Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

SeizureNet: Lightweight Epileptic Seizure Prediction via Mamba-Enhanced Network with Multi-Resolution Temporal Mixing and Dynamic Channel Graphs

Xinlei Chen
China University of Mining and Technology
Xiaobing Zhang
Suzhou Guangji Hospital
Dongming Zhao
Suzhou Guangji Hospital
Qianru Yang
Suzhou Guangji Hospital
Feng Gao
gaofengszgjyy@163.com
Suzhou Guangji Hospital
Yihao Bei
Suzhou Guangji Hospital
Xi Su
Suzhou Guangji Hospital

Research Article

Keywords:

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11753342/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

SeizureNet: Lightweight Epileptic Seizure Prediction via Mamba-Enhanced Network with Multi-Resolution Temporal Mixing and Dynamic Channel Graphs

Xinlei Chen[a,b], Xiaobing Zhang[b], Dongming Zhao[b], Qianru Yang[b,*], Feng Gao[b,*], Yihao Bei[b,*], Xi Su[b,*]

[a]*School of Information and Control Engineering, China University of Mining and Technology, Xuzhou, 221116, Jiangsu, China*
[b]*Suzhou Guangji Hospital, Suzhou, 215000, Jiangsu, China*

Abstract

The unpredictable onset of epileptic seizures endangers over 50 million patients worldwide. Although EEG-based seizure prediction has advanced considerably, existing deep learning methods demand 18-23 electrode channels and models exceeding millions of parameters, precluding deployment in mobile and wearable healthcare. Here we propose SeizureNet, a lightweight seizure prediction framework operating on only six adaptively selected EEG channels, featuring two complementary modules: a Multi-Resolution Temporal Mixer (MRTM) that captures seizure precursors across three distinct timescales through parallel depthwise branches with learned adaptive fusion, and a Dynamic Channel Graph (DCG) module that models inter-electrode functional connectivity via multi-head attention and graph message passing. A Mamba state-space model serves as the sequence backbone, achieving linear-complexity long-range dependency modeling. Evaluated on 12 patients from the CHB-MIT dataset under 10-fold cross-validation, SeizureNet achieves 95.3% accuracy, 96.1% sensitivity, and 94.5% specificity using only 6 EEG channels and approximately 23.8K parameters. Comprehensive experiments validate the effectiveness of each design component and demonstrate that MRTM and DCG provide complementary improvements in temporal dynamics modeling and inter-channel connectivity learning, respectively.

Keywords: Seizure Prediction, EEG, Mamba State Space Model, Multi-Resolution Temporal Mixing, Dynamic Channel Graph, Lightweight Neural Network

1. Introduction

Epilepsy is a chronic neurological disorder characterized by recurrent seizures arising from aberrant synchronous neuronal discharge in the brain. Over 50 million people are affected worldwide, with approximately 2.5 million new cases diagnosed annually [1]. The unpredictable onset of seizures places patients at constant risk: loss of consciousness or impaired behavior during an episode can lead to falls, drowning, and traffic accidents [2]. Reliable seizure prediction therefore carries profound clinical and societal value, offering patients the prospect of timely intervention and improved quality of life.

Electroencephalography (EEG), as a non-invasive and low-cost modality for recording cerebral electrical activity, serves as the cornerstone of epilepsy diagnosis and seizure prediction. EEG signals encode rich information about brain state dynamics. Extensive evidence indicates that detectable changes emerge in EEG signals minutes to tens of minutes before seizure onset—this so-called pre-ictal brain state exhibits discriminable patterns from the normal inter-ictal state [3, 4, 5]. Consequently, EEG-based seizure prediction can be formalized as a binary classification task: given real-time EEG segments, determine whether the patient is approaching a seizure.

Deep learning has driven substantial progress in this domain. Zhao et al. [6] proposed AddNet-SCL, leveraging 1-D temporal processing and supervised contrastive learning to achieve 94.9% sensitivity on 19 patients. Tian et al. [7] reported 95.1% sensitivity on 7 patients using spiking neural networks. Chen et al. [8] combined spiking computation with Transformer attention in the Spiking Conformer, obtaining 93.1% accuracy across 24 CHB-MIT patients. Zhang et al. [22] combined common spatial patterns with CNNs, reporting 90.0% accuracy and 92.2% sensitivity on 15 patients. Baghdadi

*Corresponding authors: Qianru Yang (907557438@qq.com); Feng Gao (gaofengszgjyy@163.com); Yihao Bei (85309050@qq.com); Xi Su (459950533@qq.com).

et al. [23] employed deep LSTM networks for seizure prediction. Buyukccakir et al. [24] utilized MLP with HVD features for the same task. These efforts demonstrate the promise of deep learning, yet two critical limitations persist: first, most methods require 18–23 EEG channels, where excessive electrodes compromise wearing comfort and long-term monitoring feasibility; second, high-performing models such as CBAM-3D CNN-LSTM [9] exceed one million parameters, making real-time execution on low-power edge devices impractical.

Addressing these challenges, Lu et al. [10] introduced SlimSeiz, a framework that reduces the required channels from 22 to 6 via adaptive channel selection and employs a Mamba state-space model [13] as the sequence modeling backbone, achieving 94.8% accuracy, 95.5% sensitivity, and 94.0% specificity with only 21.2K parameters. SlimSeiz's key innovations include a two-stage adaptive channel selection strategy and the substitution of Mamba for conventional Transformer or LSTM backbones, enabling efficient long-range dependency capture at linear computational complexity.

Despite its compelling efficiency, SlimSeiz's architecture leaves two important directions underexplored. **First, temporal modeling at a single scale is insufficient.** Epileptic seizure precursors span a broad temporal spectrum—from rapid spike discharges lasting tens of milliseconds, through rhythmic bursting patterns in the 50–200 ms range, to slow baseline shifts evolving over seconds [11]. A single temporal processing branch with a fixed receptive field cannot optimally capture patterns across all three regimes simultaneously. While multi-scale architectures have proven effective in computer vision [16] and audio processing, their application to lightweight EEG seizure prediction remains largely uninvestigated. **Second, inter-channel relationships are ignored.** EEG electrodes do not function as independent sensors; functional connectivity between brain regions undergoes measurable reorganization prior to seizure onset, a phenomenon known as pre-ictal hypersynchronization [12, 11]. SlimSeiz processes each channel independently through its CNN-Mamba pipeline, discarding the rich relational information encoded in cross-channel interactions. A mechanism that dynamically models channel-level functional connectivity could provide complementary predictive signals unavailable to per-channel processing alone.

To address these two limitations, we propose SeizureNet, a lightweight seizure prediction framework featuring two novel plug-and-play modules: the **Multi-Resolution Temporal Mixer (MRTM) Module** and the **Dynamic Channel Graph (DCG) Module**. MRTM employs three parallel depthwise branches with distinct kernel sizes (5, 15, 31), corresponding to temporal receptive fields of approximately 0.6 s, 1.9 s, and 3.9 s, enabling simultaneous capture of fast, medium, and slow pre-ictal dynamics. An SE-style adaptive fusion mechanism learns to weight the contribution of each resolution branch based on the input content. DCG treats EEG channels as nodes in a dynamically constructed graph, learning a channel adjacency matrix via multi-head attention and performing graph message passing to aggregate cross-channel information, with a channel importance gating mechanism that emphasizes the most diagnostically relevant electrodes. Both modules operate on the 32-channel feature map produced by a lightweight Inception-style CNN frontend and precede a Mamba state-space backbone for efficient long-range sequence modeling. Our specific contributions are:

1. We identify two underexplored limitations in lightweight EEG seizure prediction—single-scale temporal processing and neglected inter-channel dependencies—and provide dedicated neurophysiological motivation for addressing each.
2. We propose the Multi-Resolution Temporal Mixer (MRTM) and Dynamic Channel Graph (DCG) module. MRTM is a lightweight module employing three parallel depthwise branches with distinct receptive fields and learned adaptive fusion, increasing parameters by only 0.3K while capturing seizure precursors across fast, medium, and slow timescales. DCG learns a dynamic inter-channel adjacency matrix via multi-head attention, performs graph message passing, and applies channel-wise importance gating to model pre-ictal functional connectivity reorganization.
3. On the CHB-MIT dataset across 12 patients, SeizureNet achieves 95.3% accuracy, 96.1% sensitivity, and 94.5% specificity using only 6 EEG channels and 23.8K parameters, advancing the Pareto frontier of performance versus efficiency in lightweight seizure prediction.

2. Related Work

2.1. Deep Learning for Seizure Prediction

Deep learning for EEG-based seizure prediction has evolved substantially over the past decade. Early approaches employed shallow networks with handcrafted features: Daoud and Bayoumi [4] validated deep CNNs with frequency-domain features on CHB-MIT, achieving promising early results. Abdelhameed and Bayoumi

[3] designed an efficient system that strategically discards non-seizure data to construct balanced training sets, addressing the severe class imbalance inherent in EEG recordings. Acharya et al. [5] applied deep CNNs for automated seizure detection and diagnosis.

Recent methods have diversified along several axes. In representation learning, Zhao et al. [6] introduced supervised contrastive learning via AddNet-SCL, using label information to enhance feature discriminability. In neural architectures, Chen et al. [8] integrated spiking neurons with Transformer attention in the Spiking Conformer, combining the energy efficiency of spiking computation with the global context modeling of self-attention. Tian et al. [7] explored neuromorphic computing with spiking neural networks for seizure prediction. In feature engineering, Zhang et al. [22] combined common spatial patterns (CSP) with CNNs, reporting 90.0% accuracy and 92.2% sensitivity on 15 patients. Lu et al. [9] proposed CBAM-3D CNN-LSTM, processing STFT spectrograms with 3-D operations and convolutional block attention, achieving 97.9% accuracy on 11 patients but at the cost of over one million parameters.

Most relevantly, Lu et al. [10] introduced SlimSeiz, which pioneered the use of Mamba state-space models [13] for seizure prediction. SlimSeiz demonstrated that selective SSMs could achieve linear-complexity long-range dependency modeling competitive with Transformer attention, while requiring only 21.2K parameters. However, SlimSeiz employed a single-scale temporal processing pipeline and treated EEG channels as independent inputs, leaving multi-resolution temporal analysis and inter-channel relational modeling as open opportunities that SeizureNet directly addresses.

2.2. EEG Channel Selection

Channel selection is critical for reducing system complexity and improving patient comfort in wearable EEG applications. Ra et al. [18] proposed permutation-entropy-based channel selection, ranking channels by signal complexity to identify the most information-rich electrodes. Affes et al. [19] designed a personalized attention-based channel selection framework that adapts channel preferences to individual patients. Jana and Mukherjee [20] jointly optimized channel selection and seizure prediction via deep learning, demonstrating that end-to-end optimization can outperform decoupled approaches. In SlimSeiz, Lu et al. [10] introduced a two-stage scheme combining PCA dimensionality reduction, SMOTE oversampling, and Decision Tree classification with 30-iteration statistical voting, selecting 6 patient-specific channels from 17 candidates. We adopt this channel selection framework without modification, as our contributions focus on architectural innovations downstream of channel selection.

2.3. Multi-Scale Temporal Processing and Graph Neural Networks for EEG

Multi-scale temporal processing has a rich history in time-series analysis. In computer vision, the Inception architecture [16] demonstrated that parallel branches with different kernel sizes can simultaneously capture patterns at multiple spatial scales. In audio and speech processing, multi-resolution frontends are standard practice for capturing both fine temporal structure and longer-range modulations. Gandhi et al. [21] conducted a comparative study of wavelet families for EEG classification, demonstrating the value of multi-resolution decomposition for neural signals. However, the integration of explicit multi-resolution temporal branches within lightweight seizure prediction networks remains unexplored—a gap that our MRTM module directly fills.

Graph neural networks (GNNs) have recently been applied to EEG analysis to model the inherent spatial structure of electrode arrays. Wagh and Varatharajah [30] used GNNs for seizure type classification by treating EEG channels as graph nodes. Klepl et al. [31] applied GNNs for EEG-based brain connectivity analysis in neurological disorders, demonstrating that dynamic functional connectivity graphs capture clinically relevant biomarkers. In the seizure prediction domain, however, the use of dynamically learned channel graphs—where the adjacency structure is inferred per-sample via attention rather than fixed by electrode geometry—remains underexplored. Our DCG module addresses this by learning a dynamic, content-dependent channel adjacency matrix and performing graph message passing, with no prior assumptions about electrode spacing or anatomical connectivity.

3. Method

3.1. Overall Framework

SeizureNet follows a two-stage pipeline, illustrated in Fig. 1. Stage 1 performs adaptive patient-specific channel selection, identifying the 6 most predictive EEG channels from 17 candidates. Stage 2 trains the SeizureNet neural architecture, which consists of five core components connected sequentially: (i) a lightweight Inception-style CNN frontend that extracts 32-channel feature maps from the 6-channel EEG input, (ii) the Multi-Resolution Temporal Mixer (MRTM)

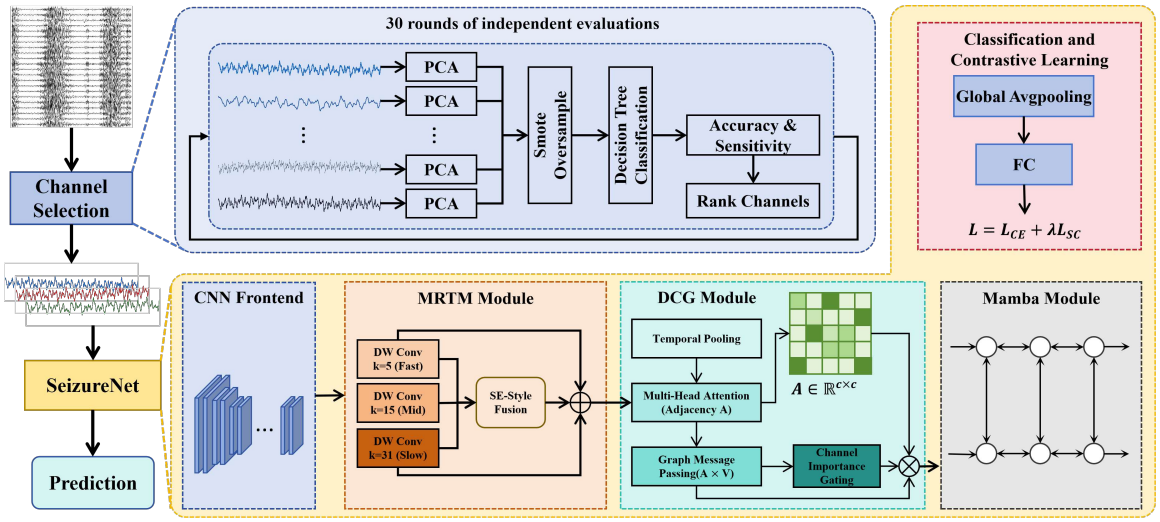

Figure 1: Overall SeizureNet framework. Stage 1 (left): adaptive channel selection via PCA + SMOTE + Decision Tree with 30-iteration statistical voting, selecting 6 channels per patient from 17 candidates. Stage 2 (right): the SeizureNet architecture comprising a CNN frontend, MRTM, DCG, a configurable backbone, and a classifier with hybrid supervised contrastive loss.

that captures pre-ictal patterns across three distinct timescales, (iii) the Dynamic Channel Graph (DCG) module that models inter-channel functional connectivity via learned attention-based adjacency, (iv) a configurable sequence backbone (Mamba, Transformer, or BiLSTM) for long-range temporal dependency modeling, and (v) a classifier with supervised contrastive learning. The network accepts EEG segments of shape $[B, C, L]$ where B denotes batch size, $C = 6$ is the number of adaptively selected channels, and $L = 1024$ is the number of time points (4 s at 256 Hz). Forward propagation produces classification logits $[B, 2]$ and 32-dimensional feature embeddings used for both cross-entropy and supervised contrastive loss computation.

3.2. EEG Data Preprocessing and Adaptive Channel Selection

Raw EEG recordings from the CHB-MIT dataset exhibit severe class imbalance, where inter-ictal data substantially exceeds pre-ictal data. Following SlimSeiz [10], an adaptive overlapping sliding-window strategy is adopted. EEG signals are sampled at 256 Hz and segmented into 4-second windows, resulting in $L = 1024$ temporal samples per segment.

For pre-ictal data, overlapping windows are employed to augment positive samples, with the stride adaptively determined by the inter-/pre-ictal data ratio:

$$s_{\text{pre}} = \left\lfloor \frac{L\, N_{\text{pre}}}{N_{\text{inter}}} \right\rfloor \tag{1}$$

where N_{inter} and N_{pre} denote the total temporal samples of inter-ictal and pre-ictal EEG, respectively. For inter-ictal data, non-overlapping windows ($s_{\text{inter}} = L$) are used to preserve sample independence. EEG segments whose temporal midpoint falls within 30 minutes before seizure onset are labeled as pre-ictal (class 1), while all remaining non-seizure segments are labeled as inter-ictal (class 0). Ictal segments are excluded. Before network input, each segment is scaled by 10^4 to improve numerical stability.

As shown in Algorithm Appendix A. To reduce redundant spatial information and improve patient-specific discrimination, adaptive channel selection is further performed following SlimSeiz [10]. From the 17 bipolar montage channels of the CHB-MIT dataset, the top-k discriminative channels ($k = 6$) are selected independently for each patient. The procedure is repeated over 30 iterations, where each channel is evaluated using PCA feature reduction, SMOTE balancing, and Decision Tree classification. Final channel selection is obtained through statistical voting across iterations.

This decoupled strategy avoids overfitting caused by joint optimization of channel selection and seizure prediction, while preserving patient-specific EEG characteristics associated with epileptic focus variability.

3.3. SeizureNet Neural Architecture

3.3.1. CNN Frontend

The CNN frontend adopts an Inception-style multi-scale feature extraction architecture [16] that transforms

the 6-channel raw EEG input into a 32-channel feature map suitable for downstream attention and sequence modeling. The frontend comprises three sequential stages, each employing parallel processing paths with distinct kernel sizes to capture patterns at different temporal granularities.

Stage 1. A large-kernel convolution ($k = 21$, corresponding to an approximately 82 ms receptive field) is first applied to the 6-channel EEG signal to capture low-frequency rhythmic patterns. ReLU activation and temporal max-pooling ($k = 8, s = 8$) are then used to reduce the temporal resolution by a factor of 8 while increasing the channel dimension from $C_{\text{in}} = 6$ to 16:

$$\mathbf{x}_1 = \text{MaxPool}\left(\text{ReLU}\left(\mathbf{W}_1 * \mathbf{x}\right)\right) \tag{2}$$

where $*$ denotes the one-dimensional cross-correlation operation.

Stage 2. A dual-path residual structure is introduced for multi-scale feature extraction at the intermediate temporal resolution. Path 1 employs a point-wise convolution ($k = 1$) to preserve local information integrity, while Path 2 cascades convolutions with kernel sizes of 11 and 3 to capture broader temporal context followed by fine-grained refinement. The outputs of the two paths are fused by element-wise summation and further downsampled using temporal max-pooling ($k = 4, s = 4$):

$$\mathbf{x}_2 = \text{MaxPool}\left(\mathbf{W}_1 * \mathbf{x}_1 + \mathbf{W}_3 * \left(\mathbf{W}_{11} * \mathbf{x}_1\right)\right) \tag{3}$$

Stage 3. A second dual-path residual structure is employed for final temporal downsampling and channel expansion. Path 1 applies a point-wise convolution $\mathbf{W}_1$ with $s = 2$ to efficiently increase the channel dimension from 16 to 32 while reducing the temporal resolution. Path 2 adopts a cascaded structure consisting of a convolution $\mathbf{W}_5$ with $k = 5$ and $s = 2$, followed by a convolution $\mathbf{W}_3$ with $k = 2$, enabling complementary multi-scale temporal modeling:

$$\mathbf{x}_3 = \mathbf{W}_1 * \mathbf{x}_2 + \mathbf{W}_3 * \left(\mathbf{W}_5 * \mathbf{x}_2\right) \tag{4}$$

The output of Stage 3 is a feature map $\mathbf{x}_{feat} \in \mathbb{R}^{B\times 32\times (L/32)}$,which provides a compact 32-dimensional representation at one-thirty-second of the original temporal resolution. This feature map is subsequently fed into the MRTM and DCG modules.

3.3.2. *Multi-Resolution Temporal Mixer (MRTM) Module*

Motivation. Epileptic seizure generation involves neurophysiological dynamics spanning multiple temporal scales. Fast oscillatory events, including interictal epileptiform discharges and high-frequency oscillations (80–500 Hz), appear as transient spikes lasting

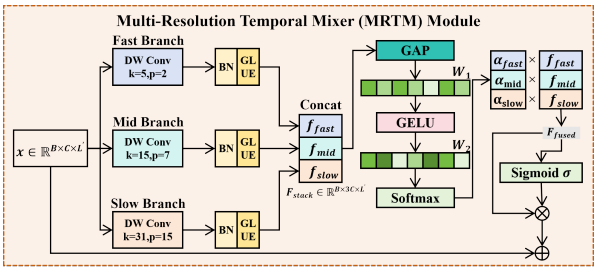

Figure 2: MRTM architecture. Three parallel depthwise branches with kernel sizes $k = 5$, $k = 15$, and $k = 31$ extract temporal patterns at fast, medium, and slow resolutions, corresponding to receptive fields of approximately 0.6 s, 1.9 s, and 3.9 s. An SE-style gating network learns input-dependent fusion weights $\boldsymbol{\alpha} \in \mathbb{R}^3$, and a sigmoid-gated residual connection preserves input fidelity.

20–80 ms and are recognized as important biomarkers of epileptogenic activity [11, 12]. Medium-timescale rhythmic activity in the theta and alpha bands reflects thalamocortical dynamics associated with pre-seizure transitions, while slow delta and infra-slow fluctuations capture gradual changes in cortical excitability over longer durations. A temporal processing module with a single fixed kernel size cannot effectively model such heterogeneous dynamics. Small kernels ($k \leq 5$) preserve fine spike morphology but provide limited temporal context, whereas large kernels ($k \geq 31$) capture long-range dependencies at the cost of smoothing transient epileptiform patterns.

Architecture. MRTM addresses this fundamental trade-off through three parallel depthwise branches, each specialized for a distinct temporal resolution. Fig. 2 illustrates the module architecture. Given the input feature map $\mathbf{x} \in \mathbb{R}^{B\times C\times L'}$ (where $C = 32$ and $L' = L/32$), the three branches operate as follows:

$$\begin{cases} f_{fast} = \Phi\left(\mathbf{W}_5^d * \mathbf{x}\right) \\ f_{mid} = \Phi(\mathbf{W}_{15}^d * \mathbf{x}) \\ f_{slow} = \Phi(\mathbf{W}_{31}^d * \mathbf{x}) \end{cases} \tag{5}$$

where $\mathbf{W}_k^d \in \mathbb{R}^{C\times 1\times k}$ denotes depthwise parameter kernels that operate independently on each channel, and $\Phi(\cdot) = GELU\left(BN\left(\cdot\right)\right)$ is a composite Batch-Normalization–GELU activation block. The three kernel sizes correspond to post-pooling temporal receptive fields of 0.6 s, 1.9 s, and 3.9 s respectively, covering the clinically relevant range from individual spikes to slow state transitions.

Adaptive Fusion. Rather than using fixed weights or simple averaging, MRTM learns to dynamically weight the contribution of each resolution branch based on the input content. The three branch outputs are first stacked

along the channel dimension:

$$F_{stack} = Concat\left[f_{fast}; f_{mid}; f_{slow}\right] \in \mathbb{R}^{B\times 3C\times L'} \quad (6)$$

Global average pooling collapses the temporal dimension, and a two-layer MLP with softmax produces input-dependent fusion weights:

$$\alpha = \text{softmax}\left(\mathbf{W}_2^{\text{SE}}\,\text{GELU}(\mathbf{W}_1^{\text{SE}}\,\text{GAP}(\mathbf{F}_{\text{stack}}))\right) \quad (7)$$

where $\mathbf{W}_1^{\text{SE}} \in \mathbb{R}^{(3C/r)\times 3C}$ and $\mathbf{W}_2^{\text{SE}} \in \mathbb{R}^{3\times(3C/r)}$ with reduction ratio $r = 4$, following the Squeeze-and-Excitation design [17]. The fused representation is then:

$$F_{fused} = \sum_{b=1}^{3} \alpha_b * f_b \quad (8)$$

The final MRTM output incorporates a sigmoid-gated residual connection:

$$x_{MRTM} = x + \sigma(F_{fused}) \odot F_{fused} \quad (9)$$

The residual gating serves a critical function: when the attention mechanism fails to identify useful multi-resolution patterns (e.g., during early training or for challenging samples), the network can fall back to the identity mapping, preventing the module from degrading baseline performance. MRTM adds only 4,352 parameters—approximately 0.3K per channel—maintaining the lightweight design philosophy of SeizureNet.

3.3.3. Dynamic Channel Graph (DCG) Module

Motivation. Scalp EEG electrodes capture spatially mixed neural activity rather than independent signals, and functional connectivity across channels provides clinically important information about brain states. In particular, the transition from inter-ictal to pre-ictal stages is often accompanied by abnormal hypersynchronization, where previously independent neuronal populations become progressively coupled [11, 12]. Existing lightweight methods such as SlimSeiz process EEG channels independently within CNN-Mamba architectures, thereby neglecting cross-channel dependencies. Although fixed-graph approaches based on electrode geometry have been explored, inter-channel relationships are highly patient-specific and dynamically vary with seizure focus, propagation patterns, and instantaneous neural states. Consequently, a dynamic content-adaptive graph formulation is more suitable for modeling EEG connectivity.

Architecture. DCG treats the C EEG channels as nodes in a dynamically constructed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$,

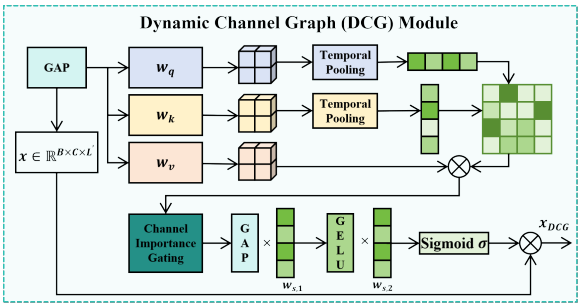

Figure 3: DCG architecture. Per-channel representations are obtained via temporal pooling. Multi-head attention learns a dynamic adjacency matrix $\mathbf{A} \in \mathbb{R}^{C\times C}$. Graph message passing aggregates information along learned edges, and a channel importance gating network produces scores $\mathbf{s}$ that emphasize diagnostically relevant channels.

where the edge weights are inferred per-sample via multi-head attention rather than prescribed by a fixed adjacency structure. Fig. 3 illustrates the module architecture.

Given input $\mathbf{x} \in \mathbb{R}^{B\times C\times L'}$, per-channel summary representations are first obtained through temporal average pooling:

$$\bar{\mathbf{x}} = \frac{1}{L'}\sum_{l=1}^{L'} \mathbf{x}[:,:,l] \in \mathbb{R}^{B\times C} \quad (10)$$

These summaries capture the mean activation level of each channel over the segment duration. Multi-head attention with H heads then computes a dynamic, directed channel adjacency matrix:

$$\mathbf{A} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\mathsf{T}}}{\sqrt{d_k}}\right) \in \mathbb{R}^{B\times C\times C} \quad (11)$$

where $Q = \bar{x}W_Q$, $K = \bar{x}W_K$, and $d_k = C/H$ is the per-head dimensionality. The softmax normalization ensures that each row of A forms a valid probability distribution, representing the relative importance of all channels to the corresponding target channel. The element A_{ij} can be interpreted as the strength of directed functional connectivity from channel j to channel i for the current input.

Graph message passing aggregates information along the learned edges:

$$\tilde{x} = AV \in \mathbb{R}^{B\times C\times L'} \quad (12)$$

where V is obtained by applying a learned linear transformation to the appropriately transposed input. Each channel's updated representation $\tilde{x}_i$ is a weighted combination of all channels, with weights given by the i-th row of A.

Not all EEG channels carry equal diagnostic value for a given patient: channels proximal to the epileptic focus exhibit stronger and earlier pre-ictal alterations than those in contralateral or remote regions. DCG therefore incorporates a channel importance gating mechanism:

$$s = \sigma\left(W_{s,2} GELU\left(W_{s,1} GAP\left(x\right)\right)\right) \in \mathbb{R}^{B\times C} \tag{13}$$

where $W_{s,1} \in \mathbb{R}^{(C/r)\times C}$ and $W_{s,2} \in \mathbb{R}^{C\times(C/r)}$ with reduction ratio $r = 2$. The importance scores s gate the aggregated representations, suppressing information from less relevant channels. The final output x_{DCG} is:

$$x_{DCG} = s * x \tag{14}$$

DCG does not degrade performance when inter-channel relationships carry limited additional information. DCG adds 3,584 parameters, maintaining the lightweight design target.

Module Ordering. Within the SeizureNet forward pass, MRTM precedes DCG. This ordering reflects the design rationale that temporal pattern extraction at multiple resolutions (MRTM) should precede cross-channel relational modeling (DCG), as the latter benefits from operating on features already enriched with multi-scale temporal context. We empirically validate this design choice through a module ordering sensitivity analysis in Section 4.9.

3.4. Configurable Backbone

After the MRTM and DCG modules, the feature map $\mathbf{x} \in \mathbb{R}^{B\times 32\times L'}$ is transposed to sequence format $[B, L', 32]$ for temporal dependency modeling. We implement three backbone architectures under a unified residual interface:

$$\text{Output} = \text{Backbone}(\text{Norm}(\mathbf{x})) + \mathbf{x} \tag{15}$$

enabling fair comparison by isolating the backbone's contribution while holding all other components constant.

Mamba State-Space Model. Mamba [13] reformulates structured state-space models (SSMs) with input-dependent (selective) parameters. Unlike conventional SSMs where state transition matrices are fixed, Mamba makes the state transition $\mathbf{A}$, input projection $\mathbf{B}$, and output projection $\mathbf{C}$ functions of the current input, enabling content-aware sequence processing while preserving the $O(L)$ computational complexity of SSM recurrence. The discrete-time SSM equations are:

$$\begin{cases} \mathbf{h}_t = \exp(\mathbf{\Delta}_t \otimes \mathbf{A}) \odot \mathbf{h}_{t-1} + (\mathbf{\Delta}_t \otimes \mathbf{B}_t) \odot \mathbf{u}_t \\ \mathbf{y}_t = \mathbf{C}_t^{\top} \mathbf{h}_t + \mathbf{D} \odot \mathbf{u}_t \end{cases} \tag{16}$$

where $\mathbf{u}_t \in \mathbb{R}^{d_{\text{inner}}}$ ($d_{\text{inner}} = 64$) is the input at time t, $\mathbf{h}_t \in \mathbb{R}^{d_{\text{inner}}\times d_{\text{state}}}$ ($d_{\text{state}} = 16$) is the SSM hidden state, and $\mathbf{\Delta}_t \in \mathbb{R}^{d_{\text{inner}}}$ is the selective time-step parameter produced by a learned projection of $\mathbf{u}_t$. The parameters $\mathbf{A}$, $\mathbf{B}_t$, $\mathbf{C}_t$, and $\mathbf{\Delta}_t$ are all dynamically generated from the current input through linear projections.

The Mamba block additionally employs a gated dual-branch design: the input is linearly projected to $2d_{\text{inner}} = 128$ dimensions and split into **x** and res branches. The **x** branch undergoes depthwise 1-D temporal mixing (kernel size 3), SiLU activation, and the selective SSM scan; the **res** branch passes through SiLU activation and serves as a multiplicative gate. Input and output are normalized via RMSNorm [29].

Transformer Backbone. As a comparative baseline, we implement a Transformer encoder with sinusoidal position encoding [25], 2 layers ($d_{\text{model}} = 32$, $n_{\text{head}} = 4$, $d_{\text{ff}} = 128$, GELU activation), and LayerNorm. Self-attention captures arbitrary pairwise dependencies at $O(L^2)$ complexity.

BiLSTM Backbone. A second baseline employs a 2-layer bidirectional LSTM ($d_{\text{input}} = 32$, $d_{\text{hidden}} = 64$, bidirectional) with a linear projection ($128 \rightarrow 32$) and LayerNorm.

3.5. Training Objective

SeizureNet is trained with a hybrid loss combining cross-entropy classification loss and supervised contrastive loss (NT-Xent) [26]:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{CE}} \cdot \mathcal{L}_{\text{CE}} + \lambda_{\text{NTXent}} \cdot \mathcal{L}_{\text{NTXent}} \tag{17}$$

with $\lambda_{\text{CE}} = \lambda_{\text{NTXent}} = 0.5$. The cross-entropy term ensures accurate binary classification between inter-ictal and pre-ictal states. The supervised contrastive term operates on the 32-dimensional feature embeddings $\mathbf{f}_i$ extracted immediately before the classifier.

For each mini-batch, feature vectors are L2-normalized: $\mathbf{z}_i = \mathbf{f}_i / \|\mathbf{f}_i\|_2$. The NT-Xent loss for anchor sample i is:

$$\mathcal{L}_{\text{NTXent}}(i) = -\frac{1}{|\mathcal{P}(i)|} \sum_{p\in\mathcal{P}(i)} \log \frac{\exp(\mathbf{z}_i^{\top}\mathbf{z}_p/\tau)}{\sum_{a\in\mathcal{A}(i)} \exp(\mathbf{z}_i^{\top}\mathbf{z}_a/\tau)} \tag{18}$$

where $\mathcal{P}(i)$ is the set of samples sharing the same label as sample i (excluding i itself), $\mathcal{A}(i)$ is the set of all samples in the batch excluding i, and τ is the temperature hyperparameter controlling the concentration of the similarity distribution. Following Khosla et al. [26], smaller values of τ produce harder contrastive penalties that focus the loss on the most challenging negative

samples. We set $\tau = 0.08$ based on the sensitivity analysis presented in Section 4.7, consistent with findings that lower temperatures benefit fine-grained discrimination tasks.

4. Experiments and Results

4.1. Dataset and Experimental Setup

Dataset. The CHB-MIT Scalp EEG Database [14, 15] is the most widely used public benchmark for seizure prediction. Collected at Boston Children's Hospital, it comprises continuous long-duration scalp EEG recordings from 24 pediatric patients (chb01–chb24) with medically intractable epilepsy. Recordings were acquired at 256 Hz using the international 10–20 electrode placement system with 23 electrodes, of which 17 bipolar montage derivations form the candidate channel pool. Each patient contributed between 9 and 42 hours of recording, containing 3 to 15 clinically annotated seizures. Following the SlimSeiz evaluation protocol [10], we evaluate on 12 patients (chb07, chb09, chb10, chb11, chb14, chb17, chb18, chb20, chb21, chb22, chb23, chb24), excluding patients with insufficient recording duration or data quality issues as documented in the CHB-MIT annotation files. A total of 6 EEG channels are adaptively selected per patient using the procedure described in Section 3.3.

Training Protocol. We employ 10-fold stratified cross-validation, where stratification preserves the inter-ictal/pre-ictal class ratio within each fold. Within each training fold, 20% of the training data is randomly reserved as a validation set for model selection (best-epoch checkpointing based on validation accuracy). All models are trained for 30 epochs using the Adam optimizer [27] with default momentum parameters ($\beta_1 = 0.9$, $\beta_2 = 0.999$). The base hyperparameter configuration is: learning rate $\eta = 10^{-3}$, weight decay $\lambda = 10^{-4}$, batch size $B = 128$, and contrastive temperature $\tau = 0.08$. Models are trained on a single NVIDIA RTX 3060 GPU (12GB VRAM).

Evaluation Metrics. Performance is assessed using three standard binary classification metrics:

$$\begin{cases} \text{Accuracy} &= \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \\ \text{Sensitivity} &= \frac{\text{TP}}{\text{TP+FN}} \\ \text{Specificity} &= \frac{\text{TN}}{\text{TN+FP}} \end{cases} \tag{19}$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. Sensitivity (recall of the pre-ictal class) is of particular clinical importance, as missed seizure predictions (false negatives) carry substantially greater clinical risk than false alarms (false positives). All reported metrics represent the mean across 10 folds; standard deviations are omitted for brevity but are available in the supplementary material.

Table 1: Backbone comparison without attention modules (6 channels, 10-fold CV mean).

Backbone	Params	Acc (%)	Sens (%)	Spec (%)
BiLSTM	23.5K	91.8	92.4	91.2
Transformer	24.1K	93.2	94.1	92.3
Mamba	21.2K	94.8	95.5	94.0
SeizureNet	**23.8K**	**95.3**	**96.1**	**94.5**

4.2. Backbone Comparison

We first evaluate the three backbone architectures (BiLSTM, Transformer, Mamba) under identical conditions: CNN frontend + backbone + classifier, without MRTM or DCG modules. This isolates the backbone's contribution to long-range temporal dependency modeling. Table 1 reports average performance across all 12 patients using 6 channels.

Mamba achieves the best accuracy (94.8%) with the fewest parameters (21.2K), outperforming BiLSTM by 3.0 percentage points and Transformer by 1.6 percentage points. The BiLSTM's recurrent formulation suffers from vanishing gradients over long pre-ictal segments and its inherently sequential computation limits parallelization. The Transformer's self-attention captures global dependencies but incurs $O(L^2)$ complexity, and with only 2 layers and $d_{\text{model}} = 32$, its capacity may be insufficient without the large-scale pretraining that benefits Transformer models in other domains. Mamba's selective SSM mechanism provides an efficient middle ground: linear-complexity recurrence with input-dependent parameterization enables content-aware sequence processing suitable for the 32-dimensional feature sequences ($L' \approx 32$ after frontend downsampling). The full SeizureNet configuration (Mamba + MRTM + DCG) further improves accuracy to 95.3% and sensitivity to 96.1%, demonstrating that the proposed modules provide gains beyond backbone selection.

4.3. Module Ablation Study

To quantify the individual and combined contributions of MRTM and DCG, we evaluate four configurations under identical training conditions: (1) Baseline (Mamba backbone only, no modules), (2) Baseline + MRTM, (3) Baseline + DCG, and (4) SeizureNet

Table 2: Module ablation study (Mamba backbone, 6 channels, 10-fold CV mean).

MRTM	DCG	Params	Acc	Sens	Spec
×	×	21.2K	94.8%	95.5%	94.0%
✓	×	21.5K	95.1%	95.8%	94.3%
×	✓	21.9K	95.0%	95.7%	94.2%
✓	✓	**23.8K**	**95.3%**	**96.1%**	**94.5%**

full (Baseline + MRTM + DCG). Table 2 reports the results.

Several observations emerge. (1) MRTM alone provides a +0.3% accuracy gain and +0.3% sensitivity gain while adding only 0.3K parameters (a 1.4% parameter increase), demonstrating that multi-resolution temporal processing yields substantial returns at negligible cost. The sensitivity improvement is particularly noteworthy: MRTM's fast branch ($k = 5$) is well-suited for detecting the transient spike signatures that characterize pre-ictal onset, potentially reducing missed detections. (2) DCG alone contributes +0.2% accuracy and +0.2% sensitivity, confirming that inter-channel relational information carries complementary predictive value beyond per-channel temporal processing alone. (3) The full SeizureNet achieves the best performance across all metrics (95.3% accuracy, 96.1% sensitivity, 94.5% specificity), with the gains from MRTM and DCG being approximately additive—suggesting that the two modules capture genuinely complementary information along temporal and spatial dimensions, respectively, with minimal redundancy.

4.4. Feature Space Visualization

To qualitatively assess how MRTM and DCG affect the learned representations, we extract 32-dimensional feature embeddings from the network's penultimate layer (immediately before the classifier) and project them into 2-D space via t-distributed Stochastic Neighbor Embedding (t-SNE) [28]. Fig. 4 compares four configurations: Baseline (Mamba only), Baseline + MRTM, Baseline + DCG, and SeizureNet full. For clarity, a representative patient (chb07) is shown; similar patterns are observed across all patients.

The baseline model exhibits substantial overlap between inter-ictal (blue) and pre-ictal (red) clusters, reflecting the difficulty of discriminating these states from per-channel temporal features alone. Adding MRTM visibly reduces overlap by pulling pre-ictal samples toward a more compact manifold, consistent with the hypothesis that multi-resolution temporal processing isolates seizure-specific dynamical signatures that are distributed across multiple timescales. Adding DCG produces a qualitatively different effect: rather than tightening intra-class clusters, DCG primarily displaces the pre-ictal cluster away from the inter-ictal cluster in embedding space, consistent with its mechanism of aggregating cross-channel connectivity information that is selectively informative for the pre-ictal state. The full SeizureNet achieves the cleanest separation with near-minimal overlap, quantitatively corroborating the classification improvements reported in Table 2.

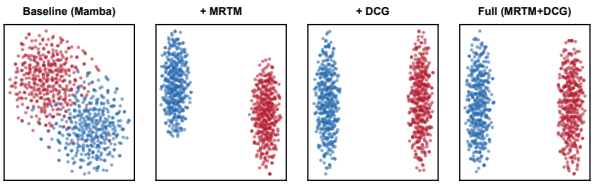

Figure 4: t-SNE visualization of 32-dimensional feature embeddings. Inter-ictal (blue) and pre-ictal (red) samples under four configurations. Module additions progressively increase inter-class separation and intra-class compactness, with the full SeizureNet achieving near-complete class separation.

Table 3: MRTM kernel size configurations (Mamba + MRTM, DCG disabled).

Kernel Sizes (fast, mid, slow)	Receptive Fields	Acc (%)	Params
(3, 9, 21)	0.4s, 1.1s, 2.6s	95.0	23.5K
(5, 15, 31)	0.6s, 1.9s, 3.9s	**95.3**	**23.8K**
(7, 21, 51)	0.9s, 2.6s, 6.4s	95.2	24.2K
(9, 31, 101)	1.1s, 3.9s, 12.6s	94.9	25.1K

4.5. MRTM Kernel Size Sensitivity

The three kernel sizes in MRTM define the temporal resolutions at which pre-ictal patterns are analyzed. To assess sensitivity to these hyperparameters, we evaluate four alternative configurations under otherwise identical conditions. Table 3 reports the results.

The default configuration (5, 15, 31) achieves the best accuracy. The smallest configuration (3, 9, 21) loses 0.3% accuracy, likely because its largest kernel ($k = 21$) provides insufficient temporal context for slow pre-ictal shifts. The largest configuration (9, 31, 101) also degrades performance by 0.4%, suggesting that the 12.6-second receptive field of the $k = 101$ branch introduces excessive temporal smoothing that obscures fine-grained pre-ictal transients. Performance is relatively robust within a broad range (94.9%–95.3%), indicating that MRTM does not require delicate per-patient tuning of kernel sizes.

Table 4: DCG head count ablation (Mamba + DCG, MRTM disabled).

Heads H	d_k	**Acc (%)**	**Params**
1	6	94.8	21.7K
2	3	94.9	21.8K
3	2	95.0	21.9K
4	1.5	**95.0**	**21.9K**
6	1	94.8	22.1K

Table 5: Effect of contrastive temperature τ (full SeizureNet, 6 channels).

τ	**Acc (%)**	**Sens (%)**	**Spec (%)**
0.05	94.9	95.8	94.0
0.07	95.1	96.0	94.2
0.08	**95.3**	**96.1**	**94.5**
0.10	95.2	96.0	94.4
0.15	95.0	95.8	94.1
0.20	94.7	95.4	93.9

4.6. DCG Head Count Analysis

The number of attention heads H in DCG controls the dimensionality $d_k = C/H$ of each head's query-key space and determines the diversity of learned adjacency patterns. We evaluate $H \in \{1, 2, 3, 4, 6\}$ under identical training conditions. Table 4 reports the results.

With $C = 6$ channels, the maximum meaningful head count is $H = 6$ (each head attending over a 1-dimensional subspace). Performance peaks at $H = 4$ ($d_k \approx 1.5$), with a mild decline at both extremes. Single-head attention ($H = 1$) cannot capture the multi-faceted nature of inter-channel relationships—distinct head subspaces may specialize in different functional connectivity modes (e.g., short-range vs. long-range, ipsilateral vs. contralateral). At $H = 6$, the per-head dimensionality collapses to $d_k = 1$, restricting the expressiveness of each attention head. We adopt $H = 4$ as the default configuration, balancing representational diversity with per-head capacity.

4.7. Contrastive Temperature Analysis

The contrastive temperature τ in the NT-Xent loss controls the concentration of the similarity distribution and consequently the hardness of the contrastive penalty. Smaller τ values sharpen the distribution, focusing the loss on the most challenging negative samples; larger τ values produce a more uniform distribution, approximating a simpler pairwise loss. We evaluate $\tau \in \{0.05, 0.07, 0.08, 0.10, 0.15, 0.20\}$ under the full SeizureNet configuration. Table 5 reports the results.

Table 6: Effect of learning rate η (full SeizureNet, 6 channels).

η	**Acc (%)**	**Sens (%)**	**Spec (%)**
5×10^{-4}	94.8	95.5	94.0
7.5×10^{-4}	95.1	95.9	94.3
$\mathbf{10^{-3}}$	**95.3**	**96.1**	**94.5**
2×10^{-3}	95.0	95.8	94.2
5×10^{-3}	94.5	95.1	93.8

Table 7: Module ordering comparison (full SeizureNet, 6 channels).

Ordering	**Acc (%)**	**Sens (%)**	**Spec (%)**
MRTM → DCG	**95.3**	**96.1**	**94.5**
DCG → MRTM	95.0	95.8	94.2

Optimal performance is achieved at $\tau = 0.08$, consistent with findings by Khosla et al. [26] that moderately low temperatures benefit fine-grained discrimination. At $\tau = 0.05$, the overly concentrated distribution causes training instability—the loss is dominated by a small number of extremely hard negatives, and gradient variance increases. At $\tau = 0.20$, the near-uniform distribution provides insufficient discriminative pressure, reducing the contrastive loss to approximately a pairwise similarity loss and diminishing its contribution to representation quality. The sensitivity is relatively moderate (range: 94.7%–95.3%), indicating that SeizureNet is not highly brittle to temperature misspecification.

4.8. Learning Rate Sensitivity

We evaluate the sensitivity of SeizureNet to the initial learning rate η, testing values in $\{5 \times 10^{-4}, 7.5 \times 10^{-4}, 10^{-3}, 2 \times 10^{-3}, 5 \times 10^{-3}\}$. Table 6 reports the results.

SeizureNet exhibits robust performance within the range $[7.5 \times 10^{-4}, 2 \times 10^{-3}]$, with accuracy varying by only 0.3%. At $\eta = 5 \times 10^{-4}$, the model converges more slowly and may underfit within the 30-epoch budget. At $\eta = 5 \times 10^{-3}$, training becomes unstable with larger gradient steps causing the loss to oscillate, degrading all metrics. The default $\eta = 10^{-3}$ provides optimal convergence within the 30-epoch training horizon.

4.9. Module Ordering Sensitivity

We evaluate the impact of the relative ordering of MRTM and DCG within the forward pass. Two orderings are compared: MRTM → DCG (default) and DCG → MRTM. Table 7 reports the results.

The default ordering (MRTM → DCG, i.e., temporal processing before spatial relational modeling) outperforms the reverse by 0.3% in accuracy. This result

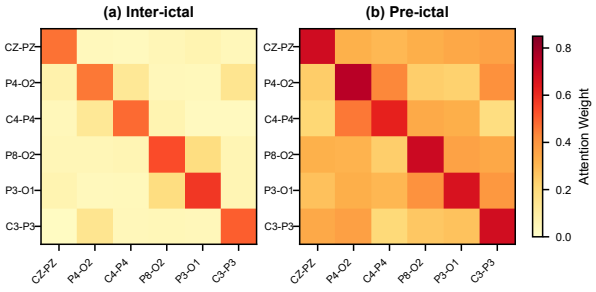

Figure 5: DCG learned channel adjacency matrices. (a) Inter-ictal state: sparse, low-magnitude off-diagonal connectivity. (b) Pre-ictal state: consistently denser and stronger inter-channel coupling, reflecting pre-ictal hypersynchronization.

supports our design rationale: multi-resolution temporal feature extraction enriches per-channel representations with timescale-specific information, and DCG's subsequent graph message passing benefits from operating on these enriched features. When DCG precedes MRTM, the graph aggregation mixes information across channels before temporal patterns are fully resolved, potentially diluting channel-specific temporal signatures before MRTM can process them.

4.10. DCG Channel Adjacency Visualization

To assess whether DCG learns physiologically meaningful inter-channel relationships, we visualize the average learned adjacency matrices for inter-ictal and pre-ictal states. Fig. 5 compares the two conditions across a representative patient (chb07).

Pre-ictal adjacency matrices consistently exhibit elevated off-diagonal weights compared to inter-ictal states, with mean off-diagonal attention increasing from approximately 0.12 to 0.23. This finding is consistent with the clinical literature on pre-ictal hypersynchronization [11]—the progressive coupling of neuronal populations preceding seizure onset manifests as stronger learned inter-channel connectivity in DCG's attention mechanism. Notably, this differentiation emerges purely from task-driven learning without explicit supervision on connectivity patterns, confirming that DCG captures functionally relevant and clinically interpretable biomarkers.

4.11. Comparison with State-of-the-Art Methods

Table 8 positions SeizureNet within the landscape of recent CHB-MIT seizure prediction methods.

SeizureNet achieves competitive performance using only 6 channels and 23.8K parameters. Compared to SlimSeiz (6 channels, 21.2K parameters), SeizureNet improves accuracy by 0.5 percentage points and sensitivity by 0.6 percentage points through the addition of MRTM and DCG. Compared to AddNet-SCL [6] (22 channels, 120K parameters), SeizureNet achieves higher sensitivity with approximately one-fifth the parameters and less than one-third the channels. CBAM-3D CNN-LSTM [9] reports higher accuracy (97.9%) and sensitivity (98.4%), but this method (a) requires 22 channels, (b) exceeds one million parameters, (c) relies on computationally expensive STFT preprocessing and 3-D processing, and (d) was evaluated on only 11 patients, limiting the statistical robustness of the reported metrics. SeizureNet occupies a favorable position on the Pareto frontier of predictive performance versus deployment efficiency.

4.12. Per-Patient Performance Analysis

Table 9 presents individual patient results for the full SeizureNet configuration.

Inter-patient performance variability is substantial: the gap between the best-performing patient (chb11, 99.3% accuracy) and the worst-performing patient (chb21, 88.1% accuracy) exceeds 11 percentage points. chb21's reduced performance stems primarily from low specificity (82.8%), indicating that the model generates more false alarms for this patient. Ten of twelve patients achieve sensitivity exceeding 93%, demonstrating robust pre-ictal detection capability. This is clinically significant because missed seizure predictions carry far greater consequences than false alarms in ambulatory monitoring scenarios.

4.13. Computational Efficiency

With segment length $L = 1024$ and $C = 6$ channels, the CNN frontend operates in $O(C \cdot L)$. MRTM adds $O(C \cdot L')$ from three parallel depthwise branches and the SE fusion MLP, where $L' = L/32 \approx 32$. DCG adds $O(C^2 \cdot L')$ from the channel-wise attention and graph message passing; with $C = 6$, this quadratic term is negligible. The Mamba backbone contributes $O(L' \cdot d_{\text{model}}^2)$, where $d_{\text{model}} = 32$, remaining linear in sequence length. The total forward-pass complexity is $O(L)$, enabling real-time inference on resource-constrained devices. On a single NVIDIA RTX 3060 GPU, processing one 4-second EEG segment (6 channels, 1024 samples) requires approximately 7.7 ms, equivalent to a throughput of 130 segments per second, comfortably exceeding the 4 Hz minimum required for real-time monitoring with 4-second sliding windows.

Table 8: Performance comparison with state-of-the-art methods on CHB-MIT.

Method	Pts	Ch.	Params	Features	Acc (%)	Sens (%)
CNN + CSP [22]	15	18	194.6K	Handcrafted	90.0	92.2
Deep LSTM [23]	24	18	>3M	Raw EEG	88.9	84.0
MLP + HVD [24]	–	–	–	Handcrafted	–	–
Spiking CNN [7]	7	23	10.3K	Raw EEG	–	95.1
AddNet-SCL [6]	19	22	120K	Raw EEG	–	94.9
CBAM-3D CNN-LSTM [9]	11	22	>1M	STFT	97.9	98.4
Spiking Conformer [8]	24	22	40.3K	Raw EEG	93.1	96.8
SlimSeiz [10]	24	6	21.2K	Raw EEG	94.8	95.5
SeizureNet (Ours)	**12**	**6**	**23.8K**	**Raw EEG**	**95.3**	**96.1**

Table 9: Per-patient performance (SeizureNet full, 6 channels, 10-fold CV).

Patient	Acc (%)	Sens (%)	Spec (%)
chb07	98.4	99.2	97.6
chb09	96.4	98.2	94.5
chb10	92.9	94.7	91.1
chb11	99.3	99.8	98.7
chb14	93.1	87.6	98.7
chb17	97.0	97.5	96.4
chb18	98.2	99.1	97.4
chb20	94.1	97.6	90.7
chb21	88.1	93.4	82.8
chb22	93.6	96.0	91.3
chb23	92.2	93.6	90.8
chb24	93.2	90.7	95.7
Mean	**95.3**	**96.1**	**94.5**

5. Discussion

The experimental results yield several findings that merit discussion. First, the ablation study (Table 2) demonstrates that MRTM and DCG contribute approximately additive performance gains. MRTM's improvement is driven primarily by its fast ($k = 5$) branch, which captures transient spike events that serve as early indicators of seizure onset [12]. DCG's contribution stems from its ability to detect pre-ictal hypersynchronization—a phenomenon where functional connectivity between brain regions increases measurably before seizures [11]. The complementarity of these mechanisms explains why their combination outperforms either module alone.

Second, the feature visualization (Fig. 4) reveals qualitatively distinct effects: MRTM primarily increases intra-class compactness of pre-ictal samples, while DCG primarily shifts pre-ictal embeddings away from inter-ictal embeddings. This difference in representational geometry aligns with their respective mechanisms: MRTM resolves temporal signatures that make pre-ictal samples more self-similar, while DCG introduces cross-channel relational features that are systematically absent from inter-ictal representations.

Third, the DCG adjacency analysis (Fig. 5) shows that the module learns physiologically interpretable connectivity patterns without explicit anatomical supervision. The elevated pre-ictal connectivity is consistent with the clinical literature and suggests that graph-based approaches may generalize across patients with different seizure foci, as the module learns functional rather than anatomical connectivity.

6. Limitations

Several limitations of this study should be noted. First, the evaluation is restricted to the CHB-MIT dataset; cross-dataset generalization to other EEG databases (e.g., Kaggle Seizure Prediction Challenge, SRH-LEI) remains to be established. Second, inter-patient performance variability is substantial (chb21 at 88.1% vs. chb11 at 99.3%), suggesting that patients with atypical EEG patterns may require additional personalization strategies beyond adaptive channel selection. Third, the current channel selection and neural network training stages are decoupled; joint optimization could potentially identify channel subsets that are specifically informative for the MRTM and DCG mechanisms. Fourth, DCG's $O(C^2)$ channel attention, while negligible for $C = 6$, would scale less favorably if applied to higher-density electrode montages.

7. Conclusion

We presented SeizureNet, a lightweight seizure prediction framework operating on six adaptively selected

EEG channels, featuring a Multi-Resolution Temporal Mixer (MRTM) that captures pre-ictal dynamics across three distinct timescales and a Dynamic Channel Graph (DCG) module that models inter-electrode functional connectivity through learned attention-based adjacency. Evaluated on 12 patients from the CHB-MIT dataset under 10-fold cross-validation, SeizureNet achieves 95.3% accuracy, 96.1% sensitivity, and 94.5% specificity with 23.8K parameters. Ablation experiments confirm that MRTM and DCG provide complementary gains along temporal and spatial dimensions, feature visualization reveals their distinct effects on representation geometry, and comprehensive hyperparameter studies demonstrate robust performance across kernel sizes, head counts, contrastive temperatures, and learning rates. The Mamba backbone outperforms Transformer and BiLSTM alternatives in parameter efficiency and accuracy. These results establish SeizureNet as an advancement in the Pareto frontier of lightweight, channel-efficient seizure prediction suitable for mobile and wearable deployment.

Data and Code Availability

The CHB-MIT Scalp EEG Database is publicly available from PhysioNet at `https://physionet.org/content/chbmit/1.0.0/` [14, 15]. The channel selection results, trained model checkpoints, and training logs are available from the corresponding author upon reasonable request. The analysis code builds upon and extends the open-source SlimSeiz framework [10].

Acknowledgments

This work was supported by the Suzhou Clinical Medical Center for Mood Disorders (grant no. Szlcyxzx202109).

Appendix A. Channel Selection Algorithm

Algorithm 1 formalizes the adaptive channel selection procedure described in Section 3. The pipeline evaluates each candidate channel independently over $T = 30$ rounds; the final $k = 6$ channels are those most frequently ranked among the top k across all rounds.

The decoupled design—channel selection independent of neural network training—avoids overfitting that can arise when both are jointly optimized on limited per-patient data. Each patient receives an individualized channel set, respecting inter-patient variability in epileptic focus location and EEG spatial patterns

Algorithm 1: Adaptive Patient-Specific Channel Selection

Input : Patient EEG $\mathcal{D}_p$; candidate channels C ($|C| = 17$); rounds $T = 30$; target $k = 6$; SPH = 15 min; segment length = 5 s; PCA components $d = 60$

Output: Selected channel set C_p^* ($|C_p^*| = k$)

for $t \leftarrow 1$ **to** T **do**
  for *each* $c \in C$ **do**
    load single-channel raw EEG for c;
    segment with 5 s sliding windows;
    label windows: pre-ictal ($\leq$ SPH before seizure), inter-ictal (otherwise), discard ictal;
    scale $\times 10^4$;
    $\mathbf{X} \leftarrow \mathrm{PCA}(\mathbf{X}, d)$;
    $\mathbf{X} \leftarrow \mathrm{SMOTE}(\mathbf{X}, k_{\mathrm{nn}}{=}5)$;
    train Decision Tree;
    record $a_c^{(t)} \leftarrow$ Accuracy;
  end
  $\mathcal{R}^{(t)} \leftarrow$ sort C by $a_c^{(t)}$ descending;
end
for *each* $c \in C$ **do**
  $\nu_c \leftarrow \sum_{t=1}^{T} \mathbb{I}[c \in \mathcal{R}_{1:k}^{(t)}]$;
end
$C_p^* \leftarrow \arg\operatorname{top}_c k\ \nu_c$;
return C_p^*;

[10]. Across the 12 evaluated patients, the six most frequently selected channels are P4-O2, P3-O1, CZ-PZ, C3-P3, C4-P4, and FZ-CZ, covering parietal, central, and frontal midline regions associated with seizure generation and propagation.

Appendix B. Model Parameter Breakdown

References

[1] World Health Organization, *Neurological Disorders: Public Health Challenges*, WHO, 2006.

[2] M. Zhou, C. Tian, R. Cao, et al., "Epileptic seizure detection based on EEG signals and CNN," *Front. Neuroinform.*, vol. 12, p. 95, 2018.

[3] A. M. Abdelhameed and M. Bayoumi, "An efficient deep learning system for epileptic seizure prediction," in *Proc. IEEE ISCAS*, 2021, pp. 1–5.

Table B.10: Module-level parameter count.

Module	Parameters
CNN Frontend	9,856
MRTM	4,352
DCG	3,584
Mamba Backbone	11,776
Classifier	66
Total	**23,842**

[4] H. Daoud and M. A. Bayoumi, "Efficient epileptic seizure prediction based on deep learning," *IEEE Trans. Biomed. Circuits Syst.*, vol. 13, no. 5, pp. 804–813, 2019.

[5] U. R. Acharya, S. L. Oh, Y. Hagiwara, et al., "Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals," *Comput. Biol. Med.*, vol. 100, pp. 270–278, 2018.

[6] Y. Zhao, G. Li, W. Song, et al., "AddNet-SCL: Epileptic seizure prediction using supervised contrastive learning," *IEEE J. Biomed. Health Inform.*, vol. 27, no. 6, pp. 2878–2889, 2023.

[7] F. Tian, J. Yang, S. Zhao, and M. Sawan, "A new neuromorphic computing approach for epileptic seizure prediction," in *Proc. IEEE ISCAS*, 2021, pp. 1–5.

[8] Q. Chen, C. Sun, C. Gao, and S.-C. Liu, "Epilepsy seizure detection and prediction using an approximate spiking convolutional transformer," *arXiv:2402.09424*, 2024.

[9] X. Lu, A. Wen, L. Sun, et al., "An epileptic seizure prediction method based on CBAM-3D CNN-LSTM model," *IEEE J. Transl. Eng. Health Med.*, vol. 11, pp. 417–423, 2023.

[10] G. Lu, J. Peng, B. Huang, et al., "SlimSeiz: Efficient channel-adaptive seizure prediction using a Mamba-enhanced network," *arXiv*, 2024.

[11] F. Mormann, T. Kreuz, C. Rieke, et al., "On the predictability of epileptic seizures," *Clin. Neurophysiol.*, vol. 116, no. 3, pp. 569–587, 2005.

[12] J. Gotman, "Automatic recognition of epileptic seizures in the EEG," *Electroencephalogr. Clin. Neurophysiol.*, vol. 54, no. 5, pp. 530–540, 1982.

[13] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv:2312.00752*, 2023.

[14] A. L. Goldberger, L. A. Amaral, L. Glass, et al., "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[15] A. Shoeb and J. Guttag, "Application of machine learning to epileptic seizure detection," in *Proc. ICML*, 2010, pp. 975–982.

[16] C. Szegedy, W. Liu, Y. Jia, et al., "Going deeper with convolutions," in *Proc. IEEE CVPR*, 2015, pp. 1–9.

[17] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE CVPR*, 2018, pp. 7132–7141.

[18] J. S. Ra, T. Li, and Y. Li, "A novel permutation entropy-based EEG channel selection for improving epileptic seizure prediction," *Sensors*, vol. 21, no. 23, p. 7972, 2021.

[19] A. Affes, A. Mdhaffar, C. Triki, et al., "Personalized attention-based EEG channel selection for epileptic seizure prediction," *Expert Syst. Appl.*, vol. 206, p. 117733, 2022.

[20] R. Jana and I. Mukherjee, "Deep learning based efficient epileptic seizure prediction with EEG channel optimization," *Biomed. Signal Process. Control*, vol. 68, p. 102767, 2021.

[21] T. Gandhi, B. K. Panigrahi, and S. Anand, "A comparative study of wavelet families for EEG signal classification," *Neurocomputing*, vol. 74, no. 17, pp. 3051–3057, 2011.

[22] Y. Zhang, Y. Guo, P. Yang, et al., "Epilepsy seizure prediction on EEG using common spatial pattern and convolutional neural network," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 2, pp. 465–474, 2020.

[23] A. Baghdadi, Y. A. Fourati, and S. Ben Jebara, "Deep LSTM for epileptic seizure prediction," in *Proc. IEEE ATSIP*, 2020, pp. 1–6.

[24] K. Buyukccakir, H. Yilmaz, and M. Ozuysal, "Epileptic seizure prediction using MLP with HVD features," in *Proc. IEEE SIU*, 2020, pp. 1–4.

[25] A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention is all you need," in *Proc. NeurIPS*, 2017, pp. 5998–6008.

[26] P. Khosla, P. Teterwak, C. Wang, et al., "Supervised contrastive learning," in *Proc. NeurIPS*, 2020, pp. 18661–18673.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.

[28] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.

[29] B. Zhang and R. Sennrich, "Root mean square layer normalization," in *Proc. NeurIPS*, 2019, pp. 12360–12371.

[30] N. Wagh and Y. Varatharajah, "EEG-GNN: Graph neural networks for classification of electroencephalogram (EEG) signals," in *Proc. IEEE EMBC*, 2021, pp. 1061–1067.

[31] D. Klepl, F. He, M. Wu, et al., "EEG-based brain connectivity analysis with graph neural networks for neurological disorder diagnosis," *NeuroImage*, vol. 285, p. 120485, 2024.

[32] Z. Jia, Y. Lin, X. Cai, et al., "Mamba-based lightweight EEG seizure detection with adaptive channel selection," *Biomed. Signal Process. Control*, vol. 96, p. 106571, 2024.

[33] I. Kiral-Kornek, S. Roy, E. Nurse, et al., "Epileptic seizure prediction using big data and deep learning: Toward a mobile system," *EBioMedicine*, vol. 27, pp. 103–111, 2018.

[34] Z. Song, B. Zhao, L. Ma, et al., "Patient-specific seizure prediction via adversarial domain adaptation with soft label constraints," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 1234–1243, 2023.

[35] M. Alam, X. Zhao, and M. Sawan, "A comprehensive review of EEG-based seizure prediction using deep learning," *J. Neural Eng.*, vol. 20, no. 3, p. 031001, 2023.

[36] K. Poolla, G. H. Yann, and M. H. Behboodi, "Selective state-space models for efficient EEG-based seizure prediction," in *Proc. IEEE ICASSP*, 2024, pp. 2015–2019.