SMEEG: Spectral Mamba-Driven EEG Temporal and Frequency Representation Learning

Bingru Lin[1,2,†], Hanran Lin[1,2,†], Bo Zheng[1,2], Shaojun Zhu[1,2], Maonian Wu[1,2,*]

[1]*School of Information Engineering, Huzhou University , Huzhou, China*

[2]*Zhejiang Province Key Laboratory of Smart Management and Application of Modern Agricultural Resources, Huzhou, China*

***Abstract*—Electroencephalogram (EEG) signals reflect the brain's functional state, providing a reliable physiological basis for applications such as disease classification, emotion detection and cognitive state assessment. However, EEG representation learning faces challenges such as low signal-to-noise ratio, inter-individual feature variability and long-sequence modeling. To address these critical challenges, this study proposes learning EEG temporal and frequency representations based on Spectral Mamba global driving (SMEEG). The SMEEG includes a temporal and frequency domain feature fusion architecture, which encodes the temporal-domain and frequency-domain signals independently, and achieves feature alignment through shared positional encoding and normalization, effectively alleviating the feature quality problem caused by low signal-to-noise ratio. Depthwise Separable Wavelet (DSW) analysis is introduced to process the multi-scale features of EEG. Through learnable wavelet filters and an inverse transform mechanism, the limitations of traditional convolution in processing EEG signals with large amplitude variations and high individual differences are effectively solved. The Spectral Mamba (SpeMamba) model is based on the linear complexity state-space model of Mamba, enabling efficient long-sequence modeling. Experiments on 5 EEG datasets show that SMEEG effectively learns temporal-frequency representations under a SpeMamba-driven self-supervised framework. The extracted features allow downstream classifiers to reach supervised-level accuracy, confirming their potential for generalizable brain-computer interface (BCI) and clinical EEG analysis. The code is available at https://github.com/huaidan666/SMEEG.**

***Index Terms*—EEG classification, EEG representation learning, Brain-computer interface, Deep learning**

I. Introduction

Electroencephalography (EEG) is a non-invasive technique that records brain electrical activity using electrodes placed on the scalp. It reflects brain functional states and is widely used in clinical and brain-computer interface applications due to its non-invasiveness, low cost and ease of acquisition [1]. Brain-computer interfaces capture, process, and convert EEG signals, enabling users to communicate with and control external devices without muscular activity. These systems provide essential foundations for decoding cognitive processes and disease monitoring [2]. Their applications span several domains, such as epilepsy detection [3], emotion recognition [4] and motor imagery classification [5]. With the adoption of deep learning methods, EEG classification has advanced significantly in recent years. This has not only enhanced the precision and real-time capabilities of mental health interventions but also accelerated the realization of highly complex brain-computer interface architectures. Furthermore, it demonstrates immense potential in improving disease detection, laying the groundwork for future medical innovation [6]. However, the advancement of EEG classification research continues to face numerous challenges and difficulties, particularly concerning model generalization capabilities and performance enhancement. These challenges mainly arise from the inherent properties of EEG signals, including low signal-to-noise ratio, inter-individual feature variability and long-sequence modeling.

These challenges have prompted researchers to adopt deep learning methods to improve EEG classification performance and generalization. However, deep learning for EEG classification typically requires large amounts of expert-annotated data. The annotation cost of EEG data is high and time-consuming, and it is extremely impractical to rely entirely on manual annotation of a large amount of data [7]. Self-supervised learning (SSL) reduces reliance on labeled data by exploiting intrinsic supervision signals. In self-supervised learning, masked autoencoders hide part of the input and reconstruct it, forcing the model to learn inherent data structure and reducing reliance on labeled samples [7], [8]. For example, TS-MAE is a masked autoencoder for learning time series representations. By masking and reconstructing EEG in the time dimension, it can learn robust feature representations [8]. The self-supervised learning method of adaptive frequency-time attention Transformer reduces the dependence on labeled data through masking and reconstruction strategies, enhances the extraction of EEG features, and thus improves the accuracy of epilepsy prediction and classification [7]. The Transformer model, with its powerful global sequence modeling capabilities, has shown great potential in small to medium scale EEG analysis, especially in processing non-stationary and high-dimensional EEG data [9]. However, its inherent quadratic computational complexity still limits its ability to process ultra-long sequences, such as those exceeding 10,000 time steps [10]. Existing Transformer models also face generalization challenges, particularly across tasks and subjects [9].

Recently, Mamba and its derivative models have shown clear advantages for long-sequence EEG modeling and have been increasingly applied to EEG analysis. This addresses the limitations of traditional Transformer models in modeling long-sequence dependencies [10]. Although performance still

† These authors contributed equally to this work.

* Maonian Wu is the corresponding author: wmn@zjhu.edu.cn.

depends strongly on the self-supervised framework design [10], the unique mechanisms and empirical success of Mamba-based models make them promising for future EEG research.

Therefore, this study proposes a novel EEG temporal and frequency representation learning approach based on Spectral Mamba global driving (SMEEG) to address challenges, including low signal-to-noise ratio, significant inter-individual variability in features and long-sequence modeling. Self-supervised learning was performed pre-training on 5 publicly available EEG datasets. This pretraining enabled more accurate representations for task extraction. Finally, a linear probing method was used for EEG classification, which achieved the ability to extract high-level abstract features across spatiotemporal dimensions.

The principal contributions of this study are as follows:

- A global EEG modeling mechanism based on Spectral Mamba (SpeMamba) is proposed. It captures long-sequence dependencies with a linear-complexity state-space structure and integrates channel grouping, residual connections, and GroupNorm-SiLU projection for efficient global and cross-channel topology modeling.
- A temporal and frequency consistency representation learning framework is designed, combining temporal and frequency branches with shared positional encoding and normalization for cross-domain feature alignment. This mechanism effectively fuses complementary temporal-frequency information, enhancing cross-task robustness and signal-to-noise ratio.
- A depthwise separable discrete wavelet convolution module is introduced to capture localized spectral-temporal features during embedding. By combining learnable filters and inverse wavelet reconstruction, the module achieves multi-scale representation while preserving channel correlation and mitigating EEG amplitude variation and individual variability.
- Extensive experiments on 5 EEG datasets demonstrate that SMEEG outperforms representative self-supervised baselines and achieves competitive performance across diverse EEG classification tasks.

II. Related Work

A. EEG Representation Learning

EEG representation learning, a key technique for understanding and interpreting brain activity, has advanced significantly in cognitive recognition, especially with deep learning models [11]. Due to its high dimensionality, non-stationarity and noise susceptibility, EEG remains challenging for effective and generalizable feature extraction [12]. Researchers have developed advanced representation learning methods, including self-supervised learning, multimodal fusion and hybrid deep learning architectures, to improve EEG feature extraction and classification performance. For example, Liu et al. proposed a contrastive self-supervised framework for robust EEG sentiment recognition. This framework effectively utilizes unlabeled and sparsely labeled EEG data and has achieved favorable results in emotion recognition tasks [13]. He et al. proposed EEG-EMG FAConformer for the fusion of EEG and EMG signals, which holds significant importance for brain-computer interface applications in motor function rehabilitation [14]. Redwan et al. proposed a hybrid CNN-BiLSTM model for EEG emotion detection that extracts features from Power Spectral Density (PSD) signals and improves contextual modeling of sequential data [15]. These studies show that different deep learning frameworks provide new solutions for EEG feature extraction and classification.

B. EEG Global Long-sequence Modeling

EEG global temporal modeling aims to capture dynamic brain activity and complex spatiotemporal dependencies in long-sequence data, which are critical for cognitive understanding, disease diagnosis and brain-computer interface development. Traditional Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are effective in local feature extraction but fail to model long-range dependencies efficiently. Recent advances in Transformer architectures and State-space Models (SSMs) have shown strong capability in global sequence modeling, motivating their application to EEG analysis.

Transformer-based global modeling relies on self-attention, the core mechanism of the Transformer architecture. This mechanism enables dependencies between arbitrary sequence positions, effectively capturing long-range dependencies and global patterns in EEG data. For example, EEG-Deformer introduced a dense convolutional Transformer structure, which adopts a top-down hierarchical temporal learning and a multi-level information purification module, significantly improving performance in attention detection, fatigue and cognitive load detection tasks [16]. Huang et al. employed a Transformer-based self-supervised method that reduces label dependence through masking reconstruction and improves epilepsy prediction and classification accuracy [7]. Li et al. proposed EEGPT, a large-scale pre-trained Transformer framework designed for generalizable and reliable EEG representation learning [17].

Based on state-space models, particularly the Mamba architecture, efficient global modeling is achieved. Mamba employs a selective state-space mechanism to dynamically adjust its parameters according to the input, providing new opportunities for global temporal modeling of EEG signals. In BCI motor imagery tasks, Yang et al. employed a spatio-temporal Mamba encoder to capture both cross-channel and cross-temporal dependencies [18]. EEGM2 integrates Mamba-2 with a self-supervised learning strategy. Through multi-branch spatiotemporal coding and a loss function that preserves spectral information, it achieves efficient long-sequence EEG modeling, improves cross-subject generalization and reduces computational overhead [10]. However, the Mamba-EEG frameworks in the aforementioned studies are limited by resolution and computational efficiency. This study uses SpeMamba to divide channels into groups and perform state-space modeling over time series, improving global-local co-representation and cross-channel dependency modeling over two-dimensional electrode topology, thereby extending the applicability of Mamba-based EEG analysis.

III. Method

A. SMEEG Architecture

This study develops a self-supervised representation learning framework called SMEEG, which integrates temporal and frequency fusion with state-space modeling. The framework includes four stages: channel reconstruction and masking, deep separable wavelet and temporal-frequency domain feature modeling, SpeMamba global modeling and encoding, and masked reconstruction. Channel reconstruction and masking use a semantic order-preserving mechanism to preprocess raw EEG and construct a more structured input representation for subsequent feature extraction. The dual-branch structure in the temporal and frequency domain processes non-stationary temporal domain signals and the results of the Fourier transform in the frequency domain. In the temporal domain, depthwise separable wavelet convolution is added to extract non-stationary features at different temporal scales, achieving complementary fusion of temporal and frequency information. The framework employs a SpeMamba global long-sequence encoder based on state-space modeling, combined with channel grouping and a linear-complexity recursive scanning mechanism to model long-range dependencies and cross-channel topology in EEG. The multi-layer perceptron reconstruction head completes the masking reconstruction task for the mask fragment. The SMEEG framework is shown in Fig. 1.

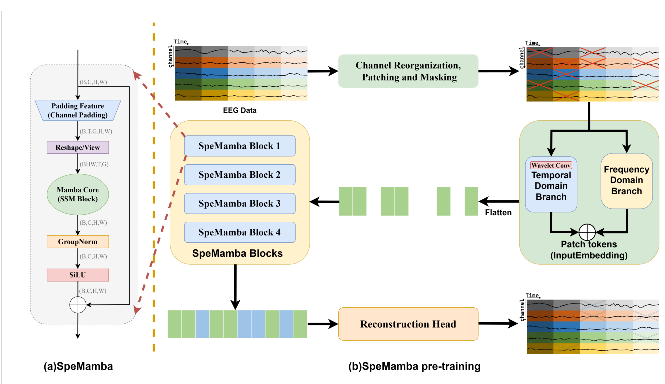

Fig. 1. Overview of SMEEG.

In the SMEEG framework, SMEEG first takes raw EEG inputs and then uses cross-channel convolution and semantic order preservation modules to reorganize, segment and mask its channels. Secondly, the temporal and frequency domain feature fusion module branches into separate temporal-domain and frequency-domain paths for feature extraction. Next, the fused temporal and frequency embeddings are fed into a multi-level SpeMamba encoder to construct global contextual representations and integrate cross-scale features. Finally, the reconstruction head, implemented as a multi-layer perceptron, performs the masked reconstruction task. The model parameters are optimized by computing the reconstruction loss, completing the self-supervised representation learning of EEG. This framework provides more discriminative features for EEG classification and recognition tasks.

The core design focuses on feature modeling and global long-sequence encoding. SMEEG introduces a cross-channel recombination mechanism and a sub-sequence masking strategy to enhance the spatial semantics of raw EEG signals and introduce localized temporal perturbations, thereby improving self-supervised representation learning. A lightweight MLP serves as the decoder for mask reconstruction, restoring masked features and reinforcing temporal-frequency consistency. The feature modeling and long-sequence encoding modules employ temporal-frequency fusion, wavelet-based convolution and state-space encoding to address low signal quality, inter-subject variability and long-range dependencies in EEG. These two stages constitute the core of the SMEEG and are essential for enhancing discriminative feature learning and improving EEG classification accuracy.

B. Temporal and Frequency Domain Feature Fusion Module

A key challenge in EEG representation learning is to effectively capture the temporal and spectral dependencies of non-stationary neural signals. To address this limitation, we introduce a temporal and frequency domain feature fusion module that jointly models the temporal dynamics and spectral structure of EEG signals. Unlike conventional single-domain approaches, the proposed module adopts a dual-branch design that separately encodes temporal-domain sequences and frequency-domain representations, followed by a fusion mechanism for cross-domain alignment. The temporal and frequency domain feature fusion module is shown in Fig. 2.

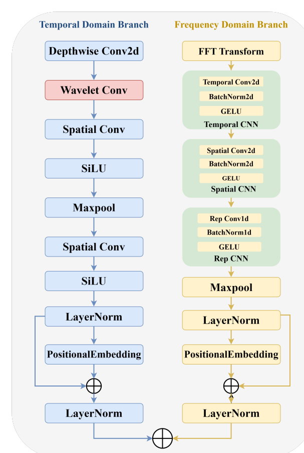

Fig. 2. Temporal and frequency domain.

Formally, given an input EEG segment $X \in \mathbb{R}^{B\times C\times T}$, temporal embeddings are derived as $Z_t = \text{Embed}_{\text{time}}(X)$, while frequency representations are obtained via a fast Fourier transform $Z_f = \text{Embed}_{\text{freq}}(\text{FFT}(X))$. To enhance local temporal sensitivity, a wavelet-based convolutional augmentation $Z_t' = \text{WTConv}(Z_t)$ is applied before concatenating the 2 representations into a unified embedding $Z_{\text{tf}} = \text{Concat}(Z_t', Z_f)$.

This fusion strategy enables the model to leverage both transient time-domain fluctuations and sustained spectral rhythms, improving robustness against noise and inter-subject variability. Furthermore, shared positional encoding and normalization across both branches ensure consistent cross-domain feature alignment, improving generalization across tasks and recording conditions. This fusion mechanism provides a foundation for learning more stable and discriminative temporal and frequency representations in self-supervised EEG modeling.

C. Depthwise Separable Wavelet Module

Traditional convolutional networks often struggle to effectively represent EEG data due to their limited receptive fields and inability to capture frequency-dependent patterns. To address this issue, we propose a Depthwise Separable Wavelet (DSW) module that introduces wavelet-domain processing into convolution. This module decomposes EEG features into multi-scale spectral sub-bands while maintaining computational efficiency through depthwise separable design. The DSW module is shown in Fig. 3.

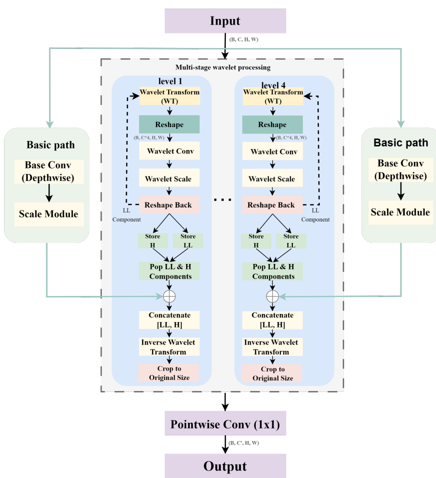

Fig. 3. Depthwise separable wavelet module.

Given an input feature tensor $X \in \mathbb{R}^{B \times C \times H \times W}$, each channel is independently processed using a Discrete Wavelet Transform (DWT) constructed with Daubechies (db1) wavelets:

$$X_{\text{WT}} = \text{DWT}(X) \rightarrow \{LL, LH, HL, HH\} \tag{1}$$

Where LL represents the low-frequency approximation component, and LH, HL and HH represent high-frequency detail components along the horizontal, vertical and diagonal orientations, respectively.

These sub-bands are concatenated and processed by grouped convolution before inverse DWT reconstruction:

$$X_{\text{rec}} = \text{IDWT}\left(\text{GroupedConv}\left([LL, LH, HL, HH]\right)\right) \tag{2}$$

A subsequent 1×1 pointwise convolution fuses channel-wise features:

$$X_{\text{out}} = \text{Conv}_{1\times 1}(X_{\text{rec}}) \tag{3}$$

By explicitly modeling spectral structures, DSW improves the representation of both low-frequency oscillatory components and high-frequency transient details inherent in EEG. The orthogonal nature of wavelet bases ensures stable decomposition and reconstruction, minimizing information loss and enhancing interpretability. As a result, the module achieves effective multi-scale representation, enhanced cross-channel correlation, and improved cross-subject generalization, addressing the limitations of standard CNN-based EEG processing.

D. SpeMamba Module

To capture long-range temporal dependencies and global contextual information in EEG sequences, we employ SpeMamba, an encoder based on a structured state-space formulation. This design balances efficiency and representational capacity by combining selective state updates with grouped channel modeling, enabling scalable modeling of long EEG sequences. The SpeMamba module is shown in Fig. 4.

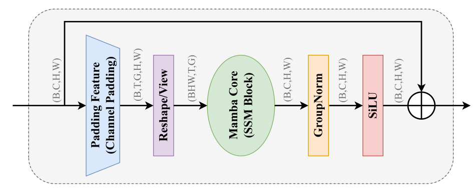

Fig. 4. SpeMamba module.

Given an input sequence x_t, SpeMamba updates the hidden state through recursive state transitions defined by:

$$y_t = C_t(A_t h_{t-1} + B_t x_t)\,, \quad t \in \{1, \ldots, T\} \tag{4}$$

Where t denotes the current time step and T is the sequence length. A_{t}, B_{t} and C_{t} are dynamic parameter matrices learned adaptively across timesteps. The projected output is then normalized and passed through a non-linear activation $x_{\text{proj}} = \text{SiLU}(\text{GroupNorm}(y_t))$.

A residual connection is further introduced to stabilize gradient propagation and preserve local consistency:

$$x_{\text{out}} = x + x_{\text{proj}} \tag{5}$$

Unlike Transformer architectures whose self-attention scales quadratically with sequence length, SpeMamba achieves linear-time modeling while retaining global dependency capture. Additionally, channel grouping allows efficient modeling of EEG electrode topologies, and the integration of residual projection enhances local representation learning. This combination ensures that the encoder can efficiently capture long-range dependencies across both time and space, making it particularly suitable for large-scale EEG analysis.

IV. Experiments

A. Datasets and Experimental Settings

This study employed 5 publicly available EEG datasets for model evaluation: STEW [19], Crowdsourced [20], DREAMER [21], TUAB [22] and PhysioNetP300 [23]. These datasets span diverse paradigms and cognitive tasks, with sampling rates of 128-256 Hz and 14-64 channels. They represent EEG application scenarios ranging from emotion detection to clinical diagnosis, providing a basis for validating the model's robustness and generalization under multi-task and multi-domain settings. Detailed descriptions of the datasets are shown in Table 1.

TABLE I
OVERVIEW OF DATASETS

Datasets	Classification Task	Freq.	Channels	Duration	Samples
STEW	Mental workload	128Hz	14	2s	26,136
Crowdsourced	Eyes open/close	128Hz	14	2s	12,296
DREAMER	Emotion detection	128Hz	14	2s	77,910
TUAB	Abnormal EEG	256Hz	16	10s	409,455
PhysioNetP300	Target detection	256Hz	64	2s	2,532

This study implemented the self-supervised representation learning model SMEEG for EEG signals in the PyTorch framework. All experiments were conducted on 2 GPUs. In terms of optimization strategy, the AdamW optimizer was adopted, with a learning rate of 1e-3, a batch size of 256 and 500 training epochs. The self-supervised masking ratio was set to 0.5, and the feedforward network dimension in the predictor was fixed at 64. All experiments were independently repeated under 5 different random seeds. The mean and standard deviation of each metric were reported, and parameter tuning were completed based on the validation-set performance.

B. Performance Metrics

To evaluate SMEEG across EEG classification tasks, this study uses 2 standard metrics: Accuracy (Acc) and Area Under the Receiver Operating Characteristic Curve (AUROC).

Accuracy is widely used in EEG classification and reflects the model's overall ability to identify all categories. It represents the proportion of correctly classified samples:

$$\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

Where TP, TN, FP and FN represent the number of true positive, true negative, false positive and false negative samples, respectively.

AUROC measures overall classification ability across different thresholds and is a key metric in classification tasks. It summarizes the ROC curve as a scalar in [0,1]. Values closer to 1 indicate stronger discrimination between positive and negative classes. AUROC is defined in integral form as:

$$\text{AUROC} = \int_0^1 \text{TPR}(\text{FPR}^{-1}(x))\, dx \tag{7}$$

Where TPR (True Positive Rate) and FPR (False Positive Rate) respectively represent the model's sensitivity and specificity under different threshold settings.

C. Experimental Results

This paper systematically compares the proposed SMEEG model with self-supervised EEG representation learners, including BIOT [24] and BENDR [25], MAEEG [26] and EEG2Rep [27], as well as the supervised classifier KnowEEG [28], on 5 publicly available EEG datasets. The comparative experiments are shown in Table 2.

TABLE II
EXPERIMENTAL RESULTS

Datasets	Methods	Acc(%)	AUROC(%)
STEW	BENDR	63.03±1.07	63.03±1.07
	MAEEG	67.99±1.86	68.58±1.86
	BIOT	67.54±2.08	67.69±3.60
	KnowEEG	**77.86±0.19**	**85.96±0.8**
	EEG2Rep	73.60±1.47	74.40±1.50
	SMEEG	75.81±2.06	83.58±3.48
Crowdsourced	BENDR	83.78±2.35	63.80±2.63
	MAEEG	86.35±3.50	86.21±3.41
	BIOT	87.95±3.52	87.78±3.09
	KnowEEG	94.53±0.35	98.27±0.13
	EEG2Rep	94.13±2.11	94.13±2.17
	SMEEG	**94.63±1.94**	**98.54±2.28**
DREAMER	BENDR	54.45±2.11	53.02±3.11
	MAEEG	53.63±2.61	52.08±2.36
	BIOT	53.45±2.01	53.53±2.13
	KnowEEG	**61.53±2.59**	55.29±2.58
	EEG2Rep	60.37±1.52	59.42±1.45
	SMEEG	60.84±2.74	**59.76±1.81**
TUAB	BENDR	76.96±3.98	83.97±3.44
	MAEEG	77.56±3.56	86.56±3.33
	BIOT	79.21±2.15	87.42±2.01
	KnowEEG	80.18±0.24	87.71±0.23
	EEG2Rep	80.52±2.22	88.43±3.09
	SMEEG	**80.78±3.41**	**88.73±3.11**
PhysioNetP300	BENDR	61.14±1.18	65.88±1.63
	MAEEG	-	-
	BIOT	54.85±3.25	53.08±3.33
	KnowEEG	-	-
	EEG2Rep	59.02±2.54	62.24±2.41
	SMEEG	**62.48±2.76**	**66.32±2.98**

On STEW, SMEEG attained 75.81% accuracy and 83.58% AUROC, which are improvements of 2.21% and 9.18%, respectively, over the EEG2Rep model, while performing marginally below KnowEEG. On Crowdsourced, SMEEG leads all models with the highest accuracy and AUROC, confirming its discriminative ability and generalization stability in noisy, weakly labeled supervision scenarios. On DREAMER, SMEEG achieved 60.84% accuracy and 59.76% AUROC, slightly outperforming EEG2Rep while remaining marginally below the supervised KnowEEG model. These results suggest that SMEEG effectively captures non-stationary EEG dynamics associated with emotional variations. On TUAB, SMEEG achieved 80.78% accuracy and 88.73% AUROC, slightly outperforming EEG2Rep by 0.26% in accuracy and 0.30% in AUROC. This indicates robust modeling of pathological EEG patterns in complex medical settings. SMEEG also achieved the best performance on PhysioNetP300, underscoring its advantages in low-sample, attention-evoked signal modeling.

Although SMEEG performs slightly below the supervised model KnowEEG on the mental workload classification task (STEW), it remains the best-performing self-supervised approach on this dataset and achieves comparable or superior results across the other 4 EEG scenarios, demonstrating stronger overall cross-task robustness. This difference reflects the distinction between the supervised and self-supervised learning paradigms. As a knowledge-driven supervised framework, KnowEEG constructs interpretable statistical features and brain-region connectivity representations, and uses a Fusion Forest classifier that is particularly effective when neural cues are spatially structured and cognitively stable. In contrast, SMEEG adopts an end-to-end self-supervised framework that integrates SpeMamba-based state-space modeling with multi-scale wavelet temporal-frequency fusion. This design enables long-range dependency modeling and robust feature learning under noisy, cross-subject and label-limited conditions without relying on explicit neuroanatomical priors. Consequently, SMEEG exhibits superior generalization and adaptability in complex EEG applications such as emotion recognition, clinical diagnosis and event-related potential detection.

In summary, SMEEG achieves the best overall performance among the compared self-supervised baselines in terms of accuracy and AUROC on most datasets, and remains competitive with the supervised baseline. The evaluated tasks include emotion recognition, resting-state consciousness classification, clinical EEG diagnosis and event-related potential detection. These results suggest the broad adaptability and cross-task transferability of the proposed structural design for multi-paradigm EEG representation learning.

D. Ablation Study

To evaluate the contribution of each component, ablation experiments were conducted on 3 datasets including STEW, Crowdsourced and PhysioNetP300. A progressive modular integration strategy was used, sequentially introducing the Wavelet Convolution Block (WCB), Frequency Domain Branch (FDB), and Spectral Mamba (SM) to assess the impact of each component. Ablation results are shown in Table 3.

TABLE III
ABLATION STUDY OF SMEEG COMPONENTS

Component			STEW		Crowdsourced		PhysioNetP300	
WCB	**FDB**	**SM**	Acc	AUROC	Acc	AUROC	Acc	AUROC
			73.60	74.40	94.13	94.13	59.02	62.24
✓			73.69	79.35	94.24	97.38	60.37	64.41
	✓		74.01	78.81	94.50	96.41	62.47	64.11
		✓	72.23	79.79	94.40	97.37	61.33	64.55
✓	✓		74.39	80.10	94.15	98.08	62.29	63.44
✓		✓	74.45	80.39	94.56	98.52	62.43	65.89
	✓	✓	74.91	83.47	94.60	**99.41**	61.83	65.02
✓	✓	✓	**75.81**	**83.58**	**94.63**	98.54	**62.48**	**66.32**

The combination of all components provides the most stable and optimal performance results in the dataset. The SM block improves the AUROC by 5.39%, 3.24% and 2.31% on the STEW, Crowdsourced and PhysioNetP300, respectively, indicating enhanced long-sequence modeling for cognitive-load and clinical EEG patterns. The WCB improves accuracy by an average of 0.73% over the baseline on the PhysioNetP300 and Crowdsourced. Its multi-resolution wavelet mechanism enhances sensitivity to rhythmic changes in resting-state EEG, such as alpha wave synchronization. The FDB block improves accuracy by up to 3.45% and AUROC by 1.87% on PhysioNetP300, significantly enhancing the stability of spectral representations in high-noise environments. Although each block contributes positively on its own, optimal performance is achieved when they are used together, improving temporal-frequency consistency, sequence representation capacity, and generalization. These suggest that the three components are complementary and that their combination leads to more stable overall performance across datasets.

E. Feature Visualizations

T-SNE was used to visualize deep feature maps learned by different models [29]. The distribution of features learned by the baseline model EEG2Rep and the proposed SMEEG model on 5 datasets is shown in Fig. 5.

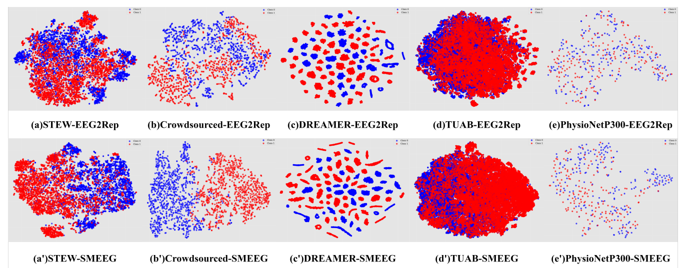

Fig. 5. T-SNE comparison of feature distributions learned by EEG2Rep and SMEEG. Different colors represent task labels in each dataset.

As shown in Fig. 5 (a-e), EEG2Rep exhibits substantial cluster overlap and dispersed intra-class distributions, indicating limited discriminability across tasks. In contrast, Fig. 5 (a'-e') shows that SMEEG produces well-separated feature clusters with compact intra-class structures, demonstrating superior separability. These visual results suggest that SMEEG, driven by temporal-frequency fusion and SpeMamba-based global modeling, effectively captures task-relevant information and learns highly discriminative EEG representations.

F. Confusion Matrix

To further evaluate classification performance, confusion matrices were computed for EEG2Rep and SMEEG on all 5 test datasets. The diagonal elements indicate correctly classified samples, whereas the off-diagonal elements indicate misclassified samples. The results are shown in Fig. 6.

As shown in Fig. 6 (a-e), EEG2Rep exhibits class confusion on multiple datasets, especially TUAB and PhysioNetP300, where high off-diagonal values indicate limited discriminative ability. In contrast, Fig. 6 (a'-e') shows the superior performance of SMEEG. Across all 5 datasets, SMEEG shows darker diagonals and lighter off-diagonal regions, implying higher true positive rates and lower false positive/false negative rates. These results indicate that SMEEG outperforms the baseline across all classification tasks. Its strong representation capability enables more accurate discrimination of EEG patterns, thereby improving classification accuracy.

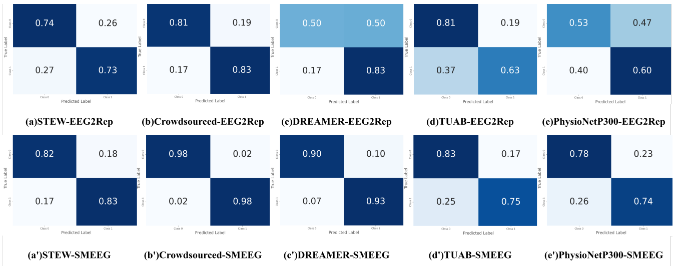

Fig. 6. Confusion matrices of EEG2Rep and SMEEG. Higher diagonal values indicate better classification accuracy.

V. Conclusion

This study introduces a self-supervised representation learning framework that integrates temporal and frequency fusion and state-space modeling. This framework extends the SMEEG backbone through joint temporal and frequency embeddings, depthwise separable wavelet convolution and a SpeMamba-based encoder. Experiments on multiple benchmark datasets show that the proposed approach achieves competitive accuracy and AUROC under self-supervised learning. The fusion of temporal and frequency features improves classification robustness and enhances the representation of complementary EEG information.

VI. Acknowledgment

This work was supported by the Postgraduate Research and Innovation Project of the College of Information Engineering, Huzhou University (No. 2025XGKC05), the Zhejiang Provincial College Students' Science and Technology Innovation Program (No. 2025R446C029) and the Science and Technology Plan Project of Huzhou (No. 2024GN04).

References

[1] J. Wang, S. Zhao, Z. Luo, Y. Zhou, S. Li, and G. Pan, "Eegmamba: An eeg foundation model with mamba," *Neural Networks*, p. 107816, 2025.

[2] R. Vishwakarma, H. Khwaja, V. Samant, P. Gaude, M. Gambhir, and S. Aswale, "Eeg signals analysis and classification for bci systems: a review," in *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*. IEEE, 2020, pp. 1–6.

[3] C. Tian and F. Zhang, "Eeg-based epilepsy detection with graph correlation analysis," *Frontiers in Medicine*, vol. 12, p. 1549491, 2025.

[4] J. Wang, X. Ning, W. Xu, Y. Li, Z. Jia, and Y. Lin, "Multi-source selective graph domain adaptation network for cross-subject eeg emotion recognition," *Neural Networks*, vol. 180, p. 106742, 2024.

[5] X. Wang, W. Yang, W. Qi, Y. Wang, X. Ma, and W. Wang, "Starnet: A spatio-temporal and riemannian network for high-performance motor imagery decoding," *Neural Networks*, vol. 178, p. 106471, 2024.

[6] X.-Y. Liu, W.-L. Wang, M. Liu, M.-Y. Chen, T. Pereira, D. Y. Doda, Y.-F. Ke, S.-Y. Wang, D. Wen, X.-G. Tong *et al.*, "Recent applications of eeg-based brain-computer-interface in the medical field," *Military Medical Research*, vol. 12, no. 1, p. 14, 2025.

[7] Y. Huang, Y. Chen, S. Xu, D. Wu, and X. Wu, "Self-supervised learning with adaptive frequency-time attention transformer for seizure prediction and classification," *Brain Sciences*, vol. 15, no. 4, p. 382, 2025.

[8] C. Pan, H. Lu, C. Lin, Z. Zhong, and B. Liu, "Set-pmae: spatial-spectral-temporal based parallel masked autoencoder for eeg emotion recognition," *Cognitive Neurodynamics*, vol. 18, no. 6, pp. 3757–3773, 2024.

[9] M. Cai and Y. Zeng, "Mae-eeg-transformer: A transformer-based approach combining masked autoencoder and cross-individual data augmentation pre-training for eeg classification," *Biomedical Signal Processing and Control*, vol. 94, p. 106131, 2024.

[10] J. Hong, G. Mackellar, and S. Ghane, "Eegm2: An efficient mamba-2-based self-supervised framework for long-sequence eeg modeling," *arXiv preprint arXiv:2502.17873*, 2025.

[11] Y. Qiao and Q. Zhao, "Sst-cram: spatial-spectral-temporal based convolutional recurrent neural network with lightweight attention mechanism for eeg emotion recognition," *Cognitive Neurodynamics*, vol. 18, no. 5, pp. 2621–2635, 2024.

[12] R. Vanidhasri and A. Kathirvel, "Improving eeg signal classification in brain-computer interfaces via grey wolf optimizer-based deep learning," in *2025 3rd International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*. IEEE, 2025, pp. 1312–1318.

[13] H. Liu, Y. Zhang, X. Chen, D. Zhang, R. Li, and T. Qin, "Self-supervised eeg representation learning for robust emotion recognition," *ACM Transactions on Sensor Networks*, vol. 20, no. 5, pp. 1–22, 2024.

[14] Z. He, M. Cai, L. Li, S. Tian, and R.-J. Dai, "Eeg-emg faconformer: Frequency aware conv-transformer for the fusion of eeg and emg," in *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2024, pp. 3258–3261.

[15] U. G. Redwan, T. Zaman, and H. B. Mizan, "Spatio-temporal cnn-bilstm dynamic approach to emotion recognition based on eeg signal," *Computers in Biology and Medicine*, vol. 192, p. 110277, 2025.

[16] Y. Ding, Y. Li, H. Sun, R. Liu, C. Tong, C. Liu, X. Zhou, and C. Guan, "Eeg-deformer: A dense convolutional transformer for brain-computer interfaces," *IEEE Journal of Biomedical and Health Informatics*, 2024.

[17] T. Yue, S. Xue, X. Gao, Y. Tang, L. Guo, J. Jiang, and J. Liu, "Eegpt: Unleashing the potential of eeg generalist foundation model by autoregressive pre-training," *arXiv preprint arXiv:2410.19779*, 2024.

[18] X. Yang and Z. Jia, "Spatial-temporal mamba network for eeg-based motor imagery classification," in *International Conference on Advanced Data Mining and Applications*. Springer, 2024, pp. 418–432.

[19] W. L. Lim, O. Sourina, and L. P. Wang, "Stew: Simultaneous task eeg workload data set," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 11, pp. 2106–2114, 2018.

[20] J. Gómez-Ramírez, S. Freedman, D. Mateos, J. L. Pérez-Velázquez, and T. Valiante, "Eyes closed or eyes open? exploring the alpha desynchronization hypothesis in resting state functional connectivity networks with intracranial eeg," *bioRxiv*, p. 118174, 2017.

[21] S. Katsigiannis and N. Ramzan, "Dreamer: A database for emotion recognition through eeg and ecg signals from wireless low-cost off-the-shelf devices," *IEEE journal of biomedical and health informatics*, vol. 22, no. 1, pp. 98–107, 2017.

[22] S. Lopez, G. Suarez, D. Jungreis, I. Obeid, and J. Picone, "Automated identification of abnormal adult eegs," in *2015 IEEE signal processing in medicine and biology symposium (SPMB)*. IEEE, 2015, pp. 1–5.

[23] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals," *circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[24] C. Yang, M. Westover, and J. Sun, "Biot: Biosignal transformer for cross-data learning in the wild," *Advances in Neural Information Processing Systems*, vol. 36, pp. 78 240–78 260, 2023.

[25] 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. 15, p. 653659, 2021.

[26] H.-Y. S. Chien, H. Goh, C. M. Sandino, and J. Y. Cheng, "Maeeg: Masked auto-encoder for eeg representation learning," *arXiv preprint arXiv:2211.02625*, 2022.

[27] N. Mohammadi Foumani, G. Mackellar, S. Ghane, S. Irtza, N. Nguyen, and M. Salehi, "Eeg2rep: enhancing self-supervised eeg representation through informative masked inputs," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024, pp. 5544–5555.

[28] A. Sahota, N. M. Foumani, R. Santos-Rodriguez, and Z. S. Abdallah, "Knoweeg: Explainable knowledge driven eeg classification," *arXiv preprint arXiv:2505.00541*, 2025.

[29] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.