CETRNet: Convolutional-EMA Temporal Residual Network for Motor Imagery Decoding

Hongyu Cai[1,a,*]

[1]Maynooth International Engineering College, Fuzhou University, Fuzhou, 350100, China
[a]cloutains1221@163.com
[*]Corresponding author

Abstract: *Brain-Computer Interfaces (BCIs) utilize wearable electroencephalography (EEG) coupled with artificial intelligence (AI) to interpret neural activity. It finds significant applications in healthcare and robotics. And Motor imagery (MI) decoding is the basis of external device control via EEG. However, achieving high decoding accuracy remains a significant challenge, hindering the widespread adoption and advancement of BCI systems. To address this challenge, the present study proposes the Convolutional-EMA Temporal Residual Network (CETRNet), a novel deep learning architecture designed to improve the classification of MI signals from EEG data. The network consists of several key modules that enhance classification performance while maintaining parameter efficiency to reduce computational requirements. The initial processing stage includes dedicated temporal and spatial blocks that capture essential spatio-temporal features, followed by channel attention mechanisms that prioritize relevant spatial information. An Exponential Moving Average (EMA) module is integrated to capture long-range temporal dependencies and detect inherent periodic patterns in the EEG data. Subsequently, higher-level temporal abstractions are derived through a temporal convolutional residual block, which also implements data augmentation using a convolutional sliding window technique. Evaluation on the BCI Competition IV-2a benchmark dataset demonstrated that CETRNet achieved a subject-specific accuracy of 83.33%, highlighting its potential for reliable classification of MI-EEG signals.*

Keywords: *Deep Learning, Attention, Dual-Branch Convolutional Network, Intelligent Healthcare, Motor Imagery, EEG, Classification*

1. Introduction

Brain-Computer Interfaces (BCIs), integrating artificial intelligence (AI) and neuroscience, translate neural activity into external device control signals. These systems show potential across applications including prosthetics, neurorehabilitation, virtual reality, and interactive gaming. Electroencephalography (EEG), a predominant non-invasive and cost-effective neural acquisition method in BCI, offers high temporal resolution. Motor Imagery (MI) is the mental simulation of movement, enabling diverse applications. Accurate decoding of MI-EEG signals remains challenging. Signal quality is degraded by artifacts (e.g., myogenic, ocular, environmental), while inherent properties like inter-subject variability, high dimensionality, inter-channel correlations, and non-stationarity impede interpretation. This necessitates robust and generalizable decoding models.

As one of the excellent methods to address these issues, Deep learning (DL) approaches directly integrate feature extraction and classification from raw EEG data. This minimizes manual preprocessing and typically enhances classification accuracy. DL's success in related domains (e.g., speech, video) has spurred its application to MI-EEG decoding. Advancements in DL-based MI classification often leverage progress from related domains. Convolutional Neural Networks (CNNs) currently represent the most widely adopted architectural approach for decoding motor imagery from EEG signals. Significant progress within this domain includes exploring the impact of network depth, as seen in Deep/Shallow ConvNets, alongside enhancements in both computational efficiency and performance achieved using separable convolutions, notably demonstrated by EEGNet [1,2]. While CNNs are prominent, the field also utilizes other deep learning approaches for MI classification. Stacked Autoencoders (SAEs) have been employed for extracting spectral features, DBNs targeted multi-class decoding challenges, and RNN variants like LSTMs have proven useful for modeling the inherent temporal dependencies present within MI-EEG recordings [3-5].

Temporal Convolutional Networks (TCNs) effectively model temporal data, capturing long-range dependencies efficiently while mitigating RNN gradient issues. Their suitability for sequential data has led to applications in MI-EEG decoding, including integration with EEGNet, feature fusion strategies, and attention-enhanced hybrid TCN-CNN frameworks [6-8].

Attention mechanisms are increasingly integrated into DL architectures for MI classification to enhance feature discrimination and performance. Examples include CNNs incorporating Squeeze-and-Excitation (SE) or Efficient Channel Attention (ECA), and frameworks utilizing Multi-Head Self-Attention (MSA) [8-10].

This study proposes the Convolutional-EMA Temporal Residual Network (CETRNet) for MI-EEG decoding. The architecture employs a three-stage process: (1) initial convolutional layers extract high-level temporal representations; (2) an Exponential Moving Average (EMA) block captures long-term signal trends; (3) a temporal convolutional layer with multi-scale fusion derives refined temporal features. CETRNet aims to improve MI classification via synergistic integration of EMA and a convolutional sliding window.

2. Material and Methods

The CETRNet model comprises three core blocks (as shown in Figure 1): convolutional block (CV),Exponential Moving Average block (EMA), and temporal convolutional residual block (TCR). The CV block, incorporating temporal and spatial sub-blocks, extracts low-level spatio-temporal MI-EEG features, while channel attention mechanisms enhance spatial information selection. The CV block outputs high-level temporal sequences, which are divided into windows and processed by the EMA block. The EMA block captures long-term trends and periodic fluctuations within each window, feeding features to the TCR block for advanced temporal feature extraction. Outputs from all windows are concatenated and classified via a SoftMax layer, generating probability predictions for MI tasks. This architecture improves data augmentation and classification accuracy.

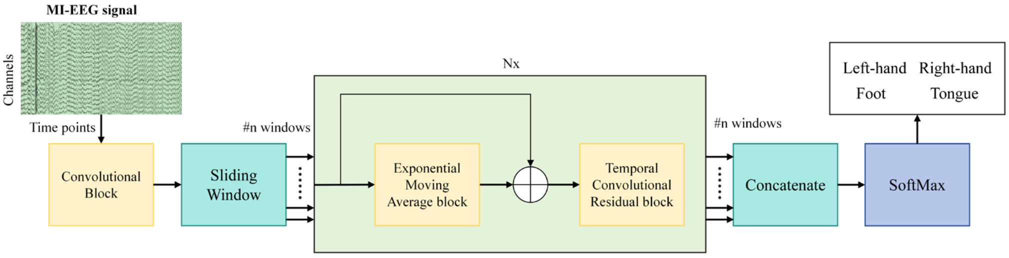

Figure 1 Components of the CETRNet architecture.

2.1 Input Representation

This study leverages the input structure of ATCNet for the CETRNet framework, utilizing the entire frequency range while retaining all artifacts [8]. The CETRNet model processes a motor imagery trial$X_i \in R^{C \times T}$, which includes *C* EEG channels and *T* time points, with the objective of mapping X_i to its associated class label y_i. The dataset $S = \{X_i, y_i\}_{i=1}^{m}$ consists of *m* trials, where $y_i \in \{1, \dots, n\}$denotes the class label, and *n* indicates the total number of distinct classes. In the BCI-2a dataset, the values are set as follows: *T* = 1125 time points, *C* = 22 EEG channels, *n* = 4 classes, and *m* = 5184 MI trials [11].

2.2 Convolutional (CV) Block

This block along with the convolutional structures described in ATCNet utilizes analogous kernel parameters and both utilize batch normalization (BN), exponential linear units (ELU), and average pooling [8,12,13]. The CV block comprises three convolutional layers (as shown in Figure 2). At the beginning, temporal convolution is set with F_1 filters (size 1 × K_c). K_c is configured as one-fourth the sampling rate, filtering out sub-4 Hz components. Then depth-wise convolution is utilized with F_2 filters (size *C* × 1) to extract spatial characteristics. *C* represents total EEG channels used. The output dimension is empirically set by *D*, typically *D* = 2. An average pooling layer (size 1 × 8) aggregates temporal information at an 8:1 ratio. Finally, the layer uses F_2 filters (size 1 × K_{c2}), with K_{c2} = 16, followed by a second average pooling layer (size 1 × P_2) to further downsample the sequence.

The CV block outputs a sequence $z_i \in \mathbb{R}^{T_c \times d}$ of temporal representations, where each vector has dimension d, empirically set to 32. The T refers to the time points of the original EEG signal. The length of the temporal sequence z_i is determined by $T_c = {}^{T}\!/_{8P_2}$.

Temporal and depthwise convolutions alone insufficiently capture spatial information in EEG signals, potentially leading to spatial feature loss. Inspired by the success of channel attention mechanisms in LMDA [14], we incorporate channel attention modules both after the temporal and spatial convolution layers of the CV block. This recalibration enhances the network's ability to model spatial and spectral information and better identify salient features within MI-EEG data.

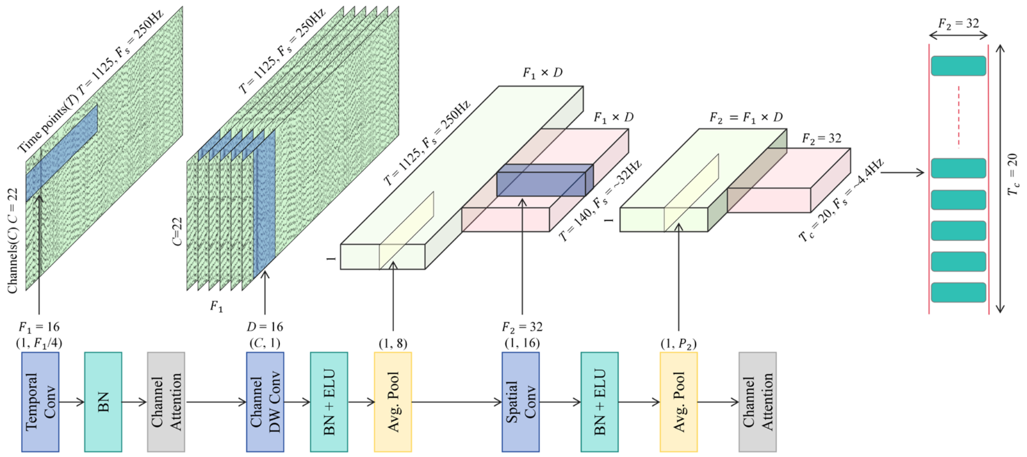

Figure 2 Convolutional (CV) block.

2.3 Sliding Window (SW)

The time series z_i is divided into multiple local sequences $z_i^w \in \mathbb{R}^{T_w \times d}$ using a sliding window (SW). This facilitates the extraction of individual local features. A SW of size $T_w = T_c$ - 5 with one element step was used. This particular configuration segmented the input sequence z_i into precisely 5 local windows. For more extensive settings and a comprehensive analysis regarding sliding window methodologies, please refer to the research detailed in reference [8].

2.4 Exponential Moving Average (EMA) Block

The multivariate time series $x = (x_1, x_1, \dots, x_L)$ is divided into M univariate sequences $x^{(i)} = (x_1^{(i)}, x_2^{(i)}, \dots, x_L^{(i)})$, where $x^{(i)} \in \mathbb{R}^L$ and L is lookback of recent historical data points. Each sequence is fed into the backbone model to generate a prediction sequence $\hat{x}^{(i)} = (\hat{x}_{L+1}^{(i)}, \hat{x}_{L+2}^{(i)}, \dots, \hat{x}_{L+T}^{(i)})$, where$\hat{x}^{(i)} \in \mathbb{R}^T$ and T is future steps observations. In Exponential Decomposition, each univariate series is decomposed into trend and seasonality components, processed separately by the dual-flow architecture, then aggregated for the final prediction (as shown in Figure 3).

(1) Non-linear Stream utilizes a CNN to model non-linear patterns.

Patching: segments each univariate time series using a sliding window, as introduced in PatchTST. Patches of length P are extracted with stride S, resulting in N two-dimensional patches $x_p^{(i)} \in \mathbb{R}^{N \times P}$, where $N = \left\lfloor \frac{L-P}{S} \right\rfloor + 2$. For consistency with PatchTST and CARD, we set P = 16 and S = 8.

Depthwise Convolution: A grouped convolution ($g = N$, kernel size $k = p$, stride $s = P$) processes each patch representation independently to extract local temporal features.

Pointwise Convolution: A 1 × 1 convolution (g = 1) aggregates features across different patches.

Each convolutional layer is followed by Batch Normalization and activation (σ). A residual connection is incorporated around the depthwise layer.

Output MLP: The output features are flattened and passed through an MLP block

(Linear-GELU-Linear projection), yielding the non-linear feature representation:

$$\hat{x}_{nonlin}^{(i)} = \mathrm{Linear}\left(\sigma\left(\mathrm{Linear}\left(\mathrm{Flatten}\left(x_p^{N\times P}\right)\right)\right)\right) \tag{1}$$

(2) Linear Stream models the trend component using an MLP-based architecture. It employs linear transformations, average pooling, and layer normalization, omitting non-linear activations to preserve linear characteristics. The output represents the linear feature prediction:

$$\hat{x}_{lin}^{(i)} = \mathrm{Linear}\left(x^{(i)}\right) \tag{2}$$

(3) Linear features and non-linear features are concatenated and fused via a final linear layer to produce the prediction:

$$\hat{x}_{lin}^{(i)} = \mathrm{Linear}\left(\mathrm{concat}\left(\hat{x}_{lin}^{(i)}, \hat{x}_{nonlin}^{(i)}\right)\right) \tag{3}$$

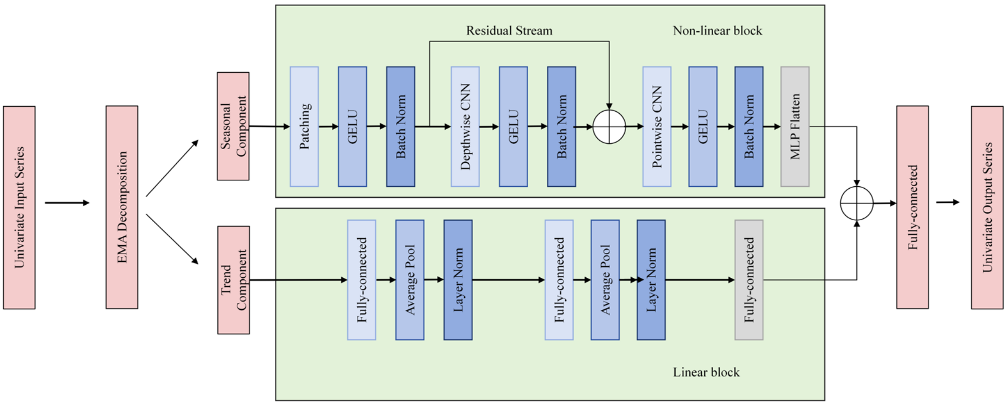

Figure 3 Exponential Moving Average block's overview.

2.5 Temporal Convolutional Residual (TCR) Block

The TCR block adopts the identical architecture and hyperparameter configuration of the TCN used in EEG-TCNet and ATCNet, but introduces a novel form of residual module. Instead of conventional shortcuts, it implements multi-level residual connections (Figure 4), which promote hierarchical feature fusion—thereby enhancing representational capacity and curbing overfitting [6,8]. Structurally, each TCR block comprises two residual blocks; within each block, two causal dilated convolutional layers are succeeded by an Exponential Linear Unit (ELU) activation and batch normalization [12] (see Figure 4). For a complete specification of this design, consult [8]. As illustrated in Figure 5, the block receives sixteen temporal vectors ($T_w = 16$), each of dimension F_2, and produces an output sequence whose final component has dimension F_T. In the work, I set $F_T = F_2 = 32$ [13].

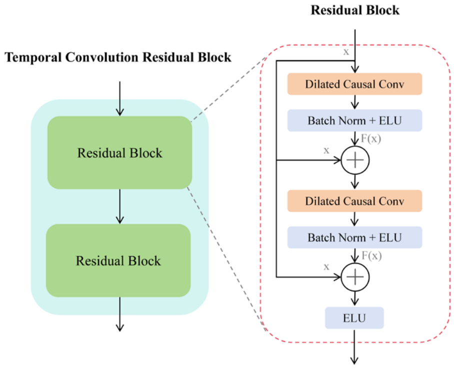

Figure 4 The architecture of temporal convolutional residual block.

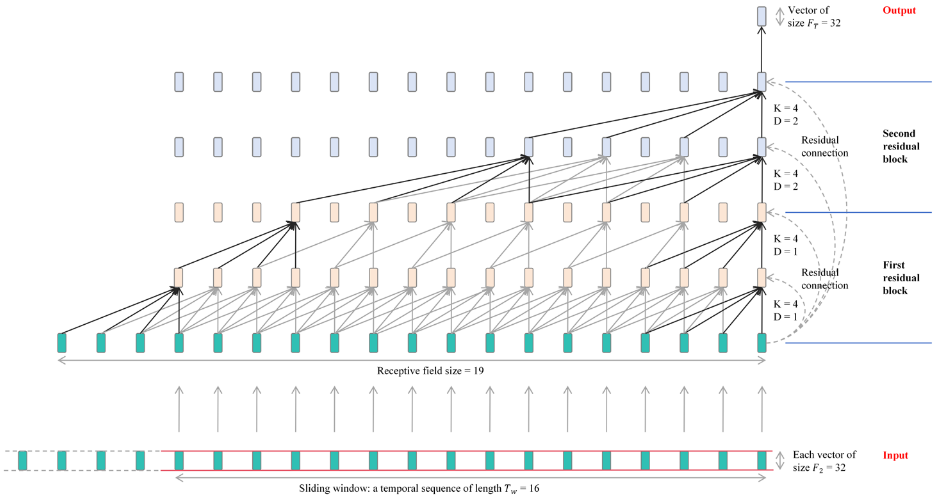
Vector of size F_T = 32
Output
K = 4 D = 2
Residual connection
Second residual block
K = 4 D = 2
K = 4 D = 1
Residual connection
First residual block
K = 4 D = 1
Receptive field size = 19
Each vector of size F_2 = 32
Input
Sliding window: a temporal sequence of length T_w = 16

Figure 5 Visualization of temporal convolutional residual block.

3. Experimental Results and Discussion

3.1 Dataset Description

The BCI Competition IV-2a (BCI-2a) dataset [11], a prominent public MI-EEG benchmark introduced by Graz University of Technology in 2008 is selected for training and evaluating the CETRNet. While widely used in MI-EEG decoding studies, this dataset is known for challenges such as limited sample size, inter-session variability, and significant artifact contamination. The dataset contains 5,184 trials across nine subjects (576 trials/subject), acquired using 22 EEG channels. Each 4 s MI trial was sampled at 250 Hz after bandpass filtering (0.5-100 Hz). The four MI tasks performed were: left-hand (class 1), right-hand (class 2), foot (class 3), and tongue (class 4) movements. Acquisition involved two sessions per subject (288 trials each), one designated for training and the other for testing.

3.2 Implementation Details

All experiments were conducted using an Nvidia GTX 3070 GPU (8GB) within the TensorFlow framework. Training was standardized across trials: weights were initialized using the Glorot uniform method, and optimization was performed with the Adam algorithm (learning rate = 0.0009, batch size = 64). Categorical cross-entropy served as the loss function. Training was constrained to a maximum of 1000 epochs, with early stopping triggered if no improvement occurred over 300 epochs.

3.3 Performance Metrics

The proposed models in this research are evaluated using accuracy, Kappa score, and standard deviat.

(1) Accuracy:

$$Accuracy = \frac{\sum_{i=1}^{n} TP_i/I_i}{n} \tag{4}$$

Let TP_i denote the number of correctly predicted instances in class i, I_i the total number of samples in class i, and n the total number of classes.

(2) Kappa score:

$$Kappa = \frac{1}{n}\sum_{a=1}^{\infty}\frac{P_a-P_e}{1-P_e} \tag{5}$$

Let P_a be the observed agreement proportion and P_e the expected chance agreement.

(3) Standard deviation:

$$Std = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Acc_i - r)^2} \tag{6}$$

N represents the total number of subjects or samples, and *r* indicates the average accuracy.

3.4 Performance Comparison

Table 1 compares the performance of the proposed CETRNet model with reimplemented versions of EEGNet, EEG-TCNet, and TCNet_Fusion. All models were trained under identical conditions using parameters from their original studies. CETRNet achieved the highest average accuracy (83.3%) and Kappa score (0.78), representing a 5.0% increase in accuracy. Confusion matrices (Figure 6) further demonstrate CETRNet's superior classification performance through higher diagonal and lower off-diagonal values. Considering the variability of EEG signals across subjects and sessions, a hold-out evaluation was conducted to assess generalization. Results (Table 2) show CETRNet maintains strong performance in cross-session settings, achieving an average accuracy of 83.3% and Kappa score of 0.78 on BCI-2a dataset, surpassing all baseline methods. These findings highlight CETRNet's effectiveness in extracting robust, transferable features and managing non-stationarities in EEG data.

Table 1 Performance (accuracy (%) and Kappa score (k)) comparison.

Sub.	CETRNet				EEGNet				EEG-TCNet				TCNet-Fusion			
	best		average		best		average		best		average		best		average	
	%	k	%	k	%	k	%	k	%	k	%	k	%	k	%	k
1	87.5	0.83	86.6	0.82	88.9	0.85	86.8	0.85	85.8	0.81	81.6	0.75	86.5	0.82	84.7	0.80
2	70.5	0.61	69.9	0.60	63.5	0.51	62.0	0.51	64.2	0.52	60.8	0.48	63.2	0.51	61.0	0.48
3	96.2	0.95	95.8	0.94	90.6	0.88	89.9	0.88	94.4	0.93	92.7	0.90	92.0	0.89	91.4	0.89
4	78.5	0.71	77.4	0.70	67.4	0.56	65.1	0.56	67.0	0.56	63.8	0.52	71.2	0.62	69.7	0.60
5	80.9	0.75	80.1	0.73	73.3	0.64	71.1	0.64	74.3	0.66	72.7	0.64	78.1	0.71	76.9	0.69
6	74.3	0.66	73.3	0.64	64.2	0.52	61.7	0.52	64.2	0.52	61.1	0.48	64.6	0.53	63.7	0.52
7	91.7	0.89	91.3	0.88	88.9	0.85	86.5	0.85	87.9	0.84	86.8	0.82	90.6	0.88	89.6	0.86
8	87.2	0.83	86.5	0.82	85.1	0.80	84.1	0.80	84.4	0.79	83.5	0.78	85.4	0.81	85.0	0.80
9	89.9	0.87	89.1	0.85	83.7	0.78	82.9	0.78	83.7	0.78	81.7	0.76	84.4	0.79	82.8	0.77
Mean	84.1	0.79	83.3	0.78	78.4	0.71	76.7	0.71	78.4	0.71	76.1	0.68	79.6	0.73	78.3	0.71
St.D.	8.5	0.11	8.7	0.11	11.3	0.15	11.6	0.15	11.3	0.15	11.8	0.16	10.9	0.14	11.1	0.15

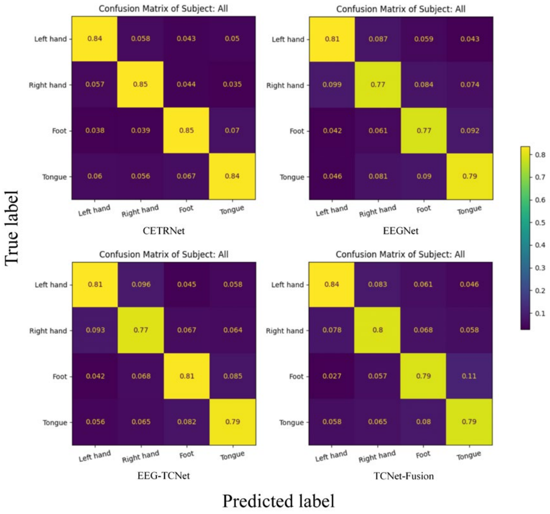

Figure 6 Confusion matrices of the CETRNet and the reproduced models.

Table 2 Performance on BCI-2a dataset using hold-out.

Method	Accuracy %	Kappa
EEGNet*	76.7	0.71
EEG-TCNet*	76.1	0.68
TCNet_Fusion*	78.3	0.71
Deep ConvNet	73.2	0.64
FBCNet	75.5	0.67
FBMSNet	76.3	0.68
Conformer	78.6	0.71
CETRNet (Proposed)	83.3	0.78

* Reproduced

3.5 Ablation Study

An ablation study was conducted to assess the contribution of individual components in the CETRNet model using the BCI-2a dataset. Table 3 shows the performance changes resulting from the removal of specific blocks prior to training and validation. The Spatial Weighting (SW) block improved accuracy by 2.5% and EMA by 1.8%, while the Frequency-Time Context (FTC) block increased accuracy by 2.6% over the Convolutional (CV)-only baseline. Results indicate that each block contributes independently to the overall performance of the model.

Table 3 Impact of each block on the classification performance of CETRNet.

Removed block	Accuracy %	Kappa
None (CETRNet)	83.3	0.78
SW	80.8	0.74
EMA	81.5	0.75
TCR	80.7	0.74
SW + EMA	78.7	0.72
SW + TCR	77.7	0.70
EMA + TCR	78.6	0.71
SW + EMA + TCR	75.8	0.68

SW: sliding window, EMA: Exponential Moving Average block, TCR: temporal convolutional residual block.

4. Conclusion

This study proposed CETRNet, a novel Convolutional-EMA Temporal Residual Network for EEG-based motor imagery (MI) classification. CETRNet integrates three core blocks: a convolutional (CV) block for encoding raw MI-EEG signals, an Exponential Moving Average (EMA) block for capturing long-term data direction and periodicity, and a temporal convolutional residual (TCR) block for extracting high-level temporal features. Performance was further enhanced by incorporating a convolutional sliding window (SW) with the CV block, enabling efficient parallel processing. Ablation analyses substantiate the efficacy of each component, revealing that the SW, EMA, and TCR blocks incrementally enhance classification accuracy by 2.5%, 1.8%, and 2.6%, respectively, relative to a baseline model comprising only the CV block. When evaluated on the publicly available BCI-2a dataset, CETRNet achieves a subject-dependent classification accuracy of 83.33%, outperforming several recent state-of-the-art methods. Remarkably, this high performance is realized with a relatively small parameter count, rendering CETRNet well-suited for deployment in resource-constrained environments, such as portable or embedded BCI systems. The model demonstrates robust feature extraction capabilities directly from raw EEG signals, requiring minimal pre-processing, even when applied to a limited dataset that may contain artifacts—a common challenge in real-world EEG applications. The consistent performance improvements observed across all MI classes and subjects highlight CETRNet's capacity to learn generalizable representations from EEG data, enhancing its applicability across diverse BCI scenarios. Future work may explore incorporating multi-domain attention mechanisms (temporal, spatial, spectral) to prioritize salient information. Additionally, performance could potentially be improved through targeted preprocessing for artifact removal and dataset augmentation via deep generative models.

References

[1] Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D. J., Glasstetter, M., Eggensperger, K., Tangermann, M., ... & Ball, T. (2017). Deep learning with convolutional neural networks for EEG decoding and visualization. Human brain Mapping, 38(11), 5391-5420.

[2] Lawhern, V.J., Solon, A.J., Waytowich, N.R., Gordon, S.M., Hung, C.P., & Lance, B. (2016). EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. Journal of Neural Engineering, 15.

[3] Hassanpour, A., Moradikia, M., Adeli, H., Khayami, R., & Babaki, P. S. (2019). A novel end-to-end deep learning scheme for classifying multi-class motor imagery electroencephalography signals. Expert Systems (7).

[4] Xu, J., Zheng, H., Wang, J., Li, D., & Fang, X. (2020). Recognition of EEG Signal Motor Imagery

Intention Based on Deep Multi-View Feature Learning. Sensors, 20(12), 3496.

[5] Luo, T. J., Zhou, C. L., & Chao, F. (2018). Exploring spatial-frequency-sequential relationships for motor imagery classification with recurrent neural network. BMC Bioinformatics, 19(1).

[6] Wang, X., Hersche, M., Tömekce, Batuhan, Kaya, B., Magno, M., & Benini, L. (2020). An accurate eegnet-based motor-imagery brain-computer interface for low-power edge computing. IEEE.

[7] Musallam, Y. K., Alfassam, N. I., Ghulam, M., Amin, S., Alsulaiman, M., & Abdul, W., et al. (2021). Electroencephalography-based motor imagery classification using temporal convolutional network fusion. Biomed. Signal Process. Control., 69, 102826.

[8] Altaheri, H., Muhammad, G., & Alsulaiman, M. (2022). Physics-Informed Attention Temporal Convolutional Network for EEG-Based motor imagery classification. IEEE Transactions on Industrial Informatics, 19(2), 2249–2258.

[9] Jia, H., Yu, S., Yin, S., Liu, L., Yi, C., Xue, K., Li, F., Yao, D., Xu, P., & Zhang, T. (2023). A Model Combining Multi Branch Spectral-Temporal CNN, Efficient Channel Attention, and LightGBM for MI-BCI Classification. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 31, 1311-1320.

[10] Altuwaijri, G. A., Muhammad, G., Altaheri, H., & Alsulaiman, M. (2022). A Multi-Branch Convolutional Neural Network with Squeeze-and-Excitation Attention Blocks for EEG-Based Motor Imagery Signals Classification. Diagnostics, 12(4), 995.

[11] Brunner, C., Leeb, R., Müller-Putz, G., Schlögl, A., & Pfurtscheller, G. (2008). BCI Competition 2008–Graz data set A: Technical Report No. 16, pp. 1–6.

[12] Ioffe, S., & Szegedy, C. (2016). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. 32nd International conference on machine learning: ICML 2015, Lile, France, 6-11 July 2015, volume 1 of 3.

[13] Zhou, K., Haimudula, A., & Tang, W. (2024). Dual-Branch Convolution Network with Efficient Channel Attention for EEG-Based Motor Imagery Classification. IEEE Access, 12, 74930–74943.

[14] Miao, Z., Zhang, X., Zhao, M., & Ming, D. (2023). LMDA-Net: A lightweight multi-dimensional attention network for general EEG-based brain-computer interface paradigms and interpretability. ArXiv, abs/2303.16407.