ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3315-6701-9/26/$31.00 ©2026 IEEE | DOI: 10.1109/ICASSP55912.2026.11463125

HIERARCHICAL CONVOLUTION MULTIBRANCH TRANSFORMER FOR EEG SIGNALS

Omid Mersa *Maiqi Jiang* *Ye Gao* *Qun Li* *Yanfu Zhang**

William & Mary

ABSTRACT

Electroencephalography (EEG) supports brain–computer interfaces (BCI), cognitive monitoring, and clinical diagnostics through millisecond-resolution neural recordings. However, effective representation learning remains challenging due to low signal-to-noise ratios, heterogeneous electrode configurations, and complex spatial–spectral–temporal dynamics. While deep models have improved performance, they often blur axis-specific structures, reducing generalizability across subjects and datasets. We introduce the Hierarchical Convolution Multibranch Transformer (HCMT), a self-supervised framework that maintains axis-aware representations via a spectral filter bank, hierarchical convolutions, and dedicated Transformer branches for spatial, spectral, and temporal modeling. HCMT is trained with a multi-objective loss that integrates masked reconstruction, branch decorrelation, and feature diversity. Across tasks including P300 detection, motor imagery, and sleep staging, HCMT consistently outperforms standard baselines, highlighting its effectiveness in robust, axis-aware EEG representation learning.

***Index Terms*—** EEG, representation learning, spatial-spectral-temporal modeling, Hierarchical CNN

1. INTRODUCTION

Electroencephalography (EEG) has the potential to transform brain–computer interfaces (BCI), cognitive monitoring, and clinical diagnostics by providing millisecond-level access to neural activity. However, progress in robust EEG modeling remains limited by the difficulty of learning representations that generalize across subjects, sessions, and recording conditions. Models that perform well on one dataset often fail to transfer, restricting their practical utility. This lack of robustness stems from three persistent challenges.

First, generalization. Handcrafted pipelines such as band power and Common Spatial Patterns (CSP) require subject-specific tuning and often degrade when applied across subjects or sessions [1, 2]. Even recent transfer-learning methods show a consistent performance gap between subject-specific and subject-independent settings [3].

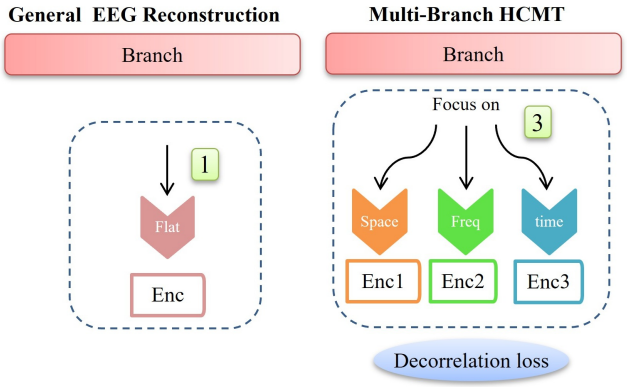

Fig. 1. HCMT comparison with common mask reconstruction

Second, long-range dependencies. EEG dynamics unfold over hundreds of milliseconds to several seconds, as seen in P300 event-related potentials (ERP) and sleep-stage transitions [4]. Convolutional networks such as EEGNet learn local spatio-spectral patterns [5, 6], but their limited receptive fields restrict temporal modeling. While Transformer-based models can capture long-range dependencies [7, 8], many flatten EEG inputs and lose axis-specific structure in the process as shown in Figure 1.

Third, axis-specific patterns. EEG signals differ across spatial locations, spectral rhythms, and temporal patterns. Classical approaches often treated these axes separately [1], but many recent end-to-end models process EEG holistically within a single encoder, obscuring structural priors and reducing robustness [7, 8]

To address these challenges, we propose the **Hierarchical Convolution Multibranch Transformer (HCMT)**, a self-supervised framework that explicitly preserves axis-aware structure. HCMT applies a spectral filter bank and 3D-to-2D convolutional stem to maintain local detail, followed by axis-specific attention and separate Transformer branches specialized for spatial, spectral, and temporal modeling. A multi-objective loss—combining masked reconstruction, branch decorrelation, and feature diversity—encourages each branch to learn complementary representations. Unlike prior work that processes EEG jointly, HCMT preserves axis structure, enabling improved generalization across subjects, long-range dependency modeling, and robustness to heterogeneous electrode layouts.

Contributions. Our work makes the following contributions:

* Corresponding author.

1. We introduce HCMT, a self-supervised EEG framework that combines a spectral filter bank, hierarchical convolutions, and multibranch Transformers for axis-aware representation learning.
2. We design a multi-objective loss that encourages complementary specialization across spatial, spectral, and temporal branches, enhancing both axis-aware modeling and cross-subject generalization.
3. We conduct comprehensive evaluations on three EEG benchmarks, including P300 detection, motor imagery, and sleep staging, demonstrating that HCMT outperforms strong convolutional and Transformer-based baselines.

2. RELATED WORK

Early EEG decoding relied on handcrafted features such as band power and CSP with shallow classifiers [2], achieving good within-subject accuracy but requiring extensive calibration and generalizing poorly across subjects [1, 9]. Deep learning enabled end-to-end feature learning. CNNs such as DeepConvNet and EEGNet [6, 5] improved local spatio-spectral extraction but have limited receptive fields, restricting long-range modeling and cross-subject transfer [10, 11]. Transformers address these limitations by modeling global dependencies. Examples include AttnSleep [7] and EEGformer [8], which integrate CNNs with self-attention or design specialized tokenization. Yet most approaches flatten EEG into 1D sequences and process spatial, spectral, and temporal cues jointly, diluting axis-specific patterns.

With success of self-supervised learning across modalities, it has been applied to EEG to reduce labeled-data requirements via masked prediction and contrastive objectives. BENDR [12] adapts wav2vec 2.0 with masked prediction, while EEGPT [13] combines masked reconstruction and feature alignment for cross-task generalization. Other methods explore contrastive predictive coding and augmentations [14], but remain constrained by architectural bottlenecks [15].

In contrast, HCMT explicitly preserves spatial, spectral, and temporal structure through attention modules and multibranch Transformers, enabling axis-aware representation learning across diverse EEG tasks.

3. METHOD

Overview. HCMT integrates self-supervised pretraining and a multibranch Transformer to capture EEG's spectral, spatial, and temporal characteristics. A sinc filter bank and hierarchical CNN feed three parallel branches whose inputs are attended along temporal, spectral, or spatial axes; we pretrain via masked reconstruction with decorrelation/diversity losses and fine-tune by aggregating branch tokens. An overview of the proposed HCMT framework is shown in Figure 2.

3.1. Hierarchical CNN–Transformer Encoder

To capture multi-scale EEG structure, we employ a hierarchical convolutional encoder followed by axis-specific Transformer branches.

3D Convolution and Pooling. The filter bank output $\mathbf{X}_{\text{fb}}$ is passed through a 3D convolution to produce $\mathbf{U}^{(3D)} \in \mathbb{R}^{B\times C_{3d}\times T\times F\times S}$. pooling reduces resolution to (T', F', S').

Axis-Specific Attention. Independent spatial, spectral, and temporal attention maps are computed by global averaging over the other two axes, passing through a two-layer MLP, and applying softmax normalization. The weighted feature maps are summed as: $\mathbf{U}_{\text{att}}^{(3D)} = \mathbf{U}_{\text{spat}}^{(3D)} + \mathbf{U}_{\text{spec}}^{(3D)} + \mathbf{U}_{\text{temp}}^{(3D)}$

2D Convolution and Axis-Specific Branches. The refined 3D volume is split into temporal–spatial and temporal–spectral branches by averaging over frequency or space, respectively. Each branch applies 2D convolution and its corresponding attention modules, producing axis-focused features $\mathbf{U}_{\text{TS}}^{\text{spatial}}, \mathbf{U}_{\text{TS}}^{\text{temporal}}, \mathbf{U}_{\text{TF}}^{\text{frequency}}, \mathbf{U}_{\text{TF}}^{\text{temporal}}$.

Compression and Axis Reconfiguration. Branch outputs are projected to match 3D attention maps. Axis-specific 2D attention maps are reshaped into 3D volumes and added to their corresponding 3D attention outputs. A $1 \times 1 \times 1$ convolution compresses channels to D, yielding $\mathbf{V} \in \mathbb{R}^{B\times D\times T'\times F'\times S'}$. Each axis-specific volume is reshaped to emphasize its primary dimension by upsampling it by a factor of four using a transposed convolution and downsampling the other two dimensions by a factor of two. For example, the frequency-emphasized output $\mathbf{Z}_{\text{spec}}$ is reshaped to $\mathbb{R}^{B\times D\times T''/2\times 4F''\times S''/2}$, with analogous transformations applied to $\mathbf{Z}_{\text{temp}}$ and $\mathbf{Z}_{\text{spat}}$.

Patch Tokenization and Masking. Each axis-specific volume $\mathbf{Z}_{\text{axis}}$ is partitioned into non-overlapping patches, with the temporal dimension divided into 8 segments and patch size along frequency/spatial dimensions set to 1. During pretraining, 50% of patches are randomly selected and within each, 50% of features are masked.

Transformer Encoding and Reconstruction. Masked patch tokens are processed by a branch-specific Transformer encoder. A lightweight Transformer decoder $d_\phi^{(i)}$ reconstructs the original unmasked patch embeddings $\hat{\mathbf{z}}_i \in \mathbb{R}^{N\times D}$ from the encoded tokens $\mathbf{z}_i \in \mathbb{R}^{N\times D}$.

3.2. Self-Supervised Pretraining and Fine-Tuning

HCMT is trained in two stages: self-supervised pretraining and task-specific fine-tuning.

Pretraining Objective. The total loss combines three components: $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{recon}} + \lambda_1\mathcal{L}_{\text{dec-trans}} + \lambda_2\mathcal{L}_{\text{div-conv}}$

(1) Reconstruction Loss. Each branch $i \in \{\text{spat}, \text{spec}, \text{temp}\}$ reconstructs original unmasked patch embeddings $\mathbf{z}^{(i)}$ from encoded tokens $\mathbf{z}_{\text{HCMT}}^{(i)}$ via a lightweight Transformer decoder $d_\phi^{(i)}$. The reconstruction loss is computed as follows:

$$\mathcal{L}_{\text{recon}} = \sum_{i\in\{\text{spat},\text{spec},\text{temp}\}} \frac{1}{N_i}\sum_{p=1}^{N_i} \frac{\left\|\hat{z}_p^{(i)} - z_p^{(i)}\right\|_2^2}{\text{RMS}(z_p^{(i)})^2 + \epsilon}$$

where N_i is the number of patch tokens and $\text{RMS}(\cdot)$ is the root-mean-square (RMS) magnitude.

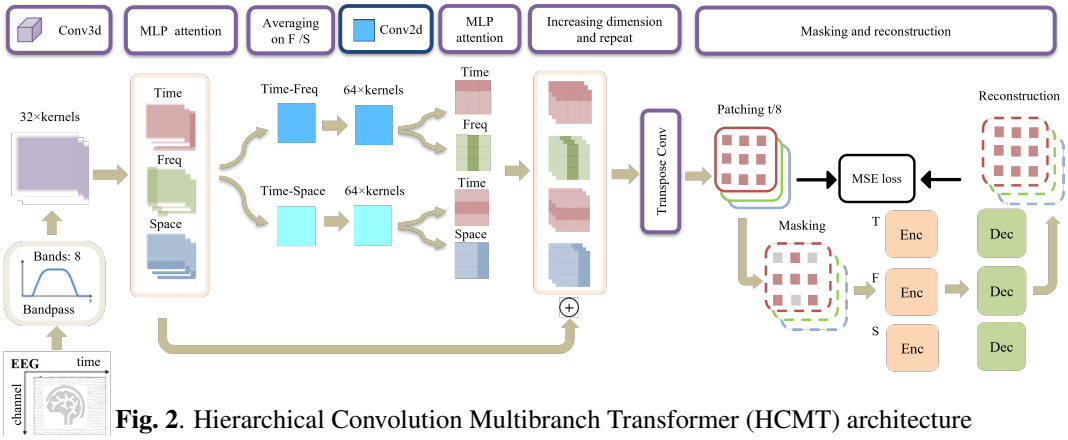

Fig. 2. Hierarchical Convolution Multibranch Transformer (HCMT) architecture

(2) Transformer Branch Decorrelation. To minimize redundancy between Transformer branches, we employ a decorrelation loss that penalizes shared covariance. To compute cross-branch decorrelation along each axis, we apply axis-specific pooling across non-target dimensions, yielding axis-aligned summaries: $\mathbf{g}_{\mathrm{F}}^{(i)} = \mathrm{Avg}_{T^{(i)},S^{(i)}}(\mathbf{z}^{(i)}) \in \mathbb{R}^{B\times F^{(i)}\times D}$, $\mathbf{g}_{\mathrm{T}}^{(i)} = \mathrm{Avg}_{F^{(i)},S^{(i)}}(\mathbf{z}^{(i)}) \in \mathbb{R}^{B\times T^{(i)}\times D}$, $\mathbf{g}_{\mathrm{S}}^{(i)} = \mathrm{Avg}_{T^{(i)},F^{(i)}}(\mathbf{z}^{(i)}) \in \mathbb{R}^{B\times S^{(i)}\times D}$. where $T^{(i)}, F^{(i)}, S^{(i)}$ denote output dimensions along each axis for branch i. To enable axis-wise comparison across branches with different resolutions, linear interpolation maps these summaries to the largest axis length. We then compute decorrelation loss:

$$\mathcal{L}_{\text{dec-trans}} = \sum_{A\in\{F,T,S\}} \sum_{\substack{i,j\in\{\text{spat,spec,temp}\}\\ i\neq j}} \mathrm{Tr}\Big(\mathrm{Cov}\big(\tilde{\mathbf{g}}_A^{(i)}, \tilde{\mathbf{g}}_A^{(j)}\big)\Big)$$

where $A \in \{F, T, S\}$ indexes frequency, temporal, and spatial axes, $\tilde{\mathbf{g}}_A^{(i)}$ denotes the axis-aligned summary for branch i after interpolation to a common resolution, $\mathrm{Cov}(\cdot,\cdot)$ is the cross-covariance, and $\mathrm{Tr}(\cdot)$ is trace operator.

(3) Convolutional Branch Diversity. The diversity loss is computed using the output of the convolutional encoder. Axis-aligned summaries are computed by mean pooling: $\mathbf{h}_{\mathrm{F}}^{(i)} = \mathrm{Avg}_{T^{(i)},S^{(i)}}(\mathbf{X}_i) \in \mathbb{R}^{B\times D\times F^{(i)}}$, $\mathbf{h}_{\mathrm{T}}^{(i)} = \mathrm{Avg}_{F^{(i)},S^{(i)}}(\mathbf{X}_i) \in \mathbb{R}^{B\times D\times T^{(i)}}$, $\mathbf{h}_{\mathrm{S}}^{(i)} = \mathrm{Avg}_{T^{(i)},F^{(i)}}(\mathbf{X}_i) \in \mathbb{R}^{B\times D\times S^{(i)}}$

then interpolated to match axis lengths. The diversity objective is defined as:

$$\mathcal{L}_{\text{div-conv}} = \sum_{A\in\{F,T,S\}} \sum_{\substack{i,j\in\{\text{spat,spec,temp}\}\\ i\neq j}} \Big(1 - \cos\big(\tilde{\mathbf{h}}_A^{(i)}, \tilde{\mathbf{h}}_A^{(j)}\big)\Big)$$

where $A \in \{F, T, S\}$ indexes frequency, temporal, and spatial axes, $\tilde{\mathbf{h}}_A^{(i)}$ denotes the axis-aligned summary from branch i after interpolation to a common resolution, and $\cos(\cdot,\cdot)$ is cosine similarity.

Fine-Tuning. For downstream tasks, token representations from each branch are globally averaged to form one vector per axis, summed across axes, and passed to a two-layer classification head.

4. EXPERIMENTS

4.1. Datasets

Pretraining. HCMT is pretrained on: (1) *PhysioNet Motor Imagery* [16], 64-channel EEG from 109 subjects performing motor tasks; (2) *SEED* [2], 62-channel EEG from 15 subjects viewing emotion-inducing videos.

Evaluation. We assess on: (1) *PhysioNet P300 Speller* [4], 64-channel ERP recordings from spelling tasks; (2) *BCI Competition IV-2a* [17], 22-channel motor imagery EEG from 9 subjects; (3) *Sleep-EDF Expanded* [18], two-channel overnight EEG from 100 subjects.

4.2. Experimental Setup

Preprocessing. EEG signals were converted to mV, and resampled to 256 Hz. Channels were reordered into 10–20 layout and zero-padded to 64.

Architecture. The sinc-based filterbank consists of 8 learnable band-pass filters with a kernel length of 251, initialized to span 0.5–40 Hz frequency range corresponding to canonical EEG rhythms. encoder has a 32-channel 3D convolution, followed by two parallel 64-channel 2D convolutional layers.

Transformer. Each branch uses 4 encoder blocks with 4 attention heads and embedding dimension $D = 64$, along with 8 temporal patches. Each branch includes a 2-layer decoder with 2 heads for reconstruction.

Training. Models are trained for 40 epochs with a batch size of 8. The loss combines reconstruction, decorrelation with $\lambda_1 = 0.1$, and diversity with $\lambda_2 = 0.1$. During self-supervised pretraining, the entire encoder is optimized end-to-end using stochastic gradient descent with a learning rate of 1e-3, while fine-tuning proceeds with a reduced learning rate of 1e-4.

Evaluation. Models are evaluated using leave-one-subject-out cross-validation for PhysioNet P300 and BCI-IV-2a, and 10-fold cross-validation for Sleep-EDF Expanded. Metrics are averaged over last 10 training epochs per subject or fold.

Table 1. Performance comparison across EEG datasets

Model	PhysioNet P300		BCI IV-2a		Sleep-EDF	
	Balanced Accuracy ↑	**Weighted F1 ↑**	**Balanced Accuracy ↑**	**Weighted F1 ↑**	**Balanced Accuracy ↑**	**Weighted F1 ↑**
DeepConvnet	64.70 ± 7.15	64.29 ± 7.32	70.88 ± 2.03	70.86 ± 2.03	68.74 ± 3.87	75.45 ± 3.83
EEGNet	64.88 ± 5.15	64.81 ± 5.24	73.13 ± 2.93	73.14 ± 2.95	71.60 ± 4.59	78.34 ± 4.88
BIOT	54.28 ± 4.62	52.10 ± 4.37	46.15 ± 3.08	43.46 ± 2.21	65.42 ± 3.64	73.78 ± 4.26
EEGPT	65.93 ± 1.42	63.73 ± 1.74	57.36 ± 1.68	57.02 ± 1.96	68.36 ± 2.82	73.17 ± 3.57
BENDR	65.12 ± 2.29	61.39 ± 2.66	49.20 ± 1.43	48.69 ± 1.78	66.74 ± 1.87	74.51 ± 1.25
HCMT	$\mathbf{72.19 \pm 7.26}$	$\mathbf{72.22 \pm 7.36}$	$\mathbf{75.37 \pm 3.14}$	$\mathbf{75.37 \pm 3.15}$	$\mathbf{73.08 \pm 6.59}$	$\mathbf{80.18 \pm 4.81}$

4.3. Performance Analysis

Comparison to baselines. We compare HCMT with convolutional models DeepConvNet [6] and EEGNet [5], transformer models EEGPT [13] and BIOT [19], and hybrid BENDR [12], in Table 1. On P300, CNNs like EEGNet and DeepConv capture short transients but miss global spatial patterns, while HCMT integrates spatial and temporal attention to better capture ERP response. On IV-2a, each trial provides a short window, limiting benefit of global attention. Transformer baselines like BENDR and BIOT transfer poorly under these conditions, while HCMT leverages spectral filtering and convolution to capture oscillations. On Sleep-EDF, EEGPT handles long two-channel sequences well, yet HCMT surpasses it by unifying temporal specialization with reconstruction. Across all datasets, HCMT achieves highest accuracy and F1, with its higher variance reflecting stronger capacity to respond to subject differences.

Prediction Visualization. For PhysioNet P300, we plot *ERP confidence curves*, which represent the model's predicted probability of a target event as a function of time. These curves provide a probabilistic analogue of the classical P300 response, where brain activity typically increases around $\sim$300 ms following a target stimulus. As shown in Fig. 3, HCMT exhibits a sharp rise in confidence near this latency, reflecting accurate capture of event-related dynamics.

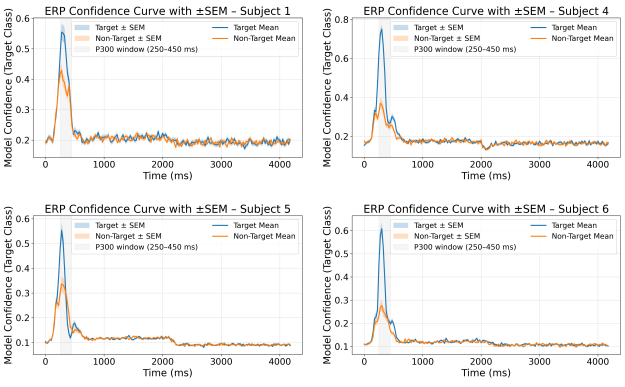

Fig. 3. ERP Confidence Curves for PhysioNet P300

4.4. Ablation Study

We assess the contribution of major components and the sensitivity of weighting parameters λ_1 and λ_2 on PhysioNet P300, BCI-IV 2a, and Sleep-EDF. All variants are pretrained on the same corpus and fine-tuned using identical protocols.

Variants. We evaluate several design choices: **Single-Branch**, where the multibranch Transformer is replaced by a single branch over concatenated axis-specific features; **No Axis Reconfiguration**, which removes axis-specific up/downsampling before Transformer encoding; **No Axis Attention**, which disables axis-specific attention in the CNN encoder; λ_1 and λ_2 reduced, where decorrelation and diversity weights are set to 0.05 and λ zero, where λ_1=0, λ_2=0.

Findings. As shown in Table 2, performance declines most under the Single-Branch variant, confirming the necessity of parallel axis-specific branches. Removing axis reconfiguration or attention also lowers accuracy, showing their effect on performance. Regularization with $\lambda_1 = \lambda_2 = 0.1$ yields the best trade-off, since both smaller and larger values degrade results and zeroing them weakens generalization. These consistent patterns across all three datasets underscore the robustness of HCMT's design.

Table 2. Ablation study. Balanced Accuracy is reported.

Variant	P300	IV-2a	Sleep-EDF
Full HCMT	**72.2**	**75.4**	**73.1**
Single-Branch	65.6	71.3	68.4
No Axis Reconfiguration	68.3	73.3	69.6
No Axis Attention	68.5	71.9	70.7
λ_1 Reduced	71.2	74.7	71.9
λ_2 Reduced	70.9	74.8	72.3
λ_1=0, λ_2=0	70.1	74.2	71.7
λ_1=0.2, λ_2=0.2	70.4	74.0	71.3

5. CONCLUSION

We introduced HCMT, a self-supervised framework that models spatial, spectral, and temporal EEG structure using axis-specific attention and Transformer branches. RMS-weighted reconstruction and dual diversity losses promote complementary representations, while spatial reordering and rotation improve robustness across montages. HCMT outperforms strong baselines across three benchmarks, highlighting the benefits of modular, axis-aware modeling. Future work will explore scaling to broader datasets and extending HCMT to multimodal neural signals.

Compliance with Ethical Standards

This is a numerical simulation study for which no ethical approval was required.

Conflicts of Interest Disclosure

The authors have no relevant financial or nonfinancial interests to disclose.

Acknowledgement

This work is supported in part by the National Science Foundation (NSF) grant IIS-2451436 and Commonwealth Cyber Initiative grant HC-4Q24-059.

6. REFERENCES

[1] F. Lotte, L. Bougrain, A. Cichocki, et al., "A review of classification algorithms for eeg-based brain–computer interfaces: a 10 year update," *Journal of Neural Engineering*, vol. 15, no. 3, pp. 031005, apr 2018.

[2] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks," *IEEE Transactions on Autonomous Mental Development*, vol. 7, no. 3, pp. 162–175, 2015.

[3] D. Wu, X. Jiang, and R. Peng, "Transfer learning for motor imagery based brain–computer interfaces: A tutorial," *Neural Networks*, vol. 153, pp. 235–253, 2022.

[4] L. Citi, R. Poli, and C. Cinel, "Documenting, modelling and exploiting p300 amplitude changes due to variable target delays in donchin's speller," *Journal of Neural Engineering*, vol. 7, no. 5, pp. 056006, sep 2010.

[5] V. J. Lawhern, A. J. Solon, N. R. Waytowich, et al., "Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces," *Journal of Neural Engineering*, vol. 15, no. 5, pp. 056013, jul 2018.

[6] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, et al., "Deep learning with convolutional neural networks for eeg decoding and visualization," *Human Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.

[7] E. Eldele, Z. Chen, C. Liu, et al., "An attention-based deep learning approach for sleep stage classification with single-channel eeg," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 809–818, 2021.

[8] Z. Wan, M. Li, S. Liu, et al., "Eegformer: A transformer–based brain activity classification method using eeg signal," *Frontiers in Neuroscience*, vol. Volume 17 - 2023, 2023.

[9] Y. Roy, H. Banville, I. Albuquerque, et al., "Deep learning-based electroencephalography analysis: a systematic review," *Journal of Neural Engineering*, vol. 16, no. 5, pp. 051001, aug 2019.

[10] F. Fahimi, Z. Zhang, W. B. Goh, et al., "Inter-subject transfer learning with an end-to-end deep convolutional neural network for eeg-based bci," *Journal of Neural Engineering*, vol. 16, no. 2, pp. 026007, jan 2019.

[11] G. Xu, X. Shen, S. Chen, et al., "A deep transfer convolutional neural network framework for eeg signal classification," *IEEE Access*, vol. 7, pp. 112767–112776, 2019.

[12] D. Kostas, S. Aroca-Ouellette, and F. Rudzicz, "Bendr: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of eeg data," *Frontiers in Human Neuroscience*, vol. Volume 15 - 2021, 2021.

[13] G. Wang, W. Liu, Y. He, et al., "Eegpt: Pretrained transformer for universal and reliable representation of eeg signals," in *Advances in Neural Information Processing Systems*. 2024, vol. 37, pp. 39249–39280, Curran Associates, Inc.

[14] M. N. Mohsenvand, M. R. Izadi, and P. Maes, "Contrastive representation learning for electroencephalogram classification," in *Proceedings of the Machine Learning for Health NeurIPS Workshop*. 2020, pp. 238–253, PMLR.

[15] A. Jaiswal, A. R. Babu, M. Z. Zadeh, et al., "A survey on contrastive self-supervised learning," *Technologies*, vol. 9, no. 1, 2021.

[16] G. Schalk, D. J. McFarland, T. Hinterberger, et al., "Bci2000: a general-purpose brain-computer interface (bci) system," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 6, pp. 1034–1043, 2004.

[17] M. Tangermann, K.-R. Müller, A. Aertsen, et al., "Review of the bci competition iv," *Frontiers in Neuroscience*, vol. Volume 6 - 2012, 2012.

[18] B. Kemp, A. H. Zwinderman, B. Tuk, et al., "Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the eeg," *IEEE Transactions on Biomedical Engineering*, vol. 47, no. 9, pp. 1185–1194, 2000.

[19] C. Yang, M. Westover, and J. Sun, "Biot: Biosignal transformer for cross-data learning in the wild," in *Advances in Neural Information Processing Systems*. 2023, vol. 36, pp. 78240–78260, Curran Associates, Inc.