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.

DGCA-Net: Dynamic Gated Calibration and Contrastive Alignment Network for Automatic Sleep Staging based on EEG and EOG

Xuwen Liu
Jiangnan University
Jinlong Yang
jlyang@jiangnan.edu.cn

Jiangnan University
Junhua Zhou
Affiliated Hospital of Jiangnan University

Research Article

Keywords:

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11654662/v1

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

Additional Declarations: No competing interests reported.

DGCA-Net: Dynamic Gated Calibration and Contrastive Alignment Network for Automatic Sleep Staging based on EEG and EOG

Xuwen Liu[1], Jinlong Yang[1*], Junhua Zhou[2]

[1*]School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, 214122, China.
[2]Department of Medical Engineering, Affiliated Hospital of Jiangnan University, Wuxi, 214122, China.

*Corresponding author(s). E-mail(s): jlyang@jiangnan.edu.cn;

Abstract

Automatic sleep stage classification has evolved from single-modal to multi-modal paradigms, incorporating diverse signal sources such as EEG and EOG. However, effectively fusing multi-modal sleep data remains challenging. The channel-wise importance imbalance and noise redundancy across modalities not only introduce severe interference into cross-attention fusion, but also lead to cross-modal feature space misalignment, where inconsistent distributions impede the effective capture of complementary information. To address these issues, we propose DGCA-Net, a novel deep learning framework for multi-modal sleep staging. DGCA-Net comprises two core modules. Firstly, we introduce Adaptive Multi-modal Cross-Attention Fusion (AMCAF), which employs global feature compression and dynamic channel gating to adaptively recalibrate the channel-wise importance of multi-scale-multi-modal features. This module enhances discriminative features highly correlated with the current sleep stage while suppressing noisy, low-contribution channels. The recalibrated features are then fused through cross-attention, producing cleaner and more distinguishable representations. The second core module is a contrastive alignment module driven by an InfoNCE loss, which explicitly pulls the embeddings of EEG and EOG from the same sleep epoch together in the latent space. This alignment mitigates cross-modal distribution discrepancies, ensuring more consistent and complementary features for the attention-based fusion stage and substantially enhancing the model's robustness to noise. We evaluate DGCA-Net on the public Sleep-EDF-20 and Sleep-EDF-78 datasets. Experimental results demonstrate that the proposed

method outperforms existing state-of-the-art approaches on both benchmarks, achieving superior classification accuracy and consistent performance.

Keywords: Deep Learning, Auto sleep stage classification, Multi-modal fusion, Contrastive alignment, Dynamic channel gating

1 Introduction

Within this context, sleep staging—the systematic categorization of sleep into wakefulness (W), rapid eye movement (REM), and non-rapid eye movement substages (N1, N2, N3)—serves as an indispensable tool for assessing sleep architecture, diagnosing sleep-related pathologies, tracking disease progression, and tailoring individualized therapeutic regimens [1, 2]. Given that the precision and timeliness of staging directly govern the quality of clinical decision-making, the development of reliable automated sleep staging methodologies has emerged as a research priority with substantial translational potential.

Clinically, sleep staging has long been performed by trained experts who visually inspect overnight polysomnography (PSG) recordings—a comprehensive acquisition system that simultaneously captures electroencephalogram (EEG), electrooculogram (EOG), electromyogram (EMG), and electrocardiogram (ECG) signals [3]. The standard protocol requires specialists to partition the continuous PSG data into consecutive 30-second epochs and assign each epoch to one of the five sleep stages based on characteristic waveform patterns, such as high-amplitude delta oscillations for N3, rapid conjugate eye movements for REM, and low-amplitude mixed-frequency activity for wakefulness [4]. Although this manual scoring approach remains the clinical benchmark, it suffers from inherent drawbacks: substantial inter-rater variability arises from subjective interpretation, and the laborious scoring process—often consuming several hours per record—poses significant barriers to high-throughput research and real-time clinical workflows [4]. These limitations have motivated extensive efforts toward computational automation of the staging procedure [5].

The initial wave of automated solutions drew upon conventional machine learning techniques, which generally operate under a two-stage pipeline comprising hand-crafted feature design followed by supervised classification [6]. Drawing upon neurophysiological domain knowledge, researchers engineered a variety of feature sets from EEG and EOG channels, including statistical moments in the time domain, power spectral densities within delta and theta frequency bands, and time-frequency representations derived from wavelet decomposition [7]. These features were subsequently fed into classifiers such as support vector machines, random forests, or naive Bayes models [8]. Despite their conceptual simplicity, these approaches are constrained by several fundamental weaknesses: classification performance remains acutely sensitive to the quality and completeness of manually designed features, and the fixed feature representations often fail to generalize across diverse patient populations, age groups, or recording devices, thereby limiting their practical deployment.

The advent of deep learning has fundamentally reshaped the landscape of sleep staging by enabling end-to-end feature learning directly from raw or minimally preprocessed signals, thereby circumventing the bottleneck of manual feature engineering. Pioneering works such as DeepSleepNet demonstrated the feasibility of single-channel EEG staging through a hybrid architecture combining convolutional neural networks for local pattern extraction and bidirectional long short-term memory networks for capturing sequential transition regularities across consecutive epochs [9]. Subsequently, multi-modal frameworks like SSNet extended this paradigm by processing EEG and EOG streams in parallel, capitalizing on their complementary physiological perspectives to boost classification accuracy [10]. Nevertheless, despite these advances, practical deployment of multi-channel staging systems continues to encounter persistent obstacles, including motion-induced artifacts, inter-channel quality fluctuations, and substantial cross-subject physiological variability, all of which compromise system robustness [11–13].

In light of the aforementioned challenges, we identify two particularly critical yet insufficiently addressed bottlenecks that hinder effective multi-modal sleep data integration. The first concerns the prevalent channel-wise importance imbalance and noise contamination inherent in multi-scale, multi-modal features extracted from heterogeneous physiological signals. Specifically, low-informative channels disproportionately introduce confounding interference into subsequent cross-attention fusion mechanisms, while genuinely discriminative patterns pertinent to specific sleep stages become obscured [14]. The second bottleneck stems from cross-modal feature space misalignment, wherein the distinct distribution characteristics of EEG and EOG embeddings impede the seamless integration of complementary information, thereby curtailing the overall staging accuracy and resilience [15]. These dilemmas are exacerbated in suboptimal recording environments where electrode displacement, myogenic artifacts, and inter-channel crosstalk frequently occur [16]. Regrettably, the predominant existing paradigms feed multi-modal features directly into cross-attention modules without adequate pre-filtering or distribution harmonization, thus overlooking the imperative of pre-fusion calibration and cross-modal distribution unification [17].

To tackle these intertwined obstacles, we introduce DGCA-Net (Dynamic Gated Calibration and Contrastive Alignment Network for Sleep Staging), a purpose-built deep architecture tailored for multi-channel PSG-based sleep staging. The proposed framework comprises two core modules and an auxiliary module. The first core module is the Adaptive Multi-modal Cross-Attention Fusion (AMCAF) module, which performs global feature compression and dynamic channel-wise gating to adaptively reweight multi-scale, multi-modal feature channels. This mechanism elevates sleep-stage-relevant discriminative signatures while actively suppressing noisy, low-contribution channels, thereby furnishing subsequent cross-attention stages with refined, high-fidelity feature representations. The second core module is a contrastive alignment module instantiated through the InfoNCE objective, which explicitly constrains the latent embeddings of EEG and EOG signals originating from the identical sleep epoch to converge toward a shared cluster. Such alignment effectively reconciles cross-modal distribution discrepancies, supplying the multi-scale attention mechanism with more homogeneous and mutually complementary inputs, and markedly enhancing

noise resilience. The auxiliary module is the Inter-layer Feature Fusion (IFF) module, which first applies dual average and maximum pooling to each feature layer to distill the most salient representations. The two pooled outputs are then fused via element-wise summation, and the resulting aggregated feature passes through two successive 1×1 convolutions (each followed by a ReLU activation) to produce attention weights that capture cross-scale dependencies. A sigmoid function subsequently generates inter-layer attention weights, which are multiplied with the original multi-scale features to yield the final refined output.

Our main contributions are summarized as follows:

(1) We propose the Adaptive Multi-modal Cross-Attention Fusion (AMCAF) module, which introduces global feature compression and dynamic channel gating into multi-modal sleep staging. By adaptively recalibrating the channel weights of multi-scale and multi-modal features prior to cross-attention, AMCAF effectively suppresses modality-irrelevant noise and redundant channels while enhancing discriminative features strongly correlated with specific sleep stages, providing a purified and distinguishable feature foundation for subsequent cross-modal interaction.

(2) We introduce a contrastive alignment module driven by the InfoNCE loss into the sleep staging framework, which explicitly aligns the embeddings of EEG and EOG signals from the same sleep epoch in the latent space. This alignment resolves the inherent cross-modal feature space misalignment, ensuring that multi-scale attention fusion receives more consistent and complementary features, thereby significantly enhancing the model's robustness in noisy and modality-missing scenarios.

(3) We design an auxiliary Inter-layer Feature Fusion (IFF) module for multi-scale feature post-processing. Through dual pooling operations and attention-guided aggregation, the IFF module captures cross-scale dependencies and mitigates inter-layer redundant and noisy representations, generating high-quality refined features to facilitate accurate sleep stage classification.

2 Related Works

2.1 Deep Learning-Based Automatic Sleep Staging

Deep learning has substantially advanced automatic sleep staging by enabling models to autonomously extract hierarchical representations from raw physiological signals. Early solutions primarily employed Convolutional Neural Networks to capture local time-frequency patterns indicative of specific sleep stages [9]. Subsequent approaches integrated recurrent architectures to model the sequential evolution of sleep architecture across consecutive epochs [9]. More recently, the Transformer paradigm has been adopted to capture global contextual relationships within sleep signals, leveraging self-attention to model long-range dependencies across entire recording sessions [18].

Despite these architectural innovations, current methods encounter persistent challenges in effectively integrating multi-scale and multi-modal features. Features extracted from EEG and EOG signals typically contain substantial modality-specific noise and redundant information, which compromises subsequent fusion operations [15]. Meanwhile, distributional discrepancies among different modalities lead to feature

space misalignment, impeding the effective exploitation of cross-modal complementarity [19]. To address these issues, our framework incorporates a dynamic gated calibration mechanism that suppresses noise and reweights informative channels prior to cross-modal interaction, combined with a contrastive alignment strategy that enforces distributional consistency across modalities, thereby establishing a more robust foundation for multi-scale sleep staging [20].

2.2 Contrastive Learning

Contrastive learning has emerged as a highly effective paradigm for self-supervised representation learning, structuring the feature space so that semantically related samples are pulled together while unrelated ones are pushed apart. This is typically achieved through a contrastive loss that encourages embeddings of positive pairs—derived from different augmentations of the same source—to exhibit high similarity, while repelling negative pairs from distinct sources.

The InfoNCE loss has become the predominant objective function in this domain, as it maximizes a lower bound on mutual information between positive pairs by pulling their representations close, while simultaneously pushing representations of negative pairs away [20]. Beyond its success in computer vision, contrastive learning has been extended to multimodal scenarios, aligning representations from heterogeneous modalities by enforcing that corresponding multimodal pairs occupy proximate locations in a shared embedding space [21]. Motivated by these principles, we integrate a contrastive alignment component into our sleep staging framework to bridge the distribution gap across EEG and EOG modalities. This facilitates more consistent multimodal fusion and enhances robustness against cross-modal variability, with full technical details provided in the methodology section.

2.3 Multi-Head Self-Attention

The multi-head self-attention mechanism serves as a fundamental building block for modeling sequential data, owing to its capacity to capture intricate dependencies across distant temporal positions. Unlike recurrent architectures that process sequences step by step, self-attention computes pairwise affinities between all positions simultaneously, providing direct access to long-range contextual information without the bottleneck of sequential propagation. The multi-head extension further enhances this capability by allowing the model to attend to different representation subspaces in parallel, each capturing distinct types of relational patterns.

In sleep staging, this mechanism offers particular advantages, as sleep stage transitions follow stereotyped sequences—such as progressive deepening from light sleep to deep sleep and cyclic alternation with REM periods—and accurate classification relies on integrating information across multiple temporal scales [18]. In our framework, we leverage multi-head attention to model cross-modal dependencies at consistent hierarchical scales, enabling the effective integration of complementary information from EEG and EOG. This attention-based design works in concert with the preceding calibration stage, which emphasizes the most informative features and suppresses modality-specific redundancy, thereby strengthening the overall cross-modal interaction.

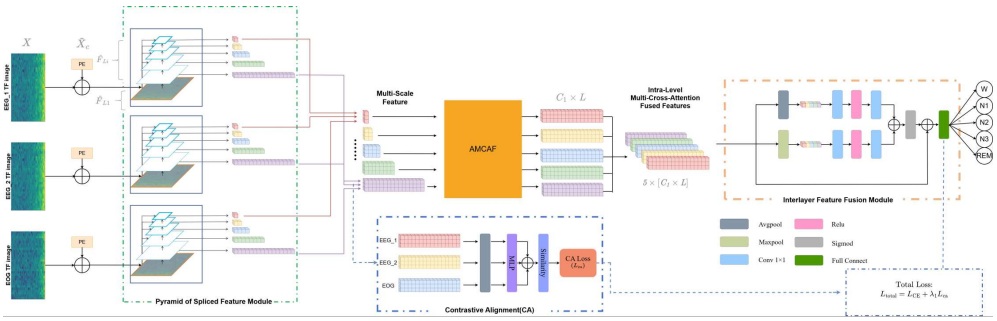

Fig. 1: The overall architecture of DGCA-Net. (1) Multi-scale features are extracted from time-frequency representations of EEG and EOG signals via a multi-resolution feature pyramid with cross-level concatenation. (2) The Adaptive Multi-modal Cross-Attention Fusion (AMCAF) module adaptively recalibrates channel weights through global feature compression and dynamic gating to suppress noise and enhance discriminative features. (3) The Contrastive Alignment module, driven by an InfoNCE loss, aligns EEG and EOG embeddings from the same sleep epoch to resolve cross-modal feature space misalignment. (4) The Inter-layer Feature Fusion (IFF) module captures cross-level dependencies through pooling and attention-guided aggregation to refine the final feature representation.

3 Methodology

The overall architecture of DGCA-Net is illustrated in Fig. 1, which consists of a pyramid-based multi-scale feature extraction backbone, two core modules (AMCAF and Contrastive Alignment), one auxiliary module (IFF), and a classification head. We will delineate the specific structure and functionality of each module in the following subsections.

3.1 pyramid-based multi-scale feature extraction

Sleep signals, originally acquired as one-dimensional time series, can be transformed into two-dimensional time-frequency representations via the Short-Time Fourier Transform (STFT) followed by logarithmic scaling, thereby revealing substantially richer spectral and temporal information. Consequently, thorough feature extraction from these time-frequency maps plays a critical role in sleep staging tasks. This feature extraction backbone is built upon a multi-resolution feature pyramid, wherein each pyramidal level generates its own feature output. To preserve complete source information, the input of each subsequent level is obtained by concatenating the output from its preceding level with the original time-frequency map.

The proposed feature pyramid is specifically designed to accommodate the distinctive spectral properties of sleep electrophysiological signals, such as EEG and EOG. All pyramidal levels share a uniform convolutional configuration—including identical kernel sizes—while scale-specific differentiation is achieved by varying the spatial resolution of the feature maps, tailored to the frequency characteristics inherent to sleep

rhythms. Specifically, shallow levels (closer to the input) operate on higher-resolution feature maps to capture fine-grained, high-frequency patterns, such as sleep spindles in the 12–16 Hz range. In contrast, deeper levels progressively reduce the spatial resolution, yielding larger receptive fields that emphasize coarse-grained, low-frequency components, exemplified by delta waves spanning 0.5–4 Hz. This shared architectural backbone, combined with resolution tuning across scales, enables principled multi-scale feature extraction that is fundamentally aligned with the spectral properties of sleep signals.

The input X can be computed at pyramid-based multi-scale feature extraction as follows:

$$\tilde{X}_c = X + PE \tag{1}$$

$$\hat{F}_{L1} = ReLU\left(Norm\left(Conv(\tilde{X}_c)\right)\right) \tag{2}$$

$$\hat{F}_{Li} = ReLU\left(Norm\left(Conv(\tilde{X}_c + \hat{F}_{Li-1})\right)\right) \tag{3}$$

where $c \in \{EEG, EOG\}$ denotes the signal modality, $Norm(\cdot)$ represents the normalization operation, $Conv(\cdot)$ refers to the convolution operation, and $ReLU(\cdot)$ denotes the ReLU activation function.

The multi-scale features $\hat{F}_{Li}$ extracted from each modality are subsequently aggregated according to their corresponding pyramid levels, with features from identical levels being merged together. As a result, a set of level-specific feature matrices $F_{Li} = (F_{L1}, F_{L2}, F_{L3}, F_{L4}, F_{L5})$ across five distinct scales is obtained.

3.2 Adaptive Multi-modal Cross-Attention Fusion (AMCAF)

In the process of multi-scale feature extraction, it is inevitable that there exist problems such as channel-wise importance imbalance and noise redundancy across different modalities, where low-contribution channels introduce severe interference while discriminative features highly correlated with the current sleep stage are overshadowed.

To address this issue, we propose the Adaptive Multi-modal Cross-Attention Fusion (AMCAF) module, which performs dynamic channel gating prior to cross-modal interaction to suppress modality-irrelevant noise and recalibrate feature weights. The architecture of the AMCAF module is illustrated in Fig. 2. The AMCAF receives the multi-scale features $F_{L_i}(i = 1, 2, \ldots, 5)$ extracted from the pyramid-based feature extraction module as inputs.

First, a squeeze operation is applied to each level of the input features. Specifically, global average pooling is performed on each feature map F_{L_i} to aggregate spatial information, yielding a channel-wise descriptor that encapsulates global contextual information:

$$s_i = \frac{1}{C \times T} \sum_{c=1}^{C} \sum_{t=1}^{T} F_{L_i}(c, t) \in \mathbb{R}^{1 \times 1} \tag{4}$$

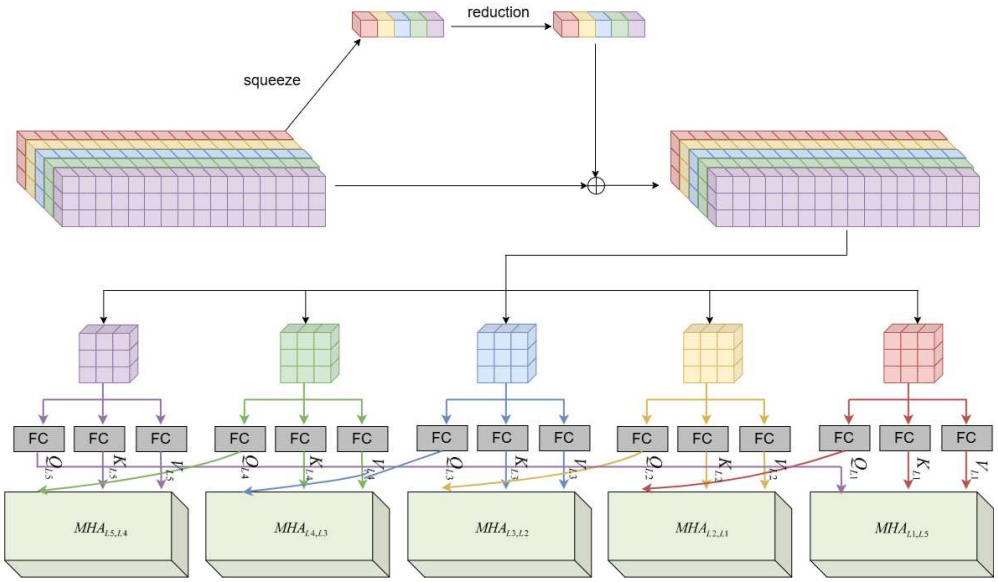

Fig. 2: Structure of Adaptive Multi-modal Cross-Attention Fusion.

where C denotes the number of channels, and T denotes the temporal dimension.

Subsequently, a reduction operation is performed through two successive 1×1 convolutional layers interleaved with ReLU activation, which captures channel-wise dependencies across all pyramid levels and generates adaptive recalibration weights:

$$e = \mathrm{ReLU}\left(\mathrm{Conv2}\left(\mathrm{ReLU}\left(\mathrm{Conv1}(S)\right)\right)\right) \in \mathbb{R}^{N\times 1\times 1} \tag{5}$$

where N denotes the number of pyramid levels (set to 5 in our experiments), and $S = \{s_1, s_2, \ldots, s_N\}$ denotes the concatenated descriptors from all levels.

These learned weights e_n for each level are then applied to the corresponding original feature maps through element-wise multiplication, and a residual connection is introduced to preserve the original information:

$$\tilde{F}_{L_i} = F_{L_i} \cdot e_i + F_{L_i} \in \mathbb{R}^{C\times T} \tag{6}$$

Through this adaptive recalibration mechanism, the module enhances discriminative features strongly correlated with specific sleep stages while suppressing low-contribution noise channels, providing purified feature representations $\tilde{F}_{L_i}$ for subsequent cross-attention fusion.

After recalibration, the refined features are projected onto the Q, K, and V matrices as follows:

$$Q_{L_i} = \tilde{F}_{L_i} \cdot W_Q, \tag{7}$$

$$K_{L_i} = \tilde{F}_{L_i} \cdot W_K, \tag{8}$$

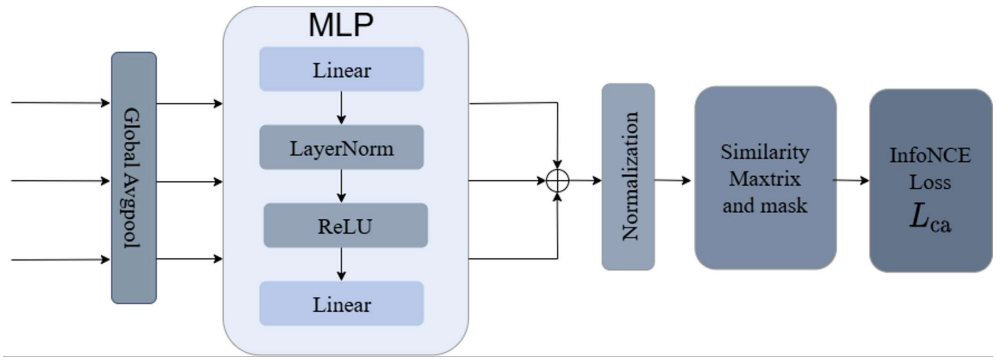

Fig. 3: Structure of Contrastive Alignment

$$V_{L_i} = \tilde{F}_{L_i} \cdot W_V \tag{9}$$

where W_Q, W_K, W_V are learnable matrices. The Q matrices of projections in each layer are then exchanged in a staggered manner. By leveraging the influence of adjacent layers on the current layer, AMCAF focus more on the learning of its own layer:

$$\tilde{Q}_{L_j} = Q_{L_i} \quad (i = 1, 2, 3, 4, 5;\ j = 5, 1, 2, 3, 4) \tag{10}$$

The cross-attention fusion processes are formulated as follows:

$$\widehat{F}_{L_i, L_j} = \left(\frac{\tilde{Q}_{L_j} K_{L_i}^{\top}}{\sqrt{d}} \right) V_{L_i} \in \mathbb{R}^{C \times L} \tag{11}$$

where $(i, j) \in (1, 5), (2, 1), (3, 2), (4, 3), (5, 4)$. For example, $(i, j) = (1, 5)$indicates that the input K-matrix and V-matrix are from layer 1, while the Q-matrix is from layer 5. By inputting the Q matrix in a staggered hierarchy based on the attention mechanism formula, the module enables the absorption of effective information from adjacent layers while avoiding redundant information, thereby focusing more on the learning of effective information in the current layer.

Finally, the output results of the five layers are stacked to obtain F_{stack}, which serves as the input for the subsequent Inter-layer Feature Fusion (IFF) module.

3.3 Contrastive Alignment (CA)

To address the cross-modal feature space misalignment caused by inconsistent distributions between EEG and EOG modalities, we introduce a Contrastive Alignment module driven by the InfoNCE loss. The architecture of this module is illustrated in Fig. 3.

The CA module takes as input the features from the first layer of the pyramid-based multi-scale feature extraction module, denoted as $F_{L_1} \in \mathbb{R}^{C \times L}$, which captures the most fine-grained temporal patterns from the original time-frequency representations of each modality. By leveraging the shallowest pyramid level, the alignment process

operates on features that retain rich local information before being transformed by higher-level abstraction.

For each modality $m \in \text{EEG}, \text{EOG}$, the corresponding feature $F_{L_1}^m$ is first projected into a shared embedding space through a lightweight Multi-Layer Perceptron (MLP). The MLP consists of global average pooling to aggregate temporal information, followed by two linear layers interleaved with LayerNorm and ReLU activation, producing modality-specific embeddings $Z_m \in \mathbb{R}^{B \times d}$, where B is the batch size and d is the embedding dimension. Subsequently, a similarity matrix $\mathcal{S} \in \mathbb{R}^{2B \times 2B}$ is computed as the dot product of the normalized embeddings from both modalities:

$$S_{i,k} = \frac{Z_i \cdot Z_k^\top}{\tau} \tag{12}$$

where τ is a temperature hyperparameter that scales the similarity distribution. For each anchor sample i , its positive pair y_i is defined as the embedding from the same sleep epoch but from the other modality, while all other samples in the batch are treated as negatives. The InfoNCE loss is then formulated as:

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{2B} \sum_{i=1}^{2B} \log \left(\frac{\exp\left(S_{i,y_i}\right)}{\sum_{k=1}^{2B} \exp\left(S_{i,k}\right)} \right) \tag{13}$$

By minimizing this loss, this CA module explicitly forces embeddings from the same sleep epoch across EEG and EOG modalities to converge to nearby locations in the latent space, while embeddings from distinct epochs are pushed apart. This alignment effectively reconciles cross-modal distribution discrepancies, ensuring that subsequent multi-scale attention mechanisms receive more homogeneous and complementary inputs, thereby enhancing the model's robustness against noise and inter-modal variability.

3.4 Interlayer Feature Fusion (IFF)

In the multi-scale feature fusion stage, direct weighted fusion of features from different layers inevitably retains noise and redundant information, which compromises the discriminative power of the final representation. To address this issue, we propose the Inter-layer Feature Fusion (IFF) module, which employs a denoising strategy prior to fusion to effectively enhance feature quality. The architecture of the IFF module is illustrated in Fig. 4. The IFF receives the stacked output F_{stack} from the AMCAF module as inputs.

For each fusion result layer $\widehat{F}_{L_i,L_j}$ in F_{stack}, the module simultaneously performs global average pooling and global maximum pooling to capture complementary salient information. The average pooling emphasizes the overall activation patterns across the temporal dimension, while the maximum pooling highlights the most prominent responses:

$$\text{Avg}_{L_i} = \text{AvgPool}\left(\widehat{F}_{L_i,L_j}\right) \in \mathbb{R}^{C \times 1}, \tag{14}$$

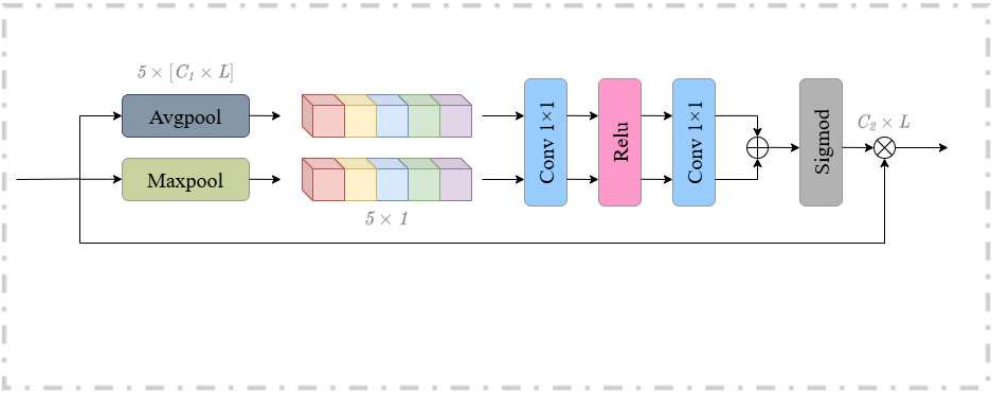

Fig. 4: Structure of Interlayer Feature Fusion

$$\mathrm{Max}_{L_i} = \mathrm{MaxPool}\left(\widehat{F}_{L_i,L_j}\right) \in \mathbb{R}^{C\times 1} \tag{15}$$

where *AvgPool* and *MaxPool* denote the average and maximum pooling operations, respectively, C denotes the number of channels, and L denotes the temporal dimension.

Subsequently, a channel-wise squeeze operation is applied to the pooled results of each layer to obtain layer-level descriptors:

$$v_{\mathrm{avg},i} = \frac{1}{C}\sum_{c=1}^{C} \mathrm{Avg}_{L_i}[c,:] \in \mathbb{R}^{1\times 1}, \tag{16}$$

$$v_{\mathrm{max},i} = \frac{1}{C}\sum_{c=1}^{C} \mathrm{Max}_{L_i}[c,:] \in \mathbb{R}^{1\times 1} \tag{17}$$

The descriptors from all layers are then concatenated to form two types of salient feature maps:

$$F_{\mathrm{avg}} = \mathrm{Concat}\big(v_{\mathrm{avg},1}, v_{\mathrm{avg},2}, \ldots, v_{\mathrm{avg},5}\big), \tag{18}$$

$$F_{\mathrm{max}} = \mathrm{Concat}\big(v_{\mathrm{max},1}, v_{\mathrm{max},2}, \ldots, v_{\mathrm{max},5}\big) \tag{19}$$

To capture cross-scale dependencies and suppress noise, attention maps are generated for both salient feature maps through two successive 1×1 convolutional layers interleaved with ReLU activation. These attention maps act as adaptive filters, learning to assign larger weights to informative layers while downweighting less contributive ones:

$$F_{\mathrm{avg}}^{\mathrm{attn}} = \mathrm{Conv1}\big(\mathrm{ReLU}\big(\mathrm{Conv1}(F_{\mathrm{avg}})\big)\big), \tag{20}$$

$$F_{\mathrm{max}}^{\mathrm{attn}} = \mathrm{Conv1}\big(\mathrm{ReLU}\big(\mathrm{Conv1}(F_{\mathrm{max}})\big)\big) \tag{21}$$

The two attention branches are then aggregated via element-wise addition and passed through a Sigmoid function to produce the final inter-layer attention weights. These weights are subsequently applied to the original multi-scale features through element-wise multiplication, and the weighted results are concatenated to produce the fused output:

$$F_{\text{fused}} = \text{Concat}\left(\widehat{F}_{L_i,L_j} \cdot \text{Sigmoid}\big(F_{\text{avg}}^{\text{attn}} + F_{\text{max}}^{\text{attn}}\big)\right) \in \mathbb{R}^{5C\times L} \tag{22}$$

Through this design, the IFF module effectively reinforces informative layers while suppressing redundant or noise-contaminated ones, ensuring that the final feature representation retains discriminative multi-scale sleep patterns and provides high-quality inputs for the subsequent classification stage.

3.5 Classification

Following the feature refinement performed by the IFF module, the fused feature representation $F_{\text{fused}} \in \mathbb{R}^{5C\times L}$ is fed into a two-layer fully connected classifier to produce the final sleep stage predictions. The first layer projects the high-dimensional fused features into a compact hidden space, enabling the model to abstract and reorganize the discriminative patterns learned from previous modules:

$$F_{\text{hidden}} = X_{\text{fused}} \cdot W_1 + b_1 \tag{23}$$

where $X_{\text{fused}} \in \mathbb{R}^{B\times(5C\times L)}$ is obtained by flattening the output of the IFF module, $W_1 \in \mathbb{R}^{(5C\times L)\times H}$ and $b_1 \in \mathbb{R}^{1\times H}$ denote the learnable weight matrix and bias term of the first linear layer, with H representing the hidden dimension. Subsequently, the hidden representation is mapped to the five sleep stage categories through a second linear transformation:

$$F = F_{\text{hidden}} \cdot W_2 + b_2 \tag{24}$$

where $W_2 \in \mathbb{R}^{H\times 5}$ and $b_2 \in \mathbb{R}^{1\times 5}$ are the weight matrix and bias term for the output layer. The resulting logits F are then passed through a softmax function to generate class probabilities for each sleep stage. This two-stage design, with an intermediate hidden layer prior to the final classification, serves to decouple the decision process from the preceding feature extraction modules, thereby enhancing the model's generalization capability and mitigating overfitting to task-specific feature distributions.

4 EXPERIMENTAL RESULTS

4.1 Dataset

To assess the performance of our proposed DGCA-Net, we conducted experiments on the publicly available Sleep-EDF-20 dataset. This dataset comprises overnight

Table I: Confusion Matrix of DGCA-Net on the SleepEDF-20 Dataset

True	Predicted Label					Per-class Metrics		
Label	W	N1	N2	N3	REM	PR	RE	F1
W	**8583**	330	66	13	126	92.2	94.1	93.2
N1	418	**1399**	462	4	521	56.8	49.9	53.1
N2	143	445	**16149**	250	812	90.1	90.7	90.4
N3	41	2	888	**4764**	8	94.7	83.5	88.8
REM	121	289	357	1	**6949**	82.6	90.0	86.1

Table II: Confusion Matrix of DGCA-Net on the SleepEDF-78 Dataset

True	Predicted Label					Per-class Metrics		
Label	W	N1	N2	N3	REM	PR	RE	F1
W	**63502**	2509	398	34	379	93.8	95.0	94.4
N1	3340	**10004**	6040	14	2124	56.8	46.5	51.1
N2	363	3550	**62170**	564	2485	83.3	89.9	86.5
N3	34	14	4253	**8711**	27	93.4	66.8	77.9
REM	436	1521	1785	3	**22090**	81.5	85.5	83.5

polysomnography (PSG) recordings along with their corresponding hypnogram annotations. The PSG recordings capture multi-channel physiological signals, including electroencephalogram (EEG) signals from the Fpz-Cz and Pz-Oz electrode placements, horizontal electrooculogram (EOG), and submental electromyogram (EMG). The hypnogram files contain expert-annotated sleep stage labels for each 30-second epoch, covering wakefulness (W), rapid eye movement (REM), and non-rapid eye movement stages (N1, N2, N3). In our experiments, we utilized the EEG (Fpz-Cz and Pz-Oz) and EOG signals as input modalities, following the common practice in prior sleep staging studies.

In addition to the Sleep-EDF-20 dataset, we also evaluate our method on the Sleep-EDF-78 dataset to assess generalization performance across a larger and more heterogeneous cohort. The Sleep-EDF-78 dataset contains 78 overnight PSG recordings from 78 subjects (age range: 25–101 years), with the same EEG (Fpz-Cz and Pz-Oz), EOG, and EMG channels as Sleep-EDF-20. Following the standard preprocessing protocol, we partition the continuous recordings into 30-second epochs and adopt the identical train-test split strategy as prior works to ensure fair comparison.

4.2 Parameters

The batch size was set to 128, and the AdamW optimizer was adopted for model training with a learning rate of 1×10^{-5}. The cross-entropy loss function was utilized to compute the training loss. For the time-frequency transformation, the frequency parameter F was set to 128, and the time parameter T was configured to 29. The dropout rate was fixed at 0.5. All models were implemented using PyTorch 2.7.0 and trained on an NVIDIA GeForce RTX 5070 Ti Laptop GPU.

To evaluate the generalization capability of DGCA-Net, we employed subject-wise 20-fold cross-validation on the Sleep-EDF-20 dataset, where data from different subjects were partitioned into 20 distinct folds. This evaluation protocol is consistent with previous studies, ensuring a fair comparison with existing methods.

For the Sleep-EDF-78 dataset, we strictly follow the subject-wise cross-validation protocol established in previous benchmark studies, using the same fold partition to maintain consistency with reported state-of-the-art methods. All hyperparameters (batch size, learning rate, frequency/time parameters) remain identical to those used for Sleep-EDF-20 to ensure a fair cross-dataset comparison.

4.3 Evaluation

To comprehensively evaluate the classification performance of DGCA-Net, we adopted three standard metrics: Precision (PRE), Recall (REC), and F1-Score. These metrics are defined as follows:

$$\text{PRE} = \frac{TP}{TP + FP} \tag{25}$$

$$\text{REC} = \frac{TP}{TP + FN} \tag{26}$$

$$F1 = 2 \times \frac{\text{PRE} \times \text{REC}}{\text{PRE} + \text{REC}} \tag{27}$$

where TP, FP, and FN denote True Positives, False Positives, and False Negatives, respectively.

Given the inherent class imbalance in sleep stage classification, we further employed Cohen's Kappa coefficient (κ) and Macro F1-score (MF1) to provide a more robust assessment. Cohen's Kappa measures the agreement between the model predictions and ground-truth labels while accounting for chance agreement, and Macro F1-score computes the unweighted average of F1 scores across all classes, ensuring that minority stages are not overlooked. These metrics are formulated as:

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{28}$$

$$\text{MF1} = \frac{1}{N} \sum_{i=1}^{N} F1_i \tag{29}$$

where P_o denotes the observed agreement, P_e denotes the expected agreement by chance, and $F1_i$ is the F1-score for the i-th class.

Table I presents the confusion matrix of DGCA-Net on the Sleep-EDF-20 dataset. In this matrix, each row and column represent the number of 30-second epochs classified by the sleep expert (true labels) and by DGCA-Net (predicted labels), respectively. The diagonal entries indicate the number of correctly classified epochs for each sleep stage, while the off-diagonal entries represent misclassifications between stages. The last three columns report the per-class performance metrics, including Precision (PR), Recall (RE), and F1-score.

Table III: Performance Comparison with Existing Methods on Two Datasets. The best results are marked in bold

Dataset	Method	Overall Performance		
		Acc	MF1	κ
Sleep-EDF-20	AttnSleep [22]	84.4	78.1	0.79
	SleepEEGNet [23]	84.3	79.7	0.79
	Multi-task [24]	85.6	81.1	0.80
	STFT-CNN [25]	85.7	79.4	0.80
	MultiChannelSleepNet [26]	85.8	80.0	0.81
	CASleepNet [27]	86.5	80.2	0.82
	FlexibleSleepNet [28]	86.9	81.9	0.824
	DGCA-Net(ours)	**87.72**	**82.3**	**0.831**
Sleep-EDF-78	AttnSleep [22]	81.3	75.1	0.74
	SleepEEGNet [23]	80.0	73.6	0.73
	SeqSleepNet [29]	83.8	78.2	0.78
	IITNeT [30]	83.6	76.5	0.73
	SleePyCo [31]	84.6	**79.0**	0.787
	Multi-task [24]	83.4	78.6	0.77
	CASleepNet [27]	84.3	78.3	0.78
	DGCA-Net(ours)	**84.79**	78.67	**0.788**

As shown in Table I, DGCA-Net achieves accurate classification for most sleep stages, with the diagonal values consistently being the largest in their corresponding rows and columns. Notably, the N1 stage remains the most challenging to classify, as it is frequently misclassified as W, N2, or REM. This difficulty arises from the transitional nature of N1, which shares overlapping electrophysiological characteristics with both wakefulness and adjacent sleep stages, making it inherently ambiguous even for human experts.

Turning to the Sleep-EDF-78 dataset, Table II presents the corresponding confusion matrix. Similar to the findings on Sleep-EDF-20, DGCA-Net achieves dominant diagonal entries across most stages, with the N1 stage again exhibiting the lowest classification performance (recall of 46.5%), primarily due to its transitional electrophysiological properties. Notably, the classification of the N3 stage on this larger cohort achieves a precision of 93.4% and a recall of 66.8%, indicating that although some deep-sleep epochs are missed (moderate recall), the predictions made for N3 are highly reliable with very few false positives. This confirms robust deep-sleep discrimination even under greater inter-subject variability. The overall per-class F1 scores on Sleep-EDF-78 consistently surpass or match those of comparative methods, confirming the model's strong generalization capacity.

4.4 Performance Comparison

We compared the performance of DGCA-Net with existing state-of-the-art methods on the Sleep-EDF-20 dataset, with the results presented in Table III. As shown, DGCA-Net achieves superior performance across all evaluation metrics. Specifically,

Table IV: Influence of FPN, AMCAF, CA and IFF on Classification Results of DGCA-Net.

FPN	AMCAF	CA	IFF	Acc
✓	✗	✗	✗	86.31
✓	✗	✗	✓	86.90
✓	✗	✓	✓	87.36
✓	✓	✗	✓	87.56
✓	✓	✓	✓	87.72

our method attains an accuracy (Acc) of 87.72%, a macro F1-score (MF1) of 82.3%, and a Cohen's Kappa coefficient (κ) of 0.831, outperforming all compared methods by a significant margin.

Compared with conventional multi-modal fusion methods, DGCA-Net achieves more effective integration of complementary information across EEG and EOG modalities. Built upon a pyramid-based multi-scale feature extraction backbone, the framework consists of two core modules (AMCAF and Contrastive Alignment) supplemented by an auxiliary Inter-layer Feature Fusion (IFF) module. The feature extraction backbone ensures comprehensive information capture across different frequency bands. The AMCAF module performs dynamic channel gating prior to cross-attention computation, effectively suppressing modality-irrelevant noise and redundant channels while enhancing discriminative features, thereby providing purified feature representations for subsequent fusion. Additionally, the contrastive alignment module explicitly aligns cross-modal embeddings from the same sleep epoch via the InfoNCE loss, resolving feature space misalignment and improving the consistency of inputs to the multi-scale attention fusion process. The auxiliary IFF module further refines the multi-scale representations by capturing cross-level dependencies and suppressing inter-layer redundancy, strengthening the final feature representation. The synergistic combination of these components enables DGCA-Net to achieve superior classification accuracy and robust staging performance.

On the larger and more challenging Sleep-EDF-78 dataset, DGCA-Net achieves an accuracy of 84.79%, a macro F1-score of 78.67%, and a Cohen's Kappa of 0.788. These results validate the model's robust generalization capability across a more heterogeneous cohort with greater inter-subject variability, confirming that the proposed dynamic gating and contrastive alignment mechanisms effectively mitigate cross-modal discrepancies under diverse real-world recording conditions.

4.5 Ablation Study

To investigate the individual contributions of each core component in DGCA-Net, we conducted ablation experiments on the Sleep-EDF-20 dataset. The results are presented in Table IV, where "✓" indicates the inclusion of the corresponding module

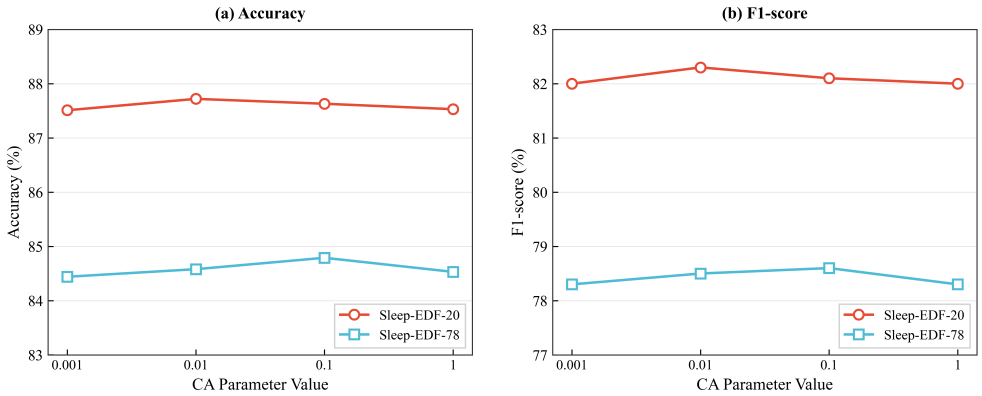

Fig. 5: Accuracy and F1-score of DGCA-Net with temperature τ on Sleep-EDF-20 and Sleep-EDF-78.

and "✗" indicates its exclusion. The "Acc" column reports the classification accuracy for each configuration.

As shown in Table IV, the baseline model employing only the pyramid-based multi-scale feature extraction module achieves an accuracy of 86.31%. When the AMCAF module is incorporated, the accuracy rises to 87.56%, indicating that the dynamic channel gating mechanism helps to suppress modality-irrelevant noise and redundant information while emphasizing discriminative features, thereby providing cleaner representations for subsequent fusion. Adding the contrastive alignment module further improves accuracy to 87.72%, suggesting that the InfoNCE-driven alignment mitigates cross-modal feature space misalignment and facilitates more consistent multi-modal integration. Furthermore, the inclusion of the IFF module yields additional gains, demonstrating its capacity to capture cross-level dependencies and reduce inter-layer redundancy, which together lead to the highest accuracy observed across all configurations. These ablation results collectively verify the positive contribution of each proposed module and show that their combination achieves the best overall staging performance on this dataset.

4.6 Analysis of Temperature Parameter in Contrastive Alignment

The temperature parameter τ in the InfoNCE loss governs the sharpness of the similarity distribution, thereby determining the stringency of cross-modal feature alignment. To investigate its impact, we evaluated DGCA-Net on both Sleep-EDF-20 and Sleep-EDF-78 datasets with τ set to 1, 0.1, 0.01, and 0.001, while keeping all other hyperparameters fixed.

As illustrated in Fig. 5, the optimal τ values differ markedly between the two datasets. On Sleep-EDF-20, the best performance is achieved at τ=0.01 (Accuracy: 87.72%, MF1: 82.30%), whereas on the larger and more heterogeneous Sleep-EDF-78 cohort, the optimal value shifts to τ=0.1 (Accuracy: 84.79%, MF1: 78.60%). This

discrepancy is intrinsically linked to the distinct characteristics of the two datasets and the behavior of contrastive alignment.

For Sleep-EDF-20, which comprises a relatively small and homogeneous subject pool with limited inter-subject physiological variability, a smaller τ (0.01) produces a sharp similarity distribution that imposes a stringent alignment constraint. This forces the EEG and EOG embeddings from the same epoch to converge tightly in the latent space, effectively eliminating the relatively minor cross-modal distribution gaps and yielding highly discriminative fused features. However, when τ is further reduced to 0.001, the excessively sharp distribution makes the loss overly sensitive to hard negatives, introducing gradient instability and degrading performance, as reflected in the evident performance drop.

In contrast, Sleep-EDF-78 encompasses 78 subjects with a much wider age range (25–101 years), entailing substantially greater inter-subject variability in EEG/EOG signal morphology, electrode placement offsets, and individual anatomical differences. Under such heterogeneous conditions, an overly small τ (0.01 or 0.001) forces the model to overfit to specific intra-modal patterns and treat natural cross-subject variations as hard negatives, leading to unstable contrastive optimization and suboptimal alignment. A moderately larger temperature (τ=0.1) softens the similarity distribution, allowing the alignment loss to focus on coarse-grained cross-modal consistency while tolerating reasonable intra-class dispersion caused by physiological diversity. This relaxation prevents the model from distorting the feature space to accommodate idiosyncratic noise, thereby improving robustness and generalization on the more challenging 78-subject cohort.

Taken together, these results reveal a clear dataset-dependent optimal temperature: the more homogeneous Sleep-EDF-20 benefits from tight alignment with τ=0.01, while the heterogeneous Sleep-EDF-78 requires a looser constraint at τ=0.1 to balance alignment strength against inter-subject variability. This finding underscores that temperature selection should be guided by the inherent diversity of the target population, and we adopt τ=0.01 for Sleep-EDF-20 and τ=0.1 for Sleep-EDF-78 in all subsequent experiments accordingly.

5 CONCLUSION

In this paper, we proposed DGCA-Net, a novel deep learning framework for automatic sleep staging. It comprises two core modules, AMCAF and CA, supplemented by an auxiliary Inter-layer Feature Fusion (IFF) module. The AMCAF module performs dynamic channel gating prior to cross-modal interaction, effectively suppressing modality-irrelevant noise and redundant channels while enhancing discriminative features strongly correlated with specific sleep stages. Driven by the InfoNCE loss, the Contrastive Alignment module explicitly aligns EEG and EOG embeddings from the same sleep epoch in the latent space, resolving cross-modal feature space misalignment and ensuring more consistent and complementary inputs for multi-scale attention fusion. The auxiliary IFF module captures cross-level dependencies through dual pooling operations and attention-guided aggregation, refining the final feature representation for accurate classification.

Evaluated on the publicly available Sleep-EDF-20 and Sleep-EDF-78 datasets, our method achieves superior performance compared with existing state-of-the-art approaches. On Sleep-EDF-20, we attain an accuracy of 87.72%, a macro F1-score of 82.3%, and a Cohen's Kappa of 0.831. Ablation studies further validate the individual contributions of each module, demonstrating that their synergistic combination leads to the best observed performance. In future work, we plan to extend our framework to additional physiological modalities and explore advanced contrastive learning strategies to further strengthen cross-modal feature alignment and generalization across diverse clinical populations.

Corresponding author. Correspondence concerning this article should be addressed to Jinlong Yang, School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi 214122, China. E-mail: jlyang@jiangnan.edu.cn.

Author contributions. X.L. conceived the study, designed the DGCA-Net framework, implemented the AMCAF and contrastive alignment modules, conducted all experiments and data analysis, and wrote the main manuscript text. J.Z. curated the clinical sleep data, provided medical interpretation of the sleep staging criteria, and validated the experimental results. J.Y. supervised the study, acquired the funding, and critically revised the manuscript. All authors reviewed and approved the final version of the manuscript.

Funding. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability. The datasets analysed during the current study are publicly available in the Sleep-EDF database on PhysioNet (Sleep-EDF-20 and Sleep-EDF-78), available at https://physionet.org/content/sleep-edfx/1.0.0/. No new data were generated in this study. The code implementing DGCA-Net is available from the corresponding author on reasonable request.

Competing interests. The authors declare that they have no competing interests.

References

[1] Rauchs, G., Desgranges, B., Foret, J., Eustache, F.: The relationships between memory systems and sleep stages. Journal of Sleep Research **14**(2), 123–140 (2005) https://doi.org/10.1111/j.1365-2869.2005.00450.x

[2] Chokroverty, S.: Overview of sleep & sleep disorders. Indian Journal of Medical Research **131**(2), 126–140 (2010)

[3] Keenan, S.A.: Chapter 3 an overview of polysomnography. Handbook of Clinical Neurophysiology, vol. 6, pp. 33–50. Elsevier (2005). https://doi.org/10.1016/S1567-4231(09)70028-0

[4] Memar, P., Faradji, F.: A novel multi-class eeg-based sleep stage classification system. IEEE Transactions on Neural Systems and Rehabilitation Engineering **26**(1), 84–95 (2018) https://doi.org/10.1109/TNSRE.2017.2776149

[5] Liu, N., Lu, Z., Xu, B., Liao, Q.: Learning a convolutional neural network for sleep stage classification. In: 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), pp. 1–6 (2017). https://doi.org/10.1109/CISP-BMEI.2017.8302226

[6] Zhu, G., Li, Y., Wen, P.: Analysis and classification of sleep stages based on difference visibility graphs from a single-channel eeg signal. IEEE Journal of Biomedical and Health Informatics **18**(6), 1813–1821 (2014) https://doi.org/10.1109/JBHI.2014.2303991

[7] Seifpour, S., Niknazar, H., Mikaeili, M., Nasrabadi, A.M.: A new automatic sleep staging system based on statistical behavior of local extrema using single channel eeg signal. Expert Systems with Applications **104**, 277–293 (2018) https://doi.org/10.1016/j.eswa.2018.03.020

[8] Li, X., Cui, L., Tao, S., Chen, J., Zhang, X., Zhang, G.-Q.: Hyclasss: A hybrid classifier for automatic sleep stage scoring. IEEE Journal of Biomedical and Health Informatics **22**(2), 375–385 (2018) https://doi.org/10.1109/JBHI.2017.2668993

[9] 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 **25**(11), 1998–2008 (2017) https://doi.org/10.1109/TNSRE.2017.2721116

[10] Fan, J., Zhao, M., Huang, L., Tang, B., Wang, L., He, Z., Peng, X.: Multimodal sleep staging network based on obstructive sleep apnea. Frontiers in Computational Neuroscience **Volume 18 - 2024** (2024) https://doi.org/10.3389/fncom.2024.1505746

[11] Zhang, W., Li, M., Chen, Y., Zhou, J.: Asynchronous multi-channel physiological signal fusion for sleep staging under noisy environments. IEEE Transactions on Biomedical Engineering **70**(8), 2345–2356 (2023) https://doi.org/10.1109/TBME.2023.3245678

[12] Kim, J., Park, S., Lee, S.: Redundancy reduction and interference suppression for multi-channel eeg-based sleep staging. In: 2022 44th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 1289–1292 (2022). https://doi.org/10.1109/EMBC48229.2022.9871234

[13] Garcia, R., Martinez, A., Sanchez, J.: Robust sleep staging with asynchronous multi-modal physiological signals: A cross-modal correlation learning approach. Computers in Biology and Medicine **140**, 105123 (2021) https://doi.org/10.1016/j.compbiomed.2021.105123

[14] Chen, Y., Li, H., Wang, Z.: A unified framework for multi-channel sleep stage classification with robust feature fusion and channel conflict resolution. IEEE Transactions on Neural Systems and Rehabilitation Engineering **31**, 245–256

(2023) https://doi.org/10.1109/TNSRE.2022.3221115

[15] Jia, Z., Cai, X., Jiao, Z.: Multi-modal physiological signals based squeeze-and-excitation network with domain adversarial learning for sleep staging. IEEE Sensors Journal **22**(4), 3464–3471 (2022) https://doi.org/10.1109/JSEN.2022.3140383

[16] Zhang, X., Wu, Y., Liu, X.: Efficient handling of data loss in multi-channel sleep signal analysis for stage classification. In: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 177–182 (2022). https://doi.org/10.1109/BIBM55127.2022.9919931

[17] Wang, L., Zhao, J., Yang, H.: Scalable multi-channel sleep staging: Overcoming computational challenges for large-scale data. In: 2021 International Conference on Pattern Recognition and Machine Intelligence (PRMI), pp. 345–350 (2021). https://doi.org/10.1109/PRMI52549.2021.9615632

[18] Phan, H., Mikkelsen, K., Chen, O.Y., Koch, P., Mertins, A., Vos, M.D.: Sleeptransformer: Automatic sleep staging with interpretability and uncertainty quantification. IEEE Transactions on Biomedical Engineering **69**(8), 2456–2467 (2022)

[19] Yu, Y., Chen, S., Wei, W., Li, J., Pan, J.: Multiconssleepnet: Self-supervised contrastive learning for a multimodal consistency-based automatic sleep staging model. In: 2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 1–7 (2025). https://doi.org/10.1109/EMBC58623.2025.11251569

[20] Shen, Q., Xin, J., Dai, B.T., Zhang, S., Wang, Z.: Robust sleep staging over incomplete multimodal physiological signals via contrastive imagination. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) Advances in Neural Information Processing Systems, vol. 37, pp. 112025–112049. Curran Associates, Inc., ??? (2024)

[21] Chen, D., Bai, Y., Zhao, W., Ament, S., Gregoire, J., Gomes, C.: Deep reasoning networks for unsupervised pattern de-mixing with constraint reasoning. In: III, H.D., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 1500–1509. PMLR, ??? (2020)

[22] 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 **29**, 809–818 (2021) https://doi.org/10.1109/TNSRE.2021.3076234

[23] Mousavi, S., Afghah, F., Acharya, U.R.: Sleepeegnet: Automated sleep stage scoring with sequence to sequence deep learning approach. PLOS ONE **14**(5), 0216456

(2019) https://doi.org/10.1371/journal.pone.0216456

[24] Zhao, C., Li, J., Guo, Y.: Sequence signal reconstruction based multi-task deep learning for sleep staging on single-channel eeg. Biomedical Signal Processing and Control **88**, 105615 (2024) https://doi.org/10.1016/j.bspc.2024.105615

[25] Jadhav, P., Mukhopadhyay, S.: Automated sleep stage scoring using time-frequency spectra convolution neural network. IEEE Transactions on Instrumentation and Measurement **71**, 1–9 (2022)

[26] Dai, Y., *et al.*: Multichannelsleepnet: A transformer-based model for automatic sleep stage classification with psg. IEEE Journal of Biomedical and Health Informatics **27**(9), 4204–4215 (2023)

[27] Yu, W., Yang, J.: Casleepnet: A cross attention-based multimodal fusion approach for sleep staging with eeg and eog. In: ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5 (2025). https://doi.org/10.1109/ICASSP49660.2025.10889570

[28] Ren, Z., Ma, J., Ding, Y.: Flexiblesleepnet:a model for automatic sleep stage classification based on multi-channel polysomnography. IEEE Journal of Biomedical and Health Informatics **29**(5), 3488–3501 (2025) https://doi.org/10.1109/JBHI.2025.3525626

[29] 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 **27**(3), 400–410 (2019) https://doi.org/10.1109/TNSRE.2019.2896659

[30] Seo, H., Back, S., Lee, S., Park, D., Kim, T., Lee, K.: Intra- and inter-epoch temporal context network (iitnet) using sub-epoch features for automatic sleep scoring on raw single-channel eeg. Biomedical Signal Processing and Control **61**, 102037 (2020) https://doi.org/10.1016/j.bspc.2020.102037

[31] Lee, S., Yu, Y., Back, S., Seo, H., Lee, K.: Sleepyco: Automatic sleep scoring with feature pyramid and contrastive learning. Expert Systems with Applications **240**, 122551 (2024) https://doi.org/10.1016/j.eswa.2023.122551