that the multi-feature fusion strategy can more comprehensively characterize the neural patterns of emotions and significantly enhance recognition performance.

3.2 Proposed Model Accuracy

To intuitively demonstrate the training dynamics and convergence of the proposed model, Fig. 4 illustrates the decreasing trend of the loss value during training and the variation trend of the accuracy on the test set with respect to epochs, respectively. The training loss curve declined rapidly in the initial phase, and its descending rate slowed down after approximately 15 epochs before gradually converging to a stable low value, indicating that the model was effectively optimized without obvious overfitting. Meanwhile, the test accuracy curve exhibited a steady upward trend; with the scale of the right vertical axis starting at 30%, the accuracy gradually increased past 40% and 50%, and finally stabilized within the high-performance range of 80%–90%. The variation trends of the two curves corroborate each other: when the training loss was effectively reduced, the generalization ability of the model improved synchronously. The test accuracy eventually stabilized at a high plateau of approximately 93.8%, which verifies that the model not only converged rapidly during training but also possessed excellent generalization performance and stability.

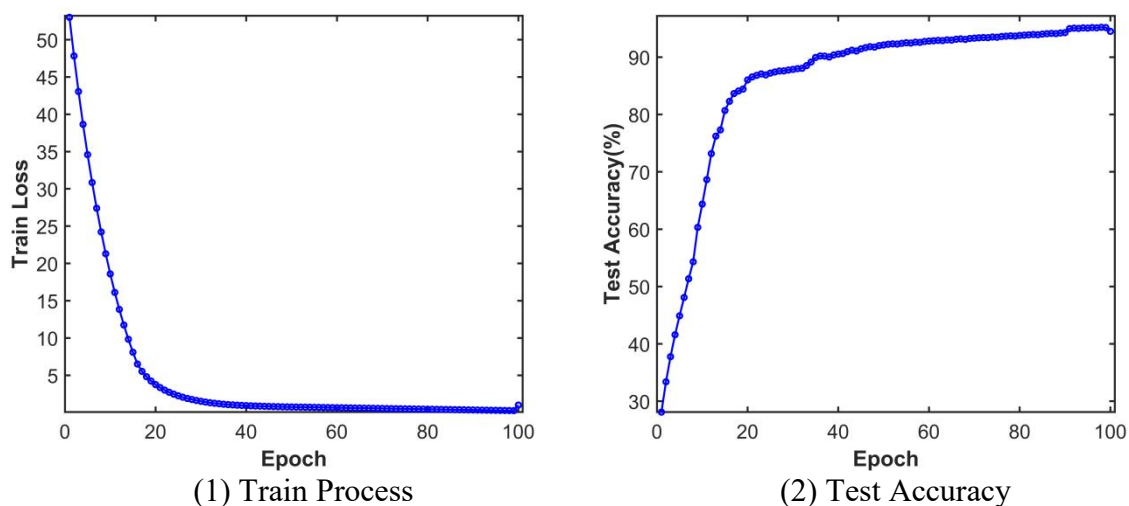


Fig.4. Results of Proposed Model

3.3 Model Comparison

Table 3. Performance comparison of different models

Model Name	Accuracy	Precision	F1-score	Recall
SVM[11]	68.2%	67.5%	66.9%	67.8%
CNN+BiLSTM[12]	82.5%	81.9%	82.1%	82.3%
HierarchicalLSTM[13]	85.7%	85.1%	85.4%	85.9%
EEGNet[14]	89.3%	88.8%	89.0%	89.5%
Proposed model	93.8%	93.5%	93.7%	94.0%

To evaluate the effectiveness of the proposed model, four state-of-the-art classifiers were selected as benchmark models in this study. The classification results and model evaluation metrics are presented in Table 3. The evaluation metrics include accuracy, precision, F1-score, and recall, respectively. For a comprehensive comparison, one traditional model and three deep learning models were chosen. As shown in the classification results in Table 3, the traditional SVM classifier exhibited relatively poor performance across multiple metrics, achieving an accuracy of 68.2%. This indicates that it is difficult for SVM to fully exploit the nonlinear discriminative information embedded in high-dimensional and complex EEG features. With the increase in model complexity, the hybrid network combining CNN and bidirectional LSTM achieved a significant accuracy improvement to 82.5%, demonstrating the advantages of deep learning in capturing the spatiotemporal patterns of EEG signals. For the hierarchical structure (Hierarchical LSTM) and EEGNet—a lightweight network specially designed for EEG analysis—the performance was further improved, with accuracies reaching 85.7% and 89.3%, respectively. The CLANet model proposed in this paper achieved the optimal results across all evaluation metrics, with an accuracy of 93.8% and an F1-score of 93.7%. This demonstrates that the model can more effectively integrate the three-dimensional spatiotemporal-spectral information of multi-band and multi-channel EEG signals, and focus on the key neural representations related to emotions through the attention mechanism, thus achieving more robust and accurate emotion recognition performance.

4. Conclusion

To address the issues of insufficient feature extraction and inadequate fusion of multi-dimensional information in existing EEG emotion recognition studies, this research explores feature fusion and model architecture. The results demonstrate that multi-domain feature fusion can effectively improve recognition performance: after fusing temporal nonlinear features and spatial functional connectivity features across the δ , θ , α , β , and γ frequency bands, the classification accuracy reaches 94.0% in the θ band. The proposed CLANet model, which integrates 3D-CNN, LSTM, and the attention mechanism, synchronously extracts multi-dimensional information and focuses on key features, achieving an accuracy of 93.8% on the SEED IV dataset—outperforming state-of-the-art models. Standardized data preprocessing ensures the reliability of the experiments.

Acknowledgements

This work was supported by the Undergraduate Innovation Training Program of Shenyang University (Project No. 202511035004) and the Doctoral Start-up Foundation of Liaoning Province (Grant No. 2025BS0872).

References

- [1] Luo Zhizeng, Lu Xianju, Zhou Ying. EEG Feature Extraction Based on Brain Function Network and Sample Entropy. *Journal of Electronics & Information Technology*, 2021, 43(2): 412-418.
- [2] Li Mingai, Zhang Yuanyuan. A Brain Functional Network Based on Continuous Wavelet Transform. *Acta Electronica Sinica*, 2022, 50(7):1600-1608.
- [3] Chen Wan, Cai Yanping, Li Aihua, Su Yanzhao, Jiang Ke. Classification of Depression Electroencephalogram Signals Based on Brain Function Network. *Journal of Rocket Force University of Engineering*, 2024, 38(3): 60-6593.
- [4] H. Aldawsari, S. Al-Ahmadi, F. Muhammad. Optimizing 1D-CNN-Based Emotion Recognition Process through Channel and Feature Selection from EEG Signals. *Diagnostics*. 2023; 13(16):2624.
- [5] X. Wang, Y. Ma, J. Cammon, F. Fang, Y. Gao and Y. Zhang. Self-Supervised EEG Emotion Recognition Models Based on CNN. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2023, 31: 1952-1962.

- [6] Mahmoud A, Amin K, Al Rahhal MM, Elkilani WS, Mekhalfi ML, Ibrahim M. A CNN Approach for Emotion Recognition via EEG. *Symmetry*, 2023, 15(10):1822.
- [7] Zhou J .EEG-Based Emotion Recognition Using Convolutional Recurrent Neural Network with Multi-Head Self-Attention[J].*Applied Sciences*, 2022, 12.
- [8] T. Song, W. Zheng, P. Song and Z. Cui. EEG Emotion Recognition Using Dynamical Graph Convolutional Neural Networks. *IEEE Transactions on Affective Computing*, 2011, 11(3): 532-541.
- [9] Zheng Wenlong, Liu Wei, Lu Yifei, Lu Baoliang, Cichocki A. EmotionMeter: A Multimodal Framework for Recognizing Human Emotions. *IEEE transactions on cybernetics*, 2019, 49(3): 1110–1122.
- [10] Feng Naishi, Zhou Bin, Zhang Qian, Hua Chengcheng, Yuan Y A comprehensive exploration of motion sickness process analysis from EEG signal and virtual reality. *Computer methods and programs in biomedicine*, 2025, 264 (2025): 108714.
- [11] Jie X, Cao R, Li L. Emotion recognition based on the sample entropy of EEG. *Biomed Mater Eng*. 2014, 24(1):1185-92.
- [12] Tang Yiheng, Wang Yongxiong, Zhang Xiaoli and Zhe Wang. STILN: A novel spatial-temporal information learning network for EEG-based emotion recognition. *Biomedical Signal Processing and Control*, 2022, 85: 104999.
- [13] Bogdan Badicu, Andreea Udrea. Cross-subjects Emotions Classification from EEG Signals using a Hierarchical LSTM based Classifier. 2019 E-Health and Bioengineering Conference (EHB), Iasi, Romania, 2019, 1-4,
- [14] Mendivil Saucedo, J. A., Marquez, B. Y., & Esqueda Elizondo, J. J. Emotion Classification from Electroencephalographic Signals Using Machine Learning. *Brain sciences*, 2024,14(12): 1211.