G-CMTF Net: Spectro-Temporal Disentanglement and Reliability-Aware Gated Cross-Modal Temporal Fusion for Robust PSG Sleep Staging

Jiongyao Ye * and Pengfei Li

Posted Date: 13 January 2026

doi: 10.20944/preprints202601.0930.v1

Keywords: automatic sleep staging; multimodal polysomnography; gated cross-modal fusion; spectrotemporal learning; noise-robust modeling

 doi:10.20944/preprints202601.0930.v1

Article

G-CMTF Net: Spectro-Temporal Disentanglement and Reliability-Aware Gated Cross-Modal Temporal Fusion for Robust PSG Sleep Staging

Jiongyao Ye * and Pengfei Li

School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China
* Correspondence: yejy@ecust.edu.cn

Abstract

Automatic sleep staging from polysomnography is challenged by marked spectro-temporal heterogeneity and non-stationary cross-channel artifacts, which often undermine naïve multimodal fusion. To address this, a Gated Cross-Modal and Temporal Fusion Network (G-CMTF Net) is proposed as an end-to-end model operating on 30-s EEG epochs and auxiliary EOG and EMG signals, in which cross-modal contributions are regulated through reliability-aware gating. A spectro-temporal disentanglement frontend learns multi-scale temporal features while incorporating FFT-derived band-power embeddings to preserve physiologically meaningful oscillatory cues. At the epoch level, gated fusion suppresses artifact-prone auxiliary inputs, thereby limiting noise transfer into a shared latent space. Long-range sleep dynamics are modeled via a convolution-augmented self-attention encoder that captures both local morphology and transition structure. On Sleep-EDF-20 and Sleep-EDF-78, G-CMTF Net achieves Macro-F1/ACC of 81.3%/85.5% and 78.2%/83.4%, respectively, while maintaining high sensitivity and geometric-mean performance on transitional epochs, consistent with the function of reliability-aware gated fusion under non-stationary auxiliary artifacts.

Keywords: automatic sleep staging; multimodal polysomnography; gated cross-modal fusion; spectro-temporal learning; noise-robust modeling

1. Introduction

Sleep is essential to normal neurophysiological function and constitutes a dynamically regulated biological state through which systemic homeostasis and neural–metabolic recovery are sustained [1]. Rather than a uniform condition, sleep is organized as a cyclic progression of distinct stages, whose coordinated alternation supports memory-related processing, metabolic equilibrium, and immune regulation [2]. For this reason, robust automatic staging is a prerequisite for deriving quantitative sleep biomarkers, thereby facilitating early screening of conditions such as chronic insomnia and obstructive sleep apnea. Conversely, insufficient or fragmented sleep is frequently accompanied by measurable impairments in executive function and emotional control [3,4]. From a mechanistic perspective, rapid eye movement (REM) sleep is closely associated with synaptic remodeling and affective processing, whereas non-rapid eye movement (NREM) sleep—particularly slow-wave activity—contributes to cerebral energy restoration and glymphatic clearance. Although differentiating these states is clinically important, the prevailing gold standard—manual PSG scoring—remains time-consuming and is subject to considerable inter-rater variability [5]. The burden and subjectivity of manual polysomnographic scoring have therefore accelerated interest in automated solutions that are both reproducible and scalable. In contrast to conventional machine-learning workflows that depend on empirically designed features, deep neural networks (DNNs) learn task-relevant representations directly from physiological waveforms in an end-to-end manner. By capturing temporal structure across multiple scales, such models have shown improved generalization to diverse subjects and recording settings [6,7]. However,

 doi:10.20944/preprints202601.0930.v1

practical deployment in real-world polysomnography remains challenged by pronounced signal heterogeneity and modality-specific artifacts, which can compromise the reliability of naive fusion strategies and motivate more structured modeling of cross-channel interactions.

Given the high annotation cost of PSG and the prevalence of distributional shifts across cohorts, a series of advanced learning paradigms has been introduced to improve both robustness and adaptability in sleep staging models. In this context, transfer learning has been widely adopted, in which representations learned from large-scale datasets are reused so that models can be adapted to smaller cohorts or population-specific recordings with reduced reliance on extensive retraining [8]. Closely related to this line of work, meta-learning approaches—exemplified by MetaSleepLearner—have been formulated to treat rapid subject-level adaptation as an explicit objective; this is typically achieved via model-agnostic optimization procedures that enable efficient updating when new subjects are encountered [9]. This emphasis on adaptability becomes particularly relevant when inter-subject variability dominates performance degradation in real-world deployments. In parallel, when computational and memory budgets constitute the primary bottleneck, knowledge distillation has been explored as a practical means of reducing model complexity, allowing compact student models to approximate the predictive behavior of larger teachers and thereby supporting deployment on wearable and embedded platforms without substantial loss of accuracy [10]. A growing body of recent work suggests that hybrid architectures combining multi-scale convolutional feature extraction with temporal attention mechanisms are effective at modeling the coupled spectral-temporal structure of sleep-related physiological signals [11–13]. These local-global modeling strategies have advanced benchmark performance; nonetheless, robustness in practical PSG settings remains limited when auxiliary channels become intermittently artifact-dominated and are fused in a uniform manner. This reliability bottleneck is further exacerbated under distribution shifts and inter-subject variability, where noisy cross-channel cues can be propagated into the shared representation and impair generalization.

These observations suggest that robust PSG-based staging requires representations that couple multi-scale temporal structure with physiologically grounded spectral cues, as well as fusion mechanisms that explicitly account for time-varying modality reliability so that artifact-prone auxiliary evidence does not corrupt shared features. Accordingly, G-CMTF Net is developed as an end-to-end framework that integrates EEG with auxiliary EOG/EMG through reliability-aware gating and models long-range sleep dynamics using convolution-augmented attention. Guided by the above considerations, G-CMTF Net is designed around three key components, each targeting a recurring source of degradation in practical PSG-based staging:

1. Spectro-temporal heterogeneity (slow waves vs. spindles). A Spectro-Temporal Disentanglement module is introduced, where multi-scale convolutions are combined with spectral prior injection (FFT-derived band-power embeddings) to preserve physiologically defined oscillatory cues across disparate time scales.
2. Non-stationary artifacts in auxiliary modalities (EOG/EMG). To avoid indiscriminate fusion, a reliability-aware Gated Cross-Modal Fusion mechanism is employed to re-weight auxiliary streams at the epoch level, attenuating artifact-dominated segments while retaining informative cross-modal context.
3. Long-range temporal dependencies. A convolution-augmented self-attention encoder (Conformer) is used to model both local morphology and global stage-transition structure, which are difficult to capture with shallow temporal models.

The proposed framework is evaluated on Sleep-EDF-20 and Sleep-EDF-78 [14,15], yielding competitive performance while preserving sensitivity on transitional and minority stages under subject-wise evaluation.

To provide an intuitive view of the proposed workflow, Figure 1 presents a block-level overview of G-CMTF Net from multimodal PSG inputs to sleep-stage prediction; the full architectural specification is detailed later in Figure 2.

 doi:10.20944/preprints202601.0930.v1

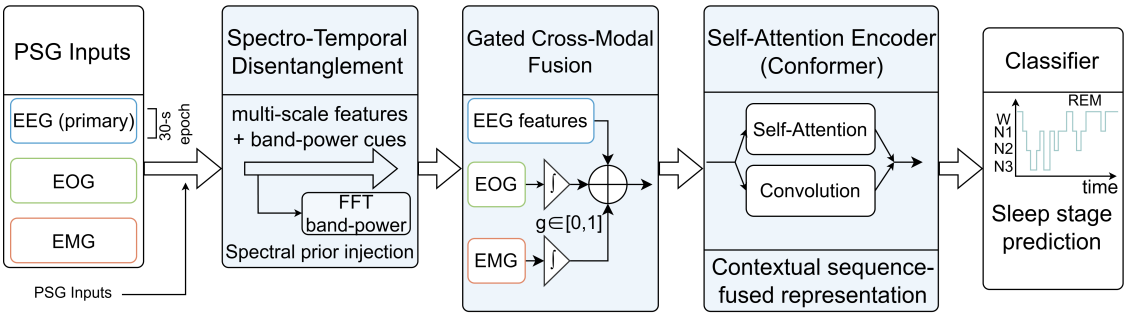


Figure 1. High-level overview of the proposed G-CMTF Net for PSG-based sleep staging. The schematic highlights the end-to-end pipeline (spectro-temporal encoding, reliability-aware cross-modal fusion, and temporal modeling).

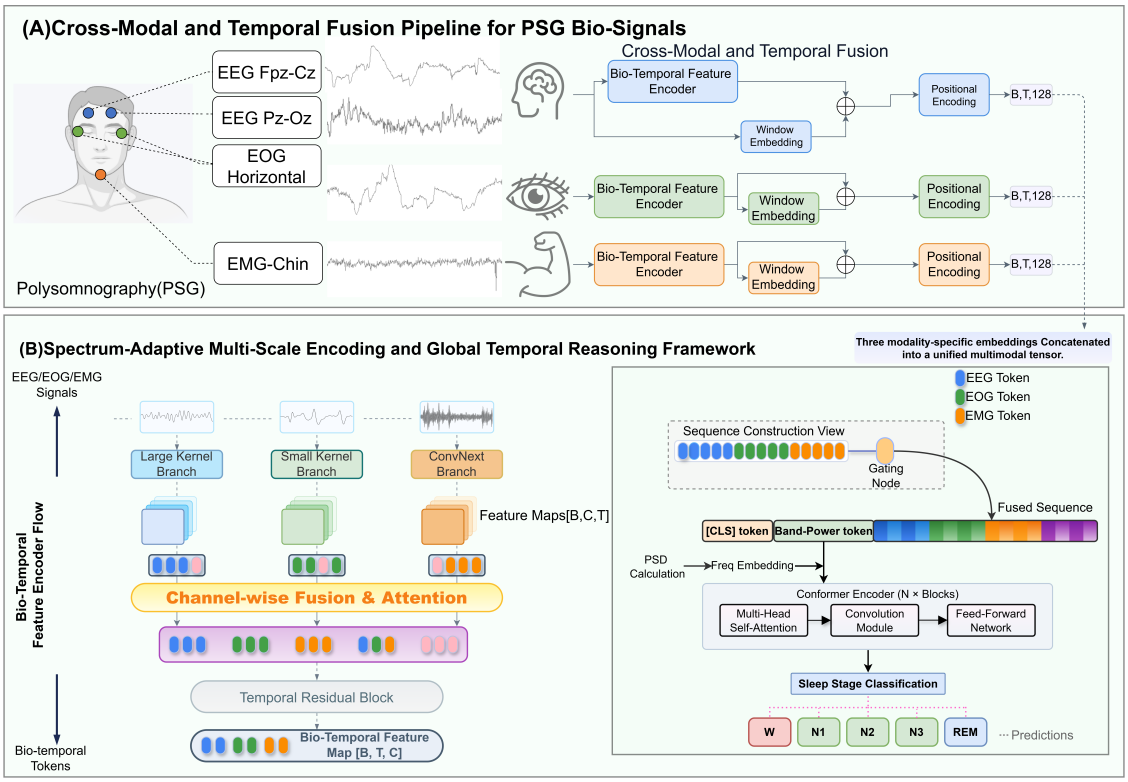


Figure 2. Overall architecture of the proposed G-CMTF Net, illustrating the cross-modal fusion pipeline, spectrum-aware bio-temporal feature encoding, and global temporal modeling with Conformer encoders.

2. Related Work

2.1. Transformers and Cross-Modal Temporal Modeling for Sleep Staging

The evolution of automated sleep staging has historically mirrored advancements in sequence modeling. Early deep learning frameworks established the efficacy of hybrid architectures that couple Convolutional Neural Networks (CNNs) for invariant feature extraction with Recurrent Neural Networks (RNNs) for transition rule learning [16]. However, LSTM-based paradigms are inherently impeded by gradient saturation and high computational latency when processing whole-night recordings [17]. To circumvent these bottlenecks, U-Time proposed a fully convolutional encoder-decoder framework, demonstrating that analyzing physiological signals via multi-scale convolutional streams could achieve state-of-the-art temporal segmentation without the computational burden of recurrent layers [18]. TinySleepNet further reported that, when the hybrid design was explicitly optimized for compression, the parameter budget could be substantially reduced while sleep scoring performance was largely preserved [19].

 doi:10.20944/preprints202601.0930.v1

As modeling requirements have evolved from short-range morphology characterization to long-range sleep-cycle dependency capture, a clear transition toward Transformer-based designs has been observed. Within this line of research, TransSleep leverages self-attention to target boundary epochs that are prone to misclassification; a transition-aware attention mechanism is introduced so that hierarchical semantic context can be encoded, which in turn reduces confusions that frequently persist in recurrent baselines during stage transitions [20]. A related yet distinct strategy has been adopted by hybrid architectures such as EEG-Conformer, where local temporal–spatial patterns are first extracted through a convolutional module and global dependencies are subsequently modeled by self-attention. By coupling these complementary inductive biases within a unified framework, competitive performance has been reported for electroencephalogram(EEG) decoding tasks [21].

Beyond single-channel pipelines, multi-view learning has been increasingly explored for sleep staging, as complementary information can be derived when signals are represented from distinct perspectives. XSleepNet exemplifies this direction by jointly encoding raw waveforms together with their time–frequency projections, thereby exploiting the complementarity between the two representations for improved stage discrimination [22]. Along similar lines, SalientSleepNet employs a U-shaped architecture to detect salient waveforms across heterogeneous modalities [23]. This trend, nevertheless, exposes a recurring limitation: multimodal performance is often contingent on heuristic fusion rules that implicitly presume consistent and non-conflicting contributions across modalities. For example, SalientSleepNet combines EEG and EOG representations using simple element-wise addition, which does not explicitly account for modality-specific corruption or cross-modal conflict. In clinical recordings, where EOG and EMG channels are frequently affected by non-stationary artifacts, the absence of a mechanism that can down-weight noise-dominant modalities on a per-epoch basis becomes a consequential gap in current representation learning practice. The above observation suggests that static arithmetic fusion should be replaced by context-aware integration strategies, in which signal reliability is estimated and fusion weights are adjusted accordingly.

2.2. Spectro-Temporal Representation and Multi-Scale Feature Dynamics

Sleep neurophysiology exhibits spectral activity that is distributed across markedly different time scales, ranging from the high-amplitude delta oscillations (0.5–4 Hz) that dominate N3 to the short-lived sleep spindles (11–16 Hz) that are characteristic of N2 [24]. Such heterogeneity imposes a practical constraint on conventional convolutional encoders, whose receptive fields are fixed once the kernel configuration is specified. Consequently, a single receptive-field scale is unlikely to represent long-duration slow waves and brief micro-events with comparable fidelity; when the field is enlarged, transient structures may be smoothed, whereas overly local fields tend to underrepresent slow-wave dynamics that unfold over extended intervals. This observation motivates multi-resolution feature extraction, as exemplified by early designs such as DeepSleepNet, which adopted a two-branch convolutional architecture with parallel pathways of different kernel sizes to separate temporal features at multiple resolutions [16]. Recent architectures such as MSTCN have adopted a genuinely multiscale temporal convolutional design, in which parallel convolutional pathways with heterogeneous kernel sizes and dilation factors are deployed to systematically capture discriminative temporal structures spanning short- and long-range dynamics [25]. Furthermore, contemporary hybrid paradigms demonstrate that aggregating these multi-scale contexts—often via attention-based or CNN-Transformer fusion—ensures that both local morphological details and broad contextual dependencies are rigorously preserved [13,26].

Closely related to multi-scale temporal modeling, the explicit incorporation of spectral priors provides a principled means of aligning learned representations with established clinical scoring conventions. Although convolutional architectures are capable of implicitly approximating spectral decompositions through learned filters, representations derived purely from data-driven optimization are not guaranteed to respect the well-defined frequency boundaries prescribed by standard scoring criteria [27]. To address this deficit, recent paradigms have moved towards Multi-View Learning, treating different signal representations as complementary input streams [28]. For example, the MST-GCN

 doi:10.20944/preprints202601.0930.v1

framework introduces a multi-view spatial–temporal graph architecture in which multiple adjacency matrices articulate distinct inter-channel and inter-epoch relationships. This formulation enables the model to integrate complementary structural dependencies that govern stage-to-stage transitions with greater fidelity [29]. Building on this line of research, architectures such as MVFSleepNet further show that coupling multi-scale convolutional features derived from raw PSG signals with time–frequency representations, and integrating them through an adaptive multi-view attention module, yields a richer and more balanced characterization of spectral and temporal dynamics. This joint representation strategy has been empirically demonstrated to enhance sleep-stage classification performance across multiple public datasets [30]. Nevertheless, a prevailing bottleneck in these multi-view frameworks is the computational redundancy associated with processing high-dimensional spectrogram images alongside raw waveforms. Furthermore, existing fusion mechanisms often treat spectral and temporal features as static parallel streams, lacking the flexibility to dynamically recalibrate spectral priors based on the signal's real-time quality. Consequently, establishing a unified, lightweight framework that explicitly disentangles spectro-temporal dynamics while maintaining computational efficiency remains an important yet insufficiently explored direction.

3. Motivation

Sleep staging from PSG is shaped by recurrent failure modes rooted in both physiology and acquisition artifacts. Figure 3 summarizes three such factors and the corresponding design choices in G-CMTF Net.

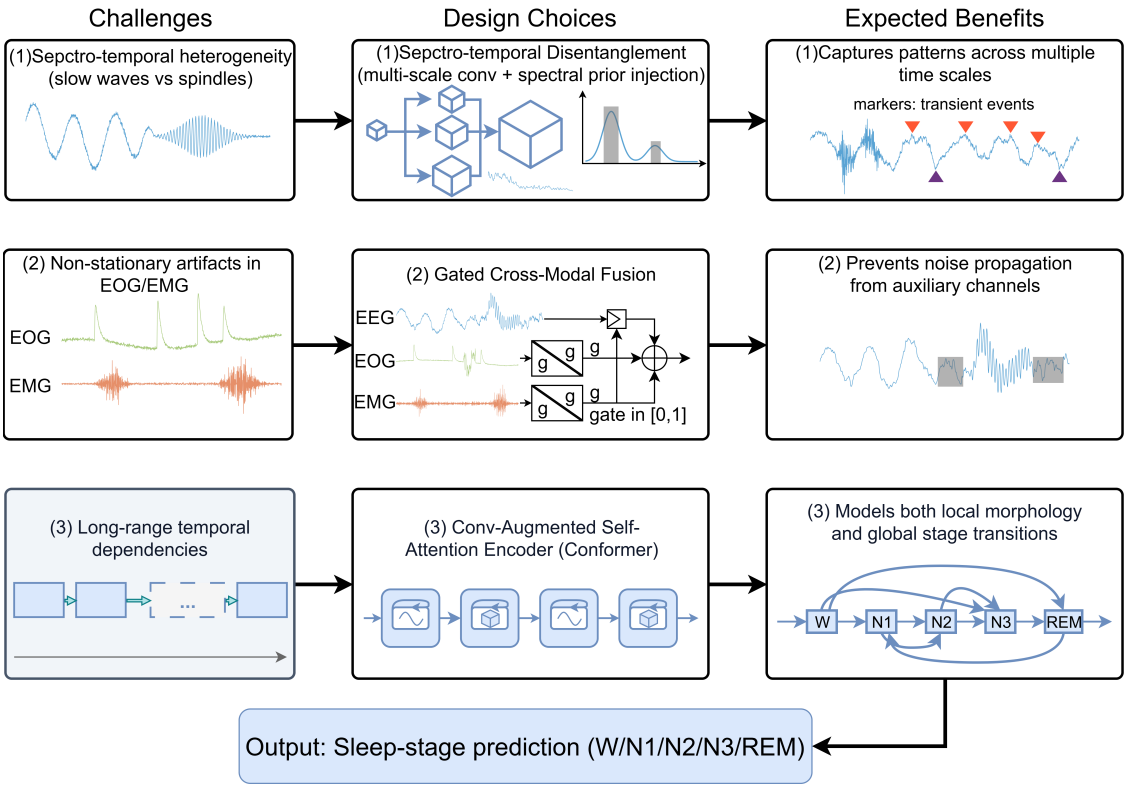


Figure 3. Motivation-driven overview of G-CMTF Net for PSG-based sleep staging. Key challenges are mapped to the corresponding modules and expected benefits.

First, discriminative signatures span disparate time scales: slow-wave activity develops over extended intervals, whereas micro-events such as spindles and K-complexes are temporally localized. Because single-scale convolutions impose an effectively fixed receptive field, a unified representation may under-resolve either long-duration dynamics or short-lived transients. Moreover, even when multi-scale encoders are adopted, learned features are not guaranteed to respect clinically defined

 doi:10.20944/preprints202601.0930.v1

frequency bands. This observation motivates injecting explicit spectral priors via FFT-derived band-power cues to steer representation learning toward physiologically interpretable regimes.

Relatedly, auxiliary channels (EOG/EMG) provide complementary evidence but exhibit time-varying reliability in practice. Existing fusion pipelines often weight modalities to optimize average discrimination (e.g., channel-wise attention in MASleepNet [31]); nevertheless, modality contributions are rarely regulated explicitly under non-stationary channel quality, allowing artifact-dominated segments to contaminate the shared latent space [32]. G-CMTF Net therefore formulates fusion as a reliability-sensitive operation and employs epoch-level gated cross-modal fusion to suppress unreliable auxiliary evidence while preserving informative context.

Finally, accurate staging depends on long-range temporal structure, including stage persistence and constrained transitions. To couple local morphology with global dynamics, a convolution-augmented self-attention encoder is adopted to model fine-grained waveform patterns together with sequence-level transition regularities.

4. G-CMTF Net: Gated Cross-Modal and Temporal Fusion for Sleep Staging

To rigorously deconvolute the spectral-temporal heterogeneity of polysomnography while mitigating cross-channel artifacts, we formulate the G-CMTF Net, which integrates gated cross-modal fusion with hierarchical temporal modeling for sleep staging. This architecture is instantiated via a hierarchical pipeline that synergizes three coupled functional blocks: (1) a Spectro-Temporal Disentanglement frontend, which augments multi-scale convolutional feature maps with explicit band-power priors to enforce frequency-domain fidelity; (2) a Gated Cross-Modal Fusion (GCMF) interface, designed to autonomously attenuate noise-corrupted modalities through learnable reliability weights before integration; and (3) a Conformer-based Temporal Encoder, which synthesizes local inductive biases with global self-attention mechanisms to reconstruct long-range sleep cycle transitions. The schematic workflow of the proposed framework is illustrated in Figure 2.

4.1. Overall Architecture

4.1.1. Problem Formulation

Formally, let the polysomnography recording for a single subject be represented as a sequence of N epochs, denoted as $\mathcal{X} = \{X_1, X_2, \ldots, X_N\}$. Each epoch $X_t \in \mathbb{R}^{C \times L}$ comprises multi-channel physiological signals, where C denotes the channel set $\{EEG, EOG, EMG\}$ and L represents the sampling length (typically $L = 3000$ for a 30s epoch at 100Hz). The objective of automatic sleep staging is to learn a non-linear mapping function $\mathcal{F} : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{Y} = \{y_1, y_2, \ldots, y_N\}$ corresponds to the sequence of discrete sleep stages $y_t \in \{W, N1, N2, N3, REM\}$ strictly strictly adhering to AASM standards [33].

4.1.2. Architectural Overview

To approximate the mapping $\mathcal{F}$ under cross-channel noise and modality-dependent corruption, the Gated Cross-Modal and Temporal Fusion Network (G-CMTF Net) is developed. As shown in Figure 2, the input X_t is processed through three hierarchically coupled stages:

1. Spectro-Temporal Feature Extraction: For each modality, raw waveforms are encoded by a Multi-Scale Sleep Stage CNN, in which multiple receptive-field scales are employed so that local sleep-related morphologies can be represented without committing to a single temporal resolution. In parallel, spectral priors are incorporated explicitly by a dedicated Band-Power Embedding module. This design is intended to preserve frequency-domain specificity and to ensure that the learned representations remain consistent with clinically defined oscillatory bands in neurophysiology.
2. Gated Cross-Modal Fusion: Since the reliability of different channels may vary across epochs, the resulting embeddings are forwarded to a Gated Cross-Modal Fusion (GCMF) module, where auxiliary modalities are re-weighted according to their estimated reliability prior to fusion.

 doi:10.20944/preprints202601.0930.v1

Drawing inspiration from channel-wise excitation principles [34], this module computes learnable reliability gates to autonomously suppress artifact-corrupted modalities (e.g., noisy EMG) before feature integration.

3. Global Temporal Context Modeling: The fused multimodal tokens are finally projected into a Conformer Encoder. By synthesizing local convolutional inductive biases with global self-attention mechanisms [35], this encoder reconstructs long-range sleep cycle transitions and inter-epoch dependencies with enhanced robustness.

The model is optimized end-to-end under a composite training objective that combines a focal loss term, introduced to mitigate the pronounced class imbalance typical of sleep-staging corpora, with auxiliary tasks tailored to waveform-level characteristics. This joint formulation encourages the network to remain sensitive to under-represented sleep stages while exploiting complementary supervisory signals.

4.2. Spectro-Temporal Feature Disentanglement

To resolve the scale-resolution uncertainty inherent in single-view modeling, the G-CMTF Net explicitly decouples feature extraction into two complementary streams: a spectral prior injection module and a multi-scale temporal encoder.

4.2.1. Explicit Spectral Prior Injection

Given an input epoch $X \in \mathbb{R}^{C\times L}$, we first transform the raw waveform into the frequency domain to capture steady-state oscillatory patterns. Let $x_c(t)$ denote the signal of the c-th channel. We apply Welch's method to compute the power spectral density (PSD), $S_c(f)$, which provides a robust estimate of spectral characteristics by averaging periodograms over overlapping segments. To align with clinical scoring rules [33], we aggregate the spectral energy into K canonical frequency bands (e.g., $\delta, \theta, \alpha, \sigma, \beta, \gamma$). The band-power vector $\mathbf{p}_c \in \mathbb{R}^K$ is derived as:

$$\mathbf{p}_{c,k} = \log\left(1 + \int_{f_{\text{low}}^{(k)}}^{f_{\text{high}}^{(k)}} S_c(f)\, df\right), \quad k = 1, \ldots, K \tag{1}$$

The logarithmic transformation mitigates the skewness of power distributions. These vectors are projected via a linear embedding to form the spectral token E_{spec}, serving as a global frequency anchor.

Physiologically, transitional stages such as N1 are typically characterized by mixed-frequency and low-amplitude activity, for which discriminative evidence is often manifested as subtle band-wise power shifts rather than salient morphology. By injecting band-power priors as an explicit global anchor, the learned representation is encouraged to preserve clinically grounded spectral contrasts (e.g., theta emergence and alpha attenuation), thereby complementing convolutional detection of transient events (e.g., spindles in the sigma band) and improving separability under inter-subject spectral variability.

4.2.2. Multi-Scale Temporal Convolution

Parallel to spectral analysis, we design a Multi-Scale Convolutional (MSC) encoder to disentangle morphological transients across varying temporal resolutions. The detailed architecture is illustrated in Figure 4. Formally, let $X \in \mathbb{R}^{C\times L}$ denote the standardized input epoch. The encoder Φ_{MSC} deploys three parallel branches $\mathcal{B} = \{\text{low}, \text{mid}, \text{high}\}$ to extract hierarchical representations. The feature map produced by branch $b \in \mathcal{B}$ is defined as

$$H_b = \mathcal{F}_b(X; \theta_b), \quad \text{with } \mathcal{F}_b \sim \text{Conv1D}(k_b, s_b). \tag{2}$$

In this setting, k_b and s_b are used to denote the kernel size and stride of branch b, respectively. Because sleep-related events are expressed over heterogeneous temporal spans, the branches are designed to impose complementary inductive biases rather than sharing an identical receptive-field configuration.

 doi:10.20944/preprints202601.0930.v1

The Low-Frequency Branch is configured with a larger kernel ($k_{\text{low}} = 31$), by which slow-wave activity that persists over longer intervals can be represented within an extended temporal context. The high-frequency branch is implemented using a depthwise–separable Conv1D with $k_{\text{high}} = 40$ (depthwise convolution with groups = in_channels followed by a 1×1 pointwise projection). An SE-style recalibration and a pooling operator are then applied to emphasize transient micro-events (e.g., spindle-/EMG-related bursts) while keeping the computational footprint bounded. The Mid-Frequency Branch is implemented with ConvNeXt blocks, where inverted bottlenecks are adopted to trade representational capacity against computational cost in a controlled manner [36].

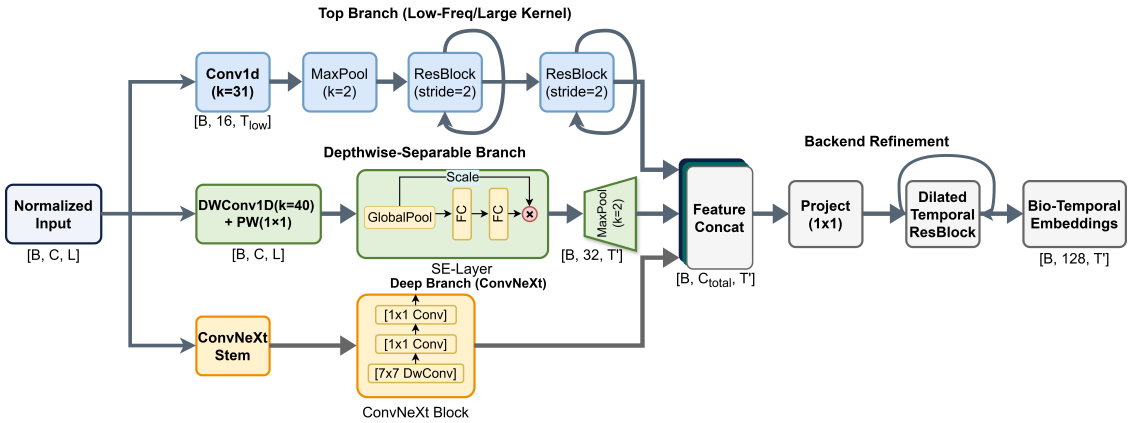


Figure 4. Architecture of the proposed multi-scale spectro-temporal CNN frontend. Three parallel convolutional branches with complementary receptive fields are employed to capture sleep-related patterns across different temporal scales. The resulting features are temporally aligned, concatenated, and refined through projection and residual temporal modeling to form bio-temporal embeddings.

Owing to the use of heterogeneous strides, the feature maps $H_{\text{low}}, H_{\text{mid}}, H_{\text{high}}$ are produced at different temporal resolutions. To enable coherent fusion across branches, a common target length is specified as $T' = \min_b(\text{len}(H_b))$, and a linear interpolation operator $\mathcal{U}(\cdot)$ is applied to align each feature map accordingly. After temporal alignment, the unified representation is formed by channel-wise concatenation followed by projection:

$$Z_{\text{agg}} = \text{BN}\Big(\mathbf{W}_p * \big[\mathcal{U}(H_{\text{low}}) \;\|\; \mathcal{U}(H_{\text{mid}}) \;\|\; \mathcal{U}(H_{\text{high}})\big]\Big), \tag{3}$$

where $\|$ denotes concatenation along the channel dimension and $\mathbf{W}_p$ denotes a 1×1 convolution used to project the aggregated features into a compact embedding space.

After projection, the fused sequence is refined by a lightweight channel–temporal gating pipeline. First, a channel-wise SE gate is applied to re-weight feature responses:

$$Z_{\text{ch}} = Z_{\text{agg}} \otimes \sigma(\mathcal{A}_{\text{ch}}(Z_{\text{agg}})), \tag{4}$$

where $\mathcal{A}_{\text{ch}}(\cdot)$ denotes a squeeze–excitation operator producing channel-wise weights, which are broadcast along the temporal axis.

Next, temporal refinement is implemented by a dilated residual block followed by a temporal SE gate. For notational compactness, their composition is denoted as $\mathcal{A}_{\text{temp}}(\cdot)$:

$$E_{\text{temp}} = \mathcal{A}_{\text{temp}}(Z_{\text{ch}}), \qquad \mathcal{A}_{\text{temp}}(Z) = \Big(Z + \mathcal{R}_{\text{dil}}(Z)\Big) \otimes \sigma\Big(\mathcal{A}_{\text{se}}\big(Z + \mathcal{R}_{\text{dil}}(Z)\big)\Big), \tag{5}$$

where $\mathcal{R}_{\text{dil}}(\cdot)$ is a dilated temporal residual block (dilation = 2 in our implementation) and $\mathcal{A}_{\text{se}}(\cdot)$ denotes the temporal SE gate. This design matches the implementation in which the fused sequence is channel-gated, refined by a dilated residual update, and finally re-weighted by a temporal SE gate to suppress artifact-amplified fluctuations while preserving transition-relevant dynamics.

 doi:10.20944/preprints202601.0930.v1

Through this joint channel–temporal recalibration, the resulting bio-temporal embedding E_{temp} is rendered less sensitive to artifactual perturbations while retaining discriminative sleep-related patterns [34].

The MSC encoder comprises three parallel branches whose outputs are aligned to a common temporal length before fusion. (1) The low-frequency branch applies a strided Conv1D (k=31, s=2, p=15) producing 16 channels, followed by MaxPool1D (k=2) and two residual Conv1D blocks with downsampling (16→32→128; each block uses stride 2). (2) The mid-frequency branch adopts a ConvNeXt-style stem (Conv1D k=7, s=2, p=3 + MaxPool1D k=2) followed by two residual blocks, yielding 64 channels. (3) The high-frequency branch uses a depthwise Conv1D (k=40, s=1, p=20; groups=in_channels) with a pointwise 1×1 projection to 32 channels, together with SE-style recalibration and MaxPool1D (k=2). Let $T_{\text{low}}, T_{\text{mid}}, T_{\text{high}}$ denote the resulting temporal lengths; the three feature maps are linearly interpolated to $T' = \min(T_{\text{low}}, T_{\text{mid}}, T_{\text{high}})$, concatenated (224 channels), and projected to 128 dimensions by a 1×1 Conv1D with BatchNorm and dropout. Finally, channel-wise gating and a dilated temporal residual refinement block are applied, followed by an SE-style recalibration, to produce the bio-temporal embeddings.

4.3. Gated Cross-Modal Fusion

In multi-channel sleep recordings, auxiliary signals such as EOG and EMG may alternately provide complementary physiological context or be dominated by artifacts, giving rise to a recurring noise–information trade-off. Unlike static fusion, GCMF operates as a reliability-aware filter. To explicitly account for this uncertainty during fusion, a Gated Cross-Modal Fusion (GCMF) mechanism is formulated, in which auxiliary information is integrated in a reliability-aware manner rather than through static aggregation.

Let $Z_{\text{main}} \in \mathbb{R}^{T\times D}$ denote the EEG feature sequence (Query) and $Z_{\text{aux}} \in \mathbb{R}^{T'\times D}$ denote the auxiliary EOG/EMG features (Key/Value). The fusion process involves two sequential steps:

First, we introduce a cross-attention module, following the encoder-decoder attention mechanism originally formalized in the Transformer architecture [17]. This enables the auxiliary modality to attend selectively to temporally informative components of the primary EEG stream, thereby achieving consistent alignment between their dynamics:

$$\hat{Z}\text{aux} = \text{Softmax}\left(\frac{Z_{\text{main}}W_Q(Z_{\text{aux}}W_K)^\top}{\sqrt{d_k}}\right)(Z_{\text{aux}}W_V). \tag{6}$$

Second, to suppress noise, a learnable Gating Network (illustrated in the bottom panel of Figure 5) computes a channel-wise reliability score $g \in [0,1]^D$ based on the global context of the aligned features:

$$g = \sigma\big(\text{Convex}\big(\delta\big(\text{Convsq}\big(\text{AvgPool}(\hat{Z}_{\text{aux}})\big)\big)\big)\big), \tag{7}$$

where $\sigma(\cdot)$ is the Sigmoid function and $\delta(\cdot)$ denotes the ReLU activation. The final fused representation is obtained by effectively filtering the auxiliary signal: $Z_{\text{fused}} = \text{Norm}(Z_{\text{main}} + g \odot \hat{Z}_{\text{aux}})$. This ensures that only high-quality auxiliary information is integrated, while artifact-heavy segments are autonomously attenuated.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 January 2026 doi:10.20944/preprints202601.0930.v1

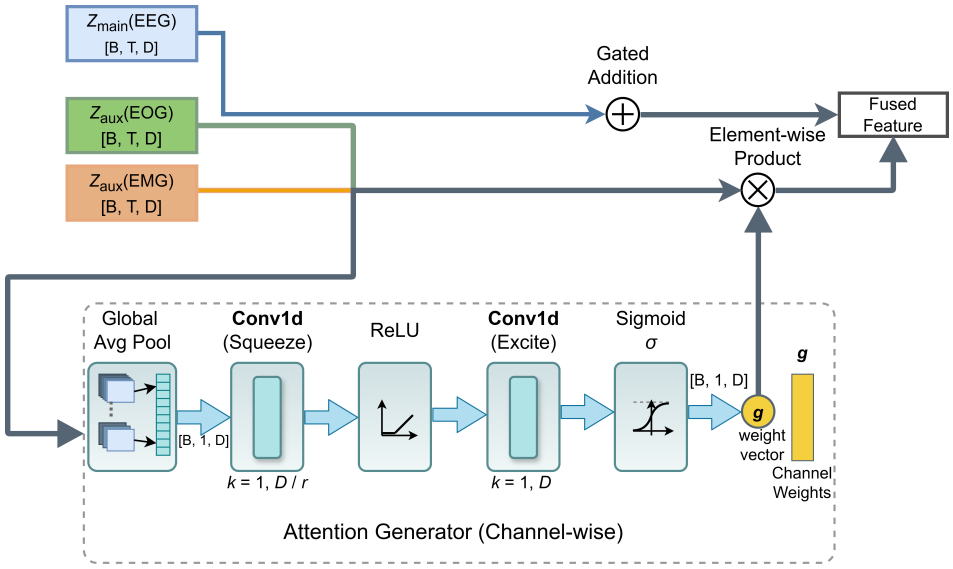


Figure 5. Gated Cross-Modal Fusion (GCMF) module. A channel-wise gating mechanism estimates modality reliability from global context and suppresses artifact-dominated auxiliary features prior to fusion with the EEG stream.

4.4. Global Temporal Context Modeling and Composite Optimization

4.4.1. Conformer-based Sequence Modeling

Following the cross-modal fusion, the integrated feature sequence $Z_{\text{fused}} \in \mathbb{R}^{T \times D}$ is passed through a Conformer Encoder to capture long-range temporal dependencies, such as the cyclic alternation between NREM and REM sleep stages. Unlike conventional Transformer architectures, the Conformer model integrates a convolution module within the self-attention block, as shown in Figure 6. This design merges the local inductive biases of convolutional neural networks (CNNs)—which are essential for capturing micro-events such as K-complexes—with the broader receptive field of Multi-Head Self-Attention (MHSA), facilitating robust modeling of both transient waveforms and prolonged sleep transitions across the night [35]. The output feature tokens are subsequently aggregated through Global Average Pooling (GAP) and passed through a linear classifier to produce the conditional probability distribution $P(y|X)$.

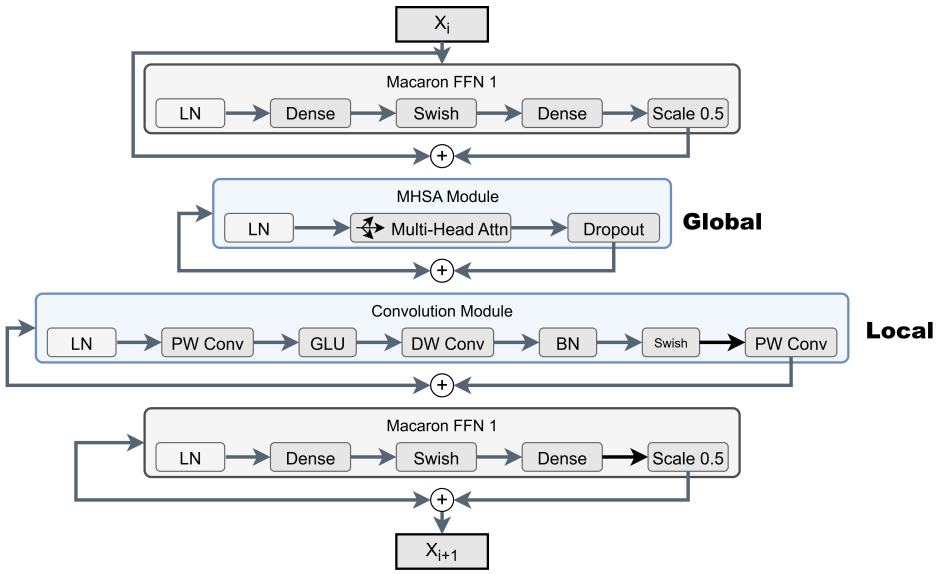


Figure 6. Internal architecture of a Conformer encoder block, in which global multi-head self-attention is integrated with a local convolutional module to jointly model long-range sleep-stage transitions and short-duration waveform events.

 doi:10.20944/preprints202601.0930.v1

4.4.2. Composite Loss Function with Deep Supervision

To mitigate the severe class imbalance inherent in sleep datasets (where N2 stage dominates), we formulate a composite objective function. The primary classification loss $\mathcal{L}_{\text{main}}$ employs a Focal Loss term modulated by Label Smoothing regularization. This combination down-weights the contribution of easy negatives while preventing the model from becoming over-confident on noisy labels:

$$\mathcal{L}_{\text{main}} = -\sum_{c=1}^{N_{\text{class}}} (1-p_c)^{\gamma} \log(p_c^{\text{LS}}), \quad p_c^{\text{LS}} = (1-\epsilon)y_c + \epsilon / N_{\text{class}} \tag{8}$$

where γ is the focusing parameter and ϵ is the smoothing factor. Furthermore, to enforce feature discriminability at intermediate layers, we introduce auxiliary supervision heads designed to regularize intermediate representations associated with characteristic sleep-related oscillatory patterns (e.g., spindles and slow-wave activity). The total optimization objective is defined as $\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{main}} + \lambda_2 \mathcal{L}_{\text{aux}}$, where λ controls the trade-off between global staging accuracy and local feature fidelity.

5. Experiments and Results

5.1. Datasets

All experiments were conducted using the Sleep-EDF Database Expanded (Sleep-EDFX), a publicly accessible polysomnography repository hosted on PhysioNet [14,15]. In this study, analysis was restricted to the Sleep Cassette subset, which contains full-night PSG recordings acquired from healthy subjects and is commonly adopted for modeling baseline sleep architecture in the absence of pharmacological intervention.

Sleep stages in Sleep-EDFX were originally annotated by clinical experts according to the Rechtschaffen and Kales (R&K) criteria. Following standard practice, these annotations were mapped to the AASM convention by merging stages N3 and N4 into a single deep sleep stage (N3). Epochs labeled as Movement (M) or Unclassified (?) were excluded from subsequent analysis to ensure label consistency.

To facilitate both benchmarking and robustness evaluation, two experimental cohorts were derived from the Sleep Cassette subset:

1. Sleep-EDF-20: which consists of 39 recordings from 20 subjects (SC4001–SC4192) and serves as a widely used reference benchmark for comparison with existing deep learning methods.
2. Sleep-EDF-78: comprising 153 recordings from 78 subjects, which introduces substantially greater inter-subject variability and provides a more stringent test of model generalization.

The class distribution statistics for both cohorts are summarized in Table 1.

Table 1. Experimental settings and dataset statistics.

Dataset	No. of subjects	Channels	Evaluation Scheme	Sampling Rate	Class Distribution					
					Wake	N1	N2	N3	REM	# Total
Sleep-EDF-20	20	EEG, EOG EMG	5-fold subject-wise CV	100 Hz	8285 (19.6%)	2804 (6.6%)	17799 (42.1%)	5703 (13.5%)	7717 (18.2%)	42308
Sleep-EDF-78	78	EEG, EOG EMG	5-fold subject-wise CV	100 Hz	65951 (33.7%)	21522 (11.0%)	69132 (35.4%)	13039 (6.7%)	25835 (13.2%)	195479

* CV: Cross Validation

5.2. Data Preprocessing and Implementation Details

In accordance with the multi-modal design of G-CMTF Net, three PSG channels were used throughout this study, namely EEG, horizontal EOG, and submental EMG, which respectively characterize cortical oscillations, ocular movements, and muscle tone. Raw signals were uniformly sampled at 100 Hz. To suppress low-frequency drift and high-frequency interference, zero-phase band-pass filtering was applied with cut-off frequencies of 0.2–40 Hz. The continuous recordings were then

 doi:10.20944/preprints202601.0930.v1

segmented into non-overlapping 30-second epochs, consistent with the AASM scoring convention. To reduce inter-subject amplitude variability, Z-score normalization was performed on a subject-wise basis for each channel, yielding zero-mean and unit-variance inputs.

The complete preprocessing workflow, including the parallel construction of time-domain waveforms and spectral priors, is summarized in Figure 7. Specifically, in addition to the normalized epoch sequences used for temporal encoding, a Welch-based power spectral density (PSD) estimate was computed per epoch to derive band-power vectors, which were subsequently embedded as explicit spectral cues. This design allows clinically interpretable oscillatory information to be retained during representation formation, rather than being inferred implicitly from raw waveforms alone.

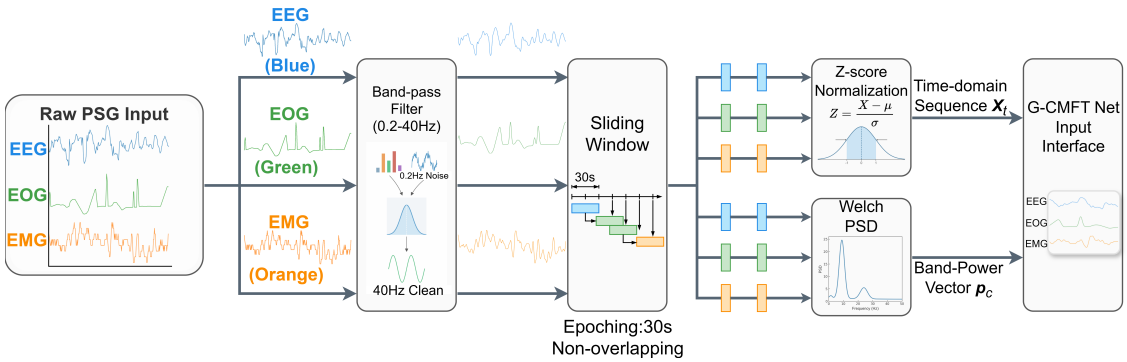


Figure 7. Overview of the spectro-temporal input construction pipeline for G-CMTF Net. Time-domain sequences and explicit spectral representations are constructed in parallel and jointly provided as model inputs.

Experiments were conducted on two commonly adopted cohorts derived from the Sleep-EDFX Sleep Cassette subset. Sleep-EDF-20 comprises 39 recordings from 20 subjects (SC4001–SC4192) and is used as a standard benchmark for comparison with prior deep learning approaches. Sleep-EDF-78 extends the evaluation to 153 recordings from 78 subjects, introducing substantially greater inter-subject heterogeneity and therefore providing a more stringent test of robustness. For both cohorts, 5-fold cross-validation was employed to assess generalization. We strictly adopted subject-wise cross-validation, ensuring that recordings from the same subject never appeared in both training and testing sets. Each fold contained mutually exclusive subject sets, thereby preventing information leakage across folds. We report the mean and 95% confidence interval (CI) of Macro-F1 over 5-fold cross-validation. The CI is computed using the t-distribution with four degrees of freedom.

Model implementation and training were carried out using PyTorch on a single NVIDIA RTX 4090 GPU. Optimization was performed using AdamW with a learning rate of 5×10^{-4} and weight decay of 0.01. A cosine-annealing learning-rate scheduler was used during training. Early stopping was applied by monitoring the validation F1-score, and training was terminated if no improvement was observed for 20 consecutive epochs.

5.3. Evaluation Metrics

To evaluate performance under class imbalance, we report Accuracy (ACC), Macro-F1 (MF1), Cohen's κ, and Macro G-Mean (MGm). Let TP_c, TN_c, FP_c, FN_c denote the confusion counts for class $c \in \{W, N1, N2, N3, REM\}$, and N be the total number of epochs.

(1) Accuracy (ACC).

$$\mathrm{ACC} = \frac{\sum_{c=1}^{C} TP_c}{N}. \tag{9}$$

(2) Macro-F1 (MF1). We compute per-class precision and sensitivity (recall) and average class-wise F1 over C classes:

$$\mathrm{Pre}_c = \frac{TP_c}{TP_c + FP_c}, \quad \mathrm{Sen}_c = \frac{TP_c}{TP_c + FN_c}, \tag{10}$$

$$\mathrm{MF1} = \frac{1}{C} \sum_{c=1}^{C} \frac{2\,\mathrm{Pre}_c \mathrm{Sen}_c}{\mathrm{Pre}_c + \mathrm{Sen}_c}. \tag{11}$$

 doi:10.20944/preprints202601.0930.v1

(3) Cohen's κ. Agreement beyond chance is measured as

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \tag{12}$$

where p_o is the observed agreement (ACC) and p_e is the chance agreement from class marginals.
(4) Macro G-Mean (MGm). To reflect balanced sensitivity–specificity under imbalance, we compute

$$\mathrm{Spec}_c = \frac{TN_c}{TN_c + FP_c}, \quad \mathrm{MGm} = \left(\prod_{c=1}^{C} \sqrt{\mathrm{Sen}_c \, \mathrm{Spec}_c} \right)^{\frac{1}{C}}. \tag{13}$$

5.4. Performance Evaluation on Sleep-EDF Benchmarks

5.4.1. Overall Performance Analysis

The overall discriminative capability of G-CMTF Net was evaluated on Sleep-EDF-20 and Sleep-EDF-78 by benchmarking against representative baselines in Table 2.

Table 2. Overall performance (%) comparison on Sleep-EDF-20 and Sleep-EDF-78. Results are taken from the corresponding publications unless otherwise stated; "–" indicates not reported. Best results are highlighted in bold. Reported results may involve different preprocessing and evaluation protocols; we quote the numbers as reported in the publications.

Dataset	Methods	ACC	MF1	κ	SEN	SPE	MGm	Channels
Sleep-EDF-20	DeepSleepNet [16]	82.0	76.9	0.76	-	-	-	EEG
	SleepEEGNet [37]	84.3	79.7	0.79	-	-	-	EEG
	MultitaskCNN [11]	83.1	75.0	0.77	-	-	83.1	EEG
	TinySleepNet [38]	85.4	80.5	0.80	-	-	-	EEG
	XSleepNet1 [22]	85.2	79.8	0.798	79.0	95.9	87.0	EEG,EOG
	XSleepNet2 [22]	86.4	80.9	0.813	79.9	**96.2**	87.6	EEG,EOG
	Naïve Fusion [22]	83.4	77.8	0.773	77.1	95.5	85.8	EEG,EOG
	FCNN+RNN [22]	83.5	77.7	0.775	77.2	95.5	85.9	EEG,EOG
	MASleepNet [31]	84.5	78.9	0.785	78.4	95.7	86.6	Multi*
	MMASleepNet [32]	**87.3**	**82.7**	**0.826**	-	-	81.7	Multi*
	G-CMTF Net (Ours)	85.5±2.0	81.3	0.802±0.028	**82.1**	96.1	**88.5**	Multi*
Sleep-EDF-78	DeepSleepNet [38]	77.1	71.2	0.69	-	-	-	EEG
	SleepEEGNet [37]	80.0	73.6	0.73	-	-	-	EEG
	MultitaskCNN [11]	79.6	72.8	0.72	-	-	82.5	EEG
	TinySleepNet [38]	83.1	78.1	0.77	-	-	-	EEG
	XSleepNet1 [22]	**84.0**	78.4	0.777	77.1	95.6	85.9	EEG,EOG
	XSleepNet2 [22]	**84.0**	**77.9**	**0.778**	77.6	**95.7**	86.2	EEG,EOG
	Naïve Fusion [22]	82.5	76.9	0.757	75.8	95.3	85.0	EEG,EOG
	FCNN+RNN [22]	82.7	76.9	0.759	75.5	95.3	84.8	EEG,EOG
	MASleepNet [31]	82.6	76.1	0.75	75.9	95.2	85.0	Multi*
	MMASleepNet [32]	82.7	77.6	0.76	-	-	76.1	Multi*
	G-CMTF Net (Ours)	83.4±1.4	78.2	0.771±0.017	**78.5**	95.6	**86.6**	**Multi***

*Multi: Indicates the use of the full core PSG channels for sleep staging (EEG + EOG + EMG).

On Sleep-EDF-20, G-CMTF Net attains a Macro-F1 of 81.3% with a Cohen's κ of 0.802, computed on pooled test predictions across the 5 folds. Across 5-fold subject-wise cross-validation, the fold-wise mean Macro-F1 is 81.29% (95% CI: 79.03–83.55), with a fold-wise mean κ of 0.802 (95% CI: 0.765–0.835). When contrasted with EEG-only pipelines such as DeepSleepNet and SleepEEGNet, the gain is consistent with the additional physiological context provided by EOG/EMG, which is not accessible to single-stream models and therefore cannot be exploited to resolve epochs whose morphology is weakly expressed in EEG alone. With that said, multi-modal inputs do not automatically translate into robustness, because auxiliary channels may intermittently become artifact-dominated; this noise–information trade-off is exactly where reliability modeling becomes consequential. In this respect, although MMASleepNet reports slightly higher aggregate scores, G-CMTF Net yields the highest

 doi:10.20944/preprints202601.0930.v1

Sensitivity (82.1%) and Macro G-Mean (88.5%), indicating fewer misses on minority and transitional stages under class imbalance. This stage-sensitive improvement aligns with the proposed design: explicit spectro-temporal priors constrain feature formation toward clinically defined oscillatory bands, while the gated cross-modal fusion suppresses unreliable auxiliary contributions before they can perturb the fused representation.

A similar trend is observed on Sleep-EDF-78, where inter-subject heterogeneity is markedly increased. G-CMTF Net reaches 83.4% ACC and 78.2% MF1 (with a Cohen's κ of 0.771), computed on pooled test predictions across the 5 folds. Across 5-fold subject-wise cross-validation, the fold-wise mean Macro-F1 is 78.16% (95% CI: 76.54–79.64), with a fold-wise mean κ of 0.771 (95% CI: 0.743–0.799). This represents a 6.3% absolute accuracy gain over DeepSleepNet. Performance remains comparable to sequence-based architectures (e.g., SeqSleepNet and XSleepNet), which explicitly model cross-epoch dependencies, while a balanced operating regime is preserved as reflected by Sensitivity (78.5%) and Specificity (95.6%). Collectively, the results suggest that combining (i) spectro-temporal priors for resolving scale-dependent ambiguity with (ii) reliability-aware gating for mitigating cross-channel artifacts improves robustness under non-stationary PSG noise, without sacrificing majority-stage specificity.

5.4.2. Class-wise Performance Analysis

To elucidate where the proposed improvements arise, stage-wise F1 scores are reported in Table 3. Across existing baselines, N1 remains the most difficult stage, largely because its low-amplitude and mixed-frequency characteristics provide weak and easily perturbed cues, especially around transitions.

Table 3. Stage-wise F1 scores (%) on Sleep-EDF-20 and Sleep-EDF-78. Best results per stage are highlighted in bold. Only methods with publicly reported per-stage F1 scores are included.

Methods	W	N1	N2	N3	REM
Dataset: Sleep-EDF-20					
DeepSleepNet [38]	84.7	46.6	85.9	84.8	82.4
TinySleepNet [38]	90.1	51.4	**88.5**	**88.3**	84.3
SeqSleepNet [38]	90.5	45.4	88.1	86.4	81.8
MultitaskCNN [11]	87.9	33.5	87.5	85.8	80.3
SleepEEGNet [37]	89.2	52.2	86.8	85.1	85.0
G-CMTF Net (Ours)	**91.0**	**56.2**	87.0	83.9	**88.3**
Dataset: Sleep-EDF-78					
DeepSleepNet [38]	90.4	46.0	79.1	68.6	71.8
TinySleepNet [37]	92.8	51.0	85.3	**81.1**	80.3
SeqSleepNet [7]	92.2	47.8	84.9	77.2	79.9
MultitaskCNN [11]	90.9	39.7	83.2	76.6	73.5
SleepEEGNet [37]	91.7	44.1	82.5	73.5	76.1
G-CMTF Net (Ours)	**93.2**	**51.4**	84.7	76.6	85.0

On Sleep-EDF-20, G-CMTF Net attains an N1 F1 score of 56.2%, exceeding DeepSleepNet (46.6%) and MultitaskCNN (33.5%), and also outperforming other competitive EEG-based baselines (e.g., SleepEEGNet at 52.2%). This gain is consistent with the proposed reliability-aware fusion: auxiliary streams are not incorporated indiscriminately, but are adaptively down-weighted when dominated by artifacts, which helps preserve subtle transition-related patterns that would otherwise be masked in the fused space. In addition, the REM stage is recognized with high fidelity on this cohort (88.3%), suggesting that the injected spectral priors and multi-scale encoding contribute useful discriminative cues for stages with distinct oscillatory structure.

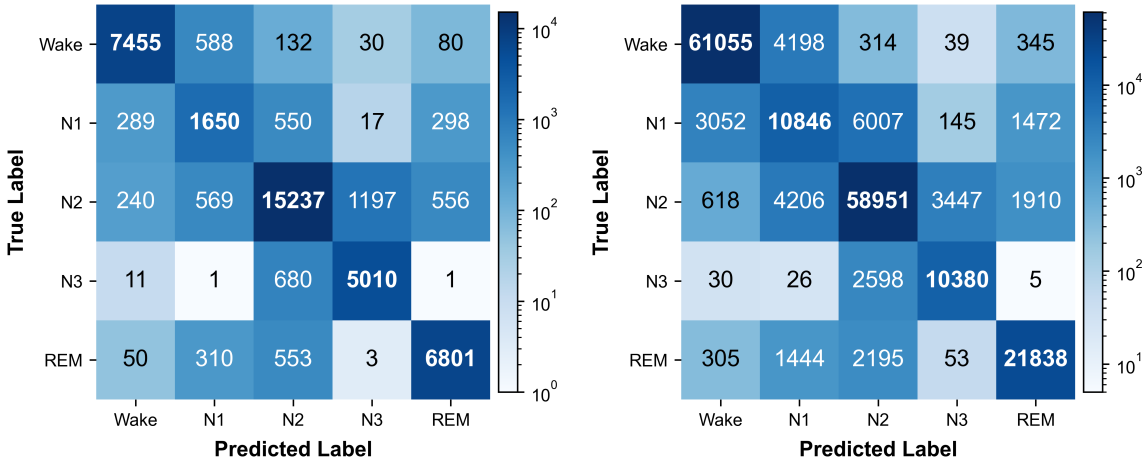
When the evaluation is extended to Sleep-EDF-78, where inter-subject heterogeneity is substantially higher, G-CMTF Net remains competitive on N1 (51.4%), ranking among the strongest reported results in the table. Meanwhile, the best scores for certain stages (e.g., N3 or REM) are achieved by

 doi:10.20944/preprints202601.0930.v1

other sequence models, indicating that stage-wise performance is not uniformly improved across all classes. Overall, the stage-wise profile suggests that the proposed design primarily benefits ambiguous and transition-prone epochs (particularly N1), while maintaining stable performance on the remaining stages under a more challenging cross-subject setting.

5.4.3. Stage-wise Error Analysis via Aggregated Confusion Matrices

To further characterize stage-wise error patterns, the confusion matrices aggregated over 5-fold cross-validation are reported in Figure 8.



(a) Sleep-EDF-20 (b) Sleep-EDF-78

Figure 8. Aggregated confusion matrices over 5-fold cross-validation on Sleep-EDF-20 and Sleep-EDF-78. Values denote epoch counts. Most confusions occur between physiologically adjacent stages, notably N1↔N2 and REM↔N2.

For both cohorts, predictions are largely concentrated on the diagonal, indicating that the dominant stages (Wake, N2, N3, and REM) are recognized with stable consistency under multimodal inputs. The residual errors are not uniformly distributed; instead, they are primarily localized to physiologically adjacent transitions. In particular, the most frequent confusions occur between N1 and N2 (e.g., N1→N2 and N2→N1), reflecting the well-known overlap of low-amplitude mixed-frequency activity at the onset of sleep. A second recurrent pattern is the interaction between N2 and N3 (notably N3→N2 and N2→N3), which is consistent with the gradual emergence of slow-wave activity and its boundary ambiguity under epoch-wise labeling. In addition, a non-negligible portion of REM epochs is mapped to N2, whereas Wake is occasionally mapped to N1, suggesting that short-lived transitions and artifact-affected segments remain challenging even with gated fusion. Overall, the confusion structure indicates that most errors are constrained to neighboring-stage substitutions rather than long-range jumps, which is aligned with the design goal of G-CMTF Net to suppress artifact-driven noise propagation while improving separability for transition-prone epochs.

5.5. Ablation Studies

To isolate the contribution of key components and to examine whether their effects are consistent across data scales, an ablation study was conducted on both Sleep-EDF-20 and Sleep-EDF-78 under an identical training protocol and subject-wise 5-fold split. The results are reported in Table 4.

Replacing the proposed Gated Cross-Modal Fusion with static concatenation (w/o Gating (Concat)) reduced MF1 by 0.5 pp on Sleep-EDF-20 (81.3% to 80.8%) and by 0.4 pp on Sleep-EDF-78 (78.2% to 77.8%), with κ decreasing from 0.802 to 0.797 and from 0.771 to 0.767, respectively. These results are consistent with the view that reliability-aware re-weighting helps limit the influence of unreliable auxiliary cues during fusion.

 doi:10.20944/preprints202601.0930.v1

Table 4. Ablation study on Sleep-EDF-20 and Sleep-EDF-78. Evaluation of key components including Gated Fusion, Band-Power Injection, Multi-Scale CNN, and auxiliary EMG.

Model Variants	Sleep-EDF-20 (5-fold)			Sleep-EDF-78 (5-fold)		
	ACC (%)	MF1 (%)	κ	ACC (%)	MF1 (%)	κ
G-CMTF Net	**85.5**	**81.3**	**0.802**	**83.4**	**78.2**	**0.771**
w/o Gating (Concat)	85.1	80.8	0.797	83.1	77.8	0.767
w/o Band-Power	**85.5**	80.8	0.800	83.0	77.2	0.765
w/o Multi-Scale	85.3	80.6	0.799	83.1	77.3	0.766
w/o EMG	85.3	81.0	0.800	83.1	77.2	0.767

Removing the Band-Power Embedding branch (w/o Band-Power) produced a larger degradation on Sleep-EDF-78 than on Sleep-EDF-20. On Sleep-EDF-78, MF1 decreased by 1.0 pp (78.2% to 77.2%) and κ decreased from 0.771 to 0.765, whereas on Sleep-EDF-20 MF1 decreased by 0.5 pp (81.3% to 80.8%) with κ decreasing slightly from 0.802 to 0.800 while ACC remained unchanged (85.5%). This cohort-dependent behavior suggests that explicitly injected spectral cues may become more beneficial when inter-subject variability is higher.

For the multi-scale CNN frontend, simplifying it to a single-scale configuration (w/o Multi-Scale) decreased MF1 by 0.7 pp on Sleep-EDF-20 (81.3% to 80.6%) and by 0.9 pp on Sleep-EDF-78 (78.2% to 77.3%), accompanied by κ reductions from 0.802 to 0.799 and from 0.771 to 0.766, respectively. This pattern supports the premise that stage-defining PSG events span multiple temporal extents, such that a single receptive-field scale provides less complete coverage.

Excluding the EMG stream (w/o EMG) leads to a cohort-dependent effect. On Sleep-EDF-20, MF1 decreases from 81.3% to 81.0% (0.3 pp), whereas on Sleep-EDF-78 MF1 decreases from 78.2% to 77.2% (1.0 pp) and κ decreases from 0.771 to 0.767. This gap is consistent with the view that EMG-derived muscle-tone cues become more consequential as cross-subject heterogeneity increases, particularly for ambiguity-prone epochs (e.g., REM/W-related decisions).

5.6. Parameter Sensitivity Analysis

To examine the role of architectural depth in temporal representation learning, a controlled sensitivity analysis was conducted by varying the number of Conformer encoder layers, with $L \in \{1, 2, 4, 6\}$. The corresponding Macro-F1 scores on Sleep-EDF-20 and Sleep-EDF-78 are reported in Table 5.

Table 5. Parameter sensitivity to Conformer depth L on Sleep-EDF-20 and Sleep-EDF-78 (5-fold). Macro-F1 scores are reported for different numbers of encoder layers.

Depth L	Sleep-EDF-20	Sleep-EDF-78
1	80.5	77.8
2	80.8	77.9
4	**81.3**	**78.2**
6	79.9	74.9

A clear trend emerges across both cohorts as model depth increases from a shallow to a moderately deep configuration. When the depth is expanded from $L = 1$ to $L = 4$, Macro-F1 improves from 80.5% to 81.3% on Sleep-EDF-20 and from 77.8% to 78.2% on Sleep-EDF-78. This progressive gain suggests that sleep-stage classification benefits from a minimum level of temporal modeling capacity, which is afforded by stacking multiple Conformer blocks and enables the encoding of long-range dependencies that are insufficiently captured by shallow architectures.

With further depth expansion, however, the performance trend reverses. At $L = 6$, Macro-F1 decreases to 79.9% on Sleep-EDF-20 and to 74.9% on Sleep-EDF-78, indicating that the additional representational capacity no longer translates into improved generalization. Such degradation is

 doi:10.20944/preprints202601.0930.v1

indicative of over-parameterization under the current data regime and is consistent with increased optimization difficulty and overfitting tendencies in deeper temporal models.

In light of these observations, $L = 4$ is adopted as the default depth throughout this study. This configuration provides sufficient expressive power for modeling long-range temporal structure while maintaining stable optimization behavior and avoiding unnecessary computational overhead.

6. Discussion

These results should not be interpreted as EMG being intrinsically uninformative; rather, its utility is reliability- and cohort-dependent, and the proposed gating is designed to exploit stable muscle-tone cues while suppressing artifact-dominated segments. While the proposed model attains competitive performance under subject-wise evaluation, several limitations should be acknowledged. Several limitations of the present study merit consideration. First, the proposed framework is developed under the assumption of a complete polysomnography (PSG) configuration, incorporating EEG, EOG, and EMG signals. In practical home-monitoring settings, however, data acquisition is often restricted to a single EEG channel, under which conditions the cross-modal fusion mechanism cannot be fully exploited. This constraint motivates future investigations into modality-dropout or modality-agnostic training paradigms, which may enable stable inference when sensor availability is incomplete. Closely related to this issue is the second limitation: the experimental evaluation is confined to recordings from healthy subjects in the Sleep-EDF dataset. Sleep data collected from patients with clinical conditions—such as obstructive sleep apnea or chronic insomnia—are known to exhibit altered time–frequency signatures and more complex artifact characteristics. Consequently, systematic validation on pathology-oriented clinical cohorts will be essential for determining the clinical robustness and generalizability of the proposed approach.

7. Conclusions

In this work, the Gated Cross-Modal and Temporal Fusion Network (G-CMTF Net) is presented as an end-to-end solution for polysomnography-based sleep staging, where scale-dependent feature ambiguity and cross-channel artifact contamination remain persistent obstacles. A multi-scale spectro-temporal frontend is coupled with an adaptive gating mechanism so that physiologically meaningful sleep dynamics can be emphasized while modality-specific noise is attenuated. Along the temporal dimension, a Conformer encoder is employed to represent long-range transition regularities, thereby supporting consistent modeling of sleep-cycle structure.

Experimental results on the Sleep-EDF-20 and Sleep-EDF-78 benchmarks indicate that G-CMTF Net performs competitively relative to state-of-the-art baselines. Of particular importance is the improved sensitivity obtained for N1 and REM, which are known to be prone to confusion under multimodal artifacts; this outcome suggests that reliability-aware fusion contributes to reducing modality-induced ambiguity. Evidence from ablation experiments further shows that explicit spectral injection and multi-scale encoding are necessary to capture the heterogeneous oscillatory characteristics that define sleep stages.

Several practical limitations should also be acknowledged. The current framework assumes the availability of a full PSG montage (EEG, EOG, and EMG), which may not be feasible in home-monitoring deployments. Future studies will therefore explore lightweight variants tailored to single-channel devices, together with self-supervised pre-training strategies, in order to improve generalization across heterogeneous clinical populations.

Author Contributions: Conceptualization, J.Y.; methodology, J.Y. and P.L.; software, J.Y. and P.L.; formal analysis, J.Y.; investigation, J.Y.; data curation, J.Y.; validation, J.Y. and P.L.; visualization, P.L.; writing—original draft preparation, P.L.; writing—review and editing, J.Y. and P.L.; supervision, J.Y. All authors have read and agreed to the published version of the manuscript.

 doi:10.20944/preprints202601.0930.v1

Funding: This work was supported by the National Natural Science Foundation of China (82374561, 82174490), the Zhejiang Provincial Natural Science Foundation of China (LY24H270003), the Key Project of Zhejiang Provincial Administration of Traditional Chinese Medicinethe(GZY-ZJ-KJ-23072), Research Project of Zhejiang Chinese Medical University (2022FSYYZZ07, 2025FSYYZY10).

Data Availability Statement: The data used in this study are publicly available from the PhysioNet repository (Sleep-EDF Expanded Dataset), including the commonly used Sleep-EDF-20 and Sleep-EDF-78 splits, at https://physionet.org/content/sleep-edfx/.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lutfy, R.H.; Ashour, A.M.; Khames, A.; Elhemiely, A.A.; Alam-ElDein, K.M.; Faraag, A.H.I.; Hamed, M.O.; Abdel Daim, Z.J.; Attia, N.I.; Gadelmawla, M.H. Targeting Oxidative Stress and Neuroinflammation: Epigallocatechin-3-gallate-Selenium Nanoparticles Mitigate Sleep Deprivation-Induced Cortical Impairment. *International Journal of Molecular Sciences* **2025**, *26*, 11173.
2. De Longis, E.; Kassis, A.; Rémond-Derbez, N.; Thota, R.; Darimont, C.; Donato-Capel, L.; Hudry, J. Cognitive benefits of sleep: A narrative review to explore the relevance of glucose regulation. *Sleep Advances* **2025**, *6*, zpae095.
3. Khan, M.A.; Al-Jahdali, H. The consequences of sleep deprivation on cognitive performance. *Neurosciences Journal* **2023**, *28*, 91–99.
4. Hyndych, A.; El-Abassi, R.; Mader Jr, E.C. The Role of Sleep and the Effects of Sleep Loss on Cognitive, Affective, and Behavioral Processes. *Cureus* **2025**, *17*.
5. Lee, Y.J.; Lee, J.Y.; Cho, J.H.; Choi, J.H. Interrater reliability of sleep stage scoring: A meta-analysis. *Journal of Clinical Sleep Medicine* **2022**, *18*, 193–202.
6. Chambon, S.; Galtier, M.N.; Arnal, P.J.; Wainrib, G.; Gramfort, A. A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **2018**, *26*, 758–769.
7. Phan, H.; Andreotti, F.; Cooray, N.; Chén, O.Y.; De Vos, M. SeqSleepNet: End-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **2019**, *27*, 400–410.
8. Phan, H.; Chén, O.Y.; Koch, P.; Mertins, A.; De Vos, M. Deep transfer learning for single-channel automatic sleep staging with channel mismatch. In Proceedings of the 2019 27th European signal processing conference (EUSIPCO). IEEE, 2019, pp. 1–5.
9. Chen, C.; Fang, H.; Yang, Y.; Zhou, Y. Model-agnostic meta-learning for EEG-based inter-subject emotion recognition. *Journal of Neural Engineering* **2025**, *22*, 016008.
10. Jia, Z.; Liang, H.; Liu, Y.; Wang, H.; Jiang, T. Distillsleepnet: Heterogeneous multi-level knowledge distillation via teacher assistant for sleep staging. *IEEE Transactions on Big Data* **2024**.
11. Eldele, E.; Chen, Z.; Liu, C.; Wu, M.; Kwoh, C.K.; Li, X.; Guan, C. An attention-based deep learning approach for sleep stage classification with single-channel EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **2021**, *29*, 809–818.
12. Phan, H.; Mikkelsen, K.; Chén, O.Y.; Koch, P.; Mertins, A.; De Vos, M. Sleeptransformer: Automatic sleep staging with interpretability and uncertainty quantification. *IEEE Transactions on Biomedical Engineering* **2022**, *69*, 2456–2467.
13. Zhang, J.; Xue, Y.; Li, Y. A Novel CNN plus Transformer Network for EEG-based Automatic Sleep Staging. In Proceedings of the 2023 4th International Conference on Computer, Big Data and Artificial Intelligence (ICCBD+ AI). IEEE, 2023, pp. 58–65.
14. Kemp, B.; Zwinderman, A.H.; Tuk, B.; Kamphuisen, H.A.; Oberye, J.J. Analysis of a sleep-dependent neuronal feedback loop: The slow-wave microcontinuity of the EEG. *IEEE Transactions on Biomedical Engineering* **2000**, *47*, 1185–1194.
15. Goldberger, A.L.; Amaral, L.A.; Glass, L.; Hausdorff, J.M.; Ivanov, P.C.; Mark, R.G.; Mietus, J.E.; Moody, G.B.; Peng, C.K.; Stanley, H.E. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *circulation* **2000**, *101*, e215–e220.
16. Supratak, A.; Dong, H.; Wu, C.; Guo, Y. DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG. *IEEE transactions on neural systems and rehabilitation engineering* **2017**, *25*, 1998–2008.

 doi:10.20944/preprints202601.0930.v1

17. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Advances in neural information processing systems* **2017**, *30*.
18. Perslev, M.; Jensen, M.; Darkner, S.; Jennum, P.J.; Igel, C. U-time: A fully convolutional network for time series segmentation applied to sleep staging. *Advances in neural information processing systems* **2019**, *32*.
19. Supratak, A.; Guo, Y. TinySleepNet: An efficient deep learning model for sleep stage scoring based on raw single-channel EEG. In Proceedings of the 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE, 2020, pp. 641–644.
20. Phyo, J.; Ko, W.; Jeon, E.; Suk, H.I. TransSleep: Transitioning-aware attention-based deep neural network for sleep staging. *IEEE Transactions on Cybernetics* **2022**, *53*, 4500–4510.
21. Song, Y.; Zheng, Q.; Liu, B.; Gao, X. EEG conformer: Convolutional transformer for EEG decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **2022**, *31*, 710–719.
22. Phan, H.; Chén, O.Y.; Tran, M.C.; Koch, P.; Mertins, A.; De Vos, M. XSleepNet: Multi-view sequential model for automatic sleep staging. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2021**, *44*, 5903–5915.
23. Jia, Z.; Lin, Y.; Wang, J.; Wang, X.; Xie, P.; Zhang, Y. SalientSleepNet: Multimodal salient wave detection network for sleep staging. *arXiv preprint arXiv:2105.13864* **2021**, [2105.13864].
24. Leach, S.; Krugliakova, E.; Sousouri, G.; Snipes, S.; Skorucak, J.; Schühle, S.; Müller, M.; Ferster, M.L.; Da Poian, G.; Karlen, W.; et al. Acoustically evoked K-complexes together with sleep spindles boost verbal declarative memory consolidation in healthy adults. *Scientific Reports* **2024**, *14*, 19184.
25. Sekaran, S.R.; Pang, Y.H.; Ling, G.F.; Yin, O.S. MSTCN: A multiscale temporal convolutional network for user independent human activity recognition. *F1000Research* **2022**, *10*, 1261.
26. Yao, Z.; Liu, X. A cnn-transformer deep learning model for real-time sleep stage classification in an energy-constrained wireless device. In Proceedings of the 2023 11th international IEEE/EMBS conference on neural engineering (NER). IEEE, 2023, pp. 1–4.
27. Liu, P.; Qian, W.; Zhang, H.; Zhu, Y.; Hong, Q.; Li, Q.; Yao, Y. Automatic sleep stage classification using deep learning: Signals, data representation, and neural networks. *Artificial Intelligence Review* **2024**, *57*, 301.
28. Yu, T.; Hu, X.; He, Y.; Wu, W.; Gu, Z.; Yu, Z.; Li, Y.; Wang, F.; Xiao, J. Multi-View Self-Supervised Learning Enhances Automatic Sleep Staging from EEG Signals. *IEEE Transactions on Biomedical Engineering* **2025**.
29. Jia, Z.; Lin, Y.; Wang, J.; Ning, X.; He, Y.; Zhou, R.; Zhou, Y.; Lehman, L.w.H. Multi-view spatial-temporal graph convolutional networks with domain generalization for sleep stage classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **2021**, *29*, 1977–1986.
30. Mao, Y.; Ma, X.; Kuang, H.; Liu, X. MVFSleepNet: Multi-View Fusion Network for Automatic Sleep Staging. In Proceedings of the 2024 IEEE 7th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC). IEEE, 2024, Vol. 7, pp. 39–43.
31. Wang, Z.; Gong, Z.; Wang, T.; Dong, Q.; Huang, Z.; Zhang, S.; Ma, Y. MASleepNet: A Sleep Staging Model Integrating Multi-Scale Convolution and Attention Mechanisms. *Biomimetics* **2025**, *10*, 642.
32. Yubo, Z.; Yingying, L.; Bing, Z.; Lin, Z.; Lei, L. MMASleepNet: A multimodal attention network based on electrophysiological signals for automatic sleep staging. *Frontiers in Neuroscience* **2022**, *16*, 973761.
33. American Academy of Sleep Medicine. *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications*, version 2.6 ed.; American Academy of Sleep Medicine: Darien, IL, USA, 2020.
34. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7132–7141.
35. Gulati, A.; Qin, J.; Chiu, C.C.; Parmar, N.; Zhang, Y.; Yu, J.; Han, W.; Wang, S.; Zhang, Z.; Wu, Y.; et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv arXiv:2005.08100* **2020**.
36. Liu, Z.; Mao, H.; Wu, C.Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A convnet for the 2020s. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 11976–11986.
37. Mousavi, S.; Afghah, F.; Acharya, U.R. SleepEEGNet: Automated sleep stage scoring with sequence to sequence deep learning approach. *PloS one* **2019**, *14*, e0216456.
38. Fiorillo, L.; Favaro, P.; Faraci, F.D. Deepsleepnet-lite: A simplified automatic sleep stage scoring model with uncertainty estimates. *IEEE transactions on neural systems and rehabilitation engineering* **2021**, *29*, 2076–2085.